University of Cincinnati

Date: 3/11/2021

I, Shana White, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Biostatistics (Environmental Health).

It is entitled:

Application and Development of Novel Methods for Pathway Analysis and Visualization of the LINCS L1000 Dataset

Student's name: **Shana White**

This work and its defense approved by:

Committee chair: Mario Medvedovic, Ph.D.

Committee member: Marepalli Rao, Ph.D.

Committee member: John Reichard, PharmD, Ph.D.

Committee member: Heidi Sucharew, Ph.D.

38157

Application and Development of Novel Methods for Pathway Analysis and Visualization of the LINCS L1000 Dataset

A dissertation submitted to the Graduate School
of the University of Cincinnati in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the Division of Biostatistics and Bioinformatics of the Department of Environmental Health
of the College of Medicine

by
Shana White

Committee Chair: Mario Medvedovic, Ph.D.

Abstract

The LINCS L1000 dataset is a large-scale compendium that contains records of the cell line specific transcriptional effects of cellular perturbation that was established to provide mechanistic and circuit-level insights with regard to cancer biology. This undertaking is a scaled-up version of the Connectivity Map (CMap) project whose goal was to connect transcriptional signatures of the downstream effects of genetic and small-molecule perturbations in a high-throughput yet cost-effective manner. This was accomplished by profiling a reduced representation of the human transcriptome – nearly 1,000 landmark transcripts whose expression is predictive of roughly 80% of non-measured genes.

Whereas the choice to measure a subset of the transcriptome was primarily cost-based, reducing the representation of transcriptional data is a common method for amplifying the signal amidst the noisy background of large datasets. It can also be a valuable tool for making data amenable to a variety of bioinformatics-based analyses, for example, when lists of genes and their direction of regulation is considered based on continuously valued measurements subjected to a significance-based threshold. In the work presented in this document, we subject the records contained in the L1000 dataset to a thresholding procedure and explore how connections between over 2,000 common genetic perturbations differ between a core set of seven cancer cell lines. Specifically, we frame the connections in the context of edges between nodes in a novel adaptation of pathway-level analysis.

We begin by conducting a simulation study in order to interrogate the data-generating mechanism best suited to reproduce our data of interest with the least amount of bias. This will be followed by a power analysis to assess the appropriate threshold for edge-based

measurements for our dataset. Then, we will demonstrate how these measurements can be incorporated into the topology of cellular signaling pathways and introduce an R Bioconductor package that easily integrates this type of data into pathways from the Kyoto Encyclopedia of Genes and Genomes or KEGG – one of the most widely known online repositories for biological pathways. Finally, we will conduct an edge set enrichment analysis of our data that applies the well-known methodology of gene set enrichment analysis to this novel edge-data type.

Acknowledgements

I would like to thank Dr. Mario Medvedovic for serving as my committee chair and offering me up challenges that ultimately guided me through this process. I would also like to thank Drs. MB Rao, John Reichard, and Heidi Sucharew for serving as members of my committee, devoting their time to numerous meetings and offering me their wisdom and encouragement. I would also like to thank Drs. Ranjan Deka and Kim Dietrich, my mentors from the Molecular Epidemiology in Children's Environmental Health fellowship that supported me for much of my tenure as a graduate student.

Dedication

This work is dedicated to my daughter Eliana White – may your curiosity lead you on many quests for many answers.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 LINCS L1000 data set

The Library of Integrated Network-Based Cellular Signatures (LINCS) consortium is an academic community of researchers supported by the NIH (National Institutes of Health) established to take on the massive project of "generating and making public data that indicates how cells respond to various *genetic* and *environmental* stressors" [1]. The overarching goal of this initiative is to establish cause-and-effect biological insight by measuring the downstream transcriptional response of genes in specific cell lines after "perturbating the [cellular] system" with shRNA interference (genetic) or chemical/pharmacological (environmental) stressors [2]. The Connectivity Map (CMap) project established by the BROAD Institute of MIT and Harvard has undertaken the ongoing task of generating and making public data obtained via the LINCS L1000 platform [2] [3].

The CMap L1000 database (from here on referred to as the L1000 data set) is a collection of gene expression signatures obtained by a high-throughput screening method developed by CMap called the L1000 assay. The "L" in L1000 stands for landmark (LM); rather than targeting the whole genome, the L1000 gene expression assay contains 1,058 probes corresponding to 978 "landmark" genes and 80 control transcripts. The 978 L1000 genes provide a reduced representation of the entire transcriptome and were chosen by CMap to be targeted for changes in expression for the following reasons [4]:

1. These genes are widely expressed/transcribed across the cancer cell lineages of interest at baseline conditions (no perturbating factors).

2. The expression of 11,350 genes that are not measured in the assay can be reliably predicted via linear regression.

The unique records curated by CMap as part of the L1000 data set are called *Consensus Genomic Signatures* (CGSs). Each CGS is the 'consensus' measurement of the expression change of L1000 genes after cancer cells undergo perturbation by either genetic "perturbagens" (genes known to be important in 'upstream' transcriptional regulation are targeted via shRNA interference – an experimental procedure intended to functionally knock-out the transcription-related-abilities of individual genes) at specific timepoints [after perturbation] and doses of the perturbagens. The term 'consensus' indicates that this data represents the on-target effects multiple of shRNAs targeting the same gene.

The L1000 data set is large and dynamic - CMap adds records to the database of over 1,000,000 L1000 profiles as data is generated. Providing reduced representation of data is an important aspect of the generation and analysis of 'big data' sets such as L1000; it has important applications not only in terms of high-throughput screening (i.e. reducing monetary cost of data collection by reducing how much data is collected) but also for reducing the noise [and thereby amplifying the signal] in the resulting data. There are two aspects reduced representation inherent in the creation of 'Level 6' CGSs. First, although records exist across multiple shRNAs, Level 6 CGSs contain one record per perturbagen at each experimental condition which makes the data more manageable from a practical standpoint. The 'consensus' signatures provide a clearer signal of transcriptional outcomes among the genes by 'averaging out' off-target shRNA effects on the transcription of LM genes across experimental replicates [5].

The notion that the L1000 genes are, in effect, the opposite of housekeeping genes is the second aspect of dimension reduction attributed to the L1000 data set. Whereas 'housekeeping'

genes are chosen for their relatively stable expression across experimental conditions, the L1000 assay measures 978 genes were chosen to provide a clear signal of changes in *transcriptionally informative* genes [6]. The discussion that follows posits the following questions: can we define a metric to describe the transcriptional information associated with a particular L1000 gene and how can the heterogeneity of transcriptional information be incorporated into the analysis of the L1000 data set?

1.2 Level 6 Consensus Genomic Signature Gene Lists

A ModZ *scores* for *all* 978 LM genes

cell_id	perturbagen	AARS_ModZ	ABCB6_ModZ	ABCC5_ModZ	ZNF586_ModZ	ZNF589_ModZ	ZW10_ModZ
A375	A2M	−0.085	0.390	−0.653	1.671	0.664	−0.661
A549	A2M	0.622	0.146	−1.447	0.219	3.073	−2.323
HA1E	A2M	−0.045	0.143	−0.368	0.346	1.642	1.451
HEPG2	A2M	−0.061	−0.646	−1.663	−0.209	−0.648	0.759
HT29	A2M	−0.451	−0.049	0.345	2.268	−0.054	−0.964
MCF7	A2M	0.839	−1.732	−2.038	0.116	−0.298	0.093
PC3	A2M	0.210	−1.595	−2.318	0.319	0.087	−1.629
A375	AARS	0.038	−1.750	−0.278	2.452	0.894	−1.063
A549	AARS	−1.209	−0.811	−1.472	−0.244	−0.517	0.534
HA1E	AARS	0.962	0.168	1.454	1.289	−1.426	1.567
HEPG2	AARS	0.840	−0.209	−1.194	0.651	−1.747	−0.444
HT29	AARS	0.428	0.048	−2.725	0.013	−0.945	−0.878
MCF7	AARS	1.278	0.793	−1.757	−1.211	−0.541	−1.592
PC3	AARS	0.468	0.068	0.073	−0.320	−1.099	0.865
A375	AATF	0.459	−0.279	−0.301	0.192	−0.790	−0.005
A549	AATF	−1.115	−0.641	−1.129	−0.116	−0.259	−0.533
HA1E	AATF	0.899	−1.845	−0.984	−0.669	1.451	−0.227
HEPG2	AATF	0.025	0.897	−0.496	0.152	0.280	−0.338
HT29	AATF	−0.178	0.855	−1.984	1.263	2.568	0.269
MCF7	AATF	−1.292	−1.593	−1.518	−1.236	−0.314	−1.233
PC3	AATF	0.521	−0.213	0.357	0.414	1.107	−0.996

B *Lists* of the *50 top/bottom ranked* LM genes

cell_id	perturbagen	UP	DOWN
A375	A2M	c("NOS3", "TBP", "ENOSF1", "OXSR1", "GHR", "RAD9A", ...	c("SLC2A6", "POLR1C", "ARFIP2", "ATP2C1", "ARNT2", "...
A549	A2M	c("BRCA1", "EED", "EML3", "STAMBP", "ETFB", "DNAJB1",...	c("TESK1", "ATP2C1", "DFFA", "CAMSAP2", "ARNT2", "...
HA1E	A2M	c("EGF", "WIPF2", "HN1L", "PHKB", "HDAC2", "KIAA010...	c("TESK1", "MLEC", "CAMSAP2", "ARNT2", "PLA2G4A", "...
HEPG2	A2M	c("HLA-DMA", "KDM5B", "AKT1", "RRP8", "HN1L", "KIA...	c("FOXO3", "SERPINE1", "RPN1", "PPP2R5A", "TESK1", "...
HT29	A2M	c("CISD1", "COG4", "RRP8", "EML3", "BTK", "LIG1", "MA...	c("IGF1R", "SLC2A6", "POLR2K", "PIK3C3", "PPP2R5A", "...
MCF7	A2M	c("CISD1", "HLA-DMA", "NOS3", "WIPF2", "ZNF274", "C...	c("SPAG7", "SYPL1", "KTN1", "CDK4", "MYCBP", "ATP1B...
PC3	A2M	c("MCM3", "FAH", "ELOVL6", "USP14", "EIF4G1", "MRPL...	c("RPN1", "PPP2R5A", "TMED10", "ATP2C1", "EDN1", "C...
A375	AARS	c("FOXO3", "SERPINE1", "SNX13", "EPRS", "PTGS2", "MA...	c("ATF1", "APP", "SLC2A6", "PPP2R5A", "RRAGA", "TME...
A549	AARS	c("BIRC5", "NFKBIB", "DNAJB1", "GNB5", "PRKCD", "ZNF...	c("SENP6", "CBLB", "PDS5A", "ECH1", "CAMSAP2", "PLA...
HA1E	AARS	c("FOXO3", "SERPINE1", "TSKU", "SLC2A6", "GABPB1", "...	c("POLR1C", "FAH", "CAMSAP2", "ARNT2", "MSH6", "PL...
HEPG2	AARS	c("FOXO3", "SPTLC2", "AKT1", "SOX4", "ELOVL6", "TWF...	c("APP", "POLR2K", "SENP6", "NMT1", "PIK3CA", "ECH1"...
HT29	AARS	c("CBLB", "USP14", "CPSF4", "APOE", "NRAS", "MTHFD2"...	c("RHEB", "RHOA", "PSMD4", "BDH1", "SOX4", "CASC3",...
MCF7	AARS	c("NFATC4", "TBP", "GABPB1", "EBP", "BHLHE40", "CLIC...	c("RHEB", "APP", "PSMD4", "DPH2", "RUVBL1", "PIK3C3",...
PC3	AARS	c("PSME1", "SPTLC2", "TMEM2", "EZH2", "EPRS", "MBTP...	c("NFKBIB", "USP14", "SCARB1", "PPARD", "COASY", "SL...
A375	AATF	c("IL18", "EPRS", "HN1L", "TICAM1", "PTGS2", "PAK6", "...	c("USP22", "TRAPPC6A", "MAT2A", "KLHDC2", "GADD4...
A549	AATF	c("GABPB1", "PTK2B", "PAX8", "PMAIP1", "PXN", "PIK3C...	c("PSME1", "BRCA1", "PSMD4", "SNX6", "RPN1", "XBP1",...
HA1E	AATF	c("NFATC4", "SOX4", "SPAG7", "PTGS2", "STX1A", "CEB...	c("PSME1", "TMEM2", "BRCA1", "EED", "NENF", "UBE2C",...
HEPG2	AATF	c("STX1A", "PNP", "GNB5", "LPAR2", "SMNDC1", "LSM6",...	c("FOXO3", "RPN1", "HSPB1", "PHKB", "MRPL19", "ETFB"...
HT29	AATF	c("SERPINE1", "ETV1", "GABPB1", "CASK", "SOX4", "PXN"...	c("SNX6", "APPBP2", "MRPL19", "CDK2", "WRB", "CASP3...
MCF7	AATF	c("WIPF2", "GABPB1", "CASK", "RRP8", "PXN", "PIK3CA", ...	c("CISD1", "IGF1R", "SNX6", "RUVBL1", "BDH1", "PPP2R...
PC3	AATF	c("TBP", "NENF", "POLR1C", "RRP8", "HDAC6", "ZNF274...	c("PSME1", "SPTLC2", "SLC2A6", "PSMD4", "SNX6", "EM...

Figure 1: Comparison of data for Level 6 CGSs ModZ (**A**) vs. Level 6 CGSs gene list (**B**) CGSs. Each Level 6 CGS (**A**) is a vector of ModZ scores for each L1000 LM gene whereas level 6 signatures are lists of LM genes that are top 50 overexpressed ("UP") or underexpressed ("DOWN") for each experimental perturbagen for each cell line.

The Level 6 CGSs in the L1000 data set are vectors of ModZ scores for all LM genes. Methods that employ ModZ CGSs for their intended purpose – to connect cellular events using measurements of their transcriptional response to stressors in the context of multiple layers of dimension reduction – have been the topic of many recently published papers both within the LINCS and among other research communities [7] [8] [9]. The discussion that follows considers

Level 6 CGS gene lists – Level 6 data that has a further reduced representation of the information measured by the L1000 genes. This novel data type - specifically provided to the LINCS community - has received little attention despite its [intended] potential to reduce the internal noise amongst expression of the L1000 genes themselves.

Here we define Level 6 CGS gene lists associated with each CGS that summarize the direction of regulatory changes of 'important' LM genes according to the CGS's internally most differentially expressed (MDE) genes. Each Level 6 CGS gene list is associated with two non-overlapping subsets: an "UP" (most up-regulated or overexpressed genes) and "DOWN" (most down-regulated or underexpressed genes) list. **Figure 1** demonstrates the differences in data structure between the original Level 6 CGSs and Level 6 CGS gene lists.

CMap has offered Level 6 CGSs defined by simple selection criteria for the two subsets of genes: "UP" lists contain the L1000 genes with the top 50 largest positive ModZ-scores and entries for the "DOWN" lists have the 50 most negative ModZ-scores for each record of Level 5 data. The 878 (978 – 50×2) LM genes that do not make either cutoff (fall in the middle) are essentially filtered out of the new Level 6 CGS. Unlike the Level 6 ModZ CGSs, Level 6 CGSs signatures do not contain expression values for each L1000 gene. Note that the directionality of a gene's expression change is only captured in a Level 6 CGS if it is among the 100 MDE genes. CMap has also generated Level 6 CGSs that first use algorithms to predict the expression of genes and select the top 100 MDE genes in both directions for either a 'best inferred gene set' ("BING_100") or for all genes across the genome ("ALL_100"). Our focus will remain on the Level 6 CGSs for the top 50 most up and down-regulated landmark genes ("LM_50") to first address the following question before using these signatures in downstream analyses; does reduced representation effectively boost the signal in our data or does the behavior of certain LM

genes overshadow the more muted, but perhaps more biologically or otherwise important transcriptional response of genes that do not make the cutoff? We will conduct simulation studies aimed to determine, in a controlled setting, the extent to which random noise 'clutters' the Level 6 signatures.

Table 1: Cell lines used for analysis

Cell Line Name	Tissue/Disease of Origin
A375	Amelanotic melanoma (skin)
A549	Lung adenocarcinoma
HA1E	Immortalized kidney epithelium
HEPG2	Hepatoblastoma (liver)
HT29	Colon adenocarcinoma
MCF7	Invasive ductal carcinoma (breast)
PC3	Prostate carcinoma

1.3 Level 6 Concordance Signatures

Lists of gene names are used in many applications of bioinformatics research, perhaps most notably as queries for enrichment analysis of gene expression data [10]. Typically, lists contain the names of genes that are *either* over or under-expressed between two phenotypic conditions (such as disease state relative to control) or across other experimental settings such as time. Instead of using the Level 6 CGSs as queries for *external* data sets, the purpose of this study is to use *internally*-derived *concordance signature*s to identify similarities/differences in gene regulation amongst 6 cancer cell lines as well as 1 immortalized cell line (**Table 1**) that each have a Level 6 CGS recorded at 96 hours across 2,042 common genetic perturbations (concentration = 1μl).

Here we define concordance signatures (CS's) between any two Level 6 CGS as the gene lists that summarize the intersection of their impact on regulatory events as follows:

For any two Level 6 CGS gene lists, let (CGS_X^{cl}, CGS_Y^{cl}), uniquely identified by one cell line (superscript; one of seven different cancer cell lines (CL's)) and two perturbated genes (subscript; two different perturbed genes (PGs) out of 2,042) let:

$CGS\_U_X^{cl}$ = 50 "UP" genes for $PG_x \times CL^{cl}$, $CGS\_U_Y^{cl}$ = 50 "UP" genes for $PG_y \times CL^{cl}$,

$CGS\_D_X^{cl}$ = 50 "DOWN" genes for $PG_x \times CL^{cl}$, $CGS\_D_Y^{cl}$= 50 "DOWN" genes for $PG_y \times CL^{cl}$.

While concordance signatures could be constructed between any two CGS gene lists, we will specifically focus on relationships that exist as edges between nodes in cellular signaling networks, as discussed in detail in the following section.

Then, the concordance signature CS_{XY}^{cl} (edge X|Y) is defined by four following sets of gene lists:

$$CS\_UU_{XY}{}^{cl} = CGS\_U_X^{cl} \cap CGS\_U_Y^{cl};\ 0 \leq \left|CS\_UU_{XY}{}^{cl}\right| \leq 50,$$

$$CS\_DD_{XY}{}^{cl} = CGS\_D_X^{cl} \cap CGS\_D_Y^{cl};\ 0 \leq \left|CS\_DD_{XY}{}^{cl}\right| \leq 50,$$

$$CS\_UD_{XY}{}^{cl} = CGS\_U_X^{cl} \cap CGS\_D_Y^{cl};\ 0 \leq \left|CS\_UD_{XY}{}^{cl}\right| \leq 50,$$

$$CS\_DU_{XY}{}^{cl} = CGS\_D_X^{cl} \cap CGS\_U_Y^{cl};\ 0 \leq \left|CS\_DU_{XY}{}^{cl}\right| \leq 50, \text{ and}$$

$$CS_{XY} = CS\_UU_{XY}{}^{cl} \cup CS_{DD\,XY}{}^{cl} \cup CS_{UD\,XY}{}^{cl} \cup CS_{DU\,XY}{}^{cl}.$$

Note that $0 \leq |CS_{XY}| \leq 100$.

Table 2: Cross-table representation of a Level 6 Concordance Signature
The four subsets of a Level 6 CS contain are themselves subsets of overlapping elements (LM genes) derived from two different Level 6 CGSs (PGs X and Y) from the same cell line (CL).

CL = *cl*		PG = X	
		Up = $CGS\_U_X^{cl}$	Down= $CGS\_D_X^{cl}$
PG = Y	Up = $CGS\_U_Y^{cl}$	$CS\_UU_{XY}{}^{cl}$ (a)	$CS\_DU_{XY}{}^{cl}$ (b)
	Down = $CGS\_D_Y^{cl}$	$CS\_UD_{XY}{}^{cl}$ (c)	$CS\_DD_{XY}{}^{cl}$ (d)

Table 2 displays these subsets in a 2×2 cross-classification table. The genes that fall in cells (a) and (d) are in *concordant* sets; if they fall within cell (a) they are among the top 50 upregulated genes for both PG_x and PG_y and if they fall into cell (d) they are among the 50 most down-regulated genes for those two perturbagens. The genes that land in cells (b) and (c) are in *discordant* sets; in cell (b) they are among the 50 most down-regulated genes for PG_x but among the 50 most upregulated genes for PG_y and vice versa for the genes in cell (c).

1.4 Proposed Analysis for Level 6 Concordance Signatures

At this point, we have formally defined Level 6 CS's but the question remains, does the decision to include only the most transcriptionally responsive genes in our summary measurements leave us with enough information to make biologically meaningful comparisons in the data set? Furthermore, can we derive statistics from these signatures that allow us to compare evidence for relationships between perturbed genes within and between cell lines based on summaries of their similarities (or differences) of downstream effects that are obtained from a reduced representation of the available data? Specifically, we will evaluate the relationships between perturbed genes that exist as edges according to the cellular-signaling pathways curated by the Kyoto Encyclopedia of Genes and Genomes (KEGG) [11].

The edges in KEGG pathways represent relationships between genes or gene products (ie. proteins) and although specific relationships have been verified by experimental results, the extent to which they translate across different types of cellular systems is not readily measured [12]. There are two main types of relationships defined in KEGG - activation and repression – but relationships such as binding, dissociation, expression and post-translational modification (either inhibiting or activating) are also present in the pathways. We will attempt to quantify heterogeneity in signaling patterns at the pathway level with regard to different cellular

phenotypes through the incorporation of LINCS L1000 data into KEGG pathway topology. The benefit of using LINCS L1000 concordance measurement data is that we can define a metric for relationships between genes when that relationship between two specific genes is not expression-based. For example, in the mTOR signaling pathway (**Figure 2**), the mTOR (mammalian target of rapamycin) protein forms two different types of multi-protein complexes that regulate protein

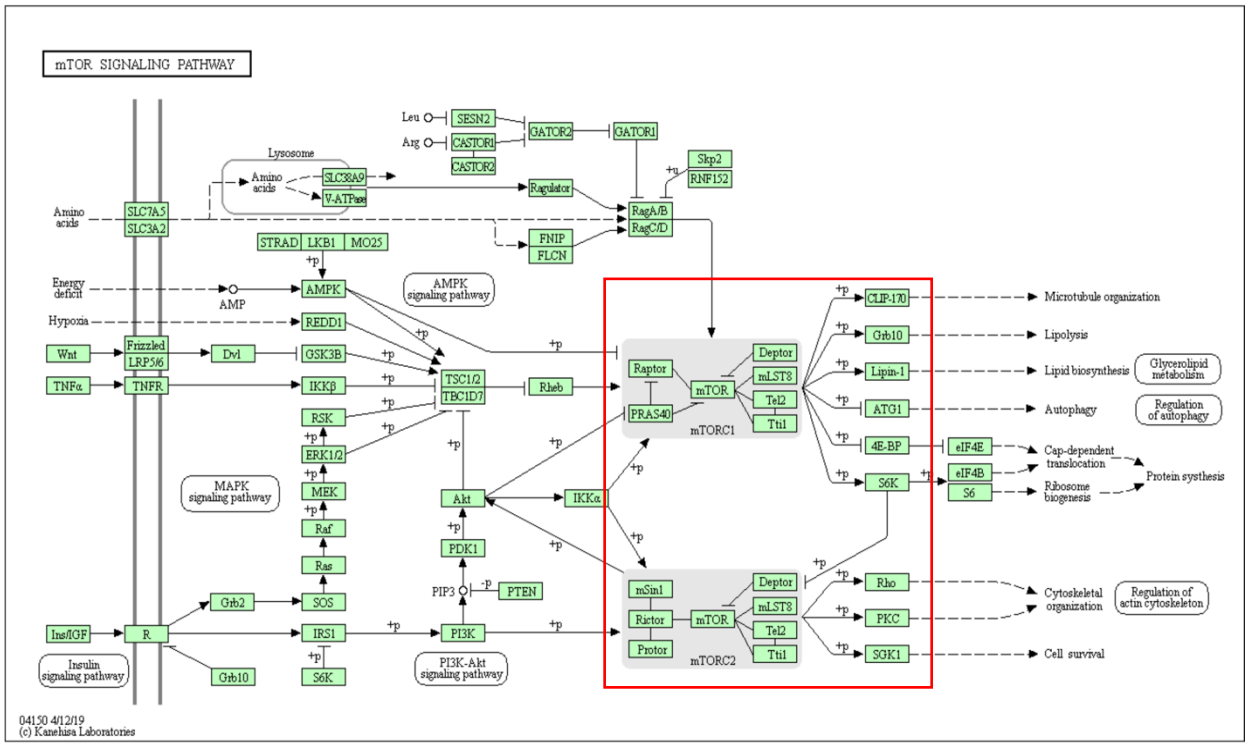
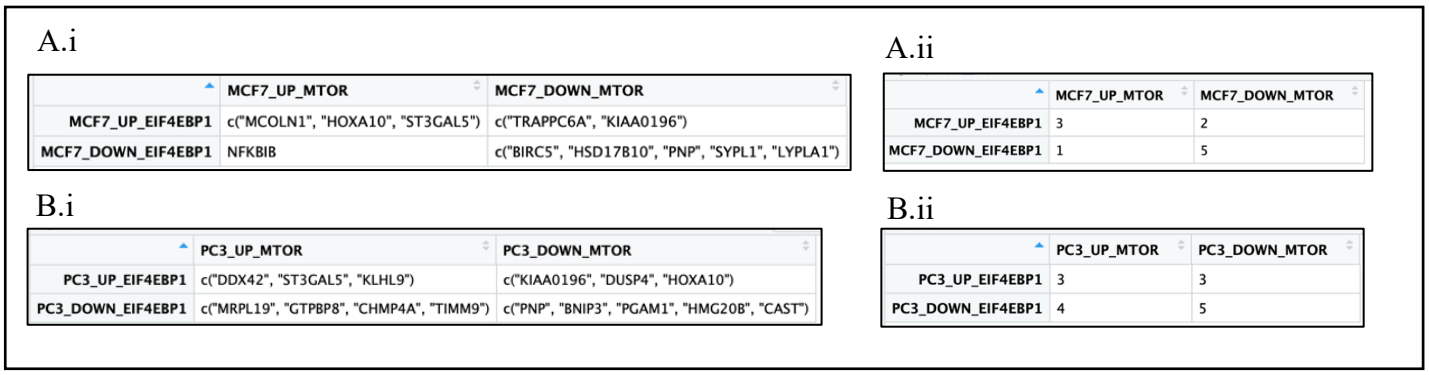


Figure 2: mTOR Signaling Pathway from KEGG. The mTOR protein complexes are upstream activators/inhibitors of proteins that impact gene expression (red box).

synthesis by interacting with intermediary proteins that directly impact gene expression.

The relationships between the mTOR complexes and its direct targets rely on the kinase activity of mTOR (activating or inhibitory phosphorylation), therefore we would not expect a change in the rate of mTOR transcription to result in transcriptional changes in its targets (increase or decrease of gene expression). On the other hand, since the mTOR complex regulates the activity of its targets, the disruption of mTOR's activity could impact downstream



A.i

	MCF7_UP_MTOR	MCF7_DOWN_MTOR
MCF7_UP_EIF4EBP1	c("MCOLN1", "HOXA10", "ST3GAL5")	c("TRAPPC6A", "KIAA0196")
MCF7_DOWN_EIF4EBP1	NFKBIB	c("BIRC5", "HSD17B10", "PNP", "SYPL1", "LYPLA1")

A.ii

	MCF7_UP_MTOR	MCF7_DOWN_MTOR
MCF7_UP_EIF4EBP1	3	2
MCF7_DOWN_EIF4EBP1	1	5

B.i

	PC3_UP_MTOR	PC3_DOWN_MTOR
PC3_UP_EIF4EBP1	c("DDX42", "ST3GAL5", "KLHL9")	c("KIAA0196", "DUSP4", "HOXA10")
PC3_DOWN_EIF4EBP1	c("MRPL19", "GTPBP8", "CHMP4A", "TIMM9")	c("PNP", "BNIP3", "PGAM1", "HMG20B", "CAST")

B.ii

	PC3_UP_MTOR	PC3_DOWN_MTOR
PC3_UP_EIF4EBP1	3	3
PC3_DOWN_EIF4EBP1	4	5

Figure 3: Example of Level 6 CS for 2 different cell lines. The cross-table views comparing Level 6 CS's (perturbagens = MTOR, EIF4EBP1) between two different cell lines (MCF7 [breast cancer], PC3 [prostate cancer]). The tables in A.i and B.i maintain the lists of genes in each subset whereas those in A.ii and B.ii summarize the counts found in each quadrant.

gene expression in a way that would mirror the disruption of one of its direct targets. As a specific example, mTORC1 (mTOR complex 1) inhibits the activity of the translational repressor 4E-BP (initiation 4E binding protein), thereby initiating the activity of eukaryotic translation initiation factor 4E (the inhibition of an inhibitor allows the downstream target to function) [13]. Thus, although mTOR does not directly impact the expression of *Eif4ebp1* (the gene that codes for 4E-BP), the relationship between mTOR and 4E-BP could be quantified by metrics that compare the similarities of the downstream impacts when either gene is functionally knocked down via shRNA as showcased for cell lines MCF7 and PC3 in **Figure 3.**

When the subsets of Level 6 CGSs are arranged as shown in **Table 2**, the data may be maintained as it is in its original form as a list (**Figure 3(A.i), 4(B.i**)) or by further summarizing as a count of the elements contained in each cell (**Figure 3(A.ii), 4(B.ii**)). When the data is maintained in this format, a different measurement of similarity can be calculated by counting the intersection of units among the cells. In this case, the genes "ST3GAL5", "KIAA0196" and "PNP" are in cells (a), (b) and (d) respectively for both cell lines. Formally, for two cell lines A

and B and perturbed genes X and Y. We define the "sum of overlaps" statistic $\sum overlaps_{XY}^{AB}$ as follows:

$$\sum overlaps_{XY}^{AB} =$$

$$\left|{CS_{UU}}_{XY}^{A} \cap {CS_{UU}}_{XY}^{B}\right| + \left|{CS_{DU}}_{XY}^{A} \cap {CS_{DU}}_{XY}^{B}\right| + \left|{CS_{UD}}_{XY}^{A} \cap {CS_{UD}}_{XY}^{B}\right| + \left|{CS_{DD}}_{XY}^{A} \cap {CS_{DD}}_{XY}^{B}\right| \quad (1)$$

In the example, $\sum overlaps_{MTOR,\mathrm{EIF4EBP1}}^{MCF7,PC3} = 3$ across the two tables. Is this value large enough or small enough to suggest significant similarity or dissimilarity between the two cell lines? The answer to this question would require a better understanding of how these values are distributed among random pairwise comparisons of CS's between perturbagens in the L1000 data set.

Data that is arranged in a 2×2 cross-classification or contingency table is amenable to the well-known odds ratio (OR) test statistic as a measurement of association between two 'treatment' variables according to binary outcomes. For a single table, in our case for a single cell line, two PGs have greater similarities in downstream effects when the OR is larger than one 1 (or equivalently a log(OR) > 0) as this means there is a higher ratio of genes that fall into the concordant cells (a) and (d) relative to the discordant cells (b) and (c). On the other hand, if the OR is between zero and 1 (or equivalently a log(OR) < 0) there are more differences in the direction of downstream gene expression than there are similarities. The procedure for deriving the OR test statistic is as follows (note that the value 0.5 is added to all of the cells of **Figure 3A.ii** and **Figure 3.B.ii** as a bias correction as recommended by Haldane [14]):

$$\hat{\theta}_{XY}^{A} = \frac{({CS_{UU}}_{XY}^{A} + 0.5) * ({CS_{DD}}_{XY}^{A} + 0.5)}{({CS_{UD}}_{XY}^{A} + 0.5) * ({CS_{DU}}_{XY}^{A} + 0.5)}, \quad \hat{\theta}_{XY}^{B} = \frac{({CS_{UU}}_{XY}^{B} + 0.5) * ({CS_{DD}}_{XY}^{B} + 0.5)}{({CS_{UD}}_{XY}^{B} + 0.5) * ({CS_{DU}}_{XY}^{B} + 0.5)} \quad (2)$$

$$SE^A_{XY} = \hat{\sigma}_{log\left(\hat{\theta}^A_{XY}\right)} = \sqrt{\frac{1}{\left(CS_{UU_{XY}}{}^A + 0.5\right)} + \frac{1}{\left(CS_{DD_{XY}}{}^A + 0.5\right)} + \frac{1}{\left(CS_{UD_{XY}}{}^A + 0.5\right)} + \frac{1}{\left(CS_{DU_{XY}}{}^A + 0.5\right)}},$$

$$SE^B_{XY} = \hat{\sigma}_{log\left(\hat{\theta}^A_{XY}\right)} = \sqrt{\frac{1}{\left(CS_{UU_{XY}}{}^B + 0.5\right)} + \frac{1}{\left(CS_{DD_{XY}}{}^B + 0.5\right)} + \frac{1}{\left(CS_{UD_{XY}}{}^B + 0.5\right)} + \frac{1}{\left(CS_{DU_{XY}}{}^B + 0.5\right)}}. \quad (3)$$

For a single cell line, we can measure the *relative* level of association/concordance under the following assumptions:

Null: assume of no association, $\theta^A_{0,XY} = 1 \rightarrow \log(\theta^A_{0,XY}) = 0$, thus H_0: $\hat{\theta}^A_{XY} = \theta^A_{0,XY} = 1$.

Under the null hypothesis the following test statistic has a standard normal distribution:

$$\hat{\delta}^A_{XY} = \frac{log\left(\hat{\theta}^A_{XY}\right) - log\left(\theta^A_{0,XY}\right)}{SE^A_{XY}} = \frac{log\left(\hat{\theta}^A_{XY}\right)}{SE^A_{XY}} \sim N(0,1) \quad (4)$$

Since we want to compare the relative association *between* two cell lines, we are more interested in the following test statistic, which also has a standard normal distribution when the data for A and B come from the same distribution:

$$\hat{\Delta}^{AB}_{XY} = \frac{log\left(\hat{\theta}^A_{XY}\right) - log\left(\hat{\theta}^B_{XY}\right)}{\sqrt{(SE^A_{XY})^2 + (SE^B_{XY})^2}} \sim N(0,1) \quad (5)$$

In our example,

$\hat{\theta}^{MCF7}_{MTOR,\text{EIF4EBP1}} = \frac{(3+0.5)*(5+0.5)}{(1+0.5)*(2+0.5)} = 5.1\overline{3}$ and $\hat{\theta}^{PC3}_{MTOR,\text{EIF4EBP1}} = \frac{(3+0.5)*(5+0.5)}{(4+0.5)*(3+0.5)} = 1.2\overline{2}$,

$SE^{MCF7}_{MTOR,\text{EIF4EBP1}} = \hat{\sigma}_{\log(\hat{\theta}^{MCF7}_{MTOR,\text{EIF4EBP1}})} = \sqrt{\frac{1}{(3+0.5)} + \frac{1}{(5+0.5)} + \frac{1}{(1+0.5)} + \frac{1}{(2+0.5)}} \approx 1.24$,

$SE^{PC3}_{MTOR,\text{EIF4EBP1}} = \hat{\sigma}_{\log(\hat{\theta}^{PC3}_{MTOR,\text{EIF4EBP1}})} = \sqrt{\frac{1}{(3+0.5)} + \frac{1}{(5+0.5)} + \frac{1}{(4+0.5)} + \frac{1}{(3+0.5)}} \approx 0.99$,

and $\hat{\Delta}^{MCF7,PC3}_{MTOR,\text{EIF4EBP1}} = \frac{log(\hat{\theta}^{A375}_{MTOR,\text{EIF4EBP1}}) - log(\hat{\theta}^{A549}_{MTOR,\text{EIF4EBP1}})}{\sqrt{\left(SE^{A375}_{MTOR,\text{EIF4EBP1}}\right)^2 + \left(SE^{A549}_{MTOR,\text{EIF4EBP1}}\right)^2}} = \frac{log(5.13) - log(1.22)}{\sqrt{(1.24)^2 + (0.99)^2}} = 0.91.$

Under the assumption that the test statistics for both cell lines are identically distributed, the delta value can be treated as a z-score and translates to a p-value of 0.18. Therefore, although there is evidence of a more concordant relationship between MTOR and EIF4EBP1 in the MCF7 cell line versus the PC3 cell line, the difference in concordance does not reach statistical significance even at the less restrictive $\alpha = 0.1$ level (ie. less restrictive than restrictive $\alpha = 0.05$).

However, this begs the question: is the data from different cell lines identically distributed or do genes in the L1000 data set have cell line specific behavior? For example, do some cell lines have genes that tend to be upregulated or downregulated across PGs leading to larger or smaller measurements of similarity on average compared to other cell lines? If so, what statistical methods can we implement that would allow us to control for differences between cell lines and possibly reconcile the differences between the different types of measurements? Before we use this "thresholded" data to compare and contrast downstream effects between cell lines, we will first consider the questions proposed at the beginning of the chapter regarding the heterogeneity of transcriptional activity inherent not only among the L1000 genes themselves but also between those same genes across cell lines. With a firm grasp on these concepts, we can implement methods that moderate either true or false positive rates for detecting similarities or differences between CS's from different cell lines. In turn, these metrics may be used to evaluate the extent to which a KEGG pathway generalizes to a range of different cell lines or, on the contrary, represents cell line-specific regulatory relationships.

A first step in understanding the impact of cell line-specific gene behavior is to conduct a simulation study. The purpose of the simulation study is to evaluate the proposed test statistics when we *know* the parameters and data-generating mechanism behind a particular distribution. The simulation study, to be explicitly described in **Chapter 2**, will involve the creation and

evaluation of Level 6 CS-type data under three different data-generating mechanisms. In **Chapter 3**, the best-performing data generating mechanism from **Chapter 2** will be used as a basis for deriving non-parametric test statistics and integrated into the package KEGGlincs for an exploratory analysis. **Chapter 4** will describe how these statistics can be incorporated into the structure of KEGG pathways to describe similarities and differences among different levels of the L1000 data set and **Chapter 5** gives an overview of the package KEGGlincs, an R Bioconductor package designed to integrate CSs in pathway analysis in an integrative visual-analytic platform. Finally, we will use differences in CSs among cell lines as our input measurement for edge set enrichment analysis (ESEA), an established method of pathway analysis, will be conducted in **Chapter 6**.

Chapter 2: Simulation Study

The level of significance for the test statistics outlined in **Chapter 1** and further described in this chapter will be determined via non-parametric permutation testing procedures whereby observed test statistics are compared to a distribution of test statistics generated under null conditions. We will propose a permutation test that generates "random" concordance signatures by utilizing information from the patterns of behavior for LM genes across cell lines. The purpose of the simulation study is to investigate the nature of our proposed test statistics when the behavior of individual entities (LM-type genes) is defined (known) rather than estimated from the data.

2.1 Motivation

The concordance signatures capture bi-directional patterns (4 combinations) of behavior for 978 LM genes between any 2 of over 2,000 perturbating factors (PGs – 2,042 common PGs to be exact). The LM genes were chosen for their combined ability to reliably predict the expression of non-measured genes across the human genome. For each LM gene at each PG×CL, we know if a given LM gene is amongst the top 50 upregulated or 50 downregulated LM genes. A given LM gene is not part of the 100 most-differentially expressed genes attributed to a PG×CL if it is neither amongst the top 50 up- nor 50 down-regulated set.

The original format of the data set (**Figure 1(B)**) comes to us as a list entries which can be converted into an information matrix (**Figure 4**) and shed light on the behavior of individual LM genes within and between cell lines. This matrix (we will denote as $\mathbf{L}_{n \times m}$) is a numerical representation of the CGS data set whereby the rows ($n = 7 \times 2{,}042$) correspond to $PG_x \times CL^{cl}$

(perturbation factor for a given cell line) and the columns (m = 978) correspond to an individual landmark gene (g_j, j = 1,2,…,978). We define each entry of the matrix in **L** by indicator variable *I* such that:

$$I_{ij}^{cl} = \begin{cases} -1, include\ gene\ g_j\ in\ CGS\_D_i^{cl} \\ 0,\ no\ membership\ in\ CGS_i^{cl} \\ +1, include\ gene\ g_j\ in\ CGS\_U_i^{cl} \end{cases}$$

	AARS	ABCB6	ABCC5	ABCF1	ABCF3	ABHD4	ABHD6	ABL1	ACAA1
A375*A2M	0	0	1	0	-1	0	0	1	
A375*AARS	-1	0	0	0	0	0	0	1	
A375*AATF	0	0	1	0	0	0	0	0	
A375*ABAT	0	0	0	0	0	0	1	0	
A375*ABCA1	0	0	0	0	0	1	0	0	
A375*ABCA3	0	0	0	0	0	0	0	-1	
A375*ABCA5	0	0	0	0	0	0	0	0	
A375*ABCB1	1	0	0	0	0	0	0	0	
A375*ABCB6	0	-1	0	0	-1	0	0	0	
A375*ABCC2	0	0	1	0	0	0	-1	1	
A375*ABCC3	0	0	0	0	0	0	-1	0	
A375*ABCC5	0	0	0	0	0	0	0	0	
A375*ABCG5	-1	0	0	0	0	0	0	0	
A375*ABCG8	-1	0	0	0	0	0	1	0	

Figure 4: Snapshot of Matrix L
Matrix **L** is a matrix with 2,042×7×978 entries for indicator variable I_{ij}^{cl}

With these I's in place in **L** we may readily calculate the following probabilities:

$pu_j{}^{cl} = \frac{\sum_{i=1}^{2,042}(I_{ij}^{cl}=+1)}{2042}$ = probability that g_j will be one of 50 genes in a random $CGS\_U_i^{cl}$;

$pd_j{}^{cl} = \frac{\sum_{i=1}^{2,042}(I_{ij}^{cl}=-1)}{2042}$ = probability that g_j will be one of 50 genes in a random $CGS\_D_i^{cl}$;

$pn_j{}^{cl} = 1 - pu_j{}^{cl} - pd_j{}^{cl}$ = probability that g_j is neither part of a random $CGS\_D_i^{cl}$ nor $CGS\_U_i^{cl}$.

Note the following relationships:

$$0 \le pd_j{}^{cl}, pn_j{}^{cl}, pu_j{}^{cl} \le 1$$
$$pu_j + pd_j + pn_j{}^{cl} = 1$$

We go on to define the following probability and conditional probabilities:

$pil_j{}^{cl} = 1 - pn_j{}^{cl} = pu_j{}^{cl} + pd_j{}^{cl} = \frac{\sum_{i=1}^{2,042}(I_{ij}^{cl}=+1\ or -1)}{2042}$

= probability that g_j is either part of a random $CGS\_D_i^{cl}$ or $CGS\_U_i^{cl}$ (CGS_i^{cl});

$cpu_j{}^{cl} = \frac{pu_j{}^{cl}}{pil_j{}^{cl}}$ = probability that g_j is part of $CGS\_U_i^{cl} \mid g_j$ is in CGS_i^{cl};

$cpd_j{}^{cl} = \frac{pd_j{}^{cl}}{pil_j{}^{cl}}$ = probability that g_j is part of $CGS\_D_i^{cl} \mid g_j$ is in CGS_i^{cl}.

2.2 Variation Across and Between LM Genes

The *behavior* of the 978 LM genes is neither uniform across nor between cell lines for individual genes. We use the following definitions for the behavior of genes within a given cell line to define the observed variability:

1. *Regulatory Responsivity*
    - This aspect of gene behavior is captured by $pil_j{}^{cl}$
    - Genes with larger $pil_j{}^{cl}$ (further from zero/closer to 1) are associated with higher levels of regulatory responsivity as they are more likely to be among the top dysregulated genes across PGs.

2. *Directional consistency*:
    - This aspect of gene behavior is measured as a factor of the difference between $cpu_j{}^{cl}$ vs. $cpd_j{}^{cl}$ or, equivalently, $pu_j{}^{cl}$ vs. $pd_j{}^{cl}$.
    - When gene g_j is significantly dysregulated, is the conditional probability that it is upregulated versus downregulated approximately equal or is it typically upregulated *or* downregulated?
    - *Directional consistency* is high if genes have a strong tendency to be either up *or* down regulated [given that they are dysregulated] and decreases as the proportion of occurrences in up vs. down lists gets closer to 1.

These factors of LM gene variability could also be used to describe the differences in gene behavior between cell lines. **Figure 5** demonstrates this variability as a function of $pu_j{}^{cl}$ vs. $pd_j{}^{cl}$ and highlights the three general categories for gene regulation. This variable behavior underlies our decision to use a permutation procedure whereby the direction and inclusion of a

LM gene in a CGS is reassigned in order to evaluate differences in concordance signatures between cell lines. Before we employ non-parametric tests with estimated parameters, we will perform a simulation study on data with known parameters to explore the effects of our chosen sampling method on the post-sampling data distribution.

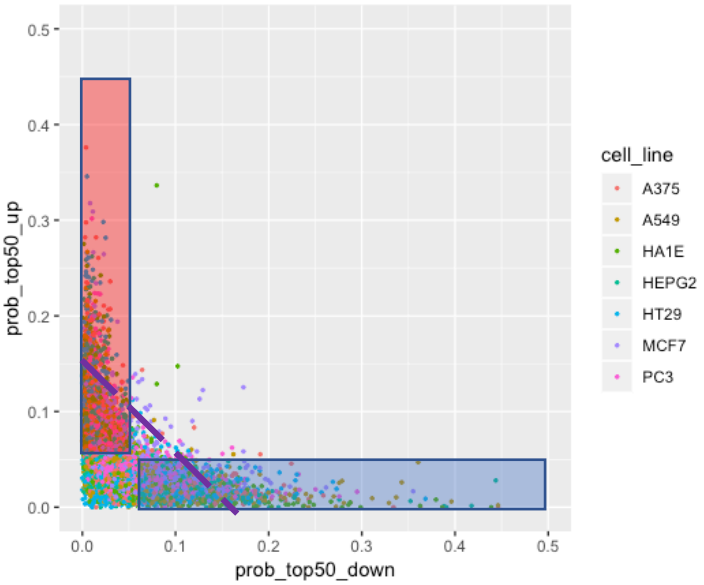


Figure 5: Regulatory patterns in L1000 genes

a) Genes likely to be upregulated vs. downregulated
   - $pu_j{}^{cl} > pd_j{}^{cl}$
   - $cpu_j{}^{cl} >> cpd_j{}^{cl}$
b) Genes likely to be downregulated vs. upregulated
   - $pu_j{}^{cl} < pd_j{}^{cl}$
   - $cpu_j{}^{cl} << cpd_j{}^{cl}$
c) Genes above the dashed line are more likely to be expressed than other genes *in general*
   - $pil_j{}^{c} > \overline{pil}$ where $\overline{pil}$ is the average marginal probability of list inclusion for genes across cell lines and perturbagens

2.3 Purpose of Simulation Study

The permutation procedure to explore similarities/differences of concordance signatures (CS) between cell lines will be conducted by comparing the observed test statistics to those generated under the null hypothesis. Our working definition of the null hypothesis is that the consensus genomic signatures are capturing random gene fluctuations within a cell line and therefor the CS's (that are based on agreement of two CGSs) are not meaningful. The permutation procedure will be conducted by generating 'random' CS's based on the concordance between two 'random' nodes – each with a CGS of the top up and down-regulated LM genes after they are selected using features of the observed LM gene behavior within a given cell line.

The simulation study will explore the statistical framework of the non-parametric testing procedure that will eventually be employed to find meaningful differences among CS's in the

L1000 data set. The distribution of tests statistics measuring differences between 'random' CS's generated under the null vs. alternate scenarios will shed light on the power of our proposed test statistics under a variety of controlled parameters. The first stage of the simulation study will evaluate different methods for generating simulated data ('random CS's) and the second stage will employ the least biased method for conducting permutation tests on the simulated data across variable parameter settings.

The premise is as follows. There is a pool of sampling units that represent genes. Each sampling unit is associated with a particular group ("Regulatory Group") that defines the unit's behavior. Sampling units in the same group have identical probabilities of overall selection (probability of being selected into either an up or down list) and identical conditional probabilities for their direction (chance of being up or down given that they are in a list); thus, the units in each group also have matching marginal probabilities for selection into an up- or down-regulated list. The purpose of grouping the units in this way is to explore the effect of the sampling algorithm on a pool of 'genes' with well-defined properties before using the method on a real data set that has many parameter estimates (as many parameters as there are genes).

2.4 Sampling Simulated Data

The end goal for each of the sampling methods described in **2.4.2** is to populate "UP" and "DOWN" lists with a prespecified number of sampling units at each run or, in other words, *synthesize* CGSs from a population of genes (referred to as sampling units or simply units) with a known distribution. After the lists are generated across different sets of parameter values under the direction of each method, the data-generating mechanisms will be evaluated for potential sampling bias as outlined in section **2.4.3**.

2.4.1 Notation

The parameters defined in this section have been chosen to reflect characteristics of the sampling distribution in accordance with the probabilities outlined in section **2.1**. Each unit g_{ij} is assigned to a specific regulatory group *i* whereby all units in that group have identical sampling properties as described by the following notation:

c = total number of groups (regulatory group)

$i =$ group index; $i = 1,2, \dots, c$

$m_i =$ number of sampling units in group i

$M = \sum_{i=1}^{c} m_i =$ total number of sampling units

$g_{ij} = i$ndividual sampling unit from group $i, j = 1, \dots, m_i, i = 1, \dots, c$

$G_i =$ collection of samlping units in group i; $\sum_{j=1}^{m_i} |g_{ij}| = |G_i|$

$G =$ collection of all sampling units; $|G| = M$

$S =$ total number of simulations

$s =$ simulation index; $s = 1,2 \dots, S$

$n_U =$ number of units in an "UP LIST"

$n_D =$ number of units in a "DOWN LIST"

$n = n_U + n_D$

$\omega_i =$ selection *weight* for genes in group i

$\pi_i = \frac{\omega_i}{\sum_{k=1}^{c} m_{k*}\omega_k} =$ selection *probability* for a single unit from group i

$\varphi_i^U =$ up-direction *weight* for units in group i

$\varphi_i^D =$ down-direction *weight* for units in group i

$\varphi_i^U + \varphi_i^D = 1$

$\varpi_i^U = \pi_i * \varphi_i^U =$ up-selection *probability* for gene in group i

$\varpi_i^D = \pi_i * \varphi_i^D =$ down-selection *probability* for gene in group i

$\varpi_i^U + \varpi_i^D = \pi_i$

At each simulation s, sampling units are either unselected or selected into L^s.

$L^s =$ collection of sampling units selected into an "UP" OR "DOWN LIST" at simulation s

If a unit is selected as part of L^s, it will be part of one but not both of the following lists:

$L_U^s =$ sampling units selected into an "UP LIST" at simulation s, $|L_U^s| = n_U$

$L_D^s =$ sampling units selected into a "DOWN LIST" at simulation s, $|L_D^s| = n_D$

$L_U^s \cup L_D^s = L^s \; ; \; L_U^s \cap L_D^s = \emptyset$

$$I_{ij}^s = \begin{cases} -1 \;\; if \;\; g_{ij} \in L_D^s \\ +1 \;\; if \;\; g_{ij} \in L_U^s \\ \quad 0 \;\; else \end{cases}$$

2.4.2 Methods

Method 1a ("Up First")

Step 1: Populate L_U^s by sampling n_U units from *G without replacement* using ϖ_i^U as weights.

Step 2: Populate L_D^s by sampling n_D units from $G - L_D^s$ *without replacement* using ϖ_i^U as weights.

Method 1b ("Down First")

Step 1: Populate L_D^s by sampling n_D units from *G without replacement* using ϖ_i^D as weights.

Step 2: Populate L_U^s by sampling n_U units from $G - L_U^s$ *without replacement* using ϖ_i^U as weights.

Method 2a ("Partition Up First")

Step 1: Sample n units *without replacement* to populate L^s from G using π_i as weights.

Step 2: Assign membership of n_U units from L^s into L_U^S using φ_i^U as weights; all other units in L^s are assigned membership into L_D^S by default.

Method 2b ("Partition Down First")

Step 1: Sample n units *without replacement* to populate L^s from G using π_{ij} as weights.

Step 2: Assign membership of n_D units from L^s into L_D^S using φ_i^D as weights; units left in L are partitioned into L_U^S by default.

Method 3 ("Random Labels")

Step 1: Label all M genes by sampling [with replacement] the labels "UP" or "DOWN" in a binomial fashion with the vector of probabilities taking the form $< \varphi_i^U, \varphi_i^D >$.

Let $\delta_U^S = \text{units labelled "UP" and } \delta_D^S = \text{units labelled "DOWN" after executing Step 1.}$

Step 2a: Sample without replacement n_U units from δ^U using π_i as weights

Step 2b: Sample without replacement n_D units from δ^D using π_i as weights

2.4.3 Measuring bias in simulated data

Notation

$$Let\ {U_i}^s = \frac{1}{n}\sum_{j=1}^{m_i}(I_{ij}^s = +1);\ \ \overline{U}_i = \frac{1}{nS}\sum_{s=1}^{S}\sum_{j=1}^{m_i}(I_{ij}^s = +1)$$

$$Let\ {D_i}^s = \frac{1}{n}\sum_{j=1}^{m_i}(I_{ij}^s = -1);\ \ \overline{D}_i = \frac{1}{nS}\sum_{s=1}^{S}\sum_{j=1}^{m_i}(I_{ij}^s = -1)$$

$$\bar{P}_i = \overline{U}_i + \overline{D}_i$$

$$\bar{\vartheta}_i^U = \overline{U}_i \Big/ (\overline{U}_i + \overline{D}_i) = \overline{U}_i \Big/ \bar{P}_i$$

$$\bar{\vartheta}_i^D = \overline{D}_i \Big/ (\overline{U}_i + \overline{D}_i) = \overline{D}_i \Big/ \bar{P}_i$$

Bias Measurements

$$\tau_i = \bar{P}_i - E[\,P_i\,]$$

$$\Delta_{ab} = \bar{P}_a - \bar{P}_b$$

$$\varsigma_i^U = \varphi_i^U - \bar{\vartheta}_i^U$$

$$\varsigma_i^D = \varphi_i^D - \bar{\vartheta}_i^D$$

The term $\bar{P}_i$ represents the average marginal 'share' of list membership by units from G_i. In other words, for every 100 entries in a randomly chosen simulated list (L^S), how many entries would we expect to be occupied by genes from G_i? $\bar{P}_i$ is a *marginal* measurement of information collapsed over lists [i.e. over L_D^S and L_U^S] whereas the term $\bar{\vartheta}_i$ is a measure of *conditional* list membership. The value of $\bar{\vartheta}_i^U$ [$\bar{\vartheta}_i^D$] is an answer to the following question: given that a unit from G_i occupies a randomly selected L^S, what is the probability that it is part of L_U^S [L_D^S]? These post-sampling quantities will be calculated to measure the relative bias between the candidate sampling methods. The least biased sampling method will adhere to the following conditions:

(1) $\bar{P}_i - E[P_i] = 0$

(2) If $\bar{\vartheta}_i^U : \bar{\vartheta}_i^D = \varphi_i^U : \varphi_i^D \implies \varsigma_i^U = \varsigma_i^D = 0$

(3) If $\omega_a = \omega_b \ \ AND \ \ m_a = m_b \implies \Delta_{ab} = 0$

Conditions 1 and 2 measure bias as it relates to within-group comparisons of parameter estimates. The statement given in Condition 3 is, perhaps, a more subtle indicator of bias whereby its measurement is a function of parameter estimates between two different groups that meet the specified criteria for similarity. Condition 3 states that groups with identical selection weights and number of sampling units should have equal marginal proportions of list membership, on average, after synthetic lists are created, even if *direction* weights differ between the groups. In each simulation scenario, Condition 3 will be evaluated between *complementary*

groups. Two groups G_a and G_b are complementary if $\omega_a = \omega_b$, $m_a = m_b$, $\varphi_b^U = \varphi_a^D$ and $\varphi_b^D = \varphi_a^U$.

2.5 Motivation

An initial 'test run' was performed prior to a large-scale simulation study in an effort to ensure that the algorithms were performing as expected. The sampling parameters were chosen in order to represent a simplified prototype of the original dataset according to gene behavior amongst the L1000 genes (via splitting the genes into groups according to their regulatory activity and group size) and the size of the CGSs ($n_U = n_D = 50$). 1000 sampling units were assigned to one of three "Regulatory Groups" (RGs) that determine the sampling parameters according to **Table 3** on the following page.

In the initial setup, three groupings were chosen as a generalization of gene behavior showcased in **Figure 5**. In a similar manner as the L1000 genes, most of the sampling units exhibit "normal behavior" (RG3, 90% of sampling units, selection weight of 1, equal chances of allocation to either an UP or DOWN list), some genes are "largely overexpressed" and usually found in an UP list (RG1, 5% of sampling units, selection weight of 2.25, 9:1 chance of inclusion in an UP vs. DOWN list), and some genes are "largely underexpressed" and usually found in an DOWN list (RG1, 5% of sampling units, selection weight of 2.25, 9:1 chance of inclusion in a DOWN vs. UP list). Each algorithm ran S = 2000 times, thereby generating 2000 "UP" and "DOWN" simulated CGSs (to induce stable estimates and to mimic the 2,042 CGSs across PGs in the L1000 data). The values for $E[P_i]$ were calculated *under the assumption* that the sampled totals would follow a multinomial distribution with $E[P_i] = m_i\pi_i$; $(E[P_1], E[P_2], E[P_3]) = (0.1, 0.1, 0.8)$ [15].

Table 3: Initial parameter settings for sampling algorithms

	G_1	G_2	G_3
m_i	50	50	900
ω_i	2.25	2.25	1
π_{ij}	0.002	0.002	0.0008888889
φ_i^U	0.9	0.1	0.5
φ_i^D	0.1	0.9	0.5
ϖ_{ij}^U	0.0018	0.0002	0.0004444444
ϖ_{ij}^D	0.0002	0.0018	0.0004444444

Table 4: Average proportion of genes from group $G_i, i = (1,2,3)$ in L^s

Method	$\bar{P}_1$	$\bar{P}_2$	$\bar{P}_3$
1a	0.0933	0.0985	0.8082
1b	0.0972	0.0935	0.8092
2a	0.0955	0.0967	0.8078
2b	0.0956	0.0961	0.8083
3	0.0955	0.0954	0.8091

Table 5: Post-sampling direction weights

Method	$\bar{\vartheta}_1^U, \bar{\vartheta}_1^D$	$\bar{\vartheta}_2^U, \bar{\vartheta}_2^D$	$\bar{\vartheta}_3^U, \bar{\vartheta}_3^D$
1a	0.9113, 0.0887	0.1053, 0.8947	0.5009, 0.4991
1b	0.8958, 0.1042	0.0995, 0.9005	0.4988, 0.5012
2a	0.7269, 0.2731	0.1336, 0.8664	0.5171, 0.4829
2b	0.8651, 0.1349	0.2628, 0.7372	0.4852, 0.5148
3	0.9052, 0.0948	0.0999, 0.9001	0.4995, 0.5005

The results of the pilot simulation are recorded in **Table 4** and **Table 5**; **Table 6** reports the bias measurements. In summary, Methods 1a/1b and Method 3 simulate data sets with more accurate post-sampling direction weights than Methods 2a/2b whereas Methods 2a/2b and Method 3 resulted in the smallest differences in Δ_{ab} between complementary groups. Although Method 3 had the best performance according to Condition 2 and Condition 3 (outlined in

Section 2.3.3), it also produced larger deviations from $E[P_i]$ than the other methods. More importantly than the individual measurements between methods was the strikingly similar pattern showing that all methods appeared to oversample from G_3 and thus under-sample from G_1 and G_2. This discrepancy was in need of resolution before the larger simulation study could be carried out. The solution was not found in a different implementation of the sampling algorithms but rather by identifying the multivariate Wallenius non-central hypergeometric distribution as opposed to the multinomial distribution as the appropriate sampling distribution for the simulated data.

Table 6: Measurements for bias across methods and groups. The minimum absolute value (or pair of values) is underlined for each column.

Method	τ_1	τ_2	τ_2	$\varsigma_1^U, \varsigma_1^D$	$\varsigma_2^U, \varsigma_2^D$	$\varsigma_3^U, \varsigma_3^D$	Δ_{12}
1a	-0.0067	<u>-0.0015</u>	0.0082	-0.0113, 0.0113	-0.0053, 0.0053	-0.0009, 0.0009	-0.0052
1b	<u>-0.0028</u>	-0.0065	0.0092	<u>0.0042, -0.0042</u>	0.0005, -0.0005	0.0012, -0.0012	0.0037
2a	-0.0045	-0.0033	<u>0.0078</u>	0.1731, -0.1731	-0.0336, 0.0336	-0.0171, 0.0171	-0.0012
2b	-0.0044	-0.0039	0.0083	0.0349, -0.0349	-0.1628, 0.1628	0.0148, -0.0148	-0.0005
3	-0.0045	-0.0046	0.0091	-0.0052, 0.0052	<u>-0.0001, 0.0001</u>	<u>0.0005, -0.0005</u>	<u>0.0001</u>

2.6 The Multivariate Wallenius Non-central Hypergeometric Distribution

2.6.1 Background

K.T. ("Ted") Wallenius first introduced the non-central hypergeometric distribution as a means to account for biased sampling between two finite populations with dichotomized attributes in his Ph.D thesis in 1967 [16]. His work was motivated by the need to measure the extent of non-randomness in a sampling population – that is – the extent to which sampling units in a population have an unequal or biased chance of being selected relative to other units being

sampled. The example featured in his paper involved measuring differences in survival vs. death between rabbits in a population that could be divided into two subsets characterized by homozygous and heterozygous blood type alleles with the notion that genetic selection for one phenotype would influence sampling from that group – i.e. the number of rabbits that were still alive in one phenotypic group instead of the other was influenced by factors above and beyond the ratio of rabbits that were alive in each group at the starting timepoint.

In 1976 Chesson extended Walleneius' definition to account for biased sampling when there are more than two subsets in the sampling population [17]. He describes a multivariate hypergeometric distribution to measure selective predation, a term to describe the degree to which a predator consumes different types of prey based in part by preferential factors in addition to the relative abundance amongst the prey. Consider the following situation: a predator's diet consists of m different species of prey in an environment with n_i $(i = 1, \dots, m)$ individual animals of each species for a starting total of $N = \sum_{i=1}^{m} n_i$ animals eligible for predation. In the case of random predation, the probability that the predator's next meal will be an animal of type i is $\frac{n_i}{\sum_{j=1}^{m} n_j} = \frac{n_i}{N}$. However, when it is feasible that other factors will have an impact on this outcome, those factors can be incorporated into a model that accounts for biased predation.

For example, let β_i be the probability that the predator will detect prey of type i at any given encounter and p_i be the probability that it will pursue that type of prey such that probability of the prey's capture and consumption takes the form $\frac{p_i \beta_i n_i}{\sum_{j=1}^{m} p_j \beta_j n_j}$ [18]. Let P_i be the probability that the predator's next meal is an animal of type i and let $\alpha_i = p_i \beta_i$, such that

$P_i = \frac{\alpha_i n_i}{\sum_{j=1}^{m} \alpha_j n_j}$. The values in vector $\boldsymbol{\alpha}$ represent the relative preference for one prey species versus the others and measure the deviation of P_i from $\frac{n_i}{N}$. P_i demonstrates the effect of preference on the outcome of an initial predatory event (one draw from a sample).

The purpose of taking into consideration selective predation is to incorporate measurements of bias into the estimation of population parameters after many selection events have taken place. Let r_i $(i = 1, \dots, m)$ be the number of prey animals belonging to species i that a predator has eaten after having r meals (or consuming a total of r animals). Note $\sum_{i=1}^{m} r_i = r$ and $r_i \leq m_i$. Chesson outlines two possible situations that will determine the distribution for the vector **R**, whose i[th] element represents the value r_i. First, consider a scenario where n_i remains relatively constant over time; either prey is added to the population at the same rate it is consumed or $r \ll N$. This is akin to a sampling with replacement scenario when there is a functionally infinite population of prey animals and **R** has the multinomial distribution where

$$P(\boldsymbol{R} = \boldsymbol{r}) = \frac{r}{\prod_{i=1}^{m} r_i} \prod_{i=1}^{m} \left[\frac{\alpha_i n_i}{\sum_{j=1}^{m} \alpha_j n_j} \right]^{r_i}. \tag{6}$$

The second scenario is one that reflects the reduction of n_i in a fixed population, as is the case in a sampling without replacement procedure. In this scenario, **R** has the multivariate non-central hypergeometric distribution and

$$P(\boldsymbol{R} = \boldsymbol{r}) = \prod_{i=1}^{m} \binom{n_i}{r_i} \int_0^1 \prod_{i=1}^{m} (1 - t^{\alpha_i c_i})^{r_i} \, dt, \tag{7}$$

where $c_i = \frac{1}{\sum_{i=1}^{m} \alpha_i (n_i - r_i)}$ and $\alpha_i \neq 1$ for at least one α_i (otherwise the distribution is central as opposed to non-central).

In his 2008 paper, Agner Fog formally identifies the 'Biased Urn' model and gives it the name 'multivariate Wallenius non-central hypergeometric distribution' in order to distinguish it from a similar model that he calls the 'multivariate Fisher non-central hypergeometric distribution' (which has also been referred to as the 'extended hypergeometric distribution') [19]. In the biased urn model, an urn contains balls with c different colors and m_i ($i \in C = \{1,2,\dots,c\}$) balls of each color for a total of $N = \sum_{i=1}^{c} m_i$ balls in the urn. The color of the ball indicates that color's ω_i, which is the probability of selecting a ball of that particular color relative to balls of different colors. ω_i can account for features that may increase/decrease the color's relative probability of selection above and beyond the ratio of balls of that type versus others such as the relative size of the ball or texture of the ball. To meet criteria for the multivariate Wallenius non-central hypergeometric distribution, a total of n balls are selected one by one, without replacement, such that the probability of selecting a ball of type i at draw v_j ($j = 0,1,2,\dots,n$) is dependent upon the combination of balls selected up to draw v_j (i.e. the composition of balls left in the urn at draw v_{j-1}). If the balls are selected simultaneously without dependence between draws, then the distribution of balls in the urn has a multivariate Fisher non-central hypergeometric distribution. In the univariate case, Fisher's non-central hypergeometric distribution can be regarded as the "conditional distribution of two independent binomial random variables, given their sum" [20]. The multivariate Fisher and Wallenius non-central hypergeometric distributions both simplify to the hypergeometric distribution when there is no bias for selection between objects of different types ($\omega_i = 1 \; \forall \; i \in C$) and reduce to the multivariate binomial distribution when n = 1 (only one draw is taken) [21].

Consider, for example, an urn with m_1 tennis balls (large balls with rough surfaces), m_2 ping-pong balls (medium size balls with slightly textured surface) and m_3 marbles (small balls

with smooth surfaces) for a total of N balls. In an experiment, the urn is placed under a metal claw which will drop down and select one ball 'at random'. If there is no effect on the size or texture of the sampling unit, the probability that the first ball is a tennis ball is $\frac{m_1}{N}$, $\frac{m_2}{N}$ that it is a ping-pong ball and $\frac{m_3}{N}$ that it is marble. However, it is reasonable to assume that, for instance, a tennis ball would be somewhat easier for the claw to grab on to versus a ping-pong ball and much easier than a marble. This experiment of a single draw could be conducted multiple (i.e. hundreds or thousands of times) to estimate values for the vector $\boldsymbol{\omega}$ so that when p_i is the first-draw probability for a ball of type i, $p_i = \frac{m_i\omega_i}{\sum_{j=1}^{c} m_j\omega_j}$. The $\boldsymbol{\omega}$ values are measurements of bias that account for all underlying (latent) mechanisms that result in biased sampling, similar to the $\boldsymbol{\alpha}$ values (where $\alpha_i = p_i\beta_i$) described by Chesson. Fog introduces notation to enumerate the dependences between draws during a selection process in terms of probabilities and expected values. Let $\boldsymbol{X_v}$ be a vector that records the number of balls of each type that have been drawn over the past v draws; $\boldsymbol{X_v} = (X_{1v}, \ldots, X_{iv}, \ldots, X_{cv})$. The *probability* that a ball of color i is selected at the next draw, draw $v + 1$, is:

$$p_{i(v+1)} = \frac{(m_i - X_{iv})\omega_i}{\sum_{j=1}^{c}(m_j - X_{jv})\omega_j}. \tag{8}$$

Now let $\boldsymbol{\mu_v} = (\mu_{1v}, \ldots, \mu_{iv}, \ldots, \mu_{cv}) = E[\boldsymbol{X_v}]$, that is, the expected count for each type of ball at draw v. Note that when $v = 0$ (before any draws have occurred) $\mu_{i0} = 0 \; \forall \; i \in C$. The means in vector $\boldsymbol{\mu_v}$ can be approximated by the recursive relationship for $v \geq 1$:

$$\mu_{iv} \approx \mu_{i(v-1)} + p_{i(v)}\boldsymbol{\mu}_{i(v)} \tag{9}$$

Equation (9) is useful in the descriptive sense and even for calculations when the number of all possible enumerated quantities is small, yet it quickly becomes unwieldy as the number of draws

(v) and/or unique groups (size of C) becomes large. Fog's recently published R package titled 'BiasedUrn' [19] has functions capable of estimating $\boldsymbol{\mu_v}$ values in an efficient and reliable manner; precisely the type of functionality we need to obtain accurate estimates for the measurements $E[P_i]$ defined in **Section 2.3.3.**

2.6.2 Estimation of Group P_i with BiasedUrn Package

The purpose of introducing the multivariate Wallenius non-central hypergeometric distribution (hereafter referred to as MWNCH) is to offer an explanation for the discrepancies and solution for the calculations of $E[P_i]$ in Section 2.4. This will afford us a more stable grounds for selecting the least biased sampling method. What follows is a description of the pilot simulation as it pertains to sampling units from a population that follows the MWNCH distribution.

The sampling space S contains $N = \sum_{i=1}^{3} m_i$ sampling units (synthetic genes) where m_i is the number of sampling units for group $i \in C = \{1,2,3\}$ and $\boldsymbol{m} = (50, 50, 100)$. Each group G_i has a corresponding weight (ω_i) in the marginal selection weight vector $\boldsymbol{\omega} = (2.25, 2.25, 1)$; in other words the [starting] odds of selecting a unit from G_1or G_2 is equal (between the groups) and 2.25 times the odds of selecting a unit from G_3 if group size is ignored. The total number of draws n is set to 100 (50 genes "UP" plus 50 genes "DOWN"). The values in $\boldsymbol{m}, \boldsymbol{\omega}$, and n are the inputs to the function "momentsMWNCHypergeo" from the BiasedUrn package which returns the following vector of values: $\mathrm{E}[\boldsymbol{X_{v=100}}] = [9.5489, 9.5489, 80.9021]$. Therefore, on the basis of a 100-unit sample, $E[\boldsymbol{P}] = [0.0955, 0.0955, 0.8090]$. The measurement in the first three columns of **Table 6** can now be adjusted to represent a more appropriate estimation of bias as shown in **Table 7.** Now it is clear that Method 3 has the least biased performance, on average,

across the performance measures. Moving forward, the central question is now: does method 3 consistently perform better across sampling scenarios with different parameters?

Table 7: Measurements for bias across methods and groups with τ values adjusted for estimation of $E[\boldsymbol{P}]$ according to MWNCH distribution. The minimum absolute value (values, or pair of values) is underlined for each column.

Method	τ_1	τ_2	τ_2	$\varsigma_1^U, \varsigma_1^D$	$\varsigma_2^U, \varsigma_2^D$	$\varsigma_3^U, \varsigma_3^D$	Δ_{12}
1a	-0.0022	0.0030	-0.0008	-0.0113, 0.0113	-0.0053, 0.0053	-0.0009, 0.0009	-0.0052
1b	0.0017	-0.0020	0.0002	0.0042, -0.0042	0.0005, -0.0005	0.0012, -0.0012	0.0037
2a	0.0000	0.0012	-0.0012	0.1731, -0.1731	-0.0336, 0.0336	-0.0171, 0.0171	-0.0012
2b	0.0001	0.0006	-0.0007	0.0349, -0.0349	-0.1628, 0.1628	0.0148, -0.0148	-0.0005
3	0.0000	-0.0001	0.00001	-0.0052, 0.0052	-0.0001, 0.0001	0.0005, -0.0005	0.0001

2.6.3 Parameters and Expected Values for Data Simulation

The purpose of the pilot simulation was to ensure that the implementation of different sampling algorithms was free from execution errors. What follows is a description of the larger simulation study which aims to draw conclusions concerning differences between the proposed sampling methods. **Table 8-11** describe the parameterizations under each scenario. The name of the scenario indicates the number of regulatory groups defined in each situation; for example, the scenario "3RG" or "RG3" has three regulatory groups (#RG and RG# are used interchangeably). In each scenario there is a "baseline" group. This group has the largest number of sampling units, a sampling weight of 1, and equal up vs. down direction (and thus selection) weights. All other groups have selection weights and up OR down direction weights with magnitude that is inversely related to the number of units in the group. With the exception of the baseline group, each group has a complementary group affiliation (see **Section 2.3.3**). For every scenario there are a total of 1000 sampling units ($\sum_{i=1}^{C} m_i = 1000$). The parameters for each scenario were fed

into each sampling algorithm across the following values of *s* (number of simulations, or in other words, number of simulated CGSs): 100, 1000, and 10000.

Table 8: Group size

Scenario	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
3RG	100	100	800						
5RG	50	50	200	200	500				
7RG	50	50	100	100	200	200	300		
9RG	50	50	75	75	100	100	125	125	300

Table 9*:* Selection weight

Scenario	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9
3RG	2	2	1						
5RG	3	3	2	2	1				
7RG	4	4	3	3	2	2	1		
9RG	5	5	4	4	3	3	2	2	1

Table 10: Direction weight

Scenario	φ_1^U	φ_2^U	φ_3^U	φ_4^U	φ_5^U	φ_6^U	φ_7^U	φ_8^U	φ_9^U
3RG	0.9	0.1	0.5						
5RG	0.9	0.1	0.7	0.3	0.5				
7RG	0.9	0.1	0.8	0.2	0.7	0.3	0.5		
9RG	0.9	0.1	0.8	0.2	0.7	0.3	0.6	0.4	0.5
Scenario	φ_1^D	φ_2^D	φ_3^D	φ_4^D	φ_5^D	φ_6^D	φ_7^D	φ_8^D	φ_9^D
3RG	0.1	0.9	0.5						
5RG	0.1	0.9	0.3	0.7	0.5				
7RG	0.1	0.9	0.2	0.8	0.3	0.7	0.5		
9RG	0.1	0.9	0.2	0.8	0.3	0.7	0.4	0.6	0.5

Table 11: $E[P_i]$

Scenario	$E[P_1]$	$E[P_2]$	$E[P_3]$	$E[P_4]$	$E[P_5]$	$E[P_6]$	$E[P_7]$	$E[P_8]$	$E[P_9]$
3RG	0.1619	0.1619	0.6762						
5RG	0.0904	0.0904	0.2489	0.2489	0.3215				
7RG	0.0918	0.0918	0.1411	0.1411	0.1929	0.1929	0.1483		
9RG	0.0963	0.0963	0.1180	0.1180	0.1204	0.1204	0.1025	0.1025	0.1256

The same scenarios have also been conducted under "complete" null conditions as well as "null selection weight" (Null SW) and "null direction weight" (Null DW) conditions (in contrast to the previously described "biased" condition) defined as follows:

Complete Null (Null)

1. $\omega_i = 1 \; \forall \; i \; \in C$
2. $\varphi_i^U = \varphi_i^D = 0.5 \; \forall \; i \; \in C$
3. $m_i = m_j \; \forall \; i, j \; \in C; \; i \neq j$

Null Selection Weight (Null SW)

1. m_i and $< \varphi_i^U, \varphi_i^D >$ are the same as those in the biased condition
2. $\omega_i = 1 \; \forall \; i \; \in C$

Null Direction Weight (Null DW)

1. ω_i and m_i are the same as those in the biased condition
2. $\varphi_i^U = \varphi_i^D = 0.5 \; \forall \; i \; \in C$

Please refer to the Appendix for the complete description of these parameter inputs.

2.7 Data Simulation Results

All of the methods proposed for data simulation are designed to accommodate two important aspects of sampling during the simulation process: they operate on the basis that there are variable probabilities for selection among the sampling units and that membership in an "UP" vs. "DOWN" list is mutually exclusive for a single sampling unit. These two aspects of data simulation violate the rules for equal probabilities and independence that are required for a simple random scheme therefore we would expect to find evidence of sampling bias in resulting data sets [22]. The identification of bias amongst the sampling methods must be addressed as a preliminary analytical measure to ensure that downstream simulation analyses and permutation

procedures employed on the actual data are conducted based on the optimal simulation and permutation methods.

Most of the bias in the distribution of the post-sampling population can be specifically attributed to the intentionally biased sampling design properties and can be predicted using the MWNCH distribution. However, as the results of the data simulation will suggest, the order in which units are selected to either an "UP" or "DOWN" list is an additional source of bias when there is heterogeneity in the direction weights amongst the sampling units – that is, in the non-central as opposed to central sampling scheme. In fact, across the factors of group size, selection weight and direction weight, non-centrality (biased-ness) among the direction weights is the only factor that separates the sampling methods' performance.

The scatterplots in **Figure A1** in the Appendix are designed to summarize the measurements of bias introduced in **Section 2.3.3** for comparison across the sampling conditions (biased, complete null, null DW and null SW), scenarios, methods and numbers of simulations; **Figure 6** is a snapshot of the biased condition run at s = 10000 simulations . In each plot, the point (τ_i, ς_i^U) is plotted in two dimensions; τ_i on the x-axis and ς_i^U along the y-axis. Given the relationship $\varsigma_i^U = \varsigma_i^D * (-1)$ (See **Table 6),** results based on ς_i^D would be mirrored across the x-axis (summaries are based on ς_i^U only for brevity). Within the figures, the plotting symbols represent the sampling method and the color represents the group. The figures are arranged by scenario (row) and number of simulations (column).

The most glaring global comparison is between the conditions with null vs. non-null direction weights. In the conditions with null direction weights there is no discernable difference between methods; the points appear to be randomly scattered across the plots and their range decreases across both axes (i.e., estimates are more precise) as the number of simulations

increases. On the other hand, in the non-null direction weight conditions, the methods separate across either the x- or y-axis. There is one commonality across scenarios for non-null direction weight conditions: as the number of simulations increase, the minimum and maximum values across the x-axis (τ_i) become smaller but the range of values remains constant for the y-axis (ς_i^U) across numbers of simulations and scenarios. This suggests that, in general, $\bar{P}_i$ gets closer to $E[P_i]$ as the number of simulations increases across methods. That being said, the discussion that follows will highlight discernable patterns of difference that point to one method as the optimal, least-biased method for data simulation.

The best performing sampling method in accordance with the three conditions introduced in **Section 2.4.3** will be the one with the majority of points centered about the origin (0, 0) across [biased] scenarios and number of simulations. With this in mind, a quick visual synopsis of **Figure 6** points to Method 3 as the best performing sampling algorithm, a claim that warrants an inspection of the performance patterns between methods across both dimensions of performance and regardless of the number of groups. The following terms will be used to identify patterns in sampling behavior:

- The term "baseline group", as previously mentioned, refers to the group in a given scenario that: 1) has the most sampling units, 2) has a sampling weight of 1 and 3) $\varphi_i^U = \varphi_i^D = 0.5$.
- Within a pair of complimentary groups ($\varphi_i^U = \varphi_j^D$ and $\varphi_i^D = \varphi_j^U, i \neq j$):
    - The group for which $\varphi_i^U > \varphi_i^D$ is the "up-dominant" group and $\varphi_i^D > \varphi_i^U$ if the group is "down-dominant".

For example, in Scenario 3 where groups 1 and 2 are the only complementary groups, group 1 is the up-dominant group ($m_1 = 100,\ \omega_1 = 2,\ \varphi_1^U = 0.9,\ \varphi_1^D = 0.1$), group 2 is the down-

dominant group $(m_2 = 100,\ \omega_2 = 2,\ \varphi_2^U = 0.1,\ \varphi_2^D = 0.9)$ and group 3 is the baseline group $(m_3 = 800,\ \omega_3 = 1, \varphi_3^U = \varphi_3^D = 0.5)$.

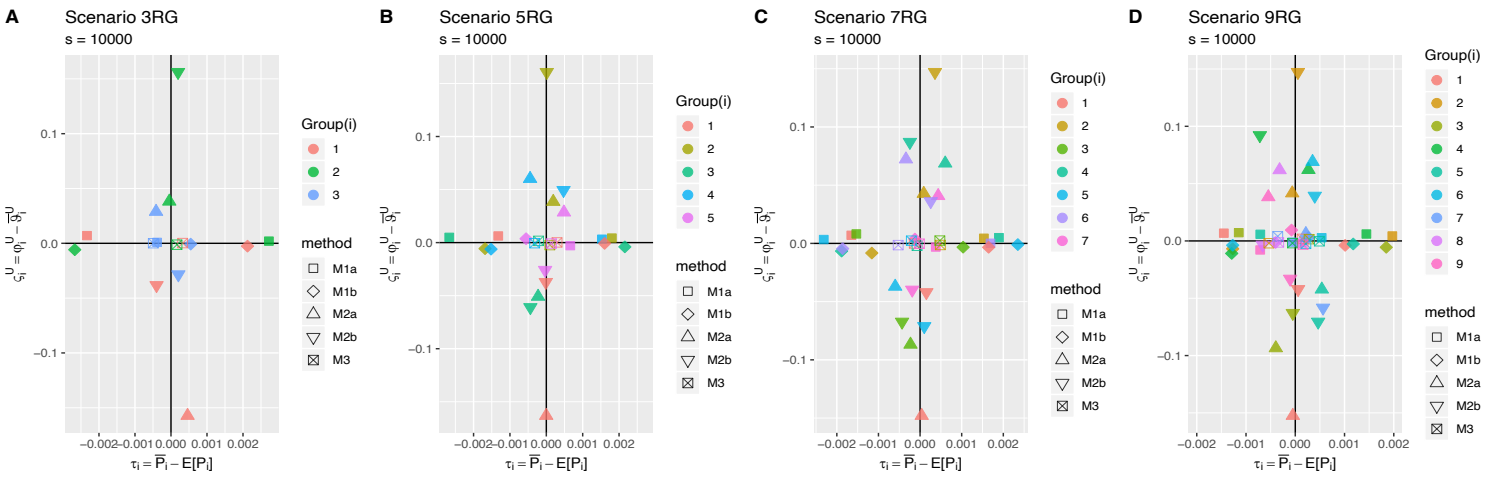


Figure 6: Scatterplots comparing sampling methods across two dimensions of bias at fixed number of simulations (s = 10000).

The plotting points for Methods 1a and 1b tend to lie close to the x-axes of the plots contained in **Figure 6** due to small values of the measurements ς_i^U (and therefore ς_i^D) across scenarios. Thus, these methods perform well in terms creating simulated data sets in which the direction weights for sampling "UP" vs. "DOWN" units from specific groups is reflected in the ratio of sampling units observed in "UP" vs. "DOWN" lists. However, the same plotting points for these methods lie further from the y-axes than those for the other methods. This observation is attributable to small but consistent patterns of bias in the measurement τ_i. On the other hand, the plotting points for Method 2a and 2b lie close to the y-axes as a result of small values of the measurements of τ_i yet are far from the x-axes due to larger values of ς_i^U.

In **Figure 7** and **Figure 8**, the two dimensions of bias are plotted separately against a variable number of simulations in order to highlight specific, consistent patterns between the methods. Note that each method is plotted individually and, in contrast to **Figure 6**, the shape of

the plotting character represents whether the group is "up-dominant" (triangle pointing upward), "down-dominant" (triangle pointing downward) or "baseline" (circular). For the clarity of discussion, Scenario 5 is presented in **Figure 7** and **Figure 8**, but the observations made regarding group patterns in this scenario extend to Scenarios 3, 7 and 9. Recall that Method 1a first selects 5 sampling units to be part of an "UP" list using the composite ϖ_{ij}^{U} weight for selection to an "UP" list and then, from the remaining 950, uses ϖ_{ij}^{D} to select 50 for the "DOWN" list - Method 1b first selects 50 sampling units to be part of an "DOWN" list using the composite ϖ_{ij}^{D} weight for selection to a "DOWN" list and then, from the remaining 950, uses ϖ_{ij}^{U} to select 50 for the "UP" list. In **Figure 7(A)** (Method 1a), the plotting points corresponding to the "up dominant" groups fall below the x-axis whereas those for the "down dominant" groups lie above it and in **Figure 7(B)** the behavior is reversed (points for the "up dominant" groups lie above the x-axis and those for the "down dominant" groups fall below the zero line).

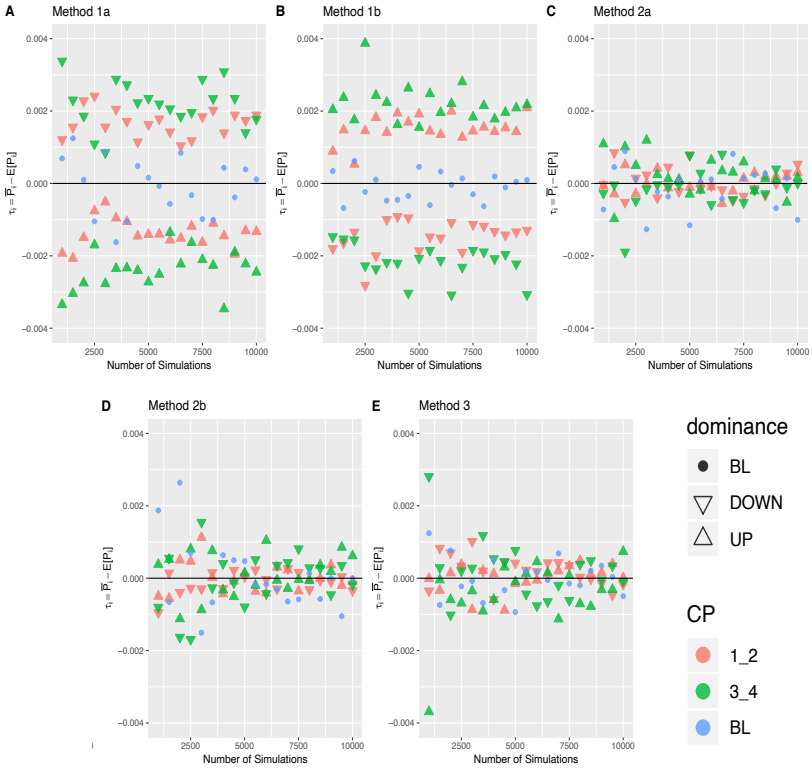


Figure 7: Scatterplots of τ_i across number of simulations for Scenario 5RG

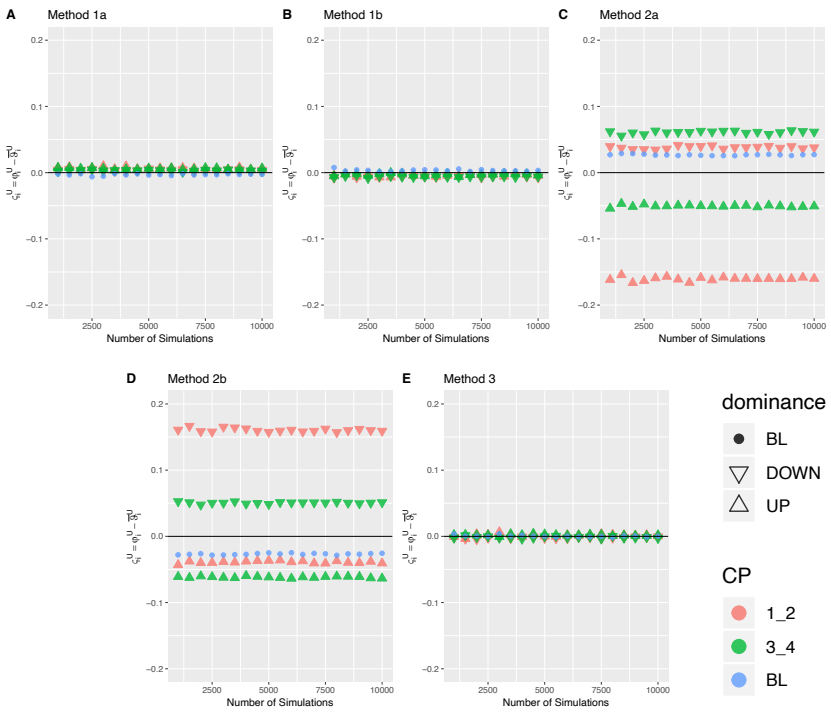


Figure 8: Scatterplots of ς_i^U across number of simulations for Scenario 5RG

The simulation results suggest that the order for selection (either "UP" or "DOWN" first) will introduce bias on the post-sampling population of sampling units for non-baseline groups in violation of bias conditions 1 and 3, thereby offering grounds for eliminating Methods 1a/1b as sampling mechanisms moving forward. On the other hand, the plots in **Figure 8(C)** and **8(D)** suggest that Methods 2a/2b violate the second condition for bias. When units are assigned to be part of an "UP" list first (after they have been selected as part of the composite "UP" and "DOWN" list at a given simulation in accordance with Method 2a), those in the up-dominant group are not assigned to the "UP" list as often as expected and the opposite is true for the down-dominant group; they are assigned to the "UP" list with greater frequency than expected. This relationship is completely reversed when units are assigned to a "DOWN" list first as mandated by Method 2b. In contrast, **Figure 8(A) (**Method 1a), **8(B) (**Method 1b) and **8(E) (**Method 3)

suggest that the corresponding methods perform well with regard to distributing the sampling units to either “UP” or “DOWN” lists in a predictable and non-biased manner.

Although Methods 1a/1b and Methods 2a/2b perform well across one dimension of non-biasedness, Method 3 stands out as the best performer across both dimensions. The design of the simulated data allows the post-sampling distribution of units to be inspected on a group-by-group basis and describe group-based trends and patterns of bias that could be difficult to identify if samplings units were considered on the individual level. The simulations were conducted across a range of different grouping scenarios in order make the general observation that Method 3 does not oversample/under-sample units that tend to be either up- or down-regulated to varying degrees. Now that a method for sampling has been chosen, we will explore the effect of list size (i.e. number of sampling units that will be included in lists) on the power of the proposed test statistics before looking at performance in permutation procedures with the LINCS L1000 data.

2.8 Simulated test statistics and Power Analysis

Simulation studies harness the power of computers to create data sets under pseudo-random sampling conditions that are defined/known to the experimenter in order to discern properties of a test statistic by observing its empirical distribution, possibly, as in our case, over a range of hypothetical scenarios [23]. Experimenters can compare the performance of methods that are either intended to generate or analyze data. In the previous section, outcome measurements pertaining to the simulated data were brought into question in order to identify the data generation mechanism that was least biased among the competing methods. In the section that follows, the ‘thresholding’ parameter (as defined below) will be varied in order to examine the performance of our proposed measurements of association with regard to their power to detect an effect under the assumption that the null hypothesis is not true.

Recall the following parameters:

$$n_U = \text{number of units in an "UP LIST"}$$

$$n_D = \text{number of units in a "DOWN LIST"}$$

$$n = n_U + n_D$$

These 'threseholding' parameters were held constant during our assessment of the data generating mechanisms ($n_U = n_D = 50;\ n = 100$).

The thresholding and discretization employed to transform the Level 5 CGSs to Level 6 CGSs is hierarchical in nature. The thresholding happens first and involves ranking the genes by the continuous-valued mod-z scores and selecting the most positive and most negative among them; up until now we have only discussed selecting genes with the 50 most positive and 50 most negative mod-z scores. The second step is the discretization of values for those genes that were retained from the original ranked list. Thresholding the data in this way is intended to filter out the effects of 'noisy' genes and 'noisy' gene behavior so that the analyses that utilize this data have more power to detect meaningful associations among elements of the data set.

2.8.1 Measurements of Association and Concordance

The purpose of the simulation study is to find the distribution of our proposed test statistics [for finding both differences and similarities among concordance measurements between different cell lines] when the sampling distributions for the primary units are defined rather than estimated from the data. The goal of the simulation study is to answer the following question: "Are the association measurements reliably able to detect dissimilarity (or similarity) in the downstream effects between two perturbating factors in one cell line vs. the other above and beyond what would be considered 'background gene discordance (or concordance)'"? The term

'background gene discordance' is meant to imply that individual L1000 genes do not have uniform probabilities for up/down list membership across perturbating factors among cell lines. The test statistics to compare measurements of association between two cell lines (A and B) at two perturbating factors (X and Y) are:

$$\sum overlaps_{XY}^{AB} \text{ and } \widehat{\Delta}_{XY}^{AB} = \frac{log(\widehat{\theta}_{XY}^{A}) - log(\widehat{\theta}_{XY}^{B})}{\sqrt{(SE_{XY}^{A})^2 + (SE_{XY}^{B})^2}}$$

The formulae for $\widehat{\theta}_{XY}^{A}$ and SE_{XY}^{A} are given in equations (2) and (3); note that under the null hypothesis $\widehat{\Delta}_{XY}^{AB} \sim N(0,1)$ when A and B have identical distributions. The statistic $\widehat{\Delta}_{XY}^{AB}$ is largely a measurement of difference; that is the magnitude of $\widehat{\Delta}_{XY}^{AB}$ in either a positive or negative direction would suggest differences in concordance between A and B at X and Y. On the other hand, we would expect that $\sum overlaps_{XY}^{AB}$ to be larger when the downstream effects of X and Y are more similar for A and B.

In the simulation study, we will observe the behavior of these test statistics when A and B are known to have different distributions by generating their empirical distribution when A and B fall under different regulatory groupings as described in **Section 2.5.3**. We will explore how varying the thresholding parameter affects the distribution of the test statistics and then conduct a power analysis in order to assess the possible benefits or drawbacks of increasing the limit of data considered for these types of tests.

2.8.2 Procedure for Simulating Edges

In order to compare concordance measurements between RGs by comparing edges, we first need to generate nodes. 2000 nodes (reflecting the number of perturbagens in the LINCS data set) will be generated according to Method 3 for each RG (as well as the Null Scenario) as outlined in **Section 2.3.2** across the following thresholding parameter t: 10, 20, 30, 40, 50, 60,

70, 80, 90, 100, 150, 200, 250, 300, 350, 400 and "complete". t represents the value $\frac{n_U + n_D}{2}$ subject to the constraint $n_U = n_D$; i.e. the length of either "Up" or "Down" list with the exception of the "complete" scenario. The "complete" parameterization, as the name implies, uses the complete set of sampling units by carrying out only Step 1 of the algorithm for Method 3 to and label each unit as "Up" or "Down"; all of the units labeled "Up" will be part of the L_U^S and all units labeled "Down" will be part of the L_D^S. On average we would expect $n_U = n_D = 500$ for the complete scenario but note that these values are actually random variables subject to the constraint $n_U + n_D = 1000$.

Let C^{Rt} be the collection of simulated nodes for a given RG scenario (R = "Null", 3, 5, 7, 9) at threshold t where C_i^{Rt} is an individual node $(i = 1, 2, \ldots, 2000)$. Then, the following algorithm will be used to generate n edges:

Input: Nodes C^{Rt} and number of edges n to generate

Initialize a data frame with n rows for results

For $1:n$:

$R \leftarrow$ integers from 1: $|C^{Rt}|$

$r^1 \leftarrow$ random sample of 1 integer from R

$R' \leftarrow R - r^1$

$r^2 \leftarrow$ random sample of 1 integer from R

Derive contingency table for $C_{r^1}^{Rt}$ and $C_{r^2}^{Rt}$ and store as edge $E_{r^1 r^2}^{Rt}$ in row n

Calculate odds ratio and SE for edge $E_{r^1 r^2}^{Rt}$ and store values in row n

In order to be able to calculate both measures of association between different RG's, the minimally sufficient data for each edge is the list of units that populate each cell of the

contingency table. n = 5000 edges will be generated for each of the five scenarios (four RG's and the Null). The value for n has been chosen to provide stable results for estimates of association between RG's but also to reflect the number of unique KEGG edges that will eventually be utilized for the analysis of the L1000 data set.

2.8.3 Generating Measurements of Association in Simulated Data

The previous section demonstrates the procedure for producing single edges with regard to parameters specific to a given RG. The next step is to generate data that will allow us to calculate outcome measurements that compare our pairwise measurements of association (the edge vs. edge summary statistics $\sum overlaps_{XY}^{AB}$ and $\widehat{\Delta}_{XY}^{AB}$ ("sum of overlaps" and "delta" values)). In the simulated data, the indexes "A" and "B" represent the RG scenario as opposed to cell lines. The indexes "X" and "Y" represent, within a given RG, two different nodes (among the 2000 simulated nodes) form any edge X|Y. Let E^{Rt} be the collection of simulated edges for a given RG scenario (R = "Null", 3, 5, 7, 9) at threshold t where E_i^{Rt} is an individual edge ($i = 1, 2, \ldots, 5000$). There will be two relevant sets of edge-based comparisons: each RG vs. the Null and each RG vs. itself (includes Null vs. Null) (referred to as the RG vs. RG scenario). The algorithm below will be used to derive measurements of association for the RG vs. Null at each threshold t:

Input: Edges $E^{R_j t}$ and $E^{R_{null} t}$;

Initialize a data frame with n rows for results

For i = 1: n:

Calculate $\sum overlaps$ and delta values between $E_i^{R_j t}$ and $E_i^{R_{null} t}$.

In the RG vs. RG scenario, the two edges for comparison are sampled in a bootstrap fashion from the same set $E^{R_j t}$ as follows:

Input: Edges $E^{R_j t}$

Initialize a data frame with n rows for results

For $1:n$:

$R \leftarrow$ integers from 1: $|E^{R_j t}|$

$r^1 \leftarrow$ random sample of one integer from R

$R' \leftarrow \ R - r^1$

$r^2 \leftarrow$ random sample of one integer from R

Calculate total overlaps and delta values between $E_{r^1}^{R_j t}$ and $E_{r^2}^{R_j t}$.

2.8.4 Distribution of Measurements of Association in Simulated Data

Now that the data has been simulated, the first task at hand is to check our distributional assumption regarding $\widehat{\Delta}_{XY}^{AB} \sim N(0,1)$ that should hold when A and B are the same. We can check this assumption by observing the behavior of $\widehat{\Delta}^{AB}$, that is, $\widehat{\Delta}_{XY}^{AB}$ across many random edges in the RG vs. RG scenario as we would expect this to reflect the null hypothesis of no difference. Graphical inspection along with summary measurements of the distributions (mean, median, standard deviation) will be utilized to verify the assumption of standard normality for $\widehat{\Delta}^{AB}$.

The plots in **Figure 9** show the density of $\widehat{\Delta}^{AB}$ as the threshold for list membership increases. When the units sampled come from the same distribution, in this example either both sampled under the "null" scenario or "RG3" scenario, $\widehat{\Delta}^{AB}$ approaches the standard normal distribution as the threshold is increased beyond a lower limit of 20 units per up and per down list (i.e. at least 30). The distributions corresponding to $\widehat{\Delta}^{AB}$ for list lengths of 20 and below fail

to converge to the approximate the standard normal curve due to the limited number of values that are possible realizations of $\widehat{\Delta}_{XY}^{AB}$ as evidenced by multiple peaks (**Figure 9: A1, B1, C1**). On the other hand, as the threshold increases to 50 and above, the distributions of $\widehat{\Delta}^{AB}$ approach the standard normal distribution (**Figure 9: A2, B2**).

The evidence for standard normality under the null as the threshold approaches 50 is echoed in **Table 12**. Although a higher threshold means little in terms of more proximity to a mean of zero, the standard deviations reported in **Table 12** suggest that list lengths of 60 and above produce $\widehat{\Delta}^{AB}$ distributions that approach the hypothetical $N(0,1)$.

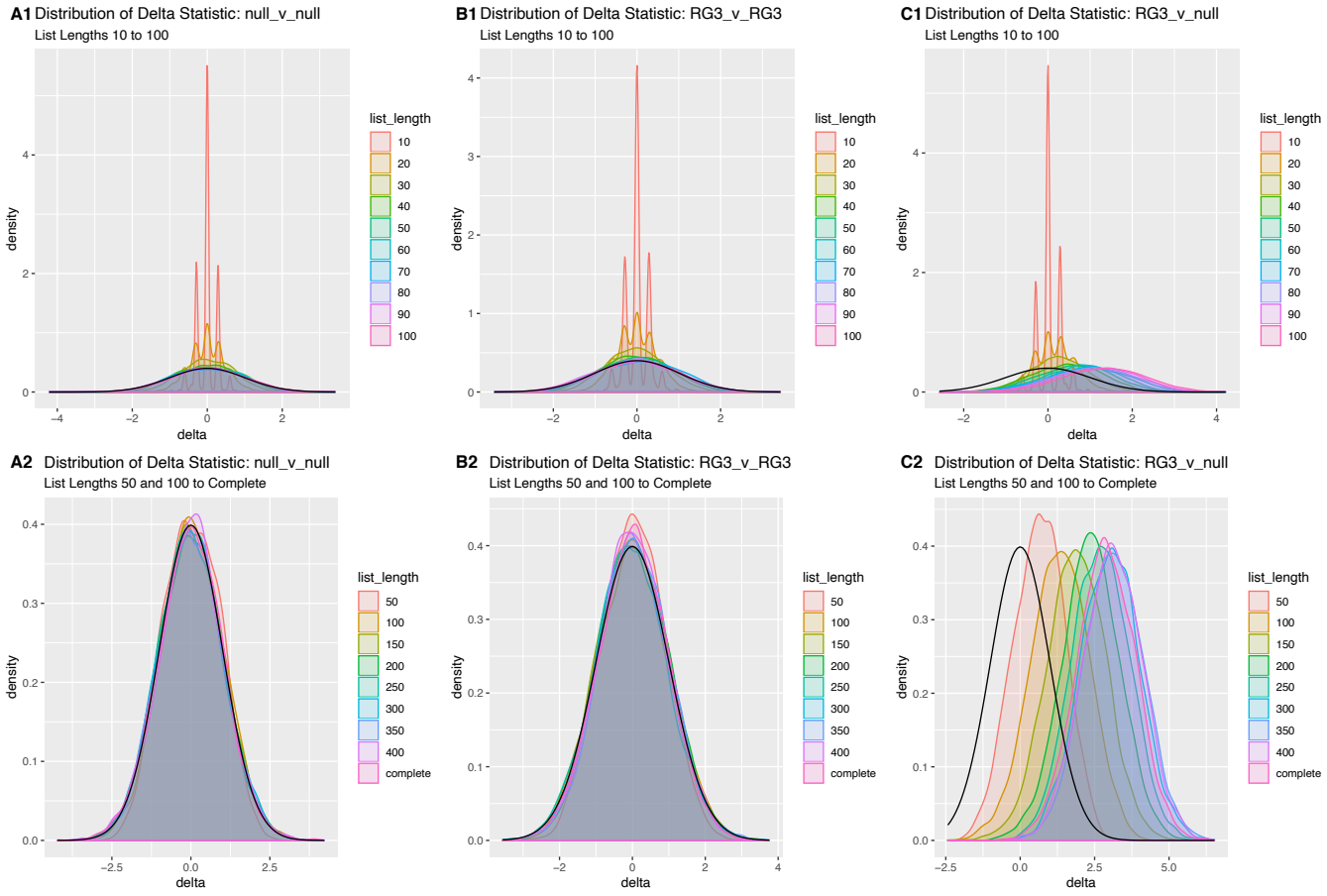


Figure 9: Distribution of $\bar{\Delta}$ across range of thresholds; from 10 – 100 in increments of 10 (A1, B1, C1) and from 50-400 plus the 'complete' scenario in increments of 50 (A2, B2, C2). The black solid line represents the standard normal curve.

The other take away from **Figure 9** is that under the alternate hypothesis, the mean of $\widehat{\Delta}^{AB}$ moves away from zero as the threshold increases. The same plots [not shown] for the RG5, RG7 and RG9 are nearly identical to those for RG3. The power analysis in the next section will address the following question: is there a point at which an increase in the threshold fails to yield additional power to detect a difference under the alternate hypothesis?

Table 12: Mean and standard deviation for distribution of $\hat{\Delta}^{AB}$ across range of thresholds for RG3.

List Length	mean	sd
10	0.003	0.275
20	-0.013	0.488
30	-0.031	0.691
40	-0.021	0.815
50	0.004	0.860
60	0.015	0.926
70	0.079	0.966
80	0.018	0.937
90	-0.060	0.995

List Length	mean	sd
100	0.026	0.975
150	-0.013	0.968
200	0.039	0.965
250	-0.037	0.981
300	0.003	0.964
350	0.061	0.955
400	0.014	0.940
complete	0.010	0.953

Whereas the test statistic Δ has a known parametric null distribution, the ad hoc test statistic $\sum overlaps$ does not have a hypothetical or theory-based null distribution. Since the purpose of this test statistic is to identify a degree of similarity between two edges that would be higher than that expected under the null hypothesis of no association, the RG vs. Null scenario will serve as the distribution for the test statistic under the null and the RG vs. RG scenario will be treated as the distribution under the alternate hypothesis of presence of association. **Figure 10** depicts the distribution of $\sum overlaps$ under the RG3 vs. Null scenario, and for comparison the Null vs. Null scenario along with the RG3 vs. RG3 as the alternate scenario.

The histograms in **Figure 10** highlight two important concepts. The first is that the threshold for list length must be greater than 100 in order to obtain a distribution that is not

dominated by zero counts in either the null or alternate scenarios. The second is that under the alternate hypothesis the distributions shift to the right; in other words, when we know that the sampling units come from similar distributions, we would expect to find more units that occupy the same cell under the alternate vs. the null scenario.

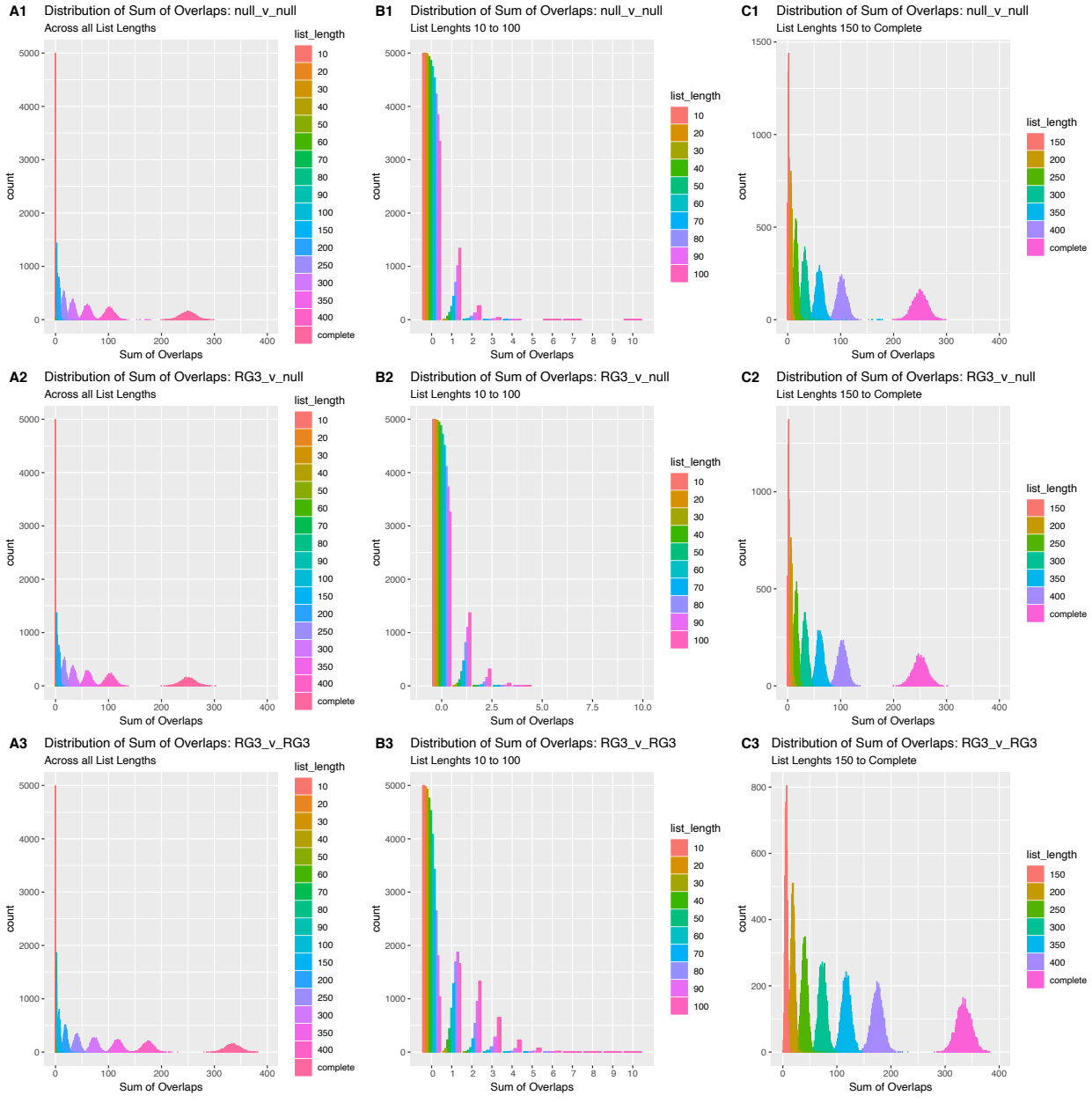


Figure 10: Distribution for the test statistics $\sum overlaps^{null,null}$, $\sum overlaps^{RG3,null}$ and $\sum overlaps^{RG3,RG3}$ cross all thresholds (**A1, A2, A3**), thresholds below 100 (**B1,B2,B3**), and thresholds above 100 (**C1,C2,C3**) under the null (top and middle) and alternate (bottom) hypothesis.

2.8.5 Power Analysis

The distributions that were introduced in Section **2.7.3** for $\sum overlaps_{XY}^{AB}$ and $\widehat{\Delta}_{XY}^{AB}$ will now be employed to conduct a power analysis of these proposed test statistics. Typically, when data is arranged in a contingency table as shown in **Table 13**, the null hypothesis of independence is that the characteristic *A* is equally distributed between the two groups in population P [24]. Under the alternate hypothesis, the "characteristic" column or "response" variable is *dependent* on the row or "explanatory" variable. In a balanced design, the marginal row totals n_{1+} and n_{2+} are equal and considered fixed whereas the marginal column totals are random variables whose value is determined by the outcome of the experiment [25]. The power of the test for independence between the row and column variable is increased by taking a larger sample from each subpopulation. In other words, if *N* is the total number of individuals in the population of interest, then the power increases as n/N increases.

Table 13: Traditional 2x2 contingency table

Population P		Characteristic		Marginal Row Totals
		A	$\overline{A}$	
Subpopulation	Group 1	n_{11}	n_{12}	n_{1+}
	Group 2	n_{21}	n_{22}	n_{2+}
Marginal Column Totals		n_{+1}	n_{+2}	n

In our analysis, we are operating based on the assumption that our sample includes the *entire* population of interest and that, for each unit in the population, we have a ranking metric that allows us to categorize that unit as "up", "down", or "not significant". As shown in **Table 14**, only units that can be cross classified between two nodes in one of the four mutually exclusive categories are part of an edge and contribute to the cell counts used to calculate

measures of association. The cells are classified as either as either concordant (CC) or discordant (DC) as follows: concordant cells are Q1 = CC:UU (units are "up" in both nodes) and Q4 = CC:DD (units are down in both nodes); discordant cells are Q2 = DC:UD (units that are "up" in node X but "down" in node Y) and Q3 = DC:DU (units that are "down" in node X but "up" in node Y. Note that units that are "down" in Node X but are "not significant" in Node Y (and vice versa) do not contribute to cell counts in the contingency table. In our sampling scenarios, we cannot *directly* manipulate *n*, which we define as the total number of sampling units that make it into a given edge/contingency table. However, we can *indirectly* influence *n* by varying the list length threshold parameter as described in the previous section.

Table 14: 2x2 contingency/directional concordance table for edge X|Y.

Regulatory Group RG		Node Y		Marginal Column Totals
		Up in Y	Down in Y	
Node X	Up in X	**Q1 (CC:UU) =** Units up in X and up in Y	**Q2 (DC:UD) =** Units up in X and down in Y	Units up in X and up or down in Y
	Down in X	**Q3 (DC:DU) =** Units down in X and up in Y	**Q4 (CC:DD) =** Units down in X and down in Y	Units down in X and up or down in Y
Marginal Row Totals		Units up in Y And up or down in X	Units down in Y and up or down in X	Concordant and Discordant Units (*n*)

The rational is as follows: as the threshold is increased, a larger portion of units will be classified as either "up" or "down" instead of "not significant" until the threshold is set to "complete" and all units are "up" or "down" (and zero remain "not significant"). Note that when all units are "up" or "down", all units will fall into the contingency table and *n* will be equal to *N*, the number of units in the population (1000 units in the simulation study and 978 genes in the L1000 data set). In all other ("non-complete") scenarios, $0 \leq n \leq t * 2$. The purpose of the

power analysis is to see how well our proposed test statistics capture the signal of "informative" units/genes against varying degrees of background noise.

2.8.5.1 Power Analysis Results: $\widehat{\Delta}^{AB}$

Under the null hypothesis, on average there should be of no difference between concordance measurements of concordance (ie. E($\widehat{\Delta}^{AB}$) = 0) and the standard deviation approaches 1 (see **Figure 9** and **Table 12**). Under non-null conditions, as we have defined for different RG scenarios, when the proportion of units that fall into the concordant cells (Q1 and Q4) of the contingency table (**Table 14**) is large relative to the discordant cells (Q2 and Q3) in a predictable manner based on the specified per-unit sampling values, it is reasonable to assume that $log\left(\hat{\theta}_{XY}^{RG\neq null}\right) - log\left(\hat{\theta}_{XY}^{RG=null}\right) > 0$. Thus, it seems reasonable to assume that increasing the threshold parameter *t* will increase the likelihood of capturing 'informative' units; that is, units that are weighted to fall into the concordant cells. By the same line of logic, however, increasing *t* will also lead to an increase in the number of units that fall into the discordant cells [or concordant cells] by random chance for any sampling scenario. Is there, in fact, an upper limit or 'tipping point' for the parameter *t* that will result in less power to detect a true difference in the non-null scenarios than lower parameter values?

In order to calculate power across different thresholds, the standard normal distribution serves as the null distribution and the alternate distributions are those described in **Section 2.7.3 (**see **Figure 9**(**C2**)). Power at a given threshold is equal to the percentage of values of the alternate distribution that are greater than the specified quantile of the null distribution for a given alpha level. The dashed line in **Figure 11** represents 80% power to detect a difference when the alternate hypothesis of a difference in concordance is true. In general, all of the graphs

show steady increases in power as the threshold *t* ("List Length", x-axis) increases to 200 after which point only modest gains or even losses in power are made by increasing *t*.

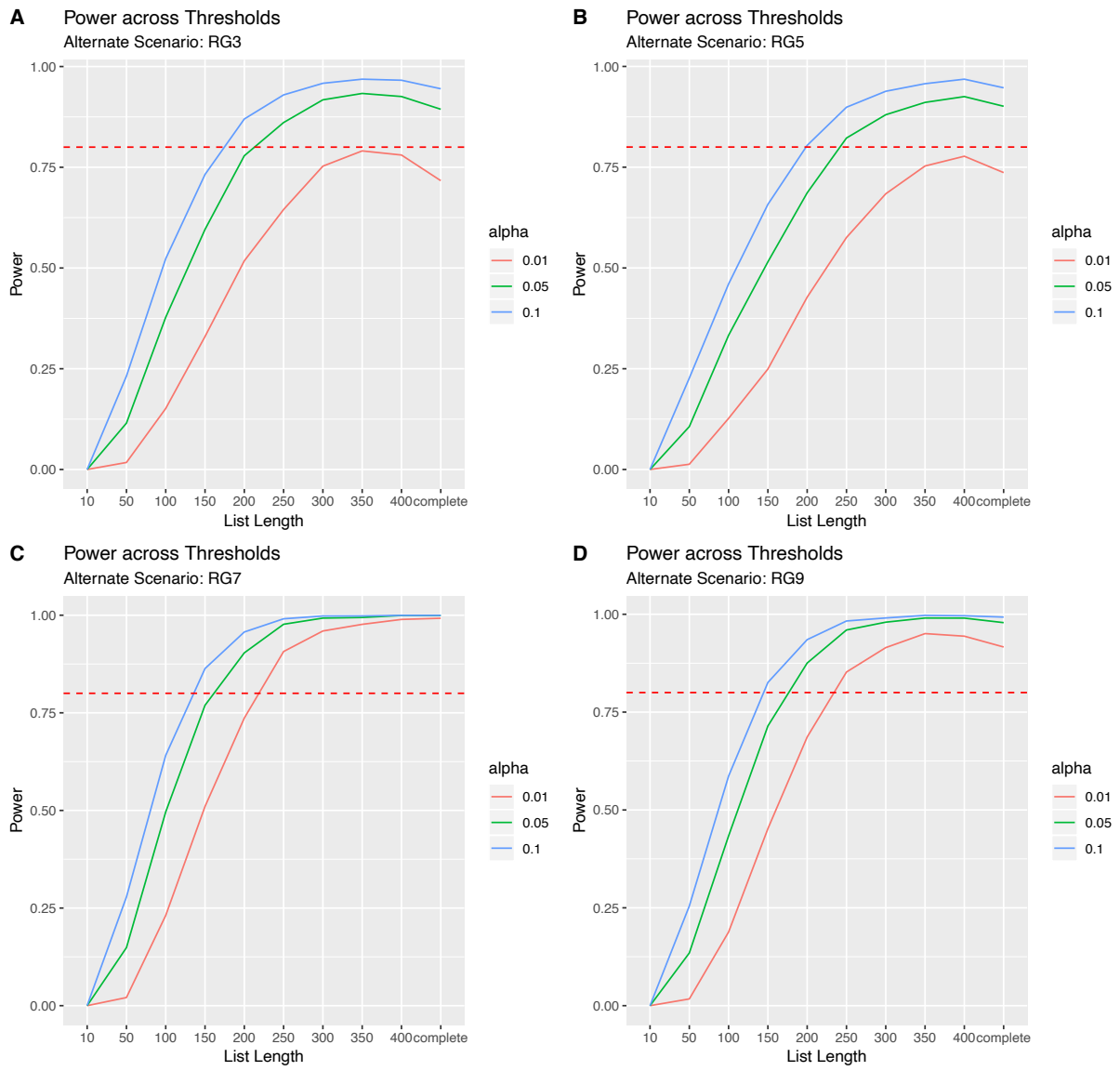


Figure 11: Power curves for Δ across threshold parameter settings for RG3(A), RG5(B), RG7(C), and RG9(D).

For scenario RG3, power decreases when the threshold is increased from 350 to 400 and "complete" across all three alpha levels. In RG5 power only decreases when the threshold goes from 400 to "complete". For RG7, there are no decreases in power across thresholds or alpha

levels. The RG9 scenario shows a very slight decrease in power when the threshold is increased beyond 400 at alpha of 0.05 as well as beyond 400 at alpha of 0.01.

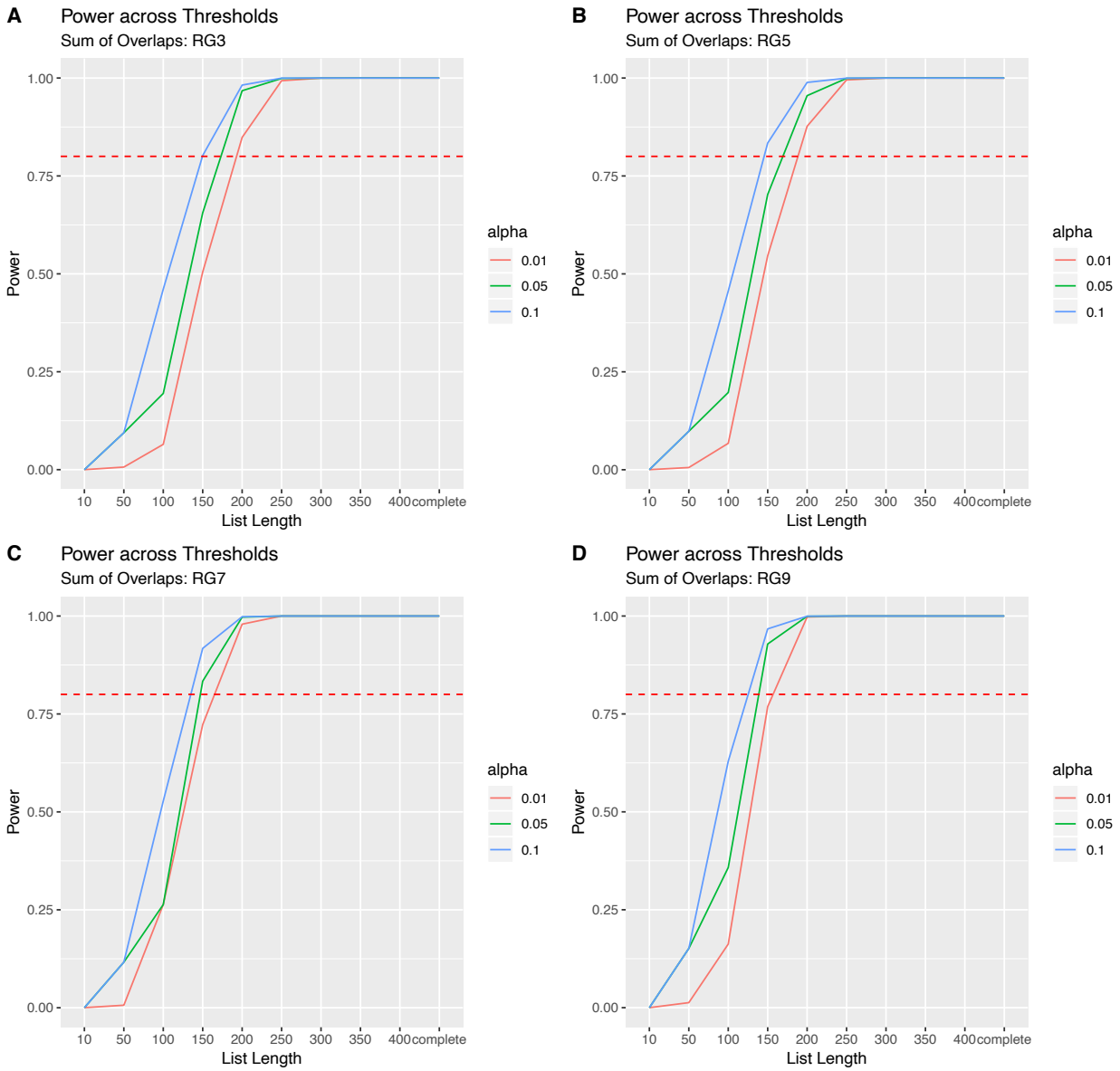


Figure 12: Power curves for $\sum overlaps$ across threshold parameter settings for RG3(A), RG5(B), RG7(C),

2.8.5.2 Power Analysis Results: $\sum overlaps$

The null hypothesis for the test statistic $\sum overlaps$ is more nuanced than that of Δ; the null distribution is obtained by comparing the counts of sampling units that show up in the same cell from a specified distribution (RG3, RG5, RG7, RG9) to a random distribution (null scenario)

instead of a known, hypothetical null such as the standard normal distribution. Power is calculated as the percentage of observations in the alternate distribution (RG vs. RG) that are greater than those of the null (RG vs. null) for quantiles at the three different alpha levels. Whereas the null distribution is identical for Δ across threshold values, there is a different null distribution at each threshold for $\sum overlaps$. The distributions at each threshold level are employed to take into consideration the fact that there will inevitably be an increase in overlapping units by random as the threshold goes up.

The curves in **Figure 13** are similar to those in **Figure 11** with regard to steady increases of power up to the threshold limit of 200. In contrast, there is not a threshold beyond which power decreases and, conversely, there are values of t for which an increase does not yield any additional power (i.e. power plateaus). For RG3, RG5, and RG7, power plateaus at t = 250, whereas for RG9 this value is t = 200.

2.8.6 Summary of Power Analysis

The plots in **Figure 14** allow for comparison of the power under different scenarios at a specified level of alpha and those in **Figure 15** are included as a visual guide to aid in possible explanations for differences in power amongst the scenarios and test statistics. For both Δ and $\sum overlaps$, the RG7 and RG9 scenarios attain higher levels of power across thresholds and values of alpha. A possible explanation for the relatively higher power to detect differences when compared to the null in scenarios RG7 and RG9 versus RG3 and RG5 is that a smaller proportion of units belong to the "baseline" group (i.e., the group for which p(UP) = p(DOWN)) in the RG7 and RG9 scenarios. The same rationale could underlie the slightly higher power to detect similarities in the RG7 and RG9 scenarios.

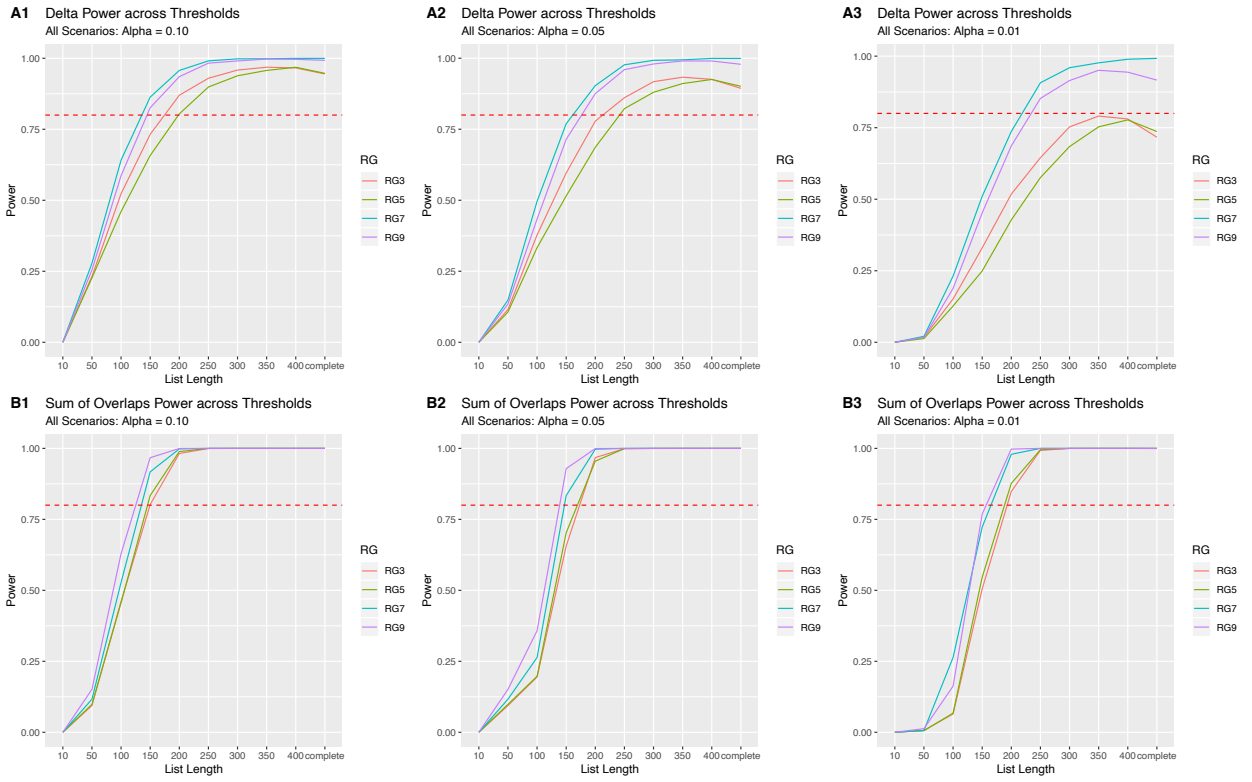


Figure 14: Power curves for alpha set to 0.1 (A1: Δ, B1: $\sum overlaps$), 0.05 (A2: Δ, B2: $\sum overlaps$) and 0.05 (A3: Δ, B3: $\sum overlaps$

The results of the power analysis suggest that a threshold of 50 units per up/down list may be insufficient to detect either differences or similarities between two entities. In order to obtain the maximum power to detect differences and similarities in the simulated scenarios, a threshold setting of 200 units per up/down list is required. Now that we know how these parameters function in a simulated setting, we will explore how the same type of analysis plays out in data specific to patterns in the L1000 data set in the next chapter.

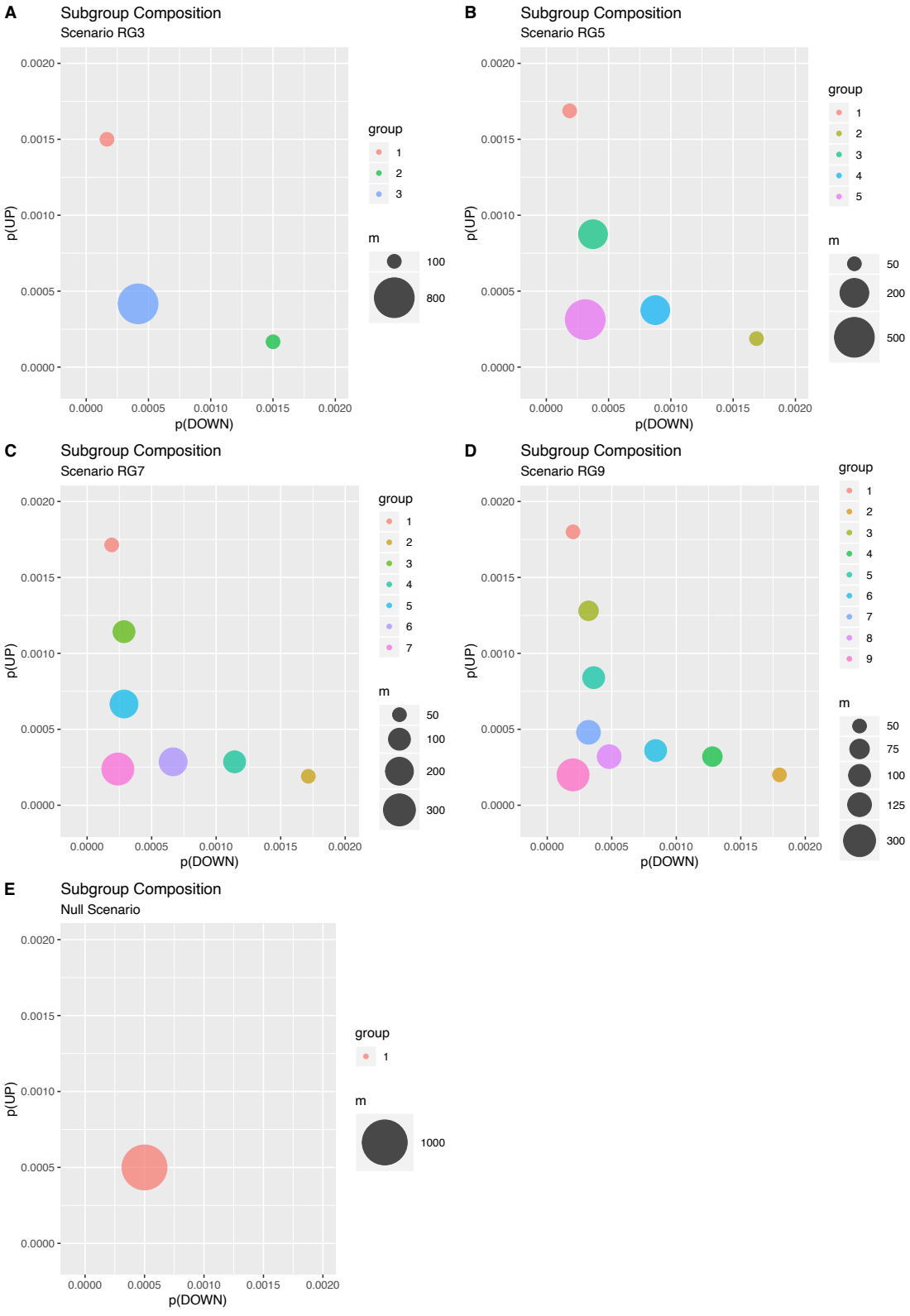


Figure 15: Size and initial p(UP) vs p(DOWN) properties for the RG3(A), RG5(B), RG7(C), and RG(9) scenarios as well as the null scenario.

Chapter 3: L1000 Power Analysis

Section 2.1 introduced a concept that will be of central importance going forward and that is the concept that the L1000 genes exhibit behavior that may be described with regard to two different dimensions; those dimensions being direction of regulation as well as cell line-specific regulation. Different patterns of gene regulation between cells are inevitable given the complex layers of epigenetic regulation even between cells of the same genotype [26]. In the L1000 data set, there are multiple cancer cell lines that are known to have genotypic differences but are similar with regard to experimental protocol; the seven core cell lines used in this analysis have over two thousand common shRNA perturbations. The analyses presented in this chapter will employ methods described in the previous chapter to model background patterns of gene regulation among different cell lines in order to identify differences and similarities between cell lines in the L1000 data set.

3.1 Motivation: Heterogeneity Among Cell Lines

The term "heterogeneity" takes on different meanings in the context of statistics and biology; here we are using it to describe both. When we make the claim that there is heterogeneity among cell lines, we a referring to differences in global gene regulation between different cancer-cell lines across perturbating factors. This difference will be quantified and then taken into consideration in the follow up analyses. Before moving on to the power analysis, a preliminary example of heterogeneity among the seven core lines will be presented to serve as a starting point for motivating the considerations that are being made when handling this diverse data set.

The data set provided by ilincs.org [27] contains Level 5 consensus genomic signatures CGSs) for 2007 common perturbating factors across seven core cell lines. Each cell line-

perturbating-factor combination contains a record of modZ scores for the 978 LM genes as well as the p-values associated with the significance of the modZ score (accounts for variation across different shRNAs).

Let $\boldsymbol{M}^{\boldsymbol{cl}}_{\boldsymbol{978\times2007}}$ be the matrix of modZ scores for cell line *cl.*

Let $\boldsymbol{P}^{\boldsymbol{cl}}_{\boldsymbol{978\times2007}}$ be the matrix of modZ p-values for cell line *cl*.

Then, the entries $m^{cl}_{lm,pf}$ and $p^{cl}_{lm,pf}$ refer to the modZ score and p-value for landmark gene *lm* (*lm* = 1,…, 978) at perturbating factor *pg* (*pg* = 1,…, 2042) within cell line *cl* (*cl* = 1,…, 7) from matrices **M** and **P.** The ranges and a few summary measurements of $m^{cl}_{lm,pf}$ and $p^{cl}_{lm,pf}$ are reported in **Table 15.** To be clear, a negative modZ score indicates that transcripts corresponding to a landmark gene are fewer in number compared to baseline conditions at a given perturbating-factor-cell line combination and a positive score means that there are more counts of the transcript in the perturbed vs. baseline condition. In this data set, there are also modZ scores of zero that reflect no significant difference between counts in the perturbed condition versus baseline.

Table 15: Summary measurements from matrices **M** and **P** across perturbagens for each cell line.

	Lowest modZ	Highest modZ	Mean modZ	Zero modZ Count	Smallest p–value	Largest p–value
A375	–10	10	–0.0037	94	1.1742881131042e–42	1
A549	–10	8.1029	–0.0244	139	1.39192301536461e–28	1
HA1E	–9.4864	8.0922	–0.0067	76	8.47284045075094e–35	1
HEPG2	–10	9.5294	0.0039	68	2.29157275174701e–27	1
HT29	–9.5979	10	0.004	129	1.02437885822332e–33	1
MCF7	–10	8.1447	–0.0187	1816	2.92367068286224e–21	1
PC3	–10	7.8563	–0.0229	1639	1.36918760797338e–32	1

The overall (across cell line) range of modZ values is [-10, 10] although the A375 (melanoma) cell line is the only entity with scores that cover this complete range. Perhaps not

coincidentally it is also the cell line with a mean modZ closest to zero followed by HEPG2 which has the second largest range. In comparison to the other cell lines in the panel, MCF7 and PC3 have modZ scores that equal zero at rates that are at least ten times that of other cell lines. The purpose for reporting the minimum p-value is not so much to show that there are differences between cell lines for this measurement but rather that there are no zero p-values; an aspect of this data set that will be important in the construction of custom-length thresholded CGSs (and in turn CSs (concordance signatures)) in the following section. It suffices to say that descriptive numerical summaries of landmark gene behavior across perturbating factors within a given cell line suggest that there may be differences in the magnitude of differential gene expression between cell lines. The remainder of this chapter will focus on how heterogeneity between cell lines may be accounted for in the analysis of the L1000 data as well as whether or not the heterogeneity might affect the interpretation of the downstream results.

3.2 Generating Data for the Power Analysis of L1000 Data

The power analysis of simulated data suggested that the proposed test statistics have little power to detect differences or similarities between two entities when the threshold parameter is set at the default value of 50 units per up/down list. In the simulated data set, although this finding was consistent across RG scenarios, different scenarios did yield consistently different (higher or lower) power estimates across threshold and alpha settings. In this section we will describe how we will generate data to conduct a parallel analysis of the L1000 data set.

The L1000 data power analysis will require cell line specific edge distributions. Specifically, three types of pair-wise edge distributions will be generated as part of the power analysis: *simulated*, *random* and *KEGG*. Simulated edges will be constructed in a manner similar to those for the specific regulatory groups in chapter 2; measurement for differential

expression of L1000 gene behavior across perturbating factors at each threshold level will be part of the "Method 3"-type data generating process. Random edges and KEGG edges will be synthesized from Level 6 CGSs in their original form across the different threshold levels. The procedures for obtaining Level 6 CGSs and then generating the different types of edges will be discussed in the remainder of this section.

3.2.1 Custom-Thresholded Level 6 CGSs

The manipulation of the thresholding parameter to produce new Level 6 CGSs will require a new data set for their de novo construction. The p-values are the primary ranking mechanism for determining a perturbating factor's inclusion at a given threshold level and the sign of their respective modZ scores accounts for the directionality of the dysregulation.

A

	A2M	AARS	AATF
AARS	0.81074	-2.41592	0.80352
ABCB6	-0.00450	0.41099	0.10218
ABCC5	1.01039	-0.08585	2.65974
ABCF1	-0.69065	-0.09388	0.05886
ABCF3	-0.91045	-0.53804	-0.15166
ABHD4	0.67756	-0.32756	-0.31919
ABHD6	0.45756	-0.49667	-0.34841
ABL1	1.16341	1.19339	-0.04573
ACAA1	0.08618	0.15777	-0.09744
ACAT2	-0.71894	-0.19292	-1.53479
ACBD3	-0.69403	-0.32066	-0.05265
ACD	-1.13797	-0.01979	-0.07719
ACLY	-0.04495	0.24140	1.02140
ACOT9	0.95737	0.17592	-0.67414
ADAM10	0.66620	0.24147	0.06493
ADAT1	0.93440	-0.22183	0.05103
ADCK3	0.17065	0.94122	-0.23139
ADGRE5	-0.47768	-0.32315	0.63217

B

	A2M	AARS	AATF
AARS	0.15680	0.00800	0.13306
ABCB6	0.99336	0.12786	0.80813
ABCC5	0.16130	0.68513	0.02934
ABCF1	0.22189	0.68865	0.87019
ABCF3	0.09911	0.03928	0.73001
ABHD4	0.28476	0.12343	0.50515
ABHD6	0.44760	0.01986	0.36238
ABL1	0.05349	0.04780	0.90906
ACAA1	0.87318	0.60504	0.81431
ACAT2	0.25072	0.33448	0.24359
ACBD3	0.20587	0.07201	0.88270
ACD	0.08608	0.91563	0.83778
ACLY	0.93264	0.26905	0.03203
ACOT9	0.09117	0.28834	0.12556
ADAM10	0.24608	0.24034	0.88458
ADAT1	0.12719	0.39924	0.90534
ADCK3	0.76209	0.03093	0.56850
ADGRE5	0.37103	0.13101	0.17562

C

	A2M	AARS	AATF
AARS	0.15680	-0.00800	0.13306
ABCB6	-0.99336	0.12786	0.80813
ABCC5	0.16130	-0.68513	0.02934
ABCF1	-0.22189	-0.68865	0.87019
ABCF3	-0.09911	-0.03928	-0.73001
ABHD4	0.28476	-0.12343	-0.50515
ABHD6	0.44760	-0.01986	-0.36238
ABL1	0.05349	0.04780	-0.90906
ACAA1	0.87318	0.60504	-0.81431
ACAT2	-0.25072	-0.33448	-0.24359
ACBD3	-0.20587	-0.07201	-0.88270
ACD	-0.08608	-0.91563	-0.83778
ACLY	-0.93264	0.26905	0.03203
ACOT9	0.09117	0.28834	-0.12556
ADAM10	0.24608	0.24034	0.88458
ADAT1	0.12719	-0.39924	0.90534
ADCK3	0.76209	0.03093	-0.56850
ADGRE5	-0.37103	-0.13101	0.17562

Figure 16: Cell line A375 modZ scores (**A**), p-values (**B**) and signed p-values (**C**) for the first twenty [rows; arranged alphabetically] L1000 LM genes across three perturbagens [columns].

The first step for generating custom-thresholded Level 6 CGSs is to obtain a metric by which to rank the L1000 LM genes within each cell line-perturbagen combination as follows:

Let $\boldsymbol{S^{cl}_{978\times 2007}}$ be $sign(\boldsymbol{M^{cl}_{978\times 2007}})$; the sign matrix of the modZ scores for cell line *cl.*

Then, $\boldsymbol{S^{cl} \odot P^{cl} = R^{cl}}$ where $\boldsymbol{R^{cl}}$ is the Hadamard product (entry-wise multiplication) of the sign and p-value matrix [28]. Each column of $\boldsymbol{R^{cl}}$ is an independent record of directional ranking metrics that will be sufficient for generating custom-thresholded Level 6 CGSs for each perturbagen-cell line-combination. **Figure 16(C)** is an example of the first twenty rows and three columns of $\boldsymbol{R^{A375}}$.

Note that some entries in $\boldsymbol{S^{cl}}$ are equal to zero; these number of these entries for given cell line are reported in the column labeled "Zero modZ Count" in **Table 15**. The zero value for these entries will result in an $\boldsymbol{R^{cl}}$ entry of zero. When an entry in $\boldsymbol{R^{cl}}$ is zero, that particular LM gene will be ineligible for a position in either an up or down CGS at any threshold. With the exception of entries equal to zero, the closer an entry's absolute value is to zero, the higher its relative (column-wise) rank within its directional category; all positive entries are eligible for membership in an "up" list and all negative entries are eligible for membership in a "down" list. With $\boldsymbol{R^{cl}}$ in place, the procedure for generating Level 6 CGSs at each cell line-perturbagen combination at each threshold level *t* is as follows:

Let $\boldsymbol{r^{cl,p}}$ be the pth column of $\boldsymbol{R^{cl}}$; the 978 signed p-values for perturbagen *p* in cell line *cl*. $\boldsymbol{r^{cl,p}}$ will then be partitioned as follows:

- $\boldsymbol{up^{cl,p}}$ contains all LM genes whose entries of $\boldsymbol{r_p^{cl}} > 0$ ranked in order of smallest value to largest value.
- $\boldsymbol{down^{cl,p}}$ contains LM genes whose entries all entries of $\boldsymbol{r_p^{cl}} < 0$ ranked in order of smallest absolute value to largest absolute value.

Then, at each threshold level *t,* the Level 6 CGS is the combined vector $< \boldsymbol{up_{1:t}^{cl,p}}, \boldsymbol{down_{1:t}^{cl,p}} >$ for each cell line-perturbagen combination.

3.2.2 Simulated L1000 Edges

A set of m = 5000 simulated edges will be generated from a set of n = 2000 simulated nodes at each threshold level t for each cell line. The values for m and n are chosen to be representative of the approximately 5000 unique edges from KEGG pathways whereby both nodes correspond to perturbagens in the L1000 data set [11]. The simulation procedure is intended to reflect what would be expected if the L1000 LM genes were behaving in a random but cell line-specific manner as predicted by their marginal counts in up and down thresholded CGSs.

The preliminary step in this procedure is to look across the 2007 cell line-perturbagen-specific CGSs at each threshold (whose generation is detailed in **Section 3.2.1**) in order to obtain probability weights similar to those that were introduced in **Section 2.3.1**. However, in this scenario, each of the 978 LM genes will be associated with unique up/down selection weights and probabilities instead of weights and probabilities specified by membership in a specific regulatory group. Recall matrix **L**, an indicator matrix for LM gene CGS membership first presented in **Figure 4**. This matrix is an indicator for LM gene membership in CGSs for the default list length [*threshold*] of 50 genes up and down (978-100 = 878 not differentially regulated). The notation used in **Section 2.1** will be modified to describe the parameterization of the cell line-specific null distribution across different threshold parameter settings as follows:

Let $\boldsymbol{L}^t$ be a 978×2007 matrix defined by the following indicator function:

$$I_{ij}^{cl,t} = \begin{cases} -1, & gene\ g_j \in CGS\_D_i^{cl,t} \subseteq CGS_i^{cl,t} \\ 0, & gene\ g_j \notin CGS_i^{cl,t} \\ +1, & gene\ g_j \in CGS\_U_i^{cl,t} \subseteq CGS_i^{cl,t} \end{cases}$$

With these I's in place in $\boldsymbol{L}^t$ we may readily calculate the following probabilities:

- $pu_j{}^{cl,t} = \frac{\sum_{i=1}^{2,007}(I_{ij}^{cl}=+1)}{2007}$ = probability that g_j will be one of *t* genes in a random $CGS\_U_i^{cl,t}$;
- $pd_j{}^{cl,t} = \frac{\sum_{i=1}^{2,007}(I_{ij}^{cl}=-1)}{2007}$ = probability that g_j will be one of *t* genes in a random $CGS\_D_i^{cl,t}$;
- $pn_j{}^{cl,t} = 1 - pu_j{}^{cl,t} - pd_j{}^{cl,t}$ = probability that g_j is neither part of a random $CGS\_D_i^{cl,t}$ nor $CGS\_U_i^{cl,t}$.

Note that in the "complete" scenario all genes will be either part of an up or down list and therefore $pn_j{}^{cl,complete} = 0 \; \forall \; i,j$ and $|CGS\_U_i^{cl,complete}|$ and $|CGS\_D_i^{cl,complete}|$ are not necessarily equal [although they could be equal by chance] but should be similar in size. With these probabilities in place for each gene, we can derive the following gene-specific sampling weights for cell line *cl* at threshold *t*:

$\pi_j{}^{cl,t} = 1 - pn_j{}^{cl,t} =$ probability that g_j is in any randomly selected up or down list

$\varphi_j^{U,cl,t} = \frac{pu_j{}^{cl,t}}{pd_j{}^{cl,t}+pu_j{}^{cl,t}} =$ <u>up-direction weight</u> for g_j

$\varphi_j^{D,cl,t} = \frac{pd_j{}^{cl,t}}{pd_j{}^{cl,t}+pu_j{}^{cl,t}} =$ <u>down-direction weight</u> for g_j

Note that $\varphi_i^U + \varphi_i^D = 1$. These weights are the inputs for the Method 3 Sampling Algorithm which, for each iteration, will produce a 'synthetic' level 6 CGS [for a given cell line at the specific threshold]. The algorithm below is slightly modified from its original version in chapter 2 to reflect the specific cell line- level specificity.

<u>Step 1: Simulate Nodes</u>

Input: Vectors with the parameters $\pi_j{}^{cl,t}$, $\varphi_j^{U,cl,t}$, $\varphi_j^{D,cl,t}$

Initialize a data frame with n rows for results

For $1:n$:

1. Label all M genes by sampling [with replacement] the labels "UP" or "DOWN" in a binomial fashion with the vector of probabilities taking the form $< \varphi_j^{U,cl,t}, \varphi_j^{D,cl,t} >$.

   $\delta_D \leftarrow$ units labelled "DOWN",

   $\delta_U \leftarrow$ units labelled "UP".

   Stop at step 1 if t = "complete".

2. Sample without replacement $t/2$ units from δ_D and $t/2$ units from δ_U using ${\pi_j}^{cl,t}$ as sampling weights.

Step 2: Simulate Edges from Simulated Nodes

Let $C^{cl,t}$ be the collection of simulated nodes for a given cell line at threshold t where $C_i^{cl,t}$ is an individual node $(i = 1, 2, \dots, 2000)$. Then, the following algorithm will be used to generate m edges:

Input: Nodes $C^{cl,t}$ and number of edges m to generate

Initialize a data frame *E* with *m* rows for results

For $\mathbf{1}$: $\boldsymbol{m}$:

1. Use random number generator to select two integers $(r_1, r_2 \in 1{:}\,2000;\ r_1 \neq r_2)$
2. Derive contingency table for $C_{r^1}^{cl,t}$ and $C_{r^2}^{cl,t}$ and store as edge $E_{r^1 r^2}^{cl,t}$ in row m
3. Calculate odds ratio and SE for edge $E_{r^1 r^2}^{cl,t}$ and store values in row m

3.2.3 Annotated KEGG L1000 Edges

The inherent assumption that underlies this analysis is that the topological annotations that connect two genes/protein products will yield insight into biological mechanisms of action that are either unique with regard to a particular cell line or ubiquitous across the board and perhaps interesting for their general applicability across biological specimens of interest. The topological annotations that we will consider for this analysis are the edge-type protein-protein

relationships extracted from the KEGG pathway database that [11]. There are 5,474 unique edges from KEGG that are amenable to our core L1000 data set; that is both proteins/gene-products are perturbed via shRNA across the seven cell lines.

Let $\boldsymbol{K}$ be a 5,474×2 matrix that contains the collection of all unique KEGG edges that have pairwise data entries in the L1000 data set such that k_{i1} and k_{i2} are the gene symbols for the two nodes of the i^{th} edge. Note that the directionality of the relationship is not a factor for this step in the analysis; an edge with the relationship B→A is equivalent to A→B and the node pair (B,A) will not be in $\boldsymbol{K}$ if the node pair (A,B) is already in $\boldsymbol{K}$. Let $\boldsymbol{CGS^{cl,t}}$ be the 2,007×2 matrix whereby $CGS\_D_i^{cl,t}$ contains $\boldsymbol{CGS_{i1}^{cl,t}}$ $CGS\_U_i^{cl,t}$ contains $\boldsymbol{CGS_{i2}^{cl,t}}$. The data set corresponding to all of the KEGG edges, $\boldsymbol{F^{cl,t}}$, is then constructed as follows:

Input: $\boldsymbol{CGS^{cl,t}}$ and $\boldsymbol{K}$

Initialize a data frame $\boldsymbol{F}$ **with** $\boldsymbol{m}$ **= 5,474 rows for results**

For j = $\boldsymbol{1:m}$**:**

1. Derive contingency table for $CGS_{i=k_{j1}}^{cl,t}$ and $CGS_{i=k_{j2}}^{cl,t}$ and store as edge $F_{k_{j1},k_{j2}}^{cl,t}$ in row j.
2. Calculate odds ratio and SE for edge $F_{k_{j1},k_{j2}}^{cl,t}$ and store values in row j.

3.2.4 Random L1000 Edges

An alternative null distribution will be synthesized from random 'real' nodes in the L1000 data set. Here, a 'real' node is a Level 6 CGS corresponding to an individual perturbagen (shRNA) at threshold level *t*. These edges are analogous to the KEGG edges described in the previous section with the exception that we will be constructing the edges from nodes stored in the matrix $\boldsymbol{R}$ instead of $\boldsymbol{K}$. The criteria for any given pair of nodes $< r_{j1}, r_{j2} >$ is that it cannot be among the pairs of nodes in $\boldsymbol{K}$ and that all pairs of nodes in $\boldsymbol{R}$ are unique amongst themselves.

The same set of random nodes $\boldsymbol{R}$ will then be used to fill the data frame $\boldsymbol{G^{cl,t}}$ in exactly the same manner that $\boldsymbol{K}$ was utilized to fill the data frame $\boldsymbol{F^{cl,t}}$.

3.2.5 Cell line vs Cell line comparisons

At this phase of the analysis, we will generate measurements of association amongst the levels of cell line-combinations and thresholds that involve the data frames $\boldsymbol{E^{cl,t}}$, $\boldsymbol{F^{cl,t}}$ and $\boldsymbol{G^{cl,t}}$. This process will be similar to the one we introduced in **Section 2.7.3**. The following procedure will be carried out across all cell line-combinations and threshold levels. The data frames $\boldsymbol{EE^{cl_a:cl_b,t}}$, $\boldsymbol{FF^{cl_a:cl_b,t}}$ and $\boldsymbol{GG^{cl_a:cl_b,t}}$ will contain the outcomes for the edge-edge comparisons for the simulated, KEGG and random edges respectively. Let the data frames $\boldsymbol{X^{cl,t}}$ and $\boldsymbol{XX^{cl_a:cl_b,t}}$ be generic versions of the edge and edge-edge data frames such that the procedure described below can be extended to the simulated, KEGG and random edges in an identical manner.

Input: Edges $\boldsymbol{X^{cl=a,t}}$ **and** $\boldsymbol{X^{cl=b,t}}$

Initialize a data frame $\boldsymbol{XX^{cl_a:cl_b,t}}$ **with** $\boldsymbol{m}$ **rows for results**

For j = 1: $\boldsymbol{m}$**:**

$XX_{j,\Delta}^{cl_a:cl_b,t}$ ← Delta value for edge *j* between cell lines *a* and *b*

$XX_{j,\sum overlaps}^{cl_a:cl_b,t}$ ← Total number of overlapping genes from the cells of the contingency tables for edge *j* between cell lines *a* and *b*

3.3 Power Analysis

The power analysis in this section will either support or fail to support the notion that the proposed test statistics [measurements of association] are most appropriate and useful when used in conjunction with cell line specific models of background gene behavior as well as custom thresholding parameters instead of the default threshold of the original Level 6 CGSs. There are

two null hypotheses to consider; the first is that the genes in the Level 6 CGSs exhibit random behavior and are therefore not able to detect meaningful or biologically relevant similarities or differences in gene-product-relationship between cell lines. This null is represented by the "simulated" edge set. The second null hypothesis is that the annotation given to edges is not predictive of a greater level of similarity or difference between cell lines; this null is represented by the "random" edge set. *In both scenarios, the alternate distributions are represented by the test statistics in the "KEGG" edge set*.

3.3.1 Calculation of Power: Delta

In the analysis of the simulated data, there was an important difference between the null scenario and the alternate scenario with regard to the delta values that does not hold true for the analysis of cell line vs. cell line behavior – this is the assumption that the alternate scenarios would show greater levels of concordance which could be captured by a delta distribution whose mean was shifted away from zero in the positive direction (i.e. shifted to the right along the x-axis). Recall that $\widehat{\Delta}_{XY}^{AB} = \frac{log(\widehat{\theta}_{XY}^{A}) - log(\widehat{\theta}_{XY}^{B})}{\sqrt{(SE_{XY}^{A})^2 + (SE_{XY}^{B})^2}}$ for edge X|Y compared in cell line A and B where cell line A is there "reference" and cell line B is the "comparator". In all cases, the relationship $\widehat{\Delta}_{XY}^{AB} = -1 * \widehat{\Delta}_{XY}^{BA}$ holds true. If $\widehat{\Delta}_{XY}^{AB} > 0$, the interpretation is that there is greater concordance between X and Y in cell line A than in cell line B whereas $\widehat{\Delta}_{XY}^{AB} < 0$ implies greater concordance in cell line B than in cell line A; thus both positive and negative values of delta suggest deviation from the null hypothesis of equality of concordance between cell lines. Therefore, under the assumption that there are in fact edges that are more concordant in the reference cell line but others that are more concordant in the comparator cell line, we would expect an alternate distribution that displays heavier tails as opposed to an absolute shift in delta.

The plots in **Figure 17** demonstrate two aspects of the data. The first is that there is indeed a higher occurrence of delta values in the extremes [tails] of the distribution under the alternate hypothesis when the simulated edge set acts as the null distribution. The second is that, by including more LM genes in concordance measurements (i.e. increasing the threshold *t*), differences in odds ratios between cell lines will be even more pronounced as evidenced by the larger spread of values in **Figure 17(A)** vs. **(B)**. It may not be readily obvious given the difference in scale between the two graphs that the shaded area (representing power at a two-tailed alpha = .10) in **Figure 17(B)** is also much greater than the shaded area in **Figure 17(A)**, 75% vs 33.5%. The power analysis in the following section will address the nature of the power gains relative to increases in list length across the spectrum of list lengths.

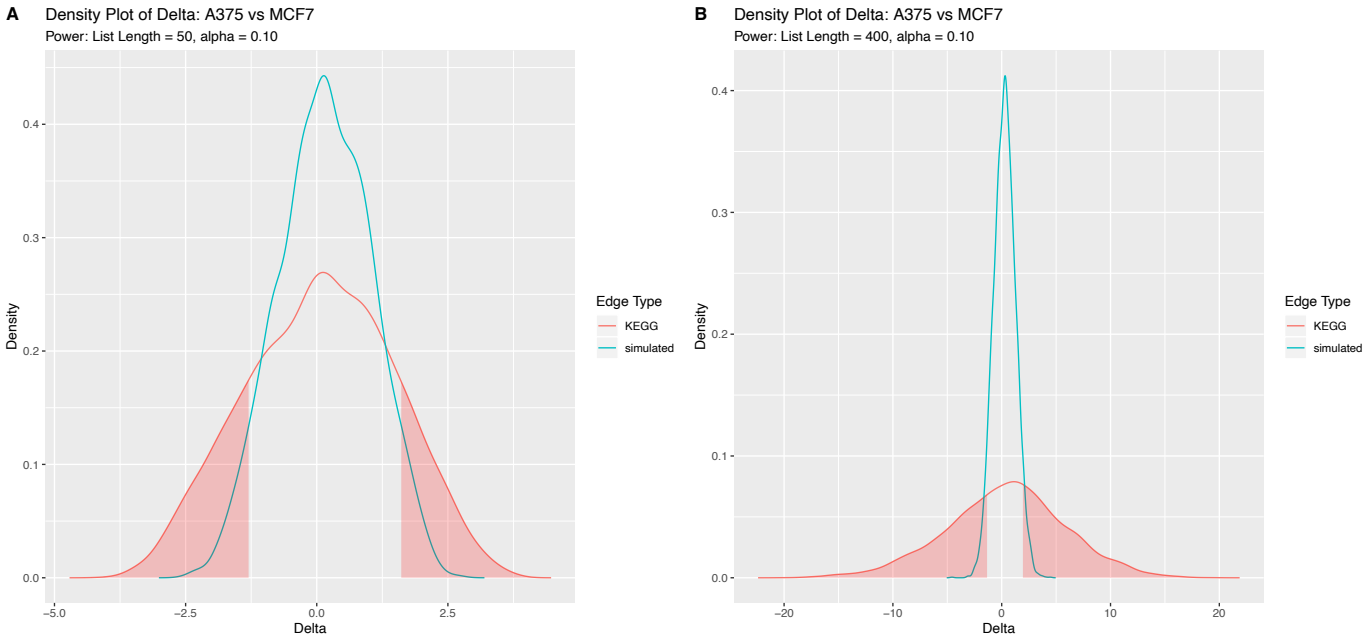


Figure 17: Density curves for Δ values with A375 as the reference cell line and MCF7 as the comparator cell line amongst simulated edges as well as KEGG annotated edges. The shaded areas on the right-hand side of the graph represent edges that are more concordant in A375 than MCF7 whereas areas on the left-hand side are more concordant in A375 than MCF7 using cutoff values for the below the 5th percentile and above the 95th percentile of simulated delta values at list lengths of 50 (A) and 400 (B).

3.3.2 Calculation of Power: $\sum overlaps$

The null hypothesis underlying the count-type data for the statistic $\sum overlaps_{XY}^{AB}$ is that the count of overlapping genes amongst the cells in the KEGG edges is not fewer or greater than we would expect under simulated or random conditions. As a consequence, the power for the $\sum overlaps$ will be handled in a two-tailed fashion akin to the treatment of the power for calculation for delta. Unlike continuous measurements of delta, the $\sum overlaps$ is discrete, and features counts that are dominated by zero at smaller list lengths. This type of behavior was characterized in the simulation study in the previous chapter. Another characteristic of the $\sum overlaps$ statistic is that there is no difference in $\sum overlaps_{XY}^{AB}$ and $\sum overlaps_{XY}^{BA}$ (recall that $\widehat{\Delta}_{XY}^{AB} = -1 * \widehat{\Delta}_{XY}^{BA}$).

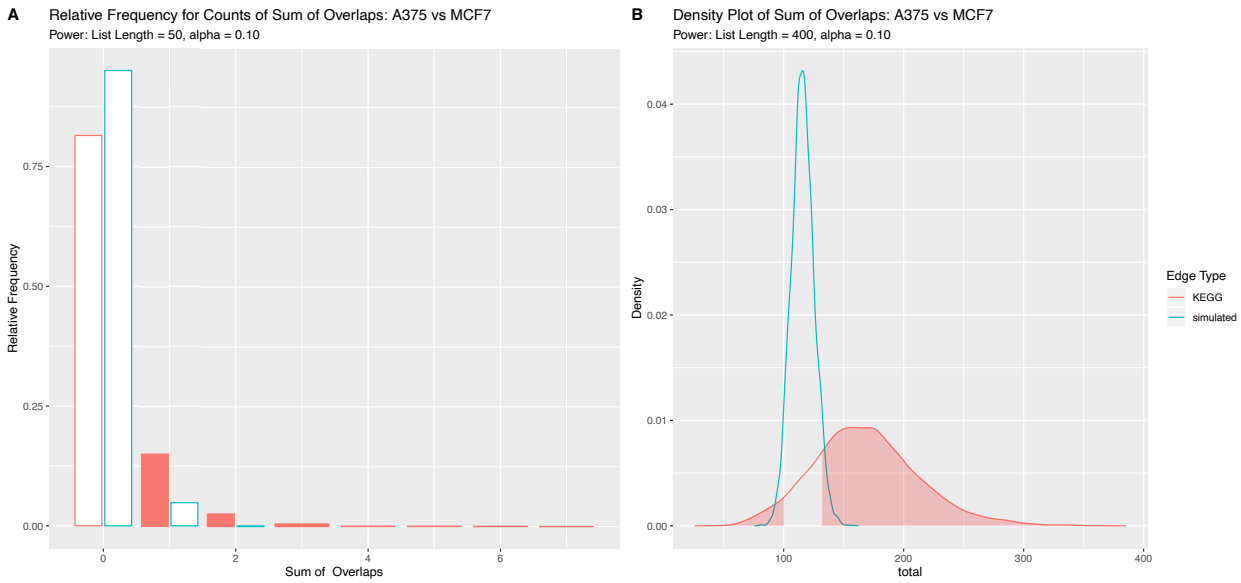


Figure 18: Relative frequency (A) and density plot (B) for the $\sum overlaps$ test statistic between the cell lines A375 and MCF7. In (A) the shaded bars correspond to counts for KEGG edges that contribute to the power calculation at a list length of 50 whereas in (B) the shaded area is a continuous approximation to the power at a list length of 400.

Figure 18 shows how drastically the distributions differ at opposite ends of the threshold spectrum. The distributions for both the null (simulated) and alternate (KEGG) delta

measurements is markedly discrete at the lower end of the threshold level (list length of 50) whereas both are well approximated by nearly normal distributions at the higher end (list length of 400). Regions of the alternate distribution that contribute to power at alpha = 0.10 is represented by the shaded regions and is equal to 18.52% in **Figure 18(A)** and 84.75% in **Figure 18(B).** As with the delta statistic, we have preliminary evidence that increasing the threshold leads to gains in power for the $\sum overlaps$ test statistic as well.

3.3.3 Power Analysis Results of Delta: Simulated Null Distribution

The power analysis of the delta test statistic shows a steady increase in power as the as the threshold of 10 (10 out of 978 LM genes are up and 10 are down) is increased to a less stringent threshold of 200 (nearly half of all LM genes). The power then plateaus as the list length approaches 250 and then begins to decrease above the threshold of 350. This pattern is evident across all levels of alpha but becomes clearer as the stringency of alpha increases (i.e. we decrease alpha).

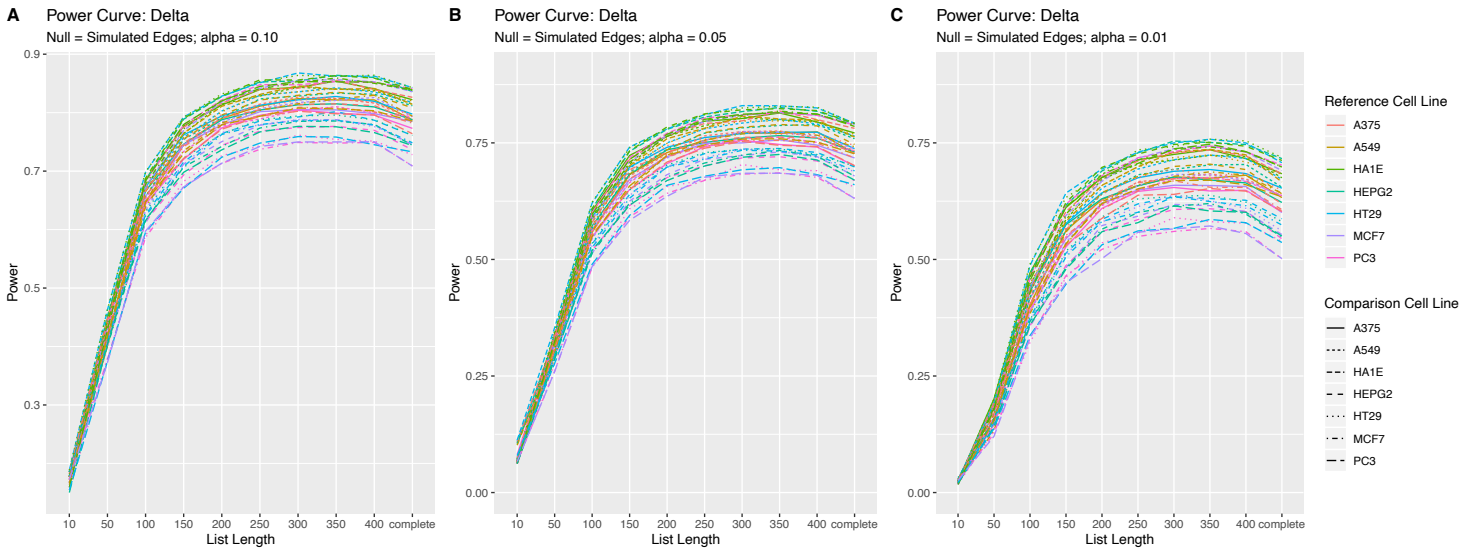


Figure 19: Power curves for Δ across all possible cell line combinations at alpha levels of 0.10 (A), 0.05 (B) and 0.01(C) with the simulated edge distributions serving as the null distribution.

Importantly, the shapes from the graphs in **Figure 19** show no discernable differences with regard to specific cell lines; that is, the choice for the most efficient threshold is independent of reference or comparison cell line. However, there are some patterns that emerge when the graphs are broken down by reference cell line (Appendix **Figure A5**). For example, comparisons between any cell line and HA1E are consistently more powerful than other two cell lines whereas the opposite is true for PC3 - in the graphs the green line (HA1E) is consistently the upper bound whereas the pink line (PC3) is consistently the lower bound for the power curves.

3.3.4 Power Analysis Results of $\sum overlaps$: Simulated Null Distribution

The power trend for the statistic $\sum overlaps$ is very similar, though not identical to the trend for delta; power increases at a steady rate to the threshold of 200, at which gains in power diminish in subsequent intervals. Unlike delta, the power for the sum of overlap statistic increases monotonically until the maximum threshold level – there is never a point at which the signal among the noise is noticeably muddied by casting a larger net via an increase in sample size.

Another commonality between the two statistics is that there are no glaring differences amongst the patterns of power increase as it relates to cell line specificity. Interestingly, the curves for the power of the $\sum overlaps$ that compare HA1E to other cell lines are clustered similarly to those of the curves for the power of the delta statistic with the exception of cell lines A375 (which trends higher on-) and HT29 (which trends lower on- the x-axis for most comparisons) [Appendix **Figure A6**].

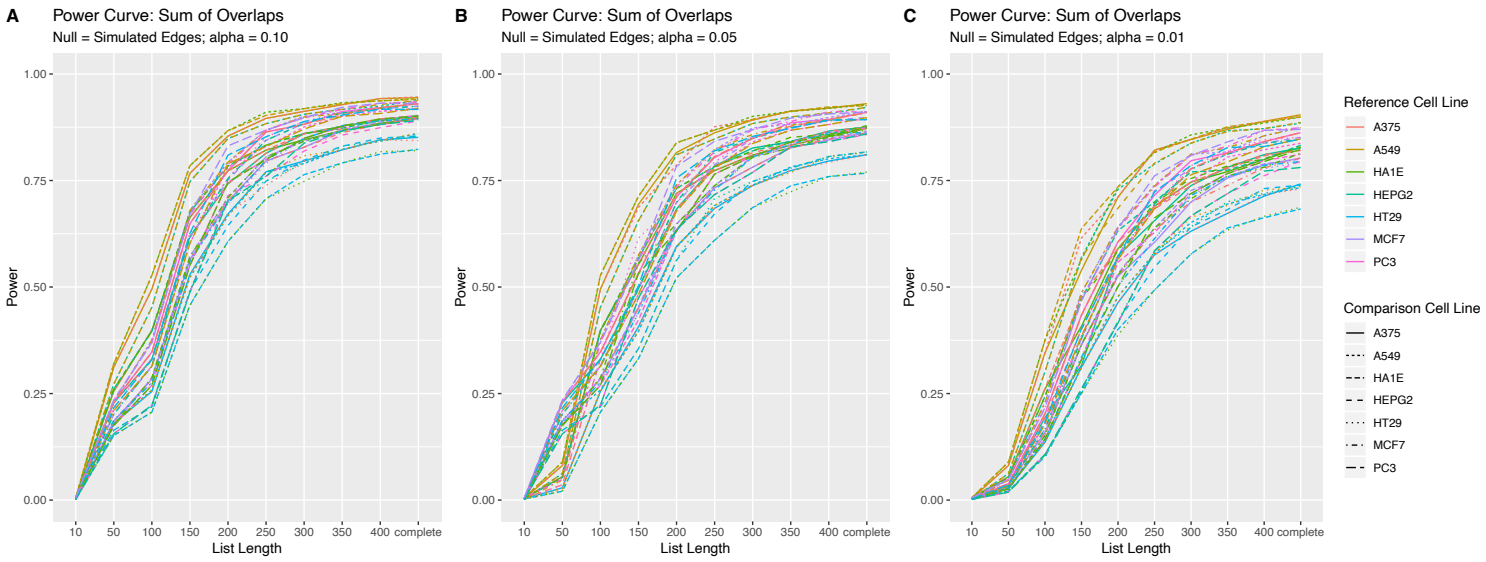


Figure 20: Power curves for $\sum overlaps$ across all possible cell line combinations at alpha levels of 0.10 (A), 0.05 (B) and 0.01(C) with the simulated edge distributions serving as the null distribution.

3.3.5 Power Analysis Results of Test Statistics: Random Null Distribution

The same power analyses of the two tests statics were performed using a null distribution as described in **Section 3.2.4.** This will allow us to get a sense of the nature of the relationship between annotated edges and their associated test statistics by answering the following questions: do we see larger differences in concordance (more extreme values of delta) when the edges being compared are documented in KEGG? And/or are the similarities more pronounced among annotated edges [as would be evidenced by different behavior of the distributions for sum of overlap statistics]?

Figure 21 and **Figure 22** depict the relationship between power and threshold for the two test statistics when the null distribution is set to random edges. The nature of the plots in **Figure 21** suggest that after the threshold of 50, there is no advantage to broaden the inclusion of genes that might contribute to the delta statistic. On the other hand, the curves for the overlap statistic in **Figure 22** have the same shape, albeit a smaller scale on the y-axis, as those in **Figure 20**.

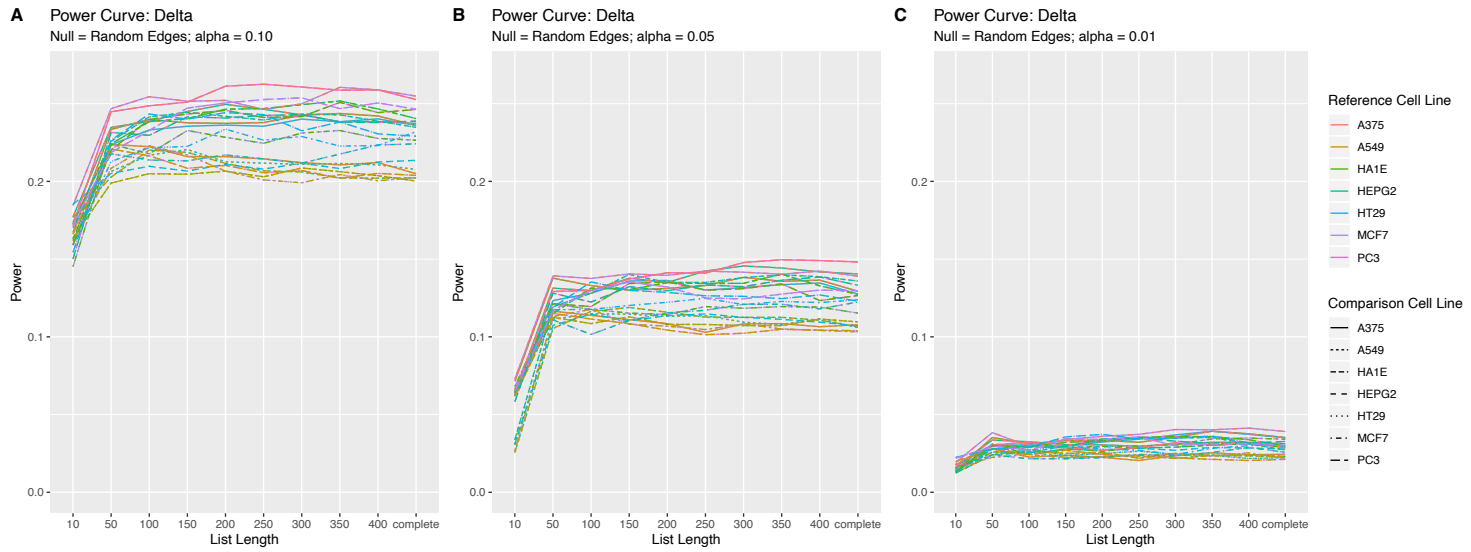


Figure 21: Power curves for delta across all possible cell line combinations at alpha levels of 0.10 (A), 0.05 (B) and 0.01(C) with the random edge distributions serving as the null distribution.

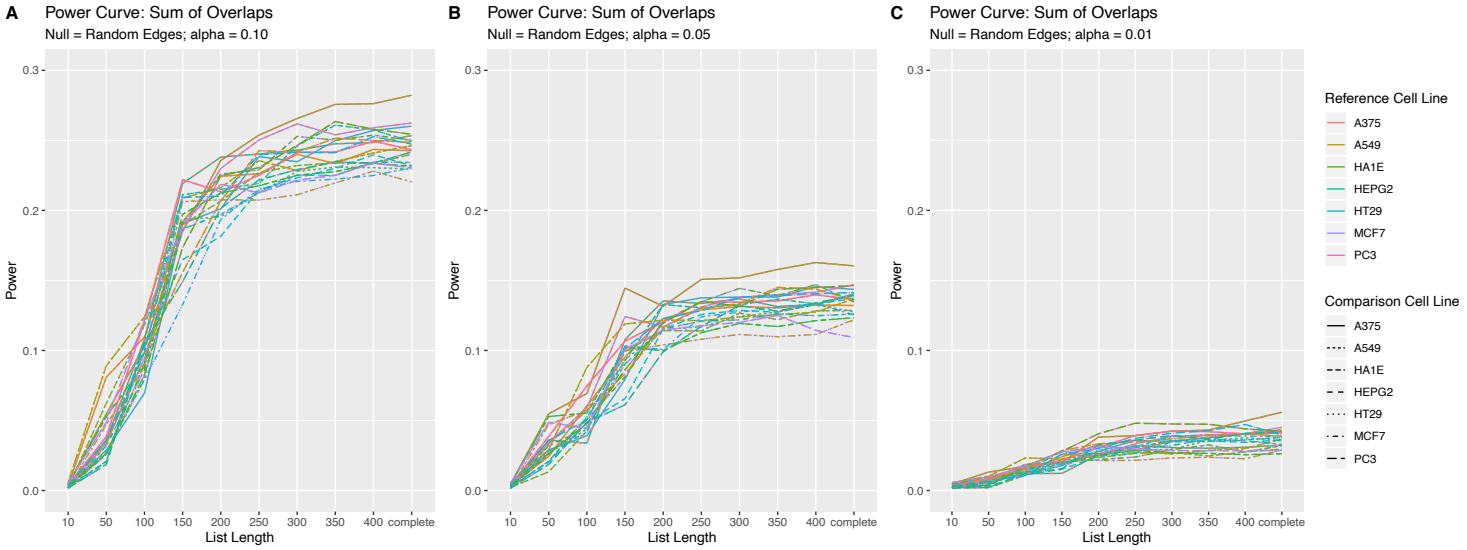


Figure 22: Power curves for $\sum overlaps$ for all possible cell line combinations at alpha levels of 0.10 (A), 0.05 (B) and 0.01(C) with the random edge distributions serving as the null distribution.

3.3.5 Power Analysis: Conclusions

The results of the power analysis suggest that the threshold of 50 may be too low to detect pairs of genes that are either similar (as measured by the $\sum overlaps$) or dissimilar (as measured by delta) between two cell lines. Although annotated edges have modestly larger sums of overlaps on average than random edges as well as delta values (**Figure 21**, **Figure 22**) there is not sufficient power to detect differences and similarities when using the random edges as the null distribution; therefore the distributions that arose from the simulated edges will serve as the null. The threshold of 200 (200 LM genes most up-regulated and 200 LM genes most down-regulated) will be used to construct test statistics.

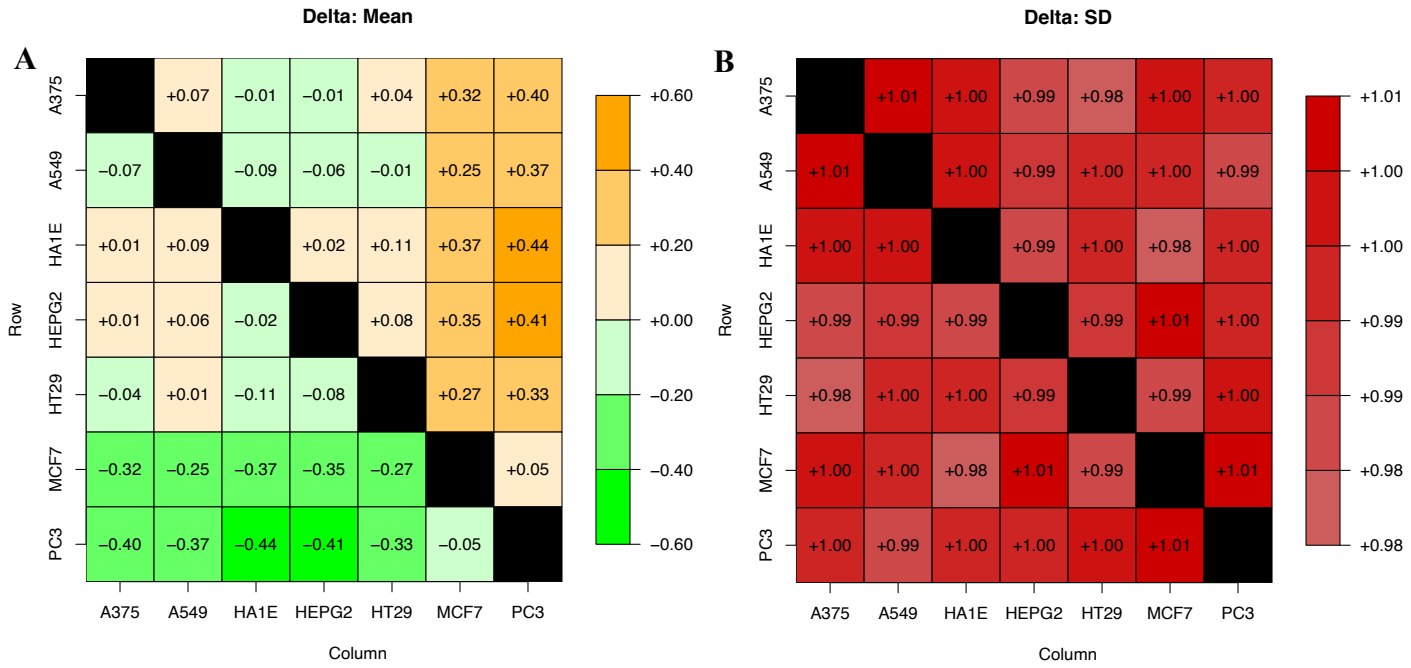


Figure 23: Matrix of mean values for delta (A) and its standard deviation (B) for cell line vs. cell line comparisons at the threshold of 200 genes up and 200 genes down.

Figure 23 provides both the quantitative summary of the distributional properties of delta along with a visual guide for interpreting differences/similarity among the values. The matrices

can be read as follows: When cell line from row X is compared to the cell line in column Y, the value is found in the entry for [X,Y] in the matrix. If the color of the cell is orange, it indicates that deltas are, on average, more positive for the cell line in the row vs. the cell line in the column, whereas if it is green the association is opposite (on average, deltas are more positive for the cell line in the column vs the cell line in the row). While minor differences in delta can be found when comparing the first few cell lines (A375, A549, HA1E, HEPG2, and HT29) to one another, the magnitude of the differences between these cell lines and MCF7 and PC3 are considerably larger. Therefore, the assumption that delta has a mean of zero is not met, most notably in comparisons involving MCF7 or PC3. The standard deviations, on the other hand, are very close [if not equal within rounding] to one – validating the assumption that delta has a standard deviation of one.

In the following chapter, edges from pathways will be summarized by their average delta in a matrix format just like those in **Figure 23(A)**; if we denote that matrix by $\boldsymbol{P}$ and the matrix in **Figure 23(A)** as $\boldsymbol{C}$, the adjusted matrix $\boldsymbol{A = P - C.}$ The adjusted matrix will then be used to describe the magnitude of the delta values. Unlike delta, which has a continuous-valued distribution, the $\sum overlaps$ statistic takes on discrete values. Therefore, we will use values that correspond to given percentiles of the simulated distributions to describe pathway similarities in the following chapter.

In figure **Figure 24(A),** the results of the simulation indicate that, with the exception of the relationship between HA1E and A549, there are no differences in median values. The matrix in **Figure 24(B)** represents the critical values for significance; at the 99th quartile 5 of the possible 21 pairwise comparisons have larger values (16) compared to the other 16 comparisons. Given the differences exhibited in **Figure 24(B)** and **Figure 23(A)** across the pairwise

comparisons, there is evidence that it may be worthwhile to take pairwise differences into consideration when describing pathway-level activity.

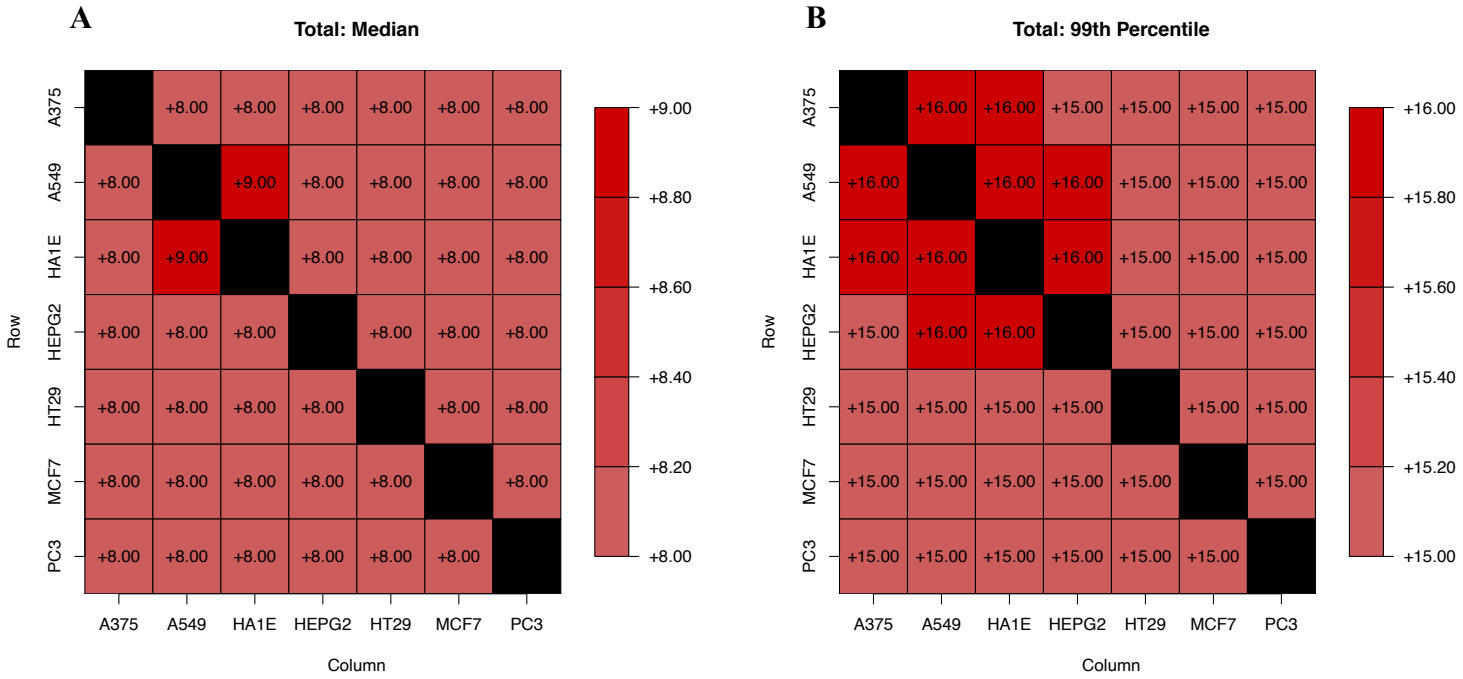


Figure 24: Matrix of median values for the $\sum overlaps$ (A) and values that represent the 99th percentile (B) for cell line vs. cell line comparisons at the threshold of 200 genes up and 200 genes down.

Chapter 4: Application in KEGG Pathway Analysis

The original purpose of the Connectivity Map or "CMap" project was to take a bottom-up approach to finding connections among genes in the human genome by virtue of similar gene expression profiles or "signatures". The goal of this project is not necessarily to find new connections, but rather to assess the nature of known connections as they relate to cell line similarities and differences. The term "known connections" refers to relationships between genes that have been derived via experimentation and documented in literature. Specifically, we are considering relationships between genes from pathways in the KEGG database.

The data generated via LINCS relies on readouts generated from cell lines. Cell lines that are developed for use in research are heterogeneous with regard to tissue/tumor of origin and specific mutations. Despite their heterogeneity, cancer and non-cancer cell lines (for example HA1E, immortalized kidney cells) have acquired mutations that allow cells to exhibit abnormal growth via evasion of apoptosis or deviant patterns of division and growth; by definition, all cell lines have achieved immortality [29] [30]. In a study comparing the gene expression profiles of cell lines, tumor-derived cells and normal tissues, cell lines clustered with one another as opposed to samples from the same tissue [31]. The phenotypic similarity amongst cell line gene expression despite their genotypic differences raises the argument that an analysis of differences among base-line gene expression would be underpowered for finding functional, pathway-level differences. This chapter will describe how the measurements of heterogeneity and similarity of the concordance signatures (CSs) introduced in Chapter 1 will be applied to a pathway-level description and then analysis (Chapter 6) of the L1000 data set using the KEGG pathway database.

4.1 The KEGG Database

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is an on-line resource that was initiated in 1995 with the goal of "computerizing the current knowledge of genetics, biochemistry, and molecular and cellular biology in terms of the pathway of interacting molecules or genes" [32]. While KEGG maintains a compendium of gene catalogs as well as catalogs for chemical elements and compounds, it is most well-known for its library of molecular biological networks or "pathways". KEGG is, if not the most, one of the most utilized biological pathway databases; it has been cited in over 15,000 publications. For the sake of comparison, the next most-often cited resource Reactome has been cited less than 2,000 times in publications since 2011 [33].

The popularity of KEGG can be attributed to its manually curated pathway maps that offer a visually interpretable representation of complex biological interactions. The linear diagrams, directed edges, and relationships based on consensus research efforts are all characteristics of "pathways" as opposed to "networks", which are typically much larger in scale, usually have undirected edges, and are compiled as a result of data-set-specific large-scale screens [34]. Each pathway map provides information on the consensus knowledge-base derived relationships between molecules interacting as part of a defined biological process. In graph terminology, the nodes in the pathways are genes/gene products (proteins) and the edges are the relationships between the genes and are classified as activating, inhibiting, or binding relationships. In addition to being either activating or inhibiting, some edges are also defined with regard to post-translational modifications such as phosphorylation (labeled as +p) and dephosphorylation (labeled as -p). As mentioned before, the edges are directed - there is a clear

to-and-from direction of action (except for in the case of binding relationships). In cases where the relationship between genes is indirect, the edge is represented with a dashed line.

There are currently over 500 pathways in KEGG's repository that span seven different categories of specification (see **Figure 25**). The KEGG database is dynamic; pathways are updated and new pathways are added as the knowledge base grows and becomes more refined. This aspect is important to keep in mind when performing analyses using resources that have static 'snapshots' of the KEGG data base that were downloaded at a given point in time and may be out of date. For example, the R package MSigDB (Molecular Signature Data Base), which contains collections of gene lists from 186 KEGG pathways, may not contain the most recent KEGG pathways or may otherwise be missing one of the roughly 300 other KEGG pathways that are excluded from MSigDB [35]. This is not necessarily a drawback of KEGG or databases that get information from KEGG; it just means that researchers should be aware of this aspect in their analyses and may need to manually add pathway data to serve their research needs.

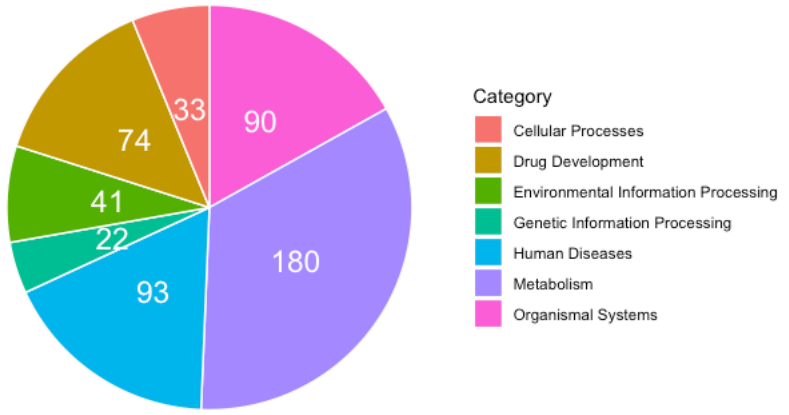


Figure 25: Pie chart representing the number of each type of pathway defined in KEGG.

4.2 Pathway-level Analysis

4.2.1 Overview of Methods for Pathway Analysis

The term 'pathway analysis' in the context of gene expression analysis has broad implications but most types of pathway analytical methods fall into one of three categories as defined by Khatri [36]:

First Generation: Over-Representation Analysis (ORA)

Also referred to as 2 x 2 table methods, these types of methods employ tests of statistical significance to identify pathways that are over-represented based on experimentally-derived gene expression profiles that surpass a given thresholds (for example, lists of differentially expressed genes between cases and controls that have p-values [adjusted for multiple testing] of less than 0.05 after conducting differential expression analysis). One examples of this type of analysis is LRpath, which uses a logistic regression-based method to identify pathways that contain significantly more genes from the input list of genes than expected [37].

Second Generation: Functional Class Scoring (FCS)

As opposed to ORA methods, FCS methods take as input gene-level statistics that that measure differential expression (i.e. t-test statistic, z-score or signal-to-noise ratio of expression of a gene between two groups) and output pathway-level measurements – that is, each pathway is "scored" based on coordinated changes in the expression of genes in that pathway. A well-known example of FCS is Gene Set Enrichment Analysis (GSEA), which uses a Kolmogorov-Smirnoff statistic to identify pathways that contain genes whose expression is similar in both direction (up-regulated or down-regulated) and magnitude of signal-to-noise ratio relative to all of the genes measured in an experiment [10] [38].

Third Generation: Pathway Topology (PT)-Based Approaches

As their name suggests, PT-Based approaches incorporate the topology pathways, i.e. features such as direction and type of relationships between genes, into their methodology. In addition to generating pathway-level measurements, PT-Based methods also produce, perhaps as an intermediary step, gene-level measurements. Gene-level measurements can either be per-gene or per-gene-pair. For example, the ScorePAGE (Scoring Pathway Activity from Gene Expression) algorithm first assigns a similarity metric for each pair of genes in a pathway and then divides that score by the graphical distance (i.e. number of nodes) between the two genes [39]. An example of the per-gene approach is SPIA (Signaling Pathway Impact Analysis) which assigns a perturbation factor (PF) to each gene in a pathway as a function of the expression of upstream genes [40]. SPIA falls into the category of impact factor (IF) approaches, a term coined by Dragici et al to describe analysis methods that take into consideration the differential expression of genes within a pathway as well as their upstream and downstream relationships before assigning a score to the pathway that is proportional to the number of differentially expressed genes in the pathway [41]. Both SPIA and ScorePAGE use the gene-level statistics to derive overall pathway-level scores.

All of the types of pathway methods described are what Khatri defines as "knowledge base-driven" methods [36]. They are also similar with regard to the starting point (gene expression data) and end goal (a list of pathways ranked by statistical significance). Other methods [not covered by Khatri] that use gene expression data in the context of pathway analysis include Ingenuity Pathway Analysis, which provides causal analytical tools to assess the upstream biological factors [genes] responsible for the observed [gene-expression-based] activity in pathways [42]. Pathway analysis can also come in the form of subgraph extraction wherein

the goal is to identify subsets of the most significant subset of connected pathway nodes and edges for a given condition [43].

4.2.2 Approach to Pathway Analysis of L1000 Data

The different generations of pathway analysis are hierarchical in nature with an additional layer of information added as the analysis moves from one level to the next. The analysis of the L1000 data set differs in one unique aspect – although we are technically working with expression-based data, it is not traditional transcription data whereby, [after pre-processing] each gene in each sample is assigned a single measurement that reflects the gene's expression in terms of both direction and magnitude. These measurements are then mapped onto their corresponding genes in a pathway to represent that gene's own expression before conducting the types of pathway analyses covered in the last section.

The per-gene consensus genomic signatures (CGSs) or gene-pair concordance signatures (CSs) that are mapped onto pathway genes are fundamentally different as the information assigned to each gene is not a reflection of its own expression. Instead, the information associated with each primary gene reflects the binary direction of expression for the L1000 genes after the primary gene is functionally knocked down in the model system. Although we are not working with traditional expression data it is useful nonetheless to frame the approach with regard to the different generations of pathway analysis. The approach that we are taking is somewhat of a hybrid of the three different approaches.

The preliminary step of our analysis, the construction of contingency tables, is akin to ORA in the sense that the genes that fall into the table are organized as a function of their binary direction but not magnitude of differential expression. Similar to the FCS approach, we are interested in both gene-level and pathway-level statistics; that is to say we are interested in

finding differentially regulated pathways among cell lines as well as the edges within pathways that suggest heterogeneity of gene-gene relationships based on quantitative measurements. Our approach can also be framed in the context of PT- based approaches since we are deriving measurements based on pairwise relationships between genes (CSs) based on the topology of a given pathway.

The sections leading up to this analysis have highlighted two important aspects of the analysis of this large, multidimensional dataset:

1) The threshold of 200 (i.e. lists of 200 upregulated and 200 downregulated L1000 landmark genes) is the optimal threshold for detecting pairwise differences and similarities between cell lines.
2) At the threshold of 200, we expect slightly more concordance by chance in the cell lines A375, A549, HA1E, HEPG2, and HT29 when compared to MCF7 and PC3 as well as slightly larger median values for sums of overlaps in 5 of the 21 pairwise comparisons (16 vs 15).

Therefore, the threshold of 200 will be used in all following descriptive analyses and the appropriate adjustments will be applied for pairwise comparisons in the formal pathway analysis conducted in **Chapter 6**.

Our approach to a pathway analysis of the L1000 is hierarchical in nature; edges will be assigned values that represent the over-all differences, similarities and level of concordance in a multi-way comparison as outlined in the following section. Pathways will be scored as a function of the of the edges within the pathway; this will give us an idea of the overall heterogeneity and/or homogeneity at the pathway level. While we use the term 'analysis', the procedures outlined in this chapter are more of a first pass means to explore the data and

demonstrate how it can be incorporated into the topology KEGG pathways. In **Chapter 6** we will conduct a formal edge set enrichment analysis (ESEA) and make conclusions regarding cell line specific pathway behavior.

4.2.3 The Breslow-Day Test for Detection of Differentially Regulated Edges and Pathways

Before diving into the specific pairwise differences among cell lines, we will calculate statistics that will allow for a global assessment of differences in concordance among all seven cell lines. The method employed to derive these values operates as a function of the 2-way tables for directional concordance (data arranged in an odds ratio table). The method of interest is the Breslow-Day test for homogeneity of odds ratios. The Breslow-Day test is preferable to the other well-known method to analyze 2×2 tables, the Cochran-Mantel-Haenszel test, which assumes that all associations are in the same direction [44]. The procedure to derive the Breslow-Day test statistic (BD) for edge M|N across j = 7 cell lines is as follows [45] [46].

Step 1: derive common odds ratio

Estimate the over-all common odds ratio ψ using the entries from **Table 16** as follows:

Table 16 : 2-way Table for Cell Line j

Cell line j		Perturbagen M		Marginal Column Totals
		Up in M	Down in M	
Perturbagen N	Up in N	a_j	b_j	n_{Uj}
	Down in N	c_j	d_j	n_{Dj}
Marginal Row Totals		m_{Uj}	m_{Dj}	T_j

$$\hat{\psi} = \frac{\sum_{j=1}^{7} {a_j d_j}/{T_j}}{\sum_{i=1}^{7} {b_j c_j}/{T_j}} \tag{10}$$

Step 2: Calculate Expected Frequencies for each cell line

Table *17* gives the expected frequencies of the contingency table for cell line j as a function of the marginal and expected frequencies for the first cell A_j:

Table 17: 2-way Table of Expected Frequencies for Cell line j

Cell line j		Perturbagen M		Marginal Column Totals
		Up in M	Down in M	
Perturbagen N	Up in N	A_j	$n_{Uj} - A_j$	n_{Uj}
	Down in N	$m_{Uj} - A_j$	$n_{Dj} - m_{Uj} + A_j$	n_{Dj}
Marginal Row Totals		m_{Uj}	m_{Dj}	T_j

A_j is derived as the positive solution to the quadratic equation:

$$A_j(\hat{\psi}) = \frac{\hat{\psi}(n_{Uj} + m_{Uj}) + (n_{Dj} - m_{Uj}) \pm \sqrt{[\hat{\psi}(n_{Uj} + m_{Uj}) + (n_{Dj} - m_{Uj})]^2 - [4(\hat{\psi} - 1)\hat{\psi}(n_{Uj}m_{Uj})]}}{2(\hat{\psi} - 1)} \quad (11)$$

The variance is given as:

$$Var(a_j; \hat{\psi}) = \frac{1}{\frac{1}{A_j} + \frac{1}{n_{Uj} - A_j} + \frac{1}{m_{Uj} - A_j} + \frac{1}{n_{Dj} - m_{Uj} + A_j}} \quad (12)$$

Step 3: Conduct Breslow-Day test for homogeneity of odd's ratio

H_0: $\psi_j = \psi$ for all cell lines (j = 1:7).

H_1: $\psi_j \neq \psi$ for at least one cell line.

$$BD = \sum_{j=1}^{7} \frac{\left(a_j - A_j(\hat{\psi})\right)^2}{Var(a_j; \hat{\psi})} \quad (13)$$

Under the null, BD has a chi-square distribution with 6 (K-1 where K is the total number of cell lines) degrees of freedom.

The Breslow-Day test statistic will be calculated for each edge in KEGG pathways that has corresponding shRNA perturbations in the L1000 data set for the core set of cell lines. Then, a pathway concordance heterogeneity score (PCH) for pathway *p* with *n* edges will be assigned to each pathway as follows:

$$PCH_p = \frac{\sum_{i=1}^{n} BD_i}{n} \tag{14}$$

The PCHs will be used a ranking metric for pathways such that those with the largest BD values contain, on average, the most heterogeneous relationships between genes across all cell lines. The most heterogeneous pathways will be examined for significant pairwise cell line comparisons by using a pairwise pathway concordance (PPC) score between cell lines A and B:

$$\overline{PPC}_{p,AB} = \frac{\sum_{i=1}^{n} \Delta_{i,AB}}{n} \tag{15}$$

Note that $PPC_{p,AB} = (-1) * PPC_{p,BA}$. If this score is positive, then, on average, the edges for pathway *p* are more concordant in cell line A vs cell lines B whereas if it is negative it implies more concordance in cell line B. Under the null, where there are not significant differences in concordance, we would expect $PPC_{p,AB} \sim N(0,1)$. Recall in the simulation study, we found that we would expect certain cell lines, namely MCF7 and PC3, to have edges that are more discordant compared to other cell lines (**Figure 23 B**). In order to take these differences into account, each $PPC_{p,AB}$ will be adjusted to reflect the distributional properties of the Δ_{AB} statistic such that

$$\widehat{PPC}_{p,AB} = \frac{\overline{PPC}_{p,AB} - \hat{\Delta}_{AB}}{SE(\hat{\Delta}_{AB})} \sim N(0,1). \tag{16}$$

Since the Metabolism pathways are designed to represent metabolic as opposed to transcription-based processes, these pathways will not be included in the following analysis. Any pathway with less than 10 edges will not be considered in the analysis (*total number of pathways considered* $N = 160$).

4.2.4 $\sum\sum overlaps$ for Detection of Similarly and Heterogeneously Regulated Edges and Pathways

Whereas the Breslow-Day statistic is a multivariate approach to find overall differences based on the odds ratios across cell lines, a chi-square-type of statistic (which we will denote as ξ) will be used as an ad-hoc measure to detect non-homogeneity amongst the $\sum overlaps$ statistic. Let c represent the 21 unique cell line pairings; this will allow us to avoid correcting for duplicate entries since $\sum overlaps$ for edge e between cell lines i and j is equal to $\sum overlaps$ for edge e between cell lines j and i. ξ^e will be calculated for each edge as follows: Let $\lambda_c^e = \sum overlaps$ for edge e between cell line pair c. Then,

$$\overline{\lambda^e} = \frac{\sum_{c=1}^{21} \lambda_c^e}{21}. \tag{17}$$

Thus, $\overline{\lambda^e}$ is the average $\sum overlaps$ for edge e across all cell line pairs. Then,

$$\xi^e = \sum_{c=1}^{21} \frac{(\overline{\lambda^e} - \lambda_c^e)^2}{\overline{\lambda^e}}. \tag{18}$$

When edges have small ξ^e values, then there are not big differences in the $\sum overlaps$ statistic across the 21 unique cell line pairings, i.e. $\lambda_c^e \approx \overline{\lambda^e}\ \forall\ c$. Large ξ^e, on the other hand, suggest $\lambda_c^e \neq \overline{\lambda^e}$ for at least one cell line-cell line pairing (see **Figure 29**).

The $\sum overlaps$ statistic will also be used to derive a statistic that describes the similarity across cell lines at a particular edge. The statistic $\sum\sum overlaps$ is simply the numerator of $\overline{\lambda^e}$ in equation 17:

$$\sum\sum overlaps^{e} = \sum_{c=1}^{21} \lambda_c^e. \quad (19)$$

Each pathway will receive a pathway overlap heterogeneity score (POH) in order to rank the pathways as was done similarly with the PCH score for the BD statistic. For pathway *p* with *n* edges the POH will be assigned to each pathway as follows:

$$POH_p = \frac{\sum_{e=1}^{n} \xi^e}{n}. \quad (20)$$

Pathways with large POH_p will then be examined for pairwise differences. Let ω_{pc} be the set of $\sum overlaps$ for cell line pair c at pathway p; note that, for a pathway with *n* edges, $|\omega_{pc}| = n$. Then, the pairwise pathway overlaps score (PPO) will be calculated as follows:

$$PPO_{pc} = median(\omega_{pc}). \quad (21)$$

If PPO_{pc} is larger than the 99th percentile for $\sum overlaps$ for cell lines pair c from the simulation study (see **Figure 24B**), then we will consider that pair to be significantly similar with regard to their $\sum overlaps$ at pathway p.

Each pathway will also receive a pathway overlap magnitude score (POM). Let ϱ_p be the set of $\sum\sum overlaps$ for pathway *p.* POM will be assigned to each pathway as follows:

$$POM_p = median(\varrho_p) \quad (22)$$

4.3 Edge-level Results

4.3.1 Distribution of Breslow-Day (BD), ξ and $\sum\sum overlaps$ Across Edges

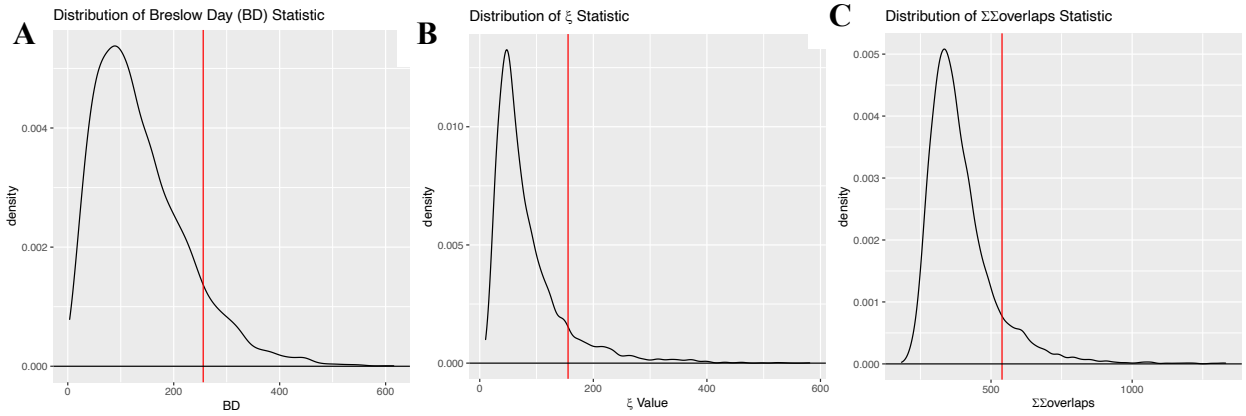


Figure 26: Density plots for the distribution of BD (**A**), $\boldsymbol{\xi}$ (**B**) and $\sum\sum overlaps$ (**C**) across all unique KEGG edges. The red line indicates 90th percentile value.

The distributions for the Breslow-Day, ξ, and $\sum\sum overlaps$ statistics are shown in **Figure 26**. All of the distributions are skewed to the right. Most (90%) BD values are less than 255.66, most (90%) $\boldsymbol{\xi}$ values are less than 155.54 and most (90%) of $\sum\sum overlaps$ are less than 539.

4.3.2 Breslow-Day Statistic Across Edges

Table 18: 20 largest BD edges

Edge	BD	A375	A549	HA1E	HEPG2	HT29	MCF7	PC3	ξ	∑overlaps
KRAS_BRAF	615.3107	9.431	8.325	7.449	5.981	8.876	9.431	-7.663	353.3155	1122
MET_ERBB3	594.6163	-6.706	-6.246	9.094	8.630	8.578	-9.048	4.716	198.6335	382
PROCR_THBD	562.6316	-8.257	5.800	7.530	9.236	-3.019	9.431	-1.620	524.8257	608
PDGFRA_KRAS	549.6416	5.935	-9.114	-7.967	-8.639	8.162	-7.383	-7.893	397.9909	661
PRKCE_BRAF	549.6001	-8.342	-8.880	9.292	-7.408	-8.575	-4.022	6.091	303.1667	468
KRAS_RPS6KA1	542.9527	-9.678	-9.159	-8.278	-3.689	7.396	-2.558	7.985	401.9721	645
MET_KRAS	529.7446	-8.844	-6.589	-8.520	7.268	9.053	-8.505	-1.625	217.2363	584
MET_PIK3CB	528.1846	8.728	8.681	-8.304	8.211	8.004	-2.811	7.950	190.9534	365
AKT3_MAP3K5	526.6204	4.929	-8.251	-8.922	6.630	8.698	-8.562	-7.268	276.0048	419
KRAS_PIK3CA	510.4201	-8.514	-8.271	-9.157	2.074	7.172	8.013	-6.357	352.8014	584
PDGFRA_JAK1	506.8152	3.588	-9.289	-7.148	-8.858	9.197	-1.461	-6.579	464.8362	409
PIK3CD_AKT3	505.9793	3.940	-6.649	-8.695	1.728	9.046	6.923	9.038	367.4886	352
MYC_CDK6	501.2416	-8.601	2.892	-7.191	7.170	5.831	6.495	7.049	236.2667	495
CDK6_RB1	501.0832	-6.935	5.450	7.910	-5.756	6.621	-8.584	3.484	218.2443	393
CDC25A_CDK6	485.1332	-8.219	-1.067	-8.765	8.990	7.626	4.110	5.166	276.6294	429
THBS1_ITGA2	483.4053	-7.985	7.671	9.399	9.795	5.802	-4.544	8.106	325.3909	614
MAP3K5_MAP2K4	482.9228	7.144	2.210	5.073	-5.238	7.914	-4.575	-3.893	191.1259	286
EP300_TCF7L1	480.9683	8.277	9.284	-6.842	-1.190	0.877	-3.144	8.469	209.7085	446
COL4A2_ITGA2	471.8633	9.121	8.609	6.600	8.480	-8.786	4.839	8.784	183.1208	621
ITGA2_PIK3CD	470.5669	4.777	-4.235	-9.126	1.980	9.197	7.138	9.615	203.7746	497

Table 18 gives us the summary of the top 20 most-differently regulated edges (MDREs) as determined by the BD test statistic as well as the individual z-scores for each cell line. Note $p < 0.01$ for all of these entries. The individual z-scores reflect the strength of association (log(OR)) as well as the size of the contingency table (SE decreases as the number of observations in the 2x2 table increase; recall $z = \log(OR)/SE$). The table is formatted so that the direction of association for each cell line should be obvious; those in red are concordant (same genes are regulated in the same directions for both perturbagens) and those in blue are discordant (genes in one perturbagen are upregulated but downregulated in the other and/or vice versa). If a

cell has a darker shade of its respective color, the individual z-score is significant at the 0.05 level (not controlling for multiple testing; $z < -1.96$ or $z > 1.96$).

From table **Table 18**, it is obvious that there are a variety of different patterns of concordance across the cell lines that will result in a large BD test statistic. The concordance pattern for the most-differentially regulated edge, KRAS-BRAF, is easily interpretable – this edge is highly concordant in all cell lines with the exception of PC3, in which case it is very discordant. The canonical relationship between KRAS and BRAF is one of activation and is part of a signaling module with two other genes, MEK and ERK [47]. In the case of KRAS-BRAF, we might be able to guess how the pairwise differences playout – we would expect delta values to be negative when PC3 is the reference cell line and positive when it is the comparison cell line and for the magnitude of the delta values to be relatively large in comparison to other cell line pairs. Also, since fewer LM genes fall in concordant cells for PC3, we might expect fewer overlaps between PC3 and other cell lines.

In **Figure 27** we have both visual and numeric representations of the pairwise comparisons and can see that there are large differences in concordance between PC3 and other

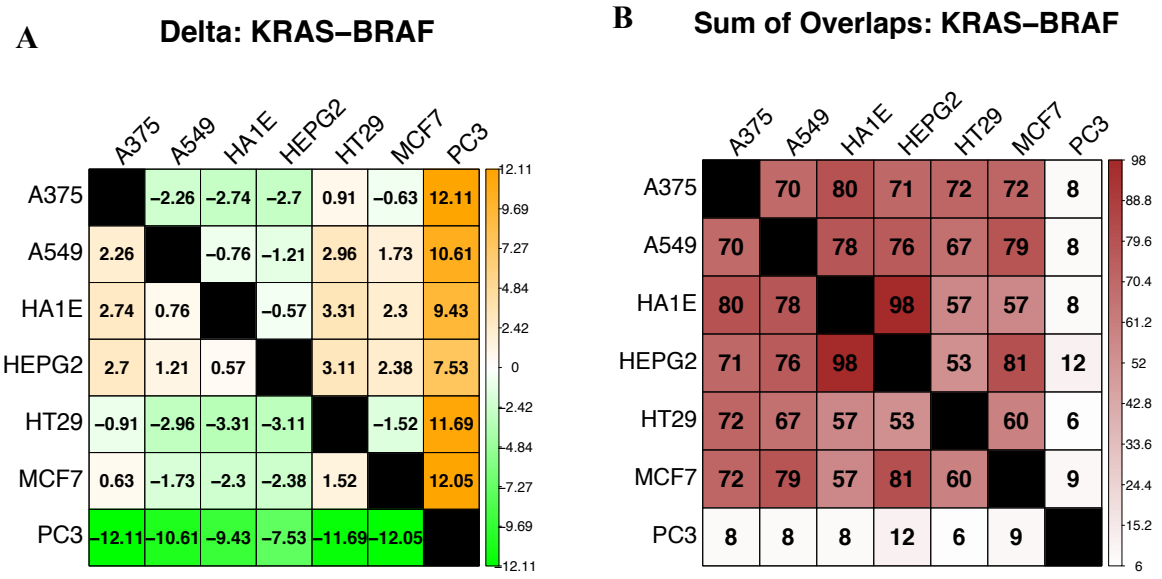


Figure 27: Pairwise values for delta (**A**) and the $\sum overlaps$ (**B**) for edge KRAS-BRAF. In (**A**) the cell lines in the rows serve as the reference cell line and the columns represent the comparison cell lines.

cell lines and minor differences when other cell lines are being compared. The other cell lines exhibit greater degree of similarity as measured by the $\sum overlaps$ whereas PC3 has very few overlapping genes with the other cell lines. As we might expect, it shares the most overlaps with HEPG2 – the comparison cell line with the smallest delta value when compared to PC3.

Table 19: 20 smallest BD edges

Edge	BD	A375	A549	HA1E	HEPG2	HT29	MCF7	PC3	ξ	∑∑overlaps
CREBBP_HK1	3.533683	0.451	-1.359	-0.143	0.097	-1.237	0.429	0.247	42.48344	302
ATP6V1A_LAMTOR3	3.848689	1.276	0.244	0.303	1.853	1.579	1.168	2.492	13.89474	342
GOT1_LDHB	3.918992	1.897	2.548	1.878	2.399	2.665	3.485	0.872	46.33557	447
ARAF_MAP2K1	4.758209	1.174	0.966	1.811	0.605	-0.487	-0.640	0.229	25.75979	383
CDO1_GOT1	4.798181	0.839	1.345	2.441	2.214	3.393	2.543	2.172	27.03650	411
BCL2L1_BAK1	5.839157	-0.645	-1.055	-0.174	-0.880	-0.173	1.882	0.069	30.17891	313
TP53_BAX	6.340465	0.343	1.354	1.133	-0.839	-0.957	-0.536	1.251	32.69969	323
EIF3J_EIF2S2	6.352989	0.290	1.200	0.549	1.525	-1.512	-0.391	-0.159	27.47059	374
PRKCA_GNG8	7.394749	-0.213	-0.669	-0.311	1.910	-1.283	0.357	1.220	44.01036	386
GNG8_MAPK13	7.469179	3.474	3.332	3.123	0.735	1.761	1.578	2.329	35.10000	420
TGFB1_FOXP3	8.073621	1.247	3.020	0.961	-0.090	0.520	2.358	0.032	15.04505	333
MAPK12_FOSL2	8.216305	-0.425	1.522	-0.816	1.198	-1.175	-0.323	-1.552	33.08649	370
ATIC_HPRT1	8.446237	1.388	3.090	1.493	2.414	1.941	0.694	4.090	32.26415	318
ARAF_MAPK1	8.550567	2.188	0.054	0.141	-0.849	1.185	0.316	2.387	26.94839	310
SMAD2_RORC	8.866169	3.879	2.198	5.368	3.325	3.882	3.100	5.166	25.69955	446
GNG8_ADCY9	9.453000	0.839	-0.615	0.311	2.394	2.785	0.253	0.422	30.05063	395
CREB3L4_BCL2	9.466745	6.751	5.708	4.605	5.469	6.773	4.028	4.705	48.01556	514
MAPK13_TP53	9.473969	2.743	-0.225	0.388	-0.826	1.376	1.344	1.889	38.59574	329
ACAT1_HADH	9.599824	3.633	4.180	4.483	3.923	3.869	0.998	4.398	16.11111	378
IFNGR2_NFKB1	9.756997	5.184	4.083	3.173	2.370	2.062	3.856	4.865	37.37265	373

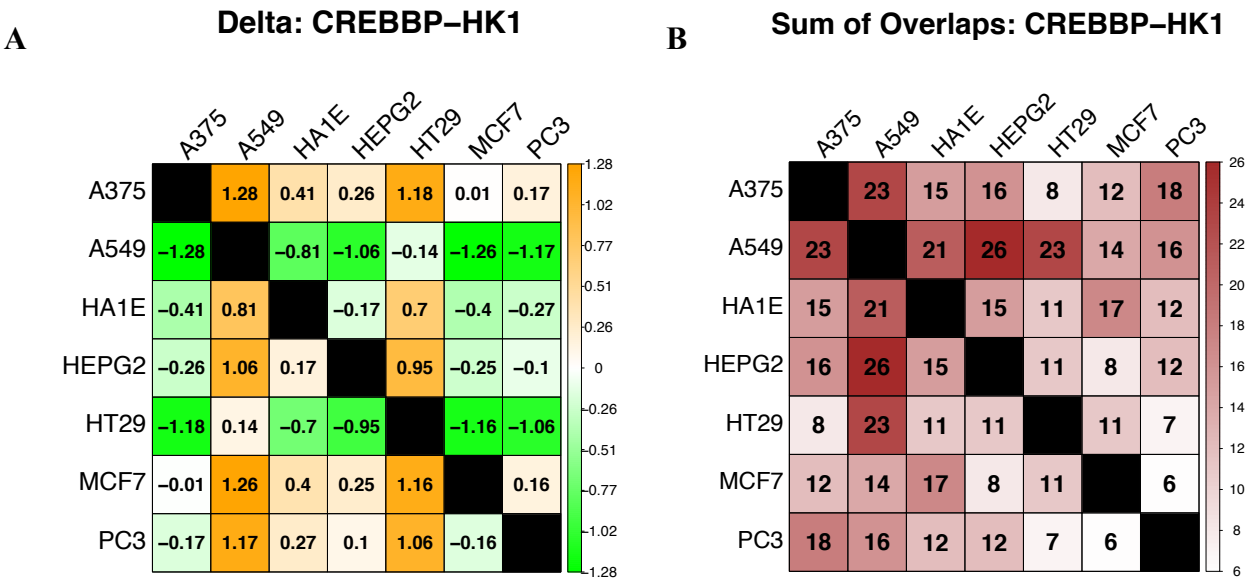


Figure 28: Pairwise values for delta (**A**) and the $\sum overlaps$ (**B**) for edge CREBBP-HK1. In (**A**) the cell lines in the rows serve as the reference cell line and the columns represent the comparison cell lines.

Figure 28 contains pair-wise data for the least differently regulated edge, CREBBP-HK1. In contrast to KRAS-BRAF (**Figure 27)** the magnitudes of both delta values and sums of overlaps are small across the board and there is no obvious discernable pattern. This edge is found in the HIF-1 (hypoxia induced factor -1) signaling pathway of KEGG's database and is the relationship is coded as "expression" – that is to say that CREBBP activity leads to the expression of HK1. This could explain the lack of a strong signal of a regulatory relationship across cell lines since HK1 expression is an outcome of a signaling cascade as opposed to a protein involved in determining the downstream transcriptional response.

4.3.3 ξ Statistic Across Edges

Edge	ξ	A375	A549	HA1E	HEPG2	HT29	MCF7	PC3	BD	∑∑overlaps
SYK_PIK3CA	580.9587	8.887	8.927	7.653	1.273	-3.774	-0.345	5.098	345.8459	436
PIK3CA_BTK	536.4660	-8.771	-7.890	-9.017	-0.484	5.656	7.680	-6.489	452.7533	412
PROCR_THBD	524.8257	-8.257	5.800	7.530	9.236	-3.019	9.431	-1.620	562.6316	608
MET_PIK3CA	522.2500	9.011	8.466	7.319	3.483	3.426	-3.532	8.989	317.7880	560
IL2_EGFR	497.0473	8.526	7.768	9.477	-5.575	7.868	2.024	1.327	379.8102	697
PDGFRA_JAK1	464.8362	3.588	-9.289	-7.148	-8.858	9.197	-1.461	-6.579	506.8152	409
PDPK1_AKT3	463.6505	3.020	6.059	8.303	1.506	2.865	3.078	9.023	266.4211	721
FADD_CFLAR	458.4468	7.227	8.418	8.161	-5.486	5.731	9.228	7.947	265.8390	658
PLCG1_PIK3CA	452.3060	9.054	9.291	8.228	2.528	0.625	1.403	8.580	169.1154	464
PDGFRA_PIK3CA	435.4271	-0.861	8.504	9.043	4.606	-1.161	-1.073	7.908	261.3040	377
SYK_BTK	434.8380	-6.945	-8.330	-8.832	5.891	-6.926	-1.692	1.886	325.1812	358
SYK_VAV3	428.6203	8.932	8.354	8.435	3.819	-3.562	4.877	1.504	294.5493	561
PIK3CA_PDPK1	427.1597	-8.432	-8.913	-8.759	1.952	8.063	1.841	-5.054	369.0411	457
PIK3CA_VAV3	406.4044	9.238	9.086	8.844	0.618	3.258	-5.788	5.027	348.0241	549
BTK_PLCG1	403.4076	-8.265	-7.260	-6.565	1.553	-7.443	2.303	-9.388	202.9278	368
KRAS_RPS6KA1	401.9721	-9.678	-9.159	-8.278	-3.689	7.396	-2.558	7.985	542.9527	645
PRKACA_BRAF	400.8151	5.277	8.371	9.104	9.169	9.332	-4.377	-0.081	314.0766	611
PDGFRA_KRAS	397.9909	5.935	-9.114	-7.967	-8.639	8.162	-7.383	-7.893	549.6416	661
STAT6_IL2	397.7846	9.209	8.989	7.933	5.426	-4.084	-0.702	3.485	279.4031	455
PLCG1_VAV3	394.6582	8.779	9.126	7.334	5.259	-5.198	6.583	5.726	282.1547	553

In contrast to the BD statistic, $\boldsymbol{\xi}$ is calculated based on pairwise rather than single cell line data; the BD is calculated based on 7 odds ratio table whereas $\boldsymbol{\xi}$ is a function of the sums of overlaps for 21 unique cell line combinations. Despite their disparate derivation, in **Table 20** we can see that edges with larger values of $\boldsymbol{\xi}$ tend to have larger BD values compared to edges with smaller BD values. The correlation between these values in our data set is 0.606. The pairwise $\sum\sum overlaps$ statistics for the single edge with the largest $\boldsymbol{\xi}$ value are summarized in **Figure 29**.

In **Figure 29(A)** we see a wide range of $\sum overlaps$ [2:90]. Cell lines A549, A375 and HA1E have large $\sum overlaps$ amongst themselves as well as moderately large $\sum overlaps$ with PC3, minor $\sum overlaps$ with HEPG2 and very small $\sum overlaps$ with HT29 and MCF7, which have small $\sum overlaps$ values across the board. In **Figure 29(B)** we can see that A375, A549 and HA1E are much more similar with regard to concordance and that set is also more concordant than cell lines HEPG2, HT29, MCF7 and PC3. On the other end of the spectrum, the edge with the smallest ξ value is depicted in **Figure 30**. In **Figure 30(B)**, the only consistent pattern is that the edge KRAS-MAPK2 in HA1E is less concordant than in other cell lines; however, evidence of this pattern is not readily reflected in **Figure 30(A).**

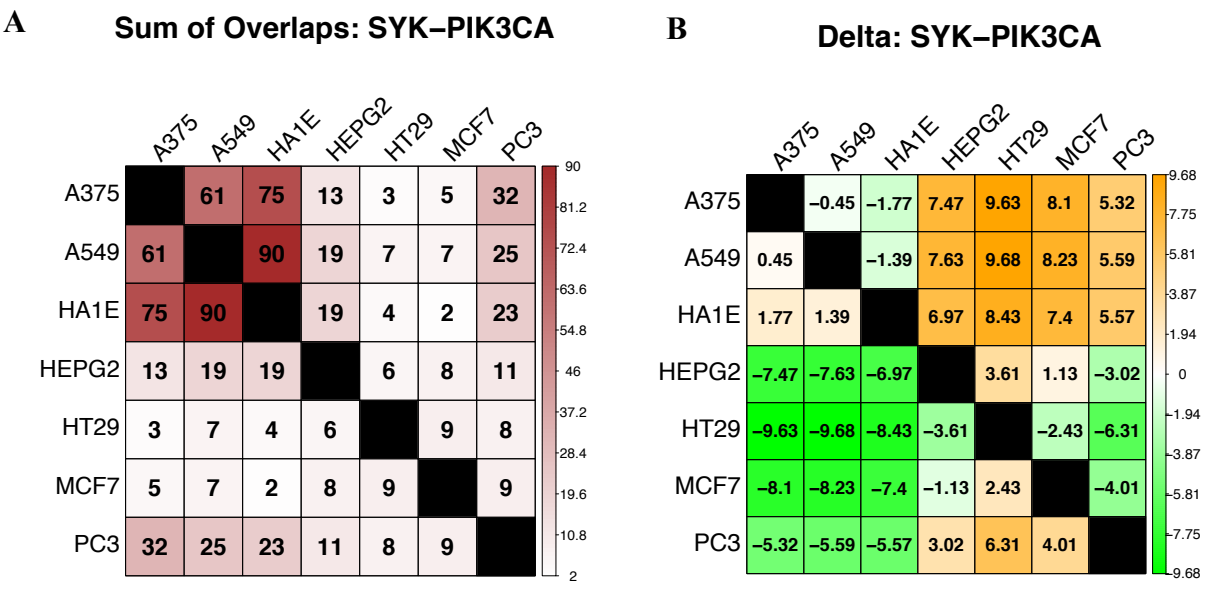


Figure 29: Pairwise values for delta (**A**) and the $\sum overlaps$ (**B**) for edge SYK-PIK3CA. In (**A**) the cell lines in the rows serve as the reference cell line and the columns represent the comparison cell lines.

Table 21 : Edges with the 20 largest ξ values

Edge	ξ	A375	A549	HA1E	HEPG2	HT29	MCF7	PC3	BD	$\sum\sum$overlaps
KRAS_MAPK12	10.27586	0.673	3.955	-7.136	-1.380	1.353	-3.321	2.450	95.263199	406
IFNGR1_FAS	11.71429	0.850	4.835	0.524	-0.326	1.335	1.591	-0.679	20.938761	392
GOT1_TST	11.95217	-0.692	1.264	5.363	1.491	3.883	1.212	4.006	29.336938	460
KCNQ1_ADCY9	12.28994	7.818	-0.318	5.036	-4.301	2.234	2.037	0.325	111.095118	338
IKBKB_NFKB1	13.08527	1.056	-3.985	0.728	-0.156	-1.001	-0.387	-4.965	34.218543	387
TAB3_MAP2K3	13.29032	4.009	0.845	6.088	1.689	0.623	-3.909	-3.881	90.030593	341
OXTR_GNAI2	13.56216	0.044	-4.877	4.164	5.659	1.623	4.883	2.500	86.971206	370
GOT1_MDH2	13.63758	-1.942	1.067	0.601	2.053	3.065	1.657	2.573	16.952847	447
ATP6V1A_LAMTOR3	13.89474	1.276	0.244	0.303	1.853	1.579	1.168	2.492	3.848689	342
MAPK12_FOS	14.15929	1.850	-1.012	-2.971	1.162	-2.325	-2.782	1.328	26.560662	339
FOXO4_CCND2	14.47839	-2.754	1.593	8.291	2.444	5.929	2.396	1.723	92.318802	347
HLA-DMA_HLA-DRB1	14.57485	2.494	-0.945	1.687	3.095	4.386	2.798	2.535	18.043879	334
BCL2L1_BAX	14.71429	0.203	2.111	0.894	2.839	1.202	0.448	-1.239	10.676158	294
RALB_MAPK9	14.72000	2.280	1.764	0.897	-5.944	5.798	-2.341	1.507	90.148770	300
LAP3_GCLM	14.75000	-2.136	2.475	-4.109	-0.324	-2.052	0.327	-0.701	27.279271	400
MAP2K6_MAPK12	14.92432	-1.901	2.823	-7.918	-1.991	-5.364	-3.611	1.717	99.725963	370
CDK2_FOXO4	14.93023	-5.233	0.409	7.062	-1.905	1.710	-0.471	1.382	94.860349	387
TGFB1_FOXP3	15.04505	1.247	3.020	0.961	-0.090	0.520	2.358	0.032	8.073621	333
ALK_NRAS	15.07077	4.013	3.086	-0.690	7.625	7.013	-4.917	1.150	136.897686	325
LSP1_IKBKE	15.07775	0.822	-5.337	-0.516	7.552	2.234	3.627	-1.249	113.253613	373

These edge-wise results for ξ are similar to those shown for the edges with the largest (**Figure 27**) and smallest (**Figure 28**) BD values; at least in the extreme cases, these statistics complement one another with regard to finding patterns of similarities and differences between pairs of cell lines in a targeted manner.

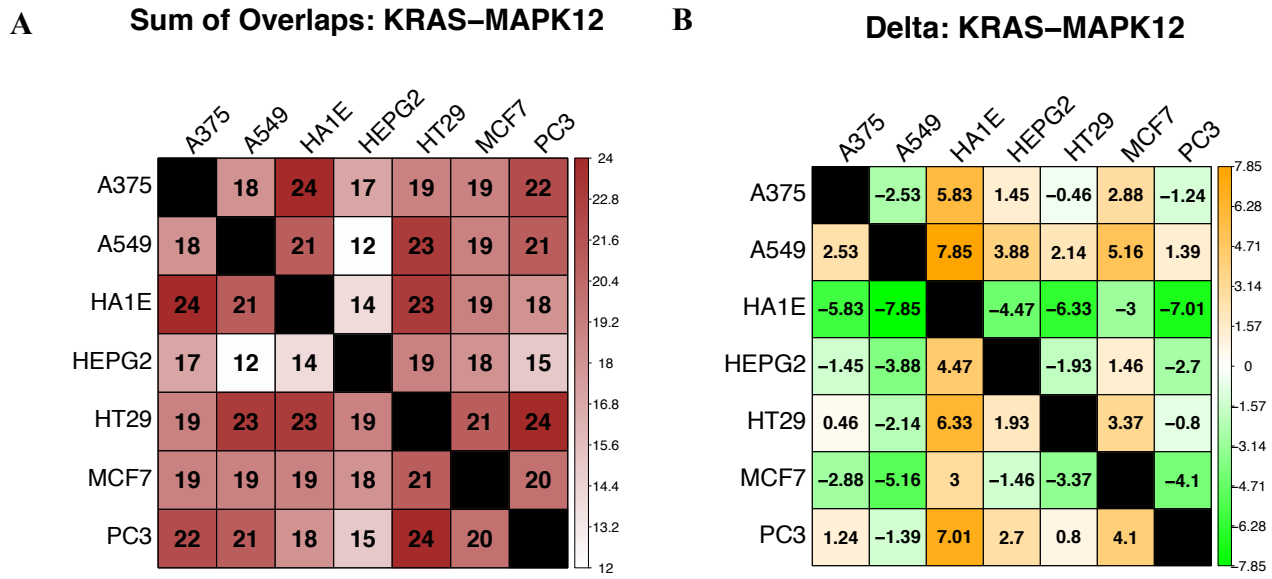


Figure 30: Pairwise values for delta (**A**) and the $\sum overlaps$ (**B**) for edge KRAS-MAPK12. In (**A**) the cell lines in the rows serve as the reference cell line and the columns represent the comparison cell lines.

4.3.4 $\sum\sum overlaps$ Statistic Across Edges

Although the $\sum\sum overlaps$ value is not directly a function of an individual edge's z-score, in **Table 22** we can see that the edges with large $\sum\sum overlaps$ values tend to have a very high degree of concordance within the cell lines nearly across the board. The reasoning behind this is straight forward in one sense; in order for there to be a high degree of overlap between two contingency tables of two cell lines, the individual contingency tables need to have a large number of genes that occupy any of the cells. Large positive z-scores indicate that an individual cell line's contingency table has many entries in the diagonal (up/up and down/down) cells but

very few entries in the off diagonal cells (up/down and down/up) whereas the opposite is true for negative z-scores of a large magnitude.

Table 22 : Edges with the 20 largest $\sum\sum overlaps$ values

Edge	∑∑overlaps	A375	A549	HA1E	HEPG2	HT29	MCF7	PC3	BD	ξ
KRAS_RAF1	1331	6.113	9.169	8.673	8.450	8.020	9.304	8.731	42.71548	163.81818
SRC_KRAS	1327	7.922	8.449	8.811	5.435	8.512	8.770	6.668	96.29924	124.75358
ICAM1_KRAS	1314	8.356	7.782	8.663	8.770	8.485	9.295	9.151	25.50174	97.41096
JUN_MYC	1292	7.811	9.009	5.404	8.971	7.756	7.627	8.716	77.95478	86.55573
STAT1_MYC	1277	9.066	8.455	5.787	7.918	8.261	7.405	7.798	41.48419	48.96006
KRAS_FOS	1223	7.730	9.623	8.232	7.542	8.122	6.788	8.933	54.32463	112.67212
KRAS_JUN	1218	7.261	9.505	5.861	9.419	6.957	9.122	8.619	69.88084	93.44828
FGFR2_KRAS	1198	7.842	9.180	7.483	8.335	6.234	9.370	9.324	69.86402	144.77295
ITGAV_KRAS	1192	8.007	8.839	9.560	8.968	9.027	7.867	8.720	30.41875	97.10403
STAT1_FOS	1175	9.182	9.301	8.007	6.717	8.870	5.083	8.745	49.08559	83.94553
EGFR_KRAS	1160	7.793	7.822	9.352	7.941	8.727	8.779	0.244	197.34496	257.68103
ITGA2_SRC	1145	7.707	8.970	9.443	9.104	8.841	7.057	4.728	100.66940	137.06376
KRAS_BRAF	1122	9.431	8.325	7.449	5.981	8.876	9.431	-7.663	615.31068	353.31551
KRAS_PDPK1	1116	8.946	5.906	7.896	9.023	3.555	4.480	9.283	194.65225	359.23118
NCOA3_MYC	1076	9.285	8.150	6.026	8.915	9.187	6.707	3.867	108.94087	81.18587
ITGA2_CRKL	1072	5.235	8.840	8.579	9.382	9.227	8.393	2.305	140.13661	101.41418
GNG5_KRAS	1071	8.701	8.929	5.284	4.190	8.452	8.910	8.413	126.56704	219.64706
CSNK1E_ARNTL	1063	7.213	8.514	7.387	8.574	9.165	5.910	5.781	91.02144	61.53716
SRC_RAF1	1059	5.433	8.991	9.111	8.509	7.626	8.898	7.557	56.54149	117.65439
EGFR_HRAS	1051	9.537	6.030	8.497	5.611	7.329	8.494	3.721	143.50600	324.27022

In **Table 22** the only edge that does not have a high degree of within-cell line concordance across the board is the edge KRAS-BRAF – this edge is highly discordant in the cell line PC3. Recall that this edge was ranked as the most different with regard to its BD value (**Table 18**) – so despite the heterogeneity between the rest of the cell lines and PC3, the similarity amongst those cell lines is so strong that it still ranks as one of the most similarly-

regulated edges with regard to $\sum\sum overlaps$. One more take-away from **Table 22** is that the individual gene KRAS has a high degree of representation – it is in over half (11) of the top 20 edges – suggesting that the downstream effects of knocking down KRAS are likely very similar across the cell lines.

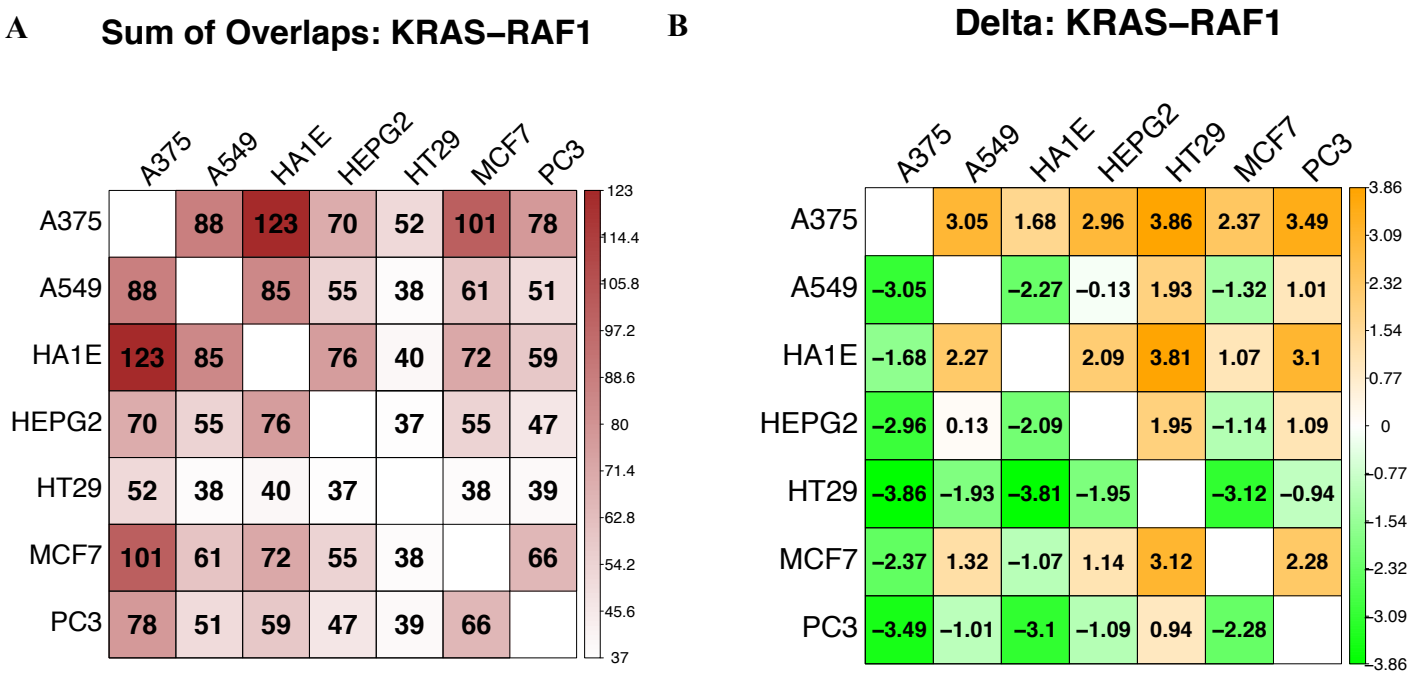


Figure 31: Pairwise values for delta (**A**) and the $\sum overlaps$ (**B**) for edge KRAS-RAF1. In (**A**) the cell lines in the rows serve as the reference cell line and the columns represent the comparison cell lines.

Figure 31 displays the pair-wise similarities/differences for the edge with the largest $\sum\sum overlaps$ – KRAS and RAF1. The canonical relationship between these proteins is direct activation of RAF1 by KRAS ultimately resulting in gene expression via the MEK/ERK pathway [48]. Note that, despite its smaller z-score, the relationship between KRAS and RAF1 is slightly more concordant in A375 than the other cell lines. The reason behind this is that although the "raw" odds ratio is very large in A375, the standard error is larger than that of other cell lines because there are so few observations in the off-diagonal cells. As demonstrated in the corresponding entry in **Table 22**, the relationship between KRAS and RAF1 is highly concordant

across cell lines which – as discussed earlier – allows for the opportunity for large $\sum overlaps$ between cell line pairs. In fact, the $\sum overlaps$ between A549 and HA1E (123) is the largest $\sum overlaps$ for all documented KEGG edge across all cell line pairs.

Table 23 : Edges with the 20 smallest $\sum\sum overlaps$ values

Edge	∑∑overlaps	A375	A549	HA1E	HEPG2	HT29	MCF7	PC3	BD	ξ
FGFR3_PIK3CD	183	0.520	3.480	-0.320	1.848	0.047	-1.756	7.768	73.01916	68.88525
PDGFRA_JAK3	191	2.471	-1.780	8.678	-2.424	4.339	0.499	-5.641	195.88322	71.66492
WNT5A_FZD2	194	5.087	0.728	7.505	0.034	0.393	0.063	0.701	65.46090	36.78351
PLCG1_CALM3	203	-0.002	-2.121	-2.145	-2.005	-7.303	-0.087	-3.335	43.72154	56.55172
ANGPT1_INSR	207	7.471	6.139	-8.034	-5.601	4.806	-1.881	-1.238	327.14182	78.57971
SETD7_FOXO1	209	4.203	6.766	-1.760	-8.223	6.326	0.372	-3.116	222.40914	74.65072
HLA-A_B2M	210	-3.790	-2.728	1.458	8.832	4.387	-4.115	6.921	202.62394	64.20000
PTPN2_JAK3	212	2.850	0.619	8.051	-0.975	1.864	1.563	-1.503	74.74068	30.88679
MCCC2_EHHADH	213	-0.221	-4.529	6.240	-5.126	-5.668	5.246	0.684	158.45640	56.64789
FGFR3_PLCG1	213	-5.630	5.893	-4.616	2.596	-3.067	4.811	-5.997	171.37812	144.19718
PIK3CD_FOXO1	214	-1.099	3.654	6.762	0.754	-8.598	0.457	-5.161	191.56079	152.02804
TFPI_F7	214	-1.349	-1.554	2.639	3.179	1.652	2.400	-2.245	32.29414	33.68224
BTRC_WWTR1	215	3.802	5.967	5.988	1.697	-2.223	2.838	2.674	55.45408	61.71163
IL11RA_JAK3	215	3.282	4.887	-9.030	3.075	1.449	2.312	5.615	203.41669	45.88837
PLCG1_PIK3CD	216	-3.545	4.532	6.184	-0.727	8.646	0.540	-8.361	273.91586	119.80556
PIP4K2B_PIK3CD	216	-2.519	2.904	-6.036	2.486	3.457	-1.269	4.327	93.94492	71.00000
ATF2_CPT1A	216	1.600	1.814	-0.643	-6.531	-4.050	-4.512	0.592	74.91524	31.33333
CTBP2_TCF7L2	217	8.607	1.697	-3.952	4.319	3.316	0.256	0.307	111.28730	58.90323
PDGFRB_SHC1	218	3.754	-2.220	3.353	1.656	-1.711	0.008	-6.378	82.58529	31.49541
GNGT2_MAPK1	218	-8.207	0.497	-4.433	8.073	-4.940	5.053	0.351	243.22543	82.93578

Unlike the edges with large $\sum\sum overlaps$, those with small $\sum\sum overlaps$ tend to have smaller within cell line z-scores or – if the z-scores are moderate – are not in the same direction across cell lines. **Figure 32** displays the same pair-wise information for FGFR3-PIK3CD, the edge with the smallest $\sum\sum overlaps$ from **Table 23**. This relationship if coded as "activation" in KEGG; the individual cell lines PC3, A549 and to a lesser extent HEPG2 support this

directionality/relationship whereas this relationship is not highly concordant or discordant in the other cell lines.

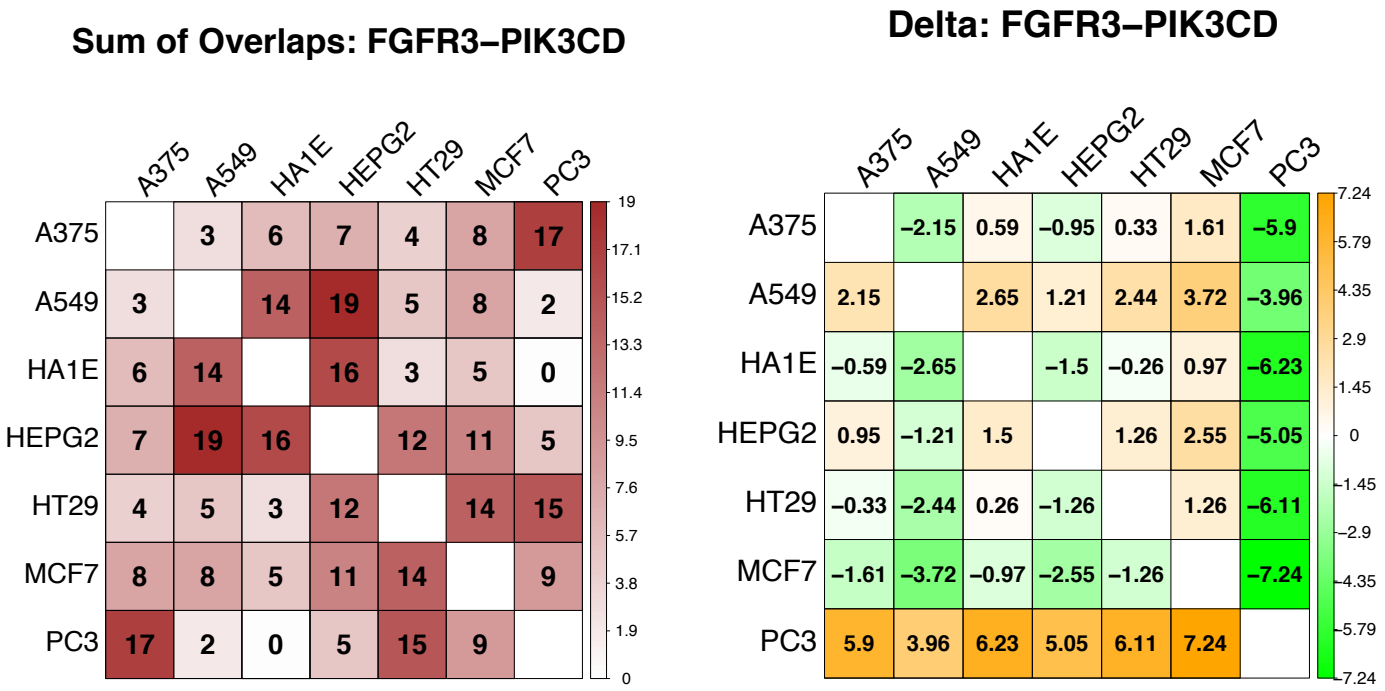


Figure 32: Pairwise values for delta (**A**) and the $\sum overlaps$ (**B**) for edge FGFR3-PIK3CD. In (**A**) the cell lines in the rows serve as the reference cell line and the columns represent the comparison cell lines.

4.4 Pathway Level Results

4.4.1 Results: Distribution of PCH and POH Across Pathways

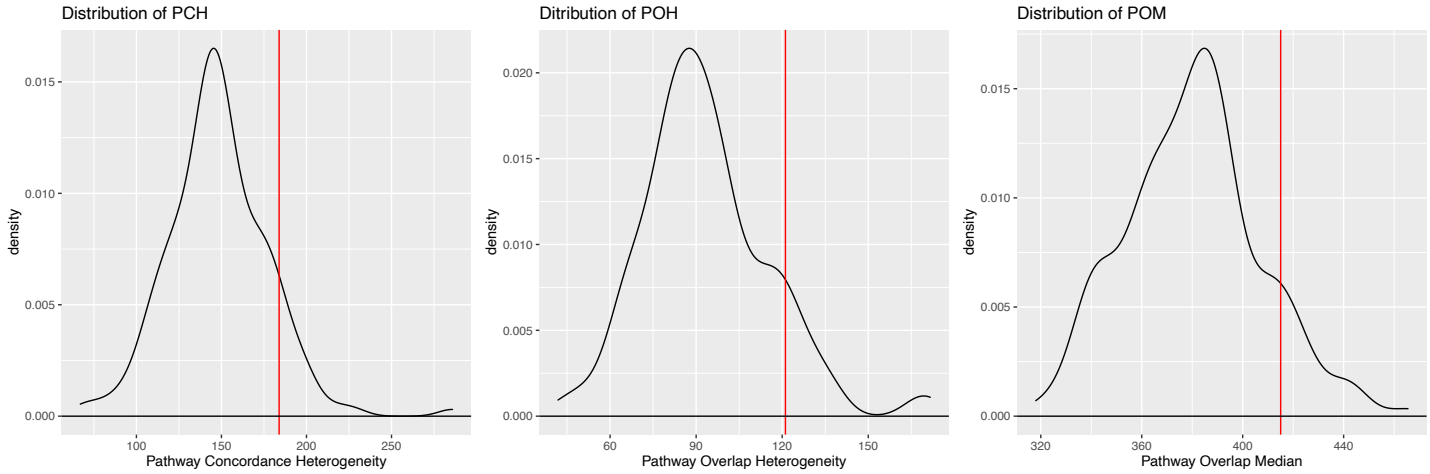


Figure 33: Distribution of Pathway Concordance Heterogeneity (PCH) score (**A**), Pathway Overlap Heterogeneity (POH) score (**B**) and Pathway Overlap Median (POM) score (**C**) across 160 KEGG pathways. Red lines indicate 90th percentile; 183.92 for PCH, 121.13 for POH and 415 for POM.

Figure 33 gives a summary of the concordance-based (PCH) and $\sum overlaps$-based pathway scores (POH and POM). All distributions are roughly bell-shaped with outliers at the right-hand tail of the distributions.

4.4.2 Results: Most Differently Regulated Pathways across Cell Lines

Both POH and PCH are measurements of differences between cell lines and complement one another, they will be observed in parallel. In **Figure 34** the distributions of these values are depicted; the correlation between the POH and PCH is 0.88. In **Table 24,** the top 20 of the 160 pathways considered in this analysis are ranked by their POH score and the same is done for PCH (**Table 25**). Notice that these tables also contain columns to indicate how many [of the 21 pair-wise] comparisons are significantly different in terms of pathway concordance (mean delta) or similarity (median $\sum overlaps$). Although the pathways may be ranked by these values

Table 25 : Top 20 KEGG Pathways ranked by pathway concordance heterogeneity (PCH)

Rank	PCH	POH	Title	Pathway Code	Category	Subcategory	#Edges	#Sig. Delta	#Sig. ∑overlaps
1	285.87	171.61	Aldosterone-regulated sodium reabsorption	hsa04960	Organismal Systems	excretory system	11	8	12
2	227.42	169.99	Spinocerebellar ataxia	hsa05017	Human Diseases	neurodegenerative disease	21	4	17
3	215.99	138.25	Platelet activation	hsa04611	Organismal Systems	immune system	36	3	17
4	203.43	116.06	Longevity regulating pathway - multiple species	hsa04213	Organismal Systems	aging	61	4	14
5	200.46	120.95	Progesterone-mediated oocyte maturation	hsa04914	Organismal Systems	endocrine system	34	1	17
6	199.64	166.05	Fc epsilon RI signaling pathway	hsa04664	Organismal Systems	immune system	57	0	12
7	197.99	129.93	Melanoma	hsa05218	Human Diseases	cancer specific types	106	1	16
8	197.01	127.95	Regulation of lipolysis in adipocytes	hsa04923	Organismal Systems	endocrine system	23	5	12
9	194.89	93.46	Type II diabetes mellitus	hsa04930	Human Diseases	endocrine and metabolic disease	26	0	10
10	190.88	120.26	Colorectal cancer	hsa05210	Human Diseases	cancer specific types	96	0	20
11	188.61	120.08	ErbB signaling pathway	hsa04012	Environmental Information Processing	signal transduction	99	0	19
12	187.65	123.15	Central carbon metabolism in cancer	hsa05230	Human Diseases	cancer overview	94	0	20
13	186.43	113.16	TNF signaling pathway	hsa04668	Environmental Information Processing	signal transduction	54	2	17
14	185.34	119.47	Neurotrophin signaling pathway	hsa04722	Organismal Systems	nervous system	143	0	15
15	184.75	126.65	Renal cell carcinoma	hsa05211	Human Diseases	cancer specific types	41	0	15
16	184.25	131.34	Endometrial cancer	hsa05213	Human Diseases	cancer specific types	63	0	20
17	183.88	127.10	Thyroid hormone signaling pathway	hsa04919	Organismal Systems	endocrine system	70	0	16
18	183.00	123.79	EGFR tyrosine kinase inhibitor resistance	hsa01521	Human Diseases	drug resistance antineplastic	118	0	16
19	181.69	137.88	Choline metabolism in cancer	hsa05231	Human Diseases	cancer overview	53	0	17
20	181.44	117.14	Non-small cell lung cancer	hsa05223	Human Diseases	cancer specific types	76	0	18

Table 25: Top 20 KEGG Pathways ranked by pathway overlap heterogeneity (POH)

Rank	POH	PCH	Title	Pathway Code	Category	Subcategory	#Edges	#Sig. Delta	#Sig. ∑overlaps
1	171.61	285.87	Aldosterone-regulated sodium reabsorption	hsa04960	Organismal Systems	excretory system	11	8	12
2	169.99	227.42	Spinocerebellar ataxia	hsa05017	Human Diseases	neurodegenerative disease	21	4	17
3	166.05	199.64	Fc epsilon RI signaling pathway	hsa04664	Organismal Systems	immune system	57	0	12
4	138.25	215.99	Platelet activation	hsa04611	Organismal Systems	immune system	36	3	17
5	137.88	181.69	Choline metabolism in cancer	hsa05231	Human Diseases	cancer overview	53	0	17
6	135.54	176.56	GnRH secretion	hsa04929	Organismal Systems	endocrine system	32	3	14
7	132.86	169.94	Fc gamma R-mediated phagocytosis	hsa04666	Organismal Systems	immune system	30	0	12
8	131.34	184.25	Endometrial cancer	hsa05213	Human Diseases	cancer specific types	63	0	20
9	129.93	197.99	Melanoma	hsa05218	Human Diseases	cancer specific types	106	1	16
10	127.95	197.01	Regulation of lipolysis in adipocytes	hsa04923	Organismal Systems	endocrine system	23	5	12
11	127.10	183.88	Thyroid hormone signaling pathway	hsa04919	Organismal Systems	endocrine system	70	0	16
12	126.65	184.75	Renal cell carcinoma	hsa05211	Human Diseases	cancer specific types	41	0	15
13	123.79	183.00	EGFR tyrosine kinase inhibitor resistance	hsa01521	Human Diseases	drug resistance antineplastic	118	0	16
14	123.23	172.09	Prostate cancer	hsa05215	Human Diseases	cancer specific types	154	0	15
15	123.15	187.65	Central carbon metabolism in cancer	hsa05230	Human Diseases	cancer overview	94	0	20
16	122.76	177.02	Glioma	hsa05214	Human Diseases	cancer specific types	99	0	14
17	120.95	200.46	Progesterone-mediated oocyte maturation	hsa04914	Organismal Systems	endocrine system	34	1	17
18	120.26	190.88	Colorectal cancer	hsa05210	Human Diseases	cancer specific types	96	0	20
19	120.17	151.04	Platinum drug resistance	hsa01524	Human Diseases	drug resistance antineplastic	50	0	15
20	120.08	188.61	ErbB signaling pathway	hsa04012	Environmental Information Processing	signal transduction	99	0	19

themselves, the discrete-ness of the data would result in many ties but could still be differentiated by POH and PCH (see **Figure 35**).

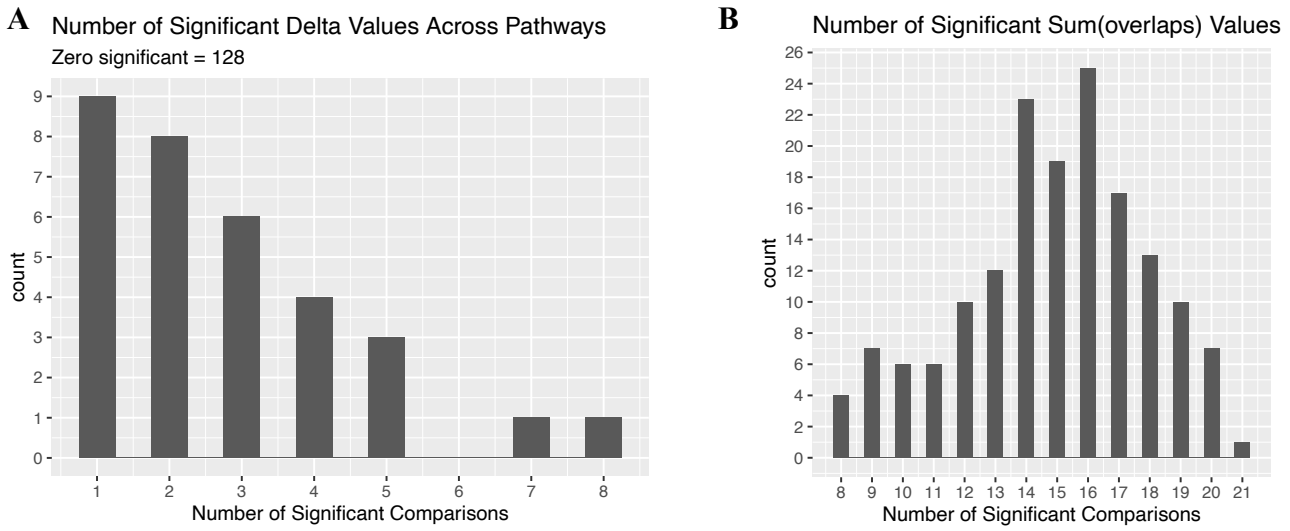


Figure 34: Discrete distribution of the number of significant delta values (**A**) and $\sum overlaps$ (**B**) across all pathways. Note that (**A**) is truncated at 1 due to a very large number (128 out of 160) of pathways that do not have any significant pathway delta values.

The first major take-away from **Table 24** and **Table 25** is the similarity of the pathways that show up in each table. In fact, the intersection of pathways in both tables is 15 (75%); those that are unique to the top 20 of either the POH or PCH ranking are highlighted in yellow (5 pathways each). While a few (2) Environmental Information Processing pathways are included in the top 20 lists (ErbB signaling pathway [both] and TNF signaling pathway [top 20 by PCH]), the majority of pathways are classified as either Human Disease (10) or Organismal Systems (8). Notice that many of the most heterogeneous pathways as defined by PCH do not have any significant delta values – in these cases, despite the individual edges showing large difference in concordance amongst all cell lines, there is not evidence to suggest that one cell line is overall more or less concordant than another cell line.

The first pathway in both lists is the Aldosterone-regulated sodium reabsorption pathway – this pathway is also the pathway with the largest number of significant delta values. From **Figure 35A**, we can see that HEPG2 and MCF7 are more concordant than cell lines A375, A549 and HA1E; HT29 and PC3 are also more concordant than HA1E. The patterns in the median $\sum overlaps$ (**Figure 35B**) are not quite as obvious, but in general HT29 and MCF7 are more different than HT29 A375, A549 and HA1E but are similar to one another – which reflects the patterns seen in **Figure 35A**. However, somewhat counter-intuitively, the pairwise comparison with the largest pathway delta, HEPG2 and A375, has the second highest number of median $\sum overlaps$.

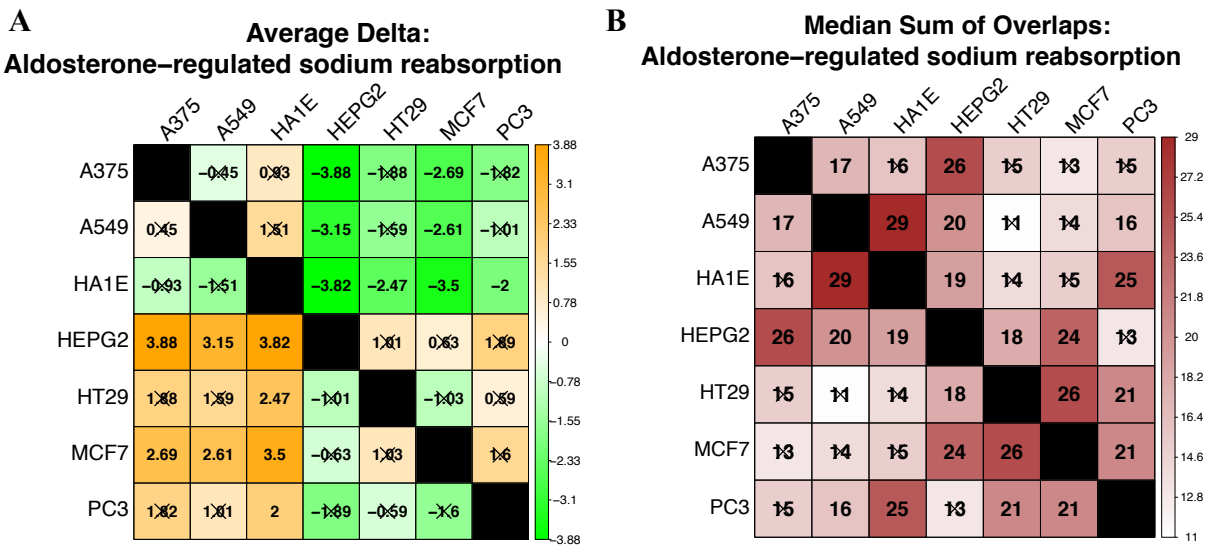


Figure 35: Matrices showing the pairwise pathway average delta (A) and median $\sum overlaps$ (B) for the aldosterone-regulated sodium reabsorption pathway.

The only issue with this particular pathway is that it just makes the criteria for inclusion in the analysis with 11 edges (recall that the criteria for inclusion is at least 10 edges) – so it could be the case that one or two edges are driving the results. Furthermore, we would like to see if the patterns of difference are similar between and across cell lines. To investigate this issue and introduce the structure of KEGG pathways, we will generate a pathway map with the

KEGGlincs package, compare it to the original KEGG rendering and take a look at a complimentary matrix-type visualization of the edge-level test statistics.

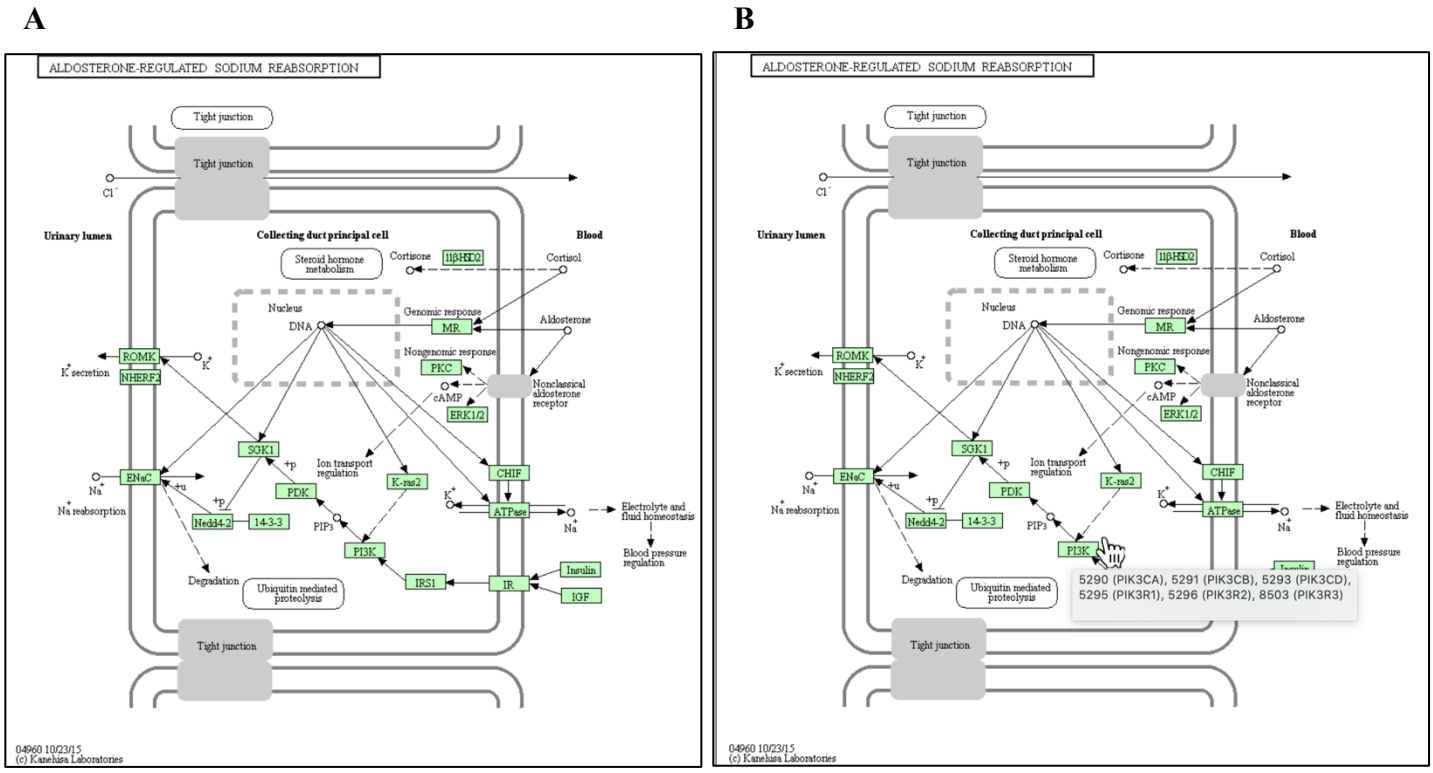


Figure 36: Renderings of the aldosterone-related sodium reabsorption pathway from KEGG from the website 'as is' (A) and with the pointer hovered above a gene with paralogues (B).

The aldosterone-regulated sodium reabsorption pathway depicted in its original form [i.e. exactly as it appears on the pathway's landing page when 'homo sapiens' is the species at KEGG's website [12]] in **Figure 36A** and again with a tiny difference in **Figure 36B** to show an important but perhaps not fully explicit feature of KEGG pathway maps. In **Figure 36B** there is a gray box in the lower right-hand corner of the map; this box contains the gene symbols as well as KEGG's own internal reference codes ("KO code") for all of the genes that are represented at a particular node. In this case, the node labeled "PI3K" is actually 6 different nodes (PIK3CA, PIK3CB, PIK3CD, and PIK3R1, PIK3R2, and PIK3R3)) that are represented by one node. One

advantage of using KEGGlincs to render the pathway maps, as detailed in **Chapter 5**, is that this information is made explicitly clear as demonstrated in **Figure 37**. Note that in an interactive Cytoscape session each edge will have its own unique paralog combination label when a cursor is hovered over that edge.

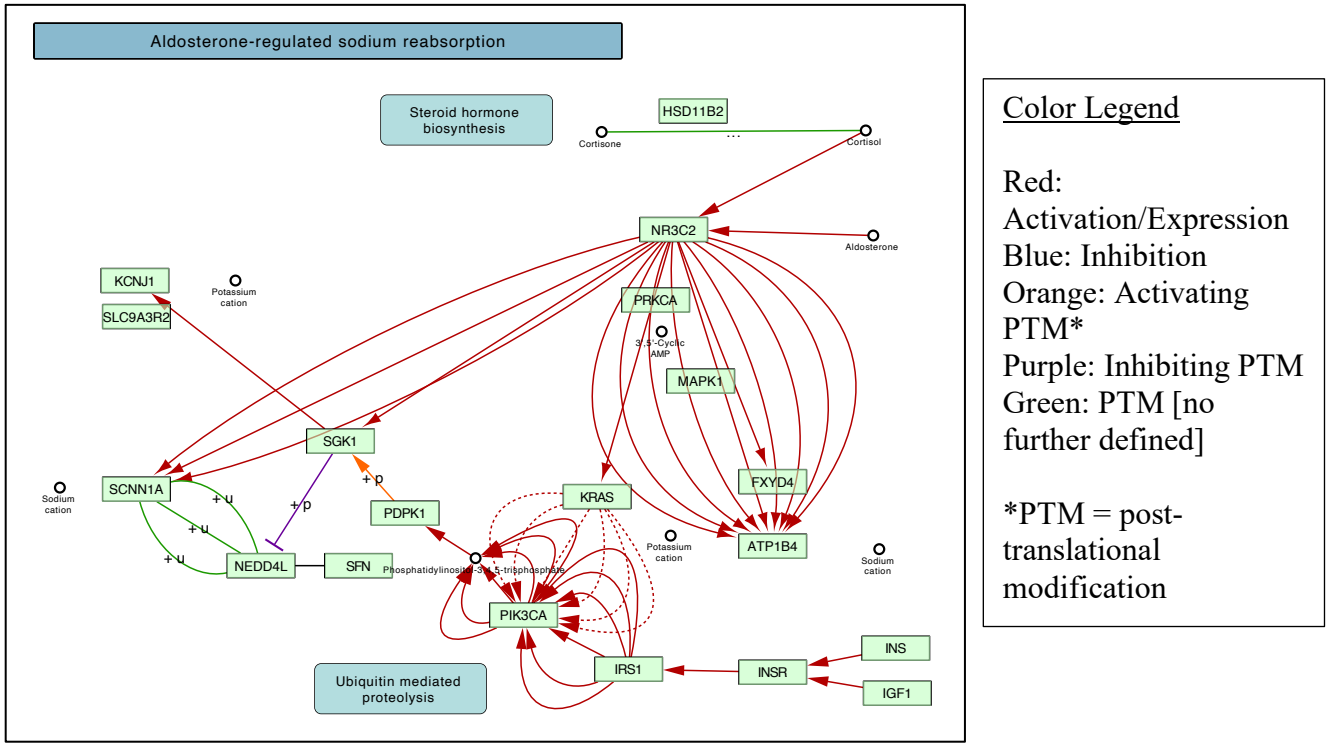


Figure 37: Rendering of the aldosterone-related sodium reabsorption pathway using KEGGlincs to visualize the exact nature (in terms of number and type) of relationships among genes in KEGG pathways.

This edge-based, or relationship-based approach to the representation of pathway edges is amenable to the type of data we have been generating throughout this project as is showcased in **Figure 38**. **Figure 38** shows which of the KEGG pathway edges appear in our data and how relationships in those edges compare between two cell lines in the L1000 data set using the delta statistic for the comparison of cell lines HEPG2 and A375 (the pairwise comparison with the largest delta value in **Figure 35A**). There are a few key points to be gleaned from the data

represented **Figure 38**. First of all, we can see that most edges are more concordant in HEPG2 (9 out of 11) and that the most concordant edge (the edge with the largest width) is between KRAS and a paralogue of PIK3CA. To figure out exactly which edge this is, and exactly which edges are available from our data set, we could scroll over it in an interactive Cytoscape session; this document includes this information in **Figure 41A**. In this case, we can see that relationships from IRS1 and KRAS to PIK3CA, PIK3CB and PIK3R1 are all more concordant in HEPG2 whereas the relationships from IRS1 and KRAS to the paralogue PIK3CD are more concordant in A375.

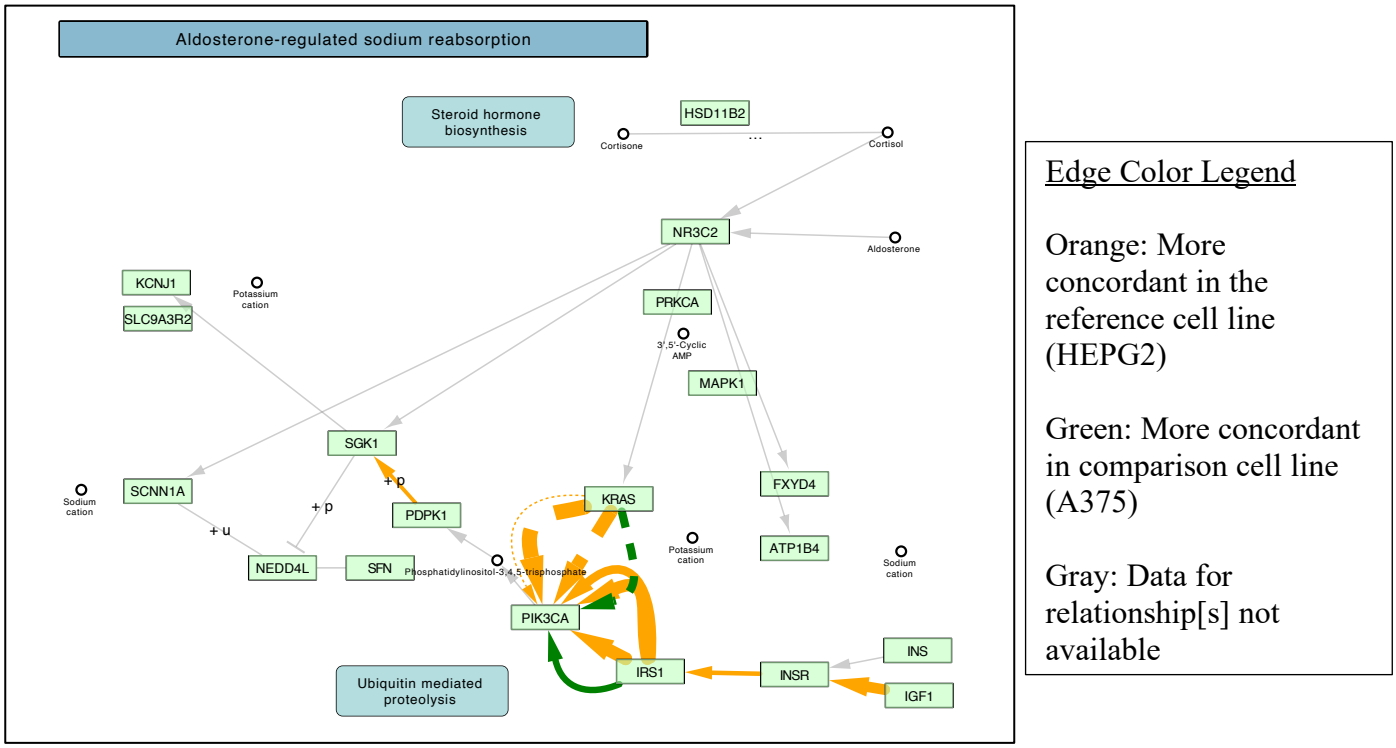


Figure 38: Rendering of the aldosterone-related sodium reabsorption pathway for HEPG2 vs A375 using KEGGlincs functions that use mapping information from KEGG and are formatted according to delta values between cell lines using L1000 data. The width of the edge represents the magnitude of delta whereas the width represents delta's direction.

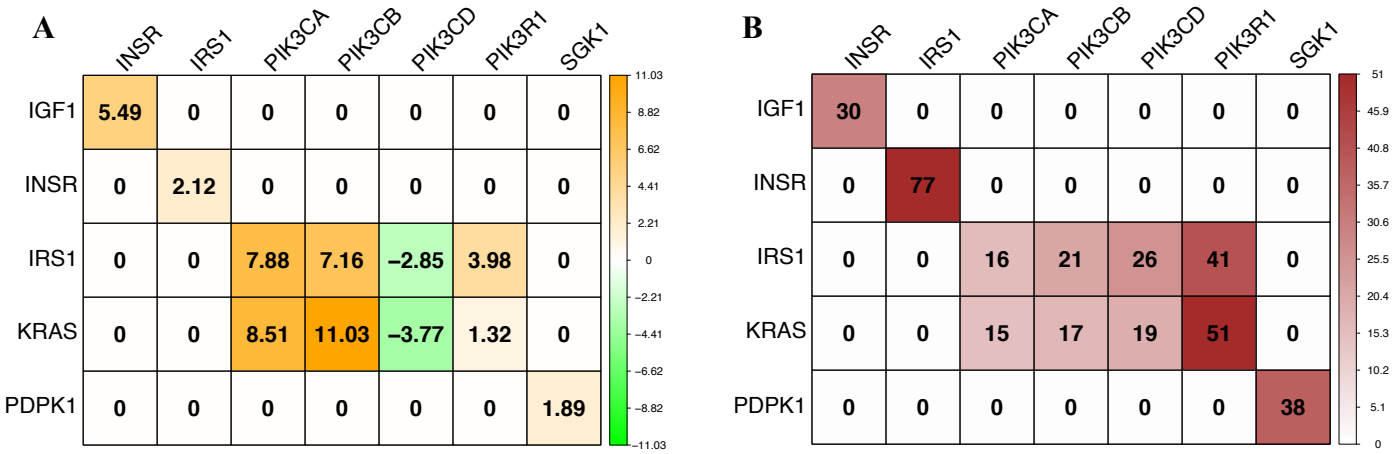


Figure 39: The edge-wise deltas (A) and $\sum overlaps$ (B) between nodes in the aldosterone-related sodium reabsorption pathway for HEPG2 vs A375. The genes in the rows are the 'from'/originating nodes and those in the columns are the 'to'/terminating nodes.

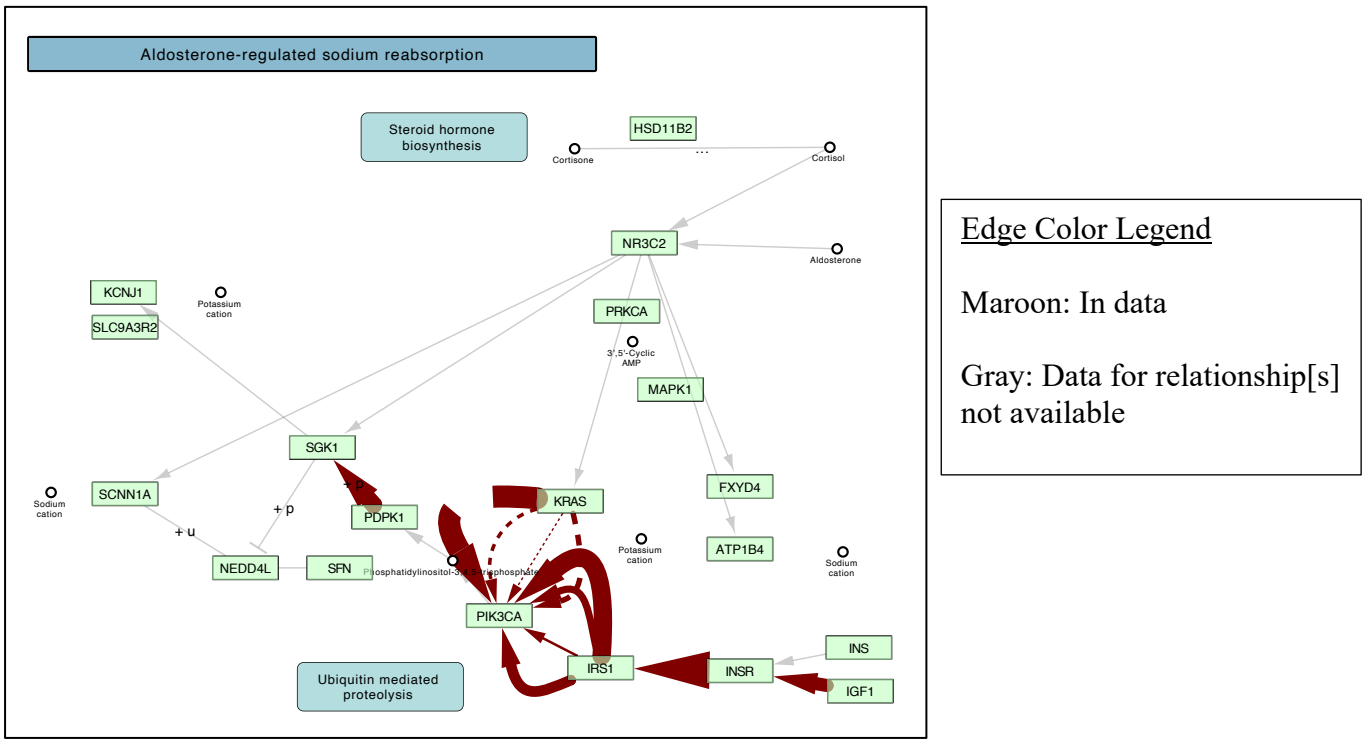


Figure 40: Rendering of the aldosterone-related sodium reabsorption pathway for HEPG2 vs A375 where the width of the edge is formatted to reflect the magnitude of $\sum overlaps$ values between cell lines using L1000 data.

In **Figure 40** the $\sum overlaps$ statistic is represented with the exact values for each edge reported in **Figure 39B**. Here, the cell lines appear to be the most similar in edge IGF1-IRS1 and least similar in the relationships between IRS1/KRAS and PIK3CA – a pattern reflected in the relatively low delta for IGF1-IRS1and relatively large deltas for IRS1/KRAS and PIK3CA. As previously mentioned, the number of pathway edges for this first example pathway is relatively low. Even though there was not one particular edge driving the difference in concordance between HEPG2 and A375, there are many edges in this pathway that are not available in our dataset. Rather than this entire pathway being characteristic of overall differences in concordance among cell lines, it is likely that this method has identified a submodule of genes that exhibit heterogeneous behavior across cell lines.

4.4.2 Results: Most Similarly Regulated Pathways across Cell Lines

Table 26: Top 20 KEGG Pathways ranked by pathway overlap median (POM) score

Rank	POM	Title	Pathway Code	Category	Subcategory	#Edges	#Sig. Delta	#Sig. ∑overlaps
1	465.5	Herpes simplex virus 1 infection	hsa05168	Human Diseases	infectious disease viral	92	0	20
2	443.5	Central carbon metabolism in cancer	hsa05230	Human Diseases	cancer overview	94	0	20
3	442.0	Estrogen signaling pathway	hsa04915	Organismal Systems	endocrine system	85	0	20
4	442.0	Spinocerebellar ataxia	hsa05017	Human Diseases	neurodegenerative disease	21	4	17
5	442.0	T cell receptor signaling pathway	hsa04660	Organismal Systems	immune system	87	0	19
6	434.0	Endocrine resistance	hsa01522	Human Diseases	drug resistance antineplastic	141	0	20
7	426.5	Colorectal cancer	hsa05210	Human Diseases	cancer specific types	96	0	20
8	425.0	Non-alcoholic fatty liver disease (NAFLD)	hsa04932	Human Diseases	endocrine and metabolic disease	44	0	18
9	423.0	Acute myeloid leukemia	hsa05221	Human Diseases	cancer specific types	98	0	21
10	423.0	Endometrial cancer	hsa05213	Human Diseases	cancer specific types	63	0	20
11	422.0	ErbB signaling pathway	hsa04012	Environmental Information Processing	signal transduction	99	0	19
12	422.0	Pertussis	hsa05133	Human Diseases	infectious disease bacterial	31	2	17
13	419.0	Yersinia infection	hsa05135	Human Diseases	infectious disease bacterial	89	0	18
14	417.0	Hepatitis C	hsa05160	Human Diseases	infectious disease viral	80	0	18
15	417.0	Thyroid hormone signaling pathway	hsa04919	Organismal Systems	endocrine system	70	0	16
16	415.5	Chemokine signaling pathway	hsa04062	Organismal Systems	immune system	136	0	18
17	415.0	Choline metabolism in cancer	hsa05231	Human Diseases	cancer overview	53	0	17
18	415.0	Small cell lung cancer	hsa05222	Human Diseases	cancer specific types	67	0	16
19	414.0	B cell receptor signaling pathway	hsa04662	Organismal Systems	immune system	61	0	19
20	413.0	VEGF signaling pathway	hsa04370	Environmental Information Processing	signal transduction	62	3	15

In **Table 26** we have the top 20 pathways ranked by the POM score – a score that indicates the over-all similarity of the downstream effects for perturbagens represented in that pathway. In this table, the pathways highlighted in orange are the pathways that also make an appearance in the top 20 lists ranked by PCH and POH. Note that are no pathways shared by POM and PCH or POM and POH that are not shared by PCH and POH. Even though there is a higher degree of overlap for the pathways represented by PCH and POH (15/20) vs POM and the combination of PCH and POH (6/20), the degree of similarity is at first surprising given that these statistics, in theory, are meant to capture different aspects of the data. However, recall that, for individual edges, the edge with the largest BD value (KRAS-BRAF) is also among the edges with the largest $\sum\sum overlaps$.

This situation is likely playing out at the level of the pathway as well; the differences attributed to certain edges may arise due to one cell line acting as the 'odd one out' vs the others that have more in common amongst themselves for that edge compared to other edges that are not different across the board. This may be the reason behind why so many pathways on this list fall under the subcategory "cancer specific subtypes" despite all of the cell lines representing different cancers or, in the case of HA1E, a cell line whose mutations have led to immortality akin to those of the cancer cell lines. Each of these pathways could be explored using the functionality provided by KEGGlincs, as is detailed in the following chapter.

Chapter 5: R Bioconductor Package: KEGGlincs

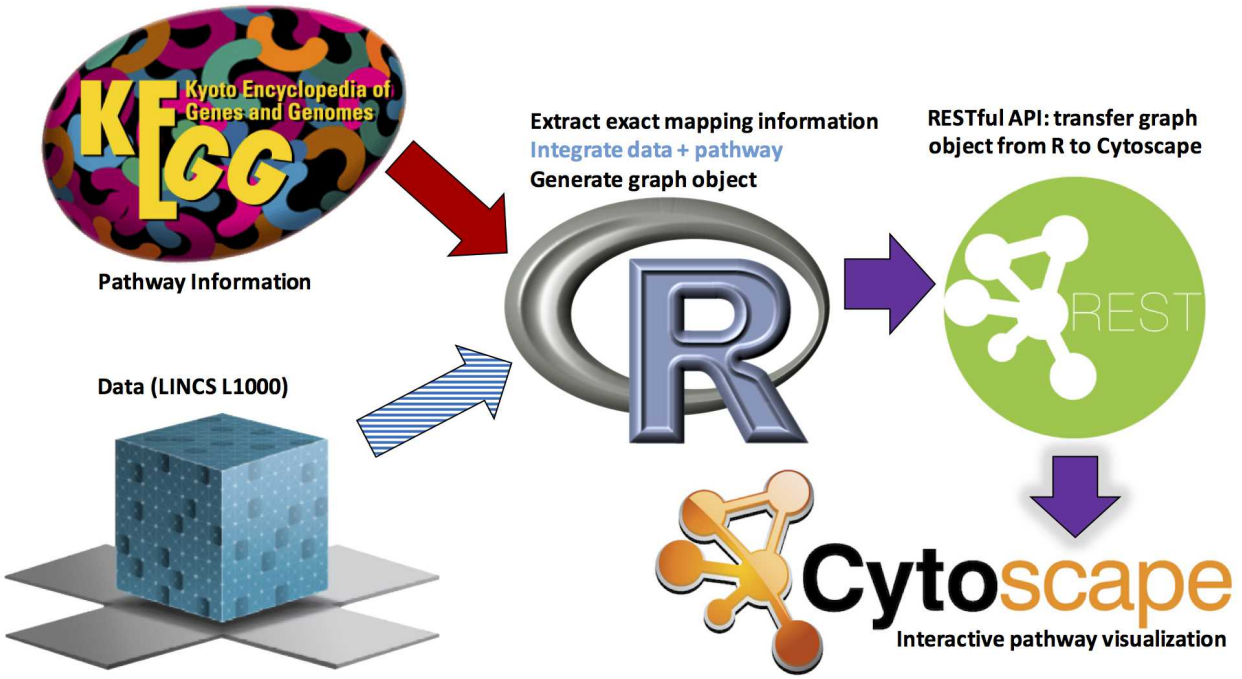


Figure 41: A summary of information retrieval and processing for KEGGlincs.

5.1 KEGGlincs Introduction

The package KEGGlincs and the functions contained within it are designed such that users can explore KEGG pathways in a more meaningful and informative manner both visually and analytically. This method of pathway analysis is approached via functions that handle the following (related) objectives:

- ‘Expanding’ node mapping for [paralogous node entries and grouped node entries
- Allowing data to be explicitly mapped to ‘expanded’ pathway edges (no summarization necessary)

The idea of ‘expanded’ nodes and edges should become very clear after reviewing the following example KEGGlincs workflows. Please keep in mind, the individual functions detailed in the

following workflows are incorporated into the KEGGlincs 'master function'; these workflows are designed to provide users a with a better understanding of how this function works, how pathway topology is represented in KGML files, and how this package could be used with non-LINCS edge data (see Workflow 2).

5.2 KEGGlincs Workflow 1: No data added

This workflow is intended to give users insight into the 'expansion' of KEGG pathway mapping via manipulation of the source KGML file. The only input required is the KEGG pathway ID for your pathway of choice. The primary goal for this method of pathway re-generation is to give users insight into the complexity that underlies many KEGG pathways but is in a sense 'hidden', yet hard-coded, in the curated KGML files. Users can also see the *exact* pathway topology that is used for input in analyses such as SPIA (Signaling Pathway Impact Analysis). The example is given for the FoxO signaling pathway and by following the steps below, users can re-create their own KEGG pathway maps as well as retrieve the information used to explicitly define any of the KEGG pathway architecture.

Step1: Initialize KEGGlincs package

```
library(KEGGlincs)
```

Step2: Download and parse the most current KGML file for Fox0 signaling pathway

```
FoxO_KGML <- get_KGML("hsa04068")

#Information from KGML can be accessed using the following syntax:
slot(FoxO_KGML, "pathwayInfo")
## [ Title ]: FoxO signaling pathway
## [ Name ]: path:hsa04068
## [ Organism ]: hsa
## [ Number ] :04068
## [ Image ] :http://www.kegg.jp/kegg/pathway/hsa/hsa04068.png
```

```
## [ Link ] :http://www.kegg.jp/kegg-bin/show_pathway?hsa04068
```

The code chunks below are useful for viewing the original pathway image within an R session:

```
#Get address for pathway with active links:
slot(slot(FoxO_KGML, "pathwayInfo"), "image")
## [1] "http://www.kegg.jp/kegg/pathway/hsa/hsa04068.png"
#Download a static pathway image (png file) to working directory:
image_link <- slot(slot(FoxO_KGML, "pathwayInfo"), "image")
download.file(image_link, basename(image_link), mode = "wb")
```

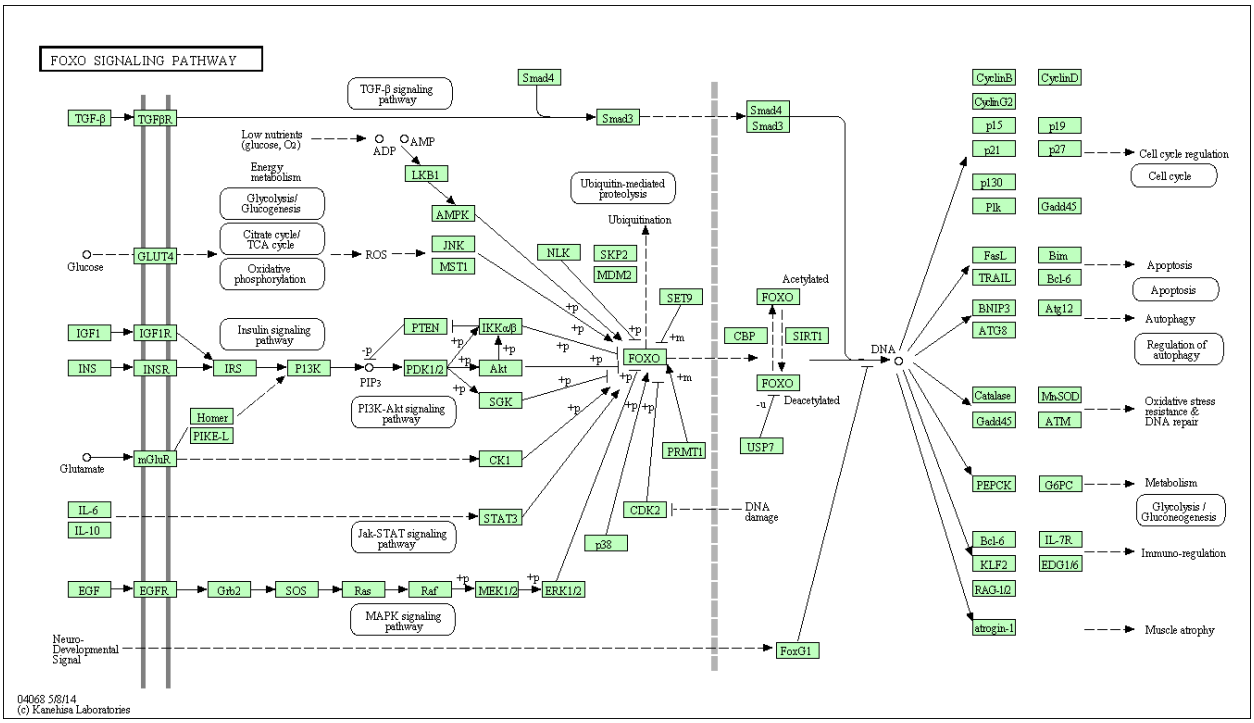


Figure 42: Rendering of the .png file for the p53 signaling pathway from KEGG

The following commands produce 'expanded' node and edge sets Note that KEGG IDs are converted to gene/compound symbols; this conversion accounts for the majority of computing time behind the `expand_KEGG_mappings` function. For quicker map generation, users may choose to change the argument `convert_KEGG_IDs` to FALSE; this will result in edges being identified by pairs of accession numbers instead of symbols in the final pathway map (example at end of this workflow using KEGG_lincs master function).

```
FoxO_KEGG_mappings <- expand_KEGG_mappings(FoxO_KGML)
FoxO_edges <- expand_KEGG_edges(FoxO_KGML, FoxO_KEGG_mappings)
```

Option - Compare counts for 'expanded' vs. 'unexpanded' nodes and edges:

```
length(graph::nodes(FoxO_KGML)) # 'Un-expanded' nodes
## [1] 98
nrow(FoxO_KEGG_mappings)        # 'Expanded' nodes
## [1] 164

length(graph::edges(FoxO_KGML)) # 'Un-expanded' edges
## [1] 78
nrow(FoxO_edges)                # 'Expanded' edges
## [1] 457
```

Step3: Add graphing information to nodes and edges and get graph object

```
#Modify existing data sets; specify as nodes and edges
FoxO_node_mapping_info <- node_mapping_info(FoxO_KEGG_mappings)
FoxO_edge_mapping_info <- edge_mapping_info(FoxO_edges)

#Create an igraph object
GO <- get_graph_object(FoxO_node_mapping_info, FoxO_edge_mapping_info)
class(GO)
## [1] "igraph"
```

Step 4: Transform graph object and send to Cytoscape

```
cyto_vis(GO, "FoxO Pathway with Expanded Edges[no data added]")
```

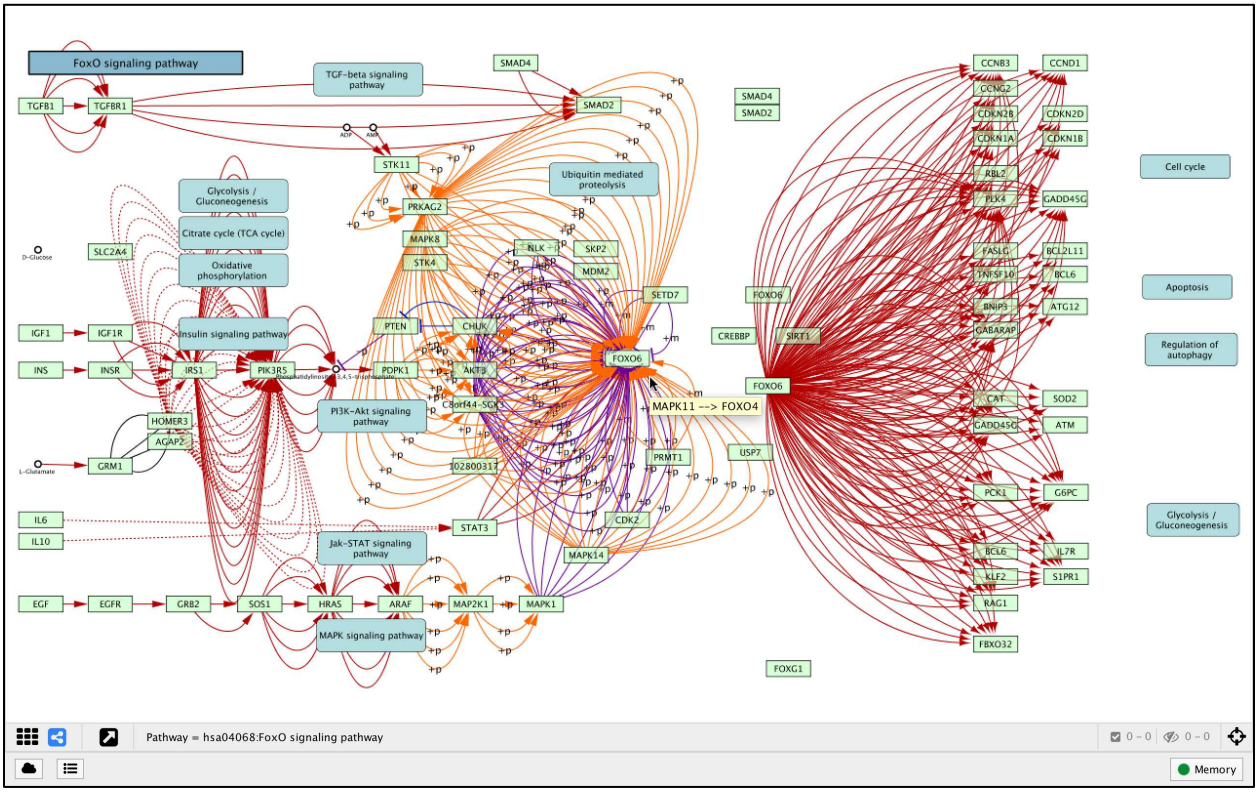


Figure 43: Rendering of the FoxO pathway via KEGGlincs

Edge Color Key:

Red: Activation or Expression *

Orange: Activating PTM **

Green: PTM (no activation/inhibition activity defined)

Blue: Inhibition

Purple: Inhibiting PTM

Black(solid): Binding/Association

Black(dashed): Indirect effect (no activation/inhibition activity defined)

*Any dashed colored line indicates that the effect is indirect

**PTM = post-translational modification or, as KEGG defines them, ‘molecular events’.

- The specific types of PTMS (indicated by edge label) include:
  - +p: phosphorylation
  - -p: dephosphorylation
  - +g: glycosylation
  - +u: ubiquitination
  - +m: methylation

Notice that the original KEGG pathway image (**Figure 42**) includes visual elements such as cellular-component-demarcations and certain edges (especially those 'connecting' genes to other pathways) that are not rendered in Cytoscape (**Figure 43**). These are features that are either not explicitly part of the pathway topology (i.e. not nodes or edges connecting nodes) or have not been hard-coded in the KGML file. The node labels may also differ between maps (KEGGlincs labels nodes as the first 'alias' in the respective KGML slot as there is no corresponding 'label' slot).

The steps above may be avoided if the user does not wish to generate intermediary files/objects by making use of the function `KEGG_lincs` as follows:

```
KEGG_lincs("hsa04068")
```

If users would like the Cytoscape-rendered map along with the detailed list of expanded edges (as an R object), `KEGG_lincs` can be invoked as follows:

```
FoxO_edges <- KEGG_lincs("hsa04068")
```

5.3 KEGGlincs Workflow 2: Overlay data to edges of KEGG pathway

Specific use case: LINCS L1000 Knock-out data

While the functions described in Workflow 1 are certainly useful for any users wishing to gain deeper insight into KEGG pathway topology and 'hard-coded' KGML information, the driving force motivating the KEGGlincs package development is the association of experimental data with pathway edges. The companion data package *KOdata* provides data for the edges rendered by the master function `KEGG_lincs`. This data package includes two unique data sets; one contains lists of significantly up- and down-regulated genes corresponding to knocked-out

genes (within individual experiments, genes are 'turned off' via shRNA) across a variety of cell lines measured at specific times and the other is a binary record of baseline gene expression (gene is either expressed or not expressed) for most cell lines from the knock-out data set. Note that the data in this set contains CGSs set at the original threshold of 50 as decided by the BROAD institute. While this package was developed primarily as a way to compare pathway topology between cell lines or within cell lines [across time] using LINCS L1000 data, this workflow will demonstrate the package's flexibility for users that wish to incorporate edge data from any source.

Example: Comparing p53 Signalling Pathway between cell lines.

As a hypothetical scenario, our goal will be to compare pathway topology between cell lines for an important cancer-related pathway: the p53 Signaling Pathway.

The 'default' pathway (with no data added to edges) can be generated either by following Workflow 1 or by using the `KEGG_lincs` master function as follows:

```
KEGG_lincs("hsa04115")
```

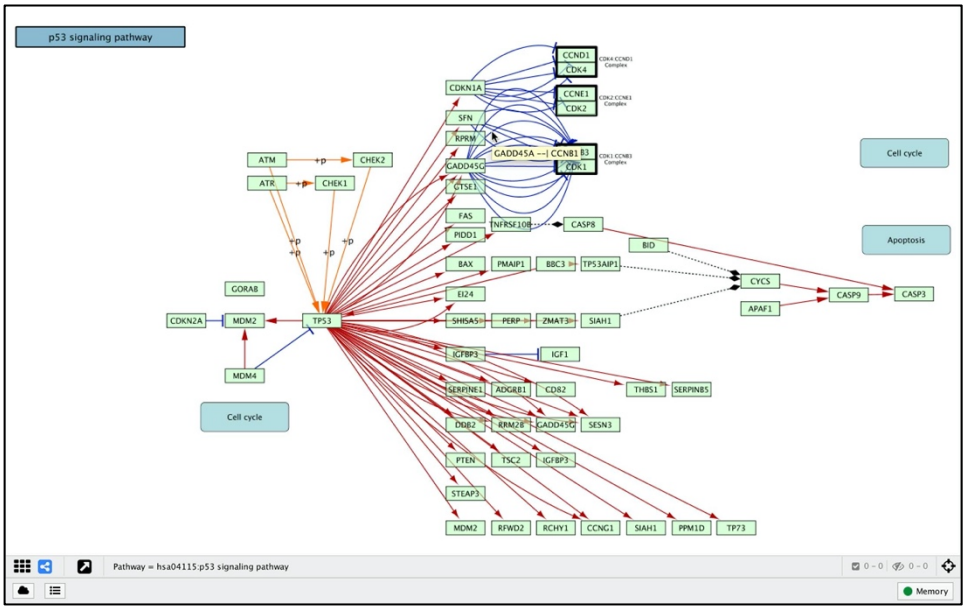


Figure 44: Rendering of the p53 signaling pathway via KEGGlincs

Note that the data available in KOdata is not limited to the 7 cell lines with 2,004 common perturbages that we have been focusing on thus far in this project. An important aspect of the L1000 knock-out and expression data is that it is incomplete; experimental data is not uniformly available for each cell line. Therefor (for this specific example with this specific data set) it is instructive to find out which cell lines make sense to compare; intuitively, cell lines with a similar percentage of pathway genes knocked out would be well suited for comparison. The following command accomplished this task in the form of an easily interpretable graphical output:

```
path_genes_by_cell_type(p53_KEGG_mappings)
```

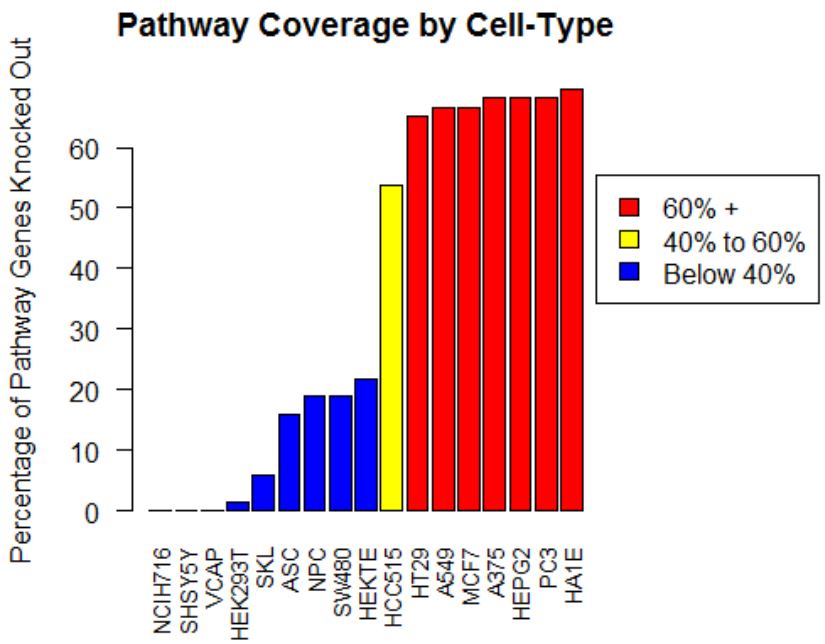


Figure 45: Bar chart of percentage of gene perturbations available for a given cell line in the p53 pathway.

The bar chart in **Figure 45** suggests that the group of cell lines colored in red have similar amounts of pathway information; for this example, we will compare the PC3 (prostate cancer) and HA1E (immortalized normal kidney epithelial) cell lines. Note that the seven cell lines that

we have used in our analyses discussed in other chapters are well represented in this pathway as well as others which is the reason that they were chosen to begin with. The following commands use the data objects generated above to generate cell line specific edge attributes corresponding to specific pathway edges and the information from the L1000 knock-out data set:

```
p53_PC3_data <- overlap_info(p53_KGML, p53_KEGG_mappings, "PC3")
## Number of genes documented in selected pathway = 72
## Number of pathway genes in dataset = 48
## Coverage = 66.67%

p53_HA1E_data <- overlap_info(p53_KGML, p53_KEGG_mappings, "HA1E")
## Number of genes documented in selected pathway = 72
## Number of pathway genes in dataset = 50
## Coverage = 69.44%
```

The following function `add_edge_data` can be used with any dataset with gene symbols in the first two columns and will append selected columns to the edge dataset. Note that the data supplied does not need to be pre-arranged in correct source-to-target order as specified by the pathway topology; the function automatically re-orients pairs correctly.

```
p53_PC3_edges <- add_edge_data(p53_edges, p53_KEGG_mappings,
                                  p53_PC3_data, only_mapped = TRUE,
                                  data_column_no = c(3,10,12))
## Number of edges documented in selected pathway = 92
## Number of edges with corresponding user data = 60
## Coverage = 65.22%
p53_HA1E_edges <- add_edge_data(p53_edges, p53_KEGG_mappings,
                                p53_HA1E_data, only_mapped = TRUE,
                                data_column_no =  c(3,10,12))
## Number of edges documented in selected pathway = 92
## Number of edges with corresponding user data = 64
## Coverage = 69.57%
```

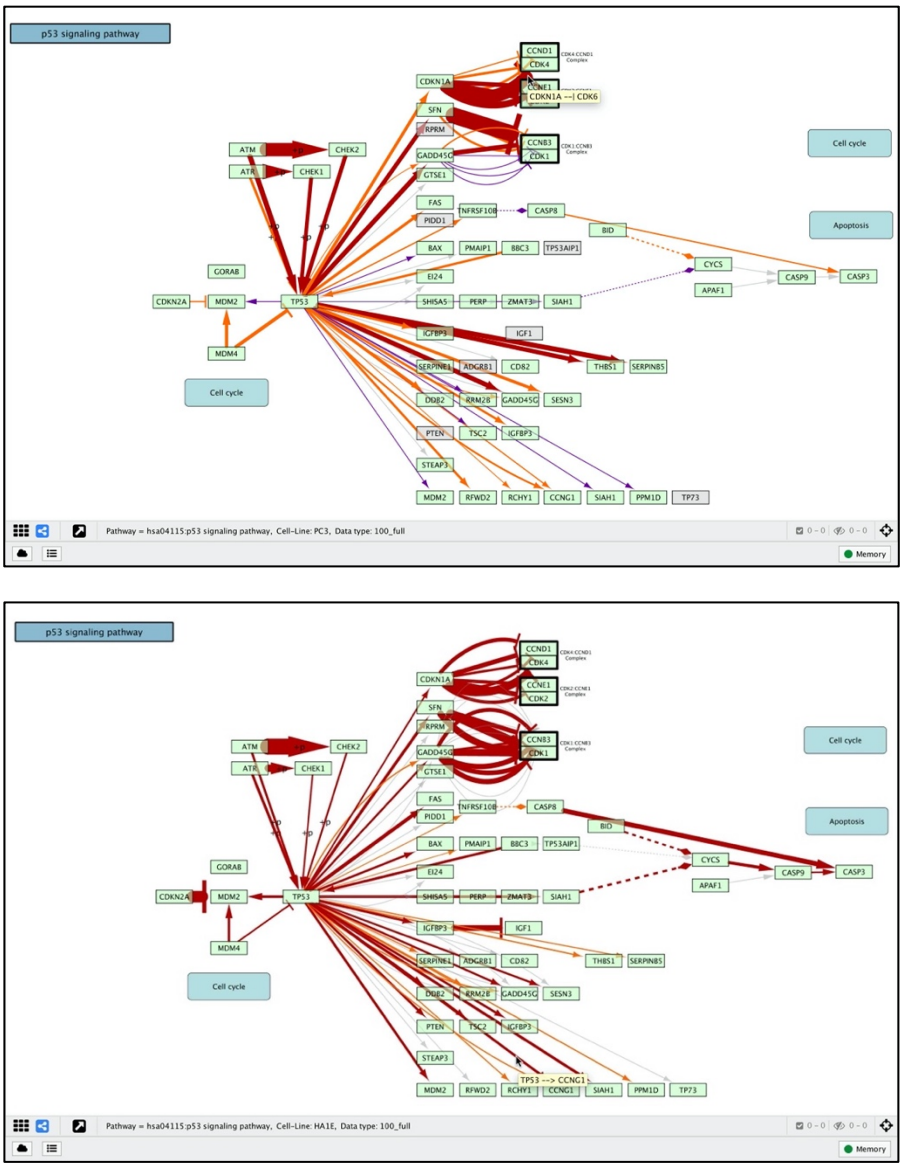
The following series of commands follow from Workflow 1 (with minor adjustments to arguments, notably ensuring that `data_added = TRUE` is specified). The edges in the resulting pathway maps are conditionally formatted to represent both the significance and magnitude of the relationship between corresponding nodes based on their concordance/discordance of up/down-regulated genes as measured by Fisher's Exact Test.

```
p53_node_map <- node_mapping_info(p53_KEGG_mappings)

p53_edge_map_PC3 <- edge_mapping_info(p53_PC3_edges, data_added =
                                      TRUE,significance_markup = TRUE)
p53_edge_map_HA1E <- edge_mapping_info(p53_HA1E_edges, data_added =
                                     TRUE, significance_markup = TRUE)


PC3_GO <- get_graph_object(p53_node_map, p53_edge_map_PC3)
HA1E_GO <- get_graph_object(p53_node_map, p53_edge_map_HA1E)

cyto_vis(PC3_GO, "Pathway = p53, Cell line = PC3")
#Option: Save PC3 as .cys file and start a fresh session in Cytoscape
cyto_vis(HA1E_GO, "Pathway = p53, Cell line = HA1E")
```

As with Workflow 1, the KEGG_lincs master function can automatically generate pathway maps identical to the final maps resulting from Workflow 2 as follows:

```
KEGG_lincs("hsa04115", "PC3", refine_by_cell_line = FALSE)
KEGG_lincs("hsa04115", "HA1E", refine_by_cell_line = FALSE)
```

These pathway maps are shown in **Figure 46**. Note that in Cytoscape graphs rendered in the same session inherit certain style elements from existing graphs that are not updated when the new graph gets pushed (such as range for conditional formatting); therefor it is best to start with a fresh session when mapping requires conditional formatting. The edge colors represent the following possible combinations of direction of Fisher's Exact Test summary scores (a modified Odd's Ratio score; either positive(+) or negative (-)) and their corresponding p-values.



Edge Color Key

Red: OR(+), pval(sig)

Orange: OR(+), pval(non-sig)

Purple: OR(-), pval(non-sig)

Blue: OR(-), pval(sig)

Figure 46: p53 signaling pathway with conditionally-formatted edges that represent the within-cell line concordance of the edges in PC3 (top) and HA1E (bottom).

Finally, the function `KL_compare` can be used to generate a graph that compares the concordance of edges between cell lines. Note that reversal of the order of cell lines will result in opposite coloration. The 'first' cell line is the 'reference' cell line (in this example, PC3) whereas the second cell line is the 'comparison' cell line (in this example HA1E). The map above was produced with the following command:

```
KL_compare("hsa04115", "PC3", "HA1E")
```

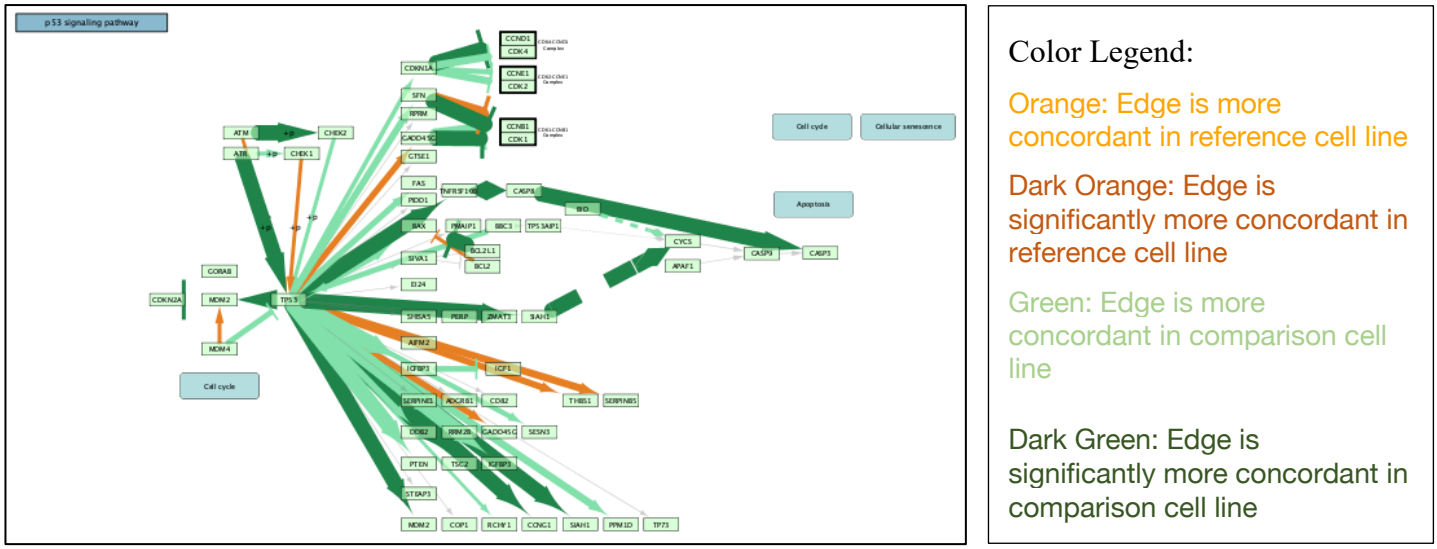


Figure 47: p53 signaling pathway with conditionally-formatted edges that represent the comparison of concordance measurements between the cell lines PC3 (reference cell line) and HA1E (comparison cell line).

5.3 KEGGlincs Example and Discussion: The ErbB Signaling Pathway

The ErbB signaling pathway – also refered to as the epidermal growth factor or EGFR pathway - describes the complex relationships involved in relaying signals from the environment to proteins in the cell's membrane that ultimately reach the nucleus and affect cellular growth. This pathway is perhaps most notably associated with breast cancer [49] but research suggests that it is involved with many different types of cancers including skin [50], colon [51], lung [52], liver [53], prostate [54] and renal [55] – ie cancers represented by all of our model cell lines. This pathway is also in the top 20 pathway lists for each of the pathway-level outcome measurements discussed in Chapter 4.

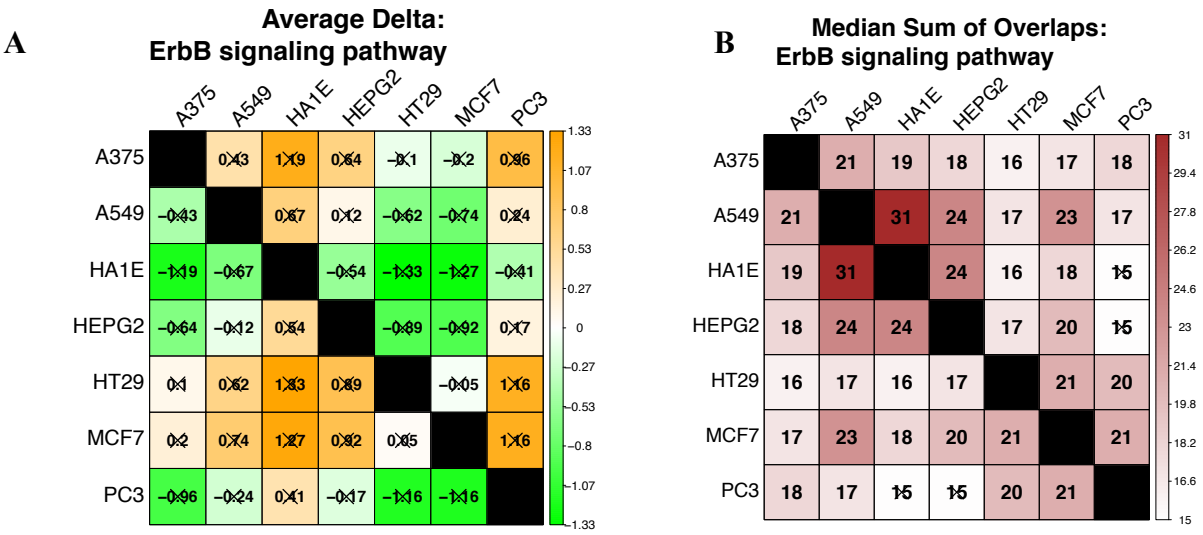


Figure 48: Matrics of the average delta values (**A**) and median $\sum overlaps$ (**B**) for the ErbB signaling pathway.

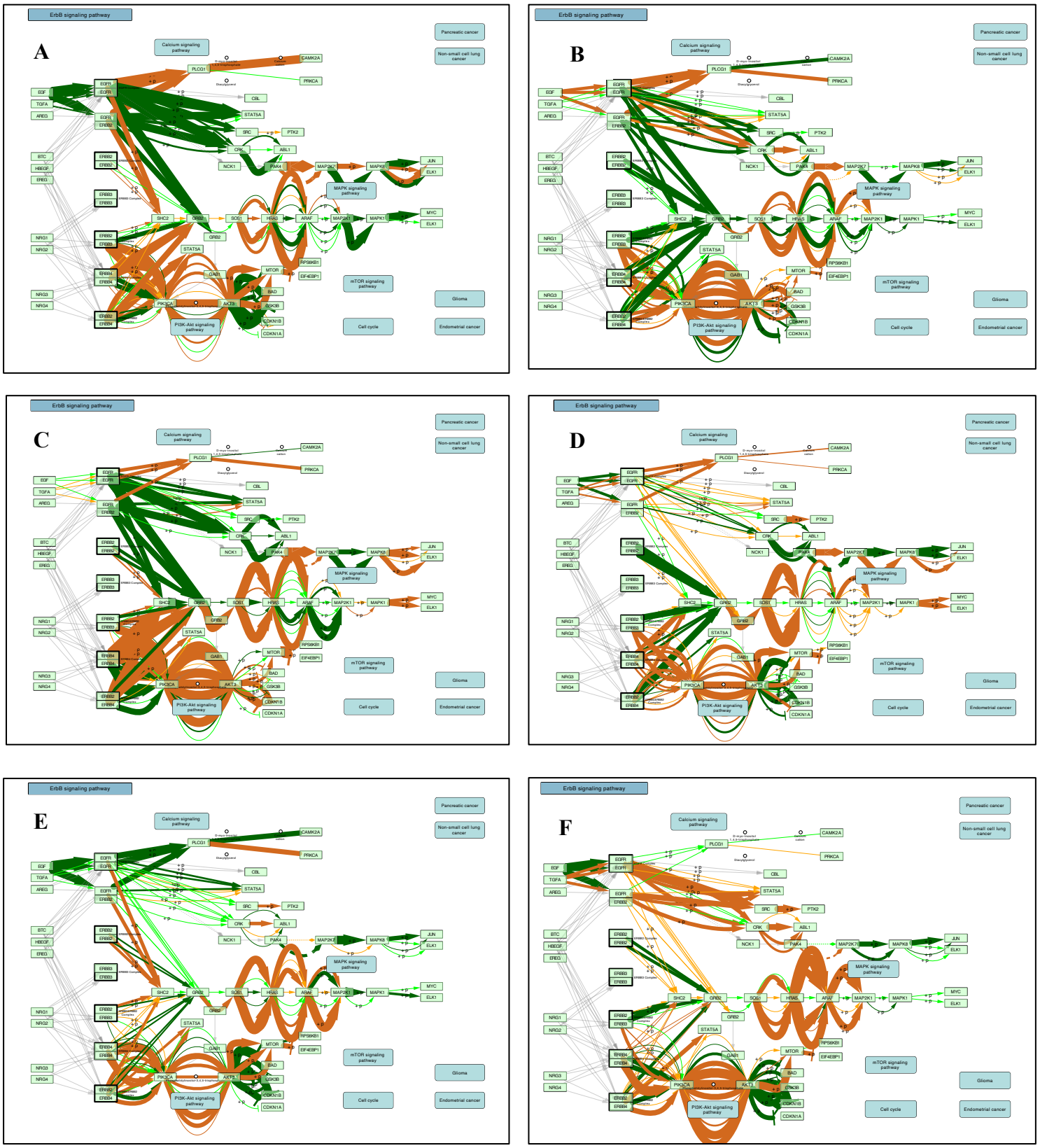


Figure 49: Edges in the ErbB signaling pathway for MCF7 compared to A375 (A), A549 (B), HA1E (C), HEPG2 (D), HT29 (E) and PC3 (F).

In **Figure 49** we can see that, depending on the cell line it is compared to, we see different patterns of differences in concordance. For example, many edges that are part of the sub-network "Calcium signaling pathway" are more concordant in HT29 vs MCF7, but in MCF7 these edges tend to be either more concordant or approximately as concordant in other cell lines. MCF7 also appears to be more concordant across paralogous relationships between PIK3 and AKT to a different degree depending on the comparison cell line. The differences between individual cell lines are descriptive, but the utility of pair-wise differences likely not as useful as observing differences across cell lines.

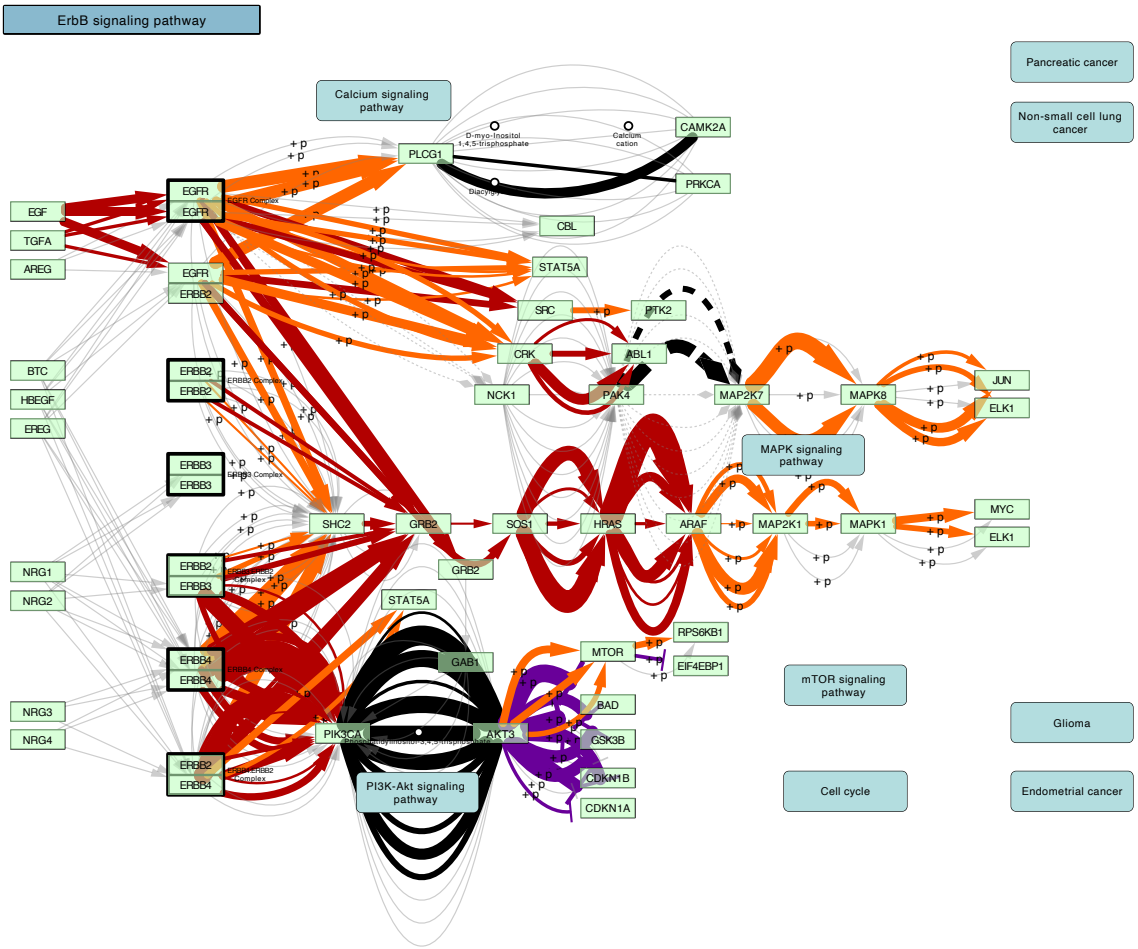


Figure 50: Edges in the ErbB signaling pathway conditionally formatted to represent the Breslow-Day statistic.

The goal of undertaking this project was to assert the notion that in a given pathway analysis, it is of utmost importance to take into consideration the type of cell line being considered for any given analysis and that any given pathway analysis may be underpowered to find meaningful results given the heterogeneity of relationships between paralogues in a given pathway. In **Figure 50**, edges in the ErbB are conditionally formatted to reflect the degree of difference in association across cell lines. We can see that there are some edge paralogues, for example between HRAS and ARAF, that are very similar across cell lines (thin line) but that many are very different between cell lines – the edge KRAS-BRAF which had the largest single edge BD value is a paralogue of this edge. In the chapter that follows, the functionality of KEGGlincs will be demonstrated as it applies to a formal analysis.

Chapter 6: Edge Set Enrichment Analysis (ESEA) of L1000 data set

6.1 GSEA

6.1.1 GSEA Introduction

Gene set enrichment analysis (GSEA) was briefly discussed in **Section 4.2.1** as an example of a second-generation or FCS (functional class scoring) approach to pathway analysis. As opposed to first-generation or ORA (over-representation analysis) approaches, the input for GSEA and other FCS methods includes gene-specific measurements – most often signal-to-noise ratio – for each gene in a pathway or list of genes of interest. It is one of the most, if not *the* most, commonly used methods for the downstream analysis of gene expression data; there have been over 20,000 citations for the methods paper since its publication in 2005 [10]. In the context of a 2-group gene expression comparison (for example mutant vs. wildtype, cases vs. controls) where the first group has the phenotype of interest (typically the "mutant" or "cases" group) the signal-to-noise ratio (S2N) for a single gene *i* is defined as follows:

$$S2N_i = \frac{\mu_i^{Group1} - \mu_i^{Group2}}{\sigma_i^{Group1} + \sigma_i^{Group2}}$$

S2N reflects the correlation (*association*) of a gene with a phenotype in terms of size and direction. In this type of comparison, a gene with a positive S2N would be correlated with Group 1 – that is – it tends to be expressed at a higher frequency in the first group as opposed to the second group. A gene with a negative S2N would be correlated with Group 2 or negatively correlated with Group 1 – the gene tends to be expressed at a higher frequency in the second group relative to the first group.

In GSEA, the focus is neither on the individual S2N measurements nor their significance as determined by traditional differential expression analysis (though p-values are an alternate

gene ranking mechanism). Instead, the goal of GSEA is to find coordinated patterns of gene expression for transcripts in an *a priori* defined sets of genes. The choice of a gene set, or more commonly gene sets, employed in GSEA depends on the research objective. In some cases, researchers want to see if a list of genes found to be differentially expressed in a previous gene expression analysis are *enriched* in their phenotype of interest. However, a more common approach is to see how their data relates to a large number of gene sets after performing GSEA with each set and then ranking them in terms of direction of enrichment and significance (the calculation of these values will be summarized shortly). The Molecular Signatures Database (*MSigDB*) maintained by the BROAD institute – the academic institution that pioneered GSEA – contains nine different collections of gene sets, defined as follows [56]:
H (hallmark gene sets), C1 (positional gene sets), C2 (curated gene sets), C3 (regulatory target gene sets), C4 (computational gene sets), C5 (ontology gene sets), C6 (oncogenic signature gene sets), C7 (immunologic signature gene sets), and C8 (cell type signature gene sets).

The hallmark gene sets are 50 gene sets that are, perhaps, the most relevant in terms of traditional gene expression analysis because the expression of the genes in those sets is related to a given biological condition. Gene sets curated from KEGG pathways are part of the CP (canonical pathways) subset of the C2 gene set (C2:CP). Note that there are nearly 3000 pathways (2868 to be exact) contained in C2:CP. While researchers are not limited in a strict sense by the number of gene sets they wish to run through in GSEA, increasing the number of sets will have an impact on the interpretation of the results both in terms of adjusting the significance for multiple hypothesis testing. Therefore, it may be in the best interest of the researcher to restrict the scope of considered gene sets before conducting GSEA and ensure that the lists are the most relevant to the research objective.

6.1.2 GSEA Methods: Enrichment Statistic

Although the per-gene measurements are not restricted to bi-directional measurements such as S2N, we will restrict our discussion to input of this type as it is the most straight-forward approach and makes the eventual output more interpretable from a directionality standpoint. That being said, let $L = (L_1, L_2, \ldots, L_N)$ be a list of all genes ranked by their S2N and let $S = (S_1, S_2, \ldots, S_k)$ be a list of genes. "Given an *a priori* defined set of genes *S* ... the goal of GSEA is to determine whether the members of *S* are randomly distributed throughout *L* or primarily found at the top or bottom" [10]. Formally,

H_0: Membership in $S \nrightarrow$ Location in L

H_A: Membership in $S \rightarrow$ Location in L (top or bottom)

The per-gene set output after conducting GSEA, namely the enrichment statistic (ES), normalized ES (NES) and its false discovery rate (FDR) adjusted p-value (or "q-value"), will be the basis for rejecting or failing to reject the null hypothesis *for each* gene set. In this section we will demonstrate how the ES is calculated. The value of the ES for gene set *S* is the maximum value of the running enrichment scores (RES), whose calculation is described below.

Let *L* and *S* be a ranked list of genes and a gene list as described above and let $m = (m_1, m_2, \ldots, m_N)$ be the vector of corresponding measurements (i.e. S2N) for all genes in *L*. Each gene is marked with two indicator variables, "Tag" and "No.Tag" whereby $Tag_i = \begin{cases} 1, & L_i \in S \\ 0, & else \end{cases}$ and $No.Tag_i = \begin{cases} 1, & L_i \notin S \\ 0, & else \end{cases}$. Then, define the following quantities:

$M = \sum_{i=1}^{k} |m_i|$ where $Tag_i = 1$ and $T = \sum_{i=1}^{\mathrm{N}} No.Tag_i$. M is the sum of the ranking metric for all genes that are members of *S* and *T* is equal to *N* – k, that is, the number of genes that are in *L* but not in *S.* The RES starts at zero and then, at the first gene will either increase if that gene is in *S* or decrease if that gene is not in that list.

At any given gene *j*,

$$RES_j = \sum_{i=1}^{j}\left(|m_j| * {}^{1}/_{M}\right) * Tag_j\,) - \sum_{i=1}^{j}(({}^{1}/_{\mathrm{T}}) * No.Tag_j) \quad (23)$$

At the final gene *N*, $RES_N = {}^{M}/_{M} - {}^{\mathrm{T}}/_{\mathrm{T}} = 0$, thus RES begins and ends at zero. Let *R* be the vector of all RES. Then, the ES is maximum deviation from zero encountered in the 'random walk' across the ranked list of genes:

$$ES = \max|R| * sign(\max|R|) \quad (24)$$

This non-parametric statistic is defined as a weighted Kolmogorov-Smirnoff-like statistic and is restricted to values between -1 and +1 [10]. Recall the group definitions whereby Group 1 has the phenotype of interest (cases, etc.) and Group 2 is the control group. If $ES > 0$, then the genes in *S* are said to be positively correlated with cases and, as a group, tend to be overexpressed or *enriched* in the cases versus the controls. If $ES < 0$, then the genes in *S* are said to be negatively correlated with cases and, as a group, tend to be under-expressed in the cases versus the controls. Note that the interpretation switches if the groups are defined in the opposite manner (i.e. controls are Group 1 and cases are Group 2).

Oftentimes, gene set results are accompanied by an enrichment plot, which plots the RES across *L* for a given gene set as is shown in **Figure 51**.

The exact visualizations included with an enrichment plot will depend on the choice of software used to conduct GSEA; the plot in **Figure 51** was generated using the JAVA package GSEA-P, a software tool provided to researchers by the BROAD institute. There are also R Bionconductor packages that conduct GSEA such as fgsea [57] and phenoTest [58] that provide similar visualizations. The dashed lines in **Figure 51** were added in to show how the leading-edge (LE) subset of genes is defined. If we define L_{ES} as the gene corresponding to the

maximum RES (represented by the x-axis in **Figure 51**), then the genes up to and including L_{ES} that are part of S are members of the LE subset. The LE subset of genes contributes to the magnitude of the ES and can be thought of as the drivers of the biological process for the phenotype being considered.

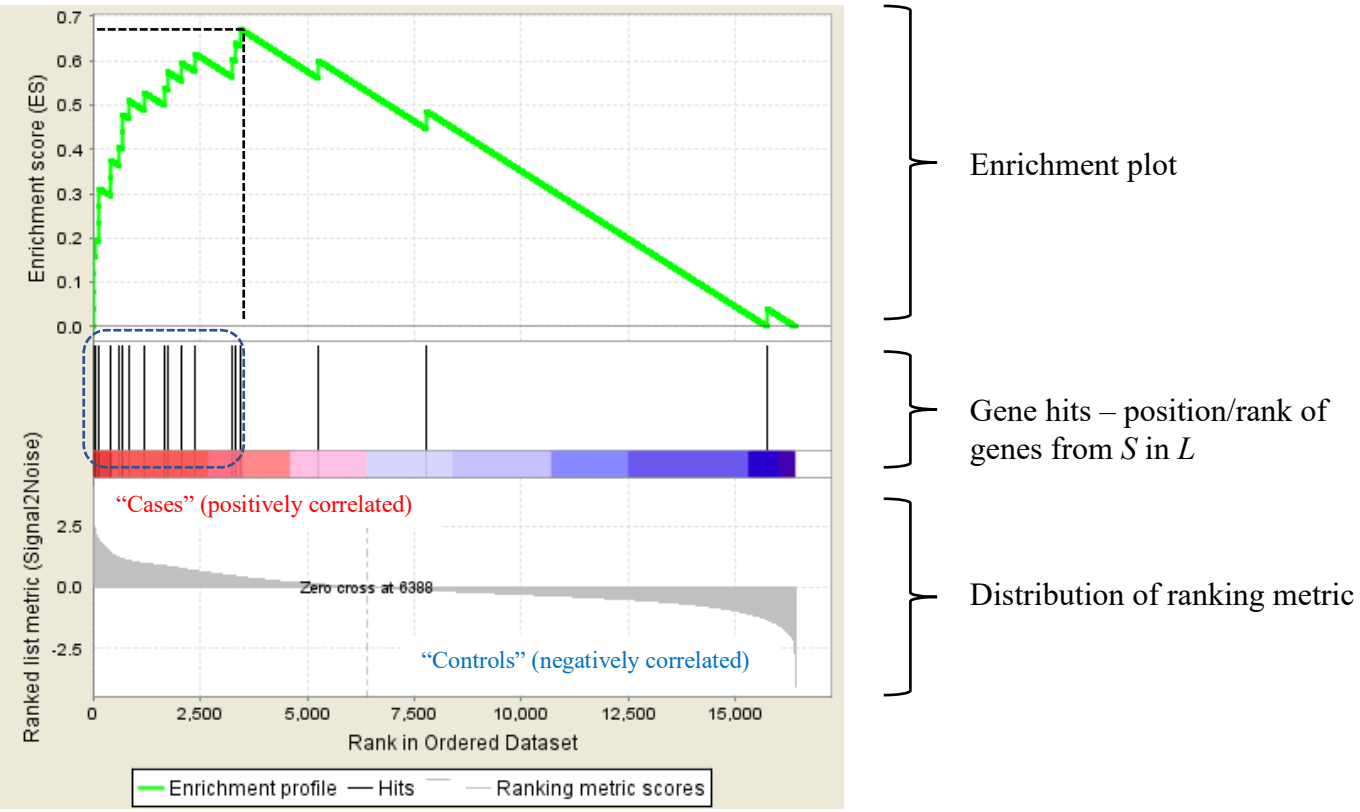


Figure 51: Example of an enrichment plot for gene set with a positive ES (ES = 0.77). The dashed line (horizontal) marks the maximum ES on the y-axis. The gene hits that fall before the vertical dashed line (circled) are part of the leading-edge subset for this gene set.

6.1.3 GSEA Methods: Empirical p-value, Normalized ES and FDR determination

The Kolmogorov-Smirnoff-like ES values require permutation tests in order to estimate an empirical p-value. There are two ways of permuting the data when the input is gene expression data. The first is by permuting or 'shuffling' sample labels; however, since our

eventual analysis of the L1000 data set is not amenable to this kind of permutation we will focus our attention on the second kind of permutation. This kind of permutation involves selecting gene sets that are the same size as *S* and calculating an ES for each permutation. Researchers are free to choose how many permutations they wish to consider, but typically 1000 is considered the standard. Let ***ESP*** be the vector of ES values obtained via permutation of genes for gene set *S*, let ES_{obs} be the ES associated with the original gene set, and let *nperm* be the number of permutations. Then, the empirical p-value is calculated as follows:

$$p = \frac{\sum_{i=1}^{nperm} ESP_i,\ ESP_i \geq\ ES_{obs}}{nperm + 1} \tag{25}$$

Figure 52 is included to show the bi-modal nature of the permuted ES values as well as how there may be a slight bias for either positive or negative ES values (there is a slight positive bias in this example).

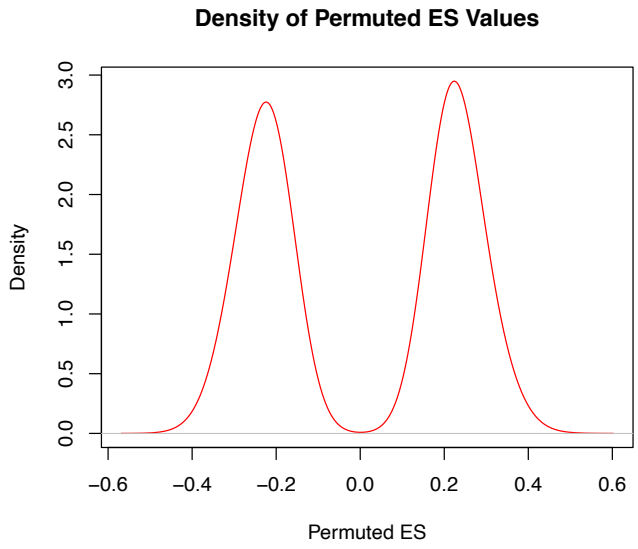


Figure 52: Example of a distribution of permuted enrichment scores

The bias in permuted ES values is taken into account when calculating the normalized enrichment score or NES for a gene set. The NES takes into account gene set size and makes

gene sets comparable to one another in the interpretation of the results. Let $\boldsymbol{ESP}^{+}$ be the positive ES values in $\boldsymbol{ESP}$, $\boldsymbol{ESP}^{-}$ be the negative ES values, and let *npos* and *nneg* be the number of positive and negative ESs respectively. Then, the NES for a given gene set is calculated as follows:

$$NES = \begin{cases} ES > 0, \dfrac{ES + \sum_{i=1}^{npos} ESP_i^{+}}{npos + 1} \\ ES < 0, \dfrac{ES + \sum_{i=1}^{nneg} ESP_i^{-}}{nneg + 1} \end{cases} \tag{26}$$

With the empirical p-values and NES values in place, the last component of the GSEA analysis is to correct for multiple hypothesis testing.

If a researcher is only interested in estimating the significance of the enrichment for one gene set, there is no need to correct for multiple hypothesis testing – sometimes referred to as multiplicity – and the empirical p-value will suffice. However, as is often the case with bioinformatics-based analyses, multiple hypotheses are tested with the same data set. The Bonferroni correction is the traditional approach to the problem of multiplicity and aims to control the familywise-error rate (FWER) across a 'family' (set) of hypotheses. The involves adjusting the significance criteria in order to maintain the Type 1 error rate, α, namely by dividing this critical value by the number of hypotheses being tested.

When the Bonferroni correction is applied, the probability of falsely rejecting *any* (one or more) of the null hypotheses is maintained at the original critical value. However, this is a very conservative approach and, depending on the research objective, is even considered to be irrelevant and counterproductive as it drastically reduces the power of a study [59]. With this in mind, Benjamini and Hochberg introduced the concept of the false discovery rate (FDR) as an alternative approach for handling multiplicity. The control of FDR as opposed to Type I error increases power and is a suggested as a multiple comparison procedure (MCP) that balances the

number of true and false positives in gene expression and other genomewide studies [60]. Benjamini and Hochberg borrow the nomenclature "discovery" from Sorić, who used the term to define hypotheses that are rejected based on making some significance threshold regardless of their nature (rejected erroneously or non-erroneously) [61]. **Table 27**, adapted from their paper describing their approach [62], allows for a straight-forward comparison of how FDR operates relative to the Bonferroni correction.

Table 27: Number of true and false rejections for m different hypotheses tested in same dataset.

	Not Rejected	Rejected	Total
True null hypotheses ("Should not" be rejected)	U	V	m_0
Non-true null hypotheses ("Should" be rejected)	T	S	$m - m_0$
	$m - R$	R	m

In this table, m is the total number of hypotheses being tested and of those tested R "discoveries" are made; S of the R are correct rejections whereas V of the R are falsely rejected. The per-comparison error rate (PCER) is the expected number of false rejections across all hypothesis tests, $E(V/m)$, and the FWER is the chance of committing at least one type one error or $P(V \geq 1)$. When the significance level is set at α and there is no correction for multiple testing, $E(V/m) \leq \alpha$. When the Bonferroni correction is applied, it ensures that the probability of making one or more false rejection is less than or equal to alpha: $P(V \geq 1) \leq \alpha$. The FDR, on the other hand, is the expectation of Q, which is equal to $V/(V + S)$, or simply V/R.

Once the NES and empirical p-values have been calculated for each gene set, the sets are separated into those with positive and negative NES values before they are assigned FDR q-values. For gene set *S*, the q-value represents the chance that *S* is a false positive finding given the distribution of empirical p-values for the other gene sets being tested at the same time. As

with alpha levels, the threshold chosen for FDR q-values in a given study is somewhat arbitrary but should help researchers filter their results in a meaningful way. For example, the authors of the seminal GSEA paper chose FDR ≤ 0.25 for their analyses but in the upcoming discussion of ESEA, those authors chose FDR ≤ 0.05 as FDR ≤ 0.25 would not have effectively filtered their results.

6.2 ESEA Overview

The 'raw' starting input for the original incarnation of edge set enrichment analysis, ESEA, is exactly the same as GSEA – the expression levels for all genes measured in an experiment across all samples in both groups (typically cases and controls). However, instead of using this data to make a background list (L) that ranks each gene by its normalized difference between cases and controls, the intermediary step of ESEA is to assign a measurement of differential correlation between cases and controls to pairs of genes. The background set is not all possible pairs of genes but rather pairs of genes with a relationship documented in one or more of seven databases (KEGG, Reactome, Biocarta, NCI, SPIKE, HumanCyc, Panther). Importantly, the databases contain pathways with genes as nodes and biological relationships as edges between genes.

The authors of the ESEA paper, titled "ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis", state that the goal of this approach is to "quantify the change of correlation between genes for each edge" in order to "identify dysregulated pathways associated with a specific phenotype by investigating the changes of biological relationships of pathways in the context of gene expression data" [63]. Although the objective is somewhat different than that of GSEA, the methodological background for ESEA is nearly identical with exception of the input for L and S. As opposed to assigning single-gene-based measurements

such as S2N to each gene *i*, an *EdgeScore* for each pair of genes, $gene_i$ and $gene_j$ or simply $EdgeScore_{ij}$, is the difference in correlation between cases and controls and is calculated as follows:

$$EdgeScore_{ij} = MI_{all}[i;j] - MI_{control}[i;j] \tag{27}$$

Here, MI stands for mutual information. MI is an information theoretic measurement, which is to say that it comes from a branch of applied mathematics concerned with the quantitative study of information, namely as it applies to communication [64]. The 'information' is the entropy of a system, or rather the balance of entropy and redundancy (non-random behavior) within that system [65]. Although this information theoretic approach to statistical systems had been introduced and described in the mid-20th century, it was not until the end of the 20th century that the technology to generate large amounts of biological data made the approach amenable to the study of systems biology. Then, the use of MI started to receive attention for its applications in reverse engineering gene networks from gene expression data [66] [67]. The measure of entropy in this context is, specifically, called Shannon's entropy and for gene *A* with *n* different expression patterns represented by *x*, the value is calculated as [65]:

$$H(A) = -\sum_{i=1}^{n} p(x_i) log_2(p(x_i)) \tag{28}$$

When $H(A)$ is large for a gene, that gene has a more random distribution of its expression values; i.e. they are more difficult to predict [68]. Note that this is technically the form for discretized data; integration is performed as opposed to summation in the continuous case [69]. Then, for two different sources of information, for example genes *A* and *B,* their mutual information is calculated as:

$$MI(A,B) = H(A) + H(B) - H(A,B) \tag{29}$$

where $H(A,B)$ is the joint entropy of genes *A* and *B*. If the joint entropy is zero, then the information on the behavior from one gene can perfectly predict the behavior of the other and vice versa; there is no randomness when they are considered together. MI will be large if one or both marginal entropies is large and the joint entropy is low. When MI is zero, this means that the entropy of the two genes considered at the same time is the same as the entropy when each gene is considered separately – they function completely independently of one another. MI is a correlation-type measurement with two key differences between itself and the traditional Pearson's correlation coefficient; it is always non-negative (positive and negative relationships of the same magnitude reduce entropy by the same amount), and it can detect strength in non-linear relationships.

Recall equation 27 where the $EdgeScore_{ij}$ for $gene_i$ and $gene_j$ is equal to the MI of all samples considered together minus the MI calculated when only the control samples are part of the MI equation. $EdgeScore_{ij} > 0$ implies that MI between $gene_i$ and $gene_j$ increases when the cases are added to the control samples whereas $EdgeScore_{ij} < 0$ suggests the opposite. After $EdgeScore_{ij}$ has been calculated for all edges in the background set *L*, ESEA proceeds in a manner identical to GSEA for each pathway with edges in *S*; the only difference is that instead of matching single genes in *S* to *L* for the calculation of an enrichment score ES, edges from pathways are matched to pairs of genes in *L* to calculate an edge enrichment score EES. *EES(P)* is used to denote the EES for pathway *P*. If *EES(P)* > 0, then the edges in pathway *P* collectively exhibit gains in MI when cases are taken into consideration along with the controls and the pathway is labelled as a gain in correlation (GoC) pathway. When *EES(P)* < 0, the opposite is true (MI collectively decreases) and the pathway is described as a loss of correlation (LoC)

pathway and in the event that *EES(P)* = 0 then pathway *P* is a no change (NC) pathway [63]. The authors also describe a core set of edges for pathway *P* that we will call the leading edge edge (LEE) subset as they consists of the edges in *P* that are part of the leading edge subset – they are the members of *P* that contribute to the magnitude of *EES(P)* and are likely most associated with the biological process of interest.

Once the empirical *EES(P)* is calculated for pathway *P*, the calculation of the empirical p-value associated with that EES proceeds in a manner analogous to the calculation of the empirical p-value for the ES in GSEA. In their paper, the authors describe a gene-based permutation test procedure that shuffles gene labels and recomputes *EES(P)* however the algorithm they employ to obtain their published results is actually an edge-based permutation procedure whereby an EES is calculated for a random list of edges (as many edges as are in pathway *P*). Normalized edge enrichment scores (NEES) are then calculated based on the permuted edge set just as NES are calculated from the permuted gene sets in GSEA to allow for inter-pathway comparisons. Finally, the same FDR correction is applied to a set of pathways with their corresponding empirical p-values to account for multiple testing. Before interpreting the results, the pathways are separated into those with positive and negative NEES and then ranked by FDR q-value. Plots of the running edge enrichment score may be produced and are identical in concept to the plot shown in **Figure 51**.

The authors give three examples of how ESEA may be applied to different gene expression datasets and compare the results to those obtained with GSEA. The most relevant with regard to our eventual analysis, which we will review to show the proof-of-concept, comes from a dataset that spans 50 cell lines classified by their p53 mutation status (17 native, 33 mutated) from the collection of NCI-60 (National Cancer Institute) panel of 60 cancer cell lines;

the rationale for the use of this subset can be found in [70]. The pathways chosen for this example are all KEGG pathways with more than 15 but less than 1000 edges; 187 pathways altogether (note: in this analysis and in the ESEA R package, the pathways are from the databases as they existed in 2015 [71]). With an FDR q-value cutoff of 0.05, five KEGG pathways are deemed significantly enriched, all in the positive (gain of correlation) direction. Note that this FDR cutoff is much more stringent than the FDR cutoff of 0.25 suggested in the GSEA paper; although the ESEA authors do not explain their choice of this threshold, an FDR of 0.25 in this first example would result in 30 pathways deemed significantly enriched. The five significant ESEA pathways are Cysteine and methionine metabolism, Alcoholism, Dilated cardiomyopathy, ECM-receptor interaction, and Colorectal cancer, all of which have at least one example in the literature of being related to p53.

Although the authors do not elaborate on the results for each pathway, they do spend time discussing the results of the colorectal pathway and map the members of the LEE subset to edges in the pathway as it appears in KEGG. As with many KEGG pathways, the Colon cancer pathway is an amalgamation of many pathways, including the Wnt, PI3K/AKT, MAPK, and TGFβ signaling pathways. Mapping the LEE subset to the KEGG pathway helps identify key GoC relationships in the PI3K/AKT sub-network that may explain why a mutation in p53 results in an overall GoC for the colon cancer pathway. When the parallel GSEA analysis is performed on the same data set with the same pathways, only the N-glycan synthesis pathway is significant at the GSEA default FDR threshold of 0.25.

When the same parallel analysis of GSEA and ESEA on the cell line data was performed with 157 Biocarta pathways as opposed to KEGG pathways with the same cutoffs applied as before, ESEA produced one significant pathway (CDK regulation of DNA replication) whereas

GSEA led to three significant pathways (hypoxia and p53 in the cardiovascular system, BCR signaling pathway and nerve growth factor pathway). The authors do not discount the results GSEA but, rather, argue that the ESEA approach identifies a "new" pathway that shows promise for further research.

The purpose of giving a detailed summary of this example is to support the notion that ESEA successfully applies existing node-based methodological architecture to an edge-based approach to data analysis which, at least in the context of gene expression data, can support hypothesis generation and bolster previous research finding. It is also to import the idea that the results are heavily dependent on the database of choice and that the biological interpretation may require a closer inspection of pathway-level results. In the p53 mutation status example, all of the significant pathways have an overall *gain* in correlation in the mutant samples but are still described as 'dysregulated'; at least in the colon cancer example, the authors are able to identify key relationships to explain why an increase in correlation for a few key relationships may provoke differential regulation in the positive direction across the pathway.

6.3 ESEA of LINCS Dataset

6.3.1 Review of Concordance Measurements

The research question behind GSEA is whether the expression of groups of genes is correlated with a with a phenotype of interest; in ESEA it is whether the edges in pathways are differently regulated in cases versus controls. Our research question when we apply ESEA to the LINCS data is, specifically, "are the edges in a given KEGG pathway more concordant [or discordant] in one cell line vs other cell lines?". Before moving on with the methods for this analysis, we will spend some time covering what it means to be concordant or discordant in the context of cell line comparisons.

Recall equation 5, where delta between the reference cell line A and comparison cell line B for edge (gene knock downs) X|Y is the standardized difference between the log of their odd's ratios, $\widehat{\Delta}_{XY}^{AB} = \frac{log(\widehat{\theta}_{XY}^{A}) - log(\widehat{\theta}_{XY}^{B})}{\sqrt{(SE_{XY}^{A})^2 + (SE_{XY}^{B})^2}}$. Briefly, let N_{XY}^{AB} be the numerator of equation 5 and D_{XY}^{AB} be the denominator such that $N_{XY}^{AB} = -1 * N_{XY}^{BA}$ and $D_{XY}^{AB} = D_{XY}^{BA}$. The denominator is a function of the number of L1000 genes (out of 200 up and 200 down, as decided based on the power analysis) that fall into each cell of the odds ratio table for the two different cell lines and will be small when there are many genes and large if there are few genes for one or both cell lines. Although D_{XY}^{AB} will have an impact on the magnitude of $\widehat{\Delta}_{XY}^{AB}$, it does not have any influence on its direction. For an individual cell line, say cell line A, the term $log(\widehat{\theta}_{XY}^{A})$ will be:

- Positive when more genes fall into concordant cells than discordant cells (i.e. L1000 genes are up or down for both gene knockdowns X and Y).
- Negative when more genes fall into discordant cells than concordant cells (i.e. L1000 genes tend to be up for knockdown X but down for knockdown Y or vice versa).
- Approaches zero when the number of genes in discordant cells is equal to the number of genes in concordant cells (*no association*).

With this in mind, $\widehat{\Delta}_{XY}^{AB}$ will be positive (*more concordant* or *less discordant* for cell line A relative to cell line B) for the following conditions:

1) $log(\widehat{\theta}_{XY}^{A})$ is positive (concordant) and $log(\widehat{\theta}_{XY}^{B})$ is negative (discordant).
2) $log(\widehat{\theta}_{XY}^{A})$ is positive (concordant) and $log(\widehat{\theta}_{XY}^{B})$ is positive but $log(\widehat{\theta}_{XY}^{A}) > log(\widehat{\theta}_{XY}^{B})$.
3) $log(\widehat{\theta}_{XY}^{A})$ is negative (discordant) and $log(\widehat{\theta}_{XY}^{B})$ is negative but $log(\widehat{\theta}_{XY}^{A}) > log(\widehat{\theta}_{XY}^{B})$.
4) $log(\widehat{\theta}_{XY}^{A})$ is positive (concordant) and $log(\widehat{\theta}_{XY}^{B})$ is close to zero.

Notice that in condition (3), the individual ORs are both in the discordant direction, but the OR for cell line A is less discordant and, hence, more concordant when compared to cell line B. For completeness, $\widehat{\Delta}_{XY}^{AB}$ will be negative (*more discordant* or *less concordant* for cell line A relative to cell line B) for the following conditions:

1) $log\left(\hat{\theta}_{XY}^{A}\right)$ is negative (discordant) and $log\left(\hat{\theta}_{XY}^{B}\right)$ is positive (concordant).
2) $log\left(\hat{\theta}_{XY}^{A}\right)$ is negative (discordant) and $log\left(\hat{\theta}_{XY}^{B}\right)$ is negative but $log\left(\hat{\theta}_{XY}^{A}\right) < log\left(\hat{\theta}_{XY}^{B}\right)$.
3) $log\left(\hat{\theta}_{XY}^{A}\right)$ is positive (concordant) and $log\left(\hat{\theta}_{XY}^{B}\right)$ is positive but $log\left(\hat{\theta}_{XY}^{A}\right) < log\left(\hat{\theta}_{XY}^{B}\right)$.
4) $log\left(\hat{\theta}_{XY}^{A}\right)$ is negative (discordant) and $log\left(\hat{\theta}_{XY}^{B}\right)$ is close to zero.

Delta measurements will be the input values for L1000 ESEA with KEGG pathways. Therefore, when we make claims about pathways being more concordant or discordant in one cell line vs another (or others), we are saying that deltas tend to be positive or negative with regard to the reference cell line, but individual relationships may fall into one of the many categories listed above.

6.3.2 Database and Input for ESEA with LINCS

The underlying mathematical and statistical methodology we use to calculate edge enrichment scores and determine statistical significance for ESEA with LINCS data is identical to the methods detailed in GSEA and ESEA – the main difference is the nature of the input used to rank edges. Also, instead of using the pathway database 'snapshot' provided by the ESEA R package from 2015, we will be using current (as of January 2021) KEGG pathway infrastructure to build our background edge set and specifically testing non-metabolic pathways with more than 5 but less than 1000 edges for concordance enrichment (162 pathways). We have already gone through our dataset and calculated pairwise (*local*, as described below) concordance-based measurements for all of the 5,604 edges in KEGG pathways that are 'in' our dataset (both nodes

[genes] have corresponding shRNA gene knockdown profiles in all 7 cell lines). In order to account for potential bias among certain cell line comparisons, we will use the matrix in **Figure 23A** to adjust the delta values as follows. Let ***A*** be the matrix in **Figure 23A** and let a_{AB} be the value corresponding to the entry for reference cell line A and comparison cell line B. Then, the adjusted delta for any edge X|Y is:

$$\tilde{\Delta}_{XY}^{AB} = \hat{\Delta}_{XY}^{AB} - a_{AB} \tag{30}$$

The experimental setup for the GSEA approach is to make pairwise comparisons, therefor in this analysis we will perform all pairwise comparisons between all seven cell lines. However, in a manner similar to the example given for p53 status among cell lines for ESEA, we will also consolidate information across cell lines in order organize the results of this analysis using a global complement to a local approach as follows:

1. Local: pairwise comparisons between all cell lines (21 unique comparison each with two different directional arrangements; i.e. A vs B and B vs A).

   Ranking metric = pairwise adjusted delta values = $\tilde{\Delta}_{XY}^{AB}$

2. Global: reference cell line vs all other cell lines (7 unique comparisons).

   Ranking metric = average adjusted delta values = $\bar{\Delta}_{XY}^{A} = \frac{1}{6}\sum_{B=1}^{6} \tilde{\Delta}_{XY}^{AB}$ (31)

   where A = (reference cell line) and B = (cell lines ≠ A)

The global comparisons will allow us to make broad claims about pathway level results which can then be interrogated at the local level to see if the differences hold across cell lines or if they are driven by specific cell lines. To be clear, let $EES(P^{AB})$ be the edge enrichment score for pathway *P* when A is the reference cell line and B is the comparison cell line for a *local comparison* and $EES(P^{A})$ be the EES when A is the reference cell line for a *global comparison*. For local comparisons, when $EES(P^{AB}) > 0$, pathway *P* is enriched with concordant edges in

cell line A relative to cell line B and when $EES(P^{AB}) < 0$ the edges in pathway P are discordant in cell line A relative to cell line B. In a global comparison, if $EES(P^{A}) > 0$ then the edges in pathway P are more concordant on average than in other cell lines whereas $EES(P^{A}) < 0$ implies more discordant edges on average for cell line A compared to other cell lines.

6.3.3 Results

The presentation of our results will proceed as follows. A summary of the global and local results will be given to describe the overall breakdown of results with regard to significance as well as direction (concordant or discordant). We will then showcase results from the MCF7 breast cancer cell line to demonstrate how this approach might be used to support existing research findings and support hypothesis generation.

6.3.3.1 Summary of Results by FDR q-value and Direction

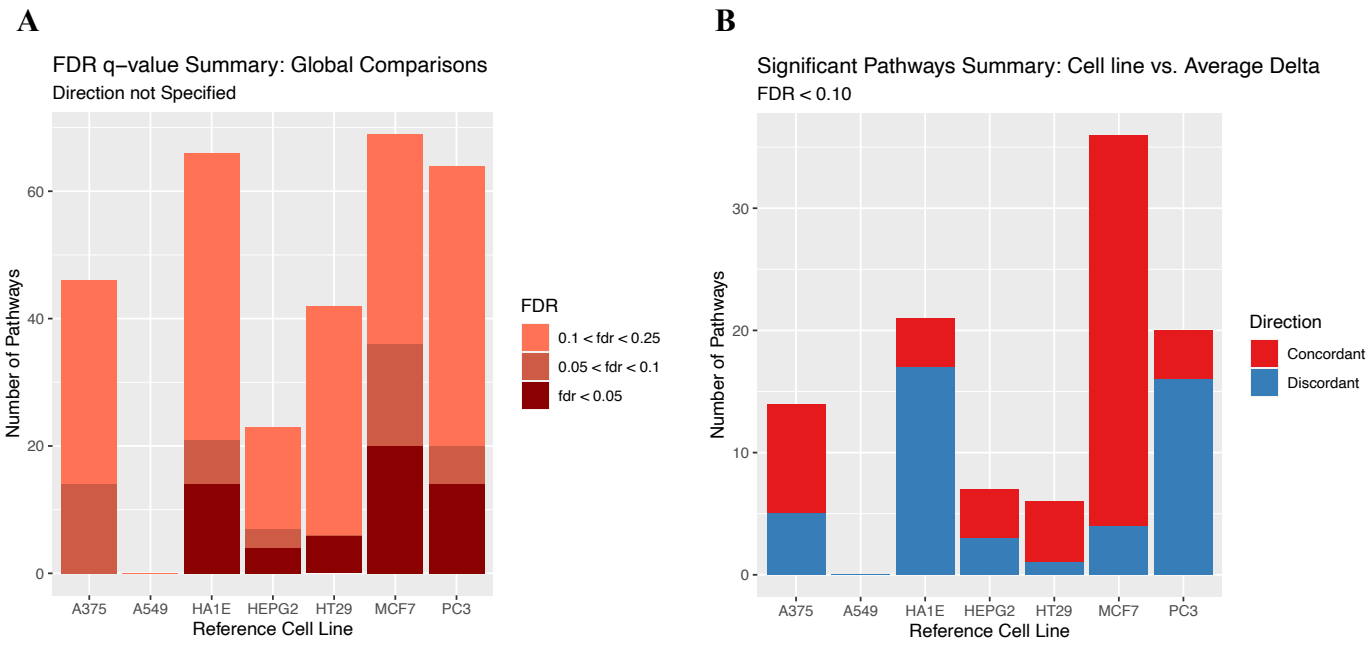


Figure 53: Breakdown of ESEA results by FDR q-value (A) and direction of results (B).

The global results across 162 are broken down by FDR q-value in **Figure 53A**. The bins for FDR q-values were chosen to demonstrate how power might be affected under the liberal

GSEA-suggested criteria (FDR ≤ 0.25), conservative ESEA-suggested criteria (FDR ≤ 0.05), and a traditional criteria level considered to be more moderate (FDR ≤ 0.10). Regardless of the criteria, the cell line A549 does not have any pathways that are significantly enriched with concordant or discordant edges in this global comparison. When FDR ≤ 0.05, the same is true for the cell line A375. When FDR ≤ 0.10, six of the seven cell lines have pathways that meet criteria for significance and the same is true for FDR ≤ 0.25. By increasing FDR from a moderate to a liberal level, however, three out of the six cell lines have significant results for over one third of the pathways that are tested (HA1E, MCF7 and PC3). **Appendix Figure A7** breaks down the local comparisons by FDR in a manner identical to **Figure 53A** for each cell line acting as a reference and similarly, across the board, FDR ≤ 0.05 yields very few results whereas FDR ≤ 0.25 leads to an unwieldy number of pathways to consider. Therefore, we will present results for pathways with FDR ≤ 0.10 as significant pathways of interest for both the global and local analyses as it provides an effective filtering mechanism for our results while still yielding meaningful outcomes.

With FDR ≤ 0.10 chosen as the thresholding level, **Figure 53B** breaks down the results by the direction of the EES value in the positive (concordant; less discordant) and negative (discordant; less concordant) directions. In the global analysis, MCF7 [breast cancer] has the most enriched pathways (n = 36), followed by HA1E [immortalized kidney] (n = 21), PC3 [prostate cancer] (n = 20), A375 [skin cancer] (n = 14), HEPG2 [liver cancer] (n = 7), HT29 [colon cancer] (n = 6) and finally A549 [lung cancer] (n = 0). In the case of MCF7, most (32/36) of the pathways are enriched in the positive direction whereas in HA1E and PC3 the opposite is true with 17/21 and 16/20 pathways respectively enriched in the negative direction.

The results for the global analysis are presented in **Appendix Tables A4a-A4f** (no global results table for A549) and the local analysis results are in **Appendix Tables A5a-A5g**. The results are ordered by FDR q-value and the magnitude of the NEES as well as by comparison cell line for the local results. ESEA was performed in both directions for the local comparisons, for example with A375 as a reference cell line and A549 as the comparison cell line as well as with A549 as a reference cell line and A375 as the comparison cell line. Although the results are somewhat redundant, having the results laid out this way helps with interpretation and showcases some particular elements of the results that are introduced when permutation procedures are used to calculate the NEES and FDR q-values.

Consider, for example, the result for the pathway Human T-cell leukemia virus 1 infection between cell lines A374 and A549 as shown in **Table 28**. This pathway is the concordant pathway with the lowest FDR q-value for the A549 vs A375 comparison (q-value = 0) and also has a very low, but not exactly equal, q-value for A375 vs A549 (q-value = 0.081). Although the EES has the same magnitude in opposite directions as we would expect (0.324 in A549 vs A375 and -0.324 in A375 vs A549), the q-values as well as the NEES (1.506 and -1.518) and nominal p-values (0, 0.004) are close, but not equivalent in magnitude since they are calculated from different permuted EES distributions.

Table 28: Results of Human T-cell leukemia virus 1 infection between A375 and A549.

Comparison	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A549 vs A375	119	0.324	1.506	0	0	Concordant
A375 vs A549	119	-0.324	-1.518	0.004	0.081	Discordant

The directional breakdowns for local comparisons at FDR $\leq$ 0.10 are included in **Appendix Figure A8**. Note that pathways are considered in this figure and are included in **Appendix Tables A5a-A5g** only if FDR $\leq$ 0.10 in both comparisons; for example, the pathway

Melanoma has FDR = 0.097 in A375 vs A549 but FDR = 0.13 in A549 vs A375 thus it is not included in **Appendix Tables A5a** or **Appendix Tables A5b** nor does it contribute to the counts in **Appendix Figure A8**. The local pathway comparisons are, as we might expect, more powerful than the global comparisons since individual cell line behavior may be more extreme than the average behavior of all cell lines. The only local comparisons with no enriched pathways are between A375 and HT29. On the other hand, A549, which had no significant pathways at the global level, has enriched pathways in both directions across the panel of other cell lines. Otherwise, the local results reflect the global results, for example, MCF7 has the most significant pathways across all cell lines and the majority of them are enriched in the positive direction.

6.3.3.1 MCF7 Results

Table 29: Top 15 pathways for MCF7 global analysis

Rank	Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
1	MCF7	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	0.667	2.134	0	0.000	Concordant
2	MCF7	AVERAGE	Platelet activation	04611	33	0.575	2.009	0	0.000	Concordant
3	MCF7	AVERAGE	Acute myeloid leukemia	05221	112	0.458	1.972	0	0.000	Concordant
4	MCF7	AVERAGE	Chemokine signaling pathway	04062	119	0.437	1.967	0	0.000	Concordant
5	MCF7	AVERAGE	mTOR signaling pathway	04150	106	0.377	1.647	0	0.000	Concordant
6	MCF7	AVERAGE	Neurotrophin signaling pathway	04722	135	0.349	1.588	0	0.000	Concordant
7	MCF7	AVERAGE	Choline metabolism in cancer	05231	50	0.464	1.800	0.001	0.020	Concordant
8	MCF7	AVERAGE	VEGF signaling pathway	04370	66	0.443	1.768	0.001	0.020	Concordant
9	MCF7	AVERAGE	B cell receptor signaling pathway	04662	60	0.410	1.642	0.002	0.032	Concordant
10	MCF7	AVERAGE	Prostate cancer	05215	150	0.330	1.530	0.002	0.032	Concordant
11	MCF7	AVERAGE	Longevity regulating pathway	04211	60	0.427	1.678	0.003	0.040	Concordant
12	MCF7	AVERAGE	Colorectal cancer	05210	106	0.385	1.646	0.004	0.040	Concordant
13	MCF7	AVERAGE	Dopaminergic synapse	04728	54	-0.411	-1.613	0.004	0.040	Discordant
14	MCF7	AVERAGE	C-type lectin receptor signaling pathway	04625	170	0.373	1.546	0.003	0.040	Concordant
15	MCF7	AVERAGE	Estrogen signaling pathway	04915	111	0.353	1.477	0.004	0.040	Concordant

Table 29 lists the results of the 15 most enriched pathways for the MCF7 cell line (breast cancer) global analysis; the full table of results for this global analysis is in **Appendix Table**

A5f. There is a six-way tie for most significantly enriched pathway; all of these pathways have nominal p-values of zero which means that in each case, the permuted EES values were never more extreme than the empirical EES values. While these pathways may have identical FDR q-values, they have their own rank due to unique NEES values. All of the pathways in this subset have positive NEES values and thus their direction is "Concordant", thus we would expect that evidence for their association with breast cancer to be in the positive direction. That is, in fact, what we find in the literature, as is summarized below for each of the pathways with FDR q-values of zero.

1) **Regulation of lipolysis in adipocytes.** A very recent review article titled 'Adipocytes in Breast Cancer, the Thick and the Thin' states that "Numerous studies demonstrated that adipocyte lipolysis stimulated by cancer cells is at the very heart of the synergy between [breast] cancer cells and adipocytes" [72]. Specifically, MCF7 cells co-cultured with adipocytes have been shown to increase the lipolytic rate of those adipocytes which in turn drives cell proliferation and migration [73].
2) **Platelet activation.** A review article states that "platelet activation has been observed for decades in women with breast cancer" [74]. Furthermore, MCF7 cells specifically have been shown to induce platelet activation/aggregation [75] [76].
3) **Acute myeloid leukemia (AML) pathway.** There are quite a few papers have been published that conclude having breast cancer [and receiving therapy] increases a patient's risk for developing AML [77] [78] [79].
4) **Chemokine signaling pathway**. This pathway has also been implicated in the progression of breast cancer and even specifically in MCF7 cells [80] [81] [82].

5) **mTOR signaling pathway**. The mTOR signaling pathway is often found to be upregulated in breast cancer due to either mutations in mTOR or increased activity of upstream receptors or pathways [83]. For this reason, mTOR is often targeted with inhibitors in certain types of breast cancer [84]. Furthermore, mTOR inhibitors have been shown to specifically inhibit cellular grown in the MCF7 cell line [85].

6) **Neurotrophin signaling pathway**. Neurotrophin nerve growth factor (NGF) receptors have been found in breast cancer cells, specifically MCF7, where NGF has been shown to be a mitogenic factor leading to cell growth [86]. Neurotrophin signaling has also been associated with an anti-apoptotic effect in the MCF7 cell line [87].

These are just a few noteworthy examples of this global analysis. Note that the Dopaminergic synapse pathway is only pathway with a negative NEES value in **Table 29** (though three more pathways with FDR ≤ 0.10 are included in **Appendix Table A5f**). While dopamine acts as a neurotransmitter in the brain, it acts as a hormone elsewhere in the body [88]. Studies have shown that the MCF7 cell line in particular does not respond to dopamine agonists, unlike other breast cancer cell lines which may mean that this pathway is less active in this cell line [89].

Table 30: Global and local results for RLA pathway with MCF7 as reference cell line.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	0.667	2.134	0	**0.000**	Concordant
MCF7	A375	Regulation of lipolysis in adipocytes	04923	23	0.495	1.591	0.015	**0.136**	Concordant
MCF7	A549	Regulation of lipolysis in adipocytes	04923	23	0.568	1.894	0.001	**0.016**	Concordant
MCF7	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.598	1.954	0.001	**0.014**	Concordant
MCF7	HEPG2	Regulation of lipolysis in adipocytes	04923	23	0.476	1.562	0.012	**0.100**	Concordant
MCF7	HT29	Regulation of lipolysis in adipocytes	04923	23	0.497	1.657	0.006	**0.122**	Concordant
MCF7	PC3	Regulation of lipolysis in adipocytes	04923	23	0.576	1.830	0.001	**0.008**	Concordant

The results for each pathway can be probed for specific pairwise differences as is demonstrated in **Table 30**. This table breaks down the results for the pathway Regulation of lipolysis in adipocytes (RLA) – the pathway with the largest NEES value when MCF7 is the reference cell line. The RLA pathway is enriched with concordant edges in MCF7 not only in the global comparison, but across the board when compared to every other cell line in this study. Furthermore, 4 of the six pairwise comparisons yield FDR q-values at or below our prespecified threshold and the remaining two are very close to this threshold as well. The same tables with the other six cell lines as the reference cell lines are included in **Appendix Tables A6a-A6f** and they show how this pathway is not significantly enriched globally for any of these cell lines and only one of the non-MCF7 local comparisons reaches the criteria for significance. Taken together, this information suggests that the edges in this pathway are indeed more concordant for MCF7 and that the results are not heavily influenced by the behavior of one cell line.

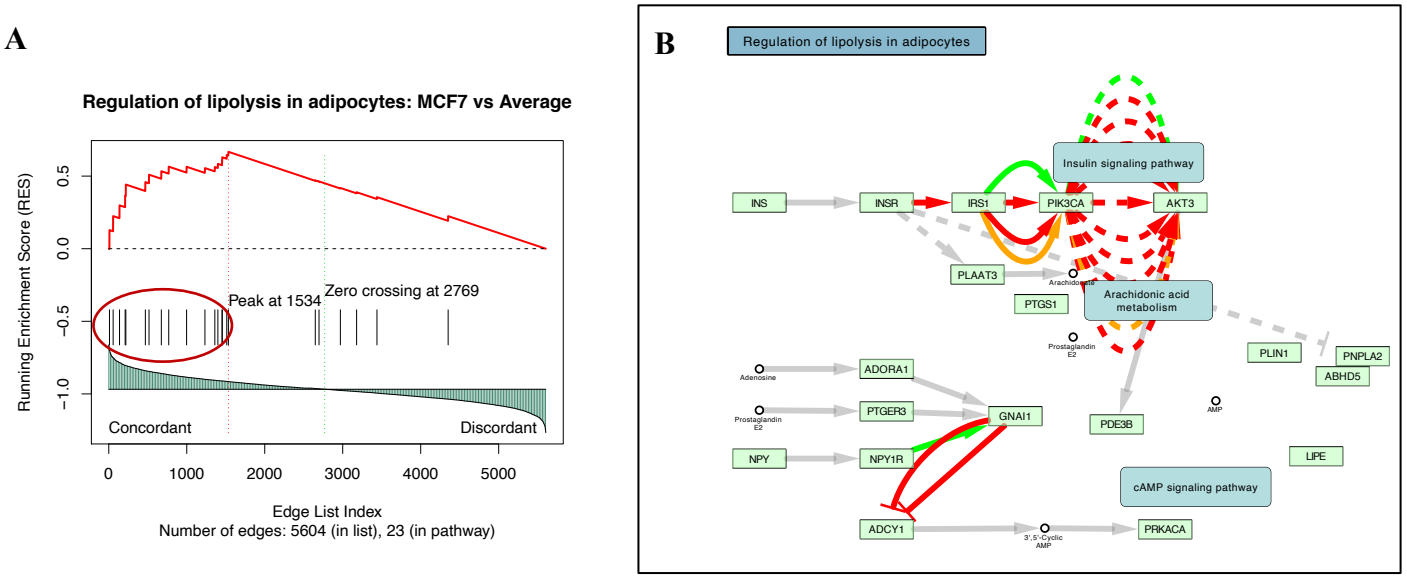


Figure 54: Visualization of results for RLA pathway via enrichment plot (A) and annotated KEGG pathway (B). In (B), the leading edge edges circled in (A) are depicted in red; those that are concordant in MCF7 but not in the LEE subset are orange and those that are discordant in MCF7 are green.

The functionality provided in the KEGGlincs package also helps to interpret results at the pathway level. With ESEA results mapped to KEGG pathways, we can answer questions such as: Are the results being driven by specific parts of the pathway? Is there agreement between paralogs? And, Are certain edges common across significant pathways? In **Figure 54B**, we see the RLA pathway with edges formatted to reflect the concordance patterns among the edges as well as their membership in the LEE subset from **Figure 54A**. This pathway is a relatively small pathway, and we can see that many of the edges that contribute to the large pathway EES are between the twelve combinations of paralogs for PI3K (PIK3CA, PIK3CB, PIK3CD, PIK3R1) and AKT (AKT1, AKT2, AKT3) that have corresponding shRNA knockdowns in the L1000 data.

This same approach is, perhaps, more useful when dealing with more complex pathways. Take, for example, the mTOR pathway with 106 edges. Similar to the RLA pathway, the direction of enrichment is concordant in MCF7 across the panel of cell lines and, in addition to reaching the criteria for significance in the global comparison, achieves FDR ≤ 0.10 in 3 of the local comparisons (A375, HEPG2 and PC3) and FDR ≤ 0.25 for the other three (A549, HA1E, and HT29). One important question that visualization with KEGGlincs for this pathway in particular can help us answer is, do we see high concordance in MCF7 edges that are part of the mTOR complexes or is ESEA picking up on less integral relationships?

In **Figure 55**, it is clear that many of the edges in the mTOR pathway are between paralogs for Wnt (WNT1, WNT5A, WNT7B, WNT9A WNT9B, WNT10B) and Frizzled (FZD1, FZD2, FZD4, FZD5, FZD7, FZD8). We would be skeptical that ESEA was picking up a true mTOR signal if most of the members in the LEE subset were part of the 36 different connections between these paralogs. Instead, only a small subset of these edges (7) contributes to the large

magnitude of the mTOR pathway NEES and most of (5) are between one Wnt paralog (WNT7B) and five of the six Frizzled paralogs (FZD1, FZD4, FZD5, FZD7, FZD8). Remarkably, the Wnt paralogs WNT7B has been singled out for being significantly up-regulated in breast cancer and suppression of its activity has been suggested to mediate breast cancer angiogenesis [90] [91].

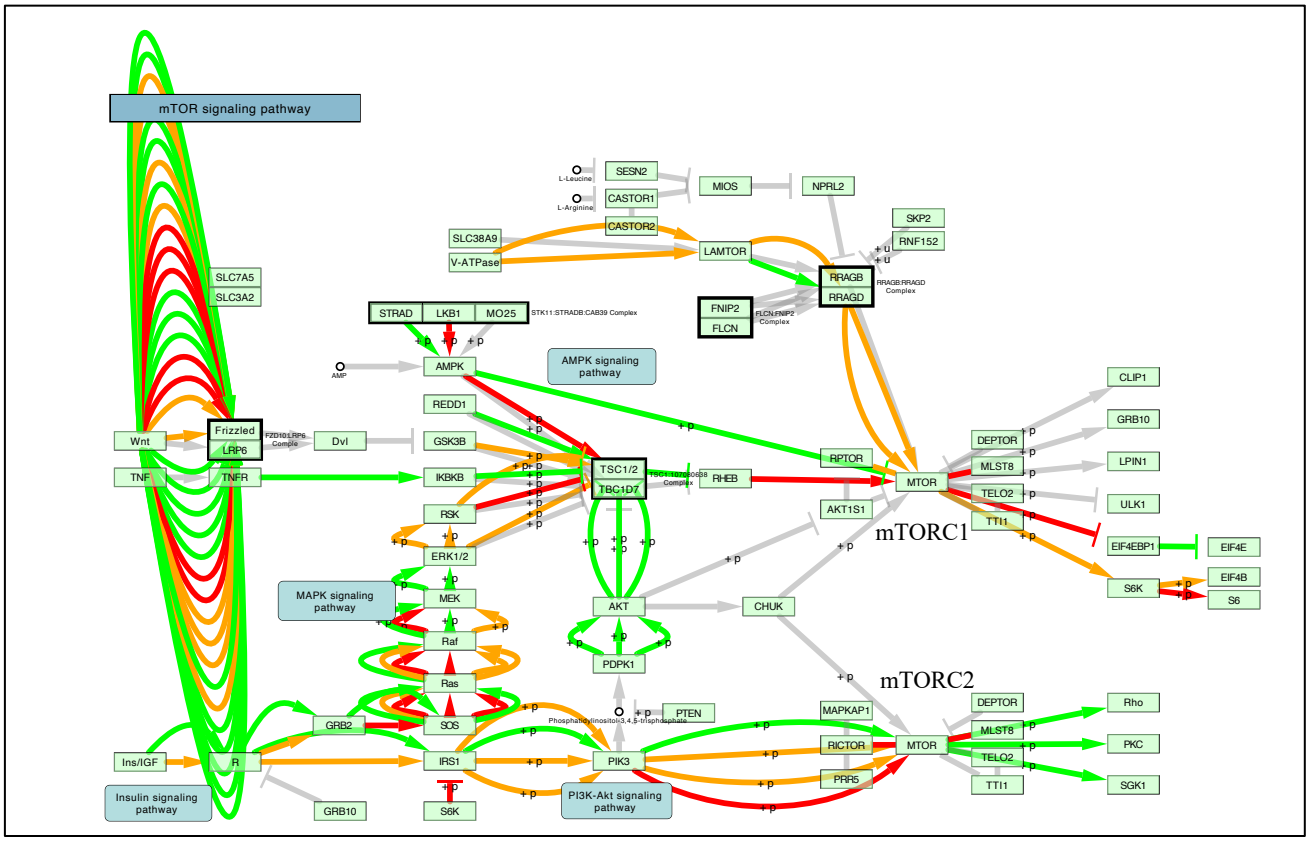


Figure 55: mTOR signaling pathway with formatting to reflect results of MCF7 global ESEA. Edges in the LEE subset are depicted in red; those that are concordant in MCF7 but not in the LEE subset are orange and those that are discordant in MCF7 are green.

Rather than the mTOR ESEA results depending on connections between, perhaps, redundant paralogs, **Figure 55** suggests that edges throughout the pathway contribute to the large NEES. Furthermore, five edges in the LEE subset involve relationships with mTOR itself. The relationship between Rheb and mTOR has been studied specifically in the MCF7 cell line and Rheb, as an activator of mTOR, has been suggested as a target for inhibition in treatment of

breast cancer [92]. Of the four PIK3 paralogs, the one with the relationship between itself and mTOR falling into the LEE subset is PIK3CA. Gain of function PIK3CA mutations specifically are among the most common in breast cancer and dual PIK/mTOR inhibitors have shown promise in the treatment of some breast cancers [93]. Two edges in the mTOR complex, between mTOR and MLST8 and RICTOR, are also in the LEE subset. RICTOR, while not frequently mutated in breast cancer, has been suggested to be more active in breast cancer via gene amplification, transcription, and/or catalytic activity and is implicated in tumor progression [94]. MLST8, which is essential in the formation of both mTORC1 and mTORC2 (mTOR complexes 1 and 2), has also been recently suggested as a therapeutic target for breast cancer treatment [95].

Table 31: Global and local results for Cell cycle pathway with MCF7 as reference cell line.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	AVERAGE	Cell cycle	04110	110	-0.304	-1.347	0.019	**0.096**	**Discordant**
MCF7	A375	Cell cycle	04110	110	-0.190	-0.873	0.328	**0.451**	**Discordant**
MCF7	A549	Cell cycle	04110	110	-0.273	-1.269	0.037	**0.171**	**Discordant**
MCF7	HA1E	Cell cycle	04110	110	-0.223	-1.071	0.12	**0.252**	**Discordant**
MCF7	HEPG2	Cell cycle	04110	110	-0.217	-0.973	0.251	**0.385**	**Discordant**
MCF7	HT29	Cell cycle	04110	110	-0.338	-1.484	0.006	**0.122**	**Discordant**
MCF7	PC3	Cell cycle	04110	110	-0.373	-1.630	0.001	**0.008**	**Discordant**

The most relevant discordant pathway that reaches significance for MCF7 in the ESEA global analysis is the cell cycle pathway (NEES = -1.347, FDR q-value = 0.96). Although **Table 31** singles out the local comparison between MCF7 and PC3 as the only significant local comparison at our threshold, a few other comparisons approach significance (A549, HT29) and across the panel the sign for the EES/NEES is negative – that is the edges in MCF7 are less

concordant/more discordant in all local comparisons. **Figure 56** shows this pathway formatted to reflect the ESEA results and, while there is a lot to untangle in this particular pathway, one notable observation is that most of the edges that make up the LEE subset (26 out of 41) are inhibitory in this case. While we have not made any claims with regard to inhibiting relationships and their association with discordance, this example demonstrates an added layer of dimensionality given to the interpretation of results.

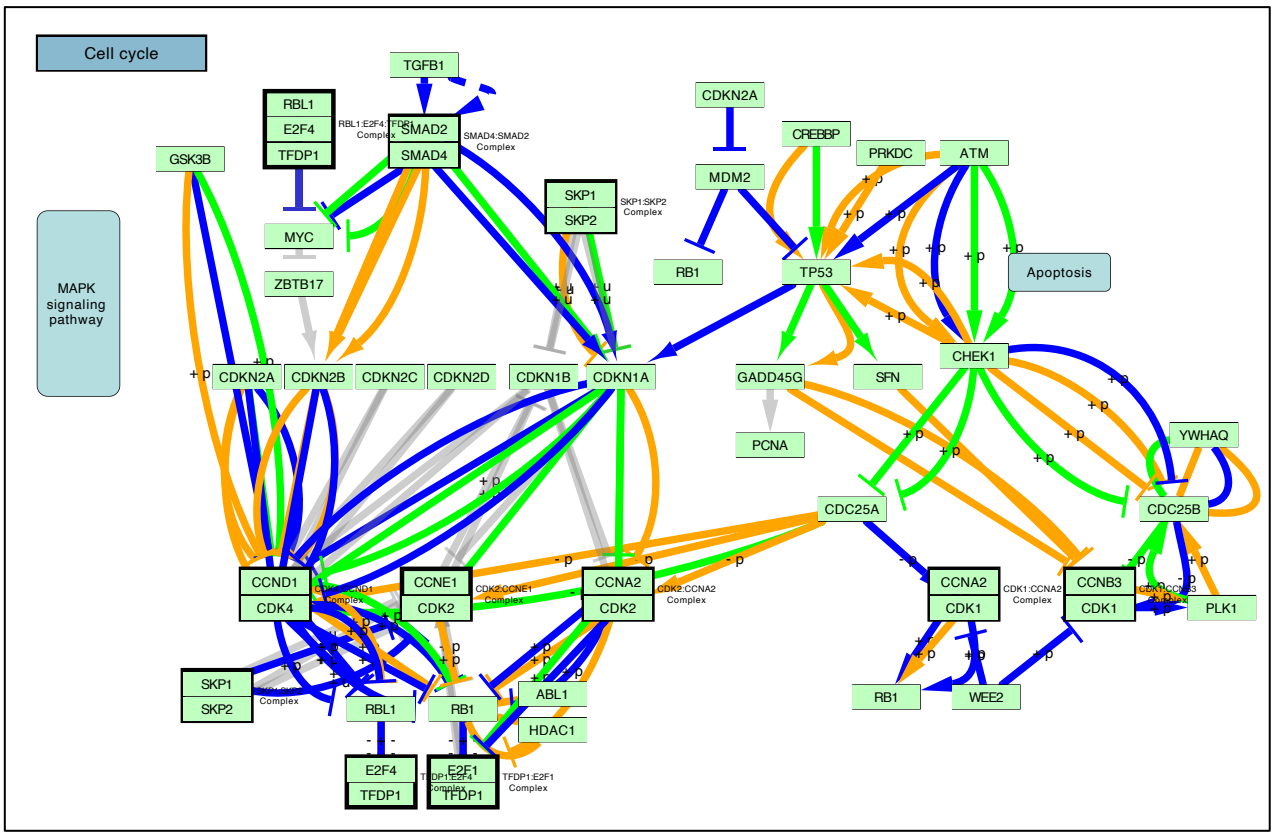


Figure 56: Cell cycle pathway with formatting to reflect results of MCF7 global ESEA. Edges in the LEE subset are depicted in blue; those that are concordant in MCF7 but not in the LEE subset are orange and those that are discordant in MCF7 are green.

Sustained proliferative signaling, evasion of growth suppressors, and resisting cell death (apoptosis) are hallmarks of cancer associated with disruption of the cell cycle program [96].

While we would expect all cell lines to exhibit abnormal behavior in this pathway when compared to normal cells, MCF7 is the only cell line in our panel with a significant ESEA result when compared to other cancer cell lines with this particular KEGG-defined pathway. 8 of the edges in the LEE subset involve relationships with RB1. RB1, first identified in retinoblastoma patients, was the first tumor suppressor gene to be molecularly defined [97]. Although its exact characterization as a tumor suppressor is complex, one of the main routes of growth inhibition is via repression of the transcription factor E2F. Though somewhat difficult to see against the blue edge connecting the RB1 and E2F1 nodes in **Figure 56**, the edge label "-+-" between these nodes is KEGG's way of labeling dissociative relationships; thus, in this pathway representation, active RB1 prevents the association of the proteins forming the complex. Although mutation of RB1 is associated with triple negative breast cancers, a subtype not characterized by the MCF7 cell line, these results support the body of evidence that RB1 function is often compromised – either by decreased expression or dysregulation of upstream regulators - across molecular subtypes [98].

6.3.3.2 Examples from Other Cell Lines

Table 32: Global and local results for Melanogenesis pathway with A375 as reference cell line.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A375	AVERAGE	Melanogenesis	04916	71	-0.395	-1.610	0.001	**0.081**	Discordant
A375	A549	Melanogenesis	04916	71	-0.308	-1.298	0.044	**0.255**	Discordant
A375	HA1E	Melanogenesis	04916	71	-0.362	-1.564	0.002	**0.027**	Discordant
A375	HEPG2	Melanogenesis	04916	71	-0.363	-1.521	0.008	**0.118**	Discordant
A375	HT29	Melanogenesis	04916	71	-0.276	-1.148	0.131	**0.383**	Discordant
A375	MCF7	Melanogenesis	04916	71	-0.370	-1.524	0.009	**0.104**	Discordant
A375	PC3	Melanogenesis	04916	71	-0.369	-1.542	0.006	**0.061**	Discordant

Although we went into detail with the results for one particular cell line, the results for other cell lines also suggest the utility of this approach. For example, as shown in **Table 32**, the Melanogenesis pathway is significantly enriched with discordant edges in the A375 skin cancer cell line global results as and the direction is discordant in all local comparisons (3 with FDR ≤ 0.10). In melanoma, loss of pigmentation is common because of dysfunction in melanogenesis proteins which, though not directly represented by edges in this pathway, may have their dysfunction attributed to deregulated proteins in this pathway. In fact, there is research to suggest that malignancy in skin cancer may be reversed when using A375 as a model cell line [99]. Another pertinent example is the Platinum drug resistance pathway in the HT29 colon cancer cell line with significant enrichment in the concordant direction in the global comparison and concordant direction in all local comparisons. HT29 is actually a model cell line for the development of platinum drug resistance in colon cancer [100].

Table 33: Global and local results for Platinum drug resistance pathway with HT29 as reference cell line.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HT29	AVERAGE	Platinum drug resistance	01524	49	0.463	1.788	0	0.000	Concordant
HT29	A375	Platinum drug resistance	01524	49	0.366	1.429	0.019	0.202	Concordant
HT29	A549	Platinum drug resistance	01524	49	0.315	1.231	0.087	0.285	Concordant
HT29	HA1E	Platinum drug resistance	01524	49	0.410	1.613	0.006	0.086	Concordant
HT29	HEPG2	Platinum drug resistance	01524	49	0.347	1.357	0.038	0.232	Concordant
HT29	MCF7	Platinum drug resistance	01524	49	0.387	1.502	0.017	0.187	Concordant
HT29	PC3	Platinum drug resistance	01524	49	0.323	1.248	0.071	0.205	Concordant

6.4 Discussion

There are two readily identifiable limitations of this approach. The first is the completeness, or perhaps lack thereof, of the L1000 data set. We are using a subset of the

available data for which all cell lines have the same shRNA perturbation. The second limitation is the dynamic nature of KEGG pathways both in terms of available pathways and pathway specification. KEGG pathways are manually curated with new pathways being added and existing pathways being modified as experimental evidence for the relationships between genes becomes available. That being said, we have demonstrated how to conduct an unsupervised analysis of L1000 data with KEGG pathways that provides meaningful results without requiring any of our own manual pathway alterations.

There are also opportunities to apply this same type of analysis to the L1000 data set with different research objectives. We have performed this analysis on a 'slice' of the data that considers records at the same dose of perturbagen (concentration = 1μl) measured at the same time (96 hours after perturbation). The same approach that we used to make comparisons across cell lines could readily be applied within cell lines to see how pathway activity changes across dose or time. Furthermore, as the database grows, new cell lines could be incorporated into the analysis.

The LINCS L1000 dataset contains millions of data points, each with its own set of attributes with regard to specific cellular perturbation, cell line, and record for L1000 gene expression. We have demonstrated a bioinformatics-based approach to the analysis of this multi-dimensional data set that leverages existing methods with the format of our data. As part of this effort, we have procured up-to-date records of pathway topology and incorporated interactive visualization tools in an effort to make the results relevant and interpretable. The results of this analysis could be used to support ongoing research efforts or aid in hypothesis generation in an effort to further enrich the field of cancer research.

References

[1] "NIH LINCS Program," BD2K-LINCS DCIC, 2019. [Online]. Available: lincsproject.org.

[2] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, G. Joshua, J. F. Davis, A. A. Tubellin, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian and M. Khan, "A Next Generation Connectivity Map: L1000 platform and the first 1,000,000 profiles," *Cell,* vol. 171, no. 6, pp. 1437-1452, 2017.

[3] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong and S. Haggarty, "The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science,* vol. 313, pp. 1929-1935, 2006.

[4] [Online]. Available: https://clue.io/connectopedia/what_are_landmark_genes.

[5] I. Smith, P. G. Greenside, T. Natoli, D. L. Lahr, D. Wadden, I. Tirosh, R. Narayan, D. E. Root, T. R. Golub, A. Subramanian and J. G. Doench, "Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map," *PLOS Biology,* vol. 15, no. 11, 2017.

[6] N. Silver, S. Best, J. Jiang and S. L. Thein, "Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR," *BMC Molecular Biology,* vol. 7, 2006.

[7] B. Szalai, V. Subramanian, R. Alfoldi, L. G. Puskas and J. Saez-Rodriguez, "Signatures of cell death and proliferation in perturbation transcriptomics data - from confounding factor to effective prediction," *preprint.*

[8] J. Peng, J. R. Scarpa, V. D. Gao, M. H. Vitaterna, A. Kasarkis and F. W. Turek, "Parkinson's disease is associated with dysregulations of a dopamine-modulated gene network relevant to sleep and affective neurobehaviors in the striatum," *Scientific Reports,* vol. 9, pp. 1-14, 2019.

[9] A. Li, X. Lu, T. Natoli, J. Bittker, N. S. Sipes, A. Subramanian, S. Aurbach, D. H. Sherr and S. Monti, "The Carcinogenome Project: In vitro gene expression profiling of chemical perturbations to predict long-term carcinogenicity," *Environmental Health Perspectives,* vol. 127, no. 4, 2019.

[10] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, E. L. Benjamin, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *PNAS,* vol. 102, no. 43, pp. 15545-15550, 2005.

[11] M. Kanehisa, Y. Sato, M. Furumichi, K. Morishima and M. Tanabe, "New approach for understanding genome variations in KEGG," *Nucleic Acids Research,* vol. 47, no. D1, pp. D590-D595, 2019.

[12] M. Kanehisa, "Toward understanding the origin and evolution of cellular organisms," *Protein Science,* pp. 1-5, 2019.

[13] R. Cao, B. Robinson, H. Xu, C. Gkogkas, A. Khoutorsky, T. Alain, A. Yanagiya, T. Nevarko, A. C. Liu, S. Amir and N. Sonenberg, "Translational control of entrainment and synchrony of the suprachiasmatic circadian clock by mTOR/4E-BP1 signaling," *Neuron,* vol. 79, no. 4, pp. 712-724, 2013.

[14] J. L. Fleiss, B. Levin and M. C. Paik, Statistical Methods for Rates and Proportions: Third Edition, Hoboken, New Jersey: John Wiley & Sns, Inc., 2003.

[15] A. Agresti, An Introduction to Categorical Data Analysis, Hoboken, New Jersey: John Wiley & Sons, Inc., 2007.

[16] K. T. Wallenius, "Biased Sampling; the Noncentral Hypergeometric Distribution. Ph D Thesis, Stanford University.," *Department of Statistcs, Stanford University,* vol. November, 1963.

[17] J. Chesson, "A Non-Central Multivariate Hypergeometric Distribution Arising from Biased Sampling with Application to Selective Predation," *Journal of Applied Probability,* vol. 13, no. 4, pp. 795-797, 1976.

[18] J. Chesson, "Measuring Preference in Selective Predation," *Ecology,* vol. 59, no. 2, pp. 211-215, 1978.

[19] A. Fog, "Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution," *Communications in Statistics—Simulation and Computation,* vol. 37, no. 2, pp. 258-273, 2008.

[20] S. Loertscher, E. Muir and P. G. Taylor, "A general noncentral hypergeometric distribution," *Communictions in Statistics - Theory and Methods,* vol. 46, no. 9, pp. 4579-4598, 2017.

[21] A. Fog, "Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric," *Communications in Statistics, Simulation and Computation,* vol. 37, no. 2, pp. 241-257, 2008.

[22] J. Sedransk and J. Meyer, "Confidence Intervals for the Quantiles of a Finite Population: Simple Random and Stratified Simple Random Sampling," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 40, no. 2, pp. 239-252, 1978.

[23] T. P. Morris, I. R. White and M. J. Crowther, "Using simulation studies to evaluate statistical methods," *Statistics in Medicine,* 2017.

[24] P. E. Cheng, M. Liou, J. A. D. Aston and A. C. Tsai, "Information identities and testing hypotheses: Power analysis for contingency tables," *Statistica Sinica,* vol. 18, p. 535–558, 2008.

[25] M. Kateri, Contingency Table Analysis: Methods and Implementation Using R, New York: Springer Science+Business Media, 2014.

[26] J. M. Raser and E. K. O'Shea, "Noise in Gene Expression: Origins, Consequences, and Control," *Science,* vol. 309, no. 5743, pp. 2010-2013, 2005.

[27] M. Medvedovic, "iLINCS," 2020. [Online]. Available: ilincs.org.

[28] E. Million, "The Hadamard Product," 12 April 2007. [Online]. Available: http://buzzard.ups.edu/courses/2007spring/projects/million-paper.pdf. [Accessed 31 January 2020].

[29] P. Mirabelli, L. Coppola and M. Salvatore, "Cancer cell lines are useful models for medical research," *Cancers,* vol. 11, no. 8, pp. 1098-1116, 2019.

[30] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, A. Bretscher, H. Ploegh and A. Amon, Molecular Cell Bioloby - 7th Edition, New York: W.H. Freeman and Company, 2000.

[31] A. Ertel, A. Verghese, S. W. Byers, M. Ochs and A. Tozeren, "Pathway-specific differences between tumor cell lines and normal and tumor tissue cells," *Molecular Cancer,* vol. 5, no. 55, 2006.

[32] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research,* vol. 27, no. 1, pp. 29-34, 1999.

[33] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gapinath, B. Jassal, S. Jupe, I. Kalatskya, S. Mahajan, B. May, N. Ndegwa, E. Schmidt and Shamovs, "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research,* vol. 39, p. D691–D697, 2011.

[34] P. Creixell, J. Reimand, S. Haider, G. Wu, T. Shibata, M. Vazquez, V. Mustonen, Gonzalez-Perez, J. Pearson, C. Sander, B. J. Raphael, D. S. Marks, B. F. F. Oullette, A. Valencia, D. G. Bader and Bo, "Pathway and network analysis of cancer genomes," *Nature Methods,* vol. 12, no. 7, pp. 615-621, 2015.

[35] A. Liberzon, A. Subramaniam, R. Pinchback, H. Thorvaldsdottir, P. Tamayo and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics,* vol. 27, no. 12, p. 1739–1740, 2011.

[36] P. Khatri, M. Sirota and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLOS Computational Biology,* vol. 8, no. 2, 2012.

[37] M. A. Sartor, G. D. Leikauf and M. Medvedovic, "LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data," *Bioinformatics,* vol. 25, no. 2, pp. 211-217, 2009.

[38] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Haustis, M. J. Daly, N. Patterson, J. P. Mesirov and Golub, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics,* vol. 34, p. 267–273, 2003.

[39] J. Rahnenfuhrer, F. S. Domingues, J. Maydt and T. Lengauer, "Calculating the statistical significance of changes in pathway activity from gene expression data," *Statistical Applications in Genetics and Molecular Biology,* vol. 3, no. 1, p. Article 16, 2004.

[40] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics,* vol. 25, no. 1, p. 75–82, 2009.

[41] S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu and R. Romero, "A systems biology approach for pathway level analysis," *Genome Research,* vol. 17, no. 10, p. 1537–1545, 2007.

[42] A. Kramer, J. Green, J. Pollard, Jr. and S. Tugendreich, "Causal analysis approaches in Ingenuity Pathway Analysis," *Bioinformatics,* vol. 30, no. 4, pp. 523-530, 2013.

[43] K. Faust, P. Dupont, J. Callut and J. van Helden, "Pathway discovery in metabolic networks by subgraph extraction," *Bioinformatics,* vol. 26, no. 9, p. 1211–1218, 2010.

[44] J. H. McDonald, Handbook of Biological Statistics (3rd Edition), Baltimore: Sparky House Publishing, 2014.

[45] P. Prieto-Marañón, M. E. Aguerri, M. S. Galibert and H. F. Attoresi, "Detection of differential item functioning - Using decision rules based on the Mantel-Haenszel procedure and breslow-day tests," *Methodology,* vol. 8, no. 2, pp. 63-70, 2012.

[46] N. E. Breslow, "Statistics in epidemiology: the case-control study," *Journal of the American Statistical Association,* vol. 91, no. 433, pp. 14-28, 1996.

[47] L. Li, G.-d. Zhao, L.-l. Qi, L.-Y. Zhou and Z.-X. Fu, "The Ras/Raf/MEK/ERK signaling pathway and its role in the occurrence and development of HCC (Review)," *Oncology Letters,* vol. 12, no. 5, p. 3045–3050, 2016.

[48] C. Muñoz-Maldonado, Y. Zimmer and M. Medova, "A comparative analysis of RAS mutations in cancer biology," *Frontiers in Onocology,* vol. 9, 2019.

[49] D. F. Stern, "Tyrosine kinase signalling in breast cancer: ErbB family receptor tyrosine kinases," *Breast Cancer Research,* vol. 2, pp. 176-183, 2000.

[50] K. Zhang, P. Wong, C. Salvaggio, A. Salhi, I. Osman and B. Bedogni, "Tyrosine kinase signalling in breast cancer: ErbB family receptor tyrosine kinases," *Journal of Investigative Dermatology,* vol. 136, no. 2, pp. 464-472, 2016.

[51] W. Ka Kei Wu, T. T. Ming Tse, J. Jao Yiu Sung, Z. Jie Li, L. Yu and C. Hin Cho, "Expression of ErbB receptors and their cognate ligands in gastric and colon cancer cell lines," *Anticancer Research,* vol. 29, pp. 229-234, 2009.

[52] J. A. Englemen and L. C. Cantley, "The role of the ErbB family members in non–small cell lung cancers sensitive to epidermal growth factor receptor kinase inhibitors," *Clinical Cancer Research,* vol. 12, no. 14, pp. 4372s-4376s, 2006.

[53] P. Huang, X. Xu, L. Wang, B. Zhu, X. Wang and J. Xia, "The role of EGF-EGFR signalling pathway in hepatocellular carcinoma inflammatory microenvironment," *Journal of Cellular and Molecular Medecine,* vol. 18, no. 2, p. 218–230, 2014.

[54] M. Wang, D. Ren, W. Guo , S. Huang, Z. Wang, Q. Li, H. Du, L. Song and X. Peng, "N-cadherin promotes epithelial-mesenchymal transition and cancer stem cell-like traits via ErbB signaling in prostate cancer cells," *International Journal of Onocology,* vol. 48, pp. 595-606, 2016.

[55] M. e. a. Hiroyuki, "HCaRG/COMMD5 inhibits ErbB receptor-driven renal cell carcinoma," *Oncotarget,* vol. 8, no. 41, pp. 69559-69576, 2017.

[56] U. S. D. a. B. Institute, "Molecular Signatures Database v7.2," Broad Institute, Inc., 2004-2020. [Online]. Available: https://www.gsea-msigdb.org/gsea/msigdb/index.jsp. [Accessed 4 December 2020].

[57] G. Korotkevich, V. Sukhov, N. Budin and A. Sergushichev, *Fast gene set enrichment analysis,* bioRxiv. doi: 10.1101/060012, 2019.

[58] E. Planet, *phenoTest: Tools to test association between gene expression and phenotype in a way that is efficient, structured, fast and scalable. We also provide tools to do GSEA (Gene set enrichment analysis) and copy number variation. R package version 1.38.0.,* 2020.

[59] T. V. Perneger, "What's wrong with Bonferroni adjustments," *BMJ,* vol. 316, no. 7139, p. 1236–1238, 1998.

[60] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *PNAS,* vol. 100, no. 16, 2003.

[61] B. Sorić, "Statistical "discoveries" and effect-size estimation," *Journal of the American Statistical Association,* vol. 84, no. 406, pp. 608-610, 1989.

[62] Y. Benjamini and Y. Hochberg, "Controling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 57, no. 1, pp. 289-300, 1995.

[63] J. Han, X. Shi, Y. Zhang, Y. Xu, Y. Jiang, C. Zhang, H. Yang, D. Shang, Z. Sun and L. X. C. Xia , "ESEA: Discovering the dysregulated pathways based on Edge Set Enrichment Analysis," *Scientific Reports,* vol. 5, 2015.

[64] A. F. Villaverde, J. Ross and J. R. Banga, "Reverse engineering cellular networks with information theoretic measuerments," *cells,* vol. 2, pp. 306-329, 2013.

[65] C. E. Shannon and W. Weaver, A Mathematical Theory of Communication, Urbana: The University of Illinois Press, 1964.

[66] S. Liang, S. Fuhrman and R. Somogyi, "REVEAL, a general reverse engineering algorithm for inferences of genetic network architectures," *Pacific Symposium on Biocomputing,* vol. 3, pp. 18-29, 1998.

[67] P. D'haeseleer, S. Fuhrman, W. Xiling and R. Smogyi, "Mining the Gene Expression Matrix: Inferring gene relationships from large scale gene expression data," in *Proceedings of the second international workshop on Information processing in cell and tissues*, Sheffield, 1998.

[68] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomics clustering using pairwise entropy measurements," *Pacific Symposium on Biocomputing,* vol. 5, pp. 415-426 , 2000.

[69] A. Kraskov, H. Stögbauer and P. Grassberger, "Estimating mutual information," *Physical Review,* vol. 69, 2004.

[70] M. Olivier, R. Eeles, M. Hollstein, M. A. Khan, C. C. Harris and P. Hainaut, "The IARC TP53 database: new online mutation analysis and recommendations to users," *Human Mutation,* vol. 19, no. 6, pp. 607-614, 2002.

[71] J. Han, X. Shi and C. Li, *ESEA: ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. R package version 1.0. https://CRAN.R-project.org/package=ESEA,* 2015.

[72] I. Rybinska, R. Agresti, A. Trapani, E. Tagliabue and T. Triulzi, "Adipocytes in breastcancer, the thick and the thin," *Cells,* vol. 9, no. 3, 2020.

[73] S. Balaban, R. F. Shearer, L. S. Lee, M. van Geldermalsen, M. Schreuder, H. C. Shtein, R. Cairns, K. C. Thomas, D. J. Fazakerley, T. Grewal, J. Holst, D. N. Saunders and A. J. Hoy, "Adipocyte lipolysis links obesity to breast cancer growth: adipocyte-derived fatty acids drive breast cancer cell proliferation and migration," *Cancer & Metabolism,* vol. 5, no. 1, 2017.

[74] I. Lal, K. Dittus and C. E. Holmes, "Platelets, coagulation and fibrinolysis in breast cancer progression," *Breast Cancer Research,* vol. 15, 2013.

[75] L. Lian, W. Li, L. Zhen-yu, Y.X. Mao, Y.T. Zhang, Y.M. Zhao, K. Chen, W.M. Duan and M. Tao, "Inhibition of MCF-7 breast cancer cell-induced platelet aggregation using a combination of antiplatelet drugs," *Oncology Letters,* vol. 5, no. 2, p. 675–680, 2013.

[76] M. Zarà, I. Canobio, C. Visconte, J. Canino, M. Torti and G. F. Guidetti, "Molecular mechanisms of platelet activation and aggregation induced by breast cancer cells," *Cellular Signalling,* vol. 48, pp. 45-53, 2018.

[77] C. G. Valenti , L. Fianchi, M. T. Voso, M. Caira, G. Leone and L. Pagano, "Incidence of myeloid leukemia after breast cancer.," *Mediterranean Journal of Hematology and Infectious Diseases,* vol. 3, no. 1, p. e2011069, 2011.

[78] H. G. Kaplan, J. A. Malmgren and M. K. Atwood, "Increased incidence of myelodysplastic syndrome and acute myeloid leukemia following breast cancer treatment with radiation alone or combined with chemotherapy: a registry cohort analysis 1990-2005," *BMC Cancer,* vol. 11, no. 260, p. 1:10, 2011.

[79] D. A. Patt, Z. Duan, S. Fang, G. N. Hortobagyi and S. H. Giordano, "Acute myeloid leukemia after adjuvant breast cancer therapy in older women: understanding risk," *Journal of Clinical Oncology,* vol. 25, no. 25, pp. 3871-3876, 2007.

[80] M. I. Palacios-Arreola, K. E. Nava-Castro, J. I. Castro, E. García-Zepeda, J. C. Carrero and J. Morales-Montor, "The role of chemokines in breast cancer pathology and its possible use as therapeutic targets," *Journal of Immunology Research,* 2014.

[81] S. J. Prest, R. C. Rees, C. Murdoch, J. F. Marshall, P. A. Cooper, M. Bibby, G. Li and S. A. Ali, "Chemokines induce the cellular migration of MCF-7 human breast carcinoma cells: Subpopulations of tumour cells display positive and negative chemotaxis and differential in vivo growth potentials," *Clinical & Experimental Metastasis,* vol. 17, pp. 389-396, 1999.

[82] S. L. Henmbruff and N. Cheng, "Chemokine signaling in cancer: Implications on the tumor microenvironment and therapeutic targeting," *Cancer Therapeutics,* vol. 7, pp. 254-267, 2009.

[83] S. H. Hare and A. J. Harvey, "mTOR function and therapeutic targeting in breast cancer," *American Journal of Cancer Research,* vol. 7, no. 3, p. 383–404, 2017.

[84] S. Vinayak and R. W. Carlson, "mTOR inhibitors in the treatment of breast cancer," *Oncology,* vol. 27, no. 1, pp. 38-44, 2013.

[85] L. Du, L. Xiamei, L. Zhen, W. Chen, L. Mu, Y. Zhang and A. Song, "Everolimus inhibits breast cancer cell growth through PI3K/AKT/mTOR signaling pathway," *Molecular Medicine Reports,* vol. 17, no. 5, pp. 7163-7169, 2018.

[86] S. Descamps, X. Lebourhis, M. Delehedde, B. Boilly and H. Hondermarck, "Nerve growth factor is mitogenic for cancerous but not normal breast epithelial cells," *Journal of Biological Chemistry,* vol. 273, pp. 16659-16662, 1998.

[87] S. Descamps, R.A. Toillon, E. Adriaenssens, V. Pawlowski, S. M. Cool, V. Nurcombe, X. Le Bourhis, B. Boilly, J.P. Peyrat and H. Hondermarck, "Nerve growth factor stimulates proliferation and survival of human breast cancer cells through two distinct signaling pathways," *The Journal of Biological Chemistry,* vol. 276, no. 21, p. 17864–17870, 2001.

[88] D. Borcherding, W. Tong, E. Hugo, D. Barnard, S. Fox, K. LaSance, E. Shaughnessy and N. Ben-Jonathan, "Expression and therapeutic targeting of dopamine receptor-1 (D1R) in breast cancer," *Oncogene,* vol. 35, p. 3103–3113, 2016.

[89] D. C. Borcherding, W. Tong, E. R. Hugo, D. F. Barnard, S. Fox, K. LaSance, E. Shaughnessy and N. Ben-Jonathan, "Expression and Therapeutic targeting of dopamine receptor-1 (D1R) in breast cancer," *Oncogene,* vol. 35, no. 24, pp. 3103-3113, June 2017.

[90] E.J. Yeo, L. Casseta, B.Z. Qian, I. Lewkowich, J.F. Li, J. A. Stephater, A. N. Smith, L. S. Wiechmann, Y. Wang, J. W. Pollard and R. A. Lang, "Myeloid WNT7b mediates the angiogenic switch and metastasis in breast cancer," *Cancer Research,* vol. 74, no. 11, p. 2962–2973, 2014.

[91] J. Chen, T.Y. Liu, H.T. Peng, Y.-Q. Wu, L.L. Zhang, X.H. Lin and Y.H. Lai, "Up-regulation of Wnt7b rather than Wnt1, Wnt7a, and Wnt9a indicates poor prognosis in breast cancer," *International Journal of Clinical and Experimental Pathology,* vol. 11, no. 9, p. 4552–4561, 2018.

[92] J. Yu and E. Petri Henske, "Estrogen induced activation of mammalian target of rapamycin is mediated via tuberin and the small GTPase Ras homologue enriched in brain," *Cancer Research,* vol. 66, no. 19, pp. 9461-9466, 2006.

[93] J. J. Lee, K. Loh and Y.-S. Yap, "PI3K/Akt/mTOR inhibitors in breast cancer," *Cancer Biology & Medicine,* vol. 12, no. 4, 2015.

[94] M. Morrison Joly, D. J. Hicks, B. Jones, V. Sanchez, M. V. Estrada, C. Young, M. Williams, B. N. Rexer, D. D. Sarbassov, W. J. Muller, D. Brantley-Sieders and R. S. Cook, "Rictor/mTORC2 drives progression and therapeutic resistance of HER2-amplified breast cancers," *Cancer Research,* vol. 76, no. 16, p. 4752–4764, 2016.

[95] S. Shin, G. Yang, D. E. James and L. K. Nguyen, "Dynamic modelling of the PI3K/mTOR signalling network uncovers biphasic dependence of mTORC1 activation on the mTORC2 subunit Sin1".

[96] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell,* vol. 144, no. 5, 2011.

[97] N. J. Dyson, "RB1: a prototype tumor suppressor and an enigma," *Genes & Development,* vol. 30, pp. 1492-1502, 2016.

[98] A. Witkiewicz and E. S. Knudsen, "Retinoblastoma tumor suppressor pathway in breast cancer: prognosis, precision medicine and therapeutic interventions," *Breast Cancer Research,* vol. 16, 2014.

[99] C. Bracalente, N. Salguero, C. Notcovich, C. B. Müller, L. L. da Motta, F. Klamt, I. L. Ibañez and H. Durán, "Reprogramming human A375 amelanotic melanoma cells by catalase overexpression: Reversion or promotion of malignancy by inducing melanogenesis or metastasis," *Oncotarget,* vol. 7, no. 27, p. 41142–41153, 2016.

[100] T. Hu, Z. Li, C.-Y. Gao and C. Hin Cho, "Mechanisms of drug resistance in colon cancer and its therapeutic strategies," *World Journal of Gastroenterology,* vol. 22, no. 30, pp. 6876-6889, 2016.

[101] Katz-Brull, D. Seger, D. Rivenson-Segal, E. Rushkin and H. Degani, "Metabolic markers of breast cancer: Enhanced choline metabolism and reduced choline-ether-phospholipid synthesis," *Cancer Research,* vol. 62, pp. 1966-1970, 2002.

Appendix Tables

Complete Null Parameterization

Table A1.a: Group Size

Scenario	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
3RG	333	333	333						
5RG	200	200	200	200	200				
7RG	142	142	142	142	142	142	142		
9RG	111	111	111	111	111	111	111	111	111

Table A1.b: Selection Weights

Scenario	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9
3RG	1	1	1						
5RG	1	1	1	1	1				
7RG	1	1	1	1	1	1	1		
9RG	1	1	1	1	1	1	1	1	1

Table A1.c: Direction Weights

Scenario	φ_1^U	φ_2^U	φ_3^U	φ_4^U	φ_5^U	φ_6^U	φ_7^U	φ_8^U	φ_9^U
3RG	0.5	0.5	0.5						
5RG	0.5	0.5	0.5	0.5	0.5				
7RG	0.5	0.5	0.5	0.5	0.5	0.5	0.5		
9RG	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Null Selection Weight

Table A2.a: Group Size

Scenario	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
3RG	100	100	800						
5RG	50	50	200	200	500				
7RG	50	50	100	100	200	200	300		
9RG	50	50	75	75	100	100	125	125	300

Table A2.b: Selection Weights

Scenario	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9
3RG	1	1	1						
5RG	1	1	1	1	1				
7RG	1	1	1	1	1	1	1		
9RG	1	1	1	1	1	1	1	1	1

Table A2.c: Direction Weights

Scenario	φ_1^U	φ_2^U	φ_3^U	φ_4^U	φ_5^U	φ_6^U	φ_7^U	φ_8^U	φ_9^U
3RG	0.9	0.1	0.5						
5RG	0.9	0.1	0.7	0.3	0.5				
7RG	0.9	0.1	0.8	0.2	0.7	0.3	0.5		
9RG	0.9	0.1	0.8	0.2	0.7	0.3	0.6	0.4	0.5

Null Direction Weight

Table A3.a: Group Size

Scenario	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
3RG	100	100	800						
5RG	50	50	200	200	500				
7RG	50	50	100	100	200	200	300		
9RG	50	50	75	75	100	100	125	125	300

Table A3.b: Selection Weights

Scenario	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9
3RG	1	1	1						
5RG	1	1	1	1	1				
7RG	1	1	1	1	1	1	1		
9RG	1	1	1	1	1	1	1	1	1

Table A3.c: Direction Weights

Scenario	φ_1^U	φ_2^U	φ_3^U	φ_4^U	φ_5^U	φ_6^U	φ_7^U	φ_8^U	φ_9^U
3RG	0.5	0.5	0.5						
5RG	0.5	0.5	0.5	0.5	0.5				
7RG	0.5	0.5	0.5	0.5	0.5	0.5	0.5		
9RG	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table A4a: Global Comparisons for A375 - FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A375	AVERAGE	Toll-like receptor signaling pathway	04620	70	0.422	1.772	0.001	0.081	Concordant
A375	AVERAGE	Melanogenesis	04916	71	-0.395	-1.610	0.001	0.081	Discordant
A375	AVERAGE	Th1 and Th2 cell differentiation	04658	87	0.360	1.588	0.002	0.081	Concordant
A375	AVERAGE	Salmonella infection	05132	124	0.293	1.403	0.002	0.081	Concordant
A375	AVERAGE	Pertussis	05133	31	0.456	1.642	0.005	0.093	Concordant
A375	AVERAGE	Longevity regulating pathway - multiple species	04213	90	-0.430	-1.623	0.004	0.093	Discordant
A375	AVERAGE	Basal cell carcinoma	05217	66	-0.381	-1.558	0.004	0.093	Discordant
A375	AVERAGE	Yersinia infection	05135	92	0.351	1.551	0.003	0.093	Concordant
A375	AVERAGE	GnRH signaling pathway	04912	59	0.369	1.518	0.006	0.093	Concordant
A375	AVERAGE	Colorectal cancer	05210	106	0.330	1.465	0.007	0.093	Concordant
A375	AVERAGE	Longevity regulating pathway	04211	60	-0.369	-1.461	0.008	0.093	Discordant
A375	AVERAGE	Human immunodeficiency virus 1 infection	05170	111	0.315	1.441	0.008	0.093	Concordant
A375	AVERAGE	HIF-1 signaling pathway	04066	163	-0.304	-1.420	0.007	0.093	Discordant
A375	AVERAGE	Pathogenic Escherichia coli infection	05130	88	0.316	1.413	0.007	0.093	Concordant

Table A4b: Global Comparisons for HA1E - FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HA1E	AVERAGE	VEGF signaling pathway	04370	66	-0.479	-1.960	0	0.000	Discordant
HA1E	AVERAGE	Osteoclast differentiation	04380	98	-0.427	-1.909	0	0.000	Discordant
HA1E	AVERAGE	Hepatitis B	05161	154	-0.400	-1.892	0	0.000	Discordant
HA1E	AVERAGE	Salmonella infection	05132	124	-0.358	-1.661	0	0.000	Discordant
HA1E	AVERAGE	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.345	-1.633	0	0.000	Discordant
HA1E	AVERAGE	Neurotrophin signaling pathway	04722	135	-0.327	-1.546	0	0.000	Discordant
HA1E	AVERAGE	HIF-1 signaling pathway	04066	163	-0.320	-1.534	0	0.000	Discordant
HA1E	AVERAGE	Apelin signaling pathway	04371	85	0.336	1.587	0.001	0.020	Concordant
HA1E	AVERAGE	TNF signaling pathway	04668	66	-0.437	-1.735	0.002	0.032	Discordant
HA1E	AVERAGE	Cushing syndrome	04934	82	0.345	1.576	0.002	0.032	Concordant
HA1E	AVERAGE	Longevity regulating pathway	04211	60	-0.406	-1.685	0.003	0.037	Discordant
HA1E	AVERAGE	Acute myeloid leukemia	05221	112	-0.361	-1.601	0.003	0.037	Discordant
HA1E	AVERAGE	Endocrine resistance	01522	177	-0.343	-1.590	0.003	0.037	Discordant
HA1E	AVERAGE	Tuberculosis	05152	88	0.303	1.420	0.004	0.046	Concordant
HA1E	AVERAGE	Toll-like receptor signaling pathway	04620	70	-0.399	-1.669	0.005	0.054	Discordant
HA1E	AVERAGE	Ras signaling pathway	04014	256	-0.271	-1.379	0.009	0.091	Discordant
HA1E	AVERAGE	Platelet activation	04611	33	-0.437	-1.560	0.01	0.095	Discordant
HA1E	AVERAGE	Adherens junction	04520	34	0.429	1.567	0.011	0.099	Concordant
HA1E	AVERAGE	Endometrial cancer	05213	71	-0.365	-1.505	0.013	0.100	Discordant
HA1E	AVERAGE	Chemokine signaling pathway	04062	119	-0.310	-1.432	0.013	0.100	Discordant
HA1E	AVERAGE	MAPK signaling pathway	04010	353	-0.248	-1.305	0.012	0.100	Discordant

Table A4c: Global Comparisons for HEPG2 - FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HEPG2	AVERAGE	Adherens junction	04520	34	-0.576	-1.946	0	0.000	Discordant
HEPG2	AVERAGE	Antigen processing and presentation	04612	77	0.537	1.915	0	0.000	Concordant
HEPG2	AVERAGE	Longevity regulating pathway - multiple species	04213	90	0.460	1.741	0.001	0.040	Concordant
HEPG2	AVERAGE	Longevity regulating pathway	04211	60	0.440	1.737	0.001	0.040	Concordant
HEPG2	AVERAGE	Thermogenesis	04714	45	-0.462	-1.747	0.002	0.065	Discordant
HEPG2	AVERAGE	Oocyte meiosis	04114	37	0.443	1.595	0.003	0.081	Concordant
HEPG2	AVERAGE	Necroptosis	04217	81	-0.360	-1.520	0.004	0.093	Discordant

Table A4d: Global Comparisons for HT29 - FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HT29	AVERAGE	Apoptosis - multiple species	04215	17	-0.692	-2.043	0	0.000	Discordant
HT29	AVERAGE	Melanoma	05218	115	0.456	1.983	0	0.000	Concordant
HT29	AVERAGE	Platinum drug resistance	01524	49	0.463	1.788	0	0.000	Concordant
HT29	AVERAGE	Prostate cancer	05215	150	0.365	1.682	0	0.000	Concordant
HT29	AVERAGE	Focal adhesion	04510	208	0.312	1.522	0	0.000	Concordant
HT29	AVERAGE	Pathways in cancer	05200	554	0.248	1.320	0.001	0.027	Concordant

Table A4e: Global Comparisons for MCF7 - FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	0.667	2.134	0	0.000	Concordant
MCF7	AVERAGE	Platelet activation	04611	33	0.575	2.009	0	0.000	Concordant
MCF7	AVERAGE	Acute myeloid leukemia	05221	112	0.458	1.972	0	0.000	Concordant
MCF7	AVERAGE	Chemokine signaling pathway	04062	119	0.437	1.967	0	0.000	Concordant
MCF7	AVERAGE	mTOR signaling pathway	04150	106	0.377	1.647	0	0.000	Concordant
MCF7	AVERAGE	Neurotrophin signaling pathway	04722	135	0.349	1.588	0	0.000	Concordant
MCF7	AVERAGE	Choline metabolism in cancer	05231	50	0.464	1.800	0.001	0.020	Concordant
MCF7	AVERAGE	VEGF signaling pathway	04370	66	0.443	1.768	0.001	0.020	Concordant
MCF7	AVERAGE	B cell receptor signaling pathway	04662	60	0.410	1.642	0.002	0.032	Concordant
MCF7	AVERAGE	Prostate cancer	05215	150	0.330	1.530	0.002	0.032	Concordant
MCF7	AVERAGE	Longevity regulating pathway	04211	60	0.427	1.678	0.003	0.040	Concordant
MCF7	AVERAGE	Colorectal cancer	05210	106	0.385	1.646	0.004	0.040	Concordant
MCF7	AVERAGE	Dopaminergic synapse	04728	54	-0.411	-1.613	0.004	0.040	Discordant
MCF7	AVERAGE	C-type lectin receptor signaling pathway	04625	170	0.373	1.546	0.003	0.040	Concordant
MCF7	AVERAGE	Estrogen signaling pathway	04915	111	0.353	1.477	0.004	0.040	Concordant
MCF7	AVERAGE	Proteoglycans in cancer	05205	221	0.311	1.435	0.004	0.040	Concordant
MCF7	AVERAGE	GnRH secretion	04929	34	0.475	1.638	0.005	0.045	Concordant
MCF7	AVERAGE	HIF-1 signaling pathway	04066	163	0.303	1.416	0.005	0.045	Concordant
MCF7	AVERAGE	Fluid shear stress and atherosclerosis	05418	109	0.337	1.488	0.006	0.049	Concordant
MCF7	AVERAGE	Signaling pathways regulating pluripotency of stem cells	04550	170	0.332	1.482	0.006	0.049	Concordant
MCF7	AVERAGE	Chagas disease (American trypanosomiasis)	05142	79	0.348	1.466	0.007	0.054	Concordant
MCF7	AVERAGE	Thyroid cancer	05216	42	-0.465	-1.579	0.008	0.056	Discordant
MCF7	AVERAGE	ErbB signaling pathway	04012	124	0.354	1.458	0.008	0.056	Concordant
MCF7	AVERAGE	Endocrine resistance	01522	177	0.323	1.450	0.009	0.061	Concordant
MCF7	AVERAGE	Spinocerebellar ataxia	05017	21	0.501	1.566	0.01	0.062	Concordant
MCF7	AVERAGE	Melanogenesis	04916	71	0.353	1.458	0.01	0.062	Concordant
MCF7	AVERAGE	Gap junction	04540	37	0.423	1.522	0.011	0.066	Concordant
MCF7	AVERAGE	Cholinergic synapse	04725	44	0.388	1.464	0.016	0.092	Concordant
MCF7	AVERAGE	Small cell lung cancer	05222	65	0.350	1.421	0.017	0.092	Concordant
MCF7	AVERAGE	Human cytomegalovirus infection	05163	208	0.282	1.307	0.017	0.092	Concordant
MCF7	AVERAGE	Thermogenesis	04714	45	0.383	1.441	0.018	0.094	Concordant
MCF7	AVERAGE	Cell cycle	04110	110	-0.304	-1.347	0.019	0.096	Discordant
MCF7	AVERAGE	Tuberculosis	05152	88	-0.307	-1.336	0.02	0.098	Discordant
MCF7	AVERAGE	Apoptosis - multiple species	04215	17	0.530	1.521	0.021	0.099	Concordant
MCF7	AVERAGE	Prolactin signaling pathway	04917	83	0.342	1.410	0.022	0.099	Concordant
MCF7	AVERAGE	Central carbon metabolism in cancer	05230	96	0.310	1.343	0.022	0.099	Concordant

Table A4f: Global Comparisons for PC3 - FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
PC3	AVERAGE	Serotonergic synapse	04726	26	-0.625	-2.077	0	0.000	Discordant
PC3	AVERAGE	Long-term depression	04730	38	-0.569	-1.997	0	0.000	Discordant
PC3	AVERAGE	Long-term potentiation	04720	35	-0.536	-1.897	0	0.000	Discordant
PC3	AVERAGE	Inflammatory mediator regulation of TRP channels	04750	29	0.555	1.865	0	0.000	Concordant
PC3	AVERAGE	Melanoma	05218	115	-0.381	-1.691	0	0.000	Discordant
PC3	AVERAGE	Hippo signaling pathway	04390	125	0.371	1.661	0	0.000	Concordant
PC3	AVERAGE	Rap1 signaling pathway	04015	209	-0.329	-1.621	0	0.000	Discordant
PC3	AVERAGE	MAPK signaling pathway	04010	353	-0.267	-1.398	0	0.000	Discordant
PC3	AVERAGE	Estrogen signaling pathway	04915	111	-0.434	-1.828	0.001	0.016	Discordant
PC3	AVERAGE	Regulation of actin cytoskeleton	04810	139	-0.362	-1.654	0.001	0.016	Discordant
PC3	AVERAGE	Thermogenesis	04714	45	0.444	1.679	0.003	0.037	Concordant
PC3	AVERAGE	Central carbon metabolism in cancer	05230	96	-0.357	-1.556	0.003	0.037	Discordant
PC3	AVERAGE	Endocrine resistance	01522	177	-0.336	-1.548	0.003	0.037	Discordant
PC3	AVERAGE	GnRH secretion	04929	34	-0.481	-1.653	0.004	0.046	Discordant
PC3	AVERAGE	cGMP-PKG signaling pathway	04022	27	-0.479	-1.597	0.006	0.065	Discordant
PC3	AVERAGE	Th1 and Th2 cell differentiation	04658	87	0.352	1.471	0.007	0.067	Concordant
PC3	AVERAGE	Ras signaling pathway	04014	256	-0.269	-1.324	0.007	0.067	Discordant
PC3	AVERAGE	Bladder cancer	05219	33	-0.471	-1.588	0.009	0.073	Discordant
PC3	AVERAGE	Natural killer cell mediated cytotoxicity	04650	68	-0.377	-1.520	0.009	0.073	Discordant
PC3	AVERAGE	Acute myeloid leukemia	05221	112	-0.319	-1.396	0.009	0.073	Discordant

Table A5a: Local Comparisons A375 (1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A375	A549	Colorectal cancer	05210	106	0.407	1.815	0	0.000	Concordant
A375	A549	Pertussis	05133	31	0.463	1.679	0.001	0.054	Concordant
A375	A549	PI3K-Akt signaling pathway	04151	417	-0.251	-1.374	0.001	0.054	Discordant
A375	A549	Toll-like receptor signaling pathway	04620	70	0.412	1.741	0.002	0.069	Concordant
A375	A549	Renal cell carcinoma	05211	40	0.458	1.678	0.003	0.069	Concordant
A375	A549	Sphingolipid signaling pathway	04071	87	-0.354	-1.560	0.003	0.069	Discordant
A375	A549	Yersinia infection	05135	92	0.344	1.507	0.003	0.069	Concordant
A375	A549	Human T-cell leukemia virus 1 infection	05166	119	-0.324	-1.518	0.004	0.081	Discordant
A375	A549	Endometrial cancer	05213	71	0.381	1.558	0.006	0.097	Concordant
A375	HA1E	Renal cell carcinoma	05211	40	0.545	1.991	0	0.000	Concordant
A375	HA1E	Toll-like receptor signaling pathway	04620	70	0.472	1.975	0	0.000	Concordant
A375	HA1E	Colorectal cancer	05210	106	0.435	1.885	0	0.000	Concordant
A375	HA1E	VEGF signaling pathway	04370	66	0.429	1.745	0	0.000	Concordant
A375	HA1E	AGE-RAGE signaling pathway in diabetic complications	04933	139	0.349	1.658	0	0.000	Concordant
A375	HA1E	Endocrine resistance	01522	177	0.348	1.641	0	0.000	Concordant
A375	HA1E	Hepatitis B	05161	154	0.392	1.869	0.001	0.020	Concordant
A375	HA1E	GnRH signaling pathway	04912	59	0.418	1.710	0.001	0.020	Concordant
A375	HA1E	Yersinia infection	05135	92	0.382	1.664	0.002	0.027	Concordant
A375	HA1E	Osteoclast differentiation	04380	98	0.359	1.609	0.002	0.027	Concordant
A375	HA1E	Melanogenesis	04916	71	-0.362	-1.564	0.002	0.027	Discordant
A375	HA1E	Salmonella infection	05132	124	0.327	1.534	0.002	0.027	Concordant
A375	HA1E	PI3K-Akt signaling pathway	04151	417	-0.234	-1.322	0.004	0.050	Discordant
A375	HA1E	TNF signaling pathway	04668	66	0.401	1.593	0.007	0.054	Concordant
A375	HA1E	B cell receptor signaling pathway	04662	60	0.378	1.567	0.006	0.054	Concordant
A375	HA1E	Acute myeloid leukemia	05221	112	0.350	1.555	0.007	0.054	Concordant
A375	HA1E	Cushing syndrome	04934	82	-0.354	-1.536	0.006	0.054	Discordant
A375	HA1E	Th1 and Th2 cell differentiation	04658	87	0.347	1.524	0.006	0.054	Concordant
A375	HA1E	Basal cell carcinoma	05217	66	-0.352	-1.508	0.005	0.054	Discordant
A375	HA1E	Endometrial cancer	05213	71	0.370	1.498	0.006	0.054	Concordant
A375	HA1E	Human immunodeficiency virus 1 infection	05170	111	0.322	1.475	0.007	0.054	Concordant
A375	HA1E	Platelet activation	04611	33	0.422	1.529	0.009	0.063	Concordant
A375	HA1E	Necroptosis	04217	81	0.336	1.477	0.009	0.063	Concordant
A375	HA1E	Progesterone-mediated oocyte maturation	04914	34	0.412	1.510	0.013	0.066	Concordant
A375	HA1E	Toxoplasmosis	05145	48	0.383	1.510	0.012	0.066	Concordant
A375	HA1E	Chagas disease (American trypanosomiasis)	05142	79	0.343	1.465	0.013	0.066	Concordant
A375	HA1E	Pathogenic Escherichia coli infection	05130	88	0.327	1.449	0.013	0.066	Concordant
A375	HA1E	Growth hormone synthesis, secretion and action	04935	116	0.308	1.425	0.013	0.066	Concordant
A375	HA1E	Prostate cancer	05215	150	0.290	1.409	0.012	0.066	Concordant
A375	HA1E	Shigellosis	05131	118	0.299	1.361	0.012	0.066	Concordant
A375	HA1E	Neurotrophin signaling pathway	04722	135	0.286	1.351	0.012	0.066	Concordant

Table A5a: Local Comparisons A375 (41 - 80): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A375	HA1E	MAPK signaling pathway	04010	353	0.240	1.288	0.011	0.066	Concordant
A375	HA1E	ErbB signaling pathway	04012	124	0.339	1.452	0.014	0.067	Concordant
A375	HA1E	Sphingolipid signaling pathway	04071	87	-0.311	-1.393	0.014	0.067	Discordant
A375	HA1E	Spinocerebellar ataxia	05017	21	0.489	1.565	0.016	0.074	Concordant
A375	HA1E	T cell receptor signaling pathway	04660	86	0.306	1.356	0.018	0.081	Concordant
A375	HA1E	Prion diseases	05020	19	0.768	1.572	0.022	0.091	Concordant
A375	HA1E	Inflammatory mediator regulation of TRP channels	04750	29	0.425	1.517	0.022	0.091	Concordant
A375	HA1E	Rap1 signaling pathway	04015	209	0.254	1.284	0.022	0.091	Concordant
A375	HA1E	C-type lectin receptor signaling pathway	04625	170	0.317	1.374	0.024	0.096	Concordant
A375	HA1E	Ras signaling pathway	04014	256	0.250	1.269	0.025	0.096	Concordant
A375	HA1E	Non-alcoholic fatty liver disease (NAFLD)	04932	46	0.371	1.410	0.028	0.099	Concordant
A375	HEPG2	Longevity regulating pathway - multiple species	04213	90	-0.512	-1.983	0	0.000	Discordant
A375	HEPG2	Th1 and Th2 cell differentiation	04658	87	0.420	1.890	0	0.000	Concordant
A375	HEPG2	Longevity regulating pathway	04211	60	-0.440	-1.784	0.001	0.023	Discordant
A375	HEPG2	Pertussis	05133	31	0.495	1.761	0.001	0.023	Concordant
A375	HEPG2	Antigen processing and presentation	04612	77	-0.476	-1.728	0.001	0.023	Discordant
A375	HEPG2	Human T-cell leukemia virus 1 infection	05166	119	-0.345	-1.565	0.001	0.023	Discordant
A375	HEPG2	HIF-1 signaling pathway	04066	163	-0.327	-1.560	0.001	0.023	Discordant
A375	HEPG2	Pathogenic Escherichia coli infection	05130	88	0.336	1.489	0.002	0.036	Concordant
A375	HEPG2	AGE-RAGE signaling pathway in diabetic complications	04933	139	0.306	1.466	0.002	0.036	Concordant
A375	HEPG2	Toll-like receptor signaling pathway	04620	70	0.400	1.694	0.003	0.049	Concordant
A375	MCF7	Longevity regulating pathway - multiple species	04213	90	-0.489	-1.867	0	0.000	Discordant
A375	MCF7	Longevity regulating pathway	04211	60	-0.508	-2.040	0.001	0.040	Discordant
A375	MCF7	mTOR signaling pathway	04150	106	-0.377	-1.656	0.001	0.040	Discordant
A375	MCF7	HIF-1 signaling pathway	04066	163	-0.320	-1.500	0.001	0.040	Discordant
A375	MCF7	Thyroid cancer	05216	42	0.508	1.780	0.002	0.065	Concordant
A375	MCF7	GnRH signaling pathway	04912	59	0.366	1.519	0.004	0.093	Concordant
A375	PC3	Long-term depression	04730	38	0.607	2.214	0	0.000	Concordant
A375	PC3	Long-term potentiation	04720	35	0.528	1.976	0	0.000	Concordant
A375	PC3	Toll-like receptor signaling pathway	04620	70	0.419	1.788	0	0.000	Concordant
A375	PC3	GnRH signaling pathway	04912	59	0.400	1.642	0.001	0.027	Concordant
A375	PC3	Regulation of actin cytoskeleton	04810	139	0.326	1.574	0.001	0.027	Concordant
A375	PC3	Endocrine resistance	01522	177	0.312	1.495	0.001	0.027	Concordant
A375	PC3	Bladder cancer	05219	33	0.517	1.827	0.002	0.036	Concordant
A375	PC3	Natural killer cell mediated cytotoxicity	04650	68	0.383	1.605	0.002	0.036	Concordant
A375	PC3	MAPK signaling pathway	04010	353	0.241	1.326	0.002	0.036	Concordant
A375	PC3	Basal cell carcinoma	05217	66	-0.389	-1.575	0.003	0.049	Discordant
A375	PC3	Longevity regulating pathway	04211	60	-0.399	-1.594	0.004	0.054	Discordant
A375	PC3	Estrogen signaling pathway	04915	111	0.352	1.530	0.005	0.054	Concordant
A375	PC3	Autophagy - animal	04140	71	-0.368	-1.505	0.005	0.054	Discordant

Table A5a: Local Comparisons A375 (81 - 88): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A375	PC3	Rap1 signaling pathway	04015	209	0.256	1.340	0.005	0.054	Concordant
A375	PC3	PI3K-Akt signaling pathway	04151	417	-0.253	-1.325	0.004	0.054	Discordant
A375	PC3	Melanogenesis	04916	71	-0.369	-1.542	0.006	0.061	Discordant
A375	PC3	Oxytocin signaling pathway	04921	35	0.425	1.565	0.007	0.067	Concordant
A375	PC3	Colorectal cancer	05210	106	0.314	1.407	0.008	0.068	Concordant
A375	PC3	Melanoma	05218	115	0.296	1.386	0.008	0.068	Concordant
A375	PC3	Serotonergic synapse	04726	26	0.451	1.546	0.012	0.093	Concordant
A375	PC3	Hippo signaling pathway	04390	125	-0.312	-1.409	0.012	0.093	Discordant

Table A5b: Local Comparisons A549 (1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A549	A375	Colorectal cancer	05210	106	-0.410	-1.811	0	0.000	Discordant
A549	A375	Toll-like receptor signaling pathway	04620	70	-0.413	-1.754	0	0.000	Discordant
A549	A375	Human T-cell leukemia virus 1 infection	05166	119	0.324	1.506	0	0.000	Concordant
A549	A375	Pertussis	05133	31	-0.463	-1.653	0.003	0.069	Discordant
A549	A375	Endometrial cancer	05213	71	-0.382	-1.615	0.003	0.069	Discordant
A549	A375	Yersinia infection	05135	92	-0.344	-1.523	0.002	0.069	Discordant
A549	A375	PI3K-Akt signaling pathway	04151	417	0.251	1.369	0.003	0.069	Concordant
A549	A375	Renal cell carcinoma	05211	40	-0.458	-1.695	0.004	0.072	Discordant
A549	A375	Sphingolipid signaling pathway	04071	87	0.354	1.550	0.004	0.072	Concordant
A549	HA1E	Osteoclast differentiation	04380	98	0.384	1.742	0	0.000	Concordant
A549	HA1E	GnRH signaling pathway	04912	59	0.421	1.727	0	0.000	Concordant
A549	HA1E	Hepatitis B	05161	154	0.349	1.666	0	0.000	Concordant
A549	HA1E	Apelin signaling pathway	04371	85	-0.378	-1.727	0.001	0.032	Discordant
A549	HA1E	Hippo signaling pathway	04390	125	-0.328	-1.570	0.001	0.032	Discordant
A549	HEPG2	Longevity regulating pathway	04211	60	-0.458	-1.888	0	0.000	Discordant
A549	HEPG2	Longevity regulating pathway - multiple species	04213	90	-0.463	-1.839	0	0.000	Discordant
A549	HT29	Progesterone-mediated oocyte maturation	04914	34	-0.530	-1.926	0	0.000	Discordant
A549	HT29	Melanoma	05218	115	-0.406	-1.817	0	0.000	Discordant
A549	HT29	Non-alcoholic fatty liver disease (NAFLD)	04932	46	-0.445	-1.696	0.002	0.072	Discordant
A549	HT29	Longevity regulating pathway - multiple species	04213	90	-0.428	-1.653	0.004	0.072	Discordant
A549	HT29	Colorectal cancer	05210	106	-0.368	-1.624	0.003	0.072	Discordant
A549	HT29	VEGF signaling pathway	04370	66	-0.389	-1.589	0.004	0.072	Discordant
A549	HT29	Chemokine signaling pathway	04062	119	-0.332	-1.535	0.002	0.072	Discordant
A549	HT29	GnRH signaling pathway	04912	59	0.391	1.633	0.005	0.074	Concordant
A549	HT29	Prostate cancer	05215	150	-0.322	-1.514	0.005	0.074	Discordant
A549	MCF7	Platelet activation	04611	33	-0.552	-2.003	0	0.000	Discordant
A549	MCF7	B cell receptor signaling pathway	04662	60	-0.450	-1.858	0	0.000	Discordant
A549	MCF7	VEGF signaling pathway	04370	66	-0.442	-1.818	0	0.000	Discordant
A549	MCF7	Chemokine signaling pathway	04062	119	-0.369	-1.730	0	0.000	Discordant
A549	MCF7	Colorectal cancer	05210	106	-0.378	-1.645	0	0.000	Discordant
A549	MCF7	Dopaminergic synapse	04728	54	0.443	1.831	0.001	0.016	Concordant
A549	MCF7	Spinocerebellar ataxia	05017	21	-0.545	-1.734	0.001	0.016	Discordant
A549	MCF7	Acute myeloid leukemia	05221	112	-0.379	-1.648	0.001	0.016	Discordant
A549	MCF7	C-type lectin receptor signaling pathway	04625	170	-0.379	-1.611	0.001	0.016	Discordant
A549	MCF7	JAK-STAT signaling pathway	04630	561	-0.288	-1.555	0.001	0.016	Discordant
A549	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.568	-1.884	0.002	0.027	Discordant
A549	MCF7	Longevity regulating pathway	04211	60	-0.414	-1.685	0.002	0.027	Discordant
A549	MCF7	T cell receptor signaling pathway	04660	86	-0.377	-1.655	0.003	0.032	Discordant
A549	MCF7	cAMP signaling pathway	04024	77	-0.365	-1.592	0.003	0.032	Discordant
A549	MCF7	Fluid shear stress and atherosclerosis	05418	109	-0.327	-1.494	0.003	0.032	Discordant

Table A5b: Local Comparisons A549 (41 - 50): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A549	MCF7	Focal adhesion	04510	208	-0.277	-1.409	0.004	0.040	Discordant
A549	MCF7	GnRH secretion	04929	34	-0.485	-1.703	0.006	0.057	Discordant
A549	MCF7	Toll-like receptor signaling pathway	04620	70	-0.367	-1.540	0.007	0.057	Discordant
A549	MCF7	Prostate cancer	05215	150	-0.299	-1.443	0.007	0.057	Discordant
A549	MCF7	Neurotrophin signaling pathway	04722	135	-0.306	-1.435	0.007	0.057	Discordant
A549	MCF7	Longevity regulating pathway - multiple species	04213	90	-0.384	-1.510	0.01	0.077	Discordant
A549	PC3	Long-term depression	04730	38	0.535	1.963	0	0.000	Concordant
A549	PC3	Long-term potentiation	04720	35	0.519	1.926	0.001	0.040	Concordant
A549	PC3	Hippo signaling pathway	04390	125	-0.353	-1.624	0.001	0.040	Discordant
A549	PC3	MAPK signaling pathway	04010	353	0.254	1.410	0.001	0.040	Concordant

Table A5c: Local Comparisons HA1E (1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HA1E	A375	Renal cell carcinoma	05211	40	-0.545	-1.954	0	0.000	Discordant
HA1E	A375	Toll-like receptor signaling pathway	04620	70	-0.472	-1.948	0	0.000	Discordant
HA1E	A375	Colorectal cancer	05210	106	-0.438	-1.915	0	0.000	Discordant
HA1E	A375	Hepatitis B	05161	154	-0.394	-1.883	0	0.000	Discordant
HA1E	A375	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.349	-1.678	0	0.000	Discordant
HA1E	A375	Yersinia infection	05135	92	-0.383	-1.676	0	0.000	Discordant
HA1E	A375	Osteoclast differentiation	04380	98	-0.359	-1.624	0	0.000	Discordant
HA1E	A375	VEGF signaling pathway	04370	66	-0.429	-1.759	0.001	0.016	Discordant
HA1E	A375	Endocrine resistance	01522	177	-0.355	-1.659	0.001	0.016	Discordant
HA1E	A375	Salmonella infection	05132	124	-0.327	-1.535	0.001	0.016	Discordant
HA1E	A375	GnRH signaling pathway	04912	59	-0.418	-1.688	0.002	0.023	Discordant
HA1E	A375	Melanogenesis	04916	71	0.362	1.598	0.002	0.023	Concordant
HA1E	A375	Acute myeloid leukemia	05221	112	-0.353	-1.567	0.002	0.023	Discordant
HA1E	A375	Cushing syndrome	04934	82	0.352	1.546	0.002	0.023	Concordant
HA1E	A375	B cell receptor signaling pathway	04662	60	-0.378	-1.571	0.004	0.036	Discordant
HA1E	A375	Th1 and Th2 cell differentiation	04658	87	-0.348	-1.540	0.004	0.036	Discordant
HA1E	A375	Endometrial cancer	05213	71	-0.371	-1.515	0.004	0.036	Discordant
HA1E	A375	PI3K-Akt signaling pathway	04151	417	0.234	1.312	0.004	0.036	Concordant
HA1E	A375	TNF signaling pathway	04668	66	-0.403	-1.598	0.007	0.054	Discordant
HA1E	A375	Pathogenic Escherichia coli infection	05130	88	-0.327	-1.457	0.007	0.054	Discordant
HA1E	A375	MAPK signaling pathway	04010	353	-0.240	-1.290	0.007	0.054	Discordant
HA1E	A375	Basal cell carcinoma	05217	66	0.352	1.527	0.009	0.056	Concordant
HA1E	A375	ErbB signaling pathway	04012	124	-0.344	-1.476	0.009	0.056	Discordant
HA1E	A375	Human immunodeficiency virus 1 infection	05170	111	-0.322	-1.460	0.009	0.056	Discordant
HA1E	A375	Growth hormone synthesis, secretion and action	04935	116	-0.308	-1.430	0.008	0.056	Discordant
HA1E	A375	Prostate cancer	05215	150	-0.291	-1.408	0.008	0.056	Discordant
HA1E	A375	Platelet activation	04611	33	-0.422	-1.518	0.011	0.066	Discordant
HA1E	A375	Toxoplasmosis	05145	48	-0.383	-1.502	0.013	0.070	Discordant
HA1E	A375	Chagas disease (American trypanosomiasis)	05142	79	-0.344	-1.465	0.013	0.070	Discordant
HA1E	A375	Necroptosis	04217	81	-0.336	-1.464	0.013	0.070	Discordant
HA1E	A375	Sphingolipid signaling pathway	04071	87	0.311	1.407	0.014	0.073	Concordant
HA1E	A375	Progesterone-mediated oocyte maturation	04914	34	-0.412	-1.483	0.015	0.074	Discordant
HA1E	A375	Ras signaling pathway	04014	256	-0.254	-1.289	0.015	0.074	Discordant
HA1E	A375	Spinocerebellar ataxia	05017	21	-0.489	-1.561	0.016	0.076	Discordant
HA1E	A375	Prion diseases	05020	19	-0.770	-1.590	0.017	0.079	Discordant
HA1E	A375	Neurotrophin signaling pathway	04722	135	-0.287	-1.354	0.018	0.081	Discordant
HA1E	A375	Inflammatory mediator regulation of TRP channels	04750	29	-0.425	-1.476	0.021	0.085	Discordant
HA1E	A375	T cell receptor signaling pathway	04660	86	-0.306	-1.365	0.021	0.085	Discordant
HA1E	A375	Shigellosis	05131	118	-0.300	-1.359	0.02	0.085	Discordant
HA1E	A375	Rap1 signaling pathway	04015	209	-0.254	-1.300	0.02	0.085	Discordant

Table A5c: Local Comparisons HA1E (41 - 80): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HA1E	A375	C-type lectin receptor signaling pathway	04625	170	-0.329	-1.413	0.023	0.091	Discordant
HA1E	A375	Non-alcoholic fatty liver disease (NAFLD)	04932	46	-0.371	-1.406	0.024	0.093	Discordant
HA1E	A549	Osteoclast differentiation	04380	98	-0.384	-1.737	0	0.000	Discordant
HA1E	A549	Hepatitis B	05161	154	-0.351	-1.662	0	0.000	Discordant
HA1E	A549	Hippo signaling pathway	04390	125	0.328	1.562	0	0.000	Concordant
HA1E	A549	Apelin signaling pathway	04371	85	0.378	1.692	0.001	0.040	Concordant
HA1E	A549	GnRH signaling pathway	04912	59	-0.421	-1.714	0.002	0.054	Discordant
HA1E	HEPG2	Longevity regulating pathway	04211	60	-0.454	-1.924	0	0.000	Discordant
HA1E	HEPG2	Adherens junction	04520	34	0.478	1.769	0.002	0.081	Concordant
HA1E	HEPG2	VEGF signaling pathway	04370	66	-0.405	-1.684	0.002	0.081	Discordant
HA1E	HEPG2	Osteoclast differentiation	04380	98	-0.343	-1.586	0.001	0.081	Discordant
HA1E	HEPG2	HIF-1 signaling pathway	04066	163	-0.309	-1.546	0.003	0.081	Discordant
HA1E	HT29	VEGF signaling pathway	04370	66	-0.470	-1.958	0	0.000	Discordant
HA1E	HT29	Longevity regulating pathway	04211	60	-0.448	-1.826	0	0.000	Discordant
HA1E	HT29	Osteoclast differentiation	04380	98	-0.408	-1.820	0	0.000	Discordant
HA1E	HT29	Melanoma	05218	115	-0.401	-1.793	0.002	0.065	Discordant
HA1E	HT29	Long-term potentiation	04720	35	0.424	1.649	0.002	0.065	Concordant
HA1E	HT29	Longevity regulating pathway - multiple species	04213	90	-0.420	-1.653	0.007	0.076	Discordant
HA1E	HT29	Retrograde endocannabinoid signaling	04723	104	-0.411	-1.608	0.006	0.076	Discordant
HA1E	HT29	TNF signaling pathway	04668	66	-0.407	-1.596	0.007	0.076	Discordant
HA1E	HT29	Salmonella infection	05132	124	-0.334	-1.561	0.005	0.076	Discordant
HA1E	HT29	Acute myeloid leukemia	05221	112	-0.346	-1.516	0.008	0.076	Discordant
HA1E	HT29	Hepatitis B	05161	154	-0.317	-1.512	0.006	0.076	Discordant
HA1E	HT29	Neurotrophin signaling pathway	04722	135	-0.316	-1.492	0.005	0.076	Discordant
HA1E	HT29	Focal adhesion	04510	208	-0.289	-1.453	0.004	0.076	Discordant
HA1E	HT29	Prostate cancer	05215	150	-0.308	-1.453	0.007	0.076	Discordant
HA1E	HT29	HIF-1 signaling pathway	04066	163	-0.291	-1.413	0.008	0.076	Discordant
HA1E	HT29	Platinum drug resistance	01524	49	-0.410	-1.586	0.01	0.085	Discordant
HA1E	HT29	Endometrial cancer	05213	71	-0.388	-1.572	0.01	0.085	Discordant
HA1E	HT29	Apelin signaling pathway	04371	85	0.292	1.376	0.013	0.100	Concordant
HA1E	MCF7	VEGF signaling pathway	04370	66	-0.557	-2.307	0	0.000	Discordant
HA1E	MCF7	Platelet activation	04611	33	-0.600	-2.164	0	0.000	Discordant
HA1E	MCF7	Acute myeloid leukemia	05221	112	-0.442	-1.984	0	0.000	Discordant
HA1E	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.598	-1.981	0	0.000	Discordant
HA1E	MCF7	Chemokine signaling pathway	04062	119	-0.383	-1.807	0	0.000	Discordant
HA1E	MCF7	Longevity regulating pathway	04211	60	-0.415	-1.753	0	0.000	Discordant
HA1E	MCF7	Osteoclast differentiation	04380	98	-0.387	-1.743	0	0.000	Discordant
HA1E	MCF7	Fc epsilon RI signaling pathway	04664	53	-0.397	-1.615	0	0.000	Discordant
HA1E	MCF7	Neurotrophin signaling pathway	04722	135	-0.339	-1.607	0	0.000	Discordant
HA1E	MCF7	Spinocerebellar ataxia	05017	21	-0.547	-1.794	0.001	0.012	Discordant

Table A5c: Local Comparisons HA1E (81 - 118): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HA1E	MCF7	Adherens junction	04520	34	0.488	1.777	0.001	0.012	Concordant
HA1E	MCF7	Endocrine resistance	01522	177	-0.357	-1.675	0.001	0.012	Discordant
HA1E	MCF7	Fluid shear stress and atherosclerosis	05418	109	-0.344	-1.600	0.001	0.012	Discordant
HA1E	MCF7	HIF-1 signaling pathway	04066	163	-0.328	-1.594	0.001	0.012	Discordant
HA1E	MCF7	GnRH secretion	04929	34	-0.505	-1.788	0.002	0.020	Discordant
HA1E	MCF7	Colorectal cancer	05210	106	-0.377	-1.656	0.002	0.020	Discordant
HA1E	MCF7	C-type lectin receptor signaling pathway	04625	170	-0.362	-1.557	0.003	0.029	Discordant
HA1E	MCF7	B cell receptor signaling pathway	04662	60	-0.402	-1.661	0.004	0.032	Discordant
HA1E	MCF7	Endometrial cancer	05213	71	-0.396	-1.633	0.004	0.032	Discordant
HA1E	MCF7	ErbB signaling pathway	04012	124	-0.370	-1.600	0.004	0.032	Discordant
HA1E	MCF7	Choline metabolism in cancer	05231	50	-0.428	-1.721	0.005	0.035	Discordant
HA1E	MCF7	Hepatitis B	05161	154	-0.314	-1.508	0.005	0.035	Discordant
HA1E	MCF7	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.303	-1.434	0.005	0.035	Discordant
HA1E	MCF7	Prostate cancer	05215	150	-0.319	-1.522	0.006	0.040	Discordant
HA1E	MCF7	Cholinergic synapse	04725	44	-0.410	-1.586	0.007	0.044	Discordant
HA1E	MCF7	Tuberculosis	05152	88	0.306	1.454	0.007	0.044	Concordant
HA1E	MCF7	Cushing syndrome	04934	82	0.312	1.418	0.008	0.046	Concordant
HA1E	MCF7	Rap1 signaling pathway	04015	209	-0.276	-1.398	0.008	0.046	Discordant
HA1E	MCF7	Progesterone-mediated oocyte maturation	04914	34	-0.444	-1.589	0.01	0.049	Discordant
HA1E	MCF7	Thermogenesis	04714	45	-0.390	-1.517	0.01	0.049	Discordant
HA1E	MCF7	Salmonella infection	05132	124	-0.310	-1.461	0.01	0.049	Discordant
HA1E	MCF7	Sphingolipid signaling pathway	04071	87	0.296	1.382	0.01	0.049	Concordant
HA1E	MCF7	Bladder cancer	05219	33	0.424	1.531	0.012	0.056	Concordant
HA1E	MCF7	Long-term potentiation	04720	35	0.394	1.517	0.012	0.056	Concordant
HA1E	MCF7	Estrogen signaling pathway	04915	111	-0.352	-1.501	0.014	0.061	Discordant
HA1E	MCF7	Focal adhesion	04510	208	-0.271	-1.381	0.014	0.061	Discordant
HA1E	MCF7	Renal cell carcinoma	05211	40	-0.426	-1.524	0.018	0.077	Discordant
HA1E	MCF7	Small cell lung cancer	05222	65	-0.343	-1.435	0.019	0.077	Discordant
HA1E	MCF7	Apelin signaling pathway	04371	85	0.293	1.347	0.019	0.077	Concordant
HA1E	MCF7	Toll-like receptor signaling pathway	04620	70	-0.347	-1.454	0.021	0.083	Discordant
HA1E	PC3	Osteoclast differentiation	04380	98	-0.388	-1.786	0	0.000	Discordant
HA1E	PC3	Hepatitis B	05161	154	-0.346	-1.663	0	0.000	Discordant
HA1E	PC3	Salmonella infection	05132	124	-0.347	-1.626	0.001	0.054	Discordant
HA1E	PC3	Serotonergic synapse	04726	26	0.487	1.745	0.002	0.065	Concordant
HA1E	PC3	Long-term potentiation	04720	35	0.454	1.786	0.003	0.076	Concordant
HA1E	PC3	Inflammatory mediator regulation of TRP channels	04750	29	-0.481	-1.713	0.004	0.076	Discordant
HA1E	PC3	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.313	-1.523	0.005	0.076	Discordant
HA1E	PC3	Rap1 signaling pathway	04015	209	0.237	1.317	0.007	0.076	Concordant

Table A5d: Local Comparisons HEPG2 (1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HEPG2	A375	Longevity regulating pathway - multiple species	04213	90	0.507	1.955	0	0.000	Concordant
HEPG2	A375	Th1 and Th2 cell differentiation	04658	87	-0.421	-1.865	0	0.000	Discordant
HEPG2	A375	HIF-1 signaling pathway	04066	163	0.327	1.548	0	0.000	Concordant
HEPG2	A375	Pertussis	05133	31	-0.495	-1.788	0.001	0.040	Discordant
HEPG2	A375	Longevity regulating pathway	04211	60	0.440	1.762	0.002	0.054	Concordant
HEPG2	A375	Toll-like receptor signaling pathway	04620	70	-0.401	-1.665	0.002	0.054	Discordant
HEPG2	A375	Human T-cell leukemia virus 1 infection	05166	119	0.345	1.570	0.003	0.069	Concordant
HEPG2	A375	Antigen processing and presentation	04612	77	0.469	1.682	0.006	0.094	Concordant
HEPG2	A375	Pathogenic Escherichia coli infection	05130	88	-0.336	-1.478	0.007	0.094	Discordant
HEPG2	A375	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.306	-1.466	0.007	0.094	Discordant
HEPG2	A549	Longevity regulating pathway	04211	60	0.458	1.866	0	0.000	Concordant
HEPG2	A549	Longevity regulating pathway - multiple species	04213	90	0.458	1.829	0.001	0.081	Concordant
HEPG2	HA1E	Longevity regulating pathway	04211	60	0.454	1.903	0.001	0.081	Concordant
HEPG2	HA1E	VEGF signaling pathway	04370	66	0.404	1.689	0.001	0.081	Concordant
HEPG2	HA1E	Adherens junction	04520	34	-0.479	-1.756	0.003	0.097	Discordant
HEPG2	HA1E	Osteoclast differentiation	04380	98	0.343	1.579	0.002	0.097	Concordant
HEPG2	HA1E	HIF-1 signaling pathway	04066	163	0.308	1.529	0.003	0.097	Concordant
HEPG2	HT29	Pertussis	05133	31	-0.525	-1.828	0	0.000	Discordant
HEPG2	HT29	Melanoma	05218	115	-0.357	-1.578	0	0.000	Discordant
HEPG2	HT29	Antigen processing and presentation	04612	77	0.500	1.822	0.002	0.081	Concordant
HEPG2	MCF7	Platelet activation	04611	33	-0.525	-1.842	0	0.000	Discordant
HEPG2	MCF7	Acute myeloid leukemia	05221	112	-0.410	-1.762	0	0.000	Discordant
HEPG2	MCF7	Antigen processing and presentation	04612	77	0.475	1.738	0.002	0.054	Concordant
HEPG2	MCF7	Necroptosis	04217	81	-0.386	-1.646	0.002	0.054	Discordant
HEPG2	MCF7	Cholinergic synapse	04725	44	-0.422	-1.629	0.002	0.054	Discordant
HEPG2	MCF7	Thermogenesis	04714	45	-0.409	-1.567	0.003	0.054	Discordant
HEPG2	MCF7	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.328	-1.550	0.001	0.054	Discordant
HEPG2	MCF7	Signaling pathways regulating pluripotency of stem cells	04550	170	-0.337	-1.544	0.003	0.054	Discordant
HEPG2	MCF7	JAK-STAT signaling pathway	04630	561	-0.257	-1.365	0.003	0.054	Discordant
HEPG2	MCF7	Gap junction	04540	37	-0.465	-1.697	0.004	0.059	Discordant
HEPG2	MCF7	Chemokine signaling pathway	04062	119	-0.323	-1.482	0.004	0.059	Discordant
HEPG2	MCF7	B cell receptor signaling pathway	04662	60	-0.361	-1.446	0.005	0.062	Discordant
HEPG2	MCF7	Neurotrophin signaling pathway	04722	135	-0.304	-1.440	0.005	0.062	Discordant
HEPG2	MCF7	Small cell lung cancer	05222	65	-0.387	-1.604	0.006	0.069	Discordant
HEPG2	MCF7	Natural killer cell mediated cytotoxicity	04650	68	-0.385	-1.535	0.007	0.071	Discordant
HEPG2	MCF7	Colorectal cancer	05210	106	-0.344	-1.500	0.007	0.071	Discordant
HEPG2	MCF7	mTOR signaling pathway	04150	106	-0.327	-1.469	0.008	0.076	Discordant
HEPG2	MCF7	Fc epsilon RI signaling pathway	04664	53	-0.384	-1.540	0.009	0.077	Discordant
HEPG2	MCF7	Relaxin signaling pathway	04926	154	-0.338	-1.425	0.01	0.077	Discordant
HEPG2	MCF7	Fluid shear stress and atherosclerosis	05418	109	-0.310	-1.408	0.01	0.077	Discordant

Table A5d: Local Comparisons HEPG2 (41 - 66): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HEPG2	MCF7	Proteoglycans in cancer	05205	221	-0.295	-1.387	0.01	0.077	Discordant
HEPG2	MCF7	VEGF signaling pathway	04370	66	-0.362	-1.461	0.011	0.081	Discordant
HEPG2	MCF7	Mitophagy - animal	04137	17	0.521	1.575	0.015	0.097	Concordant
HEPG2	MCF7	Choline metabolism in cancer	05231	50	-0.376	-1.472	0.015	0.097	Discordant
HEPG2	PC3	Long-term potentiation	04720	35	0.508	1.819	0	0.000	Concordant
HEPG2	PC3	Rap1 signaling pathway	04015	209	0.292	1.449	0	0.000	Concordant
HEPG2	PC3	Thermogenesis	04714	45	-0.501	-1.888	0.001	0.023	Discordant
HEPG2	PC3	Serotonergic synapse	04726	26	0.544	1.794	0.001	0.023	Concordant
HEPG2	PC3	Spinocerebellar ataxia	05017	21	0.540	1.732	0.001	0.023	Concordant
HEPG2	PC3	Regulation of actin cytoskeleton	04810	139	0.338	1.594	0.001	0.023	Concordant
HEPG2	PC3	Hippo signaling pathway	04390	125	-0.344	-1.575	0.001	0.023	Discordant
HEPG2	PC3	Long-term depression	04730	38	0.488	1.801	0.002	0.040	Concordant
HEPG2	PC3	Longevity regulating pathway - multiple species	04213	90	0.433	1.659	0.004	0.050	Concordant
HEPG2	PC3	Estrogen signaling pathway	04915	111	0.372	1.577	0.004	0.050	Concordant
HEPG2	PC3	Thyroid hormone signaling pathway	04919	72	0.373	1.559	0.004	0.050	Concordant
HEPG2	PC3	Melanoma	05218	115	0.339	1.539	0.004	0.050	Concordant
HEPG2	PC3	Th1 and Th2 cell differentiation	04658	87	-0.350	-1.529	0.003	0.050	Discordant
HEPG2	PC3	Pathways in cancer	05200	554	-0.236	-1.274	0.005	0.058	Discordant
HEPG2	PC3	Fc epsilon RI signaling pathway	04664	53	-0.401	-1.579	0.007	0.076	Discordant
HEPG2	PC3	Amyotrophic lateral sclerosis (ALS)	05014	15	-0.566	-1.631	0.01	0.077	Discordant
HEPG2	PC3	Necroptosis	04217	81	-0.356	-1.540	0.009	0.077	Discordant
HEPG2	PC3	Longevity regulating pathway	04211	60	0.369	1.507	0.009	0.077	Concordant
HEPG2	PC3	TNF signaling pathway	04668	66	-0.385	-1.476	0.009	0.077	Discordant
HEPG2	PC3	AGE-RAGE signaling pathway in diabetic complications	04933	139	-0.305	-1.404	0.008	0.077	Discordant
HEPG2	PC3	Relaxin signaling pathway	04926	154	-0.355	-1.507	0.012	0.085	Discordant
HEPG2	PC3	Endocrine resistance	01522	177	0.286	1.346	0.013	0.088	Concordant

Table A5e: Local Comparisons HT29 (1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HT29	A549	Progesterone-mediated oocyte maturation	04914	34	0.530	1.913	0	0.000	Concordant
HT29	A549	Melanoma	05218	115	0.405	1.835	0	0.000	Concordant
HT29	A549	Non-alcoholic fatty liver disease (NAFLD)	04932	46	0.445	1.709	0	0.000	Concordant
HT29	A549	Chemokine signaling pathway	04062	119	0.332	1.538	0.001	0.032	Concordant
HT29	A549	Prostate cancer	05215	150	0.321	1.524	0.001	0.032	Concordant
HT29	A549	Longevity regulating pathway - multiple species	04213	90	0.422	1.649	0.002	0.046	Concordant
HT29	A549	Colorectal cancer	05210	106	0.366	1.580	0.002	0.046	Concordant
HT29	A549	GnRH signaling pathway	04912	59	-0.391	-1.661	0.003	0.054	Discordant
HT29	A549	VEGF signaling pathway	04370	66	0.389	1.570	0.004	0.065	Concordant
HT29	HA1E	VEGF signaling pathway	04370	66	0.469	1.928	0	0.000	Concordant
HT29	HA1E	Longevity regulating pathway	04211	60	0.448	1.829	0	0.000	Concordant
HT29	HA1E	Osteoclast differentiation	04380	98	0.408	1.843	0.001	0.032	Concordant
HT29	HA1E	Melanoma	05218	115	0.399	1.807	0.001	0.032	Concordant
HT29	HA1E	Salmonella infection	05132	124	0.334	1.562	0.001	0.032	Concordant
HT29	HA1E	Long-term potentiation	04720	35	-0.424	-1.654	0.003	0.069	Discordant
HT29	HA1E	Hepatitis B	05161	154	0.315	1.496	0.003	0.069	Concordant
HT29	HA1E	Longevity regulating pathway - multiple species	04213	90	0.415	1.639	0.007	0.086	Concordant
HT29	HA1E	Platinum drug resistance	01524	49	0.410	1.613	0.006	0.086	Concordant
HT29	HA1E	TNF signaling pathway	04668	66	0.405	1.609	0.007	0.086	Concordant
HT29	HA1E	Endometrial cancer	05213	71	0.387	1.584	0.007	0.086	Concordant
HT29	HA1E	Retrograde endocannabinoid signaling	04723	104	0.402	1.575	0.007	0.086	Concordant
HT29	HA1E	Neurotrophin signaling pathway	04722	135	0.315	1.476	0.008	0.086	Concordant
HT29	HA1E	Focal adhesion	04510	208	0.289	1.454	0.008	0.086	Concordant
HT29	HA1E	Prostate cancer	05215	150	0.307	1.452	0.009	0.086	Concordant
HT29	HA1E	HIF-1 signaling pathway	04066	163	0.290	1.419	0.009	0.086	Concordant
HT29	HA1E	Apelin signaling pathway	04371	85	-0.292	-1.398	0.009	0.086	Discordant
HT29	HA1E	Acute myeloid leukemia	05221	112	0.343	1.495	0.011	0.099	Concordant
HT29	HEPG2	Antigen processing and presentation	04612	77	-0.507	-1.875	0	0.000	Discordant
HT29	HEPG2	Melanoma	05218	115	0.356	1.558	0	0.000	Concordant
HT29	HEPG2	Pertussis	05133	31	0.525	1.835	0.001	0.054	Concordant
HT29	MCF7	Melanoma	05218	115	0.404	1.790	0	0.000	Concordant
HT29	PC3	Melanoma	05218	115	0.523	2.276	0	0.000	Concordant
HT29	PC3	Hippo signaling pathway	04390	125	-0.394	-1.799	0	0.000	Discordant
HT29	PC3	VEGF signaling pathway	04370	66	0.411	1.623	0.001	0.040	Concordant
HT29	PC3	Prostate cancer	05215	150	0.349	1.598	0.001	0.040	Concordant
HT29	PC3	Progesterone-mediated oocyte maturation	04914	34	0.488	1.713	0.003	0.072	Concordant
HT29	PC3	Acute myeloid leukemia	05221	112	0.359	1.567	0.004	0.072	Concordant
HT29	PC3	Chemokine signaling pathway	04062	119	0.338	1.535	0.004	0.072	Concordant
HT29	PC3	Central carbon metabolism in cancer	05230	96	0.349	1.485	0.003	0.072	Concordant
HT29	PC3	MAPK signaling pathway	04010	353	0.250	1.290	0.004	0.072	Concordant

Table A5e: Local Comparisons HT29 (41 - 49): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HT29	PC3	Serotonergic synapse	04726	26	0.509	1.724	0.005	0.074	Concordant
HT29	PC3	Insulin signaling pathway	04910	98	0.321	1.406	0.005	0.074	Concordant
HT29	PC3	Inflammatory mediator regulation of TRP channels	04750	29	-0.488	-1.663	0.007	0.081	Discordant
HT29	PC3	Non-small cell lung cancer	05223	82	0.387	1.572	0.008	0.081	Concordant
HT29	PC3	Thermogenesis	04714	45	-0.403	-1.554	0.008	0.081	Discordant
HT29	PC3	Regulation of actin cytoskeleton	04810	139	0.323	1.462	0.008	0.081	Concordant
HT29	PC3	GABAergic synapse	04727	32	0.567	1.675	0.01	0.085	Concordant
HT29	PC3	EGFR tyrosine kinase inhibitor resistance	01521	158	0.350	1.497	0.01	0.085	Concordant
HT29	PC3	Glioma	05214	177	0.342	1.454	0.01	0.085	Concordant

Table A5f: Local Comparisons MCF7(1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	A375	Longevity regulating pathway	04211	60	0.508	2.038	0	0.000	Concordant
MCF7	A375	Longevity regulating pathway - multiple species	04213	90	0.483	1.836	0	0.000	Concordant
MCF7	A375	HIF-1 signaling pathway	04066	163	0.319	1.506	0	0.000	Concordant
MCF7	A375	mTOR signaling pathway	04150	106	0.377	1.658	0.001	0.040	Concordant
MCF7	A375	Thyroid cancer	05216	42	-0.510	-1.758	0.002	0.065	Discordant
MCF7	A375	GnRH signaling pathway	04912	59	-0.366	-1.510	0.006	0.097	Discordant
MCF7	A549	Platelet activation	04611	33	0.552	1.998	0	0.000	Concordant
MCF7	A549	Chemokine signaling pathway	04062	119	0.369	1.724	0	0.000	Concordant
MCF7	A549	JAK-STAT signaling pathway	04630	561	0.265	1.428	0	0.000	Concordant
MCF7	A549	Regulation of lipolysis in adipocytes	04923	23	0.568	1.894	0.001	0.016	Concordant
MCF7	A549	B cell receptor signaling pathway	04662	60	0.450	1.860	0.001	0.016	Concordant
MCF7	A549	Dopaminergic synapse	04728	54	-0.443	-1.808	0.001	0.016	Discordant
MCF7	A549	VEGF signaling pathway	04370	66	0.441	1.794	0.001	0.016	Concordant
MCF7	A549	Longevity regulating pathway	04211	60	0.414	1.670	0.001	0.016	Concordant
MCF7	A549	Acute myeloid leukemia	05221	112	0.376	1.645	0.001	0.016	Concordant
MCF7	A549	Fluid shear stress and atherosclerosis	05418	109	0.327	1.499	0.001	0.016	Concordant
MCF7	A549	GnRH secretion	04929	34	0.485	1.712	0.002	0.025	Concordant
MCF7	A549	Colorectal cancer	05210	106	0.376	1.661	0.002	0.025	Concordant
MCF7	A549	T cell receptor signaling pathway	04660	86	0.377	1.657	0.002	0.025	Concordant
MCF7	A549	Spinocerebellar ataxia	05017	21	0.545	1.759	0.003	0.035	Concordant
MCF7	A549	cAMP signaling pathway	04024	77	0.365	1.562	0.005	0.054	Concordant
MCF7	A549	C-type lectin receptor signaling pathway	04625	170	0.366	1.548	0.006	0.057	Concordant
MCF7	A549	Focal adhesion	04510	208	0.277	1.384	0.006	0.057	Concordant
MCF7	A549	Toll-like receptor signaling pathway	04620	70	0.366	1.549	0.008	0.068	Concordant
MCF7	A549	Neurotrophin signaling pathway	04722	135	0.305	1.424	0.008	0.068	Concordant
MCF7	A549	Prostate cancer	05215	150	0.298	1.412	0.01	0.081	Concordant
MCF7	A549	Longevity regulating pathway - multiple species	04213	90	0.379	1.490	0.012	0.093	Concordant
MCF7	HA1E	VEGF signaling pathway	04370	66	0.556	2.285	0	0.000	Concordant
MCF7	HA1E	Platelet activation	04611	33	0.600	2.188	0	0.000	Concordant
MCF7	HA1E	Acute myeloid leukemia	05221	112	0.439	1.950	0	0.000	Concordant
MCF7	HA1E	Chemokine signaling pathway	04062	119	0.383	1.806	0	0.000	Concordant
MCF7	HA1E	Osteoclast differentiation	04380	98	0.387	1.743	0	0.000	Concordant
MCF7	HA1E	B cell receptor signaling pathway	04662	60	0.402	1.668	0	0.000	Concordant
MCF7	HA1E	Endocrine resistance	01522	177	0.350	1.668	0	0.000	Concordant
MCF7	HA1E	Neurotrophin signaling pathway	04722	135	0.338	1.617	0	0.000	Concordant
MCF7	HA1E	HIF-1 signaling pathway	04066	163	0.327	1.587	0	0.000	Concordant
MCF7	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.598	1.954	0.001	0.014	Concordant
MCF7	HA1E	Longevity regulating pathway	04211	60	0.415	1.709	0.001	0.014	Concordant
MCF7	HA1E	Colorectal cancer	05210	106	0.375	1.637	0.001	0.014	Concordant
MCF7	HA1E	GnRH secretion	04929	34	0.505	1.796	0.002	0.022	Concordant

Table A5f: Local Comparisons MCF7(41 - 80): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	HA1E	Choline metabolism in cancer	05231	50	0.428	1.717	0.002	0.022	Concordant
MCF7	HA1E	Prostate cancer	05215	150	0.318	1.523	0.002	0.022	Concordant
MCF7	HA1E	Spinocerebellar ataxia	05017	21	0.547	1.763	0.004	0.034	Concordant
MCF7	HA1E	Adherens junction	04520	34	-0.489	-1.733	0.004	0.034	Discordant
MCF7	HA1E	Fc epsilon RI signaling pathway	04664	53	0.397	1.593	0.004	0.034	Concordant
MCF7	HA1E	Fluid shear stress and atherosclerosis	05418	109	0.344	1.580	0.004	0.034	Concordant
MCF7	HA1E	Progesterone-mediated oocyte maturation	04914	34	0.444	1.628	0.005	0.037	Concordant
MCF7	HA1E	Endometrial cancer	05213	71	0.394	1.588	0.005	0.037	Concordant
MCF7	HA1E	Tuberculosis	05152	88	-0.306	-1.437	0.005	0.037	Discordant
MCF7	HA1E	C-type lectin receptor signaling pathway	04625	170	0.349	1.528	0.006	0.039	Concordant
MCF7	HA1E	Hepatitis B	05161	154	0.312	1.496	0.006	0.039	Concordant
MCF7	HA1E	Cushing syndrome	04934	82	-0.313	-1.442	0.006	0.039	Discordant
MCF7	HA1E	ErbB signaling pathway	04012	124	0.364	1.582	0.007	0.042	Concordant
MCF7	HA1E	Rap1 signaling pathway	04015	209	0.276	1.397	0.007	0.042	Concordant
MCF7	HA1E	Estrogen signaling pathway	04915	111	0.347	1.488	0.008	0.045	Concordant
MCF7	HA1E	Focal adhesion	04510	208	0.271	1.366	0.008	0.045	Concordant
MCF7	HA1E	Bladder cancer	05219	33	-0.424	-1.588	0.009	0.047	Discordant
MCF7	HA1E	Salmonella infection	05132	124	0.309	1.450	0.009	0.047	Concordant
MCF7	HA1E	AGE-RAGE signaling pathway in diabetic complications	04933	139	0.303	1.454	0.01	0.051	Concordant
MCF7	HA1E	Cholinergic synapse	04725	44	0.410	1.593	0.011	0.054	Concordant
MCF7	HA1E	Renal cell carcinoma	05211	40	0.426	1.554	0.013	0.060	Concordant
MCF7	HA1E	Thermogenesis	04714	45	0.390	1.521	0.013	0.060	Concordant
MCF7	HA1E	Long-term potentiation	04720	35	-0.394	-1.519	0.015	0.066	Discordant
MCF7	HA1E	Toll-like receptor signaling pathway	04620	70	0.347	1.451	0.015	0.066	Concordant
MCF7	HA1E	Sphingolipid signaling pathway	04071	87	-0.296	-1.376	0.017	0.072	Discordant
MCF7	HA1E	Apelin signaling pathway	04371	85	-0.293	-1.361	0.021	0.087	Discordant
MCF7	HA1E	Small cell lung cancer	05222	65	0.343	1.446	0.022	0.089	Concordant
MCF7	HEPG2	Antigen processing and presentation	04612	77	-0.482	-1.759	0	0.000	Discordant
MCF7	HEPG2	Acute myeloid leukemia	05221	112	0.407	1.756	0	0.000	Concordant
MCF7	HEPG2	Small cell lung cancer	05222	65	0.387	1.600	0	0.000	Concordant
MCF7	HEPG2	Platelet activation	04611	33	0.525	1.884	0.001	0.032	Concordant
MCF7	HEPG2	Gap junction	04540	37	0.465	1.734	0.001	0.032	Concordant
MCF7	HEPG2	Necroptosis	04217	81	0.386	1.675	0.002	0.040	Concordant
MCF7	HEPG2	Cholinergic synapse	04725	44	0.422	1.595	0.002	0.040	Concordant
MCF7	HEPG2	Chemokine signaling pathway	04062	119	0.323	1.512	0.002	0.040	Concordant
MCF7	HEPG2	AGE-RAGE signaling pathway in diabetic complications	04933	139	0.328	1.542	0.003	0.054	Concordant
MCF7	HEPG2	Fc epsilon RI signaling pathway	04664	53	0.384	1.517	0.005	0.081	Concordant
MCF7	HEPG2	Natural killer cell mediated cytotoxicity	04650	68	0.384	1.558	0.006	0.086	Concordant
MCF7	HEPG2	Thermogenesis	04714	45	0.409	1.551	0.007	0.086	Concordant
MCF7	HEPG2	mTOR signaling pathway	04150	106	0.327	1.478	0.007	0.086	Concordant

Table A5f: Local Comparisons MCF7(81 - 120): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	HEPG2	Signaling pathways regulating pluripotency of stem cells	04550	170	0.327	1.466	0.008	0.086	Concordant
MCF7	HEPG2	Neurotrophin signaling pathway	04722	135	0.303	1.415	0.008	0.086	Concordant
MCF7	HEPG2	Colorectal cancer	05210	106	0.342	1.474	0.01	0.095	Concordant
MCF7	HEPG2	B cell receptor signaling pathway	04662	60	0.361	1.466	0.01	0.095	Concordant
MCF7	HEPG2	Mitophagy - animal	04137	17	-0.521	-1.555	0.015	0.100	Discordant
MCF7	HEPG2	VEGF signaling pathway	04370	66	0.361	1.461	0.014	0.100	Concordant
MCF7	HEPG2	Choline metabolism in cancer	05231	50	0.376	1.454	0.016	0.100	Concordant
MCF7	HEPG2	Relaxin signaling pathway	04926	154	0.329	1.404	0.015	0.100	Concordant
MCF7	HEPG2	Fluid shear stress and atherosclerosis	05418	109	0.310	1.390	0.015	0.100	Concordant
MCF7	HEPG2	Proteoglycans in cancer	05205	221	0.285	1.349	0.012	0.100	Concordant
MCF7	HEPG2	JAK-STAT signaling pathway	04630	561	0.234	1.239	0.016	0.100	Concordant
MCF7	HT29	Melanoma	05218	115	-0.406	-1.788	0	0.000	Discordant
MCF7	PC3	GnRH secretion	04929	34	0.587	2.016	0	0.000	Concordant
MCF7	PC3	Serotonergic synapse	04726	26	0.613	1.982	0	0.000	Concordant
MCF7	PC3	Platelet activation	04611	33	0.568	1.962	0	0.000	Concordant
MCF7	PC3	Acute myeloid leukemia	05221	112	0.449	1.884	0	0.000	Concordant
MCF7	PC3	Chemokine signaling pathway	04062	119	0.420	1.882	0	0.000	Concordant
MCF7	PC3	VEGF signaling pathway	04370	66	0.469	1.856	0	0.000	Concordant
MCF7	PC3	Estrogen signaling pathway	04915	111	0.448	1.852	0	0.000	Concordant
MCF7	PC3	Natural killer cell mediated cytotoxicity	04650	68	0.464	1.798	0	0.000	Concordant
MCF7	PC3	Endocrine resistance	01522	177	0.397	1.783	0	0.000	Concordant
MCF7	PC3	Colorectal cancer	05210	106	0.391	1.639	0	0.000	Concordant
MCF7	PC3	Prostate cancer	05215	150	0.348	1.616	0	0.000	Concordant
MCF7	PC3	Melanoma	05218	115	0.368	1.576	0	0.000	Concordant
MCF7	PC3	Gap junction	04540	37	0.530	1.904	0.001	0.008	Concordant
MCF7	PC3	Regulation of lipolysis in adipocytes	04923	23	0.576	1.830	0.001	0.008	Concordant
MCF7	PC3	Oxytocin signaling pathway	04921	35	0.480	1.708	0.001	0.008	Concordant
MCF7	PC3	Cell cycle	04110	110	-0.373	-1.630	0.001	0.008	Discordant
MCF7	PC3	Chagas disease (American trypanosomiasis)	05142	79	0.388	1.584	0.001	0.008	Concordant
MCF7	PC3	Neurotrophin signaling pathway	04722	135	0.350	1.574	0.001	0.008	Concordant
MCF7	PC3	Proteoglycans in cancer	05205	221	0.341	1.564	0.001	0.008	Concordant
MCF7	PC3	Fluid shear stress and atherosclerosis	05418	109	0.344	1.522	0.001	0.008	Concordant
MCF7	PC3	Rap1 signaling pathway	04015	209	0.303	1.463	0.001	0.008	Concordant
MCF7	PC3	Regulation of actin cytoskeleton	04810	139	0.350	1.558	0.002	0.015	Concordant
MCF7	PC3	Epstein-Barr virus infection	05169	119	-0.354	-1.565	0.003	0.019	Discordant
MCF7	PC3	B cell receptor signaling pathway	04662	60	0.388	1.545	0.003	0.019	Concordant
MCF7	PC3	Central carbon metabolism in cancer	05230	96	0.348	1.488	0.003	0.019	Concordant
MCF7	PC3	Glioma	05214	177	0.359	1.541	0.004	0.025	Concordant
MCF7	PC3	Renal cell carcinoma	05211	40	0.457	1.608	0.005	0.028	Concordant
MCF7	PC3	ErbB signaling pathway	04012	124	0.385	1.578	0.005	0.028	Concordant

Table A5f: Local Comparisons MCF7(121 - 139): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
MCF7	PC3	Signaling pathways regulating pluripotency of stem cells	04550	170	0.333	1.470	0.005	0.028	Concordant
MCF7	PC3	Spinocerebellar ataxia	05017	21	0.560	1.707	0.006	0.029	Concordant
MCF7	PC3	cAMP signaling pathway	04024	77	0.365	1.515	0.006	0.029	Concordant
MCF7	PC3	Chronic myeloid leukemia	05220	111	0.355	1.507	0.006	0.029	Concordant
MCF7	PC3	mTOR signaling pathway	04150	106	0.323	1.411	0.006	0.029	Concordant
MCF7	PC3	Endometrial cancer	05213	71	0.375	1.482	0.007	0.032	Concordant
MCF7	PC3	Prolactin signaling pathway	04917	83	0.357	1.472	0.007	0.032	Concordant
MCF7	PC3	Long-term potentiation	04720	35	0.429	1.514	0.012	0.053	Concordant
MCF7	PC3	Cholinergic synapse	04725	44	0.393	1.462	0.012	0.053	Concordant
MCF7	PC3	Dopaminergic synapse	04728	54	-0.387	-1.488	0.013	0.054	Discordant
MCF7	PC3	Alzheimer disease	05010	81	0.342	1.421	0.013	0.054	Concordant
MCF7	PC3	Complement and coagulation cascades	04610	26	0.459	1.522	0.014	0.057	Concordant
MCF7	PC3	Alcoholism	05034	60	0.440	1.541	0.016	0.062	Concordant
MCF7	PC3	Long-term depression	04730	38	0.452	1.526	0.016	0.062	Concordant
MCF7	PC3	Inflammatory mediator regulation of TRP channels	04750	29	-0.429	-1.467	0.022	0.081	Discordant
MCF7	PC3	T cell receptor signaling pathway	04660	86	0.320	1.348	0.022	0.081	Concordant
MCF7	PC3	C-type lectin receptor signaling pathway	04625	170	0.328	1.358	0.024	0.086	Concordant
MCF7	PC3	PPAR signaling pathway	03320	58	0.382	1.400	0.029	0.094	Concordant
MCF7	PC3	Progesterone-mediated oocyte maturation	04914	34	0.406	1.416	0.03	0.095	Concordant

Table A5g: Local Comparisons PC3(1 - 40): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
PC3	A375	Long-term depression	04730	38	-0.608	-2.170	0	0.000	Discordant
PC3	A375	Long-term potentiation	04720	35	-0.528	-1.996	0	0.000	Discordant
PC3	A375	Bladder cancer	05219	33	-0.518	-1.879	0	0.000	Discordant
PC3	A375	Toll-like receptor signaling pathway	04620	70	-0.420	-1.792	0	0.000	Discordant
PC3	A375	Endocrine resistance	01522	177	-0.318	-1.507	0	0.000	Discordant
PC3	A375	GnRH signaling pathway	04912	59	-0.400	-1.672	0.001	0.023	Discordant
PC3	A375	Regulation of actin cytoskeleton	04810	139	-0.327	-1.577	0.001	0.023	Discordant
PC3	A375	MAPK signaling pathway	04010	353	-0.241	-1.328	0.002	0.040	Discordant
PC3	A375	Longevity regulating pathway	04211	60	0.399	1.598	0.006	0.054	Concordant
PC3	A375	Oxytocin signaling pathway	04921	35	-0.425	-1.593	0.006	0.054	Discordant
PC3	A375	Natural killer cell mediated cytotoxicity	04650	68	-0.384	-1.591	0.004	0.054	Discordant
PC3	A375	Basal cell carcinoma	05217	66	0.389	1.577	0.003	0.054	Concordant
PC3	A375	Estrogen signaling pathway	04915	111	-0.357	-1.549	0.006	0.054	Discordant
PC3	A375	Melanogenesis	04916	71	0.369	1.507	0.005	0.054	Concordant
PC3	A375	Autophagy - animal	04140	71	0.368	1.505	0.006	0.054	Concordant
PC3	A375	Colorectal cancer	05210	106	-0.317	-1.417	0.005	0.054	Discordant
PC3	A375	Rap1 signaling pathway	04015	209	-0.256	-1.329	0.004	0.054	Discordant
PC3	A375	PI3K-Akt signaling pathway	04151	417	0.253	1.304	0.005	0.054	Concordant
PC3	A375	Serotonergic synapse	04726	26	-0.451	-1.571	0.009	0.077	Discordant
PC3	A375	Hippo signaling pathway	04390	125	0.311	1.396	0.013	0.088	Concordant
PC3	A375	Melanoma	05218	115	-0.297	-1.363	0.013	0.088	Discordant
PC3	A549	Long-term potentiation	04720	35	-0.519	-1.939	0	0.000	Discordant
PC3	A549	Long-term depression	04730	38	-0.536	-1.937	0.001	0.054	Discordant
PC3	A549	Hippo signaling pathway	04390	125	0.353	1.606	0.002	0.054	Concordant
PC3	A549	MAPK signaling pathway	04010	353	-0.254	-1.389	0.001	0.054	Discordant
PC3	HA1E	Osteoclast differentiation	04380	98	0.388	1.774	0	0.000	Concordant
PC3	HA1E	Salmonella infection	05132	124	0.347	1.628	0	0.000	Concordant
PC3	HA1E	Hepatitis B	05161	154	0.345	1.658	0.001	0.040	Concordant
PC3	HA1E	Long-term potentiation	04720	35	-0.454	-1.767	0.002	0.065	Discordant
PC3	HA1E	Inflammatory mediator regulation of TRP channels	04750	29	0.481	1.677	0.003	0.069	Concordant
PC3	HA1E	AGE-RAGE signaling pathway in diabetic complications	04933	139	0.313	1.498	0.003	0.069	Concordant
PC3	HA1E	Serotonergic synapse	04726	26	-0.487	-1.758	0.005	0.090	Discordant
PC3	HA1E	Rap1 signaling pathway	04015	209	-0.237	-1.303	0.005	0.090	Discordant
PC3	HEPG2	Thermogenesis	04714	45	0.501	1.908	0	0.000	Concordant
PC3	HEPG2	Long-term potentiation	04720	35	-0.508	-1.867	0	0.000	Discordant
PC3	HEPG2	Long-term depression	04730	38	-0.489	-1.743	0	0.000	Discordant
PC3	HEPG2	Hippo signaling pathway	04390	125	0.344	1.581	0	0.000	Concordant
PC3	HEPG2	Serotonergic synapse	04726	26	-0.544	-1.836	0.002	0.054	Discordant
PC3	HEPG2	Regulation of actin cytoskeleton	04810	139	-0.340	-1.562	0.002	0.054	Discordant
PC3	HEPG2	Longevity regulating pathway - multiple species	04213	90	-0.438	-1.722	0.003	0.061	Discordant

Table A5g: Local Comparisons PC3(41 - 80): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
PC3	HEPG2	Rap1 signaling pathway	04015	209	-0.292	-1.459	0.003	0.061	Discordant
PC3	HEPG2	Estrogen signaling pathway	04915	111	-0.377	-1.583	0.005	0.062	Discordant
PC3	HEPG2	Th1 and Th2 cell differentiation	04658	87	0.350	1.546	0.005	0.062	Concordant
PC3	HEPG2	Melanoma	05218	115	-0.341	-1.524	0.004	0.062	Discordant
PC3	HEPG2	Longevity regulating pathway	04211	60	-0.369	-1.516	0.005	0.062	Discordant
PC3	HEPG2	Necroptosis	04217	81	0.356	1.506	0.004	0.062	Concordant
PC3	HEPG2	Spinocerebellar ataxia	05017	21	-0.540	-1.754	0.007	0.063	Discordant
PC3	HEPG2	Thyroid hormone signaling pathway	04919	72	-0.373	-1.562	0.006	0.063	Discordant
PC3	HEPG2	Fc epsilon RI signaling pathway	04664	53	0.401	1.554	0.006	0.063	Concordant
PC3	HEPG2	Relaxin signaling pathway	04926	154	0.345	1.472	0.007	0.063	Concordant
PC3	HEPG2	Pathways in cancer	05200	554	0.231	1.258	0.007	0.063	Concordant
PC3	HEPG2	Amyotrophic lateral sclerosis (ALS)	05014	15	0.566	1.637	0.008	0.068	Concordant
PC3	HEPG2	AGE-RAGE signaling pathway in diabetic complications	04933	139	0.305	1.416	0.009	0.073	Concordant
PC3	HEPG2	TNF signaling pathway	04668	66	0.383	1.470	0.012	0.093	Concordant
PC3	HEPG2	Endocrine resistance	01522	177	-0.293	-1.375	0.013	0.096	Discordant
PC3	HT29	Melanoma	05218	115	-0.525	-2.257	0	0.000	Discordant
PC3	HT29	Hippo signaling pathway	04390	125	0.394	1.823	0	0.000	Concordant
PC3	HT29	Prostate cancer	05215	150	-0.350	-1.603	0.001	0.032	Discordant
PC3	HT29	Acute myeloid leukemia	05221	112	-0.362	-1.550	0.001	0.032	Discordant
PC3	HT29	Chemokine signaling pathway	04062	119	-0.338	-1.516	0.001	0.032	Discordant
PC3	HT29	Progesterone-mediated oocyte maturation	04914	34	-0.488	-1.734	0.002	0.054	Discordant
PC3	HT29	Inflammatory mediator regulation of TRP channels	04750	29	0.488	1.667	0.004	0.054	Concordant
PC3	HT29	VEGF signaling pathway	04370	66	-0.411	-1.635	0.004	0.054	Discordant
PC3	HT29	EGFR tyrosine kinase inhibitor resistance	01521	158	-0.357	-1.534	0.003	0.054	Discordant
PC3	HT29	Glioma	05214	177	-0.357	-1.523	0.004	0.054	Discordant
PC3	HT29	Regulation of actin cytoskeleton	04810	139	-0.324	-1.470	0.003	0.054	Discordant
PC3	HT29	MAPK signaling pathway	04010	353	-0.250	-1.300	0.004	0.054	Discordant
PC3	HT29	Non-small cell lung cancer	05223	82	-0.389	-1.579	0.005	0.058	Discordant
PC3	HT29	Central carbon metabolism in cancer	05230	96	-0.350	-1.500	0.005	0.058	Discordant
PC3	HT29	Serotonergic synapse	04726	26	-0.509	-1.686	0.007	0.071	Discordant
PC3	HT29	GABAergic synapse	04727	32	-0.570	-1.674	0.007	0.071	Discordant
PC3	HT29	Insulin signaling pathway	04910	98	-0.321	-1.391	0.009	0.081	Discordant
PC3	HT29	Thermogenesis	04714	45	0.403	1.521	0.013	0.100	Concordant
PC3	MCF7	Serotonergic synapse	04726	26	-0.613	-2.047	0	0.000	Discordant
PC3	MCF7	GnRH secretion	04929	34	-0.587	-2.025	0	0.000	Discordant
PC3	MCF7	Platelet activation	04611	33	-0.568	-1.973	0	0.000	Discordant
PC3	MCF7	Acute myeloid leukemia	05221	112	-0.452	-1.946	0	0.000	Discordant
PC3	MCF7	Gap junction	04540	37	-0.530	-1.906	0	0.000	Discordant
PC3	MCF7	Chemokine signaling pathway	04062	119	-0.420	-1.882	0	0.000	Discordant
PC3	MCF7	Estrogen signaling pathway	04915	111	-0.453	-1.858	0	0.000	Discordant

Table A5g: Local Comparisons PC3(81 - 120): FDR q-value ≤ 0.10

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
PC3	MCF7	VEGF signaling pathway	04370	66	-0.470	-1.858	0	0.000	Discordant
PC3	MCF7	Endocrine resistance	01522	177	-0.403	-1.812	0	0.000	Discordant
PC3	MCF7	Colorectal cancer	05210	106	-0.394	-1.655	0	0.000	Discordant
PC3	MCF7	Cell cycle	04110	110	0.372	1.633	0	0.000	Concordant
PC3	MCF7	Glioma	05214	177	-0.374	-1.609	0	0.000	Discordant
PC3	MCF7	Natural killer cell mediated cytotoxicity	04650	68	-0.465	-1.824	0.001	0.009	Discordant
PC3	MCF7	Melanoma	05218	115	-0.370	-1.606	0.001	0.009	Discordant
PC3	MCF7	Prostate cancer	05215	150	-0.349	-1.606	0.001	0.009	Discordant
PC3	MCF7	Proteoglycans in cancer	05205	221	-0.351	-1.606	0.001	0.009	Discordant
PC3	MCF7	Neurotrophin signaling pathway	04722	135	-0.351	-1.590	0.001	0.009	Discordant
PC3	MCF7	Regulation of actin cytoskeleton	04810	139	-0.351	-1.570	0.001	0.009	Discordant
PC3	MCF7	Fluid shear stress and atherosclerosis	05418	109	-0.344	-1.542	0.001	0.009	Discordant
PC3	MCF7	Oxytocin signaling pathway	04921	35	-0.480	-1.700	0.002	0.014	Discordant
PC3	MCF7	Chagas disease (American trypanosomiasis)	05142	79	-0.388	-1.605	0.002	0.014	Discordant
PC3	MCF7	Epstein-Barr virus infection	05169	119	0.352	1.545	0.002	0.014	Concordant
PC3	MCF7	Rap1 signaling pathway	04015	209	-0.303	-1.453	0.002	0.014	Discordant
PC3	MCF7	ErbB signaling pathway	04012	124	-0.390	-1.592	0.003	0.019	Discordant
PC3	MCF7	Signaling pathways regulating pluripotency of stem cells	04550	170	-0.343	-1.500	0.003	0.019	Discordant
PC3	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.576	-1.828	0.004	0.024	Discordant
PC3	MCF7	Spinocerebellar ataxia	05017	21	-0.560	-1.738	0.004	0.024	Discordant
PC3	MCF7	Chronic myeloid leukemia	05220	111	-0.357	-1.532	0.005	0.026	Discordant
PC3	MCF7	cAMP signaling pathway	04024	77	-0.365	-1.503	0.005	0.026	Discordant
PC3	MCF7	Central carbon metabolism in cancer	05230	96	-0.349	-1.498	0.005	0.026	Discordant
PC3	MCF7	mTOR signaling pathway	04150	106	-0.323	-1.410	0.005	0.026	Discordant
PC3	MCF7	B cell receptor signaling pathway	04662	60	-0.388	-1.542	0.006	0.030	Discordant
PC3	MCF7	Prolactin signaling pathway	04917	83	-0.358	-1.482	0.007	0.034	Discordant
PC3	MCF7	Renal cell carcinoma	05211	40	-0.457	-1.606	0.01	0.040	Discordant
PC3	MCF7	Alcoholism	05034	60	-0.443	-1.582	0.01	0.040	Discordant
PC3	MCF7	Long-term depression	04730	38	-0.453	-1.558	0.01	0.040	Discordant
PC3	MCF7	Complement and coagulation cascades	04610	26	-0.459	-1.522	0.009	0.040	Discordant
PC3	MCF7	Long-term potentiation	04720	35	-0.429	-1.508	0.009	0.040	Discordant
PC3	MCF7	Dopaminergic synapse	04728	54	0.387	1.506	0.01	0.040	Concordant
PC3	MCF7	Alzheimer disease	05010	81	-0.342	-1.425	0.009	0.040	Discordant
PC3	MCF7	Endometrial cancer	05213	71	-0.376	-1.456	0.014	0.055	Discordant
PC3	MCF7	PPAR signaling pathway	03320	58	-0.383	-1.426	0.017	0.064	Discordant
PC3	MCF7	C-type lectin receptor signaling pathway	04625	170	-0.340	-1.398	0.017	0.064	Discordant
PC3	MCF7	Cholinergic synapse	04725	44	-0.393	-1.472	0.018	0.066	Discordant
PC3	MCF7	T cell receptor signaling pathway	04660	86	-0.320	-1.355	0.019	0.068	Discordant
PC3	MCF7	Inflammatory mediator regulation of TRP channels	04750	29	0.429	1.467	0.023	0.081	Concordant
PC3	MCF7	Progesterone-mediated oocyte maturation	04914	34	-0.406	-1.425	0.026	0.090	Discordant

Table A6a: Global and local ESEA results for RLA pathway with A375 as reference.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A375	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	-0.270	-0.875	0.37	**0.460**	Discordant
A375	A549	Regulation of lipolysis in adipocytes	04923	23	0.321	1.057	0.174	**0.381**	Concordant
A375	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.429	1.430	0.025	**0.096**	Concordant
A375	HEPG2	Regulation of lipolysis in adipocytes	04923	23	-0.246	-0.812	0.403	**0.484**	Discordant
A375	HT29	Regulation of lipolysis in adipocytes	04923	23	-0.368	-1.184	0.113	**0.383**	Discordant
A375	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.495	-1.586	0.016	**0.136**	Discordant
A375	PC3	Regulation of lipolysis in adipocytes	04923	23	0.234	0.787	0.367	**0.437**	Concordant

Table A6b: Global and local ESEA results for RLA pathway with A549 as reference.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
A549	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	-0.395	-1.311	0.054	**0.307**	Discordant
A549	A375	Regulation of lipolysis in adipocytes	04923	23	-0.321	-1.071	0.169	**0.376**	Discordant
A549	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.223	0.746	0.446	**0.544**	Concordant
A549	HEPG2	Regulation of lipolysis in adipocytes	04923	23	-0.405	-1.335	0.061	**0.289**	Discordant
A549	HT29	Regulation of lipolysis in adipocytes	04923	23	-0.436	-1.432	0.033	**0.177**	Discordant
A549	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.568	-1.884	0.002	**0.027**	Discordant
A549	PC3	Regulation of lipolysis in adipocytes	04923	23	-0.313	-1.017	0.232	**0.422**	Discordant

Table A6c: Global and local ESEA results for RLA pathway with HA1E as reference.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HA1E	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	-0.442	-1.445	0.039	**0.150**	Discordant
HA1E	A375	Regulation of lipolysis in adipocytes	04923	23	-0.429	-1.401	0.044	**0.130**	Discordant
HA1E	A549	Regulation of lipolysis in adipocytes	04923	23	-0.223	-0.731	0.44	**0.544**	Discordant
HA1E	HEPG2	Regulation of lipolysis in adipocytes	04923	23	-0.435	-1.459	0.024	**0.197**	Discordant
HA1E	HT29	Regulation of lipolysis in adipocytes	04923	23	-0.453	-1.492	0.029	**0.130**	Discordant
HA1E	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.598	-1.981	0	**0.000**	Discordant
HA1E	PC3	Regulation of lipolysis in adipocytes	04923	23	-0.335	-1.101	0.182	**0.328**	Discordant

Table A6d: Global and local ESEA results for RLA pathway with HEPG2 as reference.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HEPG2	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	0.208	0.652	0.491	**0.508**	Concordant
HEPG2	A375	Regulation of lipolysis in adipocytes	04923	23	0.246	0.798	0.401	**0.481**	Concordant
HEPG2	A549	Regulation of lipolysis in adipocytes	04923	23	0.405	1.342	0.054	**0.273**	Concordant
HEPG2	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.435	1.467	0.029	**0.204**	Concordant
HEPG2	HT29	Regulation of lipolysis in adipocytes	04923	23	-0.335	-1.090	0.166	**0.320**	Discordant
HEPG2	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.476	-1.564	0.017	**0.106**	Discordant
HEPG2	PC3	Regulation of lipolysis in adipocytes	04923	23	0.321	1.045	0.189	**0.359**	Concordant

Table A6e: Global and local ESEA results for RLA pathway with HT29 as reference.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
HT29	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	0.362	1.163	0.129	**0.294**	Concordant
HT29	A375	Regulation of lipolysis in adipocytes	04923	23	0.368	1.198	0.119	**0.388**	Concordant
HT29	A549	Regulation of lipolysis in adipocytes	04923	23	0.436	1.408	0.041	**0.208**	Concordant
HT29	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.453	1.496	0.024	**0.122**	Concordant
HT29	HEPG2	Regulation of lipolysis in adipocytes	04923	23	0.335	1.087	0.175	**0.335**	Concordant
HT29	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.497	-1.646	0.006	**0.126**	Discordant
HT29	PC3	Regulation of lipolysis in adipocytes	04923	23	0.417	1.339	0.053	**0.178**	Concordant

Table A6f: Global and local ESEA results for RLA pathway with HT29 as reference.

Reference	Comparison	Pathway	hsa	# Edges	EES	NEES	NOM p-val	FDR q-val	Direction
PC3	AVERAGE	Regulation of lipolysis in adipocytes	04923	23	-0.375	-1.208	0.103	**0.253**	Discordant
PC3	A375	Regulation of lipolysis in adipocytes	04923	23	-0.234	-0.797	0.368	**0.438**	Discordant
PC3	A549	Regulation of lipolysis in adipocytes	04923	23	0.313	1.025	0.224	**0.427**	Concordant
PC3	HA1E	Regulation of lipolysis in adipocytes	04923	23	0.335	1.104	0.176	**0.333**	Concordant
PC3	HEPG2	Regulation of lipolysis in adipocytes	04923	23	-0.321	-1.051	0.191	**0.376**	Discordant
PC3	HT29	Regulation of lipolysis in adipocytes	04923	23	-0.417	-1.333	0.06	**0.197**	Discordant
PC3	MCF7	Regulation of lipolysis in adipocytes	04923	23	-0.576	-1.828	0.004	**0.024**	Discordant

Appendix Figures

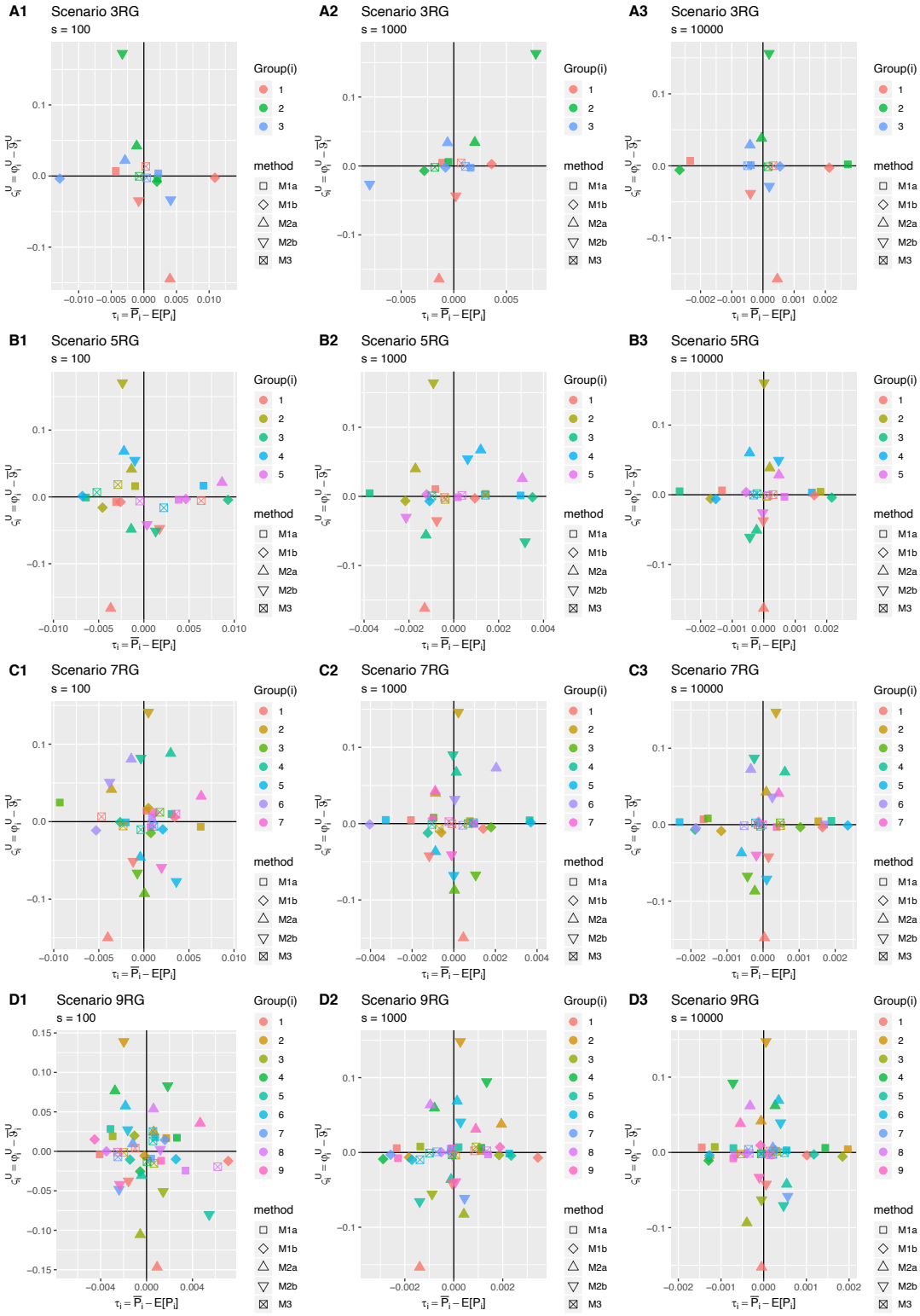


Figure A1: ***Biased Condition*** Scatterplots comparing sampling methods across two dimensions of bias at different numbers of simulations: s = 100 (A1 – D1), s = 1000 (A2-D2), s = 10000 (A3-D3).

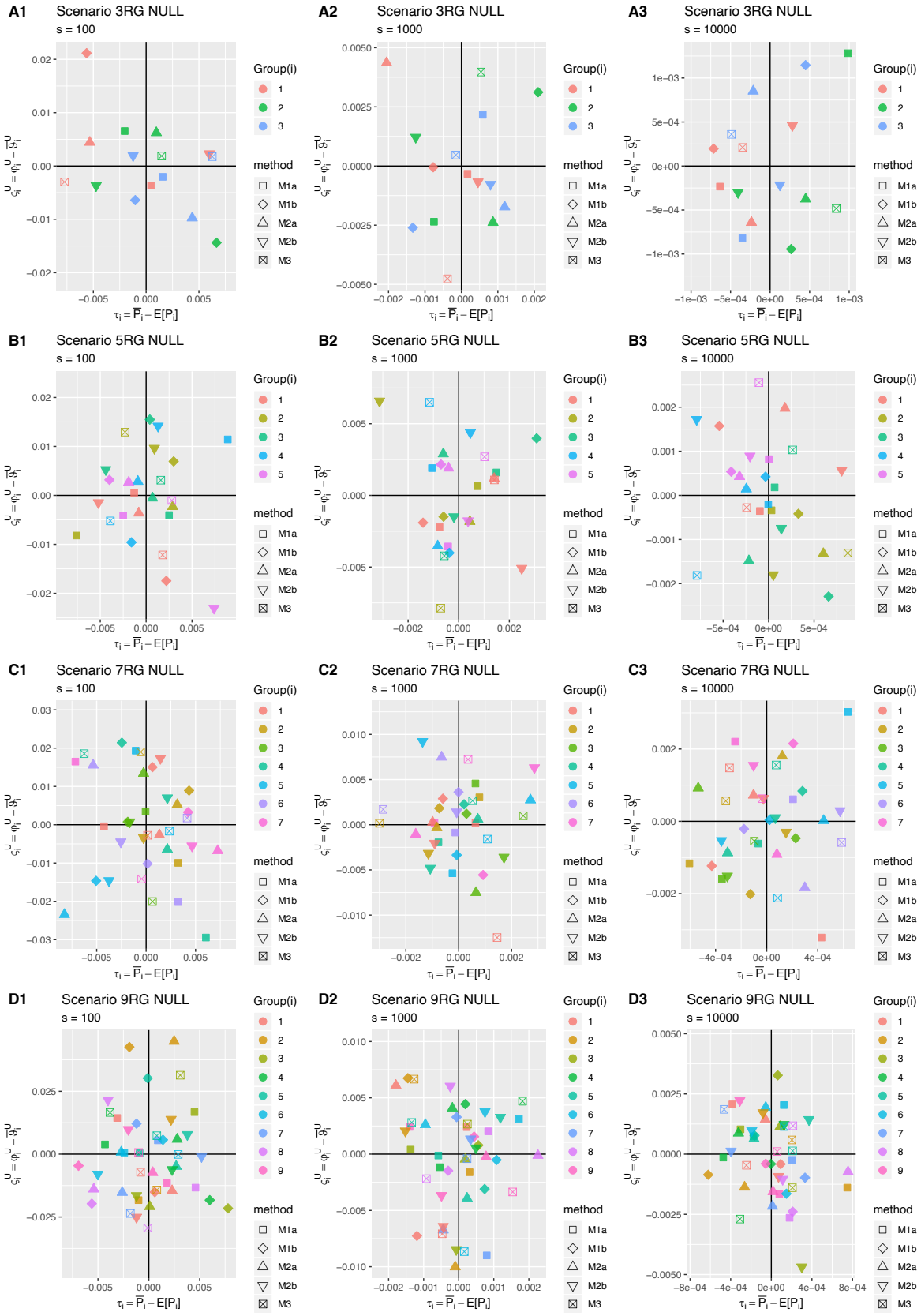


Figure A2: ***Null Condition*** Scatterplots comparing sampling methods across two dimensions of bias at different numbers of simulations: s = 100 (A1 – D1), s = 1000 (A2-D2), s = 10000 (A3-D3).

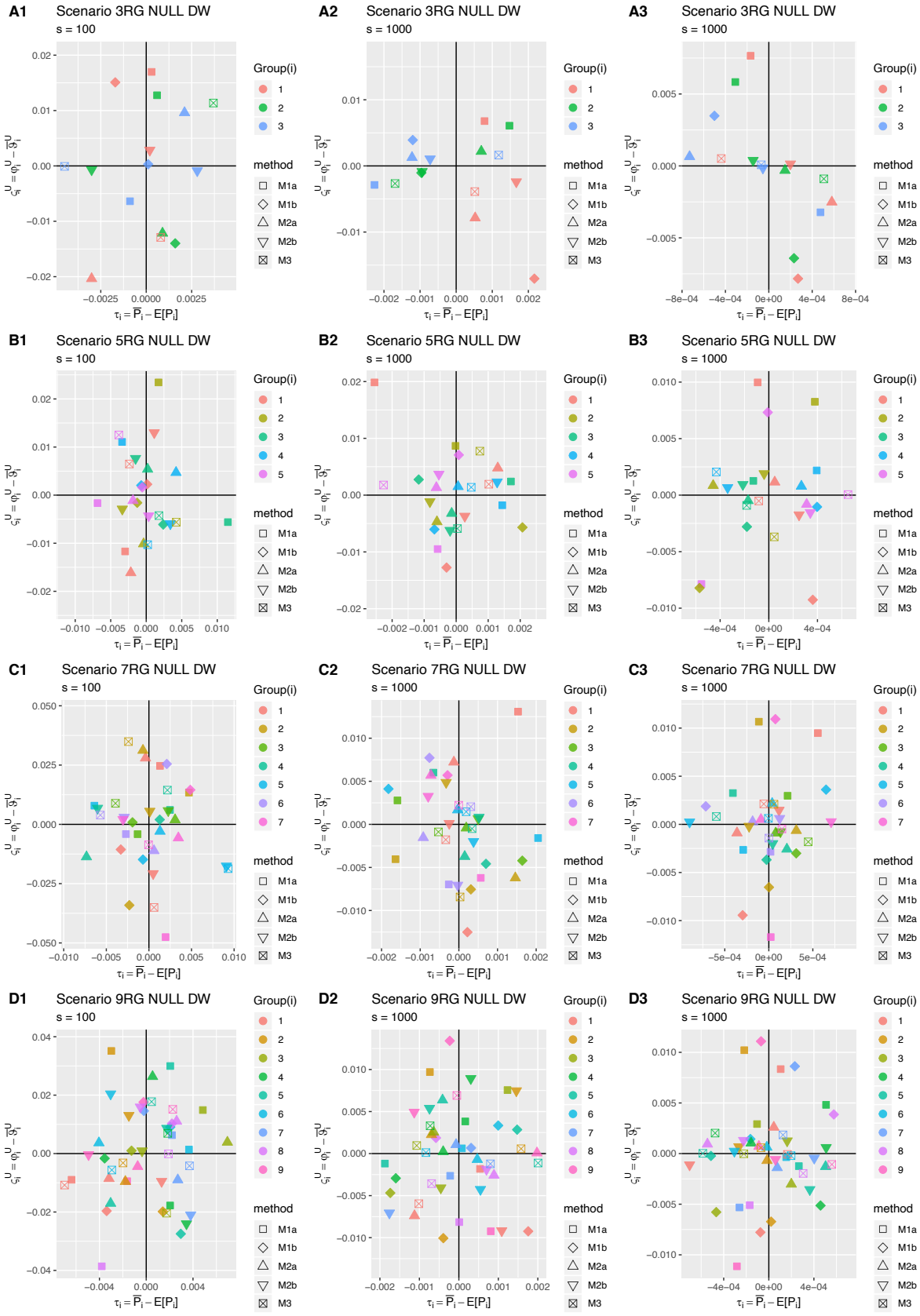


Figure A3: *Null* CW Scatterplots comparing sampling methods across two dimensions of bias at different numbers of simulations: s = 100 (A1 – D1), s = 1000 (A2-D2), s = 10000 (A3-D3).

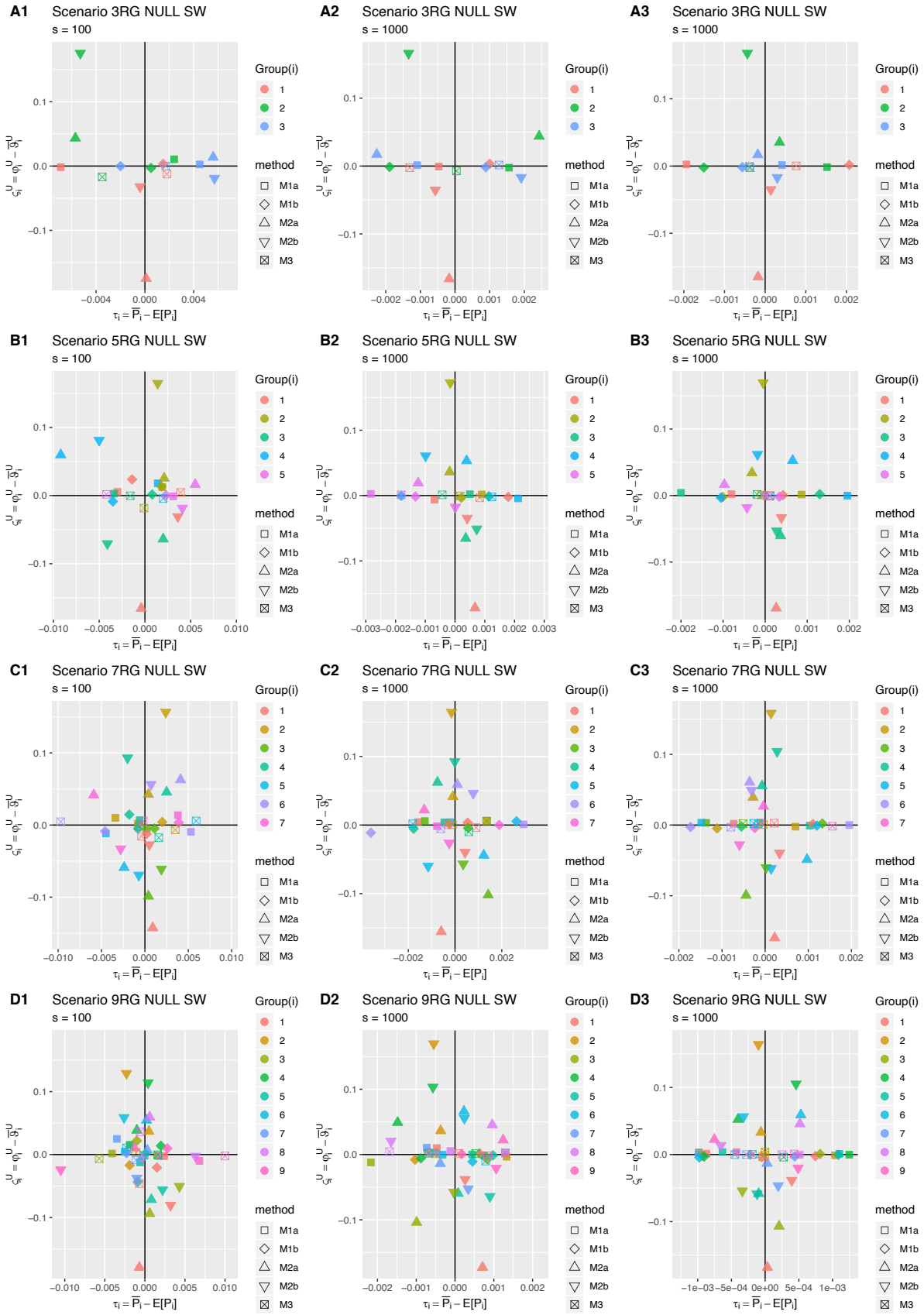


Figure A4: B*iased Condition* Scatterplots comparing sampling methods across two dimensions of bias at different numbers of simulations: s = 100 (A1 – D1), s = 1000 (A2-D2), s = 10000 (A3-D3).

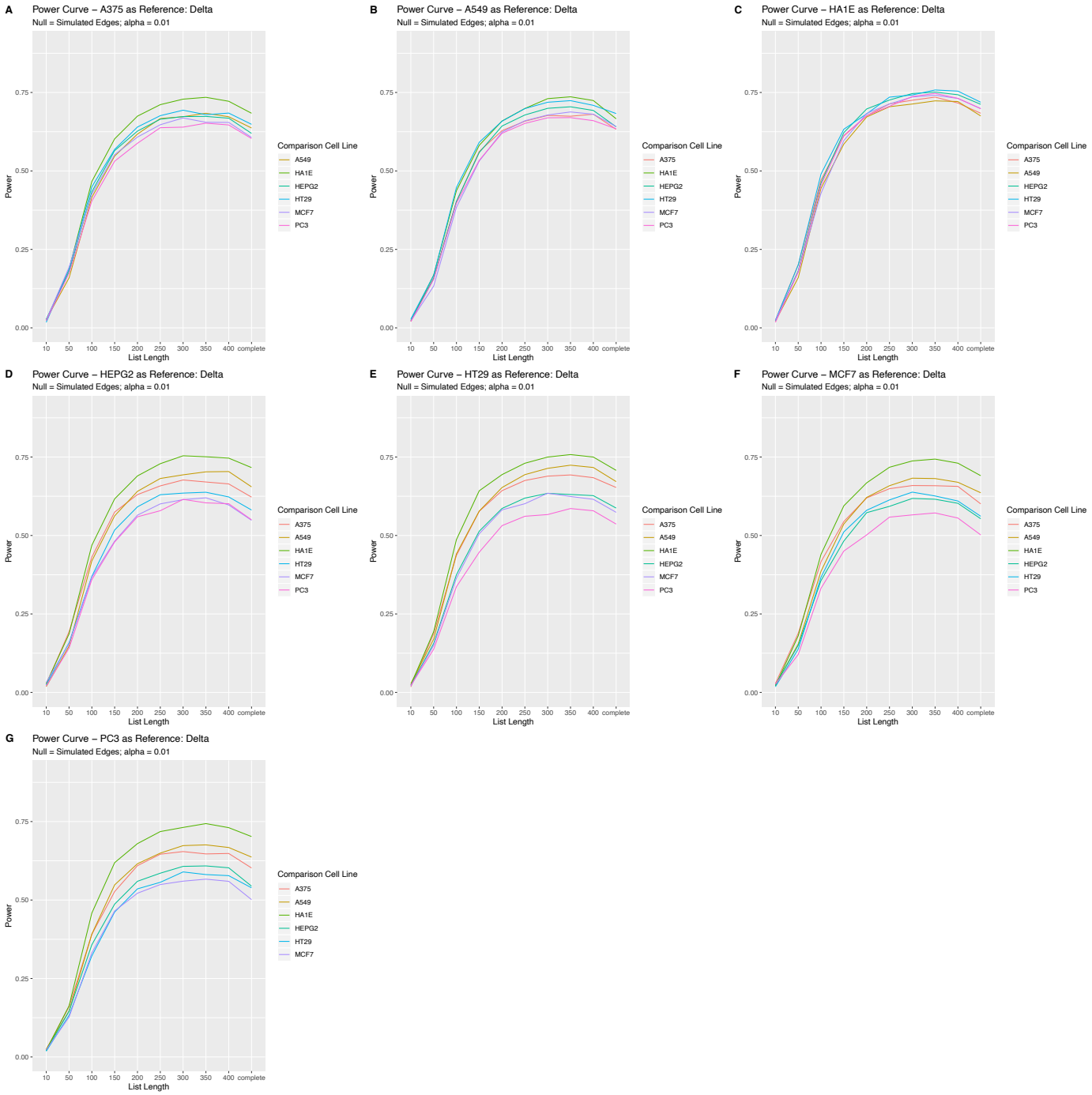


Figure A5: ***Individual cell line power plots – Delta*** These graphs represent those from **Figure 19** broken up by each cell line acting as a reference; A375(A), A549(B), HA1E(C), HEPG2(D), HT29(E), MCF7(F) and PC3(G).

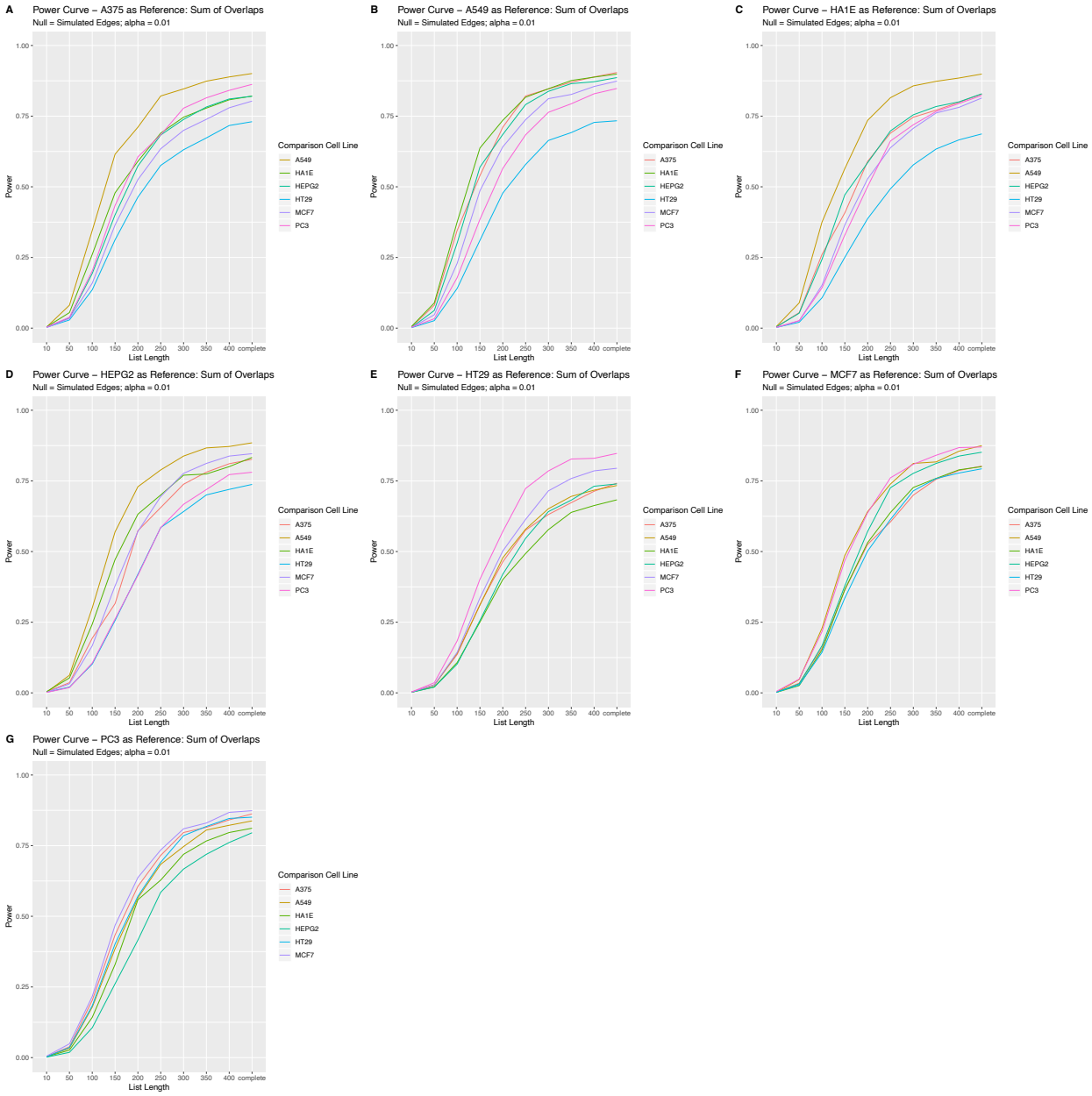


Figure A6: ***Individual cell line power plots – $\sum$ overlaps*** These graphs represent those from **Figure 20** broken up by each cell line acting as a reference; A375(A), A549(B), HA1E(C), HEPG2(D), HT29(E), MCF7(F) and PC3(G).

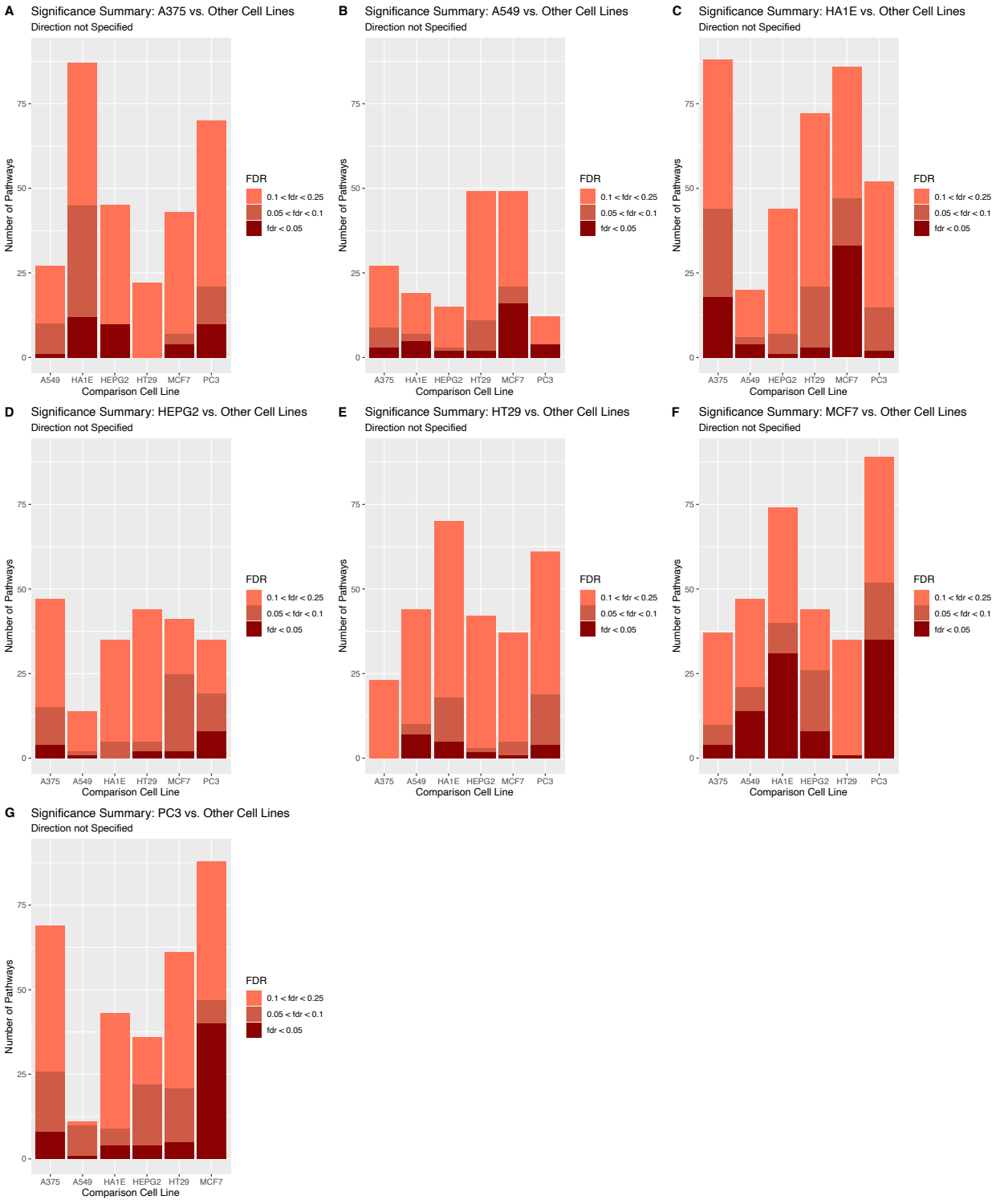


Figure A7: ***Local Results for ESEA*** Bar plots for the number of significantly enriched pathways broken down by FDR for each cell line; A375(A), A549(B), HA1E(C), HEPG2(D), HT29(E), MCF7(F) and PC3(G).

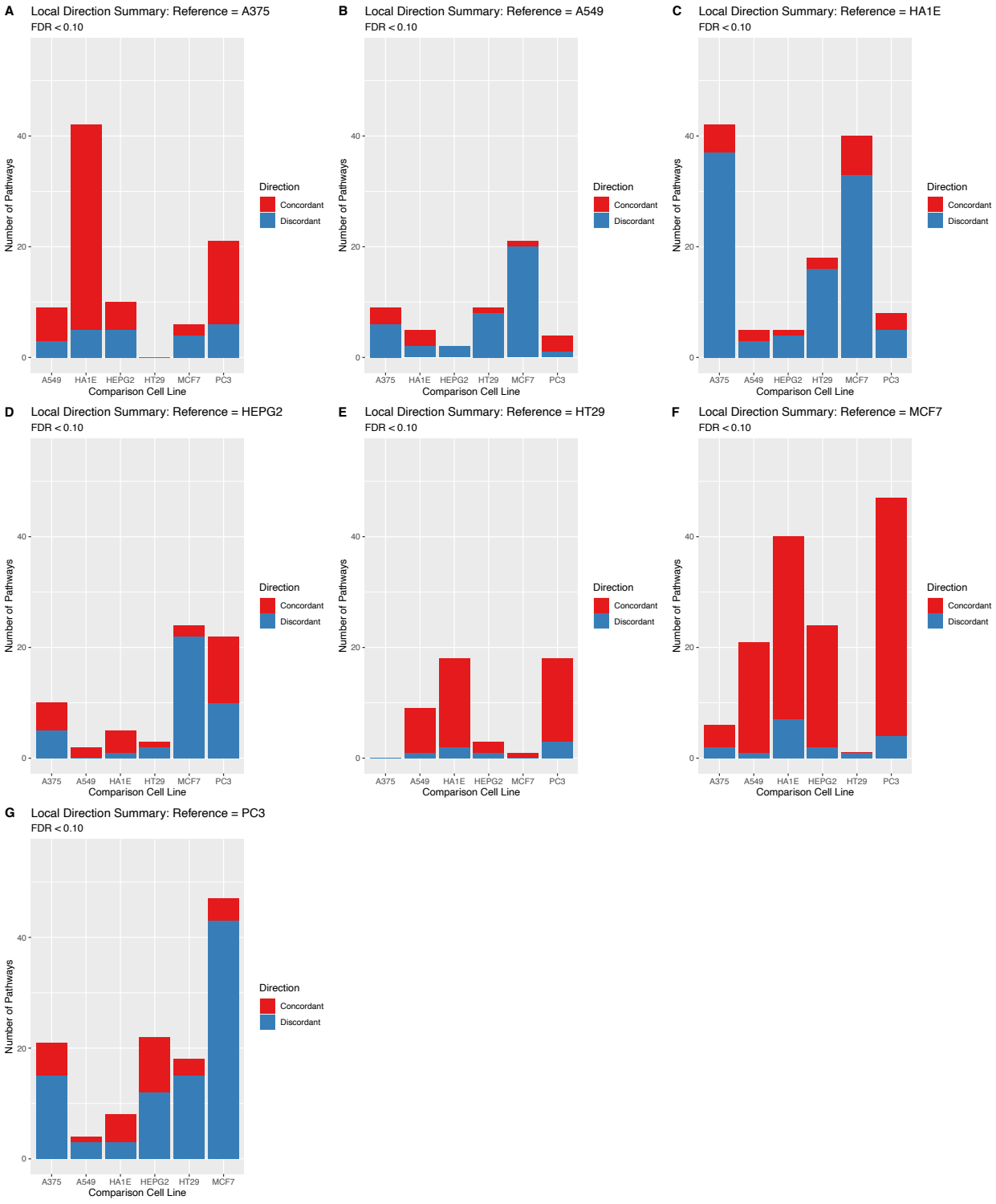


Figure A8: ***Local Results for ESEA*** Bar plots for the number of significantly enriched pathways broken down by direction for each cell line; A375(A), A549(B), HA1E(C), HEPG2(D), HT29(E), MCF7(F) and PC3(G).