Sifting the Needles in the Haystack:
Permutation Resampling Biological Pathways in Cancer Genomic Interaction Data

University of Otago
Te Whare Wānanga o Otāgo
New Zealand

Tom Kelly

Bryony Telford & Augustine Chen (experimental data)

Mik Black & Parry Guilford (PhD supervisors)

NeSI New Zealand eScience Infrastructure

NZGL New Zealand Genomics Limited

REANNZ

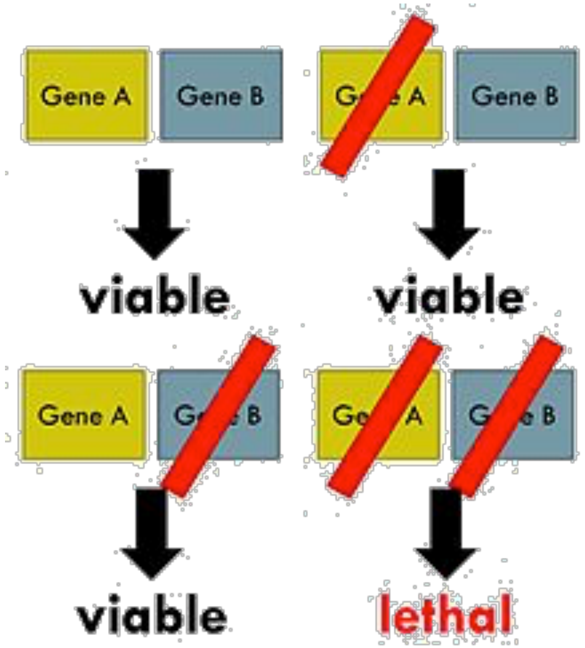
eResearch NZ 2016

9-11 February 2016 | Rydges Hotel | Queenstown

Genetics – Synthetic Lethality

Genetics – Synthetic Lethality

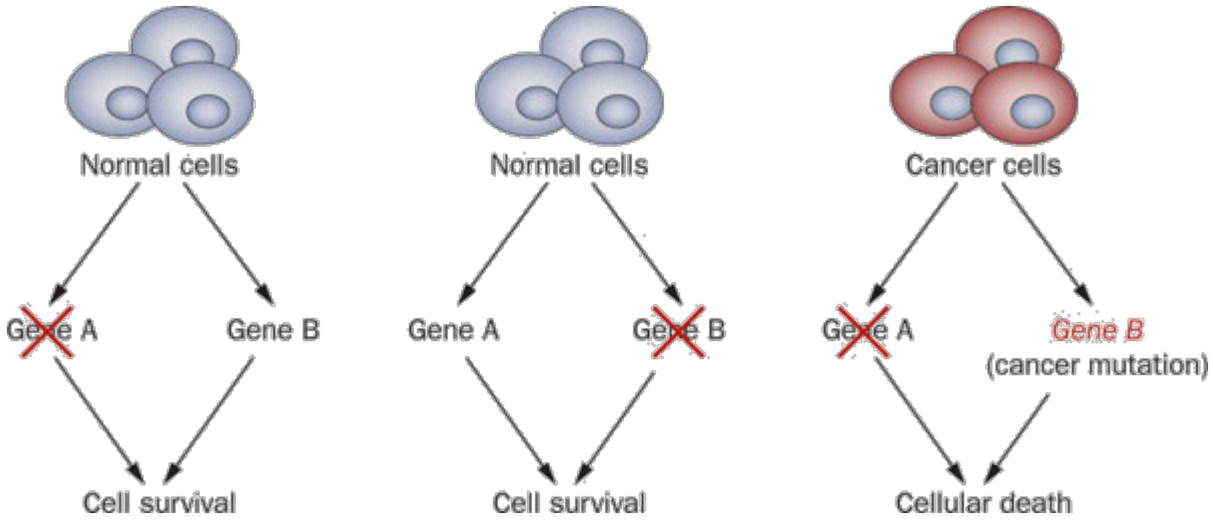
- Cell death due to inactivation of two (or more) non-essential genes
  - Loss of a shared function being lethal implies functional redundancy
  - Conserved between pathways more than individual genes



(cc) AthenaPendergrass Wikipedia

Genetics – Synthetic Lethality

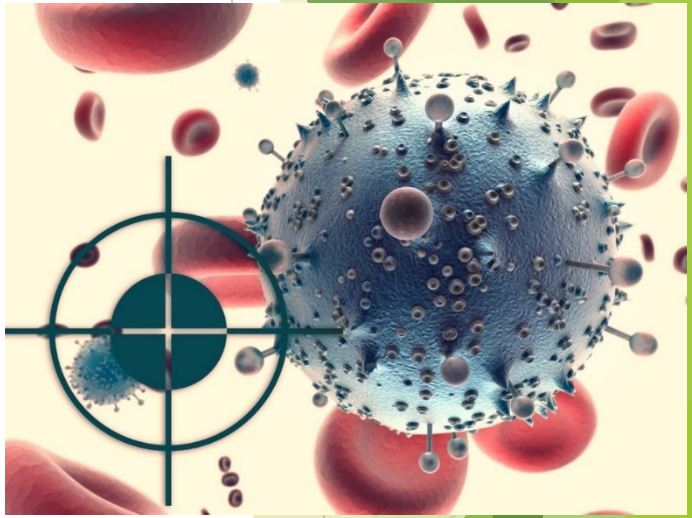
- Cell death due to inactivation of two (or more) non-essential genes
  - Loss of a shared function being lethal implies functional redundancy
  - Conserved between pathways more than individual genes



Rehman *et al.* (2010) *Nature Reviews Clinical Oncology* **7**, 718-724

Genomics – Targeted Cancer Therapy

- An appealing strategy for anti-cancer drug design
  - Specificity against genetic abnormality (even loss of function)
    - We expect low adverse effects compared to chemotherapy
  - Enables wider use of targeted therapy
    - Drugs specific against molecular changes identified by Genetics/ Genomics
  - Has been shown to be a clinically applicable strategy
    - e.g., olaparib (*BRCA* mutation, *PARP* inhibitors) successful clinical trials



http://www.oncology-central.com/2014/12/15/

Cancer Genomics – Data Sources

National Cancer Institute

National Human Genome Research Institute

The Cancer Genome Atlas Data Portal

Understanding genomics to improve cancer care

TCGA Home | Contact Us | For the Media

International Cancer Genome Consortium

cBioPortal for Cancer Genomics

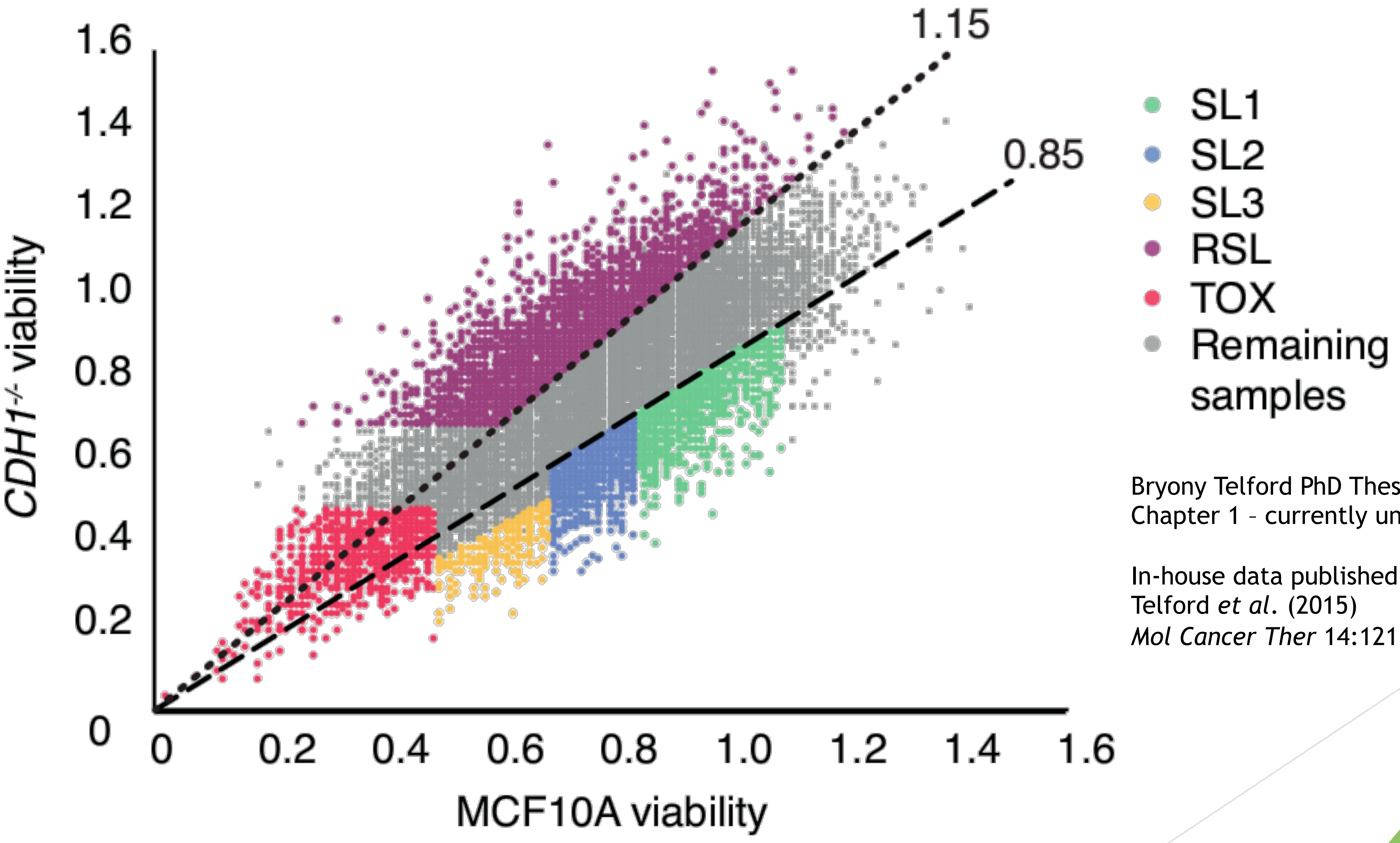
Genomic Screen – Experimental Data

- Until recently limited to a candidate approach
  - Based on known functions and studies in other species
- Screening for Synthetic Lethality has become a popular in cancer cell culture
  - Uses “ RNA interference” to knockout gene expression: screening for mutant-specific cell death
  - Combined with drug compound testing for cancer drug screening
  - Other refined gene knockdown approaches in development (e.g., CRISPR ‘genome editing’)
- Experimental screening (and validation) is costly, laborious, and prone to false positives
  - We are investigating bioinformatics analysis to assist the drug target triage process

E-cadherin (*CDH1*) – Example Gene

- E-Cadherin (encoded by the *CDH1* gene) is a cell-to-cell signalling and cell structure protein
    - Tumour suppressor (loss linked to cancer onset and progression)
- Hereditary Diffuse Gastric Cancer (Familial cancer syndrome)
    - High risk and early onset diffuse gastric cancer and lobular breast cancer
    - Current monitoring or surgery options have significant risk of patient harm
- The Cancer Genetics Lab has an ongoing project aiming to design safe drugs suitable for early stage treatment and preventative use in outwardly healthy HDGC patients / mutation carriers

E-cadherin (*CDH1*) – Example Gene



Bryony Telford PhD Thesis (2015)
Chapter 1 – currently under examination

In-house data published as:
Telford *et al*. (2015)
Mol Cancer Ther 14:1213

SLIPT - Prediction Method

- Synthetic Lethal Interaction Prediction Tool (SLIPT)
  - Score patients as low, medium or high expression for each gene (3-quantiles)
  - Chi-Square test gives significance for relationship between expression of 2 genes
  - Correct p-values for multiple tests (False Discovery Rate)
  - Score Synthetic Lethality as directional changes in expression as shown below:

		Candidate Gene (e.g. *SVIL*)		
		Low	Medium	High
Query Gene (e.g. *CDH1*)	Low	Observed less than expected	→	Observed more than expected
	Medium			
	High			

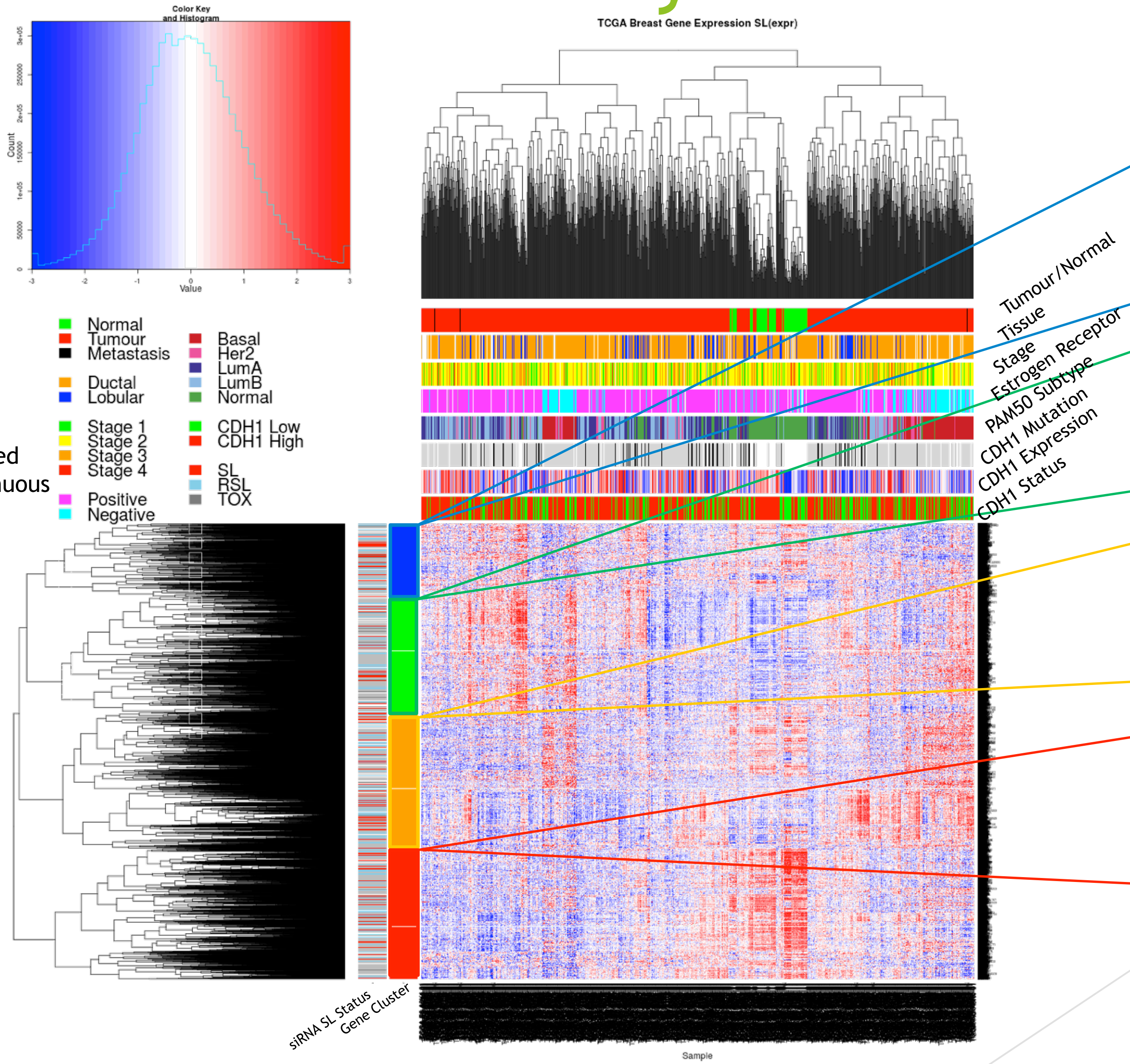
Methods – Pathway Prediction Workflow

- 1) Source data from database (and check for quality): TCGA/ICGC data portals
- 2) Predict Synthetic Lethal gene partners: SLIPT for *CDH1* in breast cancer
- 3) Gene Set over-representation analysis: ReactomeDB pathway enrichment

SLIPT – Enriched Pathways for *CDH1*

Correlation Matrix:
Voom Normalised
Plot z-transformed
Correlation Distance
Complete Linkage

Figure Legend
Key on Blue-Red
Scale if Continuous



Enriched Pathways

833

GPCR (B/2), Chaperones, Muscle Contraction, Fatty Acid Metab, G protein K+ channels, Gα(s), RAS, TGFβ, ERK, IL-6, GABAB

1307

Translation, Nonsense mediated decay, RNA metab, SRP-dep co-translation, TCA, Transcription, EFGR, Infection, Antigen

Active Immune

1547

Second messengers, TCR signal, Chemokines, PD-1, IFNγ, Peptide=ligand, GPCR (A/1), GPCR ligand, Gα(i), TLR, IL

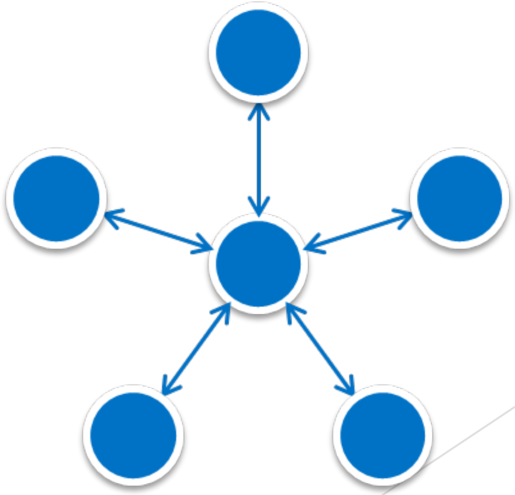
Innate Immune

1478

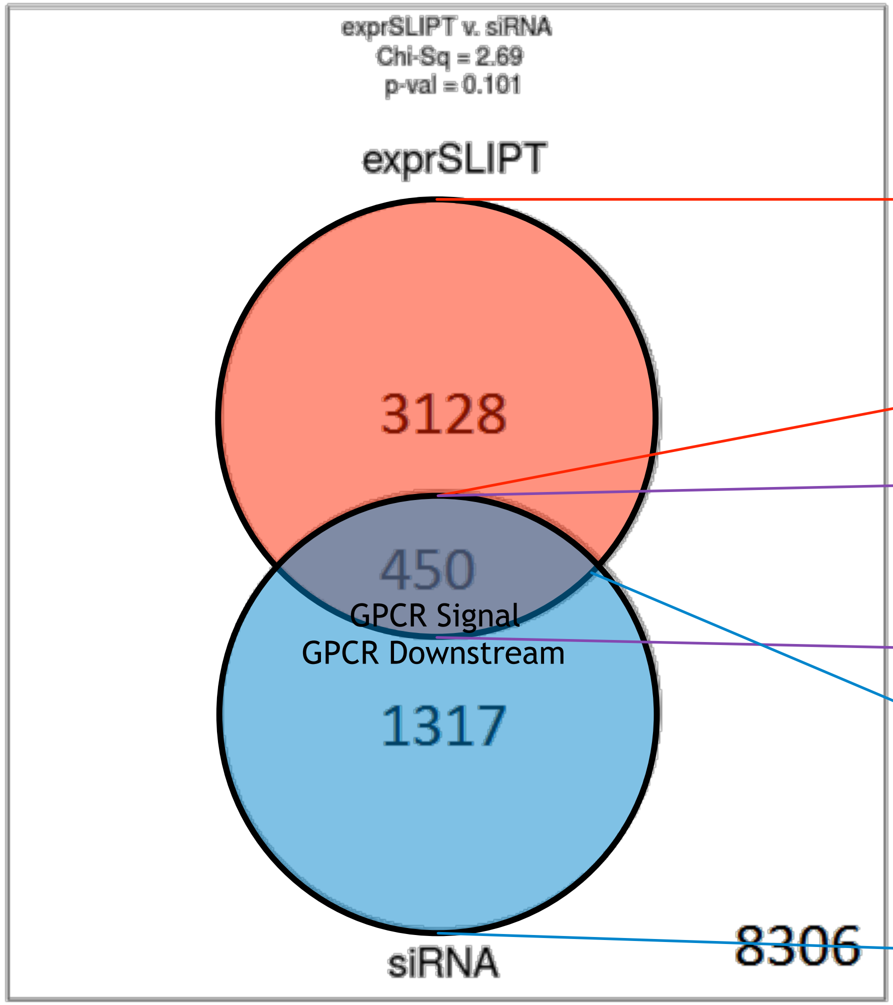
GPCR Ligand, GPCR (A/1), Gα(s), Muscle contraction, Homeostasis, Metab (ph 1), Ethanol, Develop, platelet, IGF, Gα(i), Gα(q), P2Y

Results so far

- Synthetic Lethal interactions are common across the Human Genome (used NeSI Pan cluster)
  - Consistent with scale-free networks observed in other species
- Expression of synthetic lethal partners across a patient cohort divides into several correlated clusters with:
  - Distinct functions
  - Highly expressed in different patient groups



SLIPT – Comparison to siRNA Genes



3126

Translation, 3’ UTR regulation, non-sense mediated decay, SRP-dependent co-translate, Immune, Cell Cycle, Chromatin Modifiers

450

Nicotinamide, Serotonin, GPCR Signal, Gα(s), GPCR Downstream, Innate Immune, GPCR ligand, Cytochrome P_{450}, GPCR (B/2)

1317

CaMK/CREB, Protein Metabolism, 3’ UTR Regulation, non-sense mediated decay, Translation, Phospholipase C, Calmodulin

Resampling – Permutations for Pathways

- The intersection between SLIPT and siRNA results is enriched for many of the same pathways as in the experimental siRNA data
  - Even though this differs greatly from the SLIPT results overall
  - Is this good news?
  - Or would we expect this by chance?
  - Can we explain why they overlap so poorly with siRNA hits?
- Permutation / Bootstrapping / Re-Sampling
  - The idea is to randomly sample / shuffle genes and to generate a test statistic distribution we would expect by chance
  - Then we can test whether genes are behaving as expected by chance or are we surprised by them

Resampling – Permutations for Pathways

- A random sample of the total observed size for predicted genes
  - e.g, 3576 genes predicted
- The intersection with siRNA candidates is derived from the random sample
  - Does not assume that the size of the intersection is fixed at the observed size
  - Size is not predetermined as and generates an expected intersection size
  - Observed intersection of 450 genes
- Test each sample for pathway enrichment
  - e.g., all 1652 Reactome pathways
- Rinse, repeat to generate an expect distribution (null hypothesis)

Re-sampling - Implementation

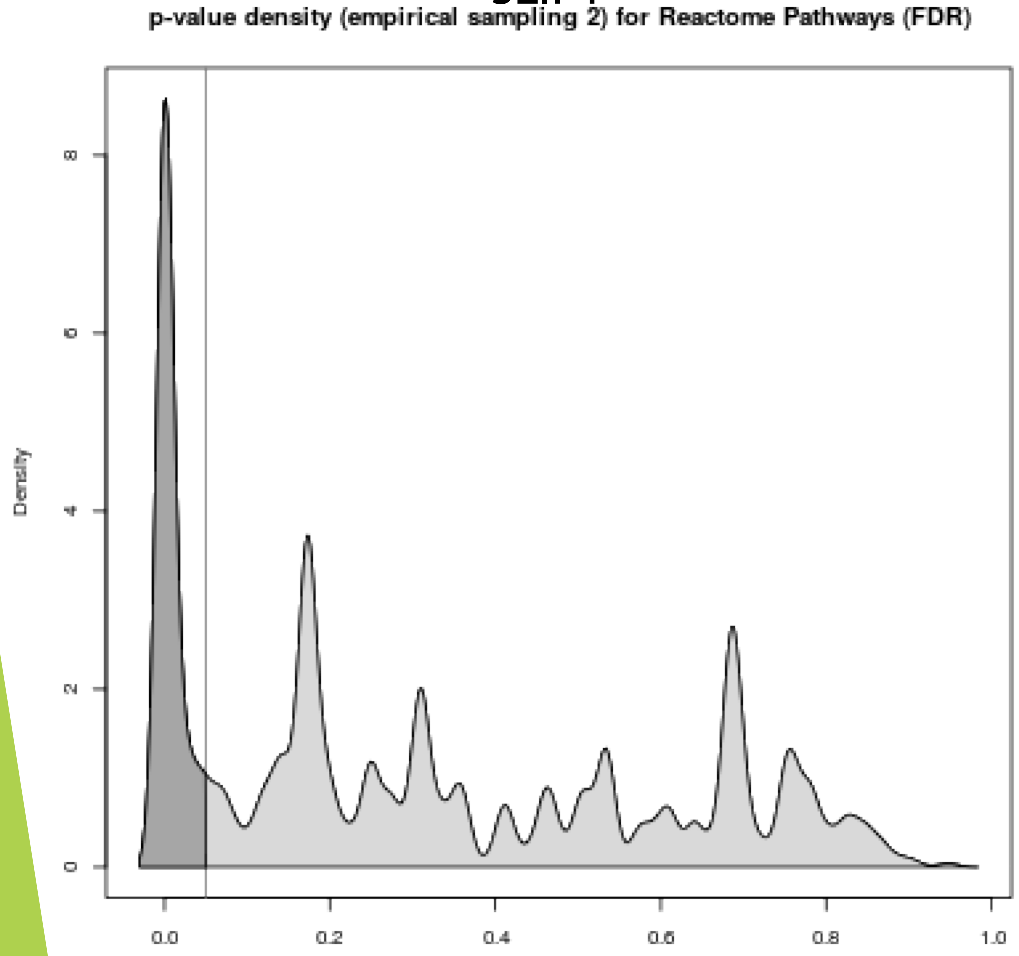
NeSI

New Zealand eScience Infrastructure

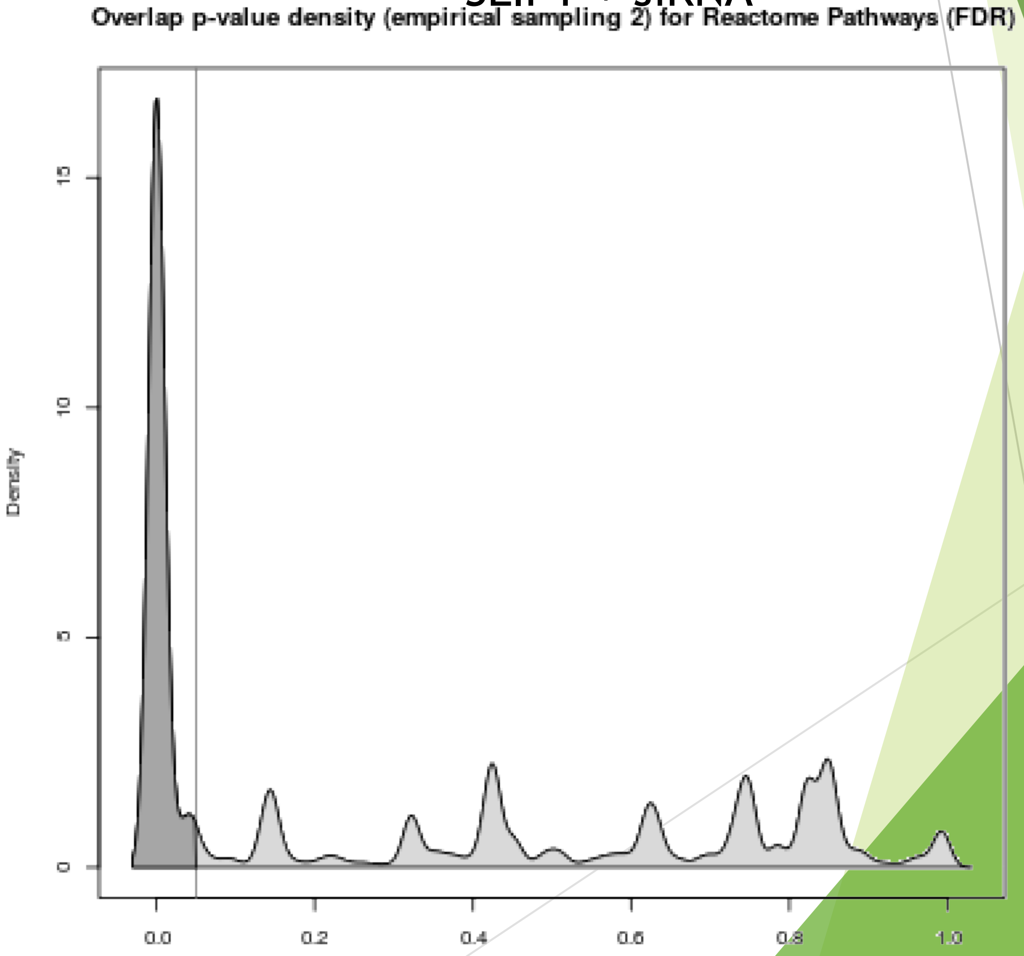
- The re-sampling approach was repeated 10,000 times
  - Running Rmpi on the New Zealand eScience Infrastructure Intel Pan Cluster
  - 1652 pathways were tested for enrichment in 10,000 simulated samples
- These were used to generate a null distribution of expected χ^2 values
  - for each Reactome pathway
  - for the SLIPT predictions and the intersection with experimental screen genes
- Empirical p-value estimates were derived from:
  - the proportion of the 10,000 null χ^2 values $\leq$ the observed χ^2 value
  - then adjusted (FDR) for multiple tests by the number of pathways
- Also preformed for the size of sampled intersections to test enrichment or depletion of siRNA candidate genes in SLIPT predictions

Re-sampling – Results (Adjusted p-value)

SLIPT

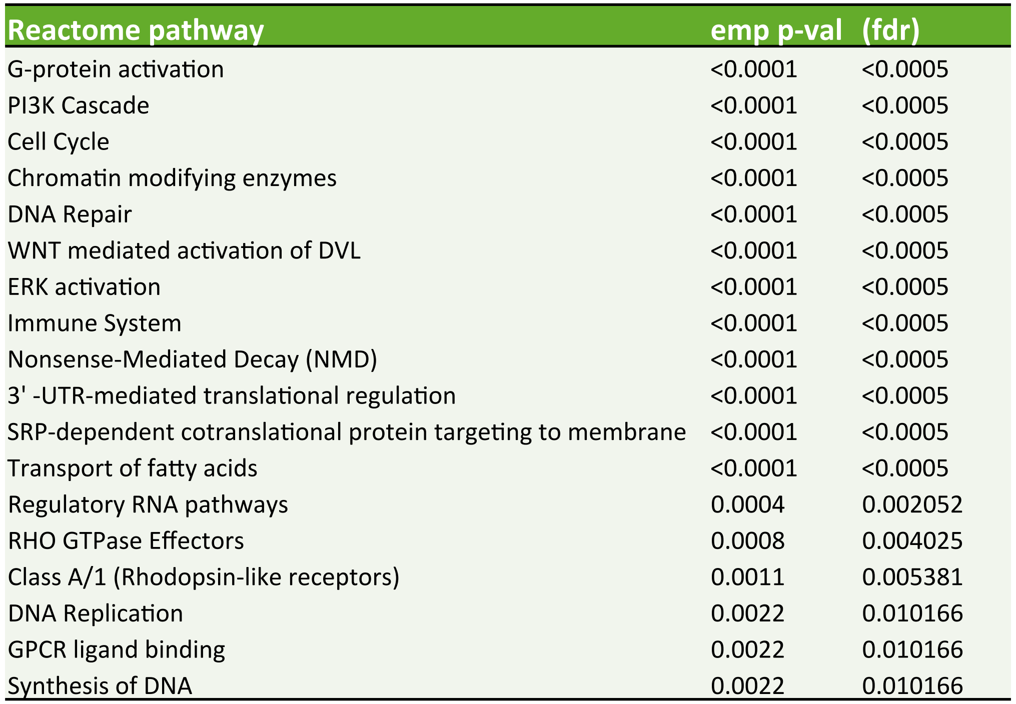


SLIPT + siRNA



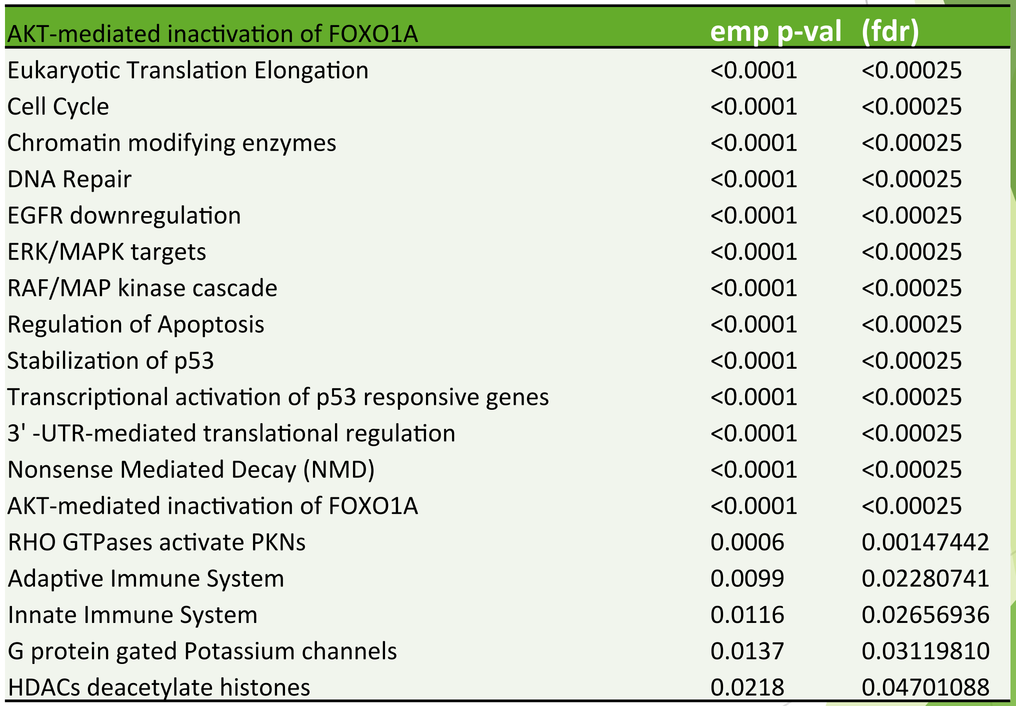
Re-sampling –Results (Key Pathways)

SLIPT



Reactome pathway	emp p-val	(fdr)
G-protein activation	<0.0001	<0.0005
PI3K Cascade	<0.0001	<0.0005
Cell Cycle	<0.0001	<0.0005
Chromatin modifying enzymes	<0.0001	<0.0005
DNA Repair	<0.0001	<0.0005
WNT mediated activation of DVL	<0.0001	<0.0005
ERK activation	<0.0001	<0.0005
Immune System	<0.0001	<0.0005
Nonsense-Mediated Decay (NMD)	<0.0001	<0.0005
3' -UTR-mediated translational regulation	<0.0001	<0.0005
SRP-dependent cotranslational protein targeting to membrane	<0.0001	<0.0005
Transport of fatty acids	<0.0001	<0.0005
Regulatory RNA pathways	0.0004	0.002052
RHO GTPase Effectors	0.0008	0.004025
Class A/1 (Rhodopsin-like receptors)	0.0011	0.005381
DNA Replication	0.0022	0.010166
GPCR ligand binding	0.0022	0.010166
Synthesis of DNA	0.0022	0.010166

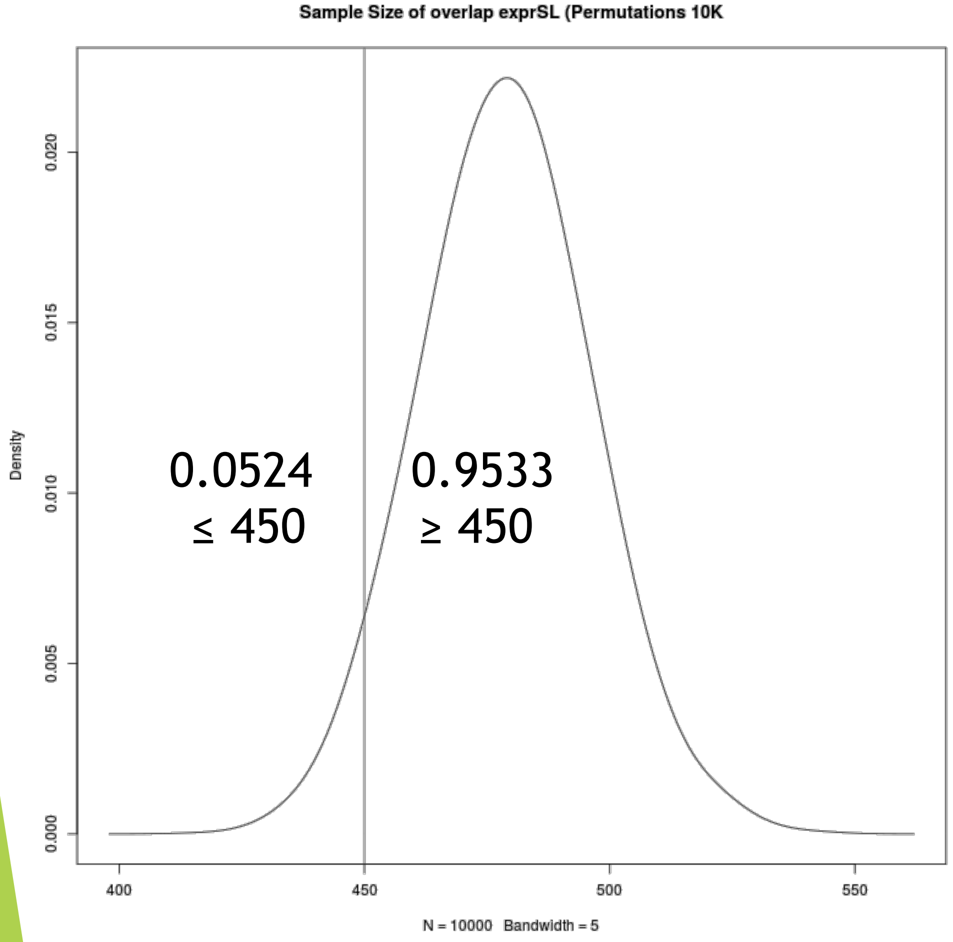
SLIPT + siRNA



AKT-mediated inactivation of FOXO1A	emp p-val	(fdr)
Eukaryotic Translation Elongation	<0.0001	<0.00025
Cell Cycle	<0.0001	<0.00025
Chromatin modifying enzymes	<0.0001	<0.00025
DNA Repair	<0.0001	<0.00025
EGFR downregulation	<0.0001	<0.00025
ERK/MAPK targets	<0.0001	<0.00025
RAF/MAP kinase cascade	<0.0001	<0.00025
Regulation of Apoptosis	<0.0001	<0.00025
Stabilization of p53	<0.0001	<0.00025
Transcriptional activation of p53 responsive genes	<0.0001	<0.00025
3' -UTR-mediated translational regulation	<0.0001	<0.00025
Nonsense Mediated Decay (NMD)	<0.0001	<0.00025
AKT-mediated inactivation of FOXO1A	<0.0001	<0.00025
RHO GTPases activate PKNs	0.0006	0.00147442
Adaptive Immune System	0.0099	0.02280741
Innate Immune System	0.0116	0.02656936
G protein gated Potassium channels	0.0137	0.03119810
HDACs deacetylate histones	0.0218	0.04701088

Re-sampling –Intersect Size

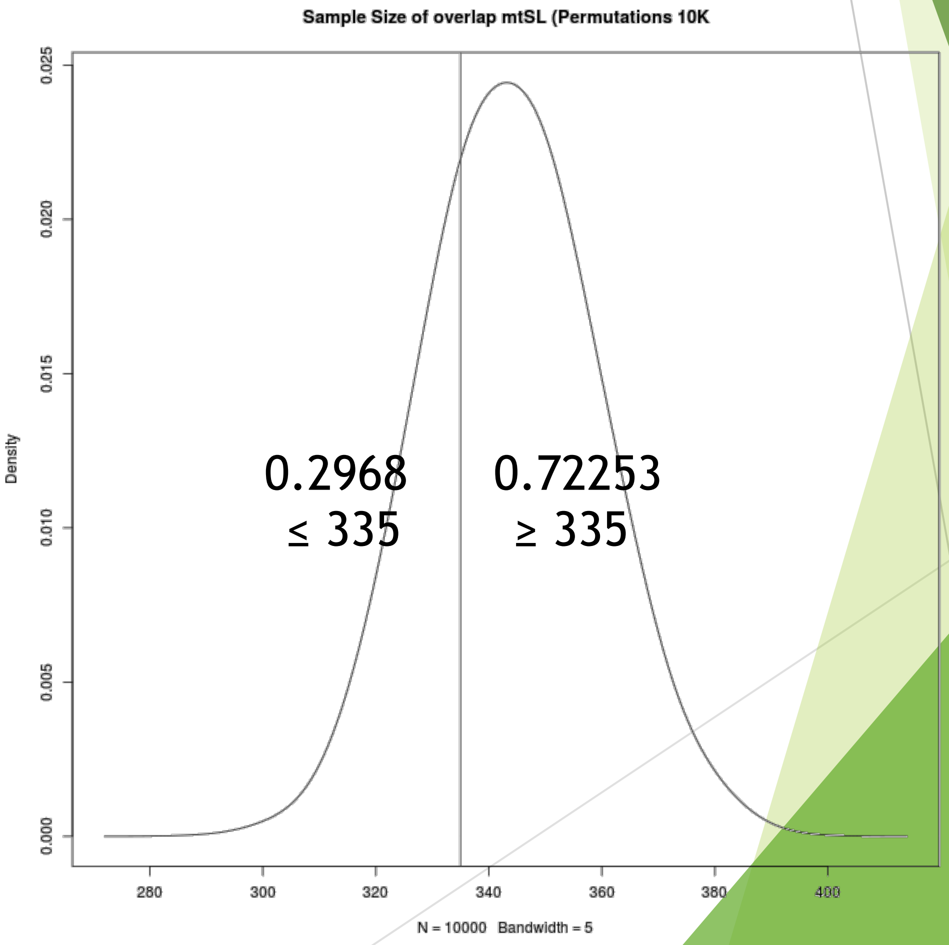
SLIPT



0.0524
≤ 450

0.9533
≥ 450

SLIPT + siRNA

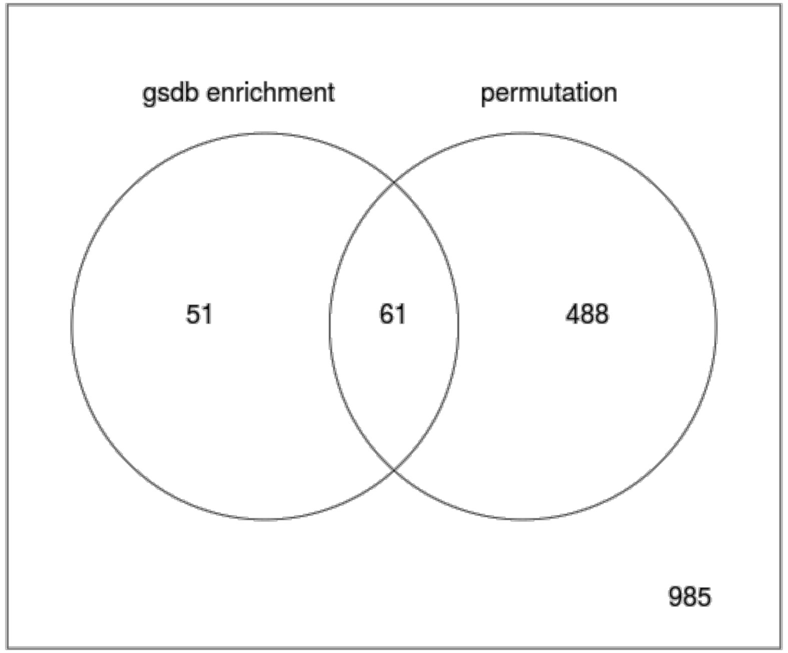


0.2968
≤ 335

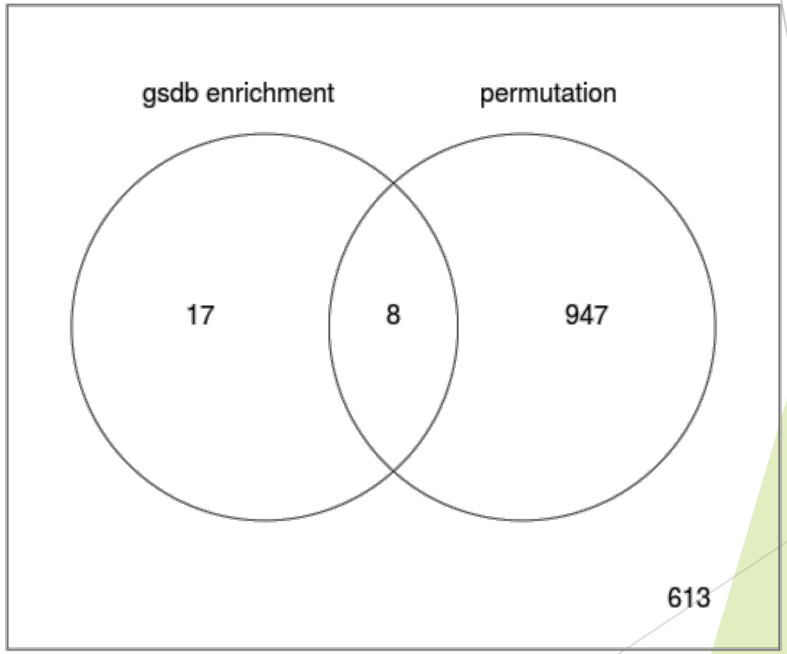
0.72253
≥ 335

Resampling – Compare to enrichment

SLIPT



SLIPT + siRNA



Resampling – Compare to enrichment

SLIPT

Reactome pathway	gsdb(fdr)	emp(fdr)
Eukaryotic Translation Elongation	2.10E-37	<0.0005
Influenza Viral RNA Transcription and Replication	6.80E-28	<0.0005
L13a-mediated translational silencing of Ceruloplasmin expression	2.20E-27	<0.0005
3' -UTR-mediated translational regulation	2.20E-27	<0.0005
Cap-dependent Translation Initiation	1.10E-23	<0.0005
SRP-dependent cotranslational protein targeting to membrane	3.20E-23	<0.0005
Translation	3.40E-19	<0.0005
Influenza Infection	4.50E-17	<0.0005
Interferon gamma signaling	4.90E-07	0.025004
Generation of second messenger molecules	9.50E-06	0.036759
GPCR ligand binding	1.90E-05	0.010256
Class A/1 (Rhodopsin-like receptors)	0.00017	0.004013
Integrin cell surface interactions	0.014	0.033305
Rho GTPase cycle	0.05	0.032987
Interferon Signaling	0.14	<0.0005
Innate Immune System	0.2	0.008019
Activation of G protein gated Potassium channels	0.25	0.045067
G protein gated Potassium channels	0.25	0.045067
PI3K Cascade	1	<0.0005
Cell Cycle	1	<0.0005
ERK/MAPK targets	1	<0.0005

SLIPT + siRNA

Reactome pathway	gsdb(fdr)	emp(fdr)
Eukaryotic Translation Elongation	1.20E-23	<0.00025
L13a-mediated translational silencing of Ceruloplasmin expression	1.30E-17	<0.00025
3' -UTR-mediated translational regulation	1.30E-17	<0.00025
Influenza Viral RNA Transcription and Replication	1.30E-17	<0.00025
SRP-dependent cotranslational protein targeting to membrane	4.20E-16	<0.00025
Cap-dependent Translation Initiation	1.20E-15	<0.00025
Translation	2.00E-12	<0.00025
Influenza Infection	1.80E-10	<0.00025
Regulation of Complement cascade	0.093	0.021758
Signaling by NOTCH3	0.14	0.027369
P2Y receptors	0.14	0.018276
G alpha (s) signalling events	0.19	0.004417
HIV Infection	1	<0.00025
Cell Cycle	1	<0.00025
DNA Replication Pre-Initiation	1	<0.00025
Cell Cycle, Mitotic	1	<0.00025
Synthesis of DNA	1	0.004417
Chromosome Maintenance	1	0.006534
Regulatory RNA pathways	1	0.011778
APC/C-mediated degradation of cell cycle proteins	1	0.025554
Apoptosis	1	0.041569

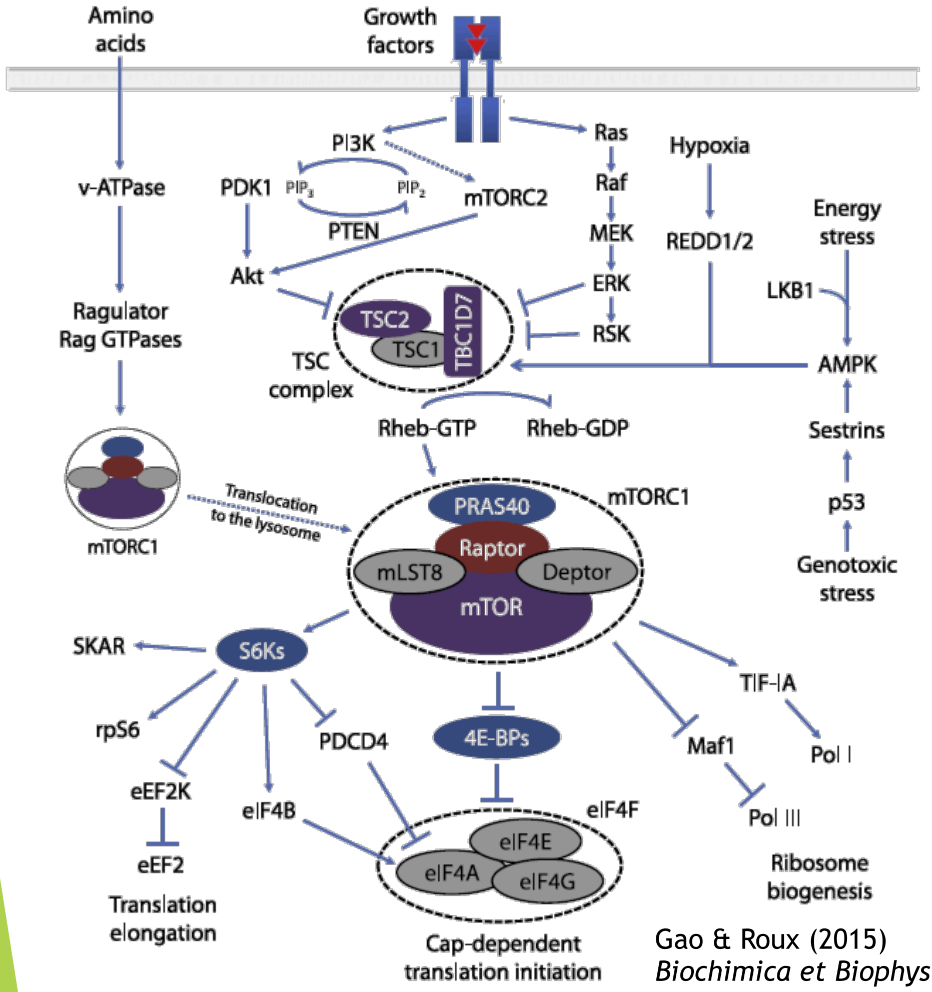
Discussion – Computational Challenges

- Each re-sample is independent
  - Simple to compute in embarrassingly parallel with Rmpi (snow R package)
- The methodology leads to a trade-off
  - Compute enrichment for every pathway for each re-sample (memory intensive)
  - Re-sample for testing one pathway many times, then do the next one… (CPU-time intensive)
- NeSI has enabled many more iterations (generating more accurate p-value estimates)
  - Especially important when multiple testing
  - Would not have been feasible to test every pathway without access to HPC
  - Simple to scale up iterations or cores
    - 10,000 Reps takes ~100min on 72 cores, 6Gb/core

Discussion – Biological Interpretations

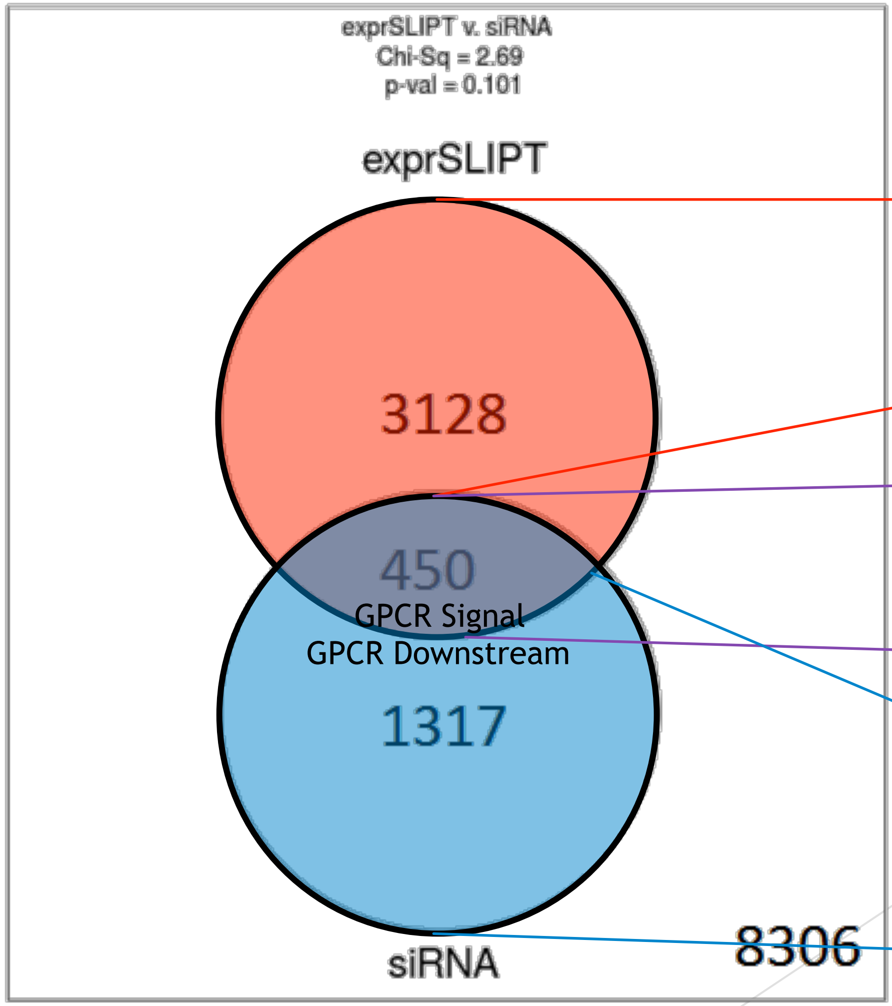
- Screening for SL needs to unexpected results in previous studies
  - Within-pathway SL
  - Between-pathway SL
  - Many molecules have unknown function or multiple functions
- Experimental screens and Bioinformatics analysis won’t detect the same genes
  - Some genes are easier to knockout in cell models (without killing all cells)
  - Genetic variation and tissue environment (e.g., immune) not tested in cell lines
- We need to understand the cell at a functional level for studying cancer
  - Many systems are dysregulated in cancer
  - Cancer cells re-wire as they develop and acquire drug resistance

Discussion – Biological Context



Gao & Roux (2015)
Biochimica et Biophysica Acta 1849: 753-765

Translation: Gene Expression and Cell Growth
Too high = cancer; Too low = cell death



3126

Translation, 3’ UTR regulation, non-sense mediated decay, SRP-dependent co-translate, Immune, Cell Cycle, Chromatin Modifiers

450

Nicotinamide, Serotonin, GPCR Signal, Gα(s), GPCR Downstream, Innate Immune, GPCR ligand, Cytochrome P_{450}, GPCR (B/2)

1317

CaMK/CREB, Protein Metabolism, 3’ UTR Regulation, non-sense mediated decay, Translation, Phospholipase C, Calmodulin

Discussion – Clinical Relevance

- Applications in cancer medicine
  - Targeted therapy against difficult molecular drivers of cancer
    - Inactivated
    - Similar to healthy (wildtype) variants
  - Chemoprevention / HDGC
    - Lower side effects would enable use against early stage cancer
    - Including preventative use in hereditary cancers before they're detected in clinic
  - Biomarkers
    - Clinical decisions based on molecular/genomic data
    - Anticipate drug resistance signatures and combination therapy (higher order interactions)
- Precision / Personalised / Genomics medicine / buzzword of the year

Discussion – Statistical Analysis

- Conservative analysis: corrected for multiple tests (false discovery rate)
  - Pathways or genes are not always independent
- Needs validation and function testing before clinical application
  - Cell line or mouse model
- Potentially vastly more effective / cheaper than experimental screens alone
  - If used in combination to select drug candidates
- Biologically consistent findings across pathways are promising
- Results support findings in experimental studies

Future Directions

- Technical
  - Refined prediction methods
  - Simulations and modelling
  - Include other data types or known pathway structure
- Biological
  - Mechanisms (molecular or cellular level)
  - Drug target triage and pre-clinical drug development
  - Combinations of mutations (e.g, CDH1, TP53, & PIK3CA)

Conclusions

- SL predictions across the human genome are valuable for cancer biologists
- Pathway predictions and candidate drug targets against *CDH1* in cancer have been found
  - Continues to inform experimental studies and drug development
- NeSI has enabled much of this work, particularly scaling up to genomics analysis and permutation re-sampling
  - Has led to statistical techniques and biological research questions not otherwise possible
- Demonstrates genomics data is a resource for biologists
  - Plenty of unexplored potential
  - Requires training next generation of researchers to utilise it
  - We need to work together (interdisciplinary skills)

Acknowledgements

- Supervisors: Mik Black & Parry Guilford
- Advisory committee: Anita Dunbier & Michael Lee
- Experimental data and advice: Cancer Genetics Lab, Bryony Telford, Augustine Chen
- Helpful discussion, advice, tech support, and proofreading: Mik’s group, collaborators, and an amazing number of people at conferences, on the web, or social media
- For making this project possible: data sources, software sources, patients, clinicians, the open science movement, and the StackOverflow/StackExchange community
- Funding source: University of Otago Postgraduate Tassel Scholarship in Cancer Research
- Compute resources: New Zealand eScience Infrastructure (NeSI) and Biochemistry Dept
- Conference funding: REANNZ, NeSI, NZGL