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

Library Declaration Form University of Otago Library

</p><table

cellpadding="0" border="0" cellspacing="0" class="tabbing">

Author’s full name and year of birth: Simon Thomas Kelly,

</p><table

cellpadding="0" border="0" cellspacing="0" class="tabbing">

   (for cataloguing purposes)

24 February 1992

</p><table

cellpadding="0" border="0" cellspacing="0" class="tabbing">

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 Gαi 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 χ2-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.

Research Contributions During Candidature

Publications Under Peer-Review

Kelly, S. T. and Spencer, H. G. (2017) Population-Genetics Models of Sex-Limited Genomic Imprinting. Theoretical Population Biology 115:35-44 doi:10.1016/j.tpb.2017.03.004

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

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. Submitted to BMC Genomics.

Community Blog Posts

Black, M. A., Kelly, S. T., and Cadzow, M. Posted on the Software Carpentry website 2016 July 4th: “Software Carpentry workshop at the University of Otago, New Zealand” https://software-\_carpentry.org/blog/2016/07/otago-\_workshop.html

Kelly, S. T., Black, M., A., Bae, S., Hayek, W., and Pawlik, A. Posted on the Software Carpentry website 2016 September 28th: “Software Carpentry Workshop Attendance: a New Zealand Perspective“ https://software-\_carpentry.org/blog/2016/09/attendance-\_nz.html

Software Packages

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.

Conference Participation

eResearch 2017 (Queenstown) Speaker February 20th-22nd “Detecting Synthetic Lethality from Cancer Gene Expression: A PhD project on genetic interactions with CDH1 inactivation in TCGA data”

RIKEN Division of Genomic Technologies (Yokohama, Japan) Seminar 2016 October 20th; National Institute of Genetics (Mishima, Japan) Seminar 2016 October 21st; Tokyo University Institute of Medical Science (Shirokanedai

Campus, Japan) Seminar 2016 October 24th; Sokendai Graduate University (Hayama, Japan) Seminar 2016 October 25th “Analysis of Synthetic Lethal Pathways in Breast Cancer: A PhD project on genetic interactions with CDH1 inactivation in TCGA data”

Next Generation Sequencing Asia 2016 (Singapore) Poster October 11th-12th “Bioinformatic Investigations of Synthetic Lethal Interactions with E-cadherin in Breast Cancer” (Supported by the University of Otago Division of Health Sciences; Maurice and Phyllis Paykel Trust)

eResearch 2016 (Queenstown, New Zealand) Speaker February 9th-11th “Sifting the Needles in the Haystack: Permutation Resampling Biological Pathways in Cancer Genomic Interaction Data” (Supported by REANNZ)

Genetics Otago Symposium 2016 (Dunedin, New Zealand) Student Speaker March 7th-8th “A Bioinformatics approach to Genetic Interactions: Synthetic Lethal Pathways with E-cadherin in Breast Cancer Genomics Data”

Research Bazaar 2015 (Melbourne, Australia) February 16th-18th “My digital research toolkit” (Supported by the New Zealand eScience Infrastructure)

QMB Cancer Drugs Satellite 2014 (Queenstown, New Zealand) Poster August 24th -25th; Otago School of Medical Sciences 2015 (Dunedin, New Zealand) Poster Postgraduate Symposium April 28th-29th Bioinformatics Prioritisation of Synthetic Lethal Targets for Drug Activity Against E-Cadherin Deficient Cancers

DunDead: Zombie Science and Popular Culture Festival 2014 (Dunedin, New Zealand) Ignite Speaker August 16th-17th “Hidden in Plain Sight - The Genetics of Zombies”

eResearch 2014 (Waikato University, Hamilton, New Zealand) Ignite Speaker June 30th-July 2nd “Bioinformatic analysis of synthetic lethal genetic interactions in breast cancer” (Supported by Google)

Acknowledgements

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

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

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

I cannot thank my friends, flatmates and family enough for their patience and support during such as massive, challenging, and (I’m sure you’ve heard too many times) stressful undertaking during both my PhD and the study leading up to it. There are too many of you to name everyone here without leaving someone out, so thank you all for everything you’ve done, both the good times and the tough. Thank you for pretending to understand when I

try to discuss complex math at the wrong moment. Thank you for checking my writing or slides, even if I should have given you more time. Thank for your time when all I really needed was a chat over a walk or a pint and a moment to think clearly.

I must also thank various organisations supported this research project:

This thesis was supported by the Postgraduate Tassell Scholarship in Cancer Research, a University of Otago Doctoral Scholarship.

The New Zealand eScience Infrastucture (NeSI) provided access to the Intel Pan high-performance computing cluster, support, and training to use it effectively. Various aspects of this thesis would not have been possible without access to such a resource.

The Health Research Council (HRC) of New Zealand provided funding for experimental research in the Cancer Genetics Laboratory. Again some aspects of this project would not have been possible without access to the data and findings funded by this grant.

The Allan Wilson Centre and Otago School of Biomedical Sciences provided funding for summer research placements which was a valuable opportunity to gain experience and training used in this thesis project.

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

Google (towards eResearch 2014 conference, Hamilton)

NeSI (towards Software Carpentry training and Research Bazaar 2015, Melbourne)

REANNZ, NZGL, and NeSI (towards eResearch 2016 conference, Queenstown)

Otago Division of Health Sciences, Department of Biochemistry, Oxford Global, and Maurice and Phyllis Paykel Trust (towards NGS Asia 2016, Singapore)

RIKEN Division of Genomics Technologies and the Okinawa Institue of Science and Technology (for hosting seminars in Japan)

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

Contents

Glossary Acronyms 1 Introduction  1.1 Cancer Research in the Post-Genomic Era   1.1.1 Cancer as a Global Health Concern    1.1.1.1 The Genetics and Molecular Biology of Cancers   1.1.2 The Human Genome Revolution    1.1.2.1 The First Human Genome Sequence    1.1.2.2 Impact of Genomics   1.1.3 Technologies to Enable Genetics Research    1.1.3.1 DNA Sequencing and Genotyping Technologies    1.1.3.2 Microarrays and Quantitative Technologies    1.1.3.3 Massively Parallel “Next Generation” Sequencing    1.1.3.4 Bioinformatics as Interdisciplinary Genomic Analysis   1.1.4 Follow-up Large-Scale Genomics Projects   1.1.5 Cancer Genomes    1.1.5.1 The Cancer Genome Atlas Project   1.1.6 Genomic Cancer Medicine    1.1.6.1 Cancer Genes and Driver Mutations    1.1.6.2 Personalised or Precision Cancer Medicine    1.1.6.3 Targeted Therapeutics and Pharmacogenomics    1.1.6.4 Systems and Network Biology  1.2 A Synthetic Lethal Approach to Cancer Medicine   1.2.1 Synthetic Lethal Genetic Interactions   1.2.2 Synthetic Lethal Concepts in Genetics   1.2.3 Studies of Synthetic Lethality    1.2.3.1 Synthetic Lethal Pathways and Networks   1.2.4 Synthetic Lethal Concepts in Cancer   1.2.5 Clinical Impact of Synthetic Lethality in Cancer   1.2.6 High-throughput Screening for Synthetic Lethality    1.2.6.1 Synthetic Lethal Screens   1.2.7 Computational Prediction of Synthetic Lethality    1.2.7.1 Bioinformatics Approaches to Genetic Interactions    1.2.7.2 Comparative Genomics    1.2.7.3 Analysis and Modelling of Protein Data    1.2.7.4 Differential Gene Expression    1.2.7.5 Data Mining and Machine Learning

   1.2.7.6 Bimodality    1.2.7.7 Rationale for Further Development  1.3 E-cadherin as a Synthetic Lethal Target   1.3.1 The CDH1 gene and it’s Biological Functions    1.3.1.1 Cytoskeleton    1.3.1.2 Extracellular and Tumour Micro-Environment    1.3.1.3 Cell-Cell Adhesion and Signalling   1.3.2 CDH1 as a Tumour (and Invasion) Suppressor    1.3.2.1 Breast Cancers and Invasion   1.3.3 Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer   1.3.4 Somatic Mutations    1.3.4.1 Mutation Rate    1.3.4.2 Co-occurring Mutations   1.3.5 Models of CDH1 loss in cell lines  1.4 Summary and Research Direction of Thesis 2 Methods and Resources  2.1 Bioinformatics Resources for Genomics Research   2.1.1 Public Data and Software Packages    2.1.1.1 Cancer Genome Atlas Data    2.1.1.2 Reactome and Annotation Data  2.2 Data Handling   2.2.1 Normalisation   2.2.2 Sample Triage   2.2.3 Metagenes and the Singular Value Decomposition    2.2.3.1 Candidate Triage and Integration with Screen Data  2.3 Techniques   2.3.1 Statistical Procedures and Tests   2.3.2 Gene Set Over-representation Analysis   2.3.3 Clustering   2.3.4 Heatmap   2.3.5 Modeling and Simulations    2.3.5.1 Receiver Operating Characteristic (Performance)   2.3.6 Resampling Analysis  2.4 Pathway Structure Methods   2.4.1 Network and Graph Analysis   2.4.2 Sourcing Graph Structure Data   2.4.3 Constructing Pathway Subgraphs   2.4.4 Network Analysis Metrics  2.5 Implementation   2.5.1 Computational Resources and Linux Utilities   2.5.2 R Language and Packages   2.5.3 High Performance and Parallel Computing 3 Methods Developed During Thesis  3.1 A Synthetic Lethal Detection Methodology  3.2 Synthetic Lethal Simulation and Modelling   3.2.1 A Model of Synthetic Lethality in Expression Data

  3.2.2 Simulation Procedure  3.3 Detecting Simulated Synthetic Lethal Partners   3.3.1 Binomial Simulation of Synthetic lethality   3.3.2 Multivariate Normal Simulation of Synthetic lethality    3.3.2.1 Multivariate Normal Simulation with Correlated Genes    3.3.2.2 Specificity with Query-Correlated Pathways  3.4 Graph Structure Methods   3.4.1 Upstream and Downstream Gene Detection    3.4.1.1 Permutation Analysis for Statistical Significance    3.4.1.2 Hierarchy Based on Biological Context   3.4.2 Simulating Gene Expression from Graph Structures  3.5 Customised Functions and Packages Developed   3.5.1 Synthetic Lethal Interaction Prediction Tool   3.5.2 Data Visualisation   3.5.3 Extensions to the iGraph Package    3.5.3.1 Sampling Simulated Data from Graph Structures    3.5.3.2 Plotting Directed Graph Structures    3.5.3.3 Computing Information Centrality    3.5.3.4 Testing Pathway Structure with Permutation Testing    3.5.3.5 Metapackage to Install iGraph Functions 4 Synthetic Lethal Analysis of Gene Expression Data  4.1 Synthetic lethal genes in breast cancer   4.1.1 Synthetic lethal pathways in breast cancer   4.1.2 Expression profiles of synthetic lethal partners    4.1.2.1 Subgroup pathway analysis  4.2 Comparison of synthetic lethal gene candidates   4.2.1 Comparison with siRNA screen candidates    4.2.1.1 Comparison with correlation    4.2.1.2 Comparison with viability    4.2.1.3 Comparison with secondary siRNA screen candidates    4.2.1.4 Comparison of screen at pathway level  4.3 Metagene Analysis   4.3.1 Pathway expression   4.3.2 Somatic mutation   4.3.3 Mutation locus   4.3.4 Synthetic lethal metagenes  4.4 Replication in stomach cancer   4.4.1 Synthetic Lethal Genes and Pathways   4.4.2 Synthetic Lethal Expression Profiles   4.4.3 Comparison to Primary Screen    4.4.3.1 Resampling Analysis   4.4.4 Metagene Analysis  4.5 Global Synthetic Lethality   4.5.1 Hub Genes   4.5.2 Hub Pathways  4.6 Replication in cell line encyclopaedia

 4.7 Discussion   4.7.1 Strengths of the SLIPT Methodology   4.7.2 Syntheic Lethal Pathways for E-cadherin   4.7.3 Replication and Validation    4.7.3.1 Integration with siRNA Screening    4.7.3.2 Replication across Tissues and Cell lines  4.8 Summary    Aims    Summary 5 Synthetic Lethal Pathway Structure  5.1 Synthetic Lethal Genes in Reactome Pathways   5.1.1 The PI3K/AKT Pathway   5.1.2 The Extracellular Matrix   5.1.3 G Protein Coupled Receptors   5.1.4 Gene Regulation and Translation  5.2 Network Analysis of Synthetic Lethal Genes   5.2.1 Gene Connectivity and Vertex Degree   5.2.2 Gene Importance and Centrality    5.2.2.1 Information Centrality    5.2.2.2 PageRank Centrality  5.3 Testing Pathway Structure of Synthetic Lethal Genes   5.3.1 Hierarchical Pathway Structure    5.3.1.1 Contextual Hierarchy of PI3K    5.3.1.2 Testing Contextual Hierarchy of Synthetic Lethal Genes   5.3.2 Upstream or Downstream Synthetic Lethality    5.3.2.1 Measuring Structure of Candidates within PI3K    5.3.2.2 Resampling for Synthetic Lethal Pathway Structure  5.4 Discussion  5.5 Summary    Aims    Summary 6 Simulation and Modeling of Synthetic Lethal Pathways  6.1 Comparing methods   6.1.1 Performance of SLIPT and χ2 across Quantiles    6.1.1.1 Correlated Query Genes affects Specificity   6.1.2 Correlation as a Synthetic Lethal Detection Strategy   6.1.3 Testing for Bimodality with BiSEp  6.2 Simulations with Graph Structures   6.2.1 Performance over a Graph Structure   6.2.2 Synthetic Lethality across Graph Stuctures   6.2.3 Performance with inhibition links   6.2.4 Performance with 20,000 genes  6.3 Simulations over pathway-based graphs  6.4 Discussion  6.5 Summary    Aims

   Summary 7 Discussion    Aims    Summary  7.1 Significance  7.2 Future Directions  7.3 Conclusion 8 Conclusion References A Sample Quality  A.1 Sample Correlation  A.2 Replicate Samples in TCGA Breast B Software Used for Thesis C Secondary Screen Data D Mutation Analysis in Breast Cancer  D.1 Synthetic Lethal Genes and Pathways  D.2 Synthetic Lethal Expression Profiles  D.3 Comparison to Primary Screen   D.3.1 Resampling Analysis  D.4 Compare SLIPT genes  D.5 Metagene Analysis  D.6 Mutation Variation   D.6.1 Mutation Frequency   D.6.2 PI3K Mutation Expression E Metagene Expression Profiles F Stomach Expression Analysis  F.1 Synthetic Lethal Genes and Pathways  F.2 Comparison to Primary Screen   F.2.1 Resampling Analysis  F.3 Metagene Analysis G Stomach Mutation Analysis  G.1 Synthetic Lethal Genes and Pathways  G.2 Synthetic Lethal Expression Profiles  G.3 Comparison to Primary Screen   G.3.1 Resampling Analysis  G.4 Metagene Analysis H Global Synthetic Lethality in Stomach Cancer  H.1 Hub Genes  H.2 Hub Pathways I Replication in cell line encyclopaedia J Synthetic Lethal Genes in Pathways K Pathway Connectivity for Mutation SLIPT L Information Centrality for Gene Essentiality M Pathway Structure for Mutation SLIPT N Simulation

List of Tables

1.1 Methods for Predicting Genetic Interactions1.2 Methods for Predicting Synthetic Lethality in Cancer1.3 Methods used by Wu et al. (2014)2.1 Excluded Samples by Batch and Clinical Characteristics.2.2 Computers used during Thesis2.3 Linux Utilities and Applications used during Thesis2.4 R Installations used during Thesis 2.5 R Packages used during Thesis2.6 R Packages Developed during Thesis4.1 Candidate synthetic lethal gene partners of CDH1 from SLIPT4.2 Pathways for CDH1 partners from SLIPT4.3 Pathway composition for clusters of CDH1 partners from SLIPT4.4 Pathway composition for CDH1 partners from SLIPT and siRNA screening4.5 Pathways for CDH1 partners from SLIPT4.6 Pathways for CDH1 partners from SLIPT and siRNA primary screen4.7 Candidate synthetic lethal metagenes against CDH1 from SLIPT4.8 Pathways for CDH1 partners from SLIPT in stomach cancer4.9 Query synthetic lethal genes with the most SLIPT partners4.10 Pathways for genes with the most SLIPT partners4.11 Pathways for CDH1 partners from SLIPT in CCLE4.12 Pathways for CDH1 partners from SLIPT in breast CCLE5.1 analysis of variance (ANOVA) for Synthetic Lethality and Vertex Degree5.2 ANOVA for Synthetic Lethality and Information Centrality5.3 ANOVA for Synthetic Lethality and PageRank Centrality5.4 ANOVA for Synthetic Lethality and PI3K Hierarchy5.5 Resampling for pathway structure of synthetic lethal detection methods B.1 R Packages used during ThesisC.1 Comparing SLIPT genes against Secondary siRNA Screen in breast cancerC.2 Comparing mtSLIPT genes against Secondary siRNA Screen in breast cancerC.3 Comparing SLIPT genes against Secondary siRNA Screen in stomach cancerD.1 Candidate synthetic lethal gene partners of CDH1 from mtSLIPTD.2 Pathways for CDH1 partners from mtSLIPTD.3 Pathway composition for clusters of CDH1 partners from mtSLIPTD.4 Pathway composition for CDH1 partners from mtSLIPT and siRNAD.5 Pathways for CDH1 partners from mtSLIPTD.6 Pathways for CDH1 partners from mtSLIPT and siRNA primary screenD.7 Candidate synthetic lethal metagenes against CDH1 from mtSLIPTF.1 Synthetic lethal gene partners of CDH1 from SLIPT in stomach cancerF.2 Pathway composition for clusters of CDH1 partners in stomach SLIPTF.3 Pathway composition for CDH1 partners from SLIPT and siRNA screeningF.4 Pathways for CDH1 partners from SLIPT in stomach cancerF.5 Pathways for CDH1 partners from SLIPT in stomach

and siRNA screenF.6 Candidate synthetic lethal metagenes against CDH1 from SLIPT in stomach cancerG.1 Synthetic lethal gene partners of CDH1 from mtSLIPT in stomach cancerG.2 Pathways for CDH1 partners from mtSLIPT in stomach cancerG.3 Pathway composition for clusters of CDH1 partners in stomach mtSLIPTG.4 Pathway composition for CDH1 partners from mtSLIPT and siRNAG.5 Pathways for CDH1 partners from mtSLIPT in stomach cancerG.6 Pathways for CDH1 partners from mtSLIPT in stomach and siRNA screenG.7 Candidate synthetic lethal metagenes against CDH1 from mtSLIPT in stomach cancerH.1 Query synthetic lethal genes with the most SLIPT partnersH.2 Pathways for genes with the most SLIPT partnersI.1 Candidate synthetic lethal gene partners of CDH1 from SLIPT in CCLEI.2 Candidate synthetic lethal gene partners of CDH1 from SLIPT in breast CCLEI.3 Candidate synthetic lethal gene partners of CDH1 from SLIPT in stomach CCLEI.4 Pathways for CDH1 partners from SLIPT in stomach CCLEI.5 Pathways for CDH1 partners from SLIPT in breast and stomach CCLEK.1 ANOVA for Synthetic Lethality and Vertex DegreeK.2 ANOVA for Synthetic Lethality and Information CentralityK.3 ANOVA for Synthetic Lethality and PageRank CentralityL.1 Information centrality for genes and molecules in the Reactome networkM.1 ANOVA for Synthetic Lethality and PI3K HierarchyM.2 Resampling for pathway structure of synthetic lethal detection methods

List of Figures

1.1 Synthetic genetic interactions1.2 Synthetic lethality in cancer2.1 Read count density2.2 Read count sample mean3.1 Framework for synthetic lethal prediction3.2 Synthetic lethal prediction adapted for mutation3.3 A model of synthetic lethal gene expression3.4 Modeling synthetic lethal gene expression3.5 Synthetic lethality with multiple genes3.6 Simulating gene function3.7 Simulating synthetic lethal gene function3.8 Simulating synthetic lethal gene expression3.9 Performance of binomial simulations3.10 Comparison of statistical performance3.11 Performance of multivariate normal simulations3.12 Simulating expression with correlated gene blocks3.13 Simulating expression with correlated gene blocks3.14 Synthetic lethal prediction across simulations3.15 Performance with correlations3.16 Comparison of statistical performance with correlation structure3.17 Performance with query correlations3.18 Statistical evaluation of directional criteria3.19 Performance of directional criteria3.20 Simulated graph structures3.21 Simulating expression from a graph structure3.22 Simulating expression from graph structure with inhibitions3.23 Demonstration of violin plots with custom features3.24 Demonstration of annotated heatmap3.25 Simulating graph structures4.1 Synthetic lethal expression profiles of analysed samples4.2 Comparison of SLIPT to siRNA4.3 Compare SLIPT and siRNA genes with correlation4.4 Compare SLIPT and siRNA genes with correlation4.5 Compare SLIPT and siRNA genes with siRNA viability4.6 Compare SLIPT and siRNA genes with viability4.7 Compare SLIPT and siRNA genes with siRNA viability4.8 Resampled intersection of SLIPT and siRNA candidates4.9 Pathway metagene expression profiles4.10 Somatic mutation against PI3K metagene4.11 Somatic mutation locus against expression4.12 Synthetic lethal expression profiles of stomach samples4.13 Synthetic lethal partners across query genes5.1 Synthetic Lethality in the PI3K Cascade5.2 Synthetic Lethality in the Elastic Fibre Formation Pathway5.3 Synthetic Lethality in the Fibrin Clot Formation5.4 Synthetic Lethality and Vertex Degree5.5 Synthetic Lethality and Centrality5.6 Synthetic Lethality and PageRank5.7 Structure of PI3K Ranking5.8 Synthetic Lethality and Hierarchy Score in PI3K5.9 Hierarchy Score in PI3K against Synthetic Lethality in PI3K5.10 Structure of Synthetic Lethality in PI3K5.11 Structure of Synthetic Lethality Resampling in PI3K6.1 Simulated graph structures6.2 Simulated graph structures6.3 Simulated graph structures6.4 Simulated

graph structures6.5 Simulated graph structures6.6 Simulated graph structures6.7 Simulated graph structures6.8 Simulated graph structures6.9 Simulated graph structures6.10 Simulated graph structuresA.1 Correlation profiles of removed samplesA.2 Correlation analysis and sample removalA.3 Replicate excluded samplesA.4 Replicate samples with all remainingA.5 Replicate samples with some excludedA.6 Replicate samples with some excludedD.1 Synthetic lethal expression profiles of analysed samplesD.2 Comparison of mtSLIPT to siRNAD.3 Compare mtSLIPT and siRNA genes with correlationD.4 Compare mtSLIPT and siRNA genes with correlationD.5 Compare mtSLIPT and siRNA genes with siRNA viabilityD.6 Somatic mutation locusD.7 Somatic mutation against PIK3CA metageneD.8 Somatic mutation against PI3K proteinD.9 Somatic mutation against AKT proteinE.1 Pathway metagene expression profilesE.2 Expression profiles for constituent genes of PI3KE.3 Expression profiles for p53 related genesE.4 Expression profiles for estrogen receptor related genesE.5 Expression profiles for BRCA related genesF.1 Comparison of SLIPT in stomach to siRNAG.1 Synthetic lethal expression profiles of stomach samplesG.2 Comparison of mtSLIPT in stomach to siRNAH.1 Synthetic lethal partners across query genesJ.1 Synthetic Lethality in the PI3K/AKT PathwayJ.2 Synthetic Lethality in the PI3K/AKT Pathway in CancerJ.3 Synthetic Lethality in the Extracellular MatrixJ.4 Synthetic Lethality in the GPCRsJ.5 Synthetic Lethality in the GPCR DownstreamJ.6 Synthetic Lethality in the Translation ElongationJ.7 Synthetic Lethality in the Nonsense-mediated DecayJ.8 Synthetic Lethality in the 3′ UTRK.1 Synthetic Lethality and Vertex DegreeK.2 Synthetic Lethality and CentralityK.3 Synthetic Lethality and PageRankL.1 Information centrality distributionM.1 Synthetic Lethality and Heirarchy Score in PI3KM.2 Heirarchy Score in PI3K against Synthetic Lethality in PI3KM.3 Structure of Synthetic Lethality in PI3KM.4 Structure of Synthetic Lethality Resampling

Glossary

</td></tr><tr

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

</td></tr><tr

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

bioinformatics

Statistical or computational approaches to biological data or research tools.

bisulfite-Seq

Methylome data from sequencing bisulfite treated DNA.

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

class="td11">

CAGE-Seq

Transcriptome data from cap analysis of gene expression.

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

class="td11">

E-cadherin

Epithelial cadherin (calcium-dependent adhesion), a cell-adhesion protein encoded by the tumour suppressor gene CDH1.

epigenome

An analysis of epigenetic modifications of all genes in the genome.

exome

An approach or technology designed to generate or use data from all genes in the genome.

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

class="td11">

genomics

An approach or technology designed to generate or use data from all genes in the genome.

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

class="td11">

metabolome

An analysis of all the metabolites and enzymes in the cell.

metagenomics

An analysis of all the genes and genomes in a community.

microRNA

Short RNA molecules generally regarded to regulate gene expression by binding to mRNA.

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

class="td11">

omics

A rationale for applying molecular studies across all of the genes in the genome or biomolecules in the cell.

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

class="td11">

Pan cancer

Study or analysis or cancers across tissue of origin.

pleiotropy

A gene which has multiple biological functions.

proteome

An analysis of all the proteins expressed from the genome.

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

class="td11">

RNA-Seq

Transcriptome data from sequencing RNA.

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

class="td11">

Sanger

A dideoxy chain termination method for DNA sequencing (named after Fred Sanger).

synthetic lethal

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

</td><td style="white-space:wrap; text-align:left;" id="TBL-2-30-2"

class="td11">

transcriptome

An analysis of all the genes expressed in the genome.

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

Acronyms

</td></tr><tr

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

</td></tr><tr

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

AMPK

AMP-activated protein kinase.

ANOVA

Analysis of Variance.

AUROC

Area under the reciever operating characteristic (curve).

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

class="td11">

BioPAX

Biological Pathway Exchange.

BiSEp

Bimodal Subsetting Expression.

BMP

Bone morphogenic protein.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-12-2"

class="td11">

CCLE

Cancer Cell Line Encyclopaedia.

cDNA

Complementary deoxyribonucleic acid (from mRNA).

ChIP-Seq

Chromatin immunopreciptation sequencing.

COSMIC

Catalogue Of Somatic Mutations In Cancer.

CXCR

Chemokine receptors.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-18-2"

class="td11">

DAISY

Data mining synthetic lethal identification pipeline.

DDBJ

DNA Data Bank of Japan.

DNA

Deoxyribonucleic acid.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-22-2"

class="td11">

EMBL

European Molecular Biology Laboratory.

EMT

Epithelial-mesenchymal transition.

ENA

The European Nucleotide Archive.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-26-2"

class="td11">

GEO

Gene Expression Omnibus.

GO

Gene Ontology.

GPCR

G protein coupled receptor.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-30-2"

class="td11">

HDAC

Histone deacetylase.

HDGC

Hereditary diffuse gastric cancer.

HLRCC

Hereditary leiomyomatosis and renal cell carcinoma.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-34-2"

class="td11">

ICGC

International Cancer Genome Consortium.

InDel

Insertion or deletion (in DNA sequence).

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-37-2"

class="td11">

JAK

Janus kinase.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-39-2"

class="td11">

lncRNA

Long non-coding ribonucleic acid.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-41-2"

class="td11">

miRNA

microRNA.

mRNA

Messenger ribonucleic acid.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-44-2"

class="td11">

NCBI

National Center for Biotechnology Information (in the USA).

NCI

National Cancer Institute (in the USA).

NGS

Next-generation sequencing.

NHGRI

National Human Genome Research Institute (in the USA).

NIG

National Institute of Genetics (in Japan).

NIH

National Institutes of Health (in the USA).

NMD

Nonsense-mediated decay.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-52-2"

class="td11">

PCR

Polymerase chain reaction.

PDE

Phosphodiesterase.

PI3K

Phosphoinositide 3-kinase.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-56-2"

class="td11">

qPCR

Quantitative (real-time) polymerase chain reaction.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-58-2"

class="td11">

RFLP

Restriction fragment length polymorphism.

RGS

G-protein signaling.

RHO

Ras Homolog Family.

RNA

Ribonucleic acid.

RNAi

Ribonucleic acid interference.

RRBS

Reduced representation bisulfite sequencing.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-65-2"

class="td11">

SGI

synthetic genetic interaction.

siRNA

Short interfering ribonucleic acid.

SLIPT

Synthetic lethal interaction prediction tool.

SNP

Single nucleotide polymorphism.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-70-2"

class="td11">

TCGA

The Cancer Genome Atlas (genomics project).

TGFβ

Transforming growth factor β.

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-73-2"

class="td11">

UTR

Untranslated region (of mRNA).

</td><td style="white-space:wrap; text-align:left;" id="TBL-4-75-2"

class="td11">

WNT

Wingless-related integration site.

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

Chapter 1Introduction

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

1.1 Cancer Research in the Post-Genomic Era

Genomics technologies have the potential to vastly impact upon various areas including health and cancer medicine. Considering the progress in recent genomics research, this technology and the findings from it have considerable potential for significant impacts in the clinic and wider applications of genetics either directly or by enabling more focused genetics research on candidates selected from genomics or bioinformatics analysis. The completion of the draft Human Genome (Lander et al., 2001) marks a major accomplishment in genetics research and raises new challenges to utilise this genomic scale information effectively. Technologies in this area have rapidly developed since completion of the human genome project and many global large-scale projects have expanded upon the human genome, to populations (1000 Genomes, 2010), to cancers (Dickson, 1999; Zhang et al., 2011), and to deeper functional understanding (Kawai et al., 2001; ENCODE, 2004). However, impact on the clinic has been slower than initially anticipated following the completion of the “draft” genome with genomics technologies yet to become widely adopted in healthcare and oncology. Here we outline the genomics technologies and bioinformatics approaches which have led to availability of genomics data and techiques used in

this thesis and potential for applications in cancer research or the clinic in the future.

1.1.1 Cancer as a Global Health Concern

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

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

While the genetic contribution to cancer risk and many of the molecular changes occurring in cancers are widely acknowledged (ASCO, 2017; Cancer Research UK, 2017; Cancer Society of NZ, 2017), the majority of these findings have yet to impact on clinical practice. Diagnostics are traditionally based on pathological examination of tissue samples where histological staining for cell type, biomolecules and biomarkers continue to be widely used, although genetic and biochemical markers are being adopted for some cancer types. In general, the current standard of care includes surgery, radiation, and cytotoxic chemotherapy, depending on whether the cancer is localised or has become systemic (via metastasis) and spread to other organ systems. These approaches can be effective against certain cancers, particularly in early stage cancer or in patients with particular subtypes (such as acute myeloid leukaemia) which respond well to modern treatment regimens. Thus early diagnosis is important to patient survival and quality of life.

National screening programs (which prioritise patients with a high risk of cancer) therefore aim to diagnose cancers early and subtypes more accurately. Therefore identification of patients with genetic variants or family histories (for inclusion in national databases) for high risk of particular cancers is an important health issue, particularly where effective interventions exist if these cancers are diagnosed early. Thus high risk individuals being regularly monitored for some cancers are sometimes offered preventative surgery or treatment for pre-cancerous tissue (Guilford et al., 2010; Scheuer et al., 2002).

Chemotherapy is a last-resort treatment for many advanced stage (systemic) cancers which is designed to inhibit the growth and spread of cancer throughout the body by targeting rapidly growing cells. However, this approach often has severe adverse effects, a narrow therapeutic window, and is not suitable for chemopreventative application in many cases (Kaelin, Jr, 2009). Since surveillence preventative surgery (removing the tissue at risk of cancer) is not completely effective at preventing cancers and may impact on quality of life, depending on the cancer tissue types they are at risk of, alternative treatment strategies based on molecular biology and other fields are being investigated. These alternatives include immunological, endocrine, and targeted therapeutics, with a particular interest in treatments with specificity against cancer cells and wider applications (i.e., tolerable effective doses in applciations as a chemopreventative or against advanced stage cancers).

1.1.1.1 The Genetics and Molecular Biology of Cancers

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

cell divisions in which mutations could occur across tissue types (Tomasetti and Vogelstein, 2015).

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

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

Despite differences in subtyping performed by different research groups and companies, there is widespread agreement that distinguishing luminal, HER2-enriched, and triple negative tumours can be performed with expression profiles and have value in our understanding of cancer progression and prognostic importance for patients Dai et al. (2015). High-throughput technologies have the potential to enable such subtyping on a vast scale in discovery of further subtypes in breast cancer or other diseases and in identification of these subtypes along with mutations in routine clinical diagnostic and prognostic testing. The “Pan cancer” approaches by the cancer genome atlas project (as discussed in more detail in Section 1.1.5.1.1) expand on the importance of molecular differences between cancers by examining molecular profiles across cancer tissue types (Weinstein et al., 2013).

The molecular variability of cancer has also been approached rationally at a pathway level with patients subgroups activating different molecular pathways reflecting differences in disease mechanisms (Gatza et al., 2010). A robust approach to measuring pathway activation in cancer is with a“metagene” which gives a consistent signal as a consensus of expression across genes even if they are inversely correlated (Anjomshoaa et al., 2008; Huang et al., 2003; Nagalla et al., 2013). These are derived from the first principal component or eigenvector of a singular matrix decomposition, capturing the the most consistent variation across genes in a pathway or gene signature. Gatza et al. (2010) used gene signatures for 18 cellular pathways in breast cancer to define subtypes with distinct molecular pathway activity. In a meta-analysis of Affymetrix microarray expression data for 1,143 samples and 50 cell lines, unsupervised hierarchical clustering robustly defined subtypes with common homogeneous pathway activity despite variation in the specific mutations giving rise to them. These subtypes with shared pathway activity have similar molecular characteristics (such as DNA copy number), clinical properties and prognosis, building upon the intrinsic subypes (Parker et al., 2009) and providing a functional interpretation for molecular stratification (Gatza et al., 2010). The pathway-based subtypes often correspond to intrinsic subtypes (Gatza et al., 2010, 2014) and provide finer molecular stratification such as environmental stress response (to hypoxia within HER2-enriched cancers) (Gatza et al., 2011) and have pathway-specfic DNA copy number variation or essential genes (Gatza et al., 2014). Gatza et al. (2014) extend these analyses include 52 pathway signatures from previous publications in breast cancer, replicating known characteristics (such as hormone re) of each subtype and identifying novel subtype-specific driver genes of proliferation by analysis of microarray

expression from 476 from The Cancer Genome Atlas (TCGA). In addition to distinct biological functions driving growth of breast cancer subtypes, these molecular subtypes provide a rational approach based on molecular properties to cancer treatment with combination and targeted therapies (Gatza et al., 2010, 2014; Hanahan and Weinberg, 2000).

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

1.1.2 The Human Genome Revolution

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

1.1.2.1 The First Human Genome Sequence

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

specific to vertebrates) was much lower than most estimates for a genome of it’s size and the number of repetitious transposon elements was very high. Even the figure of 30–40,000 genes given by the original publication is now regarded to be an overestimate (Ezkurdia et al., 2014; IHGSC, 2004).

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

1.1.2.2 Impact of Genomics

Genomics has stimulated investigations into many of these previously largely explored areas of functional genetics and thus been of immense value in genetics research, attracting high expectations for further applications. Genomics research created widespread anticipation for potential applications in healthcare, agriculture, ecology, conservation, and evolutionary biology despite few of these being realised yet.

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

1.1.3 Technologies to Enable Genetics Research

1.1.3.1 DNA Sequencing and Genotyping Technologies

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

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

1.1.3.2 Microarrays and Quantitative Technologies

Real-time or qPCR is another adaptation of genetic technologies to quantitatively study nucleic acids, often reverse transcribed “deoxyribonucleic acid (cDNA)” or messenger “messenger ribonucleic acid (mRNA)” to measure (relative) gene expression or transcript abundance. While numerous quality control measures are required to correctly interpret a qPCR experiment, these have similarly become widely adopted and are still used for smaller scale experiments and as a “gold standard” for measuring gene expression (Adamski et al., 2014). This also represents a

shift in the application of qPCR and sequencing technology, where the primary interest is quantifying the amount of input material (by the rate of amplification to a certain level) rather than the qualitative nature of the sequence itself. The more recent technologies of microarrays and RNA-Seq have similarly embraced this application to quantify DNA copy number, ribonucleic acid (RNA) expression, and DNA methylation levels. Due to results of comparable or arguably better quality from these newer technologies (Beck et al., 2016; Git et al., 2010; McCourt et al., 2013; Robin et al., 2016), this “gold standard” status has begun to come under scrutiny.

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

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

1.1.3.3 Massively Parallel “Next Generation” Sequencing

Similar to microarrays, the introduction of massively parallel sequencing technologies have further expanded the availability of high-throughput molecular studies to researchers, with corresponding availability of genomics data from these studies. This “Next-Generation Sequencing” (NGS) expands not only gene expression studies (compared to microarrays) but extends to genome sequencing de novo for previously unknown genome and transcriptome

sequences at an unprecedented scale. This has been a particularly important technological revolution in genomics, as the cost and time of genome sequencing has dropped dramatically and enabled sequencing projects of far more samples and applications beyond the Human Genome Project. Particularly, when dealing with variants in a species with an existing reference sequence such as humans, where there is a low computational cost of mapping to a reference compared to a genome assembly. However, the cost of sequencing (RNA-Seq) for gene expression or DNA methylation studies is still considerably higher than a microarray study (limiting feasible sample sizes).

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

However, the methodology itself has challenges with the sample preparation, requiring a relatively high quantity of input material and “contamination” with over abundant ribosomal rRNA taking up the majority of the sequencing if not purified correctly. This abundance of rRNA is a particularly important issue in microarray and RNA experiments in Eukaryotes where it is commonplace to target the mRNA by binding to the poly-A tail (RNA-Seq) or 5’ cap (CAGE-Seq). However, this has the potential to exclude miRNA (miRNA) and long non-coding ribonucleic acid (lncRNA) of interest unless the sample is prepared specifically to study these. Similarly capturing a subsection of the genome for exome analysis or reduced representation bisulfite sequencing (RRBS), focuses on sequencing DNA sequences and methylation levels of CpG sites near known genes to reduce cost, noise, and incidental findings.

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

Molecular Profiling with Genomics Technology

NGS is highly adaptable to different applications: DNA sequencing (whole genome or exome), DNA methylation (bisulfite-Seq), RNA-Seq, miRNA, lncRNA, or chromatin immunopreciptation sequencing (ChIP-Seq). RNA-Seq of the transcriptome is a common adaptation where RNA is reverse transcribed and sequenced from the resulting cDNA. This is utilised to quantify the levels of RNA and identify which regions of DNA are expressed. Similar bisulfite treatment converts cytosine residues to uracil (sequenced as thymidine), sparing methylated cytosine enabling it to be distinguished with bisulfite-Seq for high-throughput detection of the notable epigenetic mark and is a common procedure to generate an epigenome. Subsets of the nucleic acid may be extracted for sequencing such as the coding regions of DNA (for the “exome”), the mRNA 5′ cap (CAGE-Seq), mRNA 3′ poly-A tail (RNA-Seq), microRNA, or an enriched subset of variable regions for DNA sequencing (“genotyping by sequencing”) and methylation studies (“reduced-representation bisulfite sequencing). High-throughput gel and mass spectrometry techniques have been applied to proteins and metabolites to generate the proteome and metabolome respectively. These “omics” technologies are applicable across a wide range of biomolecules in a cell and these “molecular profiles” are produced in many experimental laboratories.

Sequencing Technologies

Illumina sequencing (developed by Solexa and later acquired by Illumina) was released in 2006. It utilises reversible terminating dyes to sequence by synthesis with a lower accuracy (98%) and read lengths of 150–250bp. Illumina more than makes up for relatively short reads (along with improving the read length of the technology) and low accuracy with high-throughput and cost effectiveness, with a Hi-Seq 4000 platform producing up to 10 billion paired-end reads (1500Gbp) in a run of appropriately 3 days, capable of sequencing 12 human genomes (30× coverage) or 100 human transcriptomes simultaneously (Illumina, 2017). Illumina has further reduced the cost of sequencing with the economies of scale with the Hi-Seq X 10 claiming to produce a human genome (with 30× coverage) for less than US$1000, the first platform to achieve this long-standing goal in genomics. The high-throughput of Illumina sequencing also makes deep sequencing for high coverage, high quality consensus reads, and sensitive RNA-Seq experiments feasible. Illumina sequencing now has a dominating market share of the NGS technologies. Their Hi-Seq platforms were used in large-scale gemomics projects (such as the cancer genome atlas discussed in Section 1.1.5.1) to generate the sequence data used throughout this thesis.

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

Due to such benefits of sequencing over previous technologies (and their continued refinement), this thesis has focused on gene expression data generated by RNA-Seq rather by microarrays. RNA-Seq data is widely available as a resource from large-scale cancer genomics projects and methods to make inferences from RNA-Seq experiments could feasibly be applied to many other studies based on these current (or similar future) technologies.

1.1.3.4 Bioinformatics as Interdisciplinary Genomic Analysis

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

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

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

1.1.4 Follow-up Large-Scale Genomics Projects

A number of projects have attempted to follow up on the human genome project to varying degrees of success. The genomes have since been sequenced for a variety of model organisms, organisms of importance in health, agriculture, metagenomics of microorganisms (microbiome), ecology and conservation. Genomics projects have also been applied functional genetics (Kawai et al., 2001; ENCODE, 2004) and to human populations with an interest

variability between individuals and health or disease risk (HapMap, 2003; 1000 Genomes, 2010).

Genomics databases have also focused on faciltating distribution of genomic data generated by researchers, rather than generating it themselves. GenBank hosted by the National Center for Biotechnology Information (NCBI) in the US, the The European Nucleotide Archive (ENA) hosted by the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ) hosted by the National Institute of Genetics (NIG) do so by serving as public repositories of DNA sequence data. Gene Expression Omnibus (GEO) (Clough and Barrett, 2016), arrayExpress (Rustici et al., 2013), and caArray (Heiskanen et al., 2014) serve a similar purpose as a resource for gene expression datasets, originally developed for microarray data but RNA-Seq data is now supported by some platforms. They are repositories for researchers to deposit, share, and access gene expression data, serving as a resource to support ongoing research where larger datasets than would were previously accessible for many individual laboratories are available (Rung and Brazma, 2013). These resources cover not only DNA sequence across the genome but also molecular profiles of other factors by adapting genomic sequencing or other high throughput technologies for quantifying gene expression or DNA methylation. Sharing the expression datasets generated in a publication is now required by some journals.

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

1.1.5 Cancer Genomes

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

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

1.1.5.1 The Cancer Genome Atlas Project

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

The TCGA projects set out to generate a molecular “profile” of the tumour (and some matched normal tissue) samples: the genotype, somatic mutations, gene expression, DNA copy number, and RNA methylation levels. While these were originally performed

largely with microarray technologies, exome and RNA-Seq has been since adopted and performed for many TCGA samples, with whole genomes being performed for some samples. Data which cannot be used to identify the patients (such as somatic mutation, expression, methylation, and various clincal factors) are publicly available.

Findings from Cancer Genomes

The Cancer Genome Atlas pilot projects (Bell et al., 2011; Hammerman et al., 2012; McLendon et al., 2008) serve to demonstrate the power of applying genomics technologies to cancer research at such as scale. In addition to sequencing the whole genome or a subset (exome), DNA copy number, gene expression, DNA methylation, and somatic mutations were also analysed. The initial projects used microarray technologies for expression and methylation data but these have since been replaced by RNA-Seq for expression. TCGA demonstrated the potential discovery of the molecular basis of cancer by analysing 206 glioblastoma brain cancer samples (McLendon et al., 2008), highlighting the roles of ERBB2, NF1, TP53, and PIK3R1 mutations, along with altered methylation of MGMT, and the core pathways of RTK, p53, and RB signalling in brain cancer. An analysis of 489 serious ovarian cancers (Bell et al., 2011) similarly reported TP53 mutations specifically over-represented in high grade tumours and reported 133 copy number variants, 168 differentially methylated regions, and recurrently somatic mutations in 9 genes in low grade tumours including NF1, BRCA1, BRCA2, RB1, and CDK12. Four transcriptional subtypes of ovarian cancers were identified, alterations in BRCA1, BRCA2, and CCLE had an impact on patient survival, and the homologous recombination, NOTCH and FOXM1 signalling pathways were involved in ovarian cancer growth. The genomics of 178 squamous cell lung cancers (Hammerman et al., 2012) were highly complex, averaging at 360 muations in coding regions. While no targeted therapies existed for this cancer subtype, 11 recurrently mutated genes were identified including TP53 and HLA-A. The pathways altered in various squamous cell lung cancers were NFE2L2, KEAP1, differentiation genes, PI3K, CDKN2A and RB1. These aberrant genes and pathways represent potential therapeutic targets which could be identified for most samples.

The TCGA breast cancer analysis (TCGA, 2012) consisted of 802 samples with exomes, copy number variants, RPPA protein quantification, and DNA methylation, mRNA, and

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

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

The TCGA stomach cancer analysis of 295 samples (Bass et al., 2014) identified 4 molecular subtypes of stomach cancers characterised by: the Epstein-Barr virus, MSI, genomics instability, and chromosomal instability. Abberrations in PD-L1, PIK3CA, and JAK2 were also identified in stomach cancers which may present therapeutic targets.

Genomic Comparisons Across Cancer Tissues

TCGA have identified various genes as recurrent, driver mutations across cancer types which are likely to have a role in driving the proliferation of these cancers and present a molecular target that could be applied across tissue types. These include TP53 (in brain, lung/head/neck squamous cell, breast, colorectal, uterine, and endometrial cancers), ERBB2 /HER2/NEU (in brain, breast, colorectal, bladder, and lung cancers), PIK3CA, PIK3R1 (in brain, breast, colorectal, endometrial, bladder, clear cell renal, and lung cancers), BRCA1/BRCA2 (in breast and ovarian cancers), NF1 (in brain, ovarian, and skin cancers), ARID1A (in colorectal, endometrial, and clear cell

renal cancers), KRAS (in colorectal, endometrial, and skin cancers), BRAF (in colorectal, thyroid, and skin cancers), EGFR (in brain, breast, and lung cancers), and PTEN (in breast, endometrial, and uterine cancers) (Agrawal et al., 2014; Akbani et al., 2015; Bass et al., 2014; Bell et al., 2011; Burk et al., 2017; Cherniack et al., 2017; Ciriello et al., 2015; Collisson et al., 2014; Creighton et al., 2013; Hammerman et al., 2012; Kandoth et al., 2013; Lawrence et al., 2015; McLendon et al., 2008; Muzny et al., 2012; TCGA, 2012; Weinstein et al., 2014). In addition to disregarding the tissue-based distinction between colon and rectal cancers based on molecular similarlity (Muzny et al., 2012), the TCGA project have observed differences within tumour types and proposed molecular subtyping for breast, clear cell renal, papillary renal, stomach, skin, bladder, and prostate cancers (Abeshouse et al., 2015; Akbani et al., 2015; Bass et al., 2014; Ciriello et al., 2015; Creighton et al., 2013; Hammerman et al., 2012; Linehan et al., 2016; Muzny et al., 2012; TCGA, 2012; Weinstein et al., 2014).

The “Pan cancer” project (Hoadley et al., 2014; Weinstein et al., 2013) analysed 3527 samples across 12 tissue types for DNA, RNA, protein, and epigenetic molecular profiles. This project was initated in 2012 to perform a comprehensive analysis of molecular data across cancer types to identify molecular simliarities and differences. Recurrent TP53 mutations characterised high grade tumours across breast, ovarian, and endometrial cancers. HER2 was identified in brain, endometrial, bladder, and lung cancers, in addition to the known role of HER2 in breast cancers. BRCA1 and BRCA2 mutations were also detected across cancers, mainly breast and ovarian cancers as expected. Microsatellite instability characterised both endometrial and colorectal cancers. The Pan cancer project (Hoadley et al., 2014) has identified 11 molecular subtypes across these tissues, 5 of corresponding to tissue cancer types and the remainder reassigned due to molecular similarities shared across cancer types. Squamous cell lung, head, and neck and a subset bladder cancers were grouped together by molecular similarities, characteried by a high frequency of TP53 mutations. Conversely, bladder cancers were divided into 3 of these molecular subtypes with distinct profiles. This project further supports the genomic stratification of patients, demonstrated in breast cancer (Parker et al., 2009; Pereira et al., 2016; Perou et al., 2000), which may apply to other cancer types and to molecular characteristics across them targeting recurrent mechanisms of cancer growth and progression (Hanahan and Weinberg, 2000, 2011).

Cancer Genomic Data Resources

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

1.1.6 Genomic Cancer Medicine

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

1.1.6.1 Cancer Genes and Driver Mutations

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

cancer either by loss of function mutations, deletion of DNA copies, repression of gene expression, or hypermethylation. Their normal functions are typically to regulate cell division, DNA repair, and cell signalling.

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

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

1.1.6.2 Personalised or Precision Cancer Medicine

The notion of using a patient’s genome to tailor healthcare to an individual has been appealing since the advent of genomics, popularised with the term “personalised medicine”. This approach was expected to span from preventative lifestyle advice to effective treatments. Personalised medicine was described in contrast with current strategies of health advice, screening, prognostics, and treatments based on what works well for the majority of the population. Adverse effects of these treatments occur in a

significant subpopulation, particularly demographics under-represented in clinical studies.

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

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

Molecular Diagnostics and Pan-Cancer Medicine

There is growing support for the use of molecular tools such as mutations or gene expression signatures to diagnose tumour subtypes addition to (or in lieu of) tissue of origin or histology. This is particularly important in breast cancer where analysis of molecular data detected several distinct “intrinsic subtypes” with differences in malignancy and patient outcome which were distinguished by molecular mechanisms rather than tissue or cellular phenotype (Parker et al., 2009; Perou et al., 2000). Conversely, common molecular mechanisms may be shared between cancers across tissue types as discovered by the “Pan cancer” studies, such as those conducted by the TCGA and International Cancer Genome Consortium (ICGC) projects, which combined molecular profiles across tissue types Weinstein et al. (2013). The molecular subtypes could feasibly be included in clinical

testing as a panel of biomarkers for diagnostics and prognosis. Such biomarkers also have the potential to monitor drug response or risk of recurrence. This is also raises the need for development of treatments that target these molecular subtypes.

1.1.6.3 Targeted Therapeutics and Pharmacogenomics

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

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

Targeting Oncogenic Driver Mutations

Oncogene targeted therapies have also been developed with some examples of effective clinical application against cancers. However, they have already begun to manifest problems with resistance, recurrence, tissue specificity, and design of inhibitors specific to oncogenic variants rather than proto-oncogene precusors. Targeted anticancer therapeutics can exploit complex interactions to distinguish normal and cancerous cells which may benefit from studies of gene regulation or interaction networks. The unexpected synergy between inhibitors of the oncogenes BRAF V600E and EGFR in colorectal cancer is an example of such a system Prahallad et al. (2012).

Despite successful application of vemurafenib against BRAFV600E in melanomas Dienstmann and Tabernero (2011); Ravnan and Matalka (2012), colorectal cancers with BRAF V600E mutations have poor prognosis and lack drug response. Prahallad et al. (2012) used an RNA interference (RNAi) screen and found that EGFR inhibition is synergistic with vemurafenib aginst BRAFV600E in colon cell lines and xenografts due to feedback activation of EGFR. Vemurafenib induced rapid reactivation of MAPK/ERK signalling via EGFR in colorectal cell lines in a tissue-specific manner Corcoran et al. (2012) and may be be relevant to acquired resistance in melanoma Sun et al. (2014). Thus combination therapies against several molecular pathways may be necessary to anticipate acquired resistance Ravnan and Matalka (2012) and targeted therapeutics may be further refined from understanding the pathway structure and functional interactions across cancer cells.

1.1.6.4 Systems and Network Biology

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

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

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

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

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

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

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

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

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

Network Medicine, and Polypharmacology

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

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

A molecular map of the interactions and pathways in the mammalian cellular network has the potential to impact upon drug design and clinical practice, particularly in treatment of cancer and infectious disease. Characterisation of the target system and impact of existing treatments, such as BRAFV600E and EGFR inhibitors, enable wider application of the mechanisms for such interventions exploiting genetic interactions or pathways. This could lead to development of more effective treatment interventions for these systems and prediction of similar molecular systems for development of novel drug targets and combinations.

1.2 A Synthetic Lethal Approach to Cancer Medicine

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

1.2.1 Synthetic Lethal Genetic Interactions

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

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

with specific inhibition of a mutant cell, including SSL interactions. Mutations are “alleviating” in positive SGIs with less deviation from the wild-type than expected. For viability, “suppression” and “rescue” are positive SGIs giving at least partial restoration of wild-type growth from single mutants with growth impairment and lethal phenotypes respectively. Negative SGIs were markedly more common than positive SGIs in a number of studies in model systems Boucher and Jenna (2013); Tong et al. (2004).

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

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

1.2.2 Synthetic Lethal Concepts in Genetics

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

Since synthetic lethal gene partners can be affected by extracellular stimuli such as chemicals, essentiality of synthetic lethal genes can be induced by the environment of a cell. An environmental stress condition may inhibit one or the other synthetic lethal gene, such as exposure to chemicals, in which case the synthetic lethal partner gene is “conditionally essential” (Hillenmeyer, 2008). Thus the evolutionary rationale for the abundance of SGIs (compared to the surprisingly low number of essential genes) in a Eukaryotic genome attributed to genetic functional redundancy and network robustness of a cell which are advantageous to survival.

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

1.2.3 Studies of Synthetic Lethality

Genetic high-throughput screens have identified unexpected, functionally informative, and clinically relevant synthetic lethal interactions; including synthetic lethal partners of genes recurrently mutated in cancer or attributed to familial early-onset cancers. While screening presents an appealing strategy for synthetic lethal discovery, computational approaches are becoming popular as an alternative or complement to experimental methods to overcome inherent bias and limitations of experimental screens. An array of recently

developed computational methods (Jerby-Arnon et al., 2014; Lu et al., 2015; Tiong et al., 2014; Wang and Simon, 2013; Wappett, 2014) show the need for synthetic lethal discovery in the fundamental genetics and translational cancer research community. However, existing computational methods are not suitable for queries of genomic data for interacting partners of a particular gene: they have been applied pairwise across the genome, do not have software released to apply the methodology, or lack statistical measures of error for further analysis. A robust prediction of gene interactions is an effective and practical approach at a scale of the entire genome for ideal translational applications, analysis of biological systems, and constructing functional gene networks.

1.2.3.1 Synthetic Lethal Pathways and Networks

SGIs are very common in genomes, with 4× more interactions detected with synthetic gene array mating screens than protein-protein interactions yeast-2-hybrid studies (Tong et al., 2004). The SGI network is scale-free with power-law vertex degree distribution and low average shortest path length (3.3) as expected for a complex biological network (Barabási and Oltvai, 2004). Highly connected “hub” genes with the highest number of links (vertex degree) are functionally important with many negative SGI hubs involved in cell cycle regulation and many positive SGI hubs involved in translation (Baryshnikova et al., 2010b; Costanzo et al., 2010). Negative SGIs were far more common than positive SGIs, with synthetic gene loss being more likely to be deleterious to cell than advantageous which indicates than synthetic lethality may be comparably easier to detect than other SGIs.

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

Evolution of Synthetic Lethality

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

A modest 5% of interactions were conserved between unicellular (S. cerevisiae) and multicellular (C. elegans) organisms. However, the nematode SGI network had similar scale-free topology and modularity despite differences in methodology: metazoan RNAi screens are incomplete knockouts whereas screening null mutations is feasible in yeast (Bussey et al., 2006). The nematode SGI screen identified network hubs with important interactions to orthologues of known human disease genes (Lehner et al., 2006). Despite the lack of direct conservation of SGIs between yeasts and nematode worms, genetic redundancy at the gene or pathway level may yet be consistent with an induced essentiality model of SGIs where gene functions are conserved with network restructuring over evolutionary change (Tischler et al., 2008). While nematode models are more closely related to human cells, cancer cells can present growth and viability phenotypes more comparable to yeast models. Therefore findings from both SGA and RNAi models are relevant to understanding cellular network structure and in healthy and cancerous human cells. RNAi has also been applied to human and mouse cancer cells in cell culture and genetic screening experiments. These findings suggest that SGI network “rewiring” is a concern for identifying specific synthetic lethal interactions in cancer and a pathway approach may be more robust in the context of evolution, patient variation, tumour heterogeneity, and disease progression.

1.2.4 Synthetic Lethal Concepts in Cancer

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

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

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

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

1.2.5 Clinical Impact of Synthetic Lethality in Cancer

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

BRCA1 or BRCA2 and PARP1 genes demonstrate the application of the synthetic lethal approach to cancer therapy Ashworth (2008); Kaelin, Jr (2005). BRCA1 and BRCA2 are homologous DNA repair genes, widely known as tumour suppressors; mutation carriers have substantially increased risk of breast (risk by age 70 of 57% for BRCA1 and 59% for BRCA2 ) and ovarian cancers (risk by age 70 of 40% for BRCA1 and 18% for BRCA2) (Chen and Parmigiani, 2007). The BRCA1 or BRCA2 genes, which usually repair DNA or destroy the cell if it cannot be repaired, have inactivating somatic mutations in some familial and sporadic cancers. Poly-ADP-ribose polymerase (PARP) genes are tumour suppressor genes involved in base excision DNA repair. Loss of PARP activity results in single-stranded DNA breaks. However, PARP1-∕- 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 BRCA1 or BRCA2 deficient embryonic stem cell mouse xenografts. They found evidence that PARP inhibition causes DNA lesions, usually repaired in wild-type cells, which lead to chromosomal instability, cell cycle arrest, and induction of apoptosis in BRCA1 or BRCA2 deficient cells. Therefore, the pathways cooperate to repair DNA giving a plausible mechanism for combined loss as an effective anti-cancer treatment.

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

A PARP inhibitor, olaparib, showed fewer adverse effects than cytotoxic chemotherapy and anti-tumour activity in phase I trials against BRCA1 or BRCA2 deficient familial breast, ovarian, and prostate cancers (Fong et al., 2009) and sporadic ovarian cancer (Fong et al., 2010). AstraZeneca has reported phase II trials showing the treatment is effective in BRCA1 or BRCA2 deficient breast (Tutt et al., 2010) and ovarian cancers (Audeh et al., 2010) with a favourable therapeutic window and similar toxicity between carriers of BRCA1 or BRCA2 mutations and sporadic cases. AstraZeneca announced that olaparib has begun phase III trials for breast and ovarian cancers in 2013. Mixed results in phase II trials in ovarian cancer are behind the delays addressed by retrospective analysis of the cohort subgroup with confirmed mutation of BRCA1 or BRCA2 genes in the tumour;

unsurprisingly these patients, benefit most from the PARP inhibitor treatment and have increased platinum sensitivity in combination treatment. These PARP inhibitors are FDA approved for some cancers McLachlan et al. (2016), are effective against germline and sporadic BRCA1 or BRCA2 mutations, and are a potential prevention alternative to prophylactic surgery for high risk mutation carriers Ström and Helleday (2012).

This demonstrates the clinical impact of a well characterised system of synthetic lethality with known cancer risk genes. Synthetic lethality has the benefit of being effective against inactivation of tumour suppressor genes by any means, broader than targeting a particular oncogenic mutation (Kaelin, Jr, 2005). The targeted therapy is effective in both sporadic and hereditary BRCA1 or BRCA2 deficient tumours acting against an oncogenic molecular aberration across several tissues.

1.2.6 High-throughput Screening for Synthetic Lethality

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

Experimental screening for synthetic lethality is an appealing strategy for wider discovery of functional interactions in vivo despite many potential sources of error which must be considered. The synthetic lethal concept has both genetic and pharmacological screening applications to cancer research. Genetic screens, with RNAi to discover the specific genes

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

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

suggest a need for complementary computational approaches to synthetic lethal discovery.

1.2.6.1 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 for synthetic lethality between PIM1 overexpression and gene knockdown in RWPE prostate cancer cell lines. PLK1 gene knockdown and drug inhibition was effective as a specific inhibitor of PIM1 overexpressing prostate cells in cell culture and mouse tumour xenografts. PLK1 inhibition reduced MYC expression in pre-clinical models, consistent with expression in human tumours in which PIM1 and PLK1 are co-expressed and correlated with tumour grade. Thus RNAi screening was valuable to identify therapeutic targets and biomarkers for patient response as demonstrated with the finding of PLK1 as a candidate drug target against prostate cancer progression.

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

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

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

Chemical genetic screens are also a viable strategy to identify therapeutically relevant synthetic lethal interactions. Bitler et al. (2015) investigated ARID1A mutations, aberrations in chromatin remodelling known to be common in ovarian cancers, for drug response. Ovarian RMG1 cells were screened for drug response specific to ARID1A knockdown cells. They used ARID1A gene knockdown for

consistent genetic background, with control experiments and 3D cell culture to ensure relevance to drug activity in the tumour micro-environment. Screening a panel of commercially available drugs targeting epigenetic regulators found ESH2 methyltransferase inhibitors effective and specific against ARID1A mutation with validation in a panel of ovarian cell lines. Synthetic lethality between ARID1A and ESH2 was supported by decreases in H3K27me3 epigenetic marks and markers of apoptosis in response to ESH2 inhibitors. This was mechanistically supported with differential expression of PIK3IP1 and association of both synthetic lethal genes with the PIK3IP1 promoter identifying the PI3K-AKT signalling pathway as disrupted when both genes were inhibited.

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

The examples above show that high-throughput screens are an effective approach to discover synthetic lethality in cancer with a wide range of applications. Screens are more comprehensive than hypothesis-driven candidate gene approaches and successfully find known and novel synthetic lethal interactions with potential for rapid clinical application. They have the power to test mode of action of drugs, find unexpected synthetic lethal interactions between pathways, or identify effective treatment strategies without needing a clear mechanism. However, synthetic lethal screens are costly, labour-intensive, error-prone, and biased towards genes with effective RNAi knockdown libraries. Limited genetic background, lethality to wild-type cell during gene knockdown, off-target effects, and difficultly replicating synthetic lethality across different cell lines, tissues, laboratories, or conditions stems from a high false positive rate and a lack of standardised thresholds to identify synthetic lethality in a high-throughput screen. Therefore there is a need for replication, validation, and alternative approaches to identify synthetic lethal candidates. Varied conditions between

experimental screens and differences between RNAi and drug screens renders meta-analysis ineffective.

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

1.2.7 Computational Prediction of Synthetic Lethality

1.2.7.1 Bioinformatics Approaches to Genetic Interactions

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

Table 1.1: Methods for Predicting Genetic Interactions

Method

</td><td style="white-space:nowrap; text-align:left;" id="TBL-5-1-2"

class="td11">

Input Data

</td><td style="white-space:nowrap; text-align:left;" id="TBL-5-1-3"

class="td11">

Species

</td><td style="white-space:nowrap; text-align:left;" id="TBL-5-1-4"

class="td11">

Source

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

class="td11">

Tool Offered

Between Pathways Model

PPI, SGI

S. cerevisiae

Kelley and Ideker (2005)

Within Pathways Model

PPI, SGI

S. cerevisiae

Kelley and Ideker (2005)

Decision Tree

PPI, expression, phenotype

S. cerevisiae

Wong et al. (2004)

2 Hop

Logistic Regression

SGI, PPI, co-expression, phenotype

C. elegans

Zhong and Sternberg (2006)

Gene Orienteer

Network Sampling

SGI, PPI, GO

S. cerevisiae

Le Meur and Gentleman (2008)

Le Meur et al. (2014)

SLGI(R)

Random Walk

GO, PPI, expression

S. cerevisiae

C. elegans

Chipman and Singh (2009)

Shared Function

Co-expression, PPI, text mining, phylogeny

C. elegans

Lee et al. (2010b)

WormNet

Logistic Regression

Co-expression, PPI, phenotype

C. elegans

Lee et al. (2010a)

GI Finder

Jaccard Index

GO, SGI, PPI, phenotype

Eukarya

Hoehndorf et al. (2013)

Machine Learning

Pandey et al. (2010)

MNMC

Machine Learning Meta-Analysis

Wu et al. (2014)

MetaSL

Flux Variability Analysis

Flux Balance Analysis

Network Simulation

Metabolism

E. coli

M. pneumoniae

Güell et al. (2014)

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

Table 1.2: Methods for Predicting Synthetic Lethality in Cancer

Method

</td><td style="white-space:nowrap; text-align:left; vertical-align:middle;" id="TBL-10-1-2"

class="td11">

Input Data

</td><td style="white-space:nowrap; text-align:left; vertical-align:middle;" id="TBL-10-1-3"

class="td11">

Source

</td><td style="white-space:nowrap; text-align:left;" id="TBL-10-1-4"

class="td11"> Tool Offered

Network Centrality

protein-protein interactions

Kranthi et al. (2013)

Differential Expression

Expression

Mutation

Wang and Simon (2013)

Comparative Genomics

Chemical-Genomics

Yeast synthetic gene interactions

Homology

Heiskanen and Aittokallio (2012)

Comparative Genomics

Yeast synthetic gene interactions

Homology

Deshpande et al. (2013)

Machine Learning

Discussed by Babyak (2004)

and Lee and Marcotte (2009)

Differential Expression

Expression

Tiong et al. (2014)

Literature Database

Li et al. (2014)

Syn-Lethality

Meta-Analysis

Meta-Analysis

Machine Learning

Wu et al. (2014)

MetaSL

Pathway Analysis

Zhang et al. (2015)

Protein Domains

Homology

Kozlov et al. (2015)

Data-Mining

Machine Learning

Expression

Somatic mutation and DNA CNV

siRNA in cell lines

Jerby-Arnon et al. (2014)

Ryan et al. (2014)

Crunkhorn (2014)

Lokody (2014)

DAISY (method)

Genome Evolution

Hypothesis Test

Machine Learning

Expression

DNA CNV

Known SL

Lu et al. (2013)

Lu et al. (2015)

Bimodality

Expression

DNA CNV

Somatic Mutation

Wappett (2014)

Wappett et al. (2016)

BImodal Subsetting ExPression (BiSEp)

Directional Chi-Square

Expression (microarray)

Somatic mutation

Kelly, S. T., Guilford, P. J., and Black, M. A.

Dissertation (Kelly, 2013) and developed here

</td></tr></table> </div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-10-15-4"

class="td11"> SLIPT

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

There are a number of existing computational methods for predicting synthetic lethal gene pairs in humans with a specfic interest in cancer (as summarised in 1.2). While these demonstrate the power and need for predictions of synthetic lethality in human and cancer contexts, limitations of previous methods could be met with a different approach. Existing computational approaches to synthetic lethal prediction are often difficult to interpret, replicate for new genes, or are reliant on data types not available for a wider range of genes to test.

1.2.7.2 Comparative Genomics

A comparative genomics approach by Deshpande et al. (2013) used the results of well characterised high-throughput mutation screens in S. cerevisiae as candidates for synthetic lethality in humans (Baryshnikova et al., 2010a; Costanzo et al., 2010, 2011; Tong et al., 2001, 2004). Yeast synthetic lethal partners were compared to human orthologues to find cancer relevant synthetic lethal candidate pairs with direct therapeutic potential. Proposed as a complementary approach to siRNA screens, approximately 24,000 of the 116,000 negative SGIs in yeast (Costanzo et al., 2011) were matched to human orthologues, with over 500 involving a cancer gene (Futreal et al., 2004). Under strict criteria of one-to-one orthologues, large effect size and significant interaction in yeast data (ϵ < -0.2, p < 0.05), 1,522 interactions were identified with 70 involving cancer genes. Of the 21 gene interactions tested with pairs of siRNA in IMR1 fibroblast cells, 6 exhibited synthetic lethal effects. The two strongest interactions (SMARCB1 with PSMA4 and ASPSCR1 with PSMC2) were successfully validated by protein analysis of human cells and replication with tetrad analysis for yeast orthologues.

Another approach to systematic synthetic lethality discovery specific to human cancer (in contrast to the plethora of yeast synthetic lethality data) was to build a database as done by 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 dedicated to human cancer synthetic lethal interactions and integrates gene function annotation, pathway and molecular mechanism data with experimental and predicted synthetic lethal gene pairs. This combination of data sources is intended to tackle

the trade-off between more conclusive synthetic lethal experiments in yeast and more clinically relevant synthetic lethal experiments in human cancer models, such as RNAi, especially when high-throughput screens are costly and prone to false positives in either system and difficult to replicate across gene backgrounds. This database centralises a wealth of knowledge scattered in the literature including cancer relevant genes (BRCA1, BRCA2, PARP1, PTEN, VHL, MYC, EGFR, MSH2, KRAS, and TP53) and is publicly available as a Java App. These included the previously mentioned interactions of BRCA1 and BRCA2 with PARP1 and TP53 with WEE1 and PLK1. However, the computational methodology was not released, so it is not possible to replicate their results, nor to add to the findings with new datasets, which are limited to 647 human genes. Suggested future directions were promising, such as constructing networks of known synthetic lethality, applying known synthetic lethality to cancer treatment, data mining, replicating the approach for synthetic lethality in model organisms, signalling pathways, and developing a complete global network in human cancer or yeast (both of which are still incomplete with experimental data), some of which has been implemented in “SynLethDB” (Guo et al., 2016).

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

Method

</td><td style="white-space:nowrap; text-align:left; vertical-align:middle;" id="TBL-28-1-2"

class="td11">

Source

</td><td style="white-space:nowrap; text-align:left;" id="TBL-28-1-3"

class="td11"> Tool Offered

Random Forest

Breiman (2001)

Random Forest

J48 (decision tree)

Bayes (Log Regression)

Bayes (Network)

PART (Rule-based)

RBF Network

Bagging / Bootstrap

Classification via Regression

Hall et al. (2009)

WEKA

Support Vector Machine (Linear)

Vapnik (1995)

Support Vector Machine (RBF – Gaussian)

Joachims (1999)

Multi-Network Multi-Class (MNMC)

Pandey et al. (2010)

MetaSL (Meta-Analysis)

Wu et al. (2014)

MetaSL

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

Machine learning approaches have also been proposed for synthetic lethal discovery (Babyak, 2004; Lee and Marcotte, 2009). Due to concerns that these may be subject to overfitting or noise, Wu et al. (2014) developed a meta-analysis method (based on the machine learning methods in Table 1.3) for synthetic lethal gene pairs relevant to developing selective drugs against human cancer, building upon their previous database (Li et al., 2014). They used training data of 10,885 synthetic lethal interactions from yeast experiments of which 7347 occurred between the 5,504 non-essential genes. Their “metaSL” approach utilises genomic, proteomic and annotation data (including GO terms Ashburner et al. (2000), PPI, protein complexes, and biological pathway) with strong statistical performance in yeast data (AUROC of 0.871). The predicted orthologous synthetic lethal partners in human data were not experimentally validated but several would be relevant to cancer such as EGFR with PRKCZ. They note that computational approaches scale-up across the genome at lower cost than experimental screen and share their most supported interactions online. However, the method is not available for analysis of other genes studied by the cancer research community. While machine learning has great potential as a predictor, the results vary greatly depending on the predictive features selected and it is not clear which threshold should be used to report reliably detected genes. Syn-Lethality (Li et al., 2014) and MetaSL (Wu et al., 2014) demonstrate the value of computational approaches to synthetic lethality but omit many genes of importance in cancer, such as CDH1, and there remains a need to enable biological researchers to query such genes in a particular tissue or genetic background.

There is also concern for analyses based on yeast data that many synthetic lethal interactions may not be conserved between species Dixon et al. (2009), although interactions between pathways may be more comparable. It is unsurprising that many of the interactions identified were not experimentally validated. There have been many gene duplications in the separate evolutionary histories of humans and yeast which may lead to differences in genetic redundancy. Yeast are not an ideal human cancer model because they are do not have tissue specificity, multicellular gene regulation, or orthologues to a number of known cancer genes such as p53. Although these studies have tried to anticipate these issues with stringent criteria such as requiring one-to-one orthologues, there remains the possibility that changes in gene function may affect whether these are solely redundant such as if functions had coevolved without sequence

homology. Many genes will also be excluded since they lack homologues in yeast, the corresponding experimental data, or having paralogues in either species. Thus conservation of yeast interactions is not an ideal strategy and analysis of human data directly for comparison with human experimental data will be the focus of this thesis.

1.2.7.3 Analysis and Modelling of Protein Data

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

However, their approach was limited to known cancer genes and is not applicable to genes that do not have PPI data. Other nucleotide sequencing data types are more commonly available for cancer studies at a genomic scale. Of further concern is that the results were enriched for p53 synthetic lethal partners which is relevant to many cancer researchers but this genome-wide approach did not detect many other cancer genes due to multiple testing. This enrichment may be due to the known drastic effect of removing p53 itself from the network as a master regulator, cancer driving tumour suppressor gene, and highly connected network “hub”. The focus on cancer genes is useful for translation into therapeutics but does not account for variable genetic backgrounds or effect of protein removal on the whole cellular network.

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

1.2.7.4 Differential Gene Expression

Differential gene expression has been explored to predict synthetic lethal pairs in cancer which would be widely applicable due to the availability of public gene expression data for a large number of samples and cancer types. Wang and Simon (2013) found differentially expressed genes (by the t-test, adjusted by FDR) between tumours with or

without functional p53 mutations in TCGA (McLendon et al., 2008) and Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012) RNA-Seq gene expression data as candidate synthetic lethal partner pathways of p53. They identified 2, 8, and 21 candidate synthetic lethal partner genes in 3 microarray datasets from the NCI60 cell lines, 31 partner genes from the CCLE RNA-Seq data, and 50 in TCGA RNA-Seq data. PLK1 was replicated across 4 of these analyses and 17 other genes were replicated across 2 analyses (including MTOR, PLK4, MAST2, MAP3K4, AURKA, BUB1 and 6 CDK genes) with many playing a role in cell cycle regulation. This was supported by a drug sensitivity experiment on the NCI60 cell lines which found that cells which lacked functional p53 were more sensitive to paclitaxel (which targets PLK1, AURKA, and BUB1). This demonstrated the potential of gene expression as a surrogate for gene function and use of public genomic data to predict synthetic lethal gene pairs in cancer. Wang and Simon (2013) advocated for pre-screening of expression profiles to augment future RNAi screens. However, the analyses were limited to kinase genes and focused on currently druggable genes, lacking wider application of synthetic lethal prediction methodology. This approach may not be feasible or applicable in cancer genes with a lower mutation rate than p53.

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

samples (from a non-caucasian population) to find those applicable to colorectal cancers.

1.2.7.5 Data Mining and Machine Learning

Recognising the utility of synthetic lethality to drug inhibition and specificity of anti-cancer treatments, Jerby-Arnon et al. (2014) also saw the need for effective prediction of gene essentiality and synthetic lethality to augment experimental studies of SL. They developed the DAta mIning SYnthetic lethal identification pipeline (DAISY), a data-driven approach for genome-wide analysis of synthetic lethality in public cancer genomics data from TCGA and CCLE (Barretina et al., 2012). DAISY is intended to predict the candidate synthetic lethal partners of a query gene such as genes recurrently mutated in cancer.

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

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

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

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

Lu et al. (2015) critique the assumption of co-expression in the DAISY methodology and propose an alternative computational prediction of synthetic lethality based on machine learning methods and a “cancer genome evolution” hypothesis. Using DNA copy number and gene expression data from TCGA patient samples, a cancer genome evolution model assumes that synthetic lethal gene pairs behave in two

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

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

Large scale, computational approaches have yet to determine whether synthetic lethal interactions are tissue-specific since Lu et al. (2015) used pan-cancer data for 14136 patients with 31 cancer types. Experimental data used for comparison was a small training dataset specific to colorectal cancer, and based on screens for other phenotypes, which may limit performance of the model or application to other cancers. Proposed expansion of the computational approach to mutation, microRNA, or epigenetic modulation of gene function and tumour micro-environment or heterogeneity suggests that synthetic lethal discovery could be widely applied to the

current challenges in cancer genomics. This approach was also based on machine learning methodology and not supported by a software release for the community to develop, contribute to, or reproduce beyond the gene pairs given in the supplementary results.

1.2.7.6 Bimodality

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

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

1.2.7.7 Rationale for Further Development

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

1.3 E-cadherin as a Synthetic Lethal Target

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

1.3.1 The CDH1 gene and it’s Biological Functions

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

cancer progression and metastasis (Becker et al., 1994; Berx et al., 1995; Christofori and Semb, 1999).

1.3.1.1 Cytoskeleton

The primary function of CDH1 is cell-cell adhesion forming the adherens junction, maintaining the cytoskeleton and mediating molecular signals between cells. The function of the adherens complex is particularly important for cell structure and regulation because it interacts with cytoskeletal actins and microtubules. The cytoskeletal role of E-cadherin maintains healthy cellular viability and growth in epithelial tissues including cellular polarity. E-cadherin is not essential to cellular viability but loss in epithelial cells does lead to defects in cytoskeletal structure and proliferation. In addition to a central role in the adherens complex, E-cadherin is involved in many other cellular functions and thus CDH1 is regarded as a highly pleiotropic gene.

1.3.1.2 Extracellular and Tumour Micro-Environment

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

1.3.1.3 Cell-Cell Adhesion and Signalling

The signals mediated by tight junctions are also passed on to intracellular signalling pathways and thus E-cadherin also has a role in maintaining cellular function and growth. One such example is the regulation of β-catenin which interacts with both the actin

cytoskeleton and acts as a transcription factor via the WNT pathway. Similarly, the Hippo and PI3K/AKT pathways are implicated in being mediated by E-cadherin, having roles in promoting cell survival, proliferation, and repressing apoptosis. E-cadherin shares several downstream pathways with signalling pathways such as integrins and thus indirectly interacts with them, particularly since feedback loops may occur in such pathways. Conversely, the multifaceted roles of E-cadherin have been shown with differing overexpression in ovarian cells promoting tumour growth, while it maintains healthy cellular functions in other cells.

1.3.2 CDH1 as a Tumour (and Invasion) Suppressor

E-cadherin has key roles in maintaining cellular structure and regulating growth, consistent with CDH1 being a tumour suppressor gene. Loss of CDH1 in epithelial tissues leads to disrupted cell polarity, differentiation, and migration. E-cadherin loss has been identified as a recurrent driver tumour suppressor mutation in sporadic cancers of many tissues including breast, stomach, lung, colon, and pancreas tissue.

1.3.2.1 Breast Cancers and Invasion

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

1.3.3 Hereditary Diffuse Gastric Cancer and Lobular Breast Cancer

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

HDGC affects less than 1 in a million people globally (Ferlay et al., 2015) and less than 1% of gastric cancers. However, HDGC is documented to affect several hundred families globally. E-cadherin mutations in the germline is implicated in 1-3% of gastric cancers presenting with a family history, varing between high and low incidence populations. E-cadherin is also mutated in 13% of sporadic gastric cancers.

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

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

1.3.4 Somatic Mutations

1.3.4.1 Mutation Rate

Estimates for the prevalence of CDH1 somatic mutations in sporadic cancers varies. The Cancer Gene Census (Futreal et al., 2004; Pleasance et al., 2010) detected 994 distinct mutations in 10,143 tumour samples (at a rate of 7.52%), COSMIC (2016) detected 632 distinct mutations in 43,865 tumour samples (at a rate of 1.71%), and detected mutations in 13.2% of 53 of the NCI60 cancer cell lines. While there is no consensus on the prevalence of CDH1 mutations, the vast variability of mutations is consistent with it’s role as a tumour supressor and it has been found to be recurrently mutated in a wide range of cancers of epithelial tissues.

COSMIC (2016) reports CDH1 mutations in 40 cancer tissue types including stomach (11.40% in 1342 samples), breast (10.29% in 3343 samples), large colon (2.87%), skin (2.83%), endometrial (2.81%), and bladder (1.9%) cancer. ICGC (2017) reports CDH1 mutations in 29 cancer tissue types including skin (23.41% in 598 samples), breast (14.50% in 1696 samples), ovary (13.98% in 93 samples), and stomach (11.41% in 289 samples) cancer samples. CDH1 mutations are reported at similar rates in breast and stomach cancer in other cancer genomics projects and studies across distinct populations. cBioPortal (2017) reports CDH1 mutation prevalence in stomach cancer at 16.7% (Kakiuchi et al., 2014, 30 samples), 15% (Wang et al., 2014, 100 samples), 14.1% (Chen et al., 2015, 78 samples), and 9.4% (TCGA, 2017b, 393 samples). cBioPortal (2017) also reports CDH1 mutation prevalence in breast cancer at 12.7% (TCGA, 2017b, 963 samples) and 10.8% (Curtis et al., 2012; Pereira et al., 2016, 2051 samples). The rare plasmacytoid bladder cancer subtype also has a high prevalence of CDH1

mutations in COSMIC (2016) at a rate of 81.8% (N=33). These demonstrate that CDH1 is important in many cancers and targeting CDH1 may be widely applied against sporadic cancers in addition to hereditary cancers. However, some of these studies have focused on disease subgroups (such as lobular subtype or estogren receptor negative breast cancers) with poor patient outcomes which may have inflated the prevalence of CDH1 mutations which are more common in some of these subtypes.

1.3.4.2 Co-occurring Mutations

Another concern is that CDH1 mutations may co-occur with other known cancer driver genes such as highly prevalent tumour suppressor gene TP53 or the proto-oncogene PIK3CA. cBioPortal (2017) reports the prevalence of the mutations in these genes at 10% for CDH1, 49% for TP53, 22% for PIK3CA in stomach cancer (TCGA, 2017b, 393 samples). There is no evidence of significant co-occurring mutations between CDH1 and PIK3CA (mutex p = 0.231) but there is evidence for significant mutually exclusive mutations for CDH1 (mutex p = 0.002) and PIK3CA (mutex p = 0.004) with TP53. cBioPortal (2017) also reports the prevalence of the mutations in these genes at 13% for CDH1, 32% for TP53, 36% for PIK3CA in breast cancer (TCGA, 2017b, 963 samples). There is evidence of significant co-occurring mutations with CDH1 and PIK3CA (mutex p < 0.0001) and evidence for significant mutually exclusive mutations for CDH1 (mutex p = 0.003) and PIK3CA (mutex p = 0.032) with TP53.

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

1.3.5 Models of CDH1 loss in cell lines

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

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

1.4 Summary and Research Direction of Thesis

Genomics technologies and the data available from them have immense potential for understanding of genetics and improving healthcare, including identification of genes altered in cancer for molecular diagnosis, prognostic biomarkers, and therapeutic targets. This has been demonstrated with the identification of cancer genes in many cancers, distinguishing tumour subtypes by expression profiles, and the development of targeted therapies against oncogenes (such as BRAF and tumour suppressors (such as BRCA1). Synthetic lethality is an important genetic interaction to study fundamental cellular functions and exploit them for biomarkers and cancer treatment. They present a means to target loss of function mutations and genetic dysregulation in tumour suppressor genes by identifying interacting partners with redundant or compensating molecular functions.

CDH1 (encoding E-cadherin) is an example of a tumour suppressor gene implicated in sporadic breast and stomach cancers. Germline mutations in CDH1 are also found in many patients with familial early onset cancers (HDGC). Discovery of synthetic lethal partners

would be contribute to an understanding on the molecular mechanisms driving the growth of CDH1 deficient tumours and identification of potential therapeutic targets or chemopreventative agents for management of HDGC. The clinical potential of the synthetic lethal approach has been demonstrated with the application of olaparib against BRCA1 and BRCA2 mutations Lord et al. (2015) but there remains the need to systematically identify synthetic lethal partner genes for other tumour suppressors such as CDH1. A synthetic lethal screen has been conducted on breast cell lines Telford et al. (2015) but computational approaches to identification of synthetic lethal partners of CDH1 remains to be done.

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

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

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

presented to investigate the relationships between synthetic lethal candidates. Release of R codes used for simulation, prediction, and analysis will enable adoption of the methodology in the cancer research community and comparison to existing methods.

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

Developed a query-based synthetic lethal detection methodology (SLIPT) for use on gene expression data

Adapt this methodology to utilise somatic mutation for query genes or candidate pathway metagenes

Apply Synthetic lethal prediction to public breast cancer genomics data from TCGA (TCGA, 2012)

Identify over-represented biological pathways using Reactome (Croft et al., 2014) among synthetic lethal candidate partner genes

Compare these at the gene and pathway level to experimental screen data in breast cell lines from Telford et al. (2015)

Replicate these analyses in stomach cancer genomics data from TCGA (Bass et al., 2014)

Determine whether synthetic lethal candidates have importance in biological networks of candidate partner pathways

Determine whether there are relationships within biological network structures between experimental and predicted gene candidate partners

Develop a statistical model of synthetic lethal gene expression

Simulate gene expression with synthetic lethal genes and pathway structures

Evaluate the effects of modification to the SLIPT procedure on it’s statistical performance

Compare the statistical performance of the SLIPT procedure to alternative statistical methods

Release a synthetic lethal prediction methodology (SLIPT) to the research community for wider application

Thesis Aims

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

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

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

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

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

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

Chapter 2Methods and Resources

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

2.1 Bioinformatics Resources for Genomics Research

2.1.1 Public Data and Software Packages

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

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

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

2.1.1.1 Cancer Genome Atlas Data

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

Performing a genomic alignment in remains a challenge in bioinformatics and methods to do so may yet be improved (Chen and Tompa, 2010). However, the statistical and biological aspects of bioinformatics are the focus of this thesis, comparing alignment methods is outside the scope of these investigations. The TCGA project (TCGA, 2012) used widely adopted tools: “Bowtie” for alignment (Langmead et al., 2009), “mapslice” to detect splice sites (Wang et al., 2010), and the Reads Per Kilobase per Million mapped reads (RPKM) approach to qualify reads per transcript as a measure of gene expression (Mortazavi et al., 2008). These are widely acceptable tools for processing RNA-Seq data which were used to produce the raw counts of mapped reads (tier 1) and normalised expression data (tier 3) publicly available from TCGA.

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

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

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

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

2.1.1.2 Reactome and Annotation Data

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

2.2 Data Handling

2.2.1 Normalisation

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

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

2.2.2 Sample Triage

The TCGA breast RNA-Seq data were assessed for batch effects using a correlation matrix of the log-transformed raw counts for which a heatmap (Euclidean distance, complete linkage) is shown in Figure A.2. While no major batch effects were detectable between the samples, 9 samples were excluded due to poor correlation with the remaining samples, as detailed in Table 2.1. These samples showed unusual density plots compared to the rest of the dataset, and exhibited low mean read count in Figures 2.1 and 2.2. A heatmap showing key clinical properties of these excluded samples and their correlation with the remainder of the samples is shown in Figure A.1, and a full correlation heatmap (Figure A.2) shows these samples as relatively poorly correlated outliers in the bottom rows and left columns. In addition to the clustering analysis (in Appendix A.1), replicate tumour samples were also examined for sample quality in Appendix A.2. After removal of these samples, the TCGA dataset used for analysis consisted of the remaining 1168 samples (from 1040 patients): 1049 tumour samples, 112 normal tissue for matched samples, and 7 metastases.

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

(a) Raw counts (log-scale)

(b) Voom normalised

Figure 2.1: Read count density. Sample density plots of raw counts on log-scale and voom normalised showing samples removed due to quality concerns.

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

(a) Mean raw counts (log-scale)

(b) Mean voom normalised

Figure 2.2: Read count sample mean. Boxplots of sample means for raw counts on log-scale and voom normalised show removed tumour samples with low mean read count.

Table 2.1: Excluded Samples by Batch and Clinical Characteristics.

Tissue Source

Type

Batch

Plate

Patient

Samples

p53

</td><td style="white-space:nowrap; text-align:center;" id="TBL-30-1-9"

class="td11">

Subtype

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

class="cmbx-12">Treatment (History)

Clinical Subtypes (Stage)

A7 Christiana

Tumour

47

A227

A0DB

1 of 3

NA

Luminal A

Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-2-10"

class="td11"> (no)

ER+

Ductal

* A7 Christiana
* Tumour
* 96
* A220
* A13D
* 1 of 3
* Wildtype
* Luminal A
* Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-3-10"
* class="td11"> (no)
* ER+
* Ductal
  + A7 Christiana
  + Tumour
  + 96
  + A227
  + A13E
  + 1 of 3
  + NA
  + Basal
  + Lumpectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-4-10"
  + class="td11"> (no)
  + ER-
  + Ductal
    - A7 Christiana
    - Tumour
    - 142
    - A277
    - A26E
    - 1 of 3
    - NA
    - Basal
    - Lumpectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-5-10"
    - class="td11"> (no)
    - ER+
    - Ductal
      * A7 Christiana
      * Tumour
      * 47
      * A277
      * A0DC
      * 1 of 2
      * NA
      * Luminal A
      * Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-6-10"
      * class="td11"> (yes)
      * ER+
      * Lobular
        + A7 Christiana
        + Tumour
        + 142
        + A220
        + A26I
        + 1 of 2
        + Mutant
        + Basal
        + Lumpectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-7-10"
        + class="td11"> (yes)
        + ER-
        + Ductal

AC Intl Genomics

Tumour

177

A18M

A2QH

2 of 2

Mutant

Basal

Radical Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-8-10"

class="td11"> (no)

ER-

Metaplastic

AC Intl Genomics

Tumour

177

A220

A2QH

2 of 2

Mutant

Basal

Radical Mastectomy </td><td style="white-space:nowrap; text-align:center;" id="TBL-30-9-10"

class="td11"> (no)

ER-

Metaplastic

GI ABS IUPUI

Normal

177

A16F

A2C8

1 of 1

NA

Luminal A

Radical Mastectomy and Neoadjuvant

(no)

ER+

Ductal

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

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

2.2.3 Metagenes and the Singular Value Decomposition

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

2.2.3.1 Candidate Triage and Integration with Screen Data

Candidate triage in combination with the experimental data was intended to integrate findings of the SLIPT analysis with an ongoing experiment project (Chen et al., 2014; Telford et al., 2015). The first procedure to compare the SLIPT gene candidates for CDH1 with an siRNA experimental screen (Telford et al., 2015) was a direct comparison of the overlapping candidates, presented in a Venn diagram and tested with the χ2 test. Since these candidates modestly overlapped at the gene level (even when excluding genes not contained in both datasets), further gene set over-representation analysis was

performed for pathways specific to each detection approach and the intersection of the two.

The pathway composition of the intersection was further verified by a permutation resampling analysis (as described in Section 2.3.6): the same number of genes detected by SLIPT were sampled randomly from the universe of genes tested by both approaches. These samplings were performed over 1 million iterations and the pathway over-representation was compared for each of the 1,652 reactome pathways. These over-representation scores (χ2) were compared the observed over-representation in the intersection of the SLIPT candidates, with the proportion of resamplings with higher χ2 values used for empirical p-values of pathway composition. The χ2 test was used as an appropriation of Fisher’s exact test on a hypergeometric distribution for resampling to computationally scale pathway over-representation tests across iterations. Pathways for which no resamplings were occurred as high as the observed were reported as p < 10-6. These empirical p-values were adjusted for multiple comparisons (FDR). Intersection size was not assumed to be constant across resamplings so similarly with the proportion of resamplings with higher or lower intersection size were used to evaluate significance of enrichment or depletion respectively (of siRNA candidate among SLIPT candidate genes).

2.3 Techniques

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

2.3.1 Statistical Procedures and Tests

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

This generates data where each row (gene) has a mean of 0 and standard deviation of 1. Where plotted as aa heatmap, any data more than 3 standard deviations above or below the mean is plotted as 3 or -3 respectively.

Empirical Bayes differential expression analysis was performed using the limma R package (Ritchie et al., 2015). Where specified, the Fisher’s exact test, χ2 test, and correlation were used to measure associations between variables (as implemented in the stats R package (R Core Team, 2016)). Unless otherwise specified, Pearson’s correlation was used for correlation analyses (r) and coefficient of determination (R2). Where these comparisons are discussed in more detail, Fisher’s exact test and χ2 tests are supported by a table or Venn diagram, rendered with the limma R package (Ritchie et al., 2015). In some analyses, correlation is furter supported by a scatter plot and a line of best fit dervied by least squares linear regression.

The t.test function (R Core Team, 2016) has also been used to implement the t-test to compare pairs of data. Where relevant, an ANOVA has been performed to report significance of multivariate predictors of outcomes, or least squares linear regression performed for the adjusted coefficient of determination (R2) and F-statistic p-value to evaluate the fit of the predictor variables. For some analyses these are supported by boxplot or violinplot visualisation (rendered in R).

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

2.3.2 Gene Set Over-representation Analysis

Gene set enrichment over-representation was performed to test whether there is an enrichment of a gene set (such as a biological pathway) among a group of input genes. Such input genes may be predicted synthetic lethal candidates or a subset defined by clustering (in Section 2.3.3) or comparison with experimental candidates (see Section 2.2.3.1). Initially, these tests were performed using the GeneSetDB web tool (Araki et al., 2012) hosted by the University of Auckland on the Reactome pathways (Croft et al., 2014). Since the GeneSetDB tool used an older version of Reactome (version 40), it was difficult to directly compare with the results of other analysis (see sections 2.2.3.1 and 2.3.6) performed on version 52 (as described in Section 2.1.1.2). Thus an implementation of the hypergeometric test in R (R Core Team, 2016) was used to test for over-representation against Reactome (version 52) pathways. Pathways containing less than 10 genes or more than 500 (as performed in GeneSetDB by Araki et al., 2012) were excluded before adjusting for multiple comparisons.

2.3.3 Clustering

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

2.3.4 Heatmap

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

as heatmap.2x). Where clearly specified, data have been split into subsets with clustering performed separately on each subset with these plotted alongside each other.

2.3.5 Modeling and Simulations

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

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

Several measures of statistical performance were used to assess the simulations. The following measures used the final classification of the detection procedure, statistical significance for χ2, significance and directional criteria met for SLIPT (see Section 3.1), and an arbitrary threshold: < -0.2 and > +0.2 for negative correlation and correlation respectively. Sensitivity (or “true positive rate”) was measured as the proportion of known synthetic lethal partners predicted to be synthetic lethal. Specificity (or “true negative rate”) was measured as the proportion of known non-synthetic lethal partners predicted not to be synthetic lethal. The “false discovery rate” (also used in adjusting for multiple comparisons) was measured here as the proportion of known non-synthetic lethal partners out of all putative partners

predicted by the detection procedure. Statistical “accuracy” is the proportion of true predictions for a detection procedure, which is both the correctly predicted known synthetic lethal partners and correctly negative known non-synthetic lethal partners.

2.3.5.1 Receiver Operating Characteristic (Performance)

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

2.3.6 Resampling Analysis

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

assessing the significance of pathway structure among synthetic lethal candidates (Section 3.4.1.1).

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

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

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

2.4 Pathway Structure Methods

2.4.1 Network and Graph Analysis

Networks are important in considering the structure of relationships in molecular biology, including gene regulation, kinase cellular signaling, and metabolic pathways (Barabási and Oltvai, 2004). Network theory is an interdisciplinary field which combines the approaches of computer science with the metrics and fundamental principles of graph theory, an area of pure mathematics dealing with relationships between sets of discrete elements. The vast amounts of molecular and cellular data from high-throughput technologies have enabled the application of network-based and genome-wide bioinformatics analysis to examine the complexity of a cell at the molecular level and understand

aberrations in cancer. This thesis uses various metrics and analysis procedures developed in Graph and Network theory to analyse graph structure of biological pathways. Where feasible, these have been implemented using the igraph R package with such procedures described below (Csardi and Nepusz, 2006). Custom R functions to perform more complex analysis and visualisation of iGraph data objects will be described later.

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

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

2.4.2 Sourcing Graph Structure Data

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

2.4.3 Constructing Pathway Subgraphs

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

2.4.4 Network Analysis Metrics

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

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

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

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

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

where G′ is the subgraph with the node removed and E is the efficiency (Latora and Marchiori, 2001) derived from shortest paths (dij between nodes i and j) as:

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

2.5 Implementation

2.5.1 Computational Resources and Linux Utilities

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

Table 2.2: Computers used during Thesis

<div class="multicolumn" style="white-space:nowrap; text-align:right;"></div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-31-1-2"

class="td11">

Viao Laptop

</td><td style="white-space:nowrap; text-align:left;" id="TBL-31-1-3"

class="td11">

Lab Machine

</td><td style="white-space:nowrap; text-align:left;" id="TBL-31-1-4"

class="td11">

Biochem Server

</td><td style="white-space:nowrap; text-align:left;" id="TBL-31-1-5"

class="td11">

NeSI Pan Cluster

Operating System (OS)

Elementary OS

Freya 0.3.2

Elementary OS

Loki 0.4

Red Hat Enterprise

Maipo 7.2

Cent OS

Final 6.4

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Upstream OS</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-31-4-2"

class="td11">

Ubuntu LTS

Trusty 14.04

Ubuntu LTS

Xenial 16.04

</td><td style="white-space:nowrap; text-align:left;" id="TBL-31-4-5"

class="td11">

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Linux Kernel</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-31-5-2"

class="td11"> 3.19.0-65-generic

4.4.0-36-generic

3.10.0-327.36.2.el7.x86\_64

2.6.32-504.16.2.el6.x86\_64

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Shell: bash</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-31-6-2"

class="td11"> 4.3.11(1)

4.3.46(1)

4.2.46(1)

4.2.1(1)

<div class="multicolumn" style="white-space:nowrap; text-align:right;">Shell: zsh</div> </td><td style="white-space:nowrap; text-align:left;" id="TBL-31-7-2"

class="td11"> 5.0.2

5.1.1

5.0.2

5.2

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

Table 2.3: Linux Utilities and Applications used during Thesis

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-1-2"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-1-3"

class="td11">

Viao Laptop

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-1-4"

class="td11">

Lab Machine

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-1-5"

class="td11">

Biochem Server

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-1-6"

class="td11">

NeSI Pan Cluster

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-3-2"

class="td11">

OS

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-3-3"

class="td11">

Elementary OS

Freya 0.3.2

Elementary OS

Loki 0.4

Red Hat Enterprise

Maipo 7.2

Cent OS

Final 6.4

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-4-2"

class="td11">

Linux Kernel

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-4-3"

class="td11"> 3.19.0-65-generic

4.4.0-36-generic

3.10.0-327.36.2.el7.x86\_64

2.6.32-504.16.2.el6.x86\_64

Scripting

Shell bash

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-5-3"

class="td11"> 4.3.11(1)

4.3.46(1)

4.2.46(1)

4.2.1(1)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-6-2"

class="td11">

Shell zsh

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-6-3"

class="td11"> 5.0.2

5.1.1

5.0.2

5.2

Programming

Python

2.7.6

2.7.12

2.7.5

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-8-2"

class="td11">

Java

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-8-3"

class="td11"> 1.8.0\_101

9-ea

1.8.0\_101

</td>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-9-2"

class="td11">

C++

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-9-3"

class="td11"> 4.8.4

5.4.0

4.8.5

4.4.7

Text Editor

nano

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-10-3"

class="td11"> 2.2.6

2.5.3

2.3.1

2.0.9

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-11-2"

class="td11">

kile (LATEX)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-11-3"

class="td11"> 2.1.3

2.1.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-11-6"

class="td11">

Version Control

git

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-12-3"

class="td11"> 1.9.1

2.11.0

1.7.1

1.8.3.1

Shell Utilities

sed

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-13-3"

class="td11"> 4.4.2

4.4.2

4.4.2

4.4.1

grep

2.16-1

2.25-1

2.20

2.6.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-15-2"

class="td11">

nohup

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-15-3"

class="td11"> 8.21

8.25

8.22

8.4

Typesetting

TEX

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-16-3"

class="td11"> 3.1415926

3.14159265

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-16-6"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-17-2"

class="td11">

TexLive (LATEX)

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-17-3"

class="td11"> 2013

2015

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-17-6"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-18-2"

class="td11">

PDFTEX

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-18-3"

class="td11"> 2.5-1

2.6

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-18-6"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-19-2"

class="td11">

pandoc

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-19-3"

class="td11"> 1.12.2.1

1.16.0.2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-19-6"

class="td11">

Remote Computing

slurm scheduler

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-20-3"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-20-5"

class="td11">

16.05.6

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-21-2"

class="td11">

OpenSSH

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-21-3"

class="td11"> 7.2p2

7.2p2

6.6.1

5.3p1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-22-2"

class="td11">

OpenSSL

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-22-3"

class="td11"> 1.0.2g

1.0.2g

1.0.01e-fips

1.0.01e-fips

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-23-2"

class="td11">

rsync

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-23-3"

class="td11"> 3.1.0p31

3.1.1p31

3.0.9p30

</td>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-24-2"

class="td11">

Globus Online Transfer

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-24-4"

class="td11">

3.1

3.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-25-2"

class="td11">

Cisco AnyConnect VPN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-25-4"

class="td11"> 3.1.05170

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-25-6"

class="td11">

Image Processing

Inkscape

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-26-3"

class="td11"> 0.48.4

0.91

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-26-6"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-27-2"

class="td11">

GIMP

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-27-3"

class="td11"> 2.8.10

2.8.16

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-27-6"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-28-2"

class="td11">

ImageMagick

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-28-3"

class="td11"> 6.7.7.10-6

</td><td style="white-space:nowrap; text-align:left;" id="TBL-38-28-5"

class="td11">

</td>

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

2.5.2 R Language and Packages

The R programming language has been used for the majority of this thesis. Current R installations across the machines used are given in Table 2.4. Local machines currently run the latest version of the R (at the time of writing) and remote machines run the versions and modules as managed by the system administrator.

Various scripts and packages in this thesis were developed or run in previous versions of RStudio and R but these run without error in the current version of R (and the older versions on remote machines). The R packages which were used throughout this thesis (as detailed in Table 2.5 with versions specified) were installed from the Comprehensive R Archive Network (CRAN, 2017), Bioconductor (Gentleman et al., 2004, version 3.4; BiocInstaller 1.24.0), or GitHub. These packages were not updated when they would change the functionality of scripts or functions in packages, in particular imported data from annotation packages (used to define gene sets) have been saved as local files to continue using stable versions of these pathway data (across machines).

This is a summary of the key packages which (in addition to their dependencies) have been used throughout this project. Where a package implementation has been central to the methods applied, they are described in more detail in the relevant section. A full table of packages used in this thesis can be found in Appendix B (Table B.1). The R packages developed during this thesis are given in Table 2.6 with the relevant sections describing their implementation and use where appropriate, in addition to further details on these functions in Section 3.5.

Table 2.4: R Installations used during Thesis

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-1-2"

class="td11">

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-1-3"

class="td11">

Viao Laptop

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-1-4"

class="td11">

Lab Machine

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-1-5"

class="td11">

Biochem Server

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-1-6"

class="td11">

NeSI Pan Cluster

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-3-2"

class="td11">

OS

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-3-3"

class="td11">

Elementary OS

Freya 0.3.2

Elementary OS

Loki 0.4

Red Hat Enterprise

Maipo 7.2

Cent OS

Final 6.4

Programming

R

</td><td style="white-space:nowrap; text-align:left;" id="TBL-43-4-3"

class="td11"> 3.3.2

3.3.2

3.3.1

3.3.0-intel (module)

Development

RStudio

1.0.136

1.0.136

1.0.136 (server)

</td>

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

Table 2.5: R Packages used during Thesis

Package

Version Used

Built

Repository

colorspace

1.3-2

3.3.1

CRAN

curl

2.3

3.3.1

CRAN

data.table

1.9.6

3.3.1

CRAN

dendextend

1.4.0

3.3.2

CRAN

DBI

0.5-1

3.3.1

CRAN

devtools

1.12.0

3.3.1

CRAN

dplyr

0.5.0

3.3.1

CRAN

ggplot2

2.2.1

3.3.1

CRAN

git2r

0.18.0

3.3.1

CRAN

gplots

3.0.1

3.3.1

CRAN

gtools

3.5.0

3.3.1

CRAN

igraph

1.0.1

3.3.1

CRAN

matrixcalc

1.0-3

3.3.1

CRAN

mclust

5.2.2

3.3.1

CRAN

mvtnorm

1.0-6

3.3.1

CRAN

org.Hs.eg.db

3.1.2

3.1.2

Bioconductor

openssl

0.9.6

3.3.1

CRAN

plyr

1.8.4

3.3.1

CRAN

purrr

0.2.2

3.3.1

CRAN

reactome.db

1.52.1

3.2.1

Bioconductor

RColorBrewer

1.1-2

3.3.1

CRAN

Rcpp

0.12.9

3.3.1

CRAN

ROCR

1.0-7

3.3.1

CRAN

roxygen2

6.0.1

3.3.2

CRAN

shiny

1.0.0

3.3.1

CRAN

snow

0.4-2

3.3.1

CRAN

testthat

1.0.2

3.3.2

CRAN

tidyr

0.6.1

3.3.2

CRAN

tidyverse

1.1.1

3.3.2

GitHub (hadley)

sm

2.2-5.4

3.3.1

CRAN

Unicode

9.0.0-1

3.3.2

CRAN

vioplot

0.2

3.3.1

CRAN

viridis

0.3.4

3.3.2

CRAN

xml2

1.1.1

3.3.2

CRAN

xtable

1.8-2

3.3.1

CRAN

zoo

1.7-14

3.3.1

CRAN

graphics

3.3.2

3.3.2

base

grDevices

3.3.2

3.3.2

base

cluster

2.0.5

3.3.1

base

graphics

3.3.2

3.3.2

base

grDevices

3.3.2

3.3.2

base

Matrix

1.2-8

3.3.1

base

stats

3.3.2

3.3.2

base

Table 2.6: R Packages Developed during Thesis

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-1-2"

class="td11">

Package Name

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-1-3"

class="td11">

Description and GitHub Repository

</td><td style="white-space:nowrap; text-align:center;" id="TBL-49-1-4"

class="td11">

Section

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-3-2"

class="td11">

slipt

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-3-3"

class="td11">

Synthetic lethal detection by SLIPT (to accompany publication)

https://github.com/TomKellyGenetics/slipt

3.1

visualisation

vioplotx

Customised violin plots (based on vioplot)

https://github.com/TomKellyGenetics/vioplotx

</td>

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-5-2"

class="td11"> heatmap.2x

Customised heatmaps (based on gplots)

https://github.com/TomKellyGenetics/heatmap.2x

2.3.4

igraph.extensions

igraph.extensions

Meta-package to install the follow iGraph functions

https://github.com/TomKellyGenetics/igraph.extensions

3.5.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-7-2"

class="td11"> plot.igraph

Custom plotting of directed graphs

https://github.com/TomKellyGenetics/plot.igraph

2.4.4

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-8-2"

class="td11"> info.centrality

Computing information centrality from network efficiency

https://github.com/TomKellyGenetics/info.centrality

3.4.2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-9-2"

class="td11"> pathway.structure.permutation

Testing pathway structure with resampling analysis

https://github.com/TomKellyGenetics/pathway.structure.permutation

3.4.1.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-49-10-2"

class="td11"> graphsim

Generating simulated expression from graph structures

https://github.com/TomKellyGenetics/graphsim

3.4.2

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

2.5.3 High Performance and Parallel Computing

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

The New Zealand eScience Infrastructure (NeSI) is a computating resource providing the Intel Pan cluster hosted by the University of Auckland (NeSI, 2017). The Pan cluster used throughout this thesis project to optimise and perform computations which would have otherwise been infeasible in the timeframe of thesis. Such technological developments and infrastructure initiatives have enabled bioinformatics research including this project. High performance computing on the Pan cluster was used extensively in this project including for resampling analysis (in sections 2.3.6 and 3.4.1.1), calculating information centrality (in Section 2.4.4), and in simulations (in sections 2.3.5, 3.2, and 3.4.2)

Scripts and data were transferred between the Pan cluster and University of Otago computing resources by rsync or the Globus file transfer service (Globus, 2017). R scripts (R Core Team, 2016) were run in parallel with the “simple network of workstations” snow R package Tierney et al. (2015). This utilised the “message passing interface” (Yu, 2002) when it was feasible with memory requirements to run in parallel across multiple compute nodes, otherwise SOCKS was used to access multiple cores within an instance of R and pass input data to them. R jobs were submitted to queue for available resources and run on the Pan cluster via the Slurm (Simple Linux Utility for Resource Management) workload manager (Slurm, 2017). When running R scripts across many parameters or for memory-intensive jobs, Slurm array job submission and independent submission of different parameters via shell commands with arguments passed to R. In some cases, this submission was automated across a range of parameters with Bash scripts.

Chapter 3Methods Developed During Thesis

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

3.1 A Synthetic Lethal Detection Methodology

The SLIPT methodology identifies gene expression patterns consistent with synthetic lethal interactions between a query gene and a panel of candidate interacting partners. Gene expression is called low, medium, or high by separating samples into tertiles (3-quantiles) for each gene. Genes with insufficient expression across all samples were excluded by requiring that the first tertile of raw counts is above zero. Then a χ2 test is performed between the query gene and each candidate partner, with the p-values for the χ2 test being corrected for multiple testing using false discovery rate (FDR) error control to reduce false positives for large candidate gene panels (Benjamini and Hochberg, 1995). Significance was called only if FDR adjusted p-values were below the threshold p < 0.05. A synthetic lethal interaction is predicted (as

shown in Figure 3.1) when (i) the χ2 test is significant; (ii) observed low-query, low-candidate samples are less frequent than expected; and (iii) observed low-query, high-candidate and high-query, low-candidate samples are more frequent than expected.

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

Figure 3.1: Framework for synthetic lethal prediction. SLIPT was designed to identify candidate interacting genes from gene expression data using the χ2 test against a query gene. Samples are sorted into low, medium, and high expression quantiles for each gene to test for a directional shift. A sample being low in both genes of a synthetic lethal pair is unlikely, since loss of both genes will be deleterious, and is expected to be statistically under-represented in a gene expression dataset. We expect a corresponding (symmetric) increase in frequency of sample with low-high gene pairs. Synthetic lethal candidate (exprSL) partners of a gene are identified by running this procedure on all possible partner genes, selecting those with an FDR-adjusted χ2 p-value of p < 0.05, and meeting the directional criteria. Since synthetic lethal genes are partners of each other commutatively, the symmetric direction criteria are all required such that synthetic lethal genes will predicted to be partners of each other.

The synthetic lethal prediction procedure has also been adapted to utilise somatic mutation data for the query gene. This is intended to utilise a query gene known to be recurrently mutated in the disease (and dataset), with the majority of mutations inactivating gene function (such as null or frameshift mutations). A synthetic lethal interaction is predicted (as shown in Figure 3.2) when (i) the χ2 test is significant; (ii) observed mutant-query, low-candidate samples are less frequent than expected; and (iii) observed mutant-query, high-candidate and wild-type-query, low-candidate samples are more frequent than expected. Unless otherwise specified, computationally predicted synthetic lethal gene candidates from SLIPT used expression data (exprSL) for both genes (as shown in Figure 3.1) rather than mutation data (mtSL) for the query gene (as shown in Figure 3.2).

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

Figure 3.2: Synthetic lethal prediction adapted for mutation. SLIPT was also adapted to identify candidate interacting genes using (somatic) mutation data of the query gene in the χ2 test. Samples are sorted into low, medium, and high expression quantiles for each candidate gene and tested for a directional shift against mutation status of the query gene. A sample having low expression or mutation for the synthetic lethal pair is expected to be unlikely with a corresponding increase in frequency of sample with mutant-high or wild-type-low gene pairs. Synthetic lethal candidate (mtSL) partners of a gene are identified by running this procedure on all possible partner genes, selecting those with an FDR-adjusted χ2 p-value of p < 0.05, and meeting the directional criteria.

The SLIPT methodology is amenable for use on expression data including pathway metagenes (as generated in Section 2.2.3). The suitability of the SLIPT methodology to application on public gene expression data will further be supported by simulation results in Section 3.3 and Chapter 6, including comparison to other statistical methods. SLIPT results for CDH1 will also compared experimental screen results in a breast cell line (Telford et al., 2015), primary screen results are discussed in Section 4.2 and secondary (validation) screen results are presented in Appendix C.

3.2 Synthetic Lethal Simulation and Modelling

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

3.2.1 A Model of Synthetic Lethality in Expression Data

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

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

Figure 3.3: A model of synthetic lethal gene expresion. A conceptual model of synthetic lethal interactions between a Query gene and partner gene (GX). Genes that are synthetic lethal may not both be non-functional in the same sample without another gene compensating for the loss of function. This is most likely to be detectable as low gene expression, whether they are lost by mutation, deletion, DNA methylation, or suppressing regulatory signals. This could manifest as coexpression, mutual exclusivity, or directional shifts in sample frequency. Thus the alternative hypothesis (HA) is that synthetic lethal genes will have a reduced frequency of co-loss samples while the null hypothesis (H0) is that non-synthetic lethal gene pairs would show no such relationship, even if they may be correlated for other means such as pathway relationships. In this model synthetic lethal genes may compensate for the loss of each other but this is not assumed, only that loss of both is unfavourable to cell viability and probability of detecting samples with combined gene loss.

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

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

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

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

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

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

A synthetic lethal pair of genes is unlikely to act in isolation, therefore higher-order synthetic lethal interactions (i.e., 3 or more genes) must be considered in the model as shown in Figure 3.5. Even when testing pairwise interactions, modelling higher level interactions that may interfere is important. If there are additional synthetic lethal partners, there are two possibilities for adding these: 1) that they are independent partners of the query genes interacting pairwise (and not with each other) or 2) that an addition partner gene interacts with both of the synthetic lethal genes already in the system and any of the three (or more) are required to be functional for the cell to survive.

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

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

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

3.2.2 Simulation Procedure

Simulations were developed to simulate normal distributions of expression data and define function with a threshold cut-off. This is the reverse to the procedure of SLIPT to predict

synthetic lethal partners (although the threshold is assumed to be unknown when testing upon simulated data). While gene function is used as an intermediary step in modelling synthetic lethal genes in expression data, the normal distribution is sampled for simulated data to represent normalised empirical gene expression data for which SLIPT (and other methods) will be applicable.

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

(a) Simulated expression matrix

(b) Corresponding gene function calls

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

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

(a) Simulated gene function with SL genes

(b) Query gene added with SL condition

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

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

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

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

(a) Initial expression matrix

(b) Simulated synthetic lethal dataset

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

This results a simulated dataset where samples with non-functional query gene have at least one functional partner gene. Similarly, the query gene is functional in all samples where all of the synthetic lethal partner genes are non-functional. Therefore a dataset has been generated with known synthetic lethal partners (see Figure 3.8) by as few assumptions about the relationships between the each synthetic lethal pair as possible (and allowing compensating functions from higher-order interactions). This has been designed to have the most stringent (least detectable) synthetic lethal relationships where higher-order interactions are possible for the purposes of testing pairwise detection procedures such as SLIPT.

3.3 Detecting Simulated Synthetic Lethal Partners

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

3.3.1 Binomial Simulation of Synthetic lethality

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

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

The simulation procedure was performed with 20,000 total genes (as feasible in the human genome and expression datasets) with a variable number of true synthetic lethal partners and sample sizes of 500, 1000, 2000, and 5000. Each ROC curve was derived from the results of 10,000 replicate simulations. The statistical performance (as shown in Figure 3.9) of such an approach based on the χ2 p-value declines towards random predictions (an AUROC of 0.5) with an increasing number of underlying true synthetic lethal partners to detect. However, increased sample size mitigates this decline to some extent, as expected with a statistical predictor, particularly for moderate numbers of synthetic lethal partners.

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

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

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

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

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

The multivariate normal simulation procedure is more representative of the (normalised) expression data SLIPT is intended for and enables the prediction procedure to be tested without changes to the methodology (presented in more detail in Section 3.3.2). Sampling continuous expression values from a normal distribution allows the expression threshold for gene function to differ from the categorical “low” and “high” expression binning performed by SLIPT (as discussed in Section 3.2.1) which represents that the SLIPT procedure does not assume a known threshold for expression but rather uses expression as an estimate of gene function. This functionality can be included in the multivariate normal simulation without compromising the statistical performance of the SLIPT, rather the performance estimates (shown in Figure 3.10) were a marked improvement over the binomial simulation procedure across simulation parameters in an equivalent simulation (without correlation structure). This imprivement may be due to binomial model defining 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 sample proportions of χ2 test.

3.3.2 Multivariate Normal Simulation of Synthetic lethality

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

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

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

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

(a) Statistical evaluation

(b) Receiver operating characteristic

(c) Statistical performance

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

As shown in Figure 3.11(a), the statistical accuracy of SLIPT as a binary classifier is considerably high across simulations of a full human dataset of 20,000 genes. However, with the χ2 p-value as a threshold for prediction, this is largely to desirable specificity: the majority of non-SL genes are distinguished from the few underlying synthetic lethal genes. In this regard, the SLIPT methodology generally performs better with larger datasets with more expected negatives and thus the results of simulations of smaller numbers of genes (such as the graph structures analysed in Chapter 6) can be applied to larger datasets where they are expected to perform comparably or better with a lower false negative rate. Accordingly, key results will be supported by replication with larger numbers of non-SL genes added to the simulations.

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

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

The receiver operating characteristic curves (in Figure 3.11(b)) demonstrate that SLIPT is subject to near equal trade-off between sensitivity and specificity across threshold values. The lower sensitivity and higher specificity with a binary classifiation (in Figure 3.11(a)) stems from stringent testing by SLIPT with (FDR) p-values adjusted for multiple tests. The area under these curves is also used to compare statistical performce (in Figure 3.11(c)), with declining performance across increased underlying synthetic lethal partners and increased performance with sample size in multivariate normal simulations.

3.3.2.1 Multivariate Normal Simulation with Correlated Genes

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

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

(a) Input Σ matrix parameter

(b) Simulated correlation matrix

(c) Simulated gene expression

(d) Simulated gene function

Figure 3.12: Simulating expression with correlated gene blocks. A Σ matrix (a) is used generate a multivariate normal distribution with of 100 genes correlated blocks of genes (correlated by 0.8) with a comparable structure (b) to the input Σ, as shown by correlation on a red–green scale. The annotation bars for genes give the χ2 (in blue if the direction of SLIPT is met or red otherwise) and the gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL). The simulated gene expression (c) and function (d) generated are ordered by χ2 showing the functional structure of synthetic lethal genes and that they are among the strongest SLIPT results.

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

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

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

(a) Input Σ matrix parameter

(b) Simulated correlation matrix

(c) Simulated gene expression

(d) Simulated gene function

Figure 3.13: Simulating expression with correlated gene blocks. Using the (a) Σ matrix, sampling from a multivariate normal distribution with of 1000 genes produced (b) correlated blocks of genes (correlated by 0.8) on a red–green scale. The simulated gene expression (c) and function (d) generated are ordered by χ2 and SLIPT direction show that synthetic lethal genes are among the strongest SLIPT results with high specficity against many potential false positives. These are annotated for χ2 (on a red–green scale) and category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) for each gene.

These simulations (on 100 genes) were repeated to examine the variation between detection on different samples and varying the number of underlying synthetic lethal partners, in simulated gene expression data with correlations structure. A small nuber (10 for each) simulations are shown in Figure 3.14 to demonstrate the variation between replicate simulations, with iterative sampling from the same multivariate normal distribution. These simulations show synthetic lethal genes are not only highly ranked by SLIPT when there are few of them but also that they are fairly consistent across replicate simulations. Whereas they become less consistent for increasing numbers of true synthetic lethal partners to detect and thus more difficult to distinguish from other geners, particularly those correlated with them. Similarly, the χ2 values show a marked stepwise increase with clear thresholds for SL and correlated genes in simple simulations, whereas these become less evident for higher numbers of SL partners.

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

(a) Gene category in simulations

(b) Corresponding χ2 values

Figure 3.14: Synthetic lethal prediction across simulations. The gene category (blue for query, cyan for query-correlated, red for SL, orange for SL-correlated, forest green for non-SL-correlated, and green for non-SL) ordered by χ2 signed by the SLIPT directional condition is shown across simulations. For each of 1–10 SL partners, 10 simulations demonstrate that the increasing numbers of SL partners become harder detect. The χ2 values show a clear threshold for SL and correlated genes when there are fewer of them, distinguishable from correlated genes in this case.

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

(a) Statistical evaluation

(b) Receiver operating characteristic

(c) Statistical performance

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

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

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

Whether the synthetic lethal genes detected in simple simulations (in Figure 3.14) are robustly detectable across greater number of simulations, in addition to further comparions, was tested with a supporting ROC analysis. These results (in Figure 3.15) are very similar to simulations without correlation structure, with SLIPT as a binary classifier having a poor sensitivity with increasing numbers of synthetic lethal partners to detect but high specificity in a total of 20,000 genes with the vast majority being true negatives. This is reflected in a similar decline in statistical performance for increasing numbers of synthetic lethal partners and a compensating increase in performance with higher sample size. Overall, the statistical performance is very similar to simulations without correlation structure (as shown in Figure 3.16).

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

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

(a) Statistical evaluation

(b) Receiver operating characteristic

(c) Statistical performance

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

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

3.3.2.2 Specificity with Query-Correlated Pathways

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

Importance of Directional Testing

It is important to notice here that the directional criteria of the SLIPT procedure is enhancing it’s performance, particularly in distinguishing positively correlated true negatives. The multivariate normal simulation results, with 20,000 genes including all of the correlation structures discussed (SL, non-SL, and query correlated genes), are compared here for SLIPT with and without (χ2) directional testing. There is a marked improvement in statistical performance with directional criteria, particularly with increased sensitivity and lower false discovery rate (as shown in Figure 3.18).

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

(a) χ2 testing without direction

(b) SLIPT with directional criteria

Figure 3.18: Statistical evaluation of directional criteria. A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT with the directional condition and and the equivalent χ2 test without. SLIPT exhibited a consistently higher sensitivity and lower false discovery rate.

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

(a) χ2 testing without direction

(b) SLIPT with directional criteria

(c) Statistical performance

Figure 3.19: Performance with directional criteria. A simulated multivariate normal dataset of 20,000 genes with correlation structures was tested by SLIPT with the directional condition and and the equivalent χ2 test without. SLIPT has higher performance across simulation parameters, clearly differing from random (grey diagonal) in ROC curves up to 100 SL genes (b). The performance (c) of SLIPT (in greyscale) was consistently higher than the χ2 test (in color).

This is encouraging for the application of SLIPT to empirical expression datasets as postively correlated genes are likely to occur and the directional condition robustly improves the performance of SLIPT across simulation parameters. Without assuming the underlying number of synthetic lethal genes, SLIPT will perform better than the χ2 test alone at detecting them. This is further supported irrespective of significance threshold for the χ2 test by the ROC analysis in Figure 3.19. The directional SLIPT methodology outperforms the ordinary χ2 test at detecting synthetic lethal partners with some predictive power (above random and AUROC of 0.5) even up to 100 synthetic lethal genes.

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

3.4 Graph Structure Methods

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

3.4.1 Upstream and Downstream Gene Detection

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

Using the Reactome version 52 data (Croft et al., 2014) as described in Section 2.4.2, genes identified by each synthetic lethal discovery approach were mapped to the graph structure for the candidate pathways identified in Chapter 4 (with subgraphs defined as described in Section 2.4.3). To test whether siRNA candidate genes were upstream of SLIPT candidate genes, shortest paths were traced between each potential pair of these genes in a directed network. The number of genes where the siRNA candidate was upstream were scored “up” and where the siRNA candidate was downstream were scored “down”. This procedure enabled counting the total number of shortest paths which supported siRNA genes being upstream or downstream of the SLIPT genes and measuring the difference between these to determine if there is an imbalance in a particular direction. While this difference is indicative of the number of paths between the gene candidate groups in either direction, alone it is not sufficient to statistically support structure or relationships between siRNA and SLIPT genes. However, it may be combined with a permutation resampling procedure (as described in Section 3.4.1.1) to test for directional relationships in either direction.

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

3.4.1.1 Permutation Analysis for Statistical Significance

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

The null hypothesis is that there is no relationship between these gene groups that would not have occurred had the genes been selected at random. Thus we can test both the

alternate hypothesis that the siRNA genes were upstream of the SLIPT genes or that they are downstream of them.

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

3.4.1.2 Hierarchy Based on Biological Context

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

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

Having identified the downstream end of the pathway, genes upstream (e.g., gene i) of

this are assigned a hierarchy by the length of their shortest path (d) to this gene, z.

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

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

This

hierarchy of pathway directionality (such as that shown in Figure 5.7) can be used for comparison with measures of the number of genes of each candidate group and the support for being synthetic lethal partners with either approach.

3.4.2 Simulating Gene Expression from Graph Structures

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

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

(a) Activating pathway structure

(b) Pathway structure including inhibitions

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

First, the graph structure is constructed for simulated data to be generated from (by sampling from a multivariate normal distribution using the mvtnorm R package (Genz and Bretz, 2009; Genz et al., 2016)). Throughout this section, the simulation procedure will be demonstrated with the relatively simple constructed graph structure shown in Figure 3.20. This graph structure visualisation was specifically developed for (directed) iGraph objects in R and has been released in the plot.igraph package and igraph.extensions library (see Table 2.6 and Section 3.5.3). The plot\_directed function allows customisation of plot parameters for each node or edge and mixed (directed) edge types for indicating activation or inhibition. These inhibition links (which often occur in biological pathways) are demonstrated in Figure 3.20(b).

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

(a) Activating pathway structure

(b) Distance matrix

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

(c) Sigma, Σ (expected correlation)

(d) Simulated correlation structure

(e) Simulated expression data

(f) Simulated gene function calls

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

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

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

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

(a) Pathway structure with inhibition

(b) Distance matrix

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

(c) Sigma, Σ (expected correlation)

(d) Simulated correlation structure

(e) Simulated expression data

(f) Simulated gene function calls

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

The simulation procedure (depicted in Figure 3.21) is amenable to pathways containing inhibition links (as shown in Figure 3.22) with several refinements. With the inhibition links (as shown in Figure 3.22(a)), distances are calculated in the same manner as before (Figure 3.22(b)) with inhibitions accounted for by iteratively multiplying downstream nodes by -1 to form blocks of negative correlations (as shown in Figures 3.22(c) and 3.22(d)). As before, a multivariate normal distribution with these negative correlations can be sampled to generate simulated data (as shown in Figures 3.22(e) and 3.22(f)).

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

3.5 Customised Functions and Packages Developed

[Move to Appendix?]

Various R packages have been developed throughout this thesis using devtools (Wickham and Chang, 2016) and roxygen (Wickham et al., 2017) to enable reproducibility of customised analysis and visualisation. Many of these have the added benefit of the functions being documented, demonstrated in example vignettes, and released on GitHub to enable the research community to access utilise them in their own analysis. These are summarised in Table 2.6, along with the corresponding urls for their GitHub repository which contains a README file with instructions for installation with the devtools R package (Wickham and Chang, 2016) and links to the relevant vignette(s) where available.

3.5.1 Synthetic Lethal Interaction Prediction Tool

The statistical methodology for detection of synthetic lethality in gene expression data (SLIPT) is one of the main novel procedures developed in this thesis, as described in Section 3.1. The slipt R package has been prepared for release to accompany a publication

demonstrating the applications of the methodology for identifying candidate interacting genes and pathways with CDH1 in breast cancer (TCGA, 2012).

SLIPT is amenable to analysis of any effectively continuous measure of gene activity (e.g., microarray, RNA-Seq, protein abundance, or pathway metagenes). Executing slipt is straightforward: the prep\_data\_for\_SL function scores samples as “low”, “medium”, or “high” for each gene, then the detect\_SL function tests a given query gene against all potential partners by performing the chi-squared test and directional conditions. This function returns a table summarising the observed and expected sample numbers used for the directional criteria, the χ2 values, and corresponding p-values including adjusting for multiple comparisons. The count\_of\_SL and table\_of\_SL functions serve to facilitate summary and extraction of the positive SLIPT hits, respectively, from the table of predictions of synthetic lethal partners.

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

3.5.2 Data Visualisation

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

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

(a) Customised violin plot

(b) Split violin plot

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

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

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

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

The heatmap.2x provides extensions for annotation colour bars for both the rows and columns (as shown in Figure 3.24). Multiple bars are enabled on both axes with matrix inputs (rather than single vector for heatmap.2) which facilitates additional plotting of gene and sample characteristics for comparison with correlation matrices, expression profiles, or pathway metagenes. Annotation bar inputs correspond to their orientation on the plot, each colour bar is provided as a column for the row annotation on the left of the heatmap and as a row for the column annotation on top of the heatmap. Row and column annotation bars are labelled with the column or row names respectively. Additional parameters enable resizing of these annotation bar labels and control of reordering columns for if samples are ordered in advance (e.g., ranked by a metagene or split into groups clustered separately). These features were used through this thesis and are provided in a package GitHub repository.

3.5.3 Extensions to the iGraph Package

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

3.5.3.1 Sampling Simulated Data from Graph Structures

The graphsim package implements the procedure for simulating gene expression from graph structures (as described in Section 3.4.2). By default, this derives a matrix with a geometrically decreasing weighting by distance (by shortest paths) between each pair of nodes with. An absolute decreasing weighting is also available with the option of to derive correlation structures from adjacency matrices or the number of links common partners (i.e., size of the shared “neighbourhood” (Hell, 1976)) between each pair of nodes. Functions to compute these are called directly by passing parameters to them when running the generate\_expression or make\_sigma\_mat commands. This package enables simulating expression data directly from a graph structure (with the intermediate steps automated) or generating Σ parameters for mvtnorm from graph structures or matrices derived from them.

These functions support assigning activating or inhibiting to each edge (with a state parameter).

3.5.3.2 Plotting Directed Graph Structures

The plot.igraph package provides the plot\_directed function specifically developed for directed graph structures and to plot activating or inhibiting for each edge (as described in Section 3.4.2). As shown in Figure 6.2, this function supports separate plotting parameters for each node, node label, and edge. This includes colours of node fill, border, label text, and edges and size of nodes, edge widths, arrowhead lengths, and font size of labels. The state parameter for assigning activating or inhibiting to each edge determines whether edges are dipected with 30° or 90° arrowheads. Colours are assigned separately and so they may be customised. Vectorised parameters are applied across each node or edge, whereas scalar parameters apply the same plotting parameters across them. The default layout function is layout.fruchterman.reingold but any layout function supported by plot function in igraph (Csardi and Nepusz, 2006) is compatible such as layout.kamada.kawai used to implement the Kamada–Kawai algorithm (Kamada and Kawai, 1989) for graph plots throughout this thesis.

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

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

3.5.3.3 Computing Information Centrality

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

3.5.3.4 Testing Pathway Structure with Permutation Testing

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

3.5.3.5 Metapackage to Install iGraph Functions

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

Chapter 4Synthetic Lethal Analysis of Gene Expression Data

Having developed a statistical synthetic lethal detection methodology (SLIPT), it was applied to empirical (publicly available) cancer gene expression datasets in this Chapter. The analysis largely focuses findings from the TCGA breast cancer data (TCGA, 2012) which covers a range of clinical subtypes and is more closely modelled by siRNA data (Telford et al., 2015) generated from screening experiments conducted in MCF10A breast cells. Although stomach cancer data will also be considered to replicate findings in an independent dataset and for it’s relevance to syndromic hereditary diffuse gastric cancer. The TCGA data also has the advantages of other clinical and molecular profiles (e.g., somatic mutation and DNA copy number) for many of the same samples, in addition to a considerable sample size for RNASeq expression data, treated with a rigorous procedure to minimise batch effects. Some findings will be replicated in the Cancer Cell Line Encyclopaedia (CCLE) (Barretina et al., 2012) which may be more comparable to the cell line experiments.

Synthetic lethal candidate partners for CDH1 will be described at both the gene and pathway level. SLIPT gene candidates will be analysed by cluster analysis for common expression profiles across samples and relationships with clinical factors and mutations in key breast cancer genes. These genes will also be compared to the gene candidates from a primary and secondary (validation) screens conducted by Telford et al. (2015) on isogenic cell lines. For comparison, an alternative SLIPT methodology which uses mutation data for CDH1 against expression of candidate partners will also be presented which may better represent the null mutations in HDGC patients and the experiment cell model (Chen et al., 2014). Pathways will be analysed by over-representation analysis (with resampling for comparisons with siRNA data) and supported by a metagene analysis of pathway gene signatures. The pathway metagene expression profiles will be used to replicate known relationships between clinical and molecular characteristics for breast cancer and to demonstrate application of SLIPT directly on metagenes to detect synthetic lethal pathways.

Together these results will demonstrate the wide range of applications for SLIPT analysis and examine the synthetic lethal partners of CDH1 in breast and stomach cancer. These synthetic lethal genes and pathways will be described in both context of the functional implications of novel synthetic lethal relationships and as potential actionable targets against CDH1 deficient tumours, in addition to replication of

established functions of E-cadherin. In particular, the focus of these analysis will be in comparisons with experimental screening data to explore the potential for SLIPT to augment such triage of candidate partners and support further experimental investigations. The key synthetic lethal partner pathways for CDH1, supported by both approaches, will be examined in more detail at the gene and pathway structure level in Chapter 5.

Some of the findings presented in this Chapter have also been included in manuscripts submitted for publication (Kelly et al., 2017a,b) and may bear similarity to them, although the results in this thesis have been edited to cohesively fit with additional findings (including consistent data versions). These findings are the result of investigations conducted throughout this thesis project and only these contributions to the articles are included in this Chapter, not that conducted by co-authors.

4.1 Synthetic lethal genes in breast cancer

The SLIPT methodology (as described in Section 3.1) was applied to the normalised TCGA breast cancer gene expression dataset (n = 1168). As shown in Table 4.1, the most significant genes had strong evidence of expression-based association with CDH1 (high χ2 values) with fewer samples exhibiting low expression of both genes than expected statistically. Eukaryotic translation gene were among the highest gene candidates, including initiation factors, elongation factors, and ribosomal proteins. These are clearly neccessary for cancer cells to grow and proliferate, with sustained gene expression needed to maintain growth signaling pathways and resist apoptosis or immune factors translation may be subject to non-oncogene addiction for CDH1-deficient cells.

While these are among the strongest synthetic lethal candidates, translational genes are cruicial to the viability of healthy cells and dosing for a selective synthetic lethal effect against these may be difficult compared to other biological functions which may also be supported among the SLIPT candidate genes. Furthermore, few known biological functions of CDH1 were among the strongest SL candidates so the remaining candidate genes may also be informative since they are likely to contain these expected functions in addition to novel relationships for CDH1. Thus further pathway level analyses were also conducted to examine biological functions over-represented among synthetic candidate genes and identify synthetic lethal pathways.

Table 4.1: Candidate synthetic lethal gene partners of CDH1 from SLIPT

Gene

Observed

Expected

χ2 value

p-value

p-value (FDR)

TRIP10

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-2-3"

class="td11"> 130

162

5.65 × 10-34

1.84 × 10-31

EEF1B2

56 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-3-3"

class="td11"> 130

158

3.10 × 10-33

9.45 × 10-31

GBGT1

61 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-4-3"

class="td11"> 131

156

1.08 × 10-32

3.14 × 10-30

ELN

81 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-5-3"

class="td11"> 130

149

3.46 × 10-31

8.82 × 10-29

TSPAN4

78 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-6-3"

class="td11"> 130

146

1.63 × 10-30

3.79 × 10-28

GLIPR2

72 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-7-3"

class="td11"> 130

146

1.68 × 10-30

3.86 × 10-28

RPS20

73 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-8-3"

class="td11"> 131

145

1.89 × 10-30

4.28 × 10-28

RPS27A

80 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-9-3"

class="td11"> 130

143

5.53 × 10-30

1.18 × 10-27

EEF1A1P9

63 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-10-3"

class="td11"> 130

141

1.91 × 10-29

3.74 × 10-27

C1R

73 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-11-3"

class="td11"> 130

141

2.05 × 10-29

3.97 × 10-27

LYL1

73 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-12-3"

class="td11"> 130

140

2.99 × 10-29

5.74 × 10-27

RPLP2

71 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-13-3"

class="td11"> 130

139

4.88 × 10-29

9.07 × 10-27

C10orf10

73 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-14-3"

class="td11"> 130

138

6.72 × 10-29

1.20 × 10-26

DULLARD

74 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-15-3"

class="td11"> 131

138

9.29 × 10-29

1.61 × 10-26

PPM1F

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-16-3"

class="td11"> 130

136

1.61 × 10-28

2.65 × 10-26

OBFC2A

69 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-17-3"

class="td11"> 130

136

2.49 × 10-28

3.93 × 10-26

RPL11

70 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-18-3"

class="td11"> 130

136

2.56 × 10-28

3.97 × 10-26

RPL18A

70 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-19-3"

class="td11"> 130

135

3.08 × 10-28

4.70 × 10-26

MFNG

76 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-20-3"

class="td11"> 131

133

7.73 × 10-28

1.12 × 10-25

RPS17

77 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-21-3"

class="td11"> 131

133

8.94 × 10-28

1.29 × 10-25

MGAT1

73 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-22-3"

class="td11"> 130

132

1.44 × 10-27

2.03 × 10-25

RPS12

72 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-23-3"

class="td11"> 130

128

8.57 × 10-27

1.12 × 10-24

C10orf54

73 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-24-3"

class="td11"> 130

127

1.37 × 10-26

1.75 × 10-24

LOC286367

72 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-25-3"

class="td11"> 130

126

2.20 × 10-26

2.70 × 10-24

GMFG

70 </td><td style="white-space:nowrap; text-align:center;" id="TBL-58-26-3"

class="td11"> 130

126

2.20 × 10-26

2.70 × 10-24

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

Strongest candidate SL partners for CDH1 by SLIPT with observed and expected samples with low expression of both genes

The modified mtSLIPT methodology (as described in Section 3.1) was also applied to the normalised TCGA breast cancer gene expression dataset, against somatic loss of function mutations in CDH1. As shown in Table D.1, the most significant genes also had strong evidence of expression associated with CDH1 mutations (high χ2 values) with fewer samples exhibiting both low expression and mutations of each gene than expected statistically. Although, these were not a strongly supported as the expression analysis (in Table 4.1) nor were as many genes detected. This is unsurprising due to the lower sample size with matching somatic mutation data and the lower frequency of CDH1 mutations compared to low expression by 1 ∕ 3 quantiles.

The mtSLIPT candidates had more genes involved in cell and gene regulation, particularly DNA and RNA binding factors. The strongest candidates also include microtubule (KIF12), microfibril (MFAP4), and cell adhesion (TENC1) genes consistent with the established cytoskeletal role of CDH1. The elastin gene (ELN) was notably strongly supported by both expression and mutation SLIPT analysis of CDH1 supporting a interactions with extracellular proteins and the tumour microenvironment.

4.1.1 Synthetic lethal pathways in breast cancer

Translational pathways were strongly over-represented in SLIPT partners, as shown in Table 4.2. These include ribosomal subunits, initiation, peptide elongation, and termination. Regulatory processes involving mRNA including 3’ untranslated region (UTR) binding, L13a-mediated translational silencing, and nonsense-mediated decay were also implicated. These are consistent with protein translation being subject to “non-oncogene addiction” (Luo et al., 2009), as a core process that is dysregulated to sustain cancer proliferation and survival (Gao and Roux, 2015).

Immune pathways, including the adaptive immune system and responses to infectious diseases were also strongly implicated as synthetic lethal with loss of E-cadherin. This is consistent with the alterations of immune response being a hallmark of cancer Hanahan and Weinberg (2000), since evading the immune system is necessary for cancer survival. Either of these systems are potential means to target CDH1 deficient cells, although these were not detected in an isolated cell line experimental screen (Telford et al., 2015) and the

differences between to findings in patient data will be described in more detail in Section 4.2.1.4.

Table 4.2: Pathways for CDH1 partners from SLIPT

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Eukaryotic Translation Elongation

86 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-2-3"

class="td11"> 81

1.3 × 10-207

Peptide chain elongation

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-3-3"

class="td11"> 78

5.6 × 10-201

Eukaryotic Translation Termination

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-4-3"

class="td11"> 77

1.2 × 10-196

Viral mRNA Translation

81 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-5-3"

class="td11"> 76

1.2 × 10-196

Formation of a pool of free 40S subunits

93 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-6-3"

class="td11"> 81

3.7 × 10-194

Nonsense Mediated Decay independent of the Exon Junction Complex

88 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-7-3"

class="td11"> 77

5.3 × 10-187

L13a-mediated translational silencing of Ceruloplasmin expression

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-8-3"

class="td11"> 82

9.6 × 10-183

3’ -UTR-mediated translational regulation

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-9-3"

class="td11"> 82

9.6 × 10-183

GTP hydrolysis and joining of the 60S ribosomal subunit

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-10-3"

class="td11"> 82

1.9 × 10-181

Nonsense-Mediated Decay

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-11-3"

class="td11"> 80

6.2 × 10-176

Nonsense Mediated Decay enhanced by the Exon Junction Complex

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-12-3"

class="td11"> 80

6.2 × 10-176

Adaptive Immune System

412 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-13-3"

class="td11"> 167

6.5 × 10-174

Eukaryotic Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-14-3"

class="td11"> 82

5.7 × 10-173

Cap-dependent Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-15-3"

class="td11"> 82

5.7 × 10-173

SRP-dependent cotranslational protein targeting to membrane

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-16-3"

class="td11"> 79

2.0 × 10-171

Translation

141 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-17-3"

class="td11"> 91

6.1 × 10-170

Infectious disease

347 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-18-3"

class="td11"> 146

1.6 × 10-166

Influenza Infection

117 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-19-3"

class="td11"> 81

1.9 × 10-163

Influenza Viral RNA Transcription and Replication

108 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-20-3"

class="td11"> 77

1.9 × 10-160

Influenza Life Cycle

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-59-21-3"

class="td11"> 77

2.5 × 10-156

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for CDH1

It is also notable that the pathways over-represented in SLIPT candidate genes have strongly significant over-representation of Reactome pathways from the hypergeometric test (as described in Section 2.3.2). Even after adjusting stringently for multiple tests, biologically related pathways give consensus support to these pathways. These pathways are further supported by testing for synthetic lethality against CDH1 mutations (mtSLIPT) with many of these pathways also among the most strongly supported in this analysis (shown in Table D.2). This analysis more closely represents the null CDH1 mutations in HDGC (Guilford et al., 1998) and the experimental MCF10A cell model (Chen et al., 2014). Although it still supports translational and immune pathways not detected in the isolated experimental system, G-protein-coupled receptors (GPCRs) were also among the most strongly supported pathways, supporting the experimental findings of Telford et al. (2015) for these intracellular signalling pathways already being targeted for other diseases.

4.1.2 Expression profiles of synthetic lethal partners

Due to the sheer number of gene candidates and to examine their expression patterns, investigations proceeded into correlation structure and pathway over-representation. This serves to explore the functional similarity of the synthetic lethal partners of CDH1, with the eventual aim to assess their utility as drug targets. As shown in Figure 4.1 (which clusters CDH1 lowly expressing samples separately), there were several large clusters of genes among the expression profiles of the CDH1 synthetic lethal candidate partners. The clustering suggests co-regulation of genes or pathway correlation between partner gene candidates. A number of candidates from an experimental RNAi screen study performed by Telford et al. (2015) were also identified by this approach. In addition, we identified novel gene candidates, which had little effect on viability in isogenic cell line experiments.

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

Figure 4.1: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance) of all samples (separated by the 1∕3 quantile of CDH1 expression) analysed in TCGA breast cancer dataset for gene expression of 5,165 candidate partners of E-cadherin (CDH1) from SLIPT prediction (with significant FDR adjusted p < 0.05). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or toxic in an siRNA screen Telford et al. (2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in CDH1 low samples, notably ‘normal-like’, basal, and estrogen receptor negative samples have elevated expression in one or more distinct clusters showing complexity and variation among candidate synthetic lethal partners. CDH1 low samples also contained most of samples with CDH1 mutations.

In these expression profiles, a gene with a moderate or high signal across samples exhibiting low CDH1 expression would represent a potential drug target. However, it appears that several molecular subtypes of cancer have elevation of different clusters of synthetic lethal candidates in samples with low CDH1. This clustering suggests that different targets or combinations could be effective in different patients suggesting potential utility for stratification. In particular, estrogen receptor negative, basal subtype, and “normal-like” samples Dai et al. (2015); Eroles et al. (2012); Parker et al. (2009) have elevation of genes specific to particular clusters which is indicative of some synthetic lethal interactions being specific to a particular molecular subtype or genetic background. Thus synthetic lethal drug therapy against these subtypes may be ineffective if it were designed against genes in another cluster.

A similar correlation structure was observed among the candidates tested against CDH1 mutation (mtSLIPT), as shown in Figure D.1. This clustering analysis similarly identified several major clusters of putative synthetic lethal partner genes. Although in this case many partner genes had consistently high expression across most of the (predominantly lobular subtype) CDH1 breast cancer samples. However, a major exception to this in the CDH1 expression analysis were the normal samples which were excluded from the mutation data (as they were not tested for tumour-specific genotypes). This supports synthetic lethal interventions being more applicable to CDH1 mutant tumours and genotyping tumours for loss of function will be essential for clinical application. There was still considerable correlation structure, particularly among CDH1 wildtype samples, sufficient to distinguish gene clusters. In contrast to the expression analysis the (predominantly ductal CDH1 wildtype) basal subtype and estrogen receptor negative samples have depleted expression among most candidate synthetic lethal partners. This is consistent with synthetic lethal interventions only being effective in lobular estrogen receptor positive breast cancers in which they are a more common, as recurrent (driver) mutation. However, the remaining samples are still informative for synthetic lethal analysis (by SLIPT) as it requires highly expressing CDH1 samples for comparison.

The CDH1 mutant samples (in Figure 4.1) were predominantly among the CDH1 lowly expressing samples and distributed throughout CDH1 samples with clustering analysis. Thus the molecular profiles of CDH1 low samples a indistinguishable from CDH1 mutant samples with the exception of normal samples (that do not have somatic mutation data as it is inferred from comparison to them to tumour-specific genotypes). Conversely, many of the

CDH1 mutant samples (in Figure D.1) had among the lowest CDH1 expression and some of the synthetic lethal partners were also highly expressed in lowly expressing CDH1 wildtype samples, despite these not being considered as “inactivated” by mtSLIPT analysis.

Together these results support the use for low CDH1 expression as a strategy for detecting CDH1 inactivation. This has the benefit of increasing sample size (including samples such as normal tissue which do not have somatic mutation data available) and increasing the expected number of mutually inactive (low-low) samples for the directional criteria of (mt)SLIPT which enabling it to better distinguish significant deviations below this (as discussed in Section 6.1). This also circumvents the assumption that all (detected) mutations are inactivating (although synonymous mutations were excluded from the analysis), which may not the case for several highly expressing CDH1 mutant samples that do not cluster together in Figures 4.1 or D.1. One of these exhibits among the lowest expression for many predicted synthetic lethal partners and would not be vulnerable to inactivation of these genes. As such correctly genotyping inactivating mutations will be essential in clinical practice for synthetic lethal targeting tumour suppressor genes, particularly for other genes such as TP53 where oncogenic and tumour suppressor mutations (with different molecular consequences) are both common in cancers. Using expression as a measure of gene expression also avoids the assumptions that mutations are somatic rather than germline and that gene inactivation is by detectable mutations rather than other mechanisms such as epigenetic changes which is supported by many lowly expressing CDH1 wildtype samples clustering with similar profiles to mutant samples.

4.1.2.1 Subgroup pathway analysis

Table 4.3: Pathway composition for clusters of CDH1 partners from SLIPT

Pathways Over-represented in Cluster 1

Pathway Size

Cluster Genes

p-value (FDR)

Collagen formation

67 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-2-3"

class="td11"> 10

4<span

class="cmmi-12">.0 × 10-11

Extracellular matrix organisation

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-3-3"

class="td11"> 21

1<span

class="cmmi-12">.8 × 10-9

Collagen biosynthesis and modifying enzymes

56 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-4-3"

class="td11"> 8

1<span

class="cmmi-12">.8 × 10-9

Uptake and actions of bacterial toxins

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-5-3"

class="td11"> 5

9<span

class="cmmi-12">.5 × 10-9

Elastic fibre formation

37 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-6-3"

class="td11"> 6

1<span

class="cmmi-12">.9 × 10-8

Muscle contraction

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-7-3"

class="td11"> 7

2<span

class="cmmi-12">.4 × 10-7

Fatty acid, triacylglycerol, and ketone body metabolism

117 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-8-3"

class="td11"> 10

4<span

class="cmmi-12">.9 × 10-7

XBP1(S) activates chaperone genes

51 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-9-3"

class="td11"> 6

6<span

class="cmmi-12">.6 × 10-7

IRE1alpha activates chaperones

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-10-3"

class="td11"> 6

1<span

class="cmmi-12">.2 × 10-6

Neurotoxicity of clostridium toxins

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-11-3"

class="td11"> 3

1<span

class="cmmi-12">.3 × 10-6

Retrograde neurotrophin signalling

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-12-3"

class="td11"> 3

1<span

class="cmmi-12">.3 × 10-6

Assembly of collagen fibrils and other multimeric structures

40 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-13-3"

class="td11"> 5

1<span

class="cmmi-12">.9 × 10-6

Collagen degradation

58 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-14-3"

class="td11"> 6

2<span

class="cmmi-12">.0 × 10-6

Arachidonic acid metabolism

41 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-15-3"

class="td11"> 5

2<span

class="cmmi-12">.1 × 10-6

Synthesis of PA

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-16-3"

class="td11"> 4

3<span

class="cmmi-12">.0 × 10-6

Signaling by NOTCH

80 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-17-3"

class="td11"> 7

3<span

class="cmmi-12">.3 × 10-6

Signalling to RAS

27 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-18-3"

class="td11"> 4

3<span

class="cmmi-12">.7 × 10-6

Integrin cell surface interactions

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-19-3"

class="td11"> 7

4<span

class="cmmi-12">.2 × 10-6

Smooth Muscle Contraction

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-20-3"

class="td11"> 4

4<span

class="cmmi-12">.4 × 10-6

ECM proteoglycans

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-21-3"

class="td11"> 6

6<span

class="cmmi-12">.3 × 10-6

</td>

Pathways Over-represented in Cluster 2

Pathway Size

Cluster Genes

p-value (FDR)

Eukaryotic Translation Elongation

86 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-24-3"

class="td11"> 75

1<span

class="cmmi-12">.1 × 10-181

Viral mRNA Translation

81 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-25-3"

class="td11"> 72

9<span

class="cmmi-12">.8 × 10-179

Peptide chain elongation

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-26-3"

class="td11"> 72

1<span

class="cmmi-12">.9 × 10-175

Eukaryotic Translation Termination

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-27-3"

class="td11"> 72

1<span

class="cmmi-12">.9 × 10-175

Formation of a pool of free 40S subunits

93 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-28-3"

class="td11"> 75

1<span

class="cmmi-12">.9 × 10-171

Nonsense Mediated Decay independent of the Exon Junction Complex

88 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-29-3"

class="td11"> 72

9<span

class="cmmi-12">.9 × 10-168

L13a-mediated translational silencing of Ceruloplasmin expression

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-30-3"

class="td11"> 75

3<span

class="cmmi-12">.0 × 10-159

3’ -UTR-mediated translational regulation

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-31-3"

class="td11"> 75

3<span

class="cmmi-12">.0 × 10-159

Nonsense-Mediated Decay

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-32-3"

class="td11"> 75

3<span

class="cmmi-12">.0 × 10-159

Nonsense Mediated Decay enhanced by the Exon Junction Complex

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-33-3"

class="td11"> 75

3<span

class="cmmi-12">.0 × 10-159

SRP-dependent cotranslational protein targeting to membrane

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-34-3"

class="td11"> 75

3<span

class="cmmi-12">.2 × 10-158

GTP hydrolysis and joining of the 60S ribosomal subunit

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-35-3"

class="td11"> 75

3<span

class="cmmi-12">.2 × 10-158

Eukaryotic Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-36-3"

class="td11"> 75

4<span

class="cmmi-12">.5 × 10-151

Cap-dependent Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-37-3"

class="td11"> 75

4<span

class="cmmi-12">.5 × 10-151

Influenza Infection

117 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-38-3"

class="td11"> 75

1<span

class="cmmi-12">.4 × 10-145

Influenza Viral RNA Transcription and Replication

108 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-39-3"

class="td11"> 72

5<span

class="cmmi-12">.7 × 10-145

Translation

141 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-40-3"

class="td11"> 81

8<span

class="cmmi-12">.0 × 10-143

Influenza Life Cycle

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-41-3"

class="td11"> 72

2<span

class="cmmi-12">.3 × 10-141

Infectious disease

347 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-42-3"

class="td11"> 103

2<span

class="cmmi-12">.2 × 10-95

Formation of the ternary complex, and subsequently, the 43S complex

47 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-43-3"

class="td11"> 33

6<span

class="cmmi-12">.8 × 10-80

</td>

Pathways Over-represented in Cluster 3

Pathway Size

Cluster Genes

p-value (FDR)

Adaptive Immune System

412 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-46-3"

class="td11"> 90

6<span

class="cmmi-12">.1 × 10-61

Chemokine receptors bind chemokines

52 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-47-3"

class="td11"> 27

6<span

class="cmmi-12">.7 × 10-56

Generation of second messenger molecules

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-48-3"

class="td11"> 21

6<span

class="cmmi-12">.5 × 10-55

Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-49-3"

class="td11"> 29

6<span

class="cmmi-12">.5 × 10-55

TCR signalling

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-50-3"

class="td11"> 27

8<span

class="cmmi-12">.9 × 10-51

Peptide ligand-binding receptors

161 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-51-3"

class="td11"> 40

1<span

class="cmmi-12">.5 × 10-45

Translocation of ZAP-70 to Immunological synapse

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-52-3"

class="td11"> 14

3<span

class="cmmi-12">.1 × 10-43

Costimulation by the CD28 family

51 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-53-3"

class="td11"> 22

4<span

class="cmmi-12">.0 × 10-43

PD-1 signalling

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-54-3"

class="td11"> 15

4<span

class="cmmi-12">.0 × 10-41

Class A/1 (Rhodopsin-like receptors)

258 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-55-3"

class="td11"> 50

6<span

class="cmmi-12">.7 × 10-41

Phosphorylation of CD3 and TCR zeta chains

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-56-3"

class="td11"> 14

1<span

class="cmmi-12">.3 × 10-40

Interferon gamma signalling

74 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-57-3"

class="td11"> 24

5<span

class="cmmi-12">.0 × 10-39

GPCR ligand binding

326 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-58-3"

class="td11"> 57

1<span

class="cmmi-12">.8 × 10-38

Cytokine Signaling in Immune system

268 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-59-3"

class="td11"> 48

8<span

class="cmmi-12">.9 × 10-37

Downstream TCR signalling

45 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-60-3"

class="td11"> 18

1<span

class="cmmi-12">.8 × 10-35

Gαi signalling events

167 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-61-3"

class="td11"> 33

2<span

class="cmmi-12">.2 × 10-33

Cell surface interactions at the vascular wall

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-62-3"

class="td11"> 21

1<span

class="cmmi-12">.3 × 10-26

Interferon Signalling

164 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-63-3"

class="td11"> 28

1<span

class="cmmi-12">.7 × 10-26

Extracellular matrix organisation

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-64-3"

class="td11"> 35

2<span

class="cmmi-12">.7 × 10-25

Antigen activates B Cell Receptor leading to generation of second messengers

32 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-65-3"

class="td11"> 12

7<span

class="cmmi-12">.2 × 10-25

</td>

Pathways Over-represented in Cluster 4

Pathway Size

Cluster Genes

p-value (FDR)

Extracellular matrix organisation

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-68-3"

class="td11"> 48

8<span

class="cmmi-12">.0 × 10-41

Class A/1 (Rhodopsin-like receptors)

258 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-69-3"

class="td11"> 47

2<span

class="cmmi-12">.8 × 10-36

GPCR ligand binding

326 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-70-3"

class="td11"> 54

2<span

class="cmmi-12">.1 × 10-34

Gαs signalling events

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-71-3"

class="td11"> 22

1<span

class="cmmi-12">.4 × 10-31

GPCR downstream signalling

472 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-72-3"

class="td11"> 68

1<span

class="cmmi-12">.1 × 10-29

Haemostasis

423 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-73-3"

class="td11"> 61

3<span

class="cmmi-12">.3 × 10-29

Platelet activation, signalling and aggregation

180 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-74-3"

class="td11"> 31

7<span

class="cmmi-12">.1 × 10-28

Binding and Uptake of Ligands by Scavenger Receptors

40 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-75-3"

class="td11"> 14

9<span

class="cmmi-12">.9 × 10-27

RA biosynthesis pathway

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-76-3"

class="td11"> 11

2<span

class="cmmi-12">.5 × 10-26

Response to elevated platelet cytosolic Ca2+

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-77-3"

class="td11"> 19

3<span

class="cmmi-12">.0 × 10-26

Developmental Biology

420 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-78-3"

class="td11"> 57

3<span

class="cmmi-12">.5 × 10-26

Gαi signalling events

167 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-79-3"

class="td11"> 28

7<span

class="cmmi-12">.3 × 10-26

Platelet degranulation

77 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-80-3"

class="td11"> 18

1<span

class="cmmi-12">.6 × 10-25

Gastrin-CREB signalling pathway via PKC and MAPK

171 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-81-3"

class="td11"> 28

2<span

class="cmmi-12">.5 × 10-25

Muscle contraction

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-82-3"

class="td11"> 16

4<span

class="cmmi-12">.7 × 10-25

Gαq signalling events

150 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-83-3"

class="td11"> 25

3<span

class="cmmi-12">.2 × 10-24

Retinoid metabolism and transport

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-84-3"

class="td11"> 12

5<span

class="cmmi-12">.0 × 10-24

Phase 1 - Functionalisation of compounds

67 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-85-3"

class="td11"> 16

6<span

class="cmmi-12">.5 × 10-24

Signalling by Retinoic Acid

42 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-86-3"

class="td11"> 13

6<span

class="cmmi-12">.7 × 10-24

Degradation of the extracellular matrix

102 </td><td style="white-space:nowrap; text-align:center;" id="TBL-60-87-3"

class="td11"> 19

1<span

class="cmmi-12">.4 × 10-22

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

Synthetic lethal gene candidates for CDH1 from SLIPT performed on RNA-Seq expression data were also used for pathway over-representation analyses (as described in Section 2.3.2). The correlation structure in the expression of candidates synthetic lethal genes in CDH1 low tumours (lowest 1 ∕ 3rd quantile of expression) was examined for distinct biological pathways in subgroups of genes elevated in different clusters of samples. These gene were highly expressed in different samples with their clinical factors including estrogen receptor status and intrinsic (PAM50) subtype (Parker et al., 2009) shown in Figure 4.1.

As shown by the most over-represented pathways in Table 4.3, each correlated cluster of candidate synthetic lethal partners of CDH1 contains functionally different genes. Cluster 1 contains genes with less evidence of over-represented pathways than other clusters, corresponding to less correlation between genes within the cluster, and to it being a relatively small group. While there is some indication that collagen biosynthesis, microfibril elastic fibres, extracellular matrix, and metabolic pathways may be over-represented in Cluster 1, these results are mainly based on small pathways containing few synthetic lethal genes. Genes in Cluster 2 exhibited low expression in normal tissue samples compared to tumour samples (see Figure 4.1) and show compelling evidence of over-representation of post-transcriptional gene regulation and protein translation processes. Similarly, Cluster 3 has over-representation of immune signalling pathways (including chemokines, secondary messenger, and TCR signaling) and downstream intracellular signalling cascades such as G protein coupled receptor (GPCR) and Gαi signalling events. While pathway over-representation was weaker among genes in Cluster 4, they contained intracellular signalling pathways and were highly expressed in normal samples (in contrast to Cluster 2). Cluster 4 also involved extracellular factors and stimuli such as extracellular matrix, platelet activation, ligand receptors, and retinoic acid signalling.

Based on these results, potential synthetic lethal partners of CDH1 include processes known to be dysregulated in cancer, such as translational, cytoskeletal, and immune processes. Intracellular signalling cascades such as the GPCRs and extracellular stimuli for these pathways were also implicated in potential synthetic lethality with CDH1.

Similar translational, cytoskeletal, and immune processes were identified among SLIPT partners with respect to CDH1 mutation, shown in Table D.3. While GPCR signalling was replicated in mtSLIPT analysis, there was also stronger over-representation for NOTCH,

ERBB2, and PI3K/AKT signalling in mutation analysis consistent with these signals being important for proliferation of CDH1 deficient tumours. The GCPR and PI3K/AKT pathways are of particular interest as pathways with oncogenic mutations that can be targeted and downstream effects on translation (a strongly supported process across analyses). Extracellular matrix pathways (such as elastic fibre formation) were also supported across analyses (in Tables 4.3 and D.3) consistent with the established cell-cell signalling role of CDH1 and the importance of the tumour microenvironment for cancer proliferation.

4.2 Comparison of synthetic lethal gene candidates

4.2.1 Comparison with siRNA screen candidates

Gene candidates were compared between computational (SLIPT in TCGA breast cancer data) and experimental (the primary siRNA screen performed by Telford et al. (2015)) approaches in Figure 4.2. The number of genes detected by both methods did not produce a significant overlap but these may be difficult to compare due to vast differences between the detection methods. There were similar issues comparison of mtSLIPT genes tested against CDH1 mutations (in Figure G.2), despite exlcuded genes not tested by both methods in either test. However, these intersecting genes may still be functionally informative or amenable to drug triage as they were replicated across both methods and pathway over-representation differed between the sections of the Venn diagram (see Figure 4.2).

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

Figure 4.2: Comparison of SLIPT to siRNA. Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ2 test suggests that the overlap is no more than would be expected by chance (p = 0.281).

4.2.1.1 Comparison with correlation

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

Figure 4.3: Compare SLIPT and siRNA genes with correlation. The χ2 p-values for genes tested by SLIPT (in TCGA breast cancer) expression analysis were compared against Pearson’s correlation of gene expression with CDH1. Genes detected by SLIPT or siRNA are coloured according to the legend.

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

Figure 4.4: Compare SLIPT and siRNA genes with correlation. Genes detected as candidate synthetic lethal partners by SLIPT (in TCGA breast cancer) expression analysis and experimental screening (with siRNA) were compared against Pearson’s correlation of gene expression with CDH1. There were no differences in correlation between gene groups detected by either approach.

Another potential means to triage drug target candidates is correlation of expression profiles with CDH1. Correlation with CDH1 was compared to SLIPT and siRNA results in Figure 4.3. The genes not detected by SLIPT (including siRNA candidates) included genes with high (insignificant) SLIPT p-values. As expected, these genes were distributed around a correlation of zero and genes with higher correlation with CDH1 (either direction) had more significant SLIPT p-values, although there were exceptions to this trend and larger positive correlations were negative correations.

The majority of SLIPT candidates appeared to have negative correlations and moreso for those genes detected by both approaches, although these were typically weak correlations and are unlikely to be sufficient to detect such genes on their own. This is supported by simulation results in Section 6.1.

There were not strong postive correlations with CDH1 among siRNA candidates, consistent with previous findings that co-expression is not predictive of synthetic lethality (Jerby-Arnon et al., 2014; Lu et al., 2015). Negative correlation may not be indicative of synthetic lethality either as many siRNA candidates also had positive correlations. The SLIPT methodology has shown to detect genes with both positive and negative correlations, although it does appear to preferentially detect negatively correlated genes to some extent. These findings were replicated with the mtSLIPT approach against CDH1 mutation (in Figure D.3), although the range of the χ2 p-values differ due to lower sample size for mutation analysis.

However, the apparent tendancy for genes detected by SLIPT or siRNA to have negative correlations with CDH1 expression may be due to the smaller number of genes in these groups. The distribution of CDH1 correations does not differ across these gene groups (as shown by Figures 4.4 and D.4). Therefore further triage of gene candidates by correaltion is not suitable, nor is use of correlation itself to preduct synthetic lethal partners in the first place.

4.2.1.2 Comparison with viability

A similar comparison of SLIPT results was made with the viability ratio (of CDH1 mutant to wildtype) in the primary siRNA screen performed by Telford et al. (2015). The significance and viability thresholds used for SLIPT and siRNA detection of synthetic lethal candidate partners of CDH1 are clear in Figure 4.5. However note that not all of the gene below these thresholds are neccessarily selected to be

candidate partners as additional criteria were used in each case: directional criteria as for SLIPT (see Section 3.1) and minimum wildtype viability for siRNA (Telford et al., 2015).

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

Figure 4.5: Compare SLIPT and siRNA genes with siRNA viability. The χ2 p-values for genes tested by SLIPT (in TCGA breast cancer) expression analysis were compared against the viability ratio of CDH1 mutant and wildtype cells in the primary siRNA screen. Genes detected by SLIPT or siRNA are coloured according to the legend.

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

Figure 4.6: Compare SLIPT and siRNA genes with viability. The χ2 p-values for genes tested by SLIPT (in TCGA breast cancer) expression analysis were compared (on a log-scale) against the viability ratio of CDH1 mutant and wildtype cells in the primary siRNA screen. Genes detected by SLIPT or siRNA are coloured according to the legend with a grey line for p = 0.05.

There does not appear to be a clear relationship between SLIPT and siRNA candidates. Many genes not detected by both approaches were numerous in Figures 4.2 and D.2. These genes detected by either are not neccessarily near the thresholds for the other. In this respect the SLIPT approach with patient data and cell line experiments are independent means to identify synthetic lethal candidates. While genes detected by both approaches were not neccessarily more strongly supported by either, the genes with a viability closer to 1 (no synthetic lethal effect) in siRNA included those with more significant SLIPT p-values whereas more extreme viability ratios tended to be less significant (as shown by a logarithmic plot in Figure 4.6). Although it should be noted that genes with more moderate viability ratios were more common and SLIPT was capable (despite adjusting for multiple testing) of detecting significant genes with extreme viability ratios, particularly those considerably lower than 1.

However, there was not support for SLIPT candidates or those detected by both approaches having considerably different viability ratios (as shown in Figures 4.7 and D.5). The difference between the gene groups stems largely from the viability thresholds used by Telford et al. (2015) to detect synthetic lethal candidates in the primary screen, rather than more extreme viability ratios for genes identified by SLIPT.

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

Figure 4.7: Compare SLIPT and siRNA genes with siRNA viability. Genes detected as candidate synthetic lethal partners by SLIPT (in TCGA breast cancer) expression analysis and experimental screening (with siRNA) were compared against the viability ratio of CDH1 mutant and wildtype cells in the primary siRNA screen. There were clear no differences in viability between genes detected by SLIPT and those not with the differences being primarily due to viability thresholds being used to detect synthetic lethality by Telford et al. (2015).

4.2.1.3 Comparison with secondary siRNA screen candidates

However, it should be noted that genes with a lower viability ratio were not necessarily the strongest supported by experimental screening. The primary screen (with 4 pooled siRNAs) has been used for the majority of comparisons in this thesis because the genome-wide panel of target genes screened enables a large number of genes to be compared with SLIPT results from gene expression and somatic mutation analysis. A secondary screen was also performed by Telford et al. (2015) on the isogenic MCF10A breast cell lines to individually validate the siRNAs separately, with the strongest candidates being those exhibiting synthetic lethal viability ratios replicated across independently targeting siRNAs. This was performed for the top 500 candidates (with the lowest viability ratio) from the primary screen and the 482 of these genes also tested by SLIPT in breast cancer (and the 486 genes tested by SLIPT in stomach cancer).

The secondary screen results are given in Appendix C which show that SLIPT candidate genes are more significantly (p = 7.49 × 10-3 by Fisher’s exact test) more likely to be validated in the secondary screen and are thus informative of more robust partner genes, in addition to providing support that these interactions are consistent with expression profiles from heterogeneous patient samples across genetic backgrounds. While the individual genes detected by either approach do not neccessarily match (and are potentially false-positives), the biological functions important in CDH1 deficient cancers and potential mechanisms for specific targeting of them can be further supported by pathway analysis of the gene detected by either method. The genes detected by both approaches may therefore be more informative at the pathway level, where it is unlikely for a pathway to be consistently detected by chance. As the SLIPT candidates differ from the siRNA candidates (and are more likely to be validated), they can provide additional mechanisms by which CDH1 deficient cancers proliferate and vulnerabilities that may be exploited against them by using the synthetic lethal pathways.

4.2.1.4 Comparison of screen at pathway level

Table 4.4: Pathway composition for CDH1 partners from SLIPT and siRNA screening

Predicted only by SLIPT (4025 genes)

Pathway Size

Genes Identified

p-value (FDR)

Eukaryotic Translation Elongation

80 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-2-3"

class="td11"> 75

1.5 × 10-182

Peptide chain elongation

77 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-3-3"

class="td11"> 72

2.9 × 10-176

Viral mRNA Translation

75 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-4-3"

class="td11"> 70

4.9 × 10-172

Eukaryotic Translation Termination

76 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-5-3"

class="td11"> 70

5.9 × 10-170

Formation of a pool of free 40S subunits

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-6-3"

class="td11"> 74

9.5 × 10-166

Nonsense Mediated Decay independent of the Exon Junction Complex

81 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-7-3"

class="td11"> 70

1.2 × 10-160

L13a-mediated translational silencing of Ceruloplasmin expression

97 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-8-3"

class="td11"> 75

3.8 × 10-155

3’ -UTR-mediated translational regulation

97 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-9-3"

class="td11"> 75

3.8 × 10-155

GTP hydrolysis and joining of the 60S ribosomal subunit

98 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-10-3"

class="td11"> 75

6.0 × 10-154

Nonsense-Mediated Decay

96 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-11-3"

class="td11"> 73

5.2 × 10-150

Nonsense Mediated Decay enhanced by the Exon Junction Complex

96 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-12-3"

class="td11"> 73

5.2 × 10-150

SRP-dependent cotranslational protein targeting to membrane

97 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-13-3"

class="td11"> 73

7.8 × 10-149

Eukaryotic Translation Initiation

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-14-3"

class="td11"> 75

4.7 × 10-146

Cap-dependent Translation Initiation

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-15-3"

class="td11"> 75

4.7 × 10-146

Translation

133 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-16-3"

class="td11"> 83

4.0 × 10-142

Influenza Viral RNA Transcription and Replication

102 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-17-3"

class="td11"> 71

2.9 × 10-137

Influenza Infection

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-18-3"

class="td11"> 74

3.7 × 10-137

Influenza Life Cycle

106 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-19-3"

class="td11"> 71

2.3 × 10-133

Infectious disease

326 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-20-3"

class="td11"> 125

4.2 × 10-120

Extracellular matrix organisation

189 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-21-3"

class="td11"> 77

5.4 × 10-95

</td>

Detected only by siRNA screen (1599 genes)

Pathway Size

Genes Identified

p-value (FDR)

Class A/1 (Rhodopsin-like receptors)

282 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-24-3"

class="td11"> 44

1.3 × 10-27

GPCR ligand binding

363 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-25-3"

class="td11"> 52

5.8 × 10-26

Gαq signalling events

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-26-3"

class="td11"> 26

6.7 × 10-23

Gastrin-CREB signalling pathway via PKC and MAPK

180 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-27-3"

class="td11"> 27

2.0 × 10-21

Gαi signalling events

184 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-28-3"

class="td11"> 27

5.3 × 10-21

Downstream signal transduction

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-29-3"

class="td11"> 23

7.6 × 10-21

Signalling by PDGF

172 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-30-3"

class="td11"> 25

4.0 × 10-20

Peptide ligand-binding receptors

175 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-31-3"

class="td11"> 25

8.5 × 10-20

Signalling by ERBB2

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-32-3"

class="td11"> 22

1.3 × 10-19

DAP12 interactions

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-33-3"

class="td11"> 23

2.6 × 10-19

DAP12 signalling

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-34-3"

class="td11"> 22

2.7 × 10-19

Organelle biogenesis and maintenance

264 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-35-3"

class="td11"> 33

5.5 × 10-19

Signalling by NGF

266 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-36-3"

class="td11"> 33

8.2 × 10-19

Downstream signalling of activated FGFR1

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-37-3"

class="td11"> 20

1.1 × 10-18

Downstream signalling of activated FGFR2

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-38-3"

class="td11"> 20

1.1 × 10-18

Downstream signalling of activated FGFR3

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-39-3"

class="td11"> 20

1.1 × 10-18

Downstream signalling of activated FGFR4

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-40-3"

class="td11"> 20

1.1 × 10-18

Signalling by FGFR

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-41-3"

class="td11"> 21

1.3 × 10-18

Signalling by FGFR1

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-42-3"

class="td11"> 21

1.3 × 10-18

Signalling by FGFR2

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-43-3"

class="td11"> 21

1.3 × 10-18

</td>

Intersection of SLIPT and siRNA screen (604 genes)

Pathway Size

Genes Identified

p-value (FDR)

Visual phototransduction

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-46-3"

class="td11"> 9

6.9 × 10-10

Gαs signalling events

48 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-47-3"

class="td11"> 7

1.6 × 10-7

Retinoid metabolism and transport

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-48-3"

class="td11"> 5

1.7 × 10-7

Acyl chain remodelling of PS

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-49-3"

class="td11"> 3

6.5 × 10-6

Transcriptional regulation of white adipocyte differentiation

51 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-50-3"

class="td11"> 6

6.5 × 10-6

Chemokine receptors bind chemokines

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-51-3"

class="td11"> 4

6.5 × 10-6

Signalling by NOTCH4

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-52-3"

class="td11"> 3

6.9 × 10-6

Defective EXT2 causes exostoses 2

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-53-3"

class="td11"> 3

6.9 × 10-6

Defective EXT1 causes exostoses 1, TRPS2 and CHDS

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-54-3"

class="td11"> 3

6.9 × 10-6

Platelet activation, signalling and aggregation

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-55-3"

class="td11"> 12

6.9 × 10-6

Phase 1 - Functionalisation of compounds

41 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-56-3"

class="td11"> 5

1.3 × 10-5

Amine ligand-binding receptors

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-57-3"

class="td11"> 3

1.7 × 10-5

Acyl chain remodelling of PE

14 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-58-3"

class="td11"> 3

2.4 × 10-5

Signalling by GPCR

300 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-59-3"

class="td11"> 23

2.4 × 10-5

Molecules associated with elastic fibres

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-60-3"

class="td11"> 4

2.6 × 10-5

DAP12 interactions

128 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-61-3"

class="td11"> 10

2.6 × 10-5

Cytochrome P450 - arranged by substrate type

30 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-62-3"

class="td11"> 4

3.2 × 10-5

GPCR ligand binding

147 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-63-3"

class="td11"> 11

3.8 × 10-5

Acyl chain remodelling of PC

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-64-3"

class="td11"> 3

4.0 × 10-5

Response to elevated platelet cytosolic Ca2+

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-61-65-3"

class="td11"> 6

4.2 × 10-5

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

These pathway over-representation analyses (performed as described in Section 2.3.2) correspond to genes separated into SLIPT or siRNA screen candidates unique to either method or detected by both (Table 4.4). The SLIPT-specific gene candidates were involved most strongly with translational and immune regulatory pathways, although extracellular matrix pathways were also supported. These pathways were largely consistent with those identified in Tabel 4.2 and in the clustering analysis (Table 4.3). The genes detected only by the siRNA screen had over-representation of cell signalling pathways, including many containing genes known to be involved in cancer (e.g., MAPK, PDGF, ERBB2, and FGFR), with the detection of Class A GPCRs supporting the independent analyses by Telford et al. Telford et al. (2015). The intersection of computational and experimental synthetic lethal partners of CDH1 has stronger evidence for over-representation of GPCR pathways and more specific subclasses, such as visual phototransduction (p = 6.9 × 10-10) and G αs signalling events (p = 1.7 × 10-7), than other signalling pathways.

The pathway analysis for mtSLIPT against CDH1 mutations (in Table D.4) had concordant results for both mtSLIPT-specific and siRNA-specific pathways. While the specific pathway composition of the intersection of these analyses differed from SLIPT against low CDH1 expression, signalling pathways including GPCRs, NOTCH, EERB2, PDGF, and SCF-KIT. These findings indicate the signalling pathways are among the most suitable vulnerability to exploit in targeting CDH1 deficient tumours as they can be detected in both a patient cohort (with TCGA expression data) and tested in a laboratory system. However, it is possible that the isolated experimental system is set up to preferentially detect kinase singalling pathways (which are amenable to pharmacological inhibition and translation to the clinic) and the other pathways identified by SLIPT may still be informative of the role of CDH1 loss of function in cancers or mechanisms by which further gene loss leads to specific inviability.

Resampling of genes for pathway enrichment

Comparing genes between experimental screen candidates and prediction from TCGA expression data has been less consistent than pathways. Although this is not unexpected since synthetic lethal pathways more more robustly conserved (Dixon et al., 2008) and the

computational approach using patient samples from complex tumour microenvironment has considerably different strengths to an experimental screen (Telford et al., 2015) based on genetically homogenous cell line models in an isolated laboratory environment. For instance, it is unlikely for immune signaling to be detected in an isolated cell culture system.

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

Figure 4.8: Resampled intersection of SLIPT and siRNA candidates. Resampling analysis of intersect size from genes detected by SLIPT and siRNA screening approaches over 1 million replicates. The proportion of expected intersection sizes for random samples below or above the observed intersection size respectively, lacking significant over-representation or depletion of siRNA screen candidates within the SLIPT predictions for CDH1.

The overlap between synthetic lethal from bioinformatics SLIPT predictions and siRNA screening has raised other questions including whether the pathways over-represented would be expected by chance. This of particular concern since the siRNA candidate genes themselves are highly over-represented for particular pathways (such as GPCRs) so selecting any intersect with them would be enriched for these pathways. Another pathway approach is to test whether pathways are over-represented in randomly sampled genes, comparing many “resamplings” 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 (as described in Section 2.3.6).

Of particular concern are the over-represented pathways in genes detected by both methods. Pathway over-representation alone does not detect whether SLIPT predicted genes or siRNA candidates are enriched within each other. This resampling analysis therefore detects whether over-represented pathways were detected by SLIPT independently of their over-representation among siRNA candidates (without assuming an underlying test statistic distribution).

A resampling approach is also applicable to testing whether the number of genes detected by each approach significantly intersected. As shown in Figure 4.8, resampling did not find evidence of significant depletion or over-representation for experimental synthetic lethal candidates in the computationally predicted synthetic lethal partners of CDH1 and the overlap may be observed by chance. This is consistent with previous findings (see Figure 4.2) and does not preclude pathway relationships being supported by resampling.

A permutation analysis was performed to resample the genes tested by both approaches to investigate whether the observed pathway over-representation could have occurred in a randomly selected sample of genes from the experimental candidates, that is, whether the pathway predictions from SLIPT could be expected by chance (as described in sections 2.2.3.1 and 2.3.6). While the number of siRNA candidate genes detected by SLIPT was not statistically significant (p = 0.281), this may be due to the vastly different limitations of the approaches and the correlation structure of gene expression not being independent (as assumed for multiple testing procedures). The intersection may still be functionally relevant to CDH1-deficient cancers, such as the pathway data in Table 4.4. The resampling analysis for pathways was compared to the pathway over-representation for SLIPT predicted synthetic lethal partners in Table 4.5. Similarly, the pathway resampling

for intersection between SLIPT predictions and experimental screen candidates was compared to pathway over-representation in Table 4.6 for intersection with siRNA data.

The pathway resampling approach for SLIPT-specific gene candidates (Table 4.5) replicates the gene set over-representation analysis for all SLIPT genes, detecting evidence of synthetic lethal pathways for CDH1 in translational, immune, and cell signalling pathways including Gαi signalling, GPCR downstream signalling, and chemokine receptor binding. While the immune and signal transduction pathways were not significantly over-represented in the resampling analysis, the results for the two approaches were largely consistent for translation and post-transcriptional gene regulation, supporting gene set over-representation of the SLIPT-specific pathways in Table 4.5. In particular, some of the most significantly over-represented pathways had higher observed χ2 values than any of the 1 million random permutations. Similar pathways were also replicated by permutation analysis for mtSLIPT candidate partners against CDH1 mutation (shown in Table D.5). This shows that many of the pathways detected specfically by SLIPT are replicated by permutation procedures and that the permutation approach is capable of detecting many of the most strongly over-represented pathways.

Table 4.5: Pathways for CDH1 partners from SLIPT

Reactome Pathway

Over-representation

Permutation

Eukaryotic Translation Elongation

1<span

class="cmmi-12">.3 × 10-207

< 1.241 × 10-5

Peptide chain elongation

5<span

class="cmmi-12">.6 × 10-201

< 1.241 × 10-5

Viral mRNA Translation

1<span

class="cmmi-12">.2 × 10-196

< 1.241 × 10-5

Eukaryotic Translation Termination

1<span

class="cmmi-12">.2 × 10-196

< 1.241 × 10-5

Formation of a pool of free 40S subunits

3<span

class="cmmi-12">.7 × 10-194

< 1.241 × 10-5

Nonsense Mediated Decay independent of the Exon Junction Complex

5<span

class="cmmi-12">.3 × 10-187

< 1.241 × 10-5

L13a-mediated translational silencing of Ceruloplasmin expression

9<span

class="cmmi-12">.6 × 10-183

< 1.241 × 10-5

3’ -UTR-mediated translational regulation

9<span

class="cmmi-12">.6 × 10-183

< 1.241 × 10-5

GTP hydrolysis and joining of the 60S ribosomal subunit

1<span

class="cmmi-12">.9 × 10-181

< 1.241 × 10-5

Nonsense-Mediated Decay

6<span

class="cmmi-12">.2 × 10-176

< 1.241 × 10-5

Nonsense Mediated Decay enhanced by the Exon Junction Complex

6<span

class="cmmi-12">.2 × 10-176

< 1.241 × 10-5

Adaptive Immune System

6<span

class="cmmi-12">.5 × 10-174

0<span

class="cmmi-12">.15753

Eukaryotic Translation Initiation

5<span

class="cmmi-12">.7 × 10-173

< 1.241 × 10-5

Cap-dependent Translation Initiation

5<span

class="cmmi-12">.7 × 10-173

< 1.241 × 10-5

SRP-dependent cotranslational protein targeting to membrane

2<span

class="cmmi-12">.0 × 10-171

< 1.241 × 10-5

Translation

6<span

class="cmmi-12">.1 × 10-170

< 1.241 × 10-5

Infectious disease

1<span

class="cmmi-12">.6 × 10-166

0<span

class="cmmi-12">.23231

Influenza Infection

1<span

class="cmmi-12">.9 × 10-163

< 1.241 × 10-5

Influenza Viral RNA Transcription and Replication

1<span

class="cmmi-12">.9 × 10-160

< 1.241 × 10-5

Influenza Life Cycle

2<span

class="cmmi-12">.5 × 10-156

< 1.241 × 10-5

Extracellular matrix organisation

1<span

class="cmmi-12">.1 × 10-152

0<span

class="cmmi-12">.071761

GPCR ligand binding

1<span

class="cmmi-12">.1 × 10-143

0<span

class="cmmi-12">.55801

Class A/1 (Rhodopsin-like receptors)

1<span

class="cmmi-12">.5 × 10-142

0<span

class="cmmi-12">.58901

GPCR downstream signalling

7<span

class="cmmi-12">.6 × 10-140

0<span

class="cmmi-12">.098357

Haemostasis

1<span

class="cmmi-12">.9 × 10-134

0<span

class="cmmi-12">.27059

Developmental Biology

2<span

class="cmmi-12">.0 × 10-123

0<span

class="cmmi-12">.52737

Metabolism of lipids and lipoproteins

3<span

class="cmmi-12">.3 × 10-120

0<span

class="cmmi-12">.724

Cytokine Signalling in Immune system

2<span

class="cmmi-12">.6 × 10-119

0<span

class="cmmi-12">.39661

Peptide ligand-binding receptors

3<span

class="cmmi-12">.7 × 10-109

0<span

class="cmmi-12">.61102

Gαi signalling events

8<span

class="cmmi-12">.9 × 10-100

< 1.241 × 10-5

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

The permutation approach was then also applied to the intersection between computational and experimental candidates. Where the permutation anlaysis is testing for consistent detection of pathways independent of their pre-existing status as experimental candidates. The pathway results for these candidate partners (in Table 4.6) differed between over-representation and resampling analyses.

Namely, many of the over-represented pathways were not significant in the resampling analysis, including visual phototransduction and retinoic acid signalling, although pathways involving defective EXT1 or EXT2 genes approach significance after FDR adjustment for multiple tests. Of the highest over-represented pathways in the intersection, only Gαs signalling events were supported by both over-representation and resampling analyses. Other pathways supported by both analyses were cytoplasmic elastic fibre formation, associated HS-GAG protein modification pathways, energy metabolism, and the fibrin clotting cascade.

Table 4.6: Pathways for CDH1 partners from SLIPT and siRNA primary screen

Reactome Pathway

Over-representation

Permutation

Visual phototransduction

6<span

class="cmmi-12">.9 × 10-10

0.91116

Gαs signalling events

1<span

class="cmmi-12">.6 × 10-7

0.012988

Retinoid metabolism and transport

1<span

class="cmmi-12">.7 × 10-7

0.20487

Transcriptional regulation of white adipocyte differentiation

6<span

class="cmmi-12">.5 × 10-6

0.38197

Acyl chain remodelling of PS

6<span

class="cmmi-12">.5 × 10-6

0.58485

Chemokine receptors bind chemokines

6<span

class="cmmi-12">.5 × 10-6

0.97255

Defective EXT2 causes exostoses 2

6<span

class="cmmi-12">.9 × 10-6

0.056437

Defective EXT1 causes exostoses 1, TRPS2 and CHDS

6<span

class="cmmi-12">.9 × 10-6

0.056437

Signalling by NOTCH4

6<span

class="cmmi-12">.9 × 10-6

0.15497

Platelet activation, signalling and aggregation

6<span

class="cmmi-12">.9 × 10-6

0.53358

Phase 1 - Functionalisation of compounds

1<span

class="cmmi-12">.3 × 10-5

0.24836

Amine ligand-binding receptors

1<span

class="cmmi-12">.7 × 10-5

0<span

class="cmmi-12">.3195

Acyl chain remodelling of PE

2<span

class="cmmi-12">.4 × 10-5

0<span

class="cmmi-12">.7307

Signalling by GPCR

2<span

class="cmmi-12">.4 × 10-5

0<span

class="cmmi-12">.9939

Molecules associated with elastic fibres

2<span

class="cmmi-12">.6 × 10-5

0.0072929

DAP12 interactions

2<span

class="cmmi-12">.6 × 10-5

0.78273

Cytochrome P450 - arranged by substrate type

3<span

class="cmmi-12">.2 × 10-5

0.87019

GPCR ligand binding

3<span

class="cmmi-12">.8 × 10-5

0.99417

Acyl chain remodelling of PC

4<span

class="cmmi-12">.0 × 10-5

0.65415

Response to elevated platelet cytosolic Ca2+

4<span

class="cmmi-12">.2 × 10-5

0.55461

Arachidonic acid metabolism

4<span

class="cmmi-12">.4 × 10-5

0.060298

Defective B4GALT7 causes EDS, progeroid type

4<span

class="cmmi-12">.9 × 10-5

0.15497

Defective B3GAT3 causes JDSSDHD

4<span

class="cmmi-12">.9 × 10-5

0.15497

Elastic fibre formation

4<span

class="cmmi-12">.9 × 10-5

0.0019227

HS-GAG degradation

6<span

class="cmmi-12">.2 × 10-5

0.017747

Bile acid and bile salt metabolism

6<span

class="cmmi-12">.2 × 10-5

0.15497

Netrin-1 signalling

7<span

class="cmmi-12">.1 × 10-5

0.95056

Integration of energy metabolism

7<span

class="cmmi-12">.1 × 10-5

0.0019287

DAP12 signalling

7<span

class="cmmi-12">.9 × 10-5

0.67835

GPCR downstream signalling

8<span

class="cmmi-12">.1 × 10-5

0.88678

Diseases associated with glycosaminoglycan metabolism

8<span

class="cmmi-12">.7 × 10-5

0.017747

Diseases of glycosylation

8<span

class="cmmi-12">.7 × 10-5

0.017747

Signalling by Retinoic Acid

8<span

class="cmmi-12">.7 × 10-5

0.13592

Signalling by Leptin

8<span

class="cmmi-12">.7 × 10-5

0.15497

Signalling by SCF-KIT

8<span

class="cmmi-12">.7 × 10-5

0.73399

Opioid Signalling

8<span

class="cmmi-12">.7 × 10-5

0.99417

Signalling by NOTCH

0<span

class="cmmi-12">.0001

0.26453

Platelet homeostasis

0<span

class="cmmi-12">.0001

0.55912

Signalling by NOTCH1

0<span

class="cmmi-12">.00011

0.13797

Class B/2 (Secretin family receptors)

0<span

class="cmmi-12">.00011

0<span

class="cmmi-12">.4659

Diseases of Immune System

0<span

class="cmmi-12">.00013

0.15497

Diseases associated with the TLR signalling cascade

0<span

class="cmmi-12">.00013

0.15497

A tetrasaccharide linker sequence is required for GAG synthesis

0<span

class="cmmi-12">.00013

0.33566

Nuclear Receptor transcription pathway

0<span

class="cmmi-12">.00016

0.22735

Formation of Fibrin Clot (Clotting Cascade)

0<span

class="cmmi-12">.00016

0.0054639

Syndecan interactions

0<span

class="cmmi-12">.00016

0<span

class="cmmi-12">.3974

Class A/1 (Rhodopsin-like receptors)

0<span

class="cmmi-12">.00016

0.99454

HS-GAG biosynthesis

0<span

class="cmmi-12">.0002

0.37199

Platelet degranulation

0<span

class="cmmi-12">.0002

0.39003

EPH-ephrin mediated repulsion of cells

0<span

class="cmmi-12">.00021

0<span

class="cmmi-12">.6193

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

Many of the pathways supported in the intersection by permutation analysis were also replicated in the mtSLIPT anlaysis of partners tested with CDH1 mutation (in Table D.6), including Gαs, elastic fibres, HS-GAG, and energy metabolism. While there were differences between the pathways Identified by over-representation anlaysis, those replicated by permutation were highly concordant supported the combined use of these pathways approaches to identify synthetic lethal gene functions and targets.

While this indicates that Gαs and GPCR class A/1 signalling events were significantly detected by both approaches, GPCR signalling pathways overall were not. It is likely that GPCRs were primarily over-represented in the intersection with the experimental candidates due to strong over-representation of these pathways in experimental candidates, rather than detection by SLIPT, which may be driven by these more specific constituent pathways.

However, several pathways, including some immune functions and neurotransmitters, were supported by the resampling analysis (in Tables 4.6 and D.6) when the initial pathway over-representation test was not significant. These functions appear to have been detected by both approaches more than expected by chance but must be interpreted with caution since they were still not common enough to be detected in pathway over-representation analysis.

Therefore computational and experimental approaches to synthetic lethal screening for CDH1 lead to a broader functional characterisation and many candidate partners, when combined, despite different strengths and limitations. Compared to candidate gene approaches, experimental genome-wide screens are an appealing unbiased strategy for identifying synthetic lethal interactions. Since these screens are costly, laborious, and subject to genetic background, computational analysis can augment candidate triage to either reduce the initial panel of screened genes or prioritise validation.

GPCR pathways were detected among both computational and experimental synthetic lethal candidates, with more support in the experimental screen (Table 4.6). The homogeneous cell line model may be more likely to detect particular pathways. For instance, SLIPT identified immune pathways, not expected to be detected in isolated cell culture. GPCR signalling was supported in experimental models Telford et al. (2015) with some of these pathways replicated in varied genetic

backgrounds of patient samples. These pathways require further investigation such as identification of more specific pathways, higher order interactions, and modes of resistance.

The pathway composition across computational and experimental synthetic lethal candidates was informative with over-representation (Table 4.4) and supported by resampling analysis (Table 4.6), despite a modest intersection of genes between them (Figure 4.2). Either approach may be significant for a pathway in this intersection without being supported by the other: resampling analysis may support pathways that were not over-represented due to small effect sizes, thus both tests are required for a candidate pathway. The pathways detected by both over-representation and resampling are the strongest candidates for further investigation, such as Gαs signalling, a strong candidate in prior analyses with a role in the regulation of translation in cancer Gao and Roux (2015), another function supported by SLIPT analysis.

The predicted synthetic lethal partners occurred across functionally distinct pathways, including characterised functions of CDH1. This diversity is consistent with the wide ranging role of CDH1 in cell-cell adhesion, cell signalling, and the cytoskeletal structure of epithelial tissues. Pathway structure may be relevant to identifying potential drug targets from gene expression signatures, indicating downstream effector genes and mechanisms leading to cell inviability. These distinct synthetic lethal gene clusters and pathways may further lead to the elucidation of drug resistance mechanisms.

4.3 Metagene Analysis

Metagenes serve as a consistent signal of pathway activity. The direction of metagenes is generally arbitrary but care has been taken to ensure that these occur in a direction which reflect overall activation of the pathway (as described in Section 2.2.3). This will be supported by examining the pathway expression of gene signatures in breast cancer to ensure they behave as expected in TCGA expression data. These metagenes were also compared to somatic mutation to show the limitations of mutation as a measure of gene activity. Having established that metagenes generated with this procedure reflect gene activity, these were then applied to the Reactome pathways for synthetic lethal analysis of pathways directly to provide an alternative approach to identifying synthetic lethal pathways with CDH1.

4.3.1 Pathway expression

Pathway metagenes (generated as described in Section 2.2.3) for gene signatures of key processes in breast cancer (Gatza et al., 2011) were used to check that metagenes were generated in the correct direction to indicate pathway activation. These gene signatures were plotting in Figure 4.9 for comparison with clinical factors and somatic mutations. The “intrinsic subtype” was computed by performing the PAM50 procedure Parker et al. (2009) for RNASeq data which was highly concordant with the subtypes provided by UCSC for TCGA samples previously analysed by microarrays (TCGA, 2012). Somatic mutations were reported for recurrently mutated genes in breast cancer, as reported by TCGA (TCGA, 2012), related genes, and those previously discussed to be important in hereditary breast cancers (BRCA1, BRCA2, and CDH1 ).

Figure 4.9: Pathway metagene expression profiles. Expression profiles for metagene signatures from Gatza et al. (2011) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples were clustered independently for each intrinsic subtype and by CDH1 expression status.

These gene signatures reflect intrinsic subtypes as expected. In particular, the estrogen and progesterone receptor signatures are low in the ER- and PR- basal subtype tumours. These tumours also had the highest frequency of TP53 mutations and a corresponding reduction of p53 metagene activity, as expected for loss of a tumour suppressor. The luminal A and luminal B tumour subtypes are the most similar, which is reflected in these metagenes signatures, although they are distinguishable molecular subtypes as shown by elevated PI3K, AKT, RAS, and β-catenin signalling in luminal B tumours. Although, these pathways were also elevated in Basal and HER2-enriched subtypes and lowly expressed in the “normal-like” subtype (which contained the normal samples). These intrinsic subtype specific gene signature profiles were further supported with metagenes for an extended set of signatures (Gatza et al., 2014), as shown in Figure E.1.

TP53 mutations were the most frequent and more common in basal subtype. Similarly, GATA3 mutations were more common in luminal subtype tumours. PI3K mutations were more frequent across breast tumours, although these were less common in the basal subtype despite an elevated metagene (this discrepancy will the discussed further in Section 4.3.2). CDH1 mutations similarly occurred across molecular subtypes with the exception of the basal subtype (as observed in gene expression with Figure 4.1). CDH1 low samples occurred in all subtypes but were predominantly lobular subtype. Apart from these genes, mutations did not show clear specficity to a particular subtype and the variation between samples was reflects the range of molecular cascades that can result in tumours with similar molecular profiles, supporting the use of expression for cancer diagnostics and identification of molecular targets.

The direction of the metagenes were also consistent with the clincal characteristics and formed a consensus of gene activity as shown in Figures E.2–E.5. In each of the examples for gene signatures for PI3K (Figure E.2), p53 (Figure E.3), estrogen receptor (Figure E.4), and BRCA (Figure E.5) genes (Gatza et al., 2011, 2014), the expression of the majority of the genes were highly concordant with the metagene, being either positively or negatively correlated. These were generally consistent with established clinical and molecular subtypes of breast cancer and the recurrent mutations shown. However, the PIK3CA and PIK3R1 mutant samples did not necessarily have elevated PI3K pathway metagene activity (as shown in Figure E.2).

4.3.2 Somatic mutation

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

(a) PIK3CA mutation

(b) PIK3CA or PIK3R1 mutation

(c) CDH1 mutation

(d) TP53 mutation

Figure 4.10: Somatic mutation against PI3K metagene. Mutations in PIK3CA, PIK3R1, CDH1, and TP53 were examined in TCGA breast cancer for their effect on the PI3K (Gatza et al., 2011) pathway metagene. The tumour suppressors CDH1 and TP53 showed an increase and decrease in the metagene respectively, whereas PIK3CA and PIK3R1 mutations had little effect on the metagene levels.

It should be noted that metagenes, while consistent with the consensus of constituent expressed genes, were not neccessarily reflecting the somatic mutation status. The PI3K (Gatza et al., 2011) metagene levels in particular, were not statistically significant in between mutant and wildtype PIK3CA samples (shown in Figure 4.10). Although the PI3K metagene differed across CDH1 and TP53 mutations, remarkably in opposite directions considering that PI3K is an oncogenic growth pathway and these are both most frequently tumour suppressors inactivated in cancers. This shows that CDH1 and TP53 deficient tumours have distinct molecular growth pathways and that synthetic lethal inactivations against CDH1 inactivation may not be applicable to other cancers with driver mutations such as TP53, although these were kept in the analysis for comparison. These differences may be related to these mutations being more frequent in tumours with difference clinical characteristics (as observed in Section 4.3.1). Thus mutations do not neccessarily have corresponding changes in pathway expression, particularly for oncogenes which may change in function rather than being upregulated.

While the more specific PIK3CA (Gatza et al., 2014) metagene showed significant differences with PIK3CA and PIK3R1 mutations (as shown in Figure D.7), this metagene replicated stronger differences for CDH1 and TP53. These differences were less pronounced in the protein levels of p110α (enocded by PIK3CA) and the downstream AKT gene (shown in Figures D.8 and D.9 respectively). Although this may be due to this regulatory cascade (kinases) being transmitted as a change in protein state (phosphorylation) rather than changes in expression levels. Another consideration is that mutations at different loci have different effects on protein function, particularly for oncogenes.

4.3.3 Mutation locus

The gene locus distribution of PIK3CA and it’s receptor PIK3R1 were consistent with oncogenic and tumour suppressor mutations, as shown in Figure D.6. PIK3CA is an has recurrent mutations in 2 hotspots, centered around the E545K and H1047R (shown in Figure D.6(a)), as expected for an oncogene. This contrasts with the tumour suppressors, PIK3R1, and CDH1 (shown in Figures D.6(b) and D.6(c) respectively), which have low frequency inactivating mutations spread across them. A notable exception is TP53 (shown in Figure D.6(d)) which displays both inactivating mutations throughout and

recurrent (oncogenic) mutations at high frequency, consistent with the complex role of TP53 in cancer biology which is outside of the scope of this thesis and shown for comparison.

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

(a) PI3KCA gene

(b) PI3KCA metagene

(c) PIK3R1 gene

(d) PIK3CA protein

(e) AKT protein

(f) CDH1 protein

(g) TP53 protein

Figure 4.11: Somatic mutation locus against expression. The recurrent E545K and H1047R oncogene mutations in PIK3CA were examined in TCGA breast cancer to show the effect of mutation locus on gene, pathway, and protein expression. While neither of these mutations had an impact of PIK3CA mRNA expression, E545K had specifically lower PI3K (Gatza et al., 2011) metagene levels and both mutations had higher PIK3R1 mRNA expression. However, these differences were not reflected in the protein expression levels.

These differences in gene locus may explain why mutations do not necessarily have corresponding changes in gene or metagene expression. Specfically, the recurrent E545K and H1047R oncogene mutations in PIK3CA did not affect PIK3CA mRNA expression but E545K had specifically lower PI3K (Gatza et al., 2011) metagene levels. Both mutations had higher PIK3R1 mRNA expression but these differences differences were not reflected in the protein expression levels of p110α protein (encoded by PIK3CA), it’s downstream target AKT, E-cadherin (encoded by CDH1), or p53 (as ashown in Figure 4.11).

While the complex effects of mutation in oncogenes such as PIK3CA are not neccessarily detected in a pathway metagene, these do capture the consensus of patheay gene expression and account for other potential means of pathway activation. Thus metagenes are sufficient as a measure of gene activity for the purposes of synthetic lethal detection with SLIPT. This approach is more applicable to tumour suppressor genes with a relationship between gene expression and activity (rather than activation at the protein level) but this is not a major concern since synthetic lethality is more clinically relevant for targeting tumour suppressor mutations than oncogenes.

4.3.4 Synthetic lethal metagenes

Pathway metagenes for Reactome pathways (generated as described in Section 2.2.3) were also used for testing synthetic lethal partner pathways with CDH1 by SLIPT. Since the metagenes have are higher when the pathway as a whole is activated, they are amenable to SLIPT analysis using low metagene levels for inactivated pathways. These synthetic lethal metagenes differed to the over-represented pathways among synthetic lethal gene candidates. However, there were some similarities to previous findings, as shown in Tables 4.7. In particular, translational pathways were replicated as observed in Table 4.2. While the specific pathways differ, immune pathways (such as NF-κB) were also supported by metagene synthetic lethal analysis.

Table 4.7: Candidate synthetic lethal metagenes against CDH1 from SLIPT

Pathway

ID

Observed

Expected

χ2value

p-value

p-value (FDR)

Glycogen storage diseases

3229121

68

130

176

6.62 × 10-37

1.82 × 10-34

Myoclonic epilepsy of Lafora

3785653

68

130

176

6.62 × 10-37

1.82 × 10-34

Diseases of carbohydrate metabolism

5663084

68

130

176

6.62 × 10-37

1.82 × 10-34

Arachidonic acid metabolism

2142753

81

130

157

8.13 × 10-33

1.49 × 10-30

Translation initiation complex formation

72649

70

130

152

7.08 × 10-32

1.17 × 10-29

Synthesis of 5-eicosatetraenoic acids

2142688

68

130

151

1.25 × 10-31

1.88 × 10-29

SRP-dependent cotranslational protein targeting to membrane

1799339

69

130

150

2.01 × 10-31

2.76 × 10-29

L13a-mediated translational silencing of Ceruloplasmin expression

156827

72

130

148

5.91 × 10-31

6.44 × 10-29

3’ -UTR-mediated translational regulation

157279

72

130

148

5.91 × 10-31

6.44 × 10-29

Activation of the mRNA upon binding of the cap-binding complex and eIFs,

and subsequent binding to 43S

72662

70

130

147

1.14 × 10-30

9.28 × 10-29

Formation of the ternary complex, and subsequently, the 43S complex

72695

70

130

147

1.14 × 10-30

9.28 × 10-29

Ribosomal scanning and start codon recognition

72702

70

130

147

1.14 × 10-30

9.28 × 10-29

Eukaryotic Translation Elongation

156842

72

130

146

1.19 × 10-30

9.28 × 10-29

Nonsense Mediated Decay independent of the Exon Junction Complex

975956

71

130

146

1.24 × 10-30

9.28 × 10-29

Viral mRNA Translation

192823

70

130

146

1.51 × 10-30

1.04 × 10-28

Eukaryotic Translation Termination

72764

70

130

146

1.51 × 10-30

1.04 × 10-28

NF-kB is activated and signals survival

209560

71

130

145

1.90 × 10-30

1.19 × 10-28

Peptide chain elongation

156902

72

130

145

1.91 × 10-30

1.19 × 10-28

Influenza Life Cycle

168255

70

130

145

1.95 × 10-30

1.19 × 10-28

Formation of a pool of free 40S subunits

72689

73

130

145

2.01 × 10-30

1.19 × 10-28

Nonsense-Mediated Decay

927802

71

130

145

2.44 × 10-30

1.34 × 10-28

Nonsense Mediated Decay enhanced by the Exon Junction Complex

975957

71

130

145

2.44 × 10-30

1.34 × 10-28

GTP hydrolysis and joining of the 60S ribosomal subunit

72706

72

130

145

2.58 × 10-30

1.37 × 10-28

Influenza Viral RNA Transcription and Replication

168273

72

130

144

4.01 × 10-30

2.07 × 10-28

Signaling by NOTCH1 HD Domain Mutants in Cancer

2691230

79

130

143

5.99 × 10-30

2.82 × 10-28

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

Strongest candidate SL partners for CDH1 by SLIPT with observed and expected samples with low expression of both genes

Signalling pathways were more strongly supported by mtSLIPT analysis of metagene pathway expression against CDH1 mutation, as shown in Table D.7. Although these results were generally less statistically significant than expression analyses. Signalling pathways detected as synthetic lethal metagenes include G˙αz, insulin-related growth factor (IGF), GABA receptor, G˙αs, S6K1 and various toxin responses mediated by GPCRs. Metabolic processes including processing of carbohydrates and fatty acids were also implicated across these analyses.

The metagene analyses differ more between expresssion and CDH1 mutation than previous analyses, with more specific signalling pathways identified in the mutation analysis. This supports the usage of a complete null mutant model in experimental testing for synthetic lethality of signalling pathways against CDH1 inactivation rather than a knockdown in expression. However, low expression of partners has been used in either case to be applicable to dose-dependent pharmacological inhibition and across genes where mutations have different functional consequences, including variants of unknown significance.

These results show an independent pathway approach to detecting synthetic lethal gene functions interacting with CDH1. Synthetic lethal metagenes, replicates support for these pathways independent of pathway size (as genes are weighted equally). The synthetic lethal analysis against low CDH1 expression support prior findings in translational and immune pathways even if they were not able to detected in an experimental screen (Telford et al., 2015). Together these findings support the role of CDH1 loss in cancer disrupting cell signalling with wider effects on protein translation and metabolism necessary for the proliferation of cancer cells. This is consistent with the GPCR pathways such as G˙αs signalling being supported by SLIPT gene candidates and the experimental primary siRNA screen, as shown by resampling in Section 4.2.1.4.1.

4.4 Replication in stomach cancer

The synthetic lethal analysis of genes and pathways (previously described for TCGA breast cancer data) was replicated in TCGA stomach cancer. The accompanying data for SLIPT and mtSLIPT analyses against CDH1 expression and mutation are in Appendices F and G respectively.

The experimental screen (Telford et al., 2015) was conducted in MCF10A breast cells so it may not be as comparable to stomach cancer. Nevertheless, CDH1 is also important in stomach cancer biology as a driver tumour suppressor gene and as a germline mutation in many cases of hereditary diffuse gastric cancer.

While the sample size was lower for TCGA stomach cancer (particularly for mutations), these results serve to support the findings in breast cancer in an independent patient cohort and tissue samples. The molecular profiling, including RNA-Seq expression, were performed by TCGA using the sample procedures as for breast cancer and the findings reported here were performed used data analysis techniques identical to those presented previously. These procedures should ensure as close comparison as feasible across cancer types for those relevant to HDGC and recurrent CDH1 mutations.

4.4.1 Synthetic Lethal Genes and Pathways

The strongest SLIPT genes for stomach cancer (shown in Table F.1) did not neccessarily directly correspond to those observed in breast cancer (shown in Table 4.1). However, several gene functions were replicated in stomach cancer. Cell membrane genes including EMP3, GYPC, LGALS1, PRR24, and FUNCD2 were among the strongest SL candidates. Similarly, cell signalling genes including PLEKHO1, RARRES2, VEGFB, HSPB2, and CREM were detected in stomach cancers. It is notable that several of these genes (EMP3, PLEKHO1, and FUNCD2) have a known role in cancer. Together these genes support the roles of CDH1 in cell membrane and signalling functions (of epithelial tissues) which are perturbed in both breast and stomach cancers.

The strongest mtSLIPT genes tested against CDH1 mutatoin for stomach cancer (shown in Table G.1) supported similar gene functions. Membrane and cell-adhesion genes including KFBP6,THY1,CLELC2B, NISCH, TSPAN1,and KCTD12 and signalling genes including ZEB2, CCND2, NEURL1B, KFBP6, and OGN were detected. Similarly, these include cancer genes such as VIM,ZEB2,BCL2,THY1, and RUNX1T1. The mtSLIPT procedure also replicated several of the strongest candidates in breast cancer (shown in Table D.1) such as NRIP2 and NISCH.

Together, these gene candidates indicate widespread functions of CDH1 and strongly detectable synthetic lethality with many genes from a strategy that can be applied across cancer types. More specifically, the signalling genes included GPCR signalling genes (such as

GNG11, GNAI1, DZIP1, PTGFR, and KCTD12), a growth signalling pathway which was one of the most supported synthetic lethal pathways in breast cancer analysis, the experimental screen (Telford et al., 2015), and has many actionable drug targets which have been applied to other diseases.

These findings were further supported by the pathways over-represented in SLIPT candidates from TCGA stomach cancer (shown in Table 4.8) which were replicated the translational and immune pathways observed in TCGA breast cancer (shown in Tabel 4.2). Further support for GCPR signalling pathways including the class A/1 receptors. The extracellular matrix was also detected at the pathway level in stomach cancer SLIPT candidates and replicated in mtSLIPT analysis for CDH1 mutation (shown in Table G.2), including elastic fibres, glycosylation, collagen, and integrin cell-surface interactions. Thus there was strong evidence for the role of extracellular matrix pathways and the tumour microenvironment in CDH1 deficient stomach cancers, in addition to cell signalling and translation pathways important in tumour growth across breast and stomach cancer.

Table 4.8: Pathways for CDH1 partners from SLIPT in stomach cancer

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Extracellular matrix organization

241 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-2-3"

class="td11"> 104

7.5 × 10-140

Hemostasis

445 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-3-3"

class="td11"> 138

1.8 × 10-121

Developmental Biology

432 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-4-3"

class="td11"> 125

9.2 × 10-107

Axon guidance

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-5-3"

class="td11"> 94

1.5 × 10-102

Eukaryotic Translation Termination

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-6-3"

class="td11"> 49

1.9 × 10-99

GPCR ligand binding

373 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-7-3"

class="td11"> 108

3.8 × 10-99

Viral mRNA Translation

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-8-3"

class="td11"> 48

3.3 × 10-98

Formation of a pool of free 40S subunits

94 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-9-3"

class="td11"> 51

3.3 × 10-98

Eukaryotic Translation Elongation

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-10-3"

class="td11"> 49

1.6 × 10-97

Peptide chain elongation

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-11-3"

class="td11"> 48

7.2 × 10-97

Class A/1 (Rhodopsin-like receptors)

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-12-3"

class="td11"> 90

2.7 × 10-96

Nonsense Mediated Decay independent of the Exon Junction Complex

89 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-13-3"

class="td11"> 49

3.0 × 10-96

Infectious disease

349 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-14-3"

class="td11"> 100

2.6 × 10-94

GTP hydrolysis and joining of the 60S ribosomal subunit

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-15-3"

class="td11"> 52

3.4 × 10-94

L13a-mediated translational silencing of Ceruloplasmin expression

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-16-3"

class="td11"> 51

2.8 × 10-92

3’ -UTR-mediated translational regulation

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-17-3"

class="td11"> 51

2.8 × 10-92

Neuronal System

272 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-18-3"

class="td11"> 84

8.4 × 10-92

SRP-dependent cotranslational protein targeting to membrane

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-19-3"

class="td11"> 51

9.5 × 10-92

Eukaryotic Translation Initiation

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-20-3"

class="td11"> 52

2.0 × 10-90

Cap-dependent Translation Initiation

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-66-21-3"

class="td11"> 52

2.0 × 10-90

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for CDH1

4.4.2 Synthetic Lethal Expression Profiles

The expression profiles of candidate synthetic lethal partners dtected by SLIPT and mtSLIPT in stomach cancer were plotted against clinical characteristics as described in section for breast cancer data in Section 4.1.2 (shown in Figures 4.12 and G.1 respectively). As expected the majority of CDH1 mutant samples had low expression of CDH1 and were the diffuse type of stomach cancer.

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

Figure 4.12: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance) of all samples (separated by the 1 ∕ 3 quantile of CDH1 expression) analysed in TCGA stomach cancer dataset for gene expression of 4,365 candidate partners of E-cadherin (CDH1) from SLIPT prediction (with significant FDR adjusted p < 0.05). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or toxic in an siRNA screen Telford et al. (2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in CDH1 low samples, notably diffuse and CDH1 mutant samples have elevated expression in one or more distinct clusters, although there was less complexity and variation among candidate synthetic lethal partners than in breast data. CDH1 low samples also contained most of samples with CDH1 mutations.

The SLIPT partners of CDH1 exhibited similar clustering in staomch cancer to breast cancer, replicating the diverse roles of elevated partner genes in different clinical samples. Specifically (in Figure 4.12), the diffuse type stomach cancers had higher expression of the candidate synthetic lethal partners (where CDH1 has a role as a driver mutation), despite an unbiased clustering. This is consistent with compensating expression of synthetic lethal partners under loss of CDH1, as suggested by Lu et al. (2015). The pathway composition of gene clusters for stomach cancer (shown in Table F.2) was also highly concordant with breast cancer findings (shown in Table 4.3). These included replicated of translation (Cluster 1), immune functions (Cluster 2), G˙αs signalling (Cluster 3), and further support for the roles of GPCRs and the extracellular matrix (Cluster 4) in the synthetic lethal partners and functions of CDH1, replicated across stomach and breast cancers. Clusters 1 and 4, which had particularly high expression of SLIPT candidate partner genes in the diffuse subtype, also had the most significant over-representation of pathways.

There was less variation between the expression profiles of mtSLIPT partners of CDH1 in stomach cancer, although clusters were still detectable (as shown in Figure G.1). While the genes and pathways detected was lewss significant (due to lower sample size), the composition of clusters was further indicative for the roles of extracellular matrix (including elastic fibres), immune functions, and the cell signalling.

4.4.3 Comparison to Primary Screen

The number of genes detected by both SLIPT in TCGA stomach cancer data and siRNA in breast cell lines (shown in Figure F.1) was also not a significant overlap (as observed for breast cancer in Figure 4.2). This was particularly the case of mtSLIPT against CDH1 mutation in stomach cancer which detected very few genes (as shown in Figure G.2) due to low sample size and mutation frequency.

This smaller overlap can also be attributed to the tissue-specific differences between the stomach cancers and the breast cells used for the experimental model (Chen et al., 2014). Nevertheless, many genes were detected across SLIPT in stomach cancers and the experimental screen (Telford et al., 2015) and the pathways detected were consistent with prior observations in breast cancer. Despite differences in the specific

genes detected, the functions of CDH1 were conserved across epithetial cancers in different tissues and synthetic lethal inhibition of interacting pathways may be effective against molecular targets such as CDH1 inactivation across tissue types.

However, the pathway composition of SLIPT-specific genes and those replicated with the siRNA primary screen (Telford et al., 2015) were highly concordant between the pathways identified by SLIPT in TCGA stomach cancer (shown in Table F.3) and pathways previously identified in TCGA breast cancer (shown in Table 4.4). In both cases, translation and immune pathways were highly over-represented in SLIPT-specific genes, which we would not expect to be detected by siRNA screening in cell lines, as discussed in Section 4.2.1.4. In addition, the extracellular matrix was supported by in stomach cancer. While the pathways identified by specifically by SLIPT in stomach cancer or siRNA screening were similar to those observed for breast cancer (in Table 4.4), the pathways over-represented in the intersection for stomach cancer SLIPT candidates and the siRNA primary screen (Telford et al., 2015) also had a clear over-representation of signalling pathways, although they differed from those observed in breast cancer SLIPT candidates. GPCR signalling was supported in genes detected in both TCGA stomach cancer and screening, including G˙αq, G˙αs, serotonin receptors, and class A signalling (shown in more detail in Table F.5). In addition MAPK and NOTCH signalling pathways were detected. These replicate the findings in breast cancer and show consistent detection of signalling pathways in stomach cancer despite less genes being detected by SLIPT and patient samples differing from the tissue in which the experiments were conducted.

Similarly, the SLIPT-specific gene candidates against CDH1 mutation (shown in Table G.4) replicated pathways observed in breast cancer (shown in Table D.4), despite a lower number of genes detected. In particular, the extracellular matrix and elastic fibres were over-represented. While the number of genes overlapping with the siRNA was too low to be amenable to pathway analysis, there is further indication that members of these genes replicated across mutation SLIPT analyses include cell-membrane, elastic fibre, and GPCR signalling genes.

4.4.3.1 Resampling Analysis

Similarly, resampling for SLIPT specific candidates (shown in Tables F.4 and G.5) replicated many of the most highly over-represented pathways in stomach cancer. These

include translational, immune, GPCR signalling, and elastic fibres, consistent with previous analyses in breast cancer (shown in Tables 4.5 and D.5).

While fewer pathways were supported by resampling for the intersection of SLIPT and experimental screen (Telford et al., 2015) candidate partners in stomach cancer than breast cancer, many of those detected (shown in Table F.5) replicate those detected in breast cancer (shown in Tables 4.6 and D.6). The pathways detected by both permutation and over-representation were more likely to be replicated across stomach and breast cancer than those detected by over-representation alone, supporting the use of this procedure to detect synthetic lethal pathways applicable across cancer types. The include G˙αs signalling and elastic fibre formation as discussed for breast cancer (in Section 4.2.1.4.1)

While many pathways were detected by resampling for mtSLIPT against CDH1 mutation in stomach cancer (shown in Table G.6), there were not enough genes detected by both mtSLIPT and the siRNA primary screen to determine over-represented pathways. Therefore this may be due to small numbers of genes which does not constitute support for pathway composition. However, this under-powered analysis does not preclude the replicated synthetic lethal pathways detected across SLIPT expression analyses in TCGA breast and stomach cancer data with an accompanying siRNA primary screen (Telford et al., 2015). Rather this further supports the use of SLIPT to test against low expression of query genes as measure of gene inactivation to avoid this issue, despite mutation (which often produces similar results) being more indicative of complete gene inactivation.

4.4.4 Metagene Analysis

Metagene analysis (as conducted in Section 4.3.4) was also performed for TCGA stomach cancer expression data, using Reactome pathways. These results (as shown in Table F.6) provided further support for signalling and extracellular processes as synthetic lethal pathways across stomach and breast cancer. Namely, cell-cell communication, VEGF signalling, and various GPCR pathways were detected.

Signalling and immune pathways were also supported by mtSLIPT analysis of metagene pathway expression against CDH1 mutation, as shown in Table G.7. Although these results were generally less statistically significant than expression analyses. Signalling pathways detected as synthetic lethal metagenes include prostacyclin, SCF-KIT, ERK, MAPK, NGF, VEGF, and PI3K/AKT. The innate immune response, the inflammasome, and integrin

signaling were also implicated to be synthetic lethal with CDH1 mutations. Cell surface interactions, cholesterol biosynthesis, and platelet homeostasis also support the role of extracellular processes in proliferation of CDH1 deficient cancers and interactions of CDH1 with the extracellular environment that was not tested in the cell line experimental screen.

4.5 Global Synthetic Lethality

Global levels of synthetic lethality were analysed to address concerns raised by the high numbers of synthetic lethal candidates for CDH1. The SLIPT procedure (as described in Section 3.1) was performed with each possible query gene from the TCGA breast cancer RNA-Seq dataset. Due to the computational demands of this procedure, it was performed on the New Zealand eScience Infrastructure Intel Pan supercomputer (as described in Section 2.5.3).

The observed number of SLIPT appears to be typical for most genes in the TCGA breast RNA- Seq dataset as shown in Figure 4.13. This figure was actually lower than the majority (95%) of genes tested, although CDH1 was ranked higher for a similar in SLIPT analysis of TCGA stomach cancer data, shown in Figure H.1. The differences in sample size make these analyses difficult to compare but (in either case), the number of partners detected for CDH1 is not unexpected, eeven when adjust for multiple comparisons across candidate partners.

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

Figure 4.13: Synthetic lethal partners across query genes. Global synthetic lethal pairs were examined across the genome in TCGA breast expression data by applying SLIPT across query genes. The high number of predicted partners for CDH1 was typical for a human gene and lower than many other genes.

The number of detected candidates reported here is higher than in Figures 4.2 and F.1 because these exlcuded genes not tested by the siRNA primary screen (Telford et al., 2015) for comparison with it. For an statistically rigorous measure of global synthetic lethality, multiple comparison procedures would need to be performed for all pairs of genes tested. However, only partner genes for each query SLIPT analysis were performed for the purposes of comparing the number of partners predicted with those observed for CDH1 throughout this thesis.

4.5.1 Hub Genes

The genes with the most synthetic lethal interactions by this SLIPT analysis are the “hub” genes of a synthetic lethal network. These genes with the highest number of candidate partners detected by SLIPT in TCGA breast cancer expression data are summarised in Table 4.9. These include several genes involved in cellular signalling such as TGFBR2, PDGFRA, FAM126A, KCTD12, MAML2, and CAV1. Gene regulation including chromatin, DNA, and RNA bindings genes were also observed as hub genes such as CELF2, PLAGL1, TSHZ2, FOXO1, and SVEP1. Genes involved in the cellular membrane such as ANXA1 and FAM171A1 were also observed in addition to genes specifically implicated in cell adhesion and tight junctions such as TNS1, BOC, AMOTL1, FAT4, and EPB41L2.

Table 4.9: Query synthetic lethal genes with the most SLIPT partners

Gene

Direction

raw p-value

p-value (FDR)

SLIPT raw p-value

SLIPT (FDR)

TGFBR2

8134

17982

17973 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-2-5"

class="td11"> 8007

8006 </td>

A2M

8571

17605

17583 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-3-5"

class="td11"> 8345

8339 </td>

TNS1

8019

17949

17934 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-4-5"

class="td11"> 7874

7873 </td>

PROS1

8539

17668

17642 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-5-5"

class="td11"> 8317

8310 </td>

ANXA1

9085

17330

17302 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-6-5"

class="td11"> 8689

8682 </td>

CELF2

8665

17406

17368 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-7-5"

class="td11"> 8370

8355 </td>

BOC

8694

17371

17348 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-8-5"

class="td11"> 8384

8381 </td>

PLAGL1

8792

17361

17327 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-9-5"

class="td11"> 8448

8436 </td>

PDGFRA

8296

17650

17621 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-10-5"

class="td11"> 8095

8087 </td>

FAM171A1

8874

17560

17533 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-11-5"

class="td11"> 8567

8562 </td>

FAM126A

8510

17383

17356 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-12-5"

class="td11"> 8184

8178 </td>

TSHZ2

7942

17983

17976 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-13-5"

class="td11"> 7787

7786 </td>

KCTD12

8366

17651

17621 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-14-5"

class="td11"> 8115

8108 </td>

MAML2

8336

17537

17503 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-15-5"

class="td11"> 8069

8061 </td>

FOXO1

8027

17753

17737 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-16-5"

class="td11"> 7840

7836 </td>

AMOTL1

8425

17388

17347 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-17-5"

class="td11"> 8147

8139 </td>

FAT4

8111

17750

17732 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-18-5"

class="td11"> 7925

7919 </td>

CAV1

8645

17491

17464 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-19-5"

class="td11"> 8342

8331 </td>

SVEP1

7945

17859

17842 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-20-5"

class="td11"> 7791

7784 </td>

EPB41L2

8415

17327

17296 </td><td style="white-space:nowrap; text-align:center;" id="TBL-67-21-5"

class="td11"> 8097

8092 </td>

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

Genes with the most candidate SL partners SLIPT in TCGA breast expression data with the number of partner genes predicted by direction criteria and χ2 testing separately and combined as a SLIPT analysis. Where specified, the p-values for the χ2 test were adjusted for multiple tests (FDR).

Genes involved in adhesion and tight junctions were also hub genes in stomach cancer (shown in Table H.1) such as HEG1, FAT4, NFASC, LAMA4, LAMC1, TNS1, and AMOTL1. These also included cytoskeletal genes such as ANK2, TTC28, and MACF1. Cancer genes were also among hub genes across breast and stomach cancer such as BOC, FAT4, and MRVI1.

It is therefore unsurprising that signalling and regulatory genes have been detected throughout this thesis. Not only are they suitable targets for anti-cancer therapy, they are also highly interacting genes themselves and so it is plausible that their interactions would be detectable by SLIPT. This is consistent with the established role of abberant signalling and gene regulation in proliferation and survival of tumours and the importance of these pathways in development with highly redundant functions across many genes under complex regulation. These are also highly amenable to detection by SLIPT analysis of expression data since their functions are dynamically regulated with corresponding changes in expression.

Cytoskeletal, membrane bound, and extracellular matrix genes are also among highly interacting synthetic lethal hubs, including focal adhesion, tight junctions, microtubules, and fibronectin. These support the use of synthetic lethal interactions to target CDH1, as a tumour suppressor gene involved in these functions. Cellular structure and cell-cell interactions are thus important functions with highly redundant genes for which there are many feasible synthetic lethal interactions by which to understand regulation of cellular functions. These functions may also be exploited as vulnerabilities in cancer as they are frequently disruped in cancers, including HDGC where loss of CDH1 is a driver of cancer proliferation and malignancy.

4.5.2 Hub Pathways

Pathways over-represented among TCGA breast cancer hub genes (as shown in Table 4.10) particularly support the importance of signalling pathways, such as the PI3K/AKT pathway, as synthetic lethal hubs. The highly redundant natures of cell-cell interaction and the extracellular matrix functions was also further supported.

Table 4.10: Pathways for genes with the most SLIPT partners

Pathways Over-represented

Pathway Size

SL Genes

p-value

p-value (FDR)

Constitutive Signaling by Aberrant PI3K in Cancer

56 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-2-3"

class="td11"> 10

8.4 × 10-16

8.7 × 10-13

PI3K/AKT Signaling in Cancer

78 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-3-3"

class="td11"> 11

2.1 × 10-14

1.1 × 10-11

Role of LAT2/NTAL/LAB on calcium mobilization

96 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-4-3"

class="td11"> 12

7.7 × 10-14

2.2 × 10-11

Complement cascade

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-5-3"

class="td11"> 7

1.2 × 10-13

2.2 × 10-11

Cell surface interactions at the vascular wall

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-6-3"

class="td11"> 12

1.6 × 10-13

2.2 × 10-11

PI3K events in ERBB4 signaling

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-7-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

PIP3 activates AKT signaling

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-8-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

PI3K events in ERBB2 signaling

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-9-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

PI-3K cascade:FGFR1

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-10-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

PI-3K cascade:FGFR2

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-11-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

PI-3K cascade:FGFR3

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-12-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

PI-3K cascade:FGFR4

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-13-3"

class="td11"> 11

2.6 × 10-13

2.2 × 10-11

Extracellular matrix organization

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-14-3"

class="td11"> 22

4.7 × 10-13

3.6 × 10-11

Muscle contraction

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-15-3"

class="td11"> 9

4.9 × 10-13

3.6 × 10-11

PI3K/AKT activation

90 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-16-3"

class="td11"> 11

5.5 × 10-13

3.8 × 10-11

GAB1 signalosome

91 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-17-3"

class="td11"> 11

7.1 × 10-13

4.6 × 10-11

Smooth Muscle Contraction

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-18-3"

class="td11"> 6

2.4 × 10-12

1.5 × 10-10

Response to elevated platelet cytosolic Ca2+

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-19-3"

class="td11"> 10

2.6 × 10-12

1.5 × 10-10

Signaling by SCF-KIT

126 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-20-3"

class="td11"> 13

3.0 × 10-12

1.6 × 10-10

Signaling by FGFR

143 </td><td style="white-space:nowrap; text-align:center;" id="TBL-68-21-3"

class="td11"> 14

5.0 × 10-12

2.2 × 10-10

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in the top 500 “hub” genes with the most candidate synthetic lethal partners by SLIPT analysis of TCGA breast expression data

Pathway over-representation for synthetic lethal hub genes was replicated in TCGA stomach cancer expression data. However, these pathways differ considerably from breast cancer, as shown in Table H.2. Cell-cell interactions and extracellular matrix pathways, including elastic fibres, were also among the hub genes for stomach cancer. The signalling pathways differ as expected in a different tissue type, although BMP and PAK signalling were detected as hub gene functions.

4.6 Replication in cell line encyclopaedia

Table 4.11: Pathways for CDH1 partners from SLIPT in CCLE

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Cell Cycle

442 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-2-3"

class="td11"> 207

1.2 × 10-215

Cell Cycle, Mitotic

365 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-3-3"

class="td11"> 180

2.9 × 10-209

Signaling by Rho GTPases

311 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-4-3"

class="td11"> 136

9.4 × 10-156

M Phase

212 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-5-3"

class="td11"> 104

8.8 × 10-145

Infectious disease

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-6-3"

class="td11"> 123

1.3 × 10-142

RHO GTPase Effectors

207 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-7-3"

class="td11"> 98

5.3 × 10-135

HIV Infection

200 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-8-3"

class="td11"> 94

2 <span

class="cmsy-10x-x-120">× 10-130

Separation of Sister Chromatids

140 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-9-3"

class="td11"> 77

5.6 × 10-128

Organelle biogenesis and maintenance

258 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-10-3"

class="td11"> 107

1.4 × 10-127

Chromatin modifying enzymes

181 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-11-3"

class="td11"> 87

4.7 × 10-126

Chromatin organization

181 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-12-3"

class="td11"> 87

4.7 × 10-126

Mitotic Metaphase and Anaphase

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-13-3"

class="td11"> 78

1.2 × 10-124

Mitotic Anaphase

148 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-14-3"

class="td11"> 77

6.3 × 10-123

Developmental Biology

421 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-15-3"

class="td11"> 142

1.6 × 10-121

RHO GTPases Activate Formins

94 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-16-3"

class="td11"> 60

5.3 × 10-118

Mitotic Prometaphase

93 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-17-3"

class="td11"> 59

5.4 × 10-116

Hemostasis

421 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-18-3"

class="td11"> 138

7.2 × 10-116

Adaptive Immune System

397 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-19-3"

class="td11"> 132

3.2 × 10-115

Assembly of the primary cilium

143 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-20-3"

class="td11"> 72

2.4 × 10-114

Transcription

133 </td><td style="white-space:nowrap; text-align:center;" id="TBL-69-21-3"

class="td11"> 68

6.2 × 10-111

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for CDH1

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.

The cancer cell line encyclopaedia provides further support for synthetic lethal genes and pathways that may be applicable across cell types and reproducible in experimental systems. In contrast to the homogeneous pooled cell samples of patients, the cell lines provide a genetically homogeneous cell population in which to examine molecular functions and as a preclinical model of cancerous disease. The complete set of 1037 cell lines was tested for synthetic lethality across tissues, in addition to the 59 breast cell lines and 38 stomach cell lines being tested separately for partners of CDH1. Synthetic lethal genes were detected by SLIPT (as described in Section 3.1) and over-represented synthetic lethal Reactome pathways (as described in Section 2.3.2).

Synthetic lethal gene candidates were detectable by SLIPT across each of these sample sets of cells lines (as shown in Tables I.1–I.3. Although these were most highly significant across the samples in the CCLE expression dataset (as shown in Table I.1) and included genes detected in prior analyses such as VIM, ZEB2, EMP3. Pathways were also highly over-represented among synthetic lethal candidates for the full CCLE dataset (as shown in Table 4.11) including Rho GTPase (GPCRs), immmune, and gene regulation (chromatin and transcription). This is unexpected since immune pathways would not be expected to be detectable in isolated cell lines, although this could be attributed to cytokine and integrin signalling occuring the cancer cells in addition to interactions with immune cells in the tumour microenvironment (which could not be distinguished in patient samples). Cell cycle and mitosis were among the highest synthetic lethal pathways across cell lines supporting CDH1 deficient cells having abberant cell signalling and consequences for proliferation such as cancer cells. However, cell cycle genes were not as strongly supported in TCGA patient samples or the siRNA screen (Telford et al., 2015) and they may not be applicable to epithelial tissues such as breast or stomach cancer or amenable to selective inhibition in experimental models.

Table 4.12: Pathways for CDH1 partners from SLIPT in breast CCLE

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Cell junction organization

71 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-2-3"

class="td11"> 5

0.006 </td>

Adherens junctions interactions

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-3-3"

class="td11"> 3

0.006 </td>

Dermatan sulfate biosynthesis

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-4-3"

class="td11"> 2

0.006 </td>

Non-integrin membrane-ECM interactions

52 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-5-3"

class="td11"> 4

0.006 </td>

Regulation of pyruvate dehydrogenase (PDH) complex

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-6-3"

class="td11"> 2

0.0069 </td>

Cell-extracellular matrix interactions

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-7-3"

class="td11"> 2

0.021 </td>

Pyruvate metabolism

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-8-3"

class="td11"> 2

0.021 </td>

Cell-cell junction organization

46 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-9-3"

class="td11"> 3

0.039 </td>

Synthesis of substrates in N-glycan biosythesis

50 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-10-3"

class="td11"> 3

0.057 </td>

Detoxification of Reactive Oxygen Species

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-11-3"

class="td11"> 2

0.082 </td>

Keratan sulfate biosynthesis

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-12-3"

class="td11"> 2

0.092 </td>

Laminin interactions

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-13-3"

class="td11"> 2

0.092 </td>

Cell-Cell communication

118 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-14-3"

class="td11"> 5

0.12 </td>

Keratan sulfate/keratin metabolism

32 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-15-3"

class="td11"> 2

0.12 </td>

Opioid Signalling

63 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-16-3"

class="td11"> 3

0.12 </td>

Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide)

and transfer to a nascent protein

63 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-17-3"

class="td11"> 3

0.12 </td>

Intraflagellar transport

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-18-3"

class="td11"> 2

0.14 </td>

Signaling by Retinoic Acid

36 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-19-3"

class="td11"> 2

0.16 </td>

Pyruvate metabolism and Citric Acid (TCA) cycle

36 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-20-3"

class="td11"> 2

0.16 </td>

Nef mediated downregulation of MHC class I complex cell surface expression

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-70-21-3"

class="td11"> 1

0.22 </td>

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for CDH1

Synthetic lethal pathways specfic to SLIPT candidates from breast cell lines (as shown in Table 4.12) were more consistent with previous obervations, particularly the established role of E-cadherin in cell junctions and the Adherens complex. Although the number of SLIPT candidate genes detected in stomach cell lines was insufficient to replicate the findings in breast cell lines to TCGA patient samples. However, SLIPT candidates across breast and stomach CCLE cell lines were over-represented (as shown in Table I.5) for similar pathways to breast cell lines with additional support for extracellular matrix pathways including elastic fibres which were replicated with resampling across breast and stomach TCGA analyses and the primary siRNA screen Telford et al. (2015).

4.7 Discussion

4.7.1 Strengths of the SLIPT Methodology

Synthetic lethal discovery with SLIPT used established statistical procedures identify putative partner genes from gene expression data. Such use of the χ2-value is amenable to pathway or permutation analyses and could feasibly be applied to other disease gene or pair-wise across the genome. Although genome-wide approaches were unable to find informative candidate genes for E-cadherin Lu et al. (2015). Synthetic lethal discovery in cancer has focused on genes with severe cellular mutant phenotypes, such as essential genes or the oncogenes TP53 and AKT Lu et al. (2015); Tiong et al. (2014); Wang and Simon (2013), with other cancer genes, such as CDH1, requiring more focused investigations. Prior computational approaches for synthetic lethal discovery, in cancer, vary widely (Jerby-Arnon et al., 2014; Lu et al., 2015; Tiong et al., 2014; Wappett et al., 2016). There is no consensus for which for the approach is more appropriate and they are difficult to compare as they either do not have a released code implementation or do not make predictions solely from normalised expression data.

However, the query-based approach demonstrated by SLIPT analysis is suitable for wider application on expression data and augmenting experimental studies such as high-throughput screens. This approach has identified biologically plausible synthetic lethal

pathways for CDH1, triaged candidates from experimenal screening (Telford et al., 2015), and replicates genes and pathways across breast and somtach cancers datasets. In addition, SLIPT avoids critical assumptions underlying the design of some prior approaches such as coexpression of synthetic candidates or that they will have known (annotated) similarities in function.

The DAISY methodology Jerby-Arnon et al. (2014), which took a similar query-based approach with the tumour suppressor VHL, has been critiqued for being too stringent Lu et al. (2015) which impedes pathway analysis. Since functional redundancy does not require genes to be expressed at the same time, the SLIPT approach does not assume co-expression of synthetic lethal genes which may enrich for synthetic lethal genes in established coregulated pathways. Rather, the interpretation of synthetic lethality for SLIPT was similar to other computational methods based on ‘co-loss under-representation’, ‘compensation’, or ‘simultaneous differential expression’ Lu et al. (2015); Tiong et al. (2014); Wang and Simon (2013).

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 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. Thus analyses throughout this thesis have focused on querying for partners of a particular gene of interest. Statistical modelling and simulations (in Section 3.3 and Chapter 6) will further support the design decisions underlying SLIPT analysis and it’s strengths over other approaches.

4.7.2 Syntheic Lethal Pathways for E-cadherin

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-representation) mutual 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 lethal detection is error-prone and identifying genes relevant for clinical application will be difficult

without a supporting biological pathway rationale. As such, investigations into the genes identified by SLIPT, correlation structure between them, and those which were validated by experimental screening (Telford et al., 2015) focused on the pathway level throughout this Chapter. Similarly, comparisons across analyses were largely made at the pathway level, including comparisons between expression and mutation, breast and stomach TCGA datasets, and patient sample data with cell line expression profiles.

Potential synthetic lethal partners of CDH1 identified by SLIPT had many distinct functions, with each gene cluster highly expressed in different patient subgroups (Figure 4.1). The expression profiles of the SL partners of CDH1 predicted from the TCGA breast cancer RNA-Seq data (expected to have compensating high or stable expression) and their corresponding functional enrichment found subgroups of genes with functional organisation particularly among CDH1 low breast tumours. 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 estrogen receptor 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.

The pathways that synthetic lethal partners of CDH1 identified by SLIPT were involved in a diverse range of biological functions and differed to those detected experimentally. This discrepancy may be accounted for by gene expression analyses detecting both synthetic lethal partners (as screened for experimentally Telford et al. (2015)) and their downstream targets (not detected by siRNA), capturing the wider pathways and mechanisms involved in synthetic lethality with CDH1 inactivation. In particular, GPCR phosphorylation cascades (which regulates gene expression and translation in cancers Gao and Roux (2015)) were predicted to be synthetic lethal with CDH1. The predicted synthetic lethal partners occurred across functionally distinct pathways, including characterised functions of CDH1. The most consistently supported pathways include elastic fibres in the extracelullar matrix, GPCR signalling, and translation presenting vulnerabilities for CDH1

deficient cancer cells from extracellular stimuli to the core growth mechanisms of a cell.

This diversity in synthetic lethal functions is consistent with the wide ranging role of CDH1 in cell-cell adhesion, cell signalling, and the cytoskeletal structure of epithelial tissues. Pathway structure may be relevant to identifying potential drug targets from gene expression signatures, indicating downstream effector genes and mechanisms leading to cell inviability. Identification of distinct synthetic lethal gene clusters may further lead to the elucidation of drug resistance mechanisms. While these pathways are indicative of the main functions of E-cadherin and synthetic lethal partners, it remains to identify the genes within these pathways that are the most actionable or supported across SLIPT analysis in patient samples and detected by experiments in preclincal models (Chen et al., 2014; Telford et al., 2015). The specific genes within key pathways will be be discussed in Chapter 5, along with further investigations into their relation to pathway structure. 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.

These sythetic lethal pathways have potential clinical implications, particularly those supported in pre-clinical models and in patient expression data. However, further validation of gene candidates will be neccessary to ensure that these are able to reproduced in further pre-clincal studies, they are applicable to tumours in vivo, and that effective inhibitory agents can be repurposed or designed against them.

4.7.3 Replication and Validation

4.7.3.1 Integration with siRNA Screening

The pathway composition across computational and experimental synthetic lethal candidates was informative with over-representation (Table 4.4) and supported by resampling analysis (Table 4.6), despite a modest intersection of genes between them (Figure 4.2). Either approach may be significant for a pathway in this intersection without being supported by the other: resampling analysis may support pathways that were not

over-represented due to small effect sizes, thus both tests are required for a candidate pathway.

The pathways detected by both over-representation and resampling are the strongest candidates for further investigation and the pathway structure analyses in Chapter 5 will focus on these pathways detected by both over-representation and resampling. Particularly, those replicated across datasets or with pathway metagenes. In addition to GCPR pathways detected across these analyses, the PI3K cascade will also be investigated in Chapter 5, this signalling pathway is a well characterised mediator between GCPR receptors and regulation of translation (Gao and Roux, 2015) (both detected throughout this Chapter) and exhibited unexpected behaviour with pathway the metagenes (in Section 4.3). This pathway is activated by protein Phosphorylation states and thus inactivatino may not be detectable with expression.

However, the SLIPT approach was shown to be predictive of which siRNA primary screen candidate partners of CDH1 were validated in a secondary screen (as shown in Appendix C). These results further support SLIPT for identifying robust synthetic lethal candidates which can be validated and as a triage approach for interpreting screening experiments.

4.7.3.2 Replication across Tissues and Cell lines

Furthermore, synthetic lethal partners identified by SLIPT were replicated across breast and stomach cancer. These were particularly concordant at the pathway level, as expected between tissues since synthetic lethal pathways have higher conservation between species (Dixon et al., 2008). These findings support gene functions conserved across CDH1 deficient cancers in breast and stomach tissues, presenting vulnerabilities that could be applied against molecular targets in both cancers. In addition, these analyses serve as a replication across independent patient cohorts from breast and stomach cancers, decreasing the likelihood of the synthetic lethal pathways detected being false positives or artifacts of either dataset.

Synthetic lethal pathways were also replicated across expression analyses of TCGA patient samples in heterogeneous tumours and homogeneous cell line isolates. This further supports that the subset of synthetic lethal functions detectable in experimental models (Chen et al., 2014; Telford et al., 2015) would be applicable tumours of patients with CDH1 deficient cancers.

There are many gene functions replicated across breast cancer gene expression analyses. Many of these were also replicated with mutation analysis and with stomach cancer or cell line expression data. These pathways were more consistent across replication analyses than previous investigations with TCGA microarray data Kelly (2013).

4.8 Summary

We have developed a simple, interpretable, computational approach to predict synthetic lethal partners from genomics data. The analyses focus on gene expression data as it is widely available for applications in other cancers and other disease genes, particularly those with malignant loss of function.

This approach has been applied to robustly detect synthetic lethal pathways for the E-cadherin (CDH1) in TCGA breast cancer molecular profiles with comparisons to experimenal screening (Telford et al., 2015) in cell lines, and replication in TCGA stomach cancer molecular profiles and across cell types in the cancer cell line encyclopaedia. The pathway replicated across several analyses included extracellular matrix pathways (such as elastic fibres formation), cell signalling (including GPCRs), and core gene regulation and translation processes crucial for the growth and proliferation of cancer cells. These pathways show evidence of non-oncogene addiction for CDH1 deficient cells and present vulnerabilities which may be exploited for specific treatment against CDH1 mutations in HCGC and sporadic cancers. There was also support for synthetic lethal pathways with CDH1 in cell adhesion and cytoskeletal processes to which CDH1 belongs, supporting the finding that synthetic lethality occurs within biological pathways (Boone et al., 2007; Kelley and Ideker, 2005).

While translational and immune pathways detected by SLIPT were not supported by primary siRNA screening (Telford et al., 2015), these were replicated across various analyses. Due to the differences between an experimenal cell line model (Barretina et al., 2012; Chen et al., 2014; Fece de la Cruz et al., 2015) and patient molecular profiles (Bass et al., 2014; TCGA, 2012), these would not be expected to be completely concordant. Furthermore, many pathways are difficult to test in an isolated experimental system. Nevertheless, many of the genes and pathways detected by SLIPT are suitable to inform further investigations and triage of therapeutic targets against CDH1 deficient tumours in combination with experimenal screening.

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. Cell surface

interactions, the extracellular matrix, and cell signalling (particularly PI3K/AKT signalling) were also found to be synthetic lethal hubs with more interactions detected than other genes. This indicates that these pathways are functionally important to survival of cancer cells since they are subject to high functional redundancy, despite frequent disruptions in cancer. These pathways being involved in a disproportionate number of synthetic lethal interactions is also consistent with their detection for CDH1.

Thus synthetic lethal pathways have been identified using TCGA patient molecular profiles, CCLE cancer cell line expression data, and experimenal screening results. Some these were robustly replicated across these datasets and against CDH1 mutation or expression analysis. However, there remains the need to identify actionable genes within these pathways, relationships with experimental candidates, and how these pathways may affect viability when lost. While the genes identified between these analyses were less concordant the results of the TCGA breast cancer analysis will be used to test pathway structure relationships and further examine the synthetic lethal genes detected in the following Chapter.

Aims

Pathway Structure of Candidate Synthetic Lethal Genes for CDH1 from TCGA breast data

Comparisons to Experimental siRNA Screen Candidates

Replication of Pathways across in TCGA Stomach data

Summary

We have developed a Synthetic Lethal detection method that generates a high number of synthetic lethal candidates

Pathways in cell signalling, extracellular matrix, and cytoskeletal functions were supported with experimental candidates and the known functions of E-cadherin

Several candidate pathways were supported by mutation analysis and replicated across breast and stomach cancer

Translation and immune functions were uniquely detected by the computational approach which may be explained by differences between patient samples and cell line models

There remains the need to identify actionable genes within these pathways, relationships with experimental candidates, and how these pathways may affect viability when lost

Chapter 5Synthetic Lethal Pathway Structure

Having identified key pathways implicated in synthetic lethal genetic interactions with CDH1, these were investigated for the underlying synthetic lethal genes within them and their relationships to pathway structure in Reactome pathways. This chapter will focus on the pathway structure of biological pathways detected across analyses in Chapter 4. The synthetic lethal genes considered here are those candidates detected by SLIPT (as described in Section 3.1) in TCGA breast cancer expression and mutation data (TCGA, 2012) in comparison to the candidate gene partners from the siRNA screening in breast cell lines (Telford et al., 2015).

The graph structure for Reactome pathways was obtained from Pathway Commons via BioPAX (as described in Section 2.4.2). The pathways describe the (directional) relationships between biomolecules, including proteins (encoded by genes), in biological pathways. These relationships include cell signalling (such as kinase phosphorylation cascades), gene regulation (such as transcription factors, chromatin modifiers, RNA binding proteins), and metabolism (such as the product of an enzyme being the substrate of another). Together these relationships describe the known functional pathways in a human cell with a reasonable resolution, from a curated database supported by publications documenting pathway relationships. While this functional pathway network encapsulates protein complexes and functional modules, protein binding or text-mining alone are not used to determine relationships between genes. The Reactome network is sufficient to test pathway relationships with directional information, although it is not documented whether these relationships are activating or inhibitory.

Pathway structures were derived from the Reactome network (as described in Section 2.4.3) for the graph structure of each biological pathway. The synthetic lethal candidate genes for notable pathways discussed in Chapter 4, including candidate synthetic lethal pathways of CDH1, were examined to show the SLIPT and siRNA candidates within these pathways. Thus synthetic lethal genes were identified within a biological context and with further investigations into their relationship with pathway structure and between synthetic lethal candidates detect by each approach. Synthetic lethal gene candidates in the context of pathway structures and additional support for belonging to a synthetic lethal pathway are ideal for triage of targets specific to CDH1 deficient tumours and for further experimental validation in preclinical models.

Network analysis metrics (as described in Sections 2.4.4 and 3.5.3) were applied to test whether gene detected by SLIPT, siRNA, or both approaches varied according to these network analysis metrics (of connectivity and importance in the network) to test whether they differed between synthetic lethal genes or approaches to detect them. Another consideration is the relationships between synthetic lethal candidates detected by either approach: these were tested by both a resampling approach (as described in Sections 3.4.1 and  3.4.1.1) and compared across a ranking based on biological context (Section 3.4.1.2). Together these approaches serve to test the pathway relationships between SLIPT and siRNA synthetic lethal gene candidate partners for CDH1 within the biological pathways identified and demonstrate a combination of network biology and statistical investigations into structural relationships between genes identified by a Bioinformatics analysis.

5.1 Synthetic Lethal Genes in Reactome Pathways

5.1.1 The PI3K/AKT Pathway

The phosphoinositide 3-kinase (PI3K) cascade signalling pathway exhibited unexpected results with metagene analyses (as discussed in Section 4.3). This pathway is also of interest because mediating signals between the G protein coupled receptors and regulation of protein translation which have both been strongly implicated to be synthetic lethal pathways with loss of CDH1 function. All three of these pathways have are also subject to dysregulation in cancer and other diseases. Thus the PI3K cascade will be examined along with the most supported synthetic lethal pathways (as identified in Chapter 4).

The PI3K pathway is also an ideal pathway to test pathway structure since it has an established direction of signal transduction from extracellular stimuli (and membrane bound receptors) to the inner mechanisms of the cell, namely the regulation of protein translation. The production of proteins is neccessary for the growth of the cell so it is reasonable to suggest that these processes may be subject to (non-oncogene) addiction in some cancer cells which rely upon them for sustained protein production and cell growth. This is also supported by the oncogenes PIK3CA and AKT1 being involved with the PI3K cascade and related PI3K/AKT pathway which may be subject to oncogene addiction when these proto-oncogenes are activated.

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

Figure 5.1: Synthetic Lethality in the PI3K Cascade. The Reactome PI3K Cascade pathway with synthetic lethal candidates coloured as shown in the Legend.

Despite the PI3K cascade not being supported across SLIPT and siRNA analysis by over-representation (in Section 4.2.1.4) or resampling (in Section 4.2.1.4.1), numerous genes were detected by either SLIPT in TCGA breast expression data or the siRNA primary screen (as shown in Figure 5.1). It is also notable, that of the few genes that were identified by both approaches, these include genes that are highly conencted in the PI3K cascade and are hubs to information transmission such as FGF9,PDE3B, and PDE4A. The key upstream genes PIK3CA and PIK3CG were detected by siRNA whereas the downstream PIK3R1 and AKT2 genes were detected by SLIPT. Gene detected by either method were also prevalent in the PI3K, phosphodiesterase (PDE), and AMP-activated protein kinase (AMPK) modules, in addition to the downstream translation factors and ribosomal genes (EIF4B, EEF2K, and RPS6). Together these suggest that there may further be structure between the SLIPT and siRNA candidates partners of CDH1 in pathways such as this example. As such, pathway structure will be tested to detect differences in the upstream and downstream gene candidates of those detected by either method. This may further explain the disparity between SLIPT and siRNA genes, even in pathways such as PI3K where they did not significantly intersect.

This disparity between SLIPT and siRNA gene candidate synthetic lethal partners of CDH1, that is a high number of genes detected by either approach with few detected by both, was replicated the related PI3K/AKT pathway and the “PI3K/AKT in cancer” pathway (shown in Figures J.1 and J.2). With many synthetic lethal candidates at the upstream core of these pathway networks and the downstream extremities. It is particularly notable that the many genes important in cell signalling and gene regulation were detected by either sytnhetic lethal detection approach. These include AKT1, AKT2, and AKT3, the Calmodulin signalling genes CALM1 and CAMK4, and the forkhead family transcription factors FOXO1 (a tumour suppressor) and FOXO4 and inhibitor of EMT.

5.1.2 The Extracellular Matrix

The extracellular pathways elastic fibre formation and fibrin clot formation (shown in Figures 5.2 and 5.3 respectively) were both supported across analyses (in Chapter 4). This includes a significant over-representation and resampling the interaction between SLIPT (for TCGA breast cancer) and siRNA gene candidates showing that SLIPT has

identified these pathways in addition to their over-representation in the siRNA screen.

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

Figure 5.2: Synthetic Lethality in the Elastic Fibre Formation Pathway. The Reactome Elastic Fibre Formation pathway with synthetic lethal candidates coloured as shown in the Legend.

Particularly for elastic fibres (in Figure 5.2), the vast majority of genes were detected by either approach in addition to a significant proportion of genes detected by both approaches (as determined in Section 4.2.1.4). The genes detected by both approaches also appeared to have a non-random distribution in the network with TFGB1, ITGB8, and MFAP2 exhibiting high connectivity and a cental role in their respective pathway modules. In addition to a structural role in the extracellular matrix and connective tissue (including the tumour microenvironment), these proteins including Furin, transforming growth factor β (TGFβ), and the bone morphogenic proteins (BMPs), are also involved in responses to endocrine signals and interacting with the cellular receptors for signalling pathways. Therefore it is plausible that CDH1 deficient tumours will be subject to non-oncogene addiction to the extracellular environment and growth signals arising from this pathway. The pathway structure is also worth further investigation into whether the genes detected by siRNA or both approaches are downstream of those detected by SLIPT in addition to whether they have higher connectivity or centrality than other genes in the pathway.

Genes detected as synthetic lethal partners of CDH1 by SLIPT or siRNA screening were also common in the Fibrin clot formation pathway (shown in Figure 5.3). This is consistent with the established pleiotropic role of CDH1 in regulating fibrin clotting. It is also notable that the genes detected by either method appear to be highly connected such as C1QBP KNG1, F8, F10, F12, F13A, and PROC (including many of the coagulation factors). Synthetic lethal candidates also include SERPINE2 and PRCP, which only affect downstream genes, in addition to PROCR and VWF, which are only affected by upstream genes.

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

Figure 5.3: Synthetic Lethality in the Fibrin Clot Formation. The Reactome Fibrin Clot Formation pathway with synthetic lethal candidates coloured as shown in the Legend.

Many of these genes are involved in the larger Extracellular Matrix pathway (shown in Figure J.3), including many of the synthetic lethal candidates discussed for elastic fibres. The number of SLIPT candidate genes outnumbers those identified by siRNA as expected from an isolated cell model. However, the endocrine response genes (such as TGFB1 and LTBP4 ) which are potentially artifacts of the cell line growth process were replicated with SLIPT analysis in patient tumours (TCGA breast cancer data). There is also additional support for synthetic elthal genes such as ITGB2, MFAP2, and SPARC being highly connected networks hubs of the pathway. Although the complexity of extracellular matrix pathway lends credence to the need for formal network analysis approaches to aid interpretation of the structure and relationships among synthetic lethal candidates in a pathway network, in addition to statistical approaches to determine whether such relationships are unlikely to be observed by sampling error.

5.1.3 G Protein Coupled Receptors

Similarly, G protein coupled receptor (GPCR) pathways are highly complex (as shown in Figures J.4 and J.5). Many of these were synthetic lethal candidates by eith SLIPT pr siRNA screening with many detected with both approaches, consistent with these pathways being supported by prior analyses (in Sections 4.2.1.4 and 4.2.1.4.1). Synthetic lethal candidates include the PDE and Calmodulin genes (as discussed in Section 5.1.3) in addition to others such as the regulators of G-protein signaling (RGS), chemokine receptors (CXCR), Janus kinase (JAK), and the Ras homolog family (RHO) genes. These are important regulatory signalling pathways necessary for cellular growth and cancer proliferation. Thus the GPCR pathways (and downstream PI3K/AKT signals) are a potentially actionable vulnerability against CDH1 deficient cancers, particularly since many existing drug targets exist among these signalling pathways, some of which have been experimentally validated (Kelly et al., 2017b; Telford et al., 2015). However, the complexity of GPCR networks containing hundreds of genes requires the relationships between SLIPT and experimental candidates to be tested with a network based statistical approach, although a statistically significant intersection of these approaches has been established (in Sections 4.2.1.4 and 4.2.1.4.1).

5.1.4 Gene Regulation and Translation

While very few synthetic lethal genes were detected in translational pathways in an experimental screen against CDH1 Telford et al. (2015), these were highly over-represented in translational elongation (as shown in Figure J.6). These SLIPT genes include many ribosomal proteins and the regulatory “elongation factors” which may be subject to responses in the upstream signalling pathways. This observation lends support the notion of pathway structure among synthetic lethal candidates detected by SLIPT in comparison with siRNA as the computational approach with SLIPT has demonstrated the ability to detect downstream genes in the core translational processes which experimental screening did not identify. Although it is possible that the experimental screening may detect upstream regulatory genes less sensitive inactivation, that is genes which are less likely to be indiscriminately lethal to both genotypes at high doses of inactivation.

Many of these SLIPT candidate genes are also among the nonsense-mediated decay (NMD) pathway (shown in Figure J.7) or 3′ untranslated region (UTR) mediated translational regulation (shown in Figure J.8). While genes in these pathways were also supported by experimental screening with siRNA, there was clear pathway structure. In particular, UPF1 was detected in the siRNA screen and is the focal downstream gene for the entire NMD pathway showing that (in this case) siRNA genes are downstream effectors of those detected by SLIPT. 3′ UTR mediated translational regulation has a similar structure with two modules connected solely by RPL13A, giving an example of SLIPT candidates genes with high connectivity, although there were many ribosomal proteins detected by SLIPT. However, EIF3K a regulatory elongation factor (not essential to ribosomal function) that was detected by SLIPT was replicated with siRNA screening while the majority of the elongation factors were not detected by either approach. Regulatory genes being more amenable to experimental validation also support further investigation into pathway structure as the SLIPT candidates may support them by structural relationships and the downstream genes not being detectable by experimental screening with high dose inhibitors may explain the greater number of SLIPT candidate partners of CDH1 than those experimentally identified.

5.2 Network Analysis of Synthetic Lethal Genes

Genes detected as synthetic lethal partners of CDH1 with the SLIPT computational approach and the siRNA screen (Telford et al., 2015) were compared across network metrics in the example of the PI3K cascade pathway (where the genes differed considerably between synthetic lethal detection methods). These were used to test whether network metrics differed between groups of genes detected by either or both approaches. These analyses serve to both test whether synthetic lethal gene candidates had higher connectivity or importance in a network and to whether either detection approach is constrained to genes with different network properties.

5.2.1 Gene Connectivity and Vertex Degree

Vertex degree (the number of connections) for each gene is a fundamental property of a network. The vast majority of genes had a relatively modest number of connections each with only a few genes in the PI3K pathway (shown in Figure 5.4) having pathway relationships with a high number of genes, consistent with the scle-free property of biological networks Barabási and Oltvai (2004). There were few differences in the number of connections between gene groups (by synthetic lethal detection). Although genes detected by siRNA included those with the fewest connections. The median connectivity of genes detected by both approaches was marginally higher.

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

Figure 5.4: Synthetic Lethality and Vertex Degree. The number of connected genes (vertex degree) was compared (on a log-scale) across genes deteced by SLIPT and siRNA screening in the Reactome PI3K cascade pathway. There were very few differences in vertex degree between the groups, although genes detected by siRNA included those with the fewest connections.

The results for the PI3K pathway were very similar when testing synthetic lethality against CDH1 mutation (mtSLIPT). In this case, there is also indication that mtSLIPT-specific genes may have higher connectivity than those detected by siRNA screening (shown in Figure K.1).

However, these apparent differences in vertex degree may be due to fewer genes being detected by either approach. There was no statistically significant effect of either computational or experimental synthetic lethal detection method on vertex degree, as determined by analysis of variance (ANOVA) (shown by Tables 5.1 and  K.1). Thus synthetic lethal detection does not discriminate among genes by their connectivity in a pathway network, nor is either approach constrained by a genes connectivity. Both approaches have been demonstrated to detect genes with many and very few connections.

Table 5.1: ANOVA for Synthetic Lethality and Vertex Degree

</td><td style="white-space:nowrap; text-align:center;" id="TBL-72-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-72-2-4"

class="td11"> 15.50

0.0134

0.9082

SLIPT

1

506 </td><td style="white-space:nowrap; text-align:center;" id="TBL-72-3-4"

class="td11"> 506.01

0.4378

0.5105

siRNA×SLIPT

1

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-72-4-4"

class="td11"> 0.05

0.0000

0.9947

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

Analysis of variance for vertex degree against synthetic lethal detection approaches (with an interaction term)

5.2.2 Gene Importance and Centrality

5.2.2.1 Information Centrality

Information centrality is a measure of the importance of nodes in a network by how vital they are to the transmission of information throughout the network. This naturally applies well to biological pathways, partcularly gene regulation and cell signalling. The nodes with the highest information centrality are not necessarily the most connected as they may also include nodes which pass signals between highly connected network hubs. Information centrality therefore provides a distinct metric for the connectivity of a gene in a pathway, which has the added benefit of being directly related to the disruption of pathway function were it to be inactivated or removed.

Information centrality has also been suggested to indicate essentiality of genes or proteins (Kranthi et al., 2013). The information centrality for was computed across the entire Reacomte network (as discussed in Appendix L). Reactome contains substrates and cofactors in addition to genes or proteins. In support of centrality as a measure of essentiality or importance to the network, a number nodes with the highest centrality (shown in Table L.1) were essential nutrients including Mg2+, Ca2+, Zn2+, and Fe.

Genes important in development of epithelial tissues and breast cancer were also detected with relatively high information centrality (as shown by the distribution across the Reactome network in Figure L.1). Interleukin 8 (encoded by IL8) is a chemokine important in epithelial cells, the innate immune system, and binding GPCRs. GATA4 is a embryonic transcription factor involved in heart development, EMT, and was reccurently mutated in in breast cancer (TCGA, 2012). β-catenin (encoded by the proto-oncogene CTNNB1) is a regulatory protein which binds E-cadherin, being involved in cell-cell adhesion and Wingless-related integration site (WNT) signalling. Together these show that information cetrality identifies nodes of importance to biological functions in pathway networks, including those relevant to CDH1 deficient breast cancers.

Within the PI3K pathway (shown in Figure 5.5), genes detected by siRNA did not include those with lower centrality, although the median information centrality across gene groups detected by either synthetic lethal approach did not differ. The gene with the highest information centrality (AKT2) was detected by SLIPT and was markedly higher than the other genes in the pathway which is consistent with the known biological role of AKT in PI3K/AKT signalling and the pathway structure (shown in Figure 5.1). The information centrality of the PI3K pathway was 1.338433.

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

Figure 5.5: Synthetic Lethality and Centrality. The information centrality was compared (on a log-scale across genes deteced by SLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by siRNA had higher connectivity than many genes not detected by either approach. The gene with the highest centrality was detected by SLIPT.

Table 5.2: ANOVA for Synthetic Lethality and Information Centrality

</td><td style="white-space:nowrap; text-align:center;" id="TBL-73-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

0.000256

0.0002561

0.1854

0.6682

SLIPT

1

0.003827

0.0038275

2.7717

0.1008

siRNA×SLIPT

1

0.000804

0.0008036

0.5820

0.4483

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

Analysis of variance for information centrality against synthetic lethal detection approaches (with an interaction term)

These findings were replicated (shown in Figure K.2) when testing synthetic lethality against CDH1 mutation (mtSLIPT). The differences in network centrality between gene groups detected by either method were not statistically significant as determined by ANOVA (shown by Tables 5.2 and  K.2). Thus neither method was unable to detect synthetic lethal genes with particular centrality constraints, although they were also not detecting genes with higher centrality than expected by chance.

5.2.2.2 PageRank Centrality

PageRank centrality is another network analysis procedure to infer a hierarchy of gene importance from a network using connections and structure (Brin and Page, 1998). In constrast to the information centrality approach of removing nodes, PageRank uses the eigenvalue properties of the adjacency matrix to rank genes according to the number of connections and paths they are involved in.

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

Figure 5.6: Synthetic Lethality and PageRank. The PageRank centrality was compared (on a log-scale across genes deteced by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by siRNA had a more restricted range of centrality values (which may be constrained experimental detection in a cell line model) than other genes not detected by either approach, although these groups also had fewer genes and a higher median.

This distinction is immediately clear within the PI3K pathway (shown in Figure 5.6), which differs considerably from the information centrality scores. While genes not detected by either method had the highest centrality, genes detected by SLIPT span the complete range of PageRank centrality values for this pathway. This was replicated (shown in Figure K.3) when testing synthetic lethality against CDH1 mutation (mtSLIPT). Thus SLIPT is not biased towards genes with more crucial role in the pathway as inferred by pathway connectivity and centrality measures and it is therefore independent of pathway structure. However, the genes detected by siRNA screening have a higher median PageRank centrality, although the differences in PageRank centrality between these methods were not statistically significant as determined by ANOVA (shown by Tables 5.2 and  K.2).

Table 5.3: ANOVA for Synthetic Lethality and PageRank Centrality

</td><td style="white-space:nowrap; text-align:center;" id="TBL-74-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

0.0002038

2.0385 × 10-4

1.1423

0.2892

SLIPT

1

0.0000208

2.0752 × 10-5

0.1163

0.7342

siRNA×SLIPT

1

0.0000137

1.3743 × 10-5

0.0770

0.7823

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

Analysis of variance for PageRank centrality against synthetic lethal detection approaches (with an interaction term)

5.3 Testing Pathway Structure of Synthetic Lethal Genes

5.3.1 Hierarchical Pathway Structure

5.3.1.1 Contextual Hierarchy of PI3K

A contextual hierarchy of genes in the PI3K pathway was performed (as described in in Secion 3.4.1.2) to assign scores for their relative order in the pathway. In the case of PI3K (shown in Figure 5.7), this orders genes from the upstream genes which respond to signals from extracellular stimuli to the downstream genes which transmit these to the gene expression (translation) responses of the cell. The directionality of this pathway is evident in transmitting signals from the PI3K complex, via AKT, PDE, and mTOR to the ribosomal regulatory proteins. This hierarchical procedure enables testing whether the biological context of a gene in a pathway is relevant to detection as a synthetic lethal candidate by either computational SLIPT analysis or experimental siRNA screening.

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

Figure 5.7: Structure of PI3K Ranking. Structure of PI3K Ranking.

5.3.1.2 Testing Contextual Hierarchy of Synthetic Lethal Genes

This pathway hierarchy in the PI3K cascade was tested for differences between genes detected across SLIPT and siRNA screening. The synthetic lethal candidates for CDH1 detected by either method (as shown by Figure 5.8) did not differ, each being distributed throughout the pathway. The SLIPT candidate genes were more numerous, there was little indication that they are more frequently upstream or downstream of siRNA candidate genes (as shown by Figure 5.9). Although SLIPT genes included more with a lower (upstream) hierarchy. Synthetic lethal candidates from both methods were less frequently detected in the downstream effectors of the pathway (such as the mTOR complex), although core pathway genes (such as AKT2 and PDE3B) were detectable as synthetic lethal candidates (as discussed for Figure 5.1).

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

Figure 5.8: Synthetic Lethality and Hierarchy Score in PI3K. The hierarchical distance scores were similarly distributed across SLIPT and siRNA genes.

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

Figure 5.9: Hierarchy Score in PI3K against Synthetic Lethality in PI3K. The number of SLIPT and siRNA genes against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

Similarly, when testing synthetic lethality against CDH1 mutation (mtSLIPT), the hierarchical score for the PI3K pathway did not differ between mtSLIPT-specific and siRNA-specific gene candidates (as shown by Figure M.1). Although the median among genes detected by both approaches was elevated, that is further downstream in the pathway that other synthetic lethal candidates partners of CDH1. This distinction is particularly notable since there were fewer genes overall with higher scores (shown in Figure M.2), while these are more frequently detected by both mtSLIPT and siRNA.

However, there was no significant effect variation in pathway hierarchy (shown by ANOVA in Tables 5.4 and M.1) accounted for by SLIPT or siRNA detection in the PI3K pathway (as shown in Figure 5.1). Thus such differences in hierarchical scores may be observed by sampling variation and there is no indication that SLIPT or siRNA detection differs along the direction of the pathway. Genes detected by either method are no more or less common among upstream or downstream of the pathway.

Table 5.4: ANOVA for Synthetic Lethality and PI3K Hierarchy

</td><td style="white-space:nowrap; text-align:center;" id="TBL-75-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

0.001 </td><td style="white-space:nowrap; text-align:center;" id="TBL-75-2-4"

class="td11"> 0.00066

0.0004

0.9842

SLIPT

1

0.456 </td><td style="white-space:nowrap; text-align:center;" id="TBL-75-3-4"

class="td11"> 0.45605

0.2740

0.6016

siRNA×SLIPT

1

0.019 </td><td style="white-space:nowrap; text-align:center;" id="TBL-75-4-4"

class="td11"> 0.01878

0.0113

0.9156

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

Analysis of variance for PI3K hierarchy score against synthetic lethal detection approaches (with an interaction term)

The pathway hierarchy may be applied here. A χ2-test was performed for the SLIPT or siRNA candidate genes upstream or downstream of each gene. It is unsurprising that these χ2 tests were more significant when the gene used as a threshold was in the middle of the pathway (as shown in Figure 5.10). However, there was no statistically significant support for pathway structure by this approach as none of the χ2 values were high enough to detect pathway structure between SLIPT and siRNA gene candidates. Nor was structure detectable for mtSLIPT testing synthetic lethality against CDH1 mutation (as shown in Figure M.3).

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

Figure 5.10: Structure of Synthetic Lethality in PI3K. The number of SLIPT and siRNA genes upstream or downstream of each gene in the Reactome PI3K pathway were tested (by the χ2-test). These are plotted as a split violin plot against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

5.3.2 Upstream or Downstream Synthetic Lethality

However, this does not ascertain whether SLIPT and siRNA candidate partners of CDH1 are upstream or downstream of one and other within a pathway such as the PI3K cascade. The hierarchical approach is designed to detected differences in pathway location between gene groups. An alternative pathway structure method has been devised to use network structures to identify directional relationships between individual SLIPT and siRNA genes. This pathway structure methodology will be applied (as described in Section 3.4.1) to detect the direction of shortest paths between SLIPT and siRNA gene candidates. This will be used to demonstrate the methodology on the PI3K pathway, to develop a statistical test for pathway structure between between SLIPT and siRNA gene candidate using resampling (as described in Section 3.4.1.1, and to apply this test for pathway structure among synthetic lethal gene candidates to the pathways identified in Chapter 4 and discussed in Section 5.1.

5.3.2.1 Measuring Structure of Candidates within PI3K

Shortest paths in a pathway network were used to devise a strategy to detect pathway structure between SLIPT and siRNA gene candidate partners of CDH1 (asdescribed in Section 3.4.1). Thus we can determine whether individual SLIPT genes have upstream or downstream siRNA candidates (scored as “up” or “down” events respectively). This procedure enables the detection of directional relationships between SLIPT and siRNA gene candidates (in contrast to the hierarchical approach).

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

Figure 5.11: Structure of Synthetic Lethality Resampling in PI3K. A null distribution with 10,000 iterations of the number of siRNA genes upstream or downstream of SLIPT genes (depicted as the difference of these) in the PI3K pathway. To assess significance, the observed events (with shortest paths) were compared to the 90% and 95% intervals for the null distribution (shown in violet). Genes detected by both methods were fixed to the same number as observed for the alternative null distribution (shown in blue), although the observed number of events (red) was not significant in either case. In both cases, these genes detected by both approaches were included in computing the number of shortest paths (in either direction) between SLIPT and siRNA genes.

The total number of gene candidate pairs in either direction can be compared within a pathway network to assess the overall directional relationships in a pathway. This directionality is detectable by the difference between the number SLIPT candidate genes with upstream and downstream siRNA gene partners. However, this measure alone is not sufficient to determine whether there is evidence of pathway structure between SLIPT and siRNA gene candidate partners of CDH1 in a pathway network. Although it does serve to measure the magnitude (and direction) of the consensus of directional relationships (upstream and downstream) between SLIPT and siRNA gene candidate partners. This measure of pathway structure can be used for testing for statistical significance of pathway structure by resampling, using a permutation procedure to test whether these relationships are detectable among randomly selected gene groups rather than the detected SLIPT and siRNA gene candidate partners (as described in Sections 2.3.6 and 3.4.1.1).

This resampling procedure was performed for the PI3K network (as shown in Figure 5.1) which generated a null distribution for the difference in the number of “up events” and “down events” for this Pathway. This provides a distribution to test whether more genes detected by SLIPT had upstream or downstream siRNA candidates. While there was modest indication that glssiRNA genes were downstream of SLIPT candidate genes, resampling for the PI3K pathway (as shown in Figure 5.11) did not detect a significant number of siRNA genes upstream or downstream.

In contrast, when testing synthetic lethality against CDH1 mutation (mtSLIPT) there was modest indication that glssiRNA genes were upstream of SLIPT candidate genes. However, resampling (as shown in Figure M.4) was also unable to detect a significant number of siRNA genes upstream or downstream of mtSLIPT candidates. Fixing the number of genes detected by both approaches (as shown by the blue line in Figures 5.11 and M.4) did not alter the findings of this approach. Nor did excluding these jointly detected genes, although these were included in the analysis since they can disproportionately count towards siRNA genes being upstream (or downstream) of SLIPT genes since they have different proportions of gene detected by either approach upstream (or downstream) of them. Furthermore, expanding the range of shortest paths to consider links in related pathways (using the “metapathways” constructed in Section 2.4.3) also had little effect on the null distribution generated, despite increasing the computational demands of the procedure.

5.3.2.2 Resampling for Synthetic Lethal Pathway Structure

The permutation procedure (as described in Section 3.4.1.1) that was performed in Section 5.3.2.1 for the PI3K cascade was also applied to other pathways identified in Chapter 4 and discussed in Section 5.1. These include extracellular matrix (with constituent elastic fibre and fibrin pathways), cell signalling (by PI3K/AKT and GCPRs), and translational pathways (with NMD and 3′UTR regulation). The resampling results across these pathways (as shown in Table 5.5) had limited support for pathway structure, with the majority of these being non-significant as shown for PI3K (in Figure M.4). However, the distribution for these pathways will differ depending on their structure, the number of genes they consist of, and the proportion of synthetic lethal candidates among them (including a higher frequency of genes detected by both methods pathways identified in Sections 4.2.1.4.1 and  4.4.3.1). This resampling is an appropriate procedure to use to detect structural relationships across pathways as it does not assume an underlying test statistic distribution.

Pathway structure was supported for the NMD pathway (which is consistent with siRNA being downstream in Figure J.7). However, this observation rest upon a single gene and was not replicated when testing synthetic lethality (mtSLIPT) against CDH1 mutation (as shown in Table M.2) or supported by the related 3′UTR regulation and translational elongation pathways.

Table 5.5: Resampling for pathway structure of synthetic lethal detection methods

<table id="TBL-76" class="tabular"

cellspacing="0" cellpadding="0" rules="groups" >

</td><td colspan="2" style="white-space:nowrap; text-align:center;" id="TBL-76-2-2"

class="td11">

Graph

</td><td colspan="2" style="white-space:nowrap; text-align:center;" id="TBL-76-2-4"

class="td11">

States

</td><td colspan="4" style="white-space:nowrap; text-align:center;" id="TBL-76-2-6"

class="td11">

Observed

</td><td colspan="2" style="white-space:nowrap; text-align:center;" id="TBL-76-2-10"

class="td11">

Permutation p-value

Pathway

Nodes

Edges

SLIPT

siRNA

Up

Down

Up-Down

Up∕Down

Up-Down

Down-Up

PI3K Cascade

138

1495

38

25

122

128

-6 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-4-9"

class="td11"> 0.953

0.5326

0.4606

PI3K/AKT Signaling in Cancer

275

12882

98

44

779

679

100 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-5-9"

class="td11"> 1.147

0.3255

0.6734

Gαi Signaling

292

22003

95

58

836

1546

-710 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-6-9"

class="td11"> 0.541

0.9971

0.0029

GPCR downstream

1270

142071

312

160

9755

9261

494 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-7-9"

class="td11"> 1.053

0.3692

0.6305

Elastic fibre formation

42

175

24

7 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-8-6"

class="td11"> 1

2

-1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-8-9"

class="td11"> 0.500

0.5461

0.3865

Extracellular matrix

299

3677

127

29

547

455

92 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-9-9"

class="td11"> 1.202

0.3351

0.6636

Formation of Fibrin

52

243

18

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-10-6"

class="td11"> 12

17

-5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-10-9"

class="td11"> 0.706

0.6198

0.3564

Nonsense-Mediated Decay

103

102

74

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-11-6"

class="td11"> 0

74

-74 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-11-9"

class="td11"> 0

1.0000

> 0.0001

3’ -UTR-mediated translational regulation

107

2860

77

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-12-6"

class="td11"> 0

0

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-12-9"

class="td11">

0.4902

0.5027

Eukaryotic Translation Elongation

92

3746

76

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-13-6"

class="td11"> 0

0

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-76-13-9"

class="td11">

0.4943

0.4933

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

Pathways in the Reactome network tested for structural relationships between SLIPT and siRNA genes by resampling. The raw p-value (computed without adjusting for multiple comparisons over pathways) is given for the difference in upstream and downstream paths from SLIPT to siRNA gene candidate partners of CDH1 with significant pathways highlighted in bold. Sampling was performed only in the target pathway and shortest paths were computed within it. Loops or paths in either direction that could not be resolved were excluded from the analysis. The gene detected by both SLIPT and siRNA (or resampling for them) were includued in the analysis and the number of these were fixed to the number observed.

There does not appear to be a consensus on the directionality of SLIPT and siRNA candidates across pathways as distinct pathways showed stronger tendency for siRNA genes to be either upstream or downstream. Even related pathways such as PI3K and PI3K/AKT signalling showed directional events in opposite directions. The strongest pathway (among those tested) with support for directional pathways structure is Gαi signaling which showed significant downstream siRNA genes for both SLIPT and mtSLIPT from a large number of shortest paths (in Tables 5.5 and  M.2).However, these results are borderline significant (with raw permutation p-values) and are unlikely to be detected after adjusting for multiple comparisons.

Therefore, there is insufficent evidence to determine whether there is pathway structure between the SLIPT and siRNA candidates observed in many pathways. In particular, directional structure among synthetic lethal candidates for CDH1 was not strongly supported in signalling pathways upon which the rationale for pathway structure hypotheses were based on. Despite the design of a robust resampling approach to test relationships between gene groups, this did not detect many structural relationships between SLIPT and siRNA gene candidates, although it may apply more broadly to gene networks. Furthermore, the pathway relationships are unlikely to be statistically supported by resampling when testing across the search space of Reactome pathways and adjusting for multiple comparisons. While there is statistically significant over-representation of many of these pathways in gene detected by both SLIPT and siRNA (as described in Chapter 4), these did not show pathway structure, nor does pathway structure account for the discrepancy between SLIPT and siRNA gene candidates which did not significantly intersect such as the PI3K cascade.

5.4 Discussion

Synthetic lethal genes and pathways (for CDH1 loss in cancer) were identified across gene expression and mutation datasets in Chapter 4. These pathway structure investigations extend those investigations into synthetic lethal gene candidates including exploring the discrepancy between SLIPT and siRNA candidates genes in a pathway such as PI3K in which they did not significantly intersect. Pathways with replicated synthetic lethal genes across these detection methods, breast and stomach cancer data, and patient and cell line

data were also investigated including pathways from the extracellular microenvironment to core translational pathway and the signalling pathways which mediate between them.

Many genes were detected by either method and the differences between the computational and experimental screening approaches could feasibly lead to differences in which genes within a synthetic lethal pathway are identified. Genes detected by synthetic lethal detection strategies included those biological importance within synthetic lethal pathways, those which are actionable drug targets, and those with functional implications for the biological growth mechanisms or vulnerabilities of CDH1 deficient tumours. It appeared that genes detected by both approaches were highly connected (or of importance) in the network structure or some pathways and that there may be some structure with SLIPT and siRNA upstream or downstream of each other. However, the complexity of biological pathways meant that they are not reliably interpretable so formal mathematical and computational approaches are needed to analyse large biological networks.

Network analysis techniques were therefore applied to formalise and quantify the connectivity and importance (centrality) of genes within pathways (using PI3K as an example). However, these network techniques were unable to identify distinct differences in the network properties of genes detected as synthetic lethal candidates by computational or experimental methods. These network metrics support the application of synthetic detection across pathways (and the findings using pathways as gene sets in Chapter 4) as neither synthetic lethal detection approach was biased towards genes of higher importance or connectivity and neither approach was insensitive to genes of lower importance or connectivity.

Similarly, a network hierarchy based on biological context (ordered from receiving extracellular stimuli to affecting downstream gene expression and cell growth) was devised to test whether PI3K genes of a particular upstream or downstream level were more frequently detected as synthetic lethal candidates. However, this approach was unable to ascertain whether genes detected by either method were further upstream or downstream in the pathway and there was no statistical evidence that either method differed in which levels of this structure were detected.

A measure of pathway structure between individual SLIPT and siRNA genes within a pathway was also devised using the direction of shortest paths in a directed graph structure. This is amenable to detecting the consensus directionality of the pathway across pairs of genes detected by either method. The pathway structure methodology

developed here is generally applicable to comparison of node groups (allowing overlapping) including genes in biological pathways and their detection by different methodologies. While the pathway structure measure alone is not able to detect structural relationships between gene groups (such SLIPT and siRNA gene candidates), it is amenable to resampling to determine whether these relationships are statistically significant.

5.5 Summary

Together these analyses of biological pathways, network metrics, and statistical procedures devised specifically for purpose were applied to Reactome pathway structures to test whether structural relationships exist between synthetic lethal candidates. Of particular interest was whether these relationships be related to the differences between the computational (SLIPT) and experimental (siRNA) synthetic lethal candidate partners of CDH1 (in the pathways discussed in Chapter 4).

While biologically relevant relationships were observed in specific pathways, there were not detectable structural relationships between SLIPT and siRNA gene candidates. These candidates did not exhibited significant differences in network connectivity or centrality measures. Network analyses were also unable to ascertain whether the candidates detected by either method stratified into upstream and downstream genes on the pathway and they likely do not.

A statistical resampling procedure was applied to shortest paths to test whether pairs of SLIPT and siRNA gene candidates were more likely to be upstream or downstream of each other. This approach detected very few structural relationships in the synthetic lethal pathways identified in Chapter 4. Overall, support for pathway structure between SLIPT and siRNA gene candidates is weak and the direction is inconsistent between pathways. Therefore pathway structure does not account for the differences between the SLIPT and siRNA gene candidates, although this does support the validity of gene set analyses in Chapter 4 and the synthetic lethal pathways identified.

Furthermore, the resampling procedure demonstrated in this Chapter is more widely applicable to gene states in network structures and may be further utility in the analysis of biological pathway or networks. This approach was able to quantify structural relationships that were otherwise difficult to interpret and to conclusively exclude many potential

relationships. In this respect, the network resampling methdology may also be applicable to triage of experimental validation.

Aims

Synthetic Lethal Genes within a Biological Pathway Structure

Importance and Connectivity of Synthetic Lethal Genes within Pathway Networks

Upstream and Downstream Relationships between SLIPT and siRNA Candidates

Summary

Synthetic Lethal genes were explored within a graph structures for key pathways identified previously

In some cases these graph structures appeared to have relationships between synthetic lethal genes

However, no existing network metrics of importance and connectivity with the networks were elevated significantly for Synthetic Lethal genes

Nor was there significant evidence of upstream and downstream relationships between SLIPT and siRNA Candidates in a shortest path permutation analysis

Chapter 6Simulation and Modeling of Synthetic Lethal Pathways

Simulation and modelling of synthetic lethality in gene expression will be revisited in greater detail in this chapter, building upon the results provided to support the use of SLIPT in Section 3.3. A simulation procedure for generating simulated data with underlying (known) synthetic lethal partners of a query gene, such as CDH1, was developed (as described in Section 3.2.2) by sampling from a Multivariate normal distribution based on a statistical model of synthetic lethality in expression data (as described in Section 3.2.1). This simulation framework was applied to simulated data (in Section 3.3), including simple correlation structures to assess the statistical performance of the SLIPT methodology and support it’s use a computational approach for detecting synthetic lethal candidates from expression data throughout this thesis (in Chapters 4 and 5).

While this basic framework was sufficent to support the use of SLIPT in prior Chapters, further investigations with simulations were conducted to assess the strengths and limitations of the SLIPT methodology, compare it to alternative statistical approaches to synthetic lethal detection, and assess it’s performance upon more complex correlation structures. Together these simulation investigations assess the performance of the SLIPT methdology, including on pathway graph structures (such as those discussed in Chapter 5) and determine whether the SLIPT methdology (or similar refined bioinformatics strategies) are statistically rigourous or suitable for wider genomics applications.

These simulation investigations continue to utilise the Multivariate Normal procedure (as applied in Section 3.3) with further refinements. The SLIPT methodology (and it’s equivalent χ2 test) were applied across a range of parameters (including altering the quantiles for detecting synthetic lethal direction and compared correlation. This was also applied to with query correlated genes (as performed in Section 3.3).

A refined simulation procedure was developed specifically to extend the simulation procedure (described in Section 3.2) to utilise pathway graph structures for the correlation structures of simulated datasets (as described in Section 3.4.2). This methdology can be applied to simulated correlation structures across simple graph structures to test specfic network modules or use pathway structures based on biological pathways (as discussed in Chapter 5). Thus graph structure and simulation approaches were combined to test whether a gene locus in a pathway affects detection by SLIPT and whether SLIPT performance is affected by pathway structure. The simulation procedure based on graph structures were applied in a computational pipeline across many parameters

with high-performance computing (as discussed in Section 2.5.3) and the core simulation functions have been released as a software package for wider use to test bioinformatics and statistical methods on graph structures (as described in Section 3.5.3).

6.1 Comparing methods

Methods were compared …

6.1.1 Performance of SLIPT and χ2 across Quantiles

Text

6.1.1.1 Correlated Query Genes affects Specificity

Text

6.1.2 Correlation as a Synthetic Lethal Detection Strategy

Text

6.1.3 Testing for Bimodality with BiSEp

Text

6.2 Simulations with Graph Structures

Text

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

(a) Activating pathway structure

(b) Pathway structure including inhibitions

Figure 6.1: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

(a) Activating pathway structure

(b) Pathway structure including inhibitions

Figure 6.2: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

(a) Activating pathway structure

(b) Pathway structure including inhibitions

Figure 6.3: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

(a) Activating pathway structure

(b) Pathway structure including inhibitions

(c) Pathway structure including inhibitions

Figure 6.4: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

Figure 6.5: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

Figure 6.6: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

Figure 6.7: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

Figure 6.8: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

Figure 6.9: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

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

Figure 6.10: Simulated graph structures. A constructed graph structure used as an example to demonstrate the simulation procedure. Activating links are denoted by blue arrows and inhibiting links by red edges.

6.2.1 Performance over a Graph Structure

Text

6.2.2 Synthetic Lethality across Graph Stuctures

Text

6.2.3 Performance with inhibition links

Text

6.2.4 Performance with 20,000 genes

Text

6.3 Simulations over pathway-based graphs

Text

6.4 Discussion

Text

6.5 Summary

Text

Aims

A Model of Synthetic Lethal Genes in Gene Expression Data

Comparison of SLIPT to Alternative Approaches

Simulations of Known Synthetic Lethal Genes within Pathway Networks

Summary

We have designed a straight-forward rational query-based synthetic lethal detection method with the example of application to CDH1 in cancer gene expression

I have developed a simulation pipeline to generate continuous gene expression with pathway structure including a procedure to simulate synthetic lethality

The simulation procedure shows that SLIPT is robust across pathway structures and has desirable performance compared to other statistical techniques

Chapter 7Discussion

Aims

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

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

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

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

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

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

Summary

We have developed a Synthetic Lethal detection method that generates a high number of synthetic lethal candidates

Pathways in cell signalling, extracellular matrix, and cytoskeletal functions were supported with experimental candidates and the known functions of E-cadherin

Several candidate pathways were supported by mutation analysis and replicated across breast and stomach cancer

Translation and immune functions were uniquely detected by the computational approach which may be explained by differences between patient samples and cell line models

There remains the need to identify actionable genes within these pathways, relationships with experimental candidates, and how these pathways may affect viability when lost

Synthetic Lethal genes were explored within a graph structures for key pathways identified previously

In some cases these graph structures appeared to have relationships between synthetic lethal genes

However, no existing network metrics of importance and connectivity with the networks were elevated significantly for Synthetic Lethal genes

Nor was there significant evidence of upstream and downstream relationships between SLIPT and siRNA Candidates in a shortest path permutation analysis

We have designed a straight-forward rational query-based synthetic lethal detection method with the example of application to CDH1 in cancer gene expression

We have developed a simulation pipeline to generate continuous gene expression with pathway structure including a procedure to simulate synthetic lethality

Our simulation procedure is robust across pathway structures and has desirable performance compared to other statistical techniques

7.1 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.

7.2 Future Directions

Such a bioinformatically-informed synthetic lethal screening and validation strategy could be integrated into existing and future screens for synthetic lethality in cancer.

Possible improvements to the SLIPT method include developing a Bayesian inference method or simulations and modelling to account for pathway structure among synthetic lethal genes. Another extension would be to test for higher order synthetic lethal interactions, where 3 or more genes perform a redundant function.

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).

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.

7.3 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 8Conclusion

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

   Burk, R.D., Chen, Z., Saller, C., Tarvin, K., Carvalho, A.L., Scapulatempo-Neto,

C., Silveira, H.C., Fregnani, J.H., Creighton, C.J., Anderson, M.L., et al. (2017) Integrated genomic and molecular characterization of cervical cancer. Nature, 543(7645): 378–384.

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

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

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

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

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

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

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

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

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

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

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

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

   Christofori, G. and Semb, H. (1999) The role of the cell-adhesion molecule E-cadherin as a tumour-suppressor gene. Trends in Biochemical Sciences, 24(2): 73 – 76.

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

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

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

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

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

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

   Collisson, E., Campbell, J., Brooks, A., Berger, A., Lee, W., Chmielecki, J., Beer, D., Cope, L., Creighton, C., Danilova, L., et al. (2014) Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511(7511): 543–550.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

consortium data portala one-stop shop for cancer genomics data. Database: The Journal of Biological Databases and Curation, 2011: bar026.

   Zhong, W. and Sternberg, P.W. (2006) Genome-wide prediction of c. elegans genetic interactions. Science, 311(5766): 1481–1484.

   Zweig, M.H. and Campbell, G. (1993) Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. Clinical Chemistry, 39(4): 561–577.

Appendix ASample Quality

A.1 Sample Correlation

Samples were excluded from expression analysis based on sample correlations and the clustering analysis presented below, as described in Section 2.2.2.

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

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

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

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

A.2 Replicate Samples in TCGA Breast

Replicate samples were picked where possible from the TCGA breast cancer gene expression data to examine for sample quality. Independent samples of the same tumour are expected to have very high Pearson’s correlation between their expression profiles unless there were issues with sample collection or preparation and are thus an indicator of sample quality. The log-transformed raw read counts for replicate samples were examined in Figures A.3–A.6. These were examined before normalisation which would be expected to increase sample concordance.

Another consideration are the samples which were removed for quality concerns (in Section 2.2.2). While these were selected by unbiased hierarchical clustering (See Figure A.2), it is notable that many of the exluded (tumour) samples were performed in replicate despite relatively few replicate samples in the overall dataset. These samples correlate poorly with the rest of the dataset, in addition to with replicate samples.

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

Figure A.3: Replicate excluded samples. Both tumour samples of patient A2QH were excluded as they were poorly correlated with other samples, although they are highly similar to each other as shown by Pearson’s correlation of log-raw counts.

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

(a) Remaining triplet

(b) Remaining triplet

(c) Remaining triplet

(d) Remaining paired samples

(e) Remaining paired samples

(f) Remaining paired samples

Figure A.4: Replicate samples with all remaining. Patient A26J was sampled 3 times and compared pairwise. Pairs of samples were also compared for other patients with replicate samples. In all cases, replicate samples remaining in the dataset were highly concordant as shown by Pearson’s correlation of log-raw counts.

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

(a) Remaining

(b) Compare with excluded

(c) Compare with excluded

(d) Remaining

(e) Compare with excluded

(f) Compare with excluded

(g) Remaining

(h) Compare with excluded

(i) Compare with excluded

Figure A.5: Replicate samples with some excluded. Patients A0DB, A13D, A13E, and A26E were each sampled 3 times and compared pairwise. Pairs of samples were also compared for other patients with replicate samples. In all cases, the replicate samples remaining in the dataset more were highly concordant (as shown by Pearson’s correlation of log-raw counts) than those excluded from the analysis.

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

(j) Remaining

(k) Compare with excluded

(l) Compare with excluded

(m) Compare with excluded

(n) Compare with excluded

Figure A.6: Replicate samples with some excluded. Patients A0DB, A13D, A13E, and A26E were each sampled 3 times and compared pairwise. Pairs of samples were also compared for other patients with replicate samples. In all cases, the replicate samples remaining in the dataset more were highly concordant (as shown by Pearson’s correlation of log-raw counts) than those excluded from the analysis.

Appendix BSoftware Used for Thesis

Table B.1: R Packages used during Thesis

Package

Repository

Laptop

Lab

Server

NeSI

base

base

3.3.2

3.3.2

3.3.1

3.3.0

abind

CRAN

1.4-5

1.4-3

acepack

CRAN

1.4.1

1.3-3.3

ade4

CRAN

1.7-5

annaffy

Bioconductor

1.46.0

AnnotationDbi

Bioconductor

1.36.0

1.36.0

1.34.4

apComplex

CRAN

2.40.0

ape

CRAN

4

3.4

arm

CRAN

1.9-3

assertthat

CRAN

0.1

0.1

0.1

0.1

backports

CRAN

1.0.5

1.0.4

1.0.5

1.0.2

base64

CRAN

2

2

base64enc

CRAN

0.1-3

0.1-3

beanplot

CRAN

1.2

1.2

1.2

BH

CRAN

1.60.0-2

1.62.0-1

1.62.0-1

1.60.0-2

Biobase

Bioconductor

2.34.0

2.34.0

2.32.0

BiocGenerics

Bioconductor

0.20.0

0.20.0

0.18.0

BiocInstaller

Bioconductor

1.24.0

1.20.3

1.22.3

BiocParallel

Bioconductor

1.8.1

1.8.1

Biostrings

Bioconductor

2.42.1

2.42.0

BiSEp

Bioconductor

2.0.1

2.0.1

2.0.1

bitops

CRAN

1.0-6

1.0-6

1.0-6

1.0-6

boot

base

1.3-18

1.3-18

1.3-18

1.3-18

brew

CRAN

1.0-6

1.0-6

1.0-6

1.0-6

broom

CRAN

0.4.1

caTools

CRAN

1.17.1

1.17.1

1.17.1

1.17.1

cgdsr

CRAN

1.2.5

checkmate

CRAN

1.8.2

1.7.4

chron

CRAN

2.3-47

2.3-48

2.3-50

2.3-47

class

base

7.3-14

7.3-14

7.3-14

7.3-14

cluster

base

2.0.5

2.0.5

2.0.5

2.0.4

coda

CRAN

0.19-1

0.18-1

codetools

base

0.2-15

0.2-15

0.2-15

0.2-14

colorRamps

CRAN

2.3

colorspace

CRAN

1.2-6

1.3-2

1.3-2

1.2-6

commonmark

CRAN

1.1

1.2

compiler

base

3.3.2

3.3.2

3.3.1

3.3.0

corpcor

CRAN

1.6.8

1.6.8

1.6.8

Cprob

CRAN

1.2.4

crayon

CRAN

1.3.2

1.3.2

1.3.2

1.3.2

crop

CRAN

0.0-2

0.0-2

curl

CRAN

1.2

2.3

2.3

0.9.7

d3Network

CRAN

0.5.2.1

data.table

CRAN

1.9.6

1.10.0

1.10.1

1.9.6

data.tree

CRAN

0.7.0

0.7.0

datasets

base

3.3.2

3.3.2

3.3.1

3.3.0

DBI

CRAN

0.5-1

0.5-1

0.5-1

0.5-1

dendextend

CRAN

1.4.0

1.4.0

1.4.0

DEoptimR

CRAN

1.0-8

1.0-8

1.0-8

1.0-4

desc

CRAN

1.1.0

1.1.0

devtools

CRAN

1.12.0

1.12.0

1.12.0

1.12.0

DiagrammeR

CRAN

0.9.0

0.9.0

dichromat

CRAN

2.0-0

2.0-0

2.0-0

2.0-0

digest

CRAN

0.6.10

0.6.11

0.6.12

0.6.9

diptest

CRAN

0.75-7

0.75-7

0.75-7

doParallel

CRAN

1.0.10

1.0.10

1.0.10

1.0.10

dplyr

CRAN

0.5.0

0.5.0

0.5.0

0.5.0

ellipse

CRAN

0.3-8

0.3-8

0.3-8

evaluate

CRAN

0.1

0.1

0.9

fdrtool

CRAN

1.2.15

fields

CRAN

8.1

flexmix

CRAN

2.3-13

2.3-13

2.3-13

forcats

CRAN

0.2.0

foreach

CRAN

1.4.3

1.4.3

1.4.3

1.4.3

foreign

base

0.8-67

0.8-67

0.8-67

0.8-66

formatR

CRAN

1.4

1.4

1.4

Formula

CRAN

1.2-1

1.2-1

fpc

CRAN

2.1-10

2.1-10

2.1-10

futile.logger

CRAN

1.4.3

1.4.3

1.4.1

futile.options

CRAN

1.0.0

1.0.0

1.0.0

gdata

CRAN

2.17.0

2.17.0

2.17.0

2.17.0

geepack

CRAN

1.2-1

GenomeInfoDb

Bioconductor

1.10.2

1.10.1

GenomicAlignments

Bioconductor

1.10.0

1.10.0

GenomicRanges

Bioconductor

1.26.2

1.26.1

ggm

CRAN

2.3

ggplot2

CRAN

2.1.0

2.2.1

2.2.1

2.1.0

git2r

CRAN

0.15.0

0.18.0

0.16.0

0.15.0

glasso

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-91-4"

class="td11"> 1.8

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-91-6"

class="td11">

GO.db

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-92-4"

class="td11"> 3.4.0

3.2.2

3.3.0

GOSemSim

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-93-4"

class="td11"> 2.0.3

1.28.2

1.30.3

gplots

CRAN

3.0.1

3.0.1

3.0.1

3.0.1

graph

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-95-4"

class="td11"> 1.52.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-95-6"

class="td11">

graphics

base

3.3.2

3.3.2

3.3.1

3.3.0

graphsim

GitHub

TomKellyGenetics

0.1.0

0.1.0

0.1.0

0.1.0

grDevices

base

3.3.2

3.3.2

3.3.1

3.3.0

grid

base

3.3.2

3.3.2

3.3.1

3.3.0

gridBase

CRAN

0.4-7

0.4-7

0.4-7

0.4-7

gridExtra

CRAN

2.2.1

2.2.1

2.2.1

2.2.1

gridGraphics

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-102-4"

class="td11"> 0.1-5

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-102-6"

class="td11">

gtable

CRAN

0.2.0

0.2.0

0.2.0

0.2.0

gtools

CRAN

3.5.0

3.5.0

3.5.0

3.5.0

haven

CRAN

1.0.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-105-5"

class="td11">

</td>

heatmap.2x

GitHub

TomKellyGenetics

0.0.0.9000

0.0.0.9000

0.0.0.9000

0.0.0.9000

hgu133plus2.db

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-107-4"

class="td11"> 3.2.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-107-6"

class="td11">

highr

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-108-4"

class="td11"> 0.6

0.6

0.6

Hmisc

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-109-4"

class="td11"> 4.0-2

4.0-2

3.17-4

hms

CRAN

0.2

0.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-110-6"

class="td11">

htmlTable

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-112-4"

class="td11"> 1.8

1.9

</td>

htmltools

CRAN

0.3.5

0.3.5

0.3.5

0.3.5

htmlwidgets

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-114-4"

class="td11"> 0.8

0.8

</td>

httpuv

CRAN

1.3.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-115-5"

class="td11"> 1.3.3

</td>

httr

CRAN

1.2.1

1.2.1

1.2.1

1.1.0

huge

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-117-4"

class="td11"> 1.2.7

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-117-6"

class="td11">

hunspell

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-118-4"

class="td11"> 2.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-118-6"

class="td11"> 2

hypergraph

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-119-4"

class="td11"> 1.46.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-119-6"

class="td11">

igraph

CRAN

1.0.1

1.0.1

1.0.1

1.0.1

igraph.extensions

GitHub

TomKellyGenetics

0.1.0.9001

0.1.0.9001

0.1.0.9001

0.1.0.9001

influenceR

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-122-4"

class="td11"> 0.1.0

0.1.0

</td>

info.centrality

GitHub

TomKellyGenetics

0.1.0

0.1.0

0.1.0

0.1.0

IRanges

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-124-4"

class="td11"> 2.8.1

2.8.1

2.6.1

irlba

CRAN

2.1.1

2.1.2

2.1.2

2.0.0

iterators

CRAN

1.0.8

1.0.8

1.0.8

1.0.8

jpeg

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-127-4"

class="td11"> 0.1-8

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-127-6"

class="td11">

jsonlite

CRAN

1.1

1.2

1.3

0.9.20

KEGG.db

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-129-4"

class="td11"> 3.2.3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-129-6"

class="td11">

kernlab

CRAN

0.9-25

0.9-25

0.9-25

</td>

KernSmooth

base

2.23-15

2.23-15

2.23-15

2.23-15

knitr

CRAN

1.15.1

1.15.1

1.14

labeling

CRAN

0.3

0.3

0.3

0.3

lambda.r

CRAN

1.1.9

1.1.9

1.1.7

lattice

base

0.20-34

0.20-34

0.20-34

0.20-33

latticeExtra

CRAN

0.6-28

0.6-28

lava

CRAN

1.4.6

lavaan

CRAN

0.5-22

lazyeval

CRAN

0.2.0

0.2.0

0.2.0

0.2.0

les

CRAN

1.24.0

lgtdl

CRAN

1.1.3

limma

Bioconductor

3.30.7

3.30.3

lme4

CRAN

1.1-12

1.1-12

lubridate

CRAN

1.6.0

magrittr

CRAN

1.5

1.5

1.5

1.5

maps

CRAN

3.1.1

markdown

CRAN

0.7.7

0.7.7

0.7.7

MASS

base

7.3-45

7.3-45

7.3-45

7.3-45

Matrix

base

1.2-7.1

1.2-7.1

1.2-8

1.2-6

matrixcalc

CRAN

1.0-3

1.0-3

1.0-3

1.0-3

mclust

CRAN

5.2

5.2.1

5.2.2

5.2

memoise

CRAN

1.0.0

1.0.0

1.0.0

1.0.0

methods

base

3.3.2

3.3.2

3.3.1

3.3.0

mgcv

base

1.8-16

1.8-16

1.8-17

1.8-12

mi

CRAN

1

mime

CRAN

0.5

0.5

0.5

0.4

minqa

CRAN

1.2.4

1.2.4

mnormt

CRAN

1.5-5

1.5-5

1.5-4

modelr

CRAN

0.1.0

modeltools

CRAN

0.2-21

0.2-21

0.2-21

multtest

Bioconductor

2.30.0

2.30.0

munsell

CRAN

0.4.3

0.4.3

0.4.3

0.4.3

mvtnorm

CRAN

1.0-5

1.0-5

1.0-6

1.0-5

network

CRAN

1.13.0

nlme

base

3.1-128

3.1-128

3.1-131

3.1-128

nloptr

CRAN

1.0.4

1.0.4

NMF

CRAN

0.20.6

0.20.6

0.20.6

0.20.6

nnet

base

7.3-12

7.3-12

7.3-12

7.3-12

numDeriv

CRAN

2016.8-1

2014.2-1

openssl

CRAN

0.9.4

0.9.6

0.9.6

0.9.4

org.Hs.eg.db

Bioconductor

3.1.2

3.3.0

org.Sc.sgd.db

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-175-4"

class="td11"> 3.4.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-175-6"

class="td11">

parallel

base

3.3.2

3.3.2

3.3.1

3.3.0

pathway.structure

.permutation

GitHub

TomKellyGenetics

0.1.0

0.1.0

0.1.0

0.1.0

pbivnorm

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-178-4"

class="td11"> 0.6.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-178-6"

class="td11">

PGSEA

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-179-4"

class="td11"> 1.48.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-179-6"

class="td11">

pkgmaker

CRAN

0.22

0.22

0.22

0.22

PKI

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-181-4"

class="td11"> 0.1-3

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-181-6"

class="td11">

plogr

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-182-4"

class="td11"> 0.1-1

0.1-1

</td>

plot.igraph

GitHub

TomKellyGenetics

0.0.0.9001

0.0.0.9001

0.0.0.9001

0.0.0.9001

plotrix

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-184-4"

class="td11"> 3.6-4

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-184-6"

class="td11">

plyr

CRAN

1.8.4

1.8.4

1.8.4

1.8.3

png

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-186-4"

class="td11"> 0.1-7

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-186-6"

class="td11"> 0.1-7

prabclus

CRAN

2.2-6

2.2-6

2.2-6

</td>

praise

CRAN

1.0.0

1.0.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-188-6"

class="td11"> 1.0.0

pROC

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-189-4"

class="td11"> 1.8

1.9.1

</td>

prodlim

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-190-4"

class="td11"> 1.5.7

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-190-6"

class="td11">

prof.tree

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-191-4"

class="td11"> 0.1.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-191-6"

class="td11">

proftools

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-192-4"

class="td11"> 0.99-2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-192-6"

class="td11">

progress

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-193-4"

class="td11">

1.1.2

</td>

psych

CRAN

1.6.12

1.6.12

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-194-6"

class="td11">

purrr

CRAN

0.2.2

0.2.2

0.2.2

0.2.2

qgraph

CRAN

1.4.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-197-6"

class="td11">

quadprog

CRAN

1.5-5

1.5-5

1.5-5

R.methodsS3

CRAN

1.7.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-199-6"

class="td11"> 1.7.1

R.oo

CRAN

1.21.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-200-6"

class="td11"> 1.20.0

R.utils

CRAN

2.5.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-201-6"

class="td11">

R6

CRAN

2.1.3

2.2.0

2.2.0

2.1.3

RBGL

CRAN

1.50.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-203-6"

class="td11">

RColorBrewer

CRAN

1.1-2

1.1-2

1.1-2

1.1-2

Rcpp

CRAN

0.12.7

0.12.9

0.12.9

0.12.7

RcppArmadillo

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-206-5"

class="td11"> 0.7.700.0.0

0.6.700.6.0

RcppEigen

CRAN

0.3.2.9.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-207-6"

class="td11"> 0.3.2.8.1

RCurl

CRAN

1.95-4.8

1.95-4.8

1.95-4.8

reactome.db

Bioconductor

1.52.1

1.52.1

</td>

reactometree

GitHub

TomKellyGenetics

0.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-210-6"

class="td11">

readr

CRAN

1.0.0

1.0.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-211-6"

class="td11">

readxl

CRAN

0.1.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-212-5"

class="td11">

</td>

registry

CRAN

0.3

0.3

0.3

0.3

reshape2

CRAN

1.4.1

1.4.2

1.4.2

1.4.1

rgexf

CRAN

0.15.3

0.15.3

</td>

rgl

CRAN

0.97.0

0.95.1441

Rgraphviz

CRAN

2.18.0

rjson

CRAN

0.2.15

RJSONIO

CRAN

1.3-0

rmarkdown

CRAN

1.3

1.3

1

Rmpi

CRAN

0.6-6

0.6-5

rngtools

CRAN

1.2.4

1.2.4

1.2.4

1.2.4

robustbase

CRAN

0.92-7

0.92-7

0.92-7

0.92-5

ROCR

CRAN

1.0-7

1.0-7

1.0-7

1.0-7

Rook

CRAN

1.1-1

1.1-1

roxygen2

CRAN

6.0.1

5.0.1

6.0.1

5.0.1

rpart

base

4.1-10

4.1-10

4.1-10

4.1-10

rprojroot

CRAN

1.2

1.1

1.2

Rsamtools

Bioconductor

1.26.1

1.26.1

rsconnect

CRAN

0.7

RSQLite

CRAN

1.1-2

1.1-2

1.0.0

rstudioapi

CRAN

0.6

0.6

0.6

0.6

rvest

CRAN

0.3.2

S4Vectors

Bioconductor

0.12.1

0.12.0

0.10.3

safe

Bioconductor

3.14.0

3.10.0

scales

CRAN

0.4.0

0.4.1

0.4.1

0.4.0

selectr

CRAN

0.3-1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-239-5"

class="td11">

</td>

sem

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-240-4"

class="td11"> 3.1-8

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-240-6"

class="td11">

shiny

CRAN

0.14

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-241-5"

class="td11"> 1.0.0

</td>

slipt

GitHub

TomKellyGenetics

0.1.0

0.1.0

0.1.0

0.1.0

sm

CRAN

2.2-5.4

2.2-5.4

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-243-6"

class="td11">

sna

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-244-4"

class="td11"> 2.4

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-244-6"

class="td11">

snow

CRAN

0.4-1

0.4-2

0.4-2

0.3-13

sourcetools

CRAN

0.1.5

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-246-5"

class="td11"> 0.1.5

</td>

SparseM

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-247-4"

class="td11"> 1.74

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-247-6"

class="td11"> 1.7

spatial

base

7.3-11

7.3-11

7.3-11

7.3-11

splines

base

3.3.2

3.3.2

3.3.1

3.3.0

statnet.common

CRAN

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-250-4"

class="td11"> 3.3.0

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-250-6"

class="td11">

stats

base

3.3.2

3.3.2

3.3.1

3.3.0

stats4

base

3.3.2

3.3.2

3.3.1

3.3.0

stringi

CRAN

1.1.1

1.1.2

1.1.2

1.0-1

stringr

CRAN

1.1.0

1.1.0

1.2.0

1.0.0

Summarized

Experiment

Bioconductor

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-255-4"

class="td11"> 1.4.0

1.4.0

</td>

survival

base

2.39-4

2.40-1

2.40-1

2.39-4

tcltk

base

3.3.2

3.3.2

3.3.1

3.3.0

testthat

CRAN

1.0.2

1.0.2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-259-6"

class="td11"> 1.0.2

tibble

CRAN

1.2

1.2

1.2

1.2

tidyr

CRAN

0.6.1

0.6.1

0.6.1

</td>

tidyverse

GitHub

hadley

1.1.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-262-5"

class="td11">

</td>

timeline

CRAN

0.9

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-263-6"

class="td11">

tools

base

3.3.2

3.3.2

3.3.1

3.3.0

tpr

CRAN

0.3-1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-265-6"

class="td11">

trimcluster

CRAN

0.1-2

0.1-2

0.1-2

</td>

Unicode

CRAN

9.0.0-1

9.0.0-1

9.0.0-1

</td>

utils

base

3.3.2

3.3.2

3.3.1

3.3.0

vioplot

CRAN

0.2

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-269-6"

class="td11">

vioplotx

GitHub

TomKellyGenetics

0.0.0.9000

0.0.0.9000

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-270-6"

class="td11">

viridis

CRAN

0.3.4

0.3.4

0.3.4

</td>

visNetwork

CRAN

1.0.3

1.0.3

</td>

whisker

CRAN

0.3-2

0.3-2

0.3-2

0.3-2

withr

CRAN

1.0.2

1.0.2

1.0.2

1.0.2

XML

base

3.98-1.3

3.98-1.1

3.98-1.5

3.98-1.4

xml2

CRAN

1.1.1

</td><td style="white-space:nowrap; text-align:left;" id="TBL-77-276-5"

class="td11"> 1.1.1

1.0.0

xtable

CRAN

1.8-2

1.8-2

1.8-2

1.8-2

XVector

Bioconductor

0.14.0

0.14.0

</td>

yaml

CRAN

2.1.14

2.1.14

2.1.13

zlibbioc

CRAN

1.20.0

1.20.0

zoo

CRAN

1.7-13

1.7-14

1.7-13

Appendix CSecondary Screen Data

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

Table C.1: Comparing SLIPT genes against Secondary siRNA Screen in breast cancer

</td><td style="white-space:nowrap; text-align:right;" id="TBL-90-1-2"

class="td11">

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

class="cmbx-12">Secondary Screen

</td>

</td><td style="white-space:nowrap; text-align:right;" id="TBL-90-3-2"

class="td11">

0/4

1/4

2/4

3/4

4/4

Total

<div class="multicolumn" style="white-space:nowrap; text-align:right;"></div> </td><td style="white-space:nowrap; text-align:right;" id="TBL-90-5-2"

class="td11">

Observed

70

46

31

8

2

SLIPT+

Expected

85

44

10

4

2

157

</td><td style="white-space:nowrap; text-align:right;" id="TBL-90-8-2"

class="td11">

Observed

190

90

31

10

4

SLIPT-

Expected

175

91

42

12

4

325

</td><td style="white-space:nowrap; text-align:right;" id="TBL-90-11-2"

class="td11">

Total

280

136

52

18

6

482

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

Similar analysis with mtSLIPT, comparing SLIPT against CDH1 somatic mutation with siRNA validation results was not significant (p = 7.02 × 10-1 by Fisher’s exact test). However, as shown in Table C.2, the observed and expected values were in a direction consistent with that observed above for SLIPT against low CDH1 expression. It is not unexpected that this result does not have comparable statistical support due to the lower sample size for mutation data.

Table C.2: Comparing mtSLIPT genes against Secondary siRNA Screen in breast cancer

</td><td style="white-space:nowrap; text-align:right;" id="TBL-91-1-2"

class="td11">

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

class="cmbx-12">Secondary Screen

</td>

</td><td style="white-space:nowrap; text-align:right;" id="TBL-91-3-2"

class="td11">

0/4

1/4

2/4

3/4

4/4

Total

<div class="multicolumn" style="white-space:nowrap; text-align:right;"></div> </td><td style="white-space:nowrap; text-align:right;" id="TBL-91-5-2"

class="td11">

Observed

54

35

17

4

6

mtSLIPT+

Expected

60

31

14

4

1

111

<div class="multicolumn" style="white-space:nowrap; text-align:right;"></div> </td><td style="white-space:nowrap; text-align:right;" id="TBL-91-8-2"

class="td11">

Observed

206

101

45

14

5

mtSLIPT-

Expected

200

105

48

14

4

371

</td><td style="white-space:nowrap; text-align:right;" id="TBL-91-11-2"

class="td11">

Total

269

143

63

19

6

482

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

This analysis was replicated on a (smaller) stomach cancer dataset but it was less conclusive (p = 2.36 × 10-1 by Fisher’s exact test). As shown in Table C.3, fewer SLIPT candidates were validated than expected statistically. However, these results in stomach cancer may not be directly comparable to experiments in a breast cell line. Genes validated by 0 or 1 siRNA behave consistently with the results above.

Table C.3: Comparing SLIPT genes against Secondary siRNA Screen in stomach cancer

</td><td style="white-space:nowrap; text-align:right;" id="TBL-92-1-2"

class="td11">

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

class="cmbx-12">Secondary Screen

</td>

</td><td style="white-space:nowrap; text-align:right;" id="TBL-92-3-2"

class="td11">

0/4

1/4

2/4

3/4

4/4

Total

<div class="multicolumn" style="white-space:nowrap; text-align:right;"></div> </td><td style="white-space:nowrap; text-align:right;" id="TBL-92-5-2"

class="td11">

Observed

67

47

13

4

1

SLIPT+

Expected

71

37

17

5

2

132

</td><td style="white-space:nowrap; text-align:right;" id="TBL-92-8-2"

class="td11">

Observed

195

90

50

14

5

SLIPT-

Expected

190

100

46

13

4

354

</td><td style="white-space:nowrap; text-align:right;" id="TBL-92-11-2"

class="td11">

Total

262

137

63

19

6

486

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

Appendix DMutation Analysis in Breast Cancer

D.1 Synthetic Lethal Genes and Pathways

SLIPT expression analysis (described in Section 3.1) on TCGA breast cancer data (n = 969) found the following genes and pathways, described in sections 4.1 and 4.1.1.

Table D.1: Candidate synthetic lethal gene partners of CDH1 from mtSLIPT

Gene

Observed

Expected

χ2 value

p-value

p-value (FDR)

TFAP2B

8 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-2-3"

class="td11"> 36.7

89.5

3.60 × 10-20

8.37 × 10-17

ZNF423

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-3-3"

class="td11"> 36.7

78.8

7.89 × 10-18

1.22 × 10-14

CALCOCO1

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-4-3"

class="td11"> 36.7

76.8

2.09 × 10-17

2.59 × 10-14

RBM5

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-5-3"

class="td11"> 36.7

75.7

3.65 × 10-17

4.00 × 10-14

BTG2

7 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-6-3"

class="td11"> 36.7

71.7

2.72 × 10-16

1.81 × 10-13

RXRA

6 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-7-3"

class="td11"> 36.7

70.5

5.00 × 10-16

2.97 × 10-13

SLC27A1

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-8-3"

class="td11"> 36.7

70.3

5.42 × 10-16

2.97 × 10-13

MEF2D

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-9-3"

class="td11"> 36.7

69.6

7.86 × 10-16

3.95 × 10-13

NISCH

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-10-3"

class="td11"> 36.7

69.6

7.86 × 10-16

3.95 × 10-13

AVPR2

9 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-11-3"

class="td11"> 36.7

69.2

9.36 × 10-16

4.58 × 10-13

CRY2

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-12-3"

class="td11"> 36.7

68.9

1.07 × 10-15

4.98 × 10-13

RAPGEF3

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-13-3"

class="td11"> 36.7

68.9

1.07 × 10-15

4.98 × 10-13

NRIP2

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-14-3"

class="td11"> 36.7

68.2

1.58 × 10-15

7.18 × 10-13

DARC

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-15-3"

class="td11"> 36.7

66.4

3.76 × 10-15

1.54 × 10-12

SFRS5

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-16-3"

class="td11"> 36.7

66.4

3.76 × 10-15

1.54 × 10-12

NOSTRIN

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-17-3"

class="td11"> 36.7

65.1

7.40 × 10-15

2.70 × 10-12

KIF13B

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-18-3"

class="td11"> 36.7

63.4

1.69 × 10-14

5.16 × 10-12

TENC1

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-19-3"

class="td11"> 36.7

62.5

2.67 × 10-14

7.40 × 10-12

MFAP4

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-20-3"

class="td11"> 36.7

60.5

7.17 × 10-14

1.67 × 10-11

ELN

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-21-3"

class="td11"> 36.7

59.7

1.07 × 10-13

2.32 × 10-11

SGK223

14 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-22-3"

class="td11"> 36.7

59

1.51 × 10-13

3.05 × 10-11

KIF12

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-23-3"

class="td11"> 36.7

58.8

1.74 × 10-13

3.34 × 10-11

SELP

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-24-3"

class="td11"> 36.7

58.8

1.74 × 10-13

3.34 × 10-11

CIRBP

9 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-25-3"

class="td11"> 36.7

58.7

1.83 × 10-13

3.41 × 10-11

CTDSP1

9 </td><td style="white-space:nowrap; text-align:center;" id="TBL-93-26-3"

class="td11"> 36.7

58.7

1.83 × 10-13

3.41 × 10-11

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

Strongest candidate SL partners for CDH1 by mtSLIPT with observed and expected mutant samples with low expression of partner genes

Table D.2: Pathways for CDH1 partners from mtSLIPT

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Eukaryotic Translation Elongation

86 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-2-3"

class="td11"> 60

2.0 × 10-128

Peptide chain elongation

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-3-3"

class="td11"> 59

2.0 × 10-128

Eukaryotic Translation Termination

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-4-3"

class="td11"> 58

2.3 × 10-125

Viral mRNA Translation

81 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-5-3"

class="td11"> 57

2.5 × 10-124

Nonsense Mediated Decay independent of the Exon Junction Complex

88 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-6-3"

class="td11"> 59

8.6 × 10-124

Nonsense-Mediated Decay

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-7-3"

class="td11"> 61

5.2 × 10-117

Nonsense Mediated Decay enhanced by the Exon Junction Complex

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-8-3"

class="td11"> 61

5.2 × 10-117

Formation of a pool of free 40S subunits

93 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-9-3"

class="td11"> 58

1.6 × 10-116

L13a-mediated translational silencing of Ceruloplasmin expression

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-10-3"

class="td11"> 59

1.3 × 10-111

3’ -UTR-mediated translational regulation

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-11-3"

class="td11"> 59

1.3 × 10-111

GTP hydrolysis and joining of the 60S ribosomal subunit

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-12-3"

class="td11"> 59

6.2 × 10-111

SRP-dependent cotranslational protein targeting to membrane

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-13-3"

class="td11"> 58

2.9 × 10-108

Eukaryotic Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-14-3"

class="td11"> 59

3.0 × 10-106

Cap-dependent Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-15-3"

class="td11"> 59

3.0 × 10-106

Influenza Viral RNA Transcription and Replication

108 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-16-3"

class="td11"> 57

5.1 × 10-103

Influenza Infection

117 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-17-3"

class="td11"> 59

1.5 × 10-102

Translation

141 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-18-3"

class="td11"> 64

3.7 × 10-101

Influenza Life Cycle

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-19-3"

class="td11"> 57

1.4 × 10-100

GPCR downstream signaling

472 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-20-3"

class="td11"> 116

1.0 × 10-80

Hemostasis

422 </td><td style="white-space:nowrap; text-align:center;" id="TBL-94-21-3"

class="td11"> 105

1.4 × 10-78

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in mtSLIPT partners for CDH1

The genes and pathways identified in Tables D.1 and D.2 were derived from comparing the expression profiles of potential partners to the mutation status of CDH1 (as shown in Figure 3.2). Thus the following analysis is only limited the samples for which TCGA provides both expression and somatic mutation data.

D.2 Synthetic Lethal Expression Profiles

Similar to the analysis of synthetic lethal partners against low CDH1 expression in 4.1.2, the partners detected from CDH1 mutation were also examined for their expression profiles and the pathway composition of gene clusters. Hierachical clustering was performed on mtSLIPT partners for CDH1 as showing in Figure D.1. Over-representation for Reactome pathways for each of the gene clusters identified is given in Table D.3.

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

Figure D.1: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance) of all samples (separated by CDH1 somatic mutation status) analysed in TCGA breast cancer dataset for gene expression of 3,743 candidate partners of E-cadherin (CDH1) from mtSLIPT prediction (with significant FDR adjusted p < 0.05). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or toxic in an siRNA screen Telford et al. (2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in CDH1 mutant samples and often lowly expressing CDH1 wildtype samples (which were not tested for), although many of the CDH1 mutant samples had among the lowest CDH1 expression. In contrast to the expression analysis the (predominantly CDH1 wildtype) basal subtype and estrogen receptor negative samples have depleted expression among most candidate synthetic lethal partners.

Table D.3: Pathway composition for clusters of CDH1 partners from mtSLIPT

Pathways Over-represented in Cluster 1

Pathway Size

Cluster Genes

p-value (FDR)

Olfactory Signaling Pathway

57 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-2-3"

class="td11"> 8

7<span

class="cmmi-12">.1 × 10-9

Assembly of the primary cilium

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-3-3"

class="td11"> 14

8<span

class="cmmi-12">.0 × 10-9

Sphingolipid metabolism

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-4-3"

class="td11"> 8

9<span

class="cmmi-12">.6 × 10-9

Signaling by ERBB4

133 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-5-3"

class="td11"> 12

5<span

class="cmmi-12">.1 × 10-8

PI3K Cascade

65 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-6-3"

class="td11"> 7

4<span

class="cmmi-12">.9 × 10-7

Circadian Clock

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-7-3"

class="td11"> 5

4<span

class="cmmi-12">.9 × 10-7

Nuclear signaling by ERBB4

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-8-3"

class="td11"> 5

4<span

class="cmmi-12">.9 × 10-7

Intraflagellar transport

35 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-9-3"

class="td11"> 5

4<span

class="cmmi-12">.9 × 10-7

PI3K events in ERBB4 signaling

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-10-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

PIP3 activates AKT signaling

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-11-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

PI3K events in ERBB2 signaling

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-12-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

PI-3K cascade:FGFR1

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-13-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

PI-3K cascade:FGFR2

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-14-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

PI-3K cascade:FGFR3

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-15-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

PI-3K cascade:FGFR4

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-16-3"

class="td11"> 8

4<span

class="cmmi-12">.9 × 10-7

Deadenylation of mRNA

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-17-3"

class="td11"> 4

5<span

class="cmmi-12">.6 × 10-7

PI3K/AKT activation

90 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-18-3"

class="td11"> 8

5<span

class="cmmi-12">.6 × 10-7

Cargo trafficking to the periciliary membrane

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-19-3"

class="td11"> 5

5<span

class="cmmi-12">.6 × 10-7

Signaling by Hedgehog

108 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-20-3"

class="td11"> 9

5<span

class="cmmi-12">.6 × 10-7

Downstream signal transduction

143 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-21-3"

class="td11"> 11

5<span

class="cmmi-12">.6 × 10-7

</td>

Pathways Over-represented in Cluster 2

Pathway Size

Cluster Genes

p-value (FDR)

Gαs signalling events

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-24-3"

class="td11"> 19

5<span

class="cmmi-12">.1 × 10-25

Extracellular matrix organization

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-25-3"

class="td11"> 30

1<span

class="cmmi-12">.4 × 10-18

Hemostasis

422 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-26-3"

class="td11"> 46

2<span

class="cmmi-12">.7 × 10-16

Aquaporin-mediated transport

32 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-27-3"

class="td11"> 9

2<span

class="cmmi-12">.7 × 10-16

Transcriptional regulation of white adipocyte differentiation

56 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-28-3"

class="td11"> 11

1<span

class="cmmi-12">.7 × 10-15

Degradation of the extracellular matrix

102 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-29-3"

class="td11"> 15

1<span

class="cmmi-12">.7 × 10-15

Integration of energy metabolism

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-30-3"

class="td11"> 13

8<span

class="cmmi-12">.8 × 10-15

GPCR downstream signaling

472 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-31-3"

class="td11"> 48

2<span

class="cmmi-12">.8 × 10-14

Gαz signalling events

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-32-3"

class="td11"> 6

5<span

class="cmmi-12">.0 × 10-14

Molecules associated with elastic fibres

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-33-3"

class="td11"> 8

5<span

class="cmmi-12">.4 × 10-14

Phase 1 - Functionalization of compounds

67 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-34-3"

class="td11"> 11

5<span

class="cmmi-12">.6 × 10-14

Platelet activation, signaling and aggregation

179 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-35-3"

class="td11"> 20

5<span

class="cmmi-12">.6 × 10-14

Vasopressin regulates renal water homeostasis via Aquaporins

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-36-3"

class="td11"> 7

6<span

class="cmmi-12">.1 × 10-14

Elastic fibre formation

37 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-37-3"

class="td11"> 8

<span

class="cmmi-12">.03 × 10-13

Calmodulin induced events

27 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-38-3"

class="td11"> 7

3<span

class="cmmi-12">.3 × 10-13

CaM pathway

27 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-39-3"

class="td11"> 7

3<span

class="cmmi-12">.3 × 10-13

cGMP effects

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-40-3"

class="td11"> 6

3<span

class="cmmi-12">.6 × 10-13

Gαi signalling events

167 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-41-3"

class="td11"> 18

6<span

class="cmmi-12">.3 × 10-13

Ca-dependent events

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-42-3"

class="td11"> 7

8<span

class="cmmi-12">.2 × 10-13

Binding and Uptake of Ligands by Scavenger Receptors

40 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-43-3"

class="td11"> 8

8<span

class="cmmi-12">.2 × 10-13

</td>

Pathways Over-represented in Cluster 3

Pathway Size

Cluster Genes

p-value (FDR)

Eukaryotic Translation Elongation

86 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-46-3"

class="td11"> 55

1<span

class="cmmi-12">.1 × 10-112

Peptide chain elongation

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-47-3"

class="td11"> 54

1<span

class="cmmi-12">.3 × 10-112

Viral mRNA Translation

81 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-48-3"

class="td11"> 53

1<span

class="cmmi-12">.6 × 10-111

Eukaryotic Translation Termination

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-49-3"

class="td11"> 53

7<span

class="cmmi-12">.1 × 10-110

Nonsense Mediated Decay independent of the Exon Junction Complex

88 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-50-3"

class="td11"> 54

1<span

class="cmmi-12">.0 × 10-108

Formation of a pool of free 40S subunits

93 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-51-3"

class="td11"> 53

4<span

class="cmmi-12">.1 × 10-102

Nonsense-Mediated Decay

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-52-3"

class="td11"> 54

3<span

class="cmmi-12">.9 × 10-98

Nonsense Mediated Decay enhanced by the Exon Junction Complex

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-53-3"

class="td11"> 54

3<span

class="cmmi-12">.9 × 10-98

L13a-mediated translational silencing of Ceruloplasmin expression

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-54-3"

class="td11"> 53

1<span

class="cmmi-12">.2 × 10-95

3’ -UTR-mediated translational regulation

103 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-55-3"

class="td11"> 53

1<span

class="cmmi-12">.2 × 10-95

SRP-dependent cotranslational protein targeting to membrane

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-56-3"

class="td11"> 53

4<span

class="cmmi-12">.3 × 10-95

GTP hydrolysis and joining of the 60S ribosomal subunit

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-57-3"

class="td11"> 53

4<span

class="cmmi-12">.3 × 10-95

Influenza Viral RNA Transcription and Replication

108 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-58-3"

class="td11"> 53

9<span

class="cmmi-12">.6 × 10-93

Eukaryotic Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-59-3"

class="td11"> 53

4<span

class="cmmi-12">.2 × 10-91

Cap-dependent Translation Initiation

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-60-3"

class="td11"> 53

4<span

class="cmmi-12">.2 × 10-91

Influenza Life Cycle

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-61-3"

class="td11"> 53

1<span

class="cmmi-12">.4 × 10-90

Influenza Infection

117 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-62-3"

class="td11"> 53

6<span

class="cmmi-12">.2 × 10-88

Translation

141 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-63-3"

class="td11"> 55

3 <span

class="cmsy-10x-x-120">× 10-81

Formation of the ternary complex, and subsequently, the 43S complex

47 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-64-3"

class="td11"> 23

2<span

class="cmmi-12">.3 × 10-48

Translation initiation complex formation

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-65-3"

class="td11"> 23

9<span

class="cmmi-12">.1 × 10-45

</td>

Pathways Over-represented in Cluster 4

Pathway Size

Cluster Genes

p-value (FDR)

ECM proteoglycans

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-68-3"

class="td11"> 10

2<span

class="cmmi-12">.9 × 10-11

deactivation of the beta-catenin transactivating complex

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-69-3"

class="td11"> 7

5<span

class="cmmi-12">.1 × 10-10

Arachidonic acid metabolism

41 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-70-3"

class="td11"> 7

1<span

class="cmmi-12">.1 × 10-9

G˙αq signalling events

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-71-3"

class="td11"> 14

4<span

class="cmmi-12">.0 × 10-9

HS-GAG degradation

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-72-3"

class="td11"> 5

4<span

class="cmmi-12">.5 × 10-9

Uptake and actions of bacterial toxins

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-73-3"

class="td11"> 5

6<span

class="cmmi-12">.1 × 10-9

Gastrin-CREB signalling pathway via PKC and MAPK

170 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-74-3"

class="td11"> 15

6<span

class="cmmi-12">.1 × 10-9

RNA Polymerase I, RNA Polymerase III, and Mitochondrial Transcription ll

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-75-3"

class="td11"> 8

6<span

class="cmmi-12">.1 × 10-9

Non-integrin membrane-ECM interactions

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-76-3"

class="td11"> 7

1<span

class="cmmi-12">.5 × 10-8

Syndecan interactions

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-77-3"

class="td11"> 5

1<span

class="cmmi-12">.5 × 10-8

NOTCH1 Intracellular Domain Regulates Transcription

40 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-78-3"

class="td11"> 6

2<span

class="cmmi-12">.3 × 10-8

Synthesis of Leukotrienes and Eoxins

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-79-3"

class="td11"> 4

3<span

class="cmmi-12">.2 × 10-8

Signaling by NOTCH1

59 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-80-3"

class="td11"> 7

5<span

class="cmmi-12">.3 × 10-8

Regulation of insulin secretion

44 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-81-3"

class="td11"> 6

6<span

class="cmmi-12">.0 × 10-8

Metabolism of lipids and lipoproteins

471 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-82-3"

class="td11"> 37

8<span

class="cmmi-12">.2 × 10-8

Signaling by NOTCH

80 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-83-3"

class="td11"> 8

1<span

class="cmmi-12">.2 × 10-7

Platelet activation, signaling and aggregation

179 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-84-3"

class="td11"> 14

1<span

class="cmmi-12">.2 × 10-7

Recruitment of mitotic centrosome proteins and complexes

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-85-3"

class="td11"> 7

1<span

class="cmmi-12">.2 × 10-7

Centrosome maturation

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-86-3"

class="td11"> 7

1<span

class="cmmi-12">.2 × 10-7

Biological oxidations

133 </td><td style="white-space:nowrap; text-align:center;" id="TBL-95-87-3"

class="td11"> 11

1<span

class="cmmi-12">.5 × 10-7

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

D.3 Comparison to Primary Screen

The mutation synthetic lethal partners with CDH1 were also compared to siRNA primary screen data (Telford et al., 2015), as performed in Section 4.2.1. These are expected to be more concordant with the experimental results performed on a null mutant, however this not the case at the gene level: less genes overlapped with experimental candidates in Figure D.2. This may be affected by lower sample size for mutations in TCGA data or lower frequency (expected value) of CDH1 mutations compared to low expression.

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

Figure D.2: Comparison of mtSLIPT to siRNA. Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ2 test suggests that the overlap is no more than would be expected by chance (p = 0.281).

Despite a lower sample size (and low number of a predicted partners) for mutation analysis, the pathway composition (Tables D.2 and D.4) is simlar to expression analysis, as described in Section 4.2.1.4. In particular, the resampling analysis (Section D.3.1) supported many of the results of expression analysis (Section 4.2.1.4.1) with Tables D.5 and D.6 detecting many of the same or functionally-related pathways.

Table D.4: Pathway composition for CDH1 partners from mtSLIPT and siRNA

Predicted only by SLIPT (2901 genes)

Pathway Size

Genes Identified

p-value (FDR)

Eukaryotic Translation Elongation

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-2-3"

class="td11"> 57

2.8 × 10-120

Peptide chain elongation

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-3-3"

class="td11"> 56

3.1 × 10-120

Eukaryotic Translation Termination

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-4-3"

class="td11"> 55

2.8 × 10-117

Viral mRNA Translation

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-5-3"

class="td11"> 54

4.1 × 10-116

Nonsense Mediated Decay independent of the Exon Junction Complex

89 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-6-3"

class="td11"> 55

3.7 × 10-113

Formation of a pool of free 40S subunits

94 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-7-3"

class="td11"> 55

2.8 × 10-109

Nonsense-Mediated Decay

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-8-3"

class="td11"> 57

8.4 × 10-108

Nonsense Mediated Decay enhanced by the Exon Junction Complex

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-9-3"

class="td11"> 57

8.4 × 10-108

L13a-mediated translational silencing of Ceruloplasmin expression

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-10-3"

class="td11"> 56

3.4 × 10-105

3’ -UTR-mediated translational regulation

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-11-3"

class="td11"> 56

3.4 × 10-105

GTP hydrolysis and joining of the 60S ribosomal subunit

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-12-3"

class="td11"> 56

1.4 × 10-104

Eukaryotic Translation Initiation

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-13-3"

class="td11"> 56

2.8 × 10-100

Cap-dependent Translation Initiation

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-14-3"

class="td11"> 56

2.8 × 10-100

SRP-dependent cotranslational protein targeting to membrane

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-15-3"

class="td11"> 54

2.2 × 10-99

Influenza Viral RNA Transcription and Replication

109 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-16-3"

class="td11"> 54

5.3 × 10-97

Influenza Life Cycle

113 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-17-3"

class="td11"> 54

9.6 × 10-95

Influenza Infection

118 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-18-3"

class="td11"> 55

1.7 × 10-94

Translation

142 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-19-3"

class="td11"> 60

3.5 × 10-94

Infectious disease

349 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-20-3"

class="td11"> 77

5.9 × 10-62

Extracellular matrix organization

241 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-21-3"

class="td11"> 54

3.0 × 10-52

</td>

Detected only by siRNA screen (1752 genes)

Pathway Size

Genes Identified

p-value (FDR)

Class A/1 (Rhodopsin-like receptors)

282 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-24-3"

class="td11"> 69

1.9 × 10-59

GPCR ligand binding

363 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-25-3"

class="td11"> 78

2.7 × 10-54

Peptide ligand-binding receptors

175 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-26-3"

class="td11"> 41

1.5 × 10-42

Gαi signalling events

184 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-27-3"

class="td11"> 41

1.1 × 10-40

Gastrin-CREB signalling pathway via PKC and MAPK

180 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-28-3"

class="td11"> 37

1.5 × 10-35

Gαq signalling events

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-29-3"

class="td11"> 34

3.7 × 10-35

DAP12 interactions

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-30-3"

class="td11"> 27

1.1 × 10-24

VEGFA-VEGFR2 Pathway

91 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-31-3"

class="td11"> 19

1.0 × 10-23

Downstream signal transduction

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-32-3"

class="td11"> 24

1.9 × 10-22

Signaling by VEGF

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-33-3"

class="td11"> 19

2.6 × 10-22

DAP12 signaling

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-34-3"

class="td11"> 24

4.2 × 10-22

Organelle biogenesis and maintenance

264 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-35-3"

class="td11"> 34

4.3 × 10-20

Downstream signaling of activated FGFR1

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-36-3"

class="td11"> 21

4.3 × 10-20

Downstream signaling of activated FGFR2

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-37-3"

class="td11"> 21

4.3 × 10-20

Downstream signaling of activated FGFR3

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-38-3"

class="td11"> 21

4.3 × 10-20

Downstream signaling of activated FGFR4

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-39-3"

class="td11"> 21

4.3 × 10-20

Signaling by ERBB2

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-40-3"

class="td11"> 22

5.3 × 10-20

Signaling by FGFR

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-41-3"

class="td11"> 22

5.3 × 10-20

Signaling by FGFR1

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-42-3"

class="td11"> 22

5.3 × 10-20

Signaling by FGFR2

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-43-3"

class="td11"> 22

5.3 × 10-20

</td>

Intersection of SLIPT and siRNA screen (450 genes)

Pathway Size

Genes Identified

p-value (FDR)

HS-GAG degradation

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-46-3"

class="td11"> 4

4.9 × 10-6

Retinoid metabolism and transport

39 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-47-3"

class="td11"> 5

4.9 × 10-6

Platelet activation, signaling and aggregation

186 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-48-3"

class="td11"> 13

4.9 × 10-6

Signaling by NOTCH4

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-49-3"

class="td11"> 3

4.9 × 10-6

Gαs signalling events

100 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-50-3"

class="td11"> 8

5.0 × 10-6

Defective EXT2 causes exostoses 2

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-51-3"

class="td11"> 3

5.0 × 10-6

Defective EXT1 causes exostoses 1, TRPS2 and CHDS

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-52-3"

class="td11"> 3

5.0 × 10-6

Class A/1 (Rhodopsin-like receptors)

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-53-3"

class="td11"> 18

2.2 × 10-5

Signaling by PDGF

173 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-54-3"

class="td11"> 11

2.9 × 10-5

Circadian Clock

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-55-3"

class="td11"> 4

2.9 × 10-5

Signaling by ERBB4

139 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-56-3"

class="td11"> 9

4.3 × 10-5

Role of LAT2/NTAL/LAB on calcium mobilization

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-57-3"

class="td11"> 7

4.4 × 10-5

Peptide ligand-binding receptors

181 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-58-3"

class="td11"> 11

4.5 × 10-5

Defective B4GALT7 causes EDS, progeroid type

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-59-3"

class="td11"> 3

4.5 × 10-5

Defective B3GAT3 causes JDSSDHD

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-60-3"

class="td11"> 3

4.5 × 10-5

Signaling by NOTCH

80 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-61-3"

class="td11"> 6

4.5 × 10-5

Gαq signalling events

164 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-62-3"

class="td11"> 10

5.1 × 10-5

Response to elevated platelet cytosolic Ca2ˆ+

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-63-3"

class="td11"> 6

7.1 × 10-5

Signaling by ERBB2

148 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-64-3"

class="td11"> 9

7.1 × 10-5

Signaling by SCF-KIT

129 </td><td style="white-space:nowrap; text-align:center;" id="TBL-96-65-3"

class="td11"> 8

8.3 × 10-5

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

D.3.1 Resampling Analysis

Table D.5: Pathways for CDH1 partners from mtSLIPT

Reactome Pathway

Over-representation

Permutation

Eukaryotic Translation Elongation

3<span

class="cmmi-12">.2 × 10-128

< 7.035 × 10-4

Peptide chain elongation

3<span

class="cmmi-12">.2 × 10-128

< 7.035 × 10-4

Eukaryotic Translation Termination

3<span

class="cmmi-12">.7 × 10-125

< 7.035 × 10-4

Viral mRNA Translation

4<span

class="cmmi-12">.1 × 10-124

< 7.035 × 10-4

Nonsense Mediated Decay independent of the Exon Junction Complex

1<span

class="cmmi-12">.4 × 10-123

< 7.035 × 10-4

Nonsense-Mediated Decay

8<span

class="cmmi-12">.4 × 10-117

< 7.035 × 10-4

Nonsense Mediated Decay enhanced by the Exon Junction Complex

8<span

class="cmmi-12">.4 × 10-117

< 7.035 × 10-4

Formation of a pool of free 40S subunits

2<span

class="cmmi-12">.6 × 10-116

< 7.035 × 10-4

L13a-mediated translational silencing of Ceruloplasmin expression

2<span

class="cmmi-12">.0 × 10-111

< 7.035 × 10-4

3’ -UTR-mediated translational regulation

2<span

class="cmmi-12">.0 × 10-111

< 7.035 × 10-4

GTP hydrolysis and joining of the 60S ribosomal subunit

9<span

class="cmmi-12">.9 × 10-111

< 7.035 × 10-4

SRP-dependent cotranslational protein targeting to membrane

4<span

class="cmmi-12">.7 × 10-108

< 7.035 × 10-4

Eukaryotic Translation Initiation

4<span

class="cmmi-12">.8 × 10-106

< 7.035 × 10-4

Cap-dependent Translation Initiation

4<span

class="cmmi-12">.8 × 10-106

< 7.035 × 10-4

Influenza Viral RNA Transcription and Replication

8<span

class="cmmi-12">.1 × 10-103

< 7.035 × 10-4

Influenza Infection

2<span

class="cmmi-12">.4 × 10-102

< 7.035 × 10-4

Translation

6<span

class="cmmi-12">.0 × 10-101

< 7.035 × 10-4

Influenza Life Cycle

2<span

class="cmmi-12">.2 × 10-100

< 7.035 × 10-4

Disease

2<span

class="cmmi-12">.1 × 10-90

0<span

class="cmmi-12">.013347

GPCR downstream signaling

1<span

class="cmmi-12">.6 × 10-80

0<span

class="cmmi-12">.095478

Hemostasis

2<span

class="cmmi-12">.1 × 10-78

0<span

class="cmmi-12">.2671

Signaling by GPCR

1<span

class="cmmi-12">.2 × 10-73

0<span

class="cmmi-12">.44939

Extracellular matrix organization

2<span

class="cmmi-12">.2 × 10-67

0<span

class="cmmi-12">.054008

Metabolism of proteins

1<span

class="cmmi-12">.4 × 10-66

0<span

class="cmmi-12">.9607

Signal Transduction

2<span

class="cmmi-12">.1 × 10-66

0<span

class="cmmi-12">.48184

Developmental Biology

2<span

class="cmmi-12">.5 × 10-66

0<span

class="cmmi-12">.54075

Innate Immune System

5<span

class="cmmi-12">.3 × 10-66

0<span

class="cmmi-12">.9589

Infectious disease

9<span

class="cmmi-12">.6 × 10-66

0<span

class="cmmi-12">.21075

Signalling by NGF

1<span

class="cmmi-12">.1 × 10-62

0<span

class="cmmi-12">.43356

Immune System

2<span

class="cmmi-12">.8 × 10-62

0<span

class="cmmi-12">.23052

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

Table D.6: Pathways for CDH1 partners from mtSLIPT and siRNA primary screen

Reactome Pathway

Over-representation

Permutation

Visual phototransduction

1<span

class="cmmi-12">.2 × 10-9

0.86279

Gαs signalling events

2<span

class="cmmi-12">.9 × 10-7

0.023066

Retinoid metabolism and transport

2<span

class="cmmi-12">.9 × 10-7

0<span

class="cmmi-12">.299

Acyl chain remodelling of PS

1<span

class="cmmi-12">.1 × 10-5

0.42584

Transcriptional regulation of white adipocyte differentiation

1<span

class="cmmi-12">.1 × 10-5

0.53928

Chemokine receptors bind chemokines

1<span

class="cmmi-12">.1 × 10-5

0.95259

Signaling by NOTCH4

1<span

class="cmmi-12">.2 × 10-5

0.079229

Defective EXT2 causes exostoses 2

1<span

class="cmmi-12">.2 × 10-5

0.22292

Defective EXT1 causes exostoses 1, TRPS2 and CHDS

1<span

class="cmmi-12">.2 × 10-5

0.22292

Platelet activation, signaling and aggregation

1<span

class="cmmi-12">.2 × 10-5

0.48853

Serotonin receptors

1<span

class="cmmi-12">.4 × 10-5

0.34596

Nicotinamide salvaging

1<span

class="cmmi-12">.4 × 10-5

0.70881

Phase 1 - Functionalization of compounds

2 <span

class="cmsy-10x-x-120">× 10-5

0.31142

Amine ligand-binding receptors

2<span

class="cmmi-12">.5 × 10-5

0.34934

Acyl chain remodelling of PE

3<span

class="cmmi-12">.8 × 10-5

0.42615

Signaling by GPCR

3<span

class="cmmi-12">.8 × 10-5

0.93888

Molecules associated with elastic fibres

3<span

class="cmmi-12">.9 × 10-5

0.017982

DAP12 interactions

3<span

class="cmmi-12">.9 × 10-5

0.71983

Beta defensins

3<span

class="cmmi-12">.9 × 10-5

0.91458

Cytochrome P450 - arranged by substrate type

4<span

class="cmmi-12">.7 × 10-5

0.83493

GPCR ligand binding

5<span

class="cmmi-12">.7 × 10-5

0.95258

Acyl chain remodelling of PC

6<span

class="cmmi-12">.1 × 10-5

0.42584

Response to elevated platelet cytosolic Ca2ˆ+

6<span

class="cmmi-12">.4 × 10-5

0.54046

Arachidonic acid metabolism

6<span

class="cmmi-12">.7 × 10-5

0.026696

Defective B4GALT7 causes EDS, progeroid type

7<span

class="cmmi-12">.3 × 10-5

0.24921

Defective B3GAT3 causes JDSSDHD

7<span

class="cmmi-12">.3 × 10-5

0.24921

Hydrolysis of LPC

7<span

class="cmmi-12">.3 × 10-5

0.80663

Elastic fibre formation

7<span

class="cmmi-12">.4 × 10-5

0.0058768

HS-GAG degradation

9<span

class="cmmi-12">.4 × 10-5

0.0083179

Bile acid and bile salt metabolism

9<span

class="cmmi-12">.4 × 10-5

0.079905

Netrin-1 signaling

0<span

class="cmmi-12">.00011

0.92216

Integration of energy metabolism

0<span

class="cmmi-12">.00011

0.011152

Dectin-2 family

0<span

class="cmmi-12">.00012

0.10385

Platelet sensitization by LDL

0<span

class="cmmi-12">.00012

0.34596

DAP12 signaling

0<span

class="cmmi-12">.00012

0.62787

Defensins

0<span

class="cmmi-12">.00012

0.77542

GPCR downstream signaling

0<span

class="cmmi-12">.00012

0.79454

Diseases associated with glycosaminoglycan metabolism

0<span

class="cmmi-12">.00013

0.065927

Diseases of glycosylation

0<span

class="cmmi-12">.00013

0.065927

Signaling by Retinoic Acid

0<span

class="cmmi-12">.00013

0.22292

Signaling by Leptin

0<span

class="cmmi-12">.00013

0.34596

Signaling by SCF-KIT

0<span

class="cmmi-12">.00013

0.70881

Opioid Signalling

0<span

class="cmmi-12">.00013

0.96053

Signaling by NOTCH

0<span

class="cmmi-12">.00015

0.26884

Platelet homeostasis

0<span

class="cmmi-12">.00015

0<span

class="cmmi-12">.4878

Signaling by NOTCH1

0<span

class="cmmi-12">.00016

0.13043

Class B/2 (Secretin family receptors)

0<span

class="cmmi-12">.00016

0.13994

Diseases of Immune System

0<span

class="cmmi-12">.0002

0<span

class="cmmi-12">.0795

Diseases associated with the TLR signaling cascade

0<span

class="cmmi-12">.0002

0<span

class="cmmi-12">.0795

A tetrasaccharide linker sequence is required for GAG synthesis

0<span

class="cmmi-12">.0002

0.42615

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

D.4 Compare SLIPT genes

The mutation synthetic lethal partners with CDH1 were also compared to siRNA primary screen data (Telford et al., 2015), by correlation and siRNA viability as described in sections 4.2.1.1 and 4.2.1.2.

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

Figure D.3: Compare mtSLIPT and siRNA genes with correlation. The mtSLIPT p-values were compared against Pearson’s correlation of expression with CDH1. Genes detected by SLIPT or siRNA are coloured according to the legend.

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

Figure D.4: Compare mtSLIPT and siRNA genes with correlation. Genes detected by mtSLIPT against CDH1 mutation and siRNA screening were compared against Pearson’s correlation of expression with CDH1. There were no differences in correlation between the gene groups.

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

Figure D.5: Compare mtSLIPT and siRNA genes with siRNA viability. Genes detected as candidate synthetic lethal partners by mtSLIPT (in TCGA breast cancer) expression analysis against CDH1 mutation and experimental screening (with siRNA) were compared against the viability ratio of CDH1 mutant and wildtype cells in the primary siRNA screen. There were clear no differences in viability between genes detected by mtSLIPT and those not with the differences being primarily due to viability thresholds being used to detect synthetic lethality by Telford et al. (2015).

D.5 Metagene Analysis

Metagene analysis was also performed for synthetic lethal candidates for CDH1 mutation. These are described and compared to expression analysis in Section 4.3.4.

Table D.7: Candidate synthetic lethal metagenes against CDH1 from mtSLIPT

Pathway

ID

Observed

Expected

χ2value

p-value

p-value (FDR)

Neurotoxicity of clostridium toxins

168799

8

36.7

79.4

5.71 × 10-18

3.14 × 10-15

Aquaporin-mediated transport

445717

8

36.7

76.3

2.73 × 10-17

9.01 × 10-15

Toxicity of botulinum toxin type G (BoNT/G)

5250989

8

36.7

76.3

2.73 × 10-17

9.01 × 10-15

ABC-family proteins mediated transport

382556

10

36.7

68.2

1.58 × 10-15

1.86 × 10-13

Gαz signalling events

418597

10

36.7

59.9

9.97 × 10-14

5.48 × 10-12

Regulation of IGF transport and uptake by IGFBPs

381426

9

36.7

56.3

5.88 × 10-13

2.11 × 10-11

GP1b-IX-V activation signalling

430116

8

36.7

55.7

8.20 × 10-13

2.76 × 10-11

GABA receptor activation

977443

12

36.7

55.1

1.07 × 10-12

3.26 × 10-11

Vasopressin regulates renal water homeostasis via Aquaporins

432040

9

36.7

54.1

1.77 × 10-12

4.88 × 10-11

Toxicity of botulinum toxin type D (BoNT/D)

5250955

14

36.7

53.4

2.54 × 10-12

6.64 × 10-11

Toxicity of botulinum toxin type F (BoNT/F)

5250981

14

36.7

53.4

2.54 × 10-12

6.64 × 10-11

STAT6-mediated induction of chemokines

3249367

16

36.7

52.2

4.72 × 10-12

1.13 × 10-10

Toxicity of botulinum toxin type B (BoNT/B)

5250958

14

36.7

50.8

9.5 × 10-12

1.98 × 10-10

S6K1 signalling

165720

12

36.7

50.2

1.24 × 10-11

2.5 × 10-10

Gαs signalling events

418555

11

36.7

49.2

2.08 × 10-11

3.85 × 10-10

RHO GTPases activate CIT

5625900

14

36.7

48.2

3.34 × 10-11

5.9 × 10-10

NADE modulates death signalling

205025

15

36.7

47.4

5.00 × 10-11

8.32 × 10-10

Keratan sulfate degradation

2022857

10

36.7

46.6

7.5 × 10-11

1.15 × 10-9

Signaling by Retinoic Acid

5362517

10

36.7

46.6

7.5 × 10-11

1.15 × 10-9

Adenylate cyclase inhibitory pathway

170670

14

36.7

45.9

1.11 × 10-10

1.59 × 10-9

Inhibition of adenylate cyclase pathway

997269

14

36.7

45.9

1.11 × 10-10

1.59 × 10-9

Fatty acids

211935

6

36.7

45.7

1.21 × 10-10

1.72 × 10-9

Ionotropic activity of Kainate Receptors

451306

13

36.7

44.6

2.03 × 10-10

2.58 × 10-9

Activation of Ca-permeable Kainate Receptor

451308

13

36.7

44.6

2.03 × 10-10

2.58 × 10-9

RA biosynthesis pathway

5365859

13

36.7

44.6

2.03 × 10-10

2.58 × 10-9

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

Strongest candidate SL partners for CDH1 by mtSLIPT with observed and expected mutant samples with low expression of partner metagenes

D.6 Mutation Variation

Mutations have different effects as shown by the following examples in cancer genes.

D.6.1 Mutation Frequency

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

(a) PI3KCA

(b) PI3KR1

(c) CDH1

(d) TP53

Figure D.6: Somatic mutation locus. Mutation frequency at each locus in TCGA breast cancer. PIK3CA shows clear recurrent E545K and H1047R oncogene mutations consistent with it being an oncogene. PIK3R1 and CDH1 are tumour suppressors with inactivating mutations distributed throughout the gene, whereas TP53 exhibits both of these properties and a very high mutation frequency compared to other genes.

D.6.2 PI3K Mutation Expression

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

(a) PIK3CA mutation

(b) PIK3CA or PIK3R1 mutation

(c) CDH1 mutation

(d) TP53 mutation

Figure D.7: Somatic mutation against PIK3CA metagene. Mutations in PIK3CA, PIK3R1, CDH1, and TP53 were examined in TCGA breast cancer for their effect on the PIK3CA (Gatza et al., 2014) pathway metagene. The tumour suppressors CDH1 and TP53 showed an increase and decrease in the metagene respectively, whereas PIK3CA and PIK3R1 mutations weaker evidence of decrease in metagene levels.

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

(a) PIK3CA mutation

(b) PIK3CA or PIK3R1 mutation

(c) CDH1 mutation

(d) TP53 mutation

Figure D.8: Somatic mutation against PI3K protein. Mutations in PIK3CA, PIK3R1, CDH1, and TP53 were examined in TCGA breast cancer for their effect on the expression of the p110α protein (encoded by PIK3CA). Protein levels were significantly elevated in samples with PIK3CA or PIK3R1 mutations and lower in samples with TP53 mutations.

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

(a) PIK3CA mutation

(b) PIK3CA or PIK3R1 mutation

(c) CDH1 mutation

(d) TP53 mutation

Figure D.9: Somatic mutation against AKT protein. Mutations in PIK3CA, PIK3R1, CDH1, and TP53 were examined in TCGA breast cancer for their effect on the expression of the AKT protein (a downstream target of PIK3CA). Protein levels were not significantly different in samples mutations in any of these cancer genes.

Appendix EMetagene Expression Profiles

Figure E.1: Pathway metagene expression profiles. Expression profiles for metagene signatures from Gatza et al. (2014) in TCGA breast data, annotated for clinical factors and cancer gene mutations.

Figure E.2: Expression profiles for constituent genes of PI3K. Expression profiles the genes contained in the PI3K gene signature from Gatza et al. (2011) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by CDH1 expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the PI3K metagene, although considerable proportion were inversely correlated with the metagene. Normal samples had low PI3K metagene expression and TP53 mutant samples had high PI3K expression. Although, oncogenic PIK3CA and tumour suppressor PIK3R1 mutations across samples including those with low metagene response.

Figure E.3: Expression profiles for p53 related genes. Expression profiles the genes contained in the TP53 gene signature from Gatza et al. (2011) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by CDH1 expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the metagene, with few very exceptions. TP53 mutant samples had low metagene expression, consistent with loss of tumour suppressor functions, and were less likely to have CDH1 or PIK3CA mutations.

Figure E.4: Expression profiles for estrogen receptor related genes. Expression profiles the genes contained in the estrogen receptor (ER) gene signature from Gatza et al. (2011) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by CDH1 expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the metagene, with very few exceptions being inversely correlated. Estrogen receptor (by antibody staining) negative samples had low metagene expression, as expected. These were more likely to be ductal and basal subtypes, lacking CDH1 or PIK3CA mutations.

Figure E.5: Expression profiles for BRCA related genes. Expression profiles the genes contained in the gene signature related to BRCA1 and BRCA2 functions from Gatza et al. (2014) in TCGA breast data, annotated for clinical factors and cancer gene mutations. Samples are separated by CDH1 expression status and sorted by the metagene. In both cases, the majority of genes were consistent with the direction of the metagene, with few very exceptions. BRCA1 and BRCA2 mutant samples had higher metagene expression than most samples for the ductal subtype, although this was not the case (for the lobular samples for which the metagene was lower). However, the metagene was higher for basal subtype and estrogen receptor negative samples.

Appendix FStomach Expression Analysis

The following results are a replication of the TCGA results (in Chapter 4) with stomach cancer data, using synthetic lethality (SLIPT) against CDH1 mutation.

F.1 Synthetic Lethal Genes and Pathways

Table F.1: Synthetic lethal gene partners of CDH1 from SLIPT in stomach cancer

Gene

Observed

Expected

χ2 value

p-value

p-value (FDR)

PRAF2

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-2-3"

class="td11"> 50.4

121

3.54 × 10-25

1.45 × 10-21

EMP3

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-3-3"

class="td11"> 50.4

115

5.06 × 10-24

1.48 × 10-20

PLEKHO1

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-4-3"

class="td11"> 50.4

112

2.14 × 10-23

4.75 × 10-20

SELM

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-5-3"

class="td11"> 50.4

111

5.13 × 10-23

8.09 × 10-20

GYPC

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-6-3"

class="td11"> 50.4

110

5.77 × 10-23

8.45 × 10-20

COX7A1

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-7-3"

class="td11"> 50.4

109

1.15 × 10-22

1.39 × 10-19

TNFSF12

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-8-3"

class="td11"> 50.4

106

4.06 × 10-22

4.38 × 10-19

SEPT4

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-9-3"

class="td11"> 50.4

106

6.58 × 10-22

5.91 × 10-19

LGALS1

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-10-3"

class="td11"> 50.4

105

6.64 × 10-22

5.91 × 10-19

RARRES2

27 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-11-3"

class="td11"> 50.4

105

8.02 × 10-22

6.85 × 10-19

VEGFB

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-12-3"

class="td11"> 50.4

104

1.19 × 10-21

9.74 × 10-19

PRR24

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-13-3"

class="td11"> 50.4

102

2.96 × 10-21

2.02 × 10-18

SYNC

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-14-3"

class="td11"> 50.4

102

3.73 × 10-21

2.39 × 10-18

MAGEH1

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-15-3"

class="td11"> 50.4

100

9.52 × 10-21

5.01 × 10-18

HSPB2

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-16-3"

class="td11"> 50.4

99.6

1.19 × 10-20

5.82 × 10-18

SMARCD3

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-17-3"

class="td11"> 50.4

99

1.59 × 10-20

7.57 × 10-18

CREM

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-18-3"

class="td11"> 50.4

98.1

2.48 × 10-20

1.13 × 10-17

GNG11

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-19-3"

class="td11"> 50.4

97.3

3.68 × 10-20

1.59 × 10-17

GNAI2

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-20-3"

class="td11"> 50.4

96.4

5.75 × 10-20

2.36 × 10-17

FUNDC2

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-21-3"

class="td11"> 50.4

95.9

7.39 × 10-20

2.91 × 10-17

CNRIP1

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-22-3"

class="td11"> 50.4

95.3

1.0 × 10-19

3.66 × 10-17

CALHM2

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-23-3"

class="td11"> 50.4

93.1

2.94 × 10-19

1.06 × 10-16

ARID5A

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-24-3"

class="td11"> 50.4

92.7

3.47 × 10-19

1.22 × 10-16

ST3GAL3

27 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-25-3"

class="td11"> 50.4

92.2

4.49 × 10-19

1.56 × 10-16

LOC339524

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-100-26-3"

class="td11"> 50.4

92.1

4.8 × 10-19

1.59 × 10-16

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

SLIPT partners of CDH1 with observed and expected mutant samples of both genes

Table F.2: Pathway composition for clusters of CDH1 partners in stomach SLIPT

Pathways Over-represented in Cluster 1

Pathway Size

Cluster Genes

p-value (FDR)

Viral mRNA Translation

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-2-3"

class="td11"> 48

1<span

class="cmmi-12">.3 × 10-97

Formation of a pool of free 40S subunits

94 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-3-3"

class="td11"> 51

1<span

class="cmmi-12">.3 × 10-97

Eukaryotic Translation Elongation

87 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-4-3"

class="td11"> 49

4<span

class="cmmi-12">.8 × 10-97

Peptide chain elongation

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-5-3"

class="td11"> 48

1<span

class="cmmi-12">.4 × 10-96

Eukaryotic Translation Termination

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-6-3"

class="td11"> 48

1<span

class="cmmi-12">.4 × 10-96

GTP hydrolysis and joining of the 60S ribosomal subunit

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-7-3"

class="td11"> 52

7<span

class="cmmi-12">.9 × 10-94

Nonsense Mediated Decay independent of the Exon Junction Complex

89 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-8-3"

class="td11"> 48

3<span

class="cmmi-12">.1 × 10-93

L13a-mediated translational silencing of Ceruloplasmin expression

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-9-3"

class="td11"> 51

5<span

class="cmmi-12">.1 × 10-92

3’ -UTR-mediated translational regulation

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-10-3"

class="td11"> 51

5<span

class="cmmi-12">.1 × 10-92

SRP-dependent cotranslational protein targeting to membrane

105 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-11-3"

class="td11"> 51

1<span

class="cmmi-12">.7 × 10-91

Eukaryotic Translation Initiation

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-12-3"

class="td11"> 52

3<span

class="cmmi-12">.3 × 10-90

Cap-dependent Translation Initiation

112 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-13-3"

class="td11"> 52

3<span

class="cmmi-12">.3 × 10-90

Translation

142 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-14-3"

class="td11"> 56

3<span

class="cmmi-12">.6 × 10-85

Nonsense-Mediated Decay

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-15-3"

class="td11"> 48

1<span

class="cmmi-12">.2 × 10-84

Nonsense Mediated Decay enhanced by the Exon Junction Complex

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-16-3"

class="td11"> 48

1<span

class="cmmi-12">.2 × 10-84

Influenza Viral RNA Transcription and Replication

109 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-17-3"

class="td11"> 48

4<span

class="cmmi-12">.1 × 10-82

Influenza Life Cycle

113 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-18-3"

class="td11"> 48

3<span

class="cmmi-12">.4 × 10-80

Influenza Infection

118 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-19-3"

class="td11"> 48

6<span

class="cmmi-12">.4 × 10-78

Infectious disease

349 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-20-3"

class="td11"> 68

1<span

class="cmmi-12">.8 × 10-50

Formation of the ternary complex, and subsequently, the 43S complex

48 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-21-3"

class="td11"> 21

3<span

class="cmmi-12">.7 × 10-43

</td>

Pathways Over-represented in Cluster 2

Pathway Size

Cluster Genes

p-value (FDR)

Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell

65 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-24-3"

class="td11"> 12

1<span

class="cmmi-12">.3 × 10-15

Phosphorylation of CD3 and TCR zeta chains

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-25-3"

class="td11"> 6

1<span

class="cmmi-12">.7 × 10-12

Generation of second messenger molecules

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-26-3"

class="td11"> 7

2<span

class="cmmi-12">.7 × 10-12

PD-1 signaling

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-27-3"

class="td11"> 6

7.4 <span

class="cmsy-10x-x-120">×10-12

TCR signaling

62 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-28-3"

class="td11"> 9

4.3 <span

class="cmsy-10x-x-120">×10-11

Translocation of ZAP-70 to Immunological synapse

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-29-3"

class="td11"> 5

1.1 <span

class="cmsy-10x-x-120">×10-10

Interferon alpha/beta signaling

68 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-30-3"

class="td11"> 9

1.6 <span

class="cmsy-10x-x-120">×10-10

Initial triggering of complement

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-31-3"

class="td11"> 5

1.6 <span

class="cmsy-10x-x-120">×10-10

IKK complex recruitment mediated by RIP1

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-32-3"

class="td11"> 5

5.1 <span

class="cmsy-10x-x-120">×10-10

TRIF-mediated programmed cell death

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-33-3"

class="td11"> 4

6.2 <span

class="cmsy-10x-x-120">×10-10

Creation of C4 and C2 activators

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-34-3"

class="td11"> 4

1.3 <span

class="cmsy-10x-x-120">×10-9

RHO GTPases Activate NADPH Oxidases

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-35-3"

class="td11"> 4

1.3 <span

class="cmsy-10x-x-120">×10-9

Interferon Signaling

175 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-36-3"

class="td11"> 15

2<span

class="cmmi-12">.3 × 10-9

Chemokine receptors bind chemokines

52 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-37-3"

class="td11"> 7

4<span

class="cmmi-12">.0 × 10-9

Interferon gamma signaling

74 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-38-3"

class="td11"> 8

1<span

class="cmmi-12">.6 × 10-8

TRAF6 mediated induction of TAK1 complex

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-39-3"

class="td11"> 4

1<span

class="cmmi-12">.6 × 10-8

Activation of IRF3/IRF7 mediated by TBK1/IKK epsilon

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-40-3"

class="td11"> 4

2<span

class="cmmi-12">.7 × 10-8

Downstream TCR signaling

45 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-41-3"

class="td11"> 6

3<span

class="cmmi-12">.5 × 10-8

Ligand-dependent caspase activation

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-42-3"

class="td11"> 4

4<span

class="cmmi-12">.2 × 10-8

Complement cascade

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-43-3"

class="td11"> 5

1<span

class="cmmi-12">.3 × 10-7

</td>

Pathways Over-represented in Cluster 3

Pathway Size

Cluster Genes

p-value (FDR)

Uptake and actions of bacterial toxins

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-46-3"

class="td11"> 4

3<span

class="cmmi-12">.5 × 10-6

Neurotoxicity of clostridium toxins

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-47-3"

class="td11"> 3

3<span

class="cmmi-12">.5 × 10-6

Activation of PPARGC1A (PGC-1alpha) by phosphorylation

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-48-3"

class="td11"> 3

3<span

class="cmmi-12">.5 × 10-6

SMAD2/SMAD3:SMAD4 heterotrimer regulates transcription

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-49-3"

class="td11"> 4

1<span

class="cmmi-12">.4 × 10-5

Assembly of the primary cilium

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-50-3"

class="td11"> 10

2<span

class="cmmi-12">.5 × 10-5

Serotonin Neurotransmitter Release Cycle

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-51-3"

class="td11"> 3

2<span

class="cmmi-12">.5 × 10-5

Glycosaminoglycan metabolism

114 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-52-3"

class="td11"> 8

3<span

class="cmmi-12">.3 × 10-5

Platelet homeostasis

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-53-3"

class="td11"> 5

3<span

class="cmmi-12">.3 × 10-5

Norepinephrine Neurotransmitter Release Cycle

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-54-3"

class="td11"> 3

3<span

class="cmmi-12">.3 × 10-5

Acetylcholine Neurotransmitter Release Cycle

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-55-3"

class="td11"> 3

3<span

class="cmmi-12">.3 × 10-5

G˙αs signalling events

100 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-56-3"

class="td11"> 7

5<span

class="cmmi-12">.5 × 10-5

GABA synthesis, release, reuptake and degradation

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-57-3"

class="td11"> 3

5<span

class="cmmi-12">.6 × 10-5

deactivation of the beta-catenin transactivating complex

39 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-58-3"

class="td11"> 4

6<span

class="cmmi-12">.7 × 10-5

Dopamine Neurotransmitter Release Cycle

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-59-3"

class="td11"> 3

6<span

class="cmmi-12">.7 × 10-5

IRS-related events triggered by IGF1R

83 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-60-3"

class="td11"> 6

7<span

class="cmmi-12">.1 × 10-5

Generic Transcription Pathway

186 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-61-3"

class="td11"> 11

7<span

class="cmmi-12">.1 × 10-5

Termination of O-glycan biosynthesis

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-62-3"

class="td11"> 3

7<span

class="cmmi-12">.4 × 10-5

Kinesins

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-63-3"

class="td11"> 3

8<span

class="cmmi-12">.5 × 10-5

Signaling by Type 1 Insulin-like Growth Factor 1 Receptor (IGF1R)

86 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-64-3"

class="td11"> 6

8<span

class="cmmi-12">.5 × 10-5

IGF1R signaling cascade

86 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-65-3"

class="td11"> 6

8<span

class="cmmi-12">.5 × 10-5

</td>

Pathways Over-represented in Cluster 4

Pathway Size

Cluster Genes

p-value (FDR)

Extracellular matrix organization

241 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-68-3"

class="td11"> 97

8<span

class="cmmi-12">.8 × 10-126

Axon guidance

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-69-3"

class="td11"> 75

8<span

class="cmmi-12">.3 × 10-72

Hemostasis

445 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-70-3"

class="td11"> 101

8<span

class="cmmi-12">.3 × 10-72

Developmental Biology

432 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-71-3"

class="td11"> 95

3<span

class="cmmi-12">.0 × 10-67

Response to elevated platelet cytosolic Ca2ˆ+

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-72-3"

class="td11"> 37

5<span

class="cmmi-12">.8 × 10-67

Platelet degranulation

79 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-73-3"

class="td11"> 36

5<span

class="cmmi-12">.8 × 10-67

Degradation of the extracellular matrix

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-74-3"

class="td11"> 39

6<span

class="cmmi-12">.7 × 10-63

Platelet activation, signaling and aggregation

186 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-75-3"

class="td11"> 52

6<span

class="cmmi-12">.6 × 10-62

ECM proteoglycans

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-76-3"

class="td11"> 31

8<span

class="cmmi-12">.1 × 10-61

Neuronal System

272 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-77-3"

class="td11"> 64

5<span

class="cmmi-12">.1 × 10-60

Signaling by PDGF

173 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-78-3"

class="td11"> 47

9<span

class="cmmi-12">.7 × 10-57

Integrin cell surface interactions

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-79-3"

class="td11"> 31

1<span

class="cmmi-12">.9 × 10-53

Collagen biosynthesis and modifying enzymes

56 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-80-3"

class="td11"> 26

1<span

class="cmmi-12">.1 × 10-52

Collagen formation

67 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-81-3"

class="td11"> 28

1<span

class="cmmi-12">.4 × 10-52

Class A/1 (Rhodopsin-like receptors)

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-82-3"

class="td11"> 61

2<span

class="cmmi-12">.3 × 10-52

GPCR ligand binding

373 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-83-3"

class="td11"> 73

2<span

class="cmmi-12">.8 × 10-52

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-84-3"

class="td11"> 22

4<span

class="cmmi-12">.7 × 10-52

Non-integrin membrane-ECM interactions

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-85-3"

class="td11"> 24

7<span

class="cmmi-12">.0 × 10-49

Glycosaminoglycan metabolism

114 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-86-3"

class="td11"> 33

4<span

class="cmmi-12">.7 × 10-47

Platelet homeostasis

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-101-87-3"

class="td11"> 23

1<span

class="cmmi-12">.0 × 10-45

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

F.2 Comparison to Primary Screen

The synthetic lethal partners with CDH1 expression in stoamch cancers were also compared to siRNA primary screen data (Telford et al., 2015), as performed in Section 4.2.1. These are expected to be more concordant with the experimental results performed on a null mutant, however this not the case at the gene level: less genes overlapped with experimental candidates in Figure F.1. This may be affected by lower sample size for mutations in TCGA data or lower frequency (expected value) of CDH1 mutations compared to low expression.

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

Figure F.1: Comparison of SLIPT in stomach to siRNA. Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ2 test suggests that the overlap is no more than would be expected by chance (p = 0.281).

Table F.3: Pathway composition for CDH1 partners from SLIPT and siRNA screening

Predicted only by SLIPT (3392 genes)

Pathway Size

Genes Identified

p-value (FDR)

Extracellular matrix organization

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-2-3"

class="td11"> 90

3.4 × 10-107

Eukaryotic Translation Termination

79 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-3-3"

class="td11"> 46

7.6 × 10-91

Viral mRNA Translation

77 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-4-3"

class="td11"> 45

1.2 × 10-89

Eukaryotic Translation Elongation

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-5-3"

class="td11"> 46

5.8 × 10-89

Peptide chain elongation

79 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-6-3"

class="td11"> 45

2.1 × 10-88

Nonsense Mediated Decay independent of the Exon Junction Complex

84 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-7-3"

class="td11"> 46

9.4 × 10-88

Formation of a pool of free 40S subunits

89 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-8-3"

class="td11"> 47

3.3 × 10-87

GTP hydrolysis and joining of the 60S ribosomal subunit

100 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-9-3"

class="td11"> 48

3.2 × 10-83

Axon guidance

284 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-10-3"

class="td11"> 84

3.9 × 10-82

Developmental Biology

426 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-11-3"

class="td11"> 111

4.2 × 10-82

L13a-mediated translational silencing of Ceruloplasmin expression

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-12-3"

class="td11"> 47

1.4 × 10-81

3’ -UTR-mediated translational regulation

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-13-3"

class="td11"> 47

1.4 × 10-81

SRP-dependent cotranslational protein targeting to membrane

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-14-3"

class="td11"> 47

1.4 × 10-81

Nonsense-Mediated Decay

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-15-3"

class="td11"> 47

1.4 × 10-81

Nonsense Mediated Decay enhanced by the Exon Junction Complex

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-16-3"

class="td11"> 47

1.4 × 10-81

Hemostasis

438 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-17-3"

class="td11"> 112

1.2 × 10-80

Eukaryotic Translation Initiation

107 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-18-3"

class="td11"> 48

8.0 × 10-80

Cap-dependent Translation Initiation

107 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-19-3"

class="td11"> 48

8.0 × 10-80

Infectious disease

338 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-20-3"

class="td11"> 90

1.6 × 10-76

Neuronal System

267 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-21-3"

class="td11"> 77

1.6 × 10-76

</td>

Detected only by siRNA screen (1803 genes)

Pathway Size

Genes Identified

p-value (FDR)

Class A/1 (Rhodopsin-like receptors)

282 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-24-3"

class="td11"> 62

8.1 × 10-50

GPCR ligand binding

363 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-25-3"

class="td11"> 71

4.9 × 10-46

Peptide ligand-binding receptors

175 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-26-3"

class="td11"> 38

7.9 × 10-38

G˙αi signalling events

184 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-27-3"

class="td11"> 37

1.1 × 10-34

Gastrin-CREB signalling pathway via PKC and MAPK

180 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-28-3"

class="td11"> 35

1.4 × 10-32

G˙αq signalling events

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-29-3"

class="td11"> 32

4.8 × 10-32

DAP12 interactions

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-30-3"

class="td11"> 29

1.4 × 10-27

Downstream signal transduction

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-31-3"

class="td11"> 26

2.4 × 10-25

DAP12 signaling

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-32-3"

class="td11"> 26

6.4 × 10-25

VEGFA-VEGFR2 Pathway

91 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-33-3"

class="td11"> 19

8.1 × 10-24

Signaling by PDGF

172 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-34-3"

class="td11"> 27

5.7 × 10-23

Signaling by ERBB2

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-35-3"

class="td11"> 24

1.4 × 10-22

Signaling by VEGF

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-36-3"

class="td11"> 19

2.0 × 10-22

Visual phototransduction

85 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-37-3"

class="td11"> 17

1.3 × 10-21

Downstream signaling of activated FGFR1

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-38-3"

class="td11"> 22

1.3 × 10-21

Downstream signaling of activated FGFR2

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-39-3"

class="td11"> 22

1.3 × 10-21

Downstream signaling of activated FGFR3

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-40-3"

class="td11"> 22

1.3 × 10-21

Downstream signaling of activated FGFR4

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-41-3"

class="td11"> 22

1.3 × 10-21

Signaling by FGFR

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-42-3"

class="td11"> 23

2.0 × 10-21

Signaling by FGFR1

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-43-3"

class="td11"> 23

2.0 × 10-21

</td>

Intersection of SLIPT and siRNA screen (547 genes)

Pathway Size

Genes Identified

p-value (FDR)

Class A/1 (Rhodopsin-like receptors)

282 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-46-3"

class="td11"> 25

3.9 × 10-9

Platelet activation, signaling and aggregation

182 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-47-3"

class="td11"> 17

3.9 × 10-9

Response to elevated platelet cytosolic Ca2+

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-48-3"

class="td11"> 9

5.5 × 10-8

Platelet homeostasis

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-49-3"

class="td11"> 7

5.7 × 10-8

Nucleotide-like (purinergic) receptors

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-50-3"

class="td11"> 4

1.8 × 10-7

Platelet degranulation

77 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-51-3"

class="td11"> 8

2.8 × 10-7

Peptide ligand-binding receptors

175 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-52-3"

class="td11"> 14

3.8 × 10-7

Molecules associated with elastic fibres

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-53-3"

class="td11"> 5

7.1 × 10-7

Amine ligand-binding receptors

35 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-54-3"

class="td11"> 5

8.6 × 10-7

G˙αi signalling events

184 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-55-3"

class="td11"> 14

9.8 × 10-7

GPCR ligand binding

363 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-56-3"

class="td11"> 27

1.1 × 10-6

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-57-3"

class="td11"> 5

1.5 × 10-6

G˙αq signalling events

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-58-3"

class="td11"> 12

1.9 × 10-6

Serotonin receptors

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-59-3"

class="td11"> 3

3.8 × 10-6

P2Y receptors

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-60-3"

class="td11"> 3

3.8 × 10-6

Signal amplification

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-61-3"

class="td11"> 3

2.3 × 10-5

Gastrin-CREB signalling pathway via PKC and MAPK

180 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-62-3"

class="td11"> 12

2.3 × 10-5

Complement cascade

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-63-3"

class="td11"> 4

2.4 × 10-5

Glycosaminoglycan metabolism

110 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-64-3"

class="td11"> 8

2.5 × 10-5

Glycogen breakdown (glycogenolysis)

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-102-65-3"

class="td11"> 3

2.7 × 10-5

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

F.2.1 Resampling Analysis

Table F.4: Pathways for CDH1 partners from SLIPT in stomach cancer

Reactome Pathway

Over-representation

Permutation

Extracellular matrix organization

7<span

class="cmmi-12">.5 × 10-140

0<span

class="cmmi-12">.070215

Hemostasis

1<span

class="cmmi-12">.8 × 10-121

0<span

class="cmmi-12">.25804

Developmental Biology

9<span

class="cmmi-12">.2 × 10-107

0<span

class="cmmi-12">.53032

Axon guidance

1<span

class="cmmi-12">.5 × 10-102

0<span

class="cmmi-12">.6704

Eukaryotic Translation Termination

1<span

class="cmmi-12">.9 × 10-99

> 1.031 × 10-5

GPCR ligand binding

3<span

class="cmmi-12">.8 × 10-99

0<span

class="cmmi-12">.54914

Viral mRNA Translation

3<span

class="cmmi-12">.3 × 10-98

> 1.031 × 10-5

Formation of a pool of free 40S subunits

3<span

class="cmmi-12">.3 × 10-98

> 1.031 × 10-5

Eukaryotic Translation Elongation

1<span

class="cmmi-12">.6 × 10-97

> 1.031 × 10-5

Peptide chain elongation

7<span

class="cmmi-12">.2 × 10-97

> 1.031 × 10-5

Class A/1 (Rhodopsin-like receptors)

2<span

class="cmmi-12">.7 × 10-96

0<span

class="cmmi-12">.58174

Nonsense Mediated Decay independent of the Exon Junction Complex

3 <span

class="cmsy-10x-x-120">× 10-96

> 1.031 × 10-5

Infectious disease

2<span

class="cmmi-12">.6 × 10-94

0<span

class="cmmi-12">.25484

GTP hydrolysis and joining of the 60S ribosomal subunit

3<span

class="cmmi-12">.4 × 10-94

> 1.031 × 10-5

L13a-mediated translational silencing of Ceruloplasmin expression

2<span

class="cmmi-12">.8 × 10-92

> 1.031 × 10-5

3’ -UTR-mediated translational regulation

2<span

class="cmmi-12">.8 × 10-92

> 1.031 × 10-5

Neuronal System

8<span

class="cmmi-12">.4 × 10-92

0<span

class="cmmi-12">.53433

SRP-dependent cotranslational protein targeting to membrane

9<span

class="cmmi-12">.5 × 10-92

> 1.031 × 10-5

Eukaryotic Translation Initiation

2<span

class="cmmi-12">.0 × 10-90

> 1.031 × 10-5

Cap-dependent Translation Initiation

2<span

class="cmmi-12">.0 × 10-90

> 1.031 × 10-5

Nonsense-Mediated Decay

7<span

class="cmmi-12">.4 × 10-90

> 1.031 × 10-5

Nonsense Mediated Decay enhanced by the Exon Junction Complex

7<span

class="cmmi-12">.4 × 10-90

> 1.031 × 10-5

Adaptive Immune System

8<span

class="cmmi-12">.1 × 10-88

0<span

class="cmmi-12">.14116

Translation

1<span

class="cmmi-12">.3 × 10-87

> 1.031 × 10-5

Platelet activation, signaling and aggregation

1<span

class="cmmi-12">.3 × 10-86

0<span

class="cmmi-12">.28959

Influenza Infection

1 <span

class="cmsy-10x-x-120">× 10-82

> 1.031 × 10-5

Influenza Viral RNA Transcription and Replication

2<span

class="cmmi-12">.4 × 10-82

> 1.031 × 10-5

Influenza Life Cycle

2 <span

class="cmsy-10x-x-120">× 10-80

> 1.031 × 10-5

Response to elevated platelet cytosolic Ca2+

4<span

class="cmmi-12">.9 × 10-78

0<span

class="cmmi-12">.50817

Signalling by NGF

1<span

class="cmmi-12">.6 × 10-75

0<span

class="cmmi-12">.38518

Rho GTPase cycle

5<span

class="cmmi-12">.1 × 10-75

0<span

class="cmmi-12">.14864

Signaling by PDGF

7<span

class="cmmi-12">.4 × 10-74

0<span

class="cmmi-12">.40493

Signaling by Rho GTPases

5<span

class="cmmi-12">.1 × 10-73

0<span

class="cmmi-12">.077217

Glycosaminoglycan metabolism

1<span

class="cmmi-12">.4 × 10-68

0<span

class="cmmi-12">.52984

G˙αi signalling events

1<span

class="cmmi-12">.8 × 10-66

0<span

class="cmmi-12">.9254

Metabolism of carbohydrates

1<span

class="cmmi-12">.1 × 10-65

0<span

class="cmmi-12">.39501

G˙αs signalling events

2<span

class="cmmi-12">.7 × 10-65

0.0050293

Potassium Channels

2<span

class="cmmi-12">.7 × 10-65

0<span

class="cmmi-12">.53359

Transmission across Chemical Synapses

1<span

class="cmmi-12">.8 × 10-64

0<span

class="cmmi-12">.81833

ECM proteoglycans

3<span

class="cmmi-12">.4 × 10-64

0<span

class="cmmi-12">.083482

Peptide ligand-binding receptors

4<span

class="cmmi-12">.8 × 10-64

0<span

class="cmmi-12">.62817

Degradation of the extracellular matrix

1<span

class="cmmi-12">.1 × 10-63

0<span

class="cmmi-12">.80879

Platelet homeostasis

5<span

class="cmmi-12">.3 × 10-63

0<span

class="cmmi-12">.53134

NGF signalling via TRKA from the plasma membrane

6<span

class="cmmi-12">.1 × 10-63

0<span

class="cmmi-12">.5717

Integration of energy metabolism

4<span

class="cmmi-12">.5 × 10-61

0<span

class="cmmi-12">.10889

Collagen formation

5<span

class="cmmi-12">.4 × 10-61

0<span

class="cmmi-12">.29896

Integrin cell surface interactions

7 <span

class="cmsy-10x-x-120">× 10-59

0<span

class="cmmi-12">.18167

Collagen biosynthesis and modifying enzymes

7 <span

class="cmsy-10x-x-120">× 10-59

0<span

class="cmmi-12">.30208

Neurotransmitter Receptor Binding And Downstream Transmission

In The Postsynaptic Cell

8<span

class="cmmi-12">.7 × 10-57

0<span

class="cmmi-12">.82522

Signaling by Wnt

8<span

class="cmmi-12">.7 × 10-57

0<span

class="cmmi-12">.25468

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

Table F.5: Pathways for CDH1 partners from SLIPT in stomach and siRNA screen

Reactome Pathway

Over-representation

Permutation

Platelet activation, signaling and aggregation

3<span

class="cmmi-12">.9 × 10-9

0<span

class="cmmi-12">.49557

Class A/1 (Rhodopsin-like receptors)

3<span

class="cmmi-12">.9 × 10-9

0<span

class="cmmi-12">.98432

Response to elevated platelet cytosolic Ca2+

5<span

class="cmmi-12">.5 × 10-8

0<span

class="cmmi-12">.54349

Platelet homeostasis

5<span

class="cmmi-12">.7 × 10-8

0<span

class="cmmi-12">.45017

Nucleotide-like (purinergic) receptors

1<span

class="cmmi-12">.8 × 10-7

0<span

class="cmmi-12">.36966

Peptide ligand-binding receptors

3<span

class="cmmi-12">.8 × 10-7

0<span

class="cmmi-12">.91294

Molecules associated with elastic fibres

7<span

class="cmmi-12">.1 × 10-7

0<span

class="cmmi-12">.0025868

Amine ligand-binding receptors

8<span

class="cmmi-12">.6 × 10-7

0<span

class="cmmi-12">.43303

G˙αi signalling events

9<span

class="cmmi-12">.8 × 10-7

0<span

class="cmmi-12">.99626

GPCR ligand binding

1<span

class="cmmi-12">.1 × 10-6

0<span

class="cmmi-12">.97733

Elastic fibre formation

1<span

class="cmmi-12">.5 × 10-6

0<span

class="cmmi-12">.0025868

G˙αq signalling events

1<span

class="cmmi-12">.9 × 10-6

0<span

class="cmmi-12">.86089

P2Y receptors

3<span

class="cmmi-12">.8 × 10-6

0<span

class="cmmi-12">.18795

Serotonin receptors

3<span

class="cmmi-12">.8 × 10-6

0<span

class="cmmi-12">.37853

Signal amplification

2<span

class="cmmi-12">.3 × 10-5

0<span

class="cmmi-12">.47856

Gastrin-CREB signalling pathway via PKC and MAPK

2<span

class="cmmi-12">.3 × 10-5

0<span

class="cmmi-12">.98567

Complement cascade

2<span

class="cmmi-12">.4 × 10-5

> 3.4628 × 10-6

Glycosaminoglycan metabolism

2<span

class="cmmi-12">.5 × 10-5

0<span

class="cmmi-12">.38953

Glycogen breakdown (glycogenolysis)

2<span

class="cmmi-12">.7 × 10-5

0<span

class="cmmi-12">.83772

Defective B4GALT7 causes EDS, progeroid type

4<span

class="cmmi-12">.9 × 10-5

0<span

class="cmmi-12">.10792

Defective B3GAT3 causes JDSSDHD

4<span

class="cmmi-12">.9 × 10-5

0<span

class="cmmi-12">.10792

Role of LAT2/NTAL/LAB on calcium mobilization

5<span

class="cmmi-12">.6 × 10-5

0<span

class="cmmi-12">.35373

Cell surface interactions at the vascular wall

5<span

class="cmmi-12">.6 × 10-5

0<span

class="cmmi-12">.47642

G˙αs signalling events

6 <span

class="cmsy-10x-x-120">× 10-5

0<span

class="cmmi-12">.019858

Signaling by NOTCH

6 <span

class="cmsy-10x-x-120">× 10-5

0<span

class="cmmi-12">.19008

A tetrasaccharide linker sequence is required for GAG synthesis

0<span

class="cmmi-12">.00017

0<span

class="cmmi-12">.47642

Extracellular matrix organization

0<span

class="cmmi-12">.00018

0<span

class="cmmi-12">.0047308

Collagen formation

0<span

class="cmmi-12">.00018

0<span

class="cmmi-12">.19245

Effects of PIP2 hydrolysis

0<span

class="cmmi-12">.0002

0<span

class="cmmi-12">.37779

Syndecan interactions

0<span

class="cmmi-12">.0002

0<span

class="cmmi-12">.37779

Diseases associated with glycosaminoglycan metabolism

0<span

class="cmmi-12">.00023

0<span

class="cmmi-12">.01028

Diseases of glycosylation

0<span

class="cmmi-12">.00023

0<span

class="cmmi-12">.01028

Chondroitin sulfate/dermatan sulfate metabolism

0<span

class="cmmi-12">.00023

0<span

class="cmmi-12">.085541

Integrin alphaIIb beta3 signaling

0<span

class="cmmi-12">.00028

0<span

class="cmmi-12">.76936

Keratan sulfate biosynthesis

0<span

class="cmmi-12">.00034

0<span

class="cmmi-12">.68744

Rho GTPase cycle

0<span

class="cmmi-12">.00034

0<span

class="cmmi-12">.15675

Creation of C4 and C2 activators

0<span

class="cmmi-12">.00035

0<span

class="cmmi-12">.12275

Abacavir transport and metabolism

0<span

class="cmmi-12">.00035

0<span

class="cmmi-12">.12443

Amine compound SLC transporters

0<span

class="cmmi-12">.00037

0<span

class="cmmi-12">.69773

FCERI mediated NF-kB activation

0<span

class="cmmi-12">.00037

0<span

class="cmmi-12">.69846

Fc epsilon receptor (FCERI) signaling

0<span

class="cmmi-12">.00056

0<span

class="cmmi-12">.43303

Defective EXT2 causes exostoses 2

0<span

class="cmmi-12">.00067

0<span

class="cmmi-12">.16053

Defective EXT1 causes exostoses 1, TRPS2 and CHDS

0<span

class="cmmi-12">.00067

0<span

class="cmmi-12">.16053

Collagen biosynthesis and modifying enzymes

0<span

class="cmmi-12">.00071

0<span

class="cmmi-12">.052911

Keratan sulfate/keratin metabolism

0<span

class="cmmi-12">.00073

0<span

class="cmmi-12">.46533

G alpha (12/13) signalling events

0<span

class="cmmi-12">.00078

0<span

class="cmmi-12">.59164

SEMA3A-Plexin repulsion signaling by inhibiting Integrin adhesion

0<span

class="cmmi-12">.00084

0<span

class="cmmi-12">.038504

Signal attenuation

0<span

class="cmmi-12">.00084

0<span

class="cmmi-12">.37779

Eicosanoid ligand-binding receptors

0<span

class="cmmi-12">.0011

0<span

class="cmmi-12">.11117

SOS-mediated signalling

0<span

class="cmmi-12">.0011

0<span

class="cmmi-12">.25387

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

F.3 Metagene Analysis

Metagene analysis was also performed for synthetic lethal candidates for CDH1 expression in stomach cancer. These are described and compared to mutation analysis in Section G.4.

Table F.6: Candidate synthetic lethal metagenes against CDH1 from SLIPT in stomach cancer

Pathway

ID

Observed

Expected

χ2value

p-value

p-value (FDR)

Cell-Cell communication

1500931

18

50.4

110

7.43 × 10-23

1.53 × 10-20

VEGFR2 mediated vascular permeability

5218920

19

50.4

109

1.36 × 10-22

2.49 × 10-20

Sema4D in semaphorin signaling

400685

20

50.4

104

1.62 × 10-21

2.12 × 10-19

Ion transport by P-type ATPases

936837

17

50.4

100

8.29 × 10-21

8.06 × 10-19

Sialic acid metabolism

4085001

19

50.4

95.3

9.95 × 10-20

7.82 × 10-18

Synthesis of pyrophosphates in the cytosol

1855167

26

50.4

94

1.86 × 10-19

1.23 × 10-17

Keratan sulfate/keratin metabolism

1638074

25

50.4

93.5

2.36 × 10-19

1.44 × 10-17

Ion channel transport

983712

19

50.4

92.8

3.37 × 10-19

1.99 × 10-17

Keratan sulfate biosynthesis

2022854

26

50.4

91.4

6.79 × 10-19

3.62 × 10-17

Arachidonic acid metabolism

2142753

22

50.4

90.6

9.81 × 10-19

5.07 × 10-17

RHO GTPases activate CIT

5625900

22

50.4

87

5.80 × 10-18

2.66 × 10-16

Stimuli-sensing channels

2672351

25

50.4

85.8

1.03 × 10-17

4.58 × 10-16

Synthesis of PI

1483226

19

50.4

85.6

1.15 × 10-17

4.89 × 10-16

G-protein activation

202040

19

50.4

85.3

1.34 × 10-17

5.53 × 10-16

NrCAM interactions

447038

22

50.4

84.3

2.1 × 10-17

8.27 × 10-16

Inwardly rectifying K+ channels

1296065

24

50.4

83.5

3.19 × 10-17

1.22 × 10-15

Calcitonin-like ligand receptors

419812

20

50.4

82.2

6.07 × 10-17

2.13 × 10-15

Prostacyclin signalling through prostacyclin receptor

392851

24

50.4

81.8

7.27 × 10-17

2.5 × 10-15

Presynaptic function of Kainate receptors

500657

26

50.4

79.7

2.00 × 10-16

6.34 × 10-15

ADP signalling through P2Y purinoceptor 12

392170

23

50.4

79.2

2.57 × 10-16

7.71 × 10-15

regulation of FZD by ubiquitination

4641263

22

50.4

78.8

3.15 × 10-16

9.3 × 10-15

Toxicity of tetanus toxin (TeNT)

5250982

27

50.4

78.7

3.36 × 10-16

9.75 × 10-15

Gap junction degradation

190873

21

50.4

78.5

3.66 × 10-16

1.04 × 10-14

Nephrin interactions

373753

25

50.4

78.2

4.21 × 10-16

1.14 × 10-14

GABA synthesis, release, reuptake and degradation

888590

26

50.4

77

7.69 × 10-16

1.95 × 10-14

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

Strongest candidate SL partners for CDH1 by SLIPT with observed and expected samples with low expression of both genes

Appendix GStomach Mutation Analysis

The following results are a replication of the TCGA results (in Appendix D) with stomach cancer data, using synthetic lethality (mtSLIPT) against CDH1 mutation.

G.1 Synthetic Lethal Genes and Pathways

Table G.1: Synthetic lethal gene partners of CDH1 from mtSLIPT in stomach cancer

Gene

Observed

Expected

χ2 value

p-value

p-value (FDR)

OLFML1

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-2-3"

class="td11"> 10.1

29.2

4.53 × 10-7

0<span

class="cmmi-12">.0031

NRIP2

6 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-3-3"

class="td11"> 10.1

25.4

3.11 × 10-6

0<span

class="cmmi-12">.00706

VIM

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-4-3"

class="td11"> 10.1

24.7

4.29 × 10-6

0<span

class="cmmi-12">.00706

TCF4

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-5-3"

class="td11"> 10.1

24.7

4.33 × 10-6

0<span

class="cmmi-12">.00706

ZEB2

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-6-3"

class="td11"> 10.1

24.7

4.33 × 10-6

0<span

class="cmmi-12">.00706

BCL2

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-7-3"

class="td11"> 10.1

22

1.66 × 10-5

0<span

class="cmmi-12">.0155

SMARCA2

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-8-3"

class="td11"> 10.1

22

1.66 × 10-5

0<span

class="cmmi-12">.0155

CCND2

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-9-3"

class="td11"> 10.1

21.1

2.61 × 10-5

0<span

class="cmmi-12">.0155

MMP19

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-10-3"

class="td11"> 10.1

21.1

2.61 × 10-5

0<span

class="cmmi-12">.0155

NEURL1B

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-11-3"

class="td11"> 10.1

21.1

2.61 × 10-5

0<span

class="cmmi-12">.0155

IGFBP6

6 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-12-3"

class="td11"> 10.1

21.1

2.65 × 10-5

0<span

class="cmmi-12">.0155

OGN

6 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-13-3"

class="td11"> 10.1

21.1

2.65 × 10-5

0<span

class="cmmi-12">.0155

THY1

6 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-14-3"

class="td11"> 10.2

21

2.7 × 10-5

0<span

class="cmmi-12">.0155

DZIP1

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-15-3"

class="td11"> 10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.0155

LOC650368

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-16-3"

class="td11"> 10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.0155

PCOLCE

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-17-3"

class="td11"> 10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.0155

PTGFR

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-18-3"

class="td11"> 10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.0155

RUNX1T1

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-19-3"

class="td11"> 10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.0155

CLEC2B

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-20-3"

class="td11"> 10.1

20.6

3.3 × 10-5

0<span

class="cmmi-12">.0155

MSC

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-21-3"

class="td11"> 10.1

20.6

3.3 × 10-5

0<span

class="cmmi-12">.0155

NISCH

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-22-3"

class="td11"> 10.1

20.6

3.3 × 10-5

0<span

class="cmmi-12">.0155

TSPAN11

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-23-3"

class="td11"> 10.1

20.6

3.3 × 10-5

0<span

class="cmmi-12">.0155

KCTD12

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-24-3"

class="td11"> 10.1

19.1

7.19 × 10-5

0<span

class="cmmi-12">.0246

LRRC55

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-25-3"

class="td11"> 10.1

19.1

7.19 × 10-5

0<span

class="cmmi-12">.0246

PCBP3

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-107-26-3"

class="td11"> 10.1

19.1

7.19 × 10-5

0<span

class="cmmi-12">.0246

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

mtSLIPT partners with observed and expected CDH1 mutant samples with low expression

Table G.2: Pathways for CDH1 partners from mtSLIPT in stomach cancer

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Extracellular matrix organization

241 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-2-3"

class="td11"> 20

9<span

class="cmmi-12">.6 × 10-9

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-3-3"

class="td11"> 6

3<span

class="cmmi-12">.7 × 10-8

Diseases associated with glycosaminoglycan metabolism

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-4-3"

class="td11"> 5

3<span

class="cmmi-12">.7 × 10-8

Diseases of glycosylation

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-5-3"

class="td11"> 5

3<span

class="cmmi-12">.7 × 10-8

Nitric oxide stimulates guanylate cyclase

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-6-3"

class="td11"> 4

3<span

class="cmmi-12">.1 × 10-6

Molecules associated with elastic fibres

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-7-3"

class="td11"> 4

3<span

class="cmmi-12">.7 × 10-5

Platelet homeostasis

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-8-3"

class="td11"> 5

3<span

class="cmmi-12">.7 × 10-5

Initial triggering of complement

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-9-3"

class="td11"> 3

3<span

class="cmmi-12">.7 × 10-5

Regulation of IGF transport and uptake by IGFBPs

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-10-3"

class="td11"> 3

3<span

class="cmmi-12">.7 × 10-5

Collagen degradation

58 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-11-3"

class="td11"> 5

5<span

class="cmmi-12">.6 × 10-5

Defective B4GALT7 causes EDS, progeroid type

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-12-3"

class="td11"> 3

5<span

class="cmmi-12">.6 × 10-5

Defective B3GAT3 causes JDSSDHD

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-13-3"

class="td11"> 3

5<span

class="cmmi-12">.6 × 10-5

Degradation of the extracellular matrix

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-14-3"

class="td11"> 7

8<span

class="cmmi-12">.0 × 10-5

ECM proteoglycans

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-15-3"

class="td11"> 5

0.00017 </td>

A tetrasaccharide linker sequence is required for GAG synthesis x

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-16-3"

class="td11"> 3

0.00025 </td>

RHO GTPases Activate WASPs and WAVEs

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-17-3"

class="td11"> 3

0.00059 </td>

Non-integrin membrane-ECM interactions

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-18-3"

class="td11"> 4

0.00065 </td>

Creation of C4 and C2 activators

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-19-3"

class="td11"> 2

0.00079 </td>

Dermatan sulfate biosynthesis

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-20-3"

class="td11"> 2

0.00079 </td>

Integrin cell surface interactions

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-108-21-3"

class="td11"> 5

0.00098 </td>

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in mtSLIPT partners for CDH1

G.2 Synthetic Lethal Expression Profiles

Similar to the analysis of synthetic lethal partners against low CDH1 expression in F.1, the partners detected from CDH1 mutation were also examined for their expression profiles and the pathway composition of gene clusters. Hierachical clustering was performed on mtSLIPT partners for CDH1 as showing in Figure G.1. Over-representation for Reactome pathways for each of the gene clusters identified is given in Table G.3.

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

Figure G.1: Synthetic lethal expression profiles of analysed samples. Gene expression profile heatmap (correlation distance) of all samples (separated by the 1 ∕ 3 quantile of CDH1 expression) analysed in TCGA stomach cancer dataset for gene expression of 257 candidate partners of E-cadherin (CDH1) from SLIPT prediction (with significant FDR adjusted p < 0.05). Deeply clustered, inter-correlated genes form several main groups, each containing genes that were SL candidates or toxic in an siRNA screen (Telford et al., 2015). Clusters had different sample groups highly expressing the synthetic lethal candidates in CDH1 low samples, notably diffuse and CDH1 mutant samples have elevated expression in one or more distinct clusters, although there was less complexity and variation among candidate synthetic lethal partners than in breast data. CDH1 low samples also contained most of samples with CDH1 mutations.

Table G.3: Pathway composition for clusters of CDH1 partners in stomach mtSLIPT

Pathways Over-represented in Cluster 1

Pathway Size

Cluster Genes

p-value (FDR)

CD28 dependent PI3K/Akt signaling

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-2-3"

class="td11"> 1

1 </td>

Hormone-sensitive lipase (HSL)-mediated triacylglycerol hydrolysis

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-3-3"

class="td11"> 1

1 </td>

CD28 co-stimulation

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-4-3"

class="td11"> 1

1 </td></tr><tr

style="vertical-align:baseline;" id="TBL-109-5-">

Lipid digestion, mobilization, and transport

48

1

1

Costimulation by the CD28 family

51 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-6-3"

class="td11"> 1

1 </td>

Dectin-1 mediated noncanonical NF-kB signaling

58 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-7-3"

class="td11"> 1

1 </td>

CLEC7A (Dectin-1) signaling

99 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-8-3"

class="td11"> 1

1 </td>

C-type lectin receptors (CLRs)

123 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-9-3"

class="td11"> 1

1 </td>

Adaptive Immune System

418 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-10-3"

class="td11"> 1

1 </td>

Metabolism of lipids and lipoproteins

494 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-11-3"

class="td11"> 1

1 </td>

Interleukin-6 signaling

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-12-3"

class="td11"> 0

1 </td>

Apoptosis

150 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-13-3"

class="td11"> 0

1 </td>

Hemostasis

445 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-14-3"

class="td11"> 0

1 </td></tr><tr

style="vertical-align:baseline;" id="TBL-109-15-">

Intrinsic Pathway for Apoptosis

36

0

1

Cleavage of Growing Transcript in the Termination Region

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-16-3"

class="td11"> 0

1 </td>

PKB-mediated events

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-17-3"

class="td11"> 0

1 </td>

PI3K Cascade

68 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-18-3"

class="td11"> 0

1 </td>

RAF/MAP kinase cascade

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-19-3"

class="td11"> 0

1 </td>

Global Genomic NER (GG-NER)

35 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-20-3"

class="td11"> 0

1 </td>

Repair synthesis for gap-filling by DNA polymerase in TC-NER

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-21-3"

class="td11"> 0

1 </td>

</td>

Pathways Over-represented in Cluster 2

Pathway Size

Cluster Genes

p-value (FDR)

Kinesins

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-24-3"

class="td11"> 1

1 </td></tr><tr

style="vertical-align:baseline;" id="TBL-109-25-">

O-linked glycosylation of mucins

49

1

1

O-linked glycosylation

59 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-26-3"

class="td11"> 1

1 </td>

MHC class II antigen presentation

85 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-27-3"

class="td11"> 1

1 </td>

Factors involved in megakaryocyte development and platelet production cellll

120 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-28-3"

class="td11"> 1

1 </td>

Post-translational protein modification

303 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-29-3"

class="td11"> 1

1 </td>

Adaptive Immune System

418 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-30-3"

class="td11"> 1

1 </td>

Hemostasis

445 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-31-3"

class="td11"> 1

1 </td>

Interleukin-6 signaling

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-32-3"

class="td11"> 0

1 </td>

Apoptosis

150 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-33-3"

class="td11"> 0

1 </td>

Intrinsic Pathway for Apoptosis

36 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-34-3"

class="td11"> 0

1 </td>

Cleavage of Growing Transcript in the Termination Region

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-35-3"

class="td11"> 0

1 </td>

PKB-mediated events

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-36-3"

class="td11"> 0

1 </td></tr><tr

style="vertical-align:baseline;" id="TBL-109-37-">

PI3K Cascade

68

0

1

RAF/MAP kinase cascade

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-38-3"

class="td11"> 0

1 </td>

Global Genomic NER (GG-NER)

35 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-39-3"

class="td11"> 0

1 </td>

Repair synthesis for gap-filling by DNA polymerase in TC-NER

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-40-3"

class="td11"> 0

1 </td>

Gap-filling DNA repair synthesis and ligation in TC-NER

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-41-3"

class="td11"> 0

1 </td>

Formation of transcription-coupled NER (TC-NER) repair complex

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-42-3"

class="td11"> 0

1 </td>

Dual incision reaction in TC-NER

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-43-3"

class="td11"> 0

1 </td></tr><tr

class="hline">

Pathways Over-represented in Cluster 3

Pathway Size

Cluster Genes

p-value (FDR)

Extracellular matrix organization

241 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-46-3"

class="td11"> 20

9<span

class="cmmi-12">.6 × 10-9

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-47-3"

class="td11"> 6

3<span

class="cmmi-12">.7 × 10-8

Diseases associated with glycosaminoglycan metabolism

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-48-3"

class="td11"> 5

3<span

class="cmmi-12">.7 × 10-8

Diseases of glycosylation

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-49-3"

class="td11"> 5

3<span

class="cmmi-12">.7 × 10-8

Molecules associated with elastic fibres

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-50-3"

class="td11"> 4

4<span

class="cmmi-12">.8 × 10-5

Initial triggering of complement

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-51-3"

class="td11"> 3

4<span

class="cmmi-12">.8 × 10-5

Regulation of IGF transport and uptake by IGFBPs

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-52-3"

class="td11"> 3

4<span

class="cmmi-12">.8 × 10-5

Collagen degradation

58 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-53-3"

class="td11"> 5

6<span

class="cmmi-12">.7 × 10-5

Defective B4GALT7 causes EDS, progeroid type

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-54-3"

class="td11"> 3

6<span

class="cmmi-12">.7 × 10-5

Defective B3GAT3 causes JDSSDHD

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-55-3"

class="td11"> 3

6<span

class="cmmi-12">.7 × 10-5

Degradation of the extracellular matrix

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-56-3"

class="td11"> 7

9<span

class="cmmi-12">.5 × 10-5

ECM proteoglycans

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-57-3"

class="td11"> 5

0.0002 </td>

A tetrasaccharide linker sequence is required for GAG synthesis

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-58-3"

class="td11"> 5 3

0.00029 </td>

Non-integrin membrane-ECM interactions

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-59-3"

class="td11"> 4

0.00079 </td>

Creation of C4 and C2 activators

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-60-3"

class="td11"> 2

0.00093 </td>

Dermatan sulfate biosynthesis

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-61-3"

class="td11"> 2

0.00093 </td>

Integrin cell surface interactions

82 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-62-3"

class="td11"> 5

0.0012 </td>

Keratan sulfate degradation

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-63-3"

class="td11"> 2

0.0012 </td>

Complement cascade

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-64-3"

class="td11"> 3

0.0013 </td>

CS/DS degradation

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-65-3"

class="td11"> 2

0.0015 </td>

</td>

Pathways Over-represented in Cluster 4

Pathway Size

Cluster Genes

p-value (FDR)

cGMP effects

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-68-3"

class="td11"> 2

0.11 </td>

Nitric oxide stimulates guanylate cyclase

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-69-3"

class="td11"> 2

0.19 </td>

Neurotoxicity of clostridium toxins

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-70-3"

class="td11"> 1

1 </td>

Platelet homeostasis

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-71-3"

class="td11"> 2

1 </td>

Eicosanoid ligand-binding receptors

14 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-72-3"

class="td11"> 1

1 </td>

Prolactin receptor signaling

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-73-3"

class="td11"> 1

1 </td>

Acyl chain remodelling of PI

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-74-3"

class="td11"> 1

1 </td>

Signaling by FGFR1 fusion mutants

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-75-3"

class="td11"> 1

1 </td>

PKA activation

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-76-3"

class="td11"> 1

1 </td>

PKA-mediated phosphorylation of CREB

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-77-3"

class="td11"> 1

1 </td>

Synthesis of glycosylphosphatidylinositol (GPI)

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-78-3"

class="td11"> 1

1 </td>

PKA activation in glucagon signalling

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-79-3"

class="td11"> 1

1 </td>

Butyrate Response Factor 1 (BRF1) destabilizes mRNA

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-80-3"

class="td11"> 1

1 </td>

Other semaphorin interactions

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-81-3"

class="td11"> 1

1 </td>

Acyl chain remodelling of PE

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-82-3"

class="td11"> 1

1 </td>

Signaling by Leptin

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-83-3"

class="td11"> 1

1 </td>

DARPP-32 events

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-84-3"

class="td11"> 1

1 </td>

Glucagon-like Peptide-1 (GLP1) regulates insulin secretion

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-85-3"

class="td11"> 1

1 </td>

Uptake and actions of bacterial toxins

22 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-86-3"

class="td11"> 1

1 </td>

Acyl chain remodelling of PC

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-109-87-3"

class="td11"> 1

1 </td>

</td>

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

G.3 Comparison to Primary Screen

The mutation synthetic lethal partners with CDH1 were also compared to siRNA primary screen data (Telford et al., 2015), as performed in Section 4.2.1. These are expected to be more concordant with the experimental results performed on a null mutant, however this not the case at the gene level: less genes overlapped with experimental candidates in Figure G.2. This may be affected by lower sample size for mutations in TCGA data or lower frequency (expected value) of CDH1 mutations compared to low expression.

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

Figure G.2: Comparison of mtSLIPT in stomach to siRNA. Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (mtSLIPT) and experimental screening (siRNA) approaches. The χ2 test suggests that the overlap is no more than would be expected by chance (p = 0.872).

Table G.4: Pathway composition for CDH1 partners from mtSLIPT and siRNA

Predicted only by SLIPT (217 genes)

Pathway Size

Genes Identified

p-value (FDR)

Diseases associated with glycosaminoglycan metabolism

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-2-3"

class="td11"> 5

1.6 × 10-7

Diseases of glycosylation

26 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-3-3"

class="td11"> 5

1.6 × 10-7

Extracellular matrix organization

238 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-4-3"

class="td11"> 18

1.7 × 10-6

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-5-3"

class="td11"> 5

4.6 × 10-6

Initial triggering of complement

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-6-3"

class="td11"> 3

7.3 × 10-5

Regulation of IGF transport and uptake by IGFBPs

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-7-3"

class="td11"> 3

8.9 × 10-5

Defective B4GALT7 causes EDS, progeroid type

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-8-3"

class="td11"> 3

0<span

class="cmmi-12">.00013

Defective B3GAT3 causes JDSSDHD

19 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-9-3"

class="td11"> 3

0<span

class="cmmi-12">.00013

Collagen degradation

57

5

0.00013

ECM proteoglycans

65 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-11-3"

class="td11"> 5

0<span

class="cmmi-12">.00039

A tetrasaccharide linker sequence is required for GAG synthesis

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-12-3"

class="td11"> 3

0<span

class="cmmi-12">.00039

Nitric oxide stimulates guanylate cyclase

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-13-3"

class="td11"> 3

0<span

class="cmmi-12">.00039

RHO GTPases Activate WASPs and WAVEs

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-14-3"

class="td11"> 3

0<span

class="cmmi-12">.00094

Creation of C4 and C2 activators

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-15-3"

class="td11"> 2

0<span

class="cmmi-12">.00098

Non-integrin membrane-ECM interactions

52

4

0.0012

Dermatan sulfate biosynthesis

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-17-3"

class="td11"> 2

0<span

class="cmmi-12">.0013

Degradation of the extracellular matrix

101 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-18-3"

class="td11"> 6

0<span

class="cmmi-12">.0016

Keratan sulfate degradation

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-19-3"

class="td11"> 2

0<span

class="cmmi-12">.0016

Complement cascade

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-20-3"

class="td11"> 3

0<span

class="cmmi-12">.0018

Molecules associated with elastic fibres

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-21-3"

class="td11"> 3

0<span

class="cmmi-12">.002

</td>

Detected only by siRNA screen (2323 genes)

Pathway Size

Genes Identified

p-value (FDR)

Class A/1 (Rhodopsin-like receptors)

282 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-24-3"

class="td11"> 86

6.5 × 10-85

GPCR ligand binding

363 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-25-3"

class="td11"> 97

9.2 × 10-79

Peptide ligand-binding receptors

175 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-26-3"

class="td11"> 52

4.5 × 10-61

Gαi signalling events

184 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-27-3"

class="td11"> 49

1.6 × 10-53

Gαq signalling events

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-28-3"

class="td11"> 43

5.2 × 10-50

Gastrin-CREB signalling pathway via PKC and MAPK

180 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-29-3"

class="td11"> 46

9.4 × 10-50

DAP12 interactions

159 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-30-3"

class="td11"> 35

8.3 × 10-37

Platelet activation, signaling and aggregation

182 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-31-3"

class="td11"> 37

2.3 × 10-35

Hemostasis

438 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-32-3"

class="td11"> 71

3.3 × 10-35

Downstream signal transduction

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-33-3"

class="td11"> 32

7.7 × 10-35

Signaling by PDGF

172 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-34-3"

class="td11"> 35

2.1 × 10-34

DAP12 signaling

149 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-35-3"

class="td11"> 32

2.7 × 10-34

Signaling by ERBB2

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-36-3"

class="td11"> 31

2.5 × 10-33

Signalling by NGF

266 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-37-3"

class="td11"> 44

5.3 × 10-31

Downstream signaling of activated FGFR1

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-38-3"

class="td11"> 28

5.3 × 10-31

Downstream signaling of activated FGFR2

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-39-3"

class="td11"> 28

5.3 × 10-31

Downstream signaling of activated FGFR3

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-40-3"

class="td11"> 28

5.3 × 10-31

Downstream signaling of activated FGFR4

134 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-41-3"

class="td11"> 28

5.3 × 10-31

Signaling by FGFR

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-42-3"

class="td11"> 29

2.0 × 10-30

Signaling by FGFR1

146 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-43-3"

class="td11"> 29

2.0 × 10-30

</td>

Intersection of SLIPT and siRNA screen (23 genes)

Pathway Size

Genes Identified

p-value (FDR)

ADP signalling through P2Y purinoceptor 1

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-46-3"

class="td11"> 1

1 </td>

G-protein beta:gamma signalling

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-47-3"

class="td11"> 1

1 </td>

G-protein activation

12 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-48-3"

class="td11"> 1

1 </td>

Eicosanoid ligand-binding receptors

14 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-49-3"

class="td11"> 1

1 </td>

Platelet homeostasis

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-50-3"

class="td11"> 2

1 </td>

Gαz signalling events

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-51-3"

class="td11"> 1

1 </td>

Signal amplification

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-52-3"

class="td11"> 1

1 </td>

Activation of Kainate Receptors upon glutamate binding

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-53-3"

class="td11"> 1

1 </td>

Thrombin signalling through proteinase activated receptors (PARs)

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-54-3"

class="td11"> 1

1 </td>

Nitric oxide stimulates guanylate cyclase

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-55-3"

class="td11"> 1

1 </td>

Activation of G protein gated Potassium channels

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-56-3"

class="td11"> 1

1 </td>

G protein gated Potassium channels

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-57-3"

class="td11"> 1

1 </td>

Inhibition of voltage gated Ca2+ channels via Gbeta/gamma subunits

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-58-3"

class="td11"> 1

1 </td>

Laminin interactions

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-59-3"

class="td11"> 1

1 </td>

Inwardly rectifying K+ channels

31 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-60-3"

class="td11"> 1

1 </td>

Glucagon signaling in metabolic regulation

33 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-61-3"

class="td11"> 1

1 </td>

Molecules associated with elastic fibres

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-62-3"

class="td11"> 1

1 </td>

Ca2+ pathway

36 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-63-3"

class="td11"> 1

1 </td>

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-64-3"

class="td11"> 1

1 </td>

GABA B receptor activation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-111-65-3"

class="td11"> 1

1 </td>

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

G.3.1 Resampling Analysis

Table G.5: Pathways for CDH1 partners from mtSLIPT in stomach cancer

Reactome Pathway

Over-representation

Permutation

Extracellular matrix organization

9<span

class="cmmi-12">.6 × 10-9

0<span

class="cmmi-12">.057678

Elastic fibre formation

3<span

class="cmmi-12">.7 × 10-8

0<span

class="cmmi-12">.033817

Diseases associated with glycosaminoglycan metabolism

3<span

class="cmmi-12">.7 × 10-8

0<span

class="cmmi-12">.049336

Diseases of glycosylation

3<span

class="cmmi-12">.7 × 10-8

0<span

class="cmmi-12">.049336

Nitric oxide stimulates guanylate cyclase

3<span

class="cmmi-12">.1 × 10-6

0<span

class="cmmi-12">.037904

Initial triggering of complement

3<span

class="cmmi-12">.7 × 10-5

0<span

class="cmmi-12">.020828

Molecules associated with elastic fibres

3<span

class="cmmi-12">.7 × 10-5

0<span

class="cmmi-12">.027865

Regulation of IGF transport and uptake by IGFBPs

3<span

class="cmmi-12">.7 × 10-5

0<span

class="cmmi-12">.069102

Platelet homeostasis

3<span

class="cmmi-12">.7 × 10-5

0<span

class="cmmi-12">.097294

Defective B4GALT7 causes EDS, progeroid type

5<span

class="cmmi-12">.6 × 10-5

0<span

class="cmmi-12">.081505

Defective B3GAT3 causes JDSSDHD

5<span

class="cmmi-12">.6 × 10-5

0<span

class="cmmi-12">.081505

Collagen degradation

5<span

class="cmmi-12">.6 × 10-5

0<span

class="cmmi-12">.1104

Degradation of the extracellular matrix

8 <span

class="cmsy-10x-x-120">× 10-5

0<span

class="cmmi-12">.43477

ECM proteoglycans

0<span

class="cmmi-12">.00017

0<span

class="cmmi-12">.06469

A tetrasaccharide linker sequence is required for GAG synthesis

0<span

class="cmmi-12">.00025

0<span

class="cmmi-12">.10536

RHO GTPases Activate WASPs and WAVEs

0<span

class="cmmi-12">.00059

0<span

class="cmmi-12">.053929

Non-integrin membrane-ECM interactions

0<span

class="cmmi-12">.00065

0<span

class="cmmi-12">.10424

Creation of C4 and C2 activators

0<span

class="cmmi-12">.00079

0<span

class="cmmi-12">.05461

Dermatan sulfate biosynthesis

0<span

class="cmmi-12">.00079

0<span

class="cmmi-12">.21163

Integrin cell surface interactions

0<span

class="cmmi-12">.00098

0<span

class="cmmi-12">.092405

Glucagon signaling in metabolic regulation

0<span

class="cmmi-12">.00098

0<span

class="cmmi-12">.13425

Keratan sulfate degradation

0<span

class="cmmi-12">.00098

0<span

class="cmmi-12">.22137

Complement cascade

0<span

class="cmmi-12">.0011

0<span

class="cmmi-12">.01552

CS/DS degradation

0<span

class="cmmi-12">.0012

0<span

class="cmmi-12">.065012

Eicosanoid ligand-binding receptors

0<span

class="cmmi-12">.0016

0<span

class="cmmi-12">.066128

Nuclear signaling by ERBB4

0<span

class="cmmi-12">.0016

0<span

class="cmmi-12">.15511

Collagen formation

0<span

class="cmmi-12">.0026

0<span

class="cmmi-12">.13447

cGMP effects

0<span

class="cmmi-12">.0041

0<span

class="cmmi-12">.020195

Voltage gated Potassium channels

0<span

class="cmmi-12">.0041

0<span

class="cmmi-12">.068923

Chondroitin sulfate biosynthesis

0<span

class="cmmi-12">.0059

> 1.5862 × 10-5

Chondroitin sulfate/dermatan sulfate metabolism

0<span

class="cmmi-12">.0065

0<span

class="cmmi-12">.087745

Heparan sulfate/heparin (HS-GAG) metabolism

0<span

class="cmmi-12">.0071

0<span

class="cmmi-12">.085622

Synthesis of substrates in N-glycan biosythesis

0<span

class="cmmi-12">.0085

0<span

class="cmmi-12">.09456

Regulation of actin dynamics for phagocytic cup formation

0<span

class="cmmi-12">.0085

0<span

class="cmmi-12">.096227

CDO in myogenesis

0<span

class="cmmi-12">.01

0<span

class="cmmi-12">.32599

Myogenesis

0<span

class="cmmi-12">.01

0<span

class="cmmi-12">.32599

Syndecan interactions

0<span

class="cmmi-12">.012

0<span

class="cmmi-12">.10975

Activation of Matrix Metalloproteinases

0<span

class="cmmi-12">.012

0<span

class="cmmi-12">.33499

Glycosaminoglycan metabolism

0<span

class="cmmi-12">.012

0<span

class="cmmi-12">.29716

Collagen biosynthesis and modifying enzymes

0<span

class="cmmi-12">.013

0<span

class="cmmi-12">.10774

Keratan sulfate biosynthesis

0<span

class="cmmi-12">.016

0<span

class="cmmi-12">.12644

O-linked glycosylation

0<span

class="cmmi-12">.016

0<span

class="cmmi-12">.65101

Laminin interactions

0<span

class="cmmi-12">.021

0<span

class="cmmi-12">.12766

Biosynthesis of the N-glycan precursor (dolichol lipid-linked oligosaccharide)

and transfer to a nascent protein

0<span

class="cmmi-12">.027

0<span

class="cmmi-12">.065782

Sialic acid metabolism

0<span

class="cmmi-12">.027

0<span

class="cmmi-12">.13413

Keratan sulfate/keratin metabolism

0<span

class="cmmi-12">.029

0<span

class="cmmi-12">.15708

Potassium Channels

0<span

class="cmmi-12">.032

0<span

class="cmmi-12">.43477

Fcgamma receptor (FCGR) dependent phagocytosis

0<span

class="cmmi-12">.042

0<span

class="cmmi-12">.15851

Ion transport by P-type ATPases

0<span

class="cmmi-12">.048

0<span

class="cmmi-12">.66686

Retinoid metabolism and transport

0<span

class="cmmi-12">.051

0<span

class="cmmi-12">.058715

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

Over-representation (hypergeometric test) and Permutation p-values adjusted for multiple tests across pathways (FDR). Significant pathways are marked in bold (FDR < 0.05) and italics (FDR < 0.1).

Table G.6: Pathways for CDH1 partners from mtSLIPT in stomach and siRNA screen

Reactome Pathway

Over-representation

Permutation

SLBP independent Processing of Histone Pre-mRNAs

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-2-3"

class="td11"> > 1.2349 × 10-5

Mitochondrial protein import

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-3-3"

class="td11"> > 1.2349 × 10-5

Voltage gated Potassium channels

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-4-3"

class="td11"> > 1.2349 × 10-5

Tandem pore domain potassium channels

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-5-3"

class="td11"> > 1.2349 × 10-5

L13a-mediated translational silencing of Ceruloplasmin expression

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-6-3"

class="td11"> > 1.2349 × 10-5

Eukaryotic Translation Elongation

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-7-3"

class="td11"> > 1.2349 × 10-5

Peptide chain elongation

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-8-3"

class="td11"> > 1.2349 × 10-5

3’ -UTR-mediated translational regulation

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-9-3"

class="td11"> > 1.2349 × 10-5

Activation of Matrix Metalloproteinases

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-10-3"

class="td11"> > 1.2349 × 10-5

HIV Infection

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-11-3"

class="td11"> > 1.2349 × 10-5

Cell Cycle

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-12-3"

class="td11"> > 1.2349 × 10-5

Influenza Infection

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-13-3"

class="td11"> > 1.2349 × 10-5

Influenza Life Cycle

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-14-3"

class="td11"> > 1.2349 × 10-5

Influenza Viral RNA Transcription and Replication

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-15-3"

class="td11"> > 1.2349 × 10-5

Neurotoxicity of clostridium toxins

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-16-3"

class="td11"> > 1.2349 × 10-5

p38MAPK events

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-17-3"

class="td11"> > 1.2349 × 10-5

SCF-beta-TrCP mediated degradation of Emi1

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-18-3"

class="td11"> > 1.2349 × 10-5

SRP-dependent cotranslational protein targeting to membrane

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-19-3"

class="td11"> > 1.2349 × 10-5

Vpu mediated degradation of CD4

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-20-3"

class="td11"> > 1.2349 × 10-5

Serotonin Neurotransmitter Release Cycle

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-21-3"

class="td11"> > 1.2349 × 10-5

Acetylcholine Binding And Downstream Events

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-22-3"

class="td11"> > 1.2349 × 10-5

Viral mRNA Translation

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-23-3"

class="td11"> > 1.2349 × 10-5

Cobalamin (Cbl, vitamin B12) transport and metabolism

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-24-3"

class="td11"> > 1.2349 × 10-5

ERK/MAPK targets

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-25-3"

class="td11"> > 1.2349 × 10-5

Vitamin B5 (pantothenate) metabolism

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-26-3"

class="td11"> > 1.2349 × 10-5

Signaling by BMP

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-27-3"

class="td11"> > 1.2349 × 10-5

Synthesis of Leukotrienes (LT) and Eoxins (EX)

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-28-3"

class="td11"> > 1.2349 × 10-5

Separation of Sister Chromatids

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-29-3"

class="td11"> > 1.2349 × 10-5

Mitotic Metaphase and Anaphase

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-30-3"

class="td11"> > 1.2349 × 10-5

TRP channels

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-31-3"

class="td11"> > 1.2349 × 10-5

Defects in cobalamin (B12) metabolism

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-32-3"

class="td11"> > 1.2349 × 10-5

Regulation by c-FLIP

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-33-3"

class="td11"> > 1.2349 × 10-5

Attenuation phase

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-34-3"

class="td11"> > 1.2349 × 10-5

Autodegradation of the E3 ubiquitin ligase COP1

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-35-3"

class="td11"> > 1.2349 × 10-5

Apoptotic cleavage of cell adhesion proteins

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-36-3"

class="td11"> > 1.2349 × 10-5

Negative regulation of TCF-dependent signaling by WNT ligand antagonists

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-37-3"

class="td11"> > 1.2349 × 10-5

PERK regulates gene expression

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-38-3"

class="td11"> > 1.2349 × 10-5

Regulation of the Fanconi anemia pathway

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-39-3"

class="td11"> > 1.2349 × 10-5

Passive transport by Aquaporins

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-40-3"

class="td11"> > 1.2349 × 10-5

Lysosome Vesicle Biogenesis

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-41-3"

class="td11"> > 1.2349 × 10-5

Zinc transporters

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-42-3"

class="td11"> > 1.2349 × 10-5

Zinc influx into cells by the SLC39 gene family

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-43-3"

class="td11"> > 1.2349 × 10-5

Asparagine N-linked glycosylation

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-44-3"

class="td11"> > 1.2349 × 10-5

AUF1 (hnRNP D0) destabilizes mRNA

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-45-3"

class="td11"> > 1.2349 × 10-5

Asymmetric localization of PCP proteins

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-46-3"

class="td11"> > 1.2349 × 10-5

degradation of DVL

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-47-3"

class="td11"> > 1.2349 × 10-5

CASP8 activity is inhibited

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-48-3"

class="td11"> > 1.2349 × 10-5

Degradation of GLI1 by the proteasome

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-49-3"

class="td11"> > 1.2349 × 10-5

BBSome-mediated cargo-targeting to cilium

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-50-3"

class="td11"> > 1.2349 × 10-5

Regulation of necroptotic cell death

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-114-51-3"

class="td11"> > 1.2349 × 10-5

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

G.4 Metagene Analysis

Metagene analysis was also performed for synthetic lethal candidates for CDH1 mutation in stomach cancer. These are described and compared to expression analysis in Section F.3.

Table G.7: Candidate synthetic lethal metagenes against CDH1 from mtSLIPT in stomach cancer

Pathway

ID

Observed

Expected

χ2value

p-value

p-value (FDR)

Prostacyclin signalling through prostacyclin receptor

392851

1

10.1

26.5

1.73 × 10-6

0<span

class="cmmi-12">.00286

Cell surface interactions at the vascular wall

202733

3

10.1

21.1

2.61 × 10-5

0<span

class="cmmi-12">.00642

The NLRP1 inflammasome

844455

3

10.1

21.1

2.61 × 10-5

0<span

class="cmmi-12">.00642

Innate Immune System

168249

6

10.1

21.1

2.65 × 10-5

0<span

class="cmmi-12">.00642

Keratan sulfate/keratin metabolism

1638074

4

10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.00642

Keratan sulfate biosynthesis

2022854

4

10.1

20.6

3.29 × 10-5

0<span

class="cmmi-12">.00642

Signaling by SCF-KIT

1433557

5

10.1

20.6

3.30 × 10-5

0<span

class="cmmi-12">.00642

VEGFA-VEGFR2 Pathway

4420097

5

10.1

20.6

3.30 × 10-5

0<span

class="cmmi-12">.00642

p130Cas linkage to MAPK signaling for integrins

372708

2

10.1

19.1

7.19 × 10-5

0<span

class="cmmi-12">.00651

cGMP effects

418457

8

10.1

19

7.46 × 10-5

0<span

class="cmmi-12">.00651

Regulation of cytoskeletal remodeling and cell spreading by IPP

complex components

446388

8

10.1

19

7.46 × 10-5

0<span

class="cmmi-12">.00651

Fcgamma receptor (FCGR) dependent phagocytosis

2029480

3

10.1

17.9

0.000127

0<span

class="cmmi-12">.00651

A third proteolytic cleavage releases NICD

157212

7

10.1

17.9

0.00013

0<span

class="cmmi-12">.00651

Signalling by NGF

166520

7

10.1

17.9

0.00013

0<span

class="cmmi-12">.00651

Signaling by VEGF

194138

7

10.1

17.9

0.00013

0<span

class="cmmi-12">.00651

Regulation of thyroid hormone activity

350864

7

10.1

17.9

0.00013

0<span

class="cmmi-12">.00651

Nitric oxide stimulates guanylate cyclase

392154

7

10.1

17.9

0.00013

0<span

class="cmmi-12">.00651

Platelet homeostasis

418346

7

10.1

17.9

0.00013

0<span

class="cmmi-12">.00651

PI3K events in ERBB4 signaling

1250342

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

PIP3 activates AKT signaling

1257604

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

GAB1 signalosome

180292

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

PI3K events in ERBB2 signaling

1963642

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

PI3K/AKT Signaling in Cancer

2219528

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

Rap1 signalling

392517

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

Lysosphingolipid and LPA receptors

419408

4

10.1

17.3

0.000179

0<span

class="cmmi-12">.00651

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

Strongest candidate SL partners for CDH1 by mtSLIPT with observed and expected mutant samples with low expression of partner metagenes

Appendix HGlobal Synthetic Lethality in Stomach Cancer

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

Figure H.1: Synthetic lethal partners across query genes. Global synthetic lethal pairs were examined across the genome in TCGA stomach expression data by applying SLIPT across query genes. The high number of predicted partners for CDH1 was typical for a human gene and lower than many other genes.

H.1 Hub Genes

Table H.1: Query synthetic lethal genes with the most SLIPT partners

Gene

Direction

raw p-value

p-value (FDR)

SLIPT raw p-value

SLIPT (FDR)

HEG1

10719

16956

16724 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-2-5"

class="td11"> 9616

9532 </td>

SYNE1

10755

17210

16984 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-3-5"

class="td11"> 9749

9676 </td>

A2M

10743

16650

16378 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-4-5"

class="td11"> 9529

9433 </td>

ANK2

11008

16616

16355 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-5-5"

class="td11"> 9764

9653 </td>

TTC28

10757

16523

16248 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-6-5"

class="td11"> 9530

9429 </td>

FAT4

10451

16286

15978 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-7-5"

class="td11"> 9225

9115 </td>

MRVI1

10904

16967

16718 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-8-5"

class="td11"> 9775

9686 </td>

PAPLN

10483

16405

16104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-9-5"

class="td11"> 9305

9193 </td>

NFASC

10773

16575

16307 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-10-5"

class="td11"> 9578

9475 </td>

MACF1

9697

16378

16058 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-11-5"

class="td11"> 8620

8540 </td>

HMCN1

10475

16101

15733 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-12-5"

class="td11"> 9156

9008 </td>

MPDZ

10878

16550

16299 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-13-5"

class="td11"> 9599

9491 </td>

FLRT2

10776

16760

16473 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-14-5"

class="td11"> 9590

9464 </td>

SETBP1

10869

16632

16349 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-15-5"

class="td11"> 9615

9489 </td>

LAMA4

10463

16447

16121 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-16-5"

class="td11"> 9273

9151 </td>

IL1R1

10611

16185

15803 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-17-5"

class="td11"> 9299

9174 </td>

ABCA6

10499

16573

16318 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-18-5"

class="td11"> 9260

9158 </td>

LAMC1

10238

15777

15392 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-19-5"

class="td11"> 8837

8691 </td>

TNS1

10920

17038

16806 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-20-5"

class="td11"> 9836

9751 </td>

AMOTL1

10612

16458

16178 </td><td style="white-space:nowrap; text-align:center;" id="TBL-117-21-5"

class="td11"> 9367

9250 </td>

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

Genes with the most candidate SL partners SLIPT in TCGA stomach expression data with the number of partner genes predicted by direction criteria and χ2 testing separately and combined as a SLIPT analysis. Where specified, the p-values for the χ2 test were adjusted for multiple tests (FDR).

H.2 Hub Pathways

Table H.2: Pathways for genes with the most SLIPT partners

Pathways Over-represented

Pathway Size

SL Genes

p-value

p-value (FDR)

Molecules associated with elastic fibres

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-2-3"

class="td11"> 10

4.6 × 10-21

2.7 × 10-18

Extracellular matrix organization

241 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-3-3"

class="td11"> 29

5.3 × 10-21

2.7 × 10-18

Smooth Muscle Contraction

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-4-3"

class="td11"> 9

5.6 × 10-20

1.6 × 10-17

Elastic fibre formation

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-5-3"

class="td11"> 10

6 × 10-20

1.6 × 10-17

Nitric oxide stimulates guanylate cyclase

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-6-3"

class="td11"> 8

6.9 × 10-19

1.4 × 10-16

Muscle contraction

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-7-3"

class="td11"> 12

8.3 × 10-19

1.4 × 10-16

Platelet homeostasis

54 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-8-3"

class="td11"> 11

1.3 × 10-18

1.9 × 10-16

cGMP effects

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-9-3"

class="td11"> 6

3.3 × 10-15

4.3 × 10-13

Laminin interactions

30 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-10-3"

class="td11"> 7

1.3 × 10-14

1.6 × 10-12

Axon guidance

289 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-11-3"

class="td11"> 25

5 × 10-13

5.2 × 10-11

Signaling by BMP

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-12-3"

class="td11"> 5

3.7 × 10-11

3<span

class="cmmi-12">.2 × 10-9

RHO GTPases activate PAKs

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-13-3"

class="td11"> 5

3.7 × 10-11

3<span

class="cmmi-12">.2 × 10-9

Non-integrin membrane-ECM interactions

53 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-14-3"

class="td11"> 7

7.2 × 10-11

5<span

class="cmmi-12">.8 × 10-9

Rho GTPase cycle

120 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-15-3"

class="td11"> 11

1.2 × 10-10

8<span

class="cmmi-12">.7 × 10-9

Degradation of the extracellular matrix

104 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-16-3"

class="td11"> 10

1.3 × 10-10

8<span

class="cmmi-12">.8 × 10-9

Netrin-1 signaling

42 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-17-3"

class="td11"> 6

2.5 × 10-10

1<span

class="cmmi-12">.6 × 10-8

Developmental Biology

432 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-18-3"

class="td11"> 32

8.3 × 10-10

5 <span

class="cmsy-10x-x-120">× 10-8

L1CAM interactions

80 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-19-3"

class="td11"> 8

8.7 × 10-10

5 <span

class="cmsy-10x-x-120">× 10-8

Semaphorin interactions

64 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-20-3"

class="td11"> 7

1.1 × 10-9

6<span

class="cmmi-12">.1 × 10-8

Cell-extracellular matrix interactions

18 </td><td style="white-space:nowrap; text-align:center;" id="TBL-118-21-3"

class="td11"> 4

1.3 × 10-9

6<span

class="cmmi-12">.6 × 10-8

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in the top 500 “hub” genes with the most candidate synthetic lethal partners by SLIPT analysis of TCGA stomach expression data

Appendix IReplication in cell line encyclopaedia

Table I.1: Candidate synthetic lethal gene partners of CDH1 from SLIPT in CCLE

Gene

Observed

Expected

χ2 value

p-value

p-value (FDR)

ZEB1

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-2-3"

class="td11"> 115

555

7.84 × 10-119

3.62 × 10-116

RP11-620J15.3

17 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-3-3"

class="td11"> 115

471

1.54 × 10-100

3.68 × 10-98

AP1S2

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-4-3"

class="td11"> 115

462

1.38 × 10-98

3.07 × 10-96

VIM

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-5-3"

class="td11"> 115

424

1.73 × 10-90

3.06 × 10-88

CCDC88A

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-6-3"

class="td11"> 115

418

3.94 × 10-89

6.86 × 10-87

RECK

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-7-3"

class="td11"> 115

416

8.23 × 10-89

1.42 × 10-86

AP1M1

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-8-3"

class="td11"> 115

414

2.42 × 10-88

4.06 × 10-86

ZEB2

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-9-3"

class="td11"> 115

396

2.32 × 10-84

3.4 × 10-82

WIPF1

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-10-3"

class="td11"> 115

390

4.9 × 10-83

6.74 × 10-81

SLC35B4

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-11-3"

class="td11"> 115

386

3.2 × 10-82

4.38 × 10-80

SACS

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-12-3"

class="td11"> 115

373

2.13 × 10-79

2.7 × 10-77

ST3GAL2

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-13-3"

class="td11"> 115

351

9.7 × 10-75

1.08 × 10-72

ATP8B2

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-14-3"

class="td11"> 115

341

1.53 × 10-72

1.61 × 10-70

IFFO1

39 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-15-3"

class="td11"> 115

332

1.66 × 10-70

1.65 × 10-68

EMP3

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-16-3"

class="td11"> 115

329

5.04 × 10-70

4.95 × 10-68

LEPRE1

40 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-17-3"

class="td11"> 115

325

5.4 × 10-69

5.22 × 10-67

STARD9

39 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-18-3"

class="td11"> 115

311

4.52 × 10-66

3.96 × 10-64

DENND5A

48 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-19-3"

class="td11"> 115

304

1.89 × 10-64

1.59 × 10-62

SYT11

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-20-3"

class="td11"> 115

300

1.21 × 10-63

9.89 × 10-62

EID2B

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-21-3"

class="td11"> 115

299

1.99 × 10-63

1.61 × 10-61

NXPE3

35 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-22-3"

class="td11"> 115

294

1.71 × 10-62

1.35 × 10-60

STX2

49 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-23-3"

class="td11"> 115

293

3.83 × 10-62

3 × 10-60

ARHGEF6

43 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-24-3"

class="td11"> 115

289

2.2 × 10-61

1.71 × 10-59

KATNAL1

50 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-25-3"

class="td11"> 115

283

4.45 × 10-60

3.38 × 10-58

ANXA6

37 </td><td style="white-space:nowrap; text-align:center;" id="TBL-119-26-3"

class="td11"> 115

282

8.92 × 10-60

6.67 × 10-58

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

Strongest candidate SL partners for CDH1 by SLIPT with observed and expected samples with low expression of both genes

Table I.2: Candidate synthetic lethal gene partners of CDH1 from SLIPT in breast CCLE

Gene

Observed

Expected

χ2 value

p-value

p-value (FDR)

MIR155HG

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-2-3"

class="td11"> 6.78

31.5

2.41 × 10-6

0.00371 </td>

ENPP2

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-3-3"

class="td11"> 6.78

30.7

3.47 × 10-6

0.00383 </td>

DCLK2

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-4-3"

class="td11"> 6.78

28.3

1.08 × 10-5

0.0071 </td>

PID1

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-5-3"

class="td11"> 6.78

27.8

1.34 × 10-5

0.00791 </td>

SCFD2

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-6-3"

class="td11"> 6.78

27.7

1.42 × 10-5

0.00791 </td>

FAT4

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-7-3"

class="td11"> 6.78

27.3

1.69 × 10-5

0.00865 </td>

ILK

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-8-3"

class="td11"> 6.78

26.9

2.04 × 10-5

0.00884 </td>

RWDD1

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-9-3"

class="td11"> 6.78

26.8

2.15 × 10-5

0.00884 </td>

RIC8A

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-10-3"

class="td11"> 6.78

26.8

2.2 × 10-5

0.00884 </td>

F2RL2

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-11-3"

class="td11"> 6.78

26.6

2.34 × 10-5

0.00901 </td>

SDCBP

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-12-3"

class="td11"> 6.78

25.9

3.26 × 10-5

0.0108 </td>

PPM1F

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-13-3"

class="td11"> 6.78

25.8

3.41 × 10-5

0.0108 </td>

IKBIP

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-14-3"

class="td11"> 6.78

25.8

3.49 × 10-5

0.0108 </td>

SPRED1

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-15-3"

class="td11"> 6.78

25.5

3.97 × 10-5

0.0108 </td>

RNH1

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-16-3"

class="td11"> 6.78

25.4

4.22 × 10-5

0.0108 </td>

SYDE1

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-17-3"

class="td11"> 6.78

25.4

4.22 × 10-5

0.0108 </td>

LINC00968

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-18-3"

class="td11"> 6.78

25.2

4.63 × 10-5

0.0109 </td>

ARHGEF10

5 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-19-3"

class="td11"> 6.78

24.5

6.22 × 10-5

0.0116 </td>

P4HA1

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-20-3"

class="td11"> 6.78

24.5

6.34 × 10-5

0.0116 </td>

AZI2

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-21-3"

class="td11"> 6.78

24.5

6.34 × 10-5

0.0116 </td>

TNFAIP6

2 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-22-3"

class="td11"> 6.78

24.5

6.34 × 10-5

0.0116 </td>

CD200

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-23-3"

class="td11"> 6.78

24.5

6.37 × 10-5

0.0116 </td>

SMPD1

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-24-3"

class="td11"> 6.78

24.4

6.67 × 10-5

0.0116 </td>

ATP6V1G2

3 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-25-3"

class="td11"> 6.78

24.2

7.33 × 10-5

0.0123 </td>

FGF2

4 </td><td style="white-space:nowrap; text-align:center;" id="TBL-120-26-3"

class="td11"> 6.78

24.1

7.49 × 10-5

0.0123 </td>

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

Strongest candidate SL partners for CDH1 by SLIPT with observed and expected samples with low expression of both genes

Table I.3: Candidate synthetic lethal gene partners of CDH1 from SLIPT in stomach CCLE

Gene

Observed

Expected

χ2 value

p - value

p-value (FDR)

ZEB1

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-2-3"

class="td11"> 4.45

36

2.84 × 10-7

0.00175 </td>

WDR47

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-3-3"

class="td11"> 4.45

26.7

2.3 × 10-5

0.013 </td>

KANK2

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-4-3"

class="td11"> 4.45

25.1

4.81 × 10-5

0.0222 </td>

LEPRE1

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-5-3"

class="td11"> 4.45

24.5

6.26 × 10-5

0.0228 </td>

KATNAL1

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-6-3"

class="td11"> 4.45

24.3

6.88 × 10-5

0.0231 </td>

TET1

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-7-3"

class="td11"> 4.45

23.9

8.23 × 10-5

0.0249 </td>

AP1S2

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-8-3"

class="td11"> 4.45

23.1

0.00012

0.0273 </td>

CDKN2C

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-9-3"

class="td11"> 4.45

22.8

0.000136

0.0292 </td>

ARMC4

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-10-3"

class="td11"> 4.45

22.4

0.000164

0.0315 </td>

CSTF3

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-11-3"

class="td11"> 4.45

22.4

0.000166

0.0315 </td>

FAM216A

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-12-3"

class="td11"> 4.45

22.4

0.000166

0.0315 </td>

ANKRD32

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-13-3"

class="td11"> 4.45

22.4

0.000166

0.0315 </td>

WDR35

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-14-3"

class="td11"> 4.45

22.4

0.000169

0.0315 </td>

ECI2

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-15-3"

class="td11"> 4.45

21.7

0.000232

0.0378 </td>

SAMD8

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-16-3"

class="td11"> 4.45

21.7

0.000232

0.0378 </td>

CHST12

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-17-3"

class="td11"> 4.45

21.7

0.000232

0.0378 </td>

RPL23AP32

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-18-3"

class="td11"> 4.45

21.7

0.000232

0.0378 </td>

STARD9

1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-19-3"

class="td11"> 4.45

21.7

0.000232

0.0378 </td>

MCM8

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-121-20-3"

class="td11"> 4.45

21.5

0.000255

0.0379 </td>

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

Strongest candidate SL partners for CDH1 by SLIPT with observed and expected samples with low expression of both genes

Table I.4: Pathways for CDH1 partners from SLIPT in stomach CCLE

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Nef mediated downregulation of MHC class I complex cell surface expression

10 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-2-3"

class="td11"> 1

1 </td>

Unwinding of DNA

11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-3-3"

class="td11"> 1

1 </td>

Processing of Intronless Pre-mRNAs

13 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-4-3"

class="td11"> 1

1 </td>

E2F mediated regulation of DNA replication

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-5-3"

class="td11"> 1

1 </td>

Chondroitin sulfate biosynthesis

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-6-3"

class="td11"> 1

1 </td>

Post-Elongation Processing of Intronless pre-mRNA

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-7-3"

class="td11"> 1

1 </td>

Nef-mediates down modulation of cell surface receptors by recruiting them

to clathrin adapters

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-8-3"

class="td11"> 1

1 </td>

Processing of Capped Intronless Pre-mRNA

21 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-9-3"

class="td11"> 1

1 </td>

Post-Elongation Processing of Intron-Containing pre-mRNA

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-10-3"

class="td11"> 1

1 </td>

Activation of the pre-replicative complex

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-11-3"

class="td11"> 1

1 </td>

mRNA 3’-end processing

23 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-12-3"

class="td11"> 1

1 </td>

Golgi Associated Vesicle Biogenesis

24 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-13-3"

class="td11"> 1

1 </td>

Lysosome Vesicle Biogenesis

25 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-14-3"

class="td11"> 1

1 </td>

Oncogene Induced Senescence

27 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-15-3"

class="td11"> 1

1 </td>

The role of Nef in HIV-1 replication and disease pathogenesis

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-16-3"

class="td11"> 1

1 </td>

Cyclin D associated events in G1

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-17-3"

class="td11"> 1

1 </td>

G1 Phase

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-18-3"

class="td11"> 1

1 </td>

Cleavage of Growing Transcript in the Termination Region

31 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-19-3"

class="td11"> 1

1 </td>

Activation of ATR in response to replication stress

31 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-20-3"

class="td11"> 1

1 </td>

DNA strand elongation

31 </td><td style="white-space:nowrap; text-align:center;" id="TBL-122-21-3"

class="td11"> 1

1 </td>

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for CDH1

Table I.5: Pathways for CDH1 partners from SLIPT in breast and stomach CCLE

Pathways Over-represented

Pathway Size

SL Genes

p-value (FDR)

Collagen formation

66 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-2-3"

class="td11"> 8

1<span

class="cmmi-12">.1 × 10-7

Glycosaminoglycan metabolism

111 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-3-3"

class="td11"> 11

1<span

class="cmmi-12">.1 × 10-7

Extracellular matrix organization

236 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-4-3"

class="td11"> 20

1<span

class="cmmi-12">.1 × 10-7

Collagen biosynthesis and modifying enzymes

55 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-5-3"

class="td11"> 7

1<span

class="cmmi-12">.7 × 10-7

Keratan sulfate biosynthesis

28 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-6-3"

class="td11"> 5

2<span

class="cmmi-12">.2 × 10-7

Keratan sulfate/keratin metabolism

32 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-7-3"

class="td11"> 5

7<span

class="cmmi-12">.5 × 10-7

ECM proteoglycans

65 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-8-3"

class="td11"> 7

1<span

class="cmmi-12">.1 × 10-6

Non-integrin membrane-ECM interactions

52 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-9-3"

class="td11"> 6

2<span

class="cmmi-12">.0 × 10-6

Cell junction organization

71 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-10-3"

class="td11"> 7

3<span

class="cmmi-12">.0 × 10-6

Assembly of collagen fibrils and other multimeric structures

39 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-11-3"

class="td11"> 5

3<span

class="cmmi-12">.6 × 10-6

Post-chaperonin tubulin folding pathway

14 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-12-3"

class="td11"> 3

1<span

class="cmmi-12">.7 × 10-5

Adherens junctions interactions

29 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-13-3"

class="td11"> 4

1<span

class="cmmi-12">.7 × 10-5

Cell-Cell communication

118 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-14-3"

class="td11"> 9

1<span

class="cmmi-12">.7 × 10-5

Sialic acid metabolism

31 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-15-3"

class="td11"> 4

2<span

class="cmmi-12">.5 × 10-5

Synthesis and interconversion of nucleotide di- and triphosphates

16 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-16-3"

class="td11"> 3

3<span

class="cmmi-12">.1 × 10-5

Transport to the Golgi and subsequent modification

34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-17-3"

class="td11"> 4

4<span

class="cmmi-12">.8 × 10-5

Asparagine N-linked glycosylation

113 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-18-3"

class="td11"> 8

7<span

class="cmmi-12">.8 × 10-5

Elastic fibre formation

37 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-19-3"

class="td11"> 4

8<span

class="cmmi-12">.5 × 10-5

L1CAM interactions

77 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-20-3"

class="td11"> 6

9<span

class="cmmi-12">.5 × 10-5

Signal transduction by L1

20 </td><td style="white-space:nowrap; text-align:center;" id="TBL-124-21-3"

class="td11"> 3

9<span

class="cmmi-12">.5 × 10-5

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

Gene set over-representation analysis (hypergeometric test) for Reactome pathways in SLIPT partners for CDH1

Appendix JSynthetic Lethal Genes in Pathways

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

Figure J.1: Synthetic Lethality in the PI3K/AKT Pathway. The Reactome PI3K/AKT pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.2: Synthetic Lethality in the PI3K/AKT Pathway in Cancer. The Reactome PI3K/AKT Pathway in Cancer pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.3: Synthetic Lethality in the Extracellular Matrix. The Reactome Extracellular Matrix pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.4: Synthetic Lethality in the GPCRs. The Reactome Gαi pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.5: Synthetic Lethality in the GPCR Downstream. The Reactome GPCR Downstream pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.6: Synthetic Lethality in the Translation Elongation. The Reactome Translation Elongation pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.7: Synthetic Lethality in the Nonsense-mediated Decay. The Reactome Nonsense-mediated Decay pathway with synthetic lethal candidates coloured as shown in the Legend.

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

Figure J.8: Synthetic Lethality in the 3′ UTR. The Reactome 3′ UTR pathway with synthetic lethal candidates coloured as shown in the Legend.

Appendix KPathway Connectivity for Mutation SLIPT

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

Figure K.1: Synthetic Lethality and Vertex Degree. The number of connected genes (vertex degree) was compared (on a log-scale across genes deteced by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. There were very few differences in vertex degree between the groups, although genes detected by siRNA included those with the fewest connections.

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

Figure K.2: Synthetic Lethality and Centrality. The information centrality was compared (on a log-scale across genes deteced by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by siRNA had higher connectivity than many genes not detected by either approach. The gene with the highest centrality was detected by mtSLIPT.

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

Figure K.3: Synthetic Lethality and PageRank. The PageRank centrality was compared (on a log-scale across genes deteced by mtSLIPT and siRNA screening in the Reactome PI3K cascade pathway. Genes detected by siRNA had a more restricted range of centrality values than other genes not detected by either approach, although these groups also had fewer genes.

Table K.1: ANOVA for Synthetic Lethality and Vertex Degree

</td><td style="white-space:nowrap; text-align:center;" id="TBL-125-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

15 </td><td style="white-space:nowrap; text-align:center;" id="TBL-125-2-4"

class="td11"> 15.50

0.0134

0.9084

mtSLIPT

1

196 </td><td style="white-space:nowrap; text-align:center;" id="TBL-125-3-4"

class="td11"> 195.94

0.1689

0.6825

siRNA×mtSLIPT

1

9 </td><td style="white-space:nowrap; text-align:center;" id="TBL-125-4-4"

class="td11"> 9.17

0.0079

0.9294

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

Analysis of variance for vertex degree against synthetic lethal detection approaches (with an interaction term)

Table K.2: ANOVA for Synthetic Lethality and Information Centrality

</td><td style="white-space:nowrap; text-align:center;" id="TBL-126-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

0.000256

0.0002561

0.1851

0.6685

mtSLIPT

1

0.003225

0.0032247

2.3308

0.1318

siRNA×mtSLIPT

1

0.001238

0.0012385

0.8952

0.3476

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

Analysis of variance for information centrality against synthetic lethal detection approaches (with an interaction term)

Table K.3: ANOVA for Synthetic Lethality and PageRank Centrality

</td><td style="white-space:nowrap; text-align:center;" id="TBL-127-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

0.0002038

2.0385 × 10-4

1.1423

0.2892

mtSLIPT

1

0.0000208

2.0752 × 10-5

0.1163

0.7342

siRNA×mtSLIPT

1

0.0000137

1.3743 × 10-5

0.0770

0.7823

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

Analysis of variance for PageRank centrality against synthetic lethal detection approaches (with an interaction term)

Appendix LInformation Centrality for 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 via Pathway Commons with Paxtools (Cerami et al., 2011; Demir et al., 2013). 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. In support of centrality as a measure of essentiality, a number nodes with the highest centrality (shown in Table L.1) were essential nutrients including Mg2+, Ca2+, Zn2+, and Fe. In addition, there were genes important in development of epithelial tissues and breast cancer such as IL8, GATA3, and CTNNB1 detected with relatively high information centrality.

Table L.1: Information centrality for genes and molecules in the Reactome network

Node

Centrality

ZNF473

0.0510

magnesium(2+)

0.0082

XBP1

0.0053

calcium(2+)

0.0050

zinc(2+)

0.0048

iron atom

0.0041

FMN

0.0040

AGT

0.0037

HSP90AA1

0.0029

phosphatidyl-L-serine

0.0029

P2RX7

0.0026

PANX1

0.0024

NCAM1

0.0022

NUDT1

0.0021

PLAUR

0.0020

IL8

0.0020

HSPA8

0.0019

TYROBP

0.0019

CASP3

0.0017

GNAL

0.0015

CBLB

0.0015

HBB

0.0014

GATA4

0.0013

TGS1

0.0013

CTNNB1

0.0012

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

Highest information centrality for genes (proteins), cofactors, and minerals in the Reactome network

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

Figure L.1: Information centrality distribution. Information centrality in the Reactome network for nodes, including genes/proteins and other biomolecules.

Appendix MPathway Structure for Mutation SLIPT

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

Figure M.1: Synthetic Lethality and Heirarchy Score in PI3K. The hierarchical distance scores were similarly distributed across mtSLIPT and siRNA genes. Genes detected by both methods had a higher (downstream) median than either group.

Table M.1: ANOVA for Synthetic Lethality and PI3K Hierarchy

</td><td style="white-space:nowrap; text-align:center;" id="TBL-129-1-2"

class="td11"> DF

Sum Squares

Mean Squares

F-value

p-value

siRNA

1

0.001 </td><td style="white-space:nowrap; text-align:center;" id="TBL-129-2-4"

class="td11"> 0.00070

0.0004

0.9841

mtSLIPT

1

0.007 </td><td style="white-space:nowrap; text-align:center;" id="TBL-129-3-4"

class="td11"> 0.0066

0.0040

0.9496

siRNA×mtSLIPT

1

3.906 </td><td style="white-space:nowrap; text-align:center;" id="TBL-129-4-4"

class="td11"> 3.9056

2.3829

0.1250

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

Analysis of variance for PI3K hierarchy score against synthetic lethal detection approaches (with an interaction term)

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

Figure M.2: Heirarchy Score in PI3K against Synthetic Lethality in PI3K. The number of mtSLIPT and siRNA genes against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

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

Figure M.3: Structure of Synthetic Lethality in PI3K. The number of mtSLIPT and siRNA genes upstream or downstream of each gene in the Reactome PI3K pathway were tested (by the χ2-test). These are plotted as a split violin plot against the hierarchical distance scores showing no significant tendency for either method to either of the pathway upstream or downstream extremities.

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

Figure M.4: Structure of Synthetic Lethality Resampling. A null distribution (10,000 iterations) of the siRNA genes upstream or downstream of mtSLIPT genes (shown by the difference) in the PI3K pathway. The observed events (red) were compared to the the distribution (violet) and were not significant. Genes detected by both methods were fixed for the distribution (blue). The genes detected by both approaches were used.

Table M.2: Resampling for pathway structure of synthetic lethal detection methods

<table id="TBL-130" class="tabular"

cellspacing="0" cellpadding="0" rules="groups" >

</td><td colspan="2" style="white-space:nowrap; text-align:center;" id="TBL-130-2-2"

class="td11">

Graph

</td><td colspan="2" style="white-space:nowrap; text-align:center;" id="TBL-130-2-4"

class="td11">

States

</td><td colspan="4" style="white-space:nowrap; text-align:center;" id="TBL-130-2-6"

class="td11">

Observed

</td><td colspan="2" style="white-space:nowrap; text-align:center;" id="TBL-130-2-10"

class="td11">

Permutation p-value

Pathway

Nodes

Edges

mtSL

siRNA

Up

Down

Up-Down

Up∕Down

Up-Down

Down-Up

PI3K Cascade

138

1495

42

25

131

123

8 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-4-9"

class="td11"> 1.065

0.4473

0.5466

PI3K/AKT Signaling in Cancer

275

12882

56

44

478

440

38 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-5-9"

class="td11"> 1.086

0.4163

0.5810

Gαi Signaling

292

22003

57

58

543

866

-323 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-6-9"

class="td11"> 0.627

0.9507

0.0488

GPCR downstream

1270

142071

218

160

7632

6500

1132

1.174

0.1707

0.8291

Elastic fibre formation

42

175

16

7

6

7

-1 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-8-9"

class="td11"> 0.857

0.5512

0.3681

Extracellular matrix

299

3677

81

29

313

347

-34 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-9-9"

class="td11"> 0.902

0.5762

0.4215

Formation of Fibrin

52

243

11

5

8

19

-11 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-10-9"

class="td11"> 0.421

0.7993

0.1800

Nonsense-Mediated Decay

103

102

56

2

0

0

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-11-9"

class="td11">

0.197

0.1373

3′ -UTR-mediated translational regulation

107

2860

56

1

52

1

51 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-12-9"

class="td11"> 52

0.1210

0.8751

Eukaryotic Translation Elongation

92

3746

57

0

0

0

0 </td><td style="white-space:nowrap; text-align:center;" id="TBL-130-13-9"

class="td11">

0.4952

0.4892

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

Pathways in the Reactome network tested for structural relationships between mtSLIPT and siRNA genes by resampling. The raw p-value (computed without adjusting for multiple comparisons over pathways) is given for the difference in upstream and downstream paths from mtSLIPT to siRNA gene candidate partners of CDH1 with significant pathways highlighted in bold. Sampling was performed only in the target pathway and shortest paths were computed within it. Loops or paths in either direction that could not be resolved were excluded from the analysis. The gene detected by both mtSLIPT and siRNA (or resampling for them) were includued in the analysis and the number of these were fixed to the number observed.

Appendix NSimulation