Contents

List of Figures

List of Tables

Chapter 4

Synthetic Lethal Analysis of Gene Expression Data

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

4.1 Synthetic lethal genes in breast cancer

- exprSL
- mtSL
- heatmap

Table 4.1: Candidate synthetic lethal genes against E-cadherin from SLIPT

Gene	**Observed Low-Low**	**Expected Low-Low**	χ^2 **value**	**raw p-val**	**p-val (FDR)**
TRIP10	62	130	162	5.65×10^{-34}	1.84×10^{-31}
EEF1B2	56	130	158	3.10×10^{-33}	9.45×10^{-31}
GBGT1	61	131	156	1.08×10^{-32}	3.14×10^{-30}
ELN	81	130	149	3.46×10^{-31}	8.82×10^{-29}
TSPAN4	78	130	146	1.63×10^{-30}	3.79×10^{-28}
GLIPR2	72	130	146	1.68×10^{-30}	3.86×10^{-28}
RPS20	73	131	145	1.89×10^{-30}	4.28×10^{-28}
RPS27A	80	130	143	5.53×10^{-30}	1.18×10^{-27}
EEF1A1P9	63	130	141	1.91×10^{-29}	3.74×10^{-27}
C1R	73	130	141	2.05×10^{-29}	3.97×10^{-27}
LYL1	73	130	140	2.99×10^{-29}	5.74×10^{-27}
RPLP2	71	130	139	4.88×10^{-29}	9.07×10^{-27}
C10orf10	73	130	138	6.72×10^{-29}	1.20×10^{-26}
DULLARD	74	131	138	9.29×10^{-29}	1.61×10^{-26}
PPM1F	64	130	136	1.61×10^{-28}	2.65×10^{-26}
OBFC2A	69	130	136	2.49×10^{-28}	3.93×10^{-26}
RPL11	70	130	136	2.56×10^{-28}	3.97×10^{-26}
RPL18A	70	130	135	3.08×10^{-28}	4.70×10^{-26}
MFNG	76	131	133	7.73×10^{-28}	1.12×10^{-25}
RPS17	77	131	133	8.94×10^{-28}	1.29×10^{-25}
MGAT1	73	130	132	1.44×10^{-27}	2.03×10^{-25}
RPS12	72	130	128	8.57×10^{-27}	1.12×10^{-24}
C10orf54	73	130	127	1.37×10^{-26}	1.75×10^{-24}
LOC286367	72	130	126	2.20×10^{-26}	2.70×10^{-24}
GMFG	70	130	126	2.20×10^{-26}	2.70×10^{-24}

Table 4.2: Candidate synthetic lethal genes against E-cadherin from mtSLIPT

Gene	Observed Low-Low	Expected Low-Low	χ^2 value	raw p-val	p-val (FDR)
TFAP2B	8	36.7	89.5	3.6×10^{-20}	8.37×10^{-17}
ZNF423	15	36.7	78.8	7.89×10^{-18}	1.22×10^{-14}
CALCOCO1	11	36.7	76.8	2.09×10^{-17}	2.59×10^{-14}
RBM5	13	36.7	75.7	3.65×10^{-17}	4.00×10^{-14}
BTG2	7	36.7	71.7	2.72×10^{-16}	1.81×10^{-13}
RXRA	6	36.7	70.5	5.00×10^{-16}	2.97×10^{-13}
SLC27A1	11	36.7	70.3	5.42×10^{-16}	2.97×10^{-13}
MEF2D	12	36.7	69.6	7.86×10^{-16}	3.95×10^{-13}
NISCH	12	36.7	69.6	7.86×10^{-16}	3.95×10^{-13}
AVPR2	9	36.7	69.2	9.36×10^{-16}	4.58×10^{-13}
CRY2	13	36.7	68.9	1.07×10^{-15}	4.98×10^{-13}
RAPGEF3	13	36.7	68.9	1.07×10^{-15}	4.98×10^{-13}
NRIP2	10	36.7	68.2	1.58×10^{-15}	7.18×10^{-13}
DARC	12	36.7	66.4	3.76×10^{-15}	1.54×10^{-12}
SFRS5	12	36.7	66.4	3.76×10^{-15}	1.54×10^{-12}
NOSTRIN	5	36.7	65.1	7.40×10^{-15}	2.70×10^{-12}
KIF13B	12	36.7	63.4	1.69×10^{-14}	5.16×10^{-12}
TENC1	10	36.7	62.5	2.67×10^{-14}	7.40×10^{-12}
MFAP4	12	36.7	60.5	7.17×10^{-14}	1.67×10^{-11}
ELN	13	36.7	59.7	1.07×10^{-13}	2.32×10^{-11}
SGK223	14	36.7	59	1.51×10^{-13}	3.05×10^{-11}
KIF12	11	36.7	58.8	1.74×10^{-13}	3.34×10^{-11}
SELP	11	36.7	58.8	1.74×10^{-13}	3.34×10^{-11}
CIRBP	9	36.7	58.7	1.83×10^{-13}	3.41×10^{-11}
CTDSP1	9	36.7	58.7	1.83×10^{-13}	3.41×10^{-11}

4.1.1 Synthetic lethal pathways in breast cancer

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

4.1.2 Expression profiles of synthetic lethal partners

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

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

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

Pathways Over-represented in Cluster 1	Total Genes in Pathway	Genes in SL Cluster	p-val (FDR)
Collagen formation	67	10	4.0×10^{-11}
Extracellular matrix organisation	238	21	1.8×10^{-9}
Collagen biosynthesis and modifying enzymes	56	8	1.8×10^{-9}
Uptake and actions of bacterial toxins	22	5	9.5×10^{-9}
Elastic fibre formation	37	6	1.9×10^{-8}
Muscle contraction	62	7	2.4×10^{-7}
Fatty acid, triacylglycerol, and ketone body metabolism	117	10	4.9×10^{-7}
XBP1(S) activates chaperone genes	51	6	6.6×10^{-7}
IRE1alpha activates chaperones	54	6	1.2×10^{-6}
Neurotoxicity of clostridium toxins	10	3	1.3×10^{-6}
Retrograde neurotrophin signalling	10	3	1.3×10^{-6}
Assembly of collagen fibrils and other multimeric structures	40	5	1.9×10^{-6}
Collagen degradation	58	6	2.0×10^{-6}
Arachidonic acid metabolism	41	5	2.1×10^{-6}
Synthesis of PA	26	4	3.0×10^{-6}
Signaling by NOTCH	80	7	3.3×10^{-6}
Signalling to RAS	27	4	3.7×10^{-6}
Integrin cell surface interactions	82	7	4.2×10^{-6}
Smooth Muscle Contraction	28	4	4.4×10^{-6}

Pathways Over-represented in Cluster 2	Total Genes in Pathway	Genes in SL Cluster	p-val (FDR)
Eukaryotic Translation Elongation	86	75	1.1×10^{-181}
Viral mRNA Translation	81	72	9.8×10^{-179}
Peptide chain elongation	83	72	1.9×10^{-175}
Eukaryotic Translation Termination	83	72	1.9×10^{-175}
Formation of a pool of free 40S subunits	93	75	1.9×10^{-171}
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	88	72	9.9×10^{-168}
L13a-mediated translational silencing of Ceruloplasmin expression	103	75	3.0×10^{-159}
3' -UTR-mediated translational regulation	103	75	3.0×10^{-159}
Nonsense-Mediated Decay (NMD)	103	75	3.0×10^{-159}
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	103	75	3.0×10^{-159}
SRP-dependent cotranslational protein targeting to membrane	104	75	3.2×10^{-158}
GTP hydrolysis and joining of the 60S ribosomal subunit	104	75	3.2×10^{-158}
Eukaryotic Translation Initiation	111	75	4.5×10^{-151}
Cap-dependent Translation Initiation	111	75	4.5×10^{-151}
Influenza Infection	117	75	1.4×10^{-145}
Influenza Viral RNA Transcription and Replication	108	72	5.7×10^{-145}
Translation	141	81	8.0×10^{-143}
Influenza Life Cycle	112	72	2.3×10^{-141}
Infectious disease	347	103	2.2×10^{-95}

Pathways Over-represented in Cluster 3	Total Genes in Pathway	Genes in SL Cluster	p-val (FDR)
Adaptive Immune System	412	90	6.1×10^{-61}
Chemokine receptors bind chemokines	52	27	6.7×10^{-56}
Generation of second messenger molecules	29	21	6.5×10^{-55}
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	64	29	6.5×10^{-55}
TCR signalling	62	27	8.9×10^{-51}
Peptide ligand-binding receptors	161	40	1.5×10^{-45}
Translocation of ZAP-70 to Immunological synapse	16	14	3.1×10^{-43}
Costimulation by the CD28 family	51	22	4.0×10^{-43}
PD-1 signalling	21	15	4.0×10^{-41}
Class A/1 (Rhodopsin-like receptors)	258	50	6.7×10^{-41}
Phosphorylation of CD3 and TCR zeta chains	18	14	1.3×10^{-40}
Interferon gamma signalling	74	24	5.0×10^{-39}
GPCR ligand binding	326	57	1.8×10^{-38}
Cytokine Signaling in Immune system	268	48	8.9×10^{-37}
Downstream TCR signalling	45	18	1.8×10^{-35}
$G_{\alpha i}$ signalling events	167	33	2.2×10^{-33}
Cell surface interactions at the vascular wall	99	21	1.3×10^{-26}
Interferon Signalling	164	28	1.7×10^{-26}
Extracellular matrix organisation	238	35	2.7×10^{-25}

Pathways Over-represented in Cluster 4	Total Genes in Pathway	Genes in SL Cluster	p-val (FDR)
Extracellular matrix organisation	238	48	8.0×10^{-41}
Class A/1 (Rhodopsin-like receptors)	258	47	2.8×10^{-36}
GPCR ligand binding	326	54	2.1×10^{-34}
$G_{\alpha s}$ signalling events	83	22	1.4×10^{-31}
GPCR downstream signalling	472	68	1.1×10^{-29}
Haemostasis	423	61	3.3×10^{-29}
Platelet activation, signalling and aggregation	180	31	7.1×10^{-28}
Binding and Uptake of Ligands by Scavenger Receptors	40	14	9.9×10^{-27}
RA biosynthesis pathway	22	11	2.5×10^{-26}
Response to elevated platelet cytosolic Ca^{2+}	82	19	3.0×10^{-26}
Developmental Biology	420	57	3.5×10^{-26}
$G_{\alpha i}$ signalling events	167	28	7.3×10^{-26}
Platelet degranulation	77	18	1.6×10^{-25}
Gastrin-CREB signalling pathway via PKC and MAPK	171	28	2.5×10^{-25}
Muscle contraction	62	16	4.7×10^{-25}
$G_{\alpha q}$ signalling events	150	25	3.2×10^{-24}
Retinoid metabolism and transport	34	12	5.0×10^{-24}
Phase 1 - Functionalisation of compounds	67	16	6.5×10^{-24}
Signalling by Retinoic Acid	42	13	6.7×10^{-24}

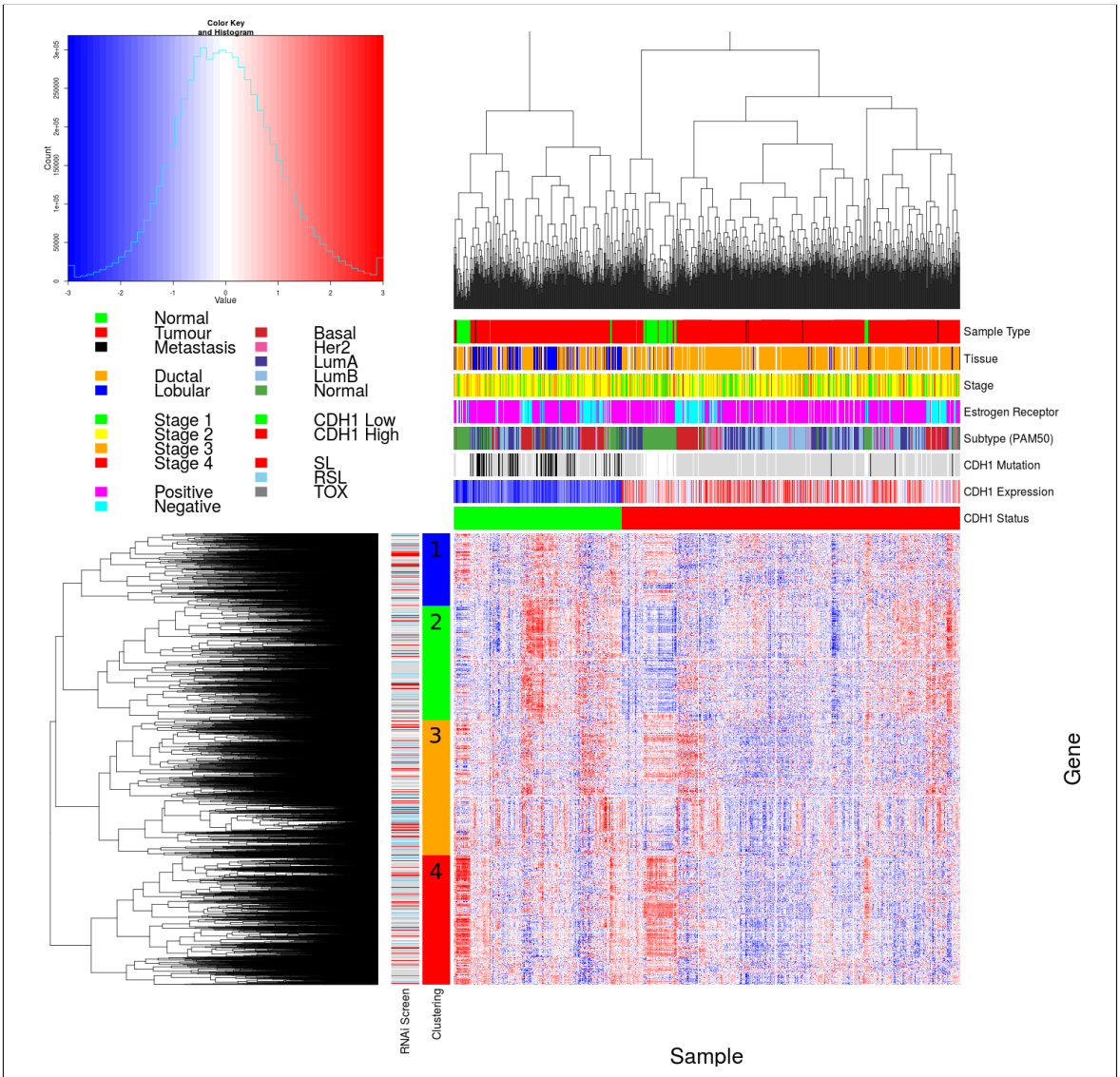

Figure 4.1: **Synthetic lethal expression profiles of analysed samples.** Gene expression profile heatmap (correlation distance) of all samples (separated by the 1/3 quantile of *CDH1* expression) analysed in TCGA breast cancer dataset for gene expression of 4,629 candidate partners of E-cadherin (*CDH1*) 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.

4.1.2.1 Subgroup pathway analysis

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

More recent analyses have also investigated intrinsic (PAM50) subtype and somatic mutation (of highest impact genes) against these gene clusters.

4.2 Comparison of synthetic lethal gene candidates

4.2.1 Comparison with differential expression

4.2.2 Comparison with correlation

4.2.3 Comparison with primary siRNA screen candidates

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

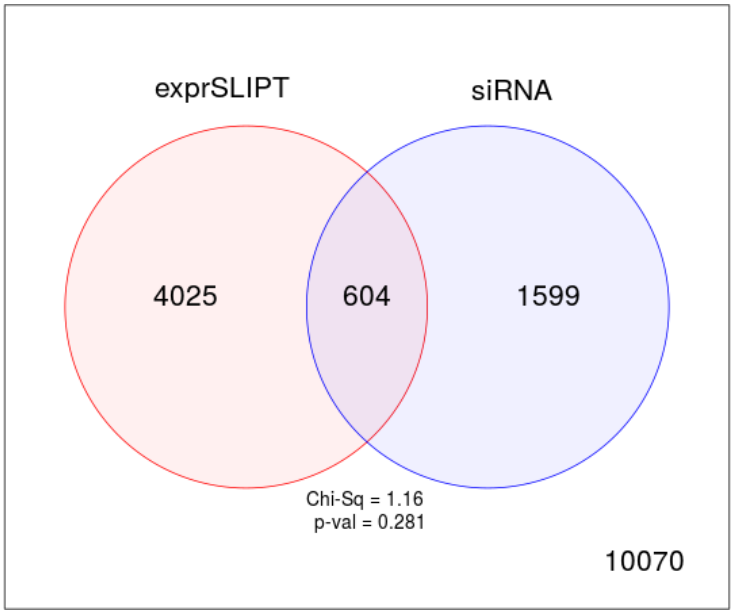

Figure 4.2: **Comparison of SLIPT to siRNA.** Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ^2 test suggests that the overlap is no more than would be expected by chance ($p = 0.281$).

pathways more than we expect by chance.

Both of these are being applied with developing a method and overcoming technical challenges for the latter being the focus of recent work. The main challenge at the moment is to compare SLIPT results to experimental candidates and explain why so few genes (and so many pathways) overlap.

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

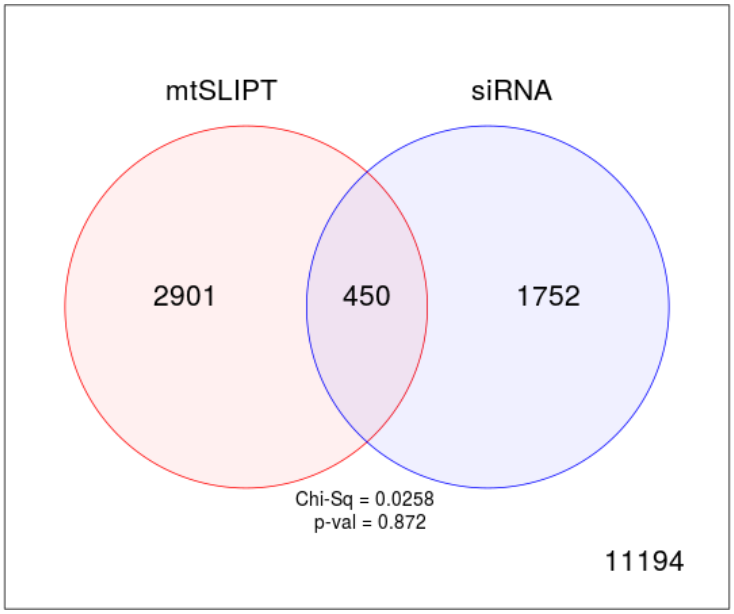

Figure 4.3: **Comparison of mtSLIPT to siRNA.** Testing the overlap of gene candidates for E-cadherin synthetic lethal partners between computational (SLIPT) and experimental screening (siRNA) approaches. The χ^2 test suggests that the overlap is no more than would be expected by chance ($p = 0.281$).

4.2.3.1 Comparison of screen at pathway level

4.2.3.1.1 Resampling of genes for pathway enrichment

4.2.4 Comparison with secondary screen siRNA screen candidates

4.2.4.1 Comparison of candidate SL Pathways

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

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

Predicted only by SLIPT (4025 genes)	**Total Genes in Pathway**	**Genes Identified**	**p-val (FDR)**
Eukaryotic Translation Elongation	80	75	1.5×10^{-182}
Peptide chain elongation	77	72	2.9×10^{-176}
Viral mRNA Translation	75	70	4.9×10^{-172}
Eukaryotic Translation Termination	76	70	5.9×10^{-170}
Formation of a pool of free 40S subunits	87	74	9.5×10^{-166}
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	81	70	1.2×10^{-160}
L13a-mediated translational silencing of Ceruloplasmin expression	97	75	3.8×10^{-155}
3' -UTR-mediated translational regulation	97	75	3.8×10^{-155}
GTP hydrolysis and joining of the 60S ribosomal subunit	98	75	6.0×10^{-154}
Nonsense-Mediated Decay (NMD)	96	73	5.2×10^{-150}
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	96	73	5.2×10^{-150}
SRP-dependent cotranslational protein targeting to membrane	97	73	7.8×10^{-149}
Eukaryotic Translation Initiation	105	75	4.7×10^{-146}
Cap-dependent Translation Initiation	105	75	4.7×10^{-146}
Translation	133	83	4.0×10^{-142}
Influenza Viral RNA Transcription and Replication	102	71	2.9×10^{-137}
Influenza Infection	111	74	3.7×10^{-137}
Influenza Life Cycle	106	71	2.3×10^{-133}
Infectious disease	326	125	4.2×10^{-120}
Extracellular matrix organisation	189	77	5.4×10^{-95}

Detected only by siRNA screen (1599 genes)	**Total Genes in Pathway**	**Genes Identified**	**p-val (FDR)**
Class A/1 (Rhodopsin-like receptors)	282	44	1.3×10^{-27}
GPCR ligand binding	363	52	5.8×10^{-26}
$G_{\alpha q}$ signalling events	159	26	6.7×10^{-23}
Gastrin-CREB signalling pathway via PKC and MAPK	180	27	2.0×10^{-21}
$G_{\alpha i}$ signalling events	184	27	5.3×10^{-21}
Downstream signal transduction	146	23	7.6×10^{-21}
Signalling by PDGF	172	25	4.0×10^{-20}
Peptide ligand-binding receptors	175	25	8.5×10^{-20}
Signalling by ERBB2	146	22	1.3×10^{-19}
DAP12 interactions	159	23	2.6×10^{-19}
DAP12 signalling	149	22	2.7×10^{-19}
Organelle biogenesis and maintenance	264	33	5.5×10^{-19}
Signalling by NGF	266	33	8.2×10^{-19}
Downstream signalling of activated FGFR1	134	20	1.1×10^{-18}
Downstream signalling of activated FGFR2	134	20	1.1×10^{-18}
Downstream signalling of activated FGFR3	134	20	1.1×10^{-18}
Downstream signalling of activated FGFR4	134	20	1.1×10^{-18}
Signalling by FGFR	146	21	1.3×10^{-18}
Signalling by FGFR1	146	21	1.3×10^{-18}
Signalling by FGFR2	146	21	1.3×10^{-18}

Intersection of SLIPT and siRNA screen (604 genes)	**Total Genes in Pathway**	**Genes Identified**	**p-val (FDR)**
Visual phototransduction	54	9	6.9×10^{-10}
$G_{\alpha s}$ signalling events	48	7	1.6×10^{-7}
Retinoid metabolism and transport	24	5	1.7×10^{-7}
Acyl chain remodelling of PS	10	3	6.5×10^{-6}
Transcriptional regulation of white adipocyte differentiation	51	6	6.5×10^{-6}
Chemokine receptors bind chemokines	22	4	6.5×10^{-6}
Signalling by NOTCH4	11	3	6.9×10^{-6}
Defective EXT2 causes exostoses 2	11	3	6.9×10^{-6}
Defective EXT1 causes exostoses 1, TRPS2 and CHDS	11	3	6.9×10^{-6}
Platelet activation, signalling and aggregation	146	12	6.9×10^{-6}
Phase 1 - Functionalisation of compounds	41	5	1.3×10^{-5}
Amine ligand-binding receptors	13	3	1.7×10^{-5}
Acyl chain remodelling of PE	14	3	2.4×10^{-5}
Signalling by GPCR	300	23	2.4×10^{-5}
Molecules associated with elastic fibres	29	4	2.6×10^{-5}
DAP12 interactions	128	10	2.6×10^{-5}
Cytochrome P_{450} - arranged by substrate type	30	4	3.2×10^{-5}
GPCR ligand binding	147	11	3.8×10^{-5}
Acyl chain remodelling of PC	16	3	4.0×10^{-5}
Response to elevated platelet cytosolic Ca^{2+}	66	6	4.2×10^{-5}

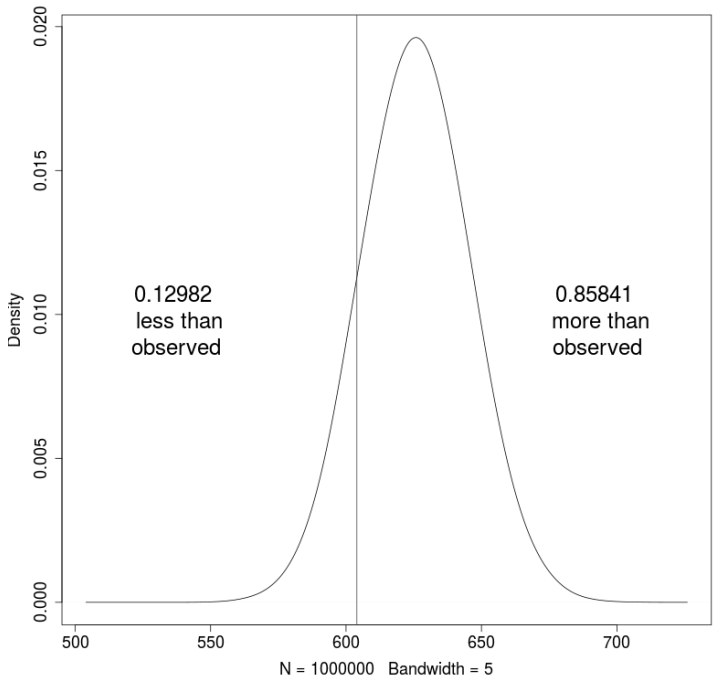

Figure 4.4: **Expected distribution of sample size for intersect with siRNA candidates.** Resampling analysis of intersect size from genes detected by SLIPT and siRNA screening approaches over 1 million replicates. The proportion of expected intersection sizes for random samples below or above the observed intersection size respectively, lacking significant over-representation or depletion of siRNA screen candidates within the SLIPT predictions for *CDH1*.

4.3 Mutation, Copy Number, and Methylation

Due to promising synthetic lethal data on mutation and DNA copy number analyses (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015), these were also investigated to compare genes for synthetic lethality in an analogous manner to expression analyses in the TCGA data. Due to the low somatic mutation rate (and lack of available) germline mutations for many genes, it was not possible to detect many double mutations with significantly under-representation in cancers. There were also concerns about using rare mutations with unknown significance or excluding functional mutations by only using those in the exons. It was possible to compare deletion and duplication of DNA copy number in a manner analogous to expression quantiles. However, these overlapped poorly with candidate interacting partners from expression analyses and concerns were raised that

Table 4.5: Pathway Analysis of *CDH1* partners from SLIPT

Reactome Pathway	Over-representation (FDR p-val)	Permutation (FDR p-val)
Eukaryotic Translation Elongation	1.3×10^{-207}	$< 1.241 \times 10^{-5}$
Peptide chain elongation	5.6×10^{-201}	$< 1.241 \times 10^{-5}$
Viral mRNA Translation	1.2×10^{-196}	$< 1.241 \times 10^{-5}$
Eukaryotic Translation Termination	1.2×10^{-196}	$< 1.241 \times 10^{-5}$
Formation of a pool of free 40S subunits	3.7×10^{-194}	$< 1.241 \times 10^{-5}$
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	5.3×10^{-187}	$< 1.241 \times 10^{-5}$
L13a-mediated translational silencing of Ceruloplasmin expression	9.6×10^{-183}	$< 1.241 \times 10^{-5}$
3' -UTR-mediated translational regulation	9.6×10^{-183}	$< 1.241 \times 10^{-5}$
GTP hydrolysis and joining of the 60S ribosomal subunit	1.9×10^{-181}	$< 1.241 \times 10^{-5}$
Nonsense-Mediated Decay (NMD)	6.2×10^{-176}	$< 1.241 \times 10^{-5}$
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	6.2×10^{-176}	$< 1.241 \times 10^{-5}$
Adaptive Immune System	6.5×10^{-174}	0.15753
Eukaryotic Translation Initiation	5.7×10^{-173}	$< 1.241 \times 10^{-5}$
Cap-dependent Translation Initiation	5.7×10^{-173}	$< 1.241 \times 10^{-5}$
SRP-dependent cotranslational protein targeting to membrane	2.0×10^{-171}	$< 1.241 \times 10^{-5}$
Translation	6.1×10^{-170}	$< 1.241 \times 10^{-5}$
Infectious disease	1.6×10^{-166}	0.23231
Influenza Infection	1.9×10^{-163}	$< 1.241 \times 10^{-5}$
Influenza Viral RNA Transcription and Replication	1.9×10^{-160}	$< 1.241 \times 10^{-5}$
Influenza Life Cycle	2.5×10^{-156}	$< 1.241 \times 10^{-5}$
Extracellular matrix organisation	1.1×10^{-152}	0.071761
GPCR ligand binding	1.1×10^{-143}	0.55801
Class A/1 (Rhodopsin-like receptors)	1.5×10^{-142}	0.58901
GPCR downstream signalling	7.6×10^{-140}	0.098357
Haemostasis	1.9×10^{-134}	0.27059
Developmental Biology	2.0×10^{-123}	0.52737
Metabolism of lipids and lipoproteins	3.3×10^{-120}	0.724
Cytokine Signalling in Immune system	2.6×10^{-119}	0.39661
Peptide ligand-binding receptors	3.7×10^{-109}	0.61102
$G_{\alpha i}$ signalling events	8.9×10^{-100}	$< 1.241 \times 10^{-5}$

they may not be relevant to *CDH1* which is typically inactivated in tumours by loss of function mutations or DNA methylation (PJ Guilford, personal communication).

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

4.3.1 Synthetic lethality by DNA copy number

4.3.2 Synthetic lethality by somatic mutation

4.3.2.1 Mutation analysis

4.3.3 ANOVA of Expression Predictors

[include?]

Another approach was to only use copy number, mutation, or hyper-methylation data for genes in which they would impact on gene function and occur frequently in tumours. Before investigating whether these impact on gene function, they were investigated as predictors of variation in gene expression. If these are not giving variation

Table 4.6: Pathway Analysis of *CDH1* partners from SLIPT and siRNA primary screen

Reactome Pathway	**Over-representation (FDR)**	**Permutation (FDR)**
Visual phototransduction	6.9×10^{-10}	0.91116
$G_{\alpha s}$ signalling events	1.6×10^{-7}	0.012988
Retinoid metabolism and transport	1.7×10^{-7}	0.20487
Transcriptional regulation of white adipocyte differentiation	6.5×10^{-6}	0.38197
Acyl chain remodelling of PS	6.5×10^{-6}	0.58485
Chemokine receptors bind chemokines	6.5×10^{-6}	0.97255
Defective EXT2 causes exostoses 2	6.9×10^{-6}	0.056437
Defective EXT1 causes exostoses 1, TRPS2 and CHDS	6.9×10^{-6}	0.056437
Signalling by NOTCH4	6.9×10^{-6}	0.15497
Platelet activation, signalling and aggregation	6.9×10^{-6}	0.53358
Phase 1 - Functionalisation of compounds	1.3×10^{-5}	0.24836
Amine ligand-binding receptors	1.7×10^{-5}	0.3195
Acyl chain remodelling of PE	2.4×10^{-5}	0.7307
Signalling by GPCR	2.4×10^{-5}	0.9939
Molecules associated with elastic fibres	2.6×10^{-5}	0.0072929
DAP12 interactions	2.6×10^{-5}	0.78273
Cytochrome P_{450} - arranged by substrate type	3.2×10^{-5}	0.87019
GPCR ligand binding	3.8×10^{-5}	0.99417
Acyl chain remodelling of PC	4.0×10^{-5}	0.65415
Response to elevated platelet cytosolic Ca^{2+}	4.2×10^{-5}	0.55461
Arachidonic acid metabolism	4.4×10^{-5}	0.060298
Defective B4GALT7 causes EDS, progeroid type	4.9×10^{-5}	0.15497
Defective B3GAT3 causes JDSSDHD	4.9×10^{-5}	0.15497
Elastic fibre formation	4.9×10^{-5}	0.0019227
HS-GAG degradation	6.2×10^{-5}	0.017747
Bile acid and bile salt metabolism	6.2×10^{-5}	0.15497
Netrin-1 signalling	7.1×10^{-5}	0.95056
Integration of energy metabolism	7.1×10^{-5}	0.0019287
DAP12 signalling	7.9×10^{-5}	0.67835
GPCR downstream signalling	8.1×10^{-5}	0.88678
Diseases associated with glycosaminoglycan metabolism	8.7×10^{-5}	0.017747
Diseases of glycosylation	8.7×10^{-5}	0.017747
Signalling by Retinoic Acid	8.7×10^{-5}	0.13592
Signalling by Leptin	8.7×10^{-5}	0.15497
Signalling by SCF-KIT	8.7×10^{-5}	0.73399
Opioid Signalling	8.7×10^{-5}	0.99417
Signalling by NOTCH	0.0001	0.26453
Platelet homeostasis	0.0001	0.55912
Signalling by NOTCH1	0.00011	0.13797
Class B/2 (Secretin family receptors)	0.00011	0.4659
Diseases of Immune System	0.00013	0.15497
Diseases associated with the TLR signalling cascade	0.00013	0.15497
A tetrasaccharide linker sequence is required for GAG synthesis	0.00013	0.33566
Nuclear Receptor transcription pathway	0.00016	0.22735
Formation of Fibrin Clot (Clotting Cascade)	0.00016	0.0054639
Syndecan interactions	0.00016	0.3974
Class A/1 (Rhodopsin-like receptors)	0.00016	0.99454
HS-GAG biosynthesis	0.0002	0.37199
Platelet degranulation	0.0002	0.39003
EPH-ephrin mediated repulsion of cells	0.00021	0.6193

independent of gene expression, expression would be a more suitable measure of gene function as it is widely generated in studies and useful as a clinical biomarker.

Globally predicting gene expression across all genes from DNA copy number and somatic mutation was attempted by ANOVA. However, this was computationally challenging and gene-specific analyses would be more informative. Gene specific ANOVA and linear regression was performed but was raised more issues than it addressed. There were issues with interaction terms and mutation data, many genes were not tested for these since there were so few mutations for these genes in the dataset. It was possible to include DNA methylation in gene-specific analyses (despite the concerns raised above) but the R^2 values for each gene were still generally very low and issues with insufficient mutant samples for interaction terms became worse. This means that the approach used differs for each gene making it difficult to compare them. The challenges raised here suggested that expression is very difficult to predict with other factors but including these other factors would be difficult and plagued by multiple-testing, particularly comparing between them with the current synthetic lethal prediction method. This led to investigations into the simulation of synthetic lethality.

4.4 Global Synthetic Lethality

[include?]

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

4.4.1 Hub Genes

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

Table 2 Hub gene function in TCGA breast cancer RNA-Seq expression SL predictions (n=878). [revise for n=1168]

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

4.5 Metagene Analysis

[include?]

4.5.1 Pathway expression

4.5.2 Somatic mutation

4.5.3 Synthetic lethal metagenes

4.6 Replication in stomach cancer

- exprSL
- mtSL
- heatmap
- Venn
- Pathway enrichment
- Permutations

4.7 Replication in cell line encyclopaedia

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

4.8 Summary

We have developed a simple, interpretable, computational approach to predict synthetic lethal partners from genomics data. Originally developed for microarray gene expression data, it has been expanded to test DNA copy number, or RNA-Seq gene expression data which are both also supported by the TCGA dataset. DNA copy number was included for comparison with the DAISY tool of Jerby-Arnon *et al.* (2014). Predictions based on microarray data were inconclusive when compared with an RNAi screen for *CDH1* in MCF10A breast cells as performed by Telford *et al.* (2015), few predictions replicated between BC2116, CCLE, or TCGA microarray datasets, results with gene expression and DNA copy number were vastly different, and predictions from TCGA microarray and RNA-Seq datasets for the same samples differed were inconsistent. The Aligent TCGA microarray data in particular is difficult to compare to other datasets and will in the future use Affymetrix microarrays or RNA-Seq platforms for predictions from gene expression data. The analyses focus on gene expression data as it is widely available for applications in other cancers and current attempts to use gene expression data for synthetic lethal discovery vary widely (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015; Tiong *et al.*, 2014). There is no consensus for which approach is more appropriate since they lack much a basis on biological experimental data or statistical modelling and often use difficult to interpret machine learning methodology.

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

other datasets, as expected from lower sample size, lower concordance of particular hits for the example query of *CDH1*, and suspected lower quality of data on the Aligent microarray platform.

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

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

Ductal breast cancers show higher expression of synthetic lethal partners suggesting treatment would be more effective in this tumour subtype. However, there is consistently low expression of SL partners in ER negative tumours, although this is independent of tumour stage and consistent with poor prognosis in these patients and could inform other treatment strategies or prevent ineffective treatment further impacting quality of life in these patients. These results suggest that synthetic lethal partner expression varies between patients; that these different tumour classes would react differently to the same treatment; that treatment of different pathways and combinations in different patients is the most effective approach to target genes compensating for *CDH1* gene loss; and the expression of synthetic partners could be a clinically impor-

tant biomarker. While these are important clinical implications, the synthetic lethal predictions lack enough confidence for direct translation into pre-clinical models or clinical applications leading to a need for statistical modelling and simulation of synthetic lethality in genomics expression data.

References

Araki, H., Knapp, C., Tsai, P., and Print, C. (2012) Genesetdb: A comprehensive meta-database, statistical and visualisation framework for gene set analysis. *FEBS Open Bio*, **2**: 76–82.

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

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.

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.

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

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.

Merlos-Suarez, A., Barriga, F.M., Jung, P., Iglesias, M., Cespedes, M.V., Rossell, D., Sevillano, M., Hernando-Momblona, X., da Silva-Diz, V., Munoz, P., *et al.* (2011) The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell*, **8**(5): 511–524.

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.

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.

Warburg, O. (1956) On the origin of cancer cells. *Science*, **123**(3191): 390–314.