

Integrating Functional Genomics with Systems Biology to Discover Drivers and
Therapeutic Targets of Human Malignancies

Jiyang Yu

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

under the Executive Committee

of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2012

© 2012

Jiyang Yu

All rights reserved

ABSTRACT

Integrating Functional Genomics with Systems Biology to Discover Drivers and Therapeutic Targets of Human Malignancies

Jiyang Yu

Genome-wide RNAi screening has emerged as a powerful tool for loss-of-function studies that may lead to therapeutic target discovery for human malignancies in the era of personalized medicine. However, due to high false-positive and false-negative rates arising from noise of high-throughput measurements and off-target effects, powerful computational tools and additional knowledge are much needed to analyze and complement it. Availability of high-throughput genomic data including gene expression profiles, copy number variations from large-sampled primary patients and cell lines allows us to tackle underlying drivers causally associated with tumorigenesis or drug-resistance.

In my dissertation, I have developed a framework to integrate functional RNAi screens with systems biology of cancer genomics to tailor potential therapeutics for reversal of drug-resistance or treatment of aggressive tumors. I developed a series of algorithms and tools to deconvolute, QC and post-analyze high-throughput shRNA screening data by next-generation sequencing technology (shSeq), particularly a novel Bayesian hierarchical modeling approach to integrate multiple shRNAs targeting the same gene, which outperforms existing

methods. In parallel, I developed a systems biology algorithm, NetBID2, to infer disease drivers from high-throughput genomic data by reverse-engineering network and Bayesian inference, which is able to detect hidden drivers that traditional methods fail to find.

Integrating NetBID2 with functional RNAi screens, I have identified known and novel driver-type therapeutic targets in various disease contexts. For example, I discovered that AKT1 is a driver for glucocorticoid (GC) resistance, a problem in the treatment of T-ALL. The inhibition of AKT1 was validated to reverse GC-resistance. Additionally, upon silencing predicted master regulators of GC resistance with shRNA screens, 13 out of 16 were validated to significantly overcome resistance. In breast cancer, I discovered that STAT3 is required for transformation of HER2+ breast cancer, an aggressive breast tumor subtype. The suppression of STAT3 was confirmed *in vitro* and *in vivo* to be an effective therapy for HER2+ breast cancer. Moreover, my analysis revealed that STAT3 silencing only works in ER- cases. Using my framework, I have also identified potential therapeutic targets for ABC or GCB-type DLBCL and subtype-based breast cancer that are currently being validated.

Table of Content

Table of Content	i
List of Tables	xvii
List of Figures	xxiii
Acknowledgements	lxiv
Chapter 1 Introduction	1
1.1 Personalized Medicine in Cancer Treatment.....	2
1.1.1 Problems of conventional cancer treatment approaches.....	2
1.1.2 Problems of targeting oncogenes for cancer treatment.....	3
1.2 Functional Genomics: Genome-wide RNAi Screening	5
1.2.1 Microarray-based shRNA screening.....	7
1.2.2 NGS-based shRNA screening.....	8
1.2.3 Limitations of high-throughput RNAi screening	9
1.3 Cancer Genomics and Systems Biology.....	9
1.3.1 Collaborative projects on cancer genomics.....	11
1.3.2 Systems biology	12
1.4 Integration of Functional Genomics with Cancer Genomics	14
1.5 Overview of the Dissertation.....	16
Chapter 2 Computational Analysis of Next Generation Sequencing-based shRNA Screening (shSeq) Data	18

2.1	Introduction.....	18
2.2	shRNA Library	21
2.3	ShortRead+: QA of Raw NGS Data.....	22
2.3.1	Format transformation of raw NGS data (QSEQ to FASTQ)	22
2.3.2	Phred quality score.....	25
2.3.3	Overall quality of raw NGS data	26
2.3.4	Cycle-based quality distribution.....	28
2.3.5	Distribution of read count.....	29
2.3.6	Cycle-based base calls.....	31
2.3.7	Example of a bad sequence run	32
2.4	shScanner: Deconvolution of Raw NGS Reads.....	34
2.5	shSEQ: Processing, Normalization and QA of Deconvoluted shSeq Data	36
2.5.1	Preprocessing	37
2.5.2	Normalization	37
2.5.3	QA of normalized data.....	43
2.6	Comparison of NGS with Microarray Platform for RNAi Screening	51
2.7	shADER: shRNA-level Differential Representation Analysis	53
2.7.1	Bayesian linear model	54
2.7.2	Summarizing differential representation results.....	56
2.8	shMA / BHM: Scoring Gene Level Activity by Integrating Multiple Hairpins Targeting the Same Gene	60
2.9	Post-Analysis.....	61

2.9.1	Functional enrichment analysis	61
2.9.2	Sensitivity analysis	63
Chapter 3 Meta-Analysis of High-throughput RNAi Screening Data (the BHM algorithm)		69
3.1	Summary	69
3.2	Introduction.....	70
3.3	Profiling Cell Essential Genes by Microarray-based RNAi Screens	73
3.4	Classical “Separate-And-Combine” Approach	76
3.5	Our “Modeling-All-Together” Approach: Bayesian Hierarchical Model ..	78
3.6	Benchmark Datasets and Validations	82
3.7	Evaluation Results	87
3.7.1	BHM dominates RIGER and RSA	87
3.8	Discussion	91
Chapter 4 NetBID2: Network-based Bayesian Inference of Disease Drivers		95
4.1	Introduction and Motivation.....	95
4.1.1	The era of post-genomics in cancer	95
4.1.2	Gene expression signature is not robust	96
4.1.3	Systems biology	101
4.2	Explanation of NetBID2 using a Social Example	102
4.3	The NetBID2 Framework	105
4.3.1	Reverse-engineering regulatory or signaling networks (Step 1) ..	105

4.3.2	Signature analysis of diseases by Bayesian Probit model (Step 2)	109
4.3.3	Gene set enrichment analysis to infer disease drivers (Step 3)...	110
4.4	Evaluation of NetBID2	112
4.4.1	NetBID2 is more robust than expression signature	112
4.4.2	Ability to identify known drivers	114
4.4.3	Ability to identify “hidden” drivers.....	115
4.4.4	High prediction rate by experimental validations	117
4.5	Evaluation of ARACNe Predictability	119
4.5.1	Using STAT3 as an example to evaluate ARACNe predictability	119
4.5.2	Overall prediction of ARACNe is good.....	120
4.5.3	The direction of interaction defined by correlation might be misleading.....	121
4.5.4	Signaling network prediction is more precise than TF network....	124
4.6	Conclusion.....	127
Chapter 5 BSEA: Bayesian Set Enrichment Analysis.....		128
5.1	Introduction.....	128
5.2	The BSEA Algorithm.....	131
5.2.1	“Maxmean” statistic	131
5.2.2	Restandarization.....	133
5.2.3	Bayesian inference	134
5.3	Benchmark and Evaluation using Meta-Analysis of RNAi Screening Data	136

5.4	Results.....	137
5.4.1	“Maxmean”-based GSA outperforms KS-based GSEA.....	137
5.4.2	BSEA \geq GSA $>$ GSEA.....	138
5.4.3	BSEA dominates GSEA.....	140
5.4.4	Bayesian vs. Frequentist.....	140
5.4.5	BSEA cannot beat BHM.....	142
5.5	Conclusion.....	143
Chapter 6 Recovering Drug-Induced Apoptosis Subnetwork from		
Connectivity Map Data via a Bayesian Network Approach.....		
6.1	Summary.....	144
6.2	Introduction.....	145
6.3	CMAP Data.....	147
6.4	Drug-Response Signature Analysis.....	148
6.5	Enrichment of Apoptosis Pathway.....	152
6.6	Bayesian Network.....	157
6.6.1	Data modeling.....	158
6.6.2	Parameter learning.....	159
6.6.3	Structure scoring and search.....	160
6.6.4	Bootstrapping and model averaging.....	161
6.7	Results.....	162
6.7.1	Known direct interactions.....	163
6.7.2	Consistency with two major cell death pathways.....	165

6.7.3	BCL2L11/BIM as a gateway gene to drug-induced intrinsic apoptosis	165
6.7.4	TNFAIP3/A20 as a gateway gene to drug-induced extrinsic apoptosis	166
6.8	Discussion	169
Chapter 7 NetBID2 Identifies AKT1 as a Therapeutic Target to Reverse Glucocorticoid Resistance in T-ALL		
173		
7.1	Summary	173
7.2	Clinical Significance	174
7.3	Introduction	174
7.4	Results	180
7.4.1	NetBID2 with signaling network identifies AKT1 as a driver of glucocorticoid resistance in T-ALL	180
7.4.2	Constitutive activation of AKT1 impairs glucocorticoid response in T-ALL	187
7.4.3	Phosphorylation of the glucocorticoid receptor (GCR) by AKT1 ..	191
7.4.4	AKT signaling inhibits NR3C1 nuclear translocation following glucocorticoid treatment	196
7.4.5	Pharmacologic inhibition of AKT reverses glucocorticoid resistance <i>in vitro</i> and <i>in vivo</i>	202
7.5	Discussion	215
7.6	Materials and Methods	217

7.6.1	Reverse engineering signaling molecule-focused network analysis of glucocorticoid resistance in T-ALL	217
7.6.2	Glucocorticoid resistance signature analysis.....	218
7.6.3	Inferring master signaling drivers of glucocorticoid resistance in T-ALL by NetBID2	218
7.6.4	Cell lines and primary leukemia samples	219
7.6.5	Inhibitors and drugs	220
7.6.6	siRNA validation of regulators of glucocorticoid resistance	220
7.6.7	Luciferase reporter assays	220
7.6.8	Quantitative real-time PCR.....	221
7.6.9	Western blotting and immunoprecipitation.....	221
7.6.10	Preparation of Cytoplasmic and Nuclear extracts.....	223
7.6.11	In vitro GST-pull down protein interaction assays	223
7.6.12	In vitro kinase assays	224
7.6.13	Mass spectrometry analysis of NR3C1 phosphorylation sites	224
7.6.14	Immunofluorescence studies.....	225
7.6.15	Cell viability assays and flow cytometric analysis.....	226
7.6.16	Retroviral and lentiviral constructs and viral production.....	227
7.6.17	Recombinant protein production.....	228
7.6.18	Mice and animal procedures	228
7.6.19	Statistical analyses	232
7.7	GC-Responsive Signature of After vs. Before Treatment to Sensitive T-ALL Patients	236

7.7.1	NetBID2 identifies AKT1 as driver of GC-responsive signature...	238
7.8	Preliminary Results of Crossing Signaling Drivers with RNAi Screens	239
7.8.1	AKT1 doesn't show up from shRNA screening as a candidate ...	240
7.8.2	PRKAR1A in PI3K pathway shows up in both driver prediction and shRNA screens.....	241
Chapter 8 Integrating Functional Genomics with Systems Biology on Discovering Therapeutics to Reverse Glucocorticoid Resistance in T-ALL		
244		
8.1	Summary	244
8.2	Introduction.....	245
8.3	Methods.....	248
8.3.1	Reverse engineering transcriptional regulatory network of T-ALL	248
8.3.2	Signature analysis of GC-resistance	249
8.3.3	GSEA of inferring regulatory drivers of GC-resistance	249
8.3.4	Pooled shRNA screening	250
8.3.5	Differential representation analysis of individual shRNA	251
8.3.6	Integration of multiple shRNAs targeting the same gene.....	251
8.3.7	Combining differential representation scores of two cell lines	252
8.3.8	Silencing by siRNA and cell apoptosis assays	252
8.4	Results.....	253

8.4.1	The framework integrating RNAi screens with regulatory driver inference by NetBID2 identifies sixteen potential therapeutic targets	253
8.4.2	13 of 16 candidates, when repressed, reverse GC-resistance in vitro	255
8.4.3	75% of top genomics-Inferred drivers show significant effects to change GC-sensitivity in vitro	260
8.4.4	Validated targets work cooperatively by forming well-connected subcircuits.....	262
8.4.5	TRIM28 is a critical master regulator of GC-resistance in T-ALL.	264
8.4.6	Silencing TRIM28 increases GC-resistance by down-regulating GR	264
8.5	Discussion	267
Chapter 9 Integrating Functional Genomics with Systems Biology to Discover Therapeutic Targets for ERBB2/HER2+ Breast Cancer.....		
276		
9.1	Summary	276
9.2	Introduction.....	277
9.3	Results.....	279
9.3.1	The integrative framework of genome-wide RNAi screening with systems biology of cancer genomics (NetBID2) to identify therapeutic targets of ERBB2+/HER2+ breast cancer.....	279
9.3.2	Integrating RNAi screens with NetBID2 identifies STAT3 and two other signaling molecules as driver-type therapeutic targets of ERBB2+ breast cancer	283

9.3.3	STAT3 and phosphorylated-STAT3 is confirmed to be active in ERBB2+ MCF10A cells.....	287
9.3.4	STAT3 is validated <i>in vitro</i> to be lethal to ERBB2+ MCF10A cells when being silenced	288
9.3.5	STAT3 is validated <i>in vivo</i> to be lethal to ERBB2+ xenograft mouse models	290
9.3.6	STAT3 inhibition is specific to ERBB2+ breast cancer	291
9.3.7	STAT3 doesn't show up as a driver for HER2+ from NetBID2 analysis on expression profiles of primary breast cancer patients	294
9.3.8	STAT3 is addicted to ER status being a driver for ER- and HER2+ breast cancer, but not for ER+ ones	296
9.3.9	Searching for downstream targets of STAT3 being involved in modulation of ER- and HER2+ breast cancer.....	299
9.3.10	STAT3 targets that are lethal to ER- and HER2+ breast cancer	302
9.3.11	Other STAT family members (STAT5A and STAT1) show up as drivers of ERBB2+ breast cancer in analysis of data from both isogenic models and primary patients.....	305
9.3.12	RNAi screening from 2D vs. 3D vs. In Vivo environment.....	311
9.4	Methods.....	316
9.4.1	Reverse engineering transcriptional regulatory or signaling networks of Breast Cancer	316
9.4.2	Signature analysis of ERBB2+ MCF10A model.....	316

9.4.3	GSEA of inferring regulatory or signaling drivers of ERBB2+ MCF10 Cells	317
9.4.4	Pooled shRNA screening of ERBB2+ MCF10A cells	317
9.4.5	Differential representation analysis of individual shRNA	318
9.4.6	Gene level activity by integration of multiple shRNAs targeting the same gene	318
9.4.7	Meta-analysis of combining differential evidences	319
9.5	Discussion	319
9.5.1	Phosphorylation of STAT3 is required for STAT3 activity	319
9.5.2	2D vs. 3D: gene expression signature and NetBID2-predicted drivers	320
9.5.3	The power of meta-analysis and integration of functional genomics with systems biology	322

Chapter 10 Integrating Functional Genomics with Systems Biology to

Discover Driver-type Therapeutic Targets for ABC or GCB-type DLBCL ..324

10.1	Summary	324
10.2	Introduction	325
10.3	Results	327
10.3.1	Pooled shRNA screens of DLBCL lines by microarray and NGS	327
10.3.2	Clustering of shRNA screening samples	329
10.3.3	Functional profiles separate well ABC from GCB, BCL2-rearranged from non-rearranged DLBCL subtypes	331
10.3.4	Differentially represented genes from shRNA screens	333

10.3.5	Top differentially represented genes cross all cell lines.....	335
10.3.6	Enriched pathways by RNAi screening identified candidates	338
10.3.7	Crossing RNAi screening with signature genes of ABC vs. GCB	341
10.3.8	Crossing RNAi screening with NetBID2-predicted drivers of ABC vs. GCB	344
10.3.9	Crossing RNAi screening with CNV data.....	347
10.3.10	Crossing RNAi screening with NetBID2-predicted drivers and amplified genes from CNV data	349
10.4	On-going and Future Work	351

Chapter 11 Integrating Functional Genomics with Systems Biology on Therapeutic Target Discovery for Subtype or Genetic-feature Specific Breast Cancer 352

11.1	Introduction.....	352
11.2	Drivers and Therapeutic Targets for Basal or Luminal type Breast Cancer.....	354
11.2.1	Summary for shSeq data of 16 breast cell lines	354
11.2.2	Results of differential representation analysis	357
11.2.3	Unsupervised clustering of functional profiles separate subtypes well	359
11.2.4	Top candidates that are common in all breast tumor lines	361
11.2.5	Sensitivity analysis: difference between depleted essential genes and enriched tumor suppressor genes	362

11.2.6	Supervised clustering analysis of functional profiles identifies subtype specific lethal genes	363
11.2.7	Functional enrichment of lethal genes specific to basal or luminal subtype	364
11.2.8	Crossing with NetBID2-predicted drivers of basal vs. luminal subtype	365
11.2.9	Consistency with drug sensitivity data	366
11.3	RNAi Screens to Search for Synthetic Lethal Partners of Genetic-features in Breast Cancer	371
11.3.1	Summary for the shSeq data of genetically-engineered models	372
11.3.2	Overall gene level activity for each genetic-feature defined breast cancer	374
11.3.3	Sensitivity difference between genetic-features in breast cancer	376
11.3.4	Unsupervised clustering genetic-features in breast cancer by functional profiles	377
11.3.5	Top candidates genetic-features in breast cancer by functional profiles	377
11.4	Ongoing and Future Work	379
Chapter 12	Additional shSeq Applications	380
12.1	Overview	380
12.2	Therapeutic Targets for MYCN-amplified Neuroblastoma	380
12.3	Overcoming Cisplatin or PARP Inhibitor Resistance in Small Cell Lung Cancer	385

12.4	Genetic Modifiers of SMN as Therapeutic Targets for Spinal Muscular Atrophy.....	389
12.5	Positive shRNA Screens to Identify Novel Modulators of P53 Pathway	391
12.6	Overcoming Resistance to Glucocorticoid or NOTCH-inhibition in T-ALL	393
Chapter 13	A Dynamic Web System for Collaboration Management ...	397
13.1	Overview.....	397
13.2	Features	397
13.2.1	Dashboard.....	398
13.2.2	Group	400
13.2.3	User Roles.....	403
13.2.4	Document / Wiki	404
13.2.5	Calendar Event.....	406
13.2.6	Blog	408
13.2.7	Project and case tracker.....	409
13.2.8	Shoutbox / Twitter.....	410
13.2.9	Images.....	410
13.2.10	Notification and others	411
13.3	Insights into Biological Systems from Software Systems.....	412
Chapter 14	Conclusions	414
14.1	Key Contributions and Findings.....	414
14.1.1	NGS-based shRNA screening (shSeq): an analytical pipeline	414

14.1.2	Systems biology of cancer genomics: NetBID2	415
14.1.3	Successful studies of integrating functional genomics with systems biology for driver-type therapeutic target discovery	416
14.1.4	Collaboration model between computational and experimental biologists.....	417
14.2	Future Directions	417
References		420
Appendix A: High-throughput RNAi Screening: Experimental Approach .		438
1.1	Introduction.....	438
1.2	Materials	442
1.2.1	shRNA Library	442
1.2.2	Bacterial media.....	443
1.2.3	Antibiotics	444
1.2.4	Linear PEI.....	444
1.2.5	DNA precipitation.....	445
1.2.6	PCR primers sequence (Table 4)	445
1.2.7	Labeling.....	446
1.3	Experimental Procedures of a RNAi Screen.....	447
1.3.1	Library preparation	447
1.3.2	Virus production	449
1.3.3	Infection (Figure 2)	451
1.3.4	Infection efficiency test	453
1.3.5	Screening	455

1.3.6	Genomic DNA	458
1.3.7	Sequencing PCR.....	459
	References.....	464
Appendix B: Book chapter – Computational Analysis of High-throughput		
RNAi screening Data		466
	Summary.....	466
1	Introduction	467
2	Materials.....	470
3	Methods	471
4	Notes.....	473
	References.....	478
	Tables and Figures.....	480

List of Tables

Table 2-1 Distribution of Number of shRNAs per Gene	21
Table 2-2 Example of raw NGS data in QSEQ format	23
Table 2-3 Example of raw NGS data in FASTQ format.....	24
Table 2-4 QSEQ format of raw NGS data: column filed descriptions.....	25
Table 2-5 A summary table of deconvolution results for 6 shSeq runs. Each run contains 6 samples (T0 and T10 in replicate A-C), in which the total identified reads (the first rows in each run) and averaged count per shRNA (the second row in each run) are reported. Identification rate is the percentage of identified reads. The numbers in red indicate a case of low signals which might cause the data noisy. T10/T0 is the ratio of total identified reads at T10 and T0.....	36
Table 4-1 Top 30 NetBID2-predicted TF or Signaling drivers for Basal vs. Luminal breast cancer, HER2+ vs. HER2- breast cancer, and ABC vs. GCB-type of DLBCL. Genes in red are known drivers of corresponding diseases reported in literature. Duplicate gene names are for different probes or transcripts.....	115
Table 5-1 A collection of available tools and methods for functional pathway enrichment analysis. Adapted from Khatri, et al, 2012 [114]......	130
Table 6-1 The 13 selected differentially-expressed or drug-responsive apoptotic genes. *: pro: annotated by GO terms: induction of apoptosis, positive regulation	

of apoptosis, negative regulation of anti-apoptosis; anti: annotated by GO terms: negative regulation of apoptosis, positive regulation of anti-apoptosis. 154

Table 7-1 All predicted signaling drivers of GC-resistance by NetBID2 with $P < 0.01$, set size > 50 , being involved in \geq known pathway. 234

Table 7-2 Gene Expression Signature of Glucocorticoid Resistance (10 Resistant vs. 22 Sensitive Primary Samples, $P < 0.01$) 235

Table 7-3 Candidates of integrating top signaling of drivers of with RNAi screening results to reverse GC-resistance upon silencing. 243

Table 8-1 Overlapped 16 candidates between RNAi screening and genomic inference of regulatory drivers. 259

Table 8-2 Additional top 30 computationally-identified regulatory drivers of GC-resistance in T-ALL but with no support from RNAi screens. 272

Table 8-3 Top computationally-inferred regulatory drivers for GC-resistance in T-ALL ($P \leq 0.05$) 273

Table 8-4 Pooled shRNA screens: depleted genes in at least one cell line ($P < 0.05$) 274

Table 8-5 Top enriched KEGG pathways by depleted TF genes in RNAi screens and genomics-inferred TF drivers. 275

Table 9-1 NetBID2 inference results of three final candidates from integrative analysis. Duplicate names for GLRX represent two probes for GLRX in the microarray data. In functional type (Func Type), TF is for transcription factor, Sig

is for signaling molecule. The column of # of pathway indicates the number of known pathway from multiple databases the candidate gene is involved in. nES.comb.Drivers is the normalized enrichment score of combing 2D and 3D NetBID2 outputted nES. Pval.comb.Drivers is based on nES.comb. The GEP signature analysis columns show fold change (FC), z score and pvalue of differential expression analysis for 2D and 3D data.285

Table 9-2 Genome-wide shRNA screening results of three final candidates from integrative analysis. In # shRNAs column, 'All' means the number of hairpins present in all platforms including hairpin-probed microarray (SH.array), barcode-probed microarray (BC.array) and sequencing. In combine Array & Seq columns, n.Pval is the number of comparisons or evidences, n.sh.Depleted is the number of hairpins showing significant depletion ($P < 0.05$).286

Table 9-3 Summary of deconvolution for NGS data of shRNA screening on ERBB2+ and wild type MCF10A cells in 2D, 3D and in vivo systems. Cells with sky blue background are data for this study. Numbers in dark red background are cases with < 1M identified reads, in light read are cases with 1-5M reads.318

Table 10-1 Summary for genome-wide shRNA screens of four DLBCL cell lines. Green check sign indicates the data quality is good while red one represents that that data is not good or missing.327

Table 10-2 Summary of deconvolution of NGS-based shRNA screening data of four DLBCL cell lines. Red ones are the run with not enough signals. SUDHL4 was run three times to get good quality data.329

Table 10-3 Summary of enriched or overrepresented and depleted or under-represented genes in shRNA screening for each cell line. “Combined” is using Stouffer’s method to integrate all four cell lines. It’s based on gene level results with selection threshold of $P < 0.05$335

Table 10-4 Number of genes depleted (under) or enriched (over) in at least 1 or 2 or 3 or 4 cell lines, based on gene level results with selection threshold of $P < 0.05$335

Table 10-5 Top genes depleted in all cell lines, and depleted in ABC or GCB cell lines only. Red indicates depletion while green indicates enrichment. “n.shRNAs” is the number of shRNAs in NGS-based data and “n.shRNAs.array” is the number of hairpins in microarray data. “numPathways” is the number of known pathways being involved in.337

Table 10-6 Top pathways enriched in shRNA screening-identified candidates. Red means genes in the pathway are significantly enriched in the under-represented genes in that cell line, while green is for enrichment in over-represented side.339

Table 11-1 Characteristics of 16 breast cell lines with shRNA screening data. BaA=Basal A; BaB=Bsal B; Lu=luminal, ER/PR/HER2/TP53 status: ER/PR positivity, HER2 overexpression, and TP53 protein levels and mutational status (obtained from the Sanger web site; M=mutant protein; WT=wild-type protein) are indicated. Square brackets indicate that levels are inferred from mRNA levels alone where protein data is not available. A/B: A is from Neve, et al, Cancer Cell,

2006; B is from Kao et al, Plos One, 2009. AC=adenocarcinoma; AnCa=anaplastic carcinoma; C Sar = carcinoma sarcoma; DC=ductal carcinoma; F=fibrocystic disease; IDC=invasive ductal carcinoma; Inf=inflammatory; Met AC = metastatic adenocarcinoma PB=primary breast; RM= reduction mammoplasty; PE=pleural effusion; BR=Brain W=White; B=Black.....355

Table 11-2 Summary for deconvolution of shSeq data for 16 breast cell lines. Numbers in red are the samples that have < 5M identified reads. The last two columns are using the default Illumina filtering criteria.....357

Table 11-3 Overlapped candidates of NetBID2-predicted drivers and RNAi Screening identified lethal candidates for basal or luminal subtype. “nES”, normalized enrichment score, indicates the driver prediction strength. “z.DE” indicates the differential expression of the gene itself. “z.BaVSLu.geneLevel” is the z score of comparing shRNA screening profiles of basal with luminal cell lines. GNA14, ATP6V1G2 are potential therapeutic targets for luminal subtype, while E2F2 is for basal subtype.366

Table 11-4 Summary of deconvolution of shSeq data for six genetic-engineered models. Numbers in red are cases with < 5M total identified reads.372

Table 11-5 Number of significant depleted or enriched genes in each of six genetically-engineered models (left) and in at least 1 to 6 these models (right). P<0.05 is defined as significant.375

Table 12-1 Summary of deconvolution for shSeq data of MYCN-amplified and non-MYCN-amplified neuroblastoma samples under normoxia and hypoxia

enrionments. “OFF” is dox-off representing MYCN-amplification, while “ON” is for non-MYCN-amplificatiion. Numbers in red are the cases with low total number of identified reads.....381

Table 12-2 Summary of shSeq data for small cell lung cancer with and without Cisplatin or PARP inhibitor treatment. Cis=Cisplatin, PARP=PARP inhibitor, PARP.C was removed for further analysis because of its low signals. 386

Table 12-3 Summary of deconvolution of shSeq data for SMA project.390

Table 12-4 Deconvolution table of shSeq data for P53 positive screen project.391

Table 12-5 Deconvolution of shSeq data for glucocorticoid or NOTCH-inhibition resistance in T-ALL. The sequence run in dark red is the re-sequenced because of a technical flaw of previous run..... 393

Table 12-6 Number of significant ($P<0.05$) candidates in each of the five cases or in at least one to five cases of shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL. 395

List of Figures

Figure 2-1 An overview of the pipeline for shSeq data analysis including a series of software packages and algorithms.....	20
Figure 2-3 Equation of transforming base-calling error rate or probability (P) to Phred Quality Score (Q).....	26
Figure 2-4 Density distribution plot of overall quality scores of all reads. The overall quality score of each read is calculated by averaging Phred scores of all nucleotides inside. A strong peak on the right indicates good overall quality of the sequencing data.....	27
Figure 2-5 Cycle-based quality distribution (boxplot) of four different sequencing runs. The dashed short line “-“ at each cycle indicates the median quality score, and the dark blue region at each cycle represents the 95% interval of quality score. Outliers are denoted by “.” or by the light blue dots at each cycle.	28
Figure 2-6 Cumulative distribution plot of read count. A point on the curve indicates the portion of reads (the score on the y axis) that has less than or equal to a certain number of count (the number on the x axis, log10 transformed). Portions on the left are for low-frequent reads and portions on the right represents reads with high count number.	30
Figure 2-7 Cycle-based base calls or base count. N is for undermined bases, which should have a low-count curve in good-quality sequencing data.	31

Figure 2-8 Example of a bad sequence run shown by (A) density distribution of averaged quality scores (B) cycle-based boxplot of quality distribution and (C) cycle-based base calls or base count.34

Figure 2-9 Sequence structure decomposition of each shSeq read. The first 6 bases in blue are from barcodes of experimental design and the 22nt bases in red are from sequences of shRNA hairpins in the library, out of which 19 nucleotides in the middle are perfectly matched to the genome sequence.34

Figure 2-10 Scatter (A), density (B) and CDF (C) plots of data before normalization. (A) Scatter plots and correlations between biological replicates. Plots in the dialogue are density distributions of data in each replicate. Texts in the upper triangle cells indicate Pearson (the first number) and Spearman correlations.40

Figure 2-11 Scatter (A), density (B) and CDF (C) plots of data after scale normalization41

Figure 2-12 Scatter (A), density (B) and Cumulative distribution function (C) plots of data after normalization by replicates42

Figure 2-13 MA plot of shSeq data of multiple samples.44

Figure 2-14 Boxplot of shSeq count in each sample.45

Figure 2-15 Density plot of hairpin count in each sample.46

Figure 2-16 Heatmap of sample similarities.48

Figure 2-17 Hierarchical clustering of samples. Each row represents a condition while boxes on each rows are replicates under that condition. Dots on the upper left plot indicates where to split the three to obtain specific number of clusters, in which the yellow one is for the current plot; colors are for different clusters.49

Figure 2-18 PCA plot of samples.50

Figure 2-19 Variance-Mean dependence plot.51

Figure 2-20 Consistence of replicates for RNAi screening data by barcode-probed (left) and hairpin-probed microarray platforms.52

Figure 2-21 An example of bad shSeq run. The numbers on the bottom-right are total number of identified reads for each replicate. Low total numbers (in red) for replicate B and C reduces the signal representation, thus making the data noisy.53

Figure 2-22 A Bayesian linear Poisson model. Y is hairpin abundance in count, which follows a Poisson distribution with a log-link. X indicates the condition, e.g. T10 or T0. The coefficient of linear model β represents the magnitude of differential representation, and α is the intercept. The noise follows a Gaussian distribution with mean 0 and standard deviation σ . Priors for this model is a conjugate one, in which coefficients, β and α use a Gaussian distribution, and variance of noise σ^2 follows Inverse Chi-square prior[89].55

Figure 2-23 A Bayesian linear Gaussian model. Y is hairpin abundance in continuous value, which follows a Gaussian distribution. X indicates the condition, e.g. T10 or T0. The coefficient of linear model β represents the magnitude of

differential representation, and α is the intercept. The noise follows a Gaussian distribution with mean 0 and standard deviation σ . Priors for this model is a conjugate one, in which coefficients, β and α use a Gaussian distribution, and variance of noise σ^2 follows Inverse Chi-square prior.55

Figure 2-24 Distribution of z-scores indicating differential representation results from four different shSeq data sets. Positive z-score means enrichment of hairpins while negative is for depletion.57

Figure 2-25 Distribution of p-values indicating differential representation results from four different shSeq data sets.58

Figure 2-26 Heatmap of z-scores of significant depleted hairpins. Euclidian or correlation can be used for distance metrics and Wald method is suggested for hierarchical clustering.59

Figure 2-27 Heatmap of shSeq data of significant hairpins in different conditions. Euclidian or correlation can be used for distance metrics and Wald method is suggested for hierarchical clustering.....60

Figure 2-28 An example GSEA plot of pathway or GO gene sets in differentially-represented shRNAs. Y axis shows the z score of differential representation at shRNA level or gene level. The red dashed lines indicate normalized Enrichment Score (nES) and P value.63

Figure 2-29 Sensitivity analysis. Number of significant depleted hairpins (y axis) in each of 16 breast cell lines at different p-value cutoffs (x axis). Cell lines

represented by lines in different colors are classified into four groups represented by shape of dots.....65

Figure 2-30 Sensitivity analysis of looking at depleted, enriched or both hairpins in the panel of 16 breast cell lines.....66

Figure 2-31 Statistical comparisons of sensitivity analysis results of 16 breast cell lines in both depleted and enriched cases. A Student's t-test is used to do the comparisons.67

Figure 2-32 The percentage (y axis) of cell lines in each group share a certain number of common depleted hairpins (x axis) at different p-value cutoff (in various colors).....68

Figure 3-1 Outline of microarray-based shRNA screens to profile cell essential genes (A) experimental procedures and (B) analysis pipelines76

Figure 3-2 Separate-and-combine approach. (A) An example of three shRNAs targeting KPNB1 gene from MCF7 dataset is selected to illustrate this approach. A Bayesian linear model is fit into data of each shRNA respectively. Estimated parameters (fitted lines in red) and summary statistical metrics are displayed on bottom left of each shRNA plot. Z scores and p-values are calculated by Wald test using a standard Gaussian as null distribution. Individual shRNA scores as input for algorithms to combine them can be calculated by t (Student's t-statistic), z (z-statistic of β in linear model), β (the coefficient in the linear model), signal-noise ratio (mean difference of TX vs. T0 over sample standard deviation), logFC (logarithm of fold change of TX vs. T0) and diff.mean (mean difference of TX vs.

T0). (B) In the linear regression model under Bayesian framework, y_i indicates time point, TX or T0, and x_i represents shRNA abundance for sample i ; m is the sample size of the corresponding shRNA; noise follows a Gaussian distribution with mean 0 and variance σ^2 ; β is the parameter of interest, indicating the silencing effects on cell viability by the shRNA in consideration. As for priors, a two-variable multi-Gaussian is set for coefficients and an Inverse-Gamma is for variance.77

Figure 3-3 Modeling-all-together approach. Bayesian hierarchical modeling (A) The data of all shRNAs targeting one gene can be fit by a hierarchical model, in which the extra level is indexed by j , indicating the shRNA group the sample belongs to. Sample index i is up to n , the total number of samples for one gene; j is up to J , the number of shRNA classes. Parameter μ , a vector of slope and intercept, reflects the gene-level activity and allows variation for each shRNA class. Conjugate priors are set for parameters. (B) The model can be rewritten to a two-component mixture model in which “fixed effect” corresponds to gene-level behavior and “random effect” indicates the noise of each shRNA group.80

Figure 3-4 Modeling-all-together approach. Bayesian hierarchical modeling. A practical application of the Bayesian hierarchical model to the example in Figure 2 is summarized in the plot. Red solid line indicates fitted gene-level/fixed effects in the model. Estimated parameters and summary statistics including z-statistic and p-value are displayed on bottom middle. Each colored dashed line reflects individual activity of each shRNA class by adding random effect to fixed gene-level effect.....81

Figure 3-5 A different representation of the hierarchical model. A middle layer is introduced to indicate the shRNA level. All data points for one shRNA are clustered together, but all shRNAs targeting the same gene are fit in the same model with allowance of their internal difference.....82

Figure 3-6 Distribution of data quality (MRC: minimum replicate correlation) for the panel of 72 shRNA screens: High (MRC>0.9): 22%, Medium (MRC in 0.8-0.9): 50%, Low (MRC<0.8): 28%.85

Figure 3-7 Quality of benchmark datasets. Each dataset has two time points (T0 and TX) and three replicates (A, B and C). Each sub-figure displays the scatter plots on bottom left and correlations (Pearson_Spearman) on upper right between any two replicates of the corresponding group, and density distribution of shRNA abundance in each replicate on the diagonal. Low variability of scatter plots, high correlations and high similarity of distribution plots indicate good consistence of replicates, thus good quality of the data. The label (High, Medium or Low) after each cell line name indicates the data quality group it belongs to, defined by MRC (minimum replicate Pearson correlation), the “bottle-neck” of each dataset. MCF7 (MRC > 0.9), HPAFII (MRC between 0.8-0.9), and OVCAR5 (MRC < 0.8) represent 22%, 50% and 28% of 72 screens respectively.87

Figure 3-8 Evaluation results of final time point data. Housekeeping or conserved ortholog genes can be used as reference gene set to evaluate algorithms to detect essential genes from RNAi screens. Each colored curve shows the percentage of each reference set (“name”_“number of genes in the set”) intersected by top 0 to 1000 hits predicted as essential genes by the

corresponding algorithm in each dataset. The slope of “Random” method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is.
.....88

Figure 3-9 Evaluation results using middle time point data. Housekeeping or conserved ortholog genes can be used as reference gene set to evaluate algorithms to detect essential genes from RNAi screens. Each colored curve shows the percentage of each reference set (“name”_“number of genes in the set”) intersected by top 0 to 1000 hits predicted as essential genes by the corresponding algorithm in each dataset. The slope of “Random” method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is.
.....90

Figure 4-1 Heatmap of example for gene expression signature genes of two phenotypes.98

Figure 4-2 Example of that master regulators are more robust than gene expression signature genes. Two groups were studying the same disease problem, metastasis in breast cancer, and each of them identified a group of signature genes, but there is only one overlap. However, in our predicted drivers for each of the datasets, the overlap improves dramatically from 1 out of 70 to 6 out of 20.99

Figure 4-3 Example of that gene expression signature is not robust. Signature genes ($z > 1.96$ or < -1.96) were plotted for both 2D (x axis) and 3D (y axis). Pearson and Spearman correlations are calculated on the top. 100

Figure 4-4 Individual examples of that gene expression signature is not robust. The first two genes are over-expressed in mutated cells with 2D culture, but are under-expressed in mutated cells with 3D culture. The last two genes are on opposite. 101

Figure 4-5 A social or non-scientific example to explain NetBID2 algorithm. 103

Figure 4-6 The NetBID2 framework. Step 1 uses reverse-engineering algorithm, ARACNe to reconstruct TF or Signaling centered networks from gene expression profiles. Step 2 utilizes phenotype information to perform signature analysis using a Bayesian Probit model approach. Step 3 applies gene set enrichment analysis for each driver candidate by taking its first neighbors as a gene set and using signature analysis results of all genes as the reference..... 108

Figure 4-7 The Bayesian Probit Model for gene expression signature analysis. Distribution details about the model is on the left and on the right is a graphical representation of the Probit model. Nodes in solid square are observation variables, in solid eclipse with white background are direct parameters of Probit model, in dashed eclipse are latent variables and the others are hyper-parameters for priors. Y is an indicator variable for phenotypes, X is expression level of gene X, Z is a latent variable in Probit model. Inside the white box is

likelihood section, while outside is for priors. Parameters are estimated by a Gibbs sampling procedure. 110

Figure 4-8 Venn diagram of NetBID2 inferred drivers (both TF and Signaling factors) from 2D or 3D expression data of ERBB2 mutated MCF10A cells. Fisher’s exact test is used to test the significance of overlaps. Total number is the number of probes for TF or signaling factors in the microarray data. 113

Figure 4-9 NetBID2 identifies AKT1 as a driver for glucocorticoid resistance in T-ALL (left), but there is no evidence from expression of AKT1 itself. 116

Figure 4-10 NetBID2 identifies STAT3 as a driver for ERBB2+ breast cancer (red line), but there is no evidence from expression of STAT3 itself (blue line). 117

Figure 4-11 Validation results by siRNA for top 30 NetBID2-predicted TF drivers of glucocorticoid resistance in T-ALL. (A) top 30 candidates (in red) together with positive controls (in blue) and negative controls (in green) are ranked by the score (central dot) for capability to reverse GC-resistance upon silencing with uncertainty (range line crossing the central dot, thick line for one standard deviation, thin line for two standard deviations corresponding to 95% confidence interval). The color of candidate label on x axis is associated with calibrated p-value: dark red for $P < 0.005$, red for $P \approx 0.05$. (B) Bar plots of apoptosis level induced by combined treatment of RNAi with DEX (in red), and control, RNAi with DMSO (in light blue) for 30 predicted candidates, positive controls (labeled in red) and negative controls (labeled in blue). All genes are ranked the same as in panel A. The label on top of bar plot represents the increased apoptosis level of

candidate gene comparing with average of negative controls (normalized by its own DMSO control and averaged over triplicates) and associated statistical significance level (** for $P < 0.01$, * for $P \approx 0.05$)..... 118

Figure 4-12 Enrichment of ARACNE-predicted targets of STAT3 in experimentally-identified targets. TF is for predicted targets in transcription regulatory network. Sig is predicted targets from signaling network. Three sets of inferred targets are for different transcripts or probes of STAT3..... 121

Figure 4-13 Enrichment of ARACNE-predicted positive targets of STAT3 in experimentally-identified targets. Positive is defined by the positive correlation between the target and STAT3 expression. TF is for predicted targets in transcription regulatory network. Sig is predicted targets from signaling network. Three sets of inferred targets are for different transcripts or probes of STAT3. Green check sign indicates $P < 0.05$ 122

Figure 4-14 Enrichment of ARACNE-predicted negative targets of STAT3 in experimentally-identified targets. Negative is defined by the negative correlation between the target and STAT3 expression. TF is for predicted targets in transcription regulatory network. Sig is predicted targets from signaling network. Three sets of inferred targets are for different transcripts or probes of STAT3. The check sign indicates non-significant..... 123

Figure 4-15 Number of target size for STAT3 (three probes at x axis) from TF (blue) or signaling (red)-centered network predicted by ARACNe. 125

Figure 4-16 Number of targets (in percentage, upper panel) from TF (blue) or signaling (red)-centered network predicted by ARACNe that are overlapped with top experiment-identified targets for STAT3 (three probes in three columns) and odds ratio of Fisher's exact test for the overlap (lower panel). The higher the overlap is, or the higher the odds ratio is, the more powerful or more precise the prediction is..... 127

Figure 5-1 Three major categories of functional enrichment analysis methods. Adapted from Khatri, et al, 2012 [114]. 129

Figure 5-2 "Maxmean" statistic developed by Efron. The genes in the set (blue bars on the bottom) are divided into positive (red on the right) and negative (blue on the left) according to the sign of their individual scores between phenotype 1 and phenotype 2 (y axis). The adjusted mean (divided by the size of the entire set) of each subset is calculated, and the one with maximum absolute value is used as the enrichment score for this set. 132

Figure 5-3 Efron's simulation results (sensitivity vs. specificity) on comparison of Maxmean statistic with KS-based GSEA. Adapted from Figure 8 in Efron, et al, 2007 [97, 98]. 133

Figure 5-4 Restandarization technique for statistical significance estimation. Adapted from Efron, et al, 2007 [97, 98]. 134

Figure 5-5 A Bayesian linear Gaussian model for individual gene scoring with Gaussian or t distribution as prior for coefficients and inverse Chi-square or Gamma distribution as prior of noise variance..... 136

Figure 5-6 “Maxmean” statistic (GSA) vs. KS statistic (GSEA) for summarization of enrichment score. Housekeeping or conserved ortholog genes can be used as reference gene set to evaluate algorithms to detect essential genes from RNAi screens. Each colored curve shows the percentage of each reference set (“name”_“number of genes in the set”) intersected by top 0 to 1000 hits predicted as essential genes by the corresponding algorithm in each dataset. The slope of “Random” method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is. 138

Figure 5-7 Evaluation results of BSEA, GSA and GSEA. Annotation is the same with Figure 62. 139

Figure 5-8 Evaluation results of BSEA vs. GSEA. Annotation is the same with Figure 62..... 140

Figure 5-9 Evaluation results of Bayesian vs. Frequentist methods for individual scoring. “Maxmean” is used for enrichment score. Annotation is the same with Figure 62..... 141

Figure 5-10 Evaluation results of BHM (Bayesian Hierarchical Model), BSEA, and GSEA. Annotation is the same with Figure 62. 142

Figure 6-1 Heatmap of top differentially-expressed genes (FDR<0.05) in drug-perturbed and control samples. The genes are ranked from most up-regulated (labeled in dark red on right panel) to most down-regulated (labeled in dark green)

in drug-perturbed samples, and the 13 selected apoptotic genes are labeled on the right with their ranks in the list..... 150

Figure 6-2 The heat map of distances between profiles of CMAP data including randomly-selected 100 control and 100 drug-perturbed samples. 151

Figure 6-3 Summary of (A) Fisher’s Exact Test and (B) Gene Set Enrichment Analysis (GSEA) to test whether apoptosis pathway with 368 apoptotic genes is enriched in drug-induced signature genes. For GSEA method, absolute mean was used to summarize the enrichment and 10,000 gene permutations were used to produce the significant level. 155

Figure 6-4 Summary of (A) Fisher’s Exact Test and (B) Gene Set Enrichment Analysis (GSEA) to test whether 207 pro-apoptotic genes are enriched in drug-induced signature genes. 156

Figure 6-5 Summary of (A) Fisher’s Exact Test and (B) Gene Set Enrichment Analysis (GSEA) to test whether 185 anti-apoptotic genes are enriched in drug-induced signature genes. 157

Figure 6-6 Marginal distributions of the 13 selected drug-responsive apoptotic genes across all samples in CMAP data..... 159

Figure 6-7 (A) Histogram and (B) Box plot of scores for best-learned graphical model in each bootstrapped sampling. 162

Figure 6-8 (A) Predicted subnetwork of 13 selected drug-responsive apoptotic genes: edges in red are validated interactions in literature, and edges in dark red are strong validated direct interactions. (B) A subnetwork from literature showing

evidences for validated interactions in predicted network including candidate genes (colored in yellow) with their validated interactants (in brown). Each validated edge in predicted network (red in A) can be mapped to one path in evidence network (B) between the two corresponding interacting candidate genes..... 167

Figure 6-9 A network from literature for 13 candidate genes (colored in yellow) with their validated interactants (in brown). Edges in red are evidences for validation of interactions in predicted apoptosis network. 168

Figure 7-1 Glucocorticoid receptor signaling pathway, adapted from SABiosciences..... 176

Figure 7-2 NetBID2 algorithm to identify drivers of GC-Resistance in T-ALL from gene expression profiles. 180

Figure 7-3 Top signaling drivers of GC-resistance inferred by NetBID2 and siRNA validation results. (a) Signaling factors associated with glucocorticoid resistance by NetBID2. (b) Apoptosis analysis in DND41 T-ALL cells electroporated with siRNA pools targeting validated candidate regulators of glucocorticoid resistance and treated with dexamethasone (1 μ M) for 48 hours. The apoptotic index indicates apoptotic cell number in gene specific siRNA dexamethasone treated samples relative to siRNA control dexamethasone treated cells. *, P <0.01; **, P <0.05. (c) Quantitative RT-PCR analysis of siRNA knockdown. 182

Figure 7-4 AKT1-subnetwork predicted by ARACNe. Out of 30 (92 genes in total) TFs (diamond shape) or signal molecules (circle shape) that are predicted to

connect with AKT1, 15 as shown are also inferred as drivers of GC-resistance. The strength of evidence (p-value) as a driver is color coded. Three signaling proteins in red are among top 9 proteins selected for validation. B3GAT3 is also confirmed to reverse GC-resistance by siRNA. 185

Figure 7-5 NetBID2 identifies AKT1 as a driver of GC-resistance in T-ALL..... 186

Figure 7-6 NetBID2 identifies AKT2 as a driver of GC-Resistance in T-ALL..... 186

Figure 7-7 mRNA expression of AKT1 in GC-resistant and GC-sensitive primary samples. AKT1 is slightly over-expressed in sensitive samples. 187

Figure 7-8 Activation of the PI3K-AKT signaling pathway via PTEN inactivation induces glucocorticoid resistance in T-ALL and blunts glucocorticoid-induced gene expression. (a) Western blot analysis of PTEN expression and AKT1 activation in DND41 T-ALL cells expressing a shRNA targeting the PTEN tumor suppressor (shRNA PTEN) compared to control cells expressing a hairpin against luciferase (shRNA LUC). (b,c) Representative plots (b) and quantification (c) of glucocorticoid-induced apoptosis in control and PTEN knockdown DND41 cells treated with dexamethasone (1 μ M) for 48 hours. Percentages of viable (lower left quadrant), apoptotic (lower right quadrant) and dead (upper right quadrant) are indicated. (d) RT-PCR analysis of glucocorticoid response gene induction in control and PTEN knockdown DND41 cells treated with dexamethasone. (e,f) Luciferase reporter analysis of dexamethasone-induced glucocorticoid receptor transactivation in U2OS cells expressing MYR-AKT1 compared with GFP only expressing controls using a synthetic glucocorticoid

response element reporter (e) and the glucocorticoid receptor promoter 1A FP11-FP12 regulatory sequence (f). 189

Figure 7-9 Inactivation of AKT by siRNA facilitates glucocorticoid-induced gene expression. RT-PCR analysis of glucocorticoid response gene induction in control and AKT1 knockdown DND41 cells treated with dexamethasone..... 190

Figure 7-10 AKT1 interacts with and directly phosphorylates the glucocorticoid receptor protein in position S134. (a) Western blot analysis of AKT1 after glucocorticoid receptor NR3C1 immunoprecipitation in 293T cells expressing Flag-tagged AKT1 and HA-tagged NR3C1. (b) Western blot analysis of glucocorticoid receptor NR3C1 protein after AKT1 immunoprecipitation in 293T cells expressing Flag-tagged AKT1 and HA-tagged NR3C1. (c) Western blot analysis of AKT1 after NR3C1 protein immunoprecipitation in DND-41 T-ALL cells. (d) Analysis of AKT1-NR3C1 interaction via AKT1 detection via Western blot analysis of protein complexes recovered after NR3C1-GST pull down of recombinant His-tagged AKT1. (e) Partial alignment of the glucocorticoid receptor protein sequence flanking S134. (f) Western blot analysis of NR3C1 phosphorylation with an anti AKT phospho-motif specific antibody in NR3C1 protein immunoprecipitates from U2OS cells expressing MYR-AKT1 together with HA-tagged wild type glucocorticoid receptor (HA-NR3C1) or an HA-tagged glucocorticoid receptor protein harboring a serine 134 to alanine substitution (HA-NR3C1 S134A). (g) In vitro kinase analysis of AKT1 phosphorylation of recombinant wild type NR3C1 (GST-NR3C1) and NR3C1 S134A mutant (GST-NR3C1 S134A) protein. Top panel shows P³² autoradiography after SDS-PAGE.

The corresponding protein loading for each reaction is shown in the Coomassie blue staining micrograph at the bottom. (h) ESI-MS/MS spectrum of monophosphorylated peptide STpS134VPENPK (S-132 to K-140) obtained after trypsin digestion of NR3C1 isolated from cells expressing constitutively active AKT1. (i) Collision induced dissociation of the molecular ion, $[M+2H]^{2+}$ at m/z 519.72 ($M = 1037.42$ Da) corresponding to S134. Characteristic b- and y-fragment ions including y7 which contains pSer and features the loss of 98 Da (elimination of phosphoric acid) are shown..... 193

Figure 7-11 AKT1 directly interacts with the glucocorticoid receptor in T-ALL cells. Western blot analysis of AKT1 after NR3C1 protein immunoprecipitation in CCRF-CEM cells..... 195

Figure 7-12 AKT1 can directly interact with both wild type and mutant S134A glucocorticoid receptor. Analysis of AKT1-NR3C1 interaction via AKT1 detection via Western blot analysis of protein complexes recovered after wild type (NR3C1-GST) or mutant (NR3C1 S134A-GST) glucocorticoid receptor GST pull down with recombinant His-tagged AKT1..... 196

Figure 7-13 AKT1-mediated S134 phosphorylation of the NRC3C1 protein impairs dexamethasone-induced glucocorticoid receptor nuclear translocation. (a) Confocal microscopy analysis and quantitation of the cellular distribution of NR3C1 cellular localization in U2OS cells expressing HA-NRC31 in basal conditions (DMSO) and after dexamethasone (Dexa) stimulation. (b) NR3C1 cellular localization in U2OS cells expressing HA-NRC31 and MYR-AKT1 in basal conditions and after dexamethasone stimulation. (c) Cellular localization of

NR3C1 in U2OS cells expressing the HA-NRC31 S134A mutant in basal conditions and after dexamethasone stimulation. (d) Cellular localization of the HA-NRC31 S134A protein in U2OS cells expressing MYR-AKT1 in basal conditions and after dexamethasone stimulation. (e) Cellular localization analysis of NR3C1 via nuclear and cytoplasmic cell fractionation and analysis of AKT1 signaling in cell lysates from CCRF-CEM T-ALL cells treated with vehicle only (DMSO), dexamethasone (Dexa), the MK2206 AKT inhibitor and MK2206 plus dexamethasone. Tubulin and MAX proteins are shown as controls for cytosolic and nuclear fractions, respectively. C: cytoplasmic fraction; N: nuclear fraction. 198

Figure 7-14 U2OS cells do not express detectable levels of endogenous NR3C1. Western blot analysis of NR3C1 expression in U2OS cells expressing pMSCV empty vector, pMSCV-HA NR3C1 or pMSCV-HA NR3C1 S134A. 200

Figure 7-15 AKT1-mediated phosphorylation of the NR3C1 protein impairs dexamethasone-induced glucocorticoid receptor nuclear translocation in T-ALL cells Cellular localization analysis of NR3C1 via nuclear and cytoplasmic cell fractionation and analysis of AKT1 signaling in cell lysates from MOLT-3 T-ALL cells treated with vehicle only (DMSO), dexamethasone (Dexa), the MK2206 AKT inhibitor and MK2206 plus dexamethasone. Tubulin and MAX proteins are shown as controls for cytosolic and nuclear fractions, respectively. C: cytoplasmic fraction; N: nuclear fraction. 201

Figure 7-16 Pharmacological inhibition of AKT synergizes with dexamethasone to increase the antileukemic effects of glucocorticoids in DND-41 T-ALL cells. (a)

Western blot analysis of AKT1 activation in DND41 T-ALL cells treated with the MK2206 AKT inhibitor. (b) Isobologram representation of cell viability results and Combination Index analysis of DND41 cells treated with dexamethasone and MK-2206 in combination.204

Figure 7-17 Pharmacologic inhibition of AKT with MK-2206 reverses glucocorticoid resistance in human T-ALL cell lines. (a) Representative plots and quantification of apoptosis and loss of cell viability in CCRF-CEM T-ALL cell line treated with vehicle only, MK2206, dexamethasone or dexamethasone plus MK2206 in combination in vitro. (b) Quantification of tumor load by bioluminescence in in vivo imaging and analysis of luciferase activity or human CD45 expressing cells in the bone marrow of CCCF-CEM T-ALL xenografted mice treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206).205

Figure 7-18 Pharmacological inhibition of AKT reverses glucocorticoid resistance in MOLT-3 T-ALL cells (a,b) Representative plots (a) and quantification (b) of apoptosis and cell viability (c) in MOLT-3 T-ALL cells for 72 hours with vehicle only, MK2206 and dexamethasone alone or dexamethasone plus MK2206 in combination.206

Figure 7-19 Pharmacologic inhibition of AKT with MK-2206 reverses glucocorticoid resistance in human T-ALL primary samples. (a,b) Representative plots (a) and quantification of loss of viability analysis (b) in primary T-ALL patient samples treated with vehicle only, MK2206 and dexamethasone alone or dexamethasone plus MK2206 in combination. Percentages of viable (PI -), and

non-viable (PI +) cells are indicated. (c-f) Representative examples of primary human T-ALL xenografted mice showing changes in tumor load assessed by *in vivo* imaging (c), spleen size (d), spleen weight (e) and luciferase activity in bone marrow cells (f) from primary human leukemia xenografted mice treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206). Scale bar: 2 cm.....207

Figure 7-20 Pharmacological inhibition of AKT *in vitro* reverses glucocorticoid resistance in primary human T-ALL xenografts. Analysis of cell viability in primary T-ALL samples treated for 72h with vehicle only, MK2206 and dexamethasone alone or dexamethasone plus MK2206 in combination.....208

Figure 7-21 Pharmacological inhibition of AKT *in vivo* reverses glucocorticoid resistance in primary human T-ALL xenografts. (a,b) Bioimaging quantification (a) and analysis (b) of tumor load changes in mice treated with vehicle (control), dexamethasone (Dexa), MK2206, MK2206 plus dexamethasone (Dexa + MK2206) for 5 days.....209

Figure 7-22 Pharmacological inhibition of AKT *in vivo* reverses glucocorticoid resistance in primary human T-ALL xenografts. (a,b) Representative images of spleens (a) and spleen weights (b) of leukemic mice treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206) for 4 days. (c,d) Quantification of tumor load by determining luciferase activity from cells isolated from the spleen (c) and bone marrow (d). (e) Quantification of tumor load changes by determining the increase in circulating CD45 positive cells in the peripheral blood of mice injected with a human

xenograft and treated with vehicle (control), dexamethasone (Dexa), MK2206, MK2206 plus dexamethasone (Dexa + MK2206) for 4 days. Scale bar: 2cm...210

Figure 7-23 Pharmacologic inhibition of AKT reverses glucocorticoid resistance in a mouse model of glucocorticoid resistant T-ALL. (a,b) Kaplan-Meier survival plot in mice treated with dexamethasone (Dexa) or vehicle (Control) after allograft transplantation of Pten-non-deleted [-Tmx (Pten f/f)] (a) or Pten-deleted [+Tmx (Pten -/-)] (b) NOTCH1-induced T-ALL tumor cells. Arrows indicate the time of drug treatment. (c,d) Representative images and changes in bioluminescence in vivo imaging (c) and analysis of treatment response in mice allografted with NOTCH1 induced Pten deleted mouse leukemia cells and treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206). (e) Kaplan-Meier plot of overall survival in mice allografted with NOTCH1 induced Pten deleted mouse leukemia cells and treated with vehicle only (control), MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206). (f) Quantification of glucocorticoid-induced loss of viability in NOTCH1 induced Pten non deleted [-Tmx (Pten f/f)] or Pten deleted [+Tmx (Pten -/-)] mouse leukemia cells infected with an empty vector control (MSCV-puro) or retroviruses expressing the wild type glucocorticoid receptor NR3C1 (MSCV HA-NR3C1) or the S134A glucocorticoid receptor NR3C1 mutant protein (MSCV HA-NR3C1 S134A).213

Figure 7-24 Overexpression of NR3C1 and NR3C1 S134A mutant in primary murine leukemia cells. (a) Western blot analysis determining the retroviral expression of NR3C1 or NR3C1 S134A in Pten non-deleted [-Tmx (Pten f/f)] and

(b) PTEN deleted [+Tmx (Pten -/-)] NOTCH1-induced T-ALL mouse leukemia cells. Beta actin is shown as loading control.214

Figure 7-25 NetBID2 results of top signaling drivers of GC-resistance in T-ALL.233

Figure 7-26 Summary of GC-Responsive signature at 6h and 24h.236

Figure 7-27 Comparison of drivers regulating or modulating early (6 or 8h) and late (24h) responsive signature genes in GC-sensitive T-ALL patients. Sign of drivers is taken into consideration. Fisher’s exact test is used to the significance of overlaps.237

Figure 7-28 Overlap of TF master regulators for GC-Resistance and GC-Response (6h or 24h). Fisher’s exact is used to test overlap significance.238

Figure 7-29 NetBID2 identifies AKT1 (P<0.001) and AKT2 (P=0.024) as drivers of early (6 or 8h) GC-responsive signature genes, but not as drivers of late (24h) responsive signature genes.239

Figure 7-30 shRNA screening results of AKT1 and AKT2 in two GC-resistant cell lines.241

Figure 7-31 PRKAR1A shows up in both NetBID2 prediction and shRNA screens.242

Figure 8-1 The framework integrating genetic RNAi screen with genomic inference of phenotype drivers to identify therapeutic targets for reversal of GC-resistance upon repression.248

Figure 8-2 Distribution setup (left) and graphical representation (right) of Probit model used for assessing association of phenotypes (GC-resistant or GC-sensitive) with gene expressions. Nodes in solid square are observation variables, in solid eclipse with white background are direct parameters of Probit model, in dashed eclipse are latent variables and the others are hyper-parameters for priors. Y is an indicator variable for phenotypes, X is expression level of gene X, Z is a latent variable in Probit model. Inside the white box is likelihood section, while outside is for priors. Parameters are estimated by a Gibbs sampling procedure.250

Figure 8-3 Summary of candidates from RNAi screening or genomic inference. (A) Only 81 out of 126 genomics-analysis identified regulatory drivers have shRNAs in RNAi screening. B-D display the distribution of 81 TF drivers (blue bars) and 26 validated ones (red bars) in RNAi screening results. All 12,049 genes are ranked from most depleted (left) to most enriched (right) using differential representation score (z score) at gene level in combination of two cell lines (B) or individual cell line (C-D). Similar summary by considering only TF genes in RNAi screening is shown in Figure 110.....254

Figure 8-4 Similar summary with Figure 109, but considering only TF genes in RNAi screen.....255

Figure 8-5 Validation results *in vitro* of 16 overlapped candidates. (A) 16 candidates (in red) together with positive controls (in blue) and negative controls (in green) are ranked by the score (central dot) for capability to reverse GC-resistance upon silencing with uncertainty (range line crossing the central dot,

thick line for one standard deviation, thin line for two standard deviations corresponding to 95% confidence interval). The color of candidate label on x axis is associated with calibrated p-value: dark red for $P < 0.005$, red for $P \approx 0.05$. (B) Bar plots of apoptosis level induced by combined treatment of RNAi with DEX (in red), and control, RNAi with DMSO (in light blue) for 16 predicted candidates, positive controls (labeled in red) and negative controls (labeled in blue). All genes are ranked the same as in panel A. The label on top of bar plot represents the increased apoptosis level of candidate gene comparing with average of negative controls (normalized by its own DMSO control and averaged over triplicates) and associated statistical significance level (***) for $P < 0.005$, * for $P \approx 0.05$).257

Figure 8-6 Validation results *in vitro* of top 30 additional genomics-predicted regulatory drivers of GC-resistance. (A) 30 candidates classified into no data (in brown-yellow), no significant evidence (in blue) and significantly over-represented (in purple) from RNAi screen, together with positive controls (in blue) and negative controls (in green) are ranked by the score for capability to reverse GC-resistance upon silencing with uncertainty. Extra annotations in panel A and B are the same with Figure 111.261

Figure 8-7 RNAi screening results and GSEA plots of CC2D1A and WHSC1, the top two validated targets *in vitro*.....263

Figure 8-8 RNAi screening result and GSEA plot of ATF6, showing the strongest effect on increasing resistance *in vitro* when silenced.263

Figure 8-9 Subnetwork from T-ALL interactome of candidates that are validated *in vitro* either to increase (blue nodes) or decrease (red nodes) sensitivity when silenced. A, B and C are three well-connected components covering only direct interactions among these candidates. Nine isolated effective candidates that have no direct interactions with other candidates are not shown. The size of node is proportional to the size of its regulons or first neighbors from T-ALL interactome. Edge in red is for positive correlation of two interactants, while blue for negative correlation.265

Figure 8-10 A TRIM28-centered subnetwork for a novel mechanism of GC-resistance and a synergistic strategy to overcome resistance. The left square-like part is extracted from Fig 5.1 including the top 3 best validated targets (blue nodes) and the clique of TRIM28 (red nodes). The right triangle-like part is from T-ALL regulatory network illustrating the mechanism of TRIM28 upregulating NR3C1 via a feed forward loop. Number on edge represents the mutual information between expression levels of two interactants. Dashed edges are recovered from false removals by DPI in ARACNe³⁵ algorithm.266

Figure 8-11 Validation results *in vitro* of all 46 selected genomics-predicted regulatory drivers of GC-resistance. (A) 46 candidates classified into no data (in brown-yellow), no significant evidence (in blue) and significantly over-represented (in purple) from RNAi screen, together with positive controls (in blue) and negative controls (in green) are ranked by the score for capability to reverse GC-resistance upon silencing with uncertainty. Extra annotations in panel A and B are the same with Figure 111.271

Figure 9-1 The integrative framework of genome-wide RNAi screening with systems biology of cancer genomics (NetBID2) to identify therapeutic targets of ERBB2+/HER2+ breast cancer.....282

Figure 9-2 Integration of RNAi screening with NetBID2 driver prediction identifies three candidates as driver-type therapeutic targets for ERBB2F+ breast cancer. The left is a summary table of candidate selection by combing evidences from both NetBID2 driver predictions (blue background) and RNAi screening (green background). For driver prediction, we have candidates based on 2D or 3D signature of ERBB2+ cells, and we also combine evidences of 2D drivers and 3D drivers, finally 137 TFs or signaling molecules are overlapped among the three driver lists. For RNAi screening, we have combined microarray or sequencing results alone in best.comb (combine hairpin from different datasets first and then select the best as representative), comb.gene (combined gene level by BHM algorithm), and comb.best (select best hairpin in each data set first and then combine them), and combined both microarray and sequencing results. Only 36 genes came out from RNAi screening. On the firth venn diagram, crossing 137 drivers and 36 candidates from RNAi screening, only three genes show up, which is not happening randomly based on Fisher's exact test.284

Figure 9-3 Western blots of phosphorylated-STAT3 (P-Stat3) and phosphorylated-STAT5 (P-Sat5) in wild type, and genetically-engineered (ERBB2+, CYCD1+, E1A+, PTEN-, and p53-) MCF10A cells. pSTAT3 and pSTAT5 are only activate in ERBB2+ MCF10A cells.....287

Figure 9-4 Validation of STAT3 *in vitro* by siRNA and shRNA. Viability assays were performed after knock-down of STAT3 by siRNA or two shRNAs.289

Figure 9-5 Validation of STAT3 *in vitro* by colony forming cell (CFC) assays with shRNA silencing. Colony assays were performed in wild type MCF10A cells with and without STAT3 silencing by shRNA, ERBB2-mutated MCF1-A with and without shSTAT3. Quantitate cell counts were measured for each colony assay.290

Figure 9-6 Validation of STAT3 *in vivo*. Xenograft mouse models were made for ERBB2+ MCF10A cells with and without STAT3 silencing by shRNA. Cell population with tumor marker were imaged and measured weekly, up to 7 weeks, when tumor size was photographed.291

Figure 9-7 Western blots of STAT3, phosphorylated STAT3, ERBB2, phosphorylated ERBB2 in different breast cancer cell lines: wild type MF10A, ERBB2+ MCF10A, three Luminal lines (SKBR3, MDAMB361, ZR7530), four Basal lines (MDAMB231, HC70, HCC1954, SUM190PT).....293

Figure 9-8 *In vitro* and *in vivo* validation of STAT3 specificity to ERBB2 using MDAMB231 cell line. MDAMB231 is a ERBB2- but STAT3+ line. *In vitro* competition assays and *in vivo* xenograft mouse models with and without STAT3 silencing by shRNA were performed to measure viability of tumor cells or tumor growth.294

Figure 9-9 Illustration of genetically-engineered isogenetic model. ERBB2 is overexpressed genetically in MCF10A cells to mimic ERBB2+ breast tumors..295

Figure 9-10 STAT3 doesn't show up as a driver of HER2+ primary samples in both transcription factor (TF)-centered network analysis and signaling protein (Sig)-focused network analysis.296

Figure 9-11 NetBID2 results of STAT3 on ER- and ER+ groups of HER2+ population. STAT3 only shows up in ER- group as a driver.297

Figure 9-12 NetBID results of STAT3 in Luminal and Basal subtype of HER2+ patients. STAT3 shows up in Luminal group but shows opposite direction pattern to the one in MCF10A model and ER- subgroup.298

Figure 9-13 Heatmap of top STAT3 targets from perturbation experiments by knocking-down STAT3 by two shRNAs or over-expressing IL6, an upstream activator of STAT3. Red stands for down-regulation while green means up-regulation, for example, in sh-STAT3 experiments (the left four samples), genes in red are under-expressed when silencing STAT3 or are potential positive targets of STAT3, while green ones are potential negative targets of STAT3. SOCS3 is a known target activated by STAT3 and STAT3 itself is another positive control.300

Figure 9-14 Subnet of STAT3 predicted by ARACNe in the signaling-centered network. Genes on the right circle are transcription factors (TF in square shape), signaling molecules (Sig in diamond shape) or both (TF_Sig in hexagon shape). Genes in dark green are also predicted as master regulators or drivers (MR) of HER2+ breast cancer. Genes on the left circle are the ones in general. Red edge is for positive correlation while blue is for negative correlation.301

Figure 9-15 Enrichment of ARACNE-predicted targets of STAT3 from transcription factor (TF)-centered or Signaling molecule (Sig)-centered network in experimentally identified targets of STAT3 by microarray profiling after knocking down STAT3 or overexpressing IL6 (activator of STAT3).302

Figure 9-16 Enrichment of STAT3 targets from perturbation experiments in shRNA screening results of ERBB2+ MCF10A cells. Reference genes are ranked from the most enriched to most depleted in ERBB2+ vs. wild type MCF10A cells. STAT3 targets are selected by $P < 0.05$ and $FC > 3$. Positive targets are defined as positive expression in STAT3-induced samples comparing with expression in STAT3-silenced samples. Negative targets are defined in the opposite way. Top lethal positive or negative targets are listed in the boxes on the bottom-right...304

Figure 9-17 Enrichment of STAT3 targets from perturbation experiments in shRNA screening results of ERBB2+ breast cancer cell lines (SKBR3, SUM190PT, MDAMB361). Reference genes are ranked from the most enriched to most depleted in three HER2+ cell lines (using a combined score). STAT3 targets are selected by $P < 0.05$ and $FC > 3$. Positive targets are defined as positive expression in STAT3-induced samples comparing with expression in STAT3-silenced samples. Negative targets are defined in the opposite way. Top lethal positive or negative targets are listed in the boxes on the bottom-right. Yellow ones are the common targets lethal to ERBB2+ MCF10A cells (Figure 133). ...305

Figure 9-18 STAT5A is predicted as a driver of ERBB2+ MCF10A cells.306

Figure 9-19 STAT5A like STAT3 is not a driver of HER2+ primary tumors.....307

Figure 9-20 STAT5A is a driver of ER- but not ER+ group of HER2+ primary patients.307

Figure 9-21 Subnet of STAT5A predicted by ARACNe in the signaling-centered network. Genes on the left circle are transcription factors (TF in square shape), signaling molecules (Sig in diamond shape) or both (TF_Sig in hexagon shape). Genes in dark green are also predicted as master regulators or drivers (MR) of HER2+ breast cancer. Genes on the right circle are the ones in general. Red edge is for positive correlation while blue is for negative correlation.308

Figure 9-22 Subnetwork of STAT3 and STAT5A predicted by ARACNe in the signaling-centered network. Annotation is the same as in Figure 137 and Figure 138..... 309

Figure 9-23 STAT1 is a driver of ERBB2+ MCF10A cells, but shows different enrichment pattern with STAT3 and STAT5A.310

Figure 9-24 STAT1, similar to STAT3 and STAT5A, is not a driver of all HER2+ primary tumors.310

Figure 9-25 STAT1, different from STAT3 and STAT5A, doesn't show any addiction to ER status being a driver of HER2+ primary tumors.....311

Figure 9-26 Overall quality (left) and cycle-based quality (right) of raw NGS shRNA screening data of cells in 2D, 3D and in vivo environments.....313

Figure 9-27 Consistence of replicates of NGS shRNA screening data in 2D, 3D and in vivo environments.314

Figure 9-28 Distribution (A: boxplot, B: density) plots, Heatmap of sample distances (C) and PCA (D) plot of 2D, 3D and in vivo data of NGS-based shRNA screening.315

Figure 9-29 Gene expression signature of ERBB2+ MCF10A cells in 2D vs. 3D environments.321

Figure 9-30 NetBID2-predicted drivers of ERBB2+ MCF10A cells in 2D vs. 3D environments.322

Figure 9-31 Heatmap of z score for top depleted (in green) or enriched (in red) candidates from shRNA screening results of ERBB2 engineered MCF10A cells in 2D, 3D and in vivo system. STAT3 is ranked 64th in the depleted gene list by the combined z-score..... 323

Figure 10-1 The integrative framework to identify therapeutic targets for ABC or GCB-type DLBCL by integrating genome-wide RNAi screens (left) with systems biology (NetBID2) of cancer genomics..... 326

Figure 10-2 Batch effects detected for shRNA hairpin-probed microarray data of SUDHL4 cell line..... 328

Figure 10-3 Clustering of T0 or T10 shRNA screening data in NGS, BC-probed microarray and shRNA hairpin-probed microarray..... 330

Figure 10-4 Heatmap with hierarchical clustering (left) and PCA plot (right) of generated functional profiles from NGS-based shRNA screening data on four DLBCL cell lines. Six profiles for cell lines with triplicates for both T10 and T0

data. For Ly7, two T10 replicates are bad and removed, therefore only three generated profiles.332

Figure 10-5 Heatmap with hierarchical clustering (left) and PCA plot (right) of generated functional profiles from NGS-based shRNA screening data on four DLBCL cell lines. All generated profiles for each cell line are averaged to produce only one profile..... 333

Figure 10-6 Histogram of p-values (left) and density plot of z-scores (right) for gene-level differential representation analysis of each cell line. The bin width for p-value histogram is 0.05..... 334

Figure 10-7 Top genes selected by a threshold of $P < 0.05$ in all four lines. Red is for depletion while green is for enrichment. 336

Figure 10-8 Top genes selected by combined p-value with a threshold of 0.001. 338

Figure 10-9 IL23-mediated signaling pathway is enriched by lethal genes in HBL1, but not by lethal genes in other cell lines. 340

Figure 10-10 Alternative NF κ B pathway is enriched by lethal genes in HBL1, but not by lethal genes in other cell lines. 341

Figure 10-11 Heatmap of top signature genes of 35 ABC vs. 50 GCB DLBCL profiles. Red means under-expression in ABC relative to GCB, while green is for over-expression in ABC. 342

Figure 10-12 Heatmap of top candidates overlapped between signature genes of ABC vs. GCB and RNAi screening identified candidates lethal to ABC or GCB. “z.GEP” is the z score indicating differential expression of ABC vs. GCB. “z.shRNA” indicates the z score of comparing RNAi screening data of ABC vs. GCB cell lines. “z.negGEP.shRNA” is a combined z score of negative z.GEP and z.shRNA. Negative z.GEP is used because negative z.shRNA for gene X indicating X is depleted in ABC samples and we expect it’s over-expressed in GEP data or positive score of z.GEP, but positive z.GEP will cancel z.shRNA out, therefore a negative z.GEP is used to indicate the evidence of being depleted and over-expressed in one subtype.343

Figure 10-13 Heatmap of top candidates crossing RNAi screening results with NetBID2-predicted drivers specific to ABC or GCB-DLBCL. “nES” is the normalized enrichment score as evidence of being a driver. “z.GEP” is the differential expression of ABC vs. GCB.345

Figure 10-14 Heatmap of top candidates crossing RNAi screening results with NetBID2-predicted drivers specific to ABC or GCB-DLBCL, based on a loosed threshold. “nES” is the normalized enrichment score as evidence of being a driver. “z.GEP” is the differential expression of ABC vs. GCB.346

Figure 10-15 Heatmap of top candidates from CNV profiles of ABC vs. GCB samples. Red means amplification while green for depletion.347

Figure 10-16 Heatmap of top candidates crossing CNV results with RNAi screening results for ABC vs. GCB-DLBCL. “z.CNV” indicates amplification (red)

or depletion (green) in ABC vs. GCB samples. “z.negCNV.shRNA” is the combined z score of CNV result with shRNA differential representation score. The negative has the same annotation as in Figure 160.....348

Figure 10-17 Selected eight candidates crossing RNAi screens with NetBID2-predicted drivers and CNV results for ABC or GCB-DLBCL. “z.MR.negCNV.shRNA” is the combined z score of driver evidence, negative CNV score and shRNA screening. “nES” is the driver score.”z.GEP” is for differential expression.350

Figure 11-1 The integrative framework of RNAi Screening with NetBID2 to identify potential therapeutic targets specific to basal or luminal type of breast cancer. On the left, we did shRNA screening on 16 breast cancer lines including 8 basal and 4 luminal by NGS. On the right, we applied NetBID2 to TCGA breast cancer gene expression profiles to identify drivers of basal vs. luminal subtype.353

Figure 11-2 Clustering of shSeq samples by normalized data. Each row is for one sample condition. Three boxes on each row are the biological triplicates. One replicate T47D.T0 is highlighted as an outlier.357

Figure 11-3 Histogram of p value at gene level activity analysis for shRNA screens of 16 breast cell lines. The bin width is 0.05.358

Figure 11-4 Density plot Histogram of z score at gene level activity analysis for shRNA screens of 16 breast cell lines.359

Figure 11-5 Clustering of 14 breast tumor cell lines by functional profiles. Functional profile is using differential representation score at gene level.360

Figure 11-6 Heatmap of functional profiles for top depleted or enriched genes ($P < 0.01$) in the majority of 12 breast tumor cell lines. The genes are selected by a combined z score of all 12 tumor lines using Stouffer's method. "12T" is the z score of combining 12 tumor lines. Similarly, "4Lu" is combining 4 luminal lines, "3BaA" for basal A lines, "5BaB" for 5 basal B lines including SUM149PT, and "2N" for two normal lines.361

Figure 11-7 Sensitivity analysis of shSeq data for 16 breast lines.362

Figure 11-8 Heatmap of top shRNA screening identified candidates using functional profile. Functional profile is defined by differential representation score at gene level. Blue stands for depletion while red for enrichment.363

Figure 11-9 Top enriched pathways by depleted genes in luminal or basal subtypes.364

Figure 11-10 Top enriched pathways by depleted genes in luminal or basal subtypes. Top lethal genes in the pathways are listed.365

Figure 11-11 Specificity analysis by linear model using drug sensitivity data and shRNA screening data, also combination. Combination using known targets of drugs matched with shRNA screening data. Combined analysis is done using Stouffer's method. Purple dashed line indicates the significance level at 0.05. 367

Figure 11-12 Number of significant drugs that are specific to basal or luminal using different data source at various p value cutoffs. "n.drugs" is using drug

sensitivity data only. “n.drugs.shRNA” is based on shRNA data of drugs’ targets. “n.overlap.drugs” is the number of overlapped drugs between “n.drugs” and “n.drugs.shRNA”. “n.consist.drugs” is the number of overlapped drugs showing consistent direction in both shRNA screening results and drug sensitivity results. Green dashed line is half of the green line, and so is the red dashed line.368

Figure 11-13 Heatmap of top drugs using their sensitivity data and their targets’ shRNA scores.369

Figure 11-14 HDAC4 is specific to luminal type breast cancers shown by sensitivity data of HDAC4 inhibitors (left) and shRNA screening data of HDAC4 hairpins (right).370

Figure 11-15 HDAC1 is specific to luminal type breast cancers shown by sensitivity data of HDAC1 inhibitors (left) and shRNA screening data of HDAC1 hairpins (right).370

Figure 11-16 IKBKB is specific to basal type breast cancers shown by sensitivity data of IKBKB inhibitor (left) and shRNA screening data of IKBKB hairpins (right).371

Figure 11-17 Distribution of hairpin count in samples of six genetic-engineered models.373

Figure 11-18 Heatmap of sample distances and PCA plot of sample of six genetic-engineered models.373

Figure 11-19 Histogram of p values for gene level differential representation analysis of shSeq data from six genetic-engineered models. The bin width is 0.05.374

Figure 11-20 Density plot of z scores for gene level differential representation analysis of shSeq data from six genetic-engineered models.375

Figure 11-21 Sensitivity analysis of shRNA screening results for six genetic-engineered models.376

Figure 11-22 Unsupervised clustering of six genetic-engineered models by their shRNA screening functional profiles.377

Figure 11-23 Heatmap of top synthetic lethal partners for each of six genetic-engineered models.378

Figure 12-1 Heatmap of shSeq profiles for top hairpins depleted or enriched in MYCN-amplified NBL lines.382

Figure 12-2 Top pathways enriched by depleted hairpins in MYCN-amplified NBL line under normoxia condition.383

Figure 12-3 MYC up-regulated targets is enriched by depleted hairpins in MYCN-amplified NBL line under hypoxia condition.383

Figure 12-4 RPL13 and HNRNPA2B1 as MYC activated targets are lethal to MYCN-amplified NBL cells.....384

Figure 12-5 HSPA8 as MYC activated target is lethal to MYCN-amplified NBL cells.....384

Figure 12-6 CDC25C as MYC activated target is lethal to MYCN-amplified NBL cells.....385

Figure 12-7 Histogram and boxplot of fold change (log2 transformed) of averaged and three individual replicates of shSeq data Cisplatin treatment vs. DMSO. ...387

Figure 12-8 Histogram and boxplot of fold change (log2 transformed) of averaged and three individual replicates of shSeq data PARP inhibitor treatment vs. DMSO.387

Figure 12-9 Cumulative distribution plot of log2-transformed fold changes of shSeq data of Cisplatin or PARP inhibitor treatment vs. DMSO.388

Figure 12-10 Heatmap of top depleted genes in Cisplatin or PARP inhibitor treated shSeq samples.388

Figure 12-11 Top enriched pathways by depleted candidates from shSeq results of Cisplatin treatment comparing DMSO control.....389

Figure 12-12 Histogram of p values and density plot of z scores from differential representation analysis at individual hairpin level.390

Figure 12-13 Consistence of replicates for shSeq data of P53 positive screen.392

Figure 12-14 Histogram of p values and density plot of z scores from differential representation analysis at individual hairpin level.392

Figure 12-15 Heatmap of sample distances and PCA plot of samples showed consistence with biological meanings.394

Figure 12-16 Unsupervised clustering of conditions in shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.....	394
Figure 12-17 Sensitivity analysis of shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.....	395
Figure 12-18 Top candidates that are depleted in at least one of five cases in shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.	396
Figure 13-1 Collaborative web system: site dashboard - home page	399
Figure 13-2 Collaborative web system: site dashboard - files.....	400
Figure 13-3 Collaborative web system: my groups	401
Figure 13-4 Collaborative web system: group list	401
Figure 13-5 Collaborative web system: users-groups	402
Figure 13-6 Collaborative web system: group dashboard – home page	402
Figure 13-7 Collaborative web system: users in different roles.....	403
Figure 13-8 Collaborative web system: document list in a group.....	404
Figure 13-9 Collaborative web system: a document	405
Figure 13-10 Collaborative web system: revision history of a document	405
Figure 13-11 Collaborative web system: document creating page	406
Figure 13-12 Collaborative web system: calendar and event	407
Figure 13-13 Collaborative web system: blog in a group	408
Figure 13-14 Collaborative web system: project case tracker.....	409

Figure 13-15 Collaborative web system: shoutbox or twitter in a group410

Figure 13-16 Collaborative web system: image or photo collection411

Figure 13-17 An analogy among software system, hardware system and biological natural system.....412

Figure 13-18 Insights from Drupal future to Pathway future: reusability, scalability and robustness.413

Acknowledgements

I am deeply grateful to the people who gave me academic, financial and spiritual support to pursue my fascinating PhD investigation over the past four years.

Andrea Califano as my advisor has greatly influenced me since I joined Columbia. During my first rotation in Andrea's lab, I found out that this lab was the right place for me to do my PhD and I was truly attracted by exciting research in the lab and Andrea's personality. So I decided to join the lab immediately and I appreciate Andrea's acceptance then. With that, I probably became the only DBMI student who didn't change rotations. Andrea guided my research, gave me innovative ideas, provided me tremendous opportunities of interesting research projects, supported me to attend conferences and short courses, and also allowed me the freedom to learn widely, collaborate broadly and pursue diversity of my academic background. I appreciate his patience and generosity to bear my poor English and stupid arguments at the beginning when I joined his lab. He has the amazing energy to supervise an overwhelming amount of research projects and handle many other responsibilities. He did everything in his power to ensure the maximum success of his people. Andrea is a model to me for my future career and I hope someday I could become a great scientist and leader like him. I have been extremely lucky to have Andrea as my PhD advisor and what I have learned from him are invaluable treasures for my future career.

Jose Silva has served as a second advisor to me and has been deeply involved in all of the work presented in this dissertation. Jose is the super expert in high-throughput RNAi screening. He developed the most widely-used shRNA library when he was a postdoc with Greg Hannon at CSHL. He is the fundamental key person to all my RNAi screening projects. Jose started his lab roughly when I joined my Columbia. In my first semester at Columbia, I heard a lot of good things about him from my friend who was doing a rotation with him. At that time, I came to Jose's lab to visit my friend occasionally and later my friend told me that Jose was asking about me and wondered if I was interested in working with him on the bioinformatics part. Until two years later, I officially met with Jose for a joint project on DLBCL, but since then, our collaboration has never stopped. Jose is a wonderful collaborator and colleague to work with. I have been enjoying very much working with him. He is also my mentor in cancer biology and experimental biology. He explained complex biological concepts in ways that made me understandable. I owe Jose a great deal for all my expertise in RNAi screening and breast cancer biology. My collaboration with Jose represents a positive model between experimental and computational specialists that is critical to succeed in the field. I believe our collaboration will continue even after I leave Columbia.

With Adolfo Ferrando, I have been working for three years on the project of glucocorticoid resistance in T-ALL, which formed two major chapters of this thesis. Adolfo is an expert in broad aspects of cancer biology and is also an expert in using modern technologies to formulate hypotheses. His long term

collaboration with our lab has produced fascinating research stories. From him I have learned tremendous amounts about cancer biology, particularly T cell leukemia. His questions in our joint meetings always guided me thinking deep and working hard to move the project forward. The data produced in his lab were among the best quality I have even seen. I thank several people in his lab including Erich Piovan for doing all the beautiful experiments in our AKT paper and Maria Sol Flaherty for doing validation experiments on candidates coming out of integrative analysis.

George Hripcsak, the department's chairperson, guided me through the dissertation process while serving as the chairman of my dissertation committee. I also thank Yufeng Shen, a young faculty member in DBMI, for agreeing to be on my committee, Raul Rabadan and Richard Friedman for giving me opportunities to assist their classes where I gained treasurable teaching experiences, especially Raul, the rising star in the department, for giving me lots of helpful and valuable suggestions on how to succeed in academic field.

Many other computational specialists at Columbia from statistics, applied math or engineering departments has taught me and made me much stronger in my computational abilities. I would like to thank Andrew Gelman and Jincheng Liu for teaching me Bayesian statistics, Tian Zheng for modern statistical learning, Zhiliang Ying for classical statistical inference, Xiaodong Wang for information theory, Predrag R. Jelenković for regulatory networks and Chris Wiggins for biophysical modeling. For many of them, I didn't register for their classes. I

appreciate their generosity to allow me sitting in their classes and to answer my questions.

I appreciate all the hard experimental work and biological knowledge from my collaborators. Ruth Rodriguez-Barrueco and Patricia Villagrasa Gonzalez from Silva lab helped me a lot understanding the protocol of RNAi screening and gave valuable suggestions for our joint ERBB2+ and other breast cancer projects. Laura Pasqualucci and Riccardo Dalla-Favera gave me good feedbacks on our joint DLBCL project. Shuobo Zhang, also a good friend, taught me a lot of biology.

Finally, many colleagues have been integral to my short but treasurable experience in Andrea's lab. Wei Keat Lim guided me during my early days in the lab; Jose Morales, a good friend as well, helped a lot to improve my English, and we have been collaborating on a medical informatics project. I also thank Mariano Alvarez, Brygida Bisikirska, Gabrielle Rieckhof, Archana Iyer, Will Shin, Celine Lefebvre, Mukesh Bansal, Seema Dhindaw, Yao Shen, Wei-Jen Chung, Pavel Sumazin for being involved in my projects and helped me in various points.

Most importantly, my parents and my younger sister in China have been always supporting my back no matter what career decision I made and how far I was away from them. My fiancé, Lai Xu, who is much smarter than me and is doing her PhD in economics at Duke, has always been with me for all these years, supporting me, encouraging me and loving me. Without them, I could not have devoted myself into this fascinating research and finished my PhD within four years. I owe them the rest of my life.

To my mother, father, sister and Lai

Chapter 1 Introduction

Personalized medicine is coming of age [1, 2]. Traditional clinical diagnosis and treatment of diseases are based on patients' phenotypic information including clinical signs and symptoms, medical and family history, and data from laboratory tests and imaging evaluation [1, 3, 4]. The phenotypic information is often too vague to make the diagnosis and treatment precise and accurate. Moreover, clinical phenotypes are often late outcomes of disease progress and development, which makes treatment starts only after the signs and symptoms appear. Advances in human genetics and molecular medicine have enabled more detailed and more personalized characteristics of disease so that diagnosis and treatment based on such information at molecular level has emerged as a new field, i.e. personalized medicine or tailed therapeutics [1-10]. Personalized medicine is rational design of therapeutic approaches based on the specific genetic and other molecular characteristic information of the patient and/or patient's disease to maximize the clinical benefits and minimize risks, or in a simple explanation, personalized medicine is healthcare targeted to you, and just you. It means your individual health interventions — prevention, diagnosis and treatment — are custom-tailored specifically for you, based on your personal DNA, the expression of powerful proteins and each of you unique biological responses. My dissertation work is in this exploding but still very young field of personalized medicine.

1.1 Personalized Medicine in Cancer Treatment

One major focus of personalized medicine or tailored therapeutics during the past one or two decades is on human cancer treatment [11-13]. In cancer treatment, we are moving from conventional approaches including chemotherapy, radiotherapy and surgery based on tumor characteristics such as size lymph node, cell grade and patient fitness such as age, weight, general health, menopausal status, which are usually vague, to the emerging targeted therapeutic approaches based on patients' genetic information and specific molecular biomarkers, which are more precise and more personalized [1, 4, 14].

1.1.1 Problems of conventional cancer treatment approaches

There are significant problems with traditional phenotype-based therapeutic approaches for cancer treatment. First, the toxicity of chemotherapy or radiotherapy is usually high [15-17]. For example, chemotherapy that uses drugs to destroy cancer cells often kills adjacent normal tissues as well. Therefore, it often makes patients suffering tremendous side effects such as nausea, vomiting, hair loss, fatigue, anemia, mouth sores, taste and smell changes, infection, diarrhea, menopause, infertility, etc. Secondly, a significant number of patients are initially resistant to chemo- or radio-therapies, or relapse to develop resistance quickly. For example, in treatment for T-cell acute lymphoblastic leukemia (T-ALL), glucocorticoids are commonly-used chemotherapeutic agents in clinic due to its inducement of apoptosis in leukemia cells [18, 19], however, over 25% of T-ALL patients are resistant to this type of treatment, in which the

majority acquires resistance after treatment [18, 20-24]. Another example is cisplatin, a widely used chemotherapeutic agent against solid tumors such as the cancers of the testis, ovary, head, neck and lung [25, 26]. Although, cisplatin shows outstanding efficacy in the treatment of testicular cancers where regimens including this drug afford cure rates of greater than 95%, its effectiveness in the treatment of other cancers is more limited because of acquired or intrinsic resistance [27-33].

Therefore, we need tumor-selective therapeutic approaches or reversal of drug resistance for cancer treatment. In Chapter 7 and Chapter 8 of this dissertation, I will demonstrate in details how we identify therapeutic targets to overcome glucocorticoid resistance in T-ALL treatment. In Chapter 12, I will also show you a study of overcoming cisplatin resistance in lung cancer.

1.1.2 Problems of targeting oncogenes for cancer treatment

In the past decade of personalized cancer medicine, one main strategy that has been developed is to target oncogenes. Oncogenes are genes that have the potential to cause cancer, and in tumor cells, they are constitutively amplified or over-expressed because of aberrant genetic alternations, for example, HER2 (Human Epidermal Growth Factor Receptor 2) in breast cancer [34], EGFR (epidermal growth factor receptor) in lung cancer [35] and NOTCH1 (Notch homolog 1, translocation-associated), a transmembrane receptor, in T-cell leukemia [36, 37]. New molecular testing methods have enabled the testing for oncogene gene, protein, and protein pathway and/or somatic mutations in cancer

cells from patients. Targeting oncogenes will most likely benefit the patients with active oncogenic proteins.

Tremendous efforts have been invested to develop drugs or small-molecules to target aberrant oncogenes in a subset of patients with a given cancer type and many drugs are approved by Food and Drug Administration (FDA). For example, trastuzumab (marketed as Herceptin) and two other drugs – pertuzumab and lapatinib – are used in the treatment of women with breast cancer in which HER2 protein is amplified or overexpressed [38]. New drugs such as cetuximab, IRESSA and Tarceva that directly target the EGFR are used for EGFR positive lung cancer patients with a 60% responsive rate [35, 39]. Gamma secretase inhibitors such as RO4929097 and MK-0752 targeting NOTCH1 are being used for treatment of T-ALL [40, 41]. Also tyrosine kinase inhibitors such as imatinib (marketed as Gleevec) blocking activity of ABL are used to treat chronic myeloid leukemia (CML), in which the BCR-ABL fusion is present in >95% of cases [42].

However, targeting oncogenes as therapeutics of cancer treatment has significant problems as well. First, patients who receive such treatment usually develop resistance very quickly. For example, 50% of HER2+ breast cancer patients are resistant to Herceptin treatment initially, and the other 50% of patients treated with Herceptin will eventually develop resistance very quickly, within one or two years [43]. Secondly, many oncogenes are undruggable, especially when they are regulatory factors, such as KRAS (V-Ki-ras2 Kirsten rat

sarcoma viral oncogene homolog), a well-studied oncogene in cancers of colon [44], pancreas [45] and lung [46], and MYC, another well-known oncogene that is constitutively activated in Burkitt's lymphoma, breast cancer, neuroblastoma and many other cancers [47]. Thirdly, the toxicity of some therapeutics targeting oncogenes is also high. For example, in treatment of T-ALL, gamma secretase inhibitors that block NOTCH1 activity have been shown to induce lethal gut toxicity [24].

The above limitations of targeting oncogenes for cancer treatment motivate us to search for alternative new therapeutic approaches or overcoming resistance of existing ones. In Chapter 9 of my dissertation, I will demonstrate an example of discovering novel therapeutic targets for HER2 positive breast cancer, and in Chapter 12, I will show you a study of searching for therapeutics to overcome PARP inhibitor resistance for BRCA1-mutated breast cancer.

1.2 Functional Genomics: Genome-wide RNAi Screening

In the era of personalized medicine, genome-wide RNA interference (RNAi) screening has been widely used to discover therapeutic targets for human malignancies. RNAi has emerged as one of the standard techniques for studying phenotype-specific gene function from plants to fungi to animals via suppression of gene expression [48-51]. RNAi-based gene silencing can be achieved by the use of short interfering RNAs (siRNAs) or short hairpin RNA (shRNA) expression vectors. Among the two approaches, shRNA is more feasible because siRNA has the problem of transient inhibition of gene expression and inefficient

transfection into non-dividing cells; however, shRNA can be stably integrated into a target cell genome via retroviral or lentiviral gene transfer, resulting in the permanent reduction of the targeted gene product. Several shRNA expression libraries targeting entire human genome have been generated to facilitate functional analysis of the whole transcriptome through loss-of-function genetic studies [52-55].

In genome-wide shRNA screening, a large population of cells is infected or transfected with a pool of different shRNA lentiviral vectors and shRNA hairpins are integrated into cell genomes. After that, there are two common applications of these transduced cells. One is growing the cells for a sufficient number of doubling times, extracting the genomic DNA at initial time (T0) and after harvesting (T10), and then comparing quantity of shRNAs in these two time-points. This usage is to identify genes that are essential for cell survival or growth, thus making potential therapeutic targets for cancer and other type of human diseases, and hairpins of those lethal genes will be depleted or under-represented in T10 population. The other application is splitting infected cells into two groups, treating the two groups differently, for example treating one group with drug and nothing to the other as control. After this selective pressure, grow cells from both populations and then compare shRNAs extracted from genomic DNA of each population. This approach is to identify genes that modulate response to the perturbation. In the example of drug treatment, this screen can help to identify genes that increase sensitivity or resistance of cells to the drug.

1.2.1 Microarray-based shRNA screening

To read out shRNA hairpins extracted from genomic DNA, microarray hybridization is commonly used with the advantage of low cost and flexibility. It employs PCR-amplified shRNA template sequence pools extracted from shRNA library-transduced cells under test as well as reference conditions. Each PCR fragment is labeled with a different fluorophore, followed by hybridization of both pools to the same array, or labeled with the same fluorophore followed by hybridization to multiplex arrays. Taking the two-color microarray as example, the ratio of signal intensities of two colors (Cy3, Cy5) for each probe sequence reflects the relative abundance of cells expressing the corresponding shRNA construct under test condition as compared to the reference. Consequently, shRNA hairpins that sensitize cells in the selective condition will be depleted from the pool, showing low values of signal ratio, whereas shRNA constructs that render cells resistant will be enriched, showing high values of signal ratio. Three types of molecular tags have been used as microarray probes, namely full-length hairpin, half hairpin, and external barcode sequence. Half hairpin is able to overcome the self-annealing problem during PCR amplification happening to full-length hairpin, and correspondingly has more efficient labeling and microarray hybridization than full-length hairpin [51, 56]. Barcodes are not necessary for enrichment screens or positive selections such as designs to detect shRNA constructs for cell proliferation [57], but are critical for depletion screens or negative selections such as studies designed to detect cell-lethal or drug-sensitive shRNAs [56, 58-60].

1.2.2 NGS-based shRNA screening

Next generation sequencing (NGS) has recently emerged as a cost-effective technology of quantitatively measuring abundance of short-length DNA or RNA in a short time. This massively parallel sequencing has been used in pooled shRNA screens[61-63], and comparing to microarray-based approaches, it offers several potential advantages in terms of coverage of targeting genes, flexibility of input library, scalability and dynamic range. Moreover, barcode-based sequencing is commonly used to increase the multiplicity and efficiency by mixing multiple samples together and sequencing at once. As the cost of NGS is rapidly decreasing, this means might dominate high-throughput shRNA screening in the near future.

In my dissertation, I will discuss computational analysis of both microarray and NGS-based shRNA screening data. In particular, I will introduce a novel computational pipeline to deconvolute and analyze NGS-based shRNA screening (shSeq) data because shSeq is relatively new comparing with microarray data analysis and there are no established tools and algorithms to analyze it. This pipeline includes software packages and algorithms for quality assessment (QA) of raw sequencing data, deconvolution of raw reads, preprocessing and normalization, quality controls (QC) of processed data, differential representation analysis at both individual shRNA level and integrated gene level, and other post-analysis of shSeq data such as functional enrichment analysis and sensitivity analysis. Especially, I will introduce a new statistical algorithm to integrate multiple shRNAs targeting the same gene in the library using Bayesian

hierarchical modeling approach, and will show that this new approach outperforms existing ones.

1.2.3 Limitations of high-throughput RNAi screening

Despite the powerfulness of high-throughput RNAi screening, there are a number of limitations and problems with current version of this emerging technology. For example, high false-positive and false-negative rates are usually associated with RNAi screening. A long list of candidates is often reported from a pooled shRNA screen, in which only a small percentage are true hits. It's impossible to follow up all identified thousands of candidates, and it's also heuristic to pick up top hits because the scores of top candidates are very close to each other. The reasons could be because of off-targets, i.e. designed shRNA construct targeting unexpected genes by sequence similarity, and low knock-down efficacy of shRNAs. Furthermore, noise and small sample size of high-throughput measurements makes the estimation of statistical metrics to score hairpins or genes inaccurate. Therefore, powerful computational analysis and additional knowledge are much needed to complement it.

1.3 Cancer Genomics and Systems Biology

Advances in human genetics and molecular medicine have driven progress in our understanding of cancer biology. Development and improvement in DNA copy number, gene expression, and next-generation sequencing (NGS) technologies have resulted in more comprehensive characterization and accurate classification of human tumors and provided insights into cancer genome

complexity and heterogeneity. This has led to the emerging field of cancer genomics to study human cancer genome. It is a systematic search within cancer families and patients for the full collection of genes and genetic or epigenetic alternations – both inherited and sporadic – that contribute to the development of a cancer cell and its progression from a localized cancer to one that grows uncontrolled and metastasizes.

Cancer genomics can also be extended and generalized to proteomics, epigenetics and epigeomics. Advances in proteomics with mass-spectrum technology have enabled comprehensive analysis and characterization of all of the proteins and protein isoforms encoded by the human genome that may have a significant impact on cancer biology as well. This is because while the DNA genome is the information archive, it is the proteins that do the work of the cell: the functional aspects of the cell are controlled by and through proteins, not genes. Progress in characterizing post-transcriptional and post-translational modifications has led to identification of epigenetic factors that governed important biological functions: growth, death, cellular movement and localization, differentiation, etc. Those proteins form a potential class of therapeutic targets for cancer treatment.

The cancer genomic data provides us significant molecule insights into genes, proteins and pathways that are causally associated with tumorigenesis, progression, or drug-resistance. We can use this information to complement genome-wide RNAi screens to shortlist candidates coming from RNAi screening,

and, more importantly, to identify novel oncogenes or tumor suppressor genes as therapeutic targets for cancer treatment.

1.3.1 Collaborative projects on cancer genomics

A significant number of community-driven collaborative research projects or programs on cancer genomics using high-throughput genomic technologies have been launched to provide systematic, comprehensive genomic characterization and sequence analysis of multiple types of human cancers, both primary samples and cell lines, and to facilitate cancer discoveries among scientists.

Below are a few examples:

- **The Cancer Genome Atlas (TCGA)** is a pilot project of the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) since 2005 [64]. It covers almost major types of human tumors including glioblastoma multiforme (GBM), squamous carcinoma of the lung, and ovarian serous cystadenocarcinoma, breast invasive carcinoma, kidney renal clear cell carcinoma, colon adenocarcinoma, etc. Data type includes copy number variants, DNA methylation, exon expression, gene expression, protein expression, miRNA expression, SNP, and somatic mutation.
- **The International Cancer Genome Consortium (ICGC)** has been organized to launch and coordinate a large number of international cancer genome research projects [65]. It provides comprehensive catalogues of genomic abnormalities including somatic mutations, gene expressions,

and epigenetic modifications in tumors from 50 different cancer types and/or subtypes.

- **Therapeutically Applicable Research to Generate Effective Treatments (TARGET)** is initiated by NCI and is dedicated to identify valid therapeutic targets in pediatric cancers including acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), neuroblastoma (NBL), high-risk Wilms tumor and osteosarcoma (OS) [66]. Data type includes gene expression, copy number variants, epigenetics, whole genome sequencing and exome sequencing.
- **The Cancer Cell Line Encyclopedia (CCLE)** project is an effort between the Broad Institute and the Novartis Institutes for Biomedical Research to conduct a detailed genetic characterization of a large panel of 1000 human cancer cell lines [67]. It includes DNA copy number, mRNA expression and mutation data.
- **The Connectivity Map (CMAP)** is a project launched by The Broad Institute to study the connections between genes, drugs and cancer [68]. It's a collection of genome-wide transcriptional expression data from human cells treated with bioactive small molecules. The second version contains more than 7,000 expression profiles representing 1,309 compounds on 5 human cancer cell lines.

1.3.2 Systems biology

To identify causally-associated genes or pathways from large-sampled high-throughput cancer genomic data, robust computational or statistical methods are

required. Because of the high dimension and noise from large-scaled genomic data, traditional methods of analysis at single factor level in such context might not work well. Taking gene expression data as an example, classical signature analysis of two phenotypes from two independent datasets designed to study the same cancer problem might produce very different results [69]. This has led to a new field of systems biology, which integrates and aggregates multiple perspectives of “omics” data to define a system and then performs analysis of the behavior of the system from a global view or network point.

Systems biology approaches have been successfully applied to high-dimensional data of cancer genomics. It has been shown that computationally inferred context-specific maps of transcriptional or post-translational molecular interactions from large-scaled gene expression profiles (GEPs) allow the elucidation of cryptic driver proteins whose gain or loss is necessary and sufficient for tumor initiation or progression [70-73]. Such master regulators or drivers are more robust than traditional signatures to distinguish phenotypes [69]. Therefore, we suggest that systematic inference of driver-type regulators from genomic data complementing with RNAi screen technology will give a more comprehensive molecular understanding of mechanisms of tumor progression or drug-resistance and provide novel targets for therapeutics.

In my dissertation, I will introduce a systems biology framework, Network-based Bayesian Inference of Disease Drivers (NetBID2), to infer disease drivers from high-throughput genomic data by reverse-engineering network and Bayesian

inference. I will demonstrate that this framework performs more robust than classical signature analysis, and is able to detect not only known drivers of various cancer contexts, but also hidden drivers that conventional methods fail to find. The prediction rate of this algorithm is also high based on experimental validations.

1.4 Integration of Functional Genomics with Cancer Genomics

As discussed above, genome-wide high-throughput RNAi screening is a powerful technology to identify therapeutic targets for cancer treatment, however, due to high false-positive and false-negative rates arising from off-target effects, low silencing efficiency, and noise, the technology itself might not be good enough to work alone. Availability of large-sampled public cancer genomic data enables us to discover tumor-associated genes or pathways. Particularly, systems biology analysis of cancer genomics, by utilizing network strength, is capable of identifying underlying drivers of tumorigenesis or drug-resistance with a high successful rate. This motivates the integration of systems biology of cancer genomics with functional RNAi screens to tailor driver-type therapeutic targets for cancer treatment.

In my dissertation, I have developed a framework to integrate functional RNAi screens with systems biology of cancer genomics to tailor potential therapeutics for reversal of drug-resistance or treatment of aggressive tumors. I have been working intensively on shRNA screening with Dr. Jose Silva, who developed GIPZ library, one of the two most popular shRNA libraries. I developed a series

of algorithms and tools to deconvolute, QC and post-analyze high-throughput shRNA screening data by next-generation sequencing technology (shSeq). My pipeline has become the standard for this type of analysis at Columbia's Genome Center. I have analyzed all shSeq data generated at Columbia so far in collaboration with over ten labs.

In parallel, I developed a systems biology algorithm, NetBID2, to infer disease drivers from high-throughput genomic data by reverse-engineering network and Bayesian inference, which is able to detect hidden drivers that traditional methods fail to find. Integrating NetBID2 with functional RNAi screens, I have identified known and novel driver-type therapeutic targets in various disease contexts.

For example, I discovered that AKT1 is a driver for glucocorticoid (GC) resistance, a problem in the treatment of T-ALL. From mass-spectrum data, we found that GC-resistance derives from AKT1's phosphorylation of the GC receptor, thereby blocking its translocation to nucleus. The inhibition of AKT1 was validated to reverse GC-resistance. Additionally, upon silencing predicted master regulators of GC resistance with shRNA screens, 13 out of 16 were validated to significantly overcome resistance.

In breast cancer collaborating with Dr. Jose Silva, I discovered that STAT3 is required for transformation of HER2+ breast cancer, an aggressive breast tumor subtype. The suppression of STAT3 was confirmed in vitro and in vivo to be an

effective therapy for HER2+ breast cancer. Moreover, my analysis revealed that STAT3 silencing only works in ER- cases.

In collaboration with Dr. Riccardo Dalla-Favera, I applied a similar approach to DLBCL. The integration of RNAi screens, ABC or GCB-type expression profiles and CNV data enabled the identification of known master regulators such as BCL6 and IRF4, as well as novel drivers specific to ABC or GCB-type. These potential therapeutic targets are currently being validated.

My integrative framework has also been applied to subtype-based breast cancer in collaboration with Dr. Jose Siva. Integrating shRNA screens of a panel of 16 breast cancer cell lines with systems biology of TCGA breast cancer expression profiling data, we identified therapeutic target candidates specific for luminal, basal A, basal B or HER2+ form of breast cancer, which are currently being validated.

1.5 Overview of the Dissertation

This dissertation is dedicated to develop algorithms and tools to integrate functional genome-wide RNAi screening data with systems biology of cancer genomics to discover drivers and therapeutic targets for human malignancies to meet the need of personalized medicine. Chapter 2 and Chapter 3 focus on genome-wide RNAi screening technology. Specifically, I will introduce a computational pipeline with a series of algorithms and tools for NGS-based shRNA screening (shSeq) data in Chapter 2, and a novel algorithm for meta-

analysis of shRNA screening data to report gene level activity in Chapter 3. In Chapter 4, I will introduce a novel systems biology framework, NetBID2, to infer disease drivers from cancer genomic data with an improved enrichment analysis method, BSEA detailed in Chapter 5. Chapter 6 is an example of using systems biology approach – Bayesian network – to infer drug-induced apoptosis pathways from CMAP data. Chapter 7 and 8 are on studies of identifying therapeutic targets to overcome glucocorticoid resistance in T-ALL. Specifically, in Chapter 7, my NetBID2 framework discovers AKT1 as a therapeutic target to reverse the resistance as validated biochemically and pharmacologically; and in Chapter 8, I will describe how we integrate genome-wide RNAi screens with systems biology to discover promising regulatory drivers as therapeutics to reverse glucocorticoid resistance in T-ALL. In Chapter 9, I will demonstrate another example of integrating RNAi screens with computational analysis of genomic data to tailor therapeutics for ERBB2/HER2+ breast cancer, in which we identify and validate STAT3 as an effective target. Chapter 10 and Chapter 11 are two more examples with preliminary results of applying my integrative framework to DLBCL and subtype-based breast cancer. In Chapter 12, I will briefly describe more examples of applying shSeq technology. In Chapter 13, I will introduce a user-friendly and dynamic web system I developed to manage collaborative projects and to facilitate and speed up our research. Finally Chapter 14 is a summary of the entire dissertation.

Chapter 2 Computational Analysis of Next Generation Sequencing-based shRNA Screening (shSeq) Data

2.1 Introduction

RNA interference (RNAi) has emerged as one of the standard techniques for phenotype-specific gene function studies from plants to fungi to animals via suppression of gene expression [48-51] and has been widely used for therapeutic target discovery [50, 74-79]. RNAi-based gene silencing can be achieved by the use of short interfering RNAs (siRNAs) or short hairpin RNA (shRNA) expression vectors, among which shRNA is more feasible than siRNA. It's because siRNA has the problem of transient inhibition of gene expression and inefficient transfection into non-dividing cells; however, shRNA can be stably integrated into a target cell genome via retroviral or lentiviral gene transfer, resulting in the permanent reduction of the targeted gene product. Several shRNA expression libraries targeting entire human genome have been generated to facilitate functional analysis of the whole transcriptome through loss-of-function genetic studies [52-55].

In genome-wide shRNA screening, a large population of cells is infected or transfected with a pool of different shRNA lentiviral vectors and shRNA hairpins are integrated into cell genomes. After that, there are two common applications of these transduced cells. One is growing the cells for a sufficient number of

doubling times, extracting the genomic DNA at initial time (T0) and after harvesting (TX), and then comparing quantity of shRNAs in these two time-points. This usage is to identify genes that are essential for cell survival or growth, thus making potential therapeutic targets for cancer and other type of human diseases, and hairpins of those lethal genes will be depleted or under-represented in T10 population. The other application is splitting infected cells into two groups, treating the two groups differently, for example treating one group with drug and nothing to the other as control. After this selective pressure, grow cells from both populations and then compare shRNAs extracted from genomic DNA of each population. This approach is to identify genes that modulate response to the perturbation. In the example of drug treatment, this screen can help to identify genes that increase sensitivity or resistance of cells to the drug.

To read out shRNA hairpins extracted from genomic DNA, microarray hybridization and Next generation sequencing (NGS) are commonly used. Microarray has a long history and is well-developed whereas NGS based on sequencing-by-synthesis has recently emerged as a cost-effective technology of quantitatively measuring abundance of short-length DNA or RNA in a short time. This massively parallel sequencing has been used in pooled shRNA screens [61-63], and comparing to microarray-based approaches, it offers several potential advantages in terms of coverage of targeting genes, flexibility of input library, scalability and dynamic range. As the cost of NGS is rapidly decreasing, this means might dominate high-throughput shRNA screening in the near future.

In this chapter, I will mainly focus on analysis of NGS-based shRNA screening (shSeq) data because this is relatively new and there are no standard tools to deconvolute and analyze such data. But for analysis of microarray-based RNAi screening data, please refer to the book chapter [80] in Appendix A I wrote for Methods in Molecular Medicine. In this analytical pipeline of shSeq data analysis (**Error! Reference source not found.**), I will introduce multiple quality assessment metrics for NGS raw

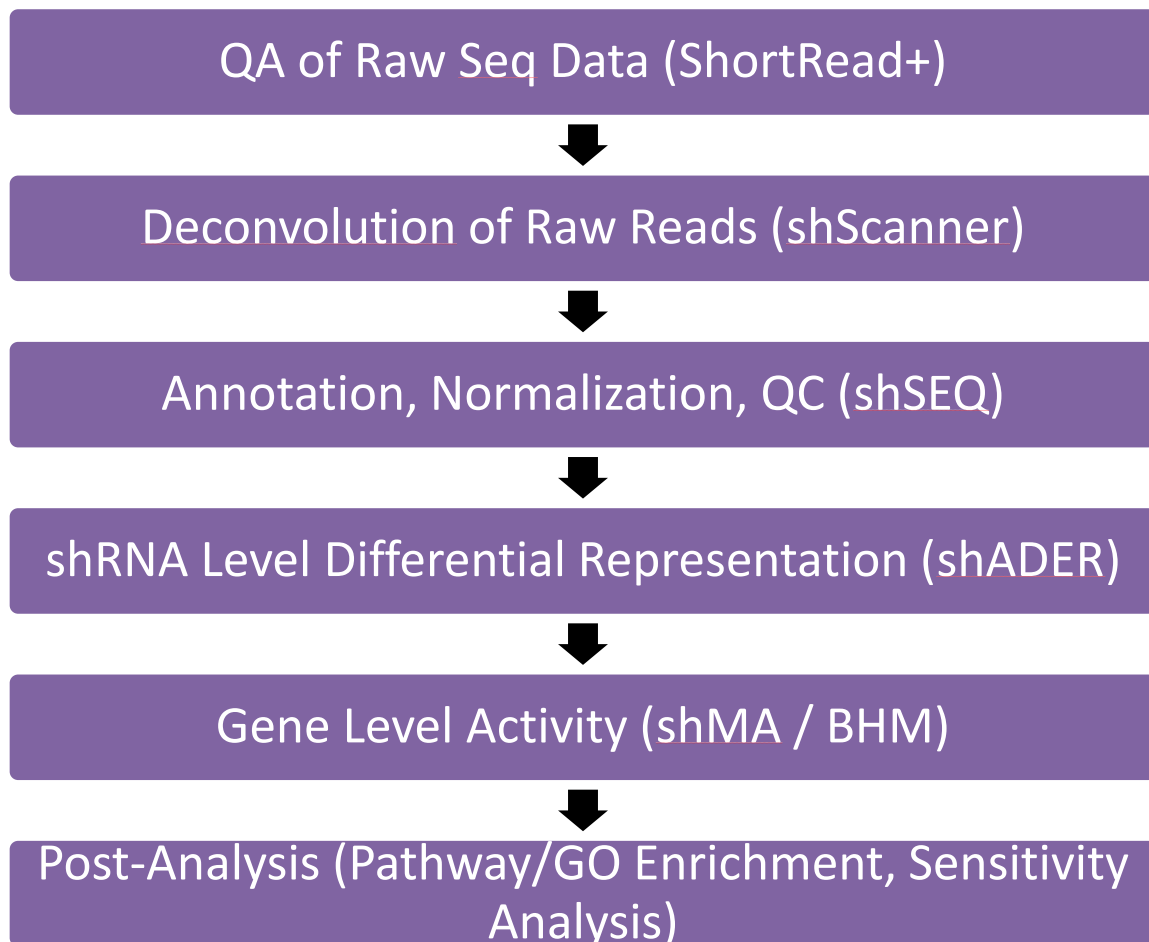


Figure 2-1 An overview of the pipeline for shSeq data analysis including a series of software packages and algorithms

data, an efficient algorithm to decode shRNA NGS data, preprocessing of screening data including background correction and normalization, quality controls of processed data to detect biological artifacts of experiments, statistical methods for differential representation analysis at individual shRNA level and gene level to identify candidates of interest and additional post-analysis of shRNA screens including functional enrichment and sensitivity analysis.

Note: all NGS-based shRNA screening data in this dissertation is on Illumina HiSeq platform and outputted by Genome Analyzer Iix.

2.2 shRNA Library

Thermo Scientific Open Biosystems GIPZ Lentiviral human shRNAmir library is used to illustrate the analysis of shSeq data. The library is composed of 58,493 hairpin constructs, in which 39,458 shRNAs are known to target 18,661 human genes, about 75% of the genome. In the GIPZ library, one gene might have multiple shRNAs and as shown in the distribution table of number of shRNAs per gene (Table 2-1), the majority of genes has 2 to 3 hairpins on average.

# shRNAs Per Gene	1	2	3	4	5	6	7	8	9	10	11	13	total
Frequency of Genes	6,931	5,986	3,635	1,355	481	168	60	24	12	4	4	1	18,661

Table 2-1 Distribution of Number of shRNAs per Gene

2.3 ShortRead+: QA of Raw NGS Data

The first thing to do after we have raw sequencing data is to check the overall quality of sequencing data and make sure there are no technical mistakes. I develop an R package, ShortRead+ to do quality assessment (QA) based on the existing ShortRead package. ShortRead package was designed for input, quality assurance, and basic manipulation of 'short read' DNA sequence produced by different platforms such as Solexa, 454, Illumina HiSeq, and related technologies. It is working well for small data set, for example, a sequence run with less than 10 million total reads, however, it's extremely time and space consuming for big data such as a sequence run with over 100 million reads, which is commonly obtained from shSeq experiments. Therefore it's not durable to process 100 or 200 million of reads by ShortRead. That's why ShortRead+ is developed. It overcomes the drawbacks and limitations of ShortRead and also improves some plotting features from classical R lattice [81] plotting to more popular ggplot2 framework [82].

2.3.1 Format transformation of raw NGS data (QSEQ to FASTQ)

ShortRead takes NGS data in FASTQ format as input. FASTQ is currently a standard format of Genome Analyzer output; however, in old version of Genome Analyzer, the default output of raw data is in QSEQ format. In this situation, we need to transform data from QSEQ to FASTQ format. Example raw NGS data in QSEQ and FASTQ formats are shown in Table 2-2 and Table 2-3. QSEQ format is in tabular text content, in which each row is a record of short read and the

columns are defined as in Table 2-4. One thing to be careful is that the quality score in QSEQ format sometimes is in Phrd64 format, which is different from the quality score (Phred33) in FASTQ format. A transformation is also needed. Functions for both transformation of QSEQ to FASTQ including quality score format are supported by ShortRead+.

HWI- ST618	80	4	1101	1081	2086	0	1	C.CG.ACTGCC CCGCTGGCAG GTAGGTGATG TTCC..GAGCGT]BYZB][[[[_ `_^____Y _`BBBBBBBB BBBBBBBBBB BBBBBBB	0
HWI- ST618	80	4	1101	1243	2093	0	1	CATCAACATGC TACTGGCGTTA GTTCCAGATCTT GAGGAAGCTAT CCCAGG	dddddaZdddcc c^d_v]ccedee ebaccWdabdc adSdbddd^^] ba]e	1

Table 2-2 Example of raw NGS data in QSEQ format

```

      @D8GSQ5P1:4:1101:1730:2234#0/1
CGATGTATCCACGCTGTTTTGACCTCCATAGAAGATTCTAGAGCTAGCGA
ATTCGCCCTTCCATGCCAAGTCAGAAGAGGTTATAATTTGGCTCTTACTCT
      +D8GSQ5P1:4:1101:1730:2234#0/1
__ceccccggfeghdgghihhddgffgfhfhhefhfhfgbdghdhfgghfgS_eagfdbdddfgbdd
eZ_bdY]Z]aaLZ_]bccccY_Y_`bbbRY`Y
      @D8GSQ5P1:4:1101:2152:2242#0/1
CAGATCATCCAAGCTGTTTTGACCTCCATAGAAGATTCTAGAGCTAGCGAA
TTCGCCCTTCTTCAAGATTTTCATCCTCCAAGTGCAGAACCAGGAAATTA
      +D8GSQ5P1:4:1101:2152:2242#0/1
[^ZZcccaeSb^eghfbadgdgagfaeb]effb_cdgcgg_fRcefX[^^eg`H[ed_ag_\_^dR\Z`
bZ^]ZZZHMZZ]]bb`Z`b_`^a[_a^bbb_

```

Table 2-3 Example of raw NGS data in FASTQ format

Field	Description
Machine Name	Identifier of the sequencer.
Run Number	Number to identify the run on the sequencer.
Lane Number	Positive integer (currently 1-8).
Title Number	Positive integer.
X	X coordinate of the spot. Integer. As of RTA 1.6, OLB 1.6, and CASAVA 1.6, the X and Y coordinates for each clusters are calculated in a way that makes sure the combination will be unique. The new coordinates are the old coordinates times 10, +1000, and then rounded.

Y coordinate of the spot. Integer. As of RTA 1.6, OLB 1.6, and CASAVA 1.6, the X and Y coordinates for each clusters are calculated in a way that makes sure the combination will be unique.

Y The new coordinates are the old coordinates times 10, +1000, and then rounded. Index Index sequence or 0. For no indexing, or for a file that has not been emultiplexed yet, this field should have a value of 0.

Read Number	1 for single reads; 1 or 2 for paired ends or multiplexed single reads; 1, 2, or 3 for multiplexed paired ends.
Sequence	Called sequence of read.
Quality	The calibrated quality string.
Filter	Did the read pass filtering? 0 - No, 1 - Yes.

Table 2-4 QSEQ format of raw NGS data: column filed descriptions

2.3.2 Phred quality score

In output of Genome Analyzer, i.e. raw NGS data, each nucleotide of a single read has a matched quality score indicating the error rate of base calling, which is in Phred format [83]. The transformation of base-calling error rate to reported quality score by Genome Analyzer follows equation in Figure 2-2. In generally, a base with a quality score over 30, corresponding base-calling error rate under 0.001 is considered as a good one.

$$Q = -10 \log_{10} P$$

Figure 2-2 Equation of transforming base-calling error rate or probability (P) to Phred Quality Score (Q)

ShortRead+ takes raw NGS data in FASTQ format (Phred33 for quality score) and checks the following quality metrics with plots and html report.

2.3.3 Overall quality of raw NGS data

The overall quality of one single read can be represented by the average Phred quality scores of all nucleotides in such read. Then we can check the distribution of overall quality of all reads. Figure 2-3 shows density distribution plot of overall quality scores from four different lanes. If the sequencing data is in good quality, you would expect a strong peak on the right, meaning that the majority of reads have a high averaged quality score, such as s_4, s_5, and s_6 in the example.

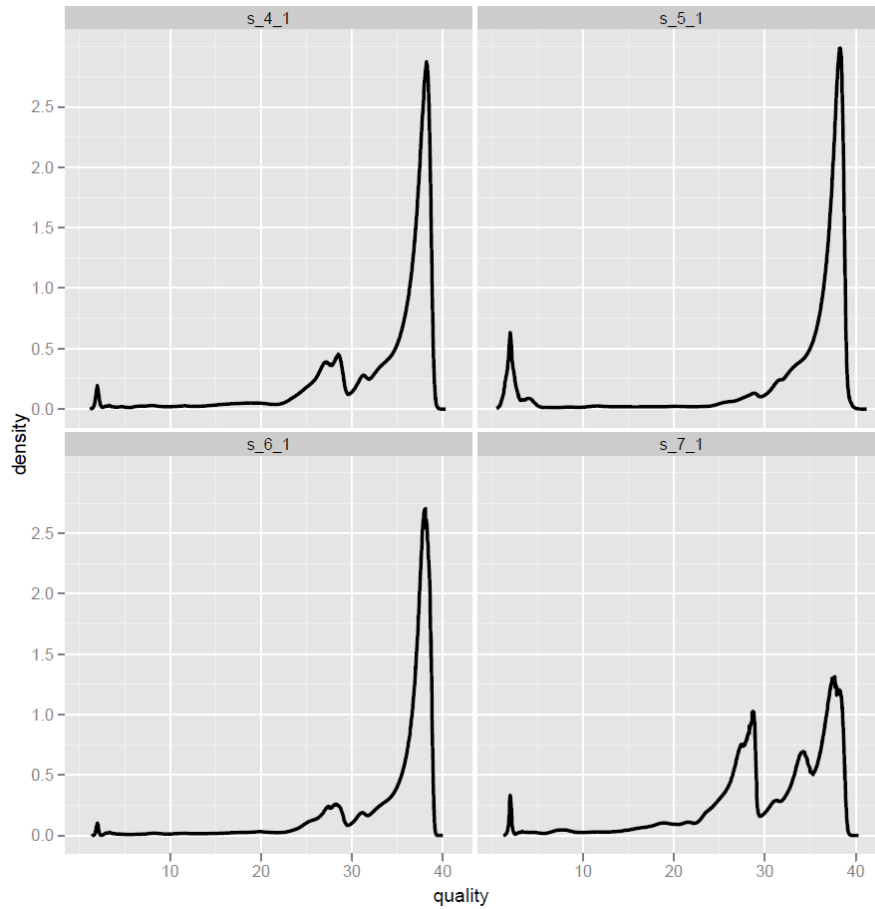


Figure 2-3 Density distribution plot of overall quality scores of all reads. The overall quality score of each read is calculated by averaging Phred scores of all nucleotides inside. A strong peak on the right indicates good overall quality of the sequencing data.

Otherwise, a peak on the left such as s_7 in the example indicates problems of the overall quality and you might want to check whether there is any flaw during the entire sequencing process.

2.3.4 Cycle-based quality distribution

We can also check the overall quality of each cycle by looking at the distribution of cycle-based or position-based quality scores. As shown in Figure 2-4,

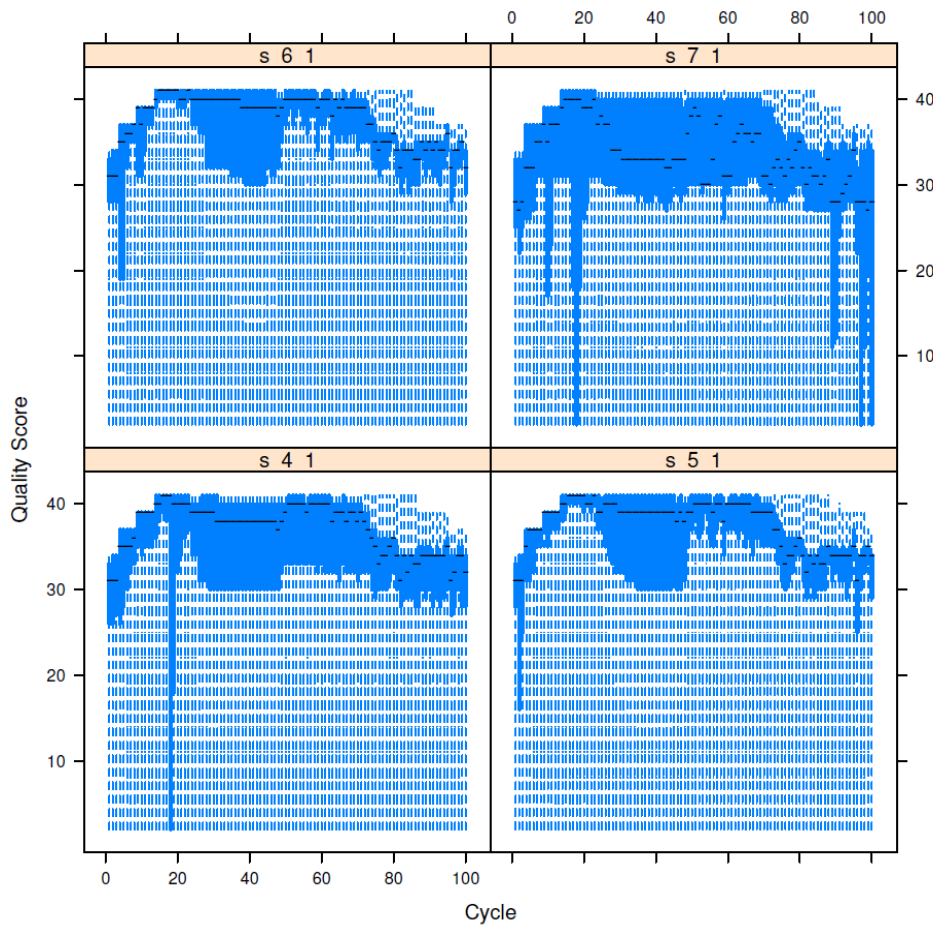


Figure 2-4 Cycle-based quality distribution (boxplot) of four different sequencing runs. The dashed short line “-” at each cycle indicates the median quality score, and the dark blue region at each cycle represents the 95% interval of quality score. Outliers are denoted by “.” or by the light blue dots at each cycle.

the quality goes up and stays stable in the middle and then goes down with increasing cycles. The bad overall quality of some cycles in s_7 lane explains the left peak in distribution plot of averaged overall quality.

2.3.5 Distribution of read count

It's also important to check the distribution of read count, especially to check the portion of low-frequent or high-frequent reads. One way to do that is to plot cumulative distribution curve (Figure 2-5) of how coverage is distributed amongst reads. Ideally, the cumulative proportion of reads will transition sharply from low to high. Portions on the left represent low-count reads and might correspond roughly to sequencing or sample processing errors. Portions to the right represent reads that are over-represented compared to expectation. Broad transitions from low to high cumulative proportion of reads might reflect sequencing bias or perhaps intentional features of sample preparation resulting in non-uniform coverage.

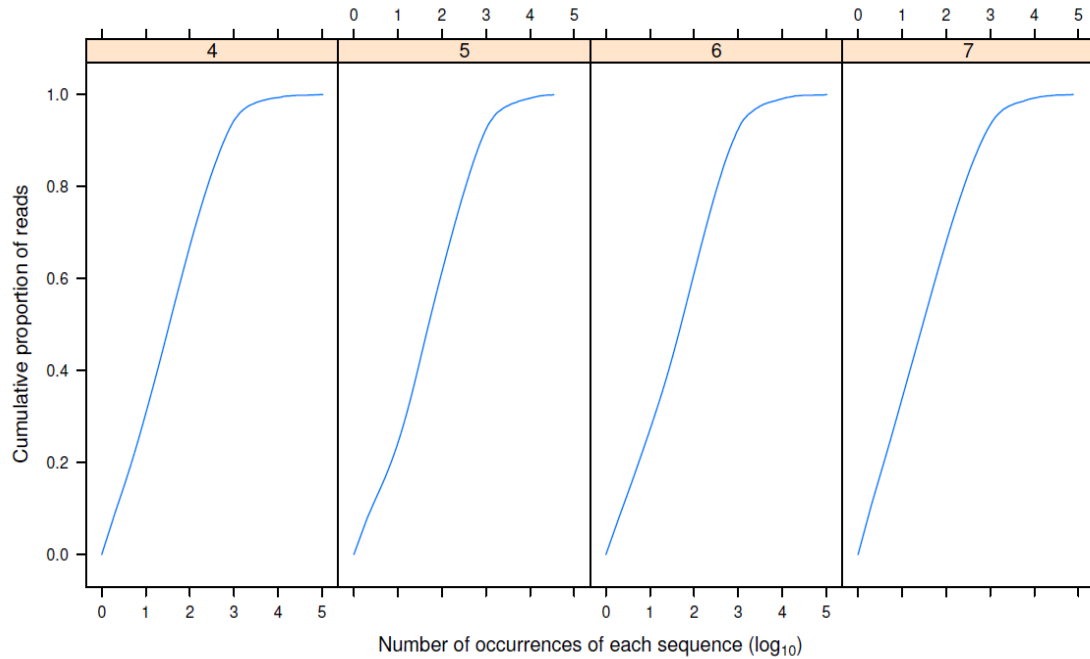


Figure 2-5 Cumulative distribution plot of read count. A point on the curve indicates the portion of reads (the score on the y axis) that has less than or equal to a certain number of count (the number on the x axis, log₁₀ transformed). Portions on the left are for low-frequent reads and portions on the right represents reads with high count number.

2.3.6 Cycle-based base calls

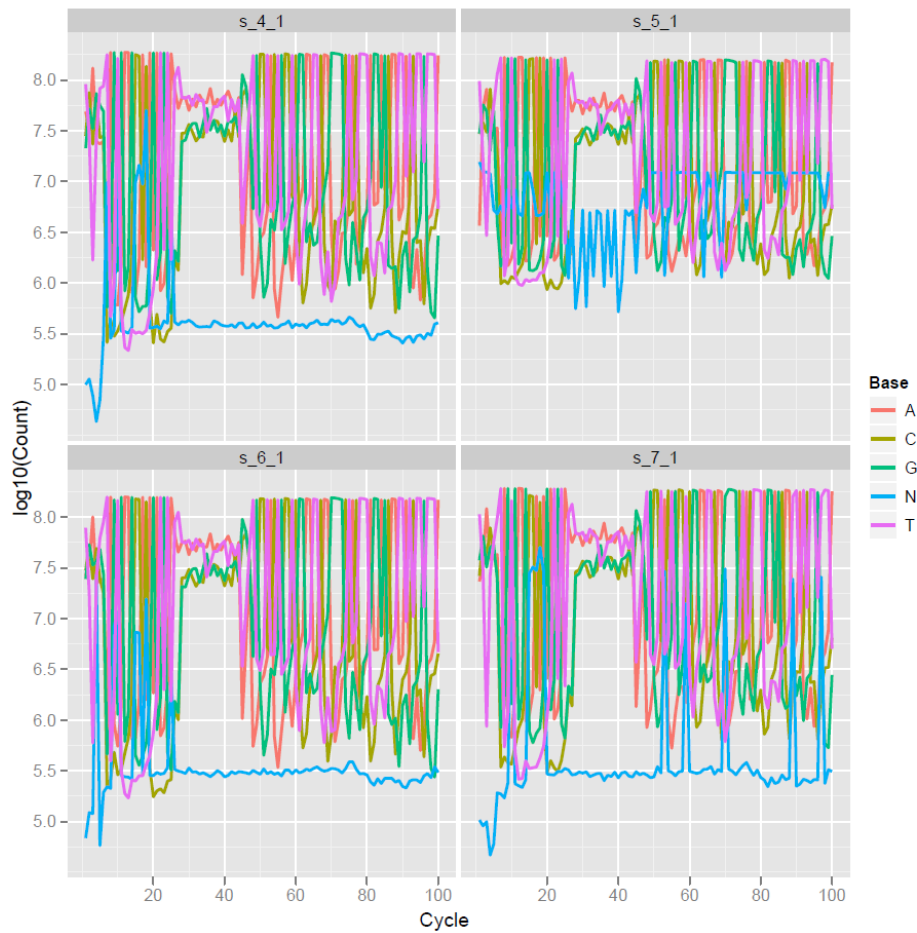


Figure 2-6 Cycle-based base calls or base count. N is for undermined bases, which should have a low-count curve in good-quality sequencing data.

Per-cycle base call should usually be approximately uniform across cycles. A constant count in the middle (around cycle 40) of the four example lanes (Figure 2-6) is because those cycles are the 19 common nucleotides of a shRNA construct (Figure 2-8) for all reads.

2.3.7 Example of a bad sequence run

Multiple QA metrics provided by ShortRead+ can help you to detect bad quality sequencing runs which might come from technical mistakes. For example in Figure 2-7, there is a significant peak on the left of averaged quality distribution (A) indicating that there is a large portion of low-quality reads. Cycle-based quality distribution and base call plots point out the reason: after cycle about 30, the quality of sequencing drops to almost 0, resulting in bad overall quality. After careful investigation of the process, it turns out that this is due to a failure of a agent kit during the sequencing. However, if you skipped this QA steps and performed analysis directly, you would not be able to identify this technical flaw.

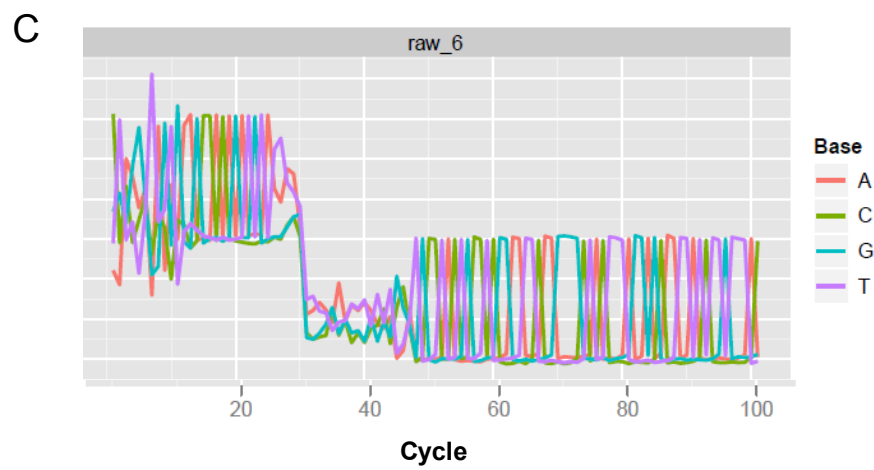
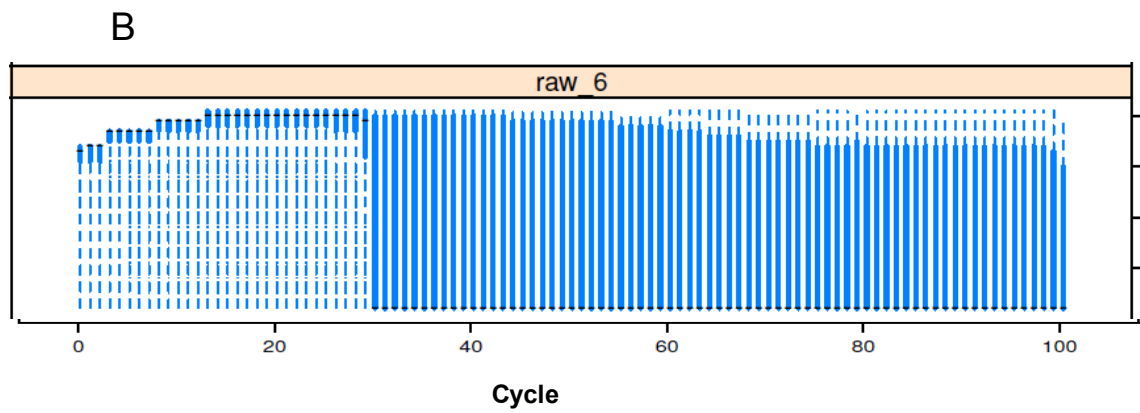
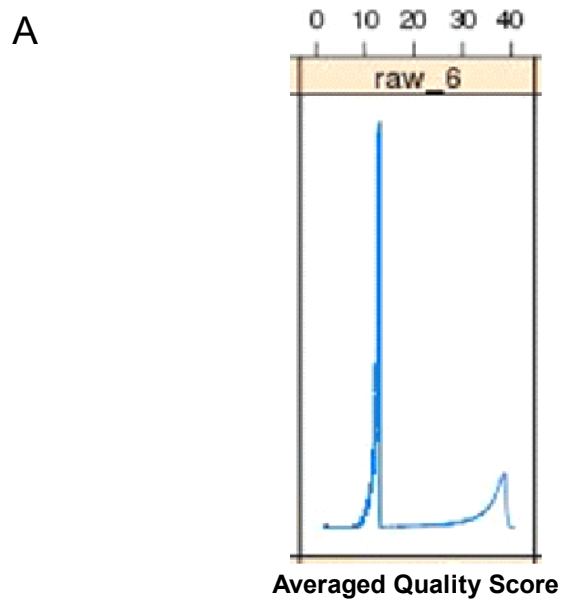


Figure 2-7 Example of a bad sequence run shown by (A) density distribution of averaged quality scores (B) cycle-based boxplot of quality distribution and (C) cycle-based base calls or base count.

2.4 shScanner: Deconvolution of Raw NGS Reads

After confirming the quality of raw sequencing reads is relatively good, the next step of shSeq analysis is to deconvolute each short read by identifying which sample it represents and which hairpin it comes from. According to the construction of each 50nt or 100nt-length sequence read (Figure 2-8), the first 6 nucleotides (in blue) are used to mapping back to the barcodes for multiple samples representing different experimental conditions, and the 22 nucleotides (in red) in the middle are used to identify shRNA hairpin in the library it belongs to.



Figure 2-8 Sequence structure decomposition of each shSeq read. The first 6 bases in blue are from barcodes of experimental design and the 22nt bases in red are from sequences of shRNA hairpins in the library, out of which 19 nucleotides in the middle are perfectly matched to the genome sequence.

Decoding a single read is easy in this context. Deconvolution of barcode and hairpin follows the same procedure. Take decoding hairpin as an example, it has the following steps:

- 1) Extract the hairpin sequence from the read (position 26 to 47)
- 2) Align the hairpin sequence with reference sequences in the library
- 3) Calculate the score of the alignment between each reference sequence and query sequence
- 4) Identify the reference sequences which has maximum alignment score
- 5) If there is only reference sequence identified, i.e. having maximum alignment score, report this sequence as the hairpin source of this read; if not, mark this read as ambiguous

The score of alignment is calculated by counting the number of exact matches between two sequences. A parameter of maximum number of mismatches (the default is 6 for barcode and 22 for hairpin) is introduced to control reads with a large number of mismatches.

However, to decode 100 million of reads, it would be extremely time and space-consuming if you do it one by one. shScanner implements a parallel computing framework under Titan clustering system [84] at Columbia C2B2 IT department. With the parallel computing technique, shScanner is able to decode 100M reads in just a few hours and only requires a 1G memory size for each job.

In deconvolution results, shScanner reports an identification table representing the count of each hairpin (rows) under each sample (column). In general, over 75% of total reads can be identified (Table 2-5). The distribution of total identified reads across barcoded samples depends on the number of cells mixed when preparing the samples, but in general they are equally distributed.

Run Name	T10.A	T10.B	T10.C	T0.A	T0.B	T0.C	total raw reads	total identified reads	identification rate	T10/T0 (total identified reads)
HS578T	14,261,572	11,276,323	3,402,531	24,198,563	18,979,220	8,794,471	103,508,686	80,912,680	78.17%	0.56
	244	193	58	414	324	150				
BT20	6,973,459	9,367,631	12,387,621	12,707,309	10,457,429	11,534,564	77,758,702	63,428,013	81.57%	0.83
	119	160	212	217	179	197				
T47D	7,742,742	16,499,446	15,292,978	8,851,725	14,139,016	11,371,252	96,503,420	73,897,159	76.57%	1.15
	132	282	261	151	242	194				
SKBR3	11,777,818	14,438,241	13,483,218	13,353,331	13,329,153	11,386,889	98,432,341	77,768,650	79.01%	1.04
	201	247	231	228	228	195				
MDAMB231	17,859,173	25,765,390	21,093,518	1,928,230	1,220,412	2,813,584	90,709,456	70,680,307	77.92%	10.85
	305	440	361	33	21	48				
MDAMB468	1,754,213	2,949,676	17,256,943	1,802,948	455,891	1,478,439	32,937,206	25,698,110	78.02%	5.88
	30	50	295	31	8	25				

Table 2-5 A summary table of deconvolution results for 6 shSeq runs. Each run contains 6 samples (T0 and T10 in replicate A-C), in which the total identified reads (the first rows in each run) and averaged count per shRNA (the second row in each run) are reported. Identification rate is the percentage of identified reads. The numbers in red indicate a case of low signals which might cause the data noisy. T10/T0 is the ratio of total identified reads at T10 and T0.

2.5 shSEQ: Processing, Normalization and QA of Deconvoluted shSeq Data

Deconvolution of raw shSeq data generates a table representing abundance of hairpins under each sample. However, before comparing different conditions, e.g. T10 vs. T0, to identify depleted or enriched candidates, we need to process the

data, normalize it and perform a secondary QA of normalized data. I develop a package, named shSEQ to do these jobs.

2.5.1 Preprocessing

Hairpin abundance in the shSeq data is count-based, thus there could be zeros, especially if the total number of identified reads is low. To obtain robust results of later comparisons and to avoid zero or infinite value when calculating fold change to represent the difference, a pseudo-count (default 1) is added to the abundance table. This is equivalent to putting a uniform or flat prior to the likelihood of such discrete data.

2.5.2 Normalization

Due to the discrete nature of shSeq count data, the normalization of such data is different from microarray data. The first step of shSeq data normalization is scaling the count to make each sample have the same total number of reads so that count between different samples is comparable. With the capacity of GIPZ library, 10 million of total reads is commonly used to do scale normalization, therefore each hairpin has about 170 reads on average. The scale-normalized count of one hairpin is proportional to the percentage of its abundance in a fixed size cell population. This step is similar to the background correction in microarray data preprocessing. After scale normalization, all replicates are scaled to have the same center roughly as shown in Figure 2-9 and Figure 2-10.

However, scale normalization doesn't reduce the variance within replicates as indicated by the correlation of replicates in Figure 2-9 and Figure 2-10. The

variance within replicates is an important source of noise for shSeq data and needs to be controlled. Therefore, a novel normalization procedure is proposed to reduce the variance of replicates by correcting outliers among replicates for each hairpin. Hairpins that have outlier count are determined by standard deviation ranking. By default, 25% of total hairpins are considered to contain outliers. For a hairpin with outlier count value, the outlier is decided as the one which causes the largest increase of standard deviation. The other two or more count values are used to fit a Gaussian distribution for the count value under this condition, which is then used to simulate a new corrected count for the outlier. If zero standard deviation occurs, the median of standard deviations of all hairpins is used instead.

After normalization by replicates (NBR), the variance within replicates is reduced as indicated by the increased correlation between replicates shown in Figure 2-11.

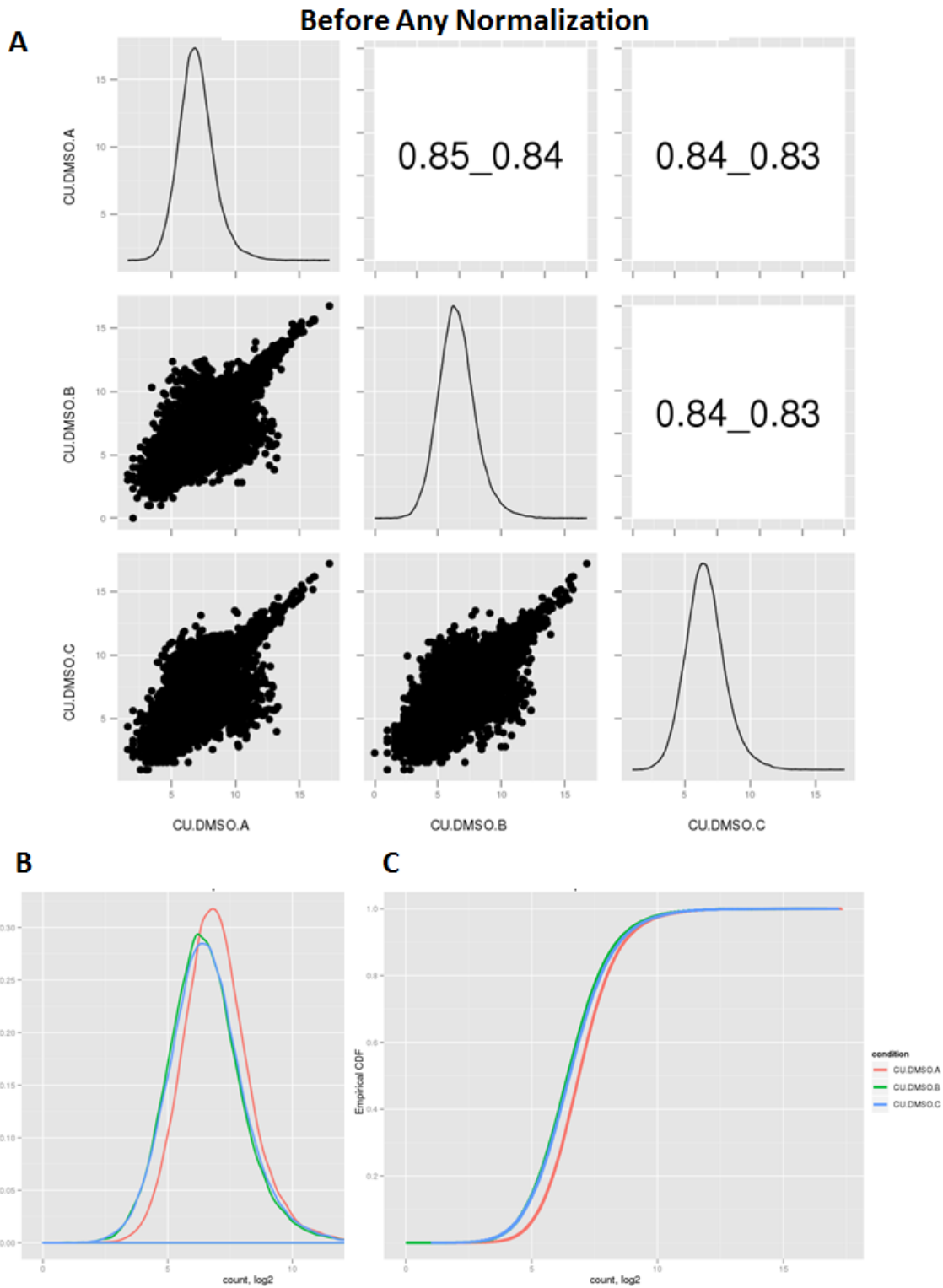


Figure 2-9 Scatter (A), density (B) and CDF (C) plots of data before normalization. (A) Scatter plots and correlations between biological replicates. Plots in the diagonal are density distributions of data in each replicate. Texts in the upper triangle cells indicate Pearson (the first number) and Spearman correlations.

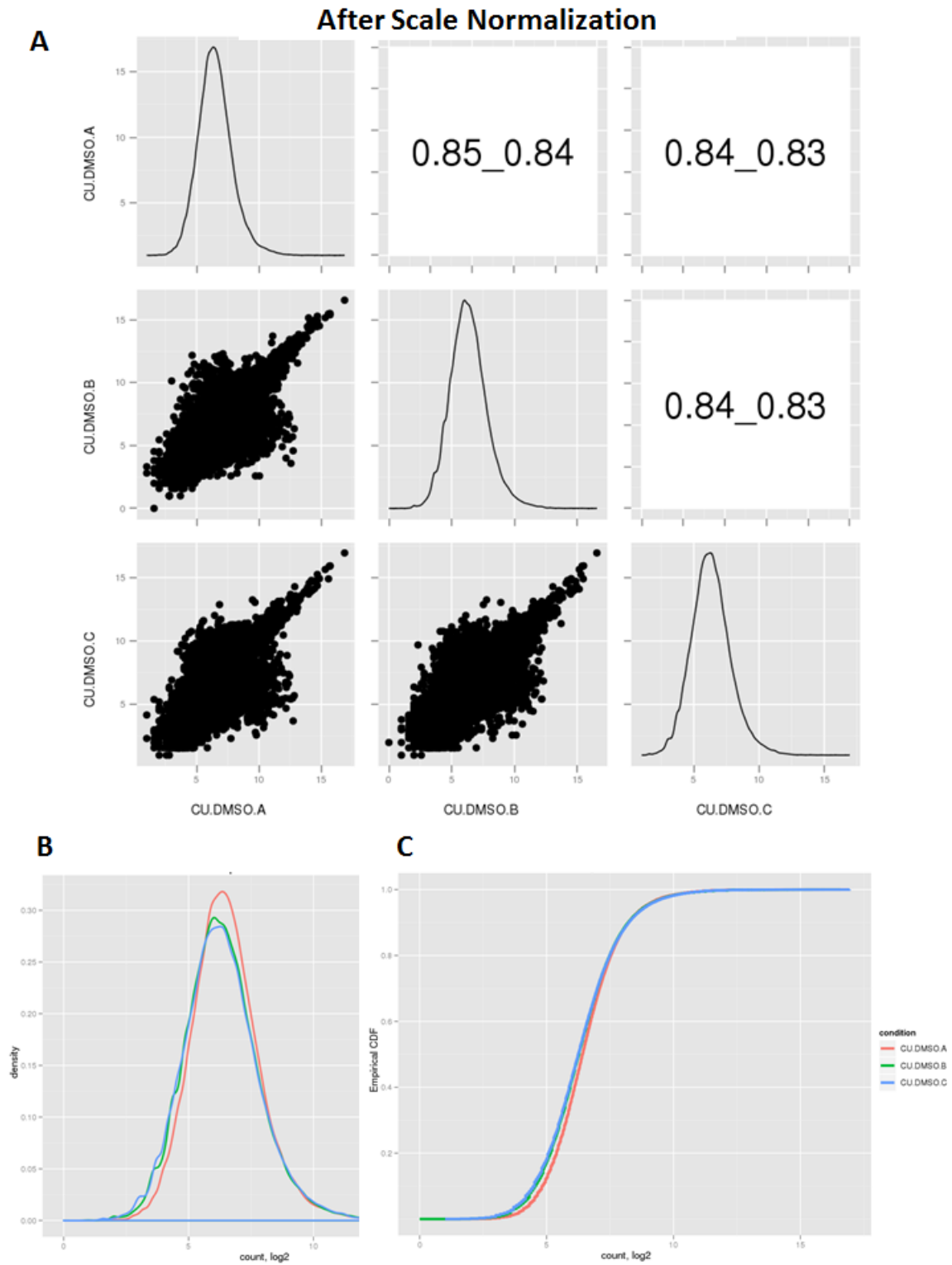


Figure 2-10 Scatter (A), density (B) and CDF (C) plots of data after scale normalization

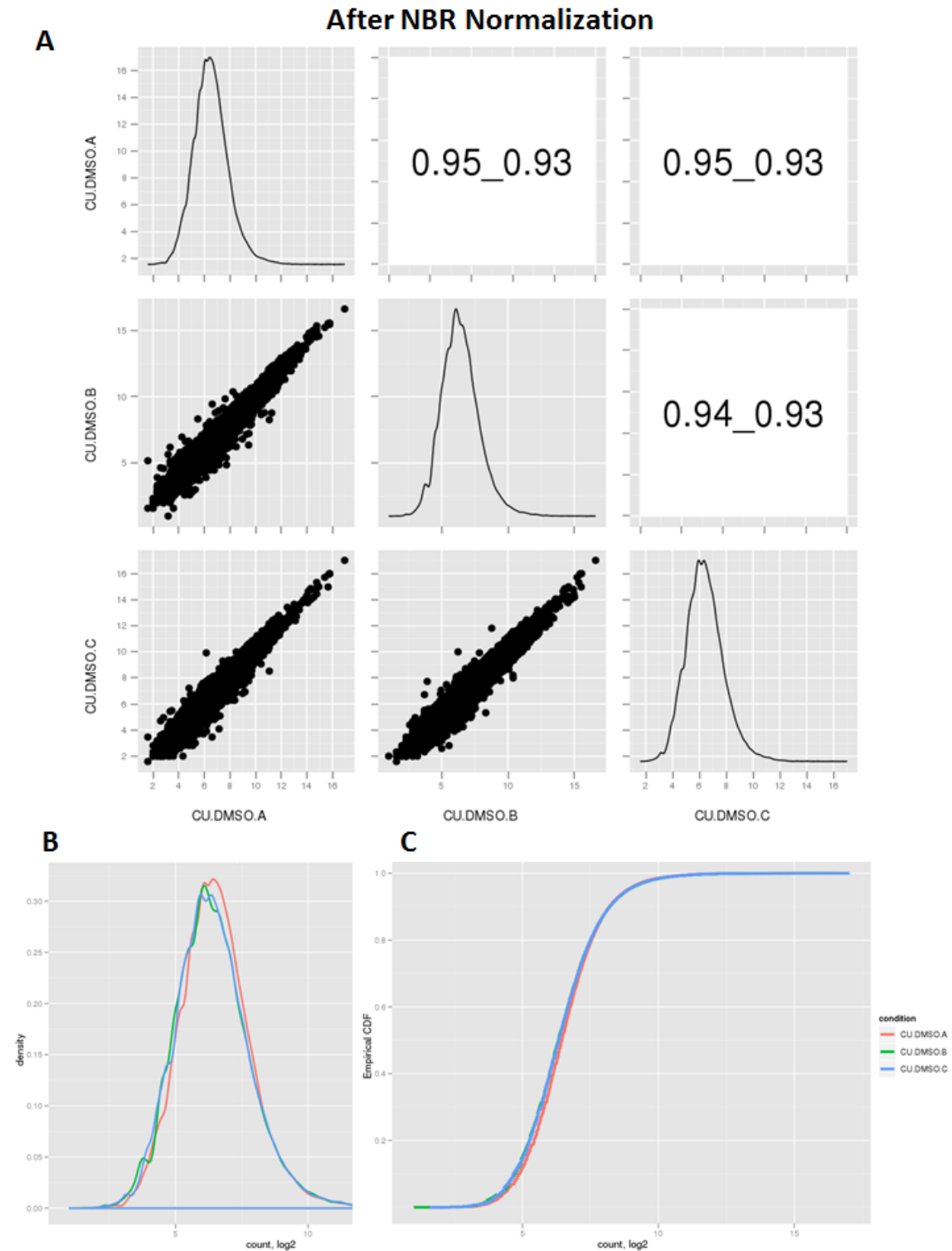


Figure 2-11 Scatter (A), density (B) and Cumulative distribution function (C) plots of data after normalization by replicates

2.5.3 QA of normalized data

After preprocessing and normalization of shSeq count data, we need to do a second QA to check the quality in biological perspective, i.e. the relations between different biological conditions and replicates.

The shSEQ package incorporates and extends QA metrics for microarray data [85] and checks the following aspects of normalized shSeq data.

2.5.3.1 MA plot

M and A are defined as:

$$M = \log_2 V_1 - \log_2 V_2$$

$$A = \frac{\log_2 V_1 + \log_2 V_2}{2}$$

V_1 is the shRNA count (sequencing data) of the sample studied, and V_2 is for a "pseudo"-sample that consists of the median across all samples. Generally, we expect the mass of the distribution in an MA plot to be concentrated along the $M = 0$ axis, and there should be no trend in M as a function of A . If there is a trend in the lower range of A , this often indicates that the samples have different background signals; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalization.

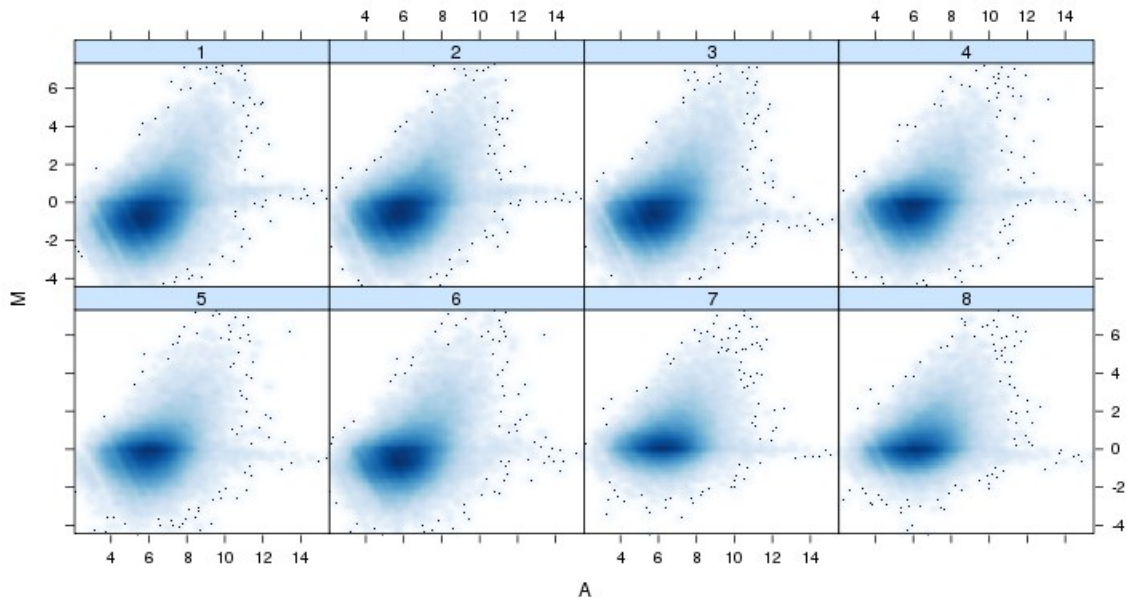


Figure 2-12 MA plot of shSeq data of multiple samples.

2.5.3.2 Distribution of hairpin count

Distribution plots including boxplot and density plot are commonly used to check the average strength of signal and noise level. Boxplots (Figure 2-13) represent summaries of the signal distributions of the samples. Each box corresponds to one sample. Typically, we expect the boxes to have similar positions and widths. If the distribution of a sample is very different from the others, this may indicate an experimental problem. Outliers based on the Kolmogorov-Smirnov statistic between each sample's distribution and the distribution of the pooled data, are marked by an asterisk (*).

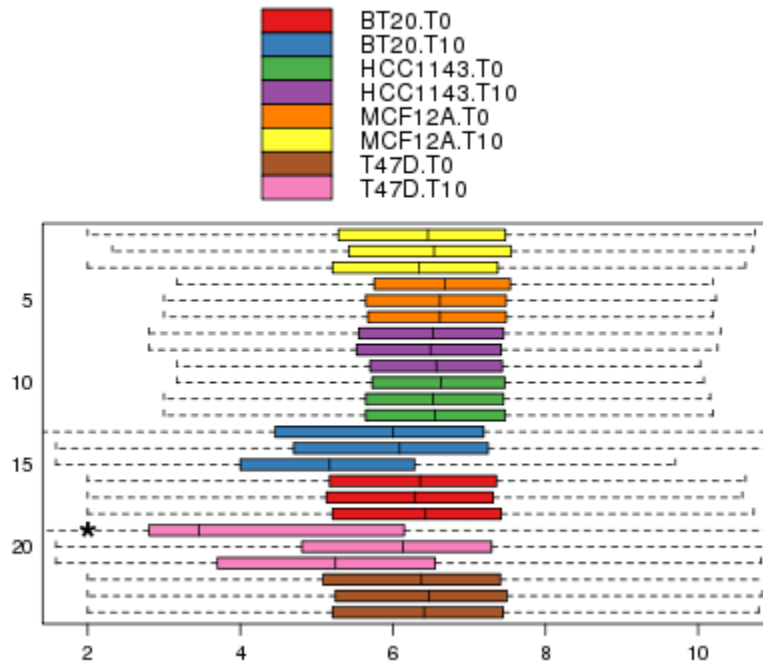


Figure 2-13 Boxplot of shSeq count in each sample.

Density plots (Figure 2-14) are smoothed histograms of the data. Typically, the distributions of the samples should have similar shapes and ranges. Outliers, according to the same criterion as in the boxplots, are highlighted by color.

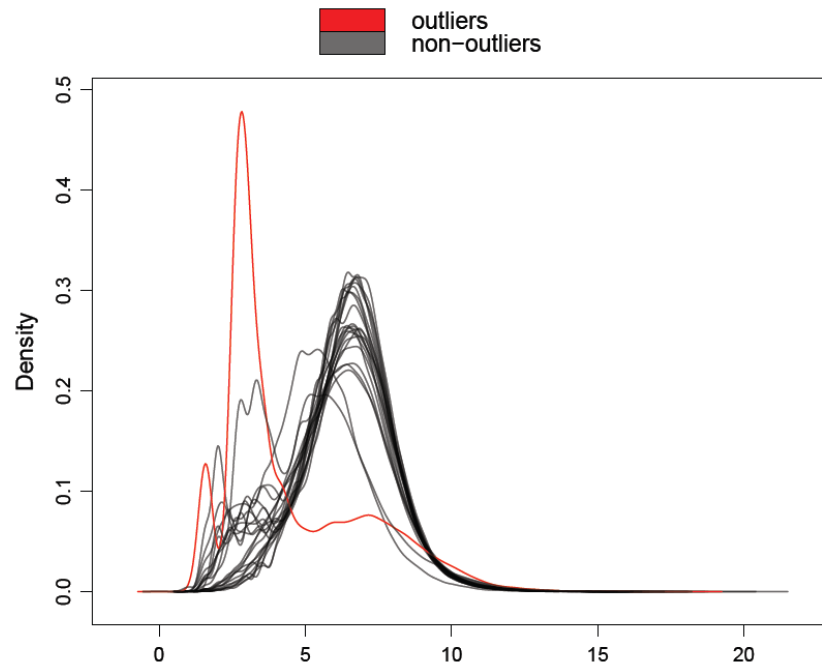


Figure 2-14 Density plot of hairpin count in each sample.

2.5.3.3 Consistence of biological replicates

One important perspective we need to check in the second QA of normalized shSeq data is the consistence of biological replicates. There are several ways of doing this supported by shSEQ package.

A straight-forward approach is to look at the correlation of replicates condition by condition. As shown in Figure 2-9, Figure 2-10 and Figure 2-11, empirical distribution of each replicate sample is plotted in the dialogue, and they are expected to have similar shape and scale as indicated in part B and part C (cumulative distribution plot). In part A, upper triangle shows the Pearson and Spearman correlation between two replicated samples without any filtering on shRNAs.

Heatmap (Figure 2-15) of between sample distances and dendrogram of sample clustering (Figure 2-16Error! Reference source not found.) can help to detect batch effects, as well as clustering of samples based on biological effects. The color scale is chosen to cover the range of distances encountered in the dataset. Datasets for which the sum of the distances to the others is much different from the others are detected as marked by * as outliers. The distance between two samples is the mean absolute difference (L1-distance) between the vectors of M-values (see 2.5.3.1) of the samples.

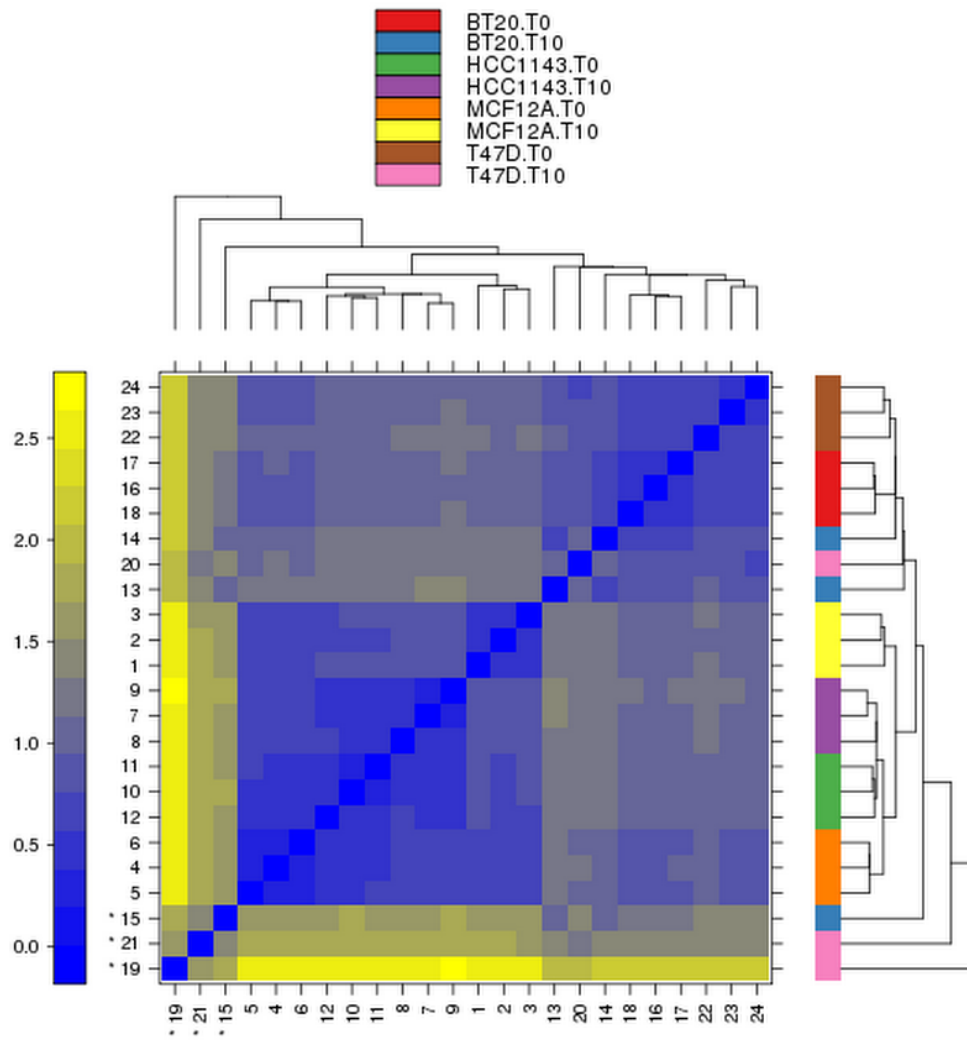


Figure 2-15 Heatmap of sample similarities.

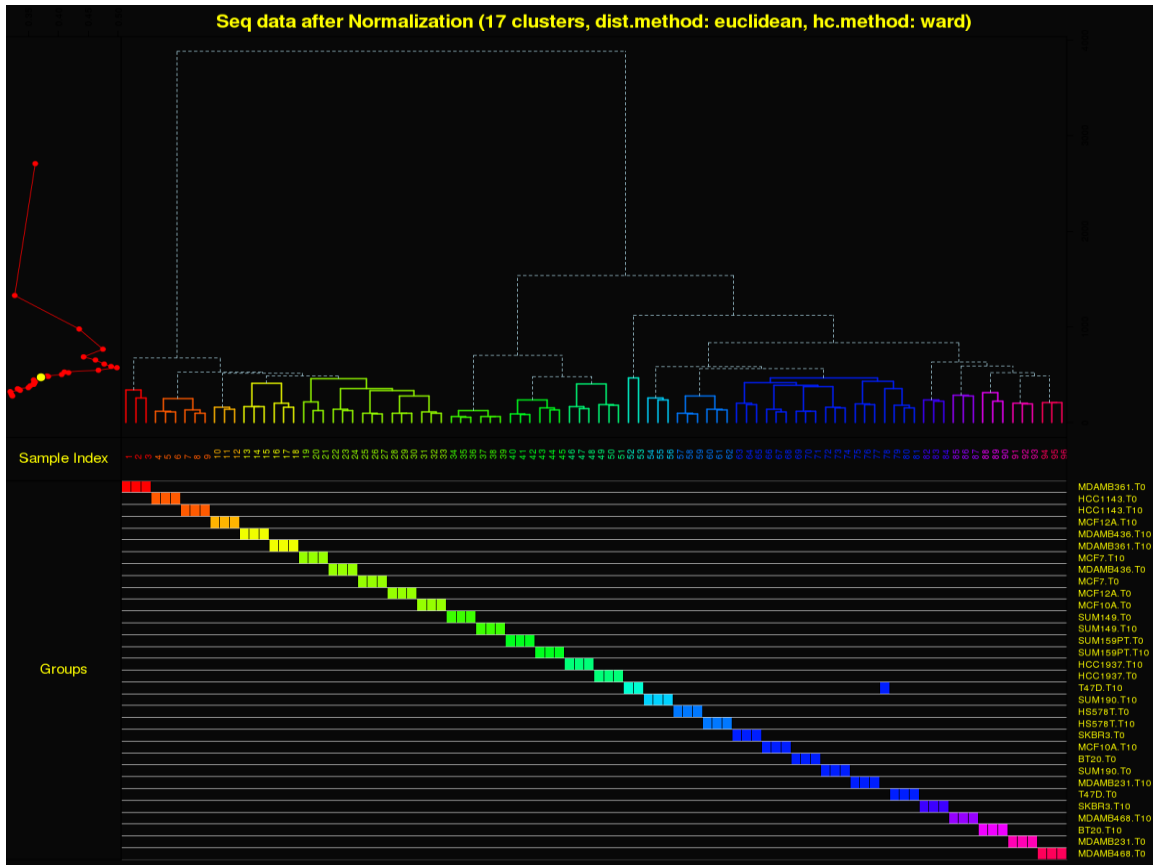


Figure 2-16 Hierarchical clustering of samples. Each row represents a condition while boxes on each rows are replicates under that condition. Dots on the upper left plot indicates where to split the three to obtain specific number of clusters, in which the yellow one is for the current plot; colors are for different clusters.

2.5.3.4 PCA plot

Scatter plot of the samples along the first two principal components (Figure 2-17) is used to check whether the samples cluster, and whether this is because of an intended biological or experimental factor, or according to unintended reasons such as "batch effects". Outliers, according to the same criterion as in the heatmap plot, are indicated by larger symbols.

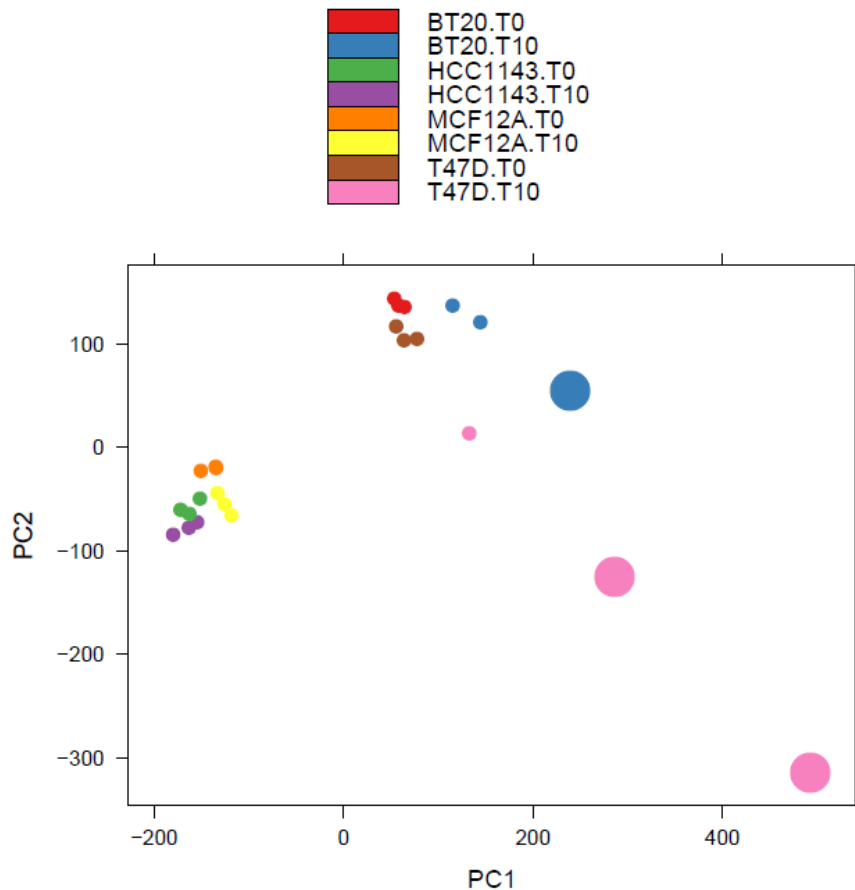


Figure 2-17 PCA plot of samples.

2.5.3.5 Variance-mean dependence plot

Variance-mean dependence plot (Figure 2-18) is the standard deviation of the representation values across samples on the y-axis versus the rank of their mean on the x-axis. The red dots, connected by lines, show the running median of the standard deviation. Typically, one expects the red line to be approximately horizontal, that is, show no substantial trend. In some cases, a hump on the right

hand of the x-axis can be observed and is symptomatic of a saturation of the measurements.

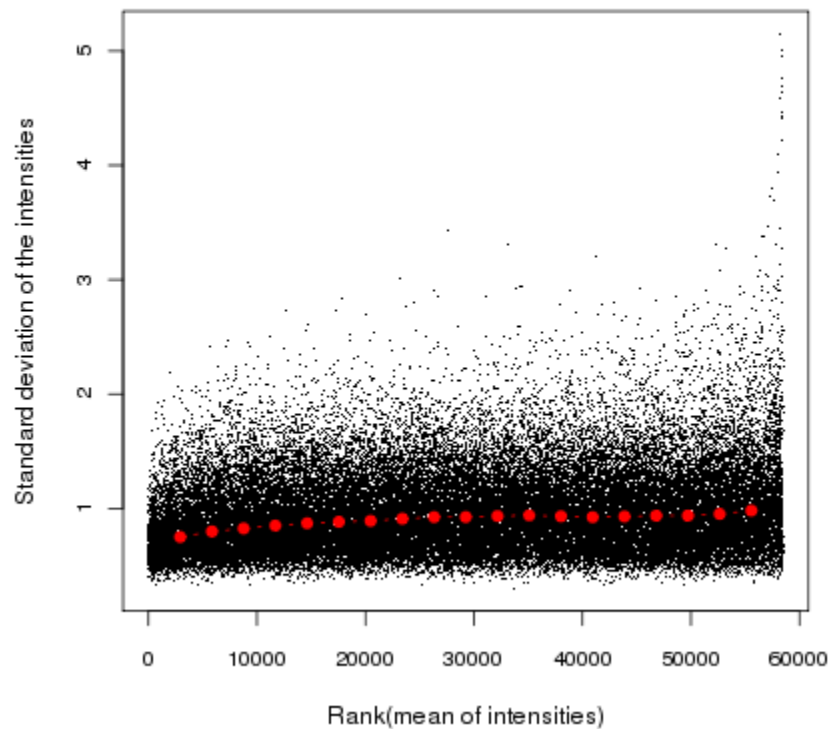


Figure 2-18 Variance-Mean dependence plot.

2.6 Comparison of NGS with Microarray Platform for RNAi Screening

With the above QA metrics, one interesting question we can ask is how shSeq data is comparing with classical microarray-based pooled shRNA screening. Using the metric of consistence between replicates, we observe that shSeq data with an average correlation of over 0.9 (Figure 2-11), is in general better than

microarray data, both barcode-probed with a correlation of 0.6 to 0.7 and hairpin-probed with a correlation of 0.7 to 0.8 (Figure 2-19). However, the shSeq data could be noisy as well, especially if the data doesn't have enough total identified reads as shown in Figure 2-20, which directly affects signal representation.

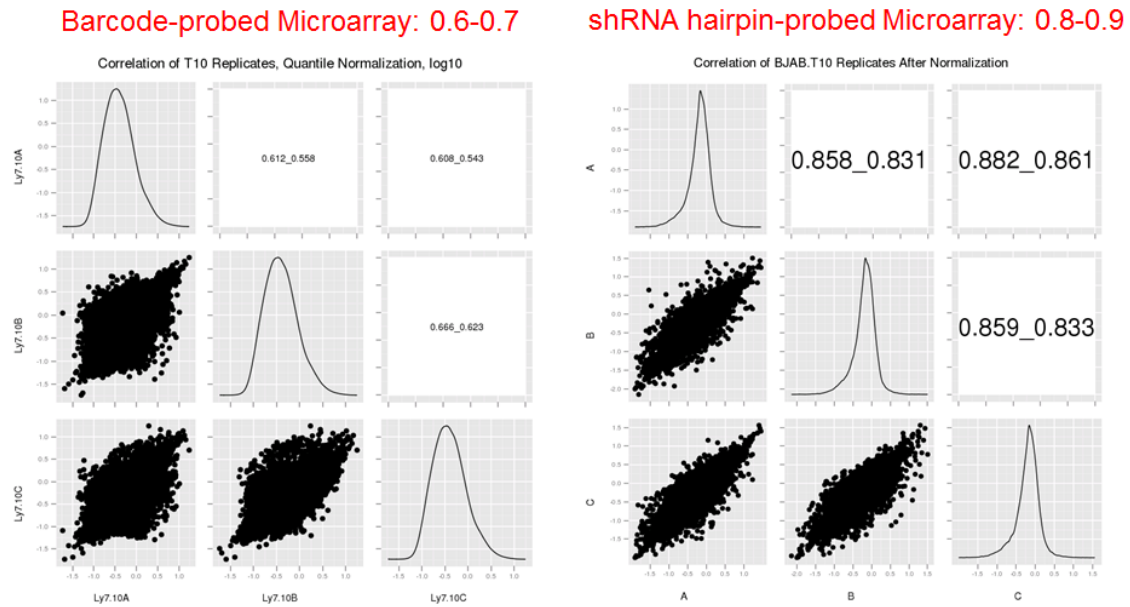


Figure 2-19 Consistency of replicates for RNAi screening data by barcode-probed (left) and hairpin-probed microarray platforms.

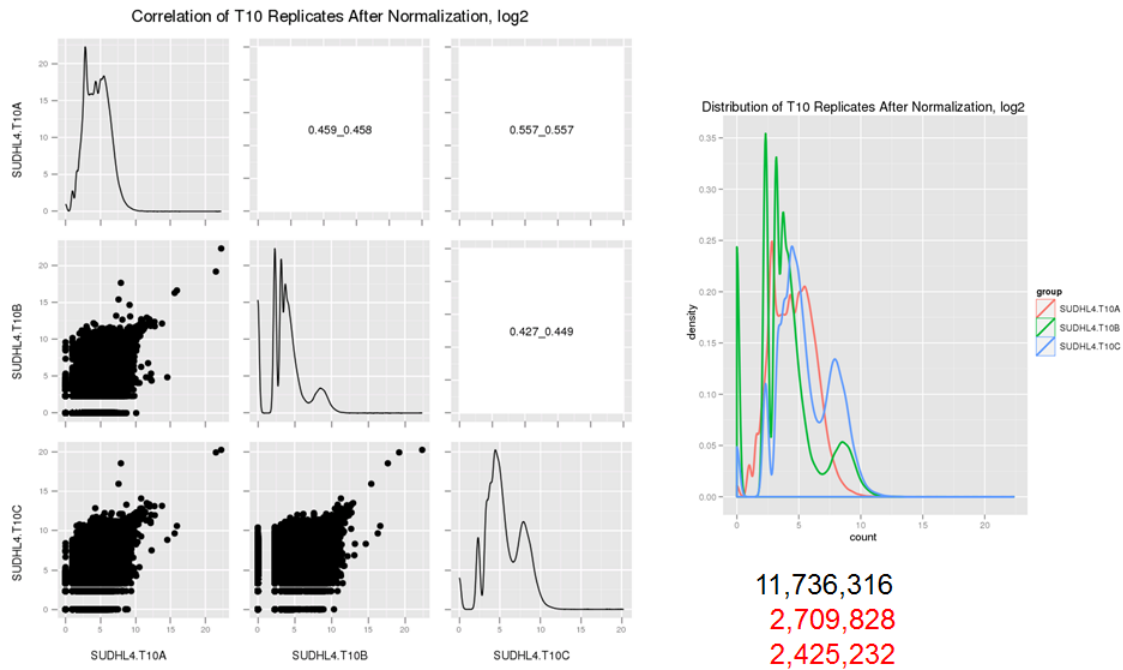


Figure 2-20 An example of bad shSeq run. The numbers on the bottom-right are total number of identified reads for each replicate. Low total numbers (in red) for replicate B and C reduces the signal representation, thus making the data noisy.

2.7 shADER: shRNA-level Differential Representation Analysis

Once shSeq is normalized and secondary QA shows good results, we are ready to conduct differential representation analysis by comparing shRNA abundance at TX with T0, to identify hairpins or genes that are either essential for cell proliferation or survival, i.e. depleted ones, or suppressors of cell growth, i.e. enriched ones. In negative pooled hairpin screens, we are more interested in depleted or under-represented candidates because those genes are potential therapeutic targets of diseases such as cancer. Because there are multiple shRNAs targeting the same gene in the library, we can do this analysis at

individual shRNA level or at gene level by integrating multiple hairpins for the same gene.

In literature, there are a number of different metrics to estimate the differential representation of individual shRNAs. For example, straight-forward such as fold change, log-transformed fold change, signal to noise ratio, difference of means can be used for simple analysis. To estimate the fold change between case and control samples, one need to calculate the mean within case or control samples. Two methods can be used: arithmetic or geometric mean, and the latter one is suggested for robustness.

2.7.1 Bayesian linear model

However, the above methods doesn't take statistical significance into account, therefore Student's t-test or moderated t-type test [86, 87] can be used to test the statistical significance, or a linear modeling approach [88] can be used to fit the data. For the modeling approach, the likelihood needs to be regularized by classical Frequentist's stabilization method [87], Bayesian or empirical Bayesian approach [88] due to small sample size issue. The regression coefficient represents the level of difference between case and control groups, and the statistical significance can be estimated by Chi-square test or Wald's z-test.

In my dissertation, I develop a method, shADER, to do shRNA level analysis of differential representation. It's essentially a linear model under Bayesian framework. The reason to do it with Bayesian inference is because shSeq data is usually noisy and the sample size is small, and Bayesian modeling [89]

overcomes those problems very well. Because of the discrete nature of shSeq count data, a Poisson distribution (Figure 2-21) is employed to model the data, but for microarray data which is continuous, a Gaussian distribution (Figure 2-22) is commonly used. In shADER, it supports multiple priors for the coefficient or the slope including Gaussian prior [89], t-prior [90] and g-prior [91]. In general, Markov Chain Monte Carlo (MCMC) computing techniques are used to simulate the posterior distribution and estimate parameters in the model by posterior mean or median.

Bayesian Linear Poisson Model	Priors
$y_i \sim \text{Poisson}(\lambda_i), \quad i = 1, \dots, m$	$(\alpha, \beta)' \mid \Sigma \sim N(\mu, \Sigma)$
$\log(\lambda_i) \sim N(\alpha + \beta x_i, \sigma^2)$	$\sigma^2 \sim \text{Inv} - \chi^2(v_0, \sigma_0^2)$

Figure 2-21 A Bayesian linear Poisson model. Y is hairpin abundance in count, which follows a Poisson distribution with a log-link. X indicates the condition, e.g.

T10 or T0. The coefficient of linear model β represents the magnitude of differential representation, and α is the intercept. The noise follows a Gaussian distribution with mean 0 and standard deviation σ . Priors for this model is a conjugate one, in which coefficients, β and α use a Gaussian distribution, and variance of noise σ^2 follows Inverse Chi-square prior[89].

Bayesian Linear Gaussian Model	Priors
$y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, \dots, m$	$(\alpha, \beta)' \mid \Sigma \sim N(\mu, \Sigma)$
OR $y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$	$\sigma^2 \sim \text{Inv} - \chi^2(v_0, \sigma_0^2)$

Figure 2-22 A Bayesian linear Gaussian model. Y is hairpin abundance in continuous value, which follows a Gaussian distribution. X indicates the condition,

e.g. T10 or T0. The coefficient of linear model β represents the magnitude of differential representation, and α is the intercept. The noise follows a Gaussian distribution with mean 0 and standard deviation σ . Priors for this model is a conjugate one, in which coefficients, β and α use a Gaussian distribution, and variance of noise σ^2 follows Inverse Chi-square prior.

2.7.2 Summarizing differential representation results

With the linear modeling approach, the slope is generally used to represent the magnitude of differential representation, but a summarized z-score is more robust to represent the differential representation results by taking the variance of the slope into account, especially when the data is noisy. Corresponding p-value will also be reported for statistical significance. The Z-score is calculated by estimate of regression coefficient over its standard deviation, which asymptotically follows a standard Gaussian distribution, therefore the two-tailed p value for statistically significance can be calculated based on this null distribution. This is essentially Wald's z-test.

FDR for correction of multiple comparisons in shADER is calculated by BH procedure [92].

The package of shADER also supports multiple visualization of the results. For example, density plot of z-scores (Figure 2-23) and histogram of p values (Figure 2-24) give overall distribution of depleted or enriched hairpins, also significance and non-significance. A uniform distribution of non-significant hairpins is expected, which is also the assumption of FDR calculation. Heatmap of selected

shRNAs using z-scores (Figure 2-25) or original shSeq data (Figure 2-26) visualizes the pattern of differentiated shRNA-silencing effects such as similarity between genes or samples.

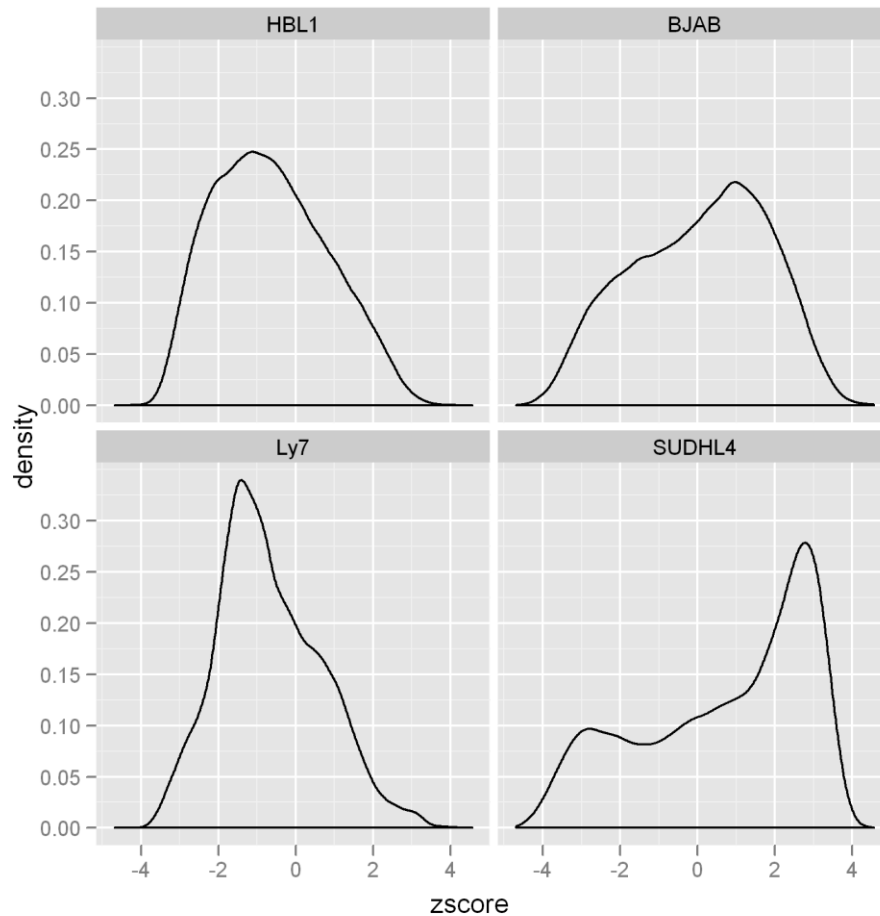


Figure 2-23 Distribution of z-scores indicating differential representation results from four different shSeq data sets. Positive z-score means enrichment of hairpins while negative is for depletion.

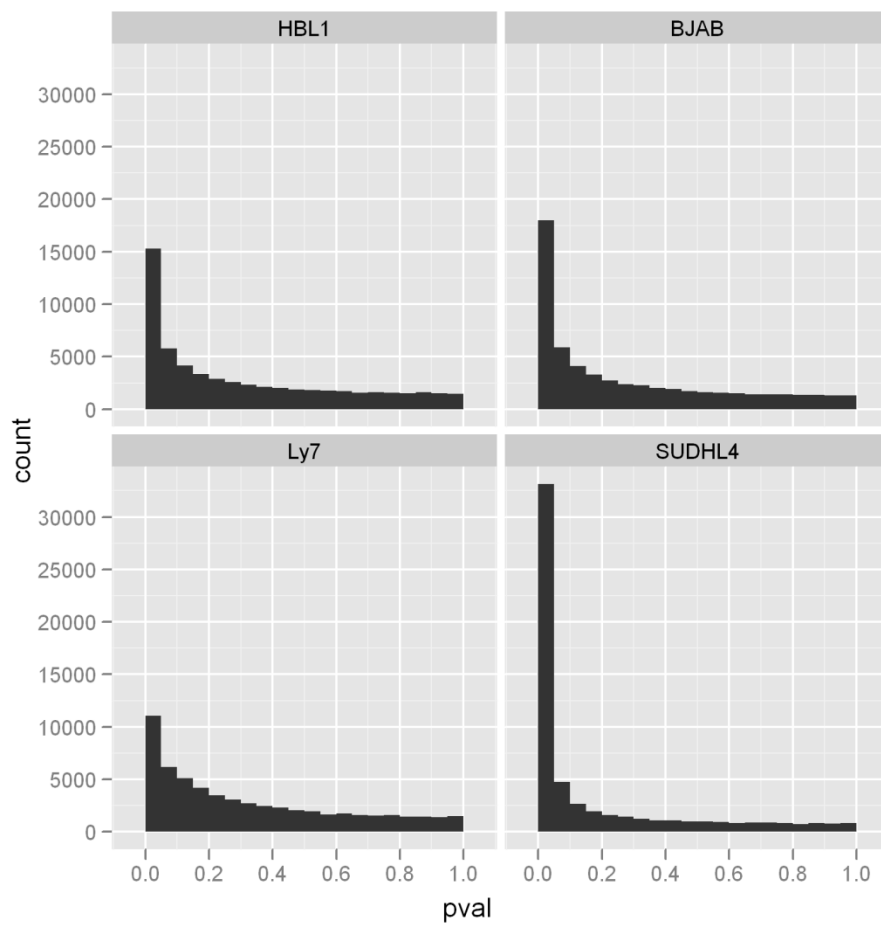


Figure 2-24 Distribution of p-values indicating differential representation results from four different shSeq data sets.

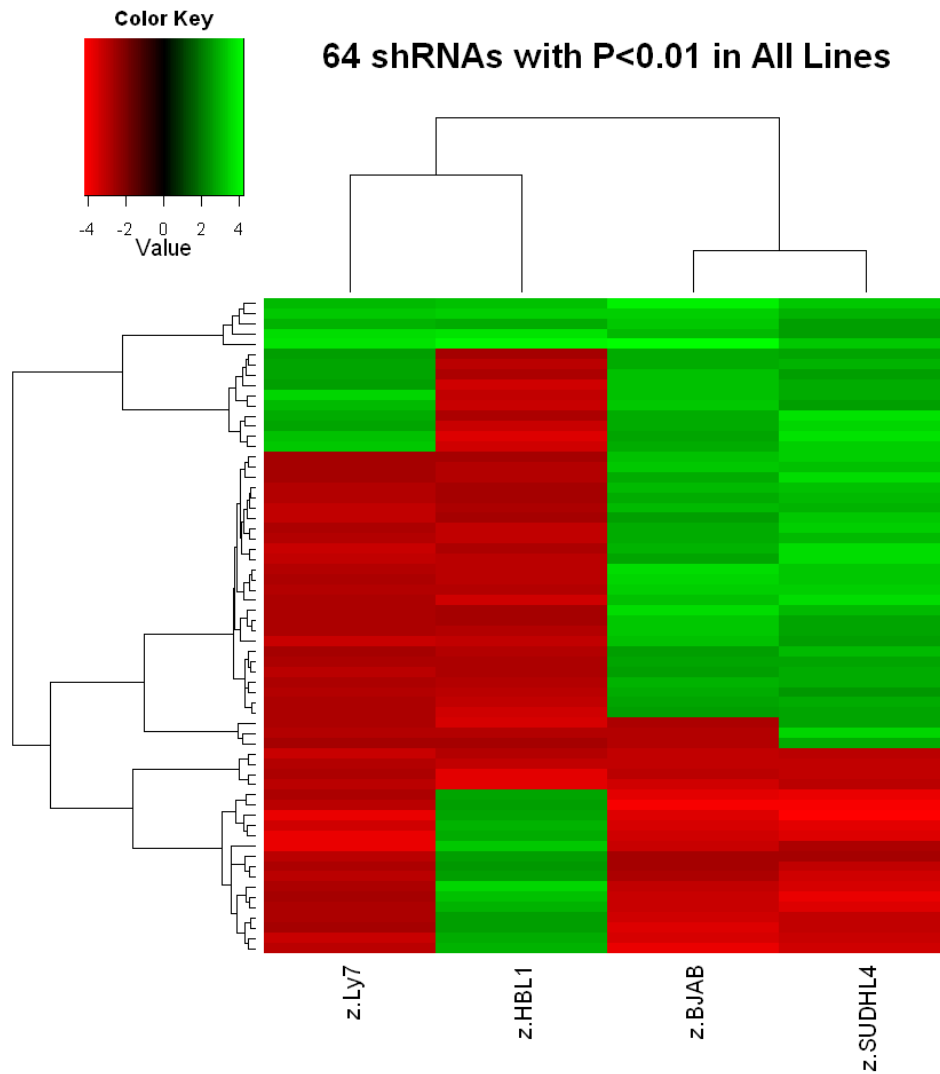


Figure 2-25 Heatmap of z-scores of significant depleted hairpins. Euclidian or correlation can be used for distance metrics and Wald method is suggested for hierarchical clustering.

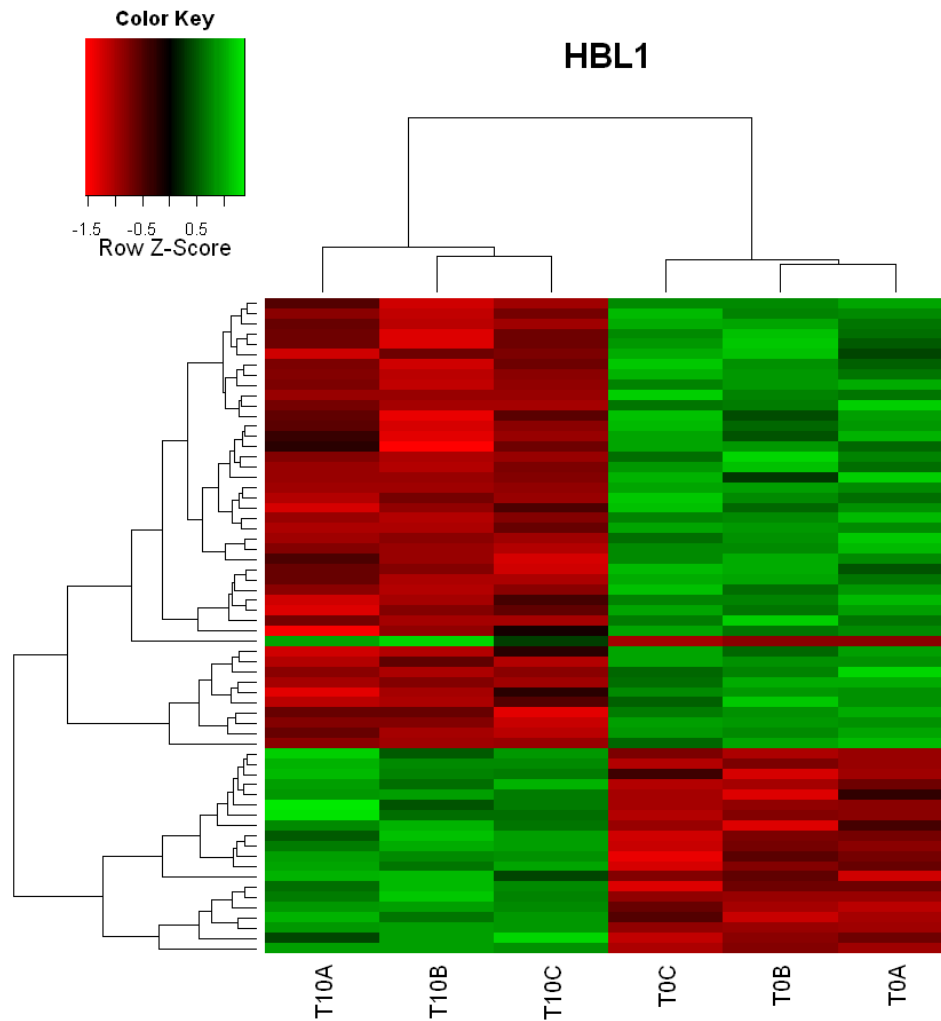


Figure 2-26 Heatmap of shSeq data of significant hairpins in different conditions. Euclidian or correlation can be used for distance metrics and Wald method is suggested for hierarchical clustering.

2.8 shMA / BHM: Scoring Gene Level Activity by Integrating Multiple Hairpins Targeting the Same Gene

The final goal of genome-wide of RNAi screening is to identify candidate genes that can be used as therapeutic targets, therefore it's more interesting to score gene

level activity from RNAi screening data. Because of the fact that multiple shRNAs targets the same gene in the shRNA library, statistical methods are needed to do gene-level differential representation analysis by integrating all hairpins for the same gene. I develop a package shMA (shRNA meta-analysis) using Bayesian hierarchical modeling (BHM) approach to combine multiple shRNAs targeting one gene. See Chapter 4 for more details about this algorithm.

2.9 Post-Analysis

With results of differential representation analysis or selected candidates, there are multiple post-analysis we can do, for example, functional enrichment analysis and sensitivity analysis.

2.9.1 Functional enrichment analysis

One interesting question we can ask about the selected candidates is that what functions or pathways they are enriched in. The way to answer it is to perform enrichment analysis in known functional categories or pathways.

There are multiple sources of functional databases we can use:

- The Gene Ontology [93] includes annotations of biological process, molecular function and cellular component for entire human or mouse genome.
- Pathway commons [94] is a collection of biological pathway information from public pathway databases including BioGRID, Nature Pathway, Reactome, KEGG, etc.

- Molecular signatures database (MSigDB) [95] is collection of annotated gene sets for use with GSEA software.

Various available methods or tools for enrichment analysis we can use include:

- DAVID [96] supports gene-annotation enrichment analysis using Fisher's exact test.
- Gene Set Enrichment Analysis (GSEA) [97] is K-S statistic based enrichment analysis method developed at Broad Institute.
- Gene Set Analysis (GSA) [98] introduces a new "maxmean" statistic for enrichment score by Brad Efron.

GSEA-type enrichment analysis of pathways or GO terms (Figure 2-27) uses differential representation results of all shRNAs or genes as the reference, for example, ranking from the most enriched to the most depleted. Classical weighted can be used to estimate the enrichment score, and gene label shuffling is commonly used to estimate significance in this small sample size situation.

I develop a new enrichment analysis method, Bayesian Set Enrichment Analysis (BSEA), using Efron's "maxmean" statistic and Bayesian inference. It outperforms GSEA and GSA. More details are in Chapter 5.

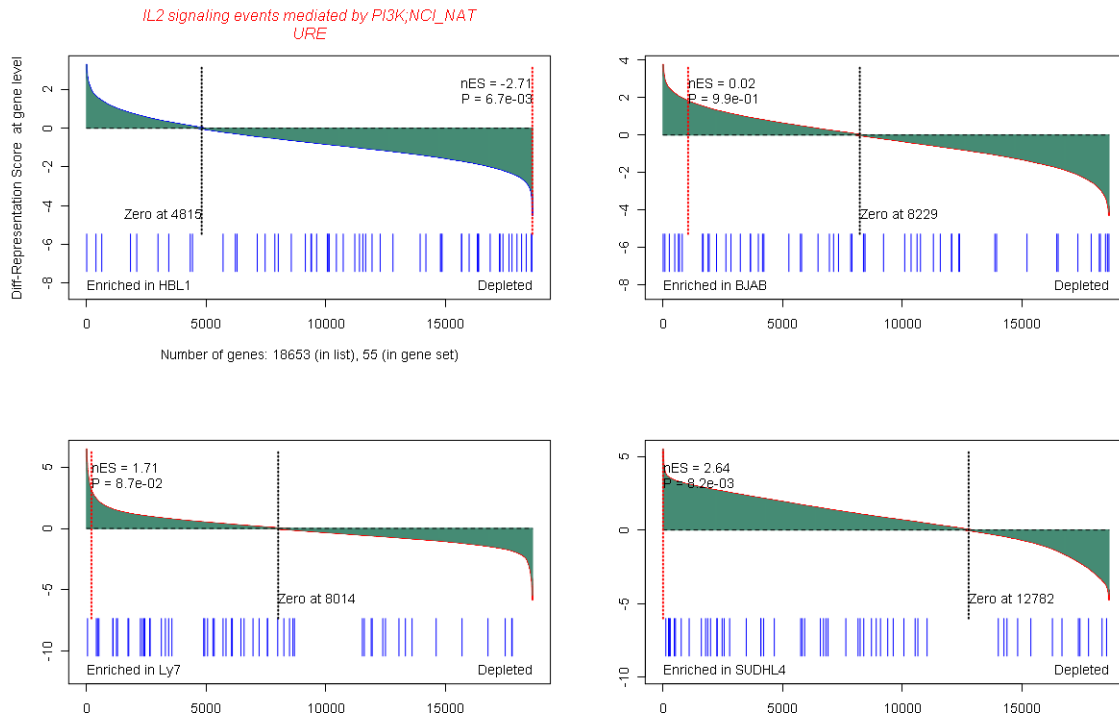


Figure 2-27 An example GSEA plot of pathway or GO gene sets in differentially-represented shRNAs. Y axis shows the z score of differential representation at shRNA level or gene level. The red dashed lines indicate normalized Enrichment Score (nES) and P value.

2.9.2 Sensitivity analysis

Another interesting analysis we can do with high-throughput shRNA screening data is sensitivity analysis, particularly when we have a number of RNAi screens across multiple cell types or disease contexts. Basically it asks a very general question: how sensitive is a cell line to respond to RNAi perturbation, or how difficult to kill certain type of cells by RNAi? The way to answer it is simply counting the number of depleted hairpins or genes in RNAi screen of each cell

line. The larger the number of significantly depleted genes, the more sensitive this cell line is to respond to RNAi perturbation, or the easier to kill this cell line by RNAi.

Based on this methodology, a sensitivity analysis plot can be generated by plotting the number of depleted hairpins of each cell line at different p-value cutoffs. Figure 2-28 shows the sensitivity analysis plot for a panel of 16 breast cancer cell lines covering major subtypes of breast cancer. The larger the area under the curve (AUC), the more sensitive the cell line is. So we can see that the most sensitive cell line is MCF10A, which is a normal line, while the most resistant one is SUM149PT, which is inflammatory breast cancer, probably the most aggressive form of breast cancer. If we group them into subtypes, there is a pattern of decreasing sensitivity from normal, to luminal to Basal A to Basal B type of breast cancer, or it is harder to kill Basal B than Basal A than Luminal than Normal cells, which is consistent with the aggressiveness we know about those forms of breast cancer. This pattern can be seen more clearly in the left panel of Figure 2-30.

Similarly, we can do the same thing for enriched hairpins which corresponds to genes that are suppressors of cell growth or survival such as tumor suppressors. However, as shown in Figure 2-29 and the right panel of Figure 2-30, there is no such pattern we observed in depleted hairpins. This might reflect the difference of essential genes (depleted) with tumor suppressor genes. There is no

preference of cell types to make the cell grow better, but to kill them depends on the cell type, which has a specific defense system.

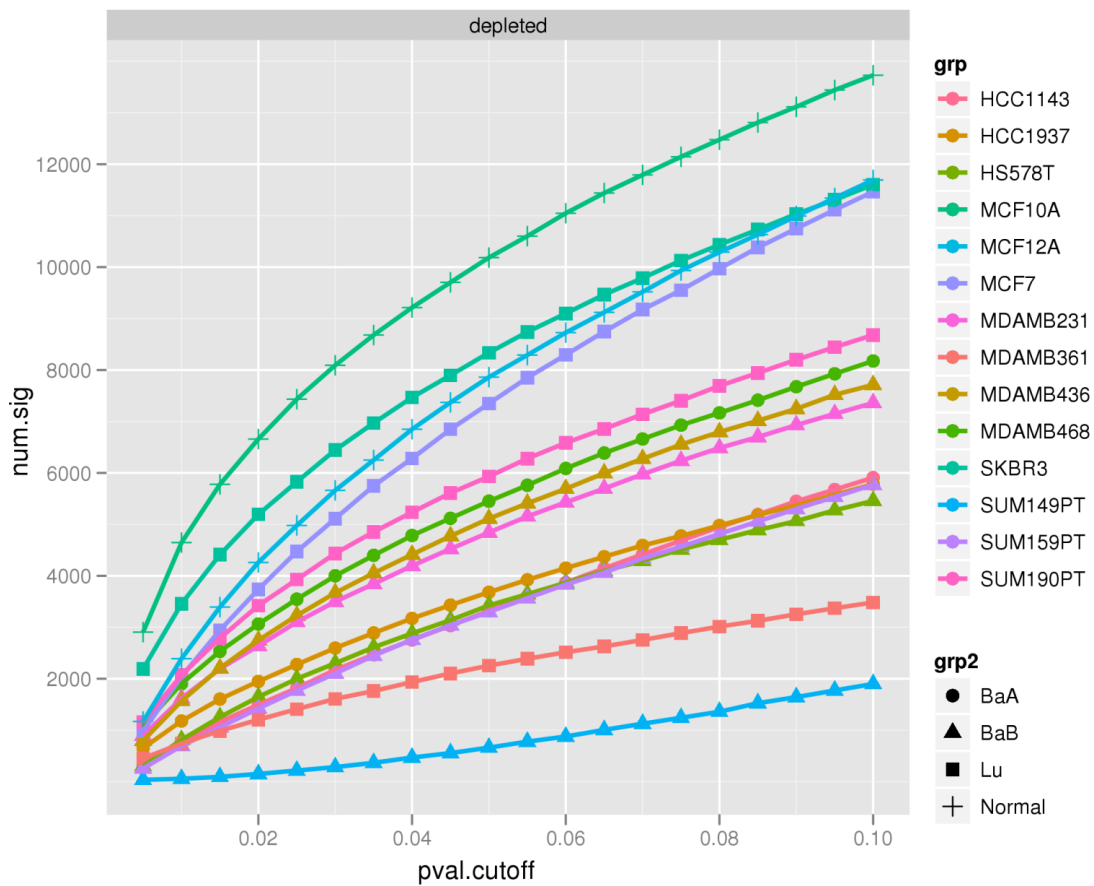


Figure 2-28 Sensitivity analysis. Number of significant depleted hairpins (y axis) in each of 16 breast cell lines at different p-value cutoffs (x axis). Cell lines represented by lines in different colors are classified into four groups represented by shape of dots.

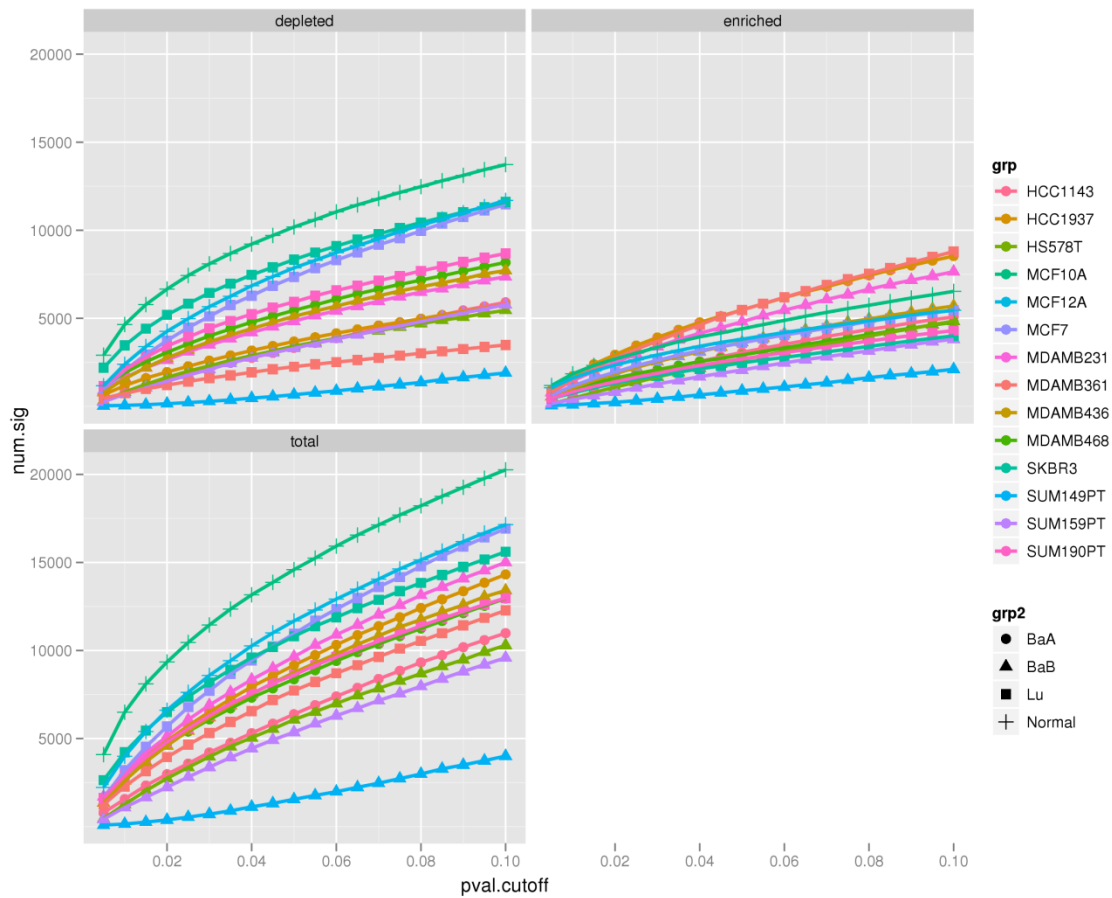


Figure 2-29 Sensitivity analysis of looking at depleted, enriched or both hairpins in the panel of 16 breast cell lines.

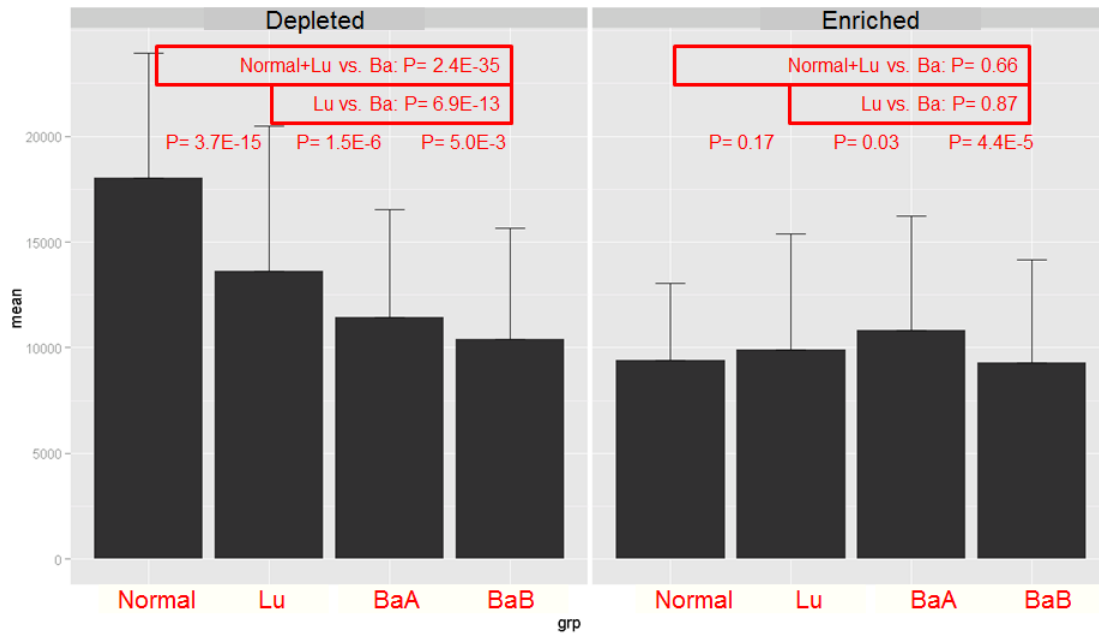


Figure 2-30 Statistical comparisons of sensitivity analysis results of 16 breast cell lines in both depleted and enriched cases. A Student's t-test is used to do the comparisons.

The sensitivity analysis of different cell types can be viewed in another perspective by counting the number of cell lines in each group that share certain number of depleted hairpins or genes (Figure 2-31). The lower the percentage at certain number of depleted genes, the more resistant this type of cells, or given a percentage of cell lines, the larger the number of depleted genes they share, the more sensitive this cell type is to respond to RNAi perturbations.

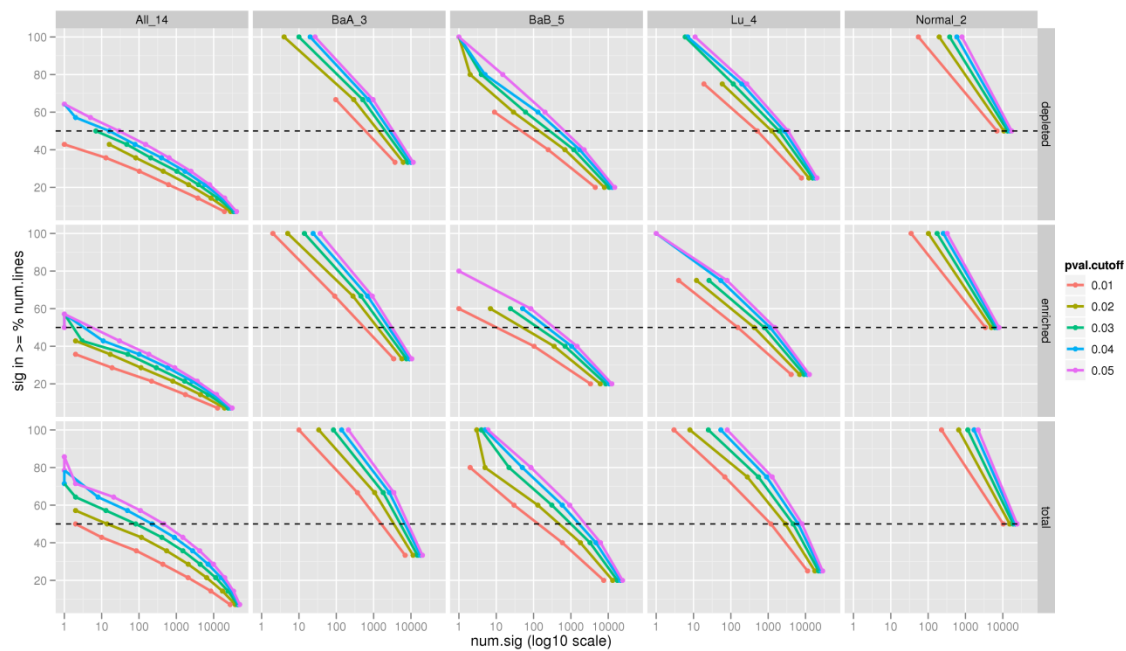


Figure 2-31 The percentage (y axis) of cell lines in each group share a certain number of common depleted hairpins (x axis) at different p-value cutoff (in various colors).

Chapter 3 Meta-Analysis of High-throughput RNAi Screening Data (the BHM algorithm)

3.1 Summary

Genome-wide RNA interference (RNAi) screening has emerged as a standard tool for systematic loss-of-function studies and therapeutic target discovery. Short hairpin RNA (shRNA) is commonly used because of its stabelness compared to small inference RNA (siRNA) that suffers transient silencing. However, due to high false-positive and false-negative rates, statistical analysis of high-throughput shRNA screening data remains challenging, particularly meta-analysis of multiple shRNAs targeting the same gene in the library to report gene-level activity. Here we propose a Bayesian hierarchical modeling (BHM) approach to tackle this challenge and validate that this novel “modeling-all-together” strategy outperforms classical “separate-and-combine” approaches. Surprisingly, our BHM algorithm works extremely well when applied to relatively low-quality screens, which is observed in about 80% of shRNA screens. Moreover, this hierarchical modeling framework can be useful in similar applications.

Keywords: Bayesian hierarchical modeling, high-throughput RNAi screening, shRNA screening, separate-and-combine, modeling-all-together

3.2 Introduction

RNA interference (RNAi) has emerged as a standard technique for studying phenotype-specific gene function in many organisms (e.g. plants, fungi and animals) via suppression of gene expression [48-51]. RNAi-based gene silencing can be achieved by the use of small interfering RNAs (siRNAs) or short hairpin RNA (shRNA) expression vectors. Between the two approaches, shRNA is more feasible due to the siRNA-specific problem of transient inhibition of gene expression and inefficient transfection into non-dividing cells. However, shRNA can be stably integrated into a target cell genome via retroviral or lentiviral gene transfer, resulting in the permanent reduction of the targeted gene product. Two major shRNA expression libraries, GIPZ library [52-54] and TRC library [55] that target the entire human genome have been generated to facilitate functional analysis of the whole transcriptome through loss-of-function genetic studies.

In a genome-wide shRNA screen, a large population of cells is transfected with a pool of different shRNA lentiviral vectors and shRNA hairpins which subsequently integrate into cells' genomes. These transduced cells can be used for two main applications. One is to identify genes that are essential for cell survival or growth, thus representing potential therapeutic targets for cancer and other human diseases. Hairpins of such essential genes will be dropped out or under-represented as time evolves. The other is to identify genes that modulate response to cell perturbation, such as chemotherapeutic agents. To do this the transfected cells are split into two groups, one treated with an agent of choice,

the other with a vehicle control. With this selective pressure, depleted or enriched hairpins will represent candidates that increase sensitivity or resistance of cells to a therapeutic agent.

To quantify and analyze shRNA hairpins extracted from genomic DNA, microarray hybridization is commonly used with the advantage of low cost and flexibility. It employs PCR-amplified shRNA template sequence pools extracted from shRNA library-transduced cells under test as well as reference conditions. Each PCR fragment is labeled with a different fluorophore, followed by hybridization of both pools to the same array, or labeled with the same fluorophore followed by hybridization to multiplex arrays. Taking the two-color microarray as example, the ratio of signal intensities of two colors (Cy3, Cy5) for each probe sequence reflects the relative abundance of cells expressing the corresponding shRNA construct under the sample condition as compared to the reference. Consequently, shRNA hairpins that target essential genes for cell viability will be depleted from the pool, showing low values of signal ratio, whereas shRNA constructs that target genes inhibiting cell growth such as tumor suppressors will be enriched, showing high values of signal ratio.

Because each gene represented in the shRNA library is targeted by an average of 2-3 hairpins in GIPZ library and 5 hairpins in TRC library, we must integrate evidences of all shRNAs for one gene to uncover the gene-level activity. Traditional methods usually employ a “separate-and-combine” approach – scoring shRNA individually first and then picking up representative shRNA with

high score or combining scores of all shRNAs targeting the same gene. Different algorithms that select or combine shRNAs have been proposed including choosing the second best or most depleted shRNA [59] (RIGER_SB), averaging the best two shRNAs[59] (RIGER_WS), performing enrichment analysis of all shRNAs targeting one gene against all shRNAs in the library [59] (RIGER_KS), or probability-based averaging of all shRNAs per gene [99] (RSA). A limitation with these types of approaches is that they rely on accurate estimation of individual shRNA activity which is very difficult to achieve in common large-scaled shRNA screens with a small sample size. Also off-target effects, low silencing efficiency, small differences among shRNAs targeting one gene, and microarray noise will make heuristic selection of shRNAs to represent gene-level behavior problematic and cause high false-positive rates.

To overcome the above drawbacks, in this study we propose a novel Bayesian hierarchical modeling (BHM) algorithm to report gene-level activity. Hierarchical modeling [89, 100], also known as multilevel modeling, has been increasingly used in large-scaled `omics studies for its robustness [101]. In this context, BHM algorithm puts all shRNAs targeting the same gene together, instead of separating them, and then fits a linear mixture model by allowing variation of activities among different hairpins, also known as random effects. This 'modeling-all-together' strategy improves parameter estimation by increasing sample size and reduces prediction error and false-positive rate by integrating information of all shRNAs. Furthermore, we employ Bayesian inference with Markov chain Monte Carlo (MCMC) techniques to further improve accuracy and robustness of

scoring metrics. Evaluation results based on benchmark shRNA screens designed for profiling essential genes suggest that our BHM method outperforms classical 'separate-and-combine' algorithms significantly on sensitivity and precision, and especially BHM dominates the others when the data is of low quality, which accounts for about 80% cases of normal high-throughput shRNA screens.

3.3 Profiling Cell Essential Genes by Microarray-based RNAi Screens

As described in the previous section, negative genome-wide RNAi screening is commonly used to identify essential genes for proliferation and viability in cancer cells. A typical procedure of microarray-based pooled shRNA screening to profile cell essential genes is shown in Figure 3-1-A. The pool of shRNA plasmid vectors is transfected into a target cell population at a multiplicity of 0.3 to achieve one shRNA per cell. Infected cells are then harvested for X doubling times in triplicates. Genomic DNAs are extracted from cells collected at T0 and TX, PCR-amplified, labeled and hybridized to multiplex microarray.

Analysis of microarray readout involves several steps (Figure 3-1-B). Signals representing shRNA relative abundance are extracted and processed with background correction and normalization. Differential representation analysis (DRA) on processed data at TX and T0 time points are performed to identify under-represented shRNAs at TX time whose target genes are potential candidates of essential genes. Because of multiple shRNA probes per gene in

the library, DRA can be conducted at individual shRNA level and integrated gene level. Classical “separate-and-combine” approaches score shRNAs for one gene separately and then combine them to derive gene level score; however, our newly-proposed “modeling-all-together” methodology skips the individually scoring step and fits a hierarchical model into all data for one gene to estimate gene level activity directly. Associated statistical metrics with gene-level behaviors including fold-change, p-value, false discovery rate or z-score will be reported as well.

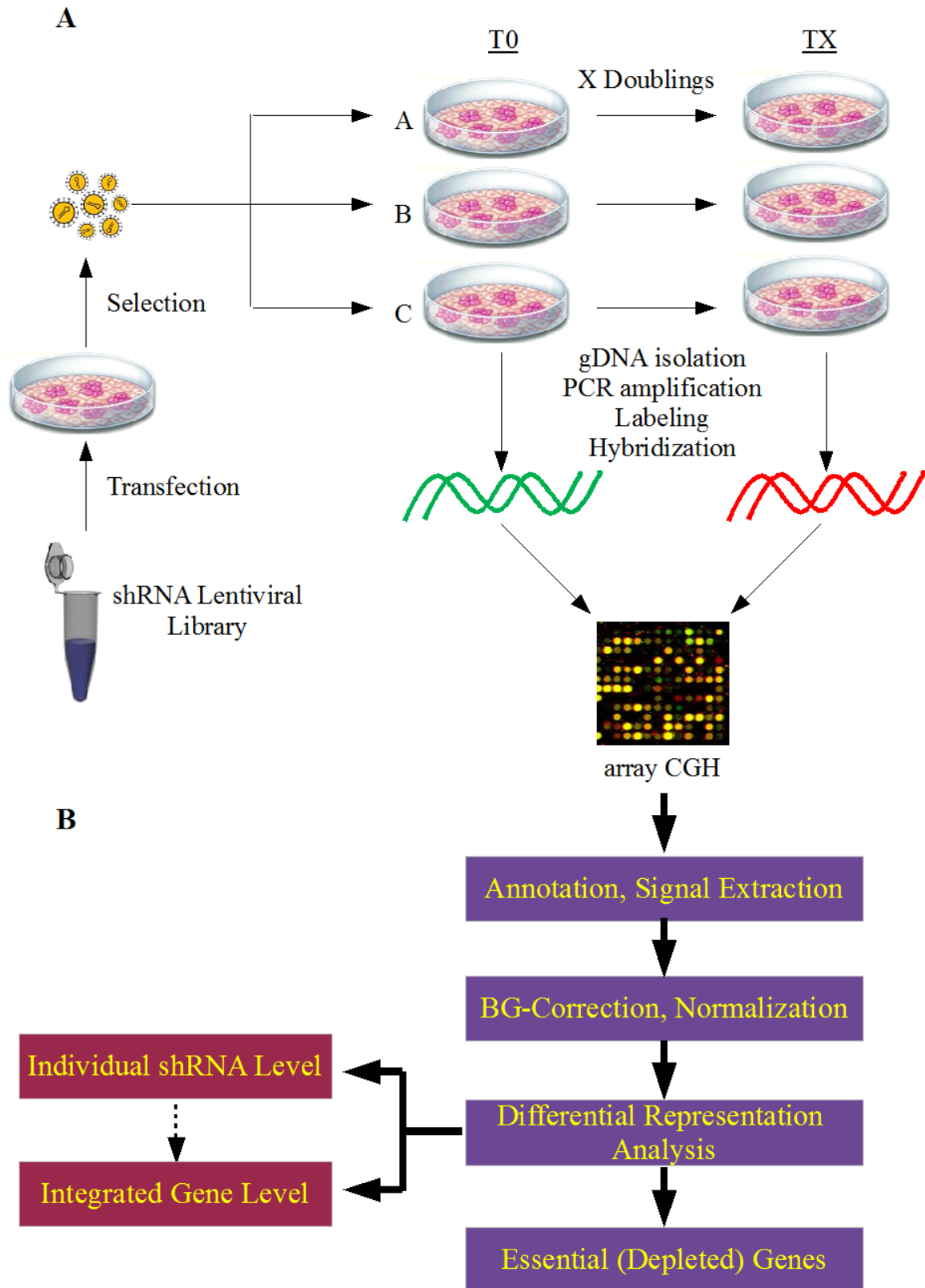


Figure 3-1 Outline of microarray-based shRNA screens to profile cell essential genes (A) experimental procedures and (B) analysis pipelines

3.4 Classical “Separate-And-Combine” Approach

All previous methods to estimate activity of genes targeted by multiple shRNAs in large-scale shRNA screening data are based on a “separate-and-combine” strategy as illustrated in an example from benchmark datasets that a gene is targeted by three shRNA clones (Figure 3-2-A). There are two well-developed algorithms for this type of approach: RNAi gene enrichment ranking [59] (RIGER) and redundant siRNA activity [99] (RSA). RIGER has three sub-algorithms to integrate multiple shRNA scores including Kolmogorov-Smirnov statistic-based enrichment analysis (RIGER_KS), weighted sum of the best two hairpins (RIGER_WS) and the second best hairpin (RIGER_SB). RSA employs a hypergeometric distribution or Fisher’s exact test-based statistical method to rank gene activities. These algorithms can be reclassified into summarizing all shRNAs targeting the same gene (RIGER_KS and RSA) and heuristic selection of representative shRNAs (RIGER_SB and RIGER_WS).

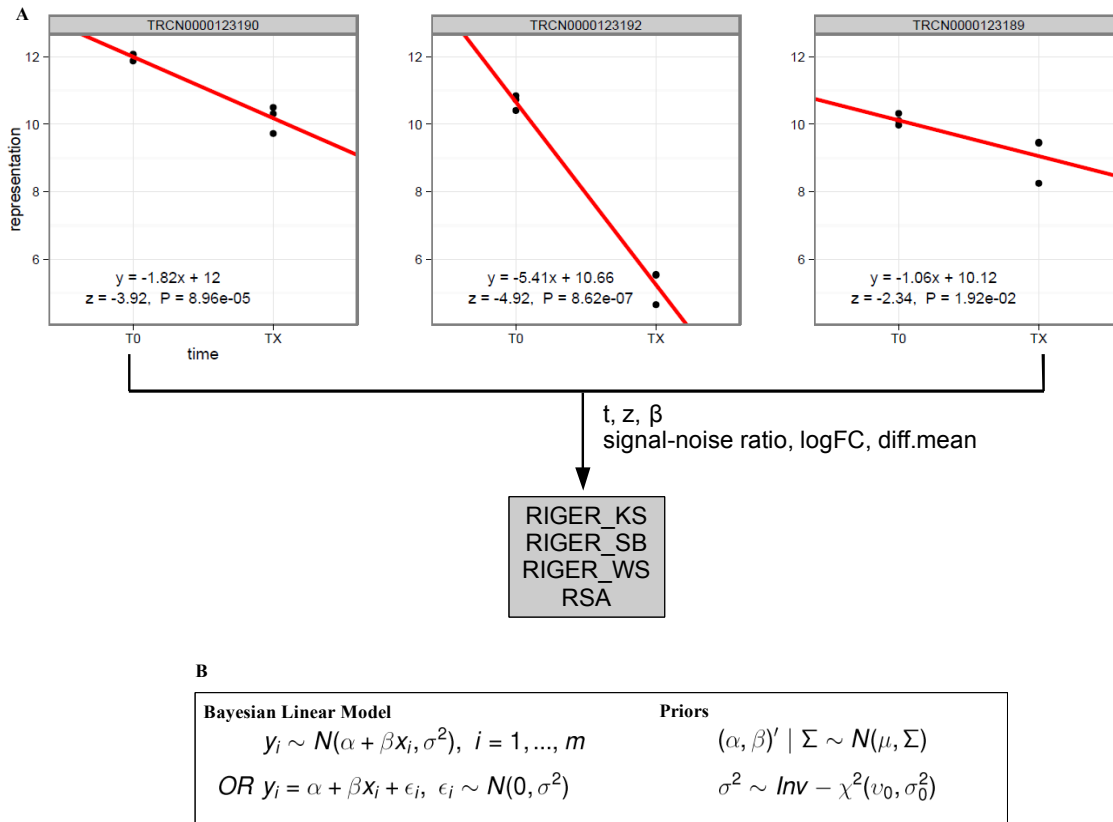


Figure 3-2 Separate-and-combine approach. (A) An example of three shRNAs targeting KPNB1 gene from MCF7 dataset is selected to illustrate this approach. A Bayesian linear model is fit into data of each shRNA respectively. Estimated parameters (fitted lines in red) and summary statistical metrics are displayed on bottom left of each shRNA plot. Z scores and p-values are calculated by Wald test using a standard Gaussian as null distribution. Individual shRNA scores as input for algorithms to combine them can be calculated by t (Student's t -statistic), z (z -statistic of β in linear model), β (the coefficient in the linear model), signal-noise ratio (mean difference of TX vs. T0 over sample standard deviation), logFC (logarithm of fold change of TX vs. T0) and diff.mean (mean difference of TX vs. T0). (B) In the linear regression model under Bayesian framework, y_i indicates time point, TX or T0, and x_i represents shRNA abundance for sample i ; m is the sample size of the corresponding shRNA; noise follows a Gaussian distribution with mean 0 and variance σ^2 ; β is the parameter of interest, indicating the silencing effects on cell viability by the shRNA in consideration. As for priors, a

two-variable multi-Gaussian is set for coefficients and an Inverse-Gamma is for variance.

Various metrics have been proposed to score individual shRNA behavior (Figure 3-2-A) at TX and T0 time points including Student's t-statistic, z-statistic, coefficient of linear regression model, signal to noise ratio, logarithm of fold change, and difference of mean. Student's t-statistic or z-statistic of coefficient in linear model is commonly used due to their statistical integration of replicate variance. Student's t-test is equivalent to a linear model with Gaussian noise. However, with the fact that the sample size in this context is usually small, a Bayesian linear model with a Gaussian prior for coefficients (Figure 3-2-B) is suggested for its robustness.

3.5 Our “Modeling-All-Together” Approach: Bayesian Hierarchical Model

Instead of two-step analysis, we propose a “modeling-all-together” approach to fit a complex hierarchical or multilevel model into data of all shRNAs targeting the same gene. Particularly, we establish the model within Bayesian framework to overcome inaccurate estimation problems from small sample size and microarray noise. Bayesian hierarchical model (BHM) introduces an additional level to the classical linear model with parameter of coefficient corresponding to silencing effect of shRNA group (indexed by j) and the parameter is assigned its own distribution (Figure 3-3-A). We have also allowed the intercept to vary across

hairpin classes in a similar manner. Additionally, our Bayesian analysis requires us to specify prior distributions for the parameters of coefficients and variance, which follow Gaussian and Inverse-Wishart or Inverse-Gamma distribution respectively (Figure 3-3-A). The multilevel model can be rewritten as a linear mixture model (Figure 3-3-B), in which “fixed effect” corresponds to gene-level activity and “random effect” reflects the variation of silencing effects from different shRNA classes targeting the same gene.

Using the same example in previous section, the data of three shRNA classes (in three different colors) is modelled together and parameters are estimated by MCMC simulations (Figure 3-4). The red solid line indicates the integrated behavior of all three shRNA groups, and each dashed line reflects individual shRNA-level behavior by adding random deviations to fixed gene-level activity on both slope and intercept.

The hierarchical model can be viewed and interpreted from another conceptual perspective. As shown in Figure 3-5, a middle layer is introduced to indicate the shRNA level, thus forming the hierarchy structure of the model. All data points for one shRNA are clustered together, but all shRNAs targeting the same gene are fit in the same model with allowance of their internal difference.

A

Model

$$y_{ij} \mid \alpha_j, \beta_j, \sigma^2 \sim N(\alpha_j + \beta_j x_{ij}, \sigma^2) \quad [\text{Level 1}]$$

$$(\alpha_j, \beta_j)' \mid \mu, \Sigma \sim N(\mu, \Sigma) \quad [\text{Level 2}]$$

$$i = 1, \dots, n, \quad j = 1, \dots, J$$

Priors

$$\mu \mid \Sigma \sim N(\mu_0, \Sigma / \kappa_0)$$

$$\Sigma \sim \text{Inv - Wishart}_{v_0}(\Lambda_0^{-1})$$

$$\sigma^2 \sim \text{Inv - } \chi^2(v_1, \sigma_1^2)$$

B

Rewriting the Model

$$y_{ij} = \underbrace{(\alpha + \beta x_{ij})}_{\text{Fixed Effect}} + \underbrace{(\alpha_j + \beta_j x_{ij})}_{\text{Random Effect}} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

Figure 3-3 Modeling-all-together approach. Bayesian hierarchical modeling (A) The data of all shRNAs targeting one gene can be fit by a hierarchical model, in which the extra level is indexed by j , indicating the shRNA group the sample belongs to. Sample index i is up to n , the total number of samples for one gene; j is up to J , the number of shRNA classes. Parameter μ , a vector of slope and intercept, reflects the gene-level activity and allows variation for each shRNA class. Conjugate priors are set for parameters. (B) The model can be rewritten to a two-component mixture model in which “fixed effect” corresponds to gene-level behavior and “random effect” indicates the noise of each shRNA group.

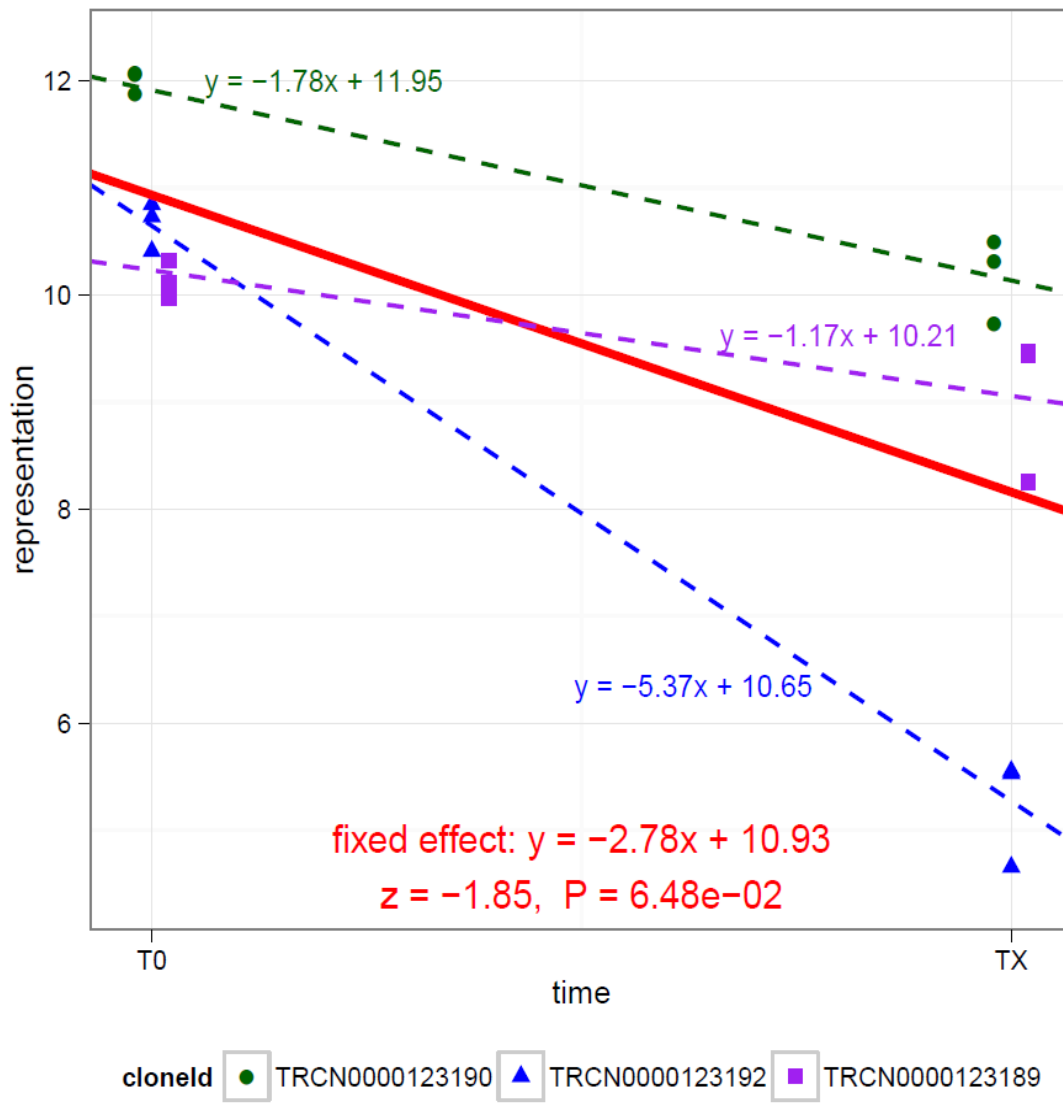


Figure 3-4 Modeling-all-together approach. Bayesian hierarchical modeling. A practical application of the Bayesian hierarchical model to the example in Figure 2 is summarized in the plot. Red solid line indicates fitted gene-level/fixed effects in the model. Estimated parameters and summary statistics including z-statistic and p-value are displayed on bottom middle. Each colored dashed line reflects individual activity of each shRNA class by adding random effect to fixed gene-level effect.

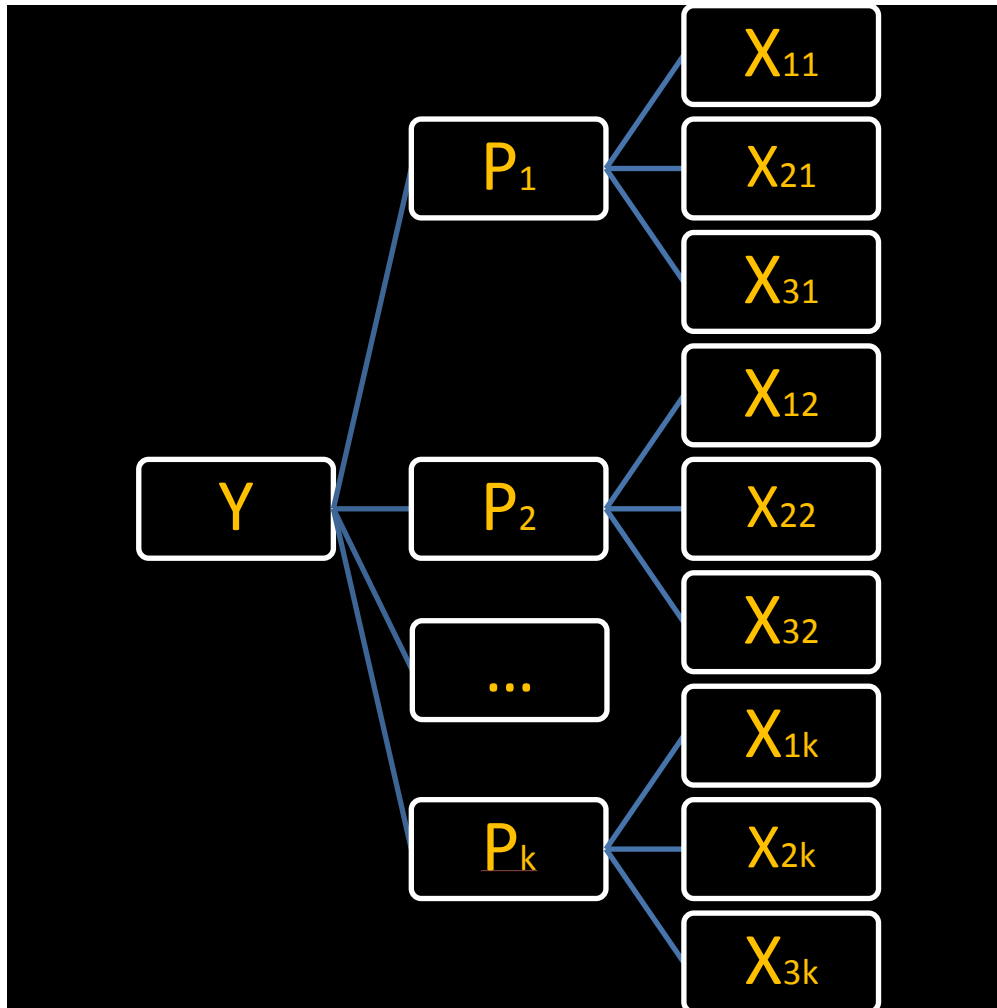


Figure 3-5 A different representation of the hierarchical model. A middle layer is introduced to indicate the shRNA level. All data points for one shRNA are clustered together, but all shRNAs targeting the same gene are fit in the same model with allowance of their internal difference.

3.6 Benchmark Datasets and Validations

To benchmark different algorithms for meta-analysis of shRNAs targeting the same gene, we selected three representative datasets from a panel of 72 RNAi screens designed for profiling essential genes in breast, pancreatic, and ovarian

cancer cell lines [102] using 80K pooled TRC library targeting 16K human genes with an average of 5 shRNAs per gene. In the original study [102], cells at three time points including T0 were collected for each cell line to investigate the dynamics of shRNA behaviors. However, due to our interest in detecting depleted genes essential for cell viability, we only selected the final time point and T0. This aids in the detection of depleted genes that are essential for cell viability, no matter whether they are dropped out early or late, and early depleted shRNAs would still remain under-represented in later time point. This two-time-point design is also generally applied in literature for cost consideration, but a long evolution time is usually required to capture late-dropped-out genes. In Figure 2, the three types of tumors selected, MCF7, HPAFII and OVCAR5 for breast, pancreatic, and ovarian cancer respectively, are represented by three lines. More importantly, they also represent high (MRC: minimum replicate correlation > 0.9), medium (MRC between 0.8 - 0.9) or low (MRC < 0.8) data quality categories in terms of the consistence between replicates (Figure 3-7), each of which accounts for 22%, 50% and 28% of the total 72 screens (Figure 3-6).

Without a gold standard in place for selecting human essential genes, housekeeping and evolutionary-conserved genes likely to be critical for cell viability were used to benchmark methods of probing essential genes [102]. We followed this validation method to compare our new BHM approach with existing algorithms of reporting gene-level potency to be essential genes from shRNA-level evidences. In our study, we collected four independent gene sets as

references – two adapted from previous study [102] and two more recent studies on human housekeeping genes [103, 104]. Housekeeping or ortholog genes that were not present in the shRNA library were filtered out. We then determined the percentage of overlapped genes of reference set with top k hits predicted as essential genes by each method. To avoid selection bias on k, we sampled k from 0 up to 1000 with a sliding widow of 5. The larger intersection with reference gene set the algorithm produces consistently, the more powerful this method is.

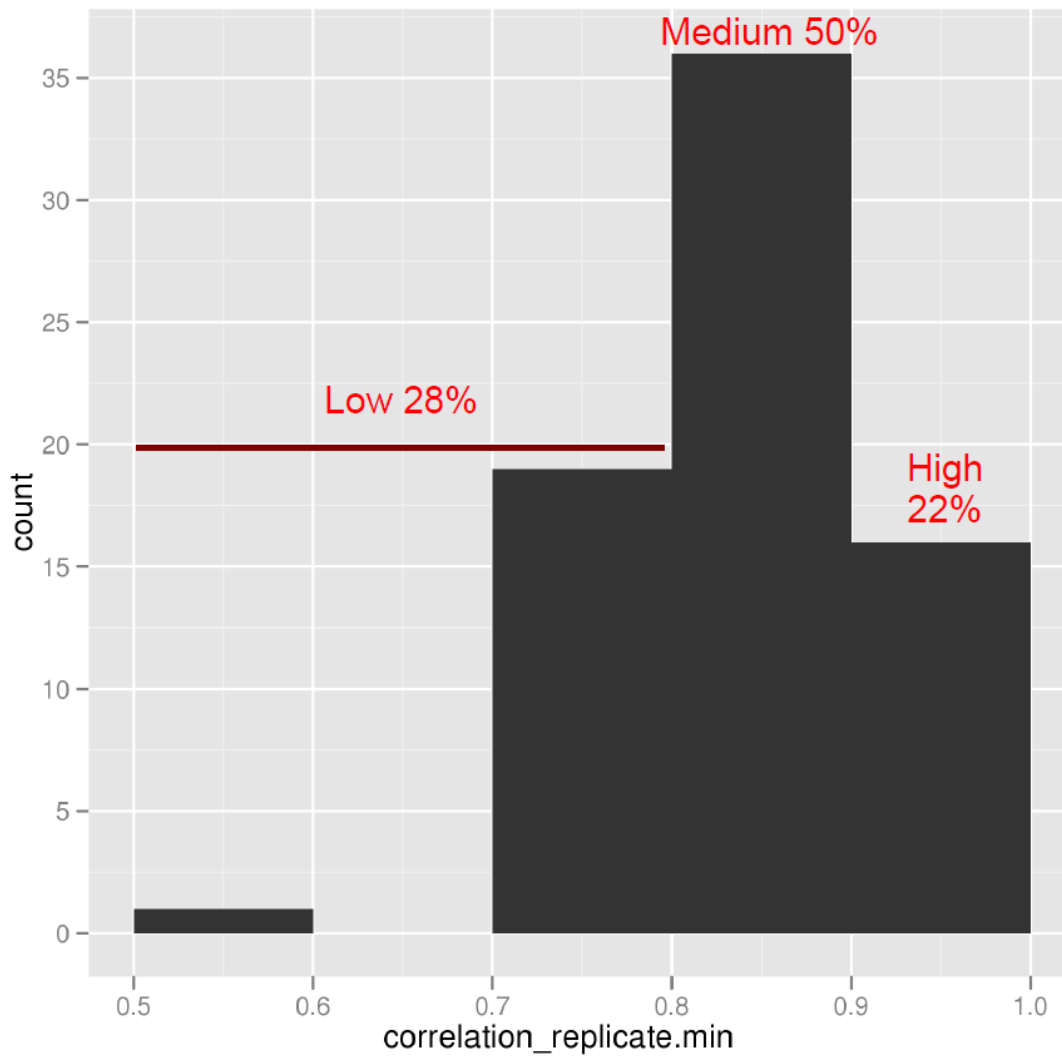


Figure 3-6 Distribution of data quality (MRC: minimum replicate correlation) for the panel of 72 shRNA screens: High (MRC>0.9): 22%, Medium (MRC in 0.8-0.9): 50%, Low (MRC<0.8): 28%.

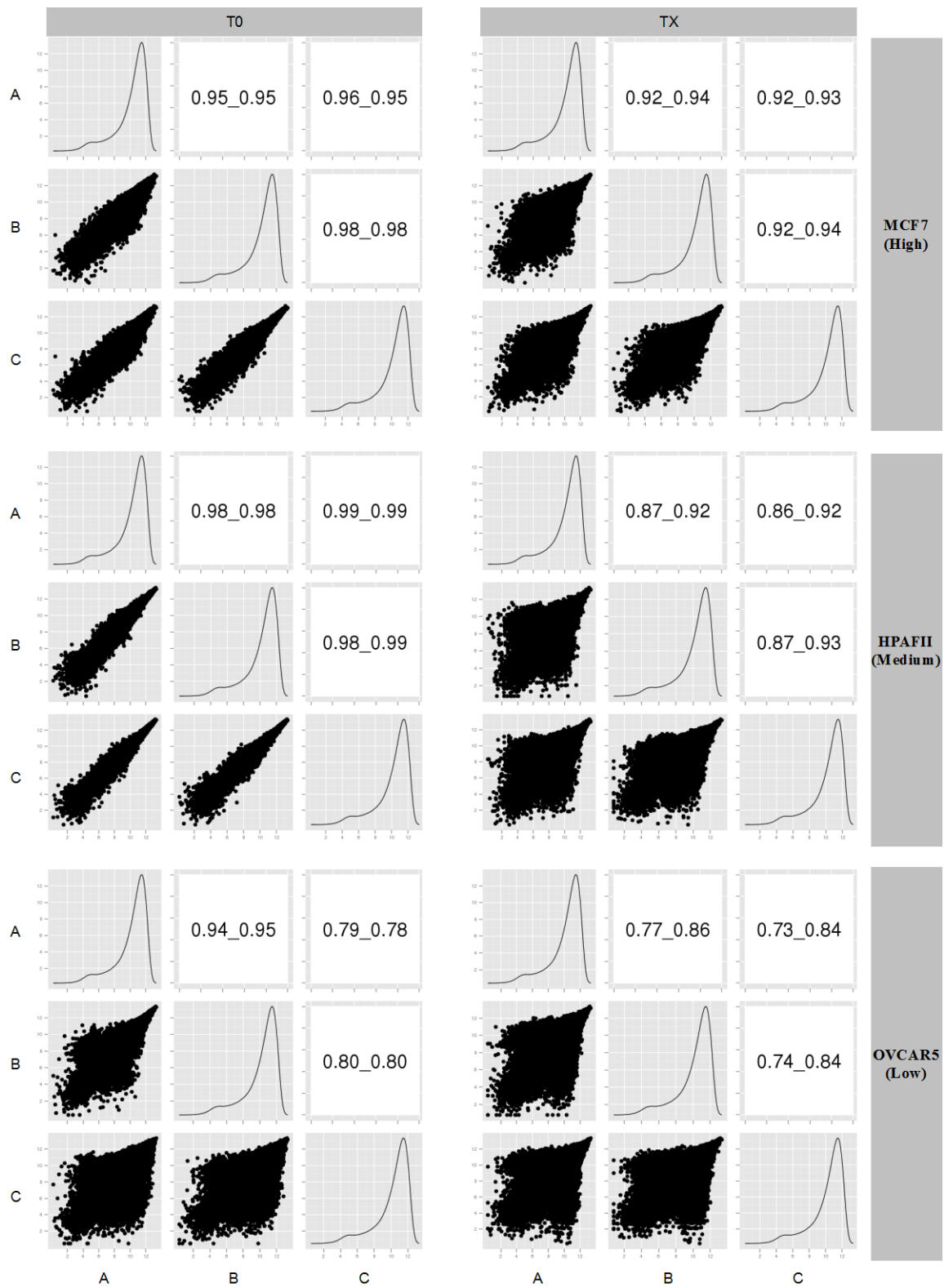


Figure 3-7 Quality of benchmark datasets. Each dataset has two time points (T0 and TX) and three replicates (A, B and C). Each sub-figure displays the scatter plots on bottom left and correlations (Pearson_Spearman) on upper right between any two replicates of the corresponding group, and density distribution of shRNA abundance in each replicate on the diagonal. Low variability of scatter plots, high correlations and high similarity of distribution plots indicate good consistence of replicates, thus good quality of the data. The label (High, Medium or Low) after each cell line name indicates the data quality group it belongs to, defined by MRC (minimum replicate Pearson correlation), the “bottle-neck” of each dataset. MCF7 (MRC > 0.9), HPAFII (MRC between 0.8-0.9), and OVCAR5 (MRC < 0.8) represent 22%, 50% and 28% of 72 screens respectively.

3.7 Evaluation Results

3.7.1 BHM dominates RIGER and RSA

To evaluate the performance of our BHM algorithm compared with classical RIGER (RIGER_KS, RIGER_SB, RIGER_WS) and RSA methods, we applied the validation strategy of using housekeeping and conserved genes as “gold standard” for essential genes to three benchmark dropout shRNA screens. In RIGER and RSA methods, we used t-statistic for individual shRNA scoring that was commonly used in their software packages. We plotted the intersection percentage of each reference gene set against top 0 up to 1000 hits inferred as essential genes by each algorithm in each testing dataset (Figure 3-8). The Y-axis corresponds to sensitivity or recall rate while the X-axis reflects precision or prediction rate. Therefore the area under the curve (AUC) of each algorithm is

proportional to its power to identify a large percentage of true hits while maintaining a high precision.

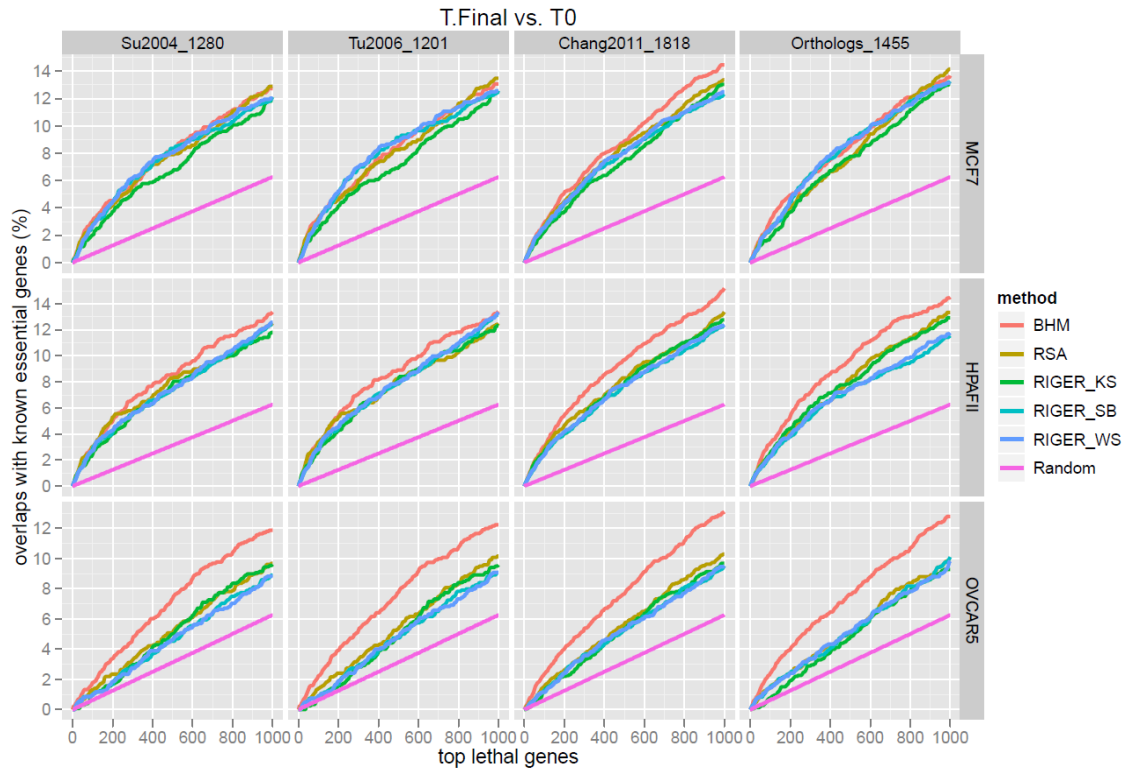


Figure 3-8 Evaluation results of final time point data. Housekeeping or conserved ortholog genes can be used as reference gene set to evaluate algorithms to detect essential genes from RNAi screens. Each colored curve shows the percentage of each reference set (“name”_“number of genes in the set”) intersected by top 0 to 1000 hits predicted as essential genes by the corresponding algorithm in each dataset. The slope of “Random” method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is.

From the evaluation plot results (Figure 3-8), first we noticed that all methods in consideration performed consistently better than random selection with p-value of

all paired comparisons against random method by Student's t-test (also used by the following similar comparisons) $< 4.4E-10$. This was expected as it indicated that all three pooled shRNA screens were informative to detect human essential genes. Second, AUC curves of RIGER_KS, RIGER_WS, RIGER_SB and RSA were mixed together in all situations without clear separation. This suggested that there was little difference among "separate-and-combine" type of methods. Third, BHM dominates RIGER and RSA in HPAFII and OVCAR5 studies (HPAFII: $P < 0.05$ in all 16 comparisons and $P < 0.01$ in 14 cases; OVCAR5: $P < 1.7E-8$ in all 16 comparisons), though it had unclear advantage in screen of MCF7 cells ($P < 0.05$ in 6 comparisons, $P > 0.05$ in the other 10 cases), which was actually explained by the association with data quality later.

The above conclusions are also supported by the analysis of the data at a different time point (Figure 3-9).

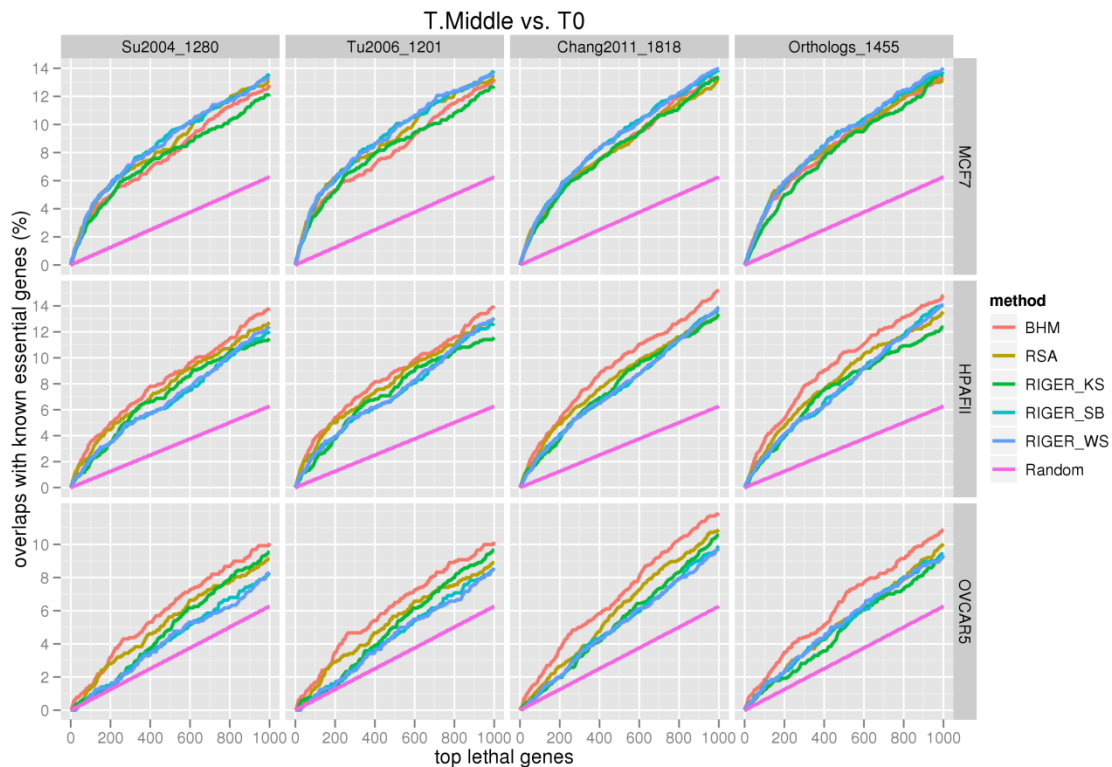


Figure 3-9 Evaluation results using middle time point data. Housekeeping or conserved ortholog genes can be used as reference gene set to evaluate algorithms to detect essential genes from RNAi screens. Each colored curve shows the percentage of each reference set (“name”_“number of genes in the set”) intersected by top 0 to 1000 hits predicted as essential genes by the corresponding algorithm in each dataset. The slope of “Random” method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is.

Interestingly, we observed that the advantages of BHM over RIGER or RSA were monotonically increasing from studies of MCF7 to HPAFII to OVCAR5, which exactly matched the decreasing pattern – from high to medium to low replicate consistence – in data quality of these three screens. This confirmed our

expectation that our “modeling-all-together” strategy with BHM algorithm consistently outperformed classical “separate-and-combine” approaches on statistical power and robustness. More importantly, about 80% of shRNA screens in the panel of 72 cancer cell lines had medium or low data quality, in which BHM dominated the other methods significantly. If the data quality of the screens were high such as MCF7 in this study, methods would not matter too much because all methods reached the optimum. However, high quality data in high-throughput RNAi screens was rarely achieved due to off-target effects, measurement noise, and small sample size, which made our improved BHM method extremely valuable to produce trustable hit decisions.

3.8 Discussion

Meta-analysis of shRNA screening data to report robust gene-level behaviors remains difficult. We proposed a novel “modeling-all-together” strategy, specifically a Bayesian hierarchical modeling algorithm, to address this problem. The evaluation results demonstrated that our BHM method outperformed traditional “separate-and-combine” approaches (RIGER and RSA) dramatically in general, and especially dominated all the other methods in about 80% cases when the data was in relatively low quality.

Hierarchical modeling, also known as partial pooling, can be viewed as a compromise between two extremes. One extreme, complete pooling, assumes the equal knockdown effect across all shRNA classes targeting the same gene. The other extreme, no pooling, ignores the similarity of the replicates within one

shRNA group and treat each hairpin replicate separately. The assumptions of these two extreme methods are too strong for shRNA screening design to be considered for integration of multiple shRNA evidences because different shRNAs targeting the same gene in the library might have significantly different silencing efficiencies. Hierarchical modeling comprises two extremes by allowing between-group variance and considering within-group effects, thus making an appropriate solution to this question.

The problem of multiple comparisons can also disappear in Bayesian hierarchical models [105]. Partial pooling in hierarchical models shifts estimates toward each other whereas classical procedures for multiple comparison correction typically adjust p-values corresponding to intervals of fixed width. Thus BHM fitting results in reliable and conservative estimates for main effects or gene-level effects in this context.

For “separate-and-combine” strategy, a few other possible algorithms might be considered to integrate shRNA-level scores for the same gene, for example, Fisher’s method [106] to combine signed p-values, or Stouffer’s method to combine z-statistics [107]. However, these integrating p-values or z-scores methods easily over-estimate the significance of gene-level activity and generate a long list of significant candidates. Also, they ignore the magnitude of knockdown effects for each hairpin by only considering the statistical significance of how the effect is away from zero, and require strong assumptions. Thereby,

these methods might not be comparable to our BHM algorithm, or could be even worse than the other “separate-and-combine” methods.

Additionally, other enrichment analysis algorithms such as GSA [98] have been used in this context [108] and might perform better than KS-based GSEA method; however, these algorithms still bear the drawbacks of “separate-and-combine” strategy, making them less powerful than BHM. The valuable point from enrichment-type methods that might improve BHM is to borrow information from all shRNAs or genes in the library because current BHM algorithm only considers shRNAs corresponding to one gene. Looking at entire list of candidates might produce more robust statistics for cut-off based hits selection, but probably would not change the rank of a gene as a potential candidate.

Next generation sequencing (NGS) has recently emerged as a cost-effective technology for quantitatively measuring the abundance of short-length DNA or RNA in a short time, and this large-scale parallel sequencing has been used in pooled shRNA screens [61, 108]. Compared to microarray-based approaches, NGS offers several potential advantages in terms of coverage of targeting genes, flexibility of input library, scalability and dynamic range, which might make it dominate the technology for RNAi screening in the near future. In this study, we only focused on microarray-based shRNA screening data. However, our approach can be extended to analyze NGS-based shRNA screens by employing different underlying models for example Poisson distribution for discrete

sequence count-based shRNA screening data, instead of Gaussian model generally used for continuous microarray data.

In summary, we developed a novel hierarchical modeling algorithm within Bayesian framework for meta-analysis of shRNA screening data. This “modeling-all-together” strategy dominates classical “separate-and-combine” methodology to analyze such noisy high-throughput data. However, this approach can be generalized and applied to any similar meta-analysis problem in which multilevel can be formulized.

Chapter 4 NetBID2: Network-based Bayesian Inference of Disease Drivers

4.1 Introduction and Motivation

4.1.1 The era of post-genomics in cancer

Advances in human genetics and molecular medicine have driven progress in our understanding of cancer biology. Development and improvement in DNA copy number, gene expression, and next-generation sequencing (NGS) technologies have resulted in more comprehensive characterization and accurate classification of human tumors and provided insights into cancer genome complexity and heterogeneity. This has led to the emerging field of cancer genomics to study human cancer genome. It is a systematic search within cancer families and patients for the full collection of genes and genetic or epigenetic alternations – both inherited and sporadic – that contribute to the development of a cancer cell and its progression from a localized cancer to one that grows uncontrolled and metastasizes.

A significant number of community-driven collaborative research projects or programs on cancer genomics using high-throughput genomic technologies have been launched to provide systematic, comprehensive genomic characterization and sequence analysis of multiple types of human cancers, both primary samples and cell lines, and to facilitate cancer discoveries among scientists. For

example, The Cancer Genome Atlas (TCGA) [64] and The International Cancer Genome Consortium (ICGC) [65] provide comprehensive characterization including gene expression, copy number variants, SNP of major types of human cancers; Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [66] focuses on pediatric cancer; The Cancer Cell Line Encyclopedia (CCLE) [67] is a collection of genomic data for a large panel of human cancer cell lines; The Connectivity Map (CMAP) [68] provides genomic profiles of perturbation experiments by thousands of small-molecules.

All these cancer genomic data provide us significant molecule insights into genes, proteins and pathways that are causally associated with tumorigenesis, progression, or drug-resistance. We can use this information to complement genome-wide RNAi screens to shortlist candidates coming from RNAi screening, and, more importantly, to identify novel oncogenes or tumor suppressor genes as therapeutic targets for cancer treatment.

The genomic data I focus on to develop computational algorithms or framework is gene expression profiles or transcriptome data which is widely available with mature microarray or NGS technologies. However, the methods or computational framework I developed can be extended to other types of genomic data.

4.1.2 Gene expression signature is not robust

To identify causally-associated genes or pathways from large-sampled high-throughput cancer genomic data such as gene expression profiles, what people usually do is to identify the so called signature genes. For example, there are two

phenotypes, tumor or normal, drug-sensitive or drug-resistant, signature genes are usually defined as differentially expressed between the two phenotypes (Figure 4-1). However, the problem with conventional signature analysis is that they are not robust. Here is an example: Two groups were studying the same disease problem, metastasis in breast cancer, and each of them identified a group of signature genes based on their own large-sampled gene expression profiles, however, there is only one overlap out of about each 70 signature genes (Figure 4-2). One paper was published in Nature [109] while the other was in Lancet [110].

Another example is the following: we aim to identify a gene expression signature for ERBB2 mutation. To do that, we generate profiles for two cell lines – ERBB2 mutated one and the wild type. However, if we culture exactly the same cells in 2D or 3D environments, we would expect to see similar results because they represent the same cell types, however, we get very different results of signature genes. For example, there are about 30% of differentially expressed genes showing opposite directions in 2D and 3D systems (Figure 4-3) with individual examples as shown in Figure 4-4. Also if we looked at top over-expressed genes in 2D and 3D signatures as in the Venn diagram of Figure 4-3, there are only about 5 or 6 percent of top genes that are overlapped. We see the same pattern

As shown in the above examples, traditional methods of signature analysis at single factor level in such context might not work well because of the high dimension and noise from large-scaled genomic data and also intrinsic noise of

gene expression itself. However, it doesn't mean the expression data is useless. We just need to develop better method to dig into it. So we have to go beyond gene expression signature and develop better methods to identify more robust biomarkers or genes that are associated with cancer progression or drug resistance. And it's more interesting to identify "driver" genes of the phenotypes instead of "passenger" ones.

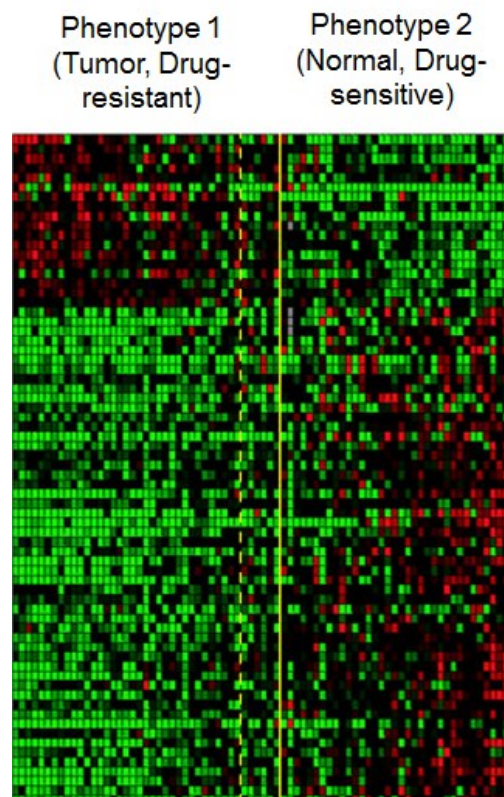


Figure 4-1 Heatmap of example for gene expression signature genes of two phenotypes.

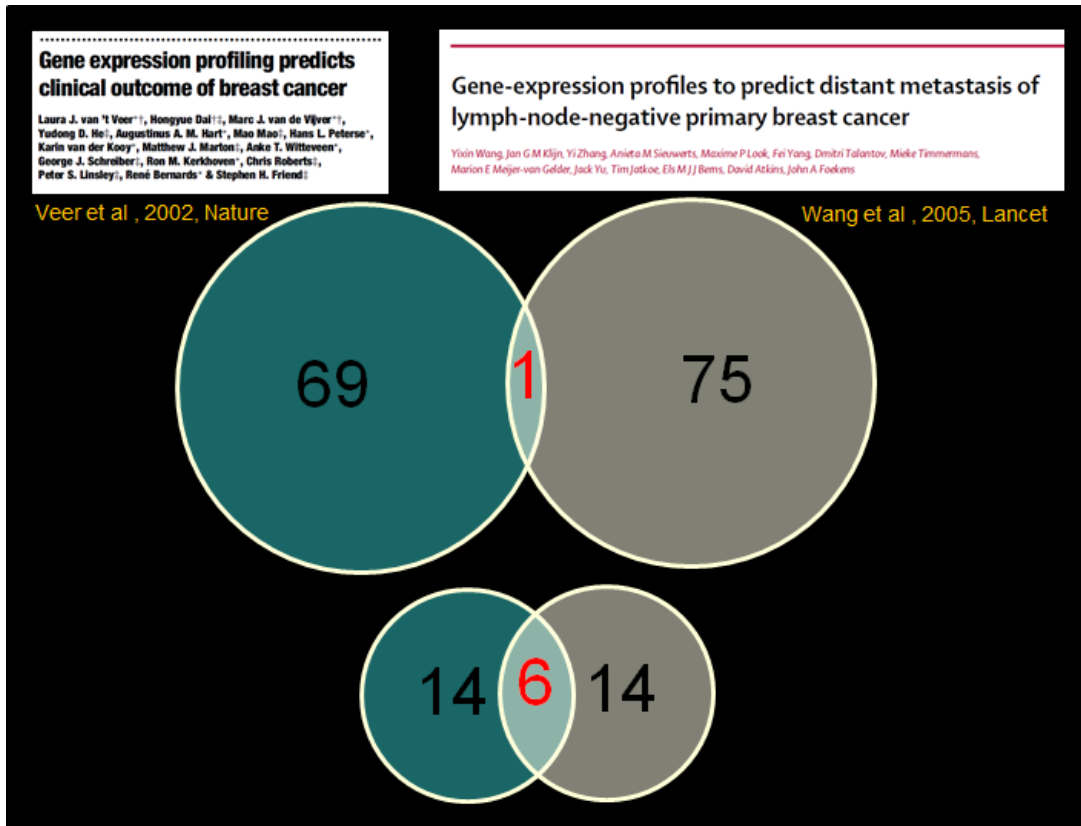


Figure 4-2 Example of that master regulators are more robust than gene expression signature genes. Two groups were studying the same disease problem, metastasis in breast cancer, and each of them identified a group of signature genes, but there is only one overlap. However, in our predicted drivers for each of the datasets, the overlap improves dramatically from 1 out of 70 to 6 out of 20.

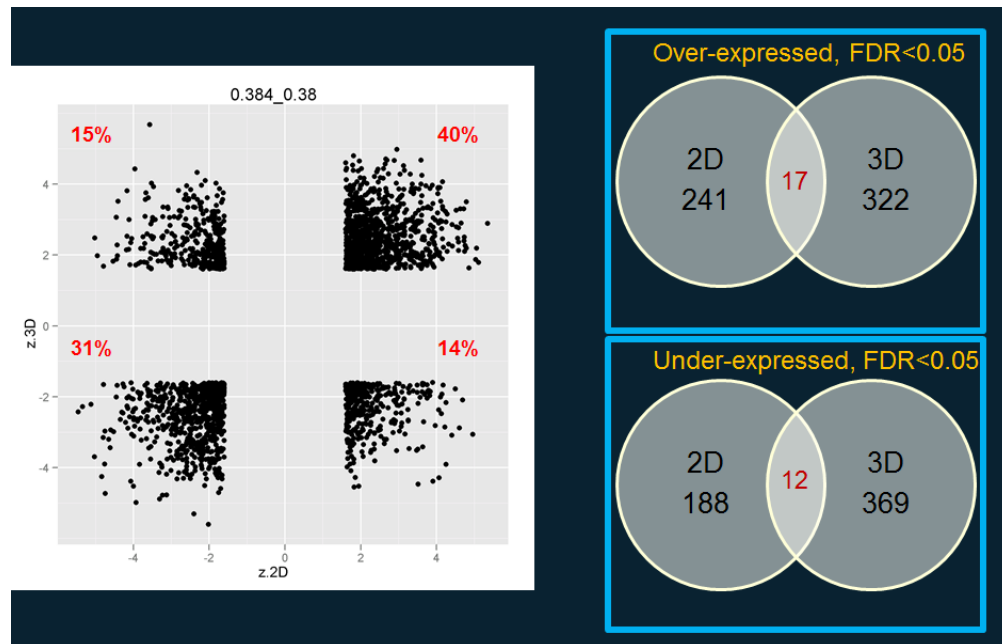


Figure 4-3 Example of that gene expression signature is not robust. Signature genes ($z > 1.96$ or $z < -1.96$) were plotted for both 2D (x axis) and 3D (y axis). Pearson and Spearman correlations are calculated on the top.

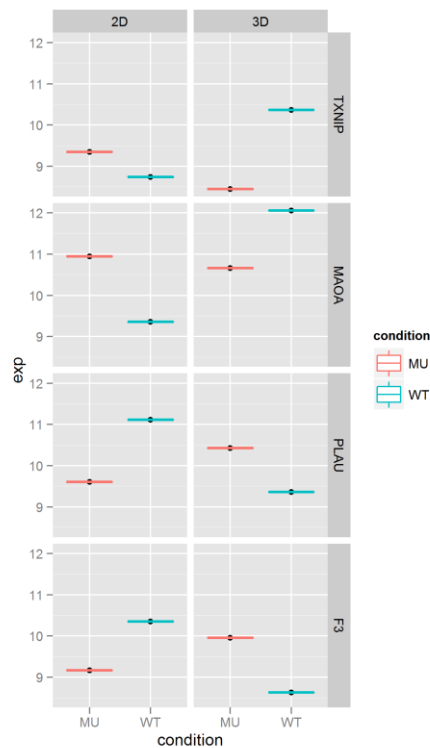


Figure 4-4 Individual examples of that gene expression signature is not robust. The first two genes are over-expressed in mutated cells with 2D culture, but are under-expressed in mutated cells with 3D culture. The last two genes are on opposite.

4.1.3 Systems biology

The problems with conventional gene expression signature have led to the development of systems biology approaches which integrate multiple pieces of information and utilize the strength of networks. Systems biology approaches have been successfully applied to high-dimensional data of cancer genomics to identify driver-type genes [70-73]. We have shown that computationally inferred context-specific maps of transcriptional or post-translational molecular interactions from large-scaled gene expression profiles (GEPs) allow the

elucidation of cryptic driver proteins whose gain or loss is necessary and sufficient for tumor initiation or progression [70-73]. Such master regulators or drivers are more robust than traditional signatures to distinguish phenotypes [69].

In this chapter, I will introduce a novel systems biology framework, Network-based Bayesian Inference of Disease Drivers (NetBID2), to infer disease drivers from high-throughput genomic data based on reverse-engineering network and Bayesian inference. It improves and extends existing MARINA algorithm [70-73]. I will demonstrate that this framework performs more robust than classical signature analysis, and is able to detect not only known drivers of various cancer contexts, but also hidden drivers that conventional methods fail to find. The prediction rate of this algorithm is also high based on experimental validation results.

4.2 Explanation of NetBID2 using a Social Example

Before going to details about NetBID2 algorithm, I would like to explain the idea using a metaphor or a social example (Figure 4-5) as below:

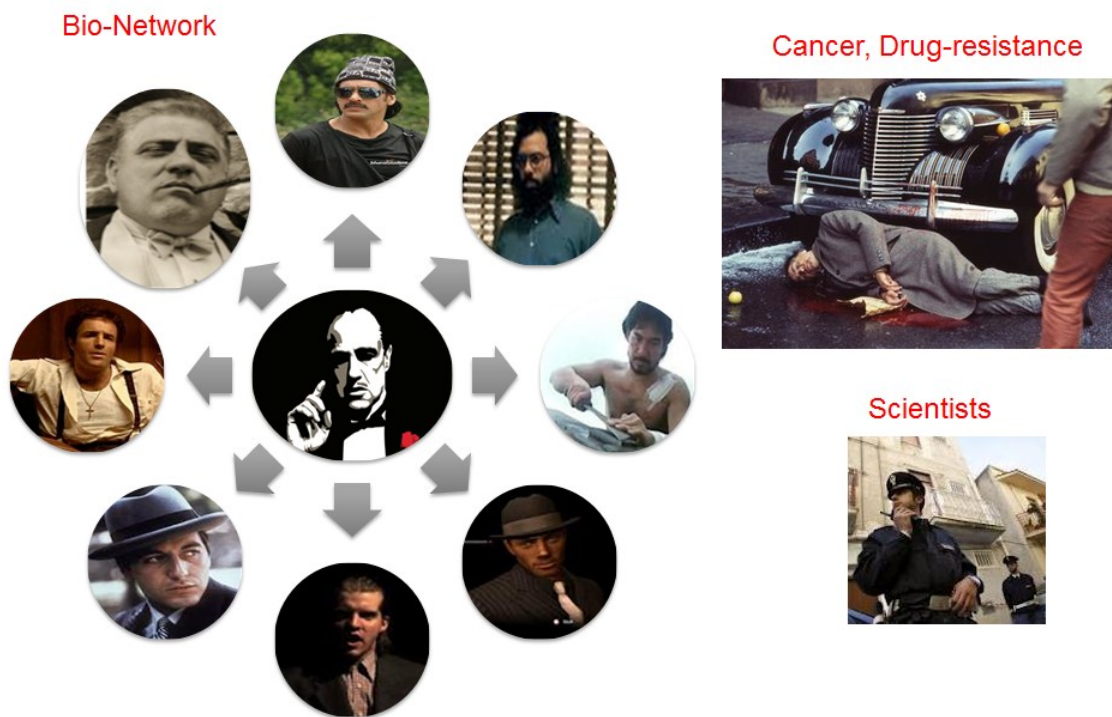


Figure 4-5 A social or non-scientific example to explain NetBID2 algorithm.

Imaging there is a gang group in the New York City (NYC) which has a relationship network structure like this. All gang members listen to the god father or Mafia. This gang group is responsible for a number of criminals such as a killing like this. Well, the New York Police Department (NYPD) wants to capture those bad guys who committed the crime. What the police would do is to investigate the crime scene and collect evidences. Actually this is very similar to what we scientists do to study cancer or drug resistance. The police profile people to identify suspects, and we do gene expression profiling to identify genes that are associated with diseases. However, if the police only look at evidences directly associated with the crime, most-likely they will identify some small guys in this gang group, however, they will never get to this Mafia, who actually drives

all these crimes because there is no evidence directly pointing to him. He is the guy who NYPD really wants to put in jail. To capture the big fish, the NYPD has to collect extra information and build a relationship network like this. Then they can borrow the strength of this network and apply the rule of “guilty-by-association” to get to him. These small players are like signature genes which are not robust because in this crime, the police might get this one, but in another crime, a different guy will be caught. What NetBID2 does is similar to this: it constructs a network first and instead of identifying those small signature genes, it looks for genes that are highly associated with so-called signature genes which are potential to be drivers.

Actually those hidden driver genes, especially signaling proteins, are promising therapeutic targets. There is an old saying in Chinese, “destroy the leader and the gang will collapse”, and the crime will stop, at least committed by this gang group. Similarly to crack cancer or drug-resistance, we would like to target those hidden drivers. And most likely, the deeper this big fish hide, the more promising it is to be a therapeutic target.

To summarize the example, the key idea of NetBID2 is to utilize the strength of networks to search for candidates that are highly connected with unstable signature genes instead of looking at individual signature genes. So we do use signature analysis results, but we go beyond them.

4.3 The NetBID2 Framework

In NetBID2 framework (Figure 4-6), there are basically three steps to infer drivers of phenotypes from gene expression data: (1) reverse-engineering network (2) signature analysis of phenotypes and (3) gene set enrichment analysis of Subnetwork of driver candidates in phenotype signature. Details about each step are discussed as below.

4.3.1 Reverse-engineering regulatory or signaling networks (Step 1)

The first step of NetBID2 is to reconstruct a biological network from gene expression data. It has been shown that context-specific maps of transcriptional or post-translational molecular interactions can be computationally inferred from large-scaled gene expression profiles (GEPs) and allow the elucidation of cryptic driver proteins whose gain or loss is necessary and sufficient for tumor initiation or progression [70-73]. In this step, we use a well-developed reverse-engineering algorithm, ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) [111]. ARACNe method uses mutual information to measure and construct pairwise connection between two genes but applies an information theory of data processing inequality (DPI) to eliminate the vast majority of indirect interactions typically inferred by pairwise analysis. On synthetic datasets ARACNE achieves extremely low error rates and significantly outperforms established methods, such as Relevance Networks and Bayesian Networks [111]. Application to the deconvolution of genetic networks in human B cells

demonstrates ARACNE's ability to infer validated transcriptional targets of the c-MYC proto-oncogene [70].

The original version of ARACNe only focused on tackling transcriptional regulatory network or transcription factor (TF)-centered network. However, in NetBID2, we extend ARACNe to reconstruct signaling molecule centered network by treating signaling protein genes as TFs in the original ARACNe setup. Actually it's more appropriate for ARACNe to reverse-engineer signaling network than transcriptional network for the following reasons.

First, the novel part of ARACNe is to use DPI to eliminate indirect directions. DPI says if three genes follow a Markov-chain network structure, i.e. $A \rightarrow B \rightarrow C$ or $C \rightarrow B \rightarrow A$, then $I(A; C) \leq I(A; B)$ or $I(B; C)$. Based on the theory, ARACNe removes edges between A and C if $I(A; C) \leq I(A; B)$ or $I(B; C)$. However, the underlying assumption is that A, B, C follows the Markov-chain structure, which is actually rare in transcriptional regulatory networks because of large amounts of feed-back and feed forward loops [112] for their importance in regulation of biological processes. Without holding the assumption, ARACNe will remove a significant number of true edges between TFs. There has been an increasing awareness of this problem [113]. However, in signaling transduction networks, Markov chain structures are everywhere whereas feed-forward loops are rare because the time scale of signaling transduction reactions is too small to have feed forward or feed-back loops. So DPI is more appropriate to remove redundant interactions between signaling factors than between transcription factors.

Second, using ARACNe to reconstruct both transcription regulatory networks and signaling networks share a common assumption: the activity of proteins such as transcription factor or signaling molecules can be inferred from their gene expression level information, which is reasonable in many cases. The more mRNAs the cell produces for a gene, the more protein copies it generates. However, there is low or no correlation between gene expressions at mRNA level with protein level due to the dynamic nature of transcription and translation processes and the internal or external noise. However, this is a common limitation using gene expression data to reconstruct transcriptional network or signaling network, which can only be overcome by improving technology and using protein level data.

In the discussion section, I will demonstrate that the targets or regulons inferred by ARACNe-predicted signaling network have a consistently higher precision than TF-centered network. Because the regulon size in TF-centered network is much bigger than signaling molecule centered network, the inferred interacting partners in signaling network are much cleaner and more informative than those in ARACNe inferred transcriptional network.

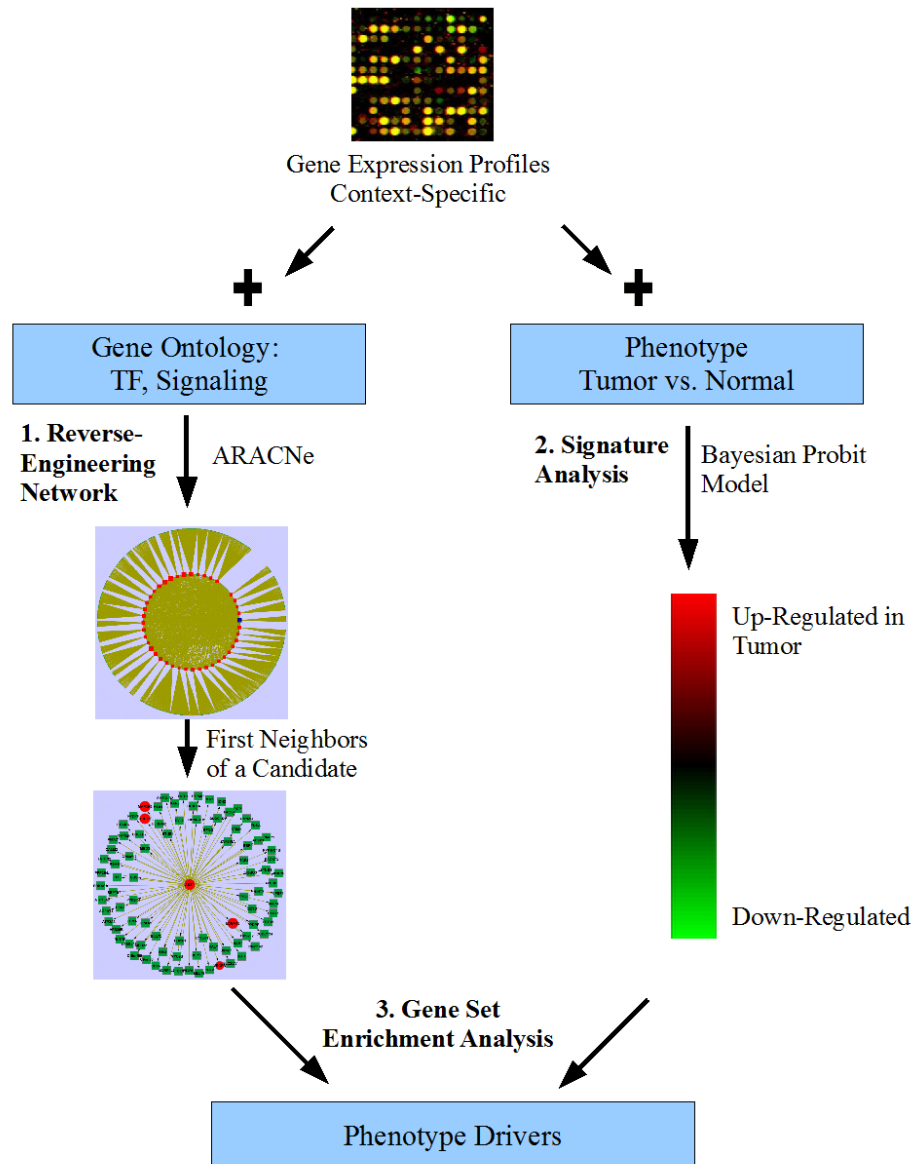


Figure 4-6 The NetBID2 framework. Step 1 uses reverse-engineering algorithm, ARACNe to reconstruct TF or Signaling centered networks from gene expression profiles. Step 2 utilizes phenotype information to perform signature analysis using a Bayesian Probit model approach. Step 3 applies gene set enrichment analysis for each driver candidate by taking its first neighbors as a gene set and using signature analysis results of all genes as the reference.

4.3.2 Signature analysis of diseases by Bayesian Probit model (Step 2)

The second step of NetBID2 is signature analysis of phenotypes. Signature analysis is to measure a correlation between each individual gene with the phenotypes such as tumor or drug resistance, and it can be performed by differential gene expression analysis between the phenotype of interest with the control phenotype, such as tumor vs. normal, drug resistant vs. sensitive samples, etc. In NetBID2, to generate a robust reference signature of these two phenotypes, we used a Bayesian Probit regression model for each individual gene [89] (Figure 4-7). Probit model has the advantage of detecting weak signals and has a nice tail behavior comparing with linear model or logistic model [89]. Bayesian inference with Markov Chain Monte Carlo (MCMC) computing techniques help to overcome sample size problem and help to estimate parameters more accurately and robustly from noisy high-throughput microarray data. For Bayesian inference of parameters, a t-distribution prior or weakly-informative prior is used due to its robustness and its ability to handle outliers [90]. Gibbs sampling of MCMC technique is used to simulate the posterior distribution of parameters and posterior mean or median can be used as estimate of parameters, especially beta, the slope of the model, which measures the correlation between its expression activities with phenotypes.

Probit Model

$$y_i \sim \text{Bernoulli}(\theta), \quad \theta = \Phi(z_i)$$

$$\text{OR } y_i = \begin{cases} 1 & \text{if } z_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \beta x + \alpha + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ i = 1, 2, \dots, n$$

Priors

$$\beta \sim (\mu_\beta, \sigma_\beta^2), \quad \sigma_\beta^2 \sim \text{Inv-}\chi^2(\nu_\beta, s_\beta^2)$$

$$\alpha \sim (\mu_\alpha, \sigma_\alpha^2), \quad \sigma_\alpha^2 \sim \text{Inv-}\chi^2(\nu_\alpha, s_\alpha^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$$

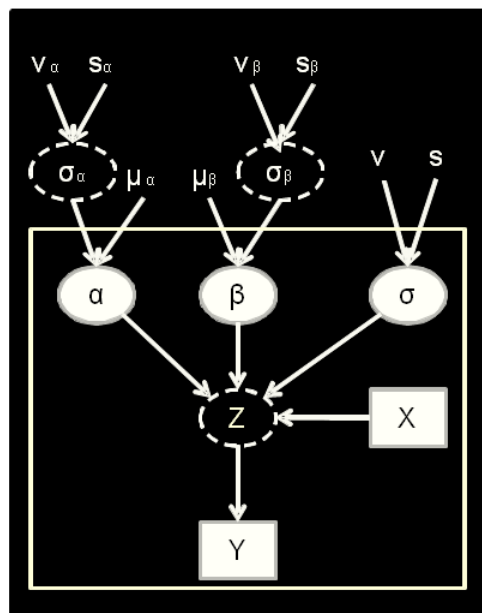


Figure 4-7 The Bayesian Probit Model for gene expression signature analysis. Distribution details about the model is on the left and on the right is a graphical representation of the Probit model. Nodes in solid square are observation variables, in solid eclipse with white background are direct parameters of Probit model, in dashed eclipse are latent variables and the others are hyper-parameters for priors. Y is an indicator variable for phenotypes, X is expression level of gene X, Z is a latent variable in Probit model. Inside the white box is likelihood section, while outside is for priors. Parameters are estimated by a Gibbs sampling procedure.

4.3.3 Gene set enrichment analysis to infer disease drivers (Step 3)

The final step of NetBID2 is to perform gene set enrichment analysis for each driver candidate and estimate their potency to be a good driver of the disease of interest. For driver candidates, NetBID2 only considers transcription factors or signaling molecules which have the function to be drivers. In the first step, transcription factor or signaling factor centered networks have been generated by

ARACNe. In the diver inference step, for each driver candidate, we take all its first neighbors from its subnetwork as a gene set and test whether this set is enriched in the signature of disease produced in the second step. If there is a significant enrichment pattern, we consider and report the driver of testing as a good driver of this disease by controlling top signature genes.

There are tons of methods to do enrichment analysis such as classical GSEA (Gene Set Enrichment Analysis) [97] developed at Broad Institute, GSA (Gene Set Analysis) [98] by Brad Efron and so on. However, I develop a new set enrichment analysis algorithm, BSEA (Bayesian Set Enrichment Analysis), using Efron's "maxmean" statistic for enrichment score and Bayesian inference, which outperforms both GSEA and GSA. See details in Chapter 5.

Enrichment analysis will report an enrichment score (ES) or normalized enrichment score (nES) with corresponding value to indicate the evidence of the candidate being a driver of the disease. Positive ES indicates the genes associate with this driver are enriched in up-regulated or over-expressed genes in the disease phenotype, vice versa, negative ES means those regulon genes of the driver are over-represented in down-regulated or under-expressed genes. However, the sign of enrichment score doesn't necessarily positively correlate with the activity of the driver gene. For example, a positive ES for a significant driver candidate doesn't necessarily mean this driver is active in the disease phenotype side. It's highly possible that the driver is repressed in the disease samples, and those associated genes are its repressed targets therefore causing

these neighbors enriched in the positive direction. What we can roughly conclude with the enrichment output is that whether this driver candidate has a strong association with the disease or not, but we cannot tell accurately whether the driver is active or inactive to cause the disease phenotype. The reason is because we are using gene expression data to infer the activity of genes at protein level. In many cases, there is a clear negative correlation of mRNA expression with protein expression due to feed-back loops and dynamics of gene transcription and translation. To tell the direction correctly, we need more information at lower genetic DNA level or upper protein level, or experimental validations of perturbation by either silencing or over-expressing driver candidates.

4.4 Evaluation of NetBID2

We evaluate the performance of this new framework, NetBID2 to discover disease drivers from gene expression data in several perspectives as shown below.

4.4.1 NetBID2 is more robust than expression signature

First, we compare NetBID2 method with conventional signature analysis to characterize the relationship between genes and the disease of study. In the example of identifying genes associated with ERBB2 mutation from expression profiles of 2D and 3D cultured cells as we discussed in introduction section (4.1.2), the signature analysis showed dramatically different results, especially the overlap of top signature genes is small, only about 5 or 6% of selected

differentially-expressed genes. So we applied NetBID2 to this example by identifying drivers that are associated with ERBB2 mutation using 2D or 3D data. Surprisingly, there is a dramatic increase of overlaps between drivers inferred from 2D expression data with 3D data (Figure 4-8), about 1/3 out of top inferred master regulators or signaling modulators. Therefore, the dramatic difference between 2D and 3D signatures cannot be explained only by the difference of environments. Actually the methods count a lot. NetBID2 is able to generate a more robust list of genes that are potential drivers of ERBB2 mutation no matter what culturing environment is used.

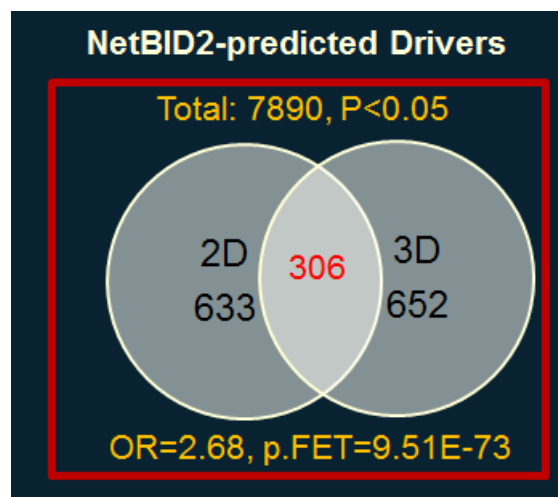


Figure 4-8 Venn diagram of NetBID2 inferred drivers (both TF and Signaling factors) from 2D or 3D expression data of ERBB2 mutated MCF10A cells. Fisher's exact test is used to test the significance of overlaps. Total number is the number of probes for TF or signaling factors in the microarray data.

4.4.2 Ability to identify known drivers

Another way to test how NetBID2 performs is to ask whether it can detect known drivers. And the answer is yes. We applied NetBID2 to predict TF or signaling drivers for Basal vs. Luminal breast cancer, HE2+ vs. HER2– breast cancer, and ABC vs. GCB-DLBCL (Diffuse large B cell lymphoma). As shown in Table 4-1, among top 30 NetBID2-predicted drivers, most known drivers from literature are identified (highlighted in red). For example, FOXA1, GATA3, ESR1, PGR are well-studied drivers for Luminal breast cancer. ER and PR are also signaling molecules, and they show up in the top list of signaling drivers. For HER2+ breast cancer, ERBB2 itself definitely should be on the top without surprise. And FGFR4, GRB7 are also commonly amplified or over-expressed with ERBB2. ZBTB4 is a newly-identified tumor-suppressor in aggressive breast cancer. It shows up in both Basal vs. Luminal, and HER2+ breast cancer. Also for DLBCL, BCL6, IRF4 are known master regulators of ABC or GCB subtype. So NetBID2 is able to detect known drivers of diseases. However, it's not surprising because the evidences for all those drivers are so strong that you don't need to do complicated analysis to identify them. Most of them are also significantly differentially-expressed in the phenotypes which can be identified by signature analysis. There are duplicated names in the list because the analysis was done at probe or transcript level. One gene could have multiple probes in expression data representing different transcripts, but more probes for the same gene showing up gives more prediction power for that gene being a driver.

Basal vs. Luminal		HER2+ vs. -		ABC vs. GCB-DLBCL	
TF	Signaling	TF	Signaling	TF	Signaling
FOXA1	ESR1	PITX1	HDAC5	FOXP1	ARHGAP25
FOXA1	ESR1	MNX1	ENTPD7	IRF4	STK24
GATA3	ESR1	POU6F1	ERBB2	FOXP1	PRKD3
ZBTB4	ESR1	CITED1	DNASE1	FOXP1	NEIL1
ESR1	ESR1	RFX3	FGFR4	FOXP1	LIMD1
ESR1	ESR1	BCL6	FGFR4	BCL6	PRKD3
ZNF540	PSAT1	ZNF557	FGFR4	BCL6	GNA13
ESR1	THSD4	DMRTC2	ENTPD7	TOX	STK39
ESR1	THSD4	KLF7	DNASE1	BCL6	PTK2
ESR1	NUDT12	SOX5	NAGS	MYBL1	DCK
ESR1	PGR	NFAT5	FGFR4	BATF	ENTPD1
ZBTB4	PGR	NFAT5	ERBB2	BPNT1	ENTPD1
E2F3	PGR	THRA	NAGS	MAML3	P2RY8
GATA3	PGR	NR3C1	IRS2	STAT3	DNASE1
ZBTB4	NAT1	NR3C1	GNPNAT1	ZNF318	ENTPD1
ZBTB4	NAT1	NR3C1	GPRC5C	IRF8	MME
E2F3	NAT1	NR3C1	ERBB2	RARA	MAP4K4
PGR	NAT1	ONECUT2	IRS2	ZNF608	HDAC1
PGR	PSAT1	NKX2-2	GNPNAT1	NFIA	RARA
PGR	NAT1	SOX5	IRS2	SMAD5	P2RY10
PGR	NOSTRIN	ZBTB4	ERBB2	CUX1	MME
ZNF396	ART3	BCL6	FGF2	BPNT1	PTK2
AFF3	IL6ST	ZBTB16	IRS2	TCF4	LIMD1
AFF3	UGCG	ZNF296	GNPNAT1	ZFP106	PPP1R16B
ZFP2	NOSTRIN	CITED1	GRB7	MEIS2	PAG1
SALL2	ZFYVE16	CITED1	GGT1	TOX	EIF2AK3
GLI3	THSD4	ZNF323	GRB7	LASS4	MAP4K4
GLI3	FBP1	EGR3	CHPT1	TCF4	PDE4B
MSL3	CREBL2	ZDHHC2	CHPT1	SPIB	S1PR2
MSL3	ARSG	PITX1	NEK6	NFIA	GSTA4

Table 4-1 Top 30 NetBID2-predicted TF or Signaling drivers for Basal vs. Luminal breast cancer, HER2+ vs. HER2- breast cancer, and ABC vs. GCB-type of DLBCL. Genes in red are known drivers of corresponding diseases reported in literature. Duplicate gene names are for different probes or transcripts.

4.4.3 Ability to identify “hidden” drivers

We showed that NetBID2 is able to find known drivers of diseases; however, the true power of this network-based framework is to detect “hidden” drivers that classical methods such as signature analysis fail to find. For example, we identified and validated AKT1 as a good driver for glucocorticoid resistance in T-ALL, but if we look at the expression of AKT1, there is no significant change or it

shows some anti-evidence to be active in resistant samples. So based on traditional differential expression analysis, AKT1 will never be identified. See details about this example in Chapter 7.

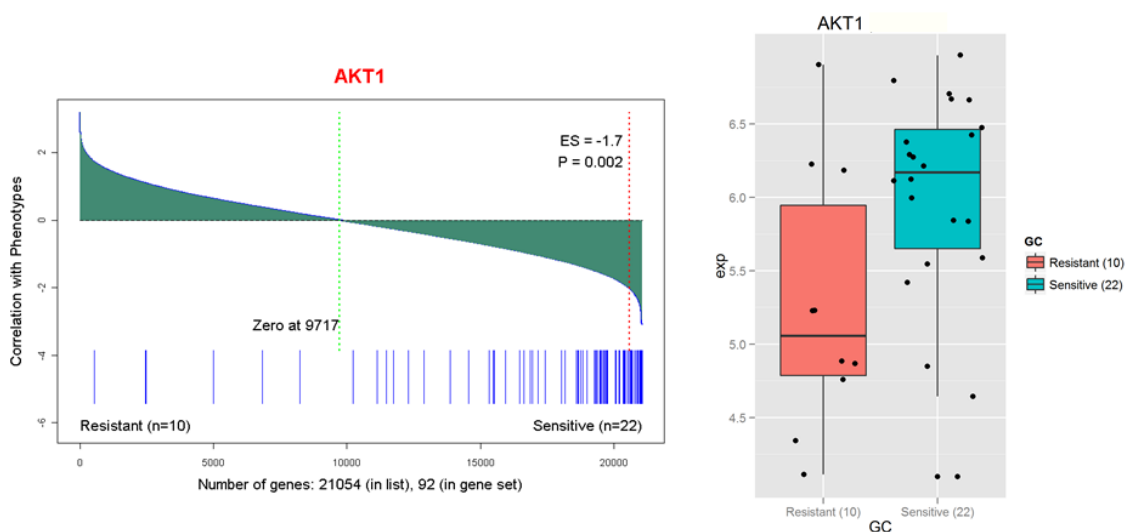


Figure 4-9 NetBID2 identifies AKT1 as a driver for glucocorticoid resistance in T-ALL (left), but there is no evidence from expression of AKT1 itself.

Another example is STAT3 for HER2+ breast cancer. By NetBID2, we identified STAT3 as a significant master regulator and signaling modulator of ERBB2 induced breast cancer, however, there is no expression change of STAT3 itself between ERBB2+ and control cells. Again, conventional signature analysis will lose STAT3. See details about this example in Chapter 9.

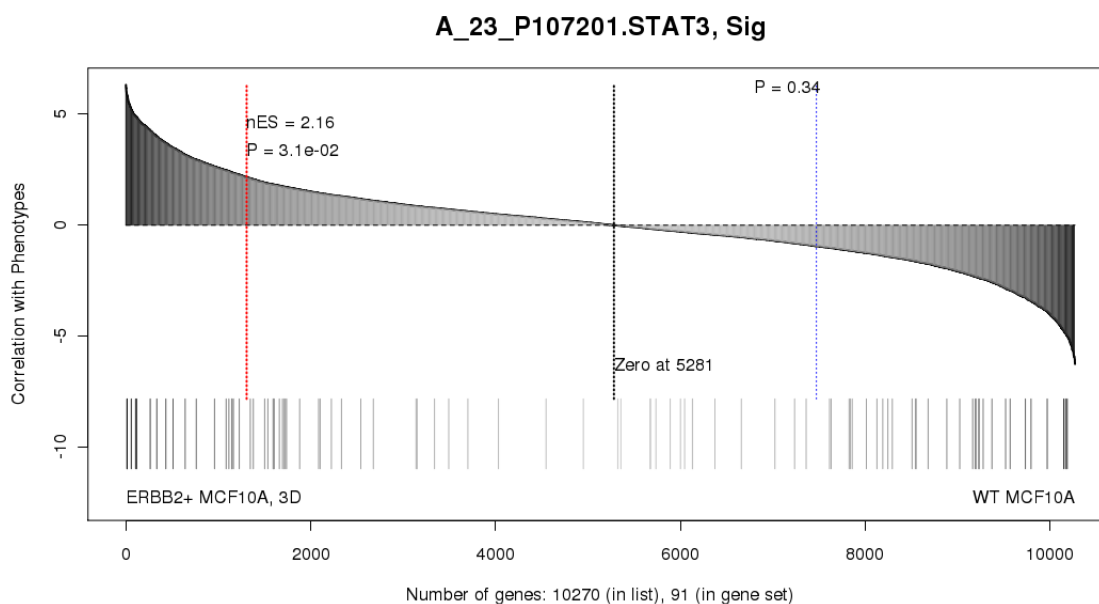


Figure 4-10 NetBID2 identifies STAT3 as a driver for ERBB2+ breast cancer (red line), but there is no evidence from expression of STAT3 itself (blue line).

4.4.4 High prediction rate by experimental validations

The most straightforward way to evaluate the prediction of NetBID2 is to do experiments. Basically we perturb the predicted driver by silencing or over-expressing it and check whether the perturbation can change the phenotype. We validated top 30 predicted TF drivers or master regulators for glucocorticoid resistance in T-ALL by siRNA. Surprisingly, only 7 (two have p value like 0.05) or 5 predicted drivers showed no effects on changing resistance upon silencing (Figure 4-11). Over 76% or 80% of predicted drivers have significant effects on either reversing resistance or increasing resistance upon silencing. The validated drivers with positive scores are potential targets to reverse resistance by silencing. Those negative ones function as suppressors of resistance, which

need to be overexpressed or activated to reverse the resistance. In summary, the prediction rate of NetBID2 for disease drivers is strikingly high.

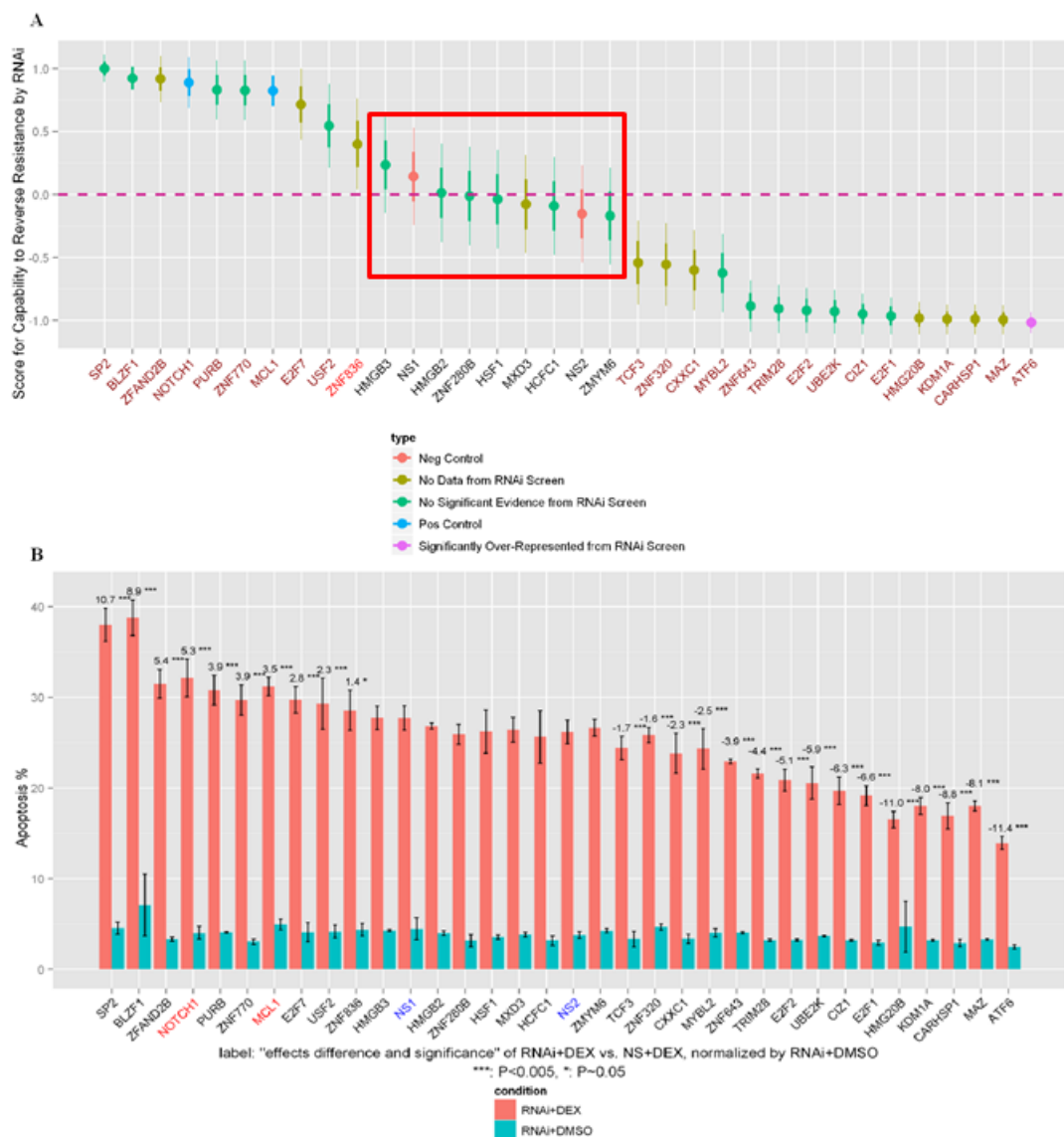


Figure 4-11 Validation results by siRNA for top 30 NetBID2-predicted TF drivers of glucocorticoid resistance in T-ALL. (A) top 30 candidates (in red) together with positive controls (in blue) and negative controls (in green) are ranked by the score (central dot) for capability to reverse GC-resistance upon silencing with uncertainty (range line crossing the central dot, thick line for one standard deviation, thin line for two standard deviations corresponding to 95% confidence

interval). The color of candidate label on x axis is associated with calibrated p-value: dark red for $P < 0.005$, red for $P \approx 0.05$. (B) Bar plots of apoptosis level induced by combined treatment of RNAi with DEX (in red), and control, RNAi with DMSO (in light blue) for 30 predicted candidates, positive controls (labeled in red) and negative controls (labeled in blue). All genes are ranked the same as in panel A. The label on top of bar plot represents the increased apoptosis level of candidate gene comparing with average of negative controls (normalized by its own DMSO control and averaged over triplicates) and associated statistical significance level (***) for $P < 0.005$, * for $P \approx 0.05$).

4.5 Evaluation of ARACNe Predictability

4.5.1 Using STAT3 as an example to evaluate ARACNe predictability

In NetBID2, we use ARACNe, an information theory-based algorithm to reconstruct transcription factor or signaling factor-centered networks. The key idea is to apply DPI to eliminate interactions between a TF and an indirect target or between a signaling molecule gene and its indirect downstream or upstream factor. We consider neighbors of a TF in the inferred regulatory network or a signaling protein in predicted signaling network as its regulons or targets or interacting partners. So how does ARACNe to predict targets of a TF or interacting proteins of a signaling factor? One way to answer this question is to do experiments to define a set of true targets and then use it as a gold standard to check the predictability of ARACNe.

We used the example of STAT3 which is both a TF and a signaling protein to check the prediction of ARACNe. We did microarray expression profiling after

silencing or activating STAT3 biochemically and defined the most changed genes after perturbation as downstream targets or effectors of STAT3. Then we used the perturbation results as a gold standard to evaluate ARACNe-predicted targets or interacting partners from either regulatory network (STAT3 as a TF) or signaling network (STAT3 as a signaling protein).

4.5.2 Overall prediction of ARACNe is good

First, we noticed that ARACNe-predicted targets, as both TF and signaling factor, are significantly enriched in experimentally-generated gold standard. Six ARACNe-inferred target sets (three probes for STAT3, two types of network) showed strong enrichment in activated target side, in consistent with STAT3 being an activator in general.

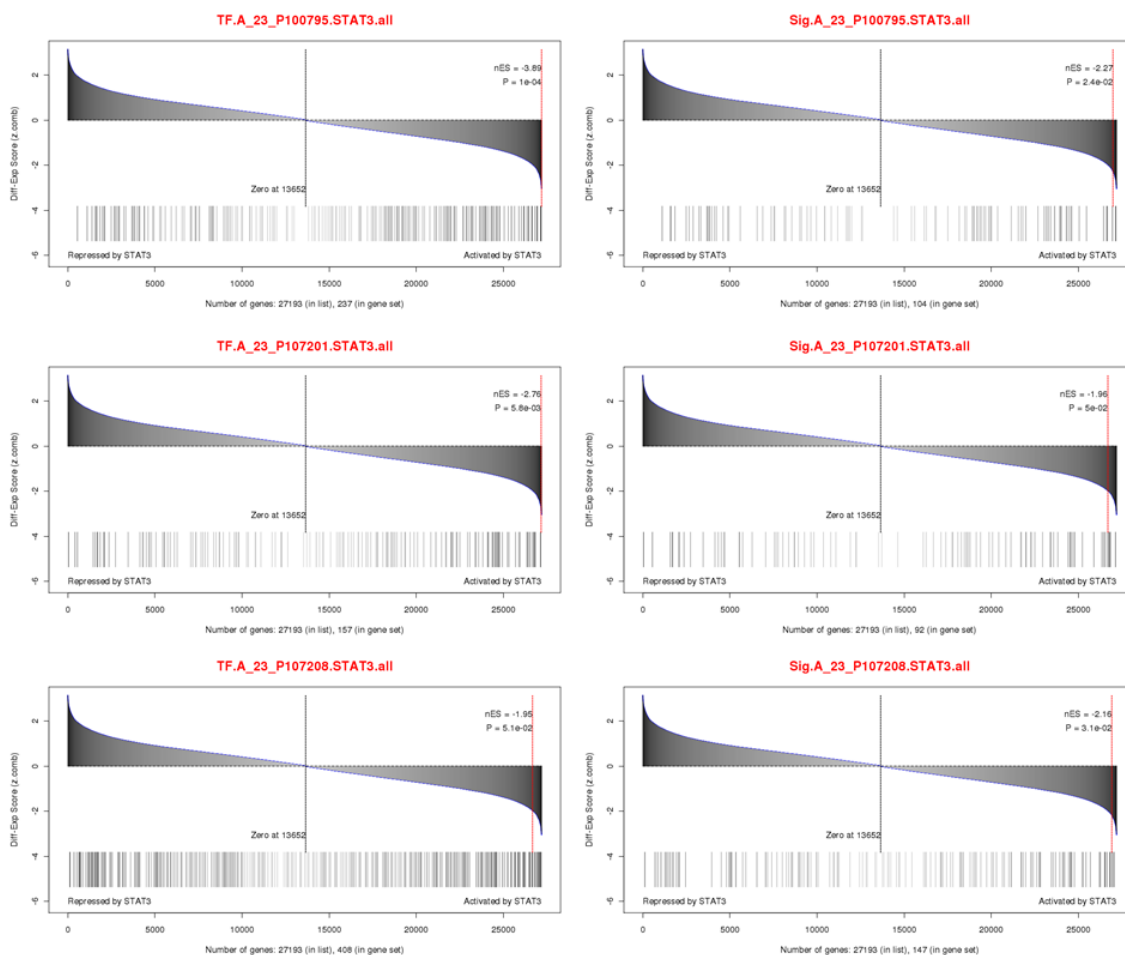


Figure 4-12 Enrichment of ARACNE-predicted targets of STAT3 in experimentally-identified targets. TF is for predicted targets in transcription regulatory network. Sig is predicted targets from signaling network. Three sets of inferred targets are for different transcripts or probes of STAT3.

4.5.3 The direction of interaction defined by correlation might be misleading

One of the key steps in MARINA [69, 71, 72] algorithm is to define positive and negative regulons of each TF after obtaining ARACNe-generated network because ARACNe uses mutual information to measure the dependence of two

distributions which is nonnegative. If the inferred a regulon or target has a significant positive correlation with the TF of study, it's defined in MARINa as an activated target. Similarly, negative correlation is to define repressed targets. However, this method might be misleading because correlation only captures linear relationship between a TF and its target which might not be true in many cases due to the dynamics of feedback or feed-forward loops.

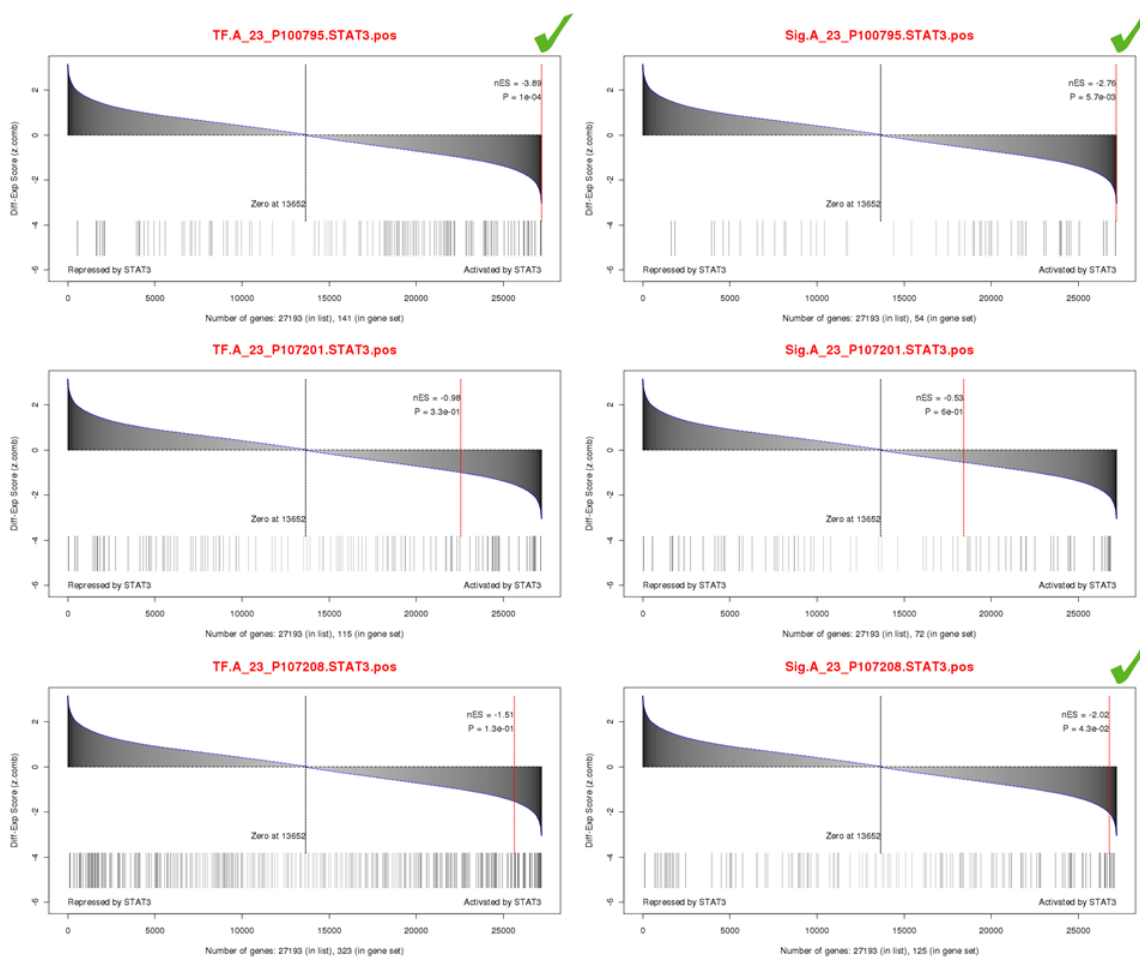


Figure 4-13 Enrichment of ARACNE-predicted positive targets of STAT3 in experimentally-identified targets. Positive is defined by the positive correlation between the target and STAT3 expression. TF is for predicted targets in transcription regulatory network. Sig is predicted targets from signaling network.

Three sets of inferred targets are for different transcripts or probes of STAT3.

Green check sign indicates $P < 0.05$.

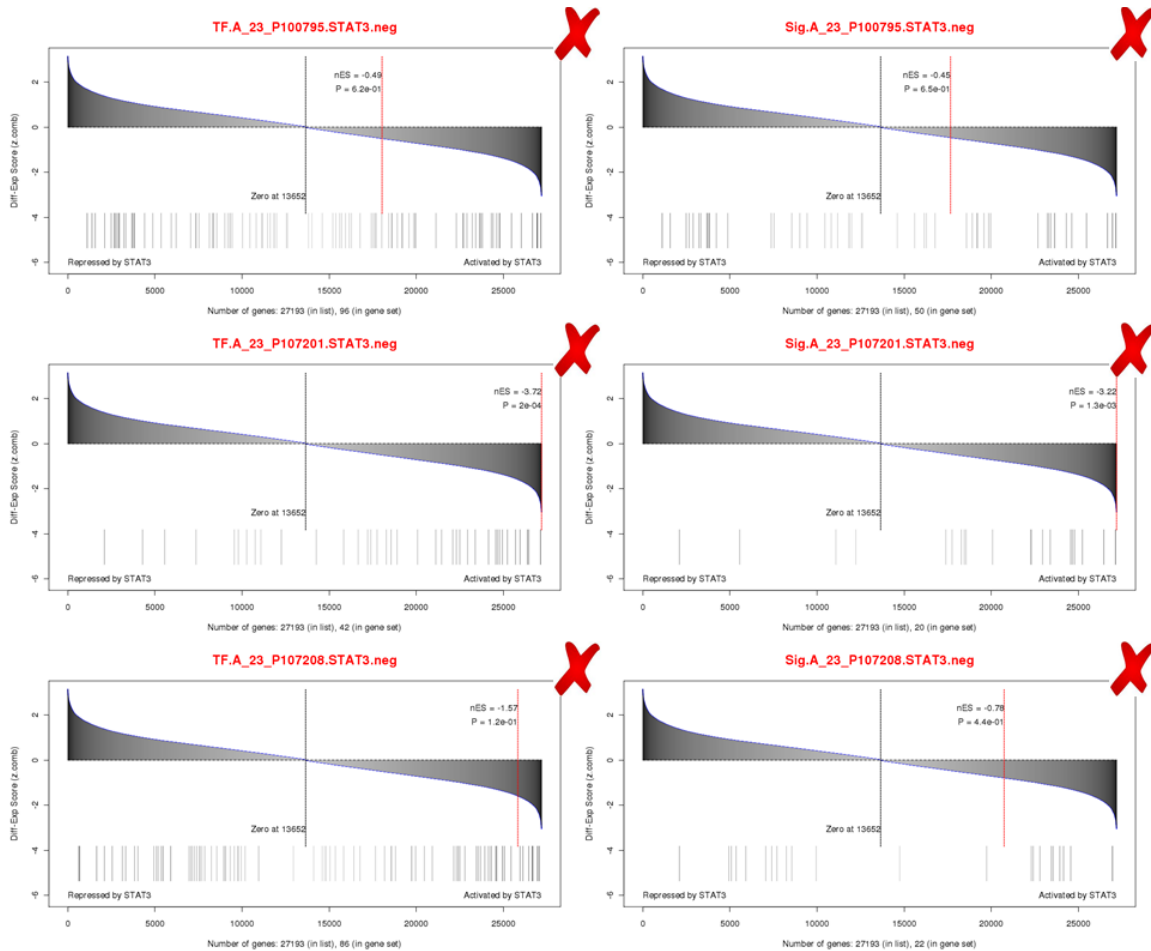


Figure 4-14 Enrichment of ARACNE-predicted negative targets of STAT3 in experimentally-identified targets. Negative is defined by the negative correlation

between the target and STAT3 expression. TF is for predicted targets in transcription regulatory network. Sig is predicted targets from signaling network.

Three sets of inferred targets are for different transcripts or probes of STAT3.

The check sign indicates non-significant.

Therefore, we use the example of STAT3 to check how good the direction defined by correlation. As shown in Figure 4-13, three out of six positive target sets predicted by ARACNe with correlation-defined signs are significantly ($P < 0.05$) enriched in experimentally identified activated targets of STAT3, out of which two are from signaling network and one is from TF network. However, as shown in Figure 4-14, all six negative target lists predicted by ARACNe with correlation post-analysis showed the wrong direction with gold standard negative targets of STAT3, or in another words, those negative targets defined by correlation are actually positive targets. This suggests that correlation-defined positive or negative targets on ARACNe-outputted network might be misleading and it's not a good idea to use mutual information to capture nonlinear relationships but then only considering linearly-related pairs.

In NetBID2, there is no classification of positive or negative sets defined by linear correlation. That's why we identified STAT3 as a driver, however, if we separated activated or repressed targets for STAT3, we would have definitely missed it because all predicted negative targets are actually positive ones and dilute the signal of STAT3 being an activator.

4.5.4 Signaling network prediction is more precise than TF network

One novel part of NetBID2 is that it extends the network from transcriptional regulatory network to signaling network by applying ARACNe against signaling proteins. With the gold standard targets from perturbation experiments for STAT3,

which is both a TF and a signaling factor, we can compare and evaluate the quality of TF network and signaling network generated by ARACNe.

First, the number of targets in TF-subnetwork of STAT3 is about twice as large as the number in signaling networks, though the number varies among the three probes of STAT3. So in signaling network, ARACNe removes more interactions than TF network.

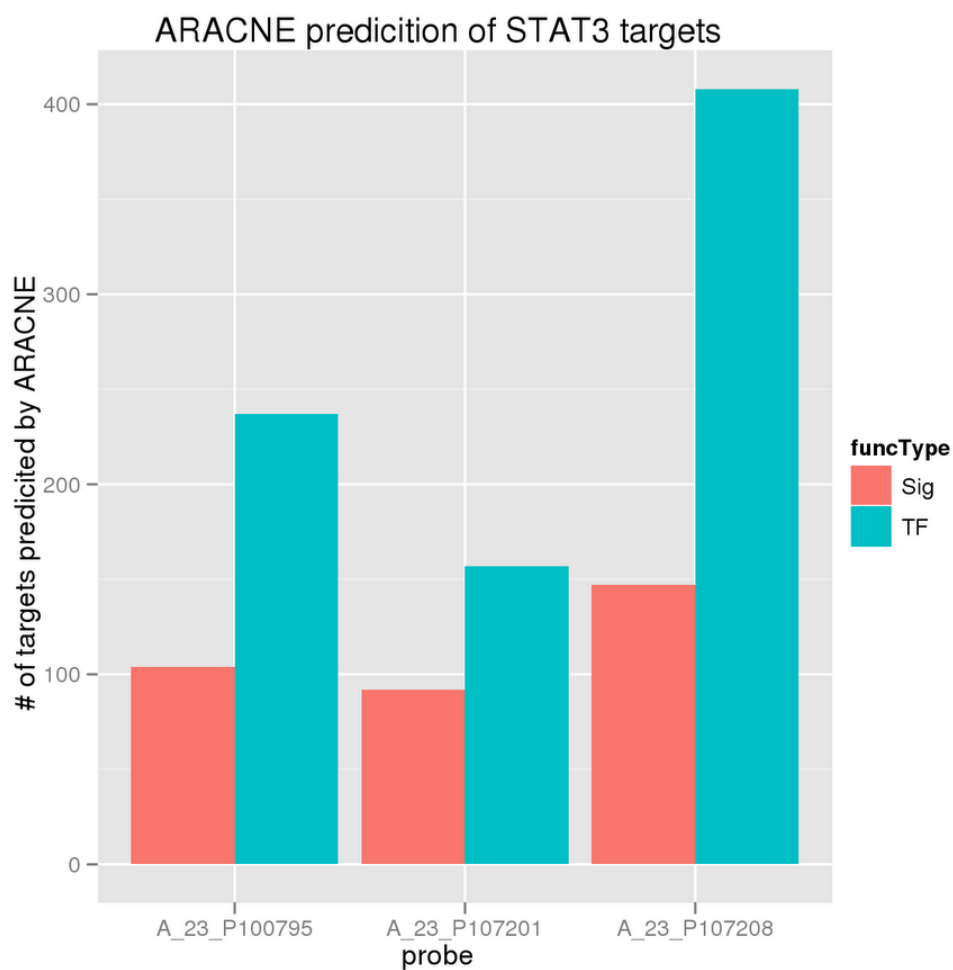


Figure 4-15 Number of target size for STAT3 (three probes at x axis) from TF (blue) or signaling (red)-centered network predicted by ARACNe.

Second, we checked the overlap of predicted targets from TF network or signaling network with top “gold standard” targets from experiments. Interestingly, the targets or interacting proteins in signaling network consistently demonstrated larger overlaps with gold standards than TF network. This suggests that signaling network is more precise than TF network. Since signaling network has a much smaller size than TF network, but identifies more true targets, it infers that signaling network has more power of identifying true positive interactions and removing true negative ones.

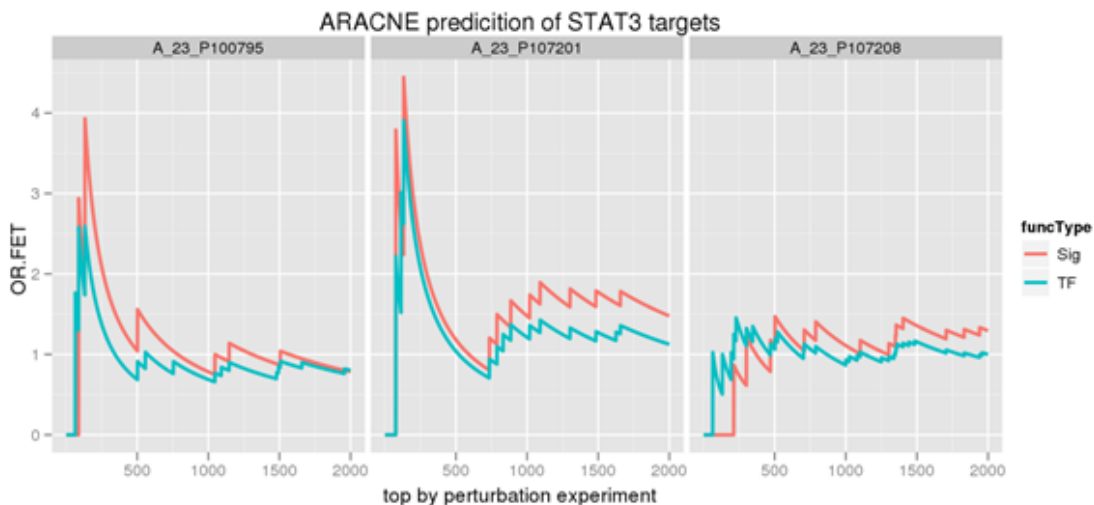
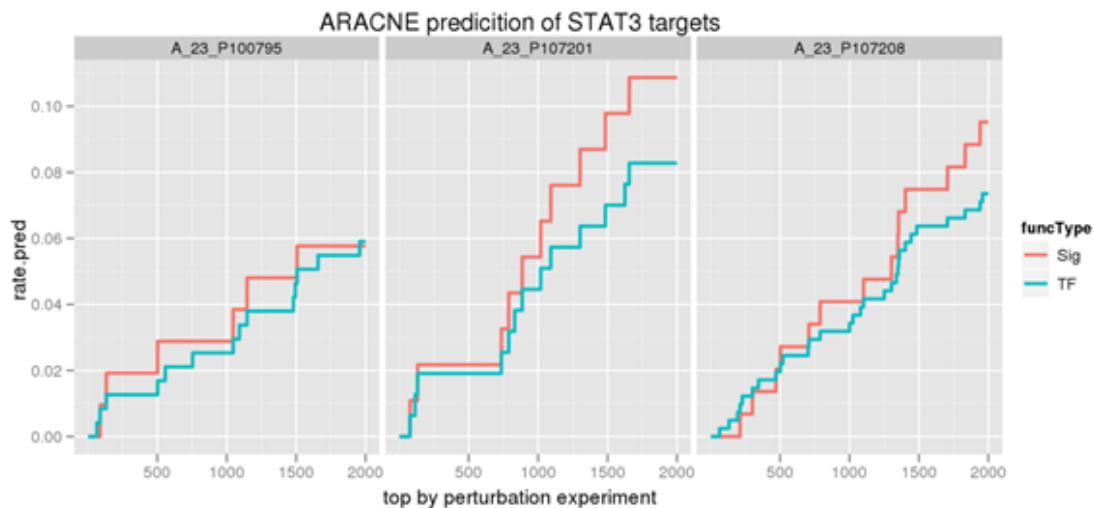


Figure 4-16 Number of targets (in percentage, upper panel) from TF (blue) or signaling (red)-centered network predicted by ARACNe that are overlapped with top experiment-identified targets for STAT3 (three probes in three columns) and odds ratio of Fisher's exact test for the overlap (lower panel). The higher the overlap is, or the higher the odds ratio is, the more powerful or more precise the prediction is.

4.6 Conclusion

In this chapter, I introduced a computational framework, NetBID2, based on network inference of both regulatory and signaling networks and Bayesian statistics, to infer disease drivers from large-scaled gene expression data. We demonstrated that this new framework is much more robust to capture disease-associated driver-type genes, is able to detect not only known drivers but more importantly, "hidden" drivers with a very high prediction rate based on validation results. I also evaluated one novel part of NetBID2, extension to signaling network, by using an experimentally-defined gold standard and confirmed that signaling network predicted by ARACNe has more power to capture true positive interactions and remove indirect true negative ones than traditional TF network.

Chapter 5 BSEA: Bayesian Set Enrichment Analysis

5.1 Introduction

Advances in techniques such as deep sequencing and high-throughput gene/protein profiling have transformed biological research by enabling comprehensive monitoring of a biological system. Analysis of such high-throughput data typically yields a list of differentially expressed genes or proteins. This list is extremely useful in identifying genes that may have roles in a given phenomenon or phenotype. However, the candidate list is usually too long to investigate all of them. Researchers are more interested in identifying underlying biological mechanisms or processes involved by a group of differentially expressed genes or proteins. That led to the development of pathway analysis or functional enrichment analysis.

Pathway enrichment analysis helps to gain insight into the underlying biology of differentially expressed genes and proteins by reducing complexity and increasing explanatory power. Additionally, identifying active pathways that differ between two conditions can have more explanatory power than a simple list of different genes or proteins.

During the evolution of pathway analysis in the past decade [114], three major classes of methods have been developed sequentially (Figure 5-1, Table 5-1). The first type of methods is over-representation analysis by doing hyper-

geometric distribution based Fisher's exact test on the overlap of top differentially-expressed genes with known member genes in the pathway. However, this type of methods gives equal weight of differential expression to the member genes in the pathway and highly depends on the selection threshold of top representative genes. That led to the second group of methods, functional class scoring or gene set enrichment analysis [97, 98]. This type of methods uses entire list of genes with differential expression scores as the reference instead of putting some threshold and selecting top differentially expressed genes and consider the genes in pathway as a gene set to test the enrichment of this set in the reference. It overcomes the unrobustness from heuristic selection and overcomes the equal weight problem by using differential expression scores weighting genes differently. A new generation of pathway analysis utilizes the pathway topology [115] to weight genes. However, this type of methods highly depends on the knowledge of the pathways. So in this chapter, we mainly focus on the second class of methods, set enrichment analysis.

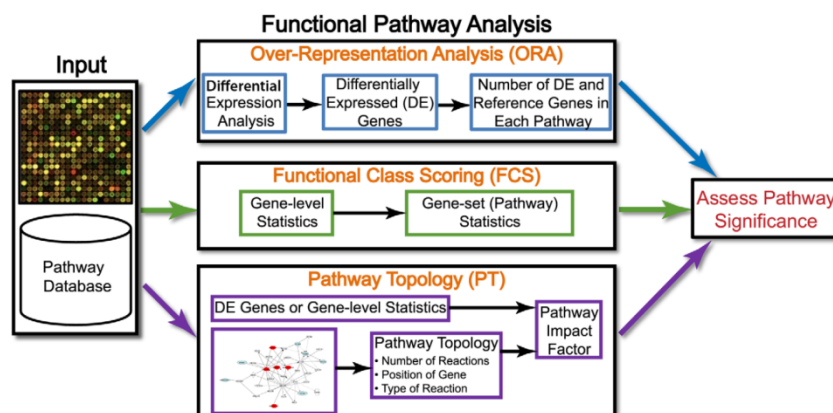


Figure 5-1 Three major categories of functional enrichment analysis methods.

Adapted from Khatri, et al, 2012 [114].

Name	Availability	Reference
ORA tools		
Onto-Express	Web (http://vortex.cs.wayne.edu)	[4,5]
GenMAPP	Standalone (http://www.genmapp.org)	[11,71]
GoMiner	Standalone, Web (http://discover.nci.nih.gov/gominer)	[72,73]
FatiGO	Web (http://babelomics.bioinfo.cipf.es)	[74]
GOstat	Web (http://gostat.wehi.edu.au)	[7]
FuncAssociate	Web (http://llama.mshri.on.ca/funcassociate/)	[6]
GOToolBox	Web (http://genome.crg.es/GOToolBox/)	[10]
GeneMerge	Standalone, Web (http://genemerge.cbc.umd.edu/)	[9]
GOEAST	Web (http://omicslab.genetics.ac.cn/GOEAST/)	[75]
ClueGO	Standalone (http://www.icl.upmc.fr/cluego/)	[76]
FunSpec	Web (http://funspec.med.utoronto.ca/)	[77]
GARBAN	Web	[78]
GO-TermFinder	Standalone (http://search.cpan.org/dist/GO-TermFinder/)	[8]
WebGestalt	Web (http://bioinfo.vanderbilt.edu/webgestalt/)	[79]
agriGO	Web (http://bioinfo.cau.edu.cn/agriGO/)	[80]
GOFFA	Standalone, Web (http://edkb.fda.gov/webstart/arraytrack/)	[81]
WEGO	Web (http://wego.genomics.org.cn/cgi-bin/wego/index.pl)	[82]
FCS tools		
GSEA	Standalone (http://www.broadinstitute.org/gsea/)	[21,29]
sigPathway	Standalone (BioConductor)	[22]
Category	Standalone (BioConductor)	[24]
SAFE	Standalone (BioConductor)	[30]
GlobalTest	Standalone (BioConductor)	[15]
PCOT2	Standalone (BioConductor)	[17]
SAM-GS	Standalone (http://www.ualberta.ca/~yyasui/software.html)	[83]
Catmap	Standalone (http://bioinfo.thep.lu.se/catmap.html)	[84]
T-profiler	Web (http://www.t-profiler.org)	[85]
FunCluster	Standalone (http://corneliu.henegar.info/FunCluster.htm)	[86]
GeneTrail	Web (http://genetrail.bioinf.uni-sb.de)	[87]
GAzer	Web	[88]
PT-based tools		
ScorePAGE	No implementation available	[37]
Pathway-Express	Web (http://vortex.cs.wayne.edu)	[38,39]
SPIA	Standalone (BioConductor)	[40]
NetGSA	No implementation available	[43]

doi:10.1371/journal.pcbi.1002375.t001

Table 5-1 A collection of available tools and methods for functional pathway enrichment analysis. Adapted from Khatri, et al, 2012 [114].

The first and the most popular enrichment analysis methods, GSEA (Gene Set Enrichment Analysis) [97, 98], was based on a signed version of the Kolmogorov-Smirnov (KS) statistic to summarize set enrichment. Later Brad Efron developed a different approach, GSA (Gene Set Analysis) [97, 98], by using a novel “maxmean” statistic to summarize enrichment and a few other statistical techniques. Multiple scoring metrics for individual differential gene

expression have been used including signal noise ratio, t-statistic, z-statistic, FC, logFC, diff mean, etc.

However, in this chapter, I will introduce a new set enrichment analysis method, BSEA (Bayesian Set Enrichment Analysis). It utilizes the “maxmean” enrichment score statistic and Bayesian modeling techniques. I will demonstrate that this new method outperforms GSEA and GSA by using the meta-analysis of RNAi screening data. As mentioned in Chapter 4, one of the key steps in NetBID2 framework is enrichment analysis, for which BSEA is the default method.

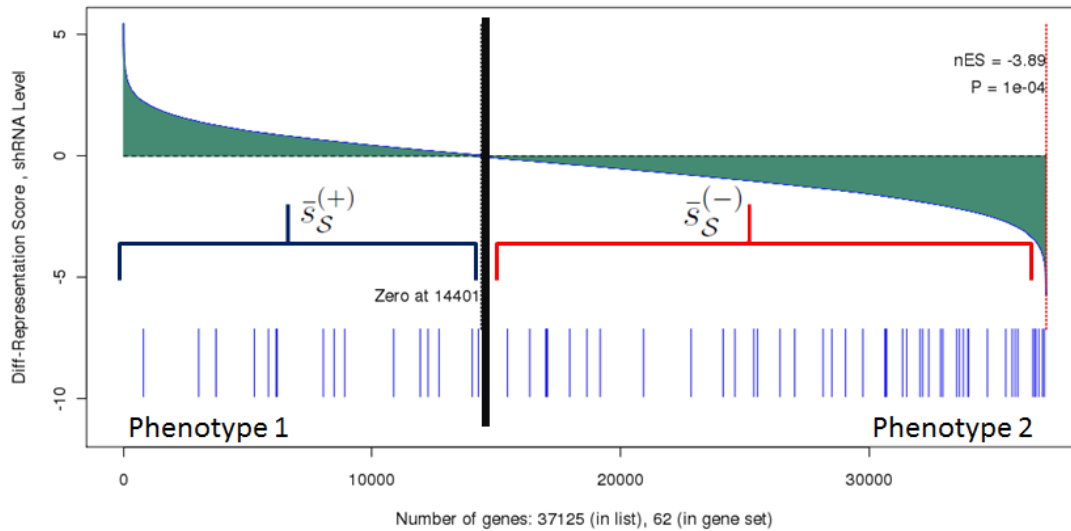
5.2 The BSEA Algorithm

The key features of BSEA algorithm include “maxmean” statistic to summarize enrichment score, restandarization for measuring statistical significance and Bayesian modeling for scoring individual gene, as discussed below.

5.2.1 “Maxmean” statistic

“Maxmean” statistic was developed by Brad Efron in his GSA method [97, 98] to summarize the enrichment of a gene set. The idea is explained in Figure 5-2. For a gene set S , we separate its member genes into positive and negative groups according to the sign of their individual scores between phenotype 1 and phenotype 2, i.e. member genes with positive scores belong to positive set while negative ones form the other set. The adjusted mean is calculated for each positive or negative subset. The adjusted mean is different from general mean in the following way: the sum of scores for genes in the positive or negative subset

is divided by the size of the entire set (union of positive or negative subsets) instead of its own subset size. And out of the two means, the one with maximum absolute value is used as the enrichment score for the full set.



$$S_{\max} = \max\{\bar{s}_S^{(+)}, \bar{s}_S^{(-)}\}$$

Figure 5-2 “Maxmean” statistic developed by Efron. The genes in the set (blue bars on the bottom) are divided into positive (red on the right) and negative (blue on the left) according to the sign of their individual scores between phenotype 1 and phenotype 2 (y axis). The adjusted mean (divided by the size of the entire set) of each subset is calculated, and the one with maximum absolute value is used as the enrichment score for this set.

According to Efron’s simulation results as shown in the ROC curves of Figure 5-3, “maxmean” statistic outperforms KS-statistic based GSEA method in both sensitivity and specificity. Later in our evaluation section using real data (5.4.1),

we also confirmed Efron's conclusion that "maxmean" method is indeed more powerful than KS statistic to summarize enrichment of a set.

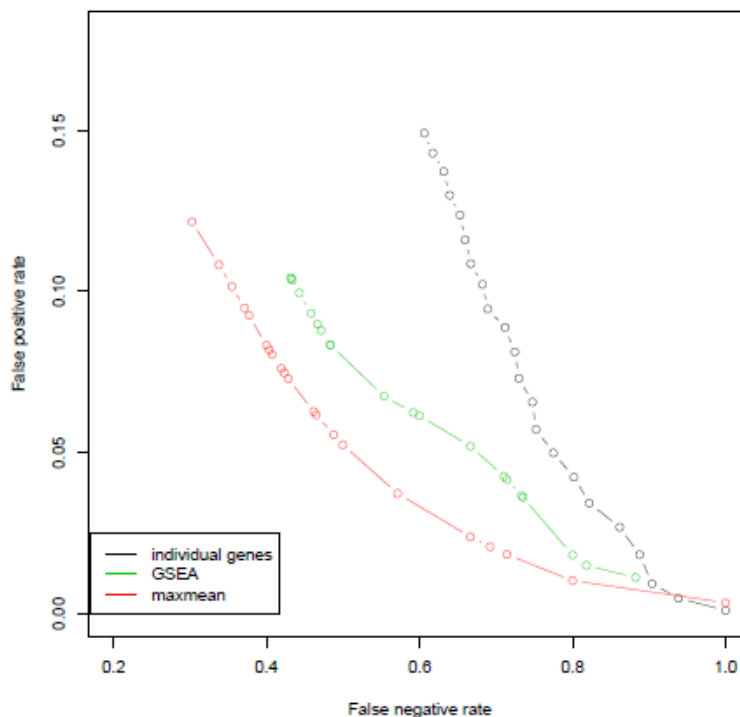


Figure 5-3 Efron's simulation results (sensitivity vs. specificity) on comparison of Maxmean statistic with KS-based GSEA. Adapted from Figure 8 in Efron, et al, 2007 [97, 98].

5.2.2 Restandardization

Another key technique BSEA adapted from GSA method is restandardization that incorporates both sample shuffling and gene shuffling for statistical significance measurement. In general to estimate a statistical significance of an enrichment score when we do gene set enrichment analysis, we use permutation test by either shuffling sample labels or shuffling gene labels. Sample permutation has been shown to have more power than gene shuffling because gene shuffling

perturbs the correlation between two genes. However, sample permutation requires a large size of samples and usually has a high false discovery rate especially when the gene sets are similar with each other. To overcome these problems, Efron developed a novel statistical technique named restandarization, to balance sample permutation and gene permutation. The basic idea is that in each sample permutation, all calculated enrichment scores are scaled according to the distribution of current enrichment scores, i.e. subtracting the mean and dividing by the standard deviation of all enrichment scores in this permutation (Figure 5-4).

$$S^{**} = \text{mean}_s + \frac{\text{stdev}_s}{\text{stdev}^*} (S^* - \text{mean}^*)$$

$$p_S = \#\{S^{**} \text{ values exceeding } S\} / B$$

Figure 5-4 Restandarization technique for statistical significance estimation.

Adapted from Efron, et al, 2007 [97, 98].

5.2.3 Bayesian inference

The only difference between BSEA and GSA lies in the usage of Bayesian statistics for individual gene scoring. The motivation is to use Bayesian techniques to overcome inaccurate estimation problems of parameters by traditional metrics to score individual genes such as signal noise ratio, t-statistic, z-statistic, FC, logFC, diff mean, etc. Those classical methods in general are

relying on large samples and good data quality, which is rare in reality. Especially in the high-throughput experiments, data is usually noisy and the sample size is small. In this situation, conventional fold change or maximum likelihood estimated parameters will be problematic. To overcome this problem, BSEA utilizes the advantage of Bayesian modeling methods for its robustness and ability to deal with noisy data, small sample size and outliers.

In NetBID2 algorithm, I already introduced a Bayesian Probit model method for differential expression analysis at individual gene level, but Probit model might require a relative large sample size. The alternative model is Bayesian linear model (Figure 5-5) and Gaussian prior or weakly-informative t-prior is commonly used for coefficients in the model. A z-score for the slope and corresponding p value will be reported to represent the statistical strength of differential expression. The slope itself can also be used to score individual gene.

<p>Bayesian Linear Gaussian Model</p> $y_i \sim N(\alpha + \beta x_i, \sigma^2), \quad i = 1, \dots, m$ <p>OR $y_i = \alpha + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$</p>	<p>Prior for Coefficients</p> <p>Gaussian Prior</p> $(\alpha, \beta)' \mid \Sigma \sim N(\mu, \Sigma)$ <p>t Prior</p> $\beta \sim (\mu_\beta, \sigma_\beta^2), \quad \sigma_\beta^2 \sim \text{Inv-}\chi^2(\nu_\beta, s_\beta^2)$ $\alpha \sim (\mu_\alpha, \sigma_\alpha^2), \quad \sigma_\alpha^2 \sim \text{Inv-}\chi^2(\nu_\alpha, s_\alpha^2)$ <hr/> <p>Prior for Variance of Noise</p> $\sigma^2 \sim \text{Inv} - \chi^2(\nu_0, \sigma_0^2)$
---	---

Figure 5-5 A Bayesian linear Gaussian model for individual gene scoring with Gaussian or t distribution as prior for coefficients and inverse Chi-square or Gamma distribution as prior of noise variance.

5.3 Benchmark and Evaluation using Meta-Analysis of RNAi Screening Data

As discussed in Chapter 3, the integration of multiple hairpins targeting the same gene to estimate gene level activity from high-throughput RNAi screening data, one category of methods to combine multiple shRNAs for a gene is to do enrichment analysis by treating all hairpins of a gene as a set and using all shRNAs in the library as the reference. And there, we introduced an evaluation strategy by using house-keeping or evolutionarily-conserved genes as a gold standard of essential genes that RNAi screening is designed to search for (3.6). So here, we follow the same idea and evaluate enrichment analysis methods only for integration of multiple shRNAs to estimate gene level activity.

Here we use exactly the same data sets, three shRNA screens (MCF7, HPAFII, OVCAR5) and four independent gene sets as references – two adapted from previous study [102] and two more recent studies on human housekeeping genes [103, 104]. Again the percentage of overlapped genes of reference set with top k hits predicted as essential genes by each enrichment method is calculated. To avoid selection bias on k, we sampled k from 0 up to 1000 with a sliding window of 5. The larger intersection with reference gene set the algorithm produces consistently, the more powerful the method is.

5.4 Results

With the benchmark RNAi screening datasets and evaluation strategy, we compare three enrichment analysis methods, my BSEA, Efron's GSA and Broad's GSEA. GSEA is equivalent to RIGER_KS method discussed in **Error! eference source not found..** GSA method was used by Allen Ashworth to report gene level activity from shRNA screens [108].

5.4.1 “Maxmean”-based GSA outperforms KS-based GSEA

First, we checked whether “maxmean” statistic proposed by Efron in his GSA method performs better to summarize enrichment score than KS statistic used by traditional GSEA, as Efron reported based on his simulation studies. We compared GSA with GSEA only first, and indeed GSA showed consistently larger AUC (Area Under the Curve) than GSEA (Figure 5-6) to identify true positives with a high precision rate.

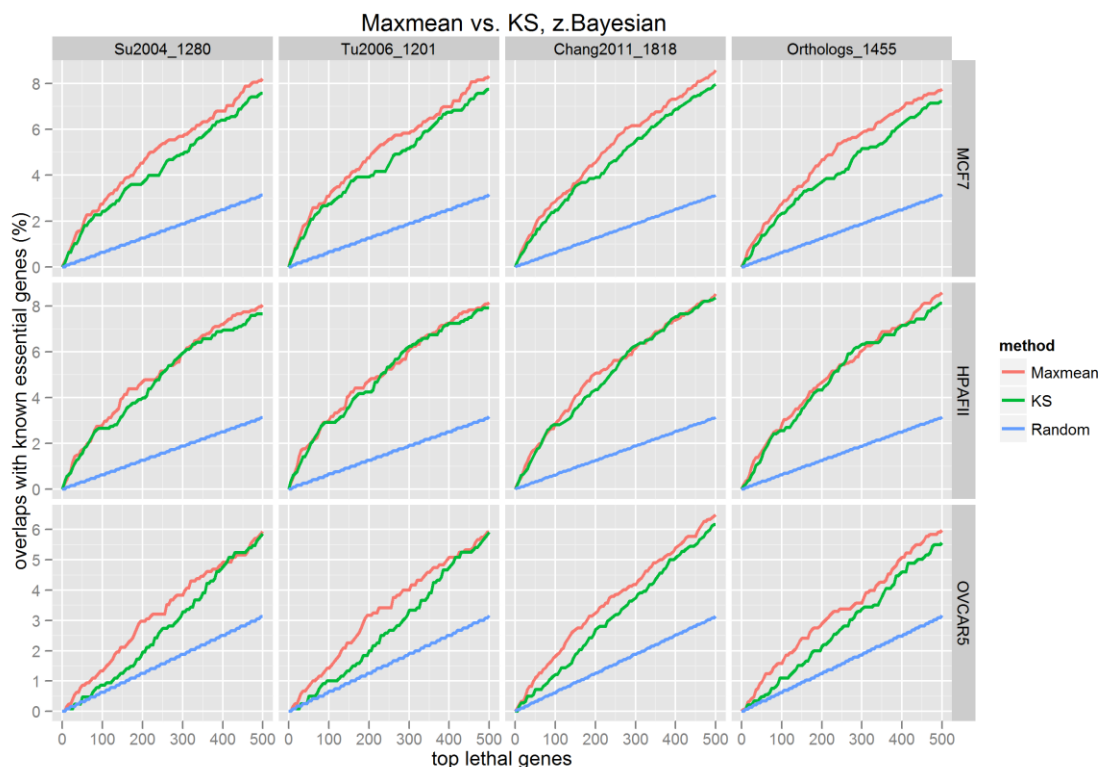


Figure 5-6 “Maxmean” statistic (GSA) vs. KS statistic (GSEA) for summarization of enrichment score. Housekeeping or conserved ortholog genes can be used as reference gene set to evaluate algorithms to detect essential genes from RNAi screens. Each colored curve shows the percentage of each reference set (“name”_“number of genes in the set”) intersected by top 0 to 1000 hits predicted as essential genes by the corresponding algorithm in each dataset. The slope of “Random” method line (in purple) is proportional to the frequency of the reference set out of all genes in the library. The greater the area under the curve, the more powerful the algorithm is.

5.4.2 BSEA \geq GSA $>$ GSEA

As shown in the comparisons of BSEA with GSA and GSEA (Figure 5-7), first, we noticed that GSEA (the blue curve) is the worst in all cases, indicating KS statistic is not a good method to summarize set enrichment. Second, in the first

two datasets, BSEA and GSA are kind of mixed together. One reason for that is because the data of the first two screens is relatively good (Figure 3-7). However, for the third one, BSEA beats GSA as expected, because BSEA uses Bayesian statistics which is much more robust than maximum likelihood statistics used by GSA. So overall, BSEA is the best enrichment analysis method comparing with GSA and GSEA.

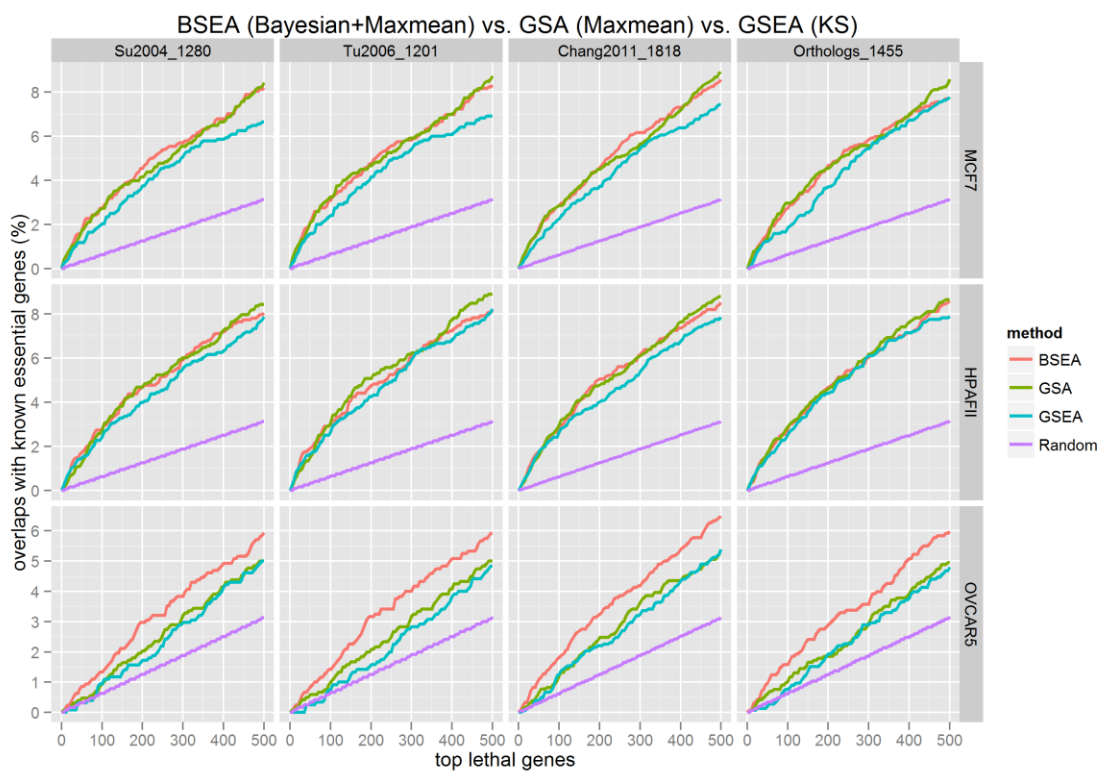


Figure 5-7 Evaluation results of BSEA, GSA and GSEA. Annotation is the same with Figure 5-6.

5.4.3 BSEA dominates GSEA

If we only compare BSEA with GSEA, as shown in Figure 5-8, GSEA dominates GSEA in all cases.

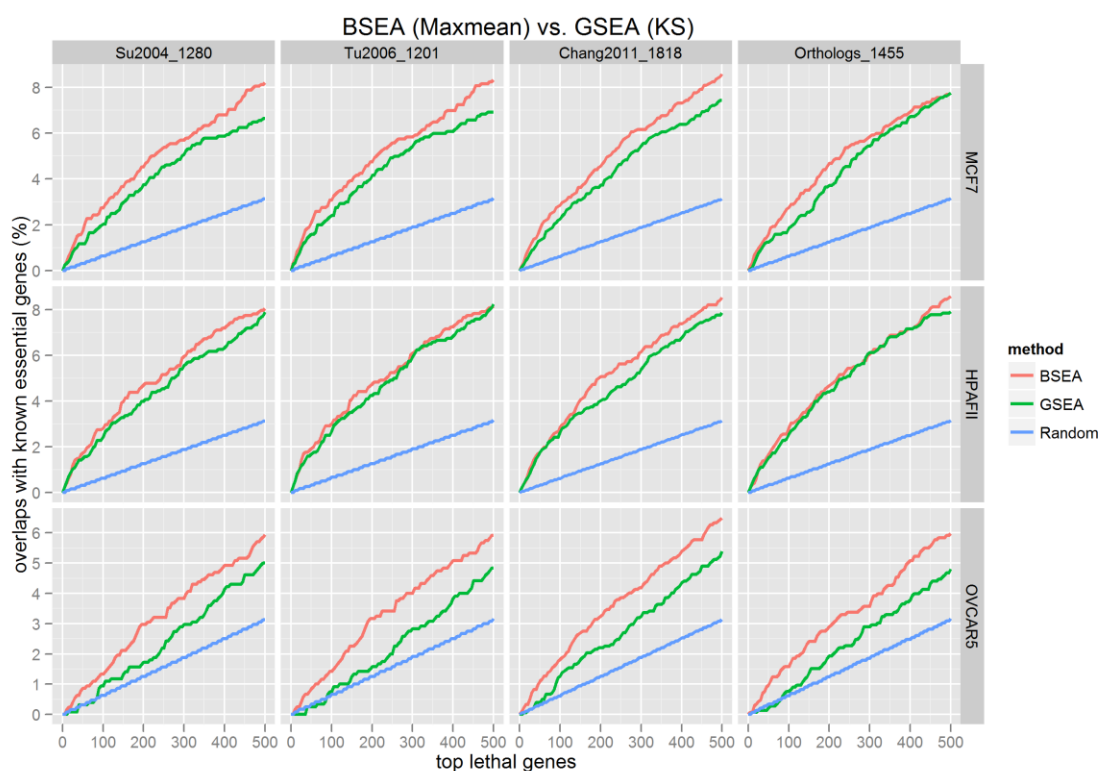


Figure 5-8 Evaluation results of BSEA vs. GSEA. Annotation is the same with Figure 5-6.

5.4.4 Bayesian vs. Frequentist

Remember that the only difference between BSEA and GSA is that BSEA uses Bayesian statistics while GSA uses classical Frequentist's maximum likelihood approaches. So we compared these two techniques by fixing all parameters but only changing individual scoring methods to Bayesian z score method or

Frequentist's t score approach. As expected shown in Figure 5-9, there is no difference between Bayesian and Frequentist's method when the data is good (the first two screens), both of which converges the optimum, however, when the data is noisy such as the third example, Bayesian shows its super power compared to Frequentist's method. So overall, Bayesian method is much more robust than classical maximum likelihood methods.

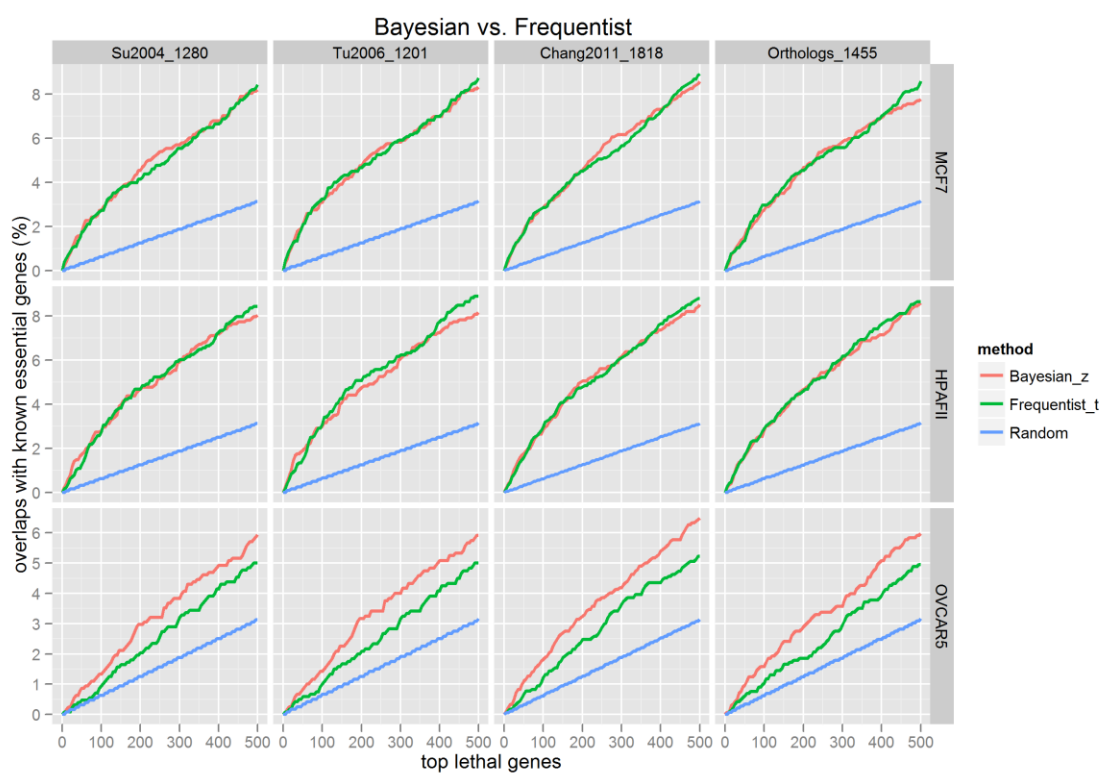


Figure 5-9 Evaluation results of Bayesian vs. Frequentist methods for individual scoring. “Maxmean” is used for enrichment score. Annotation is the same with Figure 5-6.

5.4.5 BSEA cannot beat BHM

In Chapter 3, we developed a novel Bayesian Hierarchical Modeling (BHM) approach for meta-analysis of multiple shRNAs targeting the same gene from RNAi screening data and demonstrated it's the best comparing with GSEA (RIGER). In this chapter, we developed BSEA, a better enrichment approach than GSEA, so we asked whether BSEA can beat BHM. Unfortunately, the answer is no. As shown in Figure 5-10, BSEA is getting close to BHM, but still under it. The reason might be that BHM uses the strategy of “modeling-all-together”, while BSEA is still under the framework of “separate-and-combine”.

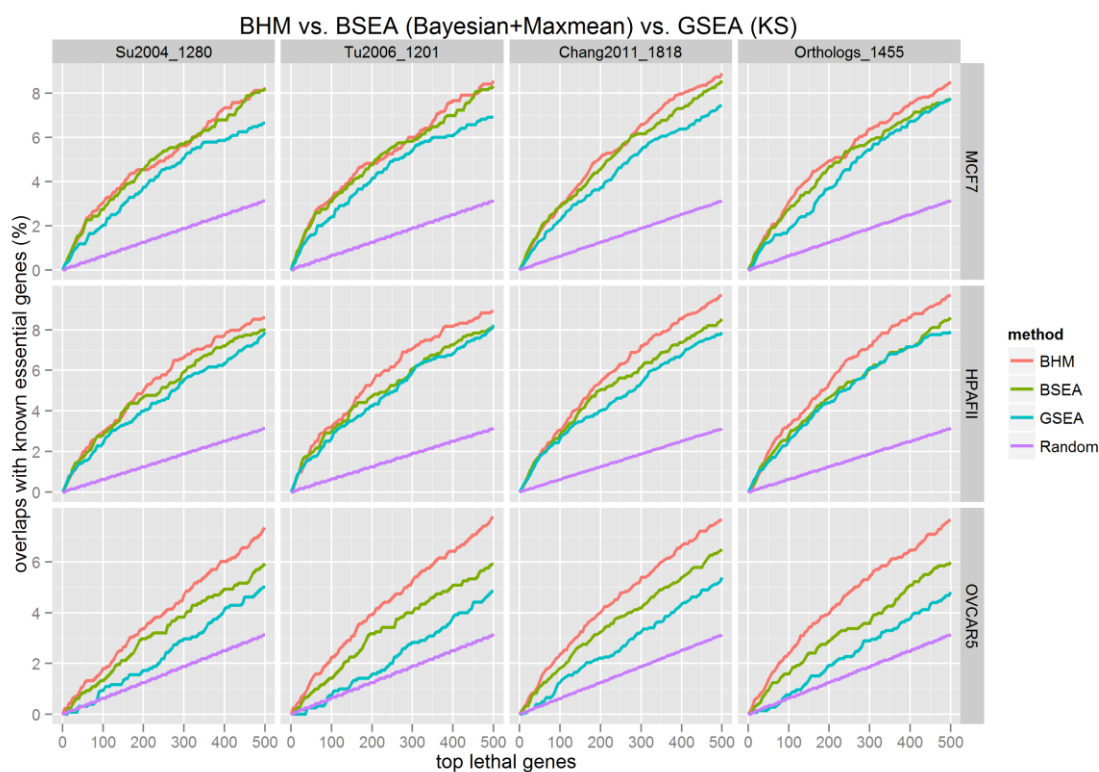


Figure 5-10 Evaluation results of BHM (Bayesian Hierarchical Model), BSEA, and GSEA. Annotation is the same with Figure 5-6.

5.5 Conclusion

To summarize this chapter, I developed a novel approach for set enrichment analysis by using “maxmean” statistic and Bayesian inference. Based on the evaluation results using RNAi screening data, I have demonstrated that BSEA outperforms existing GSA and GSEA, especially when the data is noisy. I confirmed that “maxmean” is more powerful than KS for set enrichment score, and Bayesian is more robust than classical statistic for scoring at individual level. Although BSEA cannot beat BHM for meta-analysis of RNAi screening data due to the problem of “separate-and-combine” strategy, overall, BSEA is the best algorithm for set enrichment analysis with high sensitivity and precision.

Chapter 6 Recovering Drug-Induced Apoptosis Subnetwork from Connectivity Map Data via a Bayesian Network Approach

6.1 Summary

The Connectivity Map project profiled human cancer cell lines exposed to a library of anti-cancer drugs or chemical compounds with the goal of connecting cancer with underlying genes and potential treatments. Since the therapeutic goal of most anti-cancer drugs is to induce tumor-selective apoptosis, it is critical to understand the specific cell death pathways that are activated by drugs. This can help to better understand the mechanism of how cancer cells respond to chemical stimulations and improve the treatment of aggressive human tumors. In this study, using Connectivity Map microarray data from breast cancer cell line MCF7, we applied a Gaussian Bayesian network modeling approach and identified apoptosis as a major drug-induced cellular-pathway. In order to reduce computational complexity without losing generality, we focused on 13 apoptotic genes that showed significant differential expression across all drug-perturbed samples. In our predicted subnetwork, 9 out of 15 high-confidence interactions were validated in literature, and our inferred network captured two major cell death pathways by identifying BCL2L11 and PMAIP1 as key interacting players for the intrinsic apoptosis pathway, and TAXBP1 and TNFAIP3 for the extrinsic

apoptosis pathway. Our inferred apoptosis network also suggested the role of BCL2L11 and TNFAIP3 as ‘gateway’ genes in the drug-induced intrinsic and extrinsic apoptosis pathways. Our study extended the usage of Connectivity Map data and applied a Bayesian network framework to recover underlying drug-induced biological programs for a better understanding of the mechanism of action of cancer drugs, and provided potential targets in the apoptosis pathway for better cancer treatment.

6.2 Introduction

One goal of biomedical research is to better understand human diseases such as cancer by studying gene patterns associated with diseases and using them to find the best potential treatments. Recently, Todd Golub and his colleagues at the Broad Institute initialized the “Connectivity Map” project (CMAP) [116, 117] to make these disease-gene-drug connections by utilizing microarray technology. High-throughput microarrays are able to profile gene expression at the level of the whole-genome, and can be used to detect signatures under certain perturbations or phenotypes in cells [118]. Since the therapeutic goal of most anti-cancer drugs is to induce tumor-selective cell death [119], it is reasonable to hypothesize that apoptosis may be a major cellular mechanism targeted by anti-cancer drugs. It is therefore critical to understand the specific cell death pathways that are activated by drugs. This would help to better understand the mechanism of how cancer cells respond to chemical stimulations and improve the treatment of aggressive human tumors.

Apoptosis in mammalian cells is induced by intracellular cysteine proteases known as caspases. Caspases are first synthesized as largely inactive zymogens known as procaspases, and are later activated through post-translational mechanisms. Two principal pathways of caspase activation have been recognized [120, 121]. One pathway, which is of more ancient origin and evolutionarily conserved, is known as the stress pathway, mitochondrial pathway, or intrinsic pathway [120, 121]. It is induced by developmental cues and diverse intracellular stresses. This pathway begins with the activation of caspase-9 on a scaffold formed by Apaf-1 in response to cytochrome c release from damaged mitochondria. It is known to be regulated primarily by proteins from the Bcl-2 family. The other pathway is known as the extrinsic pathway, and is triggered by so-called 'death receptors' on the cell surface. The death receptors are engaged by cognate ligands of the tumor necrosis factor (TNF) family. This pathway begins with the activation of caspase-8 (and caspase-10 in human cells), via adaptor proteins including Fas-associated death domain protein (FADD) [120, 121]. Once activated, caspase-9 in the intrinsic pathway or caspase-8 (-10) in the extrinsic pathway activates downstream 'effector caspases' including caspases-3, -6 and -7. In an expanding cascade, these caspases carry out the execution phase of cell death.

Because the CMAP database contains profiles from a large collection of human cancer cell lines that capture information of how cells respond to chemical stimulations, it can be used to test the hypothesis that the apoptosis pathway might be a major responsive program of drug perturbations in cancer cells. One

can do this by enrichment analysis of apoptotic genes in drug-responsive genes or in differentially-expressed genes in drug-exposed cancer cells. CMAP data also contains dynamic transcriptional activities of most genes across diverse conditions, giving sufficient data for associating the activities of genes of interest with each other, and for reconstructing parts of the apoptosis pathway in the context of drug-exposed cancer cells. In this study, we used CMAP gene expression profiles to test the hypothesis that apoptosis may be a major drug-induced cellular mechanism. We then employed a Gaussian Bayesian network modeling approach to reconstruct the subnetwork of the drug-induced cell death pathway. To minimize the effects of heterogeneity from different tumor types, our study focused on a single breast cancer cell line, MCF7.

To better understand whether anti-cancer drugs target the intrinsic and extrinsic apoptosis pathways, and identify specific pathways or interactions activated by anti-cancer drugs, we crossed our predicted drug-triggered apoptosis network with literature-validated interactions. We were able to identify key players as well as interactions in the drug-induced intrinsic and extrinsic pathways. Our results shed light on the mechanism of action of drugs in cancer cells and may lead to improved treatments that target key apoptotic proteins that are most related to drug response.

6.3 CMAP Data

The CMAP "build 02" gene expression dataset (<http://www.broad.mit.edu/cmap/>) contains over 7,000 profiles of cancer cells that have been exposed to

perturbations by 1,309 compounds, and contain data from five human cancer cell lines: MCF7, PC3, SKMEL5, HL60 and ssMCF7. The microarray platforms used include Affymetrix HT_HG-U133A and HT_HG-U133A_EA. To avoid the effects of tumor heterogeneity and multiple microarray platforms, to avoid the heterogeneity of different cellular contexts, we only focused on samples from the breast cancer cell line MCF7 that were profiled using the Affymetrix HT_HG-U133A platform. The dataset is composed of 404 control and 2,417 compound-perturbed samples. The HT_HG-U133A microarray platform contains 22,268 Affymetrix probe sets representing 13,262 genes. The GCRMA method [122] was used to normalize the data.

6.4 Drug-Response Signature Analysis

To identify drug-responsive signature genes at a transcriptional level in cancer cells, one approach is to perform differential gene expression analysis by comparing drug-perturbed samples with controls. However, since the dataset contains samples tested with over 1,000 chemical perturbations, it is important we take into account the diverse mechanisms of actions of the different compounds. One solution would be to perform differential expression analysis for each compound separately and then combine the results together using a p-value-based Fisher's method or Stouffer's z-score approach to obtain the overall differential expression level for each gene across all compounds. However, a limitation with this type of analysis has to do with the fact that each compound only has a limited number of perturbed samples and even smaller number of

control samples. This would cause the statistical power to be extremely low for individual compound analysis, and would result in an inaccurate estimation of parameters and a high false positive rate. In addition, another known issue with this type of 'Separate-then-Combine' analysis is a low precision rate, which means there is a high occurrence of false positives among the most differentially expressed genes or top-hits. One way to overcome this drawback is to combine all compounds together at the beginning, as known as 'complete pooling' method. Although different drugs may have distinct mechanisms of action and different target proteins, it may still be reasonable to group them together. One reason is that there are a relatively limited number of pathways or mechanisms through which cells respond to chemical stimulations. Also, compounds tested for cancer treatment are known to share some common characteristics. For example, a large number of anti-cancer drugs are known to induce cell death or repress cell growth programs. In addition, the combination or 'complete pooling' strategy increases the sample size from less than 5 to thousands, dramatically increasing the statistical power for inferring true responsive genes across all compounds. This assumption is also confirmed by the fact that most perturbed profiles are clustered together as shown in Figure 6-2. These results indicate that the variability of transcriptional profile for the same type of cell (MCF7 in this study) due to drug heterogeneity is much smaller than that caused by different chemical stimulations.

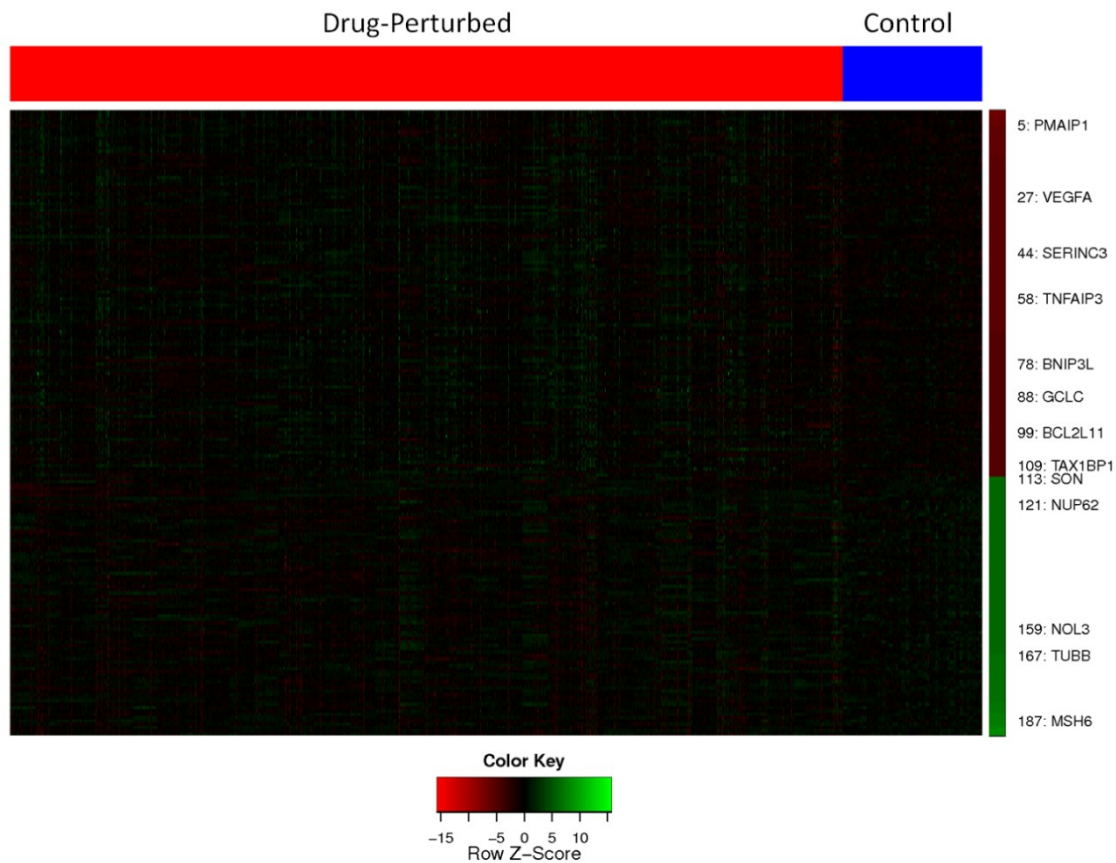


Figure 6-1 Heatmap of top differentially-expressed genes (FDR<0.05) in drug-perturbed and control samples. The genes are ranked from most up-regulated (labeled in dark red on right panel) to most down-regulated (labeled in dark green) in drug-perturbed samples, and the 13 selected apoptotic genes are labeled on the right with their ranks in the list.

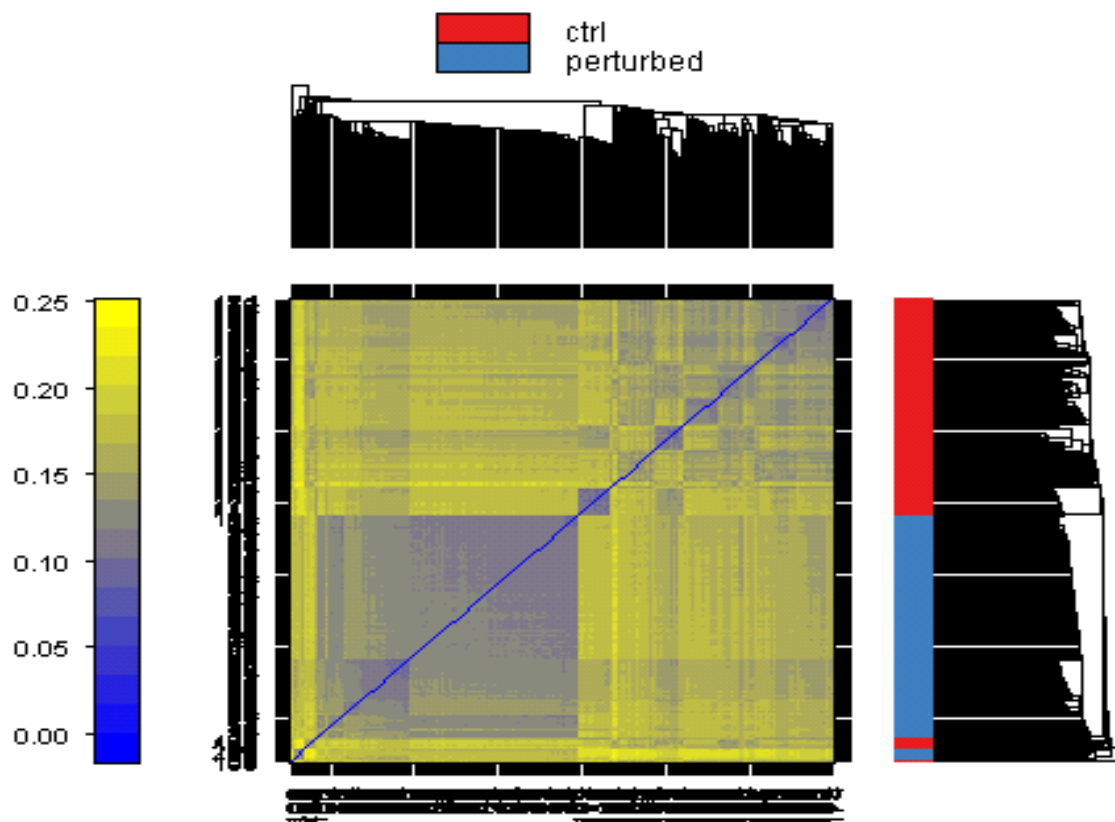


Figure 6-2 The heat map of distances between profiles of CMAP data including randomly-selected 100 control and 100 drug-perturbed samples.

To estimate the effect of each compound on gene expression and to test the significance of differential expression for each probe set, we used a linear modeling method with empirical Bayes moderated t-test [123]. A non-parametric Bonferroni procedure was employed for multiple comparison correction. Using a false-discovery-rate (FDR) threshold of 0.05, we identified 137 up-regulated and 90 down-regulated probe sets, representing 112 over-expressed and 79 under-expressed genes respectively, in drug-perturbed cancer cells.

6.5 Enrichment of Apoptosis Pathway

As described previously, one of the most important mechanisms through which cancer drugs act is the inducement of cell death programs. More specifically, we hypothesized that the apoptosis pathway may be a major drug-induced program. Enrichment analysis was proposed to validate this hypothesis. By searching the Gene Ontology database [93, 124], we obtained a list of 380 human genes that were annotated with apoptosis-related GO terms. 211 genes were annotated as pro-apoptotic by induction of apoptosis, positive regulation of apoptosis, and negative regulation of anti-apoptosis. 194 genes were annotated as anti-apoptotic by negative regulation of apoptosis and positive regulation of anti-apoptosis. 25 genes were involved in both positive and negative regulation of apoptosis. We then performed enrichment analysis with differentially-expressed genes of drug-perturbation in the apoptosis pathway. Two methods were employed to do this analysis: the first method was the Fisher's exact test to validate whether known apoptotic genes were overrepresented in a selected differentially-expressed drug-responsive gene set. The second method was to test the known apoptotic genes using Gene Set Enrichment Analysis (GSEA) which does not perform a selection on differentially-expressed genes, but instead considers the entire set of genes and their differential expression as the background. For Fisher's exact test, a set of previously-identified 191 signature genes with a threshold of $FDR < 0.05$, and all 12,632 genes in the microarray were used to fit the null hyper-geometric distribution. For GSEA, the mean of absolute value of differential expression was used as enrichment score because apoptotic

genes could be either up- or down-regulated in drug-perturbed samples. The significance of the enrichment scores were tested against 10,000 permutations of gene names.

There are 13 genes Table 6-1 that overlap between the 191 drug-inducement signature genes and the 368 human apoptotic genes in our dataset. The significance level of Fisher's exact test for this overlap is approximately 0.001 (), consistent with the result from GSEA, which had a p-value of 0.002 (Figure 6-3). Therefore, both methods confirm that the pre-identified drug-induced signature genes are significantly enriched in the human apoptosis pathway. In other words, we were able to validate our hypothesis that the apoptosis pathway is a major cellular mechanism targeted by anti-cancer drugs. Furthermore, separate analysis of pro- or anti-apoptotic genes (Figure 6-4, Figure 6-5) showed that drug-responsive genes were enriched in both positively- or negatively-regulated apoptosis gene sets. Since the analysis was done using the combination or 'complete pooling' strategy, the significance of these results suggests that 13 drug-induced apoptotic genes in our gene set are responsible for a highly conserved response to multiple chemical compounds in the context of breast cancer.

	probeld	entrezld	logFC	t	pval	FDR	annotated apoptosis type*
PMAIP1	204285_s_at	5366	0.32	7.46	1.19E-13	2.64E-09	pro
VEGFA	210512_s_at	7422	0.18	6.15	8.88E-10	1.98E-05	anti
SERINC3	221471_at	10955	0.06	5.62	2.08E-08	4.64E-04	pro
TNFAIP3	202644_s_at	7128	0.24	5.39	7.61E-08	1.70E-03	anti
BNIP3L	221479_s_at	665	0.12	5.04	4.88E-07	1.09E-02	both
GCLC	202923_s_at	2729	0.08	4.91	9.42E-07	2.10E-02	anti
BCL2L11	222343_at	10018	0.14	4.83	1.41E-06	3.14E-02	pro
TAX1BP1	200976_s_at	8887	0.07	4.76	2.01E-06	4.47E-02	anti
SON	214988_s_at	6651	-0.06	-4.75	2.11E-06	4.71E-02	anti
NUP62	202153_s_at	23636	-0.11	-4.83	1.41E-06	3.14E-02	anti
NOL3	59625_at	8996	-0.13	-5.32	1.13E-07	2.53E-03	anti
TUBB	212320_at	203068	-0.09	-5.55	3.11E-08	6.92E-04	pro
MSH6	202911_at	2956	-0.09	-6.40	1.87E-10	4.16E-06	pro

Table 6-1 The 13 selected differentially-expressed or drug-responsive apoptotic genes. *: pro: annotated by GO terms: induction of apoptosis, positive regulation of apoptosis, negative regulation of anti-apoptosis; anti: annotated by GO terms: negative regulation of apoptosis, positive regulation of anti-apoptosis.

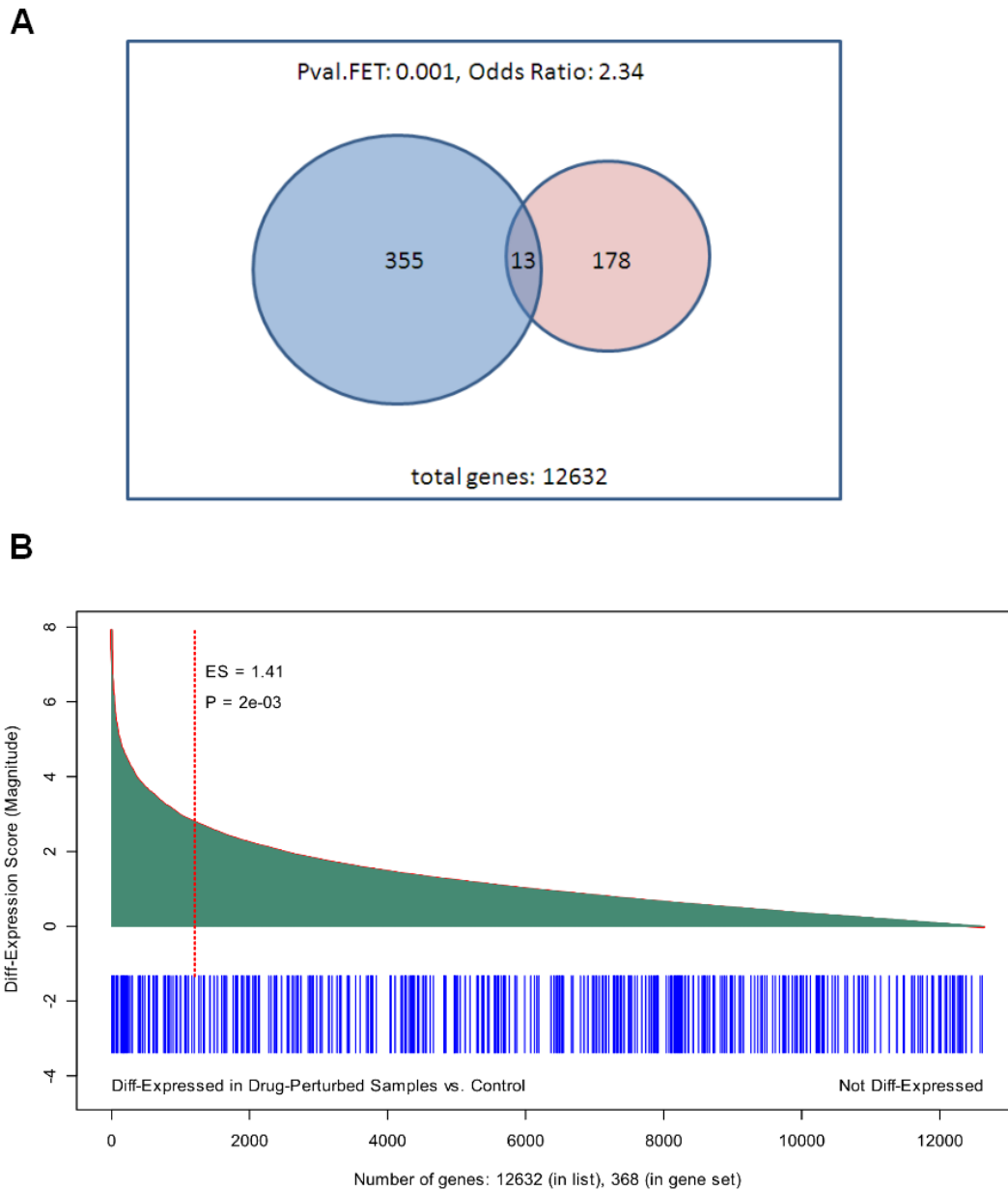


Figure 6-3 Summary of (A) Fisher's Exact Test and (B) Gene Set Enrichment Analysis (GSEA) to test whether apoptosis pathway with 368 apoptotic genes is enriched in drug-induced signature genes. For GSEA method, absolute mean was used to summarize the enrichment and 10,000 gene permutations were used to produce the significant level.

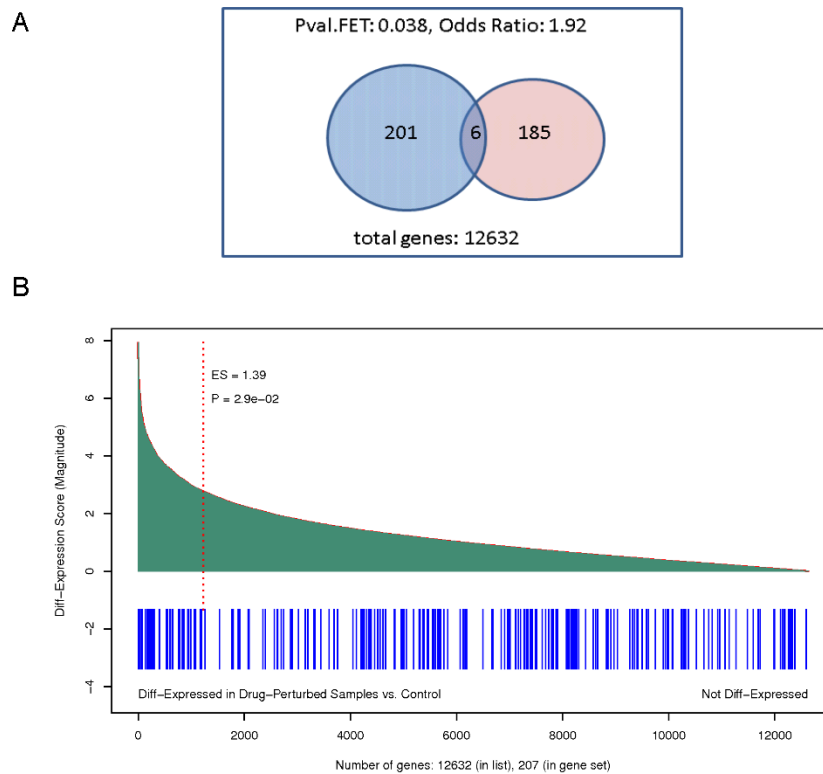


Figure 6-4 Summary of (A) Fisher's Exact Test and (B) Gene Set Enrichment Analysis (GSEA) to test whether 207 pro-apoptotic genes are enriched in drug-induced signature genes.

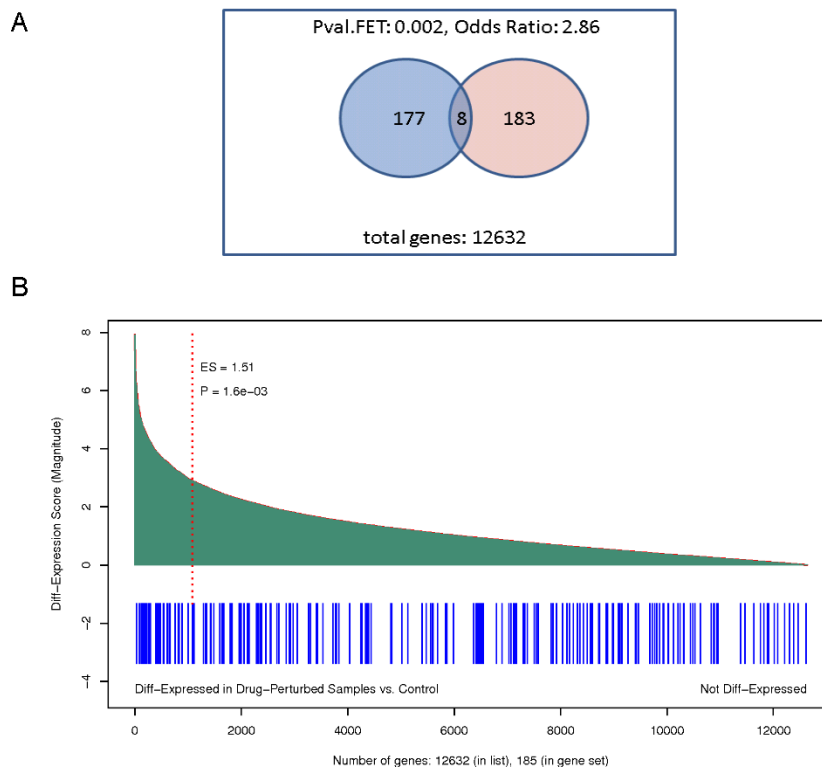


Figure 6-5 Summary of (A) Fisher's Exact Test and (B) Gene Set Enrichment Analysis (GSEA) to test whether 185 anti-apoptotic genes are enriched in drug-induced signature genes.

6.6 Bayesian Network

We next asked the question of how the 13 identified genes work together systematically and whether we can recover the underlying network structure of their interactions. This would help us to better understand the mechanism of how cancer drugs induce the apoptosis pathway at a global systems level. In order to infer the underlying signaling, transcriptional, and causality network of the 13 drug-induced apoptotic genes, we used one of the best methods for network

reconstruction in the literature, the Bayesian Network or Graphical Model [125-129]. The details of the method are described below.

6.6.1 Data modeling

A Bayesian network represents the dependence structure of a joint probability distribution of multiple variables, which can be factorized into a product of distributions of each individual node conditioning on its parents. To model the local distribution of each node conditioned on its parents, a commonly used method for continuous data is to discretize data points into bins and then fit a multinomial distribution to the discretized data. However, data discretization results in a loss of information and can be highly sensitive to the number of bins the data is split into. Furthermore, due to the continuous nature of microarray data and the marginal normality of many genes in this study as shown in Figure 6-6, we determined it would be more accurate to employ a continuous model. We therefore used a conditional linear Gaussian model [130] for the local distribution of each node as shown below:

$$(g_i | \text{parents}(g_i), \beta, \alpha_i, \sigma_i^2, G) \sim N(\sum_j \beta_j * \text{parent}_j(g_i) + \alpha_i, \sigma_i^2)$$

This model can be recognized as a linear regression model, in which node g_i is the response variable, its parents are covariates, and the noise follows a white Gaussian distribution with mean 0 and variance σ_i^2 .

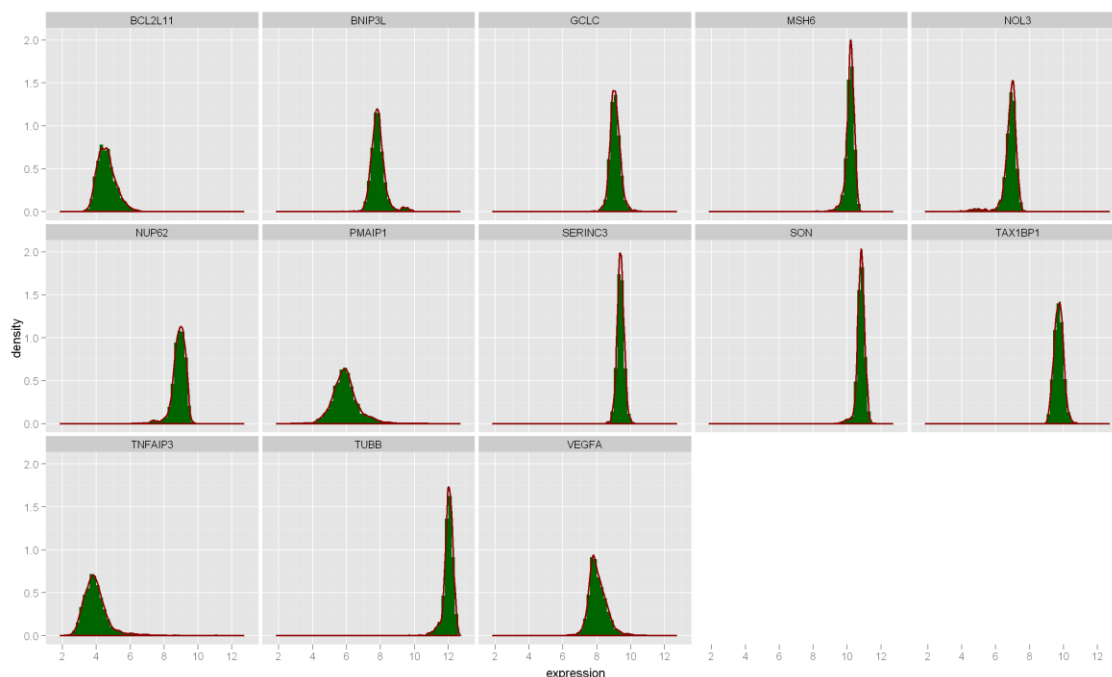


Figure 6-6 Marginal distributions of the 13 selected drug-responsive apoptotic genes across all samples in CMAP data.

6.6.2 Parameter learning

Given the linear regression model for the local distribution, a classical Maximum Likelihood or Least Squares approach can be used to estimate its parameters. However, various studies in statistics have suggested that Bayesian approaches or Bayes estimators are more robust than Frequentist maximum likelihood method [129], especially when the sample size is small or the data is noisy. Therefore a Markov Chain Monte Carlo (MCMC) simulation-based Bayesian computing method was used to estimate parameters of the model. To select the priors for the Bayesian model, two principles were followed: one is conjugation for computing easily as the posterior will fall in the same distribution family as

prior, and in our case, the prior would be Gaussian for conditional coefficient and Inv-Gamma for variance; the other is global and local parameter independence, parameter modularity and likelihood equivalence [131, 132].

6.6.3 Structure scoring and search

To determine the Bayesian network or Graphical model that can best fit the data, we needed a scoring system to compare different potential network structures. For structure learning, a Bayesian factor-based method, which compares the conditional probability of each graphical structure given observed data, was used. As shown below, according to Bayes theorem, the odds ratio between two possible structures, G_1 and G_2 can be decomposed as a product of structure prior odds ratio and the Bayesian factor, which is the ratio of the likelihoods of the two graphical models.

$$\frac{p(G_1 | D)}{p(G_2 | D)} = \frac{p(G_1)}{p(G_2)} * \frac{p(D | G_1)}{p(D | G_2)}$$

Using the uniform distribution for structure prior, which is reasonable because we have no preference on particular graphical structure, the score for a network structure, G , can be defined as the following formula, which is the log-likelihood of the graphical model.

$$score(G : D) = \log p(D | G) = \int p(D | \theta, G) p(\theta | G) d\theta$$

In our study with 13 variables, there were $1.86766e+31$ possible directed acyclic graphs [133], so it was not realistic to enumerate the entire network structure

space. To search more efficiently, we used a classical heuristic algorithm: hill climbing with random restarts [134, 135]. Using this stochastic algorithm, the search-space was reduced dramatically. Using 2 restarts, we only needed to compare 12,655 structures before reaching a maximum score. One risk was that we had found a local maximum, rather than the global maximum, but the risk would be decreased further by increasing the number of restarts.

6.6.4 Bootstrapping and model averaging

With the methods outlined above, we obtained a Bayesian network structure that best described the observed data. However, it is possible that the model may be over-fitted, which means that a small change to the dataset could make the network structure change dramatically. A way to solve this issue is to apply a re-sampling method or simulating the dataset. The method would learn the best graphical model for each sampled dataset, and generate a consensus network from the average of the sample models. This method is also known as model averaging. The simulation method we used to do model averaging was Efron's bootstrapping method [136, 137]. To increase robustness, the method only considered predicted network structures with a score within 95% of the confidence interval. The distribution of network scores is shown in Figure 6-7. In generating the final combined consensus network, edges were selected based on a confidence threshold of 75%.

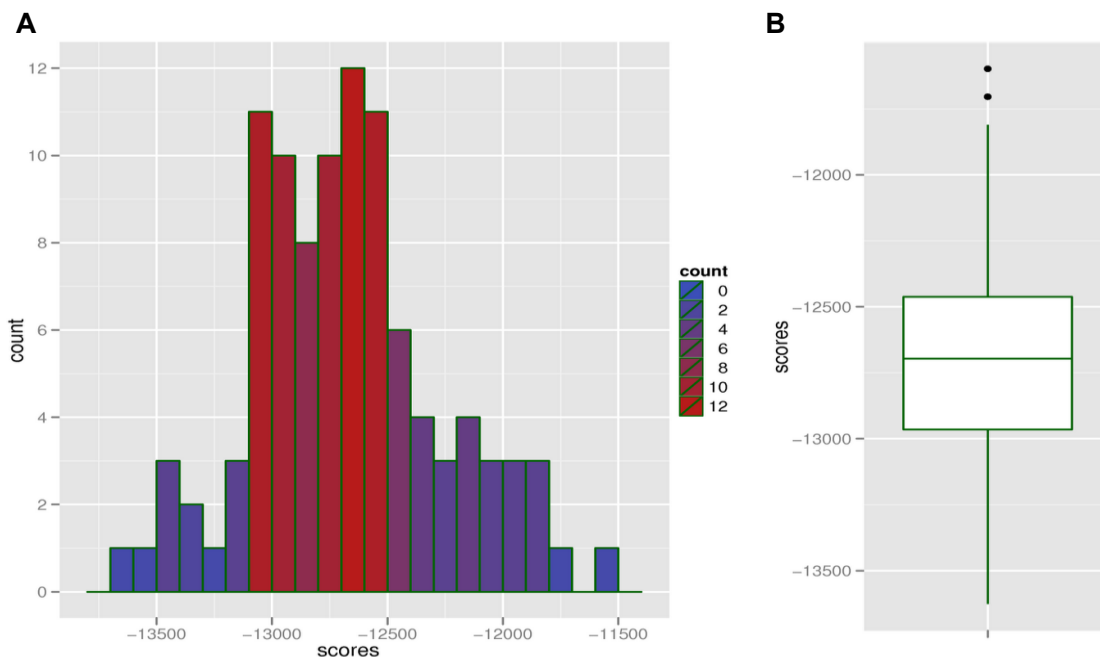


Figure 6-7 (A) Histogram and (B) Box plot of scores for best-learned graphical model in each bootstrapped sampling.

6.7 Results

Using the described Gaussian Bayesian network modeling framework, a network model was generated for the 13 identified drug-responsive apoptotic genes as shown in Figure 6-8-A. The network contains 15 interactions and each edge has a confidence of over 75%. The inferred interactions represent dependence among these 13 genes of interest, which may be due to direct or indirect protein-protein interactions, transcriptional regulation, or signal transduction. To validate the inferred interactions, we searched the Interactions component of NCBI Gene database (<http://www.ncbi.nlm.nih.gov/gene>), which contains data from multiple interaction databases such as BIND, HPRD, BioGRID, etc. We then generated a

validated interaction network of the 13 apoptotic genes using their validated interactions [Figure 6-9]. The validated network contained 216 interacting genes, including our 13 genes of interest. The network also contained 243 interactions after removing duplicate interactions (365 interactions with duplicates). When compared with our predicted network, 9 out of 15 predicted interactions were found to be direct or indirect interactions in the validated network [marked in red, Figure 6-8-A]. An indirect interaction means the network does not contain a direct edge between the two genes, but there exists a path between them via intermediate genes. Since we only considered 13 apoptotic genes in network inference, it is highly possible that the inferred interactions are indirect, but illustrate the dependence or information transmission between the two corresponding genes. More precisely a sub-validated network that includes only evidences (20 nodes and 28 interactions) for our predicted interactions was extracted as shown in Figure 6-8-B. For indirect evidences, we only counted the shortest paths between two apoptotic genes of interest.

6.7.1 Known direct interactions

Two edges in our predicted network (marked in thick red, Figure 6-8-A) have been validated as direct interactions in literature and are clearly annotated in the functional summary of corresponding genes as shown below.

TAX1BP1 -> TNFAIP3: As seen in the annotation of TAX1BP1, Tax1 (human T-cell leukemia virus type I) binding protein 1, from the NCBI Gene database, this protein inhibits TNF-induced apoptosis by mediating TNFAIP3's anti-apoptotic

activity [138, 139]. *In vivo* experiments and yeast two hybrid assays also confirm the interaction between TNFAIP3 (zinc finger protein A20) and TAX1BP1. TNFAIP3 also interacts with TXBP151, an anti-apoptotic protein and may inhibit inflammatory signaling pathways such as TNF-induced NF- κ B activation [140, 141]. TNFAIP3 and TAX1BP1 inhibit the inflammatory signaling pathway by interacting with Ubc13 and Ubch5c and triggering their ubiquitination and proteasome-dependent degradation [142].

PMAIP1 -> BCL2L11: Although there is no evidence showing direct interaction between PMAIP1 (also known as NOXA) and BCL2L11 (also known as BIM), the functional annotation of PMAIP1 [143] from the UniProtKB/Swiss-Prot database shows that the PMAIP1 competes with BCL2L11 to bind with MCL1 and can displace BCL2L11 from its binding site on MCL1. The predicted interaction between PMAIP1 and BCL2L11 may be explained by the competition between PMAIP1 and BCL2L11 in binding MCL1. The competition may occur either through a direct interaction between the two proteins, or through a third protein that is able to bind both. In addition, both PMAIP1 and BCL2L11 have been shown to interact directly with many other BCL2 protein family members including BCL2, BCL2A1, BCL2L1 and BCL2L2 [144, 145]. This indicates that NOXA and BIM may share common binding regions to BH3-only BCL2 family proteins. NOXA and BIM as BH3-only proteins have been recognized as critical mediators of anti-cancer drug- and p53-induced apoptotic responses [146, 147], which are consistent with our findings in this study that both of them are differentially expressed drug-responsive genes.

6.7.2 Consistency with two major cell death pathways

As described previously, there are two major apoptosis programs in mammalian cells: the intrinsic or mitochondrial stress-induced pathway and extrinsic or death receptor-triggered pathway. Our predicted network captures the important players and key interactions in both apoptosis programs. For the intrinsic pathway, our predicted network identifies two of the most important mediators, BCL2L11/BIM and PMAIP1/NOXA, and their competing interaction in terms of regulating many other BH3-only BCL2 family member proteins including BCL2, BCL2L1, BCL2L2, BCL2A1 and MCL1, which is illustrated as well as in the validated network (Figure 6-8-B). For the extrinsic death receptors-triggered pathway, we successfully recovered one representative of cancer-therapy or drug-induced cell death pathway: TNF-induced apoptosis. TNFAIP3/A20 and TAX1BP1/TXBP151 are two key players of this pathway, and they interact with each other to turn on the down-stream cell death machinery.

6.7.3 BCL2L11/BIM as a gateway gene to drug-induced intrinsic apoptosis

As shown in our inferred drug-induced apoptotic sub-network, BCL2L11 is located downstream of most cell death sub-pathways, which includes drug-affected apoptotic genes such as BNIP3L, NOL3, PMAIP1, NUP62, and SON. This suggests that BCL2L11 may act as a downstream gate or switch for drug- or stress-induced apoptosis programs. This finding is consistent with the main role of BCL2L11 as an apoptosis facilitator. The mechanism through which BCL2L11, a BH3-only protein, activates cell death is by inactivating Bcl-2-like proteins,

keeping them from restraining Bax and Bak. Bax or Bak can cause the outer membrane of the mitochondria to become permeable. This releases cytochrome c, which provokes Apaf-1 (apoptotic protease-activating factor 1) to activate caspase-9 [120]. The gateway role of BCL2L11 has also been illustrated in our literature-generated validation network (Figure 6-8-B).

6.7.4 TNFAIP3/A20 as a gateway gene to drug-induced extrinsic apoptosis

As shown in both our predicted network and validated network (Figure 6-8), TNFAIP3, a zinc finger protein, acts as a hub by transmitting upstream signals from cell death receptors to downstream cell death cascades. This suggests that TNFAIP3 may be a gateway protein for drug-induced extrinsic apoptosis. TNFAIP3/A20 acts as a key player in TNF-induced apoptosis by inhibiting NF- κ B activation. These results indicate that TNF-induced signaling may be the most common anti-cancer drug or chemical compound-triggered cell death program. Many studies have demonstrated the involvement of the TNF-mediated apoptosis in cancer therapies such as ionizing radiation or the chemotherapeutic agent, daunorubicin [138]. This again confirms our hypothesis that anti-cancer drugs induce apoptosis of cancer cells and confirms that apoptosis pathways can be inferred from drug-perturbed gene expression profiles.

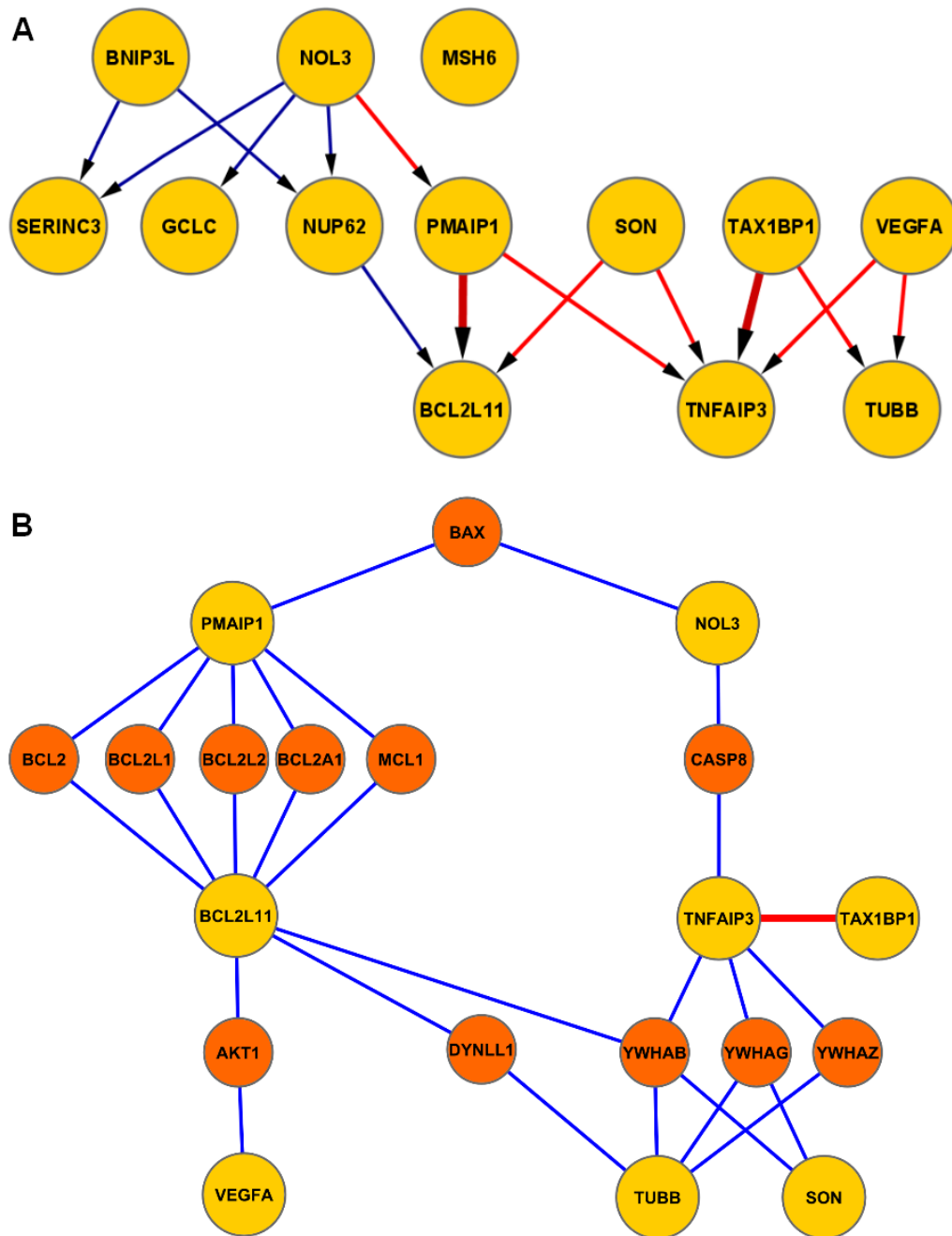


Figure 6-8 (A) Predicted subnetwork of 13 selected drug-responsive apoptotic genes: edges in red are validated interactions in literature, and edges in dark red are strong validated direct interactions. (B) A subnetwork from literature showing

evidences for validated interactions in predicted network including candidate genes (colored in yellow) with their validated interactants (in brown). Each validated edge in predicted network (red in A) can be mapped to one path in evidence network (B) between the two corresponding interacting candidate genes.

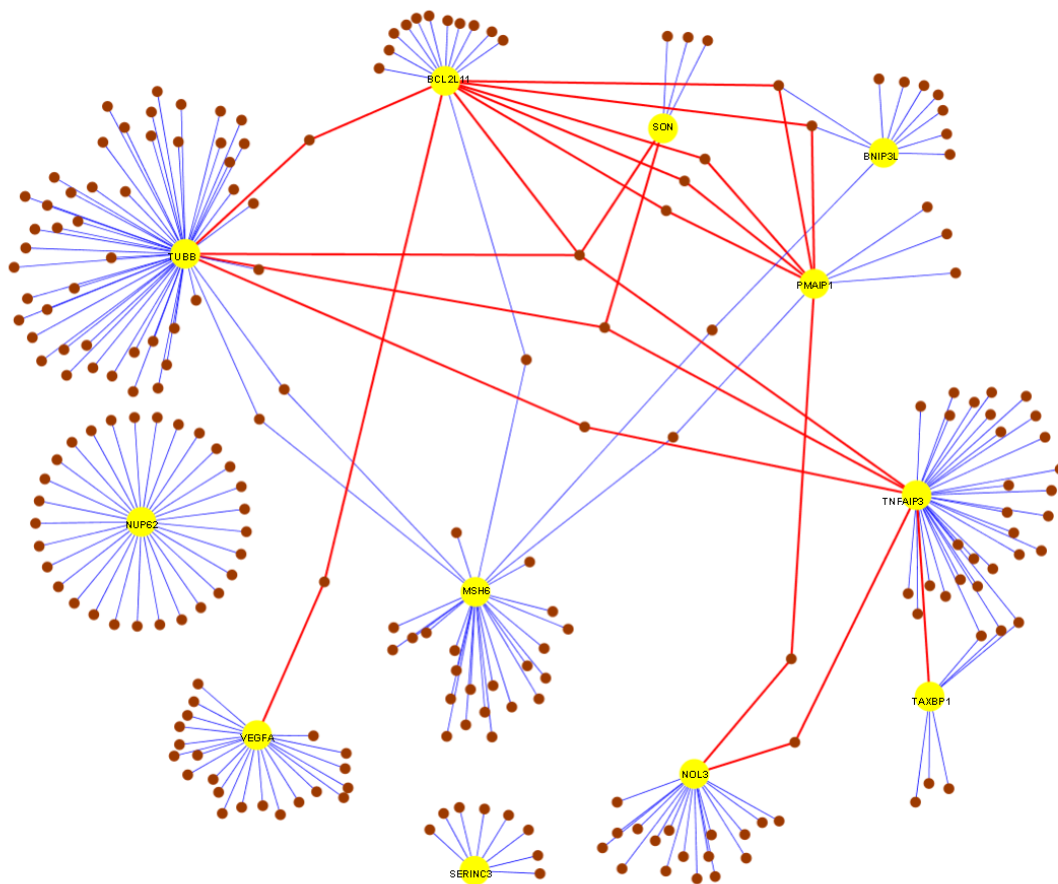


Figure 6-9 A network from literature for 13 candidate genes (colored in yellow) with their validated interactants (in brown). Edges in red are evidences for validation of interactions in predicted apoptosis network.

6.8 Discussion

We have demonstrated the value of CMAP data for studying drug-response in mammalian cancer cells. We have also validated the hypothesis that the apoptosis pathway may be a main program targeted by anti-cancer drugs. Furthermore, we have shown that CMAP data contains sufficient information about the dynamic activities of human genes for reconstructing gene-gene interactions in drug-perturbed cancer cells. We have also successfully applied a Gaussian Bayesian network framework to reconstruct a subnetwork containing validated interactions between genes with known roles in the apoptosis pathway. In addition, our network successfully predicted key players and interactions in drug-induced apoptosis, including both the intrinsic and extrinsic apoptosis pathways.

Our framework may be improved in a few ways. First, we only considered the general effects of drugs based on the assumption that cancer cells have a similar response mechanism to different drugs. However this assumption may be over-generalized, since there are some drugs to which the cells have no response. This can be clearly seen in Figure 1, which contains a heat map of signature genes across all drugs (Figure 6-1). One way to deal with this limitation may be to cluster drugs by their expression profiles or by their physical or chemical properties. A similar comparison analysis may be performed, but would take into the account the effects of different drug groups. Second, to reduce computational complexity, we limited our analysis to apoptotic genes that were differentially

expressed with a Bonferroni-corrected pvalue threshold of 0.05. This threshold might have been overly stringent and may have caused us to filter informative genes from the analysis. One way to deal with this problem might be to include more candidate genes, but this would increase complexity and computation.

We have shown that Bayesian network modeling can be a powerful tool for reconstructing biological networks from noisy high-throughput microarray data. In the Bayesian network modeling approach to network reconstruction, we have found that a linear Gaussian model for local probability distribution is able to give a more accurate description for continuous data and is also able to reduce the number of parameters when compared to discrete methods. In discrete methods, data points are separated into multiple levels, and this can result in the loss of information, especially in cases where the variable has a large range of values and has many parent variables [125-128]. However, one limitation with the linear Gaussian model is that although it works well in cases where the data fits a normal distribution and there are linear dependences between nodes and their parents, the model can easily over-fit the data if these dependencies are not met. In this study it was reasonable to apply Gaussian distribution because most candidate genes fit a normal distribution, as shown in Figure 6-2. However, a possible improvement may come from performing graphical diagnosis and doing further transformation on the data, or employing other statistical models to fit the data. An alternative approach to learning the structure of the Bayesian network is simulated annealing with Markov chain Monte Carlo (MCMC) sampling. This method may overcome the limitation of the hill-climbing method used in this study.

In hill-climbing method, the function finds the nearest optimum value. Depending on the starting point, this peak may or may not be the true optimum value. However, one limitation with MCMC sampling is that it is significantly more time-consuming than the hill-climbing method. For network comparison or scoring, other asymptotic criteria such as AIC, BIC, or DIC could be tried as well.

The two major apoptosis sub-pathways of mammalian cells are largely independent because over-expressed Bcl-2 does not protect lymphocytes from apoptosis induced by death receptor ligands. Literature has shown that in certain other cell types such as hepatocytes the two pathways intersect, because CASP8 can process the pro-apoptotic Bid into its active truncated form (tBid) and prevent catastrophic untimely cell death [121]. However, cross-talk between these two programs has been rarely studied in the context of drug-perturbations. Our computationally predicted apoptosis network might shed light on how both pathways are regulated together by identifying cross-talk interactions such as PMAIP1 and TNFAIP3, BCL2L11 and TNFAIP3 via SON.

In summary, we have extended the usage of CMAP data and reconstructed a subnetwork of drug-induced apoptosis in mammalian cancer cells using a computational statistical modeling approach. Apoptosis induction is a major theme of cancer treatment by drugs, and we confirmed that it is indeed a major drug-responsive program. Our findings have added new knowledge of how cancer cells respond to drug and provided potential specific targets in apoptosis pathway for better cancer treatment. However, cell death might not be the only

drug-induced program, so our computational framework to CMAP data could be extended to other interesting biological pathways related to cancer treatment by drugs.

Chapter 7 NetBID2 Identifies AKT1 as a Therapeutic Target to Reverse Glucocorticoid Resistance in T-ALL¹

7.1 Summary

Glucocorticoid resistance is a major driver of therapeutic failure in T-cell acute lymphoblastic leukemia (T-ALL). Here we used a systems biology approach, NetBID2, based on the reverse engineering of signaling regulatory networks, which identified the AKT1 kinase as a signaling factor driving glucocorticoid resistance in T-ALL. Indeed, activation of AKT1 in T-ALL lymphoblasts impairs glucocorticoid-induced apoptosis. Mechanistically, AKT1 directly phosphorylates the glucocorticoid receptor NR3C1 protein at position S134 and blocks glucocorticoid-induced NR3C1 translocation to the nucleus. Consistently, inhibition of AKT1 with MK-2206 increases the response of T-ALL cells to glucocorticoid therapy both in T-ALL cell lines and in primary patient samples thus effectively reversing glucocorticoid resistance in vitro and in vivo. These results warrant the clinical testing of AKT1 inhibitors and glucocorticoids, in combination, for the treatment of T-ALL.

¹ Eric Piovan from Adolfo Ferrando Lab did validation experiments. This chapter is based on our paper [148].

7.2 Clinical Significance

Glucocorticoids are central drugs in the treatment of T-ALL and glucocorticoid resistance is associated with poor outcomes in this disease. Therefore, the elucidation of molecular mechanisms contributing to glucocorticoid resistance and the identification of therapeutic targets for the treatment of glucocorticoid resistant T-ALL have become major imperatives in the field. Our identification of AKT1 as a direct inhibitor of glucocorticoid receptor function and a mediator of glucocorticoid resistance will facilitate the development of combination therapies with AKT1 inhibitors and glucocorticoids for the treatment of T-ALL. Moreover, these results further highlight the value of systems biology approaches based on reverse engineering of signaling networks to identify key modulators of drug resistance in human cancer.

Keywords: AKT1, glucocorticoid resistance, T-ALL, NetBID2, systems biology, NR3C1, phosphorylation

7.3 Introduction

Glucocorticoids (GCs) play a fundamental role in the treatment of all lymphoid tumors due to their capability to induce apoptosis in lymphoid progenitor cells (Figure 7-1) [18, 19, 149]. However, the importance of glucocorticoid therapy in lymphoid malignancies is underscored by the strong association of glucocorticoid response with prognosis in childhood acute lymphoblastic leukemia (ALL). Thus, the initial response to 7 days of glucocorticoid therapy is a strong independent

prognostic factor in this disease [150-152]. Moreover, resistance to glucocorticoids, defined as inability by lymphoblastic leukemia cells to initiate the apoptotic program in response to glucocorticoid treatment *in vitro*, is also associated with unfavorable prognosis [153, 154]. Finally, the majority of ALL patients in relapse show increased resistance to glucocorticoid therapy, suggesting glucocorticoid resistance as a potential major contributor to treatment failure [153, 155].

The transcriptional and cellular effects of glucocorticoids in leukemia cells are mediated by the glucocorticoid receptor alpha, a nuclear receptor ligand-activated transcription factor encoded by the *NR3C1* gene [156]. In its unligated state, the glucocorticoid receptor protein is located primarily in the cytoplasm as part of an inactive hetero-oligomeric complex that contains heat shock proteins and chaperones [157]. After binding to an agonist ligand, NR3C1 undergoes conformational changes, dissociates from the heat shock protein complex, partially homodimerizes, and translocates to the nucleus where it binds to DNA and activates gene expression via positive glucocorticoid response elements located in the promoters of glucocorticoid target genes [158]. In addition to its role as a transcriptional activator, the glucocorticoid receptor has also been shown to directly participate in transcriptional repression, via binding to negative glucocorticoid response elements, which mediate the assembly of cis-acting NR3C1-SMRT/NCoR repressing complexes [159] and indirectly, via interaction with other transcriptional regulators such as AP-1, NF κ B, TP53, CREBP and STAT5 [158]. Thus, activation of the glucocorticoid receptor in lymphoid cells

induces a broad transcriptional program affecting genes responsible for multiple cellular functions, including cell cycle progression, cell metabolism and the induction of apoptosis [160-163].

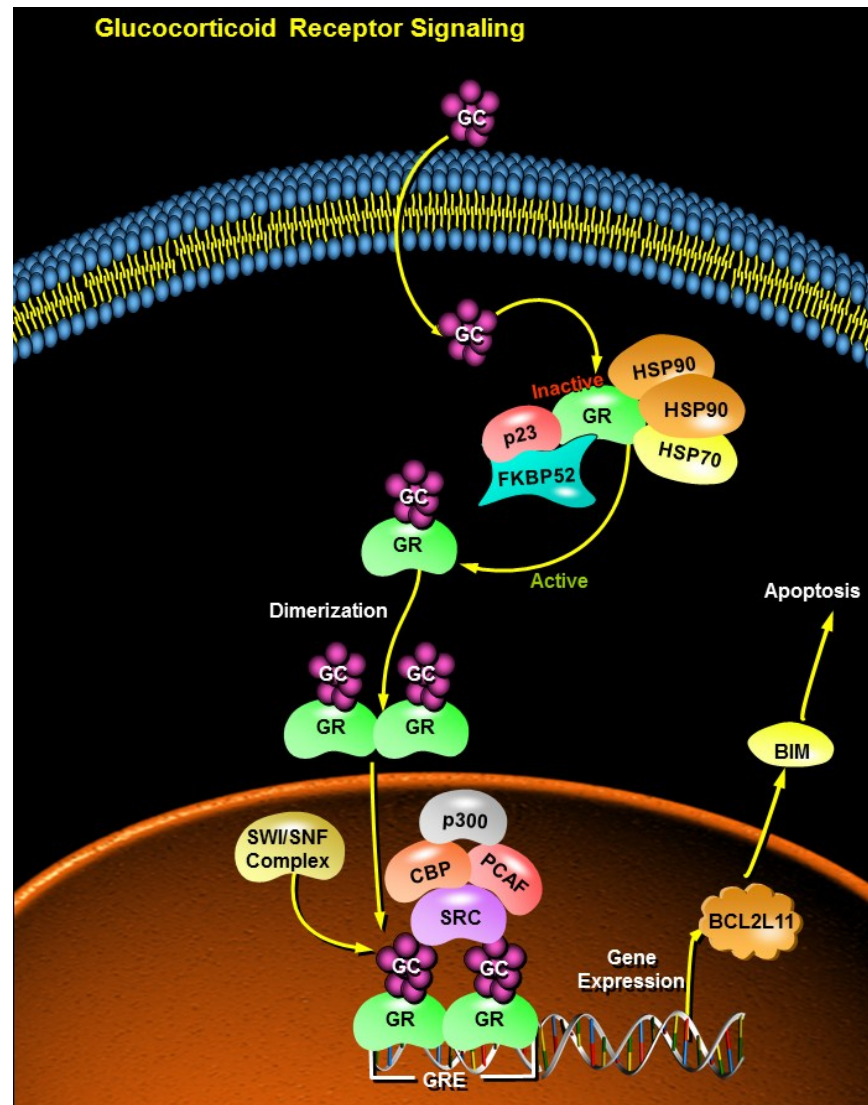


Figure 7-1 Glucocorticoid receptor signaling pathway, adapted from SABiosciences.

A number of different mechanisms have been involved in glucocorticoid resistance in ALL, including loss-of-function mutations in the glucocorticoid

receptor gene, loss of glucocorticoid receptor auto up-regulation, expression of glucocorticoid receptor splice variants, and upregulation of antiapoptotic pathways [164-173]. Overall, although multiple distinct genetic and epigenetic alterations seem to contribute to glucocorticoid resistance in ALL, complete functional loss of glucocorticoid receptor activity is rare, suggesting that strategies aimed to enhance glucocorticoid receptor expression and activity in leukemic lymphoblasts may be exploited to overcome resistance in the clinic. Moreover, even though glucocorticoid resistance is a complex phenotype, glucocorticoid resistant leukemias share a distinct gene expression signature, suggesting that common effector mechanisms may participate in blunting glucocorticoid response in resistant tumors [23]. Correspondingly, several therapeutic strategies have been proposed to overcome GC-resistance such as inhibition of MEK, HDAC, mTOR, or NOTCH1 [23, 24, 174-178]. However, due to strong toxicity of existing therapeutics [179], reversal of GC-resistance remains a clinical challenge and new therapeutic strategies are much needed.

In this chapter, we aimed to identify specific signaling proteins that directly modulate the activity of the glucocorticoid receptor and may thus be exploited for the reversal of glucocorticoid resistance. To achieve this goal, we applied NetBID2 (Figure 7-2), the systems biology framework I developed to infer disease drivers from gene expression data in couple with signaling-molecule centered network. NETBID2 is based on computationally-assembled regulatory networks from a cohort of gene expression profiles (GEPs) and Markov chain Monte Carlo (MCMC) based Bayesian modeling techniques. It extended an

existing master regulator analysis method, MARINa, which has been successful in the identification of transcription factors that are master regulators of high-grade Glioma subtypes [72] and Germinal Center formation [71], to the analysis of signaling proteins as candidate modulators of glucocorticoid resistance in T-cell acute lymphoblastic leukemia (T-ALL). This approach led to the identification of AKT1 as a master regulator of glucocorticoid resistance in T-ALL and suggested several additional potential master regulators. To validate these findings, we demonstrate that the glucocorticoid receptor, NR3C1, is a direct phosphorylation target of AKT1 at S134 and that its AKT1-mediated phosphorylation impairs glucocorticoid response via nuclear exclusion and targeted degradation of the NR3C1 protein. Consistently, and most importantly, pharmacologic inhibition of AKT1 effectively reverses glucocorticoid resistance both in T-ALL primary samples and in cell lines (i.e., *in vitro* and *in vivo*). Overall, these results show that regulatory network analysis is a valuable tool in the identification of critical modulators of therapeutic response in human cancer and identify AKT1 as an actionable therapeutic target for the reversal of glucocorticoid resistance in T-ALL.

Transcriptional Genomics: NetBID2

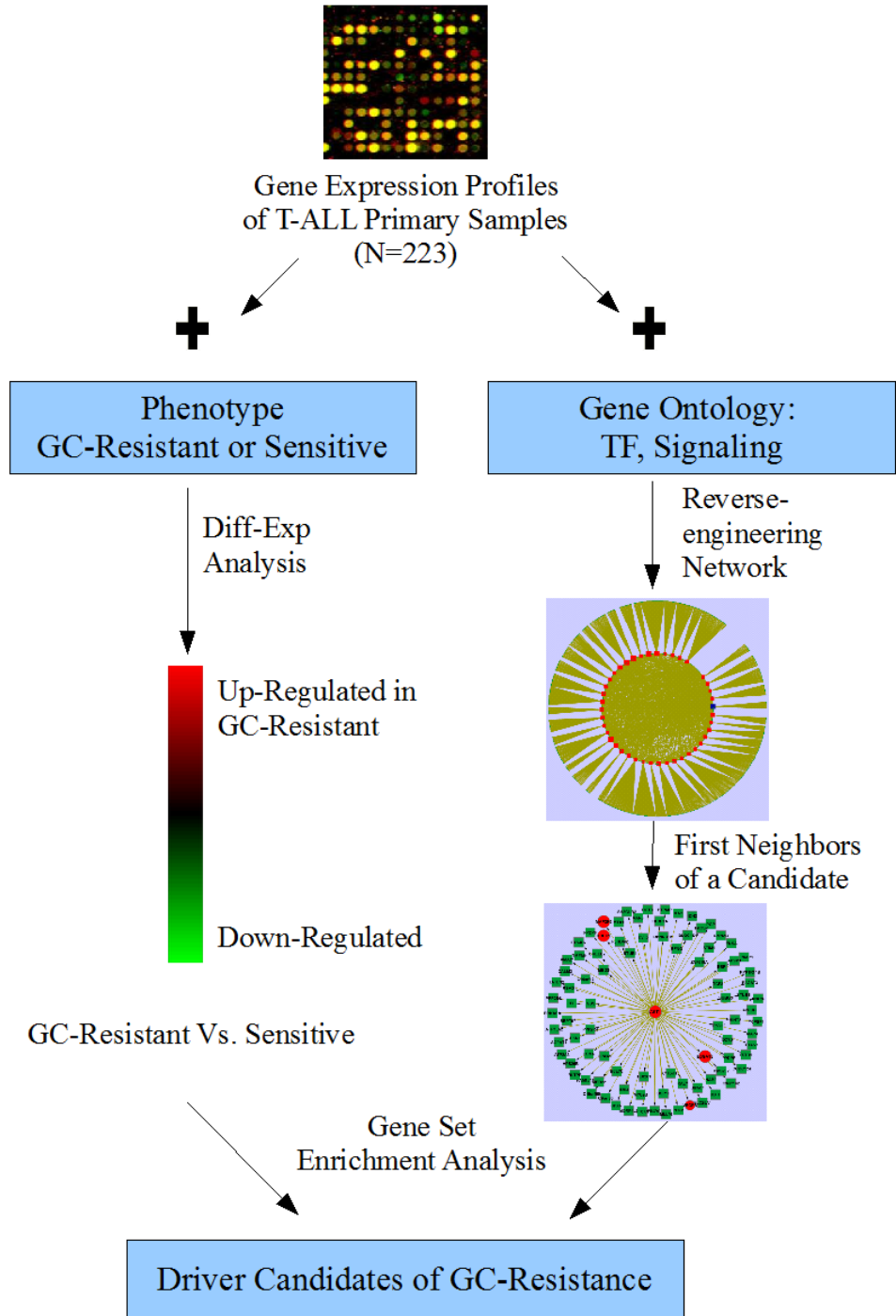


Figure 7-2 NetBID2 algorithm to identify drivers of GC-Resistance in T-ALL from gene expression profiles.

7.4 Results

7.4.1 NetBID2 with signaling network identifies AKT1 as a driver of glucocorticoid resistance in T-ALL

Reverse engineering of transcriptional regulatory or signal transduction networks has emerged as a valuable tool to identify master regulators of human phenotypes, both physiologic and pathologic [70, 180, 181]. More recently, this approach has also been successful in establishing functionally relevant, experimentally validated interactions between signaling molecules and transcription factor oncogenes [73]. Here we postulated that the gene expression signature E_{GC} , associated with glucocorticoid resistance, could be effectively used to interrogate the signaling interaction network of T-ALL to identify master regulators of resistance. Since data on signal transduction networks is too sparse and lacks context specificity, we relied instead on the fact that numerous feedback loops results in transcriptional coherence among proteins that are in the same signal transduction pathway. This suggests that candidate interactions of a signaling proteins S should be enriched among genes with a statistically significant Mutual Information with S , computed from transcriptional profiles (Figure 7-2). In addition, since these feedback loops are reasonably modeled as Markov chains, we hypothesized that using the Data Processing Inequality, a method successfully used by the ARACNe algorithm [111] to dissect direct

targets of transcription factors, could also be used to further enrich the inferred interactions in genes whose expression was more directly controlled by a signaling protein. We are cognizant that this is only an approximation. Yet, we reasoned that if the inferred regulon R_S of a signaling protein S were sufficiently enriched in genes whose expression is regulated by S , directly or indirectly, the MARINa algorithm could then be used to identify the corresponding protein's role as candidate master regulator of glucocorticoid resistance.

To define the E_{GC} signature for glucocorticoid resistant vs. glucocorticoid sensitive leukemia, we analyzed microarray data from a public series of 32 leukemias, with detailed information on glucocorticoid sensitivity [23], by Probit analysis using a Bayesian MCMC method for robust parameter estimation, see methods section. Consistent with previous reports, glucocorticoid resistant T-ALLs were characterized by a robust transcriptional signature with 53 upregulated and 73 downregulated genes in resistant patients ($P < 0.01$) (Table 7-2).

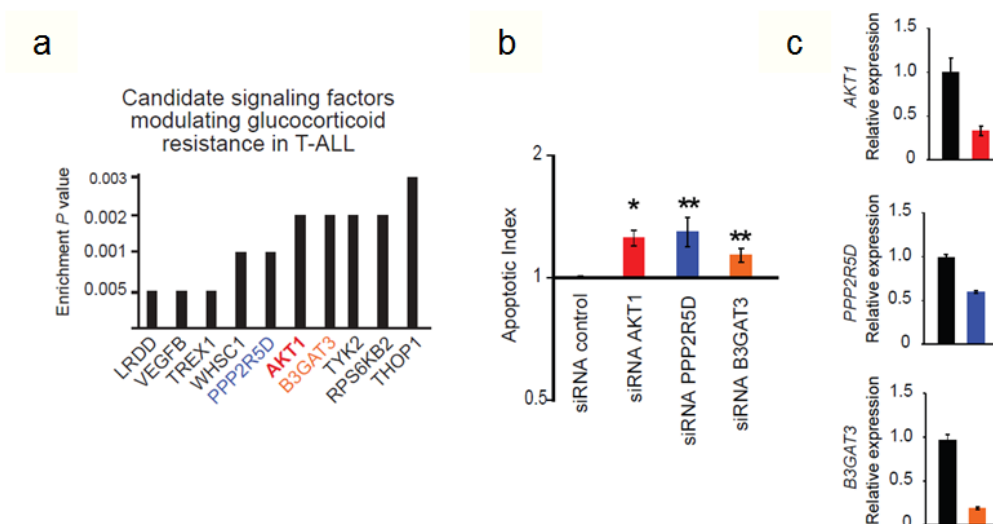


Figure 7-3 Top signaling drivers of GC-resistance inferred by NetBID2 and siRNA validation results. (a) Signaling factors associated with glucocorticoid resistance by NetBID2. (b) Apoptosis analysis in DND41 T-ALL cells electroporated with siRNA pools targeting validated candidate regulators of glucocorticoid resistance and treated with dexamethasone (1 μ M) for 48 hours. The apoptotic index indicates apoptotic cell number in gene specific siRNA dexamethasone treated samples relative to siRNA control dexamethasone treated cells. *, P <0.01; **, P <0.05. (c) Quantitative RT-PCR analysis of siRNA knockdown.

Next, to assemble a T-ALL specific signaling network to interrogate such a signature, we used gene expression profile data from a large collection of 223 T-ALL primary samples to infer groups of genes (*S*-regulons) whose expression is modulated by 2,602 proteins annotated as having signal transduction function, using the ARACNe algorithm [70, 111]. This analysis yielded a network comprising 21,033 genes and 415,424 candidate interactions.

Finally, we applied the NetBID2 algorithm to rank signaling proteins according to the enrichment of their *S*-regulon (R_S) in the glucocorticoid resistance signature, based on a two-tail Gene Set Enrichment Analysis (GSEA) [97]. Given that signaling pathways can trigger either positive or negative transcriptional feedback loops we considered that if *S* activation induced glucocorticoid resistance, then R_S genes could be enriched either among overexpressed or underexpressed genes in the glucocorticoid resistance signature. All signaling proteins were then ranked by their two-tail GSEA statistics, using the Normalized Enrichment Score, NES, and associated *P*-value. This analysis identified 42 signaling drivers of GC-resistance ($P < 0.01$, set size > 50 , involved in at least known pathway, Table 7-1, Figure 7-25). We selected top 9 signaling factor-associated gene sets with highly significant enrichment scores ($P < 0.0025$) for validation. SiRNA mediated silencing of each of these candidate glucocorticoid resistance modulators validated that inhibition of 3 out of 9 (33%) of these predicted genes can increase the response of T-ALL lymphoblasts to glucocorticoids. Thus, knockdown of PPP2R5D, a protein phosphatase 2A regulatory B subunit; the B3GAT3 glucuronosyl transferase 1; and AKT1, a central mediator in phosphatidylinositol 3-kinase (PI3K) signaling, can all enhance glucocorticoid induced apoptosis in DND41 T-ALL cells (Figure 7-3). The prominent role of the PI3K-AKT signaling pathway in the pathogenesis of T-ALL [182, 183], and the development of clinically relevant PI3K-AKT specific inhibitors for the treatment of human cancer, prompted us to analyze the mechanistic role of AKT1 in the control of glucocorticoid resistance in T-ALL.

More interestingly, in the predicted network of AKT1, three AKT1-associated genes (VEGFB, TREX1, B3GAT3) are among the top 9 signaling proteins we selected and validated. B3GAT3 is also validated by siRNA (Figure 7-3) to sensitize GC-resistant cells upon inhibition. Out of 30 transcription factors or signaling molecules (92 genes in total) connected to AKT1 in the predicted network, 15 are also inferred as drivers of GC-resistance (Figure 7-4). This may suggest that AKT pathway is highly involved in inducement of GC-resistance in T-ALL and may provide a therapeutic avenue to reverse the resistance by inhibiting AKT pathway. This also gives us more confidence and interest to follow up AKT1 to test it out and to identify the mechanism of AKT1 causing GC-resistance.

Moreover, NetBID2 also identifies AKT2 as a signal driver of GC-resistance in T-ALL (Figure 7-6), which is another important member of PI3K/AKT pathway and shares similar functions with AKT1 in many biological processes. Again this makes it more interesting to work on AKT pathway.

One point we want to highlight is that NetBID2 identifies AKT1 as a driver of GC-resistance (Figure 7-5); however, AKT1 is not a signature gene in GC-resistant samples by looking at its own expression (Figure 7-7). It's not differentially expressed in GC-resistant and sensitive samples and also shows slightly up-regulation in GC-sensitive samples. This again confirms the power of NetBID2 to identify hidden underlying drivers that classical signature method would fail to find.

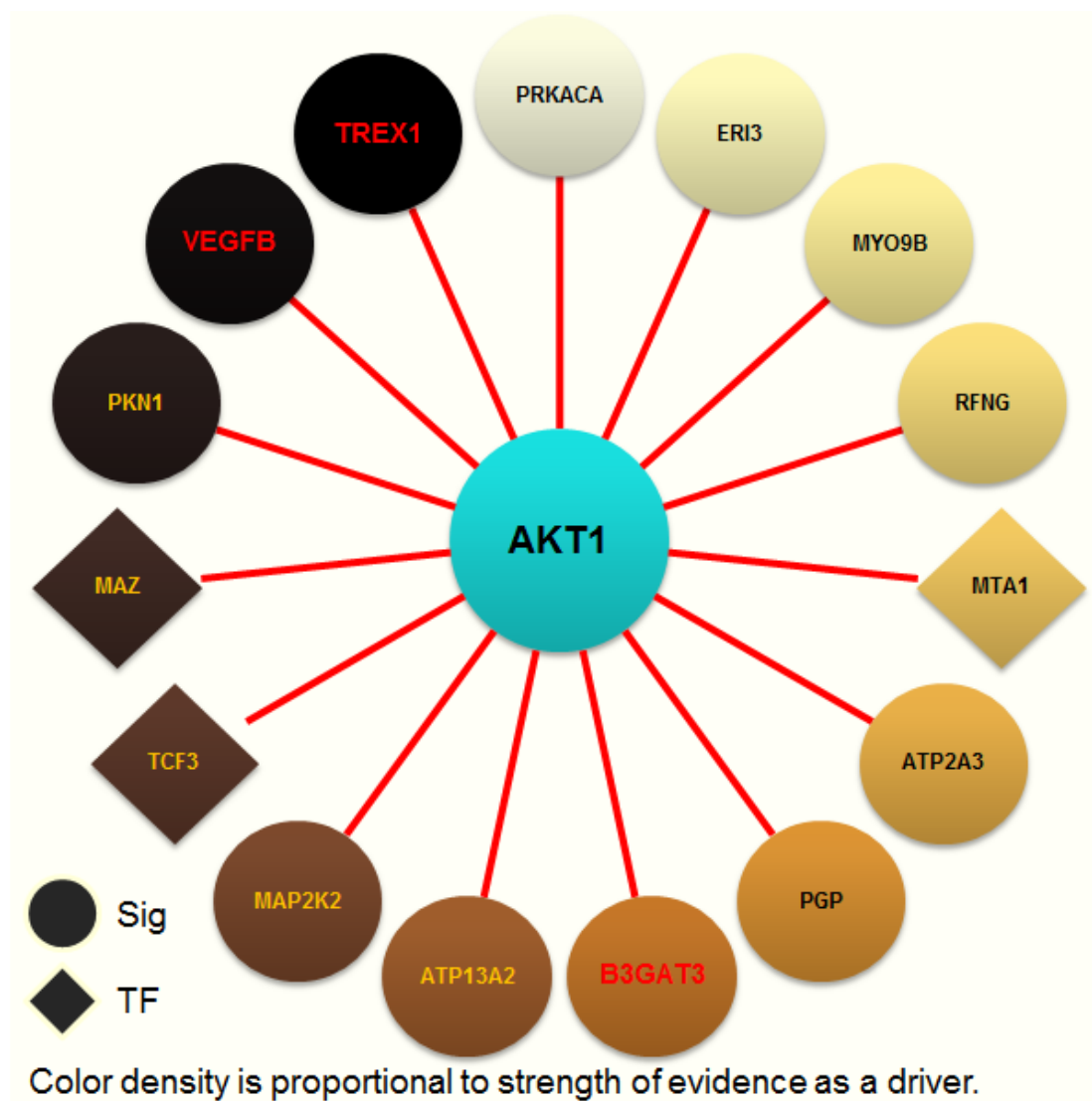


Figure 7-4 AKT1-subnetwork predicted by ARACNe. Out of 30 (92 genes in total) TFs (diamond shape) or signal molecules (circle shape) that are predicted to connect with AKT1, 15 as shown are also inferred as drivers of GC-resistance. The strength of evidence (p-value) as a driver is color coded. Three signaling proteins in red are among top 9 proteins selected for validation. B3GAT3 is also confirmed to reverse GC-resistance by siRNA.

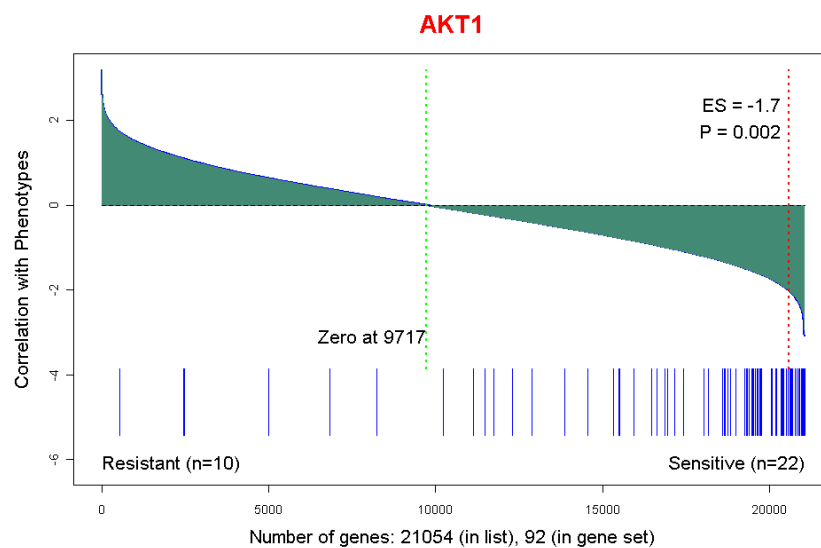


Figure 7-5 NetBID2 identifies AKT1 as a driver of GC-resistance in T-ALL.

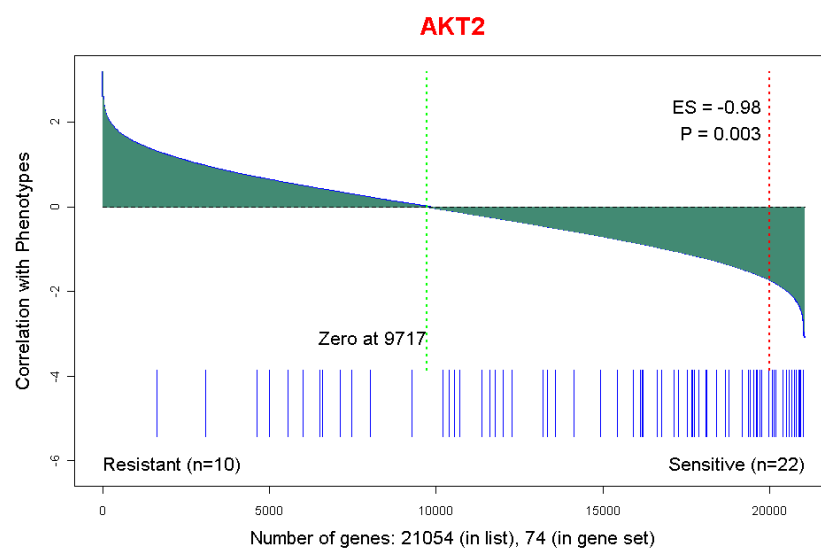


Figure 7-6 NetBID2 identifies AKT2 as a driver of GC-Resistance in T-ALL.

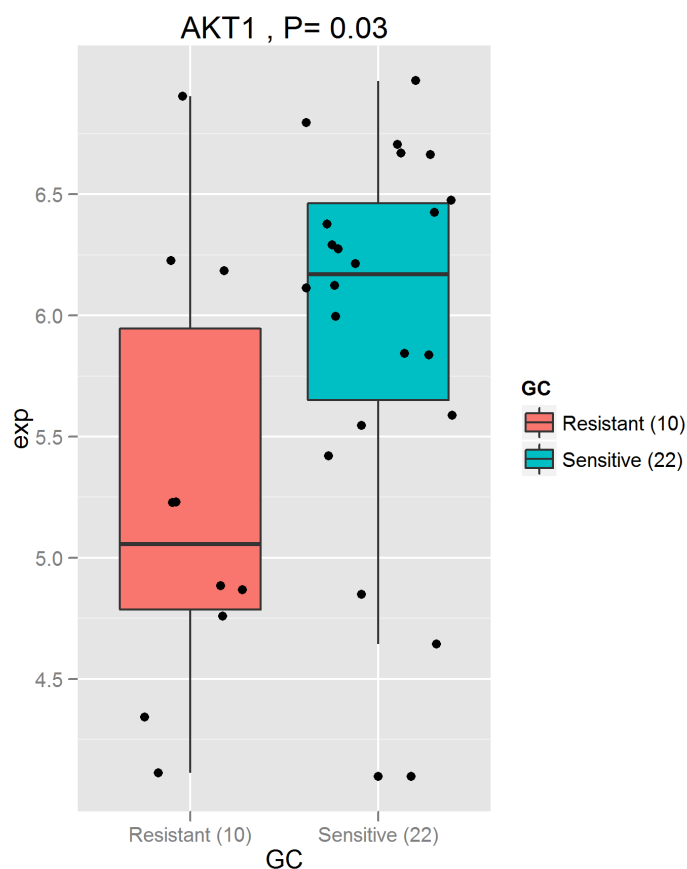


Figure 7-7 mRNA expression of AKT1 in GC-resistant and GC-sensitive primary samples. AKT1 is slightly over-expressed in sensitive samples.

7.4.2 Constitutive activation of AKT1 impairs glucocorticoid response in T-ALL

Mutations and deletions in the *PTEN* tumor suppressor gene result in constitutive activation of AKT1 in T-ALL [182, 183]. Consistently, *PTEN* inactivation in DND41 cells by shRNA knockdown resulted in drastic reduction of PTEN protein levels and increased phosphorylation of AKT1 compared to control cells infected with shRNAs targeting the Luciferase gene (Figure 7-8-a). Treatment of *PTEN*

knockdown and control DND41 cells with dexamethasone showed that loss of PTEN and consequent AKT1 activation results in blunted induction of glucocorticoid induced apoptosis in T-ALL (Figure 7-8-b,c).

The glucocorticoid receptor (*NR3C1*) functions as a ligand activated transcription factor [184]. Expression analysis of *TSC22D3*, a glucocorticoid target gene associated with inhibition of cell proliferation; *BCL2L11*, which encodes BIM a proapoptotic BH3-only factor; and the glucocorticoid receptor *NR3C1* gene itself, showed a significant reduction in activation of these glucocorticoid induced transcripts in DND41 PTEN knockdown cells treated with dexamethasone compared with controls (Figure 7-8-d). In addition, AKT1 siRNA knockdown induced a significant enhancement in the upregulation of glucocorticoid response transcripts upon dexamethasone treatment (Figure 7-9). Consistently, expression of an activated myristoylated form of AKT1 (MYR-AKT1) diminished the capacity of the glucocorticoid receptor to activate a luciferase reporter construct under the control of a synthetic glucocorticoid response element (Figure 7-8-e), and blunted the response of the physiologic AF11-AF12 glucocorticoid response element responsible for the autoupregulation of the *NR3C1* hematopoietic specific promoter [185] (Figure 7-8-f). Overall these results suggest that AKT1 could promote glucocorticoid resistance via inhibition of the glucocorticoid receptor.

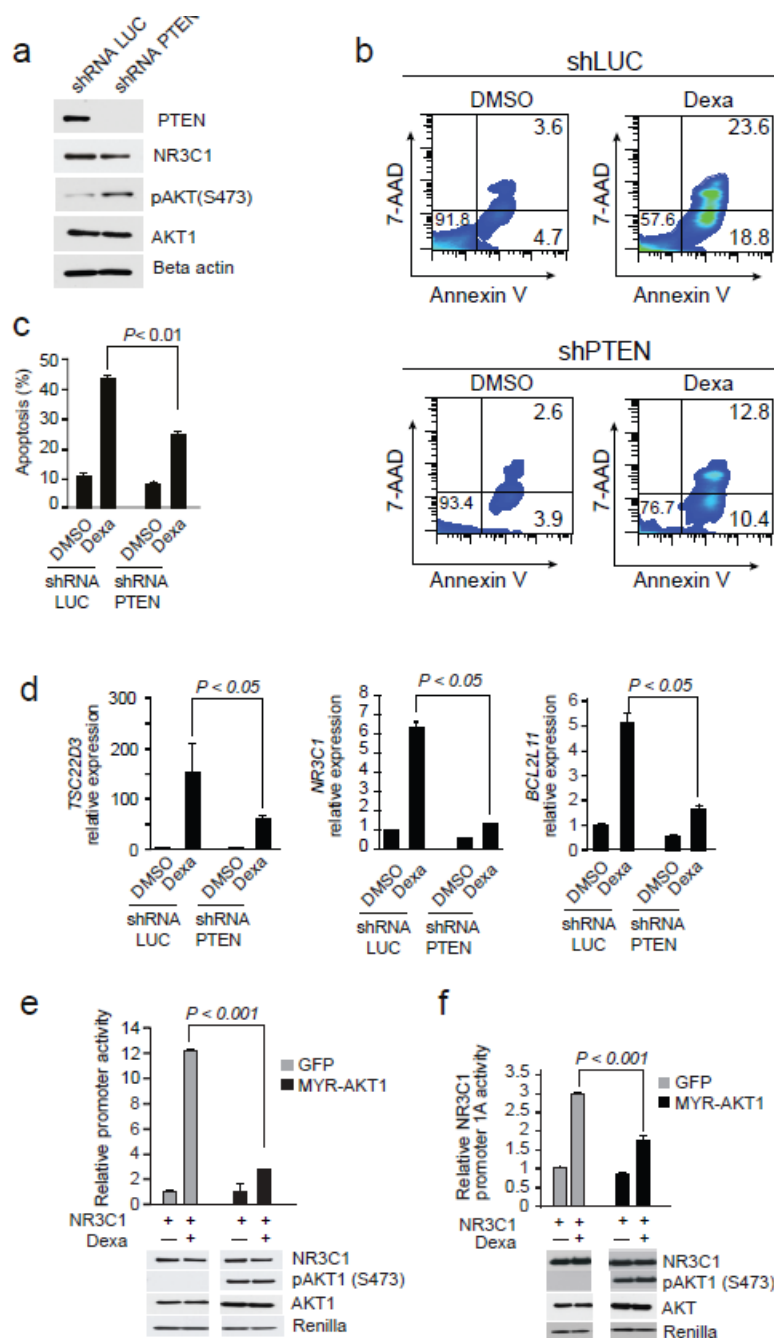


Figure 7-8 Activation of the PI3K-AKT signaling pathway via PTEN inactivation induces glucocorticoid resistance in T-ALL and blunts glucocorticoid-induced gene expression. (a) Western blot analysis of PTEN expression and AKT1 activation in DND41 T-ALL cells expressing a shRNA targeting the PTEN tumor suppressor (shRNA PTEN) compared to control cells expressing a hairpin against luciferase (shRNA LUC). (b,c) Representative plots (b) and quantification

(c) of glucocorticoid-induced apoptosis in control and PTEN knockdown DND41 cells treated with dexamethasone (1 μ M) for 48 hours. Percentages of viable (lower left quadrant), apoptotic (lower right quadrant) and dead (upper right quadrant) are indicated. (d) RT-PCR analysis of glucocorticoid response gene induction in control and PTEN knockdown DND41 cells treated with dexamethasone. (e,f) Luciferase reporter analysis of dexamethasone-induced glucocorticoid receptor transactivation in U2OS cells expressing MYR-AKT1 compared with GFP only expressing controls using a synthetic glucocorticoid response element reporter (e) and the glucocorticoid receptor promoter 1A FP11-FP12 regulatory sequence (f).

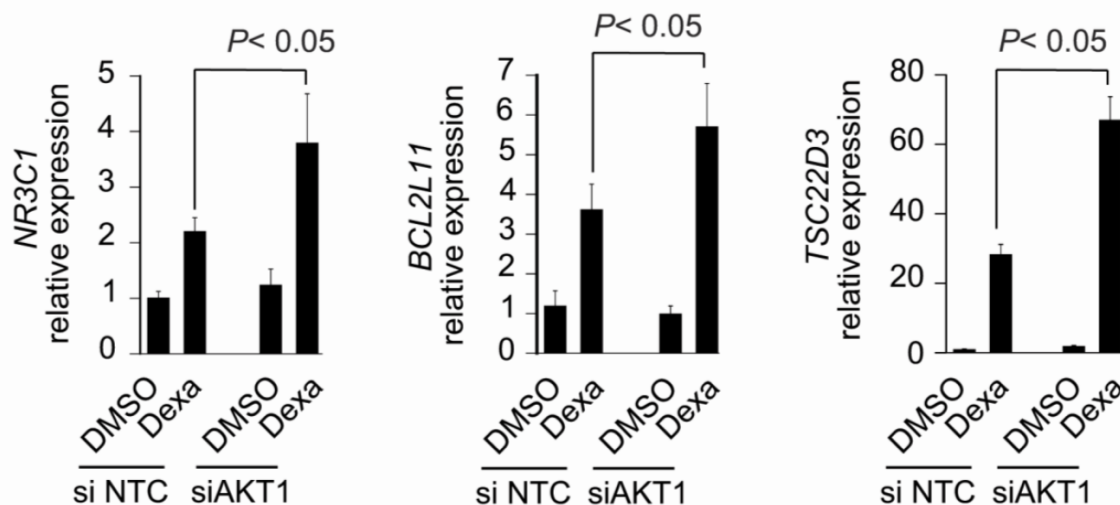


Figure 7-9 Inactivation of AKT by siRNA facilitates glucocorticoid-induced gene expression. RT-PCR analysis of glucocorticoid response gene induction in control and AKT1 knockdown DND41 cells treated with dexamethasone

7.4.3 Phosphorylation of the glucocorticoid receptor (GCR) by AKT1

Activation of gene expression by glucocorticoids is a multistep process that requires effective release of the glucocorticoid receptor from heat shock protein complexes, effective translocation to the nucleus and formation of a multiprotein transcriptional complex in the promoter of glucocorticoid target genes. To test if AKT1 could directly interact and inhibit the glucocorticoid receptor protein we transfected 293T cells with plasmid constructs driving the expression of Flag-tagged AKT1 and HA-tagged NR3C1 and isolated glucocorticoid receptor-containing protein complexes via immunoprecipitation using an anti-HA antibody. Western blot analysis demonstrated the presence of FLAG-AKT1 in HA-NR3C1 immunoprecipitates suggesting that AKT1 can interact with NR3C1 *in vivo* (Figure 7-10-a). Reciprocal immunoprecipitation experiments, confirmed the association between Flag-AKT1 and HA-NR3C1 (Figure 7-10-b). Moreover, immunoprecipitation of NR3C1 protein complexes from the T-ALL cell lines DND41 and CCRF-CEM, demonstrated that endogenous NR3C1 and AKT1 can interact in T-ALL lymphoblast cells (Figure 7-10-c, Figure 7-11). Finally, glutathione-S-transferase (GST)-pulldown assays showed that recombinant GST-NR3C1 fusion protein can directly interact with His-tagged AKT1 (Figure 7-10-d).

AKT1 kinase target proteins are typically phosphorylated by AKT at RXRXXS/T motifs [186-188]. Phospho-AKT motif scanning analysis of NR3C1 revealed a potential AKT phosphorylation motif ¹³¹RSTS¹³⁴ (Figure 7-10-e), suggesting that the glucocorticoid receptor could be an AKT1 substrate phosphorylated at serine

134. To test this possibility, we expressed HA-tagged wild type NR3C1 (HA-NR3C1) or an HA-tagged form of the glucocorticoid receptor with a serine to alanine substitution at position 134 (HA-NR3C1 S134A) in cells infected with retroviruses expressing MYR-AKT1. Protein immunoprecipitation of NR3C1 with an antibody against HA and subsequent Western blot analysis with an antibody recognizing the phospho-RXXS/T AKT phosphorylation motif showed the presence of a HA-NR3C1 phospho-AKT band in cells expressing the wild type glucocorticoid receptor, but not in cells expressing the HA-NR3C1 S134A mutant (Figure 7-10-f). Next, we performed *in vitro* kinase assays in which we analyzed the capacity of the AKT1 kinase to phosphorylate the wild type or S134A glucocorticoid receptor proteins. This assay demonstrated that AKT1 can effectively phosphorylate recombinant wild type NR3C1 protein *in vitro*, but not the serine 134 to alanine NR3C1 mutant protein (Figure 7-10-g). Importantly, this effect was not mediated by impaired interaction between AKT1 and NR3C1S134A as GST-pulldown experiments showed that GST-NR3C1 S134A mutant protein can effectively interact with AKT1 *in vitro* (Figure 7-12). Moreover, mass spectrometry analysis of HA-NR3C1 protein isolated from MYR-AKT1 expressing cells demonstrated the presence of serine phosphorylation at position 134 of the glucocorticoid receptor by mass spectrometry (Figure 7-10-h,i). Mass spectrometry of the digested peptides by nanoLC-ESI-MS/MS verified the presence of NR3C1 phosphorylation at S134 [ratio non-phosphorylated peptide: phosphorylated peptide (non-P:P)= 1.5:1] in addition to other previously characterized NR3C1 phosphosites including T8 (non-P:P= 10:1), S45 (non-P:P=

20:1), S203 (non-P:P= 1:1) and S267 (non-P:P= 25:1). Overall, these results demonstrate that the glucocorticoid receptor is a direct phosphorylation target of AKT1.

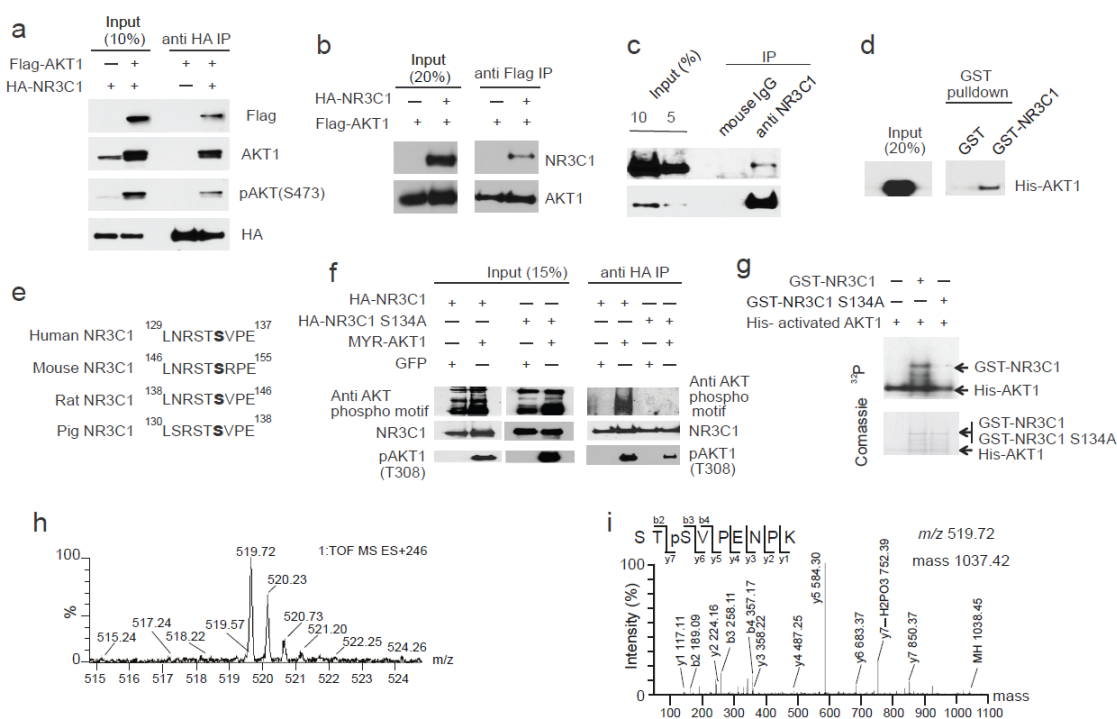


Figure 7-10 AKT1 interacts with and directly phosphorylates the glucocorticoid receptor protein in position S134. (a) Western blot analysis of AKT1 after glucocorticoid receptor NR3C1 immunoprecipitation in 293T cells expressing Flag-tagged AKT1 and HA-tagged NR3C1. (b) Western blot analysis of glucocorticoid receptor NR3C1 protein after AKT1 immunoprecipitation in 293T cells expressing Flag-tagged AKT1 and HA-tagged NR3C1. (c) Western blot analysis of AKT1 after NR3C1 protein immunoprecipitation in DND-41 T-ALL cells. (d) Analysis of AKT1-NR3C1 interaction via AKT1 detection via Western blot analysis of protein complexes recovered after NR3C1-GST pull down of recombinant His-tagged AKT1. (e) Partial alignment of the glucocorticoid receptor protein sequence flanking S134. (f) Western blot analysis of NR3C1 phosphorylation with an anti AKT phospho-motif specific antibody in NR3C1

protein immunoprecipitates from U2OS cells expressing MYR-AKT1 together with HA-tagged wild type glucocorticoid receptor (HA-NR3C1) or an HA-tagged glucocorticoid receptor protein harboring a serine 134 to alanine substitution (HA-NR3C1 S134A). (g) In vitro kinase analysis of AKT1 phosphorylation of recombinant wild type NR3C1 (GST-NR3C1) and NR3C1 S134A mutant (GST-NR3C1 S134A) protein. Top panel shows P^{32} autoradiography after SDS-PAGE. The corresponding protein loading for each reaction is shown in the Coomassie blue staining micrograph at the bottom. (h) ESI-MS/MS spectrum of monophosphorylated peptide STpS134VPENPK (S-132 to K-140) obtained after trypsin digestion of NR3C1 isolated from cells expressing constitutively active AKT1. (i) Collision induced dissociation of the molecular ion, $[M+2H]^{2+}$ at m/z 519.72 ($M = 1037.42$ Da) corresponding to S134. Characteristic b- and y-fragment ions including y7 which contains pSer and features the loss of 98 Da (elimination of phosphoric acid) are shown.

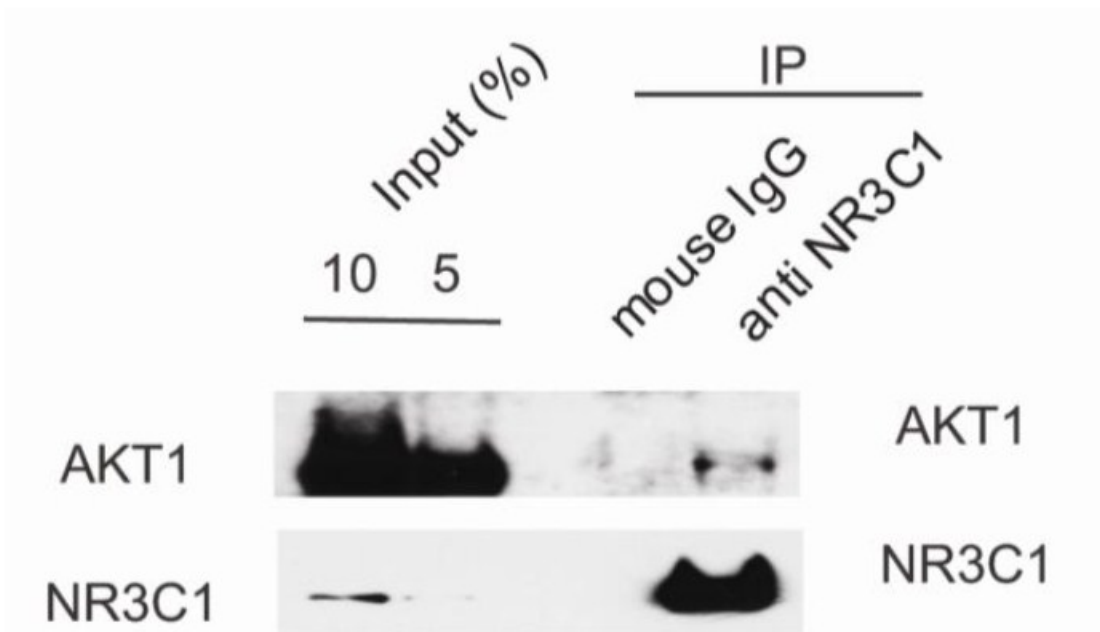


Figure 7-11 AKT1 directly interacts with the glucocorticoid receptor in T-ALL cells. Western blot analysis of AKT1 after NR3C1 protein immunoprecipitation in CCRF-CEM cells.

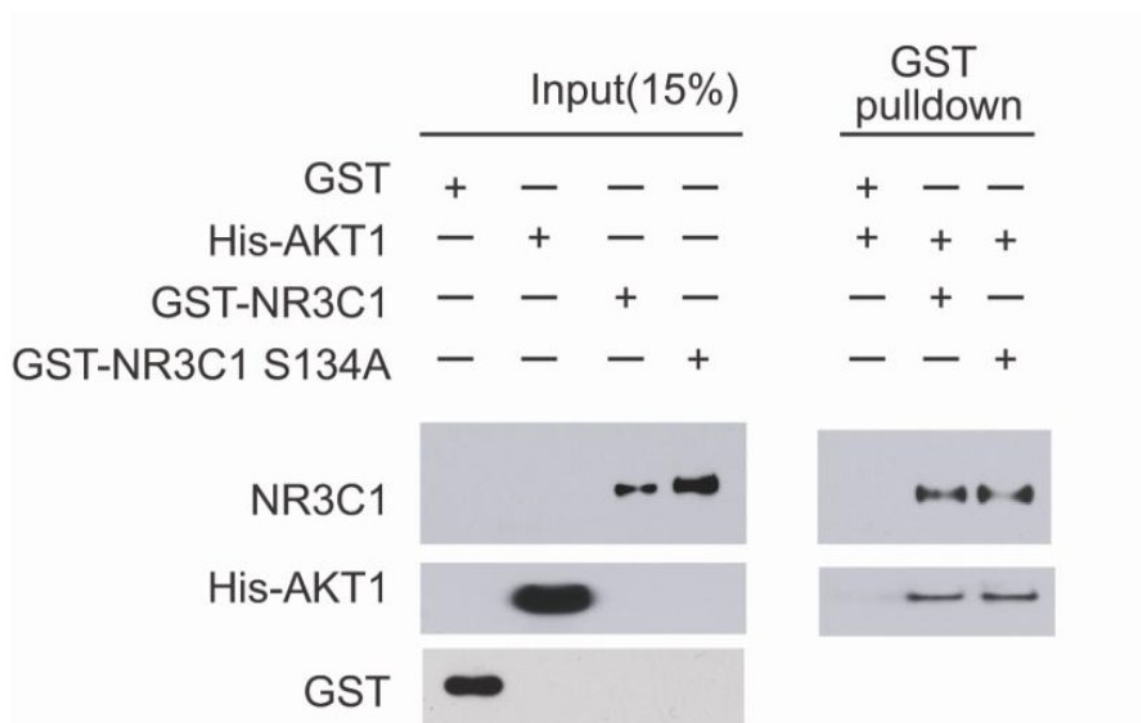


Figure 7-12 AKT1 can directly interact with both wild type and mutant S134A glucocorticoid receptor. Analysis of AKT1-NR3C1 interaction via AKT1 detection via Western blot analysis of protein complexes recovered after wild type (NR3C1-GST) or mutant (NR3C1 S134A-GST) glucocorticoid receptor GST pull down with recombinant His-tagged AKT1.

7.4.4 AKT signaling inhibits NR3C1 nuclear translocation following glucocorticoid treatment

After establishing the interaction and phosphorylation of the glucocorticoid receptor by AKT1 we aimed to elucidate the relevance of the NR3C1 S134 phosphorylation for glucocorticoid receptor function. Glucocorticoid induced cytoplasmic-nuclear shuttling is strictly required for glucocorticoid receptor

activity. U2OS cells, which express undetectable levels of endogenous NR3C1 (Figure 7-14), showed cytoplasmic localization of retrovirally expressed HA-tagged glucocorticoid receptor protein, which was completely relocalized to the nucleus upon dexamethasone treatment (Figure 7-13-a). Notably, expression of MYR-AKT1 in these cells resulted in impaired nuclear relocalization of NR3C1 after dexamethasone treatment (Figure 7-13-b). In addition, and in contrast with wild type glucocorticoid receptor, the NR3C1 S134A mutant protein showed increased nuclear localization in basal conditions and effective nuclear relocalization upon dexamethasone treatment (Figure 7-13-c), even upon expression of MYR-AKT1 (Figure 7-13-d). Next we analyzed the capacity of MK2206 a highly potent and selective AKT inhibitor [189], to modulate glucocorticoid induced translocation of NR3C1 to the nucleus in T-ALL cells. CCRF-CEM and MOLT3, two PTEN null T-ALL cell lines expressing high levels of AKT activation (Figure 7-13-e, Figure 7-15) showed cytoplasmic localization NR3C1 in basal conditions, which was only partially relocalized to the nucleus upon dexamethasone treatment (Figure 7-13-e, Figure 7-15). Inhibition of AKT with MK2206 effectively enhanced glucocorticoid-induced translocation of the NR3C1 protein to the nucleus in these cells (Figure 7-13-e, Figure 7-15).

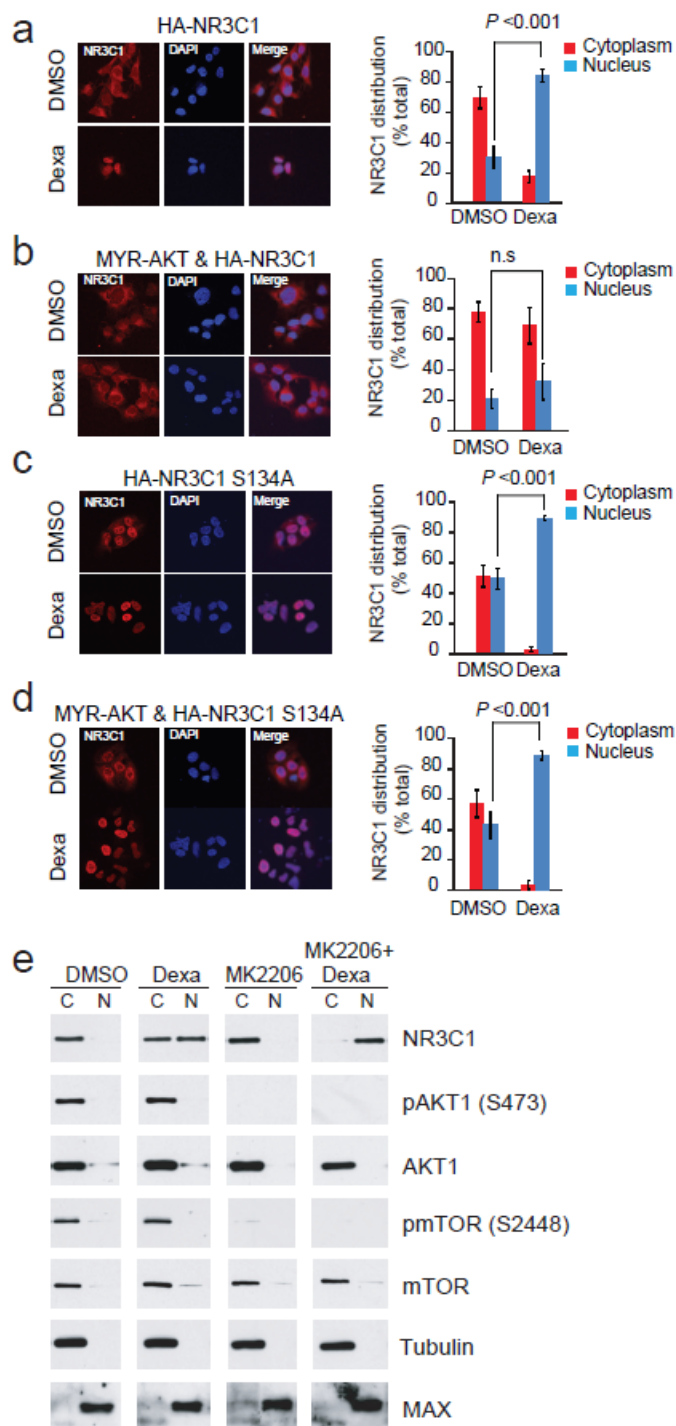


Figure 7-13 AKT1-mediated S134 phosphorylation of the NR3C1 protein impairs dexamethasone-induced glucocorticoid receptor nuclear translocation. (a) Confocal microscopy analysis and quantitation of the cellular distribution of NR3C1 cellular localization in U2OS cells expressing HA-NRC31 in basal

conditions (DMSO) and after dexamethasone (Dexa) stimulation. (b) NR3C1 cellular localization in U2OS cells expressing HA-NRC31 and MYR-AKT1 in basal conditions and after dexamethasone stimulation. (c) Cellular localization of NR3C1 in U2OS cells expressing the HA-NRC31 S134A mutant in basal conditions and after dexamethasone stimulation. (d) Cellular localization of the HA-NRC31 S134A protein in U2OS cells expressing MYR-AKT1 in basal conditions and after dexamethasone stimulation. (e) Cellular localization analysis of NR3C1 via nuclear and cytoplasmic cell fractionation and analysis of AKT1 signaling in cell lysates from CCRF-CEM T-ALL cells treated with vehicle only (DMSO), dexamethasone (Dexa), the MK2206 AKT inhibitor and MK2206 plus dexamethasone. Tubulin and MAX proteins are shown as controls for cytosolic and nuclear fractions, respectively. C: cytoplasmic fraction; N: nuclear fraction.

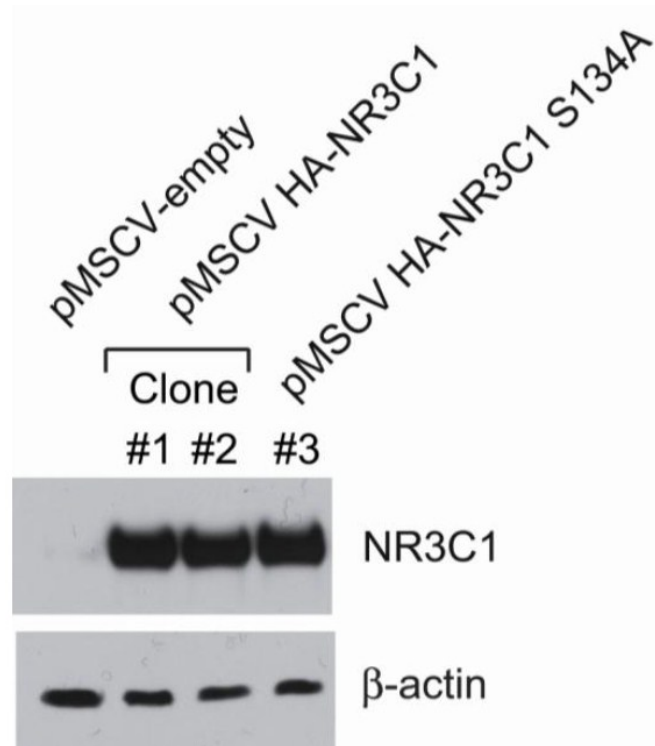


Figure 7-14 U2OS cells do not express detectable levels of endogenous NR3C1. Western blot analysis of NR3C1 expression in U2OS cells expressing pMSCV empty vector, pMSCV-HA NR3C1 or pMSCV-HA NR3C1 S134A.

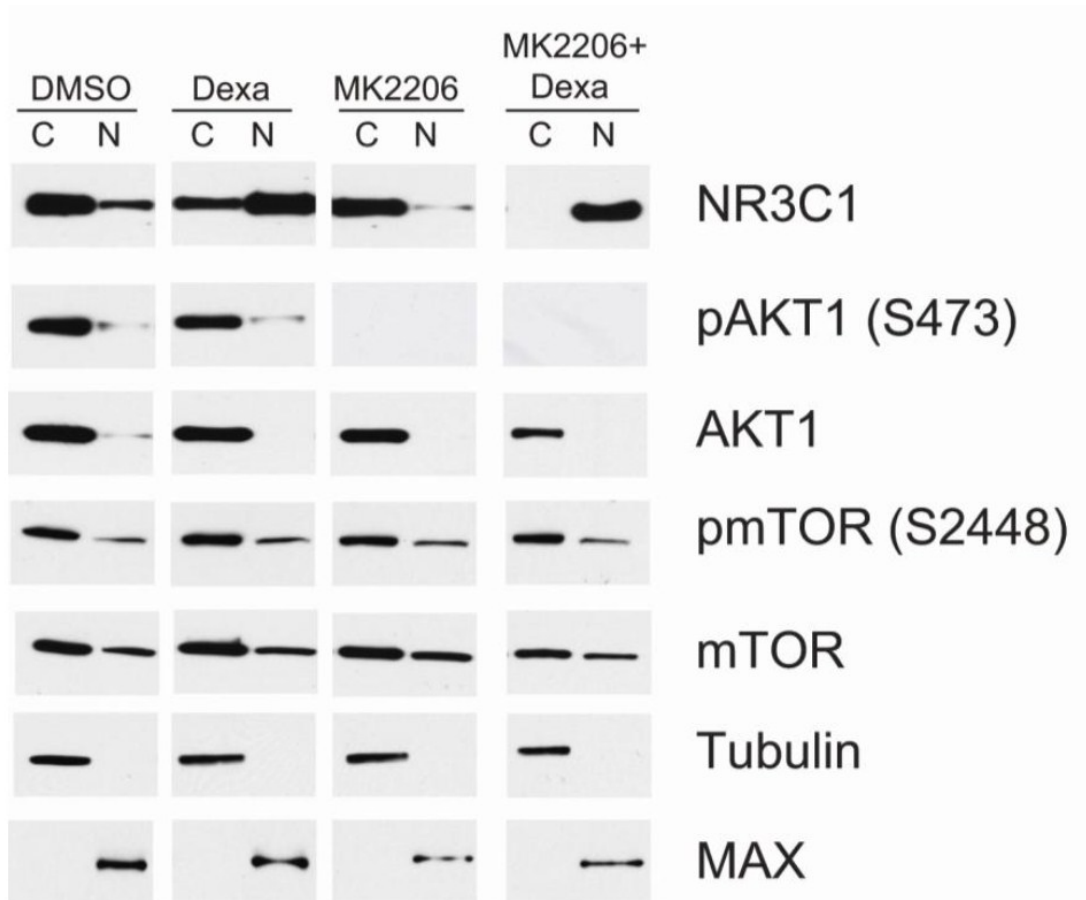


Figure 7-15 AKT1-mediated phosphorylation of the NR3C1 protein impairs dexamethasone-induced glucocorticoid receptor nuclear translocation in T-ALL cells Cellular localization analysis of NR3C1 via nuclear and cytoplasmic cell fractionation and analysis of AKT1 signaling in cell lysates from MOLT-3 T-ALL cells treated with vehicle only (DMSO), dexamethasone (Dexa), the MK2206 AKT inhibitor and MK2206 plus dexamethasone. Tubulin and MAX proteins are shown as controls for cytosolic and nuclear fractions, respectively. C: cytoplasmic fraction; N: nuclear fraction.

7.4.5 Pharmacologic inhibition of AKT reverses glucocorticoid resistance *in vitro* and *in vivo*

Next we analyzed if AKT inhibition with MK2006 could broadly enhance the antileukemic effects of glucocorticoids and reverse glucocorticoid resistance in T-ALL. Treatment of DND41 T-ALL cells with MK2206 effectively suppressed AKT1 signaling (Figure 7-16) and showed a synergistic antileukemic effect in combination with dexamethasone [MK-2206 and dexamethasone Combination Index (CI)= 0.48] (Figure 7-16). Consistently, treatment of CCRF-CEM cells with increasing doses of dexamethasone in the presence or absence of MK2206 showed effective reversal of glucocorticoid resistance upon AKT inhibition (Figure 7-17-a). Similar results were obtained in the MOLT3 cell line (Figure 7-18). Next we analyzed the effects of MK2206 and glucocorticoid *in vivo* in a xenograft model of glucocorticoid-resistant T-ALL. CCRF-CEM cells expressing the luciferase gene were injected intravenously in immunodeficient NOD SCID mice and tumor engraftment was assessed by *in vivo* bioimaging at day 18. Animals harboring homogeneous tumor burdens were treated with vehicle only (DMSO), MK2206, dexamethasone or MK2206 plus dexamethasone for 3 days. In this experiment, animals treated with dexamethasone or MK2206 showed progressive tumor growth similar to that observed in vehicle-treated controls, while mice treated with MK2206 plus dexamethasone had significant antitumor responses (Figure 7-17-b; $P < 0.05$).

Next, we evaluated the response to the combination treatment MK2206 plus dexamethasone in primary T-ALL lymphoblasts. Towards this goal we

established viable *in vitro* cultures of T-ALL leukemia samples supported by bone marrow MS5 stroma cells expressing the Delta like 1 NOTCH1 ligand [190]. Treatment of T-ALL primary leukemia cultures with MK2206 plus dexamethasone in combination showed significantly increased antileukemic effects compared with treatment with dexamethasone or MK2206 alone in 8/10 primary T-ALLs analyzed (Figure 7-22-a, b and Figure 7-19).

To further test the efficacy of this treatment combination *in vivo* we established leukemia xenografts in Rag2/gamma (c) double knockout mice using two independent primary T-ALL samples infected with lentiviruses expressing the luciferase gene. Animals harboring homogeneous tumor burdens by *in vivo* bioimaging were treated with vehicle only (DMSO), MK2206, dexamethasone or MK2206 plus dexamethasone. In this experiment, mice treated with dexamethasone or MK2206 showed progressive tumor growth similar to that observed in vehicle-treated controls, while mice treated with MK2206 plus dexamethasone showed significant antitumor responses (Figure 7-22-c, d and Figure 7-20, Figure 7-21).

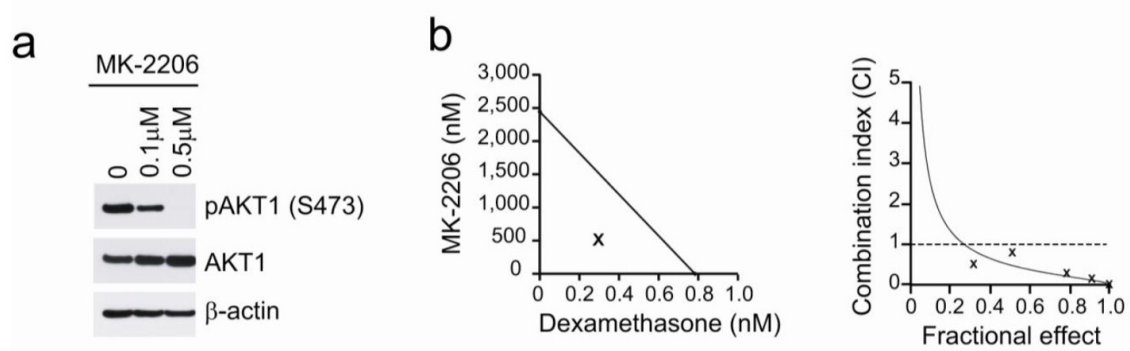


Figure 7-16 Pharmacological inhibition of AKT synergizes with dexamethasone to increase the antileukemic effects of glucocorticoids in DND-41 T-ALL cells. (a) Western blot analysis of AKT1 activation in DND41 T-ALL cells treated with the MK2206 AKT inhibitor. (b) Isobologram representation of cell viability results and Combination Index analysis of DND41 cells treated with dexamethasone and MK-2206 in combination.

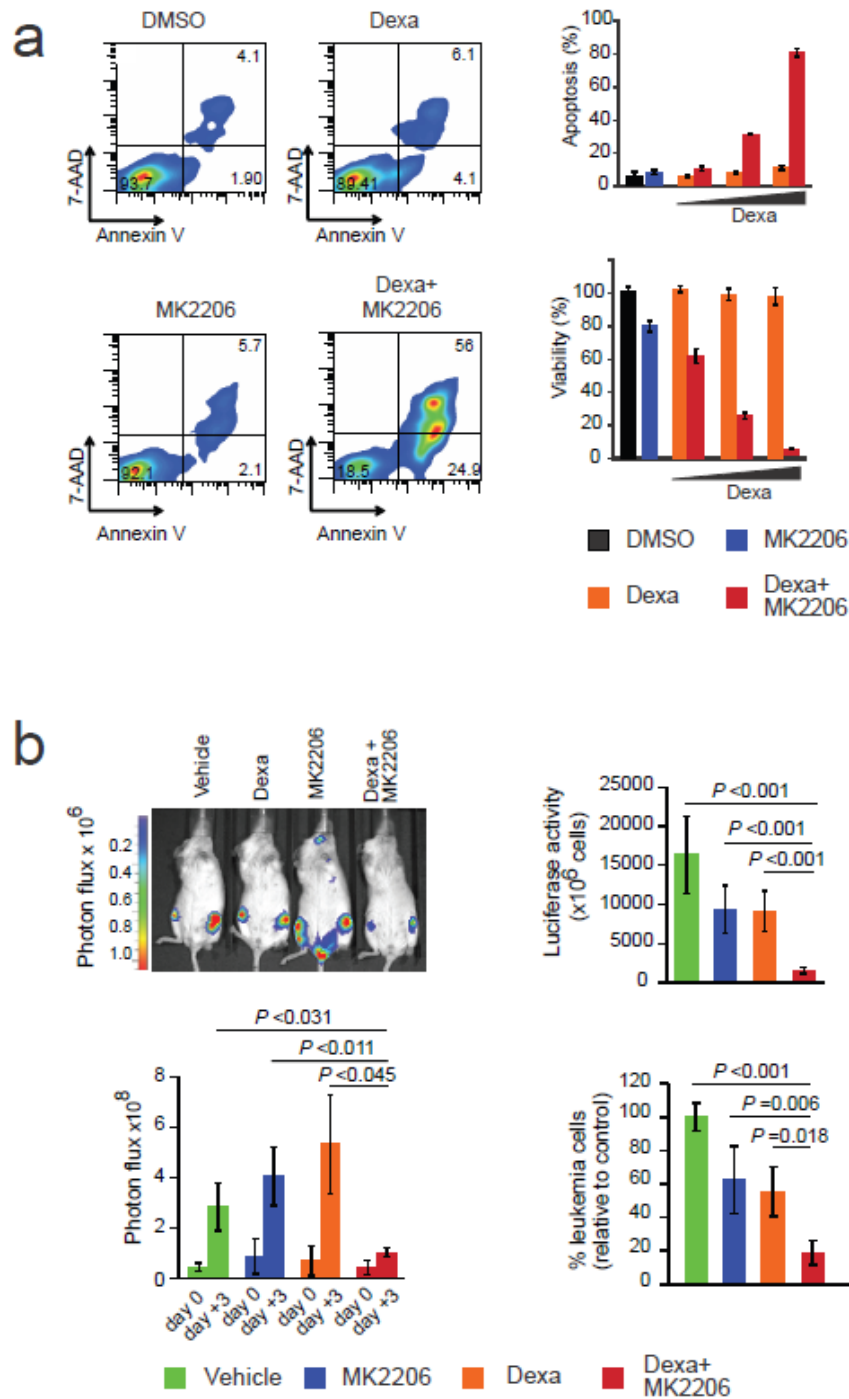


Figure 7-17 Pharmacologic inhibition of AKT with MK-2206 reverses glucocorticoid resistance in human T-ALL cell lines. (a) Representative plots and quantification of apoptosis and loss of cell viability in CCRF-CEM T-ALL cell line

treated with vehicle only, MK2206, dexamethasone or dexamethasone plus MK2206 in combination in vitro. (b) Quantification of tumor load by bioluminescence in in vivo imaging and analysis of luciferase activity or human CD45 expressing cells in the bone marrow of CCCF-CEM T-ALL xenografted mice treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206).

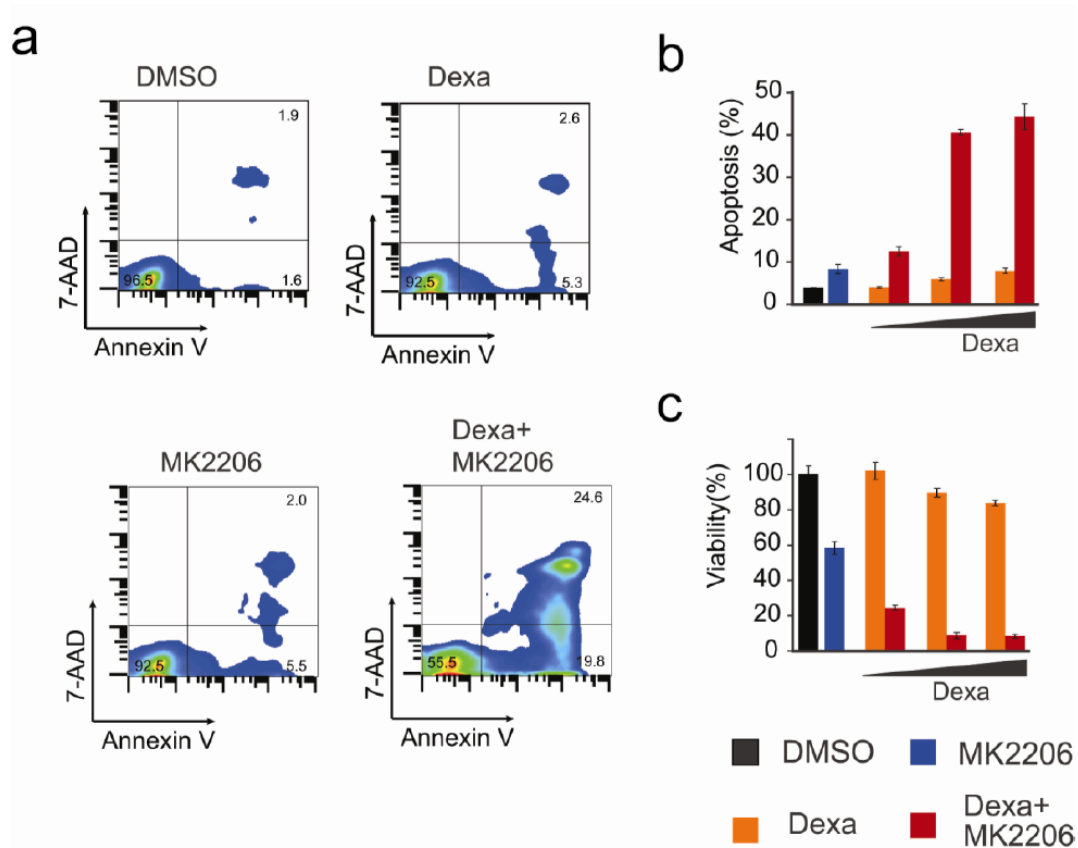


Figure 7-18 Pharmacological inhibition of AKT reverses glucocorticoid resistance in MOLT-3 T-ALL cells (a,b) Representative plots (a) and quantification (b) of apoptosis and cell viability (c) in MOLT-3 T-ALL cells for 72 hours with vehicle only, MK2206 and dexamethasone alone or dexamethasone plus MK2206 in combination.

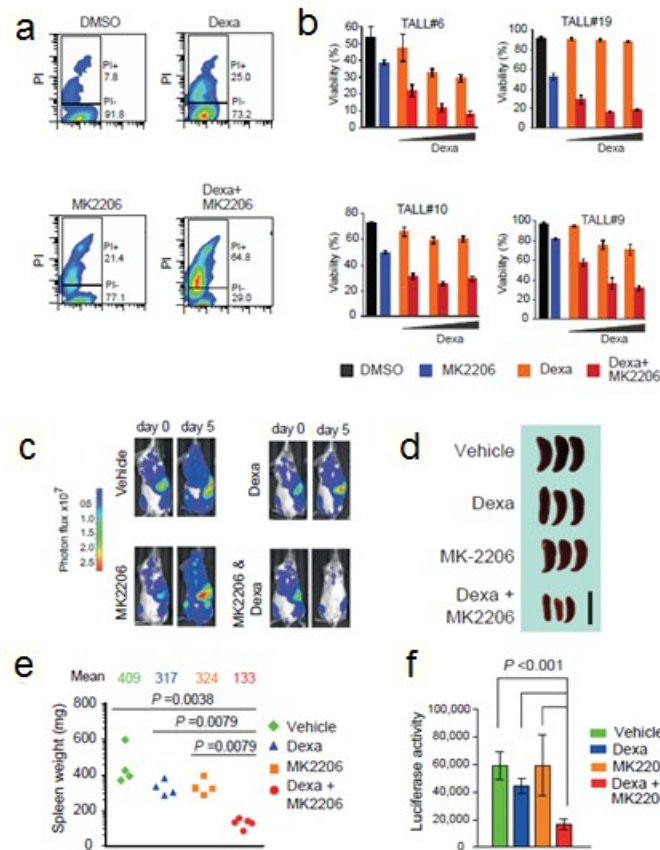


Figure 7-19 Pharmacologic inhibition of AKT with MK-2206 reverses glucocorticoid resistance in human T-ALL primary samples. (a,b) Representative plots (a) and quantification of loss of viability analysis (b) in primary T-ALL patient samples treated with vehicle only, MK2206 and dexamethasone alone or dexamethasone plus MK2206 in combination. Percentages of viable (PI -), and non-viable (PI +) cells are indicated. (c-f) Representative examples of primary human T-ALL xenografted mice showing changes in tumor load assessed by *in vivo* imaging (c), spleen size (d), spleen weight (e) and luciferase activity in bone marrow cells (f) from primary human leukemia xenografted mice treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206). Scale bar: 2 cm.

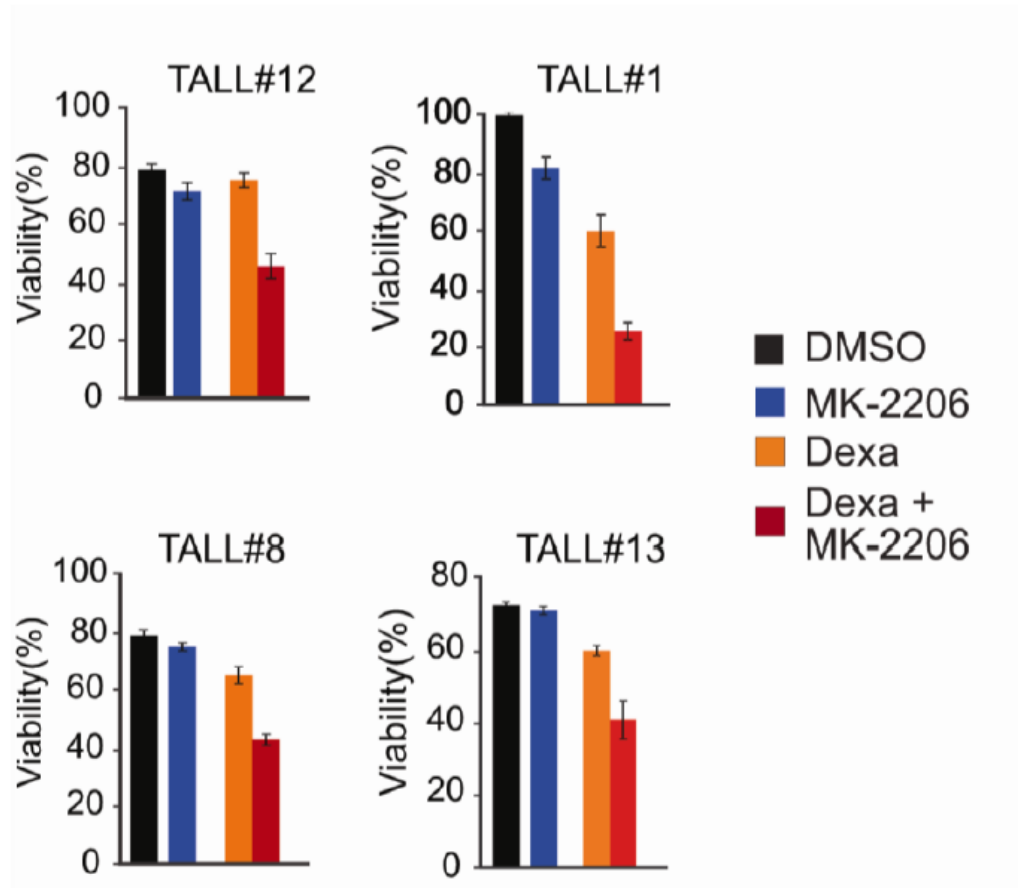


Figure 7-20 Pharmacological inhibition of AKT in vitro reverses glucocorticoid resistance in primary human T-ALL xenografts. Analysis of cell viability in primary T-ALL samples treated for 72h with vehicle only, MK2206 and dexamethasone alone or dexamethasone plus MK2206 in combination

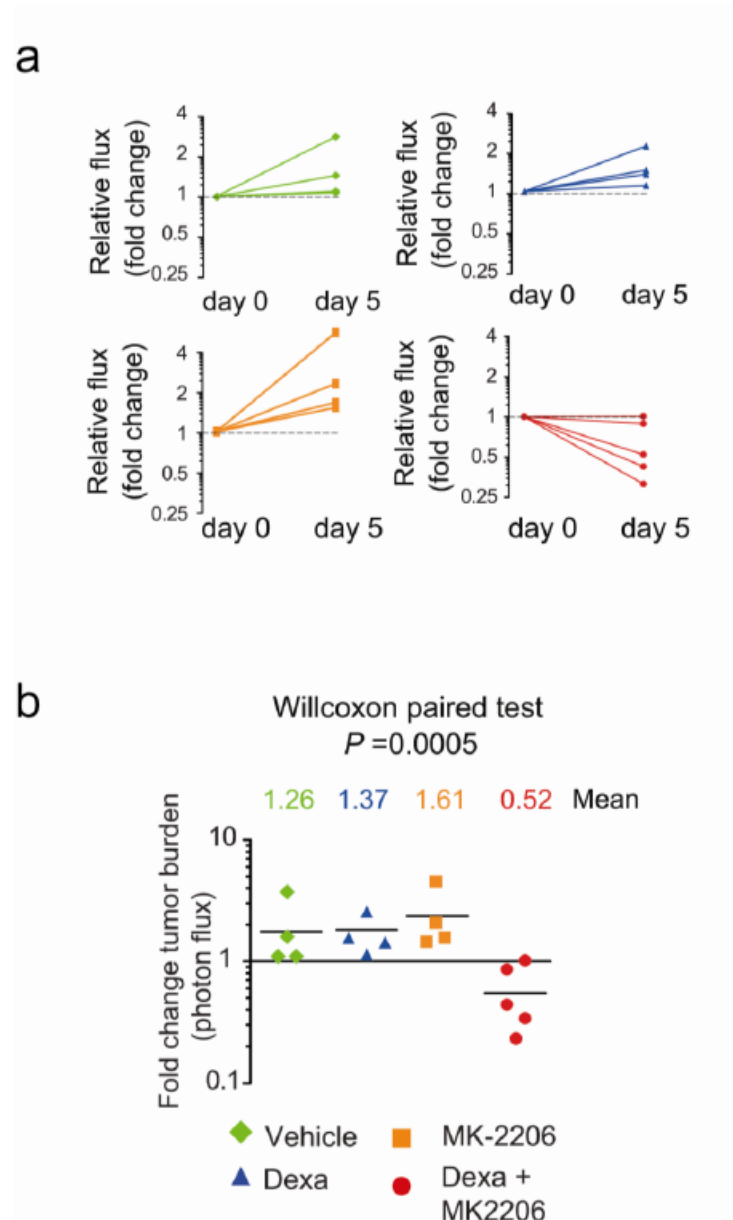


Figure 7-21 Pharmacological inhibition of AKT in vivo reverses glucocorticoid resistance in primary human T-ALL xenografts. (a,b) Bioimaging quantification (a) and analysis (b) of tumor load changes in mice treated with vehicle (control), dexamethasone (Dexa), MK2206, MK2206 plus dexamethasone (Dexa + MK2206) for 5 days.

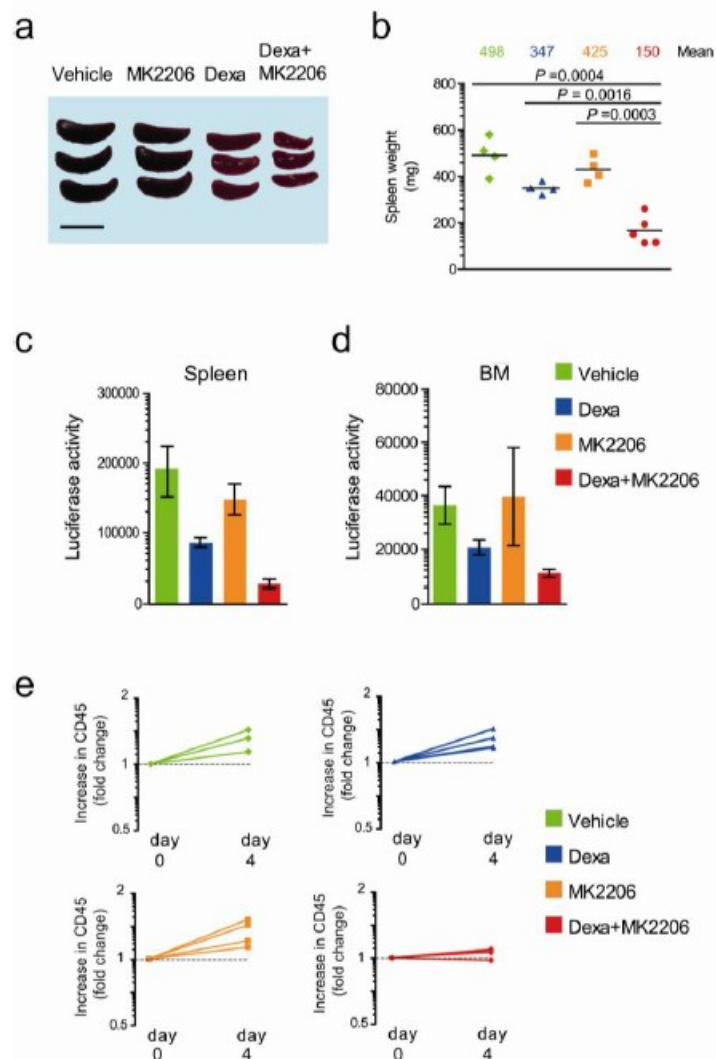


Figure 7-22 Pharmacological inhibition of AKT in vivo reverses glucocorticoid resistance in primary human T-ALL xenografts. (a,b) Representative images of spleens (a) and spleen weights (b) of leukemic mice treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206) for 4 days. (c,d) Quantification of tumor load by determining luciferase activity from cells isolated from the spleen (c) and bone marrow (d). (e) Quantification of tumor load changes by determining the increase in circulating CD45 positive cells in the peripheral blood of mice injected with a human xenograft and treated with vehicle (control), dexamethasone (Dexa), MK2206, MK2206 plus dexamethasone (Dexa + MK2206) for 4 days. Scale bar: 2cm.

Finally, we generated a mouse leukemia model in which glucocorticoid resistance is specifically driven by genetic loss of *Pten* using a well-established retroviral transduction and bone marrow transplantation protocol [191]. In this model, transplantation of tamoxifen-inducible conditional *Pten* knockout (Rosa26TMCre *Pten*^{flox/flox}) hematopoietic progenitors with retroviruses expressing a mutant and constitutively active form of the NOTCH1 receptor (*NOTCH1* L1601P ΔPEST) resulted in the development of NOTCH1 driven T-ALL tumors as previously described [191]. Next we infected *NOTCH1* Rosa26TMCre *Pten*^{flox/flox} T-ALL lymphoblasts with a luciferase expressing retrovirus and transplanted them into secondary recipients which were treated with vehicle only or tamoxifen in order to generate *Pten*-non-deleted and *Pten*-deleted isogenic tumors, respectively. Treatment of *Pten*-non-deleted tumor bearing mice with dexamethasone showed a significant improvement in survival compared with vehicle only treated controls ($P < 0.01$)(Figure 7-23-a). In contrast, and consistent with a role of *Pten* loss and AKT1 activation in promoting glucocorticoid resistance, all mice harboring *Pten*-deleted tumors failed to respond to dexamethasone treatment and showed no survival differences compared to vehicle treated controls (Figure 7-23-b).

To test the efficacy of MK2206 and glucocorticoid combination we treated mice transplanted with *NOTCH1*-induced *Pten*-deleted murine tumors expressing luciferase in secondary recipients, with vehicle only (DMSO), MK2206, dexamethasone or MK2206 plus dexamethasone and monitored their response to therapy by *in vivo* bioimaging. Animals treated with dexamethasone or

MK2206 in this experiment showed progressive tumor growth similar to that observed in vehicle-treated controls, while mice treated with MK2206 plus dexamethasone showed significant antitumor responses (Figure 7-23-c, d; $P < 0.01$) which translated in significantly improved survival in this group (Figure 7-23-e).

Finally, we analyzed the role of NR3C1 S134 phosphorylation in the therapeutic response to glucocorticoids and the effects of *Pten* loss in glucocorticoid therapy in this model. Retroviral expression of the glucocorticoid receptor in *Pten* non-deleted lymphoblasts (Figure 7-24) enhanced the response of NOTCH1-induced leukemias to glucocorticoid treatment; an effect that was effectively abrogated upon *Pten* loss (Figure 7-23-f). In contrast, expression of the AKT-resistant NR3C1 S134A mutant protein was equally effective at increasing the antileukemic effects of glucocorticoids in *Pten* non-deleted and *Pten* null lymphoblasts (Figure 7-23-f).

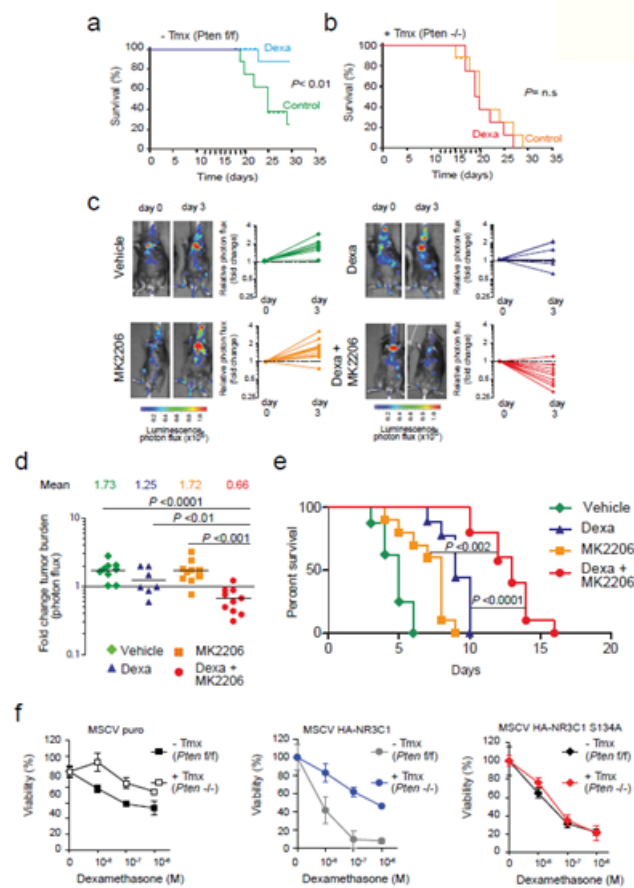


Figure 7-23 Pharmacologic inhibition of AKT reverses glucocorticoid resistance in a mouse model of glucocorticoid resistant T-ALL. (a,b) Kaplan-Meier survival plot in mice treated with dexamethasone (Dexa) or vehicle (Control) after allograft transplantation of Pten-non-deleted [-Tmx (Pten f/f)] (a) or Pten-deleted [+Tmx (Pten -/-)] (b) NOTCH1-induced T-ALL tumor cells. Arrows indicate the time of drug treatment. (c,d) Representative images and changes in bioluminescence in vivo imaging (c) and analysis of treatment response in mice allografted with NOTCH1 induced Pten deleted mouse leukemia cells and treated with vehicle only, MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206). (e) Kaplan-Meier plot of overall survival in mice allografted with NOTCH1 induced Pten deleted mouse leukemia cells and treated with vehicle only (control), MK2206, dexamethasone (Dexa) or MK2206 plus dexamethasone (Dexa + MK2206). (f) Quantification of glucocorticoid-induced loss of viability in

NOTCH1 induced Pten non deleted [-Tmx (Pten f/f)] or Pten deleted [+Tmx (Pten -/-)] mouse leukemia cells infected with an empty vector control (MSCV-puro) or retroviruses expressing the wild type glucocorticoid receptor NR3C1 (MSCV HA-NR3C1) or the S134A glucocorticoid receptor NR3C1 mutant protein (MSCV HA-NR3C1 S134A).

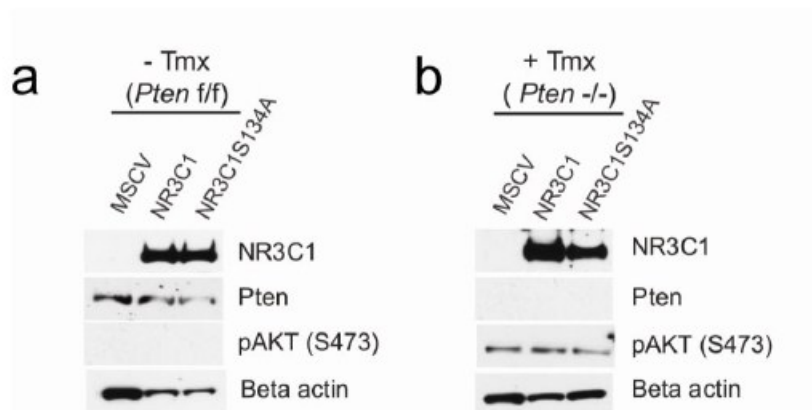


Figure 7-24 Overexpression of NR3C1 and NR3C1 S134A mutant in primary murine leukemia cells. (a) Western blot analysis determining the retroviral expression of NR3C1 or NR3C1 S134A in Pten non-deleted [-Tmx (Pten f/f)] and (b) PTEN deleted [+Tmx (Pten -/-)] NOTCH1-induced T-ALL mouse leukemia cells. Beta actin is shown as loading control.

Overall these results support a direct effect of AKT on the response to glucocorticoid therapy mediated by S134 phosphorylation of the glucocorticoid receptor protein and show that pharmacologic inhibition of AKT can effectively enhance glucocorticoid response and reverse glucocorticoid resistance in T-ALL.

7.5 Discussion

Despite much research, the molecular basis of glucocorticoid resistance in ALL remains poorly understood. Genetic abnormalities of the glucocorticoid receptor gene are rarely seen in primary glucocorticoid-resistant leukemias [19] or in ALL samples at relapse [165]. However, alternative glucocorticoid receptor-centered mechanisms of resistance including decreased glucocorticoid receptor expression [166, 192] and impaired glucocorticoid receptor auto-upregulation [167-172] have been proposed. Notably, increasing glucocorticoid receptor expression or glucocorticoid receptor autoupregulation via inhibition of NOTCH1 signaling can effectively abrogate glucocorticoid resistance in T-ALL primary samples and cell lines [193]. In addition, increased expression of antiapoptotic factors such as MCL1 [173] and epigenetic silencing of the *BCL2L11* proapoptotic gene [194] can impair glucocorticoid induced apoptosis in ALL. Thus, decreasing MCL1 expression via inhibition of mTOR with rapamycin [23], and increased *BCL2L11* levels via GSK3 inhibition [195] can enhance glucocorticoid induced cell death and reverse glucocorticoid resistance in ALL.

Our results show a new mechanistic role for AKT1 in glucocorticoid resistance in T-ALL. Notably, activation of the PI3K-AKT signaling pathway is highly prevalent in T-ALL and can result from deletions and mutations in *PTEN* [182, 183] and from activating mutations in the *PIK3CA* gene [182, 183], but also downstream of mutations and autocrine or paracrine receptor triggering the activation of cytokine receptor kinases. AKT1 is a pleiotropic factor with broad effects promoting cell

growth, metabolism and survival [23, 196-199]. Thus, constitutive activation of AKT1 can potentially antagonize glucocorticoid induced apoptosis via multiple indirect mechanisms. Still, our results demonstrate a direct effect of AKT1 in the phosphorylation and inhibition of the glucocorticoid receptor protein.

Protein phosphorylation can modulate the activity of nuclear receptors by affecting their transactivation activity, cellular localization, interaction with cofactors and stability [200]. Mechanistically, AKT1 phosphorylation of NR3C1 results in impaired glucocorticoid-induced nuclear localization. Notably, phosphorylation of S134 together with the presence of a conserved proline residue in position 136 (Figure 7-13-e), creates a potential 14-3-3 binding motif, which is a common mode of regulation of AKT1 substrates. Thus, 14-3-3 binding has been implicated in AKT mediated inhibition of the proapoptotic factor BAD and the FOXO1, FOXO3 and FOXO4 transcription factor proteins [196, 201], suggesting a potential role for 14-3-3 interaction on the inhibitory effect of AKT1 on the NR3C1 glucocorticoid receptor function.

Overall, the results presented here, strongly support that AKT inhibition may reverse glucocorticoid resistance and warrant the clinical testing of AKT inhibitors and glucocorticoids in combination for the treatment of T-ALL. Finally our results demonstrate that reverse engineering of signaling networks can be exploited to identify relevant therapeutic targets for the reversal of chemotherapy resistance in human cancer.

7.6 Materials and Methods

7.6.1 Reverse engineering signaling molecule-focused network analysis of glucocorticoid resistance in T-ALL

Here we used the mRNA expression of genes globally co-expressed with a signaling molecule (*S*) as a surrogate readout of the activity of *S* based on the assumption that such genes are enriched for both members of the signal transduction cascade that includes *S* (both upstream and downstream members) and targets of transcription factors regulated by *S*. To generate a T-ALL transcriptional network, we processed Human U133 Plus2 Affymetrix microarray gene expression data from a series of 223 T-ALL primary samples using GC-RMA normalization and non-specific filtering (removing probes with no Entrez id, Affymetrix control probes, and non-informative probes by IQR variance filtering with a cutoff of 0.5), to 21,054 probes in total. Then we run the ARACNe algorithm[180], against 4831 probe sets corresponding to 2602 genes with annotated functions in signaling transduction (with annotation of “signal transduction” (GO:0007165) in Gene Ontology as of 2009) to establish a signaling factor-centered Interactome at transcriptional level. Use of ARACNe is justified to find minimal regulatory paths, i.e., to eliminate most indirect interactions in signal transduction analysis. This produces a minimal representation such that if any interaction were removed, information transfer in the system could no longer be explained. The parameters of the algorithm were configured as below: *P*-value threshold $P = 1e-7$, DPI tolerance $e = 0$, and

number of bootstraps $N_B = 100$. We used the adaptive partitioning algorithm for mutual information estimation.

7.6.2 Glucocorticoid resistance signature analysis

Out of the 223 primary samples with gene expression profile data, twenty-two were diagnosed as glucocorticoid resistant and ten as glucocorticoid sensitive. Using this *a priori* classification, we performed differential gene expression analysis to generate an E_{GC} signature for glucocorticoid resistance. A statistical Probit model was used to infer the correlation of gene expression and phenotypes with the advantage of detecting weak effects, and Bayesian-MCMC computing was employed to estimate parameters for its robustness and accuracy, even with small sample size. In particular, a t -distribution prior and Gibbs sampling were used in this analysis [202].

7.6.3 Inferring master signaling drivers of glucocorticoid resistance in T-ALL by NetBID2

We used the NetBID2 algorithm to interrogate the ARACNe-inferred signaling network with the E_{GC} signature, to identify candidate master modulators that may induce or suppress glucocorticoid resistance. It has been shown that feedback loops in signaling pathways induce co-expression of their member proteins, once the network has relaxed to steady state [203]. Thus, for each signaling gene S , we generated a putative S -regulon R_S , from the T-ALL signaling interactome by selecting the first neighbors of S . This is based on the assumption that such genes are enriched for both members of the signal transduction cascade that

includes S (both upstream and downstream members) and targets of transcription factors regulated by S . Next, we used Gene Set Enrichment Analysis (GSEA) to test the enrichment of the R_S genes in the E_{GC} signature as previously described [204]. For GSEA method we used ‘maxmean’ statistic [205] to score the enrichment of the gene set in the E_{GC} signature and sample permutation to build the null distribution for statistical significance. To generate robust signatures, we only used signaling proteins with more than 50 genes in their S -regulon. P -values were corrected using Efron’s procedure [205].

7.6.4 Cell lines and primary leukemia samples

Human embryonic kidney (HEK) 293T and osteosarcoma U2OS (HTB-96) cells were maintained in Dulbecco’s modified Eagle’s medium (DMEM) containing 10% fetal bovine serum and 0.05 mg/ml penicillin/streptomycin. T-ALL cell lines were maintained in RPMI-1640 media supplemented with 10% FBS and 0.05mg/ml penicillin/streptomycin. T-ALL lymphoblast samples were provided by collaborating institutions in the US (Department of Pediatrics, Columbia Presbyterian Hospital, New York), Italy (Department of Pediatrics, University of Padova), the Hospital Central de Asturias (Oviedo, Spain) and the Eastern Cooperative Oncology Group (ECOG). All samples were collected with informed consent and under the supervision of local IRB committees. Primary T-ALL cells were cultured *in vitro* with MS5 stromal cells expressing the Delta-like1 NOTCH ligand protein in cytokine supplemented media as previously described [206].

7.6.5 Inhibitors and drugs

The allosteric AKT inhibitor MK2206 or 8-[4-(1-aminocyclobutyl)phenyl]-9-phenyl-1,2,4-triazolo[3,4-f] [1,6]naphthyridin-3(2H)-one hydrochloride [1:1] was obtained from Selleck Chemicals LLC. Dexamethasone and 4-Hydroxytamoxifen were from Sigma-Aldrich.

7.6.6 siRNA validation of regulators of glucocorticoid resistance

We performed siRNA experiments in the glucocorticoid sensitive T-ALL cell line, DND41. For this purpose, DND41 cells were electroporated with smartpool siRNAs (Dharmacon) targeting the top nine master regulators identified through MARINA analysis using the SF Cell line 96-well Nucleofector Kit (Lonza). Twenty-four hours after electroporation, cells were treated with Dexamethasone (1 μ M) for 48h. Cells were then collected and analyzed for apoptosis by FACS after staining membrane expression of Annexin V and 7-AAD with the PE AnnexinV Apoptosis Kit I (BD Biosciences).

7.6.7 Luciferase reporter assays

We performed NR3C1 reporter assays in U2OS cells stably expressing haemagglutinin (HA) tagged wild type or mutant S134A NR3C1 and infected with retroviruses expressing EGFP only (pMSCV IRES GFP) or myristoylated AKT1 and GFP (pMSCV MYR-AKT1 IRES GFP) and sorted for GFP expression [182]. These cells were cultured in DMEM media supplemented with 10% dialyzed fetal bovine serum in the presence or absence of increasing doses of dexamethasone (10nM to 1 μ M). In these experiments cells were co-transfected with an inducible

firefly luciferase reporter containing tandem repeats of the Glucocorticoid Responsive Elements (GRE) and a constitutively expressing Renilla construct in the Cignal GRE Reporter (luc) Kit (SABiosciences); or alternatively a luciferase reporter construct (pGL3 NR3C1 A1 FP11-FP12) containing the FP11-FP12 regulatory sequence (5'-CGTAAAATGCGCATGTGTTCCAACGGAAGCACTGG-3') responsible for autoregulation of the *NR3C1* promoter A1[24, 185] and the plasmid expressing pRL-CMV Renilla luciferase gene (Promega). NR3C1 reporter activity and Renilla luciferase activity were analyzed 40-48 hours after transfection and 24 hours after dexamethasone treatment with the Dual-Luciferase Reporter Assay kit (Promega).

7.6.8 Quantitative real-time PCR

Total RNA from T-ALL cell lines was extracted using Trizol reagent (Invitrogen). cDNA was generated with the Super Script First Strand Synthesis System for RT-PCR (Invitrogen) and analyzed by quantitative real-time PCR using SYBR Green PCR Master Mix (Applied Biosystems) and the 7300 Real-Time PCR System (Applied Biosystems). Relative expression levels were normalized with *GAPDH* expression used as a reference control.

7.6.9 Western blotting and immunoprecipitation

Total cell lysates were prepared using RIPA lysis buffer supplemented with phosphatase inhibitor cocktail set I and II (Sigma) and protease inhibitor cocktail tablets (Roche) and normalized for protein concentration using the BCA method (Pierce). For Western blotting, protein samples were separated on 4-12%

gradient Tris-Glycine SDS-PAGE (Invitrogen) and transferred to PVDF membrane (Millipore). Membranes were blocked in PBST containing 5% nonfat milk, incubated with primary antibodies according to the antibody manufacturer's instructions, followed by incubation with horseradish peroxidase-conjugated goat anti-rabbit, goat anti-mouse or donkey anti-rat IgG (Amersham) and enhanced chemiluminescence detection (Perkin Elmer). Antibodies against glucocorticoid receptor (E-20), tubulin (TU-02), beta actin (C-11) and MAX (C-17) were from Santa Cruz Biotechnology. Antibodies recognizing BIM, phospho-AKT Ser473, phospho-AKT Thr308, phospho-mTOR (S2448), mTOR, AKT and phospho-(Ser/Thr) Akt substrate were from Cell Signaling Technologies. HA epitope antibody was from Roche, FLAG epitope antibody from Sigma and an antibody against PTEN (clone 6H2.1) was obtained from Cascade Biosciences. For immunoprecipitation, cell lysates were incubated with anti-HA or anti-FLAG M2 affinity gel beads (Sigma) overnight at 4°C. Beads were washed five times with lysis buffer and proteins were eluted by incubating the beads with HA peptide (1mg/ml, Roche) or FLAG peptide (1mg/ml, Sigma). Immune complexes were analyzed by SDS-PAGE and Western blotting.

For immunoprecipitation of endogenous NR3C1 bound proteins in T-ALL cells, we lysed 100-150 million T-ALL cells for 30 min in RIPA lysis buffer supplemented with phosphatase and protease inhibitor cocktails. After centrifugation, the cell lysates were pre-cleared with TrueBlot® anti-Mouse Ig IP Beads (eBioscience) before being incubated overnight with 5µg of mouse antibody against NR3C1 (AbCam) or irrelevant mouse Ig (Santa Cruz).

Subsequently, samples were incubated 2 hours with TrueBlot® anti-Mouse Ig IP Beads and immunoprecipitates washed 5 times with CO-IP buffer (50mM Tris-HCl pH 7.9, 150 mM NaCl, 1mM EDTA, 0.1% NP-40 and protease inhibitors). Immune complexes were then analyzed by SDS-PAGE and Western blotting.

7.6.10 Preparation of Cytoplasmic and Nuclear extracts

CCRF-CEM, MOLT-3 T-ALL cells were treated with vehicle (DMSO), dexamethasone (1 μ M), MK2206 (0.5-1 μ M), or the combination dexamethasone and MK2206 for 1 hour before being harvested. Cytoplasmic and nuclear extracts were prepared using the nuclear extraction kit (Active Motif) according to the manufacturer's recommendations.

7.6.11 In vitro GST-pull down protein interaction assays

For in vitro binding assays, GST fusion proteins of NR3C1 or mutant NR3C1 S134A were expressed and purified from BL-21 bacterial cells. Approximately 2 μ g of GST fusion proteins bound to glutathione-agarose beads (Immobilized glutathione; Thermo scientific) were incubated with 1-2 μ g of Histidine-tagged activated AKT1 (His-AKT1, Millipore) in GST-lysis buffer (20 mM Tris-HCl, 200 mM NaCl, 1mM EDTA, 0.5% NP-40 and protease inhibitors) for 2 hours at 4 °C. After extensive washing in GST-lysis buffer, proteins were separated on 4-12% NuPage gradient gels, transferred to PVDF, and probed by Western blot using antibodies against AKT1 and the NR3C1 protein.

7.6.12 In vitro kinase assays

Flag-tagged recombinant GST-NR3C1 wt and GST-NR3C1 S134A mutant proteins were expressed, purified from *Escherichia coli*, and incubated with recombinant active His-AKT1 protein (Millipore) in kinase buffer (Cell Signaling) containing γ -³²P-ATP at 30°C for 30 min. The reaction was stopped by the addition of 5X SDS-Laemmli's sample buffer. Samples were separated on 3-8% Tris-Acetate SDS-PAGE (Invitrogen), and the gels subjected to autoradiography.

7.6.13 Mass spectrometry analysis of NR3C1 phosphorylation sites

U2OS cells stably expressing HA-tagged human NR3C1 and Myr-AKT were lysed in RIPA buffer and immunoprecipitated with anti-HA antibody conjugated beads (Sigma). After overnight incubation, the beads were extensively washed with BC-500 (500 mM NaCl, 20 mM Tris-Cl pH =8.0, 20% glycerol, 1% Triton-X, 1mM EDTA) and, subsequently, proteins were eluted by incubating the beads with HA peptide (1mg/ml, Roche). The eluted NR3C1 was diluted in 4X SDS-PAGE sample buffer and electrophoresed on 3-8% Tris-Acetate gels. Gel bands were stained with Simply Blue Stain (Invitrogen), excised, reduced with DTT, alkylated with iodoacetamide and digested with trypsin. Afterward, the digest was analyzed for phosphorylated peptides by nanoLC-ESI-MS/MS. MS/MS spectra were processed using ProteinLynx from the MassLynx 4.0 software and searched against the Swiss-Prot protein database using Mascot (www.matrixscience.com) with differential modifications for Ser/Thr/Tyr phosphorylation (+79.97) and the sample processing artifacts Met oxidation (+15.99) and Cys alkylation (+57.02). MS/MS spectra of phosphorylated peptides

and the corresponding non-phosphorylated peptides were manually inspected to be sure that all b- and y- fragment ions aligned with the assigned sequence and modification sites. For relative quantification of phosphorylation peptide signal levels, the total ioncurrent (TIC) for the phosphorylated peptide ion and non-phosphorylated peptide ion were integrated and calculated according to the following equation: $TIC_{PO4}/(TIC_{PO4} + TIC_{nonPO4}) = \text{ratio of phosphopeptide signal}$. Comparison of the ratio of the phosphorylated to nonphosphorylated peptide forms using this method provides an accurate measure of signal level change since the total peptide signal (modified and unmodified) is measured.

7.6.14 Immunofluorescence studies

U2OS cells stably expressing wild type or the S134A NR3C1 mutant together with MYR-AKT1 IRES EGFP or EGFP alone were plated on 35-mm dishes with glass bottom inserts and treated with vehicle only or dexamethasone (1 μ M). After 1 hour they were washed with PBS, fixed in 4% paraformaldehyde and permeabilized with NP-40 (0.1% NP-40 in PBS). We blocked the permeabilized cells with 1.5% goat serum and incubated them with antibodies against NR3C1 (1:500; Santa Cruz Biotechnology), followed by Alexa Fluor 594 (1:1000; Invitrogen) staining. We mounted the stained cells in Vectashield containing DAPI (4',6'-diamidino-2-phenylindole; Vecta Laboratories, Burlingame, CA) and analyzed them by confocal imaging on a Zeiss LSM510-NLO microscope. Quantification of the NR3C1 signal in the cytoplasmic and nuclear compartments was done using ImageJ software.

7.6.15 Cell viability assays and flow cytometric analysis

We analyzed cell viability via a metabolic colorimetric assay using the Cell Proliferation Kit I (MTT; Roche) or Cell Proliferation Reagent WST-1 (Roche). Drug concentrations used in these experiments were 10nM to 10 μ M for dexamethasone and 0.5 μ M to 5 μ M for MK2206. We analyzed apoptosis by flow cytometry (FACS) after staining membrane expression of Annexin V and 7-AAD with the PE AnnexinV Apoptosis Kit I (BD Biosciences). For primary T-ALL samples, we assessed cell viability using the BD Cell Viability kit (BD Biosciences) coupled with the use of fluorescent counting beads. In these experiments, 2×10^5 leukemic cells were plated with 4×10^4 MS5-DL1 stroma cells into 24-well plates. The next day we treated cells with vehicle only (DMSO), dexamethasone (10 nM-1 μ M), MK2206 (0.5 μ M-10 μ M) or the combination dexamethasone (10 nM-1 μ M) plus MK2206 (0.5 μ M-10 μ M). After 72 hours we harvested the treated cells, passed them through a 50 μ M Nylon mesh and stained them with an APC-conjugated antibody recognizing human CD45. After CD45 surface staining, we incubated the cells with a staining mix containing thiazole orange (TO) which labels all cells and PI which labels dead cells. Fluorescent BD Liquid counting beads were added to calculate absolute cells numbers. The viability of T-cell lymphoblasts was determined gating on CD45 positive cells and is expressed as the percentage of TO positive and PI negative cells.

7.6.16 Retroviral and lentiviral constructs and viral production

We created the retroviral construct pMSCV-HA-NR3C1 by cloning a pCMV-HA-hGR BamHI-DraI fragment containing an HA tagged full length NR3C1 cDNA, into the pMSCV-puro vector [24]. Site-directed mutagenesis was performed using the Quickchange Site Directed Mutagenesis Kit (Stratagene, Windsor, ON, USA) according to the manufacturer's protocol. Ser 134 on hNR3C1 was replaced with alanine (S134A) in the pMSCV-HA-NR3C1-puro with the following primers: forward 5'-CTCAATAGGTGACCGCCGTTCCAGAGAACCC-3' and reverse 5'-GGGTTCTCTGGAACGGCGGTGACCTATTGAG-3'. Flag-tagged constitutively active AKT (pBabe-Puro-Myr-Flag-AKT1), which lacks its pleckstrin homology domain but has a Src myristoylation signal sequence, was obtained from Addgene (plasmid number 15294). PTEN knock-down was done using pLKO-shPTEN-GFP [182] and pLKO-shLUC-GFP was used as control. Retroviral particles driving the expression of EGFP (pMSCV IRES GFP), myristoylated AKT (pMSCV MYR-AKT IRES GFP), NR3C1 (pMSCV HANR3C1 puro), NR3C1 S134A (pMSCV HA-NR3C1 S134A puro) were generated as previously described [207]. Lentiviral particles determining the knock-down of PTEN (pLKO shPTEN GFP) or the luciferase gene as control (pLKO shLUC-GFP) were generated according to standard protocols. Lentiviral particles expressing a luciferase and neomycin phosphotransferase fusion transcript were generated with the FUW-Lucneo vector [208]. Retroviral and lentiviral particles were produced and used in spin infections as previously described [209].

7.6.17 Recombinant protein production

To generate glucocorticoid receptor-GST fusion proteins we amplified the NR3C1 cDNA by PCR and subcloned it in the pGEX4T-1 prokaryotic expression vector (Amersham Biosciences). We introduced a point mutation resulting in the NR3C1 S134A substitution via site directed mutagenesis. We produced GST-NR3C1 and GST-NR3C1 S134A proteins in BL21 bacteria transformed with pGEX4T-1 NR3C1 and pGEX4T-1 NR3C1 S134A vectors. We induced GST protein synthesis in bacteria with 0.2 mM isopropyl-b-D-thiogalactopyranoside (Sigma) for 5 hours at 30°C, then harvested the bacteria cells by centrifugation, and lysed them in modified BC-500 buffer (500 mM NaCl, 20 mM Tris-Cl pH=8.0, 20% glycerol, 1% Triton-X, 1mM EDTA, 0.2% NP-40) for 1h at 4°C. Cleared bacteria lysates were subsequently incubated with glutathione-Sepharose 4B beads (Amersham Biosciences) overnight at 4°C, and the glutathione-bead bound proteins were eluted by adding 15 mM glutathione in 50 mM Tris-HCl, pH=8.0. Finally, we removed glutathione by dialysis against PBS and analyzed the recombinant proteins for yield and purity by SDS-PAGE followed by Coomassie Brilliant Blue R-250 staining.

7.6.18 Mice and animal procedures

All animals were maintained in specific pathogen-free facilities at the Irving Cancer Research Center at Columbia University Medical Campus. Animal procedures were approved by the Columbia University Institutional Animal Care and Use Committee. Rosa26 Cre-Tam mice expressing a tamoxifen-inducible form of the Cre recombinase from the ubiquitous *Rosa26* locus [210] and *Pten*

conditional knockout mice (*Pten*^{fl}) have been previously described [211]. To generate *NOTCH1*-induced T-ALL tumors in mice we performed retroviral transduction of bone marrow cells with an activated form of the *NOTCH1* oncogene (*NOTCH1* L1601P Δ PEST) and transplanted them via intravenous injection into lethally irradiated recipients as previously described [191]. Briefly, bone marrow cells were collected from the long bones of 6-9 week-old C57BL/6 *Rosa26* Cre-Tam *Pten*^{flox/flox} mice. Lin^{neg} cells were isolated using Lineage Depletion magnetic beads (Miltenyi Biotech). Purified cells were cultured in transplant medium consisting of Optimem (Gibco) supplemented with IL-3 (10ng/ml), SCF (50ng/ml), IL-6 (10ng/ml) and 5% fetal calf serum overnight, and spin infected by incubation in retroviral supernatant (MigR1-*NOTCH1* L1601P Δ PEST) containing the same cytokine cocktail and 8 μ g/ml polybrene and centrifuged at 2500 rpm for 90 minutes. A second round of spinoculation was performed after 24 hours. After washing with PBS, at least 50,000 Sca-1+GFP+ cells were injected intravenously into lethally irradiated (9.5 Gy) recipients. Mice were maintained on antibiotics in drinking water 2 weeks after bone marrow transplantation. Tumor bearing mice were euthanized and primary tumor cells extracted from the spleens of leukemic mice. These tumor cells were then infected with retroviral particles (MigR1 Cherry-LUC), expressing a fusion protein between the red cherry fluorescent protein and luciferase and re-injected in sub-lethally irradiated mice (4 Gy). After a 5 day window for tumor engraftment, secondary recipients of *NOTCH1*-induced *Pten* inducible conditional knockout cells labeled with Cherry-luciferase harboring homogeneous tumors were treated

with tamoxifen (5mg/mouse) (n=16), to induce deletion of the *Pten* locus or vehicle only (n=16) by intra-peritoneal injection. After 1 week, *Pten*-non-deleted and *Pten*-deleted mice were analyzed by luciferase bioimaging [24] and segregated into groups of isogenic leukemias containing 8 animals each and with homogeneous tumor loads. A control group of *Pten*-non-deleted animals and a control group of *Pten*-deleted mice were treated with vehicle (DMSO), while glucocorticoid treatment groups of *Pten*-non-deleted and *Pten*-deleted mice received escalating daily doses of 1mg/kg, 2mg/kg and 5mg/kg of dexamethasone. Each dose of dexamethasone was administered for three consecutive days. At the end of treatment all mice were monitored daily and animals showing overt signs of disease were euthanized following Institutional Animal Care and Use Committee guidelines. For intravenous transplantation model, we used sub-lethally irradiated C57BL/6 mice (Taconic Farms). We injected 2 million *Pten* deleted *NOTCH1* L1601P Δ PEST CHERRY-luciferase expressing cells via tail vein injection. After a 10 day window for tumor engraftment, we segregated mice with homogeneous tumor loads into treatment groups (7-10 mice per group) and treated them daily with vehicle (DMSO), dexamethasone (5 mg/kg via intraperitoneal injection), MK2206 (10 mg/kg via oral gavage twice a day) or dexamethasone (5 mg/kg) plus MK2206 (10 mg/kg) for 7 days. We evaluated disease progression and therapy response after 3 days of treatment by bioluminescence. For imaging studies, mice were anesthetized by isoflurane inhalation and injected with D-luciferin at 50 mg kg⁻¹ (Xenogen) intraperitoneally. Photonic emission was imaged with the In Vivo Imaging System

(IVIS, Xenogen) with a collection time of 5-60 seconds. Tumor bioluminescence was quantified by integrating the photonic flux (photons per second) through a region encircling each mouse as determined by the LIVING IMAGES software package (Xenogen). At the end of 7 days of treatment, the disease was allowed to progress and all mice were monitored daily and animals showing overt signs of disease were euthanized following Institutional Animal Care and Use Committee guidelines.

CCRF-CEM xenograft experiments were performed with 7 to 9-week-old female NOG (NOD/*scid*/IL-2R γ ^{null}) mice (Taconic Farms). We injected 5×10^6 CCRF-CEM cells expressing luciferase via tail vein injection. After a 15 day window for tumor engraftment, we segregated mice with homogeneous tumor burdens into treatment groups (3-4 per group) and treated them daily with vehicle (DMSO), dexamethasone (5 mg/kg via intraperitoneal injection), MK2206 (10 mg/kg via oral gavage twice a day) or dexamethasone (5 mg/kg) plus MK2206 (10 mg/kg) for three days. We evaluated disease progression and therapy response by bioluminescence (see above), luciferase activity on isolated tumor cells and by flow cytometry (CD45 staining).

For the transduction of primary T-ALL cells, freshly thawed primary T-ALL cells or tumor cells obtained from xenografts were infected with lentiviral particles expressing the red fluorescent protein CHERRY and luciferase (FUW-CHERRY-puro-LUC) by single spinoculation on retronectin coated plates. Twenty-four hours after transduction primary cells were injected intravenously into NOG recipient mice. Tumor bearing mice showing engraftment of luciferase expressing

ALL cells were euthanized and primary tumor cells extracted from the spleens of leukemic mice. Subsequently, $5-8 \times 10^6$ TALL cells were injected intravenously in NOG recipient mice. Leukemia progression was assessed by flow cytometry of mouse peripheral blood using anti-CD45 antibodies and bioimaging. When >30% human cells were detectable in blood and saturating photon emission was recorded with a collection time of 1 minute, animals were randomized into 4 treatment groups (n=4-5) and treated daily with vehicle (DMSO), dexamethasone (5 mg/ kg via intraperitoneal injection), MK2206 (10mg/kg via oral gavage twice a day) or dexamethasone (5 mg/kg) plus MK2206 (10 mg/kg) for 4-5 days. We evaluated disease progression and therapy response after 3-5 days of treatment by bioluminescence.

7.6.19 Statistical analyses

We performed statistical analysis by Student's *t*-test. We considered results with $P < 0.05$ as statistically significant. We analyzed drug synergism using the median-effect method of Chou and Talay [212] and used the CalcuSyn software (Biosoft, Great Shelford, Cambridge, UK) to calculate the combination index (CI) and perform isobologram analysis of drug interactions. CI values below 1, equal to 1, and above 1 represent synergism, additivity, and antagonism, respectively. The isobologram is formed by plotting the concentrations of each drug required for 50% inhibition (ED50) on the x-and y-axes, respectively, and connecting them to draw a line segment, which is ED50 isobologram. Combination data points that fall on, below and above the line segment represent additivity, synergy, and antagonism, respectively. Survival in animal experiments was represented with

Kaplan–Meier curves and significance was estimated with the log-rank test (Prism GraphPad).

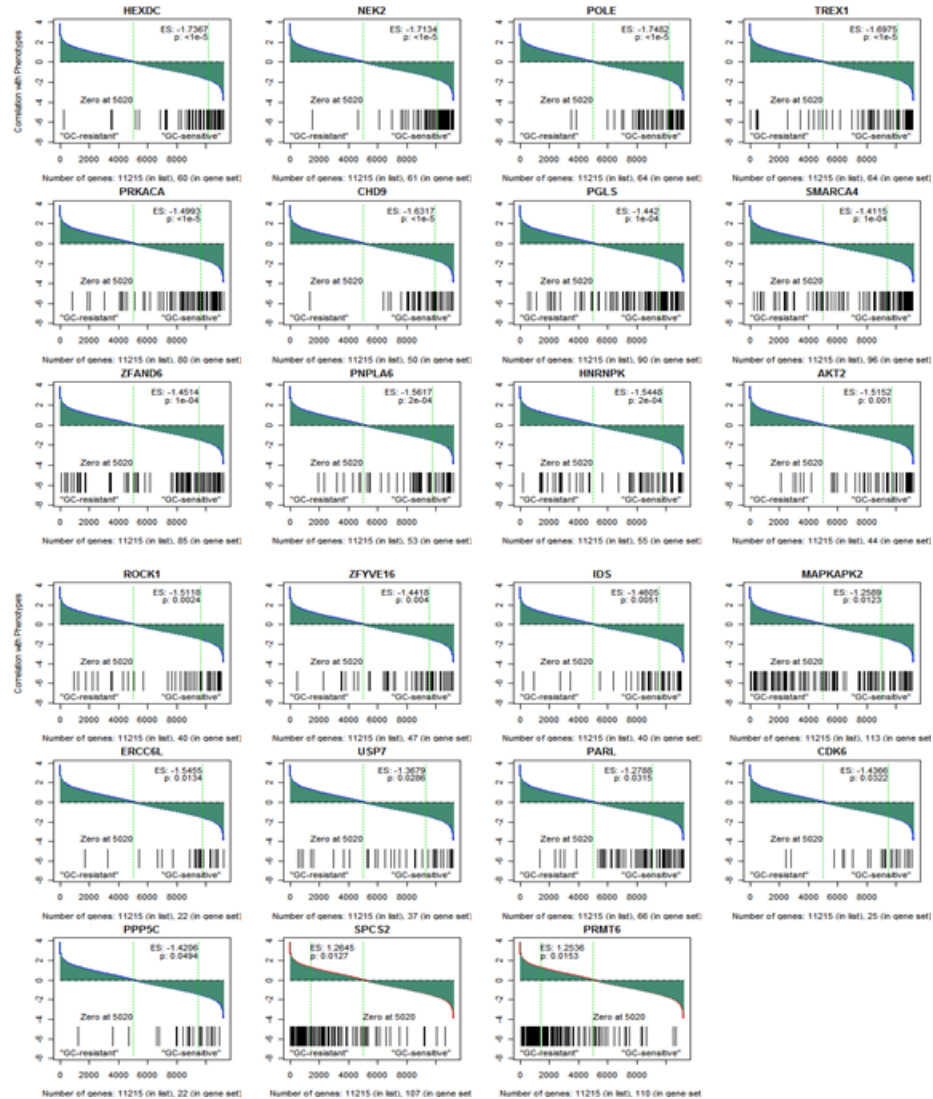


Figure 7-25 NetBID2 results of top signaling drivers of GC-resistance in T-ALL.

Probe Id	Gene Symbol	SetSize (>50)	ES	Pvalue
221640_s_at	LRDD	99	-1.96	<1E-3
203683_s_at	VEGFB	153	-1.9	<1E-3
205875_s_at	TREX1	64	-0.98	<1E-3
209053_s_at	WHSC1	188	-1.36	0.001
202513_s_at	PPP2R5D	164	-1	0.001
207163_s_at	AKT1	92	-1.7	0.002
205546_s_at	TYK2	104	-1.45	0.002
203452_at	B3GAT3	173	-1.4	0.002
203777_s_at	RPS6KB2	158	-1.26	0.002
203235_at	THOP1	82	-1.72	0.003
203709_at	PHKG2	125	-1.45	0.003
225471_s_at	AKT2	74	-0.98	0.003
203422_at	POLD1	94	-1.85	0.004
202424_at	MAP2K2	198	-1.73	0.004
203727_at	SKIV2L	67	-1.29	0.004
201407_s_at	PPP1CB	128	1.04	0.004
201598_s_at	INPPL1	60	-0.99	0.004
215054_at	EPOR	53	0.71	0.004
212218_s_at	FASN	64	-1.13	0.005
65884_at	MAN1B1	87	-1.13	0.005
213379_at	COQ2	52	1.12	0.005
214048_at	MBD4	98	0.95	0.005
202362_at	RAP1A	169	0.93	0.005
226372_at	CHST11	135	0.93	0.005
203952_at	ATF6	109	0.79	0.005
204267_x_at	PKMYT1	102	-1.67	0.006
210621_s_at	RASA1	55	1.61	0.006
218619_s_at	SUV39H1	96	-1.58	0.006
200041_s_at	BAT1	80	-1.21	0.006
212983_at	HRAS	108	-1.2	0.006
222808_at	ALG13	144	1.15	0.006
205212_s_at	ACAP1	176	-1.51	0.007
202253_s_at	DNM2	142	-1.13	0.007
1552474_a_at	GAMT	101	-1.57	0.008
219017_at	ETNK1	132	1.4	0.008
1555613_a_at	ZAP70	142	-1.15	0.008
202789_at	PLCG1	76	-1.04	0.008
201251_at	PKM2	74	-0.76	0.008
218759_at	DVL2	124	-1.45	0.009
204012_s_at	LCMT2	67	1.14	0.009
205140_at	FPGT	60	0.83	0.009
228667_at	AGPAT4	105	0.75	0.009

Table 7-1 All predicted signaling drivers of GC-resistance by NetBID2 with P<0.01, set size > 50, being involved in >= known pathway.

7.7 GC-Responsive Signature of After vs. Before Treatment to Sensitive T-ALL Patients

In addition to the expression profiles of GC-resistant and sensitive primary patients before treatment which we used to identify AKT1 as a signaling driver of resistance by NetBID2, we also had microarray data of GC-sensitive T-ALL patients before and after treatment at 6h or 8h and 224h [213]. With this information, we generated a signature of GC response by coming expression change at 6h or 24h vs. 0h (Figure 7-26). And then we applied NetBID2 to identify master regulators or signaling modulators that control GC-responsive signature genes. The drivers of early (6 or 8h) and late (24h) response to glucocorticoid show signature overlaps (Figure 7-27).

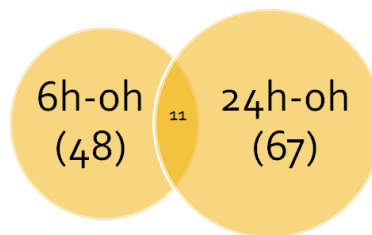


Figure 7-26 Summary of GC-Responsive signature at 6h and 24h.

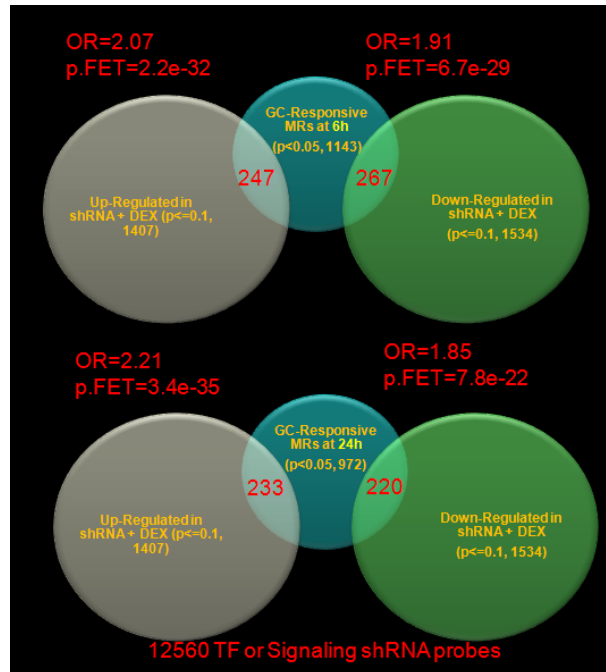


Figure 7-27 Comparison of drivers regulating or modulating early (6 or 8h) and late (24h) responsive signature genes in GC-sensitive T-ALL patients. Sign of drivers is taken into consideration. Fisher's exact test is used to the significance of overlaps.

We also checked the overlap of GC-responsive drivers, using signature treatment after 6h or 24h vs. before treatment in sensitive patients, with GC-resistant drivers, using signature of resistant vs. sensitive patients before treatment. Interestingly, there is a significant overlap between resistant drivers with early (6h) responsive drivers, but not with late (24h) responsive drivers (Figure 7-28). This may suggest that drivers that are involved in resistance mechanism are enriched in early responsive master regulators.

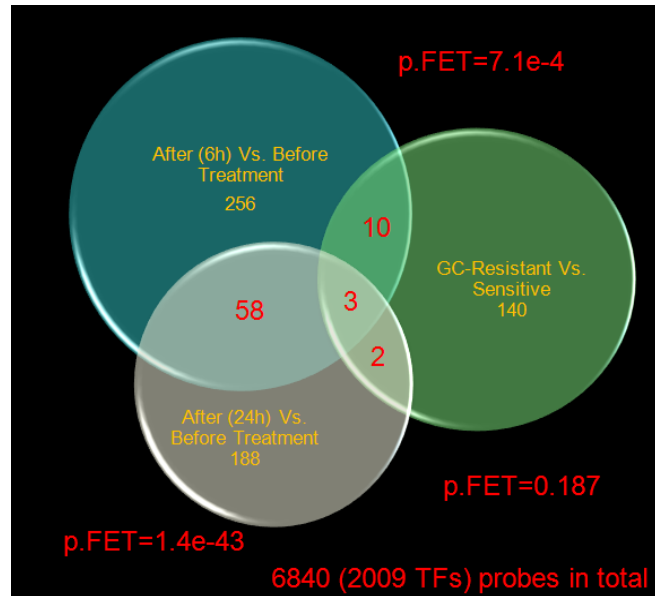


Figure 7-28 Overlap of TF master regulators for GC-Resistance and GC-Response (6h or 24h). Fisher's exact is used to test overlap significance.

7.7.1 NetBID2 identifies AKT1 as driver of GC-responsive signature

NetBID2 identifies AKT1 ($P < 0.001$) and AKT2 ($P = 0.024$) as drivers of early (6 or 8h) GC-responsive signature genes (Figure 7-29), but not as drivers of late (24h) responsive signature genes. This confirms the role of AKT1 or entire AKT pathway in glucocorticoid regulatory signaling pathway, which might explain its abnormal activation causing GC-resistance.

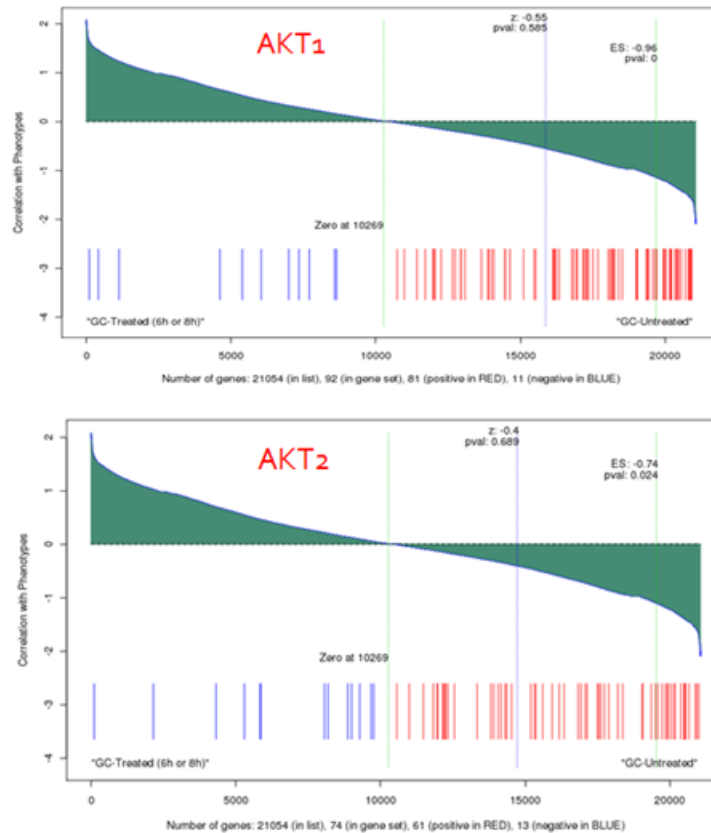


Figure 7-29 NetBID2 identifies AKT1 ($P < 0.001$) and AKT2 ($P = 0.024$) as drivers of early (6 or 8h) GC-responsive signature genes, but not as drivers of late (24h) responsive signature genes.

7.8 Preliminary Results of Crossing Signaling Drivers with RNAi Screens

In the next chapter, we will discuss the integration of NetBID2-predicted drivers with functional RNAi screening to identify potential therapeutic targets for reversal of GC-resistance, however, in the next chapter we will only focus on validation of transcription factor-type drivers while this chapter focuses on signaling drivers. And actually it's more interesting to cross signaling modulators

with shRNA-screened candidates because signaling molecules tend to be druggable. We checked the overlap of NetBID2-inferred signaling drivers with RNAi screening of two resistant T-ALL cell lines (Table 7-3) and selected candidates are to be validated by our collaborator. Here I only show you partially results relevant to AKT, the major focus of this chapter.

7.8.1 AKT1 doesn't show up from shRNA screening as a candidate

We computationally predicted and validated AKT1 as a therapeutic target to reverse GC-resistance in T-ALL. We asked whether it also shows up from shRNA screening. However, two hairpins targeting AKT1 in GIPZ library demonstrate no significant depletion of sh-AKT1 in the screens of both resistant cell lines. It's even worse that in CUTLL1, it shows some anti-evidence. However, one hairpin targeting AKT2 showed significant depletion in CUTLL1, but no evidence in HPBALL. This might be because the quality of shRNAs targeting AKT1 or AKT2 is not good and might also reflects the noisy nature of shRNA screening data.

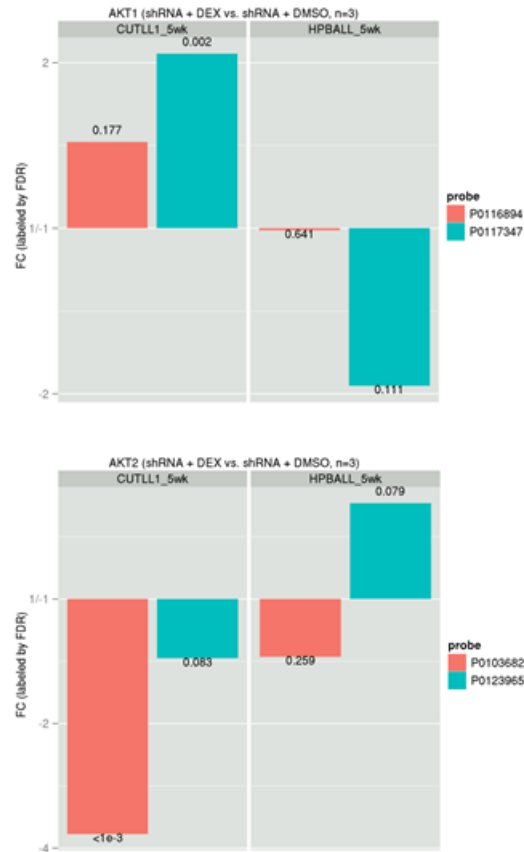


Figure 7-30 shRNA screening results of AKT1 and AKT2 in two GC-resistant cell lines.

7.8.2 PRKAR1A in PI3K pathway shows up in both driver prediction and shRNA screens

PRKAR1A, as shown in Figure 7-31, is a key downstream player of PI3K pathway, which is parallel to AKT to trigger apoptosis pathway. It is predicted by NetBID2 as a driver of GC-resistance and hairpins targeting PRKAR1A shows significant depletion in both resistant cell lines, making it a very interesting therapeutic target to reverse resistance. It might be an alternative to AKT inhibition or has synergistic effects with targeting AKT that needs to be tested out.

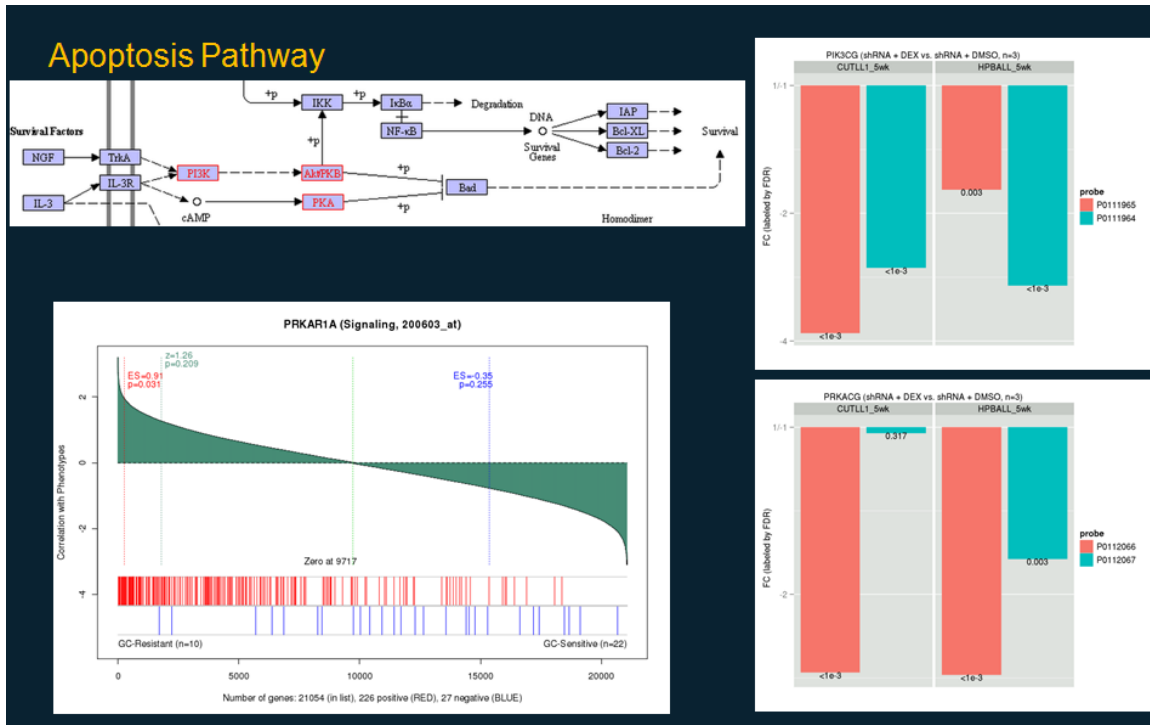


Figure 7-31 PRKAR1A shows up in both NetBID2 prediction and shRNA screens.

geneSymbol	funcType	n.probesA sMR	setSize	GSEA.pvalue	n.levels	z.CU	z.HPB	pval.CU	pval.HPB
HDAC7	Signaling	1	73	0.002	3	-4.19	-0.10	0.000	0.919
RPS6KB2	Signaling	1	158	0.002	1	-2.19	-0.45	0.029	0.652
AKT2	Signaling	2	74	0.003	2	-2.33	0.22	0.020	0.826
IMPAD1	Signaling	1	78	0.003	1	-4.00	-0.23	0.000	0.818
PPP1CB	Signaling	1	128	0.004	3	1.03	-2.33	0.301	0.020
COQ2	Signaling	1	52	0.005	2	-0.69	-2.30	0.491	0.022
PCSK7	Signaling	1	74	0.005	1	-1.97	-1.93	0.049	0.054
MTA1	TF_Sig	2	201	0.006	1	2.35	-5.84	0.019	0.000
SUV39H1	Signaling	1	96	0.006	1	-2.54	-0.73	0.011	0.463
ACAP1	Signaling	2	176	0.007	1	-1.46	-2.74	0.144	0.006
DNM2	Signaling	1	142	0.007	1	-2.30	-0.15	0.021	0.877
WHSC1	TF_Sig	3	308	0.007	5	-2.32	1.30	0.020	0.193
ETNK1	Signaling	2	132	0.008	3	-1.32	-1.87	0.185	0.062
GAA	Signaling	1	50	0.01	1	-2.45	-3.72	0.014	0.000
SMARCA4	TF_Sig	6	436	0.01	1	-5.06	-1.16	0.000	0.248
ATG4B	Signaling	1	79	0.011	1	-3.90	-4.94	0.000	0.000
BMP2K	Signaling	2	36	0.015	1	-2.10	-0.30	0.036	0.763
ITPR3	Signaling	2	98	0.015	2	0.43	-2.54	0.669	0.011
ATP13A1	Signaling	1	63	0.016	1	-2.06	-0.26	0.040	0.796
HMGB1	TF_Sig	2	257	0.016	1	-3.83	-0.85	0.000	0.396
PRKD2	Signaling	2	31	0.016	1	-3.41	0.25	0.001	0.800
TYMS	Signaling	2	147	0.016	2	-3.59	0.17	0.000	0.867
PDK3	Signaling	1	44	0.017	2	-2.46	-0.42	0.014	0.671
DTYMK	Signaling	3	128	0.019	2	-2.73	0.38	0.006	0.702
OAS3	Signaling	1	28	0.022	1	0.67	-1.98	0.505	0.048
DDX52	Signaling	1	77	0.025	8	-4.15	-1.79	0.000	0.074
COASY	Signaling	1	67	0.026	1	1.13	-4.53	0.258	0.000
NEK7	Signaling	1	125	0.026	1	-1.59	-2.51	0.111	0.012
TKT	Signaling	1	93	0.026	1	-2.39	-3.08	0.017	0.002
SMARCA2	Signaling	1	61	0.027	1	0.06	-2.74	0.949	0.006
CTBS	Signaling	1	112	0.029	4	-1.63	-1.65	0.104	0.099
PFKL	Signaling	2	102	0.029	1	3.68	-2.38	0.000	0.017
EDNRB	Signaling	1	34	0.031	1	7.54	-6.25	0.000	0.000
CAPN1	Signaling	1	134	0.034	3	-3.21	2.53	0.001	0.012
CAPN10	Signaling	1	56	0.034	1	-6.24	0.26	0.000	0.798
PA2G4	TF_Sig	1	51	0.035	1	-5.62	-0.06	0.000	0.952
USP48	Signaling	1	33	0.035	4	-2.66	0.75	0.008	0.453
FYN	Signaling	1	96	0.036	1	0.52	-3.06	0.605	0.002
ACLY	Signaling	1	77	0.037	1	-3.01	-0.26	0.003	0.793
CAMK4	Signaling	1	120	0.037	2	-3.29	-0.41	0.001	0.683
IGFBP2	Signaling	1	122	0.037	1	-7.25	0.15	0.000	0.878
NEK2	Signaling	1	88	0.037	1	-5.75	1.04	0.000	0.299
SGMS1	Signaling	1	60	0.038	1	-2.17	-0.91	0.030	0.362
ADCK2	Signaling	1	119	0.04	1	-2.89	1.54	0.004	0.124
TXN	Signaling	1	77	0.041	1	-3.55	-2.43	0.000	0.015
RAD54L	Signaling	1	106	0.043	1	2.38	-8.59	0.017	0.000
MFNG	Signaling	1	122	0.047	1	-5.51	-0.62	0.000	0.538
NAMPT	Signaling	1	53	0.047	2	-2.03	0.98	0.042	0.327
NAT15	Signaling	1	47	0.05	1	5.60	-2.30	0.000	0.022
OAT	Signaling	1	56	0.05	1	2.15	-3.59	0.031	0.000

Table 7-3 Candidates of integrating top signaling of drivers of with RNAi screening results to reverse GC-resistance upon silencing.

Chapter 8 Integrating Functional Genomics with Systems Biology on Discovering Therapeutics to Reverse Glucocorticoid Resistance in T-ALL²

8.1 Summary

Glucocorticoid (GC) resistance is strongly associated with poor prognosis in childhood acute lymphoblastic leukemia. We applied Genome-wide RNA interference (RNAi) screens, a powerful tool for systematic loss-of-function studies, to search for new therapeutic targets to reverse GC-resistance. However, due to high false positive rates of screen data, additional knowledge was needed to select candidates. In this study, we developed an integrative system biology framework, by complementing RNAi screen data with a computational algorithm inferring regulatory drivers of phenotypes, to identify therapeutic candidates for GC-resistant T-cell Acute Lymphoblastic Leukemia (T-ALL). The phenotype driver prediction algorithm (NetBID2) was based on a computationally assembled T-All specific transcriptional network from a large collection of gene expression profiles and Markov chain Monte Carlo based Bayesian modeling techniques. Our framework identified 16 transcription factors, when repressed, sensitize GC resistant cells. Out of 16 candidates, 13 were validated *in vitro*, and 10

² Maria Sol Flaherty from Ferrando lab did the validation experiments.

outperformed positive controls (NOTCH1 and MCL1). Moreover, 75% of computationally predicted drivers demonstrated significant effects on GC-sensitivity *in vitro*. Network analysis of validated targets discovered that they formed three well-connected subnetworks and might work cooperatively to induce resistance. Particularly, we identified TRIM28 as a critical master regulator of GC-resistance and a TRIM28-modulated mechanical regulatory subcircuit that gave insights on potential synergistic therapeutic strategies to rescue GC-sensitivity in T-ALL.

Keywords: glucocorticoid resistance, T-ALL, RNAi screen, regulatory driver, Bayesian, MCMC, systems biology

8.2 Introduction

Glucocorticoids (GCs) play a fundamental role in the treatment of all lymphoid tumors due to their capability to induce apoptosis in lymphoid progenitor cells [18, 19, 149]. Resistance to glucocorticoids is strongly associated with unfavorable prognosis in childhood acute lymphoblastic leukemia (ALL). Majority of ALL patients in relapse show increased resistance to GC-therapy [153, 154, 214]. Different molecular mechanisms have been elucidated for GC-resistance in ALL, including loss-of-function mutations in the glucocorticoid receptor (GR) gene, loss of GR auto upregulation, expression of GR splice variants, and upregulation of antiapoptotic pathways [20, 164, 166, 168, 171, 172, 215-219]. Correspondingly, several therapeutic strategies have been proposed to overcome GC-resistance such as inhibition of MEK, HDAC, mTOR, or NOTCH1 [23, 24, 174-178].

However, due to strong toxicity of existing therapeutics [179], reversal of GC-resistance remains a clinical challenge and new therapeutic strategies are much needed.

Genome-wide RNA interference (RNAi)-mediated genetic screen has emerged as a powerful tool for systematic loss-of-function studies in mammalian cells [50-53]. This technology can be applied to identify genes that form synthetic lethal interactions with glucocorticoids in resistant cells, thus making potential therapeutic targets to overcome GC-resistance. However, due to a high false positive rate arising from high throughput noise and off target effect, additional knowledge and powerful analysis tools are needed.

We have shown that computationally inferred context-specific maps of transcriptional or post-translational molecular interactions from large-scaled gene expression profiles (GEPs) allow the elucidation of cryptic driver proteins whose gain or loss is necessary and sufficient for tumor initiation or progression [70-73]. Such master regulators are more robust than traditional signatures to distinguish phenotypes [69]. Therefore, we suggest that systematic inference of driver-type regulators from genomic data complementing with RNAi screen technology will give a more comprehensive molecular understanding of mechanisms of GC-resistance and provide novel targets for therapeutics.

We developed a framework, NetBID2, as detailed in Chapter 4, to infer disease drivers from gene expression data based on computationally-assembled regulatory networks from a cohort of gene expression profiles (GEPs) and

Markov chain Monte Carlo (MCMC) based Bayesian modeling techniques. Integrating RNAi screens of GC-resistant cells with NetBID2 algorithm identified 16 transcription factors that, upon silencing, sensitize GC resistant T-ALL cells, out of which 13 were validated *in vitro* and 10 outperformed positive controls (NOTCH1 and MCL1). Moreover, 75% of computational-predicted regulatory drivers changed sensitivity of resistant cells *in vitro*. Network analysis of validated targets discovered that they formed three well-connected subnetworks and might work cooperatively to induce resistance. Particularly, we identified TRIM28 as a super master regulator and a TRIM28-centered mechanical regulatory subcircuit that gave insights on potential synergistic therapeutic strategies to rescue GC-sensitivity in T-ALL.

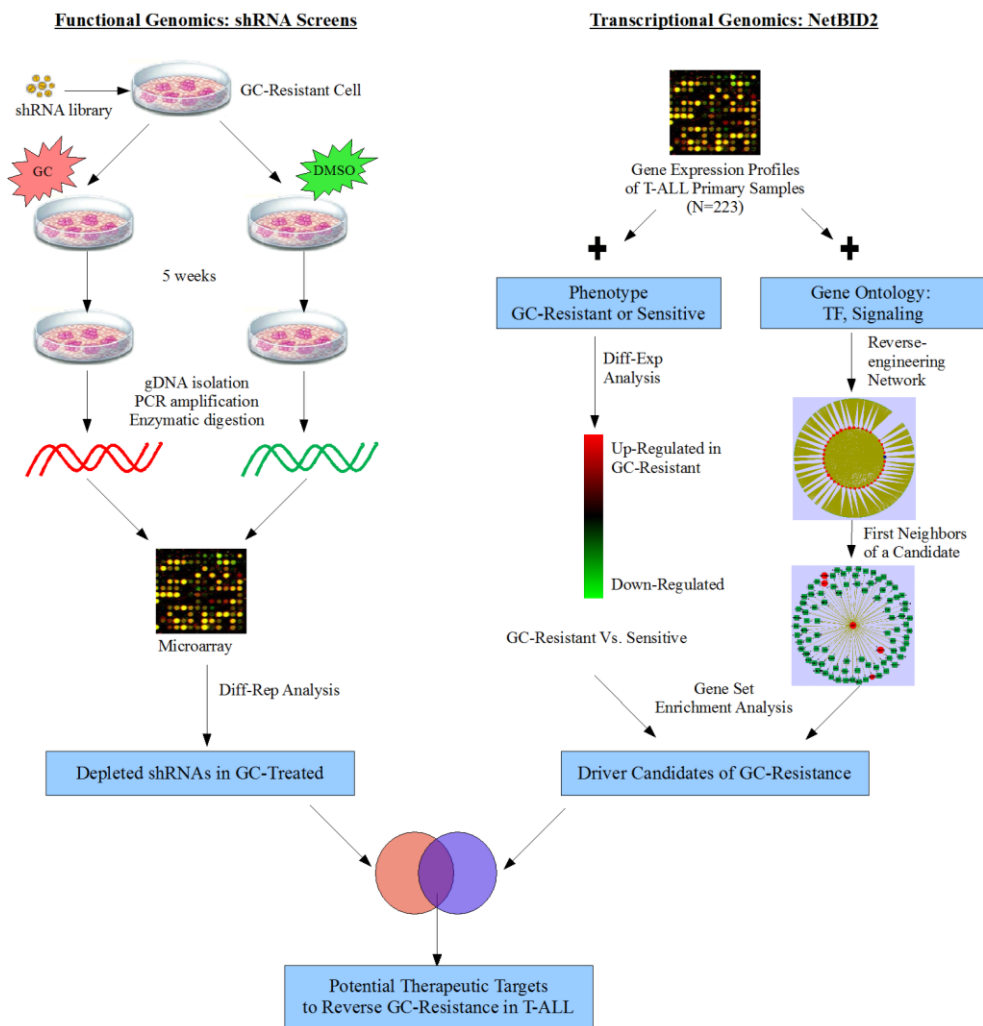


Figure 8-1 The framework integrating genetic RNAi screen with genomic inference of phenotype drivers to identify therapeutic targets for reversal of GC-resistance upon repression.

8.3 Methods

8.3.1 Reverse engineering transcriptional regulatory network of T-ALL

To generate a T-ALL transcriptional network we processed microarray gene expression data (Affymetrix HU133Plus2) of 223 T-ALL primary samples using

GC-RMA normalization and cleaned the dataset to 21,054 probe sets with non-specific filtering. Then we ran the ARACNe algorithm [111] with default parameters against 2007 probes corresponding to 1073 TFs to establish a TF-centered interactome.

8.3.2 Signature analysis of GC-resistance

Out of the 223 samples with GEPs, 22 were diagnosed as GC-resistant and 10 as sensitive. To generate a reference signature of these two phenotypes, part of our regulatory driver inference algorithm, we used a Probit regression model [89] (Figure 8-2) for its advantage of detecting weak effects. Bayesian-MCMC computing was employed to estimate parameters for its robustness and accuracy. In particular, a t-distribution prior and Gibbs sampling were used in this analysis [90].

8.3.3 GSEA of inferring regulatory drivers of GC-resistance

For GSEA method to predict regulatory drivers of GC-resistance, we used a “maxmean” statistic [98] as enrichment score and 1,000 sample permutations to build the null distribution for statistical significance.

Probit Model

$$y_i \sim \text{Bernoulli}(\theta), \quad \theta = \Phi(z_i)$$

$$\text{OR } y_i = \begin{cases} 1 & \text{if } z_i \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$z_i = \beta x + \alpha + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ i = 1, 2, \dots, n$$

Priors

$$\beta \sim (\mu_\beta, \sigma_\beta^2), \quad \sigma_\beta^2 \sim \text{Inv-}\chi^2(\nu_\beta, s_\beta^2)$$

$$\alpha \sim (\mu_\alpha, \sigma_\alpha^2), \quad \sigma_\alpha^2 \sim \text{Inv-}\chi^2(\nu_\alpha, s_\alpha^2)$$

$$\sigma^2 \sim \text{Inv-}\chi^2(\nu, s^2)$$

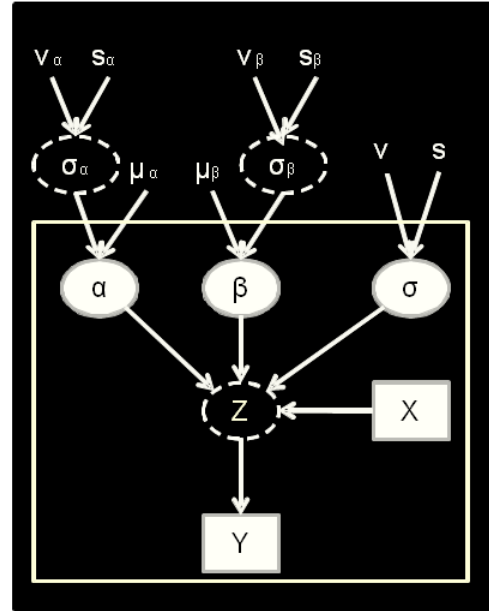


Figure 8-2 Distribution setup (left) and graphical representation (right) of Probit model used for assessing association of phenotypes (GC-resistant or GC-sensitive) with gene expressions. Nodes in solid square are observation variables, in solid eclipse with white background are direct parameters of Probit model, in dashed eclipse are latent variables and the others are hyper-parameters for priors. Y is an indicator variable for phenotypes, X is expression level of gene X, Z is a latent variable in Probit model. Inside the white box is likelihood section, while outside is for priors. Parameters are estimated by a Gibbs sampling procedure.

8.3.4 Pooled shRNA screening

We made use of the shRNAmir library [51], comprising 51,830 shRNAs targeting 12,049 genes. High titer lentiviral pools were prepared and two GC-resistant cell lines – CUTLL1 and HPBALL were infected. The infected cells were selected with puromycin (1ug/ml) for 5 days. For each infected line we treated three cultures with vehicle only (DMSO) and the other three with dexamethasone (DEX

1 μ M). Cell cultures were maintained in exponential growth and in the presence of fresh drug. Genomic DNA was extracted from samples collected after 5 weeks of treatment. PCR amplification was performed on the barcodes associated with each shRNA vector. To ensure homogeneous sampling of the library, 48 individual PCR reactions containing 2 μ g of genomic DNA each were performed. PCR products were gel purified, fluorescently labeled with Cy3 and hybridized in a custom Agilent DNA microarray together with a Cy5-labeled reference sample containing normalized amounts of all barcodes in the library.

8.3.5 Differential representation analysis of individual shRNA

To assess the effects on reversal of GC-resistance by individual shRNA, we compared abundance of shRNA in DEX-treated with DMSO control using a linear model. Bayesian-MCMC procedures were applied to overcome small sample size issue and to obtain robust estimation of parameters.

8.3.6 Integration of multiple shRNAs targeting the same gene

To estimate the overall effects of a gene targeted by multiple shRNAs on GC-sensitivity, we applied a hierarchical modeling approach [89]. This model allowed “random effects” from different shRNAs, and coefficient of ‘fixed effects’ was used to score capability of increasing sensitivity at gene level. Bayesian-MCMC computing was set up for accurate estimations.

8.3.7 Combining differential representation scores of two cell lines

To identify genes that are depleted or enriched in both RNAi-screened cell lines, we used Stouffer's z score method [107] shown in the following formula.

$$Z = \frac{Z_{\text{CU}} + Z_{\text{HPB}}}{\sqrt{2}}, Z \sim N(0,1)$$

For each gene, ZCU and ZHPB were its differential representation scores in CUTLL1 and HPBALL respectively, which followed a standard normal distribution. Combined two-tailed p value was calculated based on integrated Z score.

8.3.8 Silencing by siRNA and cell apoptosis assays

To validate 46 predicted candidates, we used siRNA (Dharmacon) to knock-down testing genes. KOPTK1, a GC-resistant cell line, was electroporated with the siRNAs using the amaxa system (Lonza, SF solution CM 150). After 24 hours of electroporation cells were treated in triplicate either with DEX (100nM) or DMSO. After 48 hours of treatment apoptosis was analyzed by annexin, PI staining (BD Biosciences). NOTCH1 and MCL1 were used as positive controls, and two non-silencing siRNAs as negative controls. Linear modeling similar to individual shRNA analysis of RNAi screen was applied to analyze these apoptosis readouts. Two negative controls were taken average to be compared with candidate genes.

8.4 Results

8.4.1 The framework integrating RNAi screens with regulatory driver inference by NetBID2 identifies sixteen potential therapeutic targets

As shown in Figure 8-1, we developed an integrative framework to identify driver-type therapeutic targets to overcome GC-resistance in T-ALL. First, we performed genome-wide, pooled short hairpin RNA (shRNA) screening on resistant T-ALL cell lines exposed to GCs. This negative genetic screen mainly aimed to identify under-represented shRNAs in GC-treated cells, whose targeting genes increased GC sensitivity. Second, we studied GEPs of large-sampled primary T-ALL patients to build a Transcription Factor (TF) centered T-ALL regulatory network using a well-developed algorithm, ARACNe [111]. Then we utilized phenotypic information, i.e. GC-resistant or sensitive, to perform signature analysis studying association between gene expression and GC-resistance. Instead of identifying classical signature genes that were not robust to characterize phenotypes [69], we developed an algorithm to discover uplevel regulatory drivers of GC-resistance. Our reasoning was that if a TF induces GC-resistance, its regulons inferred from the network should be enriched either among overexpressed or underexpressed genes in GC-resistant samples or both. Gene Set Enrichment Analysis (GSEA) was used for this analysis. Subsequently, we overlapped depleted genes in shRNA screen with inferred regulatory drivers to generate a shorter list of candidates that reverse GC-resistance in T-ALL.

Genome-wide shRNA screens on two resistant T-ALL cell lines (CUTLL1 and HPBALL) identified 1,900 genes that were significantly depleted ($P < 0.05$) in at least one cell line. Based on GEPs of 223 primary samples [220], we obtained a transcriptional interactome centered by 1,073 TFs (2,007 probe sets) comprising 21,035 transcripts and 373,327 interactions. Our regulatory driver inference algorithm yielded 126 TFs showing significant evidences (set size > 40 , $P < 0.05$) as master regulators of controlling signature genes of GC-resistance, out of which 81 had data in shRNA screens. Finally, by crossing the two candidate sets (Figure 8-3-A), we obtained 16 regulatory drivers (Table 8-1) as potential therapeutic targets with the potential to reverse GC-resistance in T-ALL when silenced.

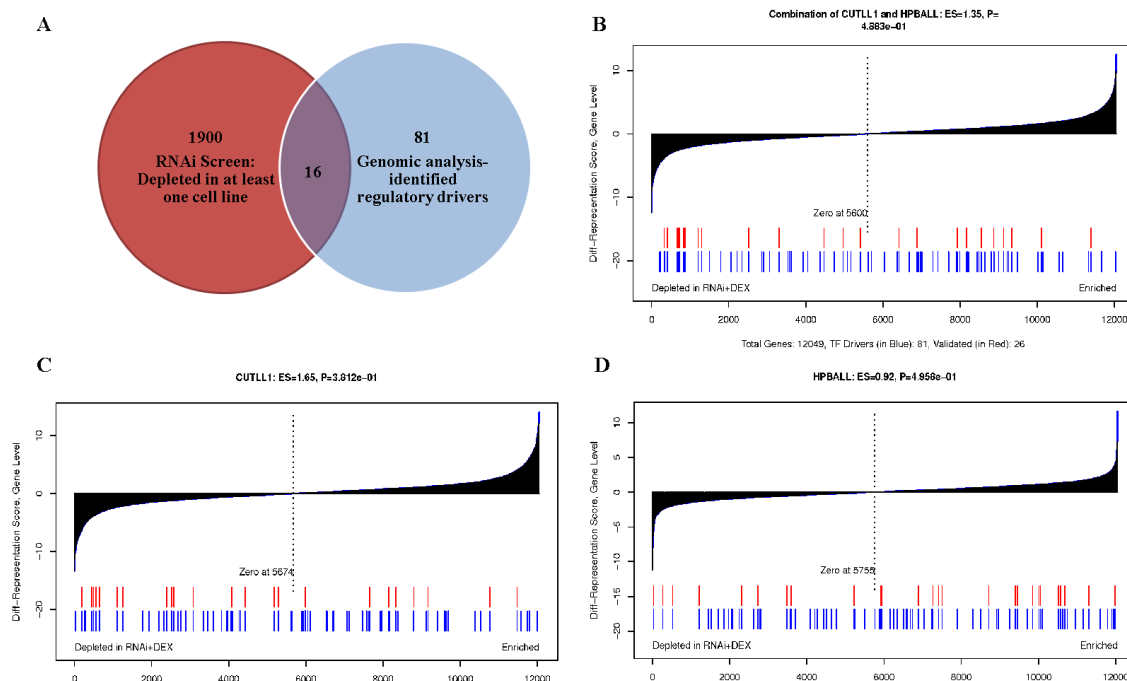


Figure 8-3 Summary of candidates from RNAi screening or genomic inference.

(A) Only 81 out of 126 genomics-analysis identified regulatory drivers have

shRNAs in RNAi screening. B-D display the distribution of 81 TF drivers (blue bars) and 26 validated ones (red bars) in RNAi screening results. All 12,049 genes are ranked from most depleted (left) to most enriched (right) using differential representation score (z score) at gene level in combination of two cell lines (B) or individual cell line (C-D). Similar summary by considering only TF genes in RNAi screening is shown in Figure 8-4.

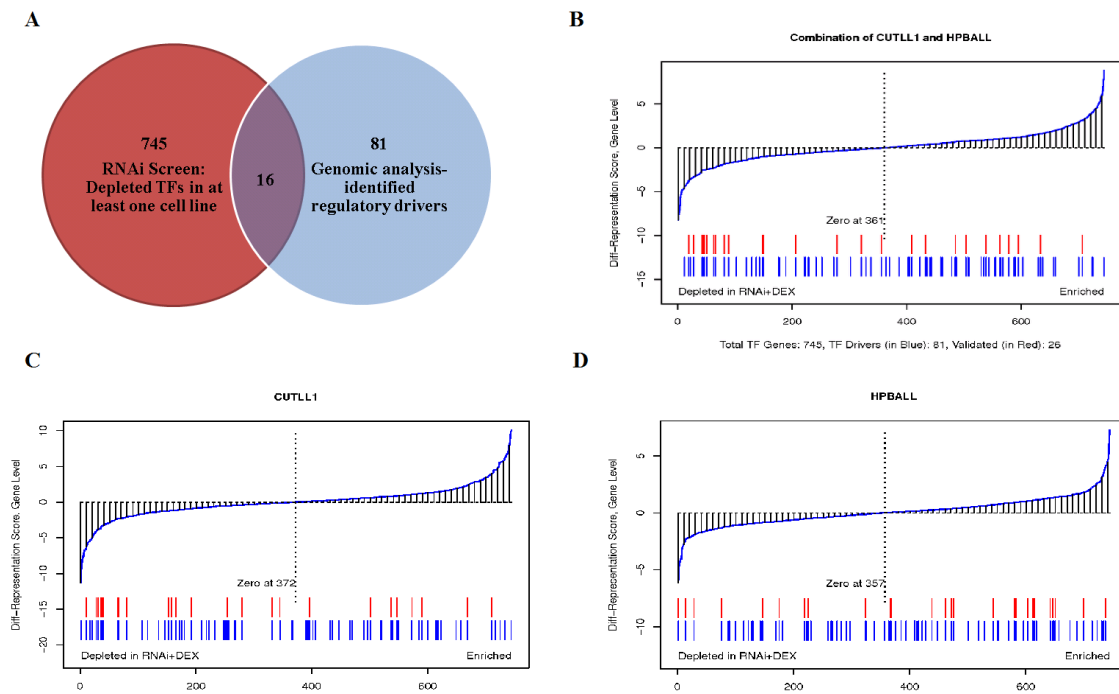


Figure 8-4 Similar summary with Figure 8-3, but considering only TF genes in RNAi screen.

8.4.2 13 of 16 candidates, when repressed, reverse GC-resistance *in vitro*

To validate these 16 candidates, we performed *in vitro* experiments by knocking down candidate genes in resistant cells and measuring cell apoptosis after being exposed to glucocorticoids. Good targets would rescue GC-sensitivity and thus

increase cell death of resistant cells. As summarized in Figure 8-5-A, all candidates including controls were ranked by the capability to reverse GC-resistance. Remarkably, 13 out of 16 showed significantly increase in apoptosis of resistant cells. Moreover, 10 out of 13 genes were more effective than positive controls, NOTCH1 and MCL1, that were previously shown to reverse GC-resistance [24, 175]. Additionally, as shown in Figure 8-5-B, top candidates (CC2D1A, WHSC1, ZHX2) showed up to 15% increase in apoptosis comparing to negative controls. All 16 predicted targets showed consistent directions with inference from RNAi screen, i.e. increase sensitivity when repressed.

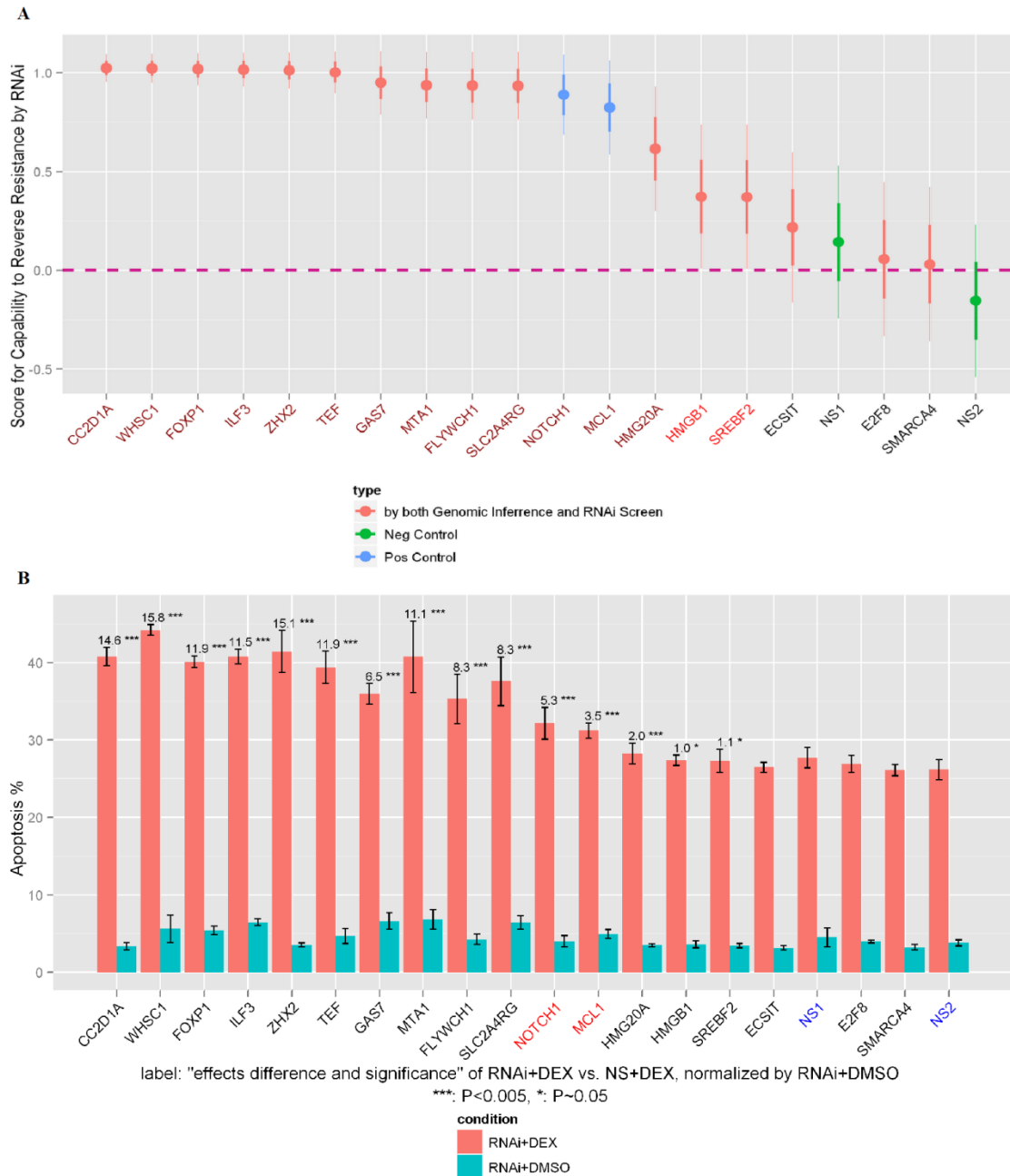


Figure 8-5 Validation results *in vitro* of 16 overlapped candidates. (A) 16 candidates (in red) together with positive controls (in blue) and negative controls (in green) are ranked by the score (central dot) for capability to reverse GC-resistance upon silencing with uncertainty (range line crossing the central dot, thick line for one standard deviation, thin line for two standard deviations corresponding to 95% confidence interval). The color of candidate label on x axis

is associated with calibrated p-value: dark red for $P < 0.005$, red for $P \approx 0.05$. (B) Bar plots of apoptosis level induced by combined treatment of RNAi with DEX (in red), and control, RNAi with DMSO (in light blue) for 16 predicted candidates, positive controls (labeled in red) and negative controls (labeled in blue). All genes are ranked the same as in panel A. The label on top of bar plot represents the increased apoptosis level of candidate gene comparing with average of negative controls (normalized by its own DMSO control and averaged over triplicates) and associated statistical significance level (***) for $P < 0.005$, * for $P \approx 0.05$.

This result with a prediction rate of 81.3% confirmed the power of our framework integrating RNAi screen with systematic analysis of genomic data to identify therapeutic targets. If we only used pooled shRNA screening and ranked all genes by ability to sensitize resistant cells from most depleted to most enriched in GC-treated case (Figure 8-3-B, C, D), the top three candidates validated *in vitro* only ranked 89th, 296th, 595th respectively (Figure 8-3-B), and among top 10 validated candidates, only two ranked within top 50, more precisely, in 44th and 46th. This suggested that high-throughput RNAi screen itself might not be sensitive and accurate enough to discover positive therapeutic targets, and our regulatory driver inference from genomic data is critical to complement it.

Symbol	V	Pooled shRNA Screens										Genomic Analysis					
		# ¹	zCU	P.CU	zHPB	P.HPB	zComb*	P.Comb*	Rank ¹	Rank ²	# ³	Size	ES	P	Rank ²	Z	Signature (Resistant vs. Sensitive)
CCND1A	Y	1	-3.59	3.27E-04	1.10	2.71E-01	-1.76	7.80E-02	1294	89	1	161	-0.55	4.70E-02	116	-0.81	4.20E-01
WHS01	Y	5	-2.32	2.04E-02	1.30	1.93E-01	-0.72	4.72E-01	3303	206	3	308	-1.07	7.00E-03	26	-2.05	4.01E-02
FOXP1	Y	1	-2.31	2.09E-02	4.01	6.19E-05	1.20	2.30E-01	9339	595	1	126	-0.39	3.50E-02	86	-1.22	2.22E-01
ILF3	Y	2	-1.24	2.15E-01	-1.94	5.21E-02	-2.25	2.45E-02	870	66	2	209	-1.14	2.00E-03	10	-1.83	6.78E-02
ZHX2	Y	1	-3.25	1.15E-03	1.81	7.09E-02	-1.02	3.07E-01	2514	148	1	145	0.98	2.70E-02	65	1.85	6.39E-02
TEF	Y	2	-1.13	2.58E-01	-2.45	1.42E-02	-2.53	1.13E-02	705	46	1	42	0.49	1.40E-02	39	-1.19	2.35E-01
GAS7	Y	3	-2.11	3.50E-02	-0.51	6.12E-01	-1.85	6.44E-02	1206	81	1	92	-0.43	3.70E-02	89	-0.84	4.01E-01
MTA1	Y	1	2.35	1.89E-02	-5.84	5.31E-09	-2.47	1.36E-02	735	51	2	201	-1.34	6.00E-03	23	-1.48	1.39E-01
FLYWCH1	Y	2	-2.29	2.18E-02	-1.32	1.85E-01	-2.56	1.05E-02	682	44	1	78	-0.88	2.00E-02	48	-1.49	1.37E-01
SLC2A4RG	Y	2	-3.31	9.17E-04	0.05	9.63E-01	-2.31	2.08E-02	832	63	1	243	-0.80	2.70E-02	67	-1.64	1.00E-01
HMG20A	Y	1	-6.28	3.44E-10	0.89	3.71E-01	-3.81	1.41E-04	327	19	1	153	0.80	4.80E-02	118	0.65	5.17E-01
HMG1	Y	1	-3.83	1.29E-04	-0.85	3.96E-01	-3.31	9.46E-04	418	28	2	257	-0.72	1.60E-02	42	-0.64	5.21E-01
SREBF2	Y	1	-4.07	4.78E-05	0.39	6.98E-01	-2.60	9.30E-03	661	42	1	183	-0.54	3.50E-02	84	-1.09	2.74E-01
ECSIT	N	1	-10.06	8.28E-24	3.47	5.18E-04	-4.66	3.18E-06	208	11	1	431	-0.90	5.00E-03	20	-2.02	4.30E-02
E2F8	N	1	-5.34	9.11E-08	0.19	8.53E-01	-3.65	2.65E-04	349	22	1	270	-1.36	3.00E-02	72	-1.74	8.12E-02
SNARCA4	N	1	-5.06	4.20E-07	-1.16	2.48E-01	-4.40	1.11E-05	237	12	6	436	-0.56	1.00E-02	31	-0.85	3.95E-01

V: whether candidate is validated to be able to sensitize resistant T-ALL cells by *in vitro* experiment; Y=Yes, N=No.

¹: Number of shRNAs targeting the gene

*: Comb: combined z-score and P value from evidences of two cell lines; CU (CUTLL1) and HPB (HPBALL) by Stouffer's method

§: Number of probes as significant driver candidates (P<0.05)

1. Rank from most under-represented genes in pooled shRNA screening
 2. Rank from most under-represented TF genes in pooled shRNA screening
 3. Rank in driver candidates ordered by P value; absolute value of ES (Enrichment Score)
- z: z-score indicating statistical difference of comparisons
P: P value of statistical significance
P.CU, P.HPB, P.Comb values in Bold indicates P<0.05.

Table 8-1 Overlapped 16 candidates between RNAi screening and genomic inference of regulatory drivers.

8.4.3 75% of top genomics-Inferred drivers show significant effects to change GC-sensitivity in vitro

In addition to validating overlapped candidates with depleted genes in RNAi screen, we also tested computationally identified regulatory drivers of GC-resistance from genomic data, because rescuing sensitivity may be achieved by both repressing and activating genes. We selected top 30 additional inferred drivers (Table 8-2) and performed the same *in vitro* experiments. Twenty three out of thirty genes, when knocked down, showed significant effects to either increase (n=8) or decrease (n=15) apoptosis of GC-exposed resistant cells (Figure 8-6). Out of the eight targets that increased apoptosis, five had no significant evidences from RNAi screen and three were not included in the shRNA library. Among the 15 targets that decreased apoptosis, effect of knocking down ATF6 (Figure 8-8) was consistent with RNAi screen finding while the other 14 genes either had no significant support (n=7) from RNAi screen or had no shRNA targeting them (n=7). The overall prediction rate of our algorithm was 75% by considering all validated genes falling in top-hits.

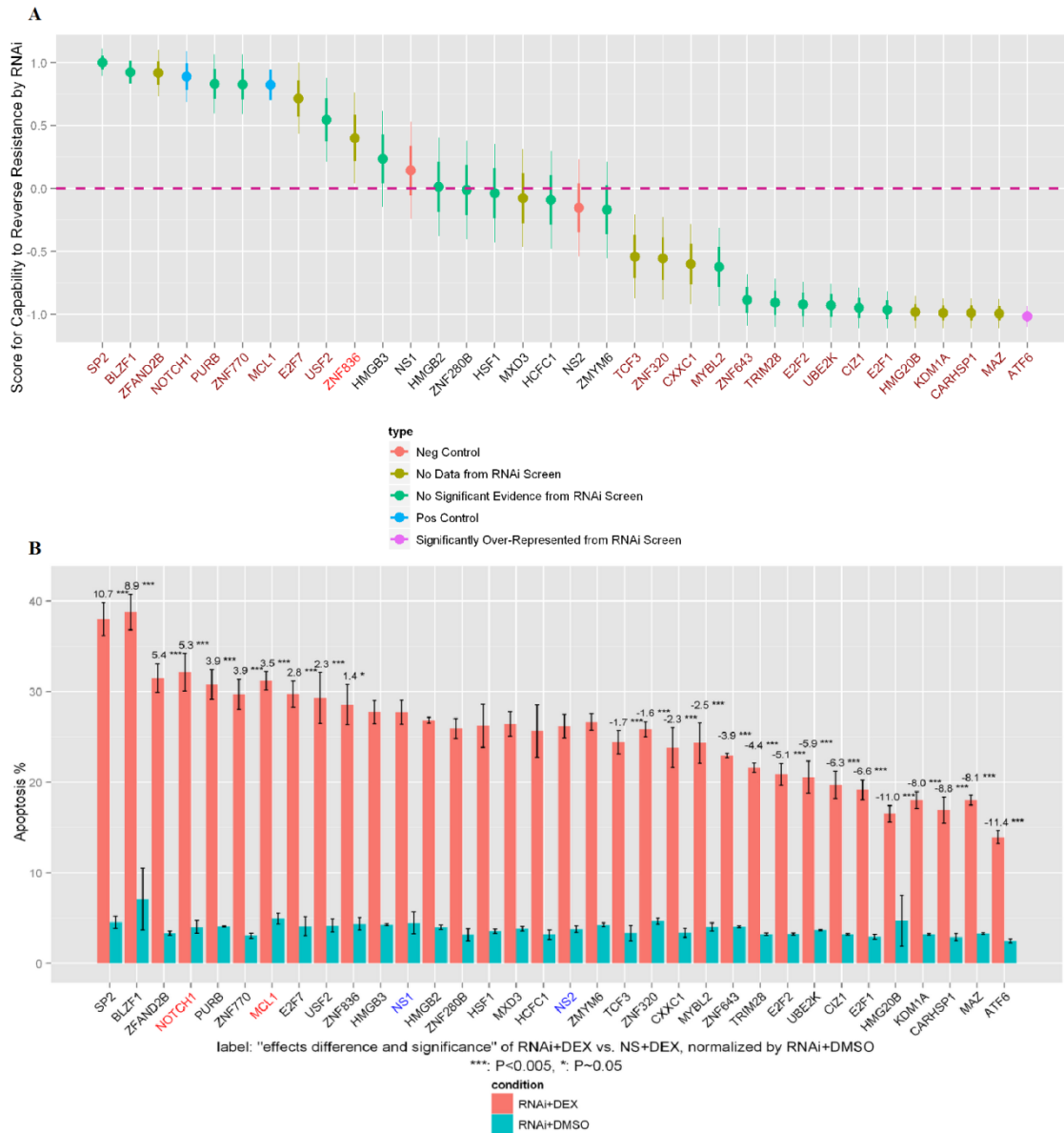


Figure 8-6 Validation results *in vitro* of top 30 additional genomics-predicted regulatory drivers of GC-resistance. (A) 30 candidates classified into no data (in brown-yellow), no significant evidence (in blue) and significantly over-represented (in purple) from RNAi screen, together with positive controls (in blue) and negative controls (in green) are ranked by the score for capability to reverse GC-resistance upon silencing with uncertainty. Extra annotations in panel A and B are the same with Figure 8-5.

8.4.4 Validated targets work cooperatively by forming well-connected subcircuits

Within 36 validated drivers of GC-resistance showing effects *in vitro* on sensitivity, we hypothesized that these key regulators might work cooperatively to induce GC-resistance. To test this hypothesis, we pulled out these candidates and their interactions from our assembled transcriptional T-ALL interactome. As shown in Figure 8-9, these regulatory drivers were separated into three well-connected subgroups denoted as A, B, and C. Size of the node represented their regulating targets in the network, while color indicated that downregulation of this gene increased (blue) or decreased (red) apoptosis of GC resistant cells.

Interestingly, candidate genes that had similar effects, especially increased resistance when silenced, tended to cluster together. For example, in component A, MYBL2, TRIM28, CXXC1 formed a clique and connected closely with E2F1, E2F2, KDM1A, and CIZ1. This indicated that these red nodes worked cooperatively as a key regulatory circuit to repress GC-resistance in T-ALL. This circuit might be a promising targeted unit to overcome GC resistance upon overexpression. Similarly, UBE2K and ZNF320 in component B, TCF3 and HMGB20 in component C represented two additional regulatory units responsible for GC-resistance and might cooperate to induce sensitivity when overexpressed.

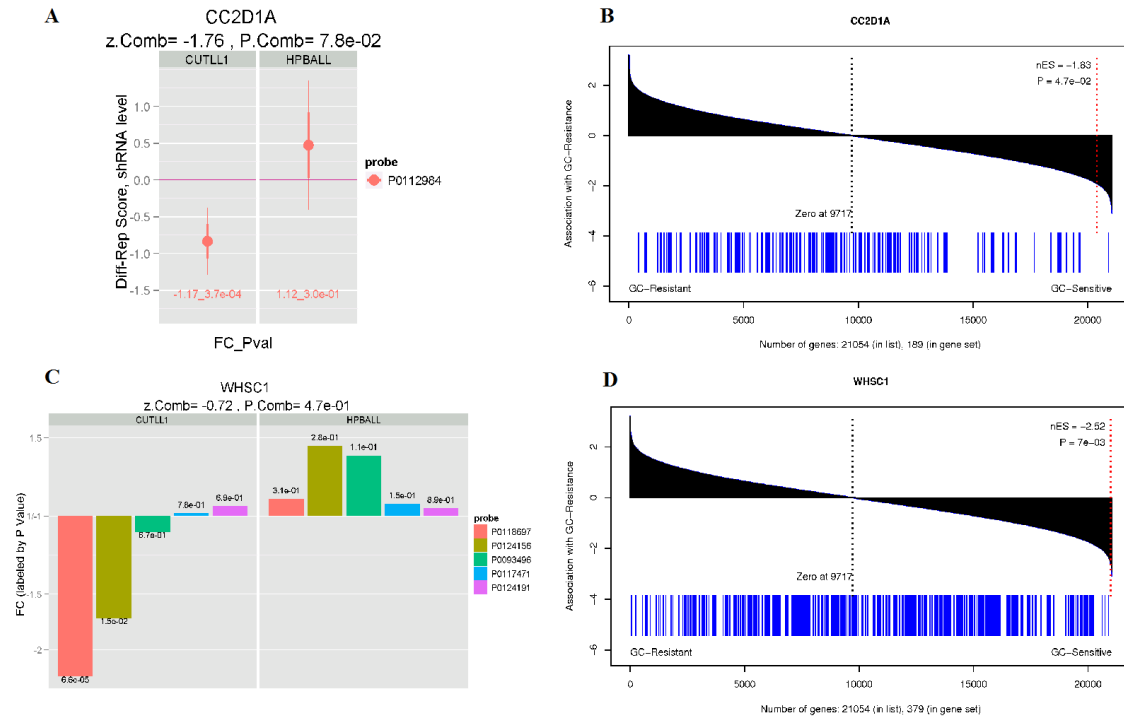


Figure 8-7 RNAi screening results and GSEA plots of CC2D1A and WHSC1, the top two validated targets *in vitro*.

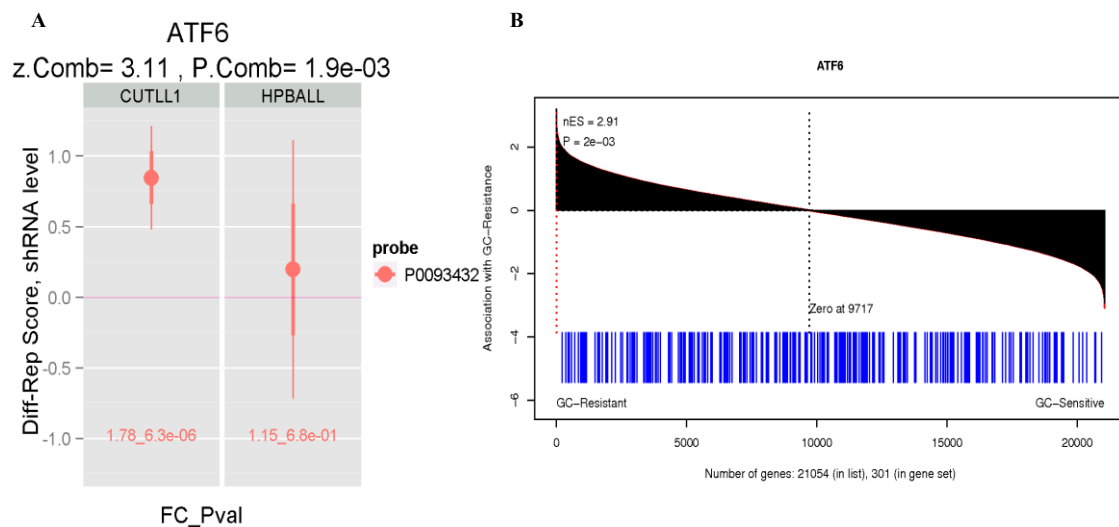


Figure 8-8 RNAi screening result and GSEA plot of ATF6, showing the strongest effect on increasing resistance *in vitro* when silenced.

8.4.5 TRIM28 is a critical master regulator of GC-resistance in T-ALL

The subnetwork of our candidates identified TRIM28 as a critical master regulator of GC-resistance in T-ALL. Firstly, TRIM28 was one of the biggest hub-type regulatory drivers among all candidates. It regulated 377 genes in our inferred transcriptional regulatory network. Secondly, eight of TRIM28 regulating genes (MYBL2, CXXC1, KDM1A, CIZ1 in red, WHSC1, CC2D1A, MTA1, SREBF2 in blue) were also confirmed to reverse GC-resistance, making TRIM28 as the largest hub in this subnetwork. Our finding was confirmed by several studies that TRIM28 epigenetically regulated a broad spectrum of genes and was involved in GR activities [221, 222].

8.4.6 Silencing TRIM28 increases GC-resistance by down-regulating GR

In subnetwork of validated targets (Figure 8-9), all interactions between red and blue nodes except MYBL2-FOXP1 were positively correlated. However, silencing them individually demonstrated opposite effects: This indicated that additional pathways were involved to cause these unexpected effects. One possible mechanism we found was that TRIM28, the critical regulator of GC-resistance, upregulated GR, or was required for GC-induced activities by interacting with GR. As shown in Figure 8-10, we recovered direct interactions between TRIM28 and NR3C1 or TRIM28 and BUD31 that were removed falsely by Data Processing Inequality in ARACNe algorithm [111]. It was observed that TRIM28 activated GR via a feed forward loop with BUD31. Validation results suggested that repressing TRIM28 or its co-activators or upstream regulators such as MYBL2

and CXXC1 would be sufficient to suppress GR expression, and therefore further reducing the sensitivity of resistant cells.

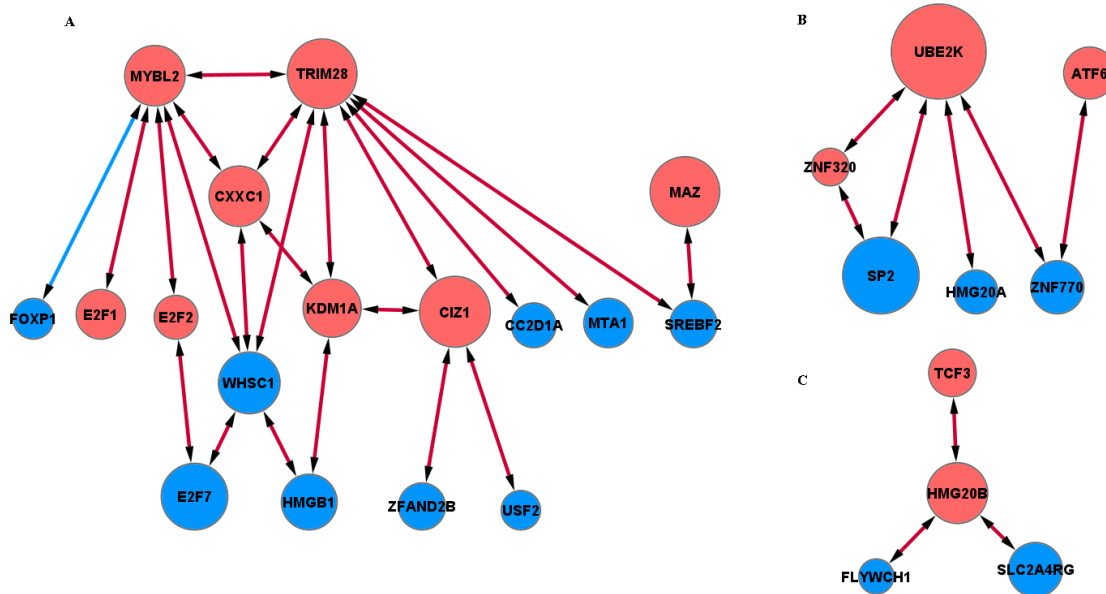


Figure 8-9 Subnetwork from T-ALL interactome of candidates that are validated *in vitro* either to increase (blue nodes) or decrease (red nodes) sensitivity when silenced. A, B and C are three well-connected components covering only direct interactions among these candidates. Nine isolated effective candidates that have no direct interactions with other candidates are not shown. The size of node is proportional to the size of its regulons or first neighbors from T-ALL interactome. Edge in red is for positive correlation of two interactants, while blue for negative correlation.

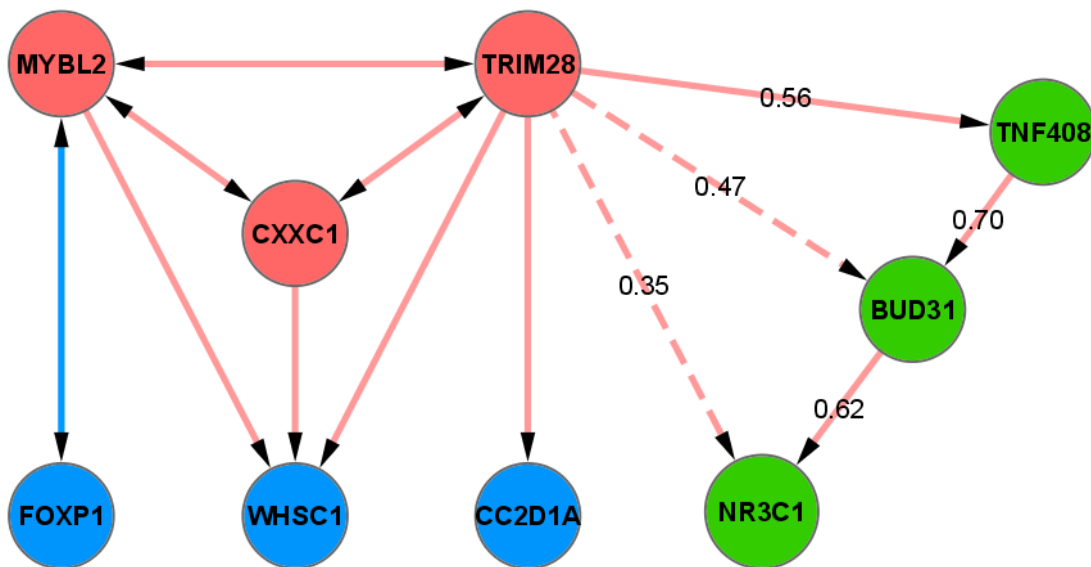


Figure 8-10 A TRIM28-centered subnetwork for a novel mechanism of GC-resistance and a synergistic strategy to overcome resistance. The left square-like part is extracted from Fig 5.1 including the top 3 best validated targets (blue nodes) and the clique of TRIM28 (red nodes). The right triangle-like part is from T-ALL regulatory network illustrating the mechanism of TRIM28 upregulating NR3C1 via a feed forward loop. Number on edge represents the mutual information between expression levels of two interactants. Dashed edges are recovered from false removals by DPI in ARACNe³⁵ algorithm.

Considering the top three most effective targets CC2D1A, WHSC1 and FOXP1, we inferred a transcriptional subcircuit from T-ALL interactome and validation results (Figure 8-10). We removed the direction from WHSC1 or CC2D1A (Figure 8-7) to the triple-clique of TRIM28 based on the observation that repressing CC2D1A or WHSC1 had no effects to upregulate GR. This subnetwork might represent a novel mechanism of GC-resistance in T-ALL and a promising

therapeutic target. In particular, inhibition of CC2D1A, WHSC1 or FOXP1 already demonstrated extraordinary capabilities to sensitize resistant cells.

8.5 Discussion

We demonstrated that integration of genome-wide RNAi screen and our computational framework to infer regulatory drivers of phenotypes was a powerful strategy to discover novel therapeutic targets. Out of 16 overlapped candidates from two approaches, 13 showed significant effects on reversal of GC resistance upon repression *in vitro*. Among them, 10 demonstrated stronger capabilities to sensitize resistant cells comparing to previously discovered targets. Moreover, 75% of top predicted drivers (Figure 8-11) showed effects on changing drug sensitivity when silenced. Network topology of all 36 effective targets identified three well-connected regulatory subcircuits that might shed lights on new mechanisms of GC-resistance and novel pathways as therapeutic targets (Table 8-5).

RNAi screening itself usually gave a long list of candidates with a high false positive rate due to off-target effects, small sample size, and noise of microarray measurements. For example, our screens identified 1,900 candidates at gene level (5,851 at individual shRNA level for 4,783 genes) significantly depleted ($P < 0.05$) in at least one cell line (Figure 8-3-A, Table 8-4). Sophisticated statistical approach such as Bayesian-MCMC method, or multiple test correction and stringent threshold would not solve the problem. The top 5 candidates from *in vitro* experiments ranked 89th, 206th, 595th, 66th, and 148th respectively in the

screen result starting from the most depleted gene, and the best rank of working candidates was 19th (Figure 8-3, Table 8-3). Moreover, 8 predicted regulatory drivers from genomic analysis showed significant effects to reduce resistance, but there was no evidence from RNAi screen. Besides, shRNAs in the screening library did not target all the genes in human genome. For example, out of 23 candidates validated *in vitro*, 10 had no shRNAs in the library. Overall, we confirmed that RNAi screen might not be sensitive enough to work alone (Figure 8-3-B, C, D).

Our inferred regulatory drivers of GC-resistance by NetBID2 showed much higher robustness and predictability than signature genes as therapeutic targets. Out of 21 effective candidates, only one gene-ZNF770, fell in the signature list ($P < 0.05$) of being overexpressed in GC-resistant samples (Table 8-2, Table 8-3).

There were limitations of our algorithm. For example, we did not identify NOTCH1, which was previously shown to reverse GC-resistance upon inhibition [24]. One reason might be that our framework assumed activity of a TF could be inferred from its transcriptional expression. However, NOTCH1 transcriptional level did not correlate to protein expression [181]. RNAi screen result did not identify NOTCH1 either. Secondly, we only focused on TFs in this study, while other types of therapeutic targets such as signaling molecules or anti-apoptotic proteins were not included. Signaling proteins might be interesting to us in the future due to their potency as drug targets. We also tested enrichment of our TF drivers in RNAi screen-identified candidates by Fisher's exact test and GSEA

(Figure 8-3, Figure 8-4), but there was no statistical significance (all $P_s > 0.1$). This suggested that other functional groups should be considered for seeking therapeutics.

Moreover, our approach could infer key regulators for GC-resistance but could not accurately predict the direction in relation to convert GC resistant phenotype. RNAi screen would show the correct directions, thus integrating two approaches would provide a much more powerful tool to predict therapeutic targets.

From the subnetwork of effect candidates (Figure 8-9), we observed that all blue node targets were well separated from red ones by sitting either upstream or downstream. This suggested these starting or ending blue proteins might modulate different subprograms that contributed to diverse mechanisms of GC-resistance in T-ALL, thus shedding light on multiple therapeutic strategies to reverse GC-resistance. For example, a subnetwork of WHSC1, E2F7, and HMGB1 in Figure 8-9-A might be an interesting therapeutic targeting program to sensitize resistant cells by RNAi.

Our network analysis identified TRIM28 as a hub master regulator of GC-resistance in T-ALL. Its centered subcircuit might represent a novel resistant mechanism. We showed that knock-down of TRIM28 induced more resistance due to subsequent downregulation of NR3C1 thus its low expression causing low production of GR that was required by glucocorticoids to induce downstream apoptosis. This also suggested that TRIM28 was required for GR activities, in consistent with literature studies showing that these two protein interacting with

each other [222]. Similar effects were also observed when silencing two other TRIM28 positively-correlated candidates, MYBL2 and CXXC1. However, overexpressing TRIM28 might not work either to reverse GC-resistance. Upregulation of TRIM28 would overexpress WHSC1 and CC2D1A which were required at low expression to rescue sensitivity. Topology of TRIM28 subnetwork indicated that down-stream players including FOXP1, WHSC1 and CC2D1A were more likely to induce resistance than their upstream regulators, and they may work synergistically to induce resistance. This leads to a potential combination therapy.

We observed dramatic difference between two RNAi-screened cell lines (Figure 8-3-C, D), probably due to heterogeneity of resistance mechanisms. Simply overlapping two cell lines for potential targets would lose a lot of true positives, therefore we considered candidates showing evidence in at least one cell line. Validations on a third cell line confirmed our strategy: 5 out of 13 validated targets including the top 3 came out in only one screened line.

Successful validation in a separate cell line of 75% genomics-inferred regulatory drivers demonstrated the power and robustness of our computational framework. However, the non-validated candidates in the cell line didn't mean that they were not important for GC-resistance. For example, SMARCA4, one of our candidates showing up in both RNAi screening and genomic analysis, was a key component of SWI/SNF complex that mediated chromatin remodeling and was required for GC transcriptional activity *in vitro*. A recent study showed that SMARCA4 was

associated with GC-resistance, but its knock-down only worked on some cell lines to reverse GC-resistance [223], in consistent with our finding. It might be because of its dependence on other factors. Thus searching for cofactors or synergistic therapeutic targets would be needed in future to overcome complicated GC-resistance in T-ALL.

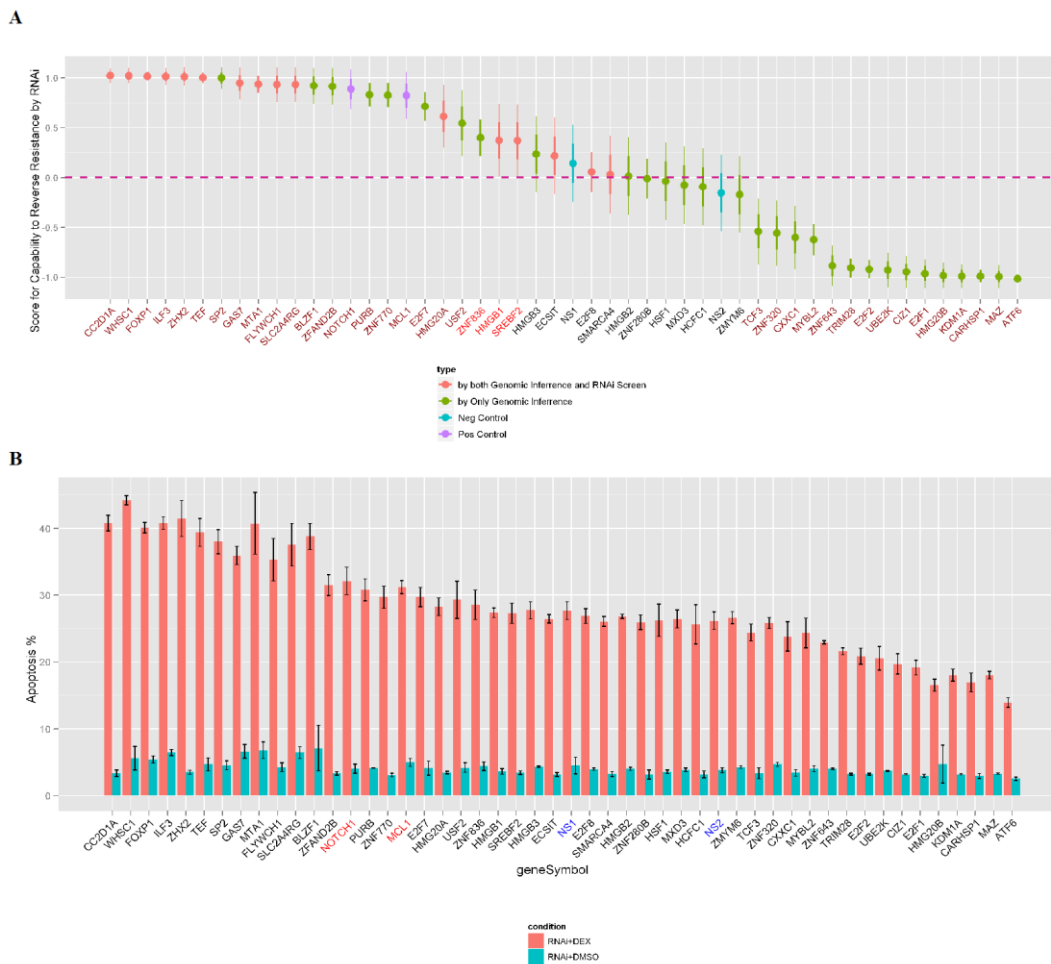


Figure 8-11 Validation results *in vitro* of all 46 selected genomics-predicted regulatory drivers of GC-resistance. (A) 46 candidates classified into no data (in brown-yellow), no significant evidence (in blue) and significantly over-represented (in purple) from RNAi screen, together with positive controls (in blue) and negative controls (in green) are ranked by the score for capability to reverse

GC-resistance upon silencing with uncertainty. Extra annotations in panel A and B are the same with Figure 8-5.

Symbol	V	Pooled shRNA Screens										Genomic Analysis						Signature (Resistant vs. Sensitive)	
		# ^f	z.CU	P.CU	z.HPB	P.HPB	z.Comb*	P.Comb*	Rank ¹	Rank ²	# ^g	Size	ES	P	Rank ³	z	P		
ATT6	Decrease	1	4.03	5.59E-05	0.36	7.15E-01	3.11	1.89E-03	11400	707	1	224	0.55	2.00E-03	12	-0.54	5.87E-01		
SP2	Increase	1	-0.52	6.02E-01	0.04	9.69E-01	-0.34	7.33E-01	4471	278	1	437	0.56	2.10E-02	52	-0.28	7.81E-01		
MAZ	Decrease										2	378	-1.4	3.00E-03	13	-1.55	1.21E-01		
CARHSP1	Decrease										1	159	-1.18	2.00E-03	9	-2.16	3.04E-02		
KDM1A	Decrease										1	276	-0.97	1.00E-03	7	-2.05	4.03E-02		
HMG20B	Decrease										2	300	-1.48	2.00E-03	8	-2.02	4.39E-02		
E2F1	Decrease	2	0.82	4.15E-01	0.25	8.01E-01	0.76	4.50E-01	8169	503	2	203	-1.53	<1E-3	2	-2.87	4.09E-03		
CI21	Decrease	2	0.64	5.20E-01	-0.12	9.04E-01	0.37	7.12E-01	6885	433	3	387	-1.28	1.00E-03	5	-1.36	1.75E-01		
UBE2K	Decrease	5	-0.12	9.04E-01	1.39	1.66E-01	0.9	3.71E-01	8551	539	1	596	1.01	<1E-3	3	2.32	2.03E-02		
BLZF1	Increase	2	-0.91	3.64E-01	-0.54	5.93E-01	-1.02	3.08E-01	2515	149	2	119	0.74	<1E-3	4	0.50	6.15E-01		
E2F2	Decrease	1	1.21	2.26E-01	1.04	2.97E-01	1.59	1.11E-01	10105	634	3	152	-0.83	3.00E-03	15	-1.96	5.00E-02		
ZFAND2B	Increase										1	181	-1.09	7.00E-03	25	-2.40	1.63E-02		
TRIM28	Decrease	1	1.06	2.90E-01	-0.73	4.64E-01	0.23	8.18E-01	6413	409	1	377	-1.14	1.00E-03	6	-2.09	3.63E-02		
ZNF643	Decrease	2	-0.41	6.84E-01	0.33	7.39E-01	-0.05	9.58E-01	5420	356	1	53	0.73	3.00E-03	16	2.40	1.64E-02		
PURB	Increase	2	0.88	3.81E-01	0.69	4.88E-01	1.11	2.67E-01	9133	579	1	94	0.36	1.30E-02	35	-0.03	9.79E-01		
ZNF770	Increase	2	-0.15	8.80E-01	1.12	2.64E-01	0.68	4.95E-01	7936	485	1	234	0.85	5.00E-03	21	2.11	3.51E-02		
E2F7	Increase										1	351	-0.86	8.00E-03	29	-1.97	4.88E-02		
MYBL2	Decrease	3	0.1	9.16E-01	1.33	1.84E-01	1.01	3.11E-01	8874	563	1	297	-1.34	3.00E-03	14	-2.64	8.39E-03		
CXXC1	Decrease										2	297	-1.04	6.00E-03	24	-2.09	3.62E-02		
ZNF320	Decrease										3	99	0.62	5.00E-03	22	-0.53	5.94E-01		
USF2	Increase	2	-1.16	2.44E-01	0.91	3.62E-01	-0.18	8.58E-01	4971	321	3	115	-0.46	7.00E-03	28	-2.10	3.57E-02		
TCF3	Decrease										6	185	-1	2.00E-03	11	-2.24	2.51E-02		
ZNF836	Increase										1	71	0.55	4.00E-03	19	-0.71	4.77E-01		
HMGB3	N	1	-0.79	4.27E-01	2.19	2.86E-02	0.99	3.24E-01	8808	556	1	221	-0.8	1.20E-02	33	-1.45	1.46E-01		
ZMYM6	N	3	-1.62	1.05E-01	0.76	4.48E-01	-0.61	5.41E-01	3619	229	1	84	-0.62	4.00E-03	18	0.90	3.68E-01		
HCFC1	N	2	0.89	3.71E-01	0.09	9.26E-01	0.7	4.85E-01	7986	487	1	336	-1.64	<1E-3	1	-2.41	1.59E-02		
MXD3	N										1	195	-1.28	1.10E-02	32	-2.69	7.17E-03		
HSF1	N	1	9.83	8.63E-23	2.63	8.47E-03	8.81	1.25E-18	12033	745	1	184	-1.45	9.00E-03	30	-1.21	2.28E-01		
HMGB2	N	2	-0.62	5.37E-01	-0.28	7.76E-01	-0.64	5.24E-01	3539	221	2	141	-0.95	4.00E-03	17	-1.36	1.75E-01		
ZNF280B	N	1	-0.53	5.97E-01	-0.36	7.21E-01	-0.63	5.31E-01	3576	224	1	171	-0.51	7.00E-03	27	-1.18	2.38E-01		

Note:
V: whether candidate is validated to be able to decrease or increase apoptosis of GC-exposed T-ALL cells when being silenced by *in vitro* experiment, Decrease=Decreasing apoptosis, Increase=Increasing apoptosis, N=No.
For the other symbol annotations, please refer to Table 1 in the paper.

Table 8-2 Additional top 30 computationally-identified regulatory drivers of GC-resistance in T-ALL but with no support from RNAi screens.

Symbol	Genomic Analysis										Pooled shRNA Screens										Rank from	
	Driver Inference					Signature (Resistan Vs. Sensitive)					mshRNAs					ZComb					Rank from Most Under-Rep in Comb	Rank from Most Under-Rep in HP
	nprobes AsMR	setSize	ES	P	FC	z	P	nshRNAs	FC_CU	z_CU	P_CU	FC_HPB	z_HPB	P_HPB	CU_HPB	CU_HP	HPB	HP	TE			
HCFE1	2	336	-1.64	<1.00E-3	-1.35	-2.41	1.59E-02	2	1.17	0.89	3.71E-01	1.02	0.09	9.26E-01	0.70	4.85E-01	7986	487				
E2F1	2	203	-1.53	<1.00E-3	-1.74	-2.87	4.09E-03	2	1.52	0.82	4.15E-01	1.02	0.25	8.01E-01	0.76	4.50E-01	8169	503				
UBE2K	1	596	1.01	<1.00E-3	1.61	2.32	2.03E-02	5	-1.05	-0.12	9.04E-01	1.26	1.39	1.66E-01	0.90	3.71E-01	8551	539				
BLZF1	2	119	0.74	<1.00E-3	1.04	0.50	6.15E-01	2	-1.47	-0.91	3.64E-01	-1.20	-0.54	5.93E-01	-1.02	3.08E-01	2515	149				
CLT1	3	387	-1.28	1.00E-03	-1.35	-1.36	1.75E-01	2	1.30	0.64	5.20E-01	-1.08	-0.12	9.04E-01	0.37	7.12E-01	6885	433				
TRIM28	1	377	-1.14	1.00E-03	-1.21	-2.09	3.65E-02	1	1.61	1.06	2.90E-01	-1.33	-0.73	4.64E-01	0.23	8.18E-01	6413	409				
KDM1A	1	276	-0.97	1.00E-03	-1.30	-2.05	4.03E-02															
HMG20B	2	300	-1.48	2.00E-03	-1.72	-2.02	4.39E-02															
CARHSP1	1	159	-1.18	2.00E-03	-1.53	-2.16	3.04E-02															
ILF3	2	209	-1.14	2.00E-03	-1.35	-1.83	6.75E-02	2	-1.40	-1.24	2.15E-01	-1.99	-1.94	5.21E-02	-2.25	2.45E-02	870	66				
TCF3	6	185	-1.00	2.00E-03	-1.45	-2.24	2.51E-02															
ATF6	1	224	0.55	2.00E-03	-1.04	-0.54	5.87E-01	1	1.78	4.03	5.59E-05	1.15	0.36	7.15E-01	3.11	1.89E-03	11400	707				
MAZ	2	378	-1.40	3.00E-03	-1.26	-1.55	1.21E-01															
MVBL2	1	297	-1.34	3.00E-03	-2.64	-2.64	8.39E-03	3	1.03	0.10	9.16E-01	1.22	1.33	1.84E-01	1.01	3.11E-01	8874	563				
E2F2	3	152	-0.83	3.00E-03	-1.13	-1.96	5.00E-02	1	1.75	1.21	2.26E-01	2.15	1.04	2.97E-01	1.59	1.11E-01	10105	634				
ZNF643	1	53	0.73	3.00E-03	1.25	2.40	1.64E-02	2	-1.54	-0.41	6.84E-01	1.12	0.33	7.39E-01	-0.05	9.58E-01	5420	356				
HMOB2	2	141	-0.95	4.00E-03	-1.12	-1.36	1.75E-01	2	-1.18	-0.62	5.37E-01	-1.09	-0.28	7.76E-01	-0.64	5.24E-01	3539	221				
ZMYM6	1	84	-0.62	4.00E-03	1.05	0.90	3.68E-01	3	-3.32	-1.62	1.05E-01	1.15	0.76	4.48E-01	-0.61	5.41E-01	3619	229				
ZNF836	1	71	0.55	4.00E-03	-1.11	-0.71	4.77E-01															
ECST	1	431	-0.90	5.00E-03	-1.23	-2.02	4.30E-02	1	-8.16	-10.06	8.28E-24	2.08	3.47	5.18E-04	-4.66	3.18E-06	208	11				
ZNF770	1	234	0.85	5.00E-03	1.54	2.11	3.51E-02	2	-1.18	-0.15	8.80E-01	1.18	1.12	2.64E-01	0.68	4.95E-01	7936	485				
ZNF330	1	99	0.62	5.00E-03	-1.02	-0.53	5.94E-01															
MTAI	2	201	-1.34	6.00E-03	-1.29	-1.48	1.39E-01	1	1.74	2.35	1.89E-02	-2.87	-5.84	5.31E-09	-2.47	1.36E-02	735	51				
CXXC1	2	297	-1.04	6.00E-03	-1.38	-2.09	3.62E-02															
ZFAND2B	1	181	-1.09	7.00E-03	-1.32	-2.40	1.63E-02															
WHSC1	3	308	-1.07	7.00E-03	-1.38	-2.05	4.01E-02	5	-1.32	-2.32	2.04E-02	1.19	1.30	1.93E-01	-0.72	4.72E-01	3303	206				
ZNF80B	1	171	-0.51	7.00E-03	-1.21	-1.18	2.38E-01	1	-1.16	-0.53	5.97E-01	-1.13	-0.36	7.21E-01	-0.63	5.31E-01	3576	224				
USP2	3	115	-0.46	7.00E-03	-1.40	-2.10	3.57E-02	2	-1.67	-1.16	2.44E-01	1.16	0.91	3.62E-01	-0.18	8.58E-01	4971	321				
E2F7	1	351	-0.86	8.00E-03	-2.24	-1.97	4.88E-02															
HSP1	1	184	-1.45	9.00E-03	-1.13	-1.21	2.28E-01	1	1.93	9.83	8.63E-23	2.73	2.63	8.47E-03	8.81	1.25E-18	12033	745				
SMARCA4	6	436	-0.56	1.00E-02	-1.23	-0.85	3.95E-01	1	-2.16	-5.06	4.20E-07	-2.37	-1.16	2.48E-01	-4.40	1.11E-05	237	12				
MXD3	1	195	-1.28	1.10E-02	-1.34	-2.69	7.17E-03															
HMOB3	1	221	-0.80	1.20E-02	-1.26	-1.45	1.46E-01	1	-1.13	-0.79	4.27E-01	1.69	2.19	2.86E-02	0.99	3.24E-01	8808	556				
ZGPAT	1	202	-0.83	1.30E-02	-1.26	-1.89	5.93E-02															
PURB	1	94	0.36	1.30E-02	-1.01	-0.03	9.79E-01	2	1.14	0.88	3.81E-01	1.11	0.69	4.88E-01	1.11	2.67E-01	9133	579				
FOXO1	1	472	-1.51	1.40E-02	-2.27	-2.25	2.42E-02															
ZFP30	1	126	0.87	1.40E-02	1.72	1.56	1.20E-01	3	-1.00	-0.02	9.84E-01	-1.30	-1.19	2.33E-01	-0.86	3.91E-01	2896	178				
ZFP106	1	79	-0.71	1.40E-02	-1.41	-1.66	1.66E-02	2	-1.23	-0.45	6.54E-01	1.35	1.14	2.56E-01	0.49	6.26E-01	7294	460				
ZFP106	1	42	0.49	1.40E-02	-1.11	-1.19	2.33E-01	2	-3.67	-1.13	2.58E-01	-2.49	-2.45	1.42E-02	-2.53	1.13E-02	705	46				
ZSCAN5A	1	66	-0.44	1.50E-02	1.07	1.16	2.45E-01	2	1.05	0.58	5.59E-01	1.06	0.28	7.77E-01	0.61	5.40E-01	7700	479				
HMGF	2	227	-0.81	1.60E-02	-1.11	-0.64	5.21E-01															
HMOB1	2	257	-0.72	1.60E-02	-1.10	-0.64	5.21E-01	1	-2.77	-3.83	1.29E-04	-1.37	-0.85	3.96E-01	-3.31	9.46E-04	418	28				
ZSCAN21	1	105	0.91	1.70E-02	1.04	0.35	7.27E-01															
NAT14	1	184	-0.65	1.80E-02	-1.10	-0.42	6.78E-01															
THAP7	1	185	-0.72	1.90E-02	-1.21	-1.69	9.01E-02															
ZBTB48	1	212	-0.59	1.90E-02	-1.26	-2.05	4.02E-02	1	1.13	1.41	1.58E-01	1.08	0.22	8.28E-01	1.15	2.49E-01	9221	587				
ZNF512B	1	88	-0.47	1.90E-02	-1.16	-1.73	8.33E-02															
FLYWCH	1	78	-0.88	2.00E-02	-1.15	-1.49	1.37E-01	2	-1.97	-2.29	2.18E-02	-1.75	-1.32	1.85E-01	-2.56	1.05E-02	682	44				

Table 8-3 Top computationally-inferred regulatory drivers for GC-resistance in T-ALL (P<=0.05).

Symbol	funcType	n.shRN As	FC.CU	z.CU	P.CU	FC.HPB	z.HPB	P.HPB	z.Comb_ CU_HPB	P.Comb_ CU_HPB
C21orf91		1	-2.42	-9.55	1.36E-21	-2.06	-6.81	1.01E-11	-11.56	6.46E-31
KCNK7		1	-4.34	-12.85	9.17E-38	-4.14	-3.50	4.59E-04	-11.56	6.56E-31
TK2		1	-4.46	-7.60	2.96E-14	-1.95	-8.38	5.14E-17	-11.30	1.28E-29
ANKRD43		1	-5.55	-4.50	6.95E-06	-8.86	-11.15	7.44E-29	-11.06	1.95E-28
MYF5		1	-4.20	-10.40	2.46E-25	-1.18	-3.40	6.70E-04	-9.76	1.68E-22
PCDHGB2		1	-2.26	-12.12	8.44E-34	-1.35	-1.11	2.67E-01	-9.35	8.45E-21
SELENBP1		1	-6.79	-12.97	1.89E-38	-1.08	-0.25	8.02E-01	-9.35	9.08E-21
ANK3		1	-3.67	-9.45	3.51E-21	-2.52	-3.44	5.90E-04	-9.11	8.30E-20
C14orf93		1	-2.51	-11.38	5.33E-30	-1.34	-1.39	1.64E-01	-9.03	1.70E-19
GNPTAB		1	-3.00	-9.12	7.58E-20	-2.07	-3.45	5.63E-04	-8.89	6.29E-19
TAC1		1	-8.19	-12.19	3.36E-34	-1.07	-0.17	8.63E-01	-8.74	2.25E-18
COLEC12		1	-5.11	-10.56	4.34E-26	-2.43	-1.77	7.60E-02	-8.73	2.66E-18
MTX2		1	-11.18	-11.28	1.57E-29	-1.35	-0.85	3.97E-01	-8.58	9.66E-18
ITGAL		1	-3.57	-9.99	1.68E-23	-1.52	-1.88	5.97E-02	-8.40	4.62E-17
FOXRED1		1	-2.68	-8.69	3.56E-18	-4.01	-3.18	1.49E-03	-8.39	4.74E-17
SMYD2		1	-2.49	-12.56	3.30E-36	1.14	0.73	4.67E-01	-8.37	5.73E-17
RASGEF1A		1	-2.03	-8.97	2.99E-19	-1.15	-2.86	4.22E-03	-8.37	6.01E-17
FKBP5		1	-4.35	-10.26	1.08E-24	-1.30	-1.55	1.20E-01	-8.35	6.67E-17
DEAF1		1	-5.83	-7.34	2.12E-13	-4.89	-4.44	9.14E-06	-8.33	8.21E-17
ODF3		1	-2.60	-10.32	5.65E-25	-1.60	-1.37	1.70E-01	-8.27	1.37E-16
ZNF44	TF	1	-3.86	-8.64	5.46E-18	-1.23	-3.00	2.68E-03	-8.23	1.81E-16
PFKM		1	-3.50	-5.50	3.73E-08	-2.07	-6.13	8.59E-10	-8.23	1.89E-16
BMP2		1	-3.80	-11.04	2.58E-28	-1.21	-0.53	5.95E-01	-8.18	2.87E-16
COX8A		1	-3.33	-4.01	6.17E-05	-1.64	-7.40	1.33E-13	-8.07	7.17E-16
OTOP2		1	-5.48	-8.57	9.96E-18	-2.39	-2.74	6.08E-03	-8.00	1.21E-15
CBX4		1	-7.44	-10.03	1.17E-23	-1.43	-1.26	2.09E-01	-7.98	1.48E-15
CACNA1F		1	-4.27	-7.57	3.87E-14	-1.56	-3.43	5.94E-04	-7.78	7.37E-15
FNDC3B		1	-4.70	-8.28	1.23E-16	-3.19	-2.45	1.41E-02	-7.59	3.19E-14
EFNA1		1	1.23	0.25	8.03E-01	-4.53	-10.93	8.18E-28	-7.55	4.26E-14
NUPR1		1	-7.84	-6.58	4.63E-11	-1.31	-4.07	4.64E-05	-7.53	4.91E-14
CDCP1		1	-5.93	-8.78	1.67E-18	-3.21	-1.71	8.64E-02	-7.42	1.18E-13
DMD		1	-1.78	-6.52	6.87E-11	-1.40	-3.96	7.45E-05	-7.41	1.23E-13
MRC2		1	-1.44	-11.14	7.60E-29	1.04	0.71	4.76E-01	-7.38	1.63E-13
BCS1L		1	-4.21	-7.47	7.93E-14	-2.48	-2.94	3.30E-03	-7.36	1.83E-13
MOS		1	-1.56	-1.60	1.10E-01	-3.40	-8.80	1.41E-18	-7.35	1.96E-13
ZNF746	TF	1	-3.25	-7.78	7.14E-15	-1.63	-2.61	8.98E-03	-7.35	1.98E-13
ZNF491		1	-5.25	-7.45	9.56E-14	-1.44	-2.94	3.32E-03	-7.34	2.10E-13
RBMS2		1	-4.70	-9.53	1.60E-21	-1.68	-0.79	4.29E-01	-7.30	2.94E-13
FBLN2		1	-2.46	-8.73	2.52E-18	-1.32	-1.59	1.12E-01	-7.30	2.96E-13
CLU		1	-6.60	-8.75	2.05E-18	-3.62	-1.56	1.19E-01	-7.29	3.05E-13
BARX2	TF	1	-7.07	-11.28	1.70E-29	1.86	0.99	3.23E-01	-7.28	3.44E-13
C15orf43		1	-3.27	-12.25	1.60E-34	1.68	1.99	4.62E-02	-7.26	4.01E-13
NUDT6		1	-2.53	-9.01	2.15E-19	-1.26	-1.23	2.17E-01	-7.24	4.47E-13
GSN		1	-5.13	-9.36	7.88E-21	-1.17	-0.85	3.95E-01	-7.22	5.19E-13
JAG1		1	-4.47	-8.28	1.28E-16	-1.68	-1.93	5.33E-02	-7.22	5.27E-13
MRPS26		1	-1.81	-8.40	4.57E-17	-1.25	-1.70	8.89E-02	-7.14	9.28E-13
NDST2		1	-3.45	-8.47	2.42E-17	-1.28	-1.62	1.06E-01	-7.13	9.86E-13
SDHC		1	-3.01	-7.64	2.18E-14	-1.43	-2.40	1.64E-02	-7.10	1.26E-12
AICF		1	-2.50	-10.07	7.21E-24	1.03	0.10	9.21E-01	-7.05	1.75E-12

Table 8-4 Pooled shRNA screens: depleted genes in at least one cell line (P<0.05).

Under-Represented TF Genes in Pooled shRNA Screens					Genomics-Inferred TF Drivers for GC-Resistance				
Pathway Name	#Genes in Pathway	#Input Genes in Pathway	corrected p-value	genesInPathway	Pathway Name	#Genes in Pathway	#Input Genes in Pathway	corrected p-value	genesInPathway
Non-small cell lung cancer	54	3	0.00	RB1, TP53, RXRA	Pancreatic cancer	72	3	0.01	STAT3, E2F1, E2F2
Maturity onset diabetes of the young	24	2	0.01	PDX1, HHEX	Chronic myeloid leukemia	75	3	0.01	CTBP1, E2F1, E2F2
Small cell lung cancer	86	3	0.01	RB1, TP53, RXRA	Small cell lung cancer	86	3	0.01	MAX, E2F1, E2F2
Thyroid cancer	29	2	0.01	TP53, RXRA	TGF-beta signaling pathway	87	3	0.01	SP1, RBL1, ZFYVE16
Bladder cancer	42	2	0.02	RB1, TP53	Bladder cancer	42	2	0.02	E2F1, E2F2
Cell cycle	118	3	0.02	RB1, TP53, PTTG1	Cell cycle	118	3	0.02	RBL1, E2F1, E2F2
Basal cell carcinoma	55	2	0.03	GLI1, TP53	Pathways in cancer	330	5	0.03	CTBP1, E2F1, E2F2, MAX, STAT3
Glioma	65	2	0.04	RB1, TP53	Non-small cell lung cancer	54	2	0.03	E2F1, E2F2
p53 signaling pathway	69	2	0.05	TP53, ZMAT3	B cell receptor signaling pathway	65	2	0.05	NFAT5, NFATC2
Melanoma	71	2	0.05	RB1, TP53	Glioma	65	2	0.05	E2F1, E2F2
Pancreatic cancer	72	2	0.05	RB1, TP53	Wnt signaling pathway	152	3	0.05	NFAT5, NFATC2, CTBP1
Chronic myeloid leukemia	75	2	0.05	RB1, TP53	Melanoma	71	2	0.05	E2F1, E2F2

Note:

Highlighted in red are common pathways shared by both groups

Pathway-Express is used for this analysis: <http://vortex.cs.wayne.edu/projects.htm#Pathway-Express>

Table 8-5 Top enriched KEGG pathways by depleted TF genes in RNAi screens and genomics-inferred TF drivers.

Chapter 9 Integrating Functional Genomics with Systems Biology to Discover Therapeutic Targets for ERBB2/HER2+ Breast Cancer³

9.1 Summary

The ERBB2/HER2 amplified breast cancers, accounting for approximately 30% of human breast cancer patients, has the worst survival and prognosis among all subtypes of breast tumors. The development of anti-HER2 therapeutic agents such as Herceptin has significantly altered the treatment of this disease in clinic. However, despite the clinical benefits of these HER2-targeted therapies, all HER2+ patients will eventually develop resistance to this therapy, in which about 50% are initially resistant to Herceptin, whereas the other half that respond to Herceptin will develop resistance within 1 to 2 years of treatment. In this study, to identify alternative therapeutic targets for HER2-amplified breast tumors, we used a genetically-engineered model cultured in 2D, 3D and in vivo xenograft environments and developed a framework to integrate genome-wide RNAi screens with NetBID2, a systems biology algorithm of inferring disease drivers from gene expression data. With the integrative approach, we discovered that STAT3 as a driver-type therapeutic target for HER2+ breast cancers, and we

³ Ruth Rodriguez-Barrueco from Silva lab did all the experiments.

biochemically validated, both in vitro and in vivo, that silencing STAT3 is indeed an effective therapeutic target to stop growth of HER2+ tumor cells. Also from analysis of primary tumor samples, we demonstrated that STAT3 inhibition for killing HER2+ breast cancers has a dependence on ER- subpopulation. We also identified downstream targets of STAT3 that are involved in HER2-triggered tumor transformation and are also potential therapeutic targets.

9.2 Introduction

The ERBB2/HER2 oncogene is overexpressed in approximately 30% of human breast cancer patients due to constitutive amplification [224]. HER2+ subtype has the worst survival and prognosis in breast cancer population [225]. The development of anti-HER2 therapeutic agents, such as trastuzumab (also known as Herceptin) and two other drugs – pertuzumab and lapatinib – have significantly altered the treatment of this disease in clinic. Despite the clinical benefits of these HER2-targeted therapies, however, about 50% of breast cancer patients with HER2-amplification have no response to Herceptin, and almost the other 50% patients that respond to Herceptin eventually develop resistance quickly, usually within 1 to 2 years after treatment [226, 227]. Therapeutics aiming to overcome resistance has been proposed such as targeting AKT pathway based on the mechanism that tumor suppressor PTEN is loss in over 40% of HER2+ breast cancer [228], or blocking Interleukins 6 (IL6) feedback loop [229] based on IL6's regulation of cancer stem cells [230-232]. However, in this

chapter, we are searching for alternative therapeutic targets for HER2-amplified breast tumors instead of reversing resistance of anti-HER2 agents.

Genome-wide RNAi screening has emerged as a powerful tool for systematic loss-of-function studies in mammalian cells [50-53] that may lead to cancer therapeutic target discovery. This technology can be applied to identify genes that form synthetic lethal interactions with ERBB2 in HER2+ breast tumor cells, thus making potential therapeutic targets for HER2+ subpopulation of breast cancers. However, due to a high false positive rate arising from high throughput noise and off target effect, additional knowledge and powerful analysis tools are needed.

We have shown that computationally inferred context-specific maps of transcriptional or post-translational molecular interactions from large-scaled gene expression profiles (GEPs) allow the elucidation of cryptic driver proteins whose gain or loss is necessary and sufficient for tumor initiation or progression [70-73]. Such master regulators are more robust than traditional signatures to distinguish phenotypes [69]. Therefore, we suggest that systematic inference of driver-type regulators from genomic data complementing with RNAi screen technology will give a more comprehensive molecular understanding of mechanisms of HER2+ breast tumors and provide novel targets for therapeutics.

We developed a framework, NetBID2, as detailed in Chapter 4, to infer disease drivers from gene expression data based on computationally-assembled regulatory networks from a cohort of gene expression profiles (GEPs) and

Markov chain Monte Carlo (MCMC) based Bayesian modeling techniques. In parallel, we performed pooled shRNA screens on a genetically-engineered model by introducing ERBB2 overexpression in MCF10A cells, using both microarray and deep sequencing techniques. We also cultured the cells in 2D, 3D and in vivo xenograft environments. Integrating NetBID2 prediction of drivers for HER2+ breast cancers using expression profiles of the isogenic model with a panel of genome-wide shRNA screens identified three candidates – STAT3, AGRN, and GLRX – that are driver-type lethal proteins to HER2-induced tumors. With a focus on STAT3, we confirmed that STAT3 is required for HE2-induced tumorigenesis and we biochemically validated, both in vitro and in vivo, that silencing STAT3 is indeed an effective therapeutic target to stop growth of HER2+ tumor cells. Also from analysis of primary tumor samples, we discovered that STAT3 inhibition for killing HER2+ breast cancers has an addiction to ER-subpopulation. We also identified downstream targets of STAT3 that are involved in HER2-triggered tumor transformation and are also potential therapeutic targets.

9.3 Results

9.3.1 The integrative framework of genome-wide RNAi screening with systems biology of cancer genomics (NetBID2) to identify therapeutic targets of ERBB2+/HER2+ breast cancer

Following the integrative framework in the previous chapter, we developed a similar approach of integrating high-throughput RNAi screens with computational

inference from genomic data to identify novel therapeutic targets for ERBB2+ breast cancer (Figure 9-1).

To identify synthetic lethal partners with ERBB2 in breast cancer, we did pooled shRNA screening on an isogenetic model, by genetically engineering a normal breast cell line, MCF10A. MCF10A is ERBB2 null, so we overexpress ERBB2 in the cells, and use wild type as control. We transduced shRNA-mir library into both mutated and wild type MCF10A cells then grow infected cells for ten doubling times. We extracted genomic DNA from both cell populations and measure hairpin abundance by using both microarray and next-generation sequencing. By comparing shRNA readout in mutated with wild type MCF10A cells, we are interested in finding hairpins or genes depleted in engineered ERBB2+ cells, which are potential therapeutic targets for ERBB2+ breast cancer.

For the isogenetic MCF10A model, we culture the mutated and wild type cells in 2D and 3D culturing system respectively *for in vitro* studies, and also make mouse xenograft models of ERBB2+ MCF10A for *in vivo* experiments.

In parallel, we applied NetBID2, the systems biology framework I developed, to infer drivers of ERBB2+ breast cancer from gene expression data and crossed computationally-predicted ERBB2+ master regulators or signaling modulators with functional shRNA screening identified candidates to produce a more promising short list of therapeutic target candidates.

To define a signature of ERBB2+ breast cancer, a key step in NetBID2, we profiled gene expression of mutated and wild type MCF10A cells using

microarray. We did this for both 2D and 3D cultured systems. The signature was generated by doing differential expression analysis of ERBB2+ vs. wild type in 2D and 3D data respectively. However, these classical signature genes were shown to be unstable to characterize phenotypes [69], we confirmed this point by looking at the overlaps of top signature genes in 2D and 3D systems (Figure 4-3). Therefore, we developed NetBID2 to go beyond expression signatures.

At the same time, we studied gene expression profiles from a cohort of 359 breast cancer patients in TCGA project and constructed both TF-centered regulatory network and signaling molecule-centered cellular networks specific to breast cancer context. The networks were built using a well-developed algorithm, ARACNe [111]. Then we performed enrichment analysis for each driver candidate, TF or signaling molecule, using BSEA algorithm to predict drivers regulating or modulating ERBB2+ breast cancer transformation.

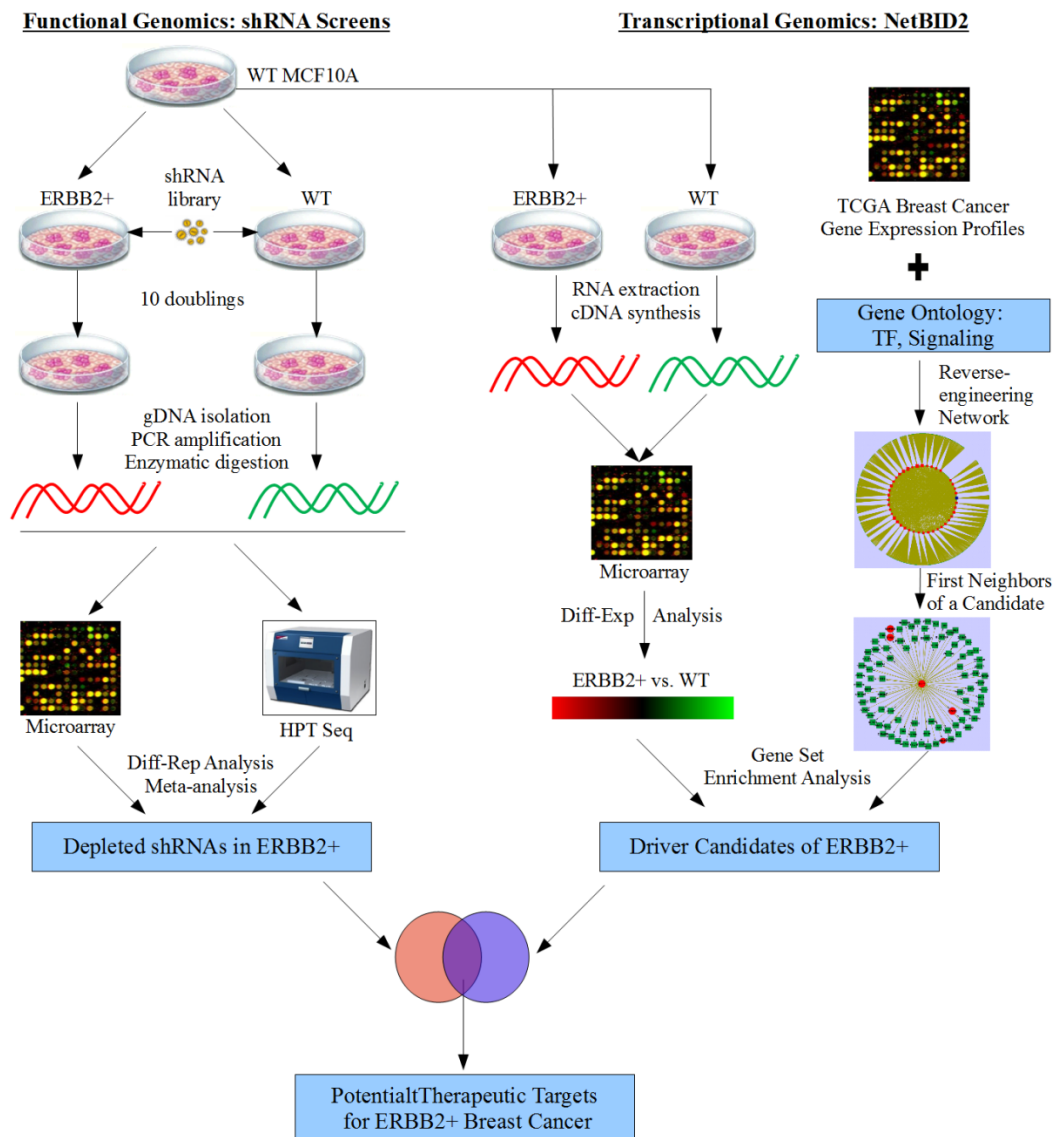


Figure 9-1 The integrative framework of genome-wide RNAi screening with systems biology of cancer genomics (NetBID2) to identify therapeutic targets of ERBB2+/HER2+ breast cancer.

Subsequently, we overlapped depleted genes in shRNA screens with inferred drivers to generate a shorter list of candidates as therapeutic targets for treatment of ERBB2+ breast cancer.

9.3.2 Integrating RNAi screens with NetBID2 identifies STAT3 and two other signaling molecules as driver-type therapeutic targets of ERBB2+ breast cancer

In pooled shRNA screens, we had data from 2D, 3D and in vivo systems using both microarray and NGS technologies, in each of which we generated a list of candidates being depleted in ERBB2+ cell population. To identify potential candidates in each separate data set, we did the differential representation analysis at both individual shRNA level and integrated gene level. The best strategy to get a robust candidate list is to combine all evidences together by doing meta-analysis. So we combined all results from microarray-based RNAi screening data in different analysis levels and identified 134 depleted genes in ERBB2+ population. Similarly by combining evidences in NGS-based data, we obtained 406 candidates. Furthermore, combining both microarray and sequencing data, we generated 355 candidates, and by crossing three separate combined results, finally we got 36 candidates (Figure 9-2) as therapeutic targets from RNAi screening only.

For driver inference using NetBID2 algorithm from gene expression data, we generated driver lists for 2D and 3D signature reference and combine them together, finally got 137 regulatory master regulators or signaling modulators.

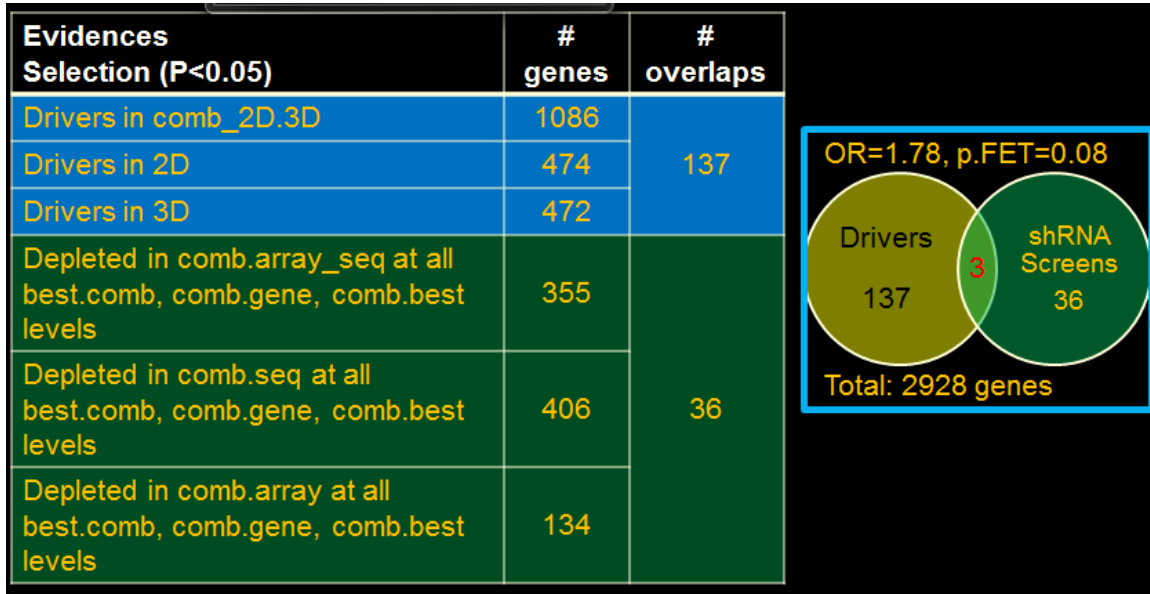


Figure 9-2 Integration of RNAi screening with NetBID2 driver prediction identifies three candidates as driver-type therapeutic targets for ERBB2F+ breast cancer. The left is a summary table of candidate selection by combining evidences from both NetBID2 driver predictions (blue background) and RNAi screening (green background). For driver prediction, we have candidates based on 2D or 3D signature of ERBB2+ cells, and we also combine evidences of 2D drivers and 3D drivers, finally 137 TFs or signaling molecules are overlapped among the three driver lists. For RNAi screening, we have combined microarray or sequencing results alone in best.comb (combine hairpin from different datasets first and then select the best as representative), comb.gene (combined gene level by BHM algorithm), and comb.best (select best hairpin in each data set first and then combine them), and combined both microarray and sequencing results. Only 36 genes came out from RNAi screening. On the first venn diagram, crossing 137 drivers and 36 candidates from RNAi screening, only three genes show up, which is not happening randomly based on Fisher's exact test.

With the integrative systems biology framework of combining functional genomics with cancer genomics, we identified three candidates, STAT3, AGRN and GLRX, that showed significant evidence of depletion in RNAi screens (Table 9-1) and were predicted as strong drivers of ERBB2+ cancer cells (Table 9-2). All three are signaling molecules which are potentially druggable.

Systems Biology of Genomics											
Symbol	Func Type	# pathway	NetBID2			GEP Signature Analysis (ErbB2.Mu vs. WT)					
			Set Size	nES.com b.Drivers	pval.com b.Drivers	FC. 2D.DE	z. 2D.DE	pval. 2D.DE	FC. 3D.DE	z. 3D.DE	pval. 3D.DE
STAT3	TF_Sig	71	396	-2.95	3.16E-03	1.22	1.37	1.70E-01	-1.05	-1.1	2.70E-01
AGRN	Sig	3	72	-3.44	5.75E-04	-1.63	-0.79	4.28E-01	1.36	1.51	1.30E-01
GLRX	Sig	2	93	3.72	1.97E-04						
GLRX	Sig	2	89	3.54	4.02E-04	2.55	3.04	2.37E-03	2.54	4.28	1.85E-05

Table 9-1 NetBID2 inference results of three final candidates from integrative analysis. Duplicate names for GLRX represent two probes for GLRX in the microarray data. In functional type (Func Type), TF is for transcription factor, Sig is for signaling molecule. The column of # of pathway indicates the number of known pathway from multiple databases the candidate gene is involved in. nES.comb.Drivers is the normalized enrichment score of combining 2D and 3D NetBID2 outputted nES. Pval.comb.Drivers is based on nES.comb. The GEP signature analysis columns show fold change (FC), z score and pvalue of differential expression analysis for 2D and 3D data.

Genome-wide shRNA Screens																
Symbol	# shRNAs				Combine Array & Seq						Combine Seq			Combine Array		
	All	SH.array	BC.array	Seq	n.Pval	hairpinId	n.sh.Depleted P<0.05	FC	z	pval	FC	z	pval	FC	z	pval
STAT3	3	3	3	2	10	v2_262105	2	-1.59	-6.65	3.02E-11	-2.75	-4.09	4.33E-05	-1.11	-5.24	1.60E-07
AGRN	1	1	1	1	10	v2_46351	1	-1.5	-4.66	3.20E-06	-2.54	-3	2.67E-03	-1.06	-3.56	3.69E-04
GLRX	1	1		1	7	v2_25191	1	-1.4	-3.3	9.71E-04	-1.64	-2.24	2.53E-02	-1.12	-2.46	1.40E-02

Table 9-2 Genome-wide shRNA screening results of three final candidates from integrative analysis. In # shRNAs column, 'All' means the number of hairpins present in all platforms including hairpin-probed microarray (SH.array), barcode-probed microarray (BC.array) and sequencing. In combine Array & Seq columns, n.Pval is the number of comparisons or evidences, n.sh.Depleted is the number of hairpins showing significant depletion ($P < 0.05$).

Among the three candidates, STAT3 is an interesting one. There are three hairpins targeting STAT3 in the shRNA library and integrated RNAi screening results showed that it's significantly depleted in both microarray and sequencing data. It's connected to 396 genes in the predicted network. STAT3 has dual roles of both transcriptional factor and signaling molecule and participated in 71 known pathways, one of which is a well-studied JAK/STAT3 pathway. STAT3 or JAK/STAT3 pathway has been shown to have oncogenetic effects in several disease contexts [233-242], but it has been associated with ERBB2+ breast cancer. All these make STAT3 as an interesting candidate to follow up.

9.3.3 STAT3 and phosphorylated-STAT3 is confirmed to be active in ERBB2+ MCF10A cells

STAT3 was predicted by NetBID2 from genomic data to be a modulator of ERBB2+ transformation, and its depletion showed lethal effects to ERBB2+ cells from RNAi screening results. So based on that, we hypothesized that STAT3 is abnormally active in ERBB2 engineered MCF10A cells. Literature also suggested that STAT3 needs to be phosphorylated to be active [243]. We checked the protein level of phosphorylated-STAT3 by doing western blot, and indeed, phosphorylated-STAT3 is only active in ERBB2+ MCF10A cells (Figure 9-3), but not in wild type and other genetically-engineered (CYCD1, E1A, PTEN, P53) MCF10A cells. This suggests that ERBB2 over-expression triggers activation of STAT3 thus transforming normal cells into tumor cells and STAT3 seems to be required by ERBB2-induced cancer transformation.

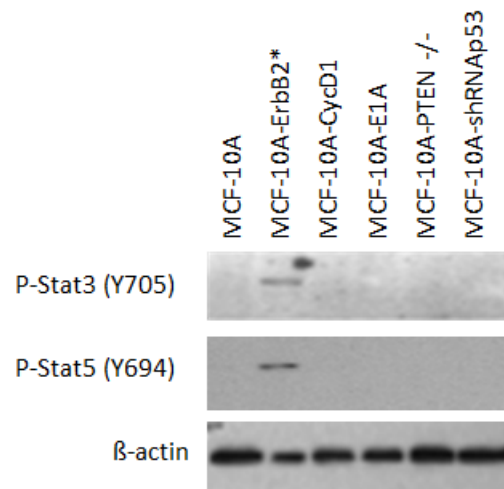


Figure 9-3 Western blots of phosphorylated-STAT3 (P-Stat3) and phosphorylated-STAT5 (P-Sat5) in wild type, and genetically-engineered

(ERBB2+, CYCD1+, E1A+, PTEN-, and p53-) MCF10A cells. pSTAT3 and pSTAT5 are only activate in ERBB2+ MCF10A cells.

9.3.4 STAT3 is validated *in vitro* to be lethal to ERBB2+ MCF10A cells when being silenced

STAT3 was identified as a synthetic lethal partner with ERBB2 in breast cancer. It was predicted as a driver of ERBB2+ MCF10A cells and was confirmed to be activated in ERBB2+ cells. We are interested in testing out whether STAT3 is a good therapeutic target to stop ERBB2+ cancer transformation. First, we validated this using in vitro system. We knocked down STAT3 by siRNA or two shRNAs in ERBB2-muated MCF10A cells. Viability assay results demonstrated that silencing STAT3 by both siRNA and shRNAs reduces growth or viability of ERBB2+ cancer cells significantly (Figure 9-4). Moreover, time-course curve of viability results showed increased reduction of ERBB2+ cell growth over time.

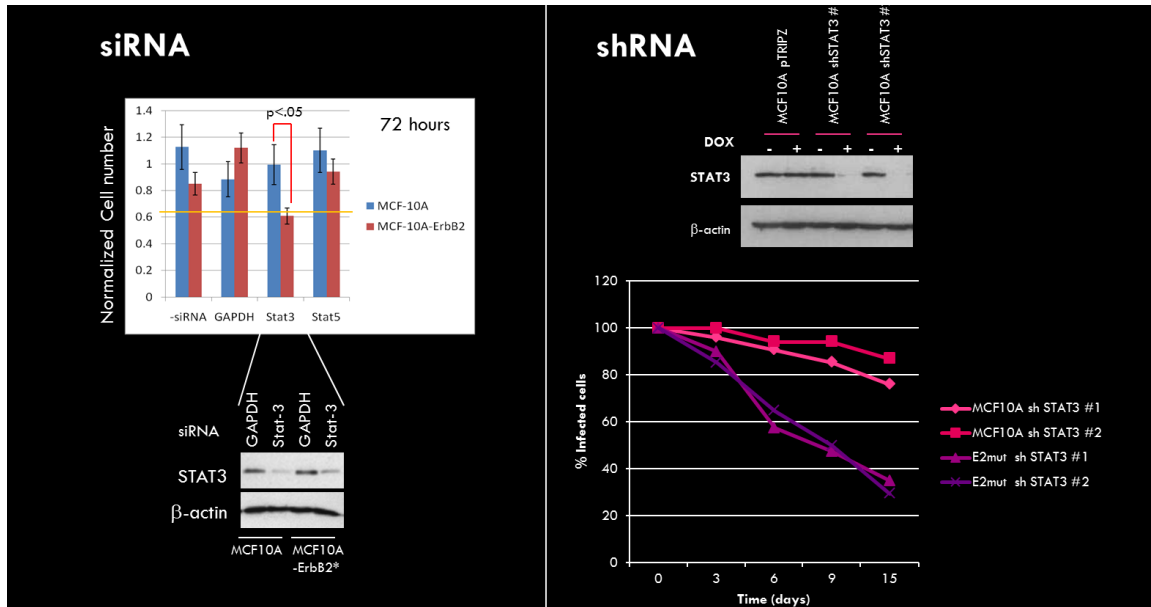


Figure 9-4 Validation of STAT3 *in vitro* by siRNA and shRNA. Viability assays were performed after knock-down of STAT3 by siRNA or two shRNAs.

Second, we performed colony forming cell (CFC) assays to validate STAT3 as a lethal gene to ERBB2+ breast cancer. Qualitative and quantitative measurements of CFC assays showed a significant reduce of cancer cell growth in ERBB2+ cells when STAT3 is inhibited by shRNA (Figure 9-5). In the control population without ERBB2 overexpression, silencing STAT3 showed no effects on tumor cell transformation and growth. If we induced ERBB2 in wild type MCF10A cells, tumor colonies were formed quickly and expanded dramatically, but however, if we silenced STAT3 in ERBB2-overexpressed MCF10 cells, there is almost zero tumor transformation or even decreased tumor formation comparing with wild type. This again confirmed that STAT3 is indeed an effective target to stop transformation and growth of ERBB2+ tumors.

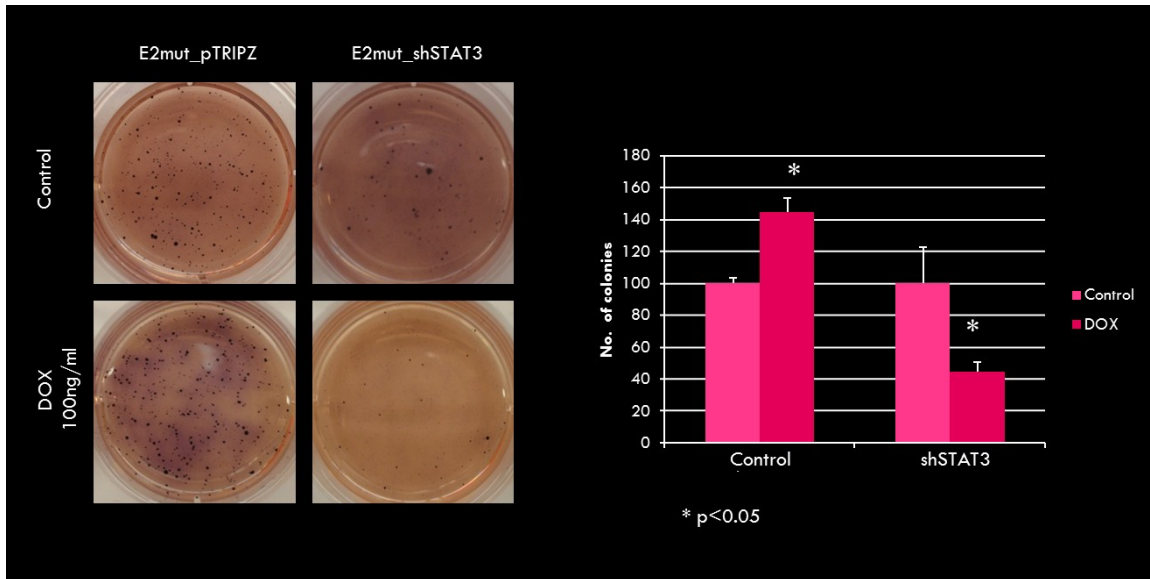


Figure 9-5 Validation of STAT3 in vitro by colony forming cell (CFC) assays with shRNA silencing. Colony assays were performed in wild type MCF10A cells with and without STAT3 silencing by shRNA, ERBB2-muanted MCF1-A with and without shSTAT3. Quantitate cell counts were measured for each colony assay.

9.3.5 STAT3 is validated *in vivo* to be lethal to ERBB2+ xenograft mouse models

Besides in vitro system, we also tested STAT3 in vivo, by making xenograft mouse models of ERBB2+ MCF10A cells with and without STAT3 silencing, More strikingly, in mouse models of ERBB2+ MCF10A cells with STAT3 silencing by shRNA, there was almost no sign of tumor formation and no growth of tumor population, comparing with the exponential growth of tumor cells in the mice without STAT3 inhibition. Again, STAT3 is indeed a valid and effect target to control tumor transformation by ERBB2 inducement upon inhibition and may

suggest potential clinical applications to treatment of HER2+ breast cancer, one of the most aggressive form in breast cancer.

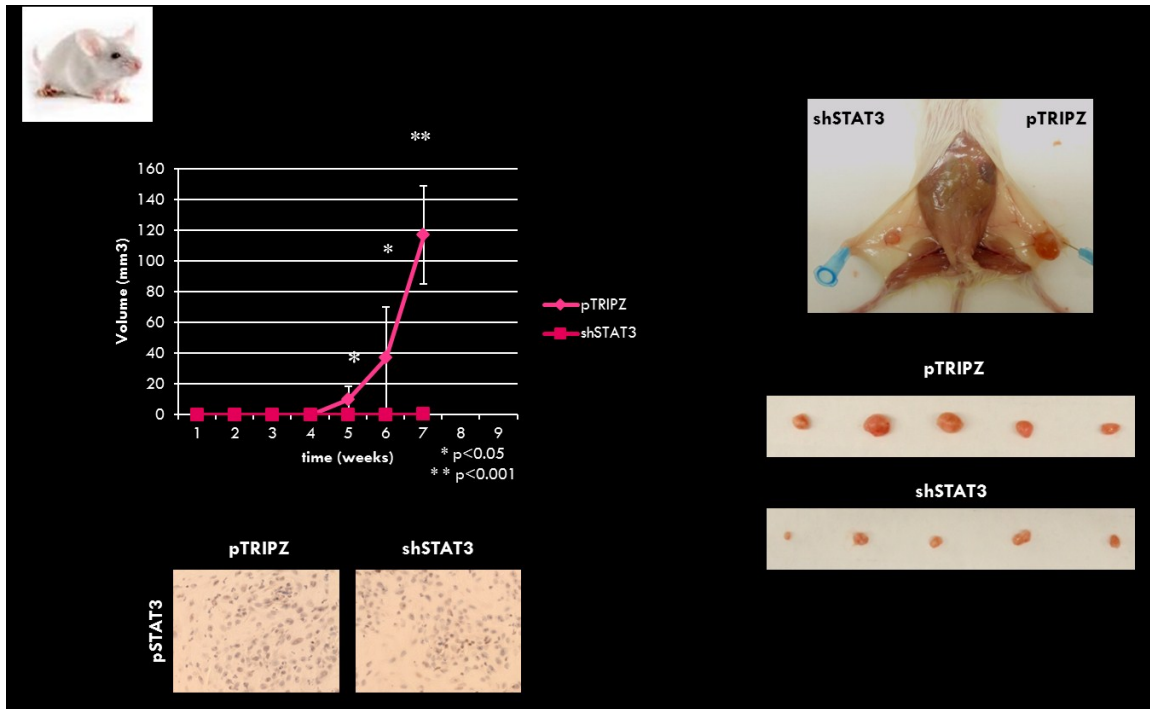


Figure 9-6 Validation of STAT3 in vivo. Xenograft mouse models were made for ERBB2+ MCF10A cells with and without STAT3 silencing by shRNA. Cell population with tumor marker were imaged and measured weekly, up to 7 weeks, when tumor size was photographed.

9.3.6 STAT3 inhibition is specific to ERBB2+ breast cancer

We have shown that silencing STAT3 dramatically stops growth of ERBB2+ cancer cells, but we have to test the specificity of STAT3 inhibition to ERBB2 overexpression in breast cancer. We selected a cell line, MDAMB231, a basal type breast cancer cell line which is ERBB2- but STAT3+ (Figure 9-7) to validate the specificity of STAT3 silencing to ERBB2 amplification. In vitro completion

assays demonstrated that there is no significant viability change of MDAMB231 cells with and without STAT3 silencing (Figure 9-8). Furthermore, in vivo mouse models of MDAMB231 cells showed almost identical tumor growth curves with and without STAT3 inhibition (Figure 9-8). Therefore, this suggested that silencing STAT3 has no effects on growth or viability of ERBB2- cancer cells though STAT3 is expressed in those cells, confirming the specificity of STAT3 knock-down to ERBB2 overexpressed breast cancer cells.

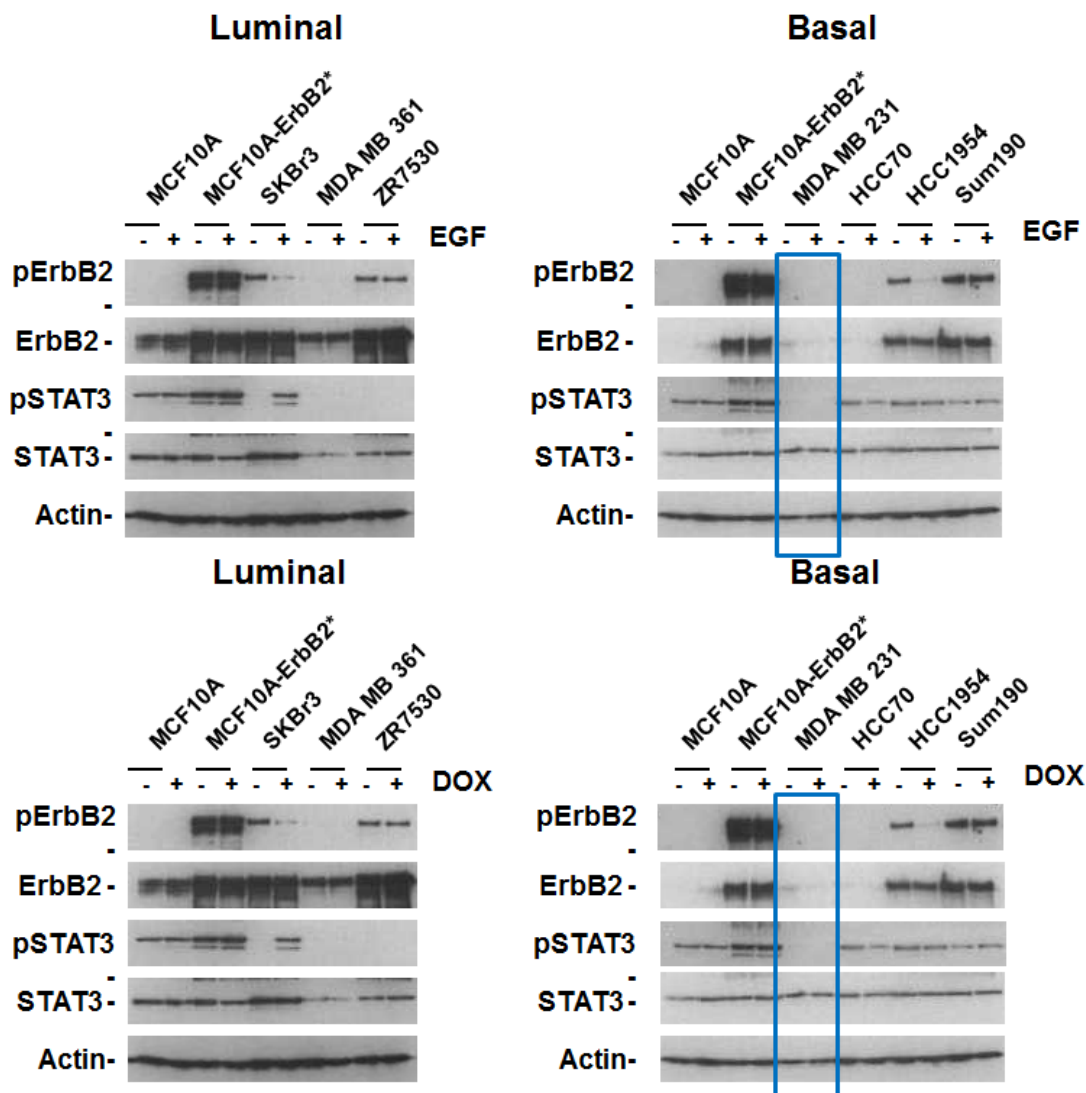


Figure 9-7 Western blots of STAT3, phosphorylated STAT3, ERBB2, phosphorylated ERBB2 in different breast cancer cell lines: wild type MF10A, ERBB2+ MCF10A, three Luminal lines (SKBR3, MDAMB361, ZR7530), four Basal lines (MDAMB231, HC70, HCC1954, SUM190PT).

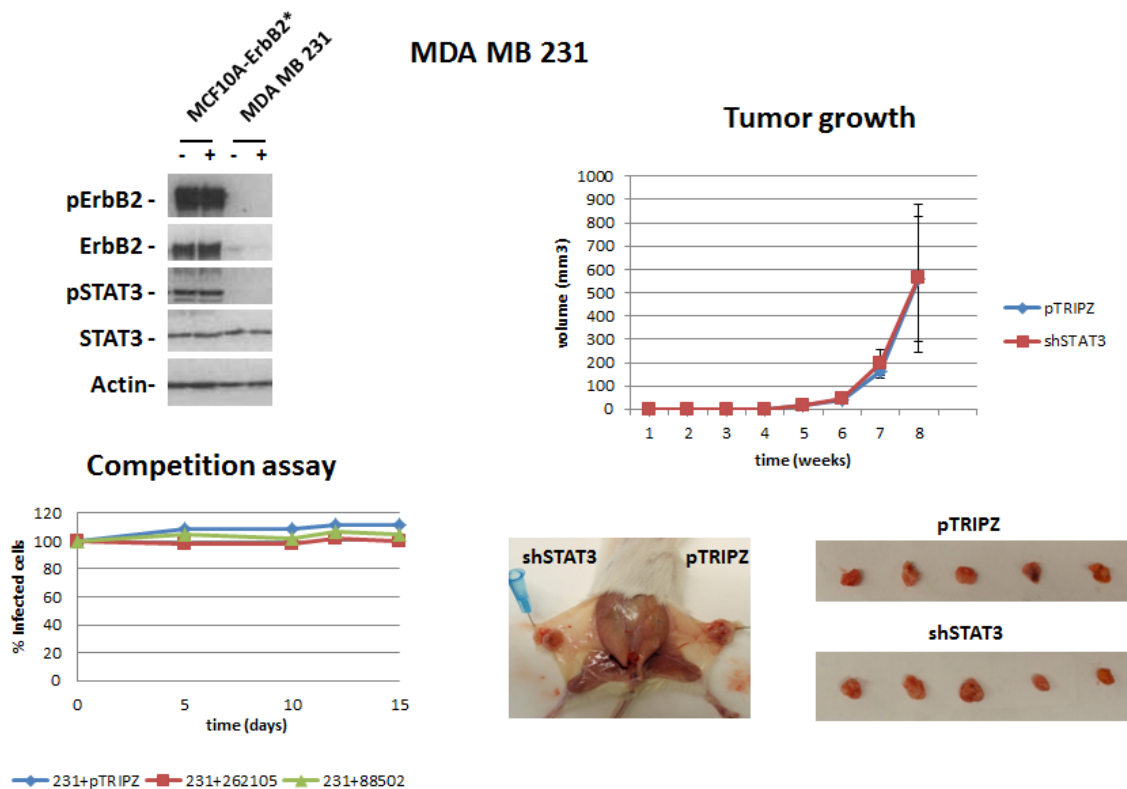


Figure 9-8 In vitro and in vivo validation of STAT3 specificity to ERBB2 using MDAMB231 cell line. MDAMB231 is a ERBB2- but STAT3+ line. In vitro competition assays and in vivo xenograft mouse models with and without STAT3 silencing by shRNA were performed to measure viability of tumor cells or tumor growth.

9.3.7 STAT3 doesn't show up as a driver for HER2+ from NetBID2 analysis on expression profiles of primary breast cancer patients

We identified STAT3 as a signaling modulator of ERBB2 induced tumorigenesis by applying NetBID2 algorithm to expression profile generated from a genetically-engineered isogenetic model, MCF10A with ERBB2 overexpression. However, we have to be careful that this type of isogenetic model is homogeneous, but

doesn't capture the heterogeneity that is commonly present in primary patients (Figure 9-9). Therefore, we asked whether we can still identify STAT3 if we applied the same computational framework (NetBID2) to expression data from primary patients.

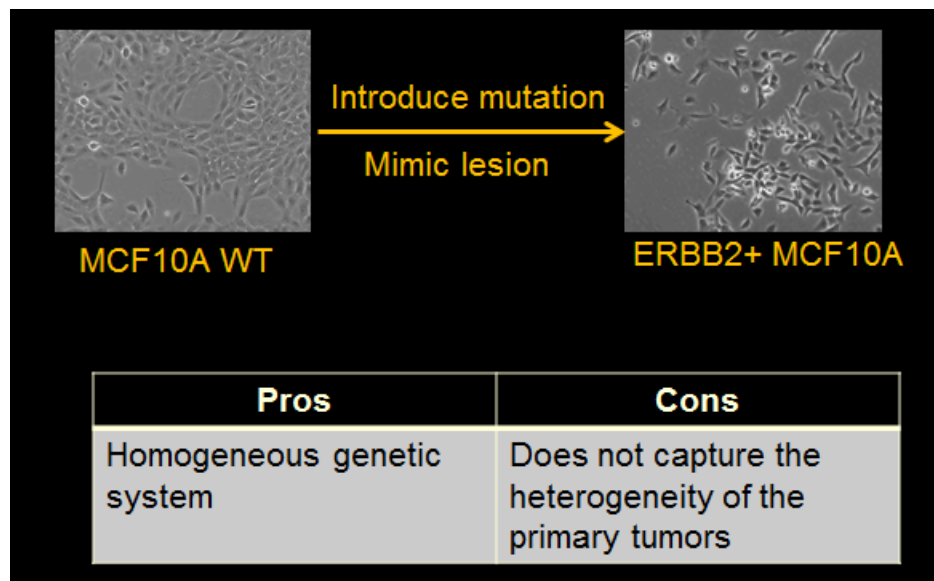


Figure 9-9 Illustration of genetically-engineered isogenetic model. ERBB2 is overexpressed genetically in MCF10A cells to mimic ERBB2+ breast tumors.

In application of NetBID2 to this project, we constructed breast cancer interactomes from gene expression profiles of a cohort of 359 primary breast cancer patients, out of which, 58 patients were clearly classified as HER2+ based on clinical lab IHC assays and 201 one were defined as HER2-. So we performed the signature analysis using those clearly defined HER2+ and HER2- patients' profiles instead of homogenous MCF10A profiles with and without ERBB2 overexpression. And then we applied NetBID2 algorithm to identify master regulators or signaling drivers of this primary tumors' signature, however,

in both transcription factor-centered network analysis and signaling protein-focused network analysis, STAT3 didn't show up as a driver candidate for HER2+ breast cancer (Figure 9-10).

This reflected the difference of isogenetic models and primary patient samples. Particularly, this inferred the heterogeneity of HER2+ breast cancer patients. It seems that STAT3 is only modulating a subset of HER2+ breast cancer population and STAT3 inhibition seems only work on the patients that are similar to the model we used, MCF10A.

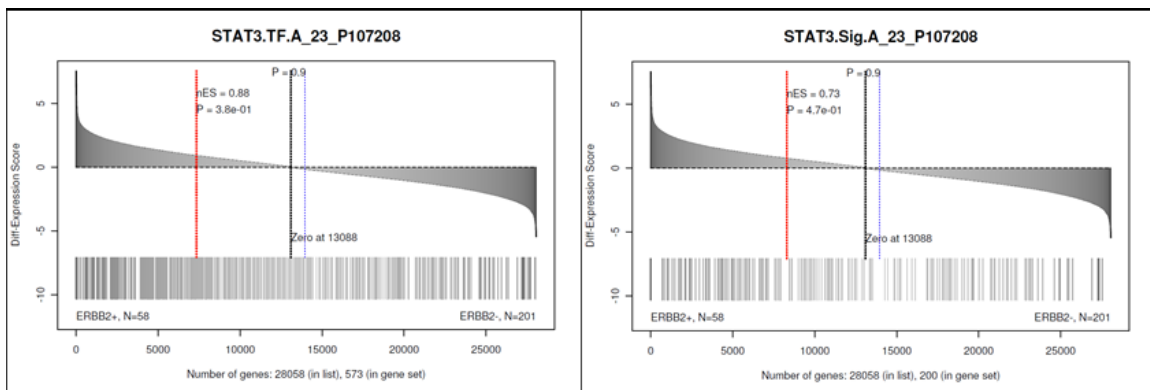


Figure 9-10 STAT3 doesn't show up as a driver of HER2+ primary samples in both transcription factor (TF)-centered network analysis and signaling protein (Sig)-focused network analysis.

9.3.8 STAT3 is addicted to ER status being a driver for ER- and HER2+ breast cancer, but not for ER+ ones

As discussed in previous section, STAT3 inhibition is not a uniform cure for all HER2+ breast cancer patients, and we need to identify the co-founding factors that determine which sub-population it works on.

Patricia Villagrasa Gonzalez, a postdoc from Silva lab reminded me of that MCF10A, the model we used to identify STAT3 as a driver of HER2+ subtype, is ER-. However, the majority of HER2+ primary patients (83%) from which we failed to find STAT3, are ER+. And we know ER is one of the most important biomarkers to classify breast cancer patients. Therefore, this motivated me to separate HER2+ patients into ER+ and ER- groups and then applied NetBID2 to each of them.

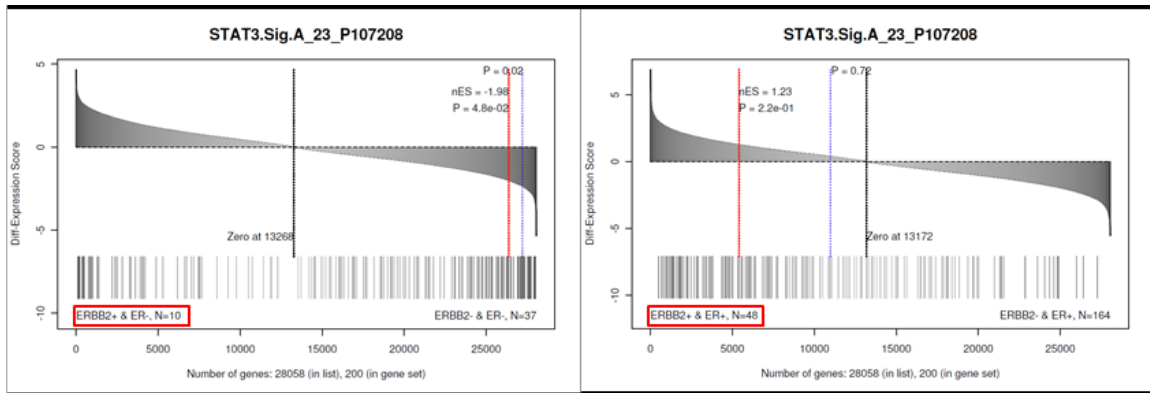


Figure 9-11 NetBID2 results of STAT3 on ER- and ER+ groups of HER2+ population. STAT3 only shows up in ER- group as a driver.

Surprisingly, after we differentiate ER+ and ER- in HER2+ population, STAT3 showed up only in ER- group with exactly the same pattern as we saw in MCF10A model (Figure 9-11). This is consistent with what we hypothesized based on the fact that MCF10A is ER-. There was no pattern for STAT3 network in ER+ and HER2+ population, which is the majority of HER2+ patients, thus diluting the signal of STAT3 as a driver if we mixed ER+ and ER- together.

We have shown that STAT3 is only modulating ER- sub-population of HER2+ patients, which is one of the most advanced forms of breast tumors. We also tried other characteristic factors to check whether we can find STAT3 in one of them. We separated HER2+ patients into Basal or Luminal type according to their gene expression profiles and applied NetBID2. Since MCF10A is classified into Basal group, we expected to STAT3 coming out from analysis on Basal-type HER2+ group, however, STAT3 didn't show up as a significant driver candidate though the direction is consistent with the one in MCF10A analysis. Interesting thing is that STAT3 showed up as a driver for Luminal type and HER2+ group, however, the pattern of direction is opposite to the pattern we observed in MCF10A model. This suggested that STAT3 is modulating Luminal type of HER2+ breast cancer patients but in a different mechanism from ER- and HER2+ group.

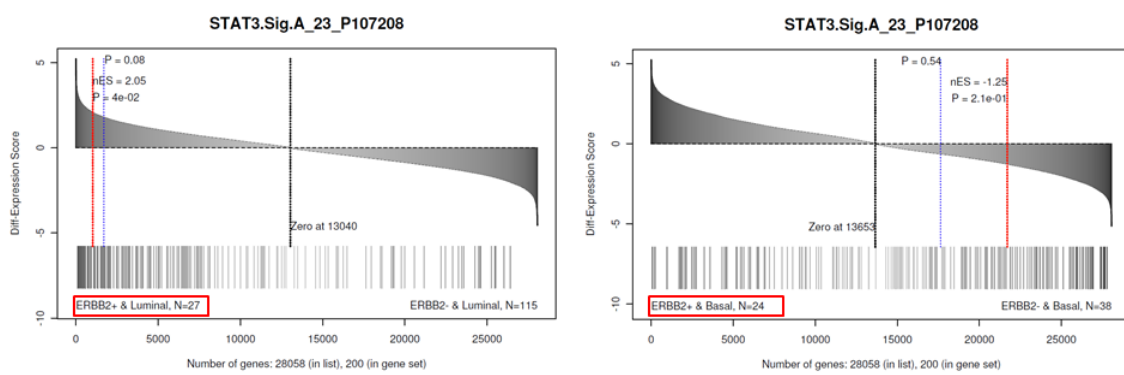


Figure 9-12 NetBID results of STAT3 in Luminal and Basal subtype of HER2+ patients. STAT3 shows up in Luminal group but shows opposite direction pattern to the one in MCF10A model and ER- subgroup.

9.3.9 Searching for downstream targets of STAT3 being involved in modulation of ER- and HER2+ breast cancer

We identified STAT3 as a modulator of ER- and HER2+ breast cancer, but we are more interested in identifying the entire pathway specifically responsible for tumorigenesis of ER- and HER2+ breast cancer. The reason is because STAT3 is a well-known upstream signaling transduction modulator and regulator of a spectrum of downstream biological processes. We asked what specific downstream players or targets of STAT3 are involved in the pathway of tumor initialization and growth induced by ERBB2, or in the pathway of inducing apoptosis in this type of breast cancer cells when we knock-down STAT3.

To identify potential targets of STAT3 experimentally in this context, we did microarray profiling of whole genome by perturbing STAT3. We did knock-down of STAT3 by two shRNAs in ERBB2+ cells and did inducement of STAT3 by over-expressing IL6, an upstream activator of STAT3, in ERBB2- cells. Positive targets of STAT3 will be down-regulated in STAT3- samples but up-regulated in STAT3+ samples, whereas negative targets will show the opposite pattern. We identified 66 targets of STAT3 (Figure 9-13) using a stringent threshold ($P < 0.05$, fold change > 4). SOCS3 is a well-known activated target of STAT3 showing up as one of the top positive targets. STAT3 itself can also be used a positive control showing down-regulation if you silence STAT3 and up-regulation if you over-express IL6.

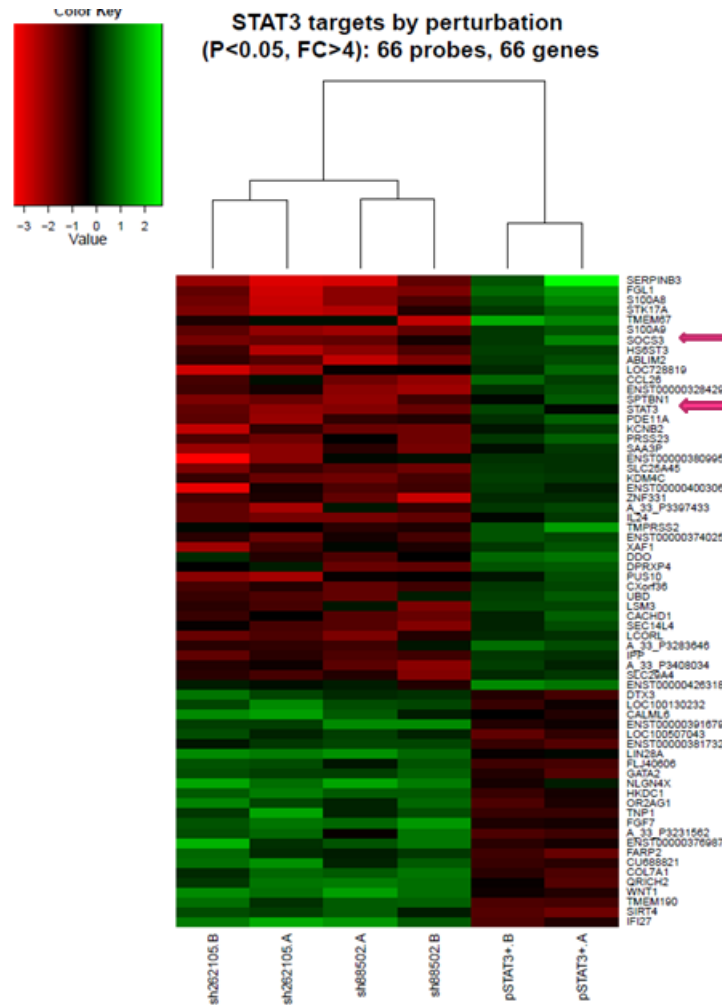


Figure 9-13 Heatmap of top STAT3 targets from perturbation experiments by knocking-down STAT3 by two shRNAs or over-expressing IL6, an upstream activator of STAT3. Red stands for down-regulation while green means up-regulation, for example, in sh-STAT3 experiments (the left four samples), genes in red are under-expressed when silencing STAT3 or are potential positive targets of STAT3, while green ones are potential negative targets of STAT3. SOCS3 is a known target activated by STAT3 and STAT3 itself is another positive control.

In addition to experimental searching for STAT3 targets, we also looked at the targets (Figure 9-14) computationally predicted by ARACNe from a large cohort

of breast cancer expression profiles. The analysis of ARACNe is on probe level and is separated for transcription factor-centered and signaling molecule-centered networks. There are three probes for STAT3 representing three different transcripts of STAT3 and STAT3 has dual role of being transcription factor and signaling protein, therefore there are 6 lists of targets predicted by ARACNe, and the number ranges from 100 to 250. And all of the six lists of predicted targets are significantly enriched in experimentally-identified targets of STAT3 (Figure 9-15).

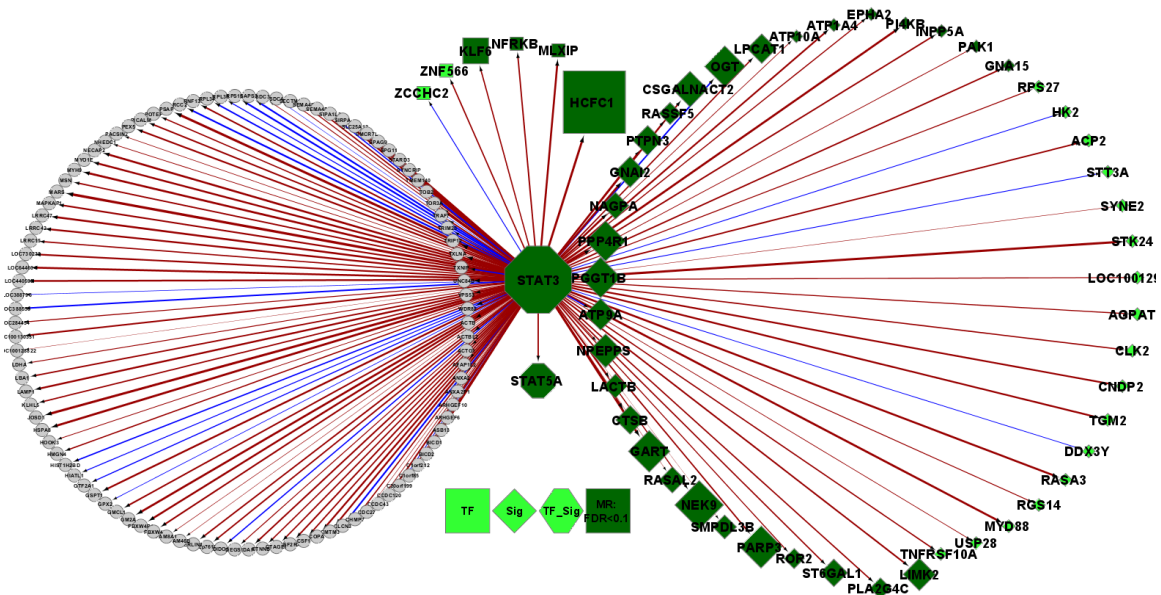


Figure 9-14 Subnet of STAT3 predicted by ARACNe in the signaling-centered network. Genes on the right circle are transcription factors (TF in square shape), signaling molecules (Sig in diamond shape) or both (TF_Sig in hexagon shape). Genes in dark green are also predicted as master regulators or drivers (MR) of HER2+ breast cancer. Genes on the left circle are the ones in general. Red edge is for positive correlation while blue is for negative correlation.

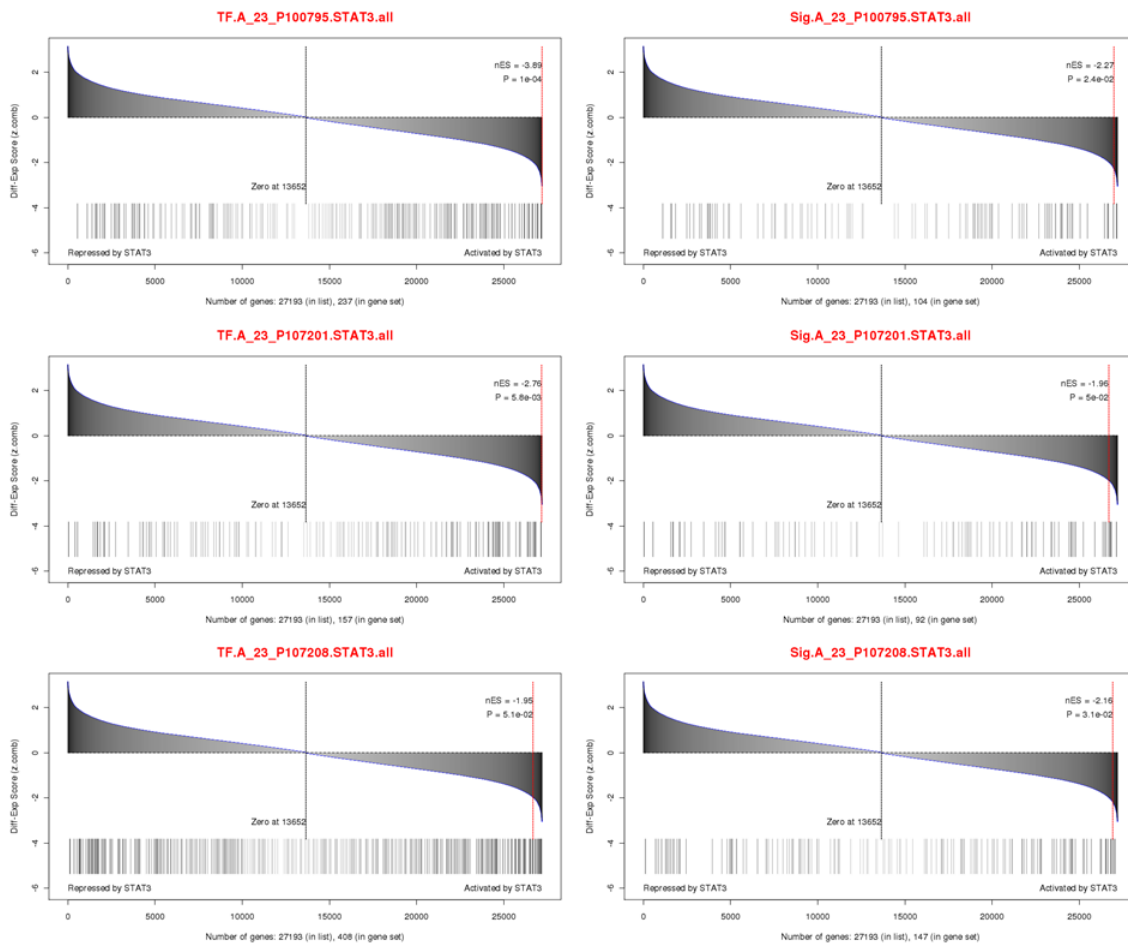


Figure 9-15 Enrichment of ARACNE-predicted targets of STAT3 from transcription factor (TF)-centered or Signaling molecule (Sig)-centered network in experimentally identified targets of STAT3 by microarray profiling after knocking-down STAT3 or overexpressing IL6 (activator of STAT3).

9.3.10 STAT3 targets that are lethal to ER- and HER2+ breast cancer

With a handful of STAT3 targets from perturbation experiments, we asked how many of them are also lethal to HER2+ cancer cells. So we did enrichment

analysis of 111 selected STAT3 targets ($P < 0.05$, $FC > 3$) in RNAi screening results of ERBB2+ MCF10A cells and there was an enrichment pattern on the depleted side, meaning that majority of STAT3 targets are also lethal to ERBB2+ MCF10A cells. If we distinguished positive and negative targets of STAT3, there was no significant enrichment for 70 positive targets, but there was a significant enrichment for 43 selected negative targets.

Besides MCF10A isogenetic model, we also did shRNA screens for three HER2+ breast cancer lines (SKBR3, SUM190PT and MDAMB361). Among these three HER2+ cell lines, SKBR3 is ER+ while the other two are ER-. Unfortunately, STAT3 didn't show up being lethal to all the three lines. MDAMB361 was expected because it is an ER+ line, but SUM190PT and SKBR3 were not expected. The reason could be either hairpins targeting STAT3 were not working well, or the data was noisy or there could be other factors making these two lines resistant to STAT3 inhibition.

However, STAT3 pathway seems to be activated in ER- and HER2+ breast tumors. So we went downstream and checked STAT3 targets that are lethal to HER2+ population. Fortunately, there are a few STAT3 targets such as S100A9, TRDMT1 (activated), FLRT1, NPAS1 (repressed) that are lethal to both HER2+ breast cell lines and genetically-engineered MCF10A cells. Those candidates might be alternative therapeutic targets for HER2+ breast cancer because of their lethal effects to both models and might be more specific than STAT3 because STAT3 is modulating or regulating a number of sub-programs and those lethal

targets might specifically involve in tumorigenesis of HER2+ breast cancer, making them more promising as novel therapeutics for HER2+ patients.

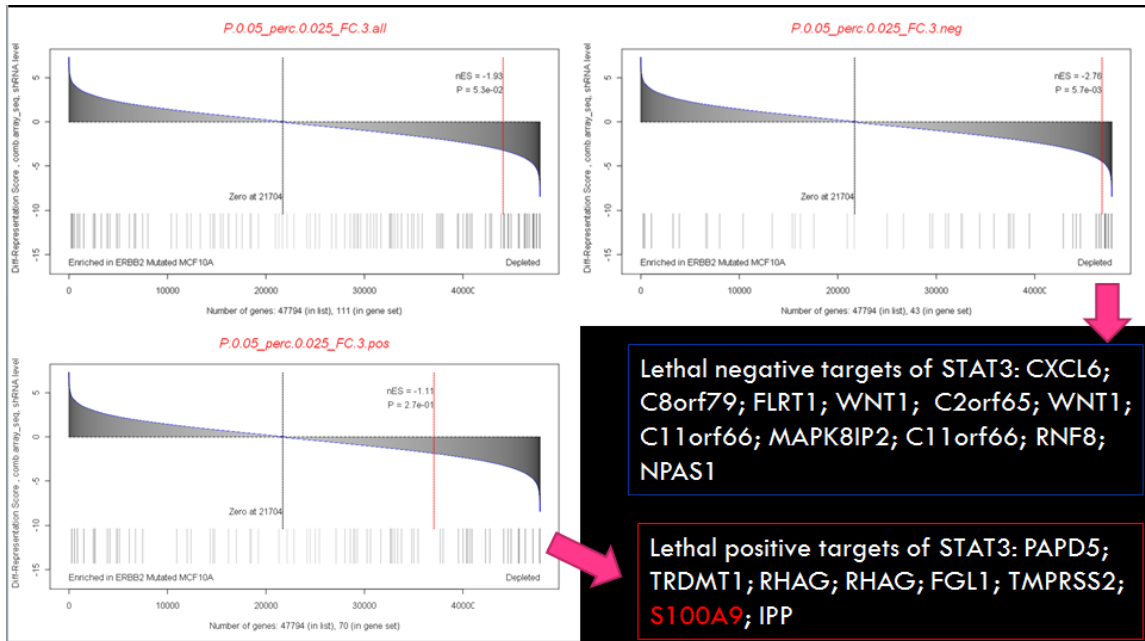


Figure 9-16 Enrichment of STAT3 targets from perturbation experiments in shRNA screening results of ERBB2+ MCF10A cells. Reference genes are ranked from the most enriched to most depleted in ERBB2+ vs. wild type MCF10A cells. STAT3 targets are selected by $P < 0.05$ and $FC > 3$. Positive targets are defined as positive expression in STAT3-induced samples comparing with expression in STAT3-silenced samples. Negative targets are defined in the opposite way. Top lethal positive or negative targets are listed in the boxes on the bottom-right.

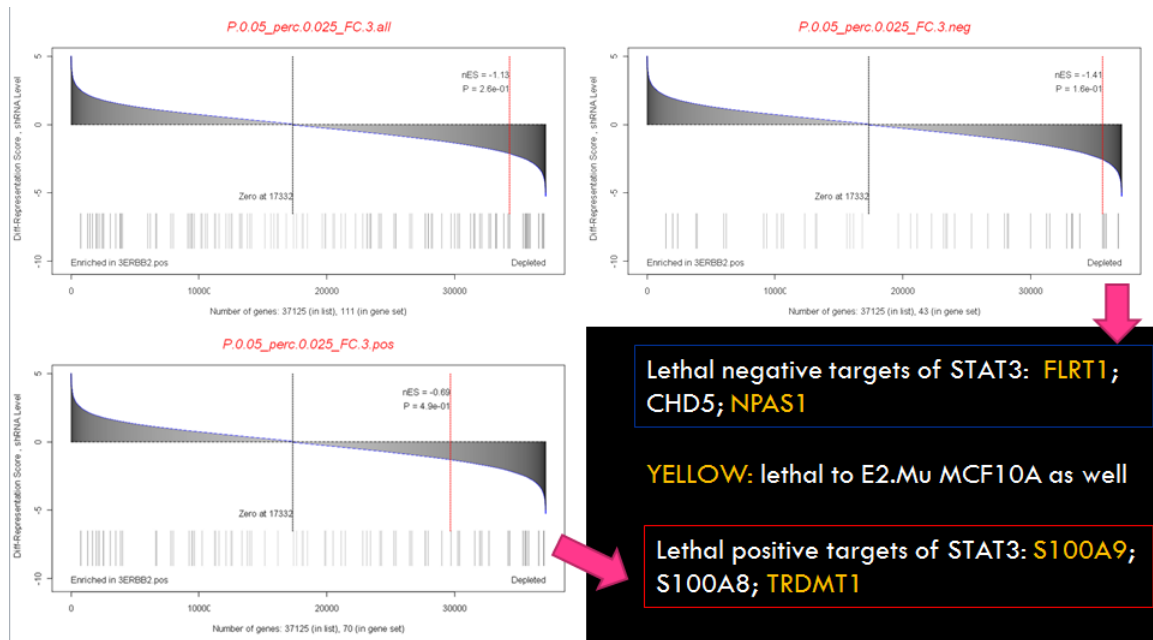


Figure 9-17 Enrichment of STAT3 targets from perturbation experiments in shRNA screening results of ERBB2+ breast cancer cell lines (SKBR3, SUM190PT, MDAMB361). Reference genes are ranked from the most enriched to most depleted in three HER2+ cell lines (using a combined score). STAT3 targets are selected by $P < 0.05$ and $FC > 3$. Positive targets are defined as positive expression in STAT3-induced samples comparing with expression in STAT3-silenced samples. Negative targets are defined in the opposite way. Top lethal positive or negative targets are listed in the boxes on the bottom-right. Yellow ones are the common targets lethal to ERBB2+ MCF10A cells (Figure 9-16).

9.3.11 Other STAT family members (STAT5A and STAT1) show up as drivers of ERBB2+ breast cancer in analysis of data from both isogenic models and primary patients

Besides STAT3, we also identified other STAT family members including STAT5A and STAT1 as drivers of HER2+ breast cancer from MCF10A isogenic model. In particular, STAT5A has similar patterns with STAT3. For example,

STAT5A was inferred as a driver from MCF10A isogenetic model (Figure 9-18), but didn't show up from all HER2+ primary tumors (Figure 9-19). However, STAT5A showed the addiction to ER status by being predicted as a driver for only ER- and HER+ primary samples, not from ER+ group (Figure 9-20). It turned out that in the predicted network, STAT5A is interacting with STAT3 in a positive manner (Figure 9-21, Figure 9-22), which might explain the above patterns. Unfortunately, STAT5A didn't show up from RNAi screening data and was confirmed by individual knock-down experiments.

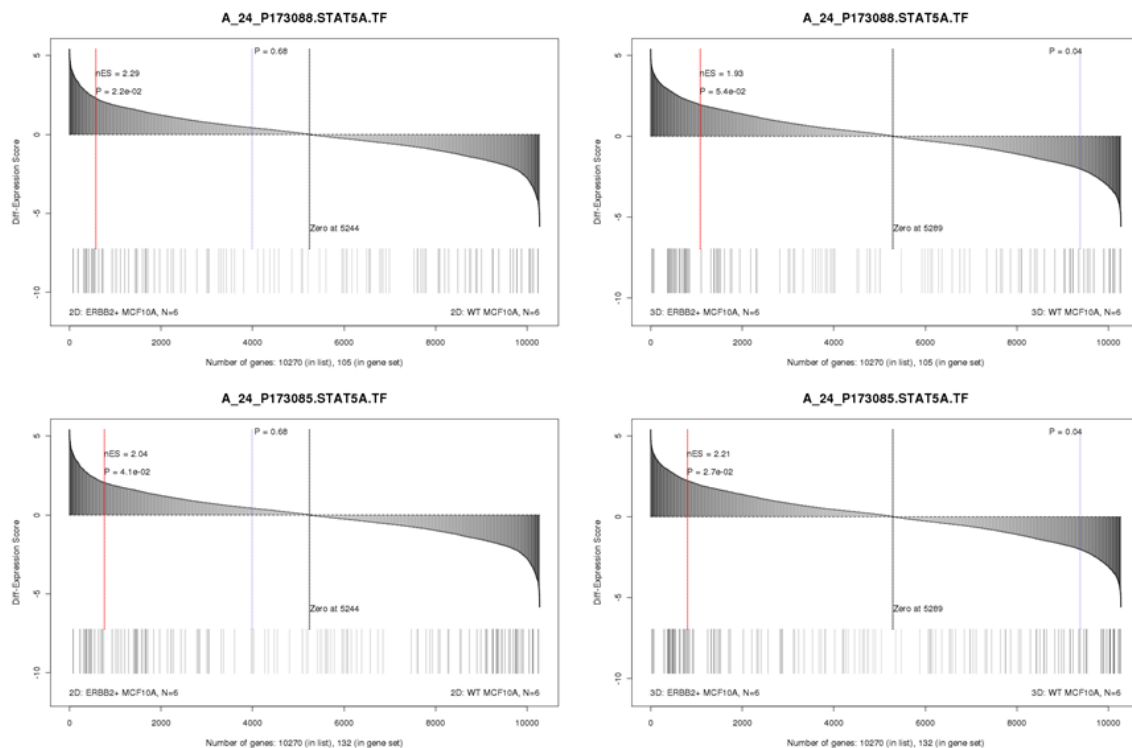


Figure 9-18 STAT5A is predicted as a driver of ERBB2+ MCF10A cells.

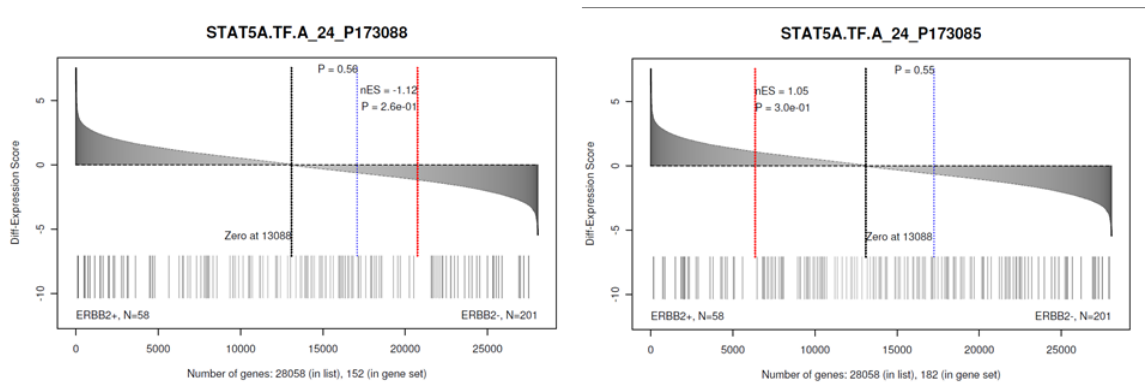


Figure 9-19 STAT5A like STAT3 is not a driver of HER2+ primary tumors.

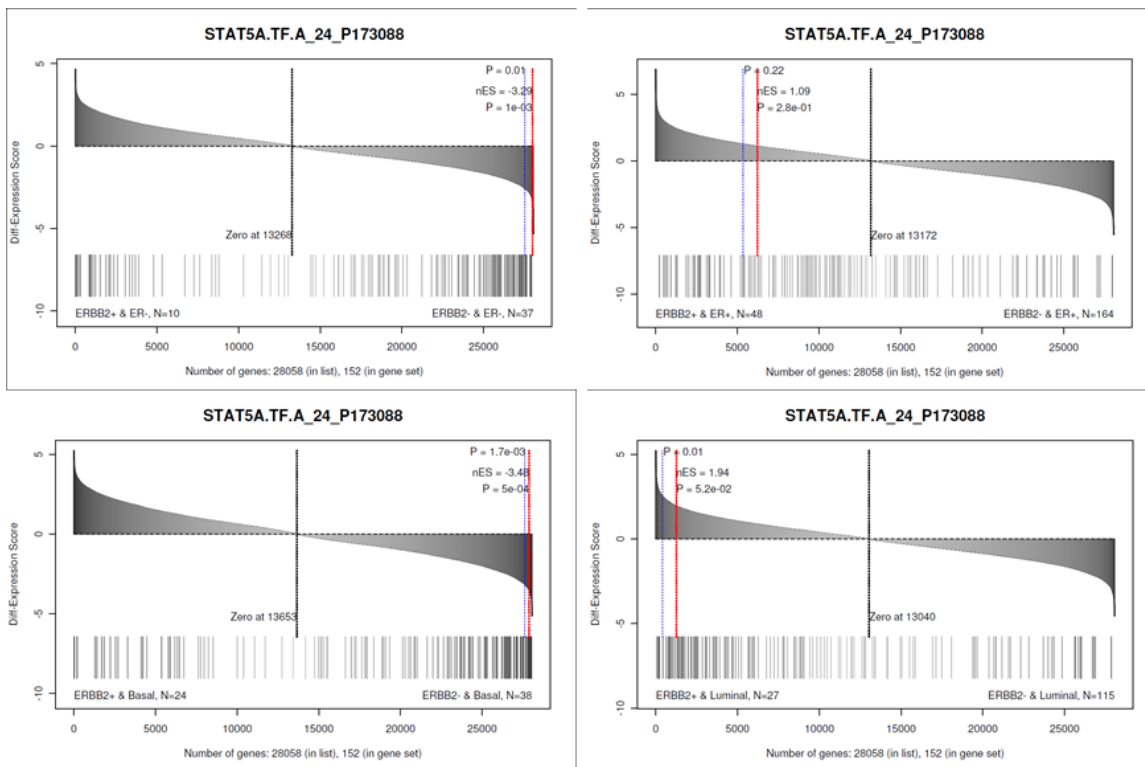


Figure 9-20 STAT5A is a driver of ER- but not ER+ group of HER2+ primary patients.

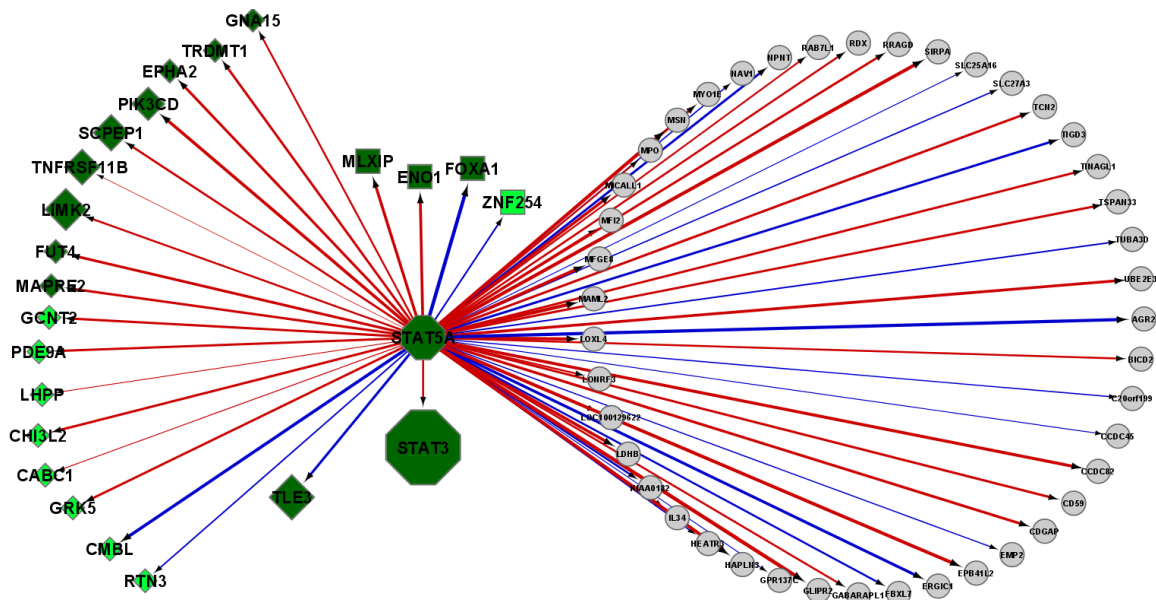


Figure 9-21 Subnet of STAT5A predicted by ARACNe in the signaling-centered network. Genes on the left circle are transcription factors (TF in square shape), signaling molecules (Sig in diamond shape) or both (TF_Sig in hexagon shape). Genes in dark green are also predicted as master regulators or drivers (MR) of HER2+ breast cancer. Genes on the right circle are the ones in general. Red edge is for positive correlation while blue is for negative correlation.

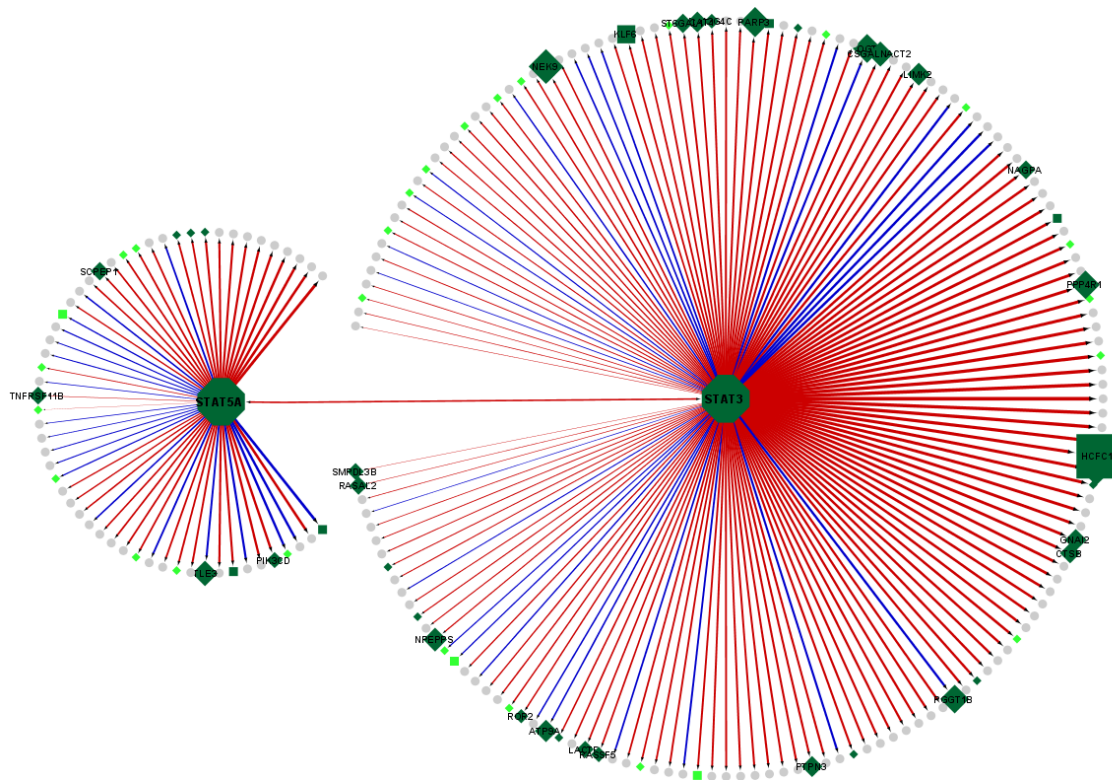


Figure 9-22 Subnetwork of STAT3 and STAT5A predicted by ARACNe in the signaling-centered network. Annotation is the same as in Figure 9-20 and Figure 9-21.

STAT1 also was predicted as a driver of ERBB2+ MCF10A cells (Figure 9-23), however it showed different enrichment pattern with STAT3 and STAT5A. Moreover, it didn't display the addiction to ER- subgroup of HER2+ primary population (Figure 9-25) though it showed no enrichment pattern in all HER2+ primary samples similar to STAT3 and STAT5A (Figure 9-24).

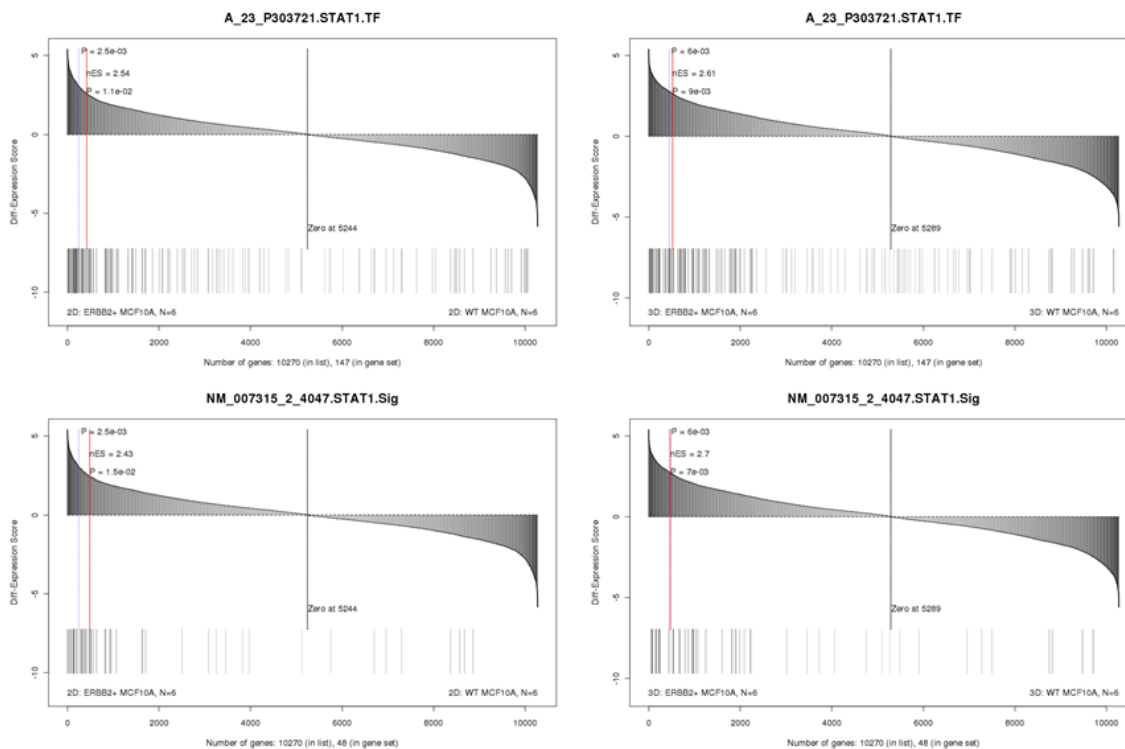


Figure 9-23 STAT1 is a driver of ERBB2+ MCF10A cells, but shows different enrichment pattern with STAT3 and STAT5A.

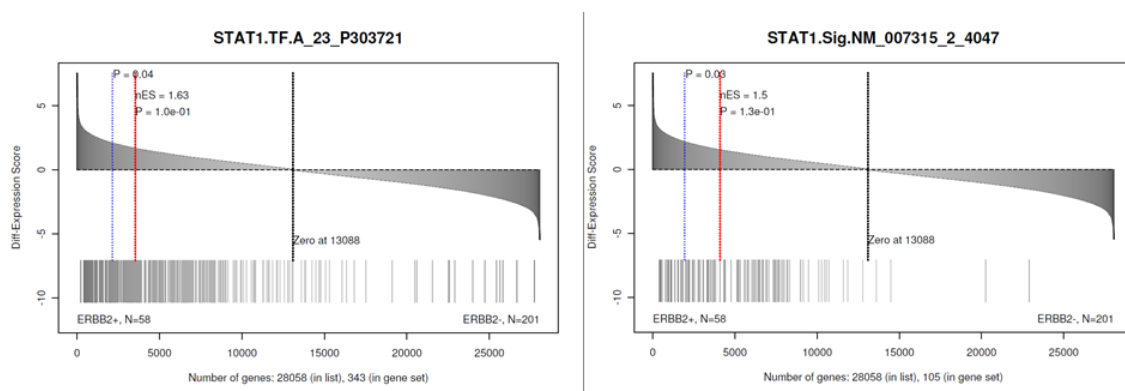


Figure 9-24 STAT1, similar to STAT3 and STAT5A, is not a driver of all HER2+ primary tumors.

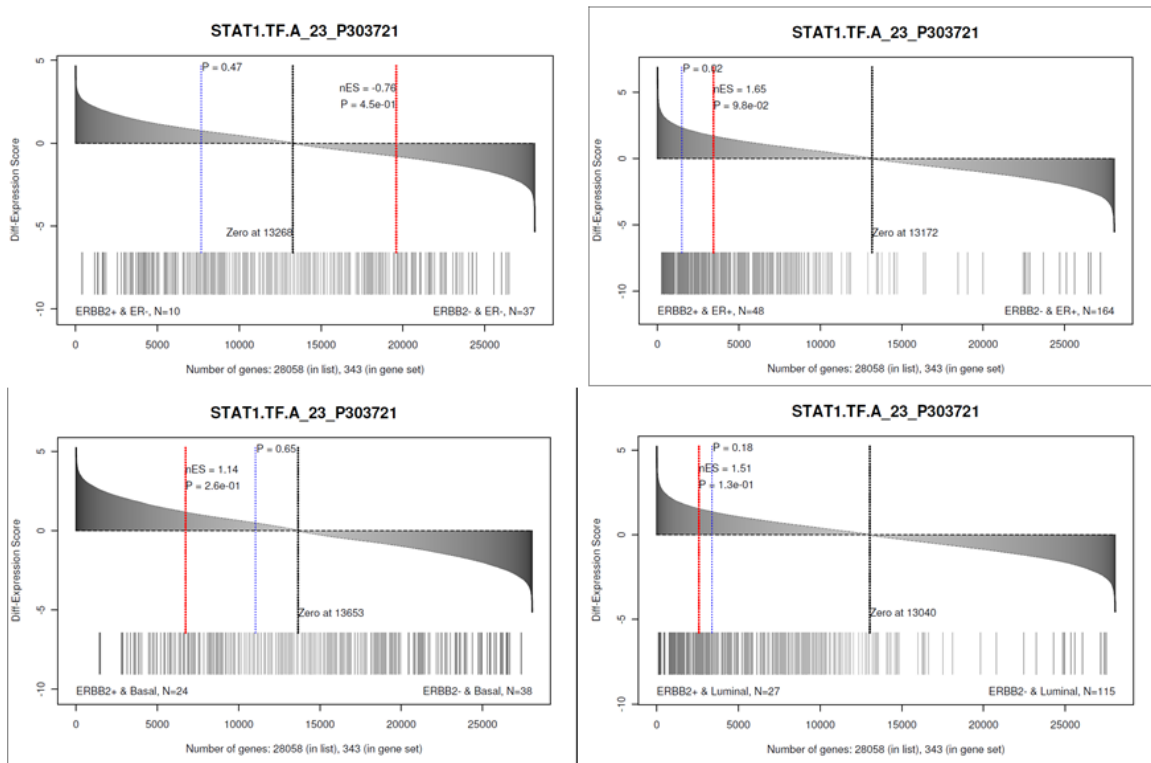


Figure 9-25 STAT1, different from STAT3 and STAT5A, doesn't show any addition to ER status being a driver of HER2+ primary tumors.

9.3.12 RNAi screening from 2D vs. 3D vs. In Vivo environment

Nowadays, there is increasing awareness of the drawbacks of 2D cell culture and the related effect on the value of the research being performed. Not surprisingly, scientists are shifting their focus to cells cultured in 3D or in vivo mouse models because cells in 3D or mouse model environments are much more similar to cells in the real environment, a living organism (in vivo) than flat, unnaturally thin, single layer cells grown on 2D plastic. So in this project to identify therapeutic targets of ERBB2+ breast cancer, we cultured our cells in 2D, 3D and in vivo environments.

From literature, we learned that there are differences between 2D and 3D culture systems such as the following:

- Shape: Cells in 3D are typically ellipsoids with dimensions of 10-30 μm , while cell cultured in 2D are flat with typical thickness of 3 μm .
- Environment: Cells in 3D usually have nearly 100% of their surface area exposed to other cells or matrix, but cells in 2D have only about 50% of their surface area exposed to fluid, approximately 50% exposed to the flat culture surface or intermediate, and a very small percent exposed to other cells.
- Behavior: Cells in 3D comparing with 2D show differences in differentiation, drug metabolism, gene and protein expression, general cell function, in vivo relevance, morphology, proliferation, response to stimuli, and viability.

However, the cons of 3D or in vivo culturing systems might be larger noise than 2D because of increased dimensions and unknown factors. So we checked the shSeq data quality from different environments. We noticed that in vivo data is much noisier than both 2D and 3D data in terms of raw NGS data quality (Figure 9-26) and consistence of replicates (Figure 9-27). There was no significant difference between 2D and 3D data quality, or in some cases, 3D data is much cleaner than 2D. We also checked the distances between samples, and interestingly, 3D and in vivo data are much closer to each other than to 2D (Figure 9-28) and the difference between mutated and wild type samples is much

smaller in 3D than that in 2D, meaning 3D data is less sensitive than 2D in terms identifying candidates from shRNA screening data.

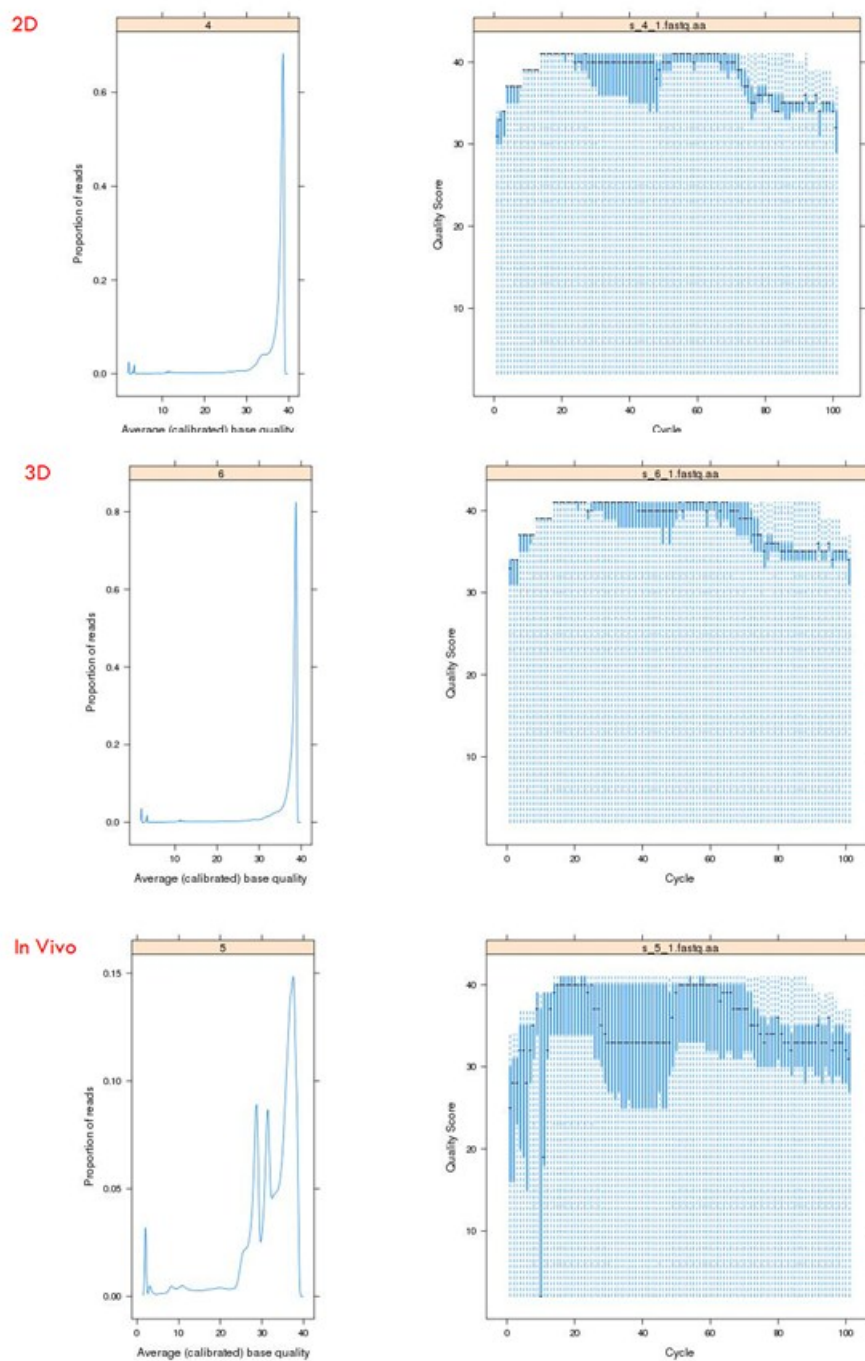


Figure 9-26 Overall quality (left) and cycle-based quality (right) of raw NGS shRNA screening data of cells in 2D, 3D and in vivo environments.

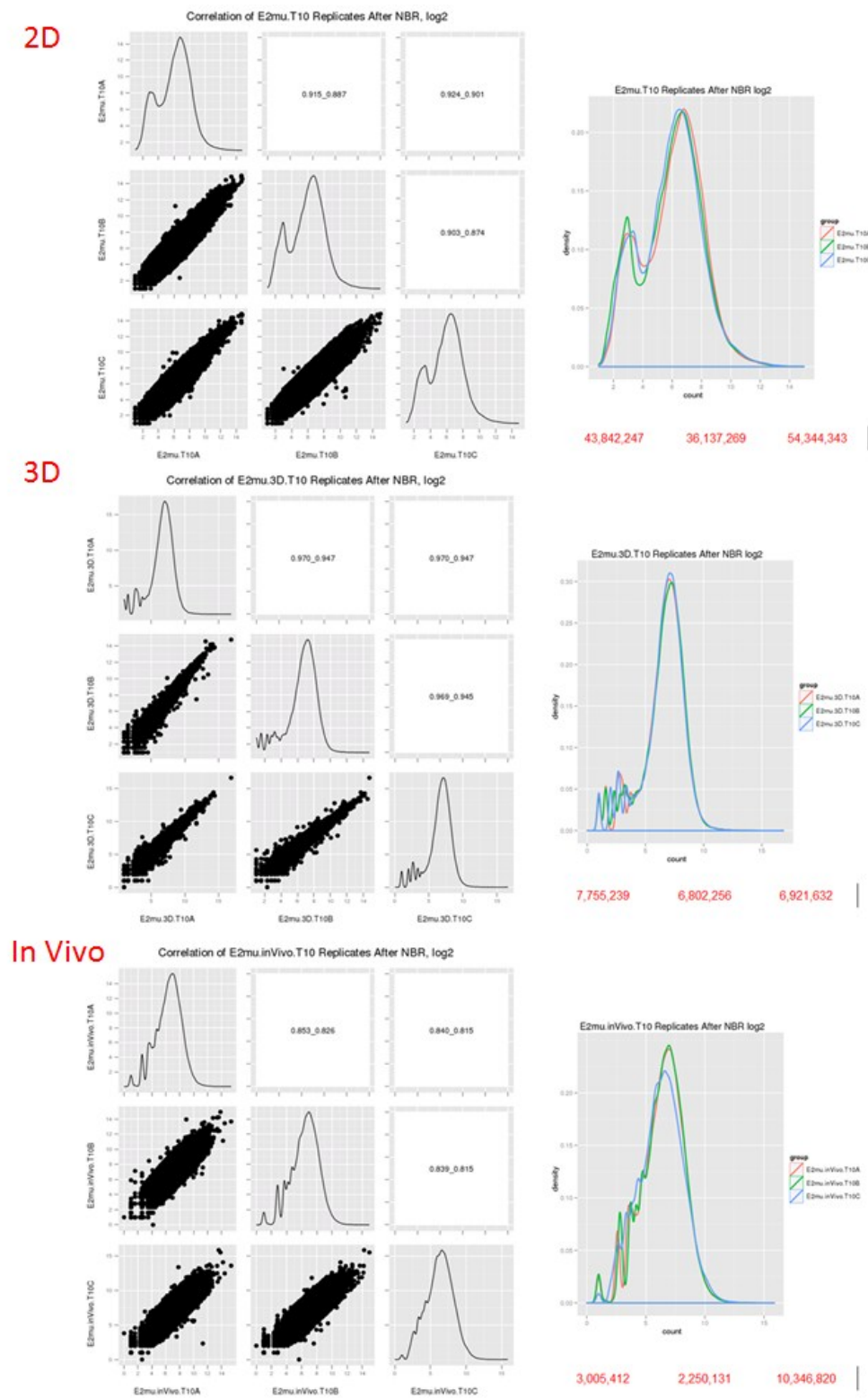


Figure 9-27 Consistency of replicates of NGS shRNA screening data in 2D, 3D and in vivo environments.

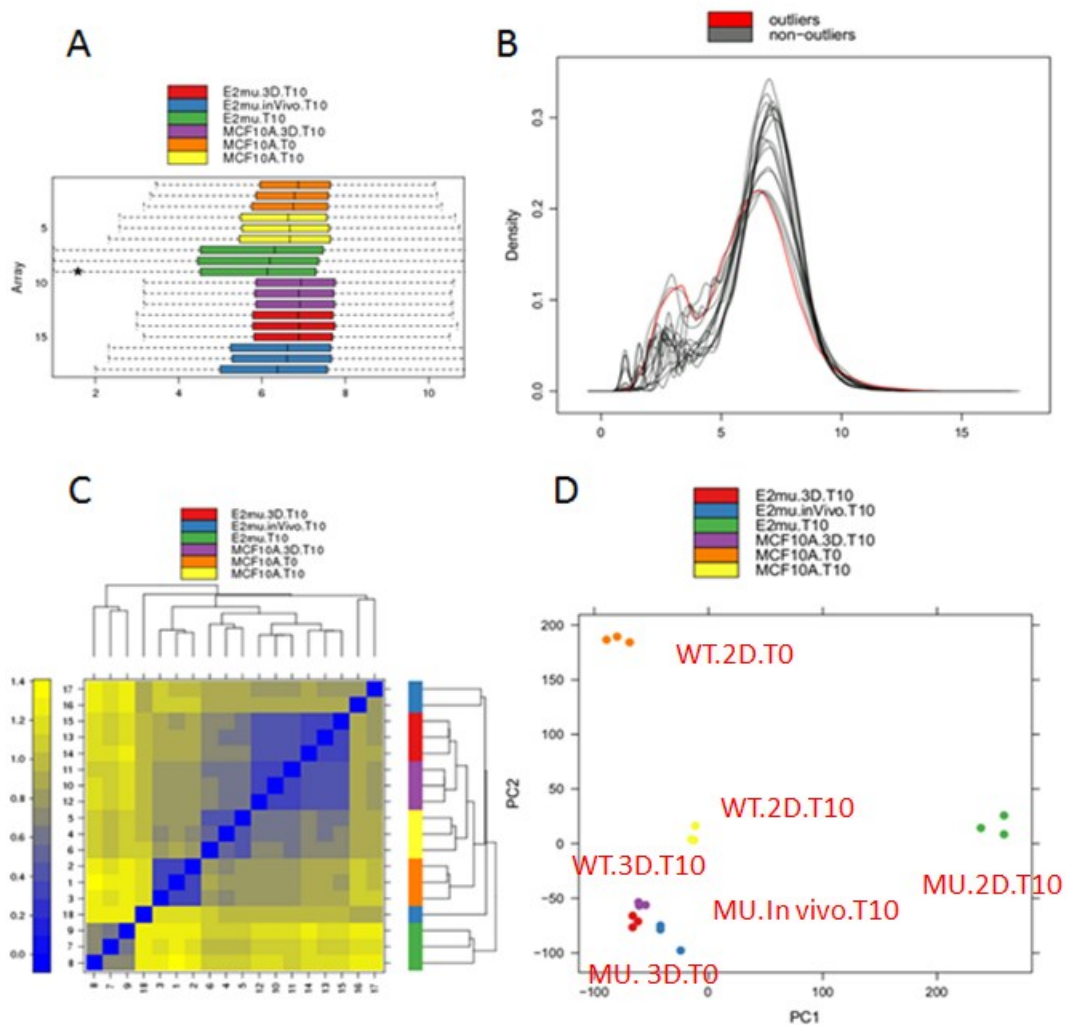


Figure 9-28 Distribution (A: boxplot, B: density) plots, Heatmap of sample distances (C) and PCA (D) plot of 2D, 3D and in vivo data of NGS-based shRNA screening.

9.4 Methods

9.4.1 Reverse engineering transcriptional regulatory or signaling networks of Breast Cancer

To generate breast cancer interactomes we processed microarray gene expression data (Agilent G4502A platform) of 359 breast cancer primary samples from TCGA project [64] using loess normalization and cleaned the dataset to 24,401 probe sets with non-specific filtering. Then we ran the ARACNe algorithm [111] with default parameters against 1775 probes corresponding to 780 TFs to establish a TF-centered interactome and against 6475 probes for 2453 signaling molecule genes to construct a signaling protein-focused network. There are 319 probes for 60 genes that are both TF and signaling proteins.

9.4.2 Signature analysis of ERBB2+ MCF10A model

We did microarray profiling (Agilent) on the isogenetic model, ERBB2 engineered and wild type MCF10A cells to generate a signature of ERBB2+. We cultured the cells in 2D and 3D systems with 6 replicates for each condition (5 replicates for 3D cells) so we can generate two signature reference for 2D and 3D data respectively. The microarray data was normalized by VSN [244] and RSN [245] methods. To generate a reference signature of ERBB2 overexpression phenotype, part of our NetBID2 driver inference algorithm, we used a Probit regression model [89] (Figure 8-2) for its advantage of detecting weak effects. Bayesian-MCMC computing was employed to estimate parameters for its

robustness and accuracy. In particular, a t-distribution prior and Gibbs sampling were used in this analysis [90].

9.4.3 GSEA of inferring regulatory or signaling drivers of ERBB2+ MCF10 Cells

For GSEA method to predict regulatory drivers or signaling modulators of ERBB2+ phenotype, we used BESA method in Chapter 5 with a “maxmean” statistic [98] as enrichment score and Bayesian statistics. 1,000 sample permutations with Efron’s restandarization technique to build the null distribution for statistical significance.

9.4.4 Pooled shRNA screening of ERBB2+ MCF10A cells

We made use of the pGIPZ shRNAmir library [51], comprising 58,493 shRNAs targeting 18,651 genes. We did shRNA screening for ERBB2 mutated and wild type MCF10A cells cultured in 2D and 3D systems separately in triplicates. We also did the screening in mouse models by injecting shRNA library-infected cells with triplicates as well. All genomic data were extracted at T0 and after 10 doubling times. Both microarray (barcode-probed and hairpin-probed) and NGS deep sequencing technologies were used to read out shRNA abundance.

In the NGS data of shRNA screening for 2D, 3D and in vivo models, over 75% of total reads were identified in each case. All samples have enough identified reads to capture signals except two replicates of in vivo mouse model (Table 9-3).

Seq Run	BC1	BC2	BC3	BC4	BC5	BC6	total raw reads	total identified reads	identification rate
1	MYC.mu.T10.A	MYC.mu.T10.B	MYC.mu.T10.C	E2.mu.T10.A	E2.mu.T10.B	E2.mu.T10.C	183,818,501	143,520,666	78.08%
	4,961,923	1,234,337	3,000,547	43,842,247	36,137,269	54,344,343			
2	ClnD1.mu.T10.A	ClnD1.mu.T10.B	ClnD1.mu.T10.C	E2.mu.InVivo.T10.A	E2.mu.InVivo.T10.B	E2.mu.InVivo.T10.C	222,698,642	160,331,245	71.99%
	41,696,724	53,308,248	49,723,910	3,005,412	2,250,131	10,346,820			
3	WT.3D.T10.A	WT.3D.T10.B	WT.3D.T10.C	E2.mu.3D.T10.A	E2.mu.3D.T10.B	E2.mu.3D.T10.C	138,825,905	104,409,077	75.21%
	28,014,416	29,067,457	25,848,077	7,755,239	6,802,256	6,921,632			
4	WT.T10.A	WT.T10.B	WT.T10.C	WT.T0.A	WT.T0.B	WT.T0.C	198,888,638	136,831,544	68.80%
	758,108	100,455	179,420	29,274,603	84,386,971	22,131,987			
5	WT.T10.A	WT.T10.B	WT.T10.C	PTEN.mu.T10.A	PTEN.mu.T10.B	PTEN.mu.T10.C	62,144,097	47,932,112	77.13%
	12,681,031	11,505,499	6,133,125	4,512,187	6,282,855	6,817,415			
MCF10A, Default culture: 2D, use WT data in Run 5 instead of Run 4									

Table 9-3 Summary of deconvolution for NGS data of shRNA screening on ERBB2+ and wild type MCF10A cells in 2D, 3D and in vivo systems. Cells with sky blue background are data for this study. Numbers in dark red background are cases with < 1M identified reads, in light red are cases with 1-5M reads.

9.4.5 Differential representation analysis of individual shRNA

To assess the effects on reversal of GC-resistance by individual shRNA, we compared abundance of shRNA in ERBB2+ with wild type control using shADER algorithm, which is essentially a Bayesian linear model as detailed in Chapter 2.7.

9.4.6 Gene level activity by integration of multiple shRNAs targeting the same gene

To estimate the gene level effects of a gene targeted by multiple shRNAs, we applied BHM algorithm, a hierarchical modeling approach as detailed in Chapter 3. This model allowed “random effects” from different shRNAs, and coefficient of

'fixed effects' was used to score capability of increasing sensitivity at gene level. Bayesian-MCMC computing was set up for accurate estimations.

9.4.7 Meta-analysis of combining differential evidences

To combine evidences from different sources for meta-analysis, for example, to identify depleted genes in ERBB2+ cells with shRNA screening results from microarray data and NGS data under 2D, 3D or in vivo systems, we used Stouffer's z score method [107] shown in the following formula.

$$Z = \frac{\sum_{i=1}^k z_i}{\sqrt{k}}, \quad z_i \sim N(0,1)$$

In the above equation, z_i is the z-score indicating the strength of evidence, for example, differential representation score of a gene or a hairpin, in one source, say number i from total number of k sources. z_i follows a standard normal distribution, so the integrated Z score also follows a standard Gaussian distribution assuming independence of all k evidences. Combined two-tailed p value was calculated based on the integrated Z score.

9.5 Discussion

9.5.1 Phosphorylation of STAT3 is required for STAT3 activity

We showed that STAT3 inhibition is specific to ERBB2+ breast cancer by measuring viability or tumor growth on MDAMB231 cell line, which is ERBB2- but STAT3+, with and without STAT3 silencing both in vitro and in vivo. One interesting point is that although STAT3 is active in MDAMB231 cells at protein

level, phosphorylation of STAT3 is not induced (Figure 9-7, Figure 9-8). This may suggest that STAT3 is actually not functionally active in MDAMB231 cells, making it resistant to STAT3 inhibition treatment. However, this might be explained that ERBB2 induces STAT3 activity by phosphorylation, most likely by indirect phosphorylation via IL6 autocrine signaling loop [230, 246]. In ERBB2-cells such as MDAMB231, STAT3 is not phosphorylated without ERBB2 inducement, thus being inactive.

9.5.2 2D vs. 3D: gene expression signature and NetBID2-predicted drivers

We also had the gene expression profiles of cells in 2D and 3D environments, giving us opportunity to check the gene expression difference between these two culturing methods. First, in gene signature results, 2D and 3D showed significantly difference. For example, the correlation of differential expression scores in 2D (ERBB2+ vs. WT) with those in 3D is poor, only about 0.1. Moreover, among top signature genes in 2D and 3D, there are only about 5-6% genes that were overlapped (Figure 9-29). However, there was a much increased correlation and overlap for NetBID2-predicted drivers for 2D and 3D data (Figure 9-30). This again confirmed the robustness of NetBID2 to detect true phenotype-associated factors.

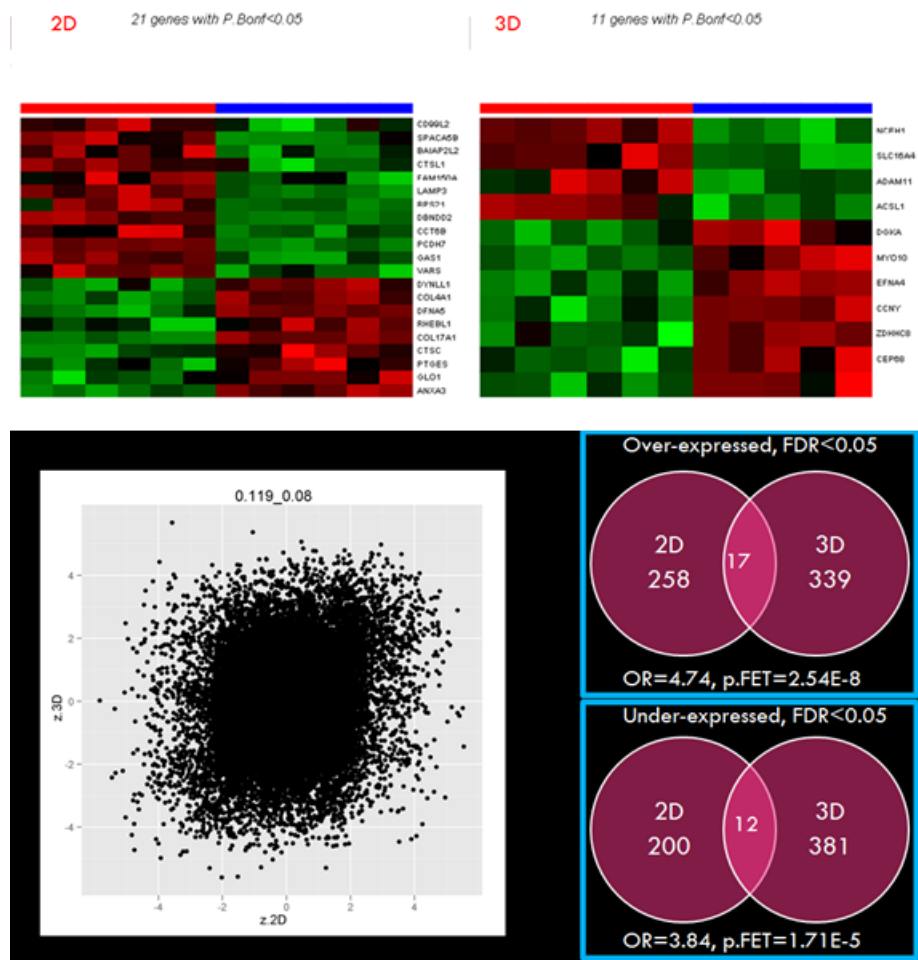


Figure 9-29 Gene expression signature of ERBB2+ MCF10A cells in 2D vs. 3D environments.

Although drivers increased the consistency between 2D and 3D inference, the correlation is still only about 0.3 and the overlap is only about 1/3, so there are still significant differences between 2D and 3D systems, which probably can be only explained by intrinsic differences between these two methods.

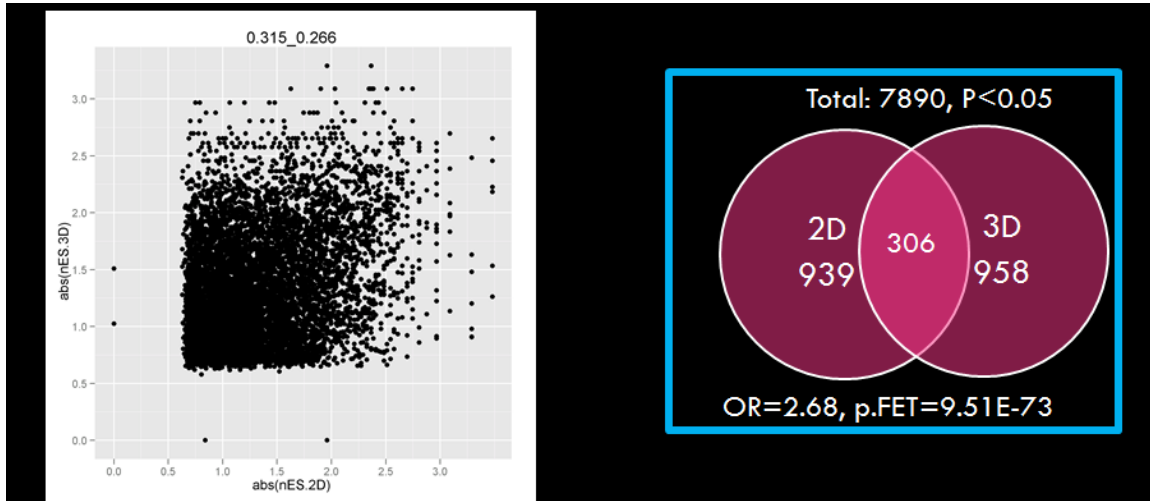


Figure 9-30 NetBID2-predicted drivers of ERBB2+ MCF10A cells in 2D vs. 3D environments.

9.5.3 The power of meta-analysis and integration of functional genomics with systems biology

In previous discussion section, we demonstrated again that NetBID2 framework is much more robust to infer disease or phenotype-associated biomarkers. However, NetBID2 is an application of meta-analysis and actually in shRNA screening data analysis of this study, we showed the power of meta-analysis as well. For example, we had multiple experiments of shRNA screening, 2D, 3D or in vivo environments, microarray or deep sequencing technologies. If we only looked at individual data set, STAT3 didn't show up as in the top candidate list for all of them. However, if we combined all evidences together by meta-analysis, STAT3 was ranked 64th in the combined results (Figure 9-31). Moreover, if we crossed with NetBID2 predictions from cancer genomic data, STAT3 was the number 1 candidate. All these again proved the power of integrating evidences

by meta-analysis. Identification of STAT3 as a validated and effective target for ER- and ERBB2+ breast cancer patients confirmed the success of our strategy by integrating noisy functional genomic RNAi screening data with systems biology inference of large-scaled cancer genomic data to tail therapeutic targets for human cancers.

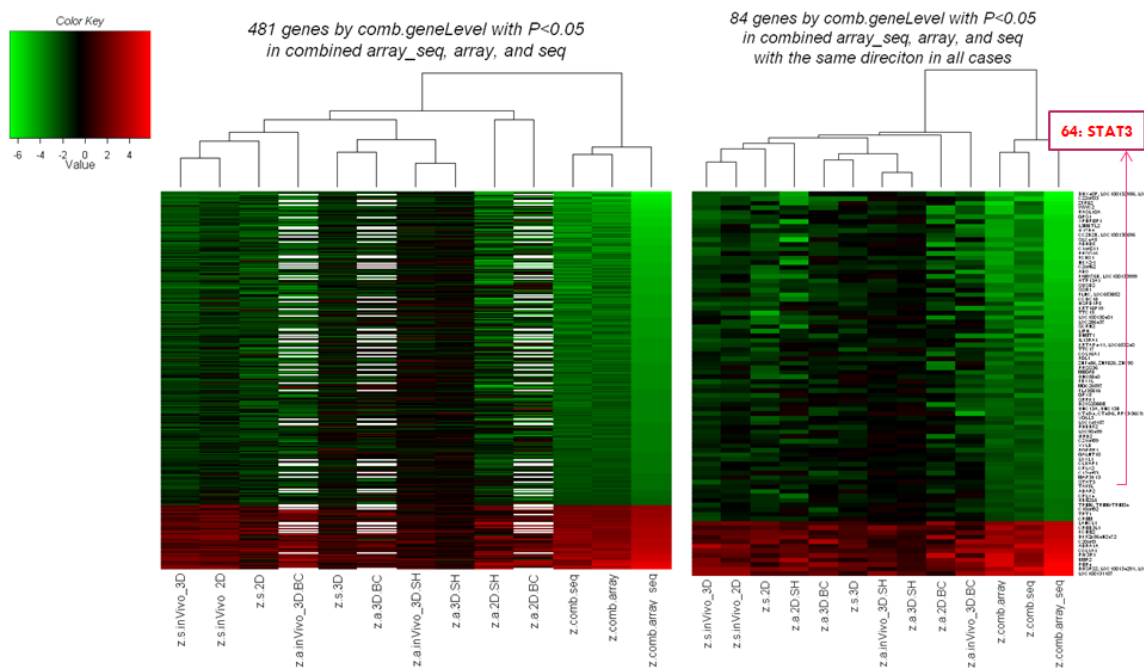


Figure 9-31 Heatmap of z score for top depleted (in green) or enriched (in red) candidates from shRNA screening results of ERBB2 engineered MCF10A cells in 2D, 3D and in vivo system. STAT3 is ranked 64th in the depleted gene list by the combined z-score.

Chapter 10 Integrating Functional Genomics with Systems Biology to Discover Driver-type Therapeutic Targets for ABC or GCB-type DLBCL

10.1 Summary

We have performed genome-wide RNAi screens on four DLBCL cell lines including one ABC-type (HBL1) and three GCB-type (BJAB, Ly7 and SUDHL4), by both microarray and deep-sequencing technologies. To obtain robust candidates from such high-throughput experiments, we designed a procedure to combine results from both microarray and sequencing data. Our analysis led to a genome-wide functional profile for each cell line, indicating gene-silencing effects on cell proliferation. Un-supervised clustering showed a clear separation between ABC and GCB. Supervised comparison of RNAi screen between ABC and GCB generated 587 candidate genes ($P < 0.001$) that are specifically lethal to ABC or GCB. Besides, based on a cohort of 260 DLBCL gene expression profiles, we built a B-cell interactome computationally and performed NetBID2 analysis to identify drivers that are specific to ABC or GCB subtype. This analysis gave us 125 master regulators or signaling factors ($P < 0.001$) mediating expression signature of ABC vs. GCB. By integrating RNAi screened candidates with genomics-inferred drivers of ABC vs. GCB, we obtained 20 transcription factors and 47 signaling molecules that are both lethal to and critical of mediating ABC-DLBCL. Out of those candidates, four genes (TCF4, ZCCHC24, CILP, and

PTPRG) also showed significant gain or amplification in ABC patients from copy number variation data, which may constitute promising therapeutic targets for ABC, which is usually associated with poor prognosis. Further biochemical experiments are being conducted to validate selected candidates both *in vitro* and *in vivo*.

10.2 Introduction

The goal of this project is to identify novel therapeutic targets that are specific to ABC or GCB-type of DLBCL. The approach we developed to address this task is similar to what we have developed in Chapter 10 and chapter 11, an integrative framework of crossing genome-wide RNAi screens with systems biology analysis of cancer genomics (Figure 10-1). On one side, we performed genome-wide shRNA screens on four DLBCL cell lines including one ABC-type (HBL1) and three GCB-type (BJAB, Ly7 and SUDHL4), by both microarray and deep-sequencing technologies. With sophisticated meta-analysis of combining microarray and sequencing data, we were able to identify candidates that are lethal to ABC or GCB-type DLBCL lines as potential therapeutic targets. However, this list is usually too long to validate all of them. On the other side, we had a cohort of 260 gene expression profiles from primary patients or cell lines in the context B cells, among which 35 are ABC and 50 are GCB-type. With this data, we applied the NetBID2 algorithm to identify drivers, both master regulators and signaling modulators, which are specific to ABC or GCB subtype. We also had copy number variants (CNV) data for 72 out of those 260 samples, which helped

us to identify genes that are amplified or depleted in ABC or GCB groups. Then we integrated all these evidences together and produced a short list of candidates as therapeutic targets for ABC or GCB subtype of DLBCL.

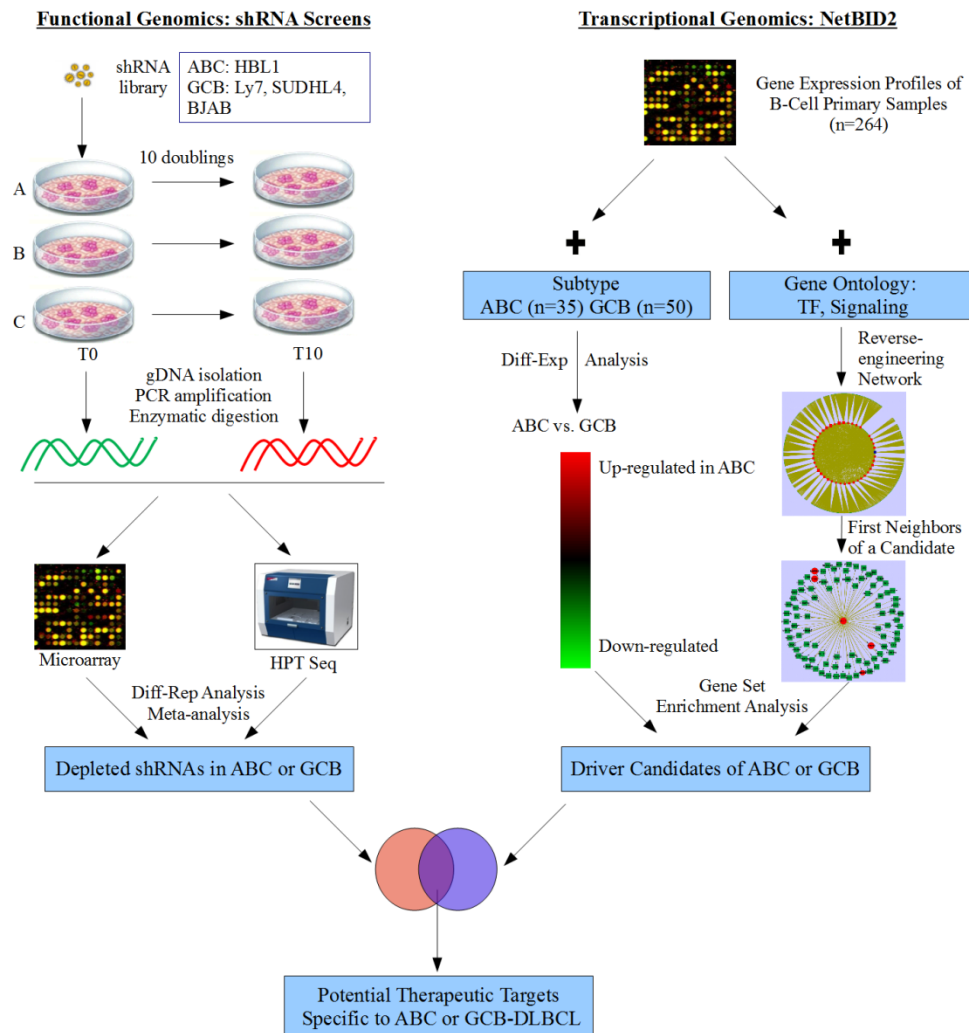


Figure 10-1 The integrative framework to identify therapeutic targets for ABC or GCB-type DLBCL by integrating genome-wide RNAi screens (left) with systems biology (NetBID2) of cancer genomics.

10.3 Results

10.3.1 Pooled shRNA screens of DLBCL lines by microarray and NGS

We did pooled shRNA screens for four DLBCL cell lines using both microarray (Barcode-probed and hairpin-probed) and NGS technologies. All screens except SUDHL4 with shRNA-probed microarray and BJAB with sequencing are in good quality (Table 10-1). There were significant batch effects for sense-probes and anti-sense probes in hairpin-probed microarray data of SUDHL4 line (Figure 10-2). Two samples of BJAB shSeq data didn't have enough total number of identified reads, which caused the data noisy (Table 10-2). The best way to get robust candidates out of all these data is to exclude the bad screening data, perform analysis on individual data set and then integrate the others together. The Stouffer's or naïve Bayesian method was used to combine multiple evidences.

Cell Line	COO Subtype	BCL2 R	BC-Array	Sense shRNA-array	Anti-sense shRNA-array	Seq
HBL1	ABC	neg	✓	✓	✓	✓
Ly7	GCB	neg	✓	✓	✓	✓
SUDHL4	GCB	t(14;18)	✓	✗	✗	✓
BJAB	GCB	t(14;18)	✓	✓	✓	✗

Table 10-1 Summary for genome-wide shRNA screens of four DLBCL cell lines. Green check sign indicates the data quality is good while red one represents that that data is not good or missing.

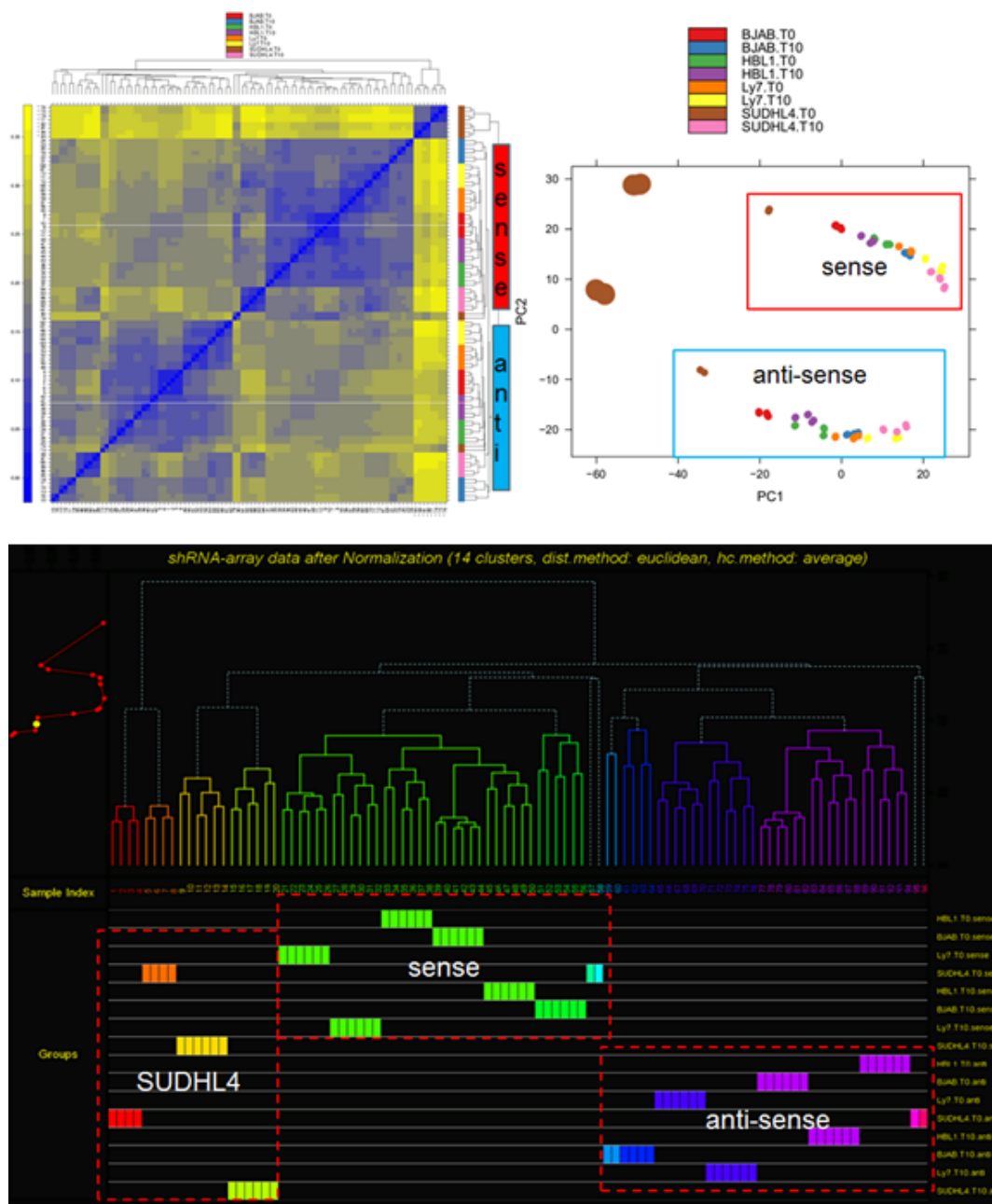


Figure 10-2 Batch effects detected for shRNA hairpin-probed microarray data of SUDHL4 cell line.

	T10.A	T10.B	T10.C	T0.A	T0.B	T0.C	total raw reads	total identified reads	identification rate	T10/T0 (total identified reads)
HBL1	9,183,261	9,649,385	7,956,819	7,799,504	8,430,612	7,186,548	66,076,022	50,206,129	75.98%	1.14
	157	165	136	133	144	123				
BJAB	8,395,783	3,893,188	7,017,524	9,986,872	9,024,446	10,515,873	64,836,232	48,833,686	75.32%	0.65
	144	67	120	171	154	180				
Ly7	2,382,978	5,790,271	3,960,236	18,924,340	13,130,577	14,916,390	77,398,420	59,094,792	76.35%	0.26
	41	99	68	324	224	255				
Old SUDHL4 @Mount Sinai	11,736,316	2,709,628	2,425,232	32,516,236	35,301,521	17,788,479	143,314,459	102,477,612	71.51%	0.20
	201	46	41	556	604	304				
Columbia_1 st	5,581,697	6,749,176	6,158,727	5,722,510	3,879,742	8,153,633	144,623,162	36,245,485	25.06%	1.04
	95	115	105	98	66	139				
Columbia_2 nd	4,419,693	4,327,785	3,895,593	4,022,553	3,453,916	4,040,091	31,673,961	24,159,631	76.52%	1.10
	76	74	67	69	59	69				
Columbia_3 rd	28,857,298	26,681,468	24,699,909	23,507,713	23,961,589	27,057,503	210,875,523	154,765,480	73.39%	1.08
	493	456	422	402	410	463				

Table 10-2 Summary of deconvolution of NGS-based shRNA screening data of four DLBCL cell lines. Red ones are the run with not enough signals. SUDHL4 was run three times to get good quality data.

10.3.2 Clustering of shRNA screening samples

First we did hierarchical clustering of all T10 or T0 samples for each of the four DLBCL lines in both NGS and microarray platforms to check the consistence of replicates and to check biological relationships among T0 and T10 data. As shown in Figure 10-3, most replicates for each condition are clustered together, indicating consistence of biological replicates and good overall quality of the data.

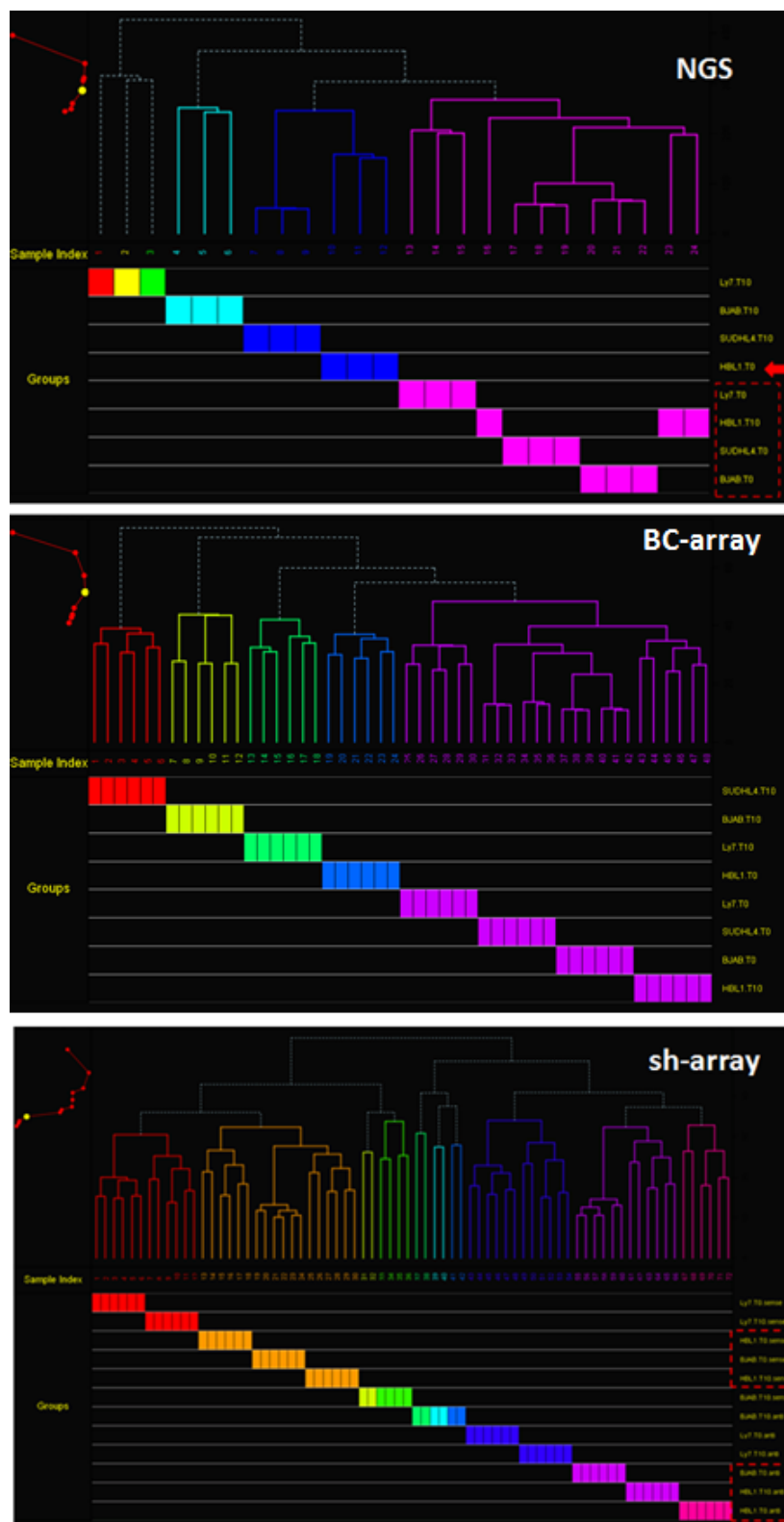


Figure 10-3 Clustering of T0 or T10 shRNA screening data in NGS, BC-probed microarray and shRNA hairpin-probed microarray.

10.3.3 Functional profiles separate well ABC from GCB, BCL2-rearranged from non-rearranged DLBCL subtypes

The four DLBCL cell lines can be classified into ABC and GCB subtype based on their gene expression profiles, or BCL2-rearranged or non-rearranged based on BCL2 translocation status. We wondered whether we can still separate those subtypes by using the functional profiles from shRNA screening data. The answer is yes.

First, we generated a new profile of differential representation for each cell line using the difference of T10 and T0 data ($\log(T10/T0)$). This profile indicates the functional effects of each hairpin or gene to cell growth or survival of the cell line, for example, positive value for enriched hairpins meaning the targeting genes are suppressors of cell growth, while negative value for depleted hairpins representing that corresponding genes are lethal to this cell line. Then we did clustering of the four lines based on their functional profiles. Here we only showed you the results using NGS-based shRNA screening data, but the results from microarray data are similar. Since there are triplicates for each cell line, we enumerated all six possible pairs of T10 vs. T0 therefore generated six new functional profiles for each cell line (except Ly7 for which we removed two bad replicates, therefore it only has three profiles).

As shown in Figure 10-4, first, we noticed that all generated profiles for each cell line are clustered together as expected; second, HBL1 in blue, the only ABC cell line is clearly separated from the other three GCB lines; third, among three GCB

lines, BJAB in red and SUDHL4 in purple stayed together showing clear difference from Ly7 because both of them are BCL2-translocated, while Ly7 is close to HBL1 as both of them are not BCL2-translocated. We saw exactly the same pattern if we averaged all generated profiles for each cell line as shown in Figure 10-5.

In conclusion, shRNA screens-produced functional profiles are able to separate ABC and GCB, the two major subtypes of DLBCL and are also able to classify BCL2-rearranged and non-rearranged samples. Also the difference of ABC with GCB subtype is larger than the signal of BCL2-rearrangement.

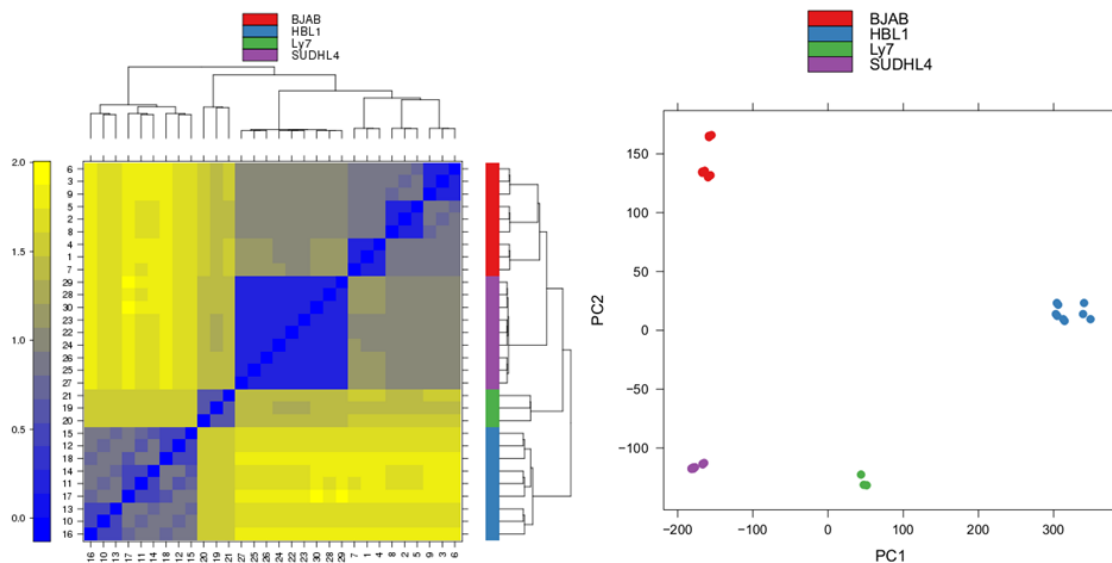


Figure 10-4 Heatmap with hierarchical clustering (left) and PCA plot (right) of generated functional profiles from NGS-based shRNA screening data on four DLBCL cell lines. Six profiles for cell lines with triplicates for both T10 and T0 data. For Ly7, two T10 replicates are bad and removed, therefore only three generated profiles.

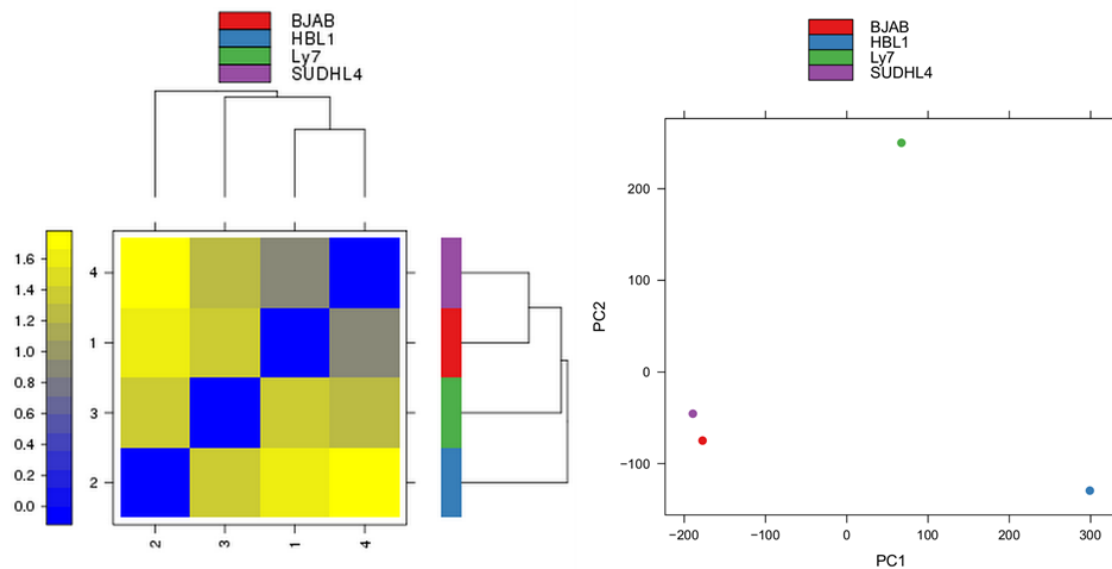


Figure 10-5 Heatmap with hierarchical clustering (left) and PCA plot (right) of generated functional profiles from NGS-based shRNA screening data on four DLBCL cell lines. All generated profiles for each cell line are averaged to produce only one profile.

10.3.4 Differentially represented genes from shRNA screens

We performed differential representation analysis at gene level using BHM algorithm for each of the four DLBCL lines to identify genes whose hairpins in the library are either enriched or depleted at T10 time. Deleted genes are of interest as they are genes that are lethal to the cells and are potential therapeutic targets to kill DLBCL cells of study. We did the analysis for individual cell lines. Statistical results including p-value and z-score are summarized in Figure 10-6.

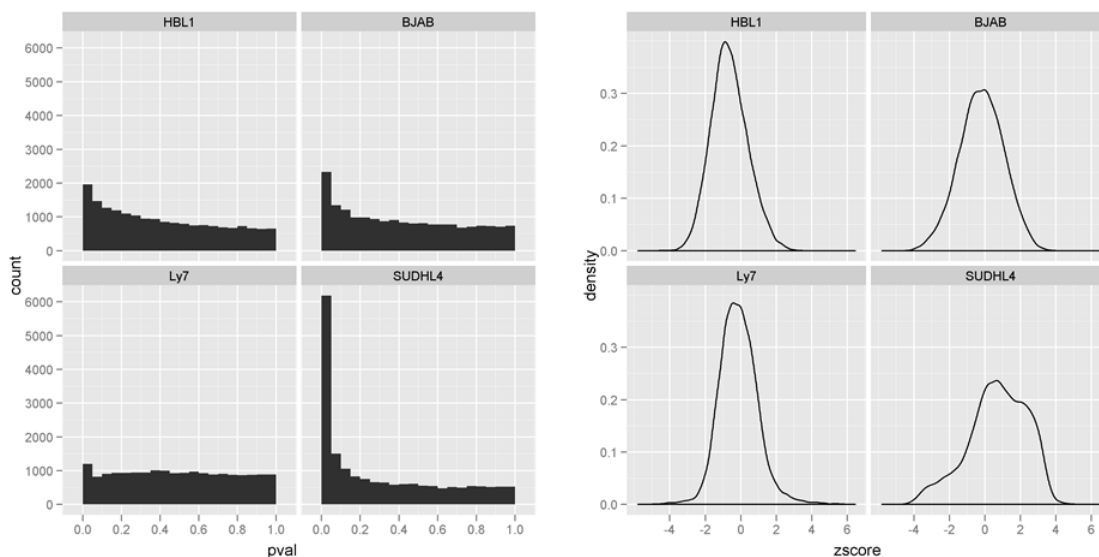


Figure 10-6 Histogram of p-values (left) and density plot of z-scores (right) for gene-level differential representation analysis of each cell line. The bin width for p-value histogram is 0.05.

We also performed a meta-analysis to combine results of four lines for each gene by Stouffer's method. The combined z score for gene X indicates the overall effects of silencing X on all four DLBCL lines. The number of significantly depleted or enriched genes in each cell line and combined analysis is summarized in Table 10-3. For example, there are 1962 significantly differentially represented genes in HBL1, in which 1783 are depleted and 179 are enriched. There is a significant bias between depleted genes and enriched ones in HBL1, BJAB and SUDHL4, which might reflect the sensitivity difference of different type of cells.

Cell line	n.total	n.over	n.under
HBL1	1962	179	1783

BJAB	2329	770	1559
Ly7	1204	627	577
SUDHL4	6184	4738	1446
Combined	2468	870	1598

Table 10-3 Summary of enriched or overrepresented and depleted or under-represented genes in shRNA screening for each cell line. “Combined” is using Stouffer’s method to integrate all four cell lines. It’s based on gene level results with selection threshold of $P < 0.05$.

We also calculated the number of genes that are depleted in at least a certain number of cell lines (Table 10-4) to indicate its lethal effects on the majority of DLBCL cells.

n.lines	1	2	3	4	sum
n.all	5936	2237	379	33	8585
n.over	4775	708	41	0	5524
n.under	3716	716	71	1	4504

Table 10-4 Number of genes depleted (under) or enriched (over) in at least 1 or 2 or 3 or 4 cell lines, based on gene level results with selection threshold of $P < 0.05$.

10.3.5 Top differentially represented genes cross all cell lines

We applied a stringent criteria and selected top depleted or enriched genes in all cell lines. With a threshold of p-value less than 0.05, 33 genes showed up as either under-represented or over-represented in all cell lines (Figure 10-7, Table

10-5). Unfortunately, there is only gene RFC3 that is depleted in all four lines. This reflects the heterogeneity of DLBCL.

We also did clustering using the profiles of selected 33 genes, and again there was a clear separation between ABC (HBL1) and GCB lines. The majority of these 33 genes are exclusively either depleted in ABC or depleted in GCB lines. However, there are a few genes that are specific to BCL2-translocation. For example, ACCS and AKT1 are depleted in BJAB and SUDHL4, two BCL2-translocated lines, while THS07A is depleted in two BCL2 non-translocated lines (Ly7 and HBL1).

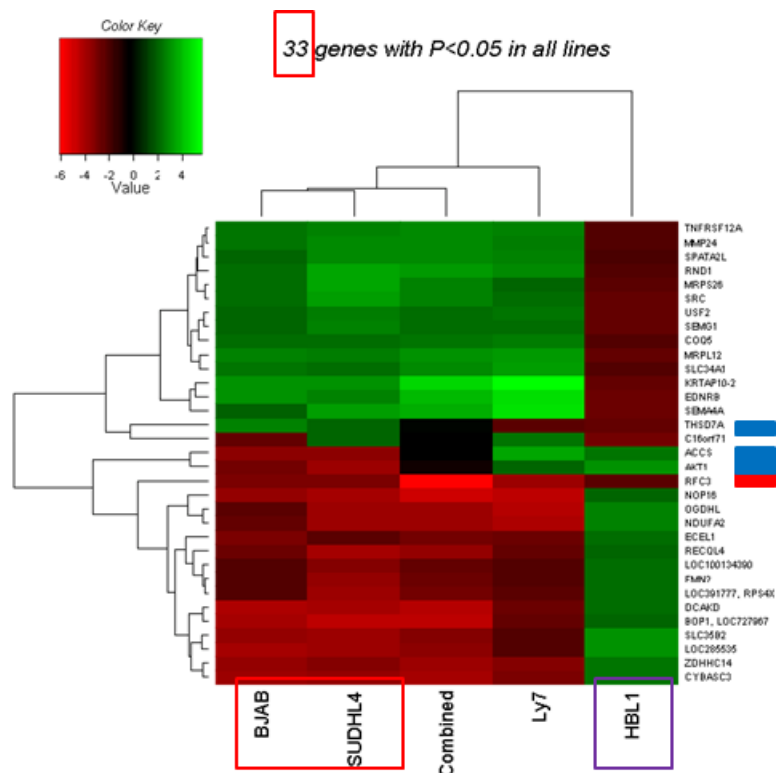


Figure 10-7 Top genes selected by a threshold of $P < 0.05$ in all four lines. Red is for depletion while green is for enrichment.

geneSymbol	n.shRNAs	n.shRNAs.array	numPathways	z.comb	p.comb	z.HBL1	z.Ly7	z.BJAB	z.SUDHL4
RFC3	1	1	26	-6.08	1.20E-09	-2.26	-3.82	-3.04	-3.03
						ABC		GCB	
NOP16	1	1		-4.92	8.78E-07	2.11	-4.48	-3.49	-3.97
BOP1, LOC727967	1			-4.42	9.79E-06	2.08	-2.54	-4.02	-4.37
DCAKD	1	1		-4.31	1.61E-05	2.23	-2.75	-4.13	-3.98
CYBASC3	1	1		-3.90	9.58E-05	2.48	-3.01	-3.81	-3.46
NDUFA2	1	1	2	-3.89	1.01E-04	2.80	-4.18	-2.63	-3.76
OGDHL	1	1		-3.75	1.77E-04	2.68	-4.31	-2.28	-3.59
ZDHC14	1	1		-3.72	2.02E-04	2.46	-3.20	-3.49	-3.20
RECQL4	1	1		-3.48	5.08E-04	2.10	-2.43	-2.67	-3.95
LOC285535	1			-3.31	9.39E-04	3.06	-2.10	-3.91	-3.67
SLC35B2	1	1	4	-3.25	1.16E-03	3.01	-2.17	-3.46	-3.87
LOC391777, RPS4X	1		30	-2.83	4.66E-03	2.33	-2.21	-2.17	-3.62
ECEL1	1	1		-2.83	4.67E-03	2.31	-2.72	-3.05	-2.20
FMN2	3	2		-2.77	5.57E-03	2.21	-2.15	-2.17	-3.44
LOC100134390	1			-2.59	9.71E-03	2.31	-2.13	-2.15	-3.20
SEMG1	1	1		2.21	0.03	-2.58	2.21	2.10	2.69
USF2	1	2	43	2.33	0.02	-2.51	2.49	2.08	2.60
COQ5	1	1		2.46	0.01	-2.17	2.61	2.26	2.22
SRC	2	3	113	2.67	7.64E-03	-2.41	2.24	2.19	3.31
MRPS26	1	1		2.70	7.01E-03	-2.33	2.04	2.25	3.44
SPATA2L	1	1		2.84	4.48E-03	-1.96	2.67	2.04	2.94
SLC34A1	1	1	5	2.87	4.13E-03	-2.14	3.30	2.38	2.19
MMP24	1	1		2.87	4.05E-03	-2.12	2.56	2.44	2.87
TNFRSF12A	3	4		2.92	3.54E-03	-2.04	2.67	2.43	2.77
MRPL12	1	1		3.04	2.33E-03	-2.54	3.20	2.80	2.63
RND1	1	1	6	3.27	1.07E-03	-2.08	2.86	2.19	3.57
SEMA4A	1	1	3	3.75	1.79E-04	-2.74	4.88	2.03	3.33
EDNRB	3	1	45	3.97	7.06E-05	-2.69	4.78	3.06	2.79
KRTAP10-2	1	1		4.67	3.07E-06	-2.43	5.62	3.07	3.08

Table 10-5 Top genes depleted in all cell lines, and depleted in ABC or GCB cell lines only. Red indicates depletion while green indicates enrichment. “n.shRNAs” is the number of shRNAs in NGS-based data and “n.shRNAs.array” is the number of hairpins in microarray data. “numPathways” is the number of known pathways being involved in.

We also tried to loosen the threshold and selected genes by combined statistics. The goal was to identify genes that showed consistent effects cross all cell lines, either lethal to all of them or suppressing their growth upon silencing. With a combined p-value threshold of 0.001, 293 genes were selected (Figure 10-8) and clustering analysis showed the same pattern as we observed using all genes or top differentially-represented genes.

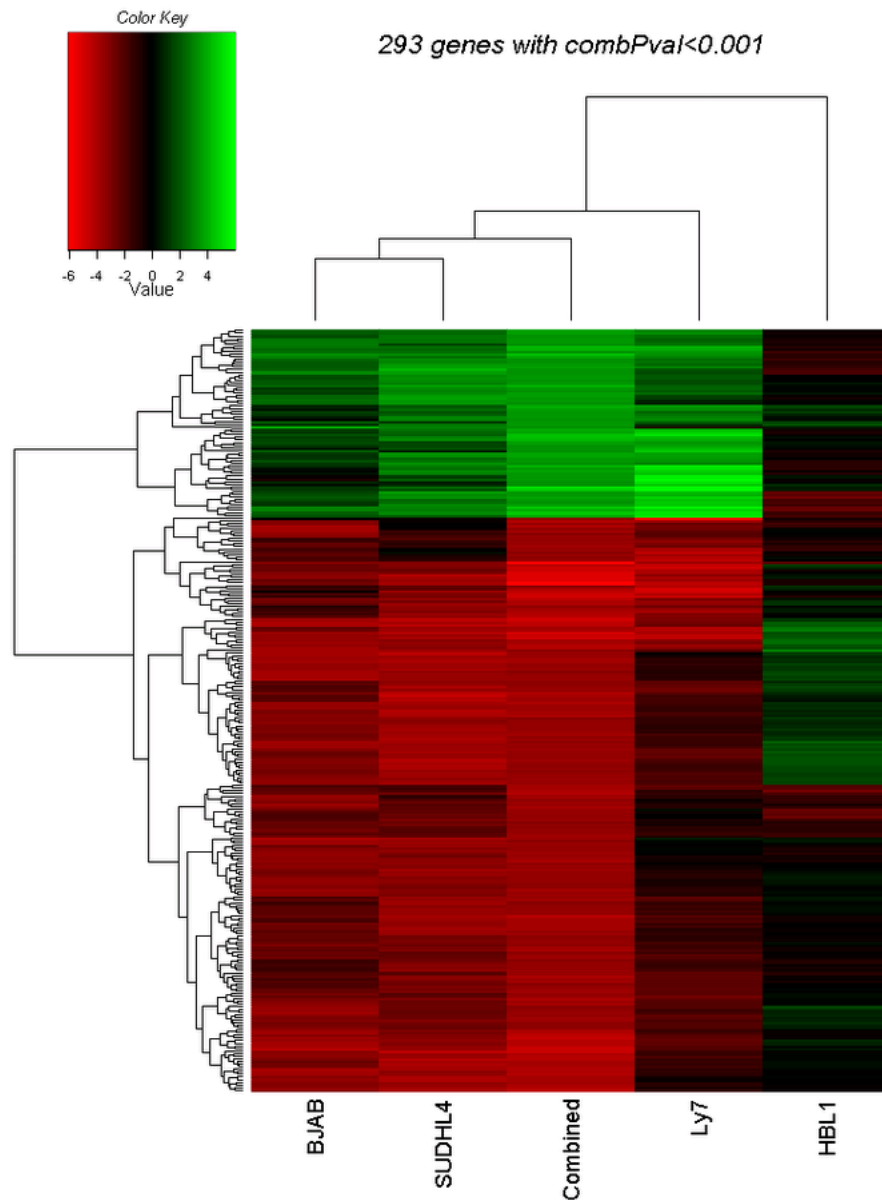


Figure 10-8 Top genes selected by combined p-value with a threshold of 0.001.

10.3.6 Enriched pathways by RNAi screening identified candidates

With a long list of candidates for each cell line, we were interested in whether those lethal candidates are enriched in any known pathways, or whether there are common lethal pathways shared by different subtypes of DLBCL. Therefore,

we performed the functional enrichment analysis by using my BSEA (Chapter 5) algorithm. As shown in the top enriched pathways (Table 10-6) in at least one of the four cell lines, there are a significant number of pathways that are lethal to HBL1, the ABC type of DLBCL. Details about top two pathways that are lethal to HBL1 are shown in Figure 10-9 and Figure 10-10.

setName	setSize	hitSize	z.HBL1	z.BJAB	z.Ly7	z.SUDHL4
IL23-mediated signaling events;NCL_NATURE	66	62	-3.89	-0.10	1.20	2.95
Alternative NF-kappaB pathway;NCL_NATURE	6	6	-3.72	1.74	0.83	2.40
NGF signalling via TRKA from the plasma membrane;REACTOME	52	42	-3.14	0.40	-0.29	2.13
Other semaphorin interactions;REACTOME	14	11	-3.09	0.33	2.24	0.85
Smooth Muscle Contraction;REACTOME	11	10	-3.01	0.77	1.33	2.04
Vitamins B6 activation to pyridoxal phosphate;REACTOME	2	2	-3.00	1.42	1.15	1.97
Vif-mediated degradation of APOBEC3G;REACTOME	53	46	-2.86	-1.07	2.05	1.22
APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1;REACTOME	65	56	-2.82	-1.03	2.13	1.37
CD40/CD40L signaling;NCL_NATURE	53	46	-2.76	-0.02	0.95	2.02
IL2 signaling events mediated by PI3K;NCL_NATURE	63	55	-2.71	0.02	1.71	2.64
Degradation of beta-catenin by the destruction complex;REACTOME	49	43	-2.71	-0.97	2.09	1.17
Regulation of activated PAK-2p34 by proteasome mediated degradation;REACTOME	50	44	-2.70	-1.28	2.08	0.87
Signalling events mediated by TCPTP;NCL_NATURE	90	80	-2.68	-0.30	2.01	1.15
Signalling by Wnt;REACTOME	49	43	-2.68	-0.99	2.04	1.18
Canonical NF-kappaB pathway;NCL_NATURE	35	32	-2.66	0.10	1.16	2.11
SCF-beta-TrCP mediated degradation of Emi1;REACTOME	49	43	-2.65	-0.98	2.07	1.15
p53-Independent DNA Damage Response;REACTOME	54	47	-2.64	-1.40	2.01	0.78
Vpu mediated degradation of CD4;REACTOME	50	44	-2.63	-0.96	2.21	1.12
p53-Independent G1/S DNA damage checkpoint;REACTOME	54	47	-2.62	-1.40	2.02	0.78
Ubiquitin Mediated Degradation of Phosphorylated Cdc25A;REACTOME	53	46	-2.56	-1.53	2.15	0.79
PI-3K cascade;REACTOME	3	3	-2.40	1.30	-0.70	2.43
FRS2-mediated cascade;REACTOME	3	3	-2.39	1.31	-0.70	2.41
Opioid Signalling;REACTOME	13	13	-2.38	-0.53	-0.17	2.21
G-protein activation;REACTOME	1	1	-2.36	-0.83	0.91	2.16
Ubiquitin-dependent degradation of Cyclin D;REACTOME	52	46	-2.31	-1.66	1.97	0.60
Destabilization of mRNA by AUF1 (hnRNP D0);REACTOME	53	45	-2.30	-1.06	2.27	1.04
IL12-mediated signaling events;NCL_NATURE	104	90	-2.29	0.21	1.16	2.59
Ion transport by P-type ATPases;REACTOME	4	3	-2.26	0.72	2.02	1.50
Signalling by BMP;REACTOME	4	4	-2.18	0.89	0.20	2.59
Signalling by FGFR;REACTOME	10	9	-2.07	0.74	-0.38	2.60
Endogenous TLR signaling;NCL_NATURE	56	51	-2.06	0.00	1.42	2.32
GPVI-mediated activation cascade;REACTOME	15	13	-2.06	1.04	0.74	2.06
Platelet Adhesion to exposed collagen;REACTOME	14	12	-2.04	2.06	0.85	1.32
Downstream signaling of activated FGFR;REACTOME	6	6	-2.03	1.03	-0.08	2.51
Integrins in angiogenesis;NCL_NATURE	61	57	-1.97	0.04	2.66	1.34
HIV-1 Nef: Negative effector of Fas and TNF-alpha;NCL_NATURE	35	33	-1.97	1.29	1.33	3.01
Collagen adhesion via alpha 2 beta 1 glycoprotein;REACTOME	4	4	-1.96	1.86	1.01	1.98

Table 10-6 Top pathways enriched in shRNA screening-identified candidates.

Red means genes in the pathway are significantly enriched in the under-represented genes in that cell line, while green is for enrichment in over-represented side.

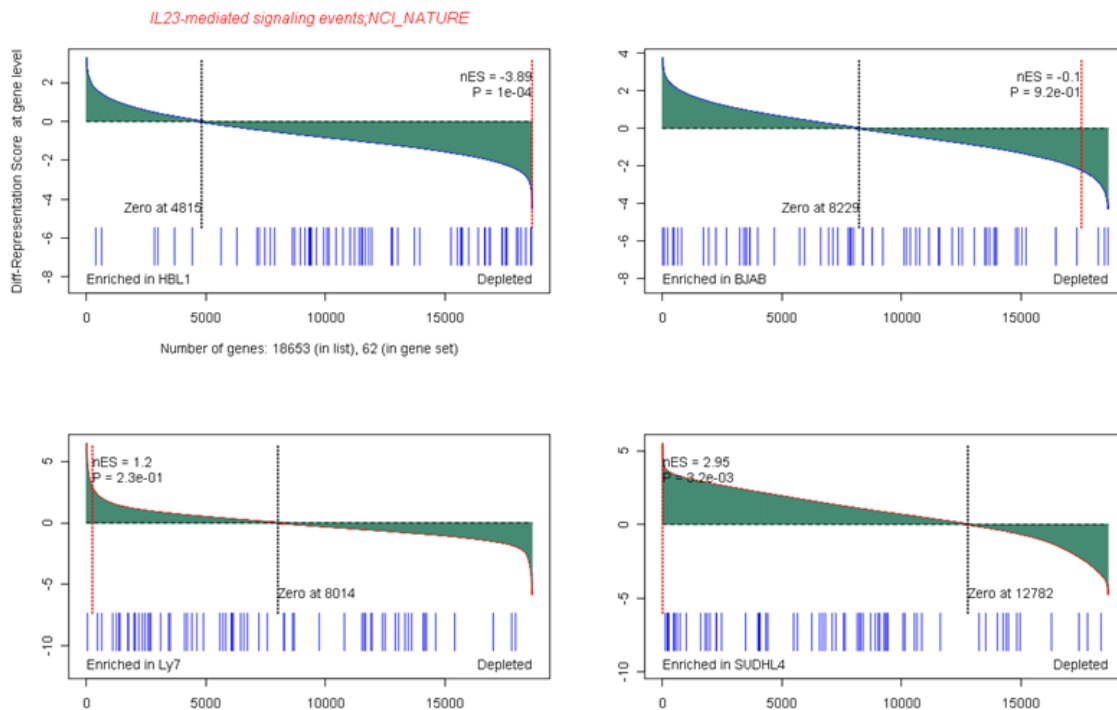


Figure 10-9 IL23-mediated signaling pathway is enriched by lethal genes in HBL1, but not by lethal genes in other cell lines.

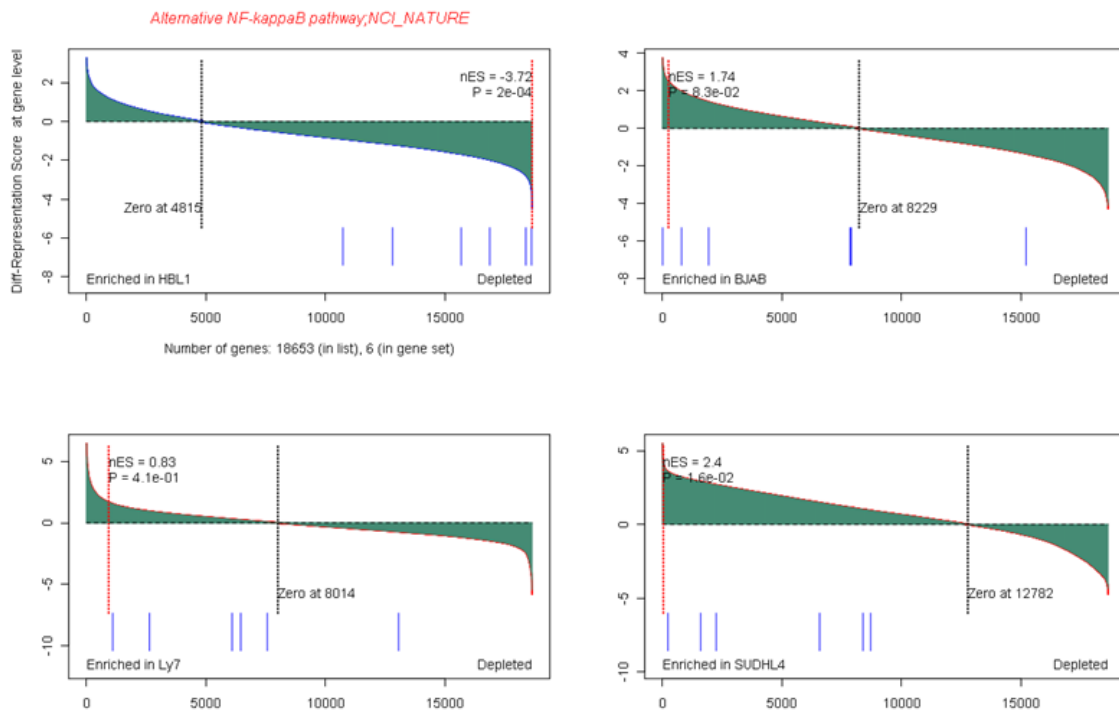


Figure 10-10 Alternative NF κ B pathway is enriched by lethal genes in HBL1, but not by lethal genes in other cell lines.

10.3.7 Crossing RNAi screening with signature genes of ABC vs. GCB

Out of the cohort of 230 gene expression profiles from DLBCL primary samples and cell lines, 35 are ABC type and 50 GCB type. Although I mentioned the chapter of NetBID2 algorithm that gene expression signature is not robust, the signature for ABC and GCB might be an exception because these two subtypes are classified based on gene expression signature data. So using the profiles of 35 ABC and 50 GCB samples, we generated a signature for ABC vs. GCB type of DLBCL (Figure 10-11). And crossed top signature genes with RNAi screening identified candidates. The selection criteria we applied are the following:

- GEPs: as a signature, $P < 0.05$
- RNAi Screen:
 - Significant in combined ABC vs. GCB: $P < 0.05$
 - Depleted in either only HBL1 (ABC) or only all three GCB lines
- Significant in combined GEP and shRNA: $P < 0.05$

With the above selection, 141 genes showed up as shown in Figure 10-12.

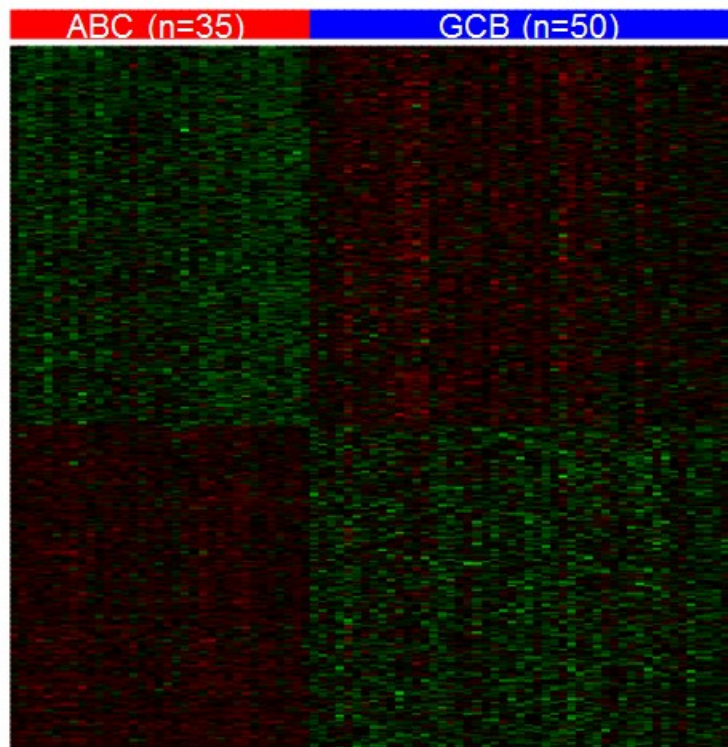


Figure 10-11 Heatmap of top signature genes of 35 ABC vs. 50 GCB DLBCL profiles. Red means under-expression in ABC relative to GCB, while green is for over-expression in ABC.

negative z.GEP is used to indicate the evidence of being depleted and over-expressed in one subtype.

10.3.8 Crossing RNAi screening with NetBID2-predicted drivers of ABC vs. GCB

More interesting and more robust integration is to cross our NetBID2-predicted drivers with RNAi screening results to identify driver-type therapeutic targets for ABC or GCB-type DLBCL. We used the cohort of 230 DLBCL samples to build a B-cell interactome and applied NetBID2 on the signature of 35 ABC vs. 50 GCB samples. We applied the following selection criteria and identified 64 driver-type therapeutic targets for ABC or GCB-DLBCL (Figure 10-13).

- GEPs: as a MR, $P < 0.05$
- RNAi Screening:
 - Significant in combined ABC vs. GCB: $P < 0.05$
 - Depleted in either only HBL1 (ABC) or only all three GCB lines

If we loosed the selection criteria from requiring depletion in all GCB lines to being depleted in at least one GCB line, we got 35 more candidates (Figure 10-14).

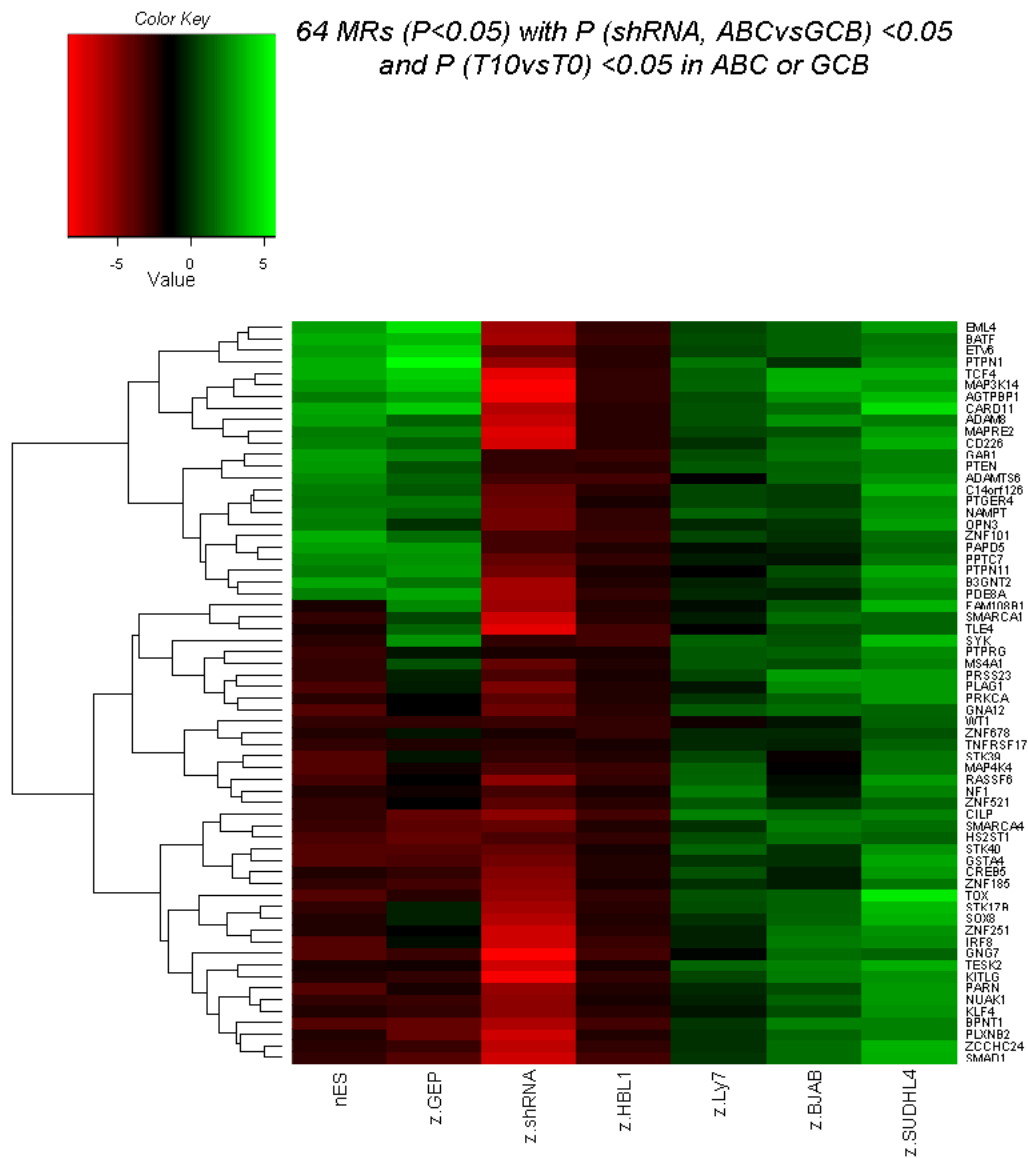


Figure 10-13 Heatmap of top candidates crossing RNAi screening results with NetBID2-predicted drivers specific to ABC or GCB-DLBCL. “nES” is the normalized enrichment score as evidence of being a driver. “z.GEP” is the differential expression of ABC vs. GCB.

10.3.9 Crossing RNAi screening with CNV data

In this study, we also had copy number variants (CNV) data for 29 ABC and 26 GCB DLBCL primary samples. So with CNV data, we identified a list of genes that are amplified or depleted in ABC vs. GCB samples (Figure 10-15). Then we crossed with RNAi screening results and identified 48 candidates (Figure 10-16) with the following criteria:

- CNVs: as a signature, $P < 0.05$
- RNAi Screen:
 - Significant in combined ABC vs. GCB: $P < 0.05$
 - Depleted in either only HBL1 (ABC) or only all three GCB lines
- Significant in combined CNV and shRNA: $P < 0.05$

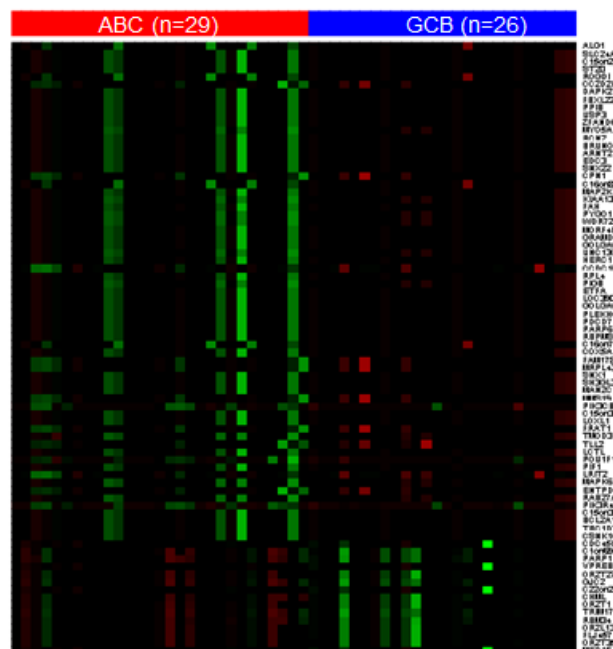


Figure 10-15 Heatmap of top candidates from CNV profiles of ABC vs. GCB samples. Red means amplification while green for depletion.

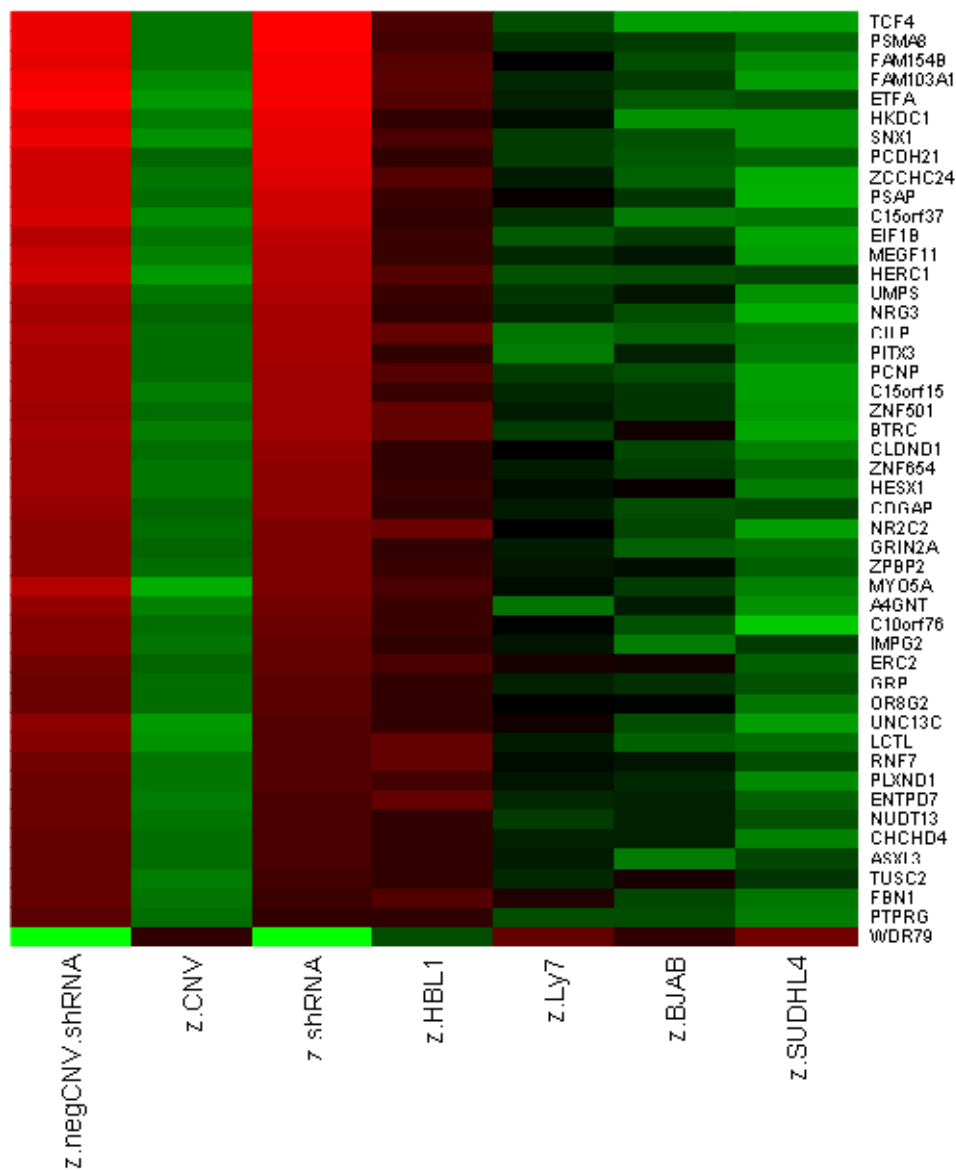


Figure 10-16 Heatmap of top candidates crossing CNV results with RNAi screening results for ABC vs. GCB-DLBCL. “z.CNV” indicates amplification (red) or depletion (green) in ABC vs. GCB samples. “z.negCNV.shrRNA” is the combined z score of CNV result with shRNA differential representation score. The negative has the same annotation as in Figure 10-12.

10.3.10 Crossing RNAi screening with NetBID2-predicted drivers and amplified genes from CNV data

We further filtered the candidates by integrating RNAi screens with NetBID2-predicted driver and with CNV results, and finally selected only eight driver-type candidates (Figure 10-17) that are lethal to ABC or GCB type DLBCL and show evidence of amplification in corresponding subtype with the following criteria:

- GEPs: as a MR, $P < 0.05$
- CNVs: as a signature gene, $P < 0.05$
- RNAi screens:
 - Significant in combined ABC vs. GCB: $P < 0.05$
 - Depleted in either only HBL1 (ABC) or only at least one GCB lines
- Significant in combined GEP, CNV and shRNA: $P < 0.05$

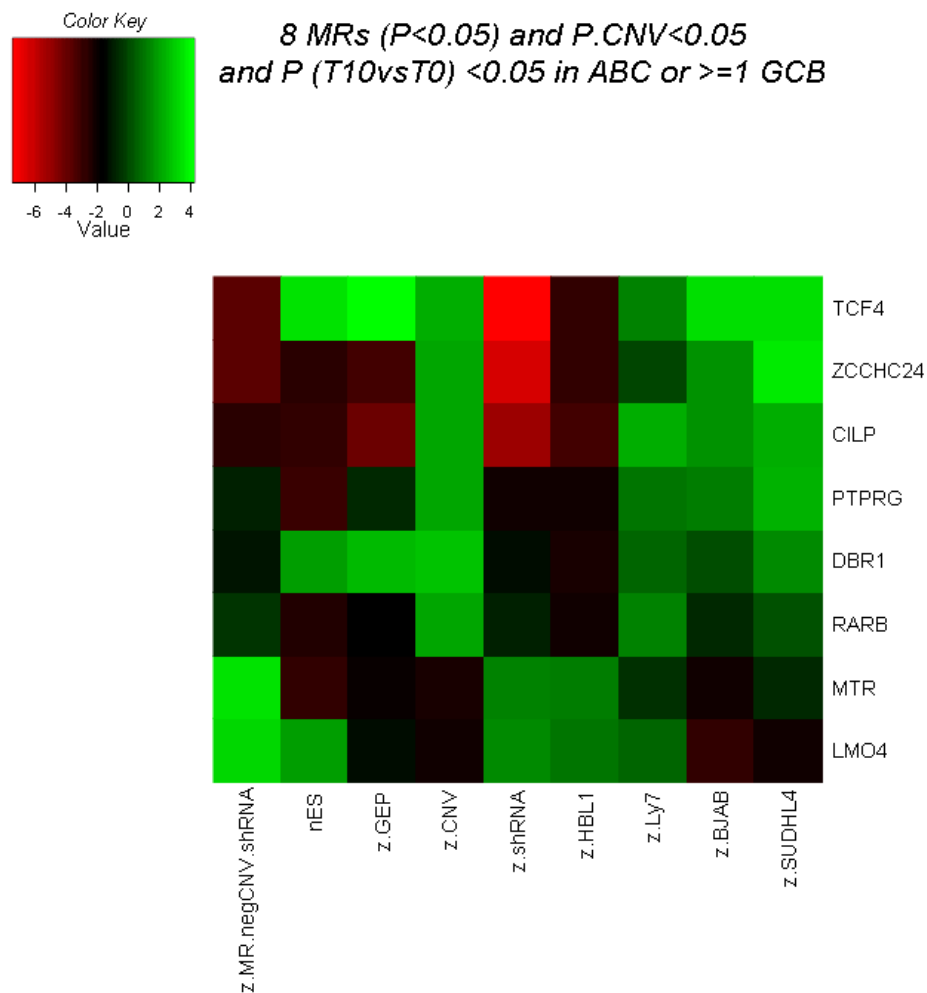


Figure 10-17 Selected eight candidates crossing RNAi screens with NetBID2-predicted drivers and CNV results for ABC or GCB-DLBCL.

“z.MR.negCNV.shRNA” is the combined z score of driver evidence, negative CNV score and shRNA screening. “nES” is the driver score. “z.GEP” is for differential expression.

10.4 On-going and Future Work

In the results section, I only showed the prediction results and selected promising candidate therapeutic targets specific to ABC or GCB type DLBCL. All selected targets are being validated by our collaborators.

Another interesting genetic feature we were interested in is BCL2 translocation. We also talked a little about this in this chapter. Literature suggested different clinical outcomes of BCL2-translocated and non-translocated patients and needs to be treated differently. We have two BCL2-rearranged cell lines (BJAB and SUDHL4) and clustering analysis based on functional profiles showed significant difference of these two BCL2-rearranged lines with the other two non-rearranged cell lines. We have performed similar analysis with ABC vs. GCB subtype and identified potential therapeutic targets specific for BCL2-rearrangement, which are being under validation as well.

Chapter 11 Integrating Functional Genomics with Systems Biology on Therapeutic Target Discovery for Subtype or Genetic-feature Specific Breast Cancer

11.1 Introduction

The goal of this chapter is to apply the integrative framework of crossing RNAi screening with systems biology to identification of therapeutic targets that are specific to subtype-based breast cancer and genetically-defined breast cancer.

For subtype-based breast cancer, we focused on luminal and basal types which are two well-classified forms of breast tumors based on transcriptional profiles. We did genome-wide shRNA screens using NGS technology on 16 breast cancer cell lines in which four are luminal subtype and eight are basal subtype. From RNAi screens, we identified candidates that are lethal to each cell line and lethal to only luminal or basal subtype cell lines. In parallel, we applied our NetBID2 algorithm to TCGA gene expression data of breast cancer and identified drivers for basal vs. luminal subtype. Then integrating RNAi results with computationally-predicted drivers suggested potential therapeutic target candidates for luminal or basal type of breast cancer. The graphical explanation of the integration framework is shown in Figure 11-1.

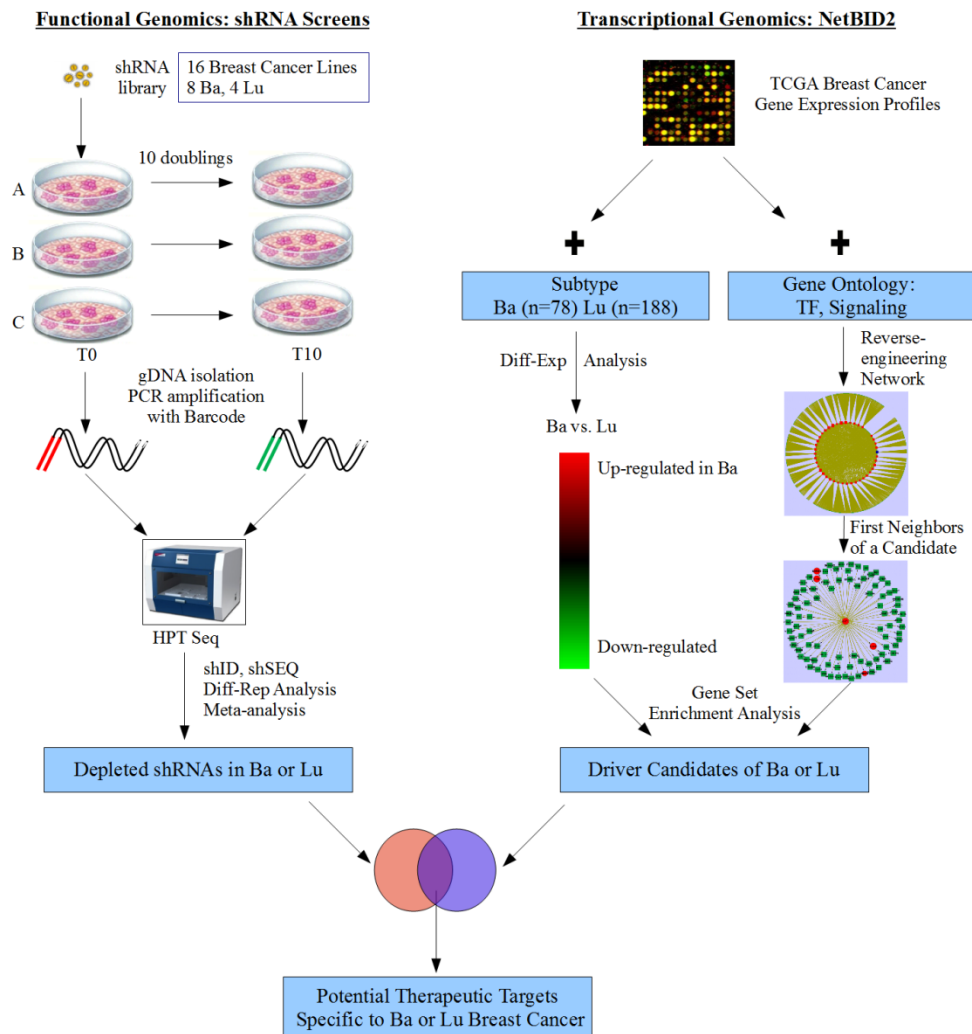


Figure 11-1 The integrative framework of RNAi Screening with NetBID2 to identify potential therapeutic targets specific to basal or luminal type of breast cancer. On the left, we did shRNA screening on 16 breast cancer lines including 8 basal and 4 luminal by NGS. On the right, we applied NetBID2 to TCGA breast cancer gene expression profiles to identify drivers of basal vs. luminal subtype.

Breast cancer can also be classified by various genetic features such as amplified oncogenes or depleted tumor suppressors that are causally associated

with tumorigenesis or progression. In this study, we applied the same idea in Chapter 9 by employing an isogenetic model and mimicking genetic alternations in a normal breast cell line, MCF10A. We focused on oncogenes of PI3K, MYC, ERBB2, CCND1, and tumor suppressors of E1A (RB1), PTEN. We did shRNA screens on those genetically-engineered models. Comparing shRNA abundance in mutated cells with the control of wild type, we identified candidates that are functionally related to the corresponding oncogene or tumor suppressor, forming a synthetic lethal pair.

11.2 Drivers and Therapeutic Targets for Basal or Luminal type Breast Cancer

11.2.1 Summary for shSeq data of 16 breast cell lines

We did shRNA screens using high-throughput sequencing or NGS on 16 breast cell lines, which are characterized in Table 11-1. Among 16 breast cell lines, there are four basal A, four basal B, two normal but basal B like based on transcription profiles, and two inflammatory breast cancer (IBC) lines. The deconvolution of shSeq data for 16 cell lines is summarized in Table 11-2. In generally, over 75% of total reads can be identified in terms of sample conditions and shRNA hairpins, and most samples have enough identified reads (over 5 million in total or 80 per hairpin on average). The default Illumina filtering procedure might be useless (last two columns in Table 11-2) because the sequence we care about is the first 6 nucleotides for barcodes of samples and the 22 nucleotides for hairpin sequence.

The shSeq data for all 16 cell lines are in good quality, except T47D highlighted as an outlier in Figure 11-2.

Cell Line	subType	ER	PR	HER2	TP53	source	isTumor	tumorType	Age	Ethnicity
BT20	BaA	-	-	-	++.WT	PB	Yes	IDC	74	W
HCC1143	BaA	-	-	-	++.M	PB	Yes	DC	52	W
HCC1937	BaA	-	-	-	[-]	PB	Yes	DC	24	W
MDAMB468	BaA	-	-	-	[+]	PE	Yes	AC/Met AC	51	B
HS578T	BaB	-	-	-	+.M	PB	Yes	IDC/C.Sar	74	W
SUM159PT	BaB	[-]	[-]	-	[-]	PB	Yes	AnCar		
MCF10A	BaB	-	-	-	+/-WT	PB/RM	No	F	36	W
MCF12A	BaB	-	[-]	-	+	PB	No	F	60	W
MDAMB231	BaB	-	-	-	++.M	PE	Yes	AC/Met AC	51	W
MDAMB436	BaB	-	-	-	[-]	PE	Yes	IDC/AC	43	W
MCF7	Lu	+	+	-	+/-WT	PE	Yes	IDC/Met AC	69	W
MDAMB361	Lu	+	[-]/+	+	-.WT	PB/BR	Yes	AC/Met AC	40	W
SKBR3	Lu	-	-	+	+	PE	Yes	AC	43	W
T47D	Lu	+	+	-	++.M	PE	Yes	IDC	54	
SUM149PT	BaB	-	-	-	[+]	PB/PE	Yes	Inf DC/Inf		
SUM190PT	BaA/L	-	-	+	[+/-]	PB	Yes	Inf		

Table 11-1 Characteristics of 16 breast cell lines with shRNA screening data.

BaA=Basal A; BaB=Bsal B; Lu=luminal, ER/PR/HER2/TP53 status: ER/PR positivity, HER2 overexpression, and TP53 protein levels and mutational status (obtained from the Sanger web site; M=mutant protein; WT=wild-type protein) are indicated. Square brackets indicate that levels are inferred from mRNA levels alone where protein data is not available. A/B: A is from Neve, et al, Cancer Cell, 2006; B is from Kao et al, Plos One, 2009. AC=adenocarcinoma; AnCa=anaplastic carcinoma; C Sar = carcinoma sarcoma; DC=ductal carcinoma; F=fibrocystic disease; IDC=invasive ductal carcinoma; Inf=inflammatory; Met AC = metastatic adenocarcinoma PB=primary breast; RM= reduction mammoplasty; PE=pleural effusion; BR=Brain W=White; B=Black.

Run 12	T10.A	T10.B	T10.C	T0.A	T0.B	T0.C	total raw reads	total identified reads	identification rate	T10/T0 (total identified reads)	reads > Illumina Filter	percentage of high-quality reads
MDAMB468	1,754,213	2,949,676	17,256,943	1,802,948	455,891	1,478,439	32,937,206	25,698,110	78.02%	5.88	32,357,018	98.24%
	30	50	295	31	8	25						
HS57BT	14,261,572	11,276,323	3,402,531	24,198,563	18,979,220	8,794,471	103,508,686	80,912,680	78.17%	0.56	99,391,525	96.02%
	244	193	58	414	324	150						
BT20	6,973,459	9,367,631	12,387,621	12,707,309	10,457,429	11,534,564	77,758,702	63,428,013	81.57%	0.83	75,940,895	97.66%
	119	160	212	217	179	197						
T47D	7,742,742	16,499,446	15,292,978	8,851,725	14,139,016	11,371,252	96,503,420	73,897,159	76.57%	1.15	92,867,820	96.23%
	132	282	261	151	242	194						
SKBR3	11,777,818	14,438,241	13,483,218	13,353,331	13,329,153	11,386,889	98,432,341	77,768,650	79.01%	1.04	94,076,155	95.57%
	201	247	231	228	228	195						
MDAMB231	17,859,173	25,765,390	21,093,518	1,928,230	1,920,412	2,813,584	90,709,456	70,680,307	77.92%	10.85	87,526,651	96.49%
	305	440	361	33	21	48						
Run 13	T10.A	T10.B	T10.C	T0.A	T0.B	T0.C	total raw reads	total identified reads	identification rate	T10/T0 (total identified reads)	reads > Illumina Filter	percentage of high-quality reads
MDAMB436	25,482,266	26,261,321	24,346,592	31,264,391	34,332,466	28,975,782	255,000,397	170,762,818	66.97%	0.80	0	0.00%
	436	449	416	536	587	495						
MCF7	14,340,516	28,069,129	39,963,895	18,972,635	32,688,501	26,542,675	249,302,521	160,577,351	64.41%	1.05	4,419,501	1.77%
	245	480	683	324	559	454						
MDAMB361	51,984,610	1,898,961	53,870,807	748,200	10,501,461	17,771,791	217,593,362	136,575,830	62.77%	3.71	25,733,328	11.83%
	889	29	921	13	180	304						
MCF12A	25,031,463	31,708,390	28,861,748	20,056,529	27,702,982	33,319,063	260,187,826	166,680,175	64.06%	1.06	4,216,162	1.62%
	428	542	493	343	474	570						
HCC1143	26,195,474	17,935,349	24,925,799	17,261,596	30,610,993	25,217,165	256,128,123	142,146,376	55.50%	0.94	28,837,961	11.26%
	448	307	426	295	523	431						
MCF10A	738,108	100,453	179,420	29,274,603	84,386,971	22,131,987	198,888,638	136,831,544	68.80%	0.01	86,794,917	43.64%
	13	2	3	500	1,443	378						
	T10.A	T10.B	T10.C	T0.A	T0.B	T0.C	total raw reads	total identified reads	identification rate	T10/T0 (total identified reads)	reads > Illumina Filter	percentage of high-quality reads
HCC1937	36,801,671	26,468,050	30,038,449	17,036,962	18,918,400	21,842,185	246,575,078	151,105,717	61.28%	1.61	0	0.00%
	629	452	514	291	323	373						
SUM159PT	25,549,573	10,752,389	24,429,136	32,627,012	37,739,830	38,130,894	253,020,220	169,228,834	66.88%	0.56	0	0.00%
	437	184	418	558	645	652						
SUM149	5,146,443	8,587,854	2,390,297	7,287,859	4,867,167	7,963,681	49,401,202	36,243,301	73.37%	0.80	47,828,674	96.82%
	88	147	41	125	83	136						
SUM190	2,285,708	14,421,723	334,022	4,675,186	7,438,550	862,966	42,016,238	29,718,153	70.73%	1.34	39,472,053	93.95%
	39	247	6	80	127	10						
MDAMB468_reRun_MountSinal	4,652,228	7,283,637	65,762,618	7,232,370	1,718,249	5,010,299	118,000,000	91,659,401	77.68%	5.57		
80	125	1,124	124	29	86							

Table 11-2 Summary for deconvolution of shSeq data for 16 breast cell lines. Numbers in red are the samples that have < 5M identified reads. The last two columns are using the default Illumina filtering criteria.

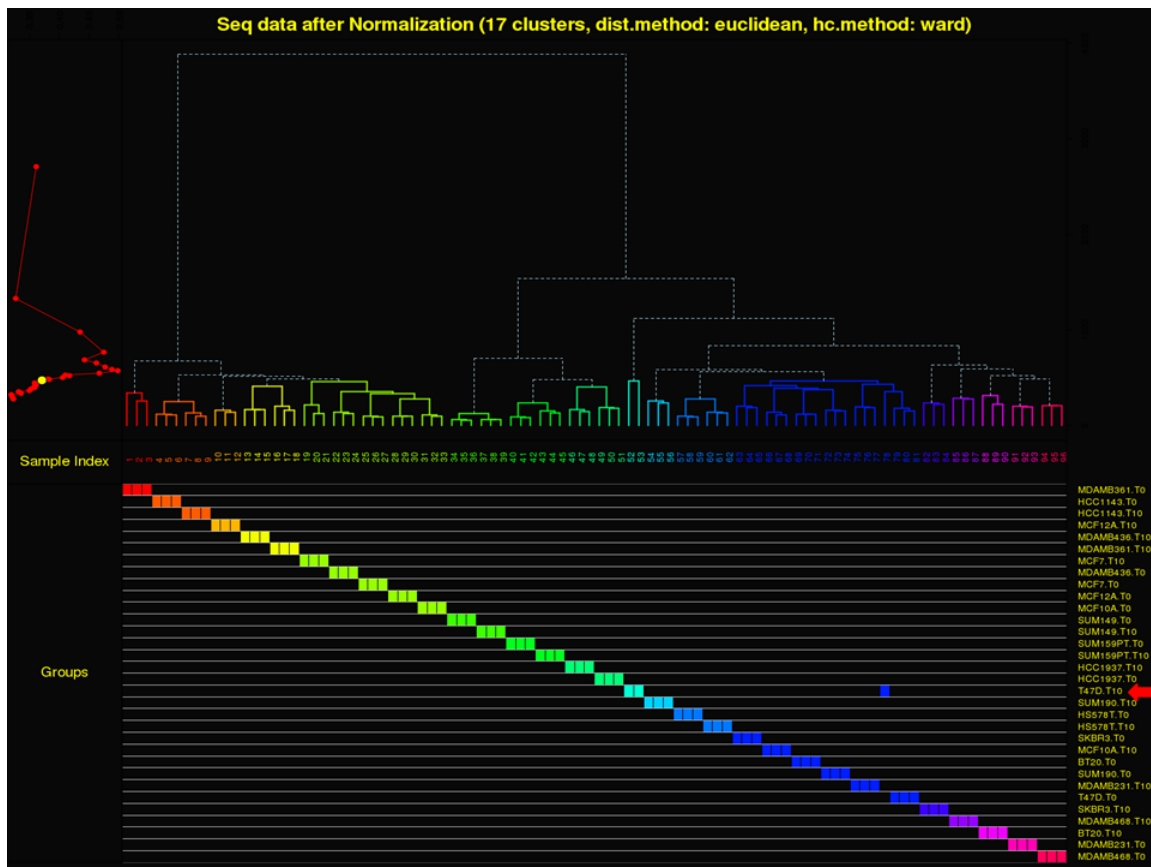


Figure 11-2 Clustering of shSeq samples by normalized data. Each row is for one sample condition. Three boxes on each row are the biological triplicates. One replicate T47D.T0 is highlighted as an outlier.

11.2.2 Results of differential representation analysis

For each cell line, we performed differential representation analysis comparing T10 with T0 data at both individual shRNA level by shADER and gene level by

BHM algorithm. Summarized statistics including p value and z score are plotted in Figure 11-3 and Figure 11-4.

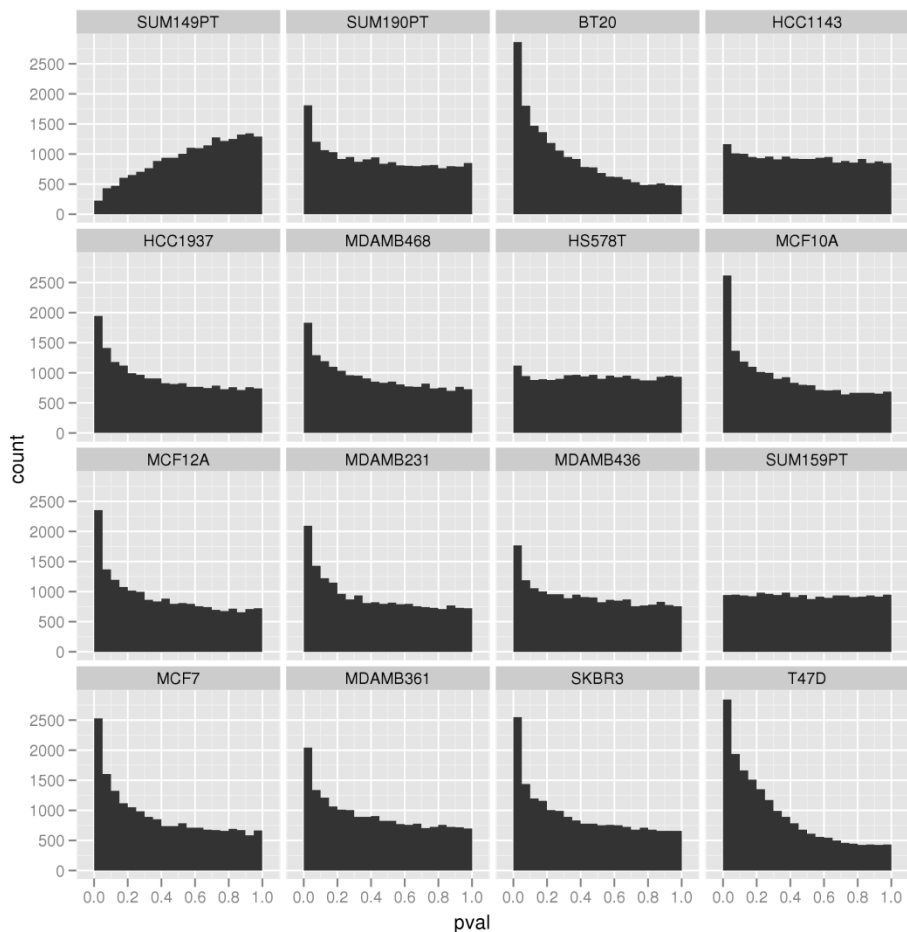


Figure 11-3 Histogram of p value at gene level activity analysis for shRNA screens of 16 breast cell lines. The bin width is 0.05.

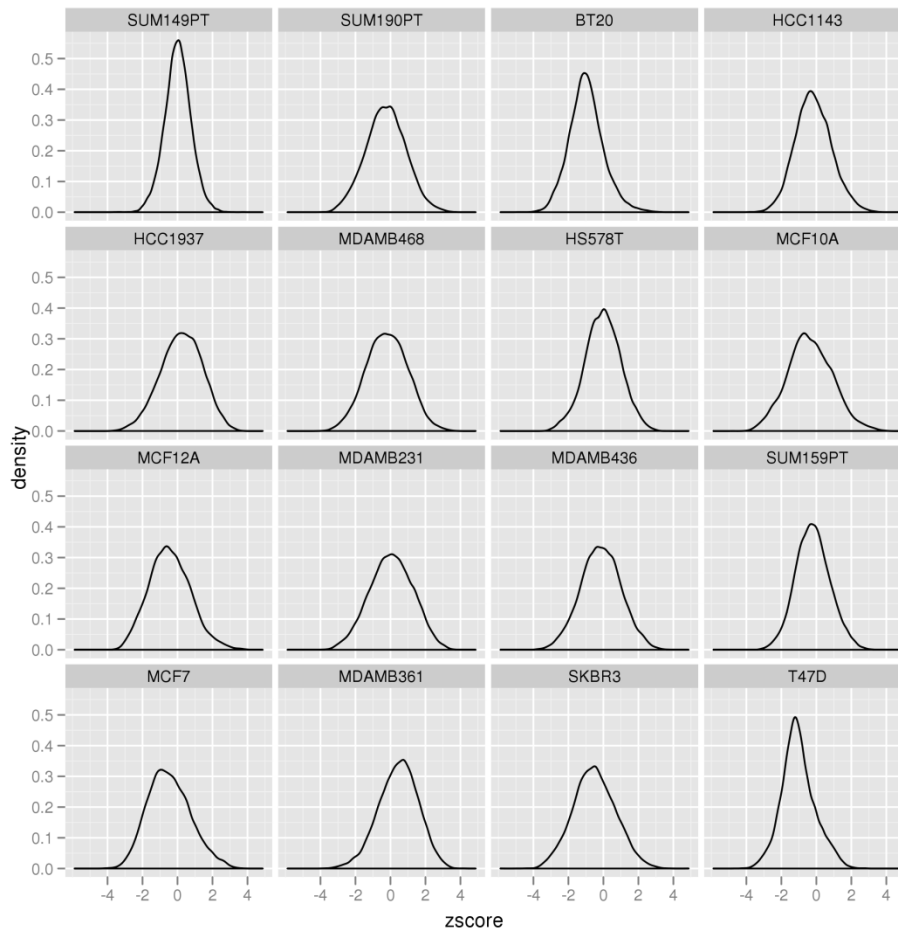


Figure 11-4 Density plot Histogram of z score at gene level activity analysis for shRNA screens of 16 breast cell lines.

11.2.3 Unsupervised clustering of functional profiles separate subtypes well

With functional profiles indicating silencing effects in each line, we asked whether the clustering of functional profiles showed consistence with the clustering of gene expression profiles that defined different subtypes of breast cancer. So we performed unsupervised clustering analysis of 14 breast tumor cell lines using their z scores at gene level. As shown in Figure 11-5, these 14 tumor lines can

be separated into five clusters: the red one on the left is for HER2+ subtype; the yellow one is for luminal; the green one is IBC, the most aggressive form of breast cancer, showing very different performance with the others; the blue one is for basal B, and the purple one is for basal A. If we only considered two clusters, one big cluster on the left is more like luminal subtype including HER2+, and the other on the right is more like basal subtype. This suggested that functional profiles from shRNA screening are able to separate different subtypes well.

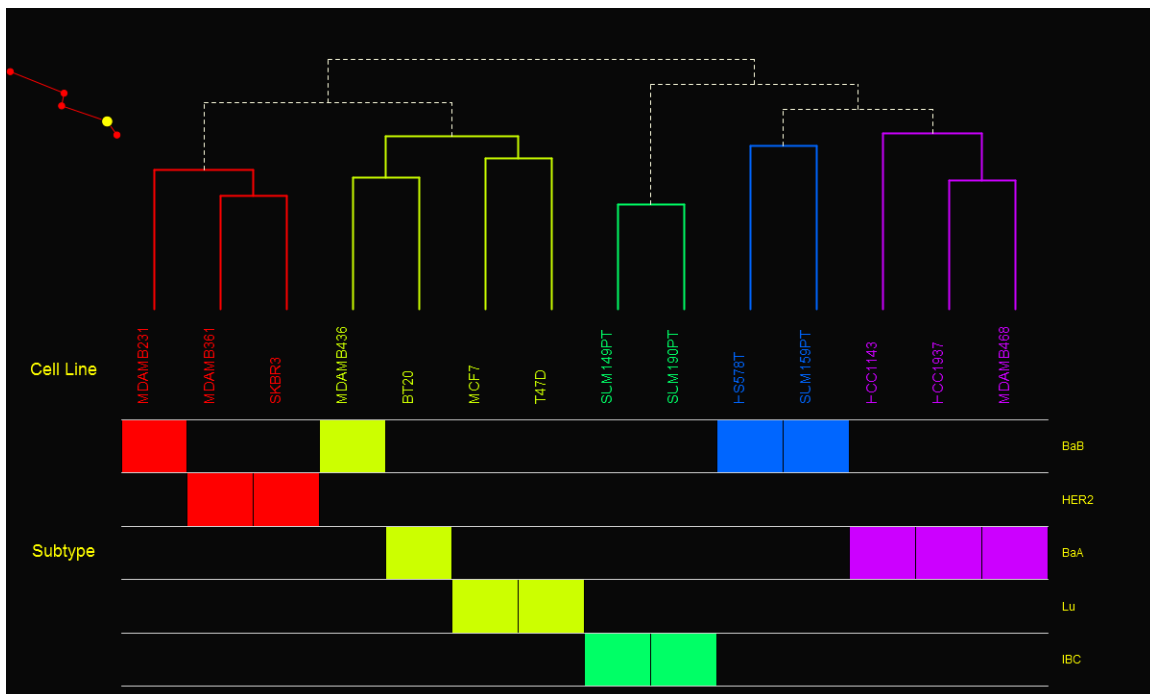


Figure 11-5 Clustering of 14 breast tumor cell lines by functional profiles. Functional profile is using differential representation score at gene level.

11.2.4 Top candidates that are common in all breast tumor lines

Before discussing about essential genes specific to any subtype of breast cancer, we first asked whether there is any common gene that is depleted or enriched in all breast tumor lines. Those candidates might be potential targets for the majority of breast tumors. We used Stouffer's method to combine shRNA screening data of 12 tumor lines with good data and 912 genes were selected with a combined p value less than 0.01 (Figure 11-6).

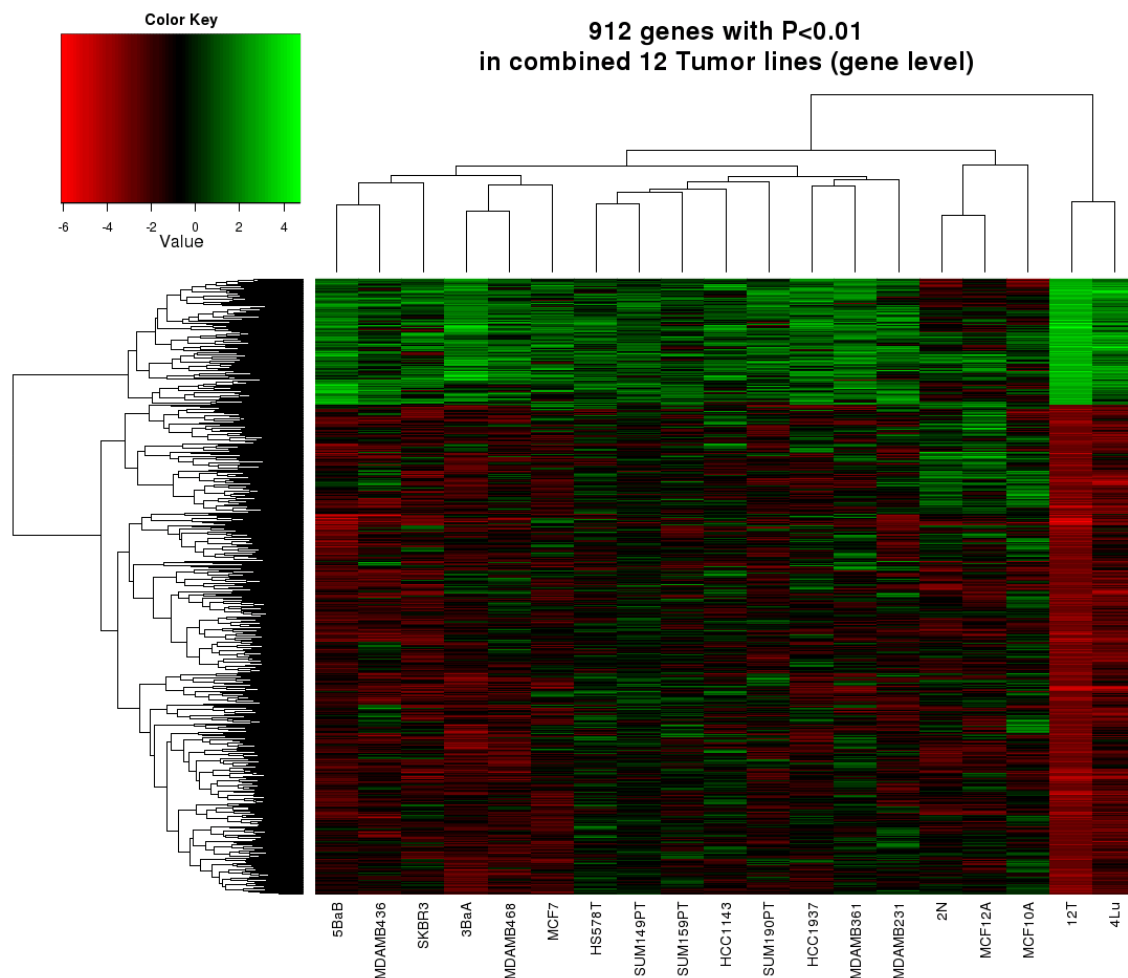


Figure 11-6 Heatmap of functional profiles for top depleted or enriched genes ($P < 0.01$) in the majority of 12 breast tumor cell lines. The genes are selected by

a combined z score of all 12 tumor lines using Stouffer's method. "12T" is the z score of combining 12 tumor lines. Similarly, "4Lu" is combining 4 luminal lines, "3BaA" for basal A lines, "5BaB" for 5 basal B lines including SUM149PT, and "2N" for two normal lines.

11.2.5 Sensitivity analysis: difference between depleted essential genes and enriched tumor suppressor genes

With shRNA profiles of a panel of 16 breast cancer lines, we have the ability to do so-called "sensitivity analysis" (Figure 11-7) to address the question of how sensitive each cell line is to respond RNAi perturbation or how easily to kill a cell line by RNAi. Details were discussed in 2.9.2.

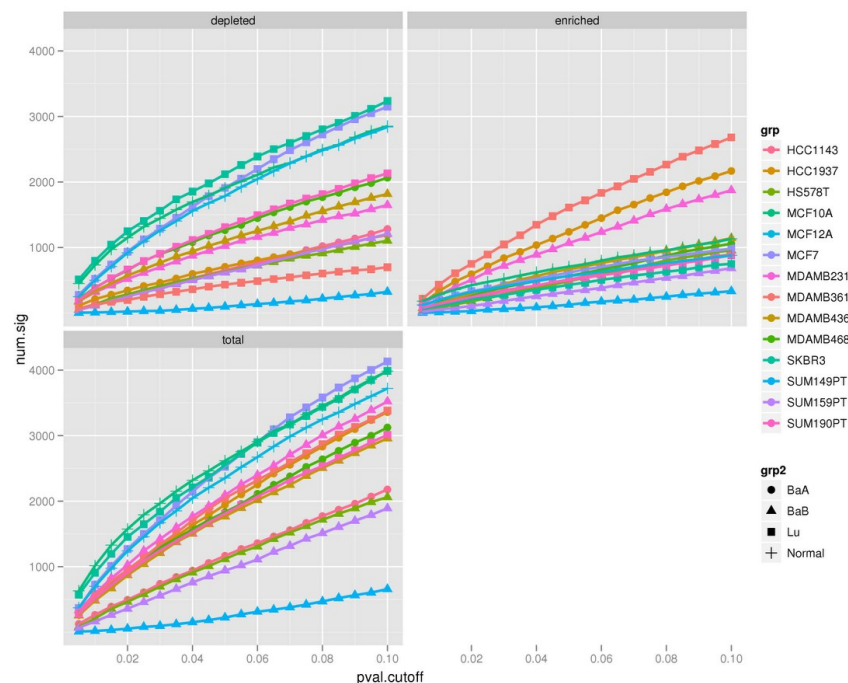


Figure 11-7 Sensitivity analysis of shSeq data for 16 breast lines.

11.2.6 Supervised clustering analysis of functional profiles identifies subtype specific lethal genes

We performed supervised clustering analysis on basal vs. luminal breast cancer cell lines using their functional profiles. Interestingly, top differentially-behaved genes in basal vs. luminal types also classify other subtypes well such as HER2+, IBC, basal A, and basal B (Figure 11-8).

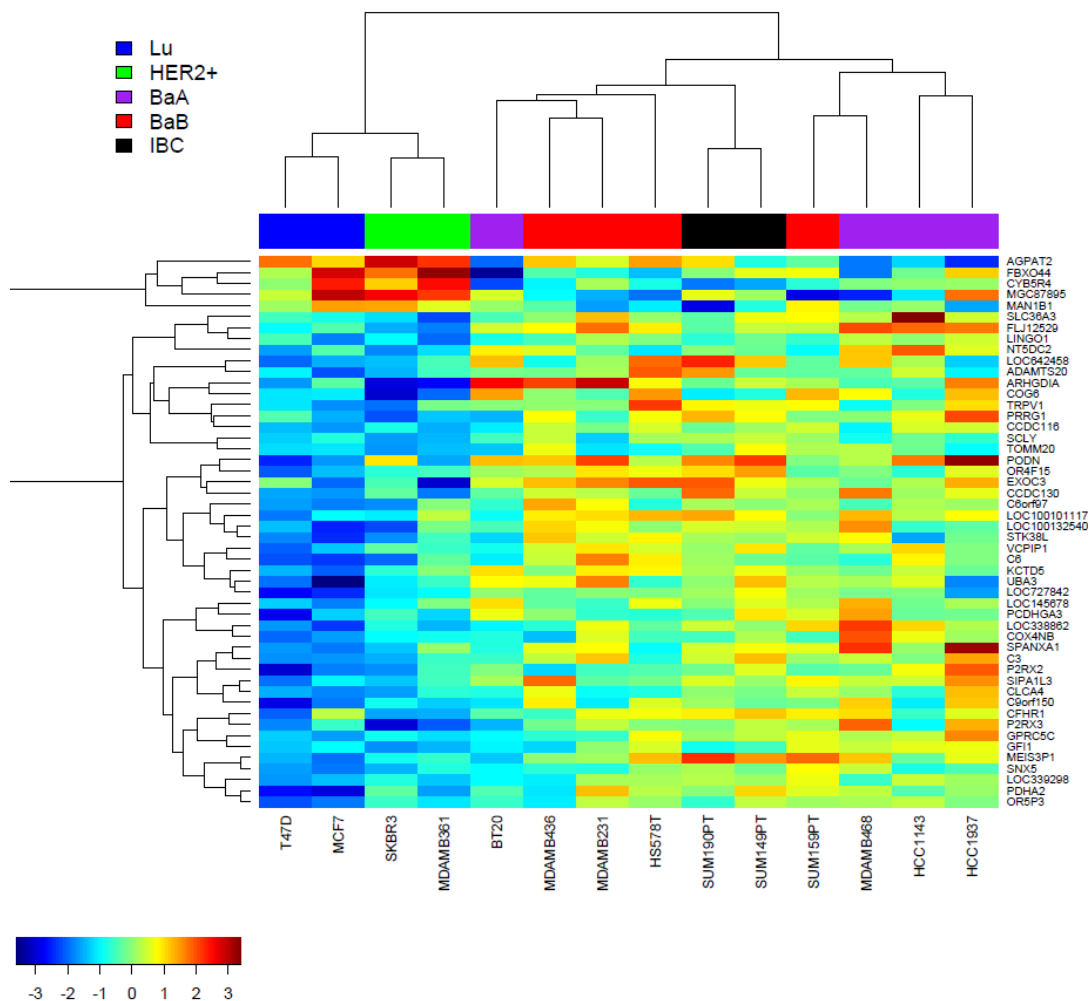


Figure 11-8 Heatmap of top shRNA screening identified candidates using functional profile. Functional profile is defined by differential representation score at gene level. Blue stands for depletion while red for enrichment.

11.2.7 Functional enrichment of lethal genes specific to basal or luminal subtype

We performed the functional enrichment analysis to identify potential pathways that are essential to either basal or luminal type of breast cancer. A combined z score was generated for each gene for all luminal lines or basal lines. BSEA method was used to estimate the enrichment of known pathways. We identified that ESRRA up-regulated targets are lethal to luminal type, ESRRA down-regulated targets, COPI mediated transport pathway, ribosome pathway are lethal to basal type (Figure 11-9). More interestingly, E2F family proteins with their targets showed lethal effects to basal type of breast cancer (Figure 11-10).

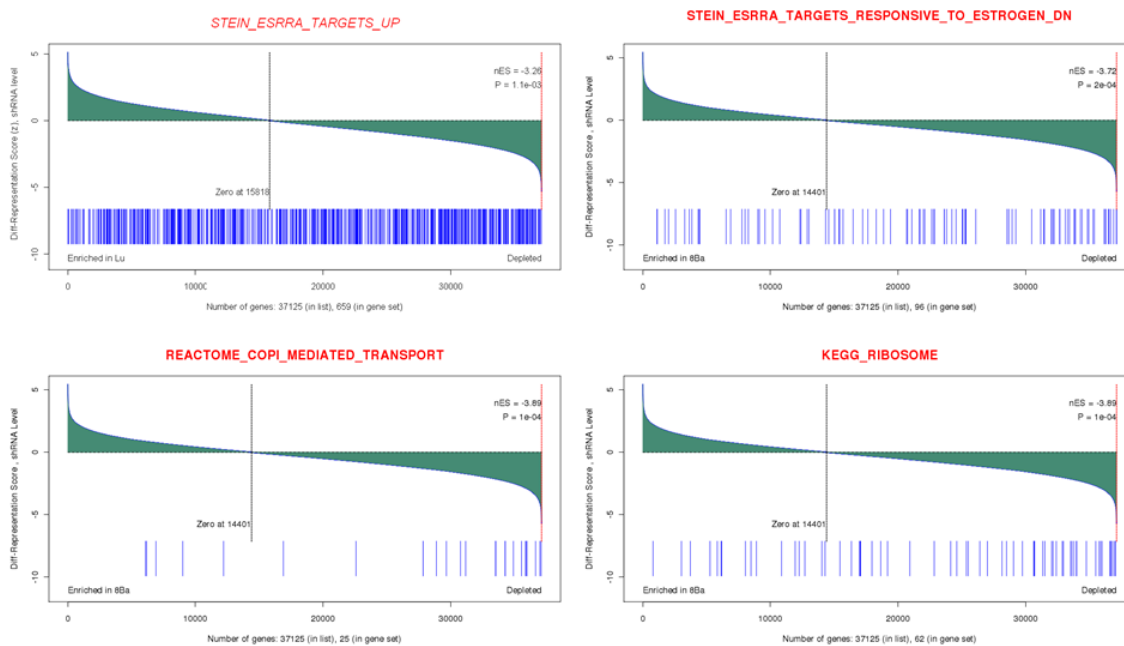


Figure 11-9 Top enriched pathways by depleted genes in luminal or basal subtypes.

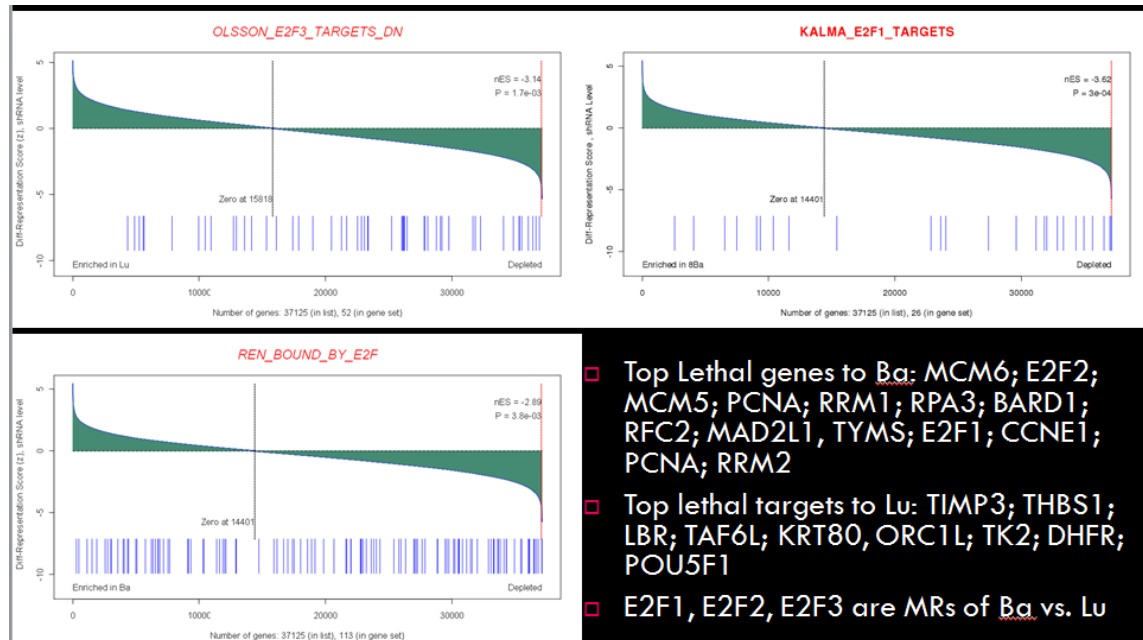


Figure 11-10 Top enriched pathways by depleted genes in luminal or basal subtypes. Top lethal genes in the pathways are listed.

11.2.8 Crossing with NetBID2-predicted drivers of basal vs. luminal subtype

Out of 359 TCGA primary breast cancer patients with gene expression profiles, 78 are classified as basal type and 188 are luminal. With these data, we applied the NetBID2 algorithm to predict drivers of basal or luminal breast cancer, and then crossed with shRNA screening identified basal or luminal specific lethal genes. The integrative analysis with stringent threshold yielded only three candidates, in which two signaling proteins, GNA14 and ATP6V1G2, were predicted as drivers and showed lethal effects in luminal breast cancers, while E2F2 was a therapeutic target candidate for basal type breast tumors.

geneSymbol	funcType	n.shRNAs	nES	z.DE	z.BaVSLu.geneLevel
GNA14	Sig	3	-2.23	-3.28	4.29
ATP6V1G2	Sig	3	-2.75	-6	2.26
E2F2	TF	1	2.39	7.51	-2.34

Table 11-3 Overlapped candidates of NetBID2-predicted drivers and RNAi Screening identified lethal candidates for basal or luminal subtype. “nES”, normalized enrichment score, indicates the driver prediction strength. “z.DE” indicates the differential expression of the gene itself. “z.BaVSLu.geneLevel” is the z score of comparing shRNA screening profiles of basal with luminal cell lines. GNA14, ATP6V1G2 are potential therapeutic targets for luminal subtype, while E2F2 is for basal subtype.

11.2.9 Consistency with drug sensitivity data

Out of the 16 breast cell lines we did shRNA screening, we also had the sensitivity data for 12 lines treated by 74 small molecules or compounds [247]. Known targets for those 74 drugs were also collected from drugbank database [248]. We used the drug targets to connect shRNA screening data with drug sensitivity data and checked the consistence of drug effects with shRNA silencing effects on cell viability.

Here we only focused on a specificity study of basal and luminal, the two major subtypes of breast cancer. The basic idea was to perform specificity analysis of comparing basal vs. luminal cell lines using drug sensitivity data only, shRNA screening data of drug targets only and a combined meta-analysis of both. The results (Figure 11-11, Figure 11-12) suggested that targets of over half of top

subtype-specific drugs also show up in the specificity analysis of shRNA screen data and half of which show the same direction.

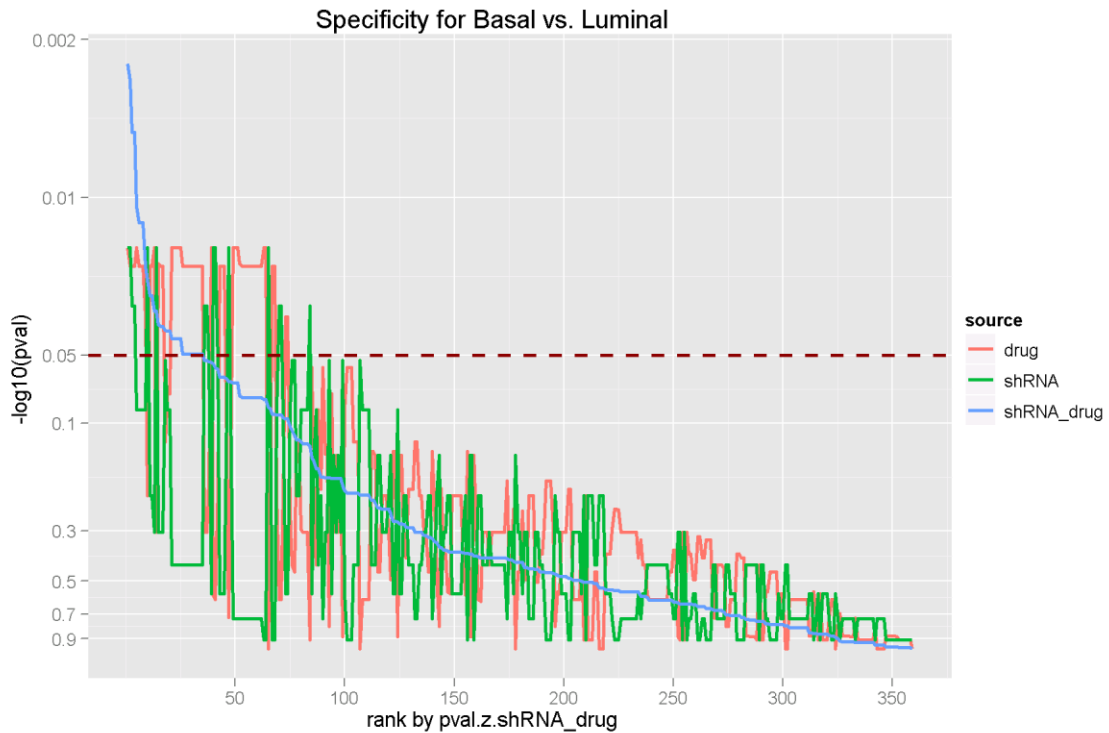


Figure 11-11 Specificity analysis by linear model using drug sensitivity data and shRNA screening data, also combination. Combination using known targets of drugs matched with shRNA screening data. Combined analysis is done using Stouffer's method. Purple dashed line indicates the significance level at 0.05.

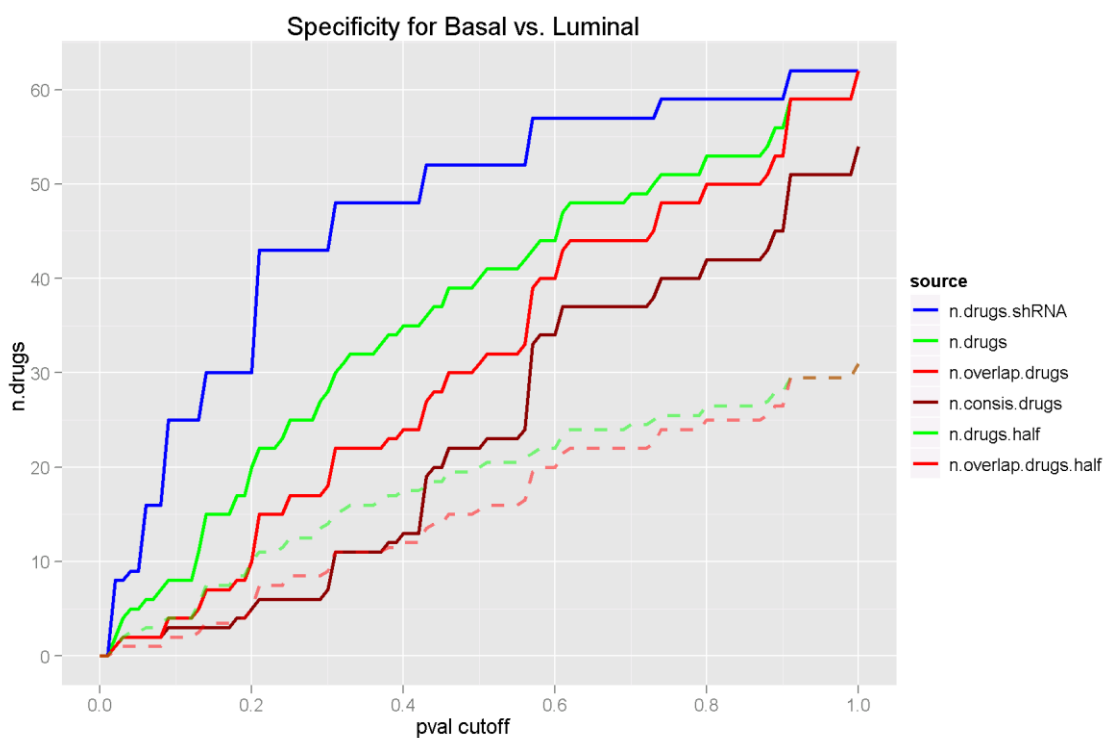


Figure 11-12 Number of significant drugs that are specific to basal or luminal using different data source at various p value cutoffs. “n.drugs” is using drug sensitivity data only. “n.drugs.shRNA” is based on shRNA data of drugs’ targets. “n.overlap.drugs” is the number of overlapped drugs between “n.drugs” and “n.drugs.shRNA”. “n.consist.drugs” is the number of overlapped drugs showing consistent direction in both shRNA screening results and drug sensitivity results. Green dashed line is half of the green line, and so is the red dashed line.

We used the functional profiles of targets from top drugs that are specific to basal or luminal type to classify basal and luminal breast cancer cell lines, and they separated the two subtypes well (Figure 11-13). This again confirmed a large consistence of drug data and shRNA screening data.

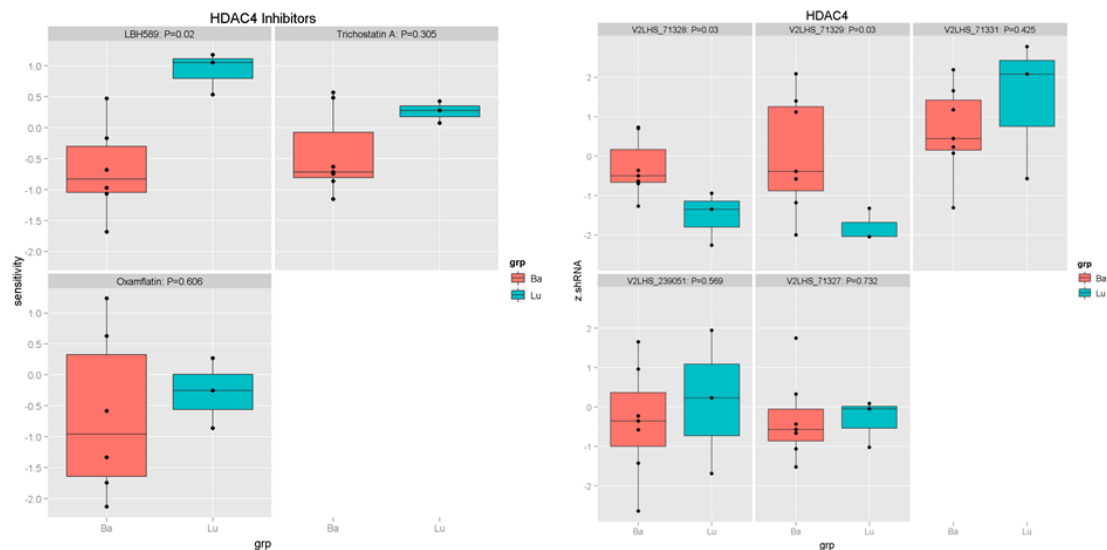


Figure 11-14 HDAC4 is specific to luminal type breast cancers shown by sensitivity data of HDAC4 inhibitors (left) and shRNA screening data of HDAC4 hairpins (right).

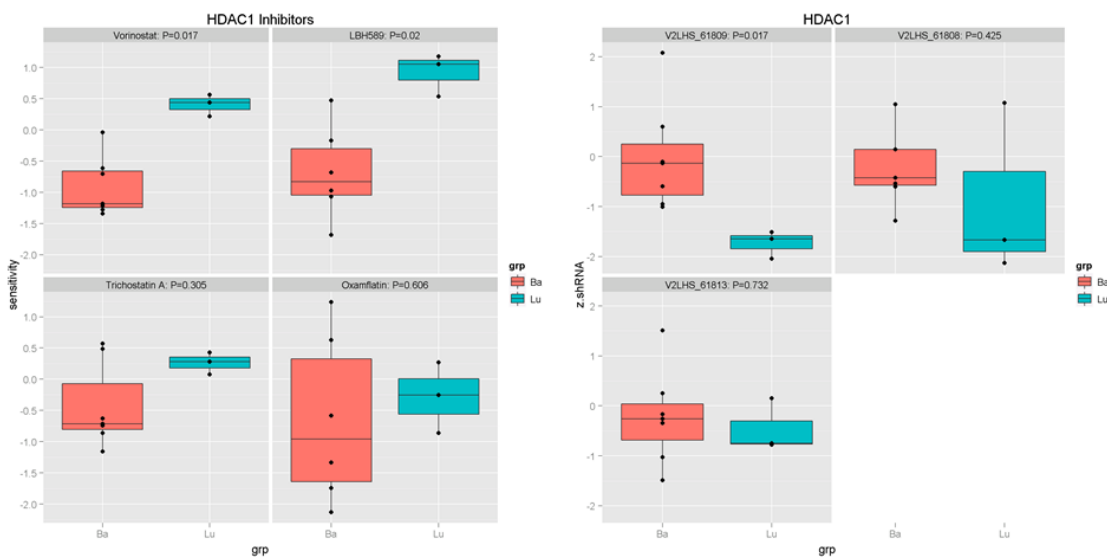


Figure 11-15 HDAC1 is specific to luminal type breast cancers shown by sensitivity data of HDAC1 inhibitors (left) and shRNA screening data of HDAC1 hairpins (right).

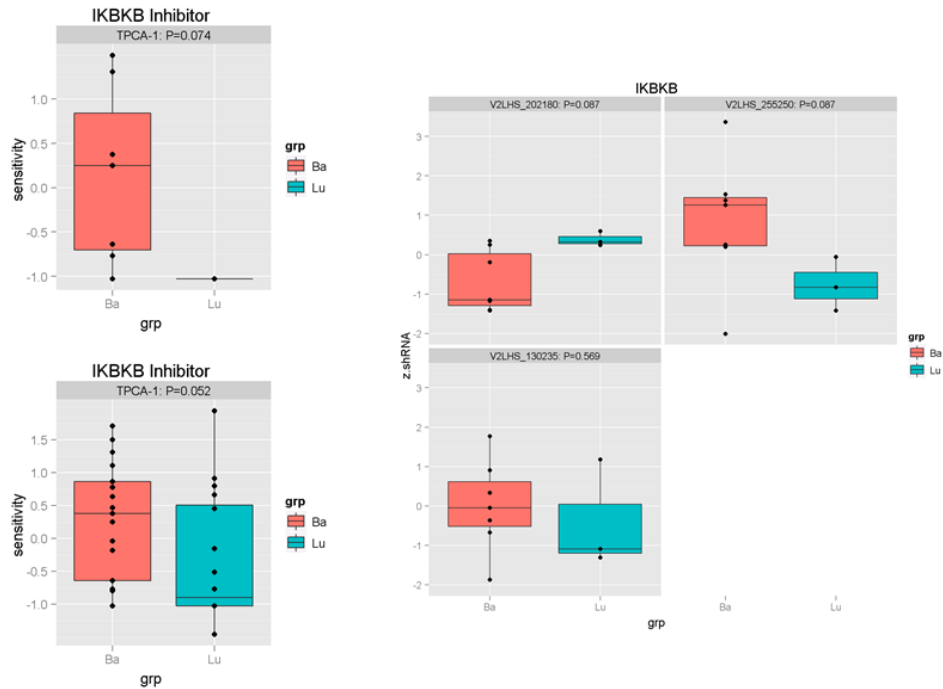


Figure 11-16 IKBKB is specific to basal type breast cancers shown by sensitivity data of IKBKB inhibitor (left) and shRNA screening data of IKBKB hairpins (right).

However, there might be some issues with the drug sensitivity data we have to be careful with. For example, drugs targeting the same targets might show different sensitivity. The sample size of cell lines in each subtype might make the analysis less powerful. More data might produce more accurate results.

11.3 RNAi Screens to Search for Synthetic Lethal Partners of Genetic-features in Breast Cancer

Breast cancer are also characterized by a number of genetic features such as PI3K, MYC, CCND1, ERBB2 oncogenes and PTEN, RB1 tumor suppressors. We were also interested in identifying synthetic lethal partners with those genetic

features in breast cancer. We used shRNA screening to search for lethal genes that are specific to genetic-feature defined breast cancer. To avoid heterogeneity, we used an isogenetic model – MCF10A cell line – and genetically introduced engineered amplification of PI3K, ERBB2, MYC, CCND1 and depletion of PTEN and E1A individually. Then we did shRNA screening using NGS for each of those engineered models.

11.3.1 Summary for the shSeq data of genetically-engineered models

We performed shRNA screening using NGS technology for six genetically-engineered MCF10A models including PI3K, ERBB2, MYC, CCND1, PTEN and E1A. The cell culture was under 2D system and the infected cells were harvested for 10 generations. Most of the shSeq data had enough signals expect MYC (Table 11-4). And QC report of normalized data showed all the shSeq data were in good shape (Figure 11-17, Figure 11-18).

Sample	F1	F2	F3	F4	F5	F6	total raw reads	total identified reads	identification rate
WT.T10_TO	758,108 13	100,455 2	179,420 3	29,274,603 500	84,386,971 1,443	22,131,987 378	198,888,638	136,831,544	68.80%
MYC.T10_ERBB2.T10	4,961,923 85	1,234,337 21	3,000,547 51	43,842,247 750	36,137,269 618	54,344,343 929	183,818,501	143,520,666	78.08%
CCND1.T10_ERBB2.InVi vo.T10	41,696,724 713	53,308,248 911	49,723,910 850	3,003,412 51	2,250,131 38	10,346,820 177	222,698,642	160,331,245	71.99%
WT.3D.T10_ERBB2.3D.T 10	28,014,416 479	29,067,457 497	25,848,077 442	7,755,239 133	6,802,256 116	6,921,632 118	138,825,905	104,409,077	75.21%
WT.T10_PTEN.T10	12,681,031 217	11,505,499 197	6,133,125 105	4,512,187 77	6,282,855 107	6,817,415 117	62,144,097	47,932,112	77.13%
E1A.T10_PI3K.T10	26,780,841 458	4,544,639 78	1,424,443 24	20,613,030 352	43,957,770 752	17,875,969 306	150,166,672	115,196,692	76.71%

Table 11-4 Summary of deconvolution of shSeq data for six genetic-engineered models. Numbers in red are cases with < 5M total identified reads.

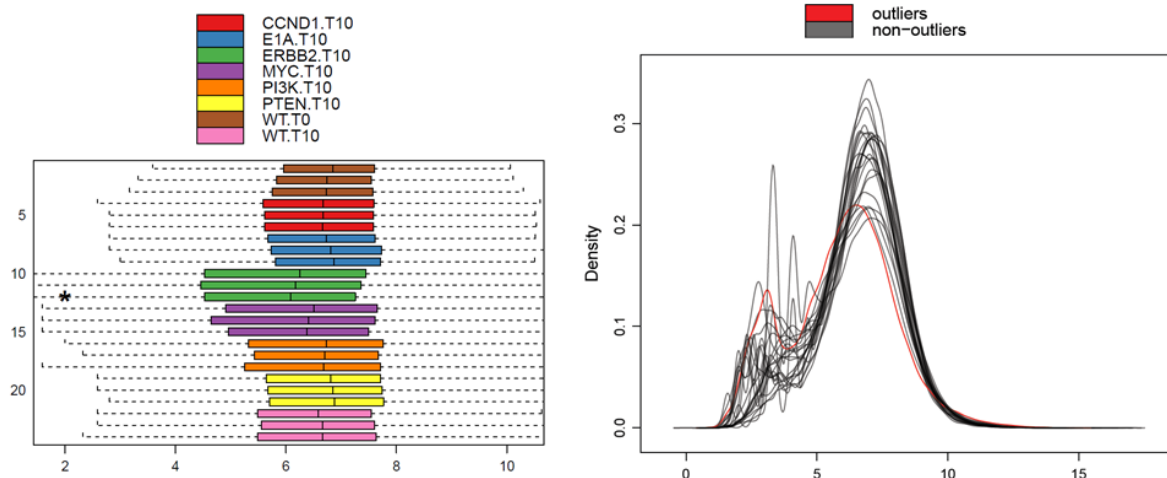


Figure 11-17 Distribution of hairpin count in samples of six genetic-engineered models.

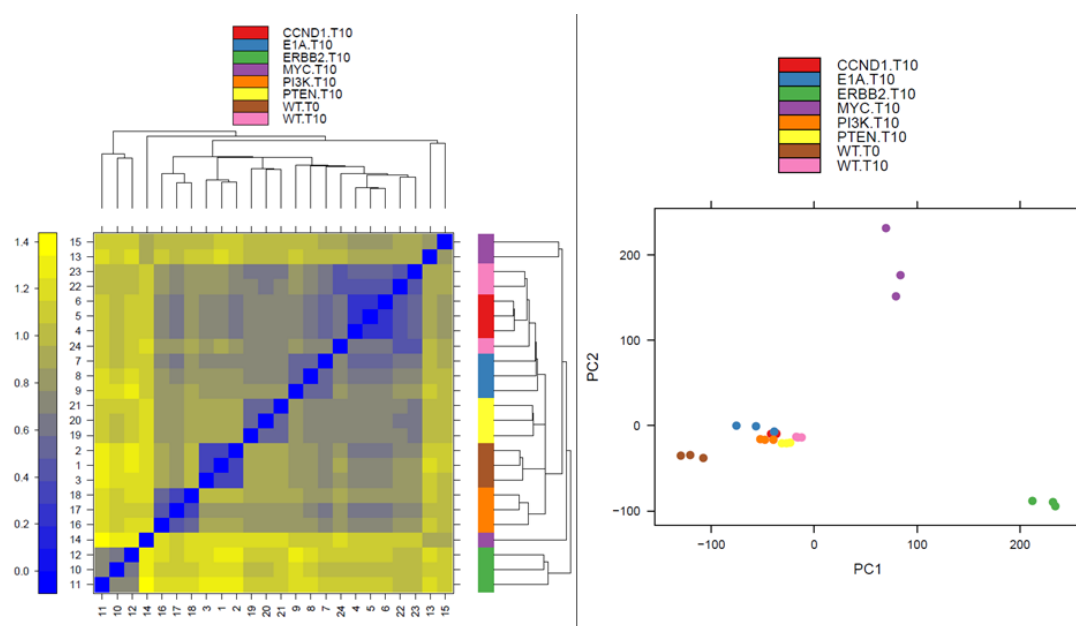


Figure 11-18 Heatmap of sample distances and PCA plot of sample of six genetic-engineered models.

11.3.2 Overall gene level activity for each genetic-feature defined breast cancer

As usual, we performed differential representation analysis at both individual shRNA level and gene level for each of the six shSeq datasets. Summary statistics including p value (Figure 11-19) and z score (Figure 11-20) showed reasonable results as the uniform distribution of non-significant p values. With a p value threshold of 0.05, numbers of significant candidates in each of six models or in at least 1 to 6 models were summarized in Table 11-5. One interesting thing we noticed is that there is a significant bias between number of depleted and enriched genes in most of the genetic features: number of depleted genes in CCND1, PTEN, PI3K and E1A is much smaller than enriched ones, while ERBB2 and MYC are on the opposite.

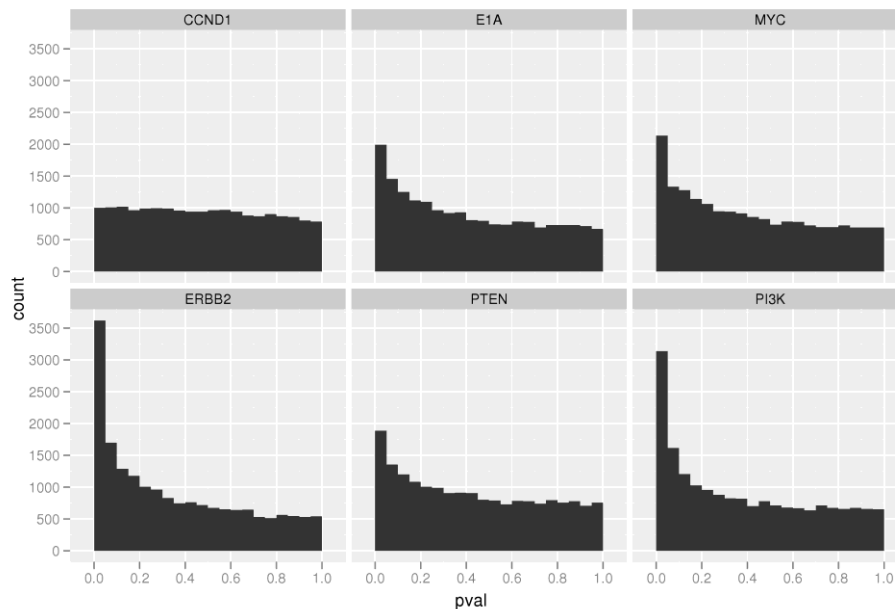


Figure 11-19 Histogram of p values for gene level differential representation analysis of shSeq data from six genetic-engineered models. The bin width is 0.05.

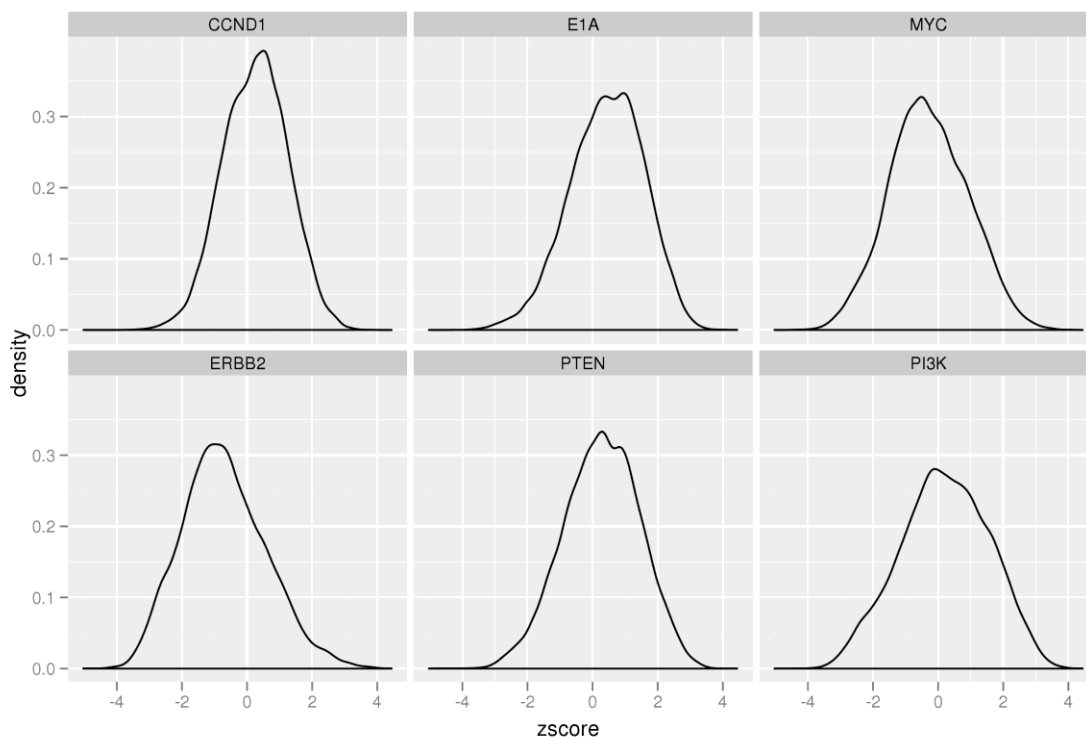


Figure 11-20 Density plot of z scores for gene level differential representation analysis of shSeq data from six genetic-engineered models.

# of Significant Genes				Distribution of Significant Genes			
condition	total	n.over	n.under	Sig in at least # cases	Total	Enriched	Depleted
CCND1	1004	749	255	>=1	8772	4480	5189
E1A	1995	1551	444	>=2	3349	1472	1369
ERBB2	3617	518	3099	>=3	1179	514	384
MYC	2138	628	1510	>=4	362	155	117
PI3K	3137	1899	1238	>=5	94	35	42
PTEN	1884	1313	571	=6	19	2	16

Table 11-5 Number of significant depleted or enriched genes in each of six genetically-engineered models (left) and in at least 1 to 6 these models (right).

P<0.05 is defined as significant.

11.3.3 Sensitivity difference between genetic-features in breast cancer

We were interested in the sensitivity difference of various genetic backgrounds, especially the difference between oncogenes and tumor suppressors. Therefore, we performed sensitivity analysis by counting the number of depleted or enriched genes in each case. As shown in Figure 11-21, for the depleted genes or synthetic lethal partner genes, there is a significantly decrease from ERBB2 to MYC to PI3K to PTEN to E1A and to CCND1, which can be grouped into ERBB2 class, PI3K and MYC class, PTEN, E1A and CCNC1 class. In general, oncogenes (ERBB2, MYC, PI3K) are much more sensitive than tumor suppressors (PTEN, E1A). However, for the enriched genes, there is no such clear pattern, consistency with what we have seen in other screening data.

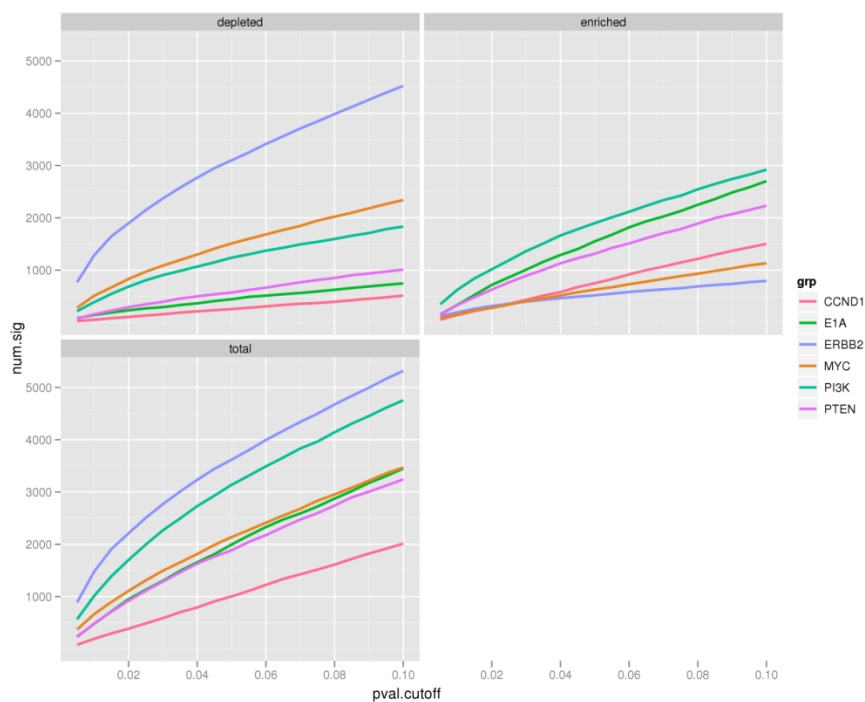


Figure 11-21 Sensitivity analysis of shRNA screening results for six genetic-engineered models.

11.3.4 Unsupervised clustering genetic-features in breast cancer by functional profiles

We also did unsupervised clustering of six genetic features using their functional profiling scores (Figure 11-22). Interestingly, there is a clear separation between tumor suppressors and oncogenes, which might reflect intrinsic and fundamental difference between these two types of tumor casual genes.

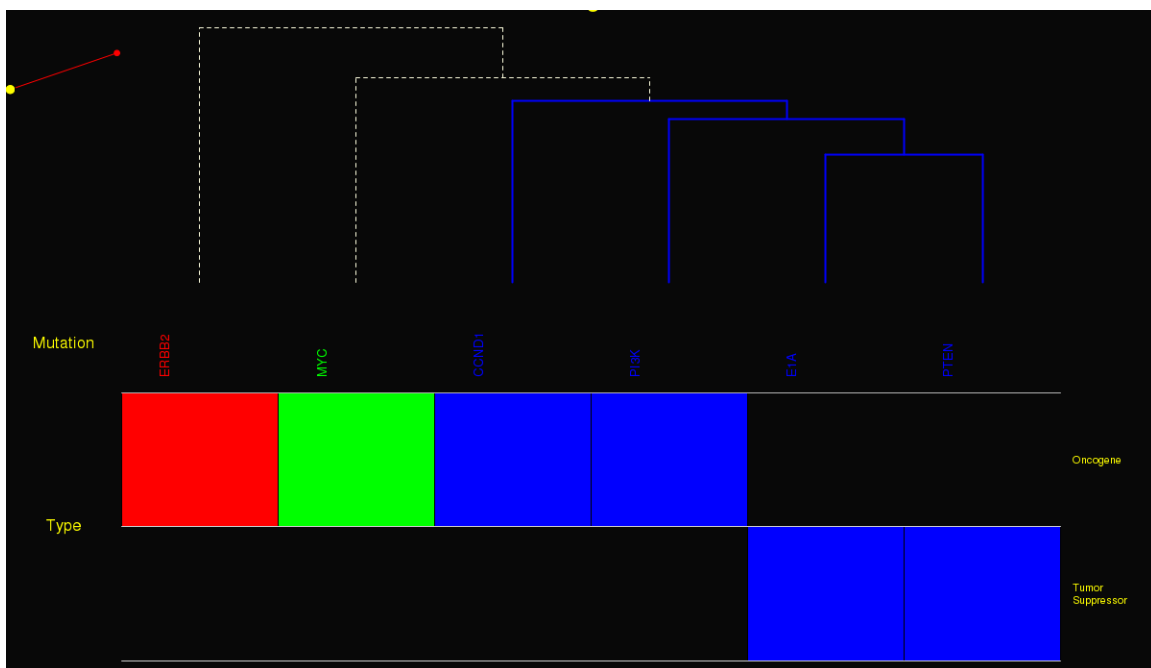


Figure 11-22 Unsupervised clustering of six genetic-engineered models by their shRNA screening functional profiles.

11.3.5 Top candidates genetic-features in breast cancer by functional profiles

Top candidates or synthetic lethal pair genes were selected by the criteria of p value < 0.05 and showing depletion in ≥ 1 genetic-engineered models (Figure

11-23). There are only 16 genes ($P < 0.05$) essential for all 6 driver mutations. The clustering based on these depleted candidates also classified tumor suppressor genes and oncogenes well.

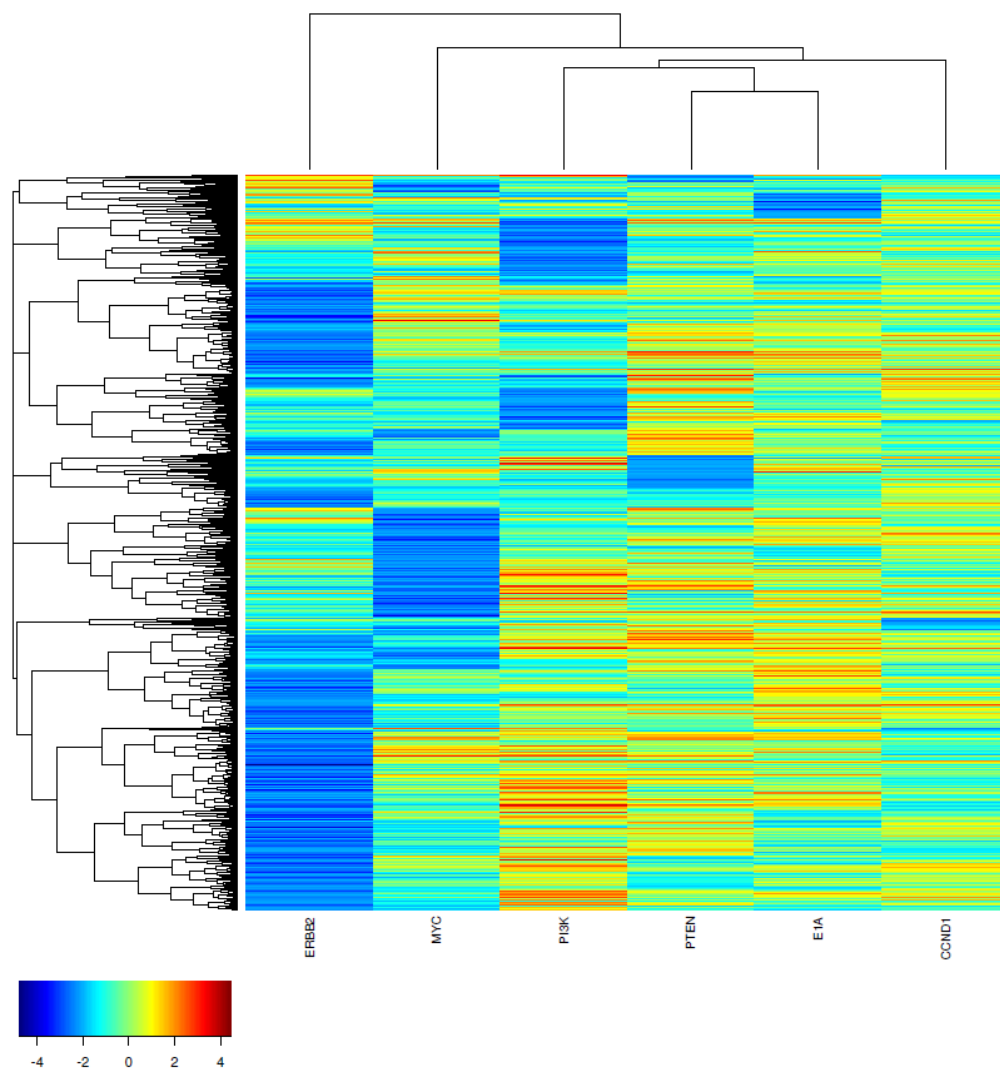


Figure 11-23 Heatmap of top synthetic lethal partners for each of six genetic-engineered models.

11.4 Ongoing and Future Work

Identified candidates for basal or luminal type of breast cancer are being under validation. Other analysis to identify drivers and therapeutic targets specific to basal A, basal B, IBC type of breast cancer will be conducted. This is a joint project with Archana Iyer, Celine Lefebvre, Mariano Alvarez and Yao Shen from our lab and Jose Silva lab.

Chapter 12 Additional shSeq Applications

12.1 Overview

In addition to all the projects that employed the NGS-based shRNA screening (shSeq) technology for therapeutic target discovery as I demonstrated previously, I have also been involved in many other projects by applying shSeq technology in different contexts for different purposes. These additional applications are in collaboration with over ten labs at Columbia. I have analyzed all shSeq data generated at Columbia Genome Center. In total there are over 50 screens with over 10 billion of reads, and the data size is over 5 TB data. The following are a few examples.

12.2 Therapeutic Targets for MYCN-amplified Neuroblastoma

Amplification of MYCN is one of the most important genetic prognosis factors for neuroblastoma (NBL). Neuroblastoma patients with amplification of MYCN have much worse survival than non-MYCN-amplified patients. In this project, collaborating with labs of Darrell Yamashiro and Jose Silva, we aimed to search for therapeutic targets for MYCN-amplified NBL by using shRNA screening. We used an isogenetic model by introducing MYCN amplification in a MYCN- NBL cell line under normoxia and hypoxia environments and did shRNA screening using both microarray and NGS technologies. The results shown below were based on shSeq data (Table 12-1). Top depleted candidates (Figure 12-1) that

are lethal to MYCN-amplified NBL cells were selected from individual analysis of normoxia and hypoxia conditions. Pathway analysis revealed that PI3K and GSK pathways (Figure 12-2) seemed to mediate MYCN amplification and serve as potential targeting avenues to stop MYCN+ NBL cells. Interestingly, known up-regulated targets of MYC showed a significant enrichment in depleted genes of MYCN+ NBL cells (Figure 12-3) and top depleted MYC activated targets (Figure 12-4, Figure 12-5, Figure 12-6) might be interesting therapeutic targets for treatment of MYCN amplified NBL.

Cell Line	BC1	BC2	BC3	BC4	BC5	BC6	total raw reads	total identified reads	identification rate
Normoxia	OFF.A	ON.A	OFF.B	ON.B	OFF.C	ON.C	113,133,354	87,445,922	77.29%
	14,265,751	21,013,556	15,575,280	21,651,813	3,270,776	11,668,746			
	244	359	266	370	56	199			
Hypoxia	OFF.A	ON.A	OFF.B	ON.B	shRNALib	noDNA	161,427,274	124,429,550	77.08%
	19,294,613	235,380	16,686,011	39,818,796	48,176,571	218,179			
	330	4	285	681	824	4			

Table 12-1 Summary of deconvolution for shSeq data of MYCN-amplified and non-MYCN-amplified neuroblastoma samples under normoxia and hypoxia environments. “OFF” is dox-off representing MYCN-amplification, while “ON” is for non-MYCN-amplification. Numbers in red are the cases with low total number of identified reads.

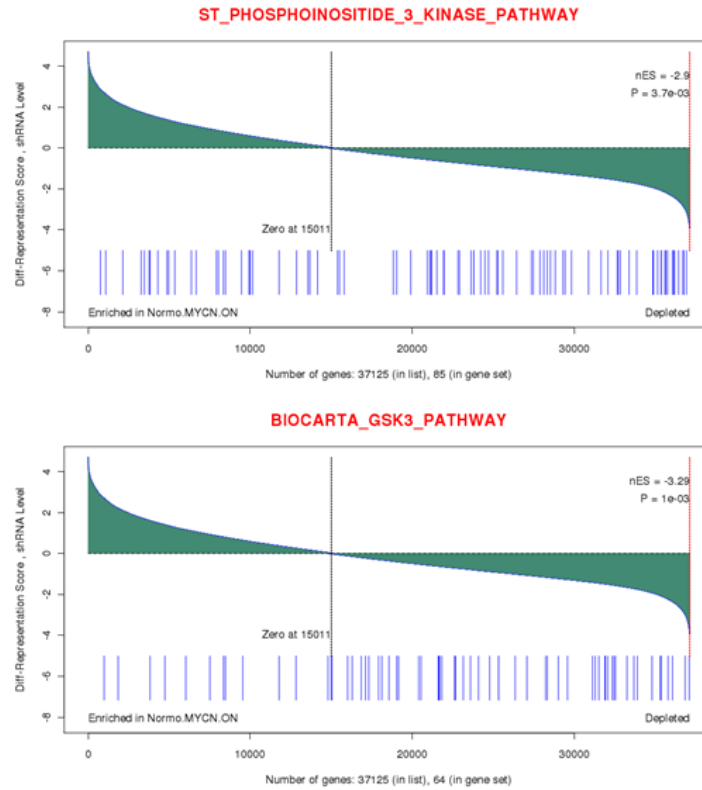


Figure 12-2 Top pathways enriched by depleted hairpins in MYCN-amplified NBL line under normoxia condition.

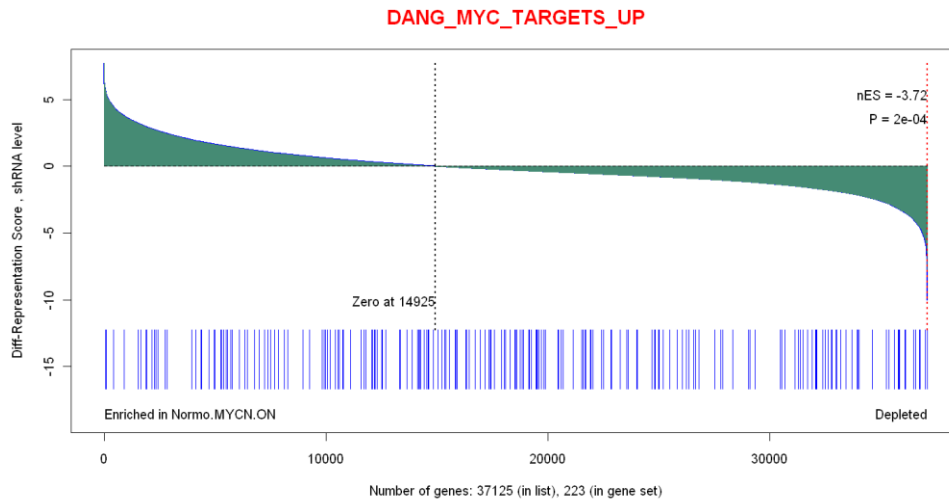


Figure 12-3 MYC up-regulated targets is enriched by depleted hairpins in MYCN-amplified NBL line under hypoxia condition.

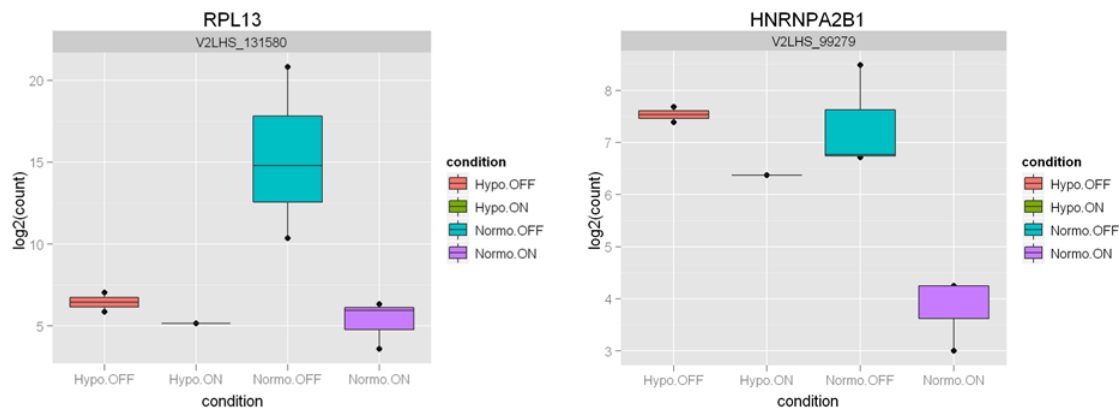


Figure 12-4 RPL13 and HNRNPA2B1 as MYC activated targets are lethal to MYCN-amplified NBL cells.

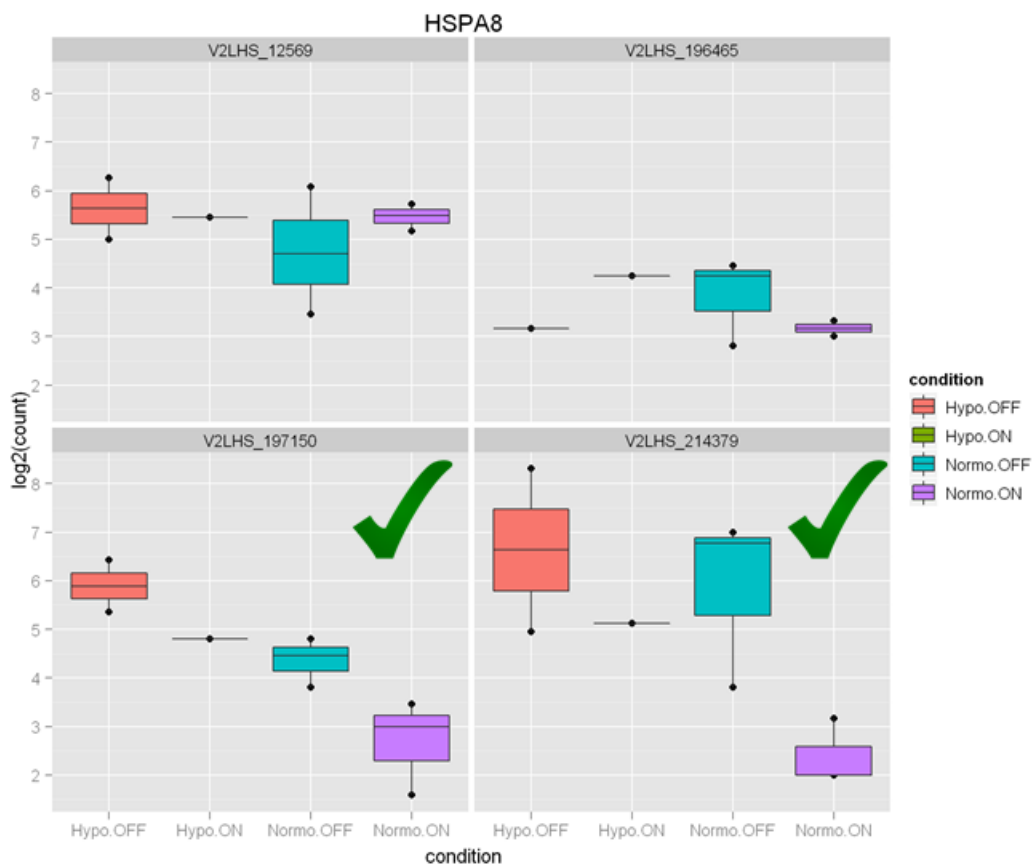


Figure 12-5 HSPA8 as MYC activated target is lethal to MYCN-amplified NBL cells.

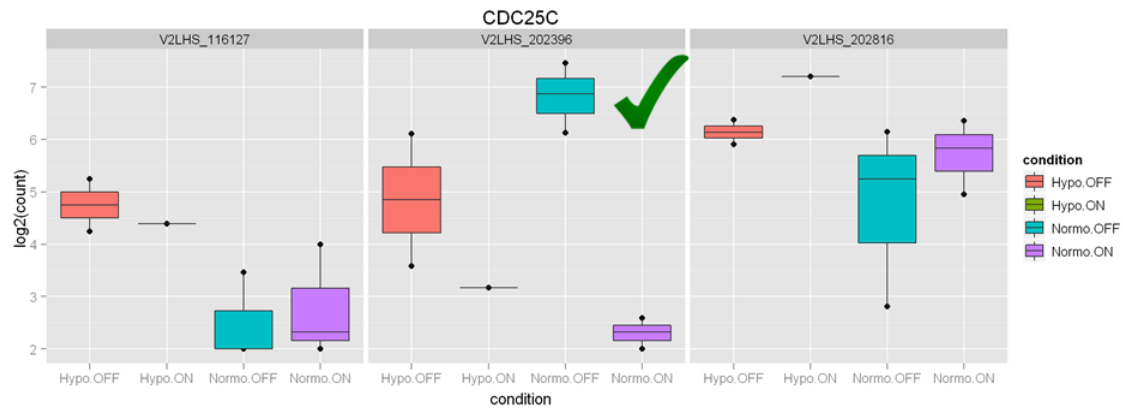


Figure 12-6 CDC25C as MYC activated target is lethal to MYCN-amplified NBL cells.

12.3 Overcoming Cisplatin or PARP Inhibitor Resistance in Small Cell Lung Cancer

Cisplatin is one of the most commonly used chemotherapeutic agents for solid tumor, such as small cell lung cancer because Cisplatin is able to trigger apoptosis by causing crosslinking of DNA. However, over 70% of patients with Cisplatin treatment are resistant or relapse to develop resistance. PARP1 is important for repairing single-strand break and its inhibition causes multiple double strand breaks. PARP inhibitor is a new on-trial drug for treatment of lung cancer and it has been that shown to have dependence on BRCA-mutated cancer cells. However, a large percentage of lung cancer patients have no response to PARP inhibitor treatment. Collaborating with Haiying Cheng and Jose Silva, this project aimed to search for therapeutic targets or modifiers to overcome resistance of Cisplatin and PARP inhibition treatments in small cell

lung cancer. Pooled shRNA screens by NGS were performed for a lung cancer cell line with and without Cisplatin (IC20 dose) treatment or PARP inhibitor (IC50 dose) treatment. DMSO was used as control. Potential modifiers of resistance would be dropped out in the treated cells comparing with DMSO control. In the NGS data, nine samples were mixed together for barcode sequencing and all samples except one were successfully deconvoluted (Table 12-2). Analysis of individual hairpins was done using both statistical shADER method and simple fold change analysis (Figure 12-7, Figure 12-8 and Figure 12-9). Top candidates to reverse resistance to Cisplatin or PARP inhibition treatment were selected (Figure 12-11), and top enriched pathways were identified (Figure 12-10).

BC1	BC2	BC3	BC4	BC5	BC6	BC7	BC8	BC9	total raw reads	total identified reads	identification rate
DMSO.A	DMSO.B	DMSO.C	Cis.A	Cis.B	Cis.C	PARP.A	PARP.B	PARP.C			
11,271,492	13,666,953	11,096,133	14,480,241	11,764,091	15,304,810	23,848,369	21,771,492	306,068	159,921,212	123,509,649	77.23%
193	234	190	248	201	262	408	372	5			

Table 12-2 Summary of shSeq data for small cell lung cancer with and without Cisplatin or PARP inhibitor treatment. Cis=Cisplatin, PARP=PARP inhibitor, PARP.C was removed for further analysis because of its low signals.

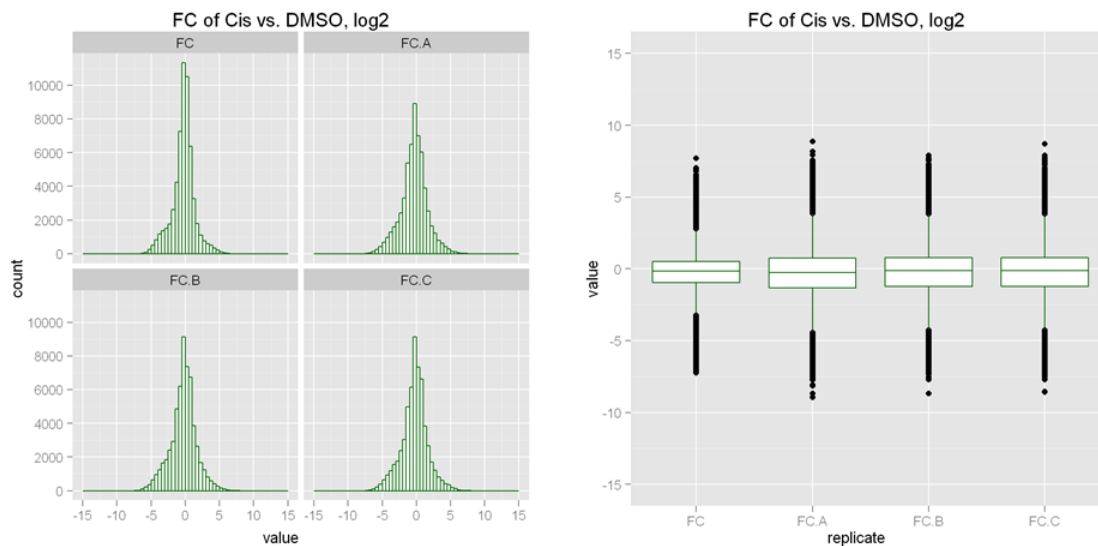


Figure 12-7 Histogram and boxplot of fold change (log₂ transformed) of averaged and three individual replicates of shSeq data Cisplatin treatment vs. DMSO.

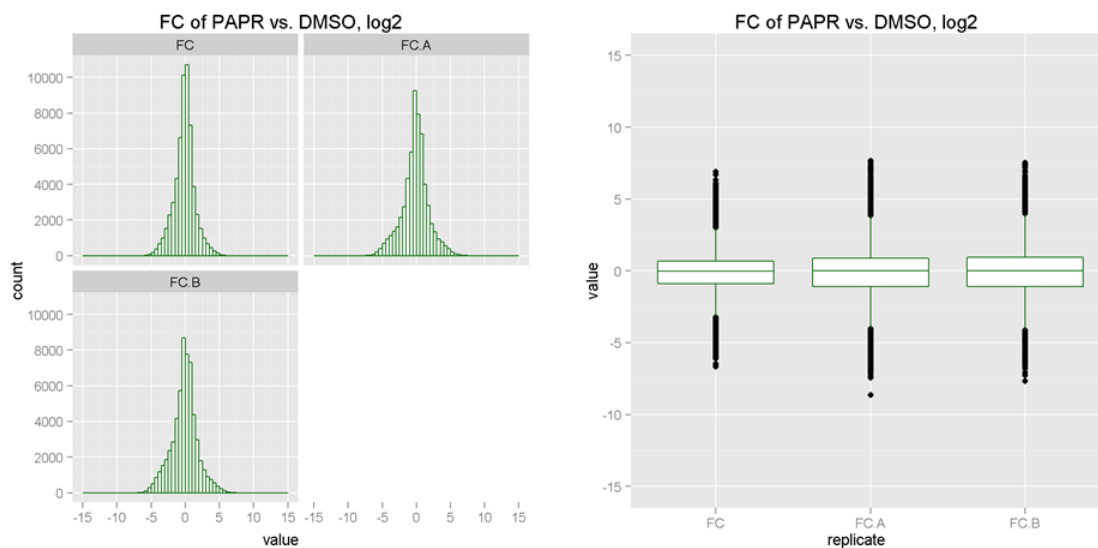


Figure 12-8 Histogram and boxplot of fold change (log₂ transformed) of averaged and three individual replicates of shSeq data PARP inhibitor treatment vs. DMSO.

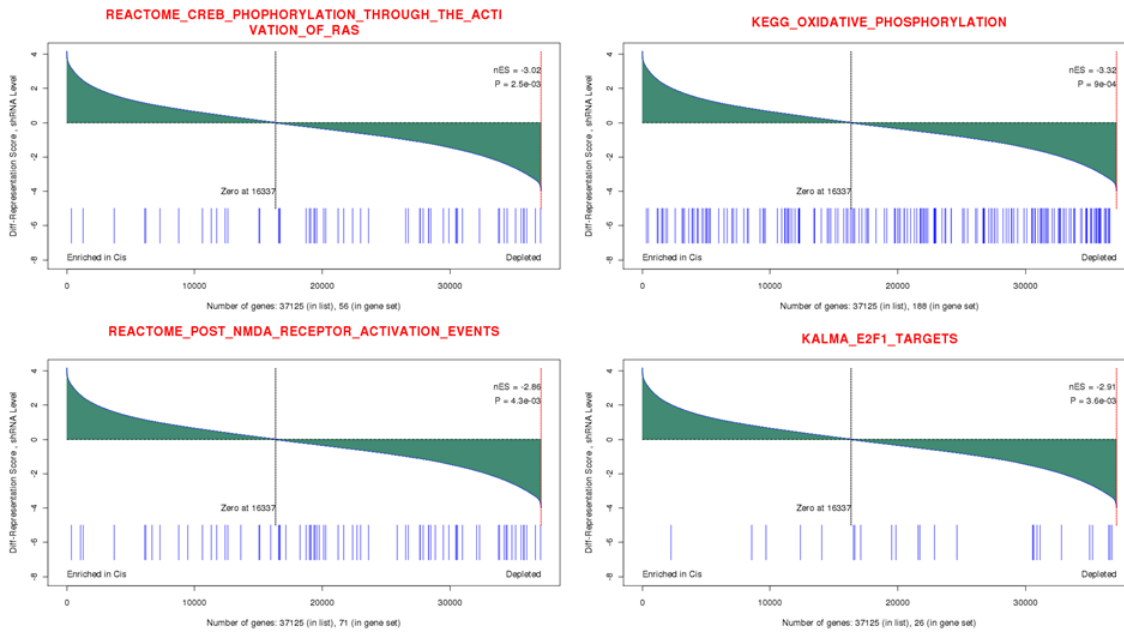


Figure 12-11 Top enriched pathways by depleted candidates from shSeq results of Cisplatin treatment comparing DMSO control.

12.4 Genetic Modifiers of SMN as Therapeutic Targets for Spinal Muscular Atrophy

Spinal muscular atrophy (SMA) – the most common genetic cause of death in infancy – is a motor neuron disease caused by reduced expression of the survival motor neuron (SMN) protein. SMA patients have homozygous loss of the *SMN1* gene and retain at least one copy of the nearly identical *SMN2* gene. Currently, there is no effective treatment for SMA and most therapeutic efforts focus on identifying strategies that enhance expression of SMN from the *SMN2* gene. However, discovery efforts for SMA therapeutics are hampered by the limited knowledge of suitable targets. Collaborating with Livio Pellizzoni and Jose Silva, this project aims to identify and characterize cellular factors that control the

expression and function of the SMN protein with the ultimate goal of identifying novel avenues of therapeutic intervention for SMA via genome-wide RNAi screening. Preliminary results were reported in Table 12-3 and Figure 12-12.

Lane	CellLine	Time	F10	F2	F3	F4	F5	F6	total raw reads	total identified reads	identification rate
			-Dox	-Dox	-Dox	+Dox	+Dox	+Dox			
4	NIH3T3 mSmni	D7	29,979,667	24,369,195	23,630,324	20,525,747	42,247,245	19,208,095	189,486,269	159,960,273	84.42%
			368	299	290	252	519	236			
5	NIH3T3 mSmni	D14	43,796,015	26,260,187	15,040,358	25,979,578	2,418,773	26,671,798	176,197,502	140,166,709	79.55%
			538	322	185	319	30	327			
6	NIH3T3 mSmni/hSMN2 _{Low}	D7	30,064,708	25,991,899	15,581,857	20,796,464	23,110,646	21,101,929	159,301,217	136,647,503	85.78%
			369	319	191	255	284	259			
7	NIH3T3 mSmni/hSMN2 _{Low}	D14	29,137,770	21,198,521	23,496,439	35,344,603	19,269,471	32,186,798	193,827,729	160,633,602	82.87%
			358	260	288	434	237	395			

Table 12-3 Summary of deconvolution of shSeq data for SMA project.

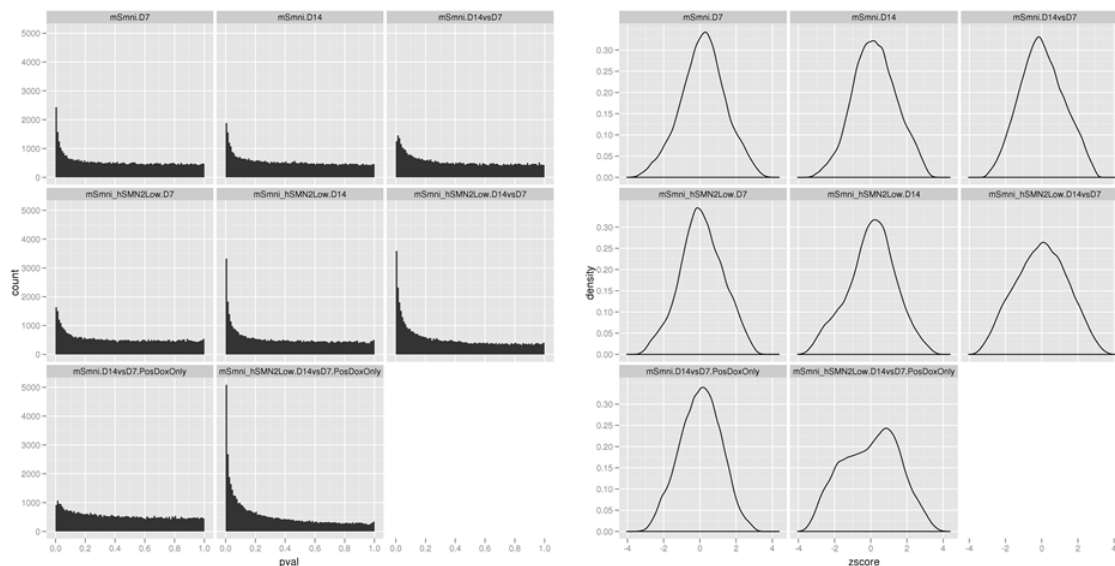


Figure 12-12 Histogram of p values and density plot of z scores from differential representation analysis at individual hairpin level.

12.5 Positive shRNA Screens to Identify Novel Modulators of P53 Pathway

In collaboration with Wei Gu and Jose Silva, this project aims to identify genes that are able to survive P53 induced apoptosis as potential modulators of P53 pathway. A positive shRNA screen using NGS was carried out on a P53 null cell line with and without P53 inducement in four replicates. The deconvolution of shSeq data looked fine (Table 12-4), however, the distribution of read counts (Figure 12-13) and results of differential representation analysis (Figure 12-14) reflected some flaws of this experiment. For example, there is one or two hairpins which have a count of over 2M, about 1/5 of total reads in each sample, and the majority of hairpins have zero or very low count. Also the distribution of non-significant p values showed abnormal behavior.

BC1	BC2	BC3	BC4	BC5	BC6	BC7	BC8	total raw reads	total identified reads	identification rate
A+	B+	C+	D+	A-	B-	C-	D-			
15,728,627	14,146,810	13,988,996	12,946,455	13,988,372	14,299,239	12,471,239	13,598,000			
269	242	239	221	239	244	213	232	134,923,102	111,167,738	82.39%

Table 12-4 Deconvolution table of shSeq data for P53 positive screen project.

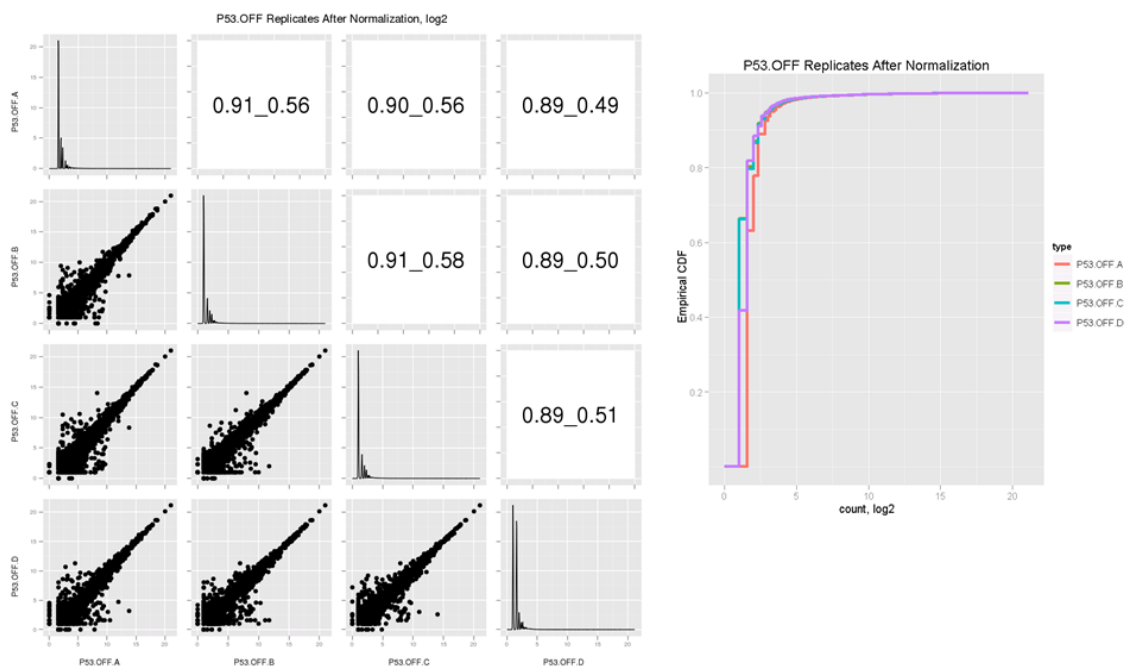


Figure 12-13 Consistency of replicates for shSeq data of P53 positive screen.

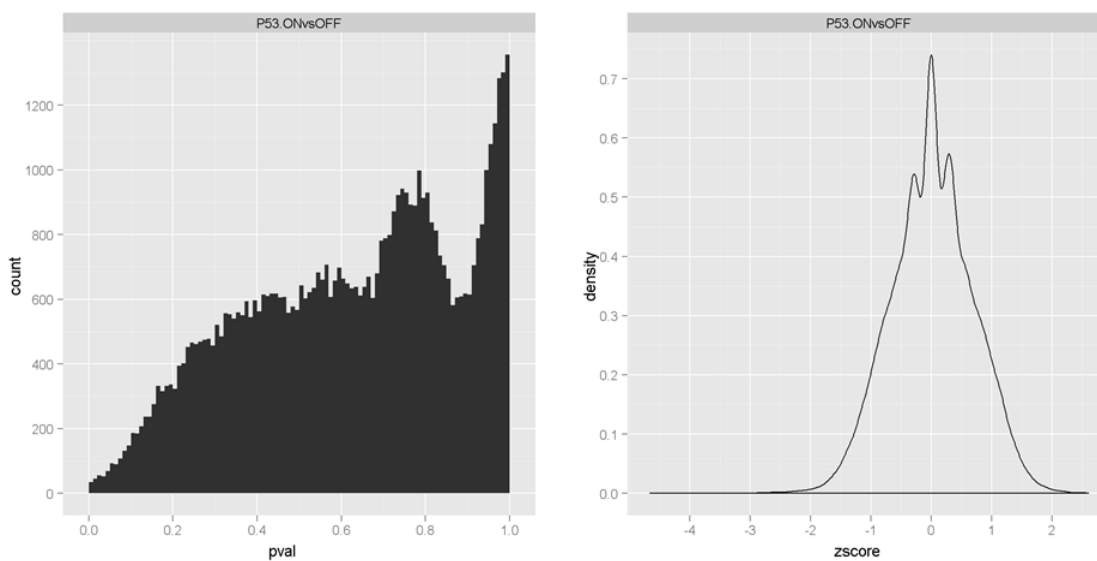


Figure 12-14 Histogram of p values and density plot of z scores from differential representation analysis at individual hairpin level.

12.6 Overcoming Resistance to Glucocorticoid or NOTCH-inhibition in T-ALL

This study continued the project of glucocorticoid resistance in Chapter 7 and Chapter 8 by using NGS technology and screening one more cell line, and also aimed to identify therapeutic targets to reverse resistance of NOTCH-inhibition, another major strategy for treatment of NOTCH1-mutated T-ALL. Genome-wide shRNA screens by NGS were performed on three cell lines with and without glucocorticoid or NOTCH-inhibitor treatment, in which all three are resistant to glucocorticoid and two are resistant to treatment of NOTCH inhibition.

Pool	Samples	Barcodes	Duplication	Total Raw Reads	Algorithm 1				Algorithm 2			
					Identified Reads	Avg Identified Reads/shRNA	Total Identified Reads	Identification Rate	Identified Reads.2nd	Avg Identified Reads/shRNA	Total Identified Reads	Identification Rate
1	CUTLL1_DBZ_1	F1	Yes	180,019,354	23,642,987	404	72,970,038	40.54%	12,633,795	216	44,911,753	24.95%
	CUTLL1_DBZ_2	F2	Yes		26,977,346	461			17,140,726	293		
	CUTLL1_DBZ_3	F3	No		22,349,705	382			15,137,232	259		
2	CUTLL1_DEXA_1	F1	No	166,678,423	21,941,225	375	64,626,269	38.17%	12,219,989	209	38,122,564	22.87%
	CUTLL1_DEXA_2	F2	No		23,179,775	396			14,415,942	246		
	CUTLL1_DEXA_3	F3	No		19,505,269	333			11,486,633	196		
3	CUTLL1_DMSO_1	F4	No	161,753,666	21,907,690	375	64,537,024	33.98%	13,674,390	234	36,722,629	22.70%
	CUTLL1_DMSO_2	F5	No		18,743,561	320			11,164,848	191		
	CUTLL1_DMSO_3	F6	No		23,885,773	408			11,883,391	203		
4	HPB ALL_DEXA_1	F1	No	184,632,267	26,702,515	457	79,423,460	43.03%	17,437,585	298	53,215,671	28.82%
	HPB ALL_DEXA_2	F2	No		26,979,015	461			17,522,670	300		
	HPB ALL_DEXA_3	F3	No		25,741,930	440			18,255,416	312		
5	HPB ALL_DMSO_1	F4	Yes	187,538,533	26,534,757	454	80,484,028	42.37%	18,732,282	320	57,207,016	30.50%
	HPB ALL_DMSO_2	F5	Yes		26,791,548	458			20,028,672	342		
	HPB ALL_DMSO_3	F6	Yes		27,157,723	464			18,446,062	315		
6	JURKAT_DBZ_1	F1	Yes	200,312,575	14,546,661	249	43,123,283	21.53%	8,691,724	149	27,070,384	13.51%
	JURKAT_DBZ_2	F2	Yes		15,197,419	260			9,248,846	158		
	JURKAT_DBZ_3	F3	No		13,379,203	229			9,129,814	156		
6 re run	JURKAT_DBZ_1	F1	Yes	151,884,986	17,959,340	307	52,746,999	34.73%	10,005,116	171	30,939,086	20.37%
	JURKAT_DBZ_2	F2	Yes		18,745,965	320			10,634,982	182		
	JURKAT_DBZ_3	F3	No		16,041,694	274			10,298,988	176		
7	JURKAT_DEXA_1	F1	No	201,711,983	25,616,627	438	72,188,720	35.79%	12,745,890	218	37,273,962	18.48%
	JURKAT_DEXA_2	F2	No		25,425,212	435			13,296,623	227		
	JURKAT_DEXA_3	F3	No		21,146,881	362			11,231,449	192		
8	JURKAT_DMSO_1	F4	Yes	192,322,214	22,649,098	387	68,880,936	35.52%	12,869,426	220	36,784,686	19.13%
	JURKAT_DMSO_2	F5	No		18,056,173	309			9,612,494	164		
	JURKAT_DMSO_3	F6	No		28,175,665	482			14,302,766	245		

Table 12-5 Deconvolution of shSeq data for glucocorticoid or NOTCH-inhibition resistance in T-ALL. The sequence run in dark red is the re-sequenced because of a technical flaw of previous run.

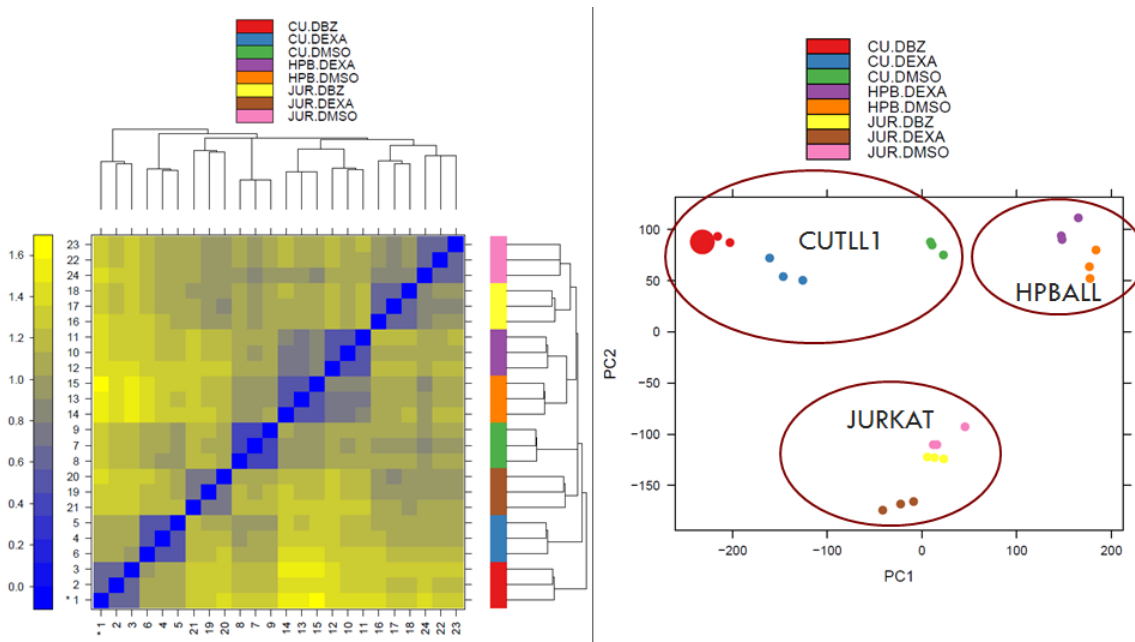


Figure 12-15 Heatmap of sample distances and PCA plot of samples showed consistence with biological meanings.

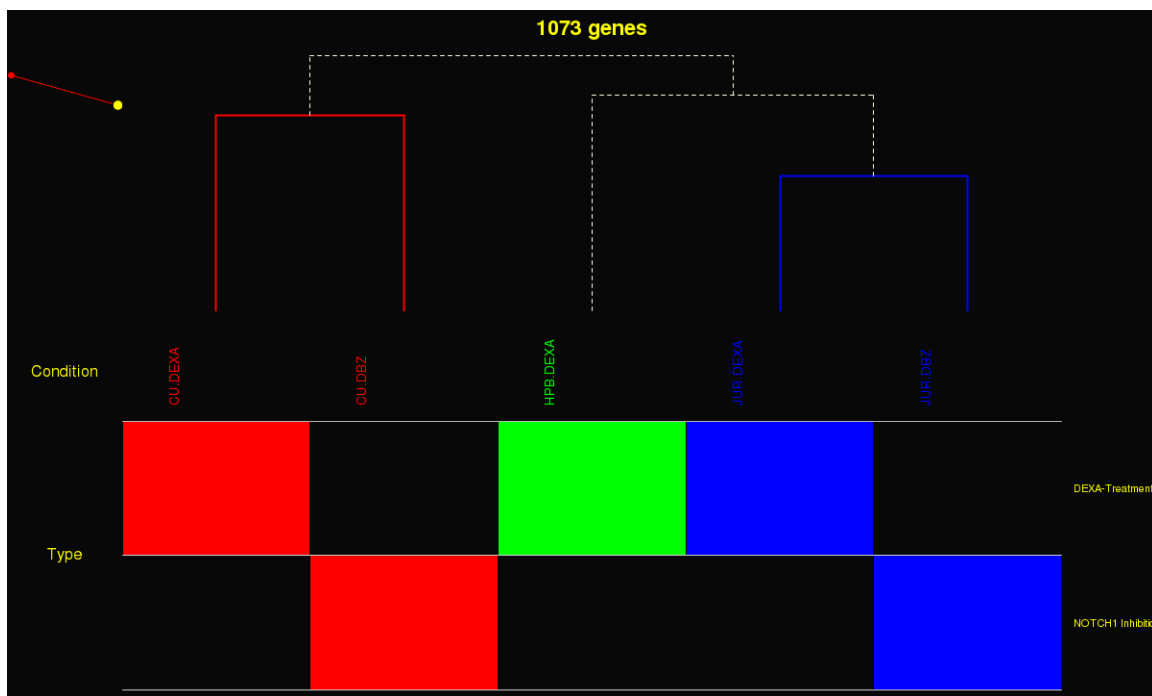


Figure 12-16 Unsupervised clustering of conditions in shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.

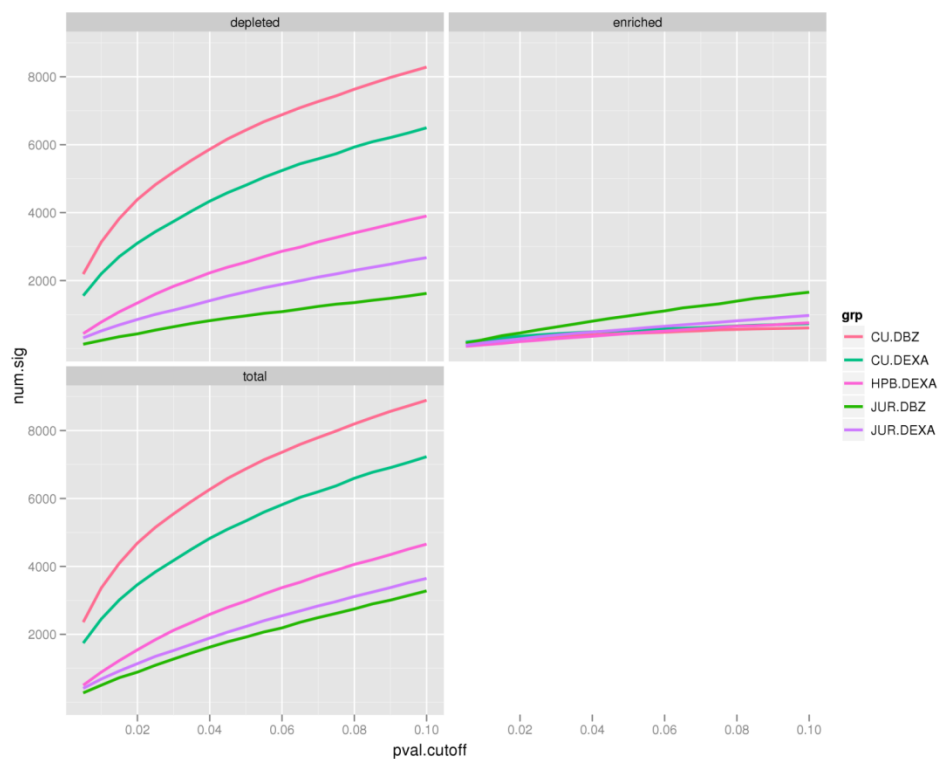


Figure 12-17 Sensitivity analysis of shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.

# of Significant Genes				Distribution of Significant Genes			
condition	total	n.over	n.under	Sig in at least # cases	Total	Enriched	Depleted
CU.DBZ	6872	442	6430	>=1	11922	2512	10739
CU.DEXA	5339	532	4807	>=2	5469	375	4436
HPB.DEXA	2979	439	2540	>=3	1609	44	1063
JUR.DBZ.old	2637	547	2090	>=4	301	5	149
JUR.DBZ	1919	958	961	=5	38	0	16
JUR.DEXA	2230	565	1665				

Table 12-6 Number of significant ($P < 0.05$) candidates in each of the five cases or in at least one to five cases of shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.

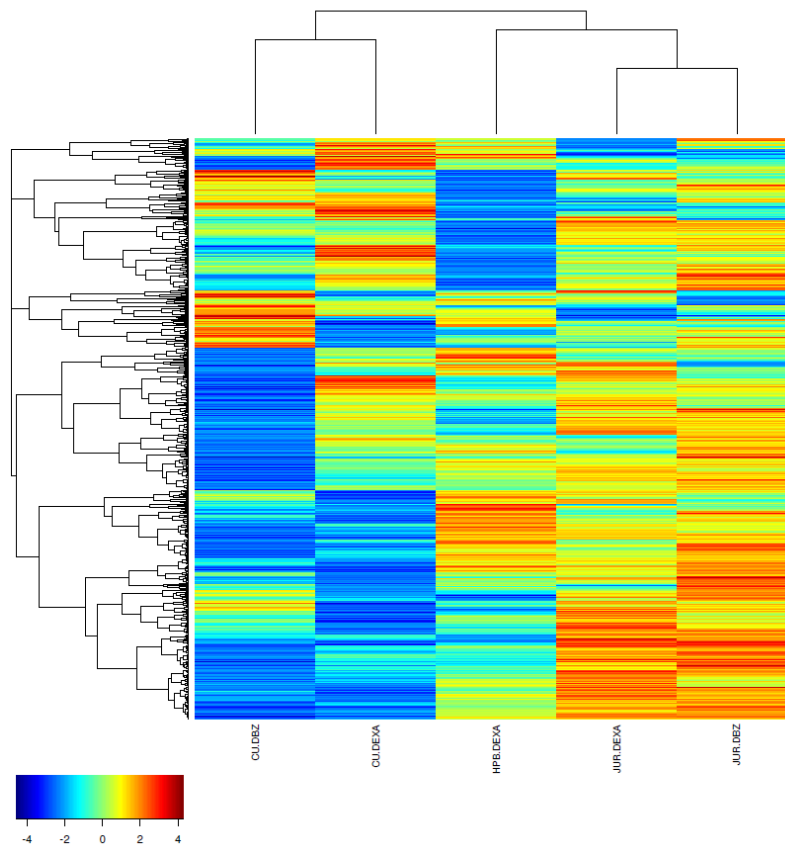


Figure 12-18 Top candidates that are depleted in at least one of five cases in shSeq studies for glucocorticoid or NOTCH-inhibition resistance in T-ALL.

Deconvolution (Table 12-5) and QA (Figure 12-15) showed all shSeq data were in good shape. In unsupervised clustering (Figure 12-16), functional profiles from the same cell line were clustered together, revealing intrinsic difference between the three cell lines. Sensitivity analysis (Figure 12-17) showed a decrease of sensitivity from CUTLL1 to HPBALL to JUKAT. Top candidates (Table 12-6, Figure 12-18) identified to reverse either of the two resistances are being under validation.

Chapter 13 A Dynamic Web System for Collaboration Management

13.1 Overview

To facilitate our research and to manager a large number of collaborative projects, in winter 2009, I developed a user-friendly and dynamic web system for the lab (<http://califano.c2b2.columbia.edu>). This system is under Apache-PHP-MySQL framework. It is based on Drupal [249] an open-sourced content management system and Open Atrium [250], an Drupal-enabled team collaboration tool. It is project or group-oriented: each project is a group and within the group, you can create accounts for your collaborators and add existing users into this group. You can share documents, blogs, events, tweets, comments, or anything within the group. You can make the group public or private. This system also supports different user roles with different level of permissions. So far over 60 groups have been created and over 120 users, of which half are collaborators, have been added.

13.2 Features

In the following, I will demonstrate with snapshots a few representative features for collaboration management in this web system.

13.2.1 Dashboard

Dashboard is the default home page after you log into the system (Figure 13-1). It shows the latest activities in your groups including shoutbox messages, blog teasers and list of group-tagged activity titles. “All activity” shows all activities you are permitted to view. “My threads” shows the posts that you are subscribed to. “Files” shows you any files that have been attached to new posts (Figure 13-2).

Home My groups Jiyang Yu Need help?

Califano Lab, Columbia University

Home Create content

View All activity My threads Files

Shoutbox

Twitter for Lab General

Jiyang Yu 12:17pm Wed Jun 20 George Casella passed away because of cancer... if you have ever read his statistics books... RP...

Jiyang Yu 11:37am Mon May 21 HCCC Annual Symposium: Epigenetics And Cancer
Sat-6pm, May 23, Ross Berne Pavilion, Room 1A2. Registration is not needed. I checked with Maria Baton.

Mariano Alvarez 7:35am Tue May 1 Changes to MARNA package:
1) Shadow function was added that performs shadow analysis on a MARNA object and return an updated regulon object corrected for shadow effect

Blog

Oops forgot! Back from Eye surgery

Jose Morales 9:59pm Aug 11, 2012 Add new comment

- 4:56pm Aug 10, 2012 Back from eye surgery! Jose Morales 0 comment(s)
- 11:39am Fri Jul 27 LYMPS v 1.2 UPDATE: Current edition matched LYMPS v 1.0 best (0.37) Jose Morales 1 comment(s)
- 11:52am Thu Jul 19 SUV, Troponin, Chaitan and me: The happy tale of the winding path of a recent result Jose Morales 0 comment(s)
- 2:16pm Wed Jul 18 Journal Briefing July 18th 2012 - data manipulation from the command line Yishai Shimon 1 comment(s)
- 1:56pm Wed Jul 18 SUV and Troponin: The amazing information content of MDS Jose Morales 0 comment(s)

Recent activity

Thursday, Aug 16

- 2:06pm LYMPS Collaborative Project Jiyang Yu

Saturday, Aug 11

- 9:59pm Oops forgot! Back from Eye surgery Jose Morales

Friday, Aug 10

- 4:56pm Back from eye surgery! Jose Morales

Friday, Jul 27

- 4:56pm SNP in NBL maria
- 4:54pm SNP in NBL maria

Thursday, Jul 19

- 3:08pm Presha's Presentations Presha Rajbhandari
- 3:06pm SNP in NBL maria

About

This system is designed for Califano lab internal **COMMUNICATION, project MANAGEMENT, and COLLABORATIONS!**
Learn more in the **PRESENTATION** of introduction to this system, and **SUGGESTIONS** for beginners.

Twitter for Lab General

Show

Mariano Alvarez 7:35am Tue May 1 Changes to MARNA package:
1) Shadow function was added that performs shadow analysis on a MARNA object and return an updated regulon object corrected for shadow effect
2) TF-distance function was upgraded to include the Jaccard's dissimilarity metric method
The updated package can be obtained from `ifs\data\c2b2\ac_lab\malvarez\packages/`
Please, report any bug you find.

Jiyang Yu 11:08am Fri Apr 13 discussion continued on Nature Biotech about the Science paper

Mariano Alvarez 10:27am Fri Apr 13 I fixed a small bug in function `marinaModule` from package `marina`. So please, get the last version from the packages repository: `ifs\data\c2b2\ac_lab\malvarez\packages/`

Calendar

August 2012

M	T	W	T	F	S	S
30	31	1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2

Upcoming events

Developed and powered by Jiyang Yu

Figure 13-1 Collaborative web system: site dashboard - home page

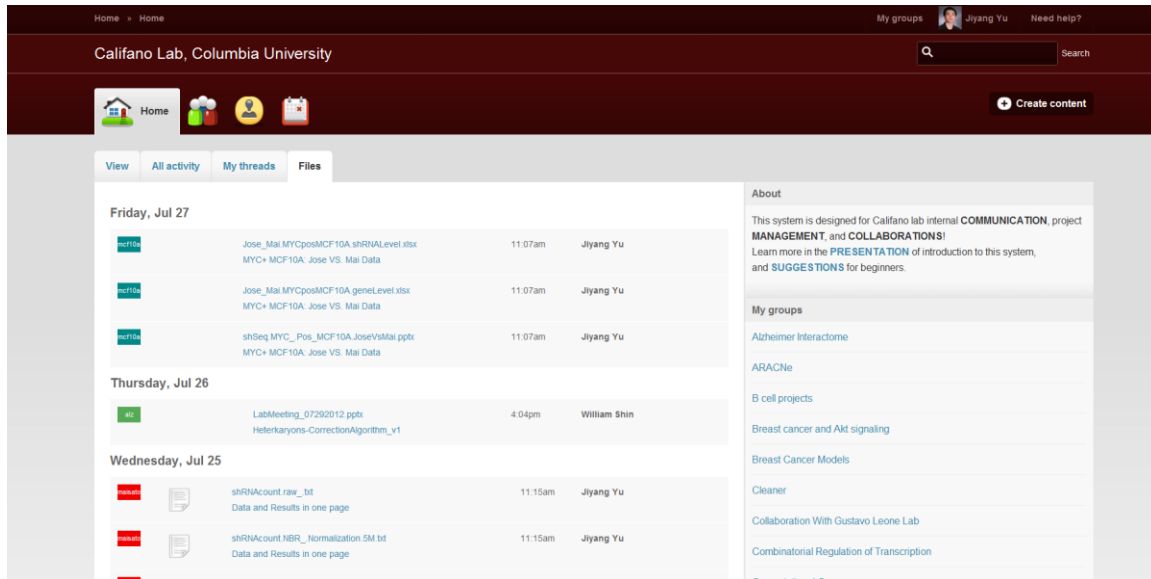


Figure 13-2 Collaborative web system: site dashboard - files

13.2.2 Group

The key feature of this system is to organize contents in groups (Figure 13-3, Figure 13-4, Figure 13-5, and Figure 13-6). A group could be created for (only by current member)

- an on-going project by who is mainly in charge (group admin)
- a topic of interest: e.g. Computational Group
- general membership: Lab General Activities.

The user who creates the group is automatically as group admin, which can add/approve new members, give member admin permission, add new account for collaborator and modify group settings, etc.

There are two types of groups supported:

- Public group: everything is public to all users in the system
- Private group: only available for group members.

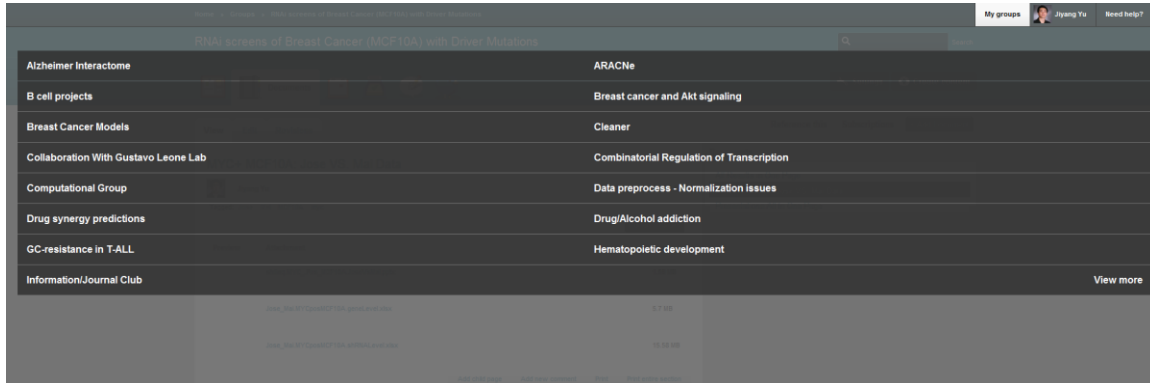


Figure 13-3 Collaborative web system: my groups

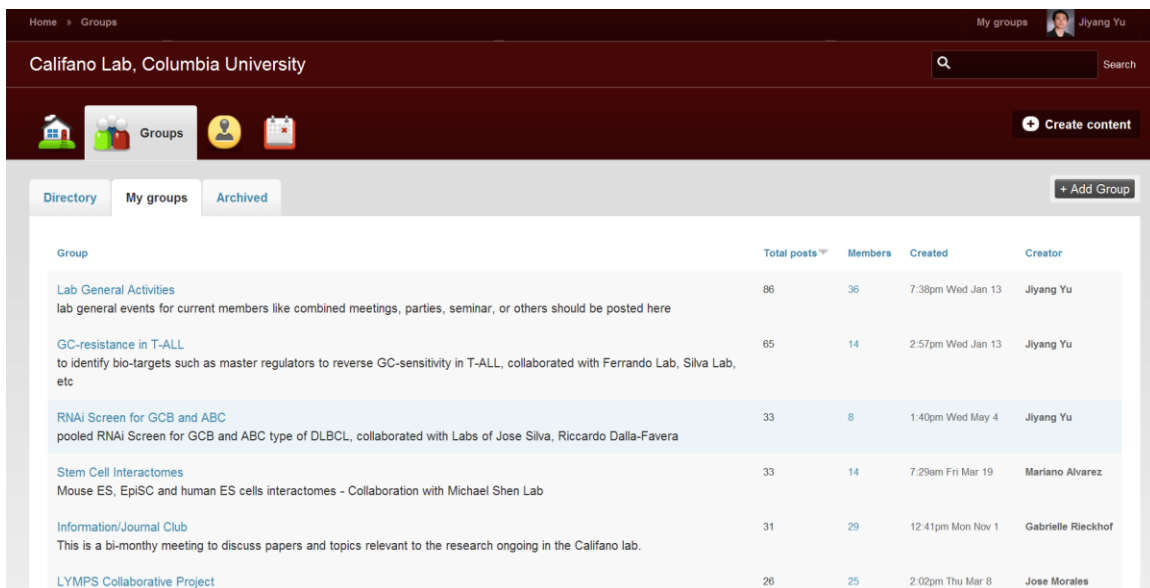


Figure 13-4 Collaborative web system: group list

	In Sock Jang	js113@c2b2.columbia.edu	Computational Group, Lab General Activities, ARACNe, Cleaner, B cell projects, Motif Discovery, Prediction for Causal Effects of Perturbation, MINDy, MaRNA, Drug/Alcohol addiction, Reverse Engineering Signaling Networks, Prostate Cancer Interactome, Alzheimer Interactome, Ovarian Cancer, Integrated Interactome (naive bayes classification), Data preprocess - Normalization issues, Information/Journal Club, NBL, Hematopoietic development, PrePP	past member
	James Chen	jc2162@columbia.edu	Computational Group, Lab General Activities, PDx or Personalized Medicine, Combinatorial Regulation of Transcription, Motif Discovery, Glioblastoma Multiforme, Synthetic Dosage Lethality Screening, MINDy, MaRNA, Drug/Alcohol addiction, Reverse Engineering Signaling Networks, Prostate Cancer Interactome, Alzheimer Interactome, Ovarian Cancer, Integrated Interactome (naive bayes classification), Data preprocess - Normalization issues, Information/Journal Club, NBL, FANTOMS, Wet Lab Inventories, Hematopoietic development, PrePP, LNCs, R packages, Motor/Neuron Interactome, Tamoxifen resistance in ESR1+ breast carcinoma, Subtype-specific addiction points in breast carcinoma	current member
	Jiyang Yu	jj232@c2b2.columbia.edu	Computational Group, GC-resistance in T-ALL, Lab General Activities, ARACNe, Cleaner, Phenomics, PDx or Personalized Medicine, B cell projects, Combinatorial Regulation of Transcription, Motif Discovery, MINDy, MaRNA, Stem Cell Interactomes, Drug/Alcohol addiction, Reverse Engineering Signaling Networks, Prostate Cancer Interactome, Alzheimer Interactome, Breast cancer and Akt signaling, Ovarian Cancer, Breast Cancer Models, Integrated Interactome (naive bayes classification), Data preprocess - Normalization issues, Information/Journal Club, NBL, Optimal shRNA Design, RNAi Screen for GCB and ABC, RNAi screen of Breast Cancer Cells (MCF10A) with PTEN or PI3K Mutations, RNAi screen of Neuroblastoma (NB) with MYCN Amplification, RNAi screen for Breast Cancer, Hematopoietic development, PrePP, Search for Therapeutics for ERBB2/HER2+ Breast Cancer, RNAi Screen of Drug-resistant Lung Cancer, Positive RNAi Screen of PS3-induced Cells, R packages, LYMP5 Collaborative Project, RNAi Screen on Mouse NIH3T3 line w/w/o edr/mak/m2/Low, Drug synergy predictors, RNAi screens of Breast Cancer (MCF10A) with Driver Mutations, Motor/Neuron Interactome, Tamoxifen resistance in ESR1+ breast carcinoma, Subtype-specific addiction points in breast carcinoma, RNAi Screening for MYC+ Xenopus lewis cells, Collaboration With Gustavo Leone Lab	current member
	John Dick	jdick@uhresearch.ca	Hematopoietic development	collaborator
	Alla Babina	ab7001@bmi.columbia.edu	LYMP5 Collaborative Project	collaborator
	Alvaro Aytes Meneses	aam2177@columbia.edu	Information/Journal Club	collaborator
	am3489	antonina@c2b2.columbia.edu	Computational Group, Lab General Activities, Cleaner, Combinatorial Regulation of Transcription, Motif Discovery, MINDy, MaRNA, Drug/Alcohol addiction, Reverse Engineering Signaling Networks, Prostate Cancer Interactome, Alzheimer Interactome, Ovarian Cancer, Integrated Interactome (naive bayes classification), Data preprocess - Normalization issues, Sage Federation, Information/Journal Club, NBL, FANTOMS, Hematopoietic development, PrePP, R packages, Motor/Neuron Interactome, Tamoxifen resistance in ESR1+ breast carcinoma, Subtype-specific addiction points in breast carcinoma	current member
	Andrea Califano	califano@c2b2.columbia.edu	Computational Group, GC-resistance in T-ALL, Lab General Activities, ARACNe, Cleaner, Phenomics, PDx or Personalized Medicine, B cell projects, MYC and BCL6 pathological interaction, Combinatorial Regulation of Transcription, Motif Discovery, Glioblastoma Multiforme, HTS, Prediction for Causal Effects of Perturbation, MINDy, Hamilton Robot, Germ Cell Tumors, MaRNA, Stem Cell Interactomes, Drug/Alcohol addiction, Validations of the Predicted Modulators of Master Regulators, CST, Reverse Engineering Signaling Networks, Prostate Cancer Interactome, Alzheimer Interactome, Breast cancer and Akt signaling, Ovarian Cancer, Breast Cancer Models, Integrated Interactome (naive bayes classification), microRNA biogenesis and targeting, Data preprocess - Normalization issues, Colon and Breast carcinoma METs, Sage Federation, ALS - SCOT, NBL, ALS, Transcriptome Analysis, FANTOMS, STAT3/CEBPb Broad collaboration, Wet Lab Inventories, RNAi Screen for GCB and ABC, RNAi screen of Breast Cancer Cells (MCF10A) with PTEN or PI3K Mutations, RNAi screen of Neuroblastoma (NB) with MYCN Amplification, Chronic Lung Disease, RNAi screen for Breast Cancer, Hematopoietic development, PrePP, LNCs, Search for Therapeutics for ERBB2/HER2+ Breast Cancer, R packages, LYMP5 Collaborative Project, Hepatocellular carcinoma metastasis and prognosis, SAGE breast cancer challenge, Drug synergy predictors, RNAi screens of Breast Cancer (MCF10A) with Driver Mutations, Tamoxifen resistance in ESR1+ breast carcinoma, Subtype-specific addiction points in breast carcinoma	manager
	Antonella Galli	ag2994@columbia.edu	Stem Cell Interactomes	collaborator

Figure 13-5 Collaborative web system: users-groups

The screenshot shows a user interface for a group dashboard. At the top, the group name "GC-resistance in T-ALL" is displayed. Below this, there is a navigation bar with "Dashboard" selected. The main content area is divided into two columns. The left column, titled "Recent activity", lists several posts with timestamps and descriptions, such as "Deep Sequencing of pooled shRNA screen on TALL lines with DBZ, DEXA treatment" and "paper discussion". The right column features a "Calendar" for August 2012 and a section for "Upcoming events". The interface includes a search bar, a "Settings" button, and a "Create content" button.

Figure 13-6 Collaborative web system: group dashboard – home page

13.2.3 User Roles

This system is customized to have different roles for different levels of permission to control the management and collaboration (Figure 13-7). It supports the following roles:

- **Manager:** manage all projects and all members
- **Current Member:** create groups, access their own and all public content, access view profiles
- **Collaborator:** only access their groups and public content, cannot create group and view group directory, cannot view other people's profiles
- **Intern or Rotation Student:** similar with collaborator except they can view group directory
- **Past Member:** similar with collaborator.

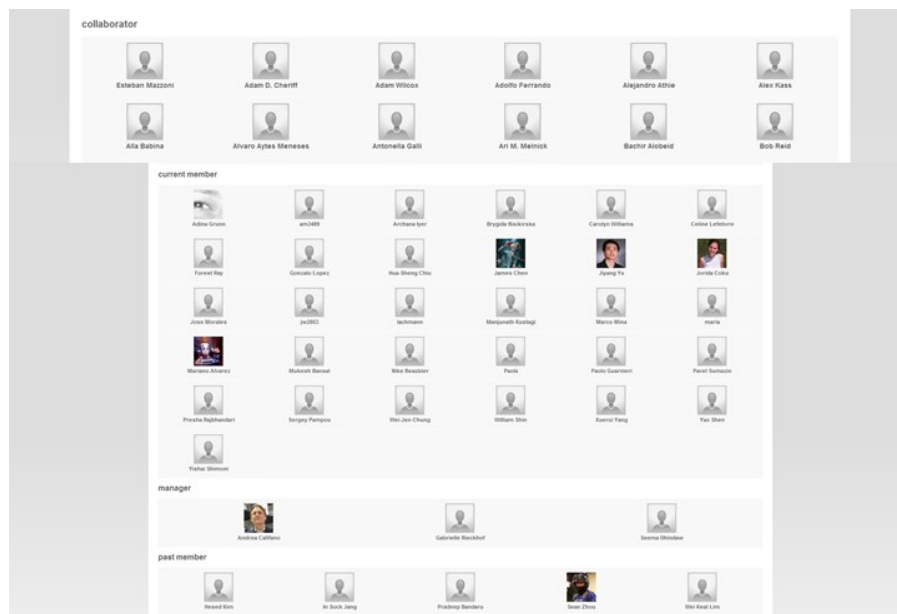


Figure 13-7 Collaborative web system: users in different roles

13.2.4 Document / Wiki

The most commonly used content type to share and communicate within a group is document or wiki page (Figure 13-8, Figure 13-9). A document could be

- a web page for project description, software usage, useful resources
- a collection of presentation slides (in the attachment)
- a manuscript by many authors.

The document content type has the following features:

- Organized by the format of book (Figure 13-11)
- Wiki: editable by any people in the group
- revision control: revert, show diff (Figure 13-10).

The screenshot displays a web interface for a group named 'GC-resistance in T-ALL'. The top navigation bar includes 'Home > Groups > GC-resistance in T-ALL', a search bar, and user information for 'Jiyang Yu'. Below the navigation bar, there are icons for various content types and a 'Documents' tab. The main content area is divided into two sections: 'Overview' and 'Archived'. The 'Overview' section contains a table of documents with the following data:

Updated	Title	Changed by
9:02pm Mon Jun 11	Deep Sequencing of pooled shRNA screen on TALL lines with DBZ, DEXA treatment	Jiyang Yu
12:51pm Tue Dec 20	Reply to Cancer Cell Reviewers on AKT paper	Jiyang Yu
1:27pm Fri Nov 18	computational strategy to integrate shRNA screening with genomic analysis	Jiyang Yu
4:07pm Fri Jun 17	Results Apoptosis Assay on KOPTK1 of all TF MRs	Jiyang Yu
12:39pm Fri Jun 17	viability assays by silencing TF MRs and GC-treatment	Jiyang Yu
3:44pm Wed Jun 8	full results of two cell lines at gene level by HM	Jiyang Yu

The right sidebar, titled 'Documents', lists the following document titles:

- Deep Sequencing of pooled shRNA screen on TALL lines with DBZ, DEXA treatment
- GC-Resistant and GC-sensitive pre-treated T-ALL patient GEP data
- GC-Sensitive T-ALL patient data: DE genes and Master Regulators
- GR regulatory pathway
- group meeting presentation collection
- MIRs identified by GSEA with two sets (positive and negative) for each MIR
- NOTCH data
- Overlapped GC-Responsible and GC-Responsive (6h or 24h) MRs
- pooled shRNA data

Figure 13-8 Collaborative web system: document list in a group

Home > Groups > Lab General Activities

My groups Jiyang Yu Need help?

Lab General Activities

Documents

Settings Create content

View Edit Revisions Reference this Subscriptions Add Document

Lab Meeting Rotation List & Schedule (a reference to check presenters)

Jiyang Yu 10:51am Tue Mar 23

Tagged: lab meeting rotation list

Highlight changes

By Gabrielle:

Roughly every week we will have 8 assigned presentations according to the order below. There will always be open slots for people to present who are not on the rotation for that week.

To summarize:

1. Everyone is on the list regardless of whether they are doing theoretical and/or applied work.
2. It is the expectation that everyone whose turn it is, will present something even if it is a brief update.
3. If you cannot present for any reason, it is your responsibility to find a replacement, and then switch into your replacement's presentation slot.

Documents

- Lab Meeting Rotation List & Schedule (a reference to check presenters)
- Andrea's 50th Birthday Party
- Cosma Update 0701
- Cosma Update 07122011
- Handbook of this Web System
- HTS report 09/03/2010
- Lab Meeting Presentation 04/06/2010
- Lab Meeting Presentation 4-26-2010
- Lab Meeting Presentation 4-6-2010
- lab retreat
- lab retreat
- Lab Retreat - Sean
- Lab retreat + lab meeting

Figure 13-9 Collaborative web system: a document

Lab General Activities

Documents

Settings Create content

View Edit Revisions Reference this Subscriptions

Revision	Show diff	Operations
Jul 26, 2012 by Seema Dhindaw	<input type="radio"/>	<input checked="" type="radio"/> current revision
Jul 23, 2012 by Seema Dhindaw	<input checked="" type="radio"/>	<input type="radio"/>
Jul 23, 2012 by Seema Dhindaw	<input type="radio"/>	<input type="radio"/>
Jul 17, 2012 by Seema Dhindaw	<input type="radio"/>	<input type="radio"/>
Jul 17, 2012 by Seema Dhindaw	<input type="radio"/>	<input type="radio"/>

Figure 13-10 Collaborative web system: revision history of a document

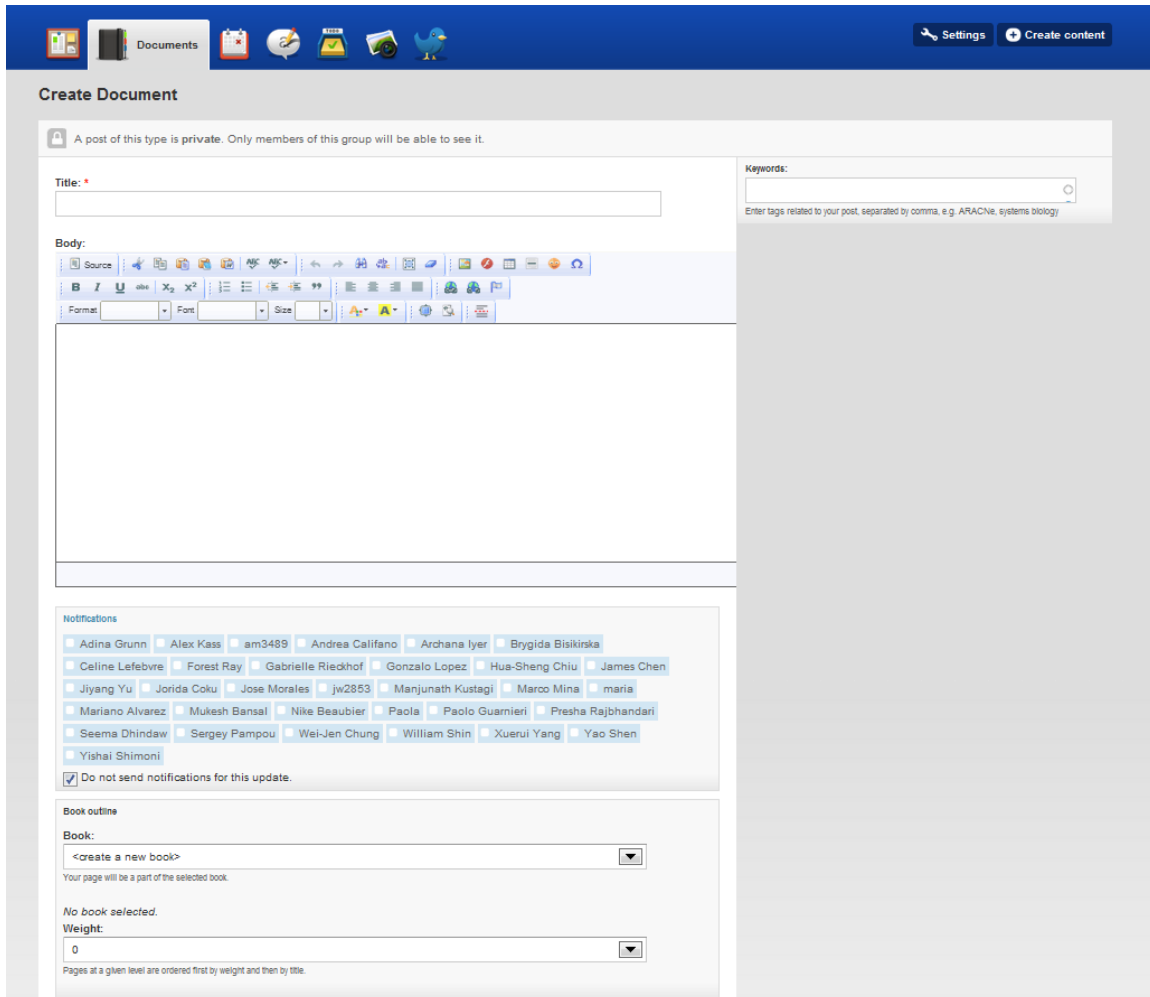



Figure 13-11 Collaborative web system: document creating page





13.2.5 Calendar Event

The calendar event is used for manage and record lab meetings or meetings with collaborators (Figure 13-12). An event could be

- a lab meeting
- a seminar or talk you would like to share with people
- a conference outside.

Home My groups  Need help?

Califano Lab, Columbia University Search

    Create content

Calendar Upcoming ICal Feeds

May 2012 Previous Next






Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
30	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

BSEA: Bayesian Set Enrichment Analysis (Jiyang)
3:00pm - 4:00pm
Wed May 30, 2012

BSEA: Bayesian Set

Upcoming events

GC-resistance in T-ALL Search

     Settings Create content

Calendar Upcoming ICal Feeds + Add Event







November 2011 Previous Next

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
21	22	23	24	25	26	27
28	29	30	1	2	3	4

GC-resistance
11:00am - 12:00pm
Wed Nov 16, 2011


Upcoming events

Information/Journal Club Search

      Create content

View Edit Reference this Subscriptions + Add Event

BSEA: Bayesian Set Enrichment Analysis (Jiyang)

 Jiyang Yu 1:01pm Wed May 30

Tagged: BSEA enrichment Jiyang

When: 3:00pm - 4:00pm Wed May 30, 2012

Where: Conference Room 913, 9th floor, KRC Bldg, 1130 St. Nicholas Ave

[Add new comment](#) [Print](#)

May 2012

M	T	W	T	F	S	S
30	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

Upcoming events

Developed and powered by Jiyang Yu

Figure 13-12 Collaborative web system: calendar and event

13.2.6 Blog

A blog entry (Figure 13-13) is generally used for

- a discussion for anything about the project or group
- a review or comment on interesting papers
- a further discussion about a seminar talk or a presentation
- any suggestion to your group or the lab
- anything about your life and research.

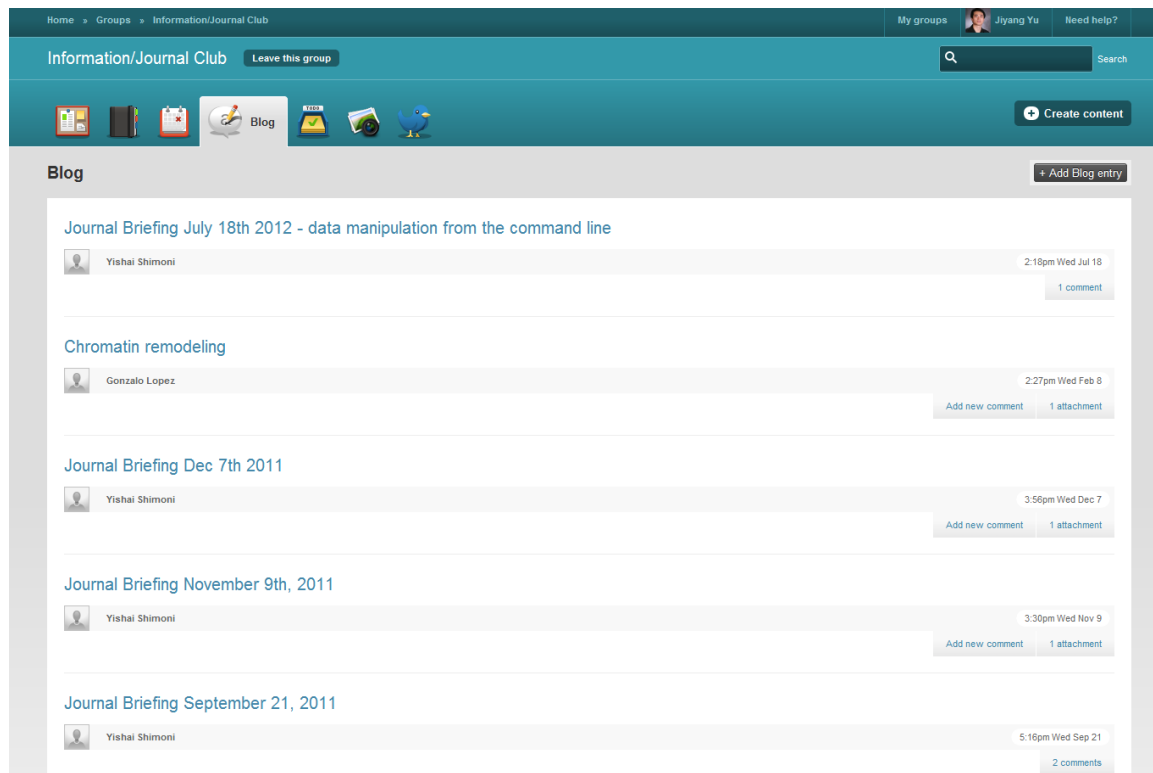


Figure 13-13 Collaborative web system: blog in a group

13.2.7 Project and case tracker

You can create projects within a group, assign it to any group members and keep track of the project status using the feature of “Case Tracker” (Figure 13-14). The status of a case can be Open, Resolved, Deferred, Duplicated, Closed. The priority can be defined as High, Normal, or Low. You can also define case types.

The screenshot displays the 'Case Tracker' interface within a web application. The top navigation bar includes 'Lab General Activities' and 'Case Tracker' tabs. The main content area is divided into two sections: a list of cases and a detailed view of a selected case.

Case List:

Priority	Title	Assigned	Status	Last post
High	Open access to all Califano lab members	Jiyang Yu	Open	1:44pm Mon Feb 22
High	change william's status to member	Jiyang Yu	Resolved	8:41pm Tue Mar 23
High	group all users by their roles so that group admin can select all users with one role like current_user	Jiyang Yu	Resolved	7:22pm Thu Mar 11
High	a bug when creating new gallery with keywords	Jiyang Yu	Resolved	7:22pm Thu Mar 11

Case Detail View:

Title: a bug when creating new gallery with keywords
Status: Resolved
Priority: Normal
Type: Bug
Last updated: 2 years 24 weeks ago

Comments:

9:41pm Mon Jan 18
 Jiyang Yu
 Web System issues assigned to Jiyang Yu

7:22pm Thu Mar 11
 Jiyang Yu
 Status: Open → Resolved
 solved

Form for new comment:

Assign to: Jiyang Yu
 Status: Resolved
 Priority: Normal
 Type: Bug

Figure 13-14 Collaborative web system: project case tracker

13.2.8 Shoutbox / Twitter

The shoutbox feature (Figure 13-15) is like the twitter or facebook status function to share a short message, a link, a comment, a word about your mood, etc.

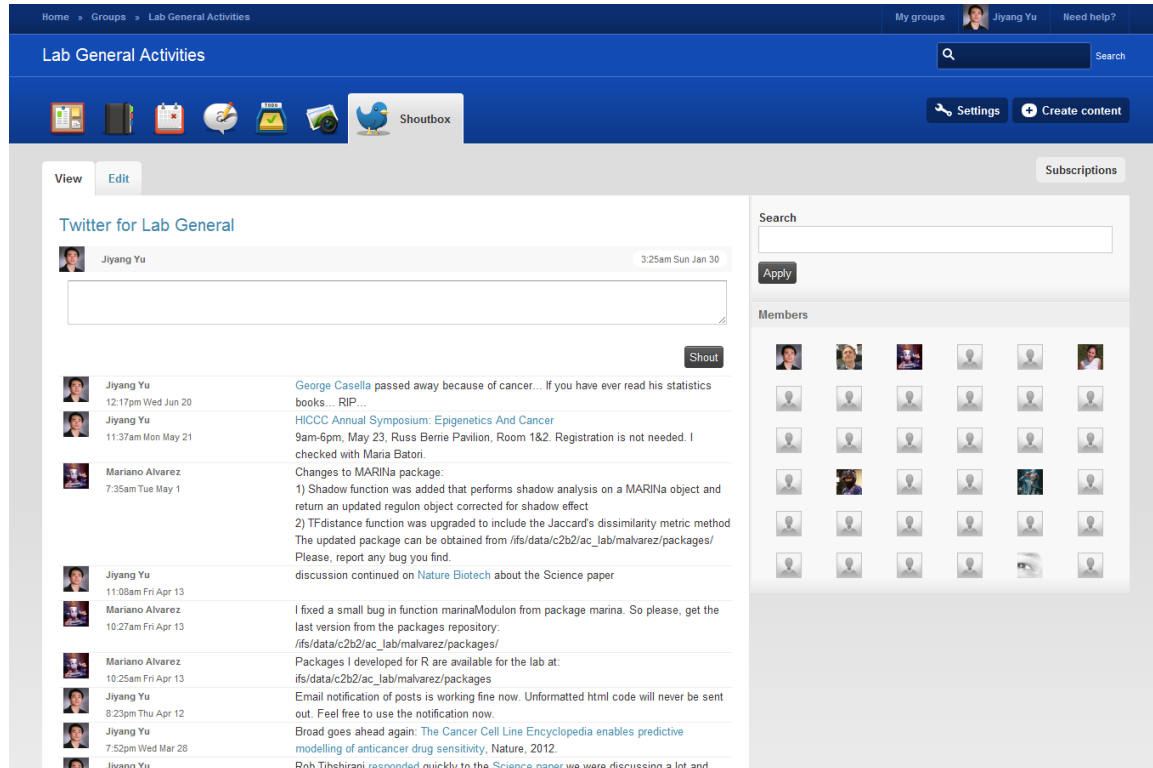


Figure 13-15 Collaborative web system: shoutbox or twitter in a group

13.2.9 Images

Image feature is used to share photos or images organized in galleries (Figure 13-16).

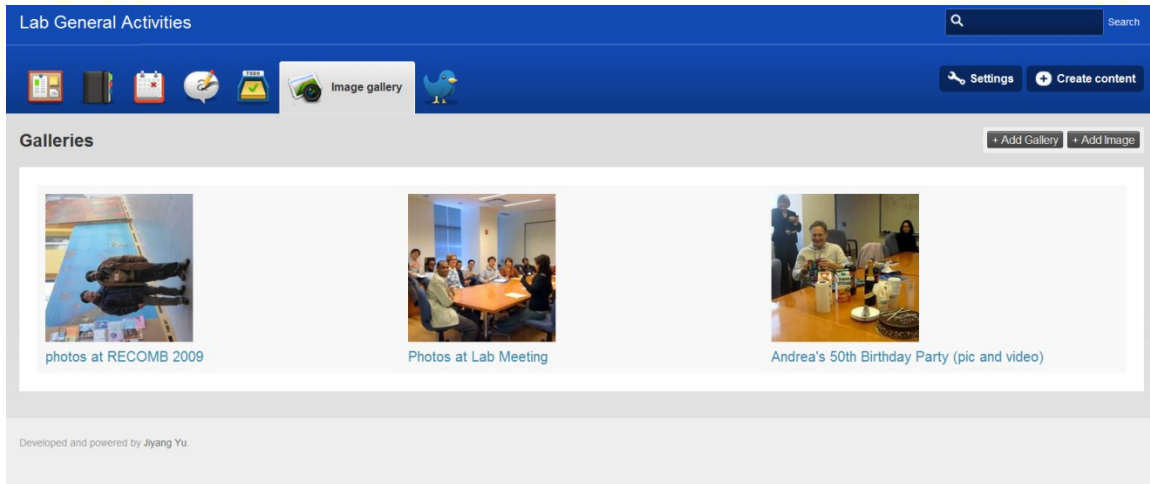


Figure 13-16 Collaborative web system: image or photo collection

13.2.10 Notification and others

Notification feature allows the user to send out email notification when a content entry is created or updated. The user can choose subscribe to all content type in the group or to author so that the notification will be automatically sent out.

There are many other web 2.0 or web 3.0 features that this system supports, such as personalization, in which you can customize background color, logo, layout of your group and customize your personal page including background color, avatar, profiles, editor, etc. The content in the system is context-dependent, which only shows related links or tabs when you are view a page. It also supports WYSIWYG (What You See Is What You Get) html editor for easier content writing.

13.3 Insights into Biological Systems from Software Systems

This website I created for lab collaboration management is a software engineering system which is built by different elements on different layers. Actually from the architecture of this engineering system, we might gain some insights into reverse-engineering and better understanding the nature or biological system (Figure 13-17). For example, the data content in this web system is similar to DNA, RNA or protein in the cells of biological system; the underlying modules in the web system are like the signaling transduction,

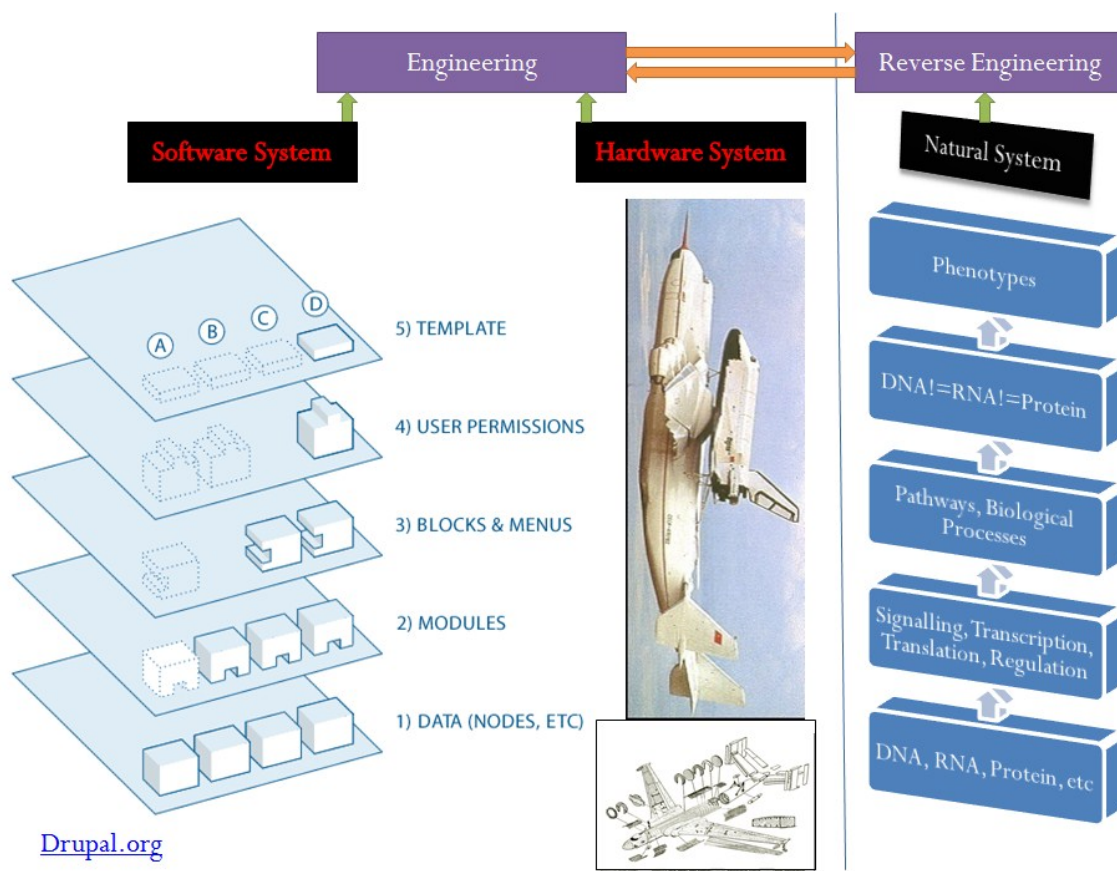


Figure 13-17 An analogy among software system, hardware system and biological natural system.

transcription, translation, regulation, and other basic and fundamental units in the living cells; blocks or menus that are built on those underlying modules in the web system are the functional pathways or biological processes that are formed by basic units; also in the web system different users or units have various permissions to function normally and similarly in the cellular system, different players have specific rules to function well; finally the beautiful web styles and themes we view are like the phenotypes of biological system. The “bottom-up” construction framework to engineer a functional software system helps us on “up-down” reverse-engineering of biological system. Additionally, the improvement of modules in current system on reusability, scalability and robustness also shed light on how to decipher underlying pathways in biological system (Figure 13-18).

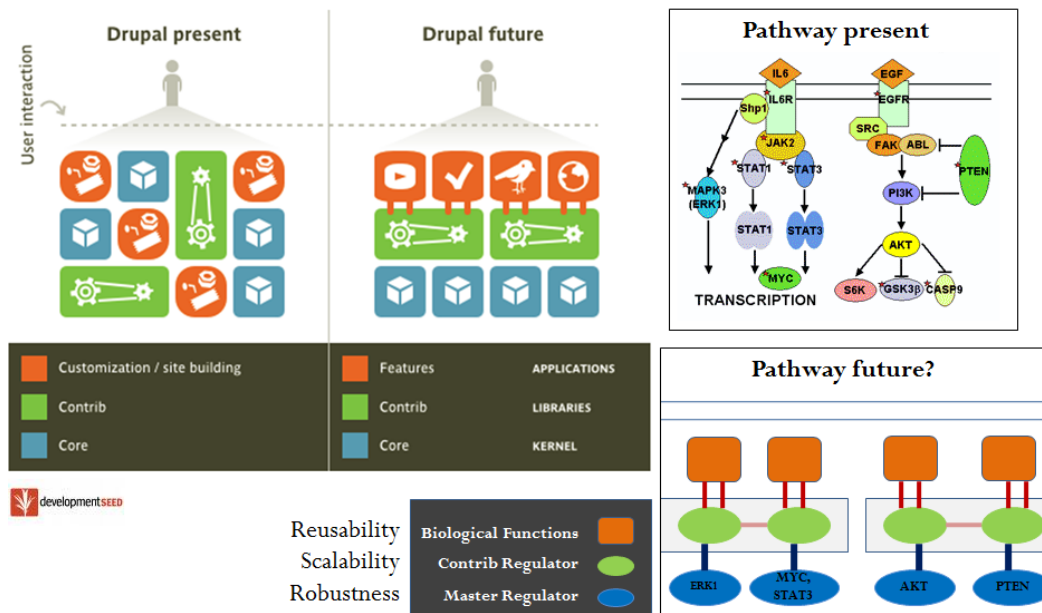


Figure 13-18 Insights from Drupal future to Pathway future: reusability, scalability and robustness.

Chapter 14 Conclusions

14.1 Key Contributions and Findings

In this dissertation, I have showed you the power of integrating functional genome-wide RNAi screens with systems biology of cancer genomics to discover driver-type therapeutic targets for reversal of drug-resistance or treatment of aggressive human tumors.

14.1.1 NGS-based shRNA screening (shSeq): an analytical pipeline

For high-throughput RNAi screening, I have focused on NGS-based pooled shRNA screening (shSeq). The shSeq technology is new and there are no established tools to analyze such new data. I have developed a computational pipeline with a series of algorithms and software packages to deconvolute, QC and post-analyze shSeq data as detailed in Chapter 2. Specifically, the ShortRead+ package is used to do QA of raw NGS data; the shScanner package is for deconvolution of shSeq raw reads; shSEQ package is to annotate, normalize and do a secondary QC of shRNA abundance profiles; the method of shADER is for differential representation analysis at individual shRNA level, and the package of shMA/BHM is at gene level to identify potential depleted or enriched hits or candidates.

In particular, I have developed a novel “Modeling-All-Together” strategy using Bayesian hierarchical modeling (BHM) approach as shown in Chapter 3 to

integrate multiple shRNAs targeting the same gene and estimate gene level activity. I have demonstrated that this algorithm consistently outperforms existing “Separate-And-Combine” methods such as RIGER, RSA, etc, especially when the data is in low-quality.

14.1.2 Systems biology of cancer genomics: NetBID2

In parallel, for systems biology of cancer genomics, I have developed a computational algorithm, Network-based Bayesian Inference of Disease Drivers (NetBID2), to infer regulatory or signaling drivers of diseases from high-throughput transcriptomic data or gene expression profiles as detailed in Chapter 4. NetBID2 is based on reverse-engineering cellular network and Bayesian inference. I have demonstrated that NetBID2 is more robust than conventional expression signature analysis; it is able to detect not only known drivers of human cancers but also, more importantly, the hidden drivers that traditional methods fail to find. Furthermore, experimental validation results have showed high prediction rates (>75%) of NetBID2.

Additionally, for a key step of NetBID2 framework, I have developed a new enrichment analysis algorithm, Bayesian Set Enrichment Analysis (BSEA), as described in Chapter 5. BSEA uses “maxmean” statistic to summarize enrichment score but is under Bayesian framework. According to evaluation results, BSEA consistently outperforms existing GSEA and GSA methods.

14.1.3 Successful studies of integrating functional genomics with systems biology for driver-type therapeutic target discovery

By integrating functional RNAi screens with NetBID2, I have identified known and novel driver-type therapeutic targets in various disease contexts. For example, in Chapter 7, I have discovered that AKT1 is a driver for glucocorticoid (GC) resistance, a significant clinical problem in the treatment of T-ALL. We have validated, both biochemically and pharmacologically, that the inhibition of AKT1 is able to reverse GC-resistance in T-ALL. Additionally, integration of systems biology predictions with shRNA screens (Chapter 8) identified 16 master regulators of GC resistance, out of which 13 have been validated to significantly overcome resistance upon silencing, and more surprisingly, 10 have showed stronger effects than positive controls to sensitize GC-resistant T-ALL cells.

In breast cancer, I have discovered that STAT3 is required for transformation of HER2+ breast cancer, an aggressive breast tumor subtype (Chapter 9). The suppression of STAT3 has been confirmed *in vitro* and *in vivo* to be an effective therapy for HER2+ breast cancer. Moreover, my analysis has revealed that STAT3 silencing has a co-dependence on ER-.

With my integrative framework, I have also identified potential therapeutic targets for ABC or GCB-type DLBCL (Chapter 10) and subtype-based breast cancer (Chapter 11) that are currently being validated.

14.1.4 Collaboration model between computational and experimental biologists

This dissertation has demonstrated a perfect collaboration model between computational and experimental biologists. As a computational biologist, I have been interacting well with my biologist collaborators and have achieved a few successful interdisciplinary stories as shown in previous section. Additionally, I have been collaborating with over ten biological labs (Chapter 12) to apply my computational framework of shSeq technology for therapeutic target discovery or genetic modifier identification. Besides, I have developed a user-friendly and dynamic web system to manage collaborative projects (Chapter 13), which has helped to facilitate and speed up the communication between computational biologists and experimental biologist collaborators.

14.2 Future Directions

In this dissertation, I have shown that genome-wide RNAi screening technology, especially NGS-based pooled shRNA screening (shSeq), is indeed a powerful tool for therapeutic target discovery. However, there is much space to improve this technology and analysis of shSeq data. First, design of high quality shRNA library is much needed due to low silencing efficiency and off-target effects of a significant number of hairpin constructs in current libraries. A recent study [61] of enumerating RNAi performance of all possible hairpins sequences using a tiled sensor assay approach provides an opportunity to learn and develop an algorithm of optimal shRNA design based on sequence features of targeting

genes, therefore yielding a quality-improved and coverage-enlarged shRNA library. Second, with a design algorithm of shRNA construct, we can predict the knock-down efficiency or quality of existing hairpins in current library and filter or weight them when estimating gene level activity by integrating all shRNAs targeting the same gene. Third, the deconvolution of shSeq raw reads can be improved, in both time and space efficiency, by using optimal data structure such as suffix array or suffix tree. Additionally, Poisson distribution is currently used to model shSeq count data in differential representation analysis; however, negative binomial distribution might be a better try as widely used in RNA-Seq data analysis. Besides, it's challenging to extend the biological models of shRNA screening from 2D to 3D or in vivo systems, which are more close to the true living system.

In this dissertation, we mainly focus on transcriptomic data or gene expression profiles, and use them to integrate with RNAi screening data. However, there are many other types of cancer genomic data such as copy number variations, SNPs, epigenomic, proteomic profiles and microRNA expression data, which are widely-available and might be mined to cross with functional RNAi screens for discovery of underlying genetic mutation causes or dependence or translational and regulatory modifiers.

Additionally, large amounts of context-specific small-molecule screens provide valuable resources and information for study of mechanism of actions for small molecules. Therefore, integration of small-molecule screens with functional RNAi

screens would be able to identify drugs or compounds targeting RNAi screening identified therapeutic targets in specific tumor contexts, which will boost up cancer drug discovery.

References

1. Vaidyanathan, G., *Redefining clinical trials: the age of personalized medicine*. Cell, 2012. **148**(6): p. 1079-80.
2. Ng, P.C., et al., *An agenda for personalized medicine*. Nature, 2009. **461**(7265): p. 724-6.
3. Kaiser, J., *Personalized medicine. New cystic fibrosis drug offers hope, at a price*. Science, 2012. **335**(6069): p. 645.
4. Leek, J.T., R.D. Peng, and R.R. Anderson, *Personalized medicine: Keep a way open for tailored treatments*. Nature, 2012. **484**(7394): p. 318.
5. Pennisi, E., *Genomic medicine. Gene sequence study takes a stab at personalized medicine*. Science, 2005. **308**(5725): p. 1102.
6. Liebman, M., *Personalized medicine: a perspective on the patient, disease and causal diagnostics*. Personalized Medicine, 2007. **4**(2): p. 171-174.
7. Collins, F., *Francis Collins interview. Departing U.S. genome institute director takes stock of personalized medicine. Interview by Jocelyn Kaiser*. Science, 2008. **320**(5881): p. 1272.
8. Virgin, H.W. and J.A. Todd, *Metagenomics and personalized medicine*. Cell, 2011. **147**(1): p. 44-56.
9. Lyon, G.J., *Personalized medicine: Bring clinical standards to human-genetics research*. Nature, 2012. **482**(7385): p. 300-1.
10. Schreiber, S.L., et al., *Towards patient-based cancer therapeutics*. Nat Biotechnol, 2010. **28**(9): p. 904-6.
11. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer*. Cell, 2000. **100**(1): p. 57-70.
12. Shay, J.W. and I.B. Roninson, *Hallmarks of senescence in carcinogenesis and cancer therapy*. Oncogene, 2004. **23**(16): p. 2919-33.
13. Hanahan, D. and R.A. Weinberg, *Hallmarks of cancer: the next generation*. Cell, 2011. **144**(5): p. 646-74.
14. van't Veer, L.J. and R. Bernards, *Enabling personalized cancer medicine through analysis of gene-expression patterns*. Nature, 2008. **452**(7187): p. 564-70.

15. Jones, B., *Toxicity after cervical cancer treatment using radiotherapy and chemotherapy*. Clin Oncol (R Coll Radiol), 2009. **21**(1): p. 56-63.
16. Steingart, R., *Mechanisms of late cardiovascular toxicity from cancer chemotherapy*. J Clin Oncol, 2005. **23**(36): p. 9051-2.
17. Kelemen, G., et al., *Long-term efficiency and toxicity of adjuvant dose-dense sequential adriamycin-Paclitaxel-cyclophosphamide chemotherapy in high-risk breast cancer*. Oncology, 2010. **78**(3-4): p. 271-3.
18. Kaspers, G.J., et al., *Glucocorticoid resistance in childhood leukemia*. Leuk Lymphoma, 1994. **13**(3-4): p. 187-201.
19. Tissing, W.J., et al., *Molecular determinants of glucocorticoid sensitivity and resistance in acute lymphoblastic leukemia*. Leukemia, 2003. **17**(1): p. 17-25.
20. Catts, V.S., et al., *High level resistance to glucocorticoids, associated with a dysfunctional glucocorticoid receptor, in childhood acute lymphoblastic leukemia cells selected for methotrexate resistance*. Leukemia, 2001. **15**(6): p. 929-35.
21. Bachmann, P.S., et al., *Dexamethasone resistance in B-cell precursor childhood acute lymphoblastic leukemia occurs downstream of ligand-induced nuclear translocation of the glucocorticoid receptor*. Blood, 2005. **105**(6): p. 2519-26.
22. Tissing, W.J., et al., *Glucocorticoid-induced glucocorticoid-receptor expression and promoter usage is not linked to glucocorticoid resistance in childhood ALL*. Blood, 2006. **108**(3): p. 1045-9.
23. Wei, G., et al., *Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance*. Cancer Cell, 2006. **10**(4): p. 331-42.
24. Real, P.J., et al., *Gamma-secretase inhibitors reverse glucocorticoid resistance in T cell acute lymphoblastic leukemia*. Nat Med, 2009. **15**(1): p. 50-8.
25. Gong, J.G., et al., *The tyrosine kinase c-Abl regulates p73 in apoptotic response to cisplatin-induced DNA damage*. Nature, 1999. **399**(6738): p. 806-9.
26. Ohndorf, U.M., et al., *Basis for recognition of cisplatin-modified DNA by high-mobility-group proteins*. Nature, 1999. **399**(6737): p. 708-12.
27. Sakai, W., et al., *Secondary mutations as a mechanism of cisplatin resistance in BRCA2-mutated cancers*. Nature, 2008. **451**(7182): p. 1116-20.
28. Turchi, J.J., *Nitric oxide and cisplatin resistance: NO easy answers*. Proc Natl Acad Sci U S A, 2006. **103**(12): p. 4337-8.

29. Scanlon, K.J., et al., *Cisplatin resistance in human cancers*. *Pharmacol Ther*, 1991. **52**(3): p. 385-406.
30. Itamochi, H., et al., *Mechanisms of cisplatin resistance in clear cell carcinoma of the ovary*. *Oncology*, 2002. **62**(4): p. 349-53.
31. Siddik, Z.H., *Cisplatin: mode of cytotoxic action and molecular basis of resistance*. *Oncogene*, 2003. **22**(47): p. 7265-79.
32. Ibanez de Caceres, I., et al., *IGFBP-3 hypermethylation-derived deficiency mediates cisplatin resistance in non-small-cell lung cancer*. *Oncogene*, 2010. **29**(11): p. 1681-90.
33. Galluzzi, L., et al., *Molecular mechanisms of cisplatin resistance*. *Oncogene*, 2012. **31**(15): p. 1869-83.
34. Coussens, L., et al., *Tyrosine kinase receptor with extensive homology to EGF receptor shares chromosomal location with neu oncogene*. *Science*, 1985. **230**(4730): p. 1132-9.
35. Jackman, D.M., et al., *Impact of epidermal growth factor receptor and KRAS mutations on clinical outcomes in previously untreated non-small cell lung cancer patients: results of an online tumor registry of clinical trials*. *Clin Cancer Res*, 2009. **15**(16): p. 5267-73.
36. Weng, A.P., et al., *Activating mutations of NOTCH1 in human T cell acute lymphoblastic leukemia*. *Science*, 2004. **306**(5694): p. 269-71.
37. Grabher, C., H. von Boehmer, and A.T. Look, *Notch 1 activation in the molecular pathogenesis of T-cell acute lymphoblastic leukaemia*. *Nat Rev Cancer*, 2006. **6**(5): p. 347-59.
38. Le, X.F., F. Pruefer, and R.C. Bast, Jr., *HER2-targeting antibodies modulate the cyclin-dependent kinase inhibitor p27Kip1 via multiple signaling pathways*. *Cell Cycle*, 2005. **4**(1): p. 87-95.
39. Zhang, H., et al., *ErbB receptors: from oncogenes to targeted cancer therapies*. *J Clin Invest*, 2007. **117**(8): p. 2051-8.
40. Palomero, T. and A. Ferrando, *Therapeutic targeting of NOTCH1 signaling in T-cell acute lymphoblastic leukemia*. *Clin Lymphoma Myeloma*, 2009. **9 Suppl 3**: p. S205-10.
41. Pui, C.H., *T cell acute lymphoblastic leukemia: NOTCHing the way toward a better treatment outcome*. *Cancer Cell*, 2009. **15**(2): p. 85-7.
42. Saglio, G., et al., *Rational approaches to the design of therapeutics targeting molecular markers: the case of chronic myelogenous leukemia*. *Ann N Y Acad Sci*, 2004. **1028**: p. 423-31.

43. Lan, K.H., C.H. Lu, and D.H. Yu, *Mechanisms of trastuzumab resistance and their clinical implications*. Tumor Progression and Therapeutic Resistance, 2005. **1059**: p. 70-75.
44. Burmer, G.C. and L.A. Loeb, *Mutations in the KRAS2 oncogene during progressive stages of human colon carcinoma*. Proc Natl Acad Sci U S A, 1989. **86**(7): p. 2403-7.
45. Almoguera, C., et al., *Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes*. Cell, 1988. **53**(4): p. 549-54.
46. Tam, I.Y., et al., *Distinct epidermal growth factor receptor and KRAS mutation patterns in non-small cell lung cancer patients with different tobacco exposure and clinicopathologic features*. Clin Cancer Res, 2006. **12**(5): p. 1647-53.
47. Soucek, L., et al., *Modelling Myc inhibition as a cancer therapy*. Nature, 2008. **455**(7213): p. 679-83.
48. Cheung, H.W., et al., *Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer*. Proc Natl Acad Sci U S A, 2011.
49. Hammond, S.M., et al., *An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells*. Nature, 2000. **404**(6775): p. 293-6.
50. Silva, J.M., et al., *Cyfp1 is a putative invasion suppressor in epithelial cancers*. Cell, 2009. **137**(6): p. 1047-61.
51. Silva, J.M., et al., *Profiling essential genes in human mammary cells by multiplex RNAi screening*. Science, 2008. **319**(5863): p. 617-20.
52. Silva, J.M., et al., *Second-generation shRNA libraries covering the mouse and human genomes*. Nat Genet, 2005. **37**(11): p. 1281-8.
53. Silva, J.M., et al., *RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells*. Proc Natl Acad Sci U S A, 2004. **101**(17): p. 6548-52.
54. Paddison, P.J., et al., *A resource for large-scale RNA-interference-based screens in mammals*. Nature, 2004. **428**(6981): p. 427-31.
55. Moffat, J., et al., *A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen*. Cell, 2006. **124**(6): p. 1283-98.
56. Schlabach, M.R., et al., *Cancer proliferation gene discovery through functional Genomics*. Science, 2008. **319**(5863): p. 620-624.

57. Berns, K., et al., *A large-scale RNAi screen in human cells identifies new components of the p53 pathway*. Nature, 2004. **428**(6981): p. 431-7.
58. Brummelkamp, T.R., et al., *An shRNA barcode screen provides insight into cancer cell vulnerability to MDM2 inhibitors*. Nature Chemical Biology, 2006. **2**(4): p. 202-206.
59. Luo, B., et al., *Highly parallel identification of essential genes in cancer cells*. Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(51): p. 20380-20385.
60. Possemato, R., et al., *Functional genomics reveal that the serine synthesis pathway is essential in breast cancer*. Nature, 2011. **476**(7360): p. 346-50.
61. Fellmann, C., et al., *Functional Identification of Optimized RNAi Triggers Using a Massively Parallel Sensor Assay*. Molecular cell, 2011. **41**(6): p. 733-746.
62. Bassik, M.C., et al., *Rapid creation and quantitative monitoring of high coverage shRNA libraries*. Nat Methods, 2009. **6**(6): p. 443-5.
63. Burgess, D.J., et al., *Topoisomerase levels determine chemotherapy response in vitro and in vivo*. Proc Natl Acad Sci U S A, 2008. **105**(26): p. 9053-8.
64. The Cancer Genome Atlas (TCGA) : <https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>.
65. The International Cancer Genome Consortium (ICGC): <http://www.icgc.org/>.
66. Therapeutically Applicable Research to Generate Effective Treatments (TARGET) : <http://target.cancer.gov/>.
67. The Cancer Cell Line Encyclopedia (CCLE) : <http://www.broadinstitute.org/ccle/home>.
68. The Connectivity Map (CMAP) : <http://www.broadinstitute.org/cmap/>.
69. Lim, W.K., E. Lyashenko, and A. Califano, *Master regulators used as breast cancer metastasis classifier*. Pac Symp Biocomput, 2009: p. 504-15.
70. Basso, K., et al., *Reverse engineering of regulatory networks in human B cells*. Nat Genet, 2005. **37**(4): p. 382-90.
71. Lefebvre, C., et al., *A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers*. Mol Syst Biol, 2010. **6**: p. 377.

72. Carro, M.S., et al., *The transcriptional network for mesenchymal transformation of brain tumours*. Nature, 2010. **463**(7279): p. 318-25.
73. Wang, K., et al., *Genome-wide identification of post-translational modulators of transcription factor activity in human B cells*. Nat Biotechnol, 2009. **27**(9): p. 829-39.
74. Zuber, J., et al., *RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia*. Nature, 2011.
75. Hahn, W.C., et al., *Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1*. Nature, 2009. **462**(7269): p. 108-U122.
76. Hahn, C.K., et al., *Proteomic and Genetic Approaches Identify Syk as an AML Target*. Cancer Cell, 2009. **16**(4): p. 281-294.
77. Tyner, J.W., et al., *RNAi screening of the tyrosine kinome identifies therapeutic targets in acute myeloid leukemia*. Blood, 2008. **111**(4): p. 2238-45.
78. Ngo, V.N., et al., *A loss-of-function RNA interference screen for molecular targets in cancer*. Nature, 2006. **441**(7089): p. 106-10.
79. Kolfschoten, I.G.M., et al., *A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity*. Cell, 2005. **121**(6): p. 849-858.
80. Yu, J., et al., *Computational Analysis of High-Throughput RNAi Screening Data (in print)*, in *Methods in Molecular Medicine*. 2012.
81. Lattice: <http://lmdvr.r-forge.r-project.org/figures/figures.html>.
82. Wickham, H., ggplot2: <http://had.co.nz/ggplot2/>.
83. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment*. Genome Res, 1998. **8**(3): p. 175-85.
84. Columbia C2B2 Titan Cluster: http://wiki.c2b2.columbia.edu/systems/index.php/Documentation/Titan_cluster.
85. Kauffmann, A., R. Gentleman, and W. Huber, *arrayQualityMetrics--a bioconductor package for quality assessment of microarray data*. Bioinformatics, 2009. **25**(3): p. 415-6.
86. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes*. Bioinformatics, 2001. **17**(6): p. 509-19.

87. Tibshirani, R., <http://www-stat.stanford.edu/~tibs/SAM/>.
88. Diboun, I., et al., *Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma*. *Bmc Genomics*, 2006. **7**: p. 252.
89. Gelman, A., et al., *Bayesian Data Analysis*. 2nd edition ed. Texts in Statistical Science, ed. C. Chatfield, M. Tanner, and J. Zidek. 2004: Chapman & Hall.
90. Gelman, A., et al., *A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models*. *Annals of Applied Statistics*, 2008. **2**(4): p. 1360-1383.
91. Zellner, A., *On assessing prior distributions and Bayesian regression analysis with g-prior distributions*, in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, P.K.G.a.A. Zellner, Editor. 1986. p. 233-243.
92. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing*. *Journal of the Royal Statistical Society Series B-Methodological*, 1995. **57**(1): p. 289-300.
93. *The Gene Ontology*: <http://www.geneontology.org/>.
94. *Pathway Commons*: <http://www.pathwaycommons.org>.
95. *MSiDB*:
<http://www.broadinstitute.org/gsea/msigdb/index.jsp>.
96. *DAVID*: <http://david.abcc.ncifcrf.gov/>.
97. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. *Proc Natl Acad Sci U S A*, 2005. **102**(43): p. 15545-50.
98. Efron, B. and R. Tibshirani, *On Testing the Significance of Sets of Genes*. *Annals of Applied Statistics*, 2007. **1**(1): p. 107-129.
99. Konig, R., et al., *A probability-based approach for the analysis of large-scale RNAi screens*. *Nat Methods*, 2007. **4**(10): p. 847-9.
100. Gelman, A. and J. HILL, *Data Analysis using Regression and Multilevel/Hierarchical Models*. 2007: Cambridge University Press.
101. Ji, H.K. and X.S. Liu, *Analyzing 'omics data using hierarchical models*. *Nature Biotechnology*, 2010. **28**(4): p. 337-340.

102. Marcotte, R., et al., *Essential Gene Profiles in Breast, Pancreatic, and Ovarian Cancer Cells*. *Cancer Discovery*, 2012. **2**(2): p. 172-189.
103. Tu, Z.D., et al., *Further understanding human disease genes by comparing with housekeeping genes and other genes*. *Bmc Genomics*, 2006. **7**.
104. Chang, C.W., et al., *Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis*. *PLoS One*, 2011. **6**(7): p. e22859.
105. Gelman, A., J. Hill, and M. Yajima, *Why we (usually) don't have to worry about multiple comparisons*. Technical Report, 2011.
106. Fisher, R.A., *Notes on Combining Idependent Tests of Significance*. 1948.
107. Stouffer, S.A., et al., *Adjustment During Army Life*. Vol. 1. 1949, Princeton: Princeton University Press.
108. Sims, D., et al., *High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing*. *Genome Biology*, 2011. **12**(10): p. R104.
109. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer*. *Nature*, 2002. **415**(6871): p. 530-6.
110. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*. *Lancet*, 2005. **365**(9460): p. 671-9.
111. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*. *BMC Bioinformatics*, 2006. **7 Suppl 1**: p. S7.
112. Alon, U., *An Introduction to Systems Biology: Design Principles of Biological Circuits*. 2006: Chapman & Hall/CRC Mathematical & Computational Biology.
113. Belcastro, V., et al., *Transcriptional gene network inference from a massive dataset elucidates transcriptome organization and gene function*. *Nucleic Acids Res*, 2011. **39**(20): p. 8677-88.
114. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges*. *PLoS Comput Biol*, 2012. **8**(2): p. e1002375.
115. Draghici, S., et al., *A systems biology approach for pathway level analysis*. *Genome Res*, 2007. **17**(10): p. 1537-45.
116. Lamb, J., *The Connectivity Map: a new tool for biomedical research*. *Nat Rev Cancer*, 2007. **7**(1): p. 54-60.

117. Lamb, J., et al., *The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease*. Science, 2006. **313**(5795): p. 1929-35.
118. Lander, E.S., *Array of hope*. Nat Genet, 1999. **21**(1 Suppl): p. 3-4.
119. Sellers, W.R. and D.E. Fisher, *Apoptosis and cancer drug targeting*. J Clin Invest, 1999. **104**(12): p. 1655-61.
120. Adams, J.M. and S. Cory, *The Bcl-2 apoptotic switch in cancer development and therapy*. Oncogene, 2007. **26**(9): p. 1324-37.
121. Adams, J.M., *Ways of dying: multiple pathways to apoptosis*. Genes Dev, 2003. **17**(20): p. 2481-95.
122. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays*. Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.
123. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
124. Ashburner, M., et al., *Gene ontology: tool for the unification of biology*. The Gene Ontology Consortium. Nat Genet, 2000. **25**(1): p. 25-9.
125. Pe'er, D., *Bayesian network analysis of signaling networks: a primer*. Sci STKE, 2005. **2005**(281): p. pl4.
126. Friedman, N., *Inferring cellular networks using probabilistic graphical models*. Science, 2004. **303**(5659): p. 799-805.
127. Pe'er, D., et al., *Inferring subnetworks from perturbed expression profiles*. Bioinformatics, 2001. **17 Suppl 1**: p. S215-24.
128. Friedman, N., et al., *Using Bayesian networks to analyze expression data*. Journal of Computational Biology, 2000. **7**(3-4): p. 601-620.
129. Ellis, B. and W.H. Wong, *Learning causal Bayesian network structures from experimental data*. Journal of the American Statistical Association, 2008. **103**(482): p. 778-789.
130. Bøttcher, S., *Learning Bayesian networks with mixed variables*. In Proceedings of the Eighth International Workshop in Artificial Intelligence and Statistics, 2001.
131. Geiger, D. and D. Heckerman, *Learning Gaussian networks*. Technical Report MSRTR-94-10. Microsoft Research, 1994.

132. Heckerman, D., D. Geiger, and D. Chickering, *Learning Bayesian networks: The combination of knowledge and statistical data*. Machine Learning, 1995.
133. Chickering, D., *Learning Bayesian networks is NP-complete.*, in *Learning from Data: Artificial Intelligence and Statistics*, D.F. V and H.-J. Lenz, Editors. 1996, Springer-Verlag: New York.
134. Bøttcher, S. and C. Dethlefsen, *deal: A Package for Learning Bayesian Networks*. Journal of Statistical Software, 2003a. **8(20)**: p. 1-40.
135. Bøttcher, S. and C. Dethlefsen, *Learning Bayesian networks with R, in Proceedings of the 3rd international workshop on distributed statistical computing*. K. Hornik, F. Leisch and A. Zeileis Eds. ISSN 1609-395X. 2003b.
136. Efron, B. and R. Tibshirani, *An Introduction to the Bootstrap*. 1993, London: Chapman & Hall.
137. Friedman, N., M. Goldszmidt, and A. Wyner, *Data analysis with Bayesian networks: A bootstrap approach*. In *Proceedings of the Fifteenth Annual Conference on uncertainty in Artificial Intelligence pp. 206 - 215 (Morgan Kaufmann, San Francisco)*. 1999.
138. Wang, C.Y., M.W. Mayo, and A.S. Baldwin, Jr., *TNF- and cancer therapy-induced apoptosis: potentiation by inhibition of NF-kappaB*. Science, 1996. **274(5288)**: p. 784-7.
139. <http://www.ncbi.nlm.nih.gov/gene/8887>.
140. Beyaert, R., et al., *Functional redundancy of the zinc fingers of A20 for inhibition of NF-kappa B activation and protein-protein interactions*. Febs Letters, 2001. **498(1)**: p. 93-97.
141. Beyaert, R., et al., *The zinc finger protein A20 interacts with a novel anti-apoptotic protein which is cleaved by specific caspases*. Oncogene, 1999. **18(29)**: p. 4182-4190.
142. Shembade, N., A. Ma, and E.W. Harhaj, *Inhibition of NF-kappa B Signaling by A20 Through Disruption of Ubiquitin Enzyme Complexes*. Science, 2010. **327(5969)**: p. 1135-1139.
143. <http://www.uniprot.org/uniprot/Q13794>.
144. <http://thebiogrid.org/115335/summary/homo-sapiens/bcl2l11.html>.
145. <http://thebiogrid.org/111379/summary/homo-sapiens/pmaip1.html>.

146. Villunger, A., et al., *BH3-only proteins Puma and Bim are rate-limiting for gamma-radiation- and glucocorticoid-induced apoptosis of lymphoid cells in vivo*. *Blood*, 2005. **106**(13): p. 4131-4138.
147. Villunger, A., et al., *p53- and drug-induced apoptotic responses mediated by BH3-only proteins Puma and Noxa*. *Science*, 2003. **302**(5647): p. 1036-1038.
148. Piovan, E., et al., *Reversal of glucocorticoid resistance by AKT inhibition in T-ALL*. *Nature* (In review), 2012.
149. Pui, C.H., M.V. Relling, and J.R. Downing, *Acute lymphoblastic leukemia*. *N Engl J Med*, 2004. **350**(15): p. 1535-48.
150. Dordelmann, M., et al., *Prednisone response is the strongest predictor of treatment outcome in infant acute lymphoblastic leukemia*. *Blood*, 1999. **94**(4): p. 1209-1217.
151. Schrappe, M., et al., *Philadelphia chromosome-positive (Ph+) childhood acute lymphoblastic leukemia: good initial steroid response allows early prediction of a favorable treatment outcome*. *Blood*, 1998. **92**(8): p. 2730-41.
152. Schrappe, M., et al., *Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: results of trial ALL-BFM 90. German-Austrian-Swiss ALL-BFM Study Group*. *Blood*, 2000. **95**(11): p. 3310-22.
153. Klumper, E., et al., *In vitro cellular drug resistance in children with relapsed/refractory acute lymphoblastic leukemia*. *Blood*, 1995. **86**(10): p. 3861-8.
154. Hongo, T., et al., *In vitro drug sensitivity testing can predict induction failure and early relapse of childhood acute lymphoblastic leukemia*. *Blood*, 1997. **89**(8): p. 2959-2965.
155. Kaspers, G.J., et al., *Immunophenotypic cell lineage and in vitro cellular drug resistance in childhood relapsed acute lymphoblastic leukaemia*. *Eur J Cancer*, 2005. **41**(9): p. 1300-3.
156. Ashwell, J.D., F.W. Lu, and M.S. Vacchio, *Glucocorticoids in T cell development and function**. *Annu Rev Immunol*, 2000. **18**: p. 309-45.
157. Golonzhka, O., et al., *Ctip2/Bcl11b controls ameloblast formation during mammalian odontogenesis*. *Proc Natl Acad Sci U S A*, 2009. **106**(11): p. 4278-83.
158. Schmidt, S., et al., *Glucocorticoid-induced apoptosis and glucocorticoid resistance: molecular mechanisms and clinical relevance*. *Cell Death Differ*, 2004. **11 Suppl 1**: p. S45-55.
159. Surjit, M., et al., *Widespread negative response elements mediate direct repression by agonist- liganded glucocorticoid receptor*. *Cell*. **145**(2): p. 224-41.

160. Obexer, P., et al., *Expression profiling of glucocorticoid-treated T-ALL cell lines: rapid repression of multiple genes involved in RNA-, protein- and nucleotide synthesis*. *Oncogene*, 2001. **20**(32): p. 4324-36.
161. Tissing, W.J., et al., *Genomewide identification of prednisolone-responsive genes in acute lymphoblastic leukemia cells*. *Blood*, 2007. **109**(9): p. 3929-35.
162. Tonko, M., et al., *Gene expression profiles of proliferating vs. G1/G0 arrested human leukemia cells suggest a mechanism for glucocorticoid-induced apoptosis*. *FASEB J*, 2001. **15**(3): p. 693-9.
163. Wang, Z., et al., *Microarray analysis uncovers the induction of the proapoptotic BH3-only protein Bim in multiple models of glucocorticoid-induced apoptosis*. *J Biol Chem*, 2003. **278**(26): p. 23861-7.
164. Hillmann, A.G., et al., *Glucocorticoid receptor gene mutations in leukemic cells acquired in vitro and in vivo*. *Cancer Res*, 2000. **60**(7): p. 2056-62.
165. Irving, J.A., et al., *Loss of heterozygosity and somatic mutations of the glucocorticoid receptor gene are rarely found at relapse in pediatric acute lymphoblastic leukemia but may occur in a subpopulation early in the disease course*. *Cancer Res*, 2005. **65**(21): p. 9712-8.
166. Tissing, W.J., et al., *Expression of the glucocorticoid receptor and its isoforms in relation to glucocorticoid resistance in childhood acute lymphocytic leukemia*. *Haematologica*, 2005. **90**(9): p. 1279-81.
167. Eisen, L.P., M.S. Elsasser, and J.M. Harmon, *Positive regulation of the glucocorticoid receptor in human T-cells sensitive to the cytolytic effects of glucocorticoids*. *J Biol Chem*, 1988. **263**(24): p. 12044-8.
168. Ramdas, J., W. Liu, and J.M. Harmon, *Glucocorticoid-induced cell death requires autoinduction of glucocorticoid receptor expression in human leukemic T cells*. *Cancer Res*, 1999. **59**(6): p. 1378-85.
169. Levine, E.G., et al., *Glucocorticoid receptors in chronic lymphocytic leukemia*. *Leuk Res*, 1985. **9**(8): p. 993-9.
170. Leventhal, B.G., *Glucocorticoid receptors in lymphoid tumors*. *Cancer Res*, 1981. **41**(11 Pt 2): p. 4861-2.
171. Pedersen, K.B. and W.V. Vedeckis, *Quantification and glucocorticoid regulation of glucocorticoid receptor transcripts in two human leukemic cell lines*. *Biochemistry*, 2003. **42**(37): p. 10978-90.
172. Pedersen, K.B., C.D. Geng, and W.V. Vedeckis, *Three mechanisms are involved in glucocorticoid receptor autoregulation in a human T-lymphoblast cell line*. *Biochemistry*, 2004. **43**(34): p. 10851-8.

173. Holleman, A., et al., *Gene-expression patterns in drug-resistant acute lymphoblastic leukemia cells and response to treatment*. N Engl J Med, 2004. **351**(6): p. 533-42.
174. Tsapis, M., et al., *HDAC inhibitors induce apoptosis in glucocorticoid-resistant acute lymphatic leukemia cells despite a switch from the extrinsic to the intrinsic death pathway*. Int J Biochem Cell Biol, 2007. **39**(7-8): p. 1500-9.
175. Gu, L., et al., *Rapamycin reverses NPM-ALK-induced glucocorticoid resistance in lymphoid tumor cells by inhibiting mTOR signaling pathway, enhancing G1 cell cycle arrest and apoptosis*. Leukemia, 2008. **22**(11): p. 2091-6.
176. Rambal, A.A., et al., *MEK inhibitors potentiate dexamethasone lethality in acute lymphoblastic leukemia cells through the pro-apoptotic molecule BIM*. Leukemia, 2009. **23**(10): p. 1744-1754.
177. Stam, R.W., et al., *Association of high-level MCL-1 expression with in vitro and in vivo prednisone resistance in MLL-rearranged infant acute lymphoblastic leukemia*. Blood, 2010. **115**(5): p. 1018-1025.
178. Dong, H., et al., *Inhibition of PDE3, PDE4 and PDE7 potentiates glucocorticoid-induced apoptosis and overcomes glucocorticoid resistance in CEM T leukemic cells*. Biochemical Pharmacology, 2010. **79**(3): p. 321-9.
179. Gutierrez, A. and A.T. Look, *Molecular Targeted Therapies in T-Cell Acute Lymphoblastic Leukemia*, in *Molecularly Targeted Therapy for Childhood Cancer*, P.J. Houghton and R.J. Arceci, Editors. 2010, Springer New York. p. 19-30.
180. Margolin, A.A., et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context*. BMC Bioinformatics, 2006. **7 Suppl 1**: p. S1-7.
181. Palomero, T., et al., *NOTCH1 directly regulates c-MYC and activates a feed-forward-loop transcriptional network promoting leukemic cell growth*. Proc Natl Acad Sci U S A, 2006. **103**(48): p. 18261-6.
182. Palomero, T., et al., *Mutational loss of PTEN induces resistance to NOTCH1 inhibition in T-cell leukemia*. Nat Med, 2007. **13**(10): p. 1203-10.
183. Gutierrez, A., et al., *High frequency of PTEN, PI3K, and AKT abnormalities in T-cell acute lymphoblastic leukemia*. Blood, 2009. **114**(3): p. 647-50.
184. Heitzer, M.D., et al., *Glucocorticoid receptor physiology*. Rev Endocr Metab Disord, 2007. **8**(4): p. 321-30.
185. Geng, C.D. and W.V. Veddeckis, *c-Myb and members of the c-Ets family of transcription factors act as molecular switches to mediate opposite steroid*

regulation of the human glucocorticoid receptor 1A promoter. J Biol Chem, 2005. **280**(52): p. 43264-71.

186. Mok, C.L., et al., *Bad can act as a key regulator of T cell apoptosis and T cell development.* J Exp Med, 1999. **189**(3): p. 575-86.

187. Zimmermann, S. and K. Moelling, *Phosphorylation and regulation of Raf by Akt (protein kinase B).* Science, 1999. **286**(5445): p. 1741-4.

188. Ozes, O.N., et al., *NF-kappaB activation by tumour necrosis factor requires the Akt serine-threonine kinase.* Nature, 1999. **401**(6748): p. 82-5.

189. Hirai, H., et al., *MK-2206, an allosteric Akt inhibitor, enhances antitumor efficacy by standard chemotherapeutic agents or molecular targeted drugs in vitro and in vivo.* Mol Cancer Ther. **9**(7): p. 1956-67.

190. Armstrong, F., et al., *NOTCH is a key regulator of human T-cell acute leukemia initiating cell activity.* Blood, 2009. **113**(8): p. 1730-40.

191. Chiang, M.Y., et al., *Leukemia-associated NOTCH1 alleles are weak tumor initiators but accelerate K-ras-initiated leukemia.* J Clin Invest, 2008. **118**(9): p. 3181-94.

192. Geley, S., et al., *Resistance to glucocorticoid-induced apoptosis in human T-cell acute lymphoblastic leukemia CEM-C1 cells is due to insufficient glucocorticoid receptor expression.* Cancer Res, 1996. **56**(21): p. 5033-8.

193. Real, P.J. and A.A. Ferrando, *NOTCH inhibition and glucocorticoid therapy in T-cell acute lymphoblastic leukemia.* Leukemia, 2009.

194. Bachmann, P.S., et al., *Epigenetic silencing of BIM in glucocorticoid poor-responsive pediatric acute lymphoblastic leukemia, and its reversal by histone deacetylase inhibition.* Blood. **116**(16): p. 3013-22.

195. Spokoini, R., et al., *Glycogen synthase kinase-3 plays a central role in mediating glucocorticoid-induced apoptosis.* Mol Endocrinol. **24**(6): p. 1136-50.

196. Datta, S.R., et al., *Akt phosphorylation of BAD couples survival signals to the cell-intrinsic death machinery.* Cell, 1997. **91**(2): p. 231-41.

197. del Peso, L., et al., *Interleukin-3-induced phosphorylation of BAD through the protein kinase Akt.* Science, 1997. **278**(5338): p. 687-9.

198. Cross, D.A., et al., *Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B.* Nature, 1995. **378**(6559): p. 785-9.

199. Bornhauser, B.C., et al., *Low-dose arsenic trioxide sensitizes glucocorticoid-resistant acute lymphoblastic leukemia cells to dexamethasone via an Akt-dependent pathway.* Blood, 2007. **110**(6): p. 2084-91.

200. Faus, H. and B. Haendler, *Post-translational modifications of steroid receptors*. Biomed Pharmacother, 2006. **60**(9): p. 520-8.
201. Tran, H., et al., *The many forks in FOXO's road*. Sci STKE, 2003. **2003**(172): p. RE5.
202. Gelman, A., et al., *A weakly informative default prior distribution for logistic and other regression models*. Ann Appl Stat, 2008. **2**(4): p. 1360-1383.
203. Stuart, J.M., et al., *A gene-coexpression network for global discovery of conserved genetic modules*. Science, 2003. **302**(5643): p. 249-55.
204. Carro, M.S., et al., *The transcriptional network for mesenchymal transformation of brain tumours*. Nature. **463**(7279): p. 318-25.
205. Efron, B. and R. Tibshirani, *On testing the significance of sets of genes*. The Annals of Applied Statistics, 2007. **1**(1): p. 107-129.
206. Armstrong, F., et al., *NOTCH is a key regulator of human T-cell acute leukemia initiating cell activity*. Blood, 2009. **113**(8): p. 1730-40.
207. Kelly, E., et al., *IL-2 and related cytokines can promote T cell survival by activating AKT*. J Immunol, 2002. **168**(2): p. 597-603.
208. Ferrando, A.A., et al., *Gene expression signatures in MLL-rearranged T-lineage and B-precursor acute leukemias: dominance of HOX dysregulation*. Blood, 2003. **102**(1): p. 262-8.
209. Pear, W.S., et al., *Exclusive development of T cell neoplasms in mice transplanted with bone marrow expressing activated Notch alleles*. J Exp Med, 1996. **183**(5): p. 2283-91.
210. Guo, K., et al., *Disruption of peripheral leptin signaling in mice results in hyperleptinemia without associated metabolic abnormalities*. Endocrinology, 2007.
211. Trotman, L.C., et al., *Pten dose dictates cancer progression in the prostate*. PLoS Biol, 2003. **1**(3): p. E59.
212. Chou, T.C. and P. Talalay, *Quantitative analysis of dose-effect relationships: the combined effects of multiple drugs or enzyme inhibitors*. Adv Enzyme Regul, 1984. **22**: p. 27-55.
213. Schmidt, S., et al., *Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia*. Blood, 2006. **107**(5): p. 2061-2069.
214. Kaspers, G.J.L., et al., *Immunophenotypic cell lineage and in vitro cellular drug resistance in childhood relapsed acute lymphoblastic leukaemia*. European Journal of Cancer, 2005. **41**(9): p. 1300-1303.

215. Leventhal, B.G., *Glucocorticoid Receptors in Lymphoid Tumors*. Cancer Research, 1981. **41**(11): p. 4861-4862.
216. Levine, E.G., et al., *Glucocorticoid Receptors in Chronic Lymphocytic-Leukemia*. Leukemia Research, 1985. **9**(8): p. 993-999.
217. Eisen, L.P., M.S. Elsassser, and J.M. Harmon, *Positive Regulation of the Glucocorticoid Receptor in Human T-Cells Sensitive to the Cytolytic Effects of Glucocorticoids*. Journal of Biological Chemistry, 1988. **263**(24): p. 12044-12048.
218. Holleman, A., et al., *Gene-Expression Patterns in Drug-Resistant Acute Lymphoblastic Leukemia Cells and Response to Treatment*. New England Journal of Medicine, 2004. **351**(6): p. 533-542.
219. Irving, J.A.E., et al., *Loss of heterozygosity and somatic mutations of the Glucocorticoid receptor gene are rarely found at relapse in pediatric acute lymphoblastic leukemia but may occur in a subpopulation early in the disease course*. Cancer Research, 2005. **65**(21): p. 9712-9718.
220. Van Vlierberghe, P., et al., *The recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia*. Blood, 2008. **111**(9): p. 4668-4680.
221. Iyengar, S. and P.J. Farnham, *KAP1 protein: an enigmatic master regulator of the genome*. J Biol Chem, 2011. **286**(30): p. 26267-76.
222. Chang, C.J., Y.L. Chen, and S.C. Lee, *Coactivator TIF1beta interacts with transcription factor C/EBPbeta and glucocorticoid receptor to induce alpha1-acid glycoprotein gene expression*. Mol Cell Biol, 1998. **18**(10): p. 5880-7.
223. Pottier, N., et al., *The SWI/SNF chromatin-remodeling complex and glucocorticoid resistance in acute lymphoblastic leukemia*. J Natl Cancer Inst, 2008. **100**(24): p. 1792-803.
224. Slamon, D.J., et al., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene*. Science, 1987. **235**(4785): p. 177-82.
225. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups*. Nature, 2012. **advance online publication**.
226. Wang, Y.C., et al., *Different mechanisms for resistance to trastuzumab versus lapatinib in HER2-positive breast cancers--role of estrogen receptor and HER2 reactivation*. Breast Cancer Res, 2011. **13**(6): p. R121.
227. Lan, K.H., C.H. Lu, and D. Yu, *Mechanisms of trastuzumab resistance and their clinical implications*. Ann N Y Acad Sci, 2005. **1059**: p. 70-5.

228. Nagata, Y., et al., *PTEN activation contributes to tumor inhibition by trastuzumab, and loss of PTEN predicts trastuzumab resistance in patients*. *Cancer Cell*, 2004. **6**(2): p. 117-27.
229. Korkaya, H., et al., *Activation of an IL6 Inflammatory Loop Mediates Trastuzumab Resistance in HER2+ Breast Cancer by Expanding the Cancer Stem Cell Population*. *Molecular cell*, 2012. **47**(4): p. 570-584.
230. Hartman, Z.C., et al., *HER2 overexpression elicits a proinflammatory IL-6 autocrine signaling loop that is critical for tumorigenesis*. *Cancer Res*, 2011. **71**(13): p. 4380-91.
231. Ginestier, C., et al., *CXCR1 blockade selectively targets human breast cancer stem cells in vitro and in xenografts*. *Journal of Clinical Investigation*, 2010. **120**(2): p. 485-497.
232. Iliopoulos, D., et al., *Inducible formation of breast cancer stem cells and their dynamic equilibrium with non-stem cancer cells via IL6 secretion*. *Proceedings of the National Academy of Sciences of the United States of America*, 2011. **108**(4): p. 1397-1402.
233. Deng, J., et al., *S1PR1-STAT3 Signaling Is Crucial for Myeloid Cell Colonization at Future Metastatic Sites*. *Cancer Cell*, 2012. **21**(5): p. 642-654.
234. Hedvat, M., et al., *The JAK2 inhibitor AZD1480 potently blocks Stat3 signaling and oncogenesis in solid tumors*. *Cancer Cell*, 2009. **16**(6): p. 487-97.
235. Lee, H., et al., *Persistently activated Stat3 maintains constitutive NF-kappaB activity in tumors*. *Cancer Cell*, 2009. **15**(4): p. 283-93.
236. Kortylewski, M., R. Jove, and H. Yu, *Targeting STAT3 affects melanoma on multiple fronts*. *Cancer Metastasis Rev*, 2005. **24**(2): p. 315-27.
237. Scuto, A., et al., *STAT3 inhibition is a therapeutic strategy for ABC-like diffuse large B-cell lymphoma*. *Cancer Res*, 2011. **71**(9): p. 3182-8.
238. Bromberg, J.F., et al., *Stat3 as an oncogene*. *Cell*, 1999. **98**(3): p. 295-303.
239. Marotta, L.L.C., et al., *The JAK2/STAT3 signaling pathway is required for growth of CD44(+)CD24(-) stem cell-like breast cancer cells in human tumors*. *Journal of Clinical Investigation*, 2011. **121**(7): p. 2723-2735.
240. Lee, H., et al., *STAT3-induced S1PR1 expression is crucial for persistent STAT3 activation in tumors*. *Nat Med*, 2010. **16**(12): p. 1421-8.
241. Yu, H. and R. Jove, *The STATs of cancer--new molecular targets come of age*. *Nat Rev Cancer*, 2004. **4**(2): p. 97-105.

242. Garcia, R., et al., *Constitutive activation of Stat3 by the Src and JAK tyrosine kinases participates in growth regulation of human breast carcinoma cells*. *Oncogene*, 2001. **20**(20): p. 2499-513.
243. Yuan, Z.L., et al., *Central role of the threonine residue within the p+1 loop of receptor tyrosine kinase in STAT3 constitutive phosphorylation in metastatic cancer cells*. *Mol Cell Biol*, 2004. **24**(21): p. 9390-400.
244. Huber, W., et al., *Variance stabilization applied to microarray data calibration and to the quantification of differential expression*. *Bioinformatics*, 2002. **18 Suppl 1**: p. S96-104.
245. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray*. *Bioinformatics*, 2008. **24**(13): p. 1547-8.
246. Korkaya, H., et al., *Activation of an IL6 Inflammatory Loop Mediates Trastuzumab Resistance in HER2+ Breast Cancer by Expanding the Cancer Stem Cell Population*. *Molecular cell*, 2012(0).
247. Heiser, L.M., et al., *Subtype and pathway specific responses to anticancer compounds in breast cancer*. *Proc Natl Acad Sci U S A*, 2011.
248. DrugBank: <http://www.drugbank.ca/>.
249. Drupal: <http://drupal.org/>.
250. OptenAtrium: <http://openatrium.com>.

Appendix A: High-throughput RNAi Screening: Experimental Approach⁴

1.1 Introduction

RNA interference (RNAi), a cellular process that regulates gene expression, has emerged as a powerful genetic venue to functionally interrogate the entire genome by loss-of-function studies. Small regulatory RNAs (siRNAs and miRNAs) bind to the enzymatic RNAi machinery and suppress the expression level of targeting mRNAs [2]. This process can be experimentally controlled to knock down the expression of any specific gene.

Response to transfected siRNAs is transient, from 3 to 7 days, making this approach unsuitable for the analysis of silencing long-term effects. The search for a sustained silencing response has resulted in the development of a class of RNAi triggers denominated short hairpin RNAs (shRNAs). Plasmids expressing shRNAs are integrated into the cell genome, so that by continuously supplying the RNAi trigger, stable gene silencing can be achieved [3].

Several groups have previously described the construction of shRNA libraries that cover a significant fraction of all human genes [4-7]. In this chapter, we use

⁴ This chapter is co-authored by my collaborator Ruth Rodriguez-Barrueco from Jose Silva lab, based on her book chapter [1].

Thermo Scientific Open Biosystems GIPZ Lentiviral Human shRNAmir Library as a model to illustrate the accomplishment of genome-wide RNAi screening. The library is composed of 58,493 hairpin constructs, in which 39,458 shRNAs are known to target 18,661 human genes, about 75% of the human genome. These shRNA-mirs are modeled after endogenous miRNAs, specifically contained in the backbone of the primary miR-30 microRNA [6]. Additionally, targeting sequences were selected, by a mathematical algorithm, to fit thermodynamic asymmetry rules. Overall, these shRNA libraries offer a convenient, flexible, and highly effective tool for studying gene function in human cells.

There are two main approaches for screening large collections of silencing triggers for their effects on mammalian cells: cell-based assay and pooled screening. In the first method, individual siRNA or shRNA is transfected and screened in a multi-well format for the activation or repression of a reporter [8, 9]. In this format, individual genes are transiently suppressed 'one-by-one' and analysis is carried out in a high-throughput manner using a robotic platform. However, this method is expensive, labor intensive, and time consuming. On the other hand, pooled screening allows the ability to analyze the effects of the whole library at once. If, for example, a gene targeted by a particular shRNA induces response to a growth inhibitor stimulus, then its representation should increase after treatment. If a given shRNA sensitized a population to a specific stress, then, the relative abundance this shRNA should diminish after the stress. Using RNAi libraries in a pooled fashion provides the opportunity to investigate the

entire genome in loss-of-function studies and to find genes relevant to any biological process.

The RNAi library can be used to perform either positive or negative screening (Figure 1). In a positive screen, cells that survive a selection pressure or that show a differentiating phenotype are selected. The specific shRNA leading to the recognizable cell population can be sequenced from the genomic DNA of isolated colonies or analyzed in a high throughput fashion. When performing a negative screen, populations that are sensitive to a selection pressure or that show an impaired growth are selected. In this case, the integrated shRNA will be depleted in the final population; microarray hybridization or next generation sequencing (NGS) technologies can be used to read out hairpin abundance. Careful design of the experimental pipeline is essential to answer the biological question. In the following sections, a protocol for conducting both positive and negative screens is described.

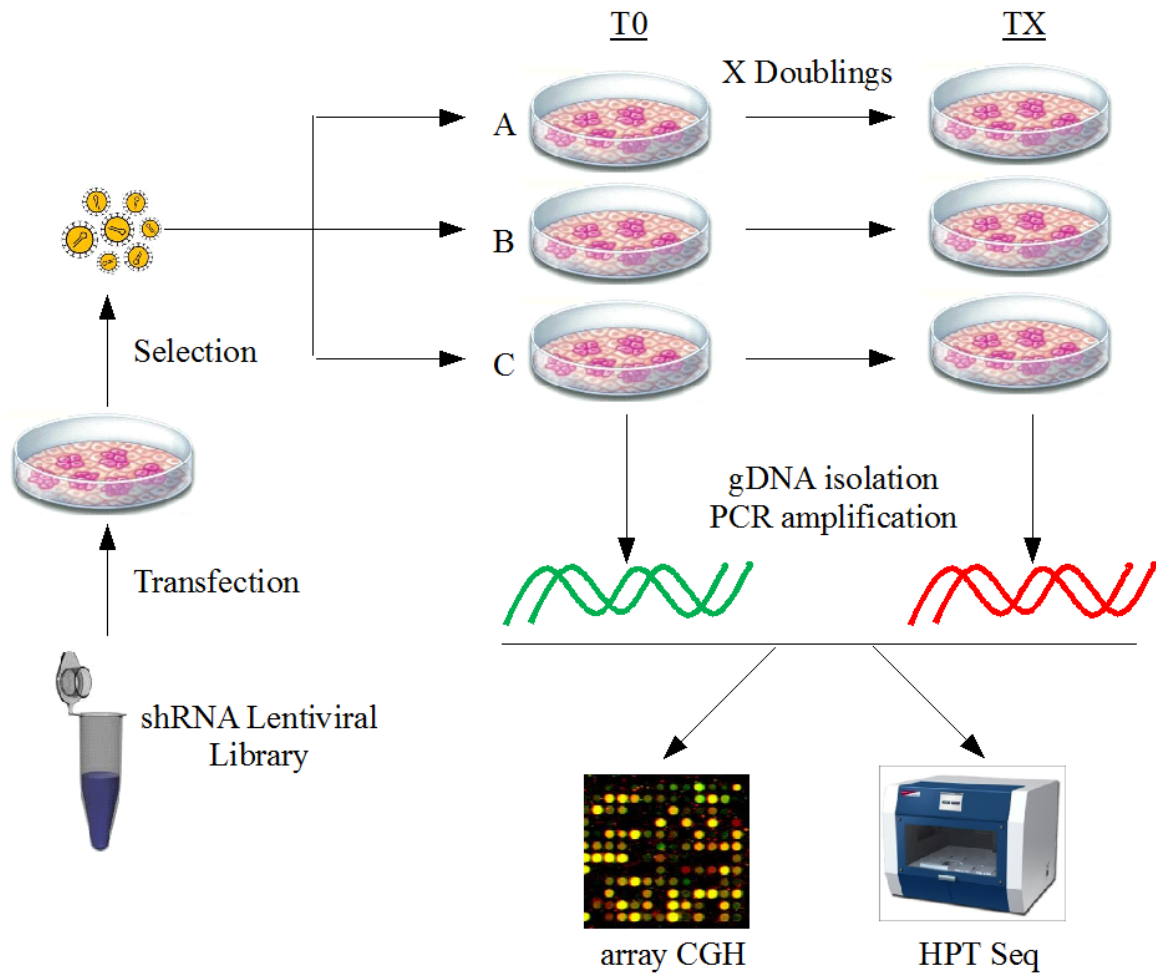


Figure 1 Procedure of a shRNA screen: transduce shRNA lentiviral library into cells of study with a MOI of ~ 0.3 , filter out uninfected cells, grow cells for X doubling times, extract genomic DNA from T0 and TX, PCR amplify them, and then measure the shRNA abundance either by microarray or next generation sequencing.

1.2 Materials

1.2.1 shRNA Library

The library pool consists of 58,493 shRNAs integrated into the backbone of miR30 and cloned into the pGIPZ lentiviral vector (Open Biosystems GIPZ Lentiviral Human shRNA Library). It is known that 39,458 of these shRNAs target 18,661 human genes, which accounts for about 75% of the human genome. Number of hairpins varies in different platforms (Table 1-3), but on average, each gene has two to three shRNAs. Combining RNA polymerase II promoters with shRNAs in miR30-backbones permits efficient suppression even with the integration of a single copy. Each shRNA cassette contains two unique identifiers: the shRNA itself and a random 60-nucleotide barcode that was determined for the identification of a single shRNA amongst the human shRNA library. Overall, the shRNA library offers a convenient, flexible, and effective tool for studying gene function in human cells [10].

#	1	2	3	4	5	6	7	8	9	10	13	total
shRNAs												
Freq (genes)	4092	4238	2792	1065	347	110	30	10	2	2	1	12689

Table 1 Barcode-probed microarray platform: number of shRNAs frequency.

#	1	2	3	4	5	6	7	8	9	10	11	12	13	15	total
shRNAs															
Freq (gene)	2801	4362	4278	1837	678	248	88	27	13	8	2	1	1	1	14345

Table 2 Hairpin-probed microarray platform: number of shRNAs frequency.

#	1	2	3	4	5	6	7	8	9	10	11	13	Total
shRNAs													
Freq (genes)	6934	5983	3628	1355	480	167	60	24	12	4	4	1	18652

Table 3 Next generation sequencing platform: number of shRNAs frequency.

1.2.2 Bacterial media

Prepare, sterilize, and store the solutions at room temperature.

- 1) SOC media: In a glass beaker, mix 0.5 g of NaCl, 5.0 g of yeast extract, 20 g of tryptone, 2.5 mL of 1M KCl, and 10 mL of 1M MgCl₂. Add about 900 mL of ultrapure water and adjust the pH to 7.0 with NaOH. Using a cylinder, complete with water to one liter. Autoclave the solution, and while still warm add 10 mL of 40% Glucose.
- 2) LB low salt media: In a beaker mix, 5 g of NaCl, 10 g of tryptone, and 5 g of yeast extract. Add 900 mL of ultrapure water and adjust the pH to 7.5 with 10 M NaOH. Using a cylinder, complete with water to one liter. Before using, autoclave the solution.
- 3) LB Ampicillin plates: In a beaker, mix 10 g of NaCl, 10 g of tryptone, 5 g of yeast extract, and 15 g of agar. Add 900 mL of ultrapure water and adjust the pH to 7.4. Using a cylinder, complete with water to one liter and autoclave the solution in a bottle. Once the LB agar is warm, add 1 mL of Ampicillin prepared as described in section 2.2.3.
- 4) Following sterile procedures pour the LB agar in to plates.

1.2.3 Antibiotics

Prepare the solutions as described below. Once resuspended, filter and then prepare 1 mL aliquots. Store at -20 °C.

- 1) Ampicillin: Weigh 1 g of Ampicillin and put into a conical tube. Add 10 mL of ultrapure water and dissolve by shaking.
- 2) Zeocin: Weigh 1 g of Zeocin, put into a conical tube, and dissolve in 10 mL of ultrapure water.
- 3) Puromycin: On a precision scale, weigh 20 mg of Puromycin. Add 10 mL of ultrapure water and dissolve by gently shaking the tube.

1.2.4 Linear PEI

Polyethylenimine (PEI) [11] is available in branched and linear forms and can be found in different molecular size polymers. However, low molecular Linear PEI (25kDa) is the one that shows the best transfection efficiency in our hands. The preparation of Linear PEI may be difficult because it is not highly soluble in water. To mix the linear PEI, it is necessary to stir and heat for a long time. As described in the Methods section, the drop by drop addition of HCl to the water is necessary to fully dissolve the Linear PEI.

1. On a precision scale, weigh 323 mg of Linear PEI 25,000 Da.
2. In a glass beaker, add 90 mL of ultrapure water and mix by warming and stirring the solution. Add HCl drop-wise until the product is completely dissolved. Make sure to mix between each drop.

3. Using a cylinder, adjust the volume of the solution to 100 mL. Store at -20 °C as a concentrated stock.
4. Prepare the working Linear PEI solution by diluting the stock 1/10: In a conical tube, mix 9 mL of ultrapure water and 1 mL of concentrated Linear PEI. Mix by inverting the tube several times. Store the transfection reagent at 4 °C until use.

1.2.5 DNA precipitation

3M NaOAc: In a glass beaker, dissolve 24.6 g of NaOAc in 80 mL of ultrapure water. Adjust the pH to 4.8 with Glacial Acetic Acid. Complete to 100 mL with ultrapure water. Autoclave the solution and store at room temperature.

1.2.6 PCR primers sequence (Table 4)

Name	Sequence
Illumi na Fw1	5'- AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTatcacgT AGTGAAGCCACAGATGTA - 3'
Illumi na Fw2	5'- AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTcgatgtT AGTGAAGCCACAGATGTA - 3'
Illumi na Fw3	5'- AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTttaggcT AGTGAAGCCACAGATGTA - 3'
Illumi na Fw4	5'- AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTtgaccaT AGTGAAGCCACAGATGTA - 3'
Illumi na Fw5	5'- AATGATACGGCGACCACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTacagtgT AGTGAAGCCACAGATGTA - 3'

Illumi	5'-
na	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTggcaat
Fw6	AGTGAAGCCACAGATGTA – 3'
Illumi	5'- CAAGCAGAAGACGGCATAACGAGCTCTTCCGATCTGTAATCCAGAGGTTGATTGTTCCA -3'
na	
Rv1	

Table 4 Sequence of example PCR primers including Illumina barcode for PCR preparation of RNAi screening

1.2.7 Labeling

- 1) Random primers: A random sequence of 6 nucleotides is used (NNNNNN). Dissolve dry hexamers in ultrapure water to a concentration of 100 OD units/mL; this corresponds with 15X concentrated primers. Make 5X primer stock by mixing 200 μ L of ultrapure water with 100 μ L of stock primers. Prepare aliquots and store at -20°C.
- 2) Labeling Buffer: In a conical tube, mix 250 μ L of 1 M Tris pH 7.5, 125 μ L of 1 M MgCl₂, 187.5 μ L of 1 M DTT, 500 μ L of 10 mM dATP, 500 μ L of 10 mM dGTP, 50 μ L of 10 mM dTTP, 12.5 μ L of 10 mM dCTP, and 425.5 μ L of ultrapure water. Mix by inverting the tube several times. Prepare 1 mL aliquots and store at -20 °C.

1.3 Experimental Procedures of a RNAi Screen

1.3.1 Library preparation

1.3.1.1 Bacterial transformation

- 1) Thaw two vials of high efficiency electrocompetent bacteria on ice (100 μ l bacteria per vial).
- 2) Pipette 1 μ g of the plasmid library directly into each bacteria vial. Mix by tapping.
- 3) Put 50 μ l of bacteria into a previously chilled electroporation cubette. Be careful not to generate bubbles and be careful to keep the cubette on ice at all the times.
- 4) Prepare four polypropylene round-bottom tubes with 5-10 mL of SOC media (each tube will correspond with an electroporation cubette).
- 5) Dry the cubette, put into the electroporator, and switch the electroporator on. Immediately, add 1 mL of SOC media to the cubette and pipette up and down gently. Collect the bacteria and pipette in to the previously prepared SOC media. Note that having an optimal bacterial transformation is critical to maintain the representation of the library. To ensure this it is important to perform every step gently and to keep the bacteria on ice at all times. After electroporation, the bacteria have to be pipetted in to SOC media immediately. This works better if the cubettes are electroporated one by one.
- 6) Let the bacteria recover for 1 hour in a 37°C shaker.

7) Pipette the recovered bacteria into 1 L of low salt LB containing 25 μ g/mL of Zeocin and 100 μ g/mL of Ampicillin. Take a 1 mL aliquot to test the efficiency of the transformation. Testing the efficiency of the transformation is an easy control for the efficiency of this process. The library is composed of almost 60,000 shRNA and it has been determined that a good representation is having at least a representation of 1,000 times per shRNA; thus 6×10^7 bacteria have to be transformed. From the 1 mL of bacteria that you remove, prepare the following dilutions:

- a. Dilution 1- 90 μ L of LB + 10 μ L bacteria: You expect, at least, 600 colonies.
- b. Dilution 2- 90 μ L of LB + 10 μ L bacteria (from dilution 1): You expect about 60 colonies.
- c. Dilution 3- 90 μ L of LB + 10 μ L bacteria (from dilution 2): You expect more than 6 colonies.

Plate 100 μ L of each dilution in a LB Ampicillin plate and incubate it at 37 $^{\circ}$ C overnight. Count the colonies the next day.

8) Put to grow in a 37 $^{\circ}$ C shaker until saturation-usually about 24 hours.

1.3.1.2 Plasmidic DNA extraction

- 1) Harvest the bacterial cells by centrifugation at 6,000 x g for 15 minutes at 4 $^{\circ}$ C.
- 2) Using the QIAGEN Plasmid Mega Kit, follow the manufacturer's instructions to extract the plasmidic DNA.

1.3.1.3 Library validation

It is necessary to confirm that the representation of each shRNA in the new plasmid library is equal to the representation of each shRNA in the original one. This can be checked by either hybridization to custom microarrays or by NGS of the PCR product (as described in section 2.3.7).

1.3.2 Virus production

Following sterile procedures, carry out all of the cellular manipulation in a hood.

1.3.2.1 Plate Phoenix cells (Day 1)

Aspirate the media from the Phoenix cells plate and wash with sterile PBS. Remove the PBS and add 1.5 mL of trypsin to cover the plate. Place 3 mL of PBS + 20% FBS into a conical tube. Once the cells have detached from the surface of the plate, take the liquid with a pipette and add to the conical tube. Centrifuge the cells at 1,200 rpm for 5 minutes at room temperature. Aspirate the media and resuspend the pellet in 5 mL of DMEM + 10% FBS. Count the cells and plate the amount necessary to have a confluency of 50-70% the next day. The number of cells plated the day before depends on the size of the plate, being optimal:

- a. In a 6-well plate: 6×10^5 - 8×10^5 cells
- b. For a 10 cm plate: 3×10^6 - 4×10^6 cells
- c. In a 15 cm plate: 6×10^6 - 8×10^6 cells

1.3.2.2 Transfection (Day 2)

The following protocol has been constructed to perform a transfection in a 10 cm plate. If there is a need to increase the volumes, follow the instructions as described in Table 5 for DNA amount and Table 6 for linear PEI amount.

Plate	Total	Lentiviral	PMD	CMV	NaCl
6-well	4 g	2 g	1 g	1 g	100 l
10 cm	12 g	6 g	3 g	3 g	250 l
15 cm	24 g	12 g	6 g	6 g	500 l

Table 5 DNA amount reference for plate volume adjustment during transfection of RNAi screening

Plate	Linear PEI volume	NaCl
6-well	16 l	100 l
10 cm	48 l	250 l
15 cm	96 l	500 l

Table 6 Linear PEI amount reference for plate volume adjustment during transfection of RNAi screening

- 1) In a 1 mL tube, mix 6 g of library DNA with 3 g of pMD.G plasmid and 3 g of pCMVR8.91 plasmid [12], and then pipette up and down several times. To the DNA mix, add 150 mM NaCl to a final volume of 250 L. Vortex gently and spin down briefly.
- 2) In a different tube, put 200 L of 150 mM NaCl. Add 50 L of Linear PEI directly to the liquid. Vortex gently and spin down briefly.

- 3) Add the 250 μ L of Linear PEI dilution to the 250 μ L of DNA solution. Mixing the solution in the reverse order may reduce the transfection efficiency.
- 4) Vortex the solution immediately and spin down briefly.
- 5) Incubate for 20 minutes at room temperature
- 6) Add the complete 500 μ L of Linear PEI/DNA mix drop-wise to the cells in a 10 cm plate containing 10 mL of serum-containing medium and homogenize by gently swirling the plate. Return the plates to the cell incubator.

1.3.2.3 Change media (Day 3)

Carefully remove the media from Phoenix cells and replace with fresh DMEM + 10% FBS media. The efficiency of the transfection can be assessed by visualizing the green cells with a fluorescence microscope.

1.3.2.4 Collect the virus (Day 4)

Collect the virus-containing media from the plates and filter using a 0.45 μ m sterile filter. This virus-containing media can either be used to directly infect the cells or can be concentrated.

1.3.3 Infection (Figure 2)

Following sterile procedures carry out all of the cellular manipulation in a hood. The conditions for infection must be determined for each different cell line and batch of virus prepared, as described in section 2.3.4.

- 1) Day 1: In a 6-well plate, plate the appropriate number of cells that is needed to perform the screen.

- 2) Day 2: In a conical tube, prepare a mix of fresh media and virus-containing media at the ratio calculated to produce a low infection efficiency. Add Polybrene to the mix to a final concentration of 10 g/mL. Remove the media from the 6-well plates and add the previously prepared mix. Centrifuge the plates at 1,000 rpm for one hour at room temperature. After the centrifugation, place the plates back in to the cell incubator.
- 3) Day 3: Change the media of the infected cells.

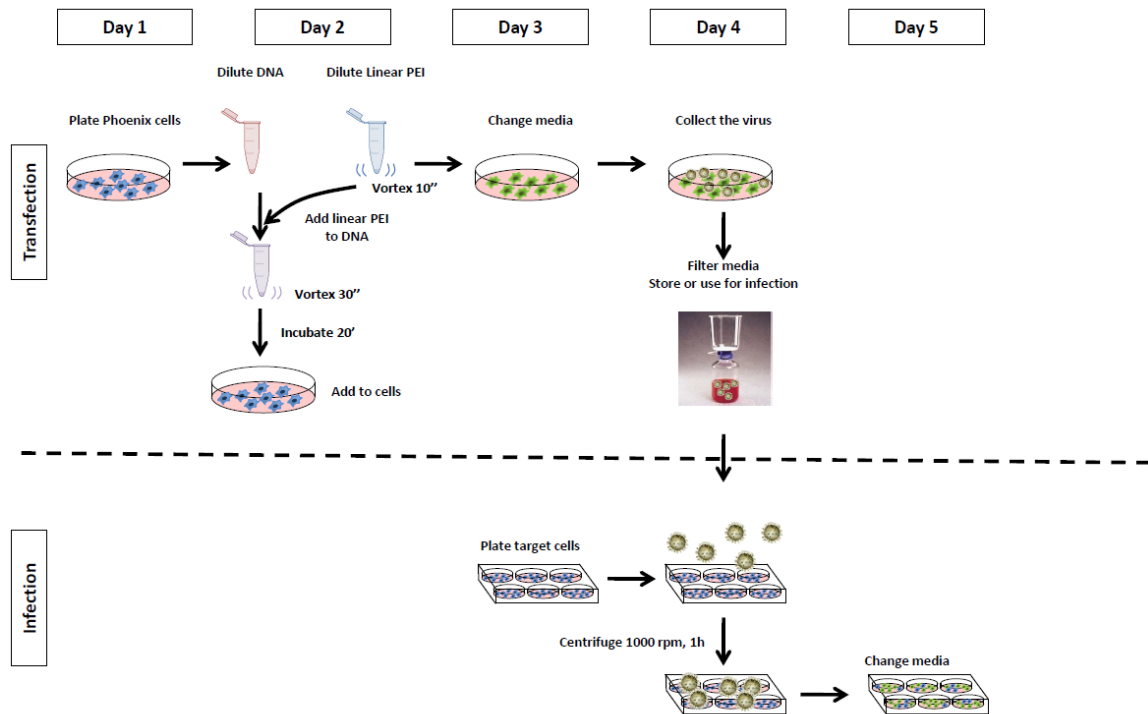


Figure 2 Pipeline for transfection and infection protocol in RNAi screening.

Transfection procedure is schematized in the upper part of the panel. Plate Phoenix cells for a confluency of about 50 -70% the next day. On the second day follow transfection protocol and change the media 24 hours after. Finally, collect and filter the virus-containing media, this media can be directly used to infect the target cells or stored at -20 °C. The infection protocol is outlined under the dashed line. Plate the target cells for 30% confluency and allow them to attach.

The next day, remove the media and replace with media/virus-containing media, centrifuge the plate, and place back into the incubator. Change the media the next day and allow the cells to recover. Infection efficiency can be tested the next day or antibiotic can be added this same day to remove the uninfected cells.

1.3.4 Infection efficiency test

It is crucial to have a multiplicity of infection (MOI) lower than 1 (usually between 0.1 and 0.3) to ensure that any observed effects are in response to the effects of a single shRNA. To ensure this, the number of cells and the ratio media to virus must be adjusted to have an efficiency of infection of 10 - 30% (see Note 7). In all the procedures, the manipulation of the cells must be performed in sterile conditions.

The success of the screen depends on the fact that only single viral particle infects every cell, thus transductions should be performed at multiplicity of infection of less than 1 (typically between 0.1 and 0.3) having in account the Poisson distribution (Table 7).

MOI	0	1	2	3	4
0.1	0.90	0.09	0.00	0.00	0.00
0.2	0.82	0.16	0.02	0.00	0.00
0.3	0.74	0.22	0.03	0.00	0.00
0.4	0.67	0.27	0.05	0.01	0.00
0.5	0.61	0.30	0.08	0.01	0.00
0.6	0.55	0.33	0.10	0.02	0.00
0.7	0.5	0.35	0.12	0.03	0.00
0.8	0.45	0.36	0.14	0.04	0.01

0.9	0.41	0.37	0.16	0.05	0.01
1.0	0.37	0.37	0.18	0.06	0.02

Table 7 Poisson Distribution Reference of Multiplicity Of Infection (MOI)

1.3.4.1 FACS determination of infected cells

- 1) Infect the target cells as described in section 2.3.3.
- 2) 48 hours after infection collect the cells and spin down at 2,000 rpm for 5 minutes at room temperature. Resuspend in 1 mL of PBS + 2% of serum and centrifuge again. Aspirate the PBS and resuspend the pellet in 500 Lof PBS + 2% serum. Pass the suspension of cells through a round-bottom tube with cell-strainer cap.
- 3) Analyze the efficiency of infection by Cytometer by quantifying the percentage of cells that express green fluorescence protein (GFP).

1.3.4.2 Cell number

The efficiency of infection will change depending on multiple aspects of the target cell line; therefore it is necessary to determine the conditions for each case and for each new batch of virus prepared. The main factors to be determined are the number of plated cells and the dilution of the virus. Follow the following instructions as a first test and modify depending on the results attained.

- 1) Day 1: Plate different amount of cells in every well of a 6-well plate. As an example, $1 \cdot 10^5 / 2 \cdot 10^5 / 3 \cdot 10^5 / 5 \cdot 10^5 / 7 \cdot 10^5 / 1 \cdot 10^6$ cell per well can be seeded in a plate.

- 2) Day 2: Infect the cells as described in section 3.3. The infection mix may be done in a ratio 1/1 or 2/1 (media/virus-containing media).
- 3) Day 3: Change the media.
- 4) Day 4: Determine the efficiency of the infection in every well by FACS.

1.3.4.3 Ratio media/virus

- 1) Day 1: In a 6-well plate, plate the number of cells to ensure an infection efficiency of around 50%.
- 2) Day 2: In 3 conical tubes prepare 3 different infection mixes:
- 3) Tube 1: 1 mL of fresh media + 1 mL of virus-containing media
- 4) Tube 2: 1.5 mL of fresh media + 0.5 mL of virus-containing media
- 5) Tube 3: 1.75 mL of fresh media + 0.25 mL of virus-containing media
 - a. Add polybrene to each tube to a final concentration of 10 μ g/mL.
 - b. Aspirate the media from the 6-well plate and add a different mix to each well. Centrifuge the plate at 1,000 rpm for 1 hour at room temperature. Return the plates to the cell incubator
- 6) Day 3: Change the media on the 6-well plates
- 7) Day 4: Determine the efficiency of the infection in each well by FACS.

1.3.5 Screening

1.3.5.1 Infection

The infection of a single particle of virus per cell is critical for de-convoluting the resulting phenotype, but it is also necessary to ensure that a minimum number of cells are infected with each shRNA to ensure the reliability of the screen. It is

accepted that having a minimal representation of 50 - 100 times the number of shRNAs is enough for performing a positive screen, while a representation of 500 – 1,000 is necessary for a negative screening. Thus, for positive screen, the minimum number of infected cells has to be 6×10^6 cells (while 6×10^7 in a negative screen). You must keep at least this same number in every passage, freezing aliquots or preparing pellets as you go.

- 1) Day 1: Plate the previously determined number of target cells in 6-well plates (Section 2.3.4.2). Prepare as many plates needed to maintain the minimal representation of the library and prepare an additional plate to be used as an uninfected control. The number of plated cells has to be calculated taking in to account the minimal representation of the library and that not all of the cells will be infected. Typically, assuming a MOI of 0.3, it is necessary to plate 1.8×10^7 cells to obtain the 6×10^6 infected cells needed for a positive screen (1.8×10^8 in a negative screening).
- 2) Day 2: In a conical tube, mix fresh media with the virus-containing media in accordance with the ratio established in section 3.4.3. Add polybrene to a final concentration of 10 μ g/mL. Mix by inverting the tubes several times. Aspirate the media from the 6-well plates and add 2 mL of the diluted virus prepared in point 2. Add only fresh media to the uninfected control plate. Centrifuge the plates at 1,000 rpm for 1 hour at room temperature. Return the plates to the incubator after centrifugation.
- 3) Day 3: Collect the infected cells from the 6-well plates and mix together in a tube. Mix well and plate the necessary number of 15 cm plates to reach a

confluency of about 70%. In a separate tube, mix the uninfected cells and plate in a separate 15 cm plate.

- 4) Day 4: Add Puromycin to the 15 cm plates in the appropriate concentration to kill the uninfected cells in 2-4 days.

Monitor the cells daily. When all the cells in the uninfected plate have died, the selection has been completed.

1.3.5.2 Passages

Once the 15 cm dishes are subconfluent, collect the cells from all the plates and mix them in a tube. Count the live cells and split them in to three tubes that contain at least the minimum number of cells to maintain the representation of the library:

- 1) Tube 1: Time 0 pellet. Spin down the cells. Aspirate the media and wash the pellet with PBS. Centrifuge and aspirate the media again. Resuspend the pellet in 1 mL of fresh PBS and transfer the suspension to a 1.5 mL tube. Centrifuge at 2,000 rpm for 2 minutes. Aspirate the supernatant and freeze the pellet in liquid nitrogen. This is the t=0 of the screening.
- 2) Tube 2: Storage. Spin down the cells and resuspend the cells in 1 mL of serum + 10% DMSO. Place the cells on dry ice immediately and store in liquid nitrogen.
- 3) Tube 3: Screening. Split the cells in to three replicas that each contains the minimum representation of the library. Replate the cells following the pipeline that you have designed for you screening.

1.3.6 Genomic DNA

1.3.6.1 Cell pellet

- 1) Collect the cells and spin them down at 1,200 rpm for 5 minutes at room temperature.
- 2) Aspirate the supernatant and resuspend the pellet in 1 mL of PBS. Transfer the cells to a previously labeled 1.5 mL tube. Centrifuge in a microcentrifuge at 2,000 rpm for 5 minutes at room temperature. When preparing the cells for the genomic DNA extraction, it is recommended to save to aliquots of the final time point, as a backup in every pellet, remember to freeze the minimum number of cells for maintaining the representation of the library.
- 3) Aspirate the supernatant and lyse or freeze the cell pellet.

1.3.6.2 Genomic DNA extraction

- 1) Resuspend the cell pellet in 10 mL of cold PBS
- 2) Extract genomic DNA using QIAGEN Blood & Cell Culture DNA Kit (Genomic Tip 500/G) following the manufacturer's instructions.
- 3) Quantify the DNA concentration and the quality of the DNA and store at -20 °C. The optimal concentration of the DNA to have a good PCR quality is around 0.5 – 1 mg/mL. If the resulting DNA is more concentrated, increase the volume by adding nuclease-free water.

1.3.7 Sequencing PCR

1.3.7.1 PCR conditions

The conditions for PCR have been set up with using the FastStart Taq DNA Polymerase, dNTP Pack from Roche, and may have to be modified if using a different polymerase. To avoid the contamination of the reaction, it is critical to set up the PCR in a clean hood and to never expose to the library plasmid. To ensure the reliability of the PCR, always include a negative control. Include also a positive control in which the reaction is performed used the original library (10 ng of the plasmidic DNA). A minimal representation of each shRNA is also required when setting up the PCR. The DNA from 6×10^6 cells has to be amplified when doing a positive screen (6×10^7 in a negative one). It is calculated that every cell contains about 3 pg of DNA, therefore to get the minimal representation, it is necessary to use 180 ng of DNA in a positive screen (1800 ng in a negative). This procedure results in a PCR product of 490-500 bp depending on the primers used.

- 1) Use 400 ng of genomic DNA per PCR reaction and only 10 ng of DNA for the library plasmid.
- 2) PCR Mix (Table 8). Per reaction prepare. Frequently, primer dimers occur when running the PCR; to avoid primer dimers it is recommended to keep the reaction on ice all the times.

Reagent	Amount
H ₂ O	Adjust to 100 L

PCR Buffer 10 X+	10	L
MgCl₂		
DMSO	2	L
dNTPS	2	L
Fw primer (100 M)	1	L
Rv primer (100 M)	1	L
FastStart Taq Polymerase	0.8	L
DNA	4	g or 10 ng

Table 8 PCR mix in PCR step of RNAi screening.

3) PCR Conditions (Table 9)


95°C	5 minutes	 35 cycles
95°C	45 seconds	
57°C	30 seconds	
72°C	45 seconds	
72°C	10 minutes	
4°C	∞	

Table 9 PCR conditions of PCR step in RNAi screening

- 4) In a conical tube, mix all of the PCR products obtained from the same sample.
- 5) Prepare a 1.5% agarose gel and run 20 L of the PCR product to ensure that the reaction worked properly. To visualize the PCR prepare a 1.5 % agarose gel in TAE and run an aliquot of the reaction. It is necessary to let it

run for a long time with a low voltage to avoid the generation of secondary structures and separate the amplified product from the dimers.

1.3.7.2 DNA precipitation

- 1) Take the conical tube containing the total PCR product for every sample and add 1/10 volume of 3M NaOAc pH 4.8 and 1/100 volume of Glycogen. Mix it well by inversion.
- 2) Add 0.8 volume of Isopropanol and mix well by balancing gently. At this point, a turbid solution should appear.
- 3) Spin down at 4,000 rpm for 30 minutes.
- 4) Decant the supernatant, resuspend the DNA in 300 μ L of nuclease-free water and transfer to a 1.5 mL tube.
- 5) Precipitate the DNA again by adding 1 mL of 100% Ethanol to the tube.
- 6) Pellet DNA at 13,000 rpm for 30 minutes.
- 7) Aspirate the supernatant and wash the pellet with 1 mL of 70% Ethanol and spin at 13,000 rpm for 15 minutes.
- 8) Aspirate supernatant and air dry DNA. Do not over dry pellet because it will be difficult to resuspend.
- 9) Add 100 μ L of nuclease-free water to the DNA and resuspend at room temperature for at least 2 hours.

1.3.7.3 PCR product purification

- 1) Once resuspended, run the PCR product in a 1.5% agarose gel. Loading 100 μ L to gel-purify the PCR product is not easy depending on the comb format.

It is possible to connect 2 or 3 wells with tape to get a bigger well. Also, the product could be split in 3 wells and mixed at the end.

- 2) Cut with a razor the band that appears at 490-500 bp size and pass it to a previously labeled 1.5 mL tube.
- 3) Extract the DNA from the agarose using a DNA gel extraction kit and following the manufacturer instructions. NOTE: Elute DNA in the minimum amount permitted.

1.3.7.4 DNA quantification

The quantification of DNA concentration has to be very precise to keep the correct proportion between conditions when preparing the samples. To ensure this, it is recommended to quantify the DNA concentration by 2 different methods, as for example:

- 1) Quantification of DNA using an accurate equipment, such as Nanodrop or Qubit.
- 2) Quantification in gel. Prepare a 1.5% agarose gel and load 200 ng of each sample according to the concentration that has been previously measured. The intensity of all of the bands must be equal.
- 3) The resulting PCR product can be analyzed by hybridization of customized microarray (2.3.7.5) or by NGS (section 2.3.7.6).

1.3.7.5 Sample preparation for hybridization

- 1) Place 1-2 μ g of the PCR product into a 0.2 mL PCR tube and adjust the volume to 74.5 μ L with nuclease-free water.

- 2) Add 10 μ L of 5X random primer. Mix with a pipette a few times and spin down. Incubate the mixture at 98 $^{\circ}$ C for 5 minutes and immediately place on ice for 5 minutes.
- 3) Add 10 μ L of 10X Labeling reaction buffer to each reaction mixture.
- 4) Add 4.5 μ L of Cy5-dCTP to each experimental sample and 4.5 μ L of Cy3-dCTP to each reference sample. Pipette the mix up and down.
- 5) Add 1 μ L of Kleenow enzyme to each tube. Mix gently and spin down briefly. Incubate at 37 $^{\circ}$ C for 4 hours in a thermocycler. Let it cool down to 4 $^{\circ}$ C.
- 6) Add 5 μ L of 0.5 M EDTA pH 8.0 to stop the labeling reaction (optional).
- 7) Increase the volume of the sample by adding 400 μ L of nuclease free-water. Pipette the mix into an Amicon Ultra -0.5 mL 30 K disposal to clean the unbound dye. Centrifuge at 10,000 X g for 15 minutes.
- 8) Remove the liquid present in the lower receptacle. Add 500 μ L of nuclease-free water to the sample. Centrifuge the tube at 10,000 x g for 15 minutes. Repeat this procedure until the drained liquid appears clear.
- 9) Collect the sample and store in a new previously labeled 1.5 mL tube. Quantify the DNA amount and the incorporation of the dye using a Nanodrop device.
- 10) In a previously labeled 1.5 mL tube, mix pipette 1 μ g of your sample and 1 μ g of the reference sample. Complete to a volume of 158 μ L by adding nuclease-free water. Store the sample at 4 $^{\circ}$ C or freeze it.

1.3.7.6 Sample preparation for sequencing

The conditions of the sample to be deep sequenced may vary between different facilities, but, as a general statement, in a 1.5 mL, mix 100 ng of each sample. Add H₂O to reach a volume of 20 μ L and keep frozen or at 4 °C depending on the instructions received.

References

1. Rodriguez-Barrueco, R., N. Marshall, and J. Silva, Pooled ShRNA screenings: Experimental approach (in print), in *Methods in Molecular Medicine*. 2012.
2. Dykxhoorn, D.M., C.D. Novina, and P.A. Sharp, Killing the messenger: short RNAs that silence gene expression. *Nat Rev Mol Cell Biol*, 2003. 4(6): p. 457-67.
3. Paddison, P.J. and G.J. Hannon, RNA interference: the new somatic cell genetics? *Cancer Cell*, 2002. 2(1): p. 17-23.
4. Paddison, P.J., et al., A resource for large-scale RNA-interference-based screens in mammals. *Nature*, 2004. 428(6981): p. 427-31.
5. Moffat, J., et al., A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 2006. 124(6): p. 1283-98.
6. Silva, J.M., et al., Second-generation shRNA libraries covering the mouse and human genomes. *Nat Genet*, 2005. 37(11): p. 1281-8.
7. Berns, K., et al., A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, 2004. 428(6981): p. 431-7.
8. Willingham, A.T., et al., RNAi and HTS: exploring cancer by systematic loss-of-function. *Oncogene*, 2004. 23(51): p. 8392-400.
9. Silva, J.M., et al., RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells. *Proc Natl Acad Sci U S A*, 2004. 101(17): p. 6548-52.
10. Silva, J.M., et al., Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, 2008. 319(5863): p. 617-20.

11. Reed, S.E., et al., Transfection of mammalian cells using linear polyethylenimine is a simple and effective means of producing recombinant adeno-associated virus vectors. *Journal of Virological Methods*, 2006. 138(1-2): p. 85-98.
12. Salmon, P., et al., High-level transgene expression in human hematopoietic progenitors and differentiated blood lineages after transduction with improved lentiviral vectors. *Blood*, 2000. 96(10): p. 3392-8.

Appendix B: Book chapter – Computational Analysis of High-throughput RNAi screening Data

Jiyang Yu¹⁻², Preeti Putcha³⁻⁴, Jose M. Silva³⁻⁴, Andrea Califano¹⁻³

¹ Department of Biomedical Informatics, Columbia University, New York, NY, 10032, USA.

² Joint Centers for Systems Biology, Columbia University, New York, NY, 10032, USA.

³ Institute for Cancer Genetics, Columbia University, New York, NY, 10032, USA.

⁴ Department of Pathology, Columbia University Medical Center, New York, NY, 10032, USA.

Summary

Genome-wide RNA interference (RNAi) screening has emerged as a powerful tool for functional genomic studies of disease-related phenotypes and discovery of molecular therapeutic targets for human diseases. Commercial short hairpin RNA (shRNA) libraries are commonly used in this area and state-of-the-art technologies including microarray and next generation sequencing are available to read out shRNA-triggered phenotypes. However, computational analysis of this complex data remains challenging due to noise and small sample size from such large-scaled experiments. In this chapter we discuss the pipelines and

statistical methods of processing, quality assessment and post-analysis for both microarray and sequencing-based screening data.

Key Words: genome-wide, pooled shRNA screen, microarray, Next-Generation Sequencing, barcode, QA, normalization, decoding, differential representation, GSEA

1 Introduction

RNA interference (RNAi) has emerged as one of the standard techniques for studying phenotype-specific gene function from plants to fungi to animals via suppression of gene expression [1-4]. RNAi-based gene silencing can be achieved by the use of short interfering RNAs (siRNAs) or short hairpin RNA (shRNA) expression vectors. Among the two approaches, shRNA is more feasible because siRNA has the problem of transient inhibition of gene expression and inefficient transfection into non-dividing cells; however, shRNA can be stably integrated into a target cell genome via retroviral or lentiviral gene transfer, resulting in the permanent reduction of the targeted gene product. Several shRNA expression libraries targeting entire human genome have been generated to facilitate functional analysis of the whole transcriptome through loss-of-function genetic studies [5-8].

In genome-wide shRNA screening, a large population of cells is infected or transfected with a pool of different shRNA lentiviral vectors and shRNA hairpins are integrated into cell genomes. After that, there are two common applications

of these transduced cells. One is growing the cells for a sufficient number of doubling times, extracting the genomic DNA at initial time (T0) and after harvesting (T10), and then comparing quantity of shRNAs in these two time-points. This usage is to identify genes that are essential for cell survival or growth, thus making potential therapeutic targets for cancer and other type of human diseases, and hairpins of those lethal genes will be depleted or under-represented in T10 population. The other application is splitting infected cells into two groups, treating the two groups differently, for example treating one group with drug and nothing to the other as control. After this selective pressure, grow cells from both populations and then compare shRNAs extracted from genomic DNA of each population. This approach is to identify genes that modulate response to the perturbation. In the example of drug treatment, this screen can help to identify genes that increase sensitivity or resistance of cells to the drug.

To read out shRNA hairpins extracted from genomic DNA, microarray hybridization is commonly used with the advantage of low cost and flexibility. It employs PCR-amplified shRNA template sequence pools extracted from shRNA library-transduced cells under test as well as reference conditions. Each PCR fragment is labeled with a different fluorophore, followed by hybridization of both pools to the same array, or labeled with the same fluorophore followed by hybridization to multiplex arrays. Taking the two-color microarray as example, the ratio of signal intensities of two colors (Cy3, Cy5) for each probe sequence reflects the relative abundance of cells expressing the corresponding shRNA construct under test condition as compared to the reference. Consequently,

shRNA hairpins that sensitize cells in the selective condition will be depleted from the pool, showing low values of signal ratio, whereas shRNA constructs that render cells resistant will be enriched, showing high values of signal ratio. Three types of molecular tags have been used as microarray probes, namely full-length hairpin, half hairpin, and external barcode sequence. Half hairpin is able to overcome the self-annealing problem during PCR amplification happening to full-length hairpin, and correspondingly has more efficient labeling and microarray hybridization than full-length hairpin [4, 9]. Barcodes are not necessary for enrichment screens or positive selections such as designs to detect shRNA constructs for cell proliferation [10], but are critical for depletion screens or negative selections such as studies designed to detect cell-lethal or drug-sensitive shRNAs [9, 11-13].

Next generation sequencing (NGS) has recently emerged as a cost-effective technology of quantitatively measuring abundance of short-length DNA or RNA in a short time. This massively parallel sequencing has been used in pooled shRNA screens [14-16], and comparing to microarray-based approaches, it offers several potential advantages in terms of coverage of targeting genes, flexibility of input library, scalability and dynamic range. As the cost of NGS is rapidly decreasing, this means might dominate high-throughput shRNA screening in the near future.

In this chapter, we will discuss computational analysis of both microarray and NGS-based shRNA screening data. In particular, we will introduce multiple

quality assessment metrics for raw data of microarray and NGS respectively, the pipeline to decode shRNA NGS data, preprocessing of screening data including background correction and normalization, quality controls of processed data to detect biological artifacts of experiments, statistical methods for differential representation analysis at individual shRNA level and gene level to identify candidates of interest and functional enrichment of selected candidates.

2 Materials

2.1 shRNA Library: Thermo Scientific Open Biosystems GIPZ Lentiviral human shRNAmir library is used to illustrate the analysis of RNAi screening data.

The library is composed of 58,493 hairpin constructs, in which 39,458 shRNAs are known to target 18,661 human genes, about 75% of the genome. In the GIPZ library, one gene might have multiple shRNAs and as shown in the distribution table of number of shRNAs per gene (Table 1), the majority of genes has at least 2 or 3 shRNAs.

2.2 Microarray Data: Two types of microarray probe designs based on the structure of shRNA construct are introduced in this chapter: barcode and half shRNA hairpin. For each type of design, the oligonucleotide probes for hybridization have both sense and anti-sense sequences of the molecular tags.

2.3 Next-Gen Sequencing: NGS-based screening data used in this chapter is generated from Illumina HiSeq 2000.

2.4 Software: all analysis is performed under the platform of R [17] and related Bioconductor [18] packages.

2.5 Experimental Design: A negative RNAi screen experiment in both microarray and NGS platforms is employed (see Note 1) to illustrate the computational analysis procedures.

3 Methods

3.1 Preprocess of Microarray Data: extract intensity signals of two-colors (red for sample, green for reference) from DNA microarray readout, and use the score of $\log_{10}(I_{\text{red}}/I_{\text{green}})$ as representation of shRNAs, separate data for barcode or shRNA probes and control probes, and then perform background correction (see **Note 2**) for each array.

3.2 Normalization of Microarray Data: assemble the microarray data of all samples together and conduct quantile normalization (see **Note 3**) across all arrays.

3.3 QA of Raw Sequencing Data: transform millions of short-reads generated from NGS machine into FASTQ format, and conduct quality assessment (QA) using different metrics (see **Note 4**) before any further analysis.

3.4 Decoding of shRNA Sequencing Data: decode each short read by identifying its experimental condition and represented shRNA construct (see **Note 5**), and count number of short reads for each shRNA in each sample.

3.5 Normalization of Processed NGS Data: normalize the profiles of shRNA count (see **Note 6**) in order to compare shRNA representations under different conditions.

3.6 QA of Normalized Data: perform advanced quality assessment on normalized microarray or sequencing data to identify outlier experiments using MA plot (see **Note 7**), variance-mean dependence plot (see **Note 8**), distribution plot (see **Note 9**), clustering of samples (see **Note 10**), Principle Component Analysis (PCA) and plot of the first two components, (see **Note 11**).

3.7 Consistence of Replicates: plot and calculate correlations (see **Note 12**) between biological replicates of the same experiment to further check consistence of experiments and to detect outlier samples.

3.8 Differential Representation Analysis: conduct case-control comparison (T10 vs. T0) to identify differentially-represented shRNAs, either depleted or enriched in case samples. Depletion or under-representation of a shRNA means the targeting gene is lethal to experimental cells, thus making it a good candidate as potential therapeutic targets of diseases. Due to the fact that multiple shRNAs could target the same gene in the library, silencing effects of shRNA on cell viability are estimated at individual shRNA level (see **Note 13**), best shRNA level (see **Note 14**), and integrated gene level (see **Note 15**). Corresponding statistics including fold change (FC) (see **Note 16**), Z-score (see **Note 17**), p-value, False Discovery Rate (FDR) (see **Note 18**) are reported.

3.9 Heatmap of Selected shRNAs/Genes: to visualize the pattern of differentiated shRNA-silencing effects such as similarity between genes or samples, plot clustering-enabled heatmap of z score and microarray or NGS data of pre-selected shRNAs (see **Note 19**).

3.10 Functional Enrichment Analysis: to identify functional similarities of genes identified by RNAi screens, perform Gene Set Enrichment Analysis (GSEA) using public available functional database including Gene Ontology, KEGG pathways, etc (see **Note 20**).

4 Notes

4.1 Four Diffuse-Large B-Cell Lymphoma (DLBCL) cell lines are prepared for shRNA screening. For each cell line, genomic DNA is extracted at T0, the initial time after transduction of lentiviral shRNA library and at T10 when cells are harvested after 10 doubling times. Triplicates are conducted for each time point.

4.2 Negative control probes that targeting no genes are used to estimate the background signal of each array, which is proportional to the total amount of sample DNAs. Background correction is performed by subtracting the mean/median of negative controls from signal of each shRNA within the same microarray.

4.3 Quantile-normalization method [19] is suggested to normalize microarray data across multiple arrays to preserve the rank of shRNAs and to make it comparable between conditions by forcing the same distribution of each array.

4.4 An R Bioconductor package 'ShortRead' [20] is used to do quality assessment of Genome Analyzer output. It takes FASTQ format as input. The report summary includes read distribution (**Figure 1A**), read count (**Figure 1B**), read overall quality (**Figure 1C**) and cycle-specific base quality (**Figure 1D**).

4.5 According to the construction of each 51nt-length sequence read (**Figure 2**), the first 6 nucleotides (in blue) are used to mapping back to the barcodes for 6 experimental conditions, and the 22 nucleotides (in red) in the middle are used to identify shRNA hairpin in the library it belongs to.

4.6 The normalization of sequencing count profile is scaling reads of each shRNA to equalize the total number of reads for all samples, which is proportional to the cell population size of each experiment.

4.7 M and A are defined as:

$$M = \log_2 V_1 - \log_2 V_2$$

$$A = \frac{\log_2 V_1 + \log_2 V_2}{2}$$

V_1 is the intensity ratio (microarray data) or shRNA count (sequencing data) of the sample studied, and V_2 is for a "pseudo"-sample that consists of the median across all samples. Generally, we expect the mass of the distribution in an MA plot (**Figure 3A**) to be concentrated along the $M = 0$ axis, and there should be no trend in M as a function of A . If there is a trend in the lower range of A , this often indicates that the samples have different background signals; this may be addressed by background correction. A trend in the upper range of A can indicate saturation of the measurements; in mild cases, this may be addressed by non-linear normalization.

4.8 Variance-mean dependence plot (**Figure 3B**) is the standard deviation of the representation values across samples on the y-axis versus the rank of their mean on the x-axis. The red dots, connected by lines, show the running median of the standard deviation. Typically, one expects the red line to be approximately

horizontal, that is, show no substantial trend. In some cases, a hump on the right hand of the x-axis can be observed and is symptomatic of a saturation of the measurements.

- 4.9 Boxplots (**Figure 3C**) represent summaries of the signal distributions of the samples. Each box corresponds to one sample. Typically, we expect the boxes to have similar positions and widths. If the distribution of a sample is very different from the others, this may indicate an experimental problem. Outliers based on the Kolmogorov-Smirnov statistic between each sample's distribution and the distribution of the pooled data, are marked by an asterisk (*). Density plots (**Figure 3D**) are smoothed histograms of the data. Typically, the distributions of the samples should have similar shapes and ranges. Outliers, according to the same criterion as in the boxplots, are highlighted by color.
- 4.10 Heatmap (**Figure 4A**) of between sample distances and dendrogram of sample clustering (**Figure 4B**) can help to detect batch effects, as well as clustering of samples based on biological effects. The color scale is chosen to cover the range of distances encountered in the dataset. Datasets for which the sum of the distances to the others is much different from the others are detected as marked by * as outliers. The distance between two samples is the mean absolute difference (L1-distance) between the vectors of M-values (see **Note 7**) of the samples.
- 4.11 Scatter plot (**Figure 4C**) of the samples along the first two principal components is used to check whether the samples cluster, and whether this is because of an intended biological or experimental factor, or according to

unintended reasons such as "batch effects". Outliers, according to the same criterion as in the heatmap plot, are indicated by larger symbols.

- 4.12 Scatter plot (**Figure 5**) of between biological or technical replicates is another quick visualization method to check the consistence of experimental replicates. Empirical distribution of each replicate sample is plotted in the dialogue, and they are expected to have similar shape and scale. Upper triangle shows the Pearson and Spearman correlation between two replicated samples without any filtering on shRNAs.
- 4.13 To estimate the differential representation of individual shRNAs, a moderated t-type test [21, 22] can be used to test the statistical significance, or a linear modeling approach [23] can be used to fit the data. For the modeling approach, the likelihood needs to be regularized by classical Frequentist's stabilization method [22], Bayesian or empirical Bayesian approach [23] due to small sample size issue. The regression coefficient represents the level of difference between case and control groups, and the statistical significance can be estimated by Chi-square test or Wald's z-test.
- 4.14 To obtain the effects of shRNA on cell viability at gene level, one simple approach is to perform individual shRNA analysis first, and from all shRNAs targeting the same gene, select the one showing the most significant depletion or enrichment, namely the best shRNA for the corresponding gene. However, this approach is heuristic and might cause a high false positive rate.
- 4.15 Another idea to estimate the effects at gene level is to integrate all shRNA data for the same gene. For this type of method, one can perform individual

shRNA analysis first and then combine statistics of all shRNAs for the same gene by either signed Fisher's method [24] or Stouffer's method [25], or one can use hierarchical modeling approach [26, 27] by introducing an indicator variable for shRNA level and allowing random effects along with different shRNAs and use the fixed effects to estimate the overall gene level effects. The first separate-and-combine approach might introduce many false positives due to inaccurate estimation of individual shRNAs effects from small sample size and noisy nature of high-throughput design, but the second modeling-together method might overcome this problem by increasing sample size. Also, the likelihood of statistical model needs to be penalized by Bayesian approach to obtain robust estimation of parameters.

- 4.16 To estimate the fold change between case and control samples, one need to calculate the mean within case or control samples. Two methods can be used: arithmetic or geometric mean, and the latter one is suggested for robustness.
- 4.17 For the linear modeling approach, the Z score (**Figure 6A**) is calculated by estimate of regression coefficient over its standard deviation, which asymptotically follows a standard Gaussian distribution, therefore the two-tailed p value (**Figure 6B**) for statistically significance can be calculated based on this null distribution.
- 4.18 FDR for correction of multiple comparisons is calculated by BH procedure [28].
- 4.19 The Heatmap with both row and column clustering of Z scores (**Figure 7A**) and normalized data (**Figure 7B**) can be used to visualize the similarity pattern

between shRNAs or sample conditions. Euclidian or correlation can be used for distance metrics and Wald method is suggested for hierarchical clustering.

4.20 Gene Set Enrichment Analysis of pathways (**Figure 8**) or GO terms uses differential representation results of all shRNAs or genes as the reference, for example, ranking from the most enriched to the most depleted. Classical weighted K-S statistic [29] or Maxmean statistic [30] can be used to estimate the enrichment score, and gene label shuffling is commonly used to estimate significance in this small sample size situation.

References

1. Cheung, H.W., et al., Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci U S A*, 2011.
2. Hammond, S.M., et al., An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature*, 2000. 404(6775): p. 293-6.
3. Silva, J.M., et al., Cyfip1 is a putative invasion suppressor in epithelial cancers. *Cell*, 2009. 137(6): p. 1047-61.
4. Silva, J.M., et al., Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science*, 2008. 319(5863): p. 617-20.
5. Silva, J.M., et al., Second-generation shRNA libraries covering the mouse and human genomes. *Nat Genet*, 2005. 37(11): p. 1281-8.
6. Silva, J.M., et al., RNA interference microarrays: high-throughput loss-of-function genetics in mammalian cells. *Proc Natl Acad Sci U S A*, 2004. 101(17): p. 6548-52.
7. Paddison, P.J., et al., A resource for large-scale RNA-interference-based screens in mammals. *Nature*, 2004. 428(6981): p. 427-31.
8. Moffat, J., et al., A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell*, 2006. 124(6): p. 1283-98.

9. Schlabach, M.R., et al., Cancer proliferation gene discovery through functional Genomics. *Science*, 2008. 319(5863): p. 620-624.
10. Berns, K., et al., A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature*, 2004. 428(6981): p. 431-7.
11. Brummelkamp, T.R., et al., An shRNA barcode screen provides insight into cancer cell vulnerability to MDM2 inhibitors. *Nature Chemical Biology*, 2006. 2(4): p. 202-206.
12. Luo, B., et al., Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences of the United States of America*, 2008. 105(51): p. 20380-20385.
13. Possemato, R., et al., Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature*, 2011. 476(7360): p. 346-50.
14. Fellmann, C., et al., Functional Identification of Optimized RNAi Triggers Using a Massively Parallel Sensor Assay. *Molecular cell*, 2011. 41(6): p. 733-746.
15. Bassik, M.C., et al., Rapid creation and quantitative monitoring of high coverage shRNA libraries. *Nat Methods*, 2009. 6(6): p. 443-5.
16. Burgess, D.J., et al., Topoisomerase levels determine chemotherapy response in vitro and in vivo. *Proc Natl Acad Sci U S A*, 2008. 105(26): p. 9053-8.
17. <http://www.r-project.org>.
18. Gentleman, R.C., et al., Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 2004. 5(10): p. R80.
19. Bolstad, B.M., et al., A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003. 19(2): p. 185-93.
20. Morgan, M., et al., ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, 2009. 25(19): p. 2607-8.
21. Baldi, P. and A.D. Long, A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 2001. 17(6): p. 509-19.
22. Tibshirani, R., <http://www-stat.stanford.edu/~tibs/SAM/>.
23. Diboun, I., et al., Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. *Bmc Genomics*, 2006. 7: p. 252.

24. Fisher, R.A., Notes on Combining Independent Tests of Significance. 1948.
25. Stouffer, S.A., et al., Adjustment During Army Life. Vol. 1. 1949, Princeton: Princeton University Press.
26. Gelman, A., et al., Bayesian Data Analysis. 2nd edition ed. Texts in Statistical Science, ed. C. Chatfield, M. Tanner, and J. Zidek. 2004: Chapman & Hall.
27. Ji, H. and X.S. Liu, Analyzing 'omics data using hierarchical models. Nat Biotech, 2010. 28(4): p. 337-340.
28. Benjamini, Y. and Y. Hochberg, Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological, 1995. 57(1): p. 289-300.
29. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A, 2005. 102(43): p. 15545-50.
30. Efron, B. and R. Tibshirani, On Testing the Significance of Sets of Genes. Annals of Applied Statistics, 2007. 1(1): p. 107-129.

Tables and Figures

Table 1: Distribution of Number of shRNAs per Gene

# shRNAs Per Gene	1	2	3	4	5	6	7	8	9	10	11	13	total
Freq of Genes	6,931	5,986	3,635	1,355	481	168	60	24	12	4	4	1	18,661

sequences of shRNA hairpins in the library, out of which 19 nucleotides are perfectly matched to the genome sequence.

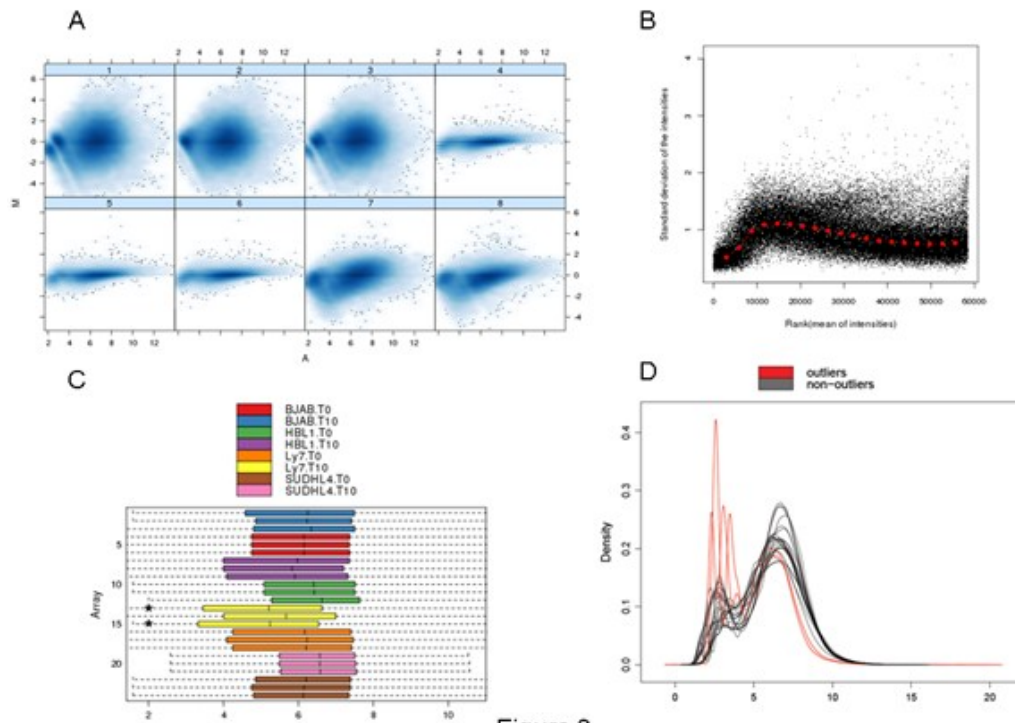


Figure 3

Figure 3: Quality assessment part 1 of normalized microarray or NGS data. (A) MA plot (B) Variance-mean dependence plot (C) Boxplot (D) Density distribution plot

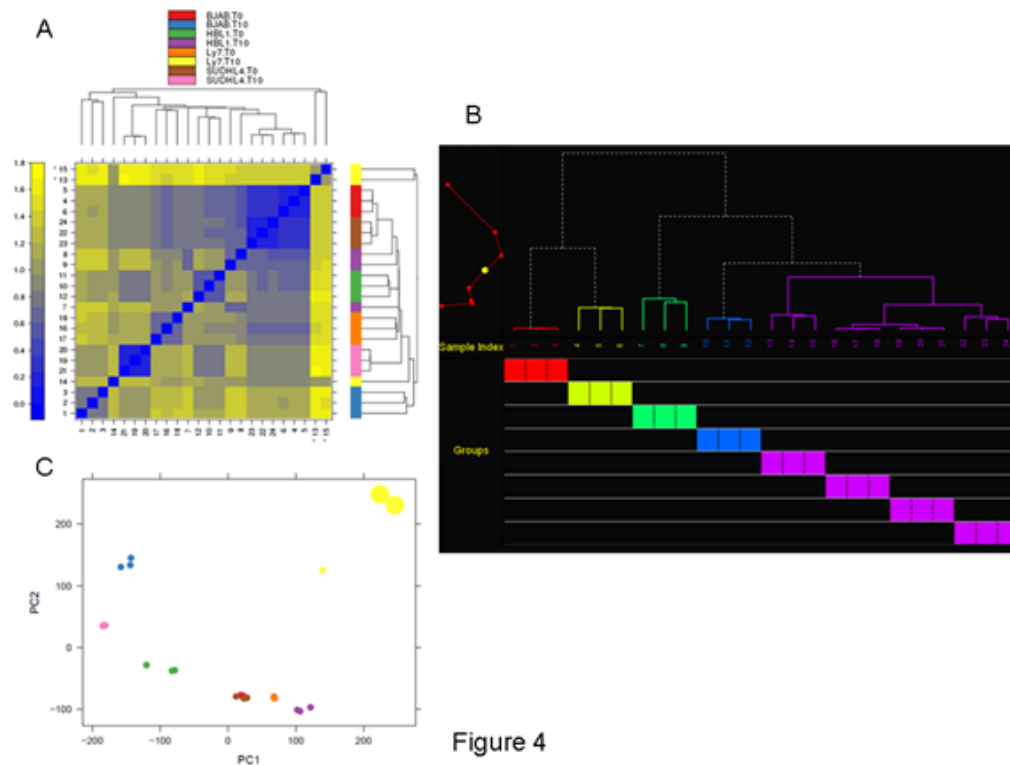


Figure 4

Figure 4: Quality assessment part 2 of normalized microarray or NGS data. (A) Heatmap of sample distances (B) Scatter plot the first two principal components (C) Hierarchical clustering of all samples: dots on the upper left plot indicates where to split the three to obtain specific number of clusters, in which the yellow one is for the current plot; colors are for different clusters.

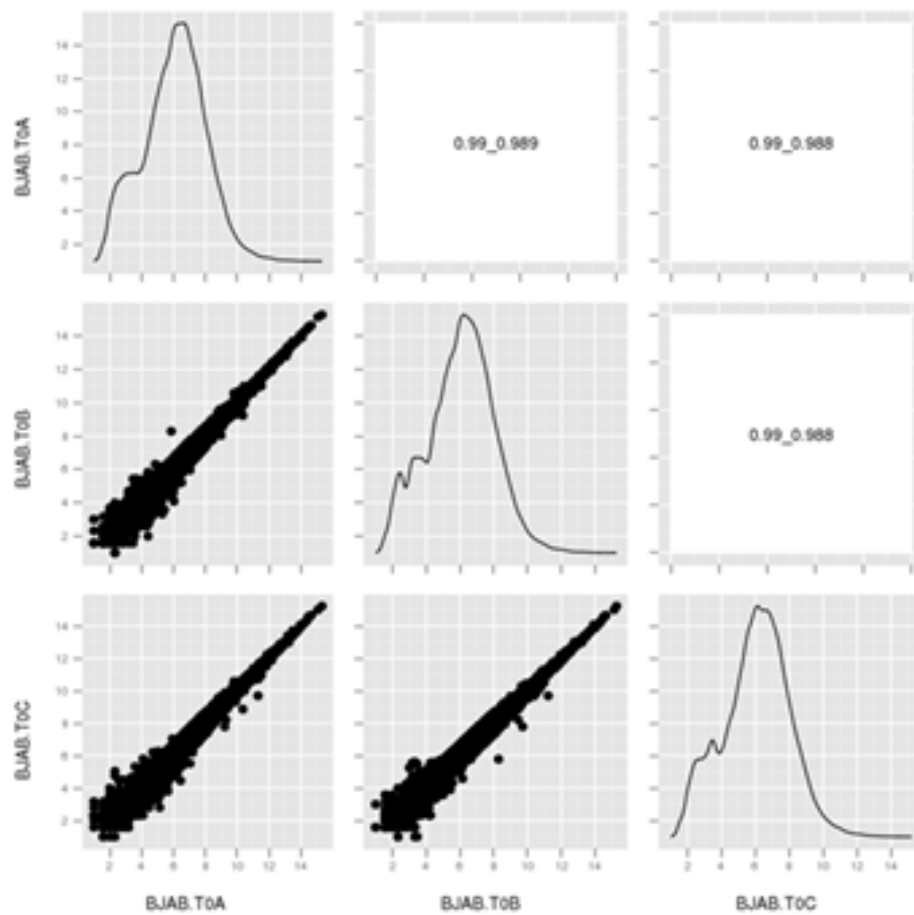


Figure 5

Figure 5: Scatter plots and correlations between biological replicates. Plots in the dialogue are density distributions of data in each replicate. Texts in the upper triangle cells indicate Pearson (the first number) and Spearman correlations.

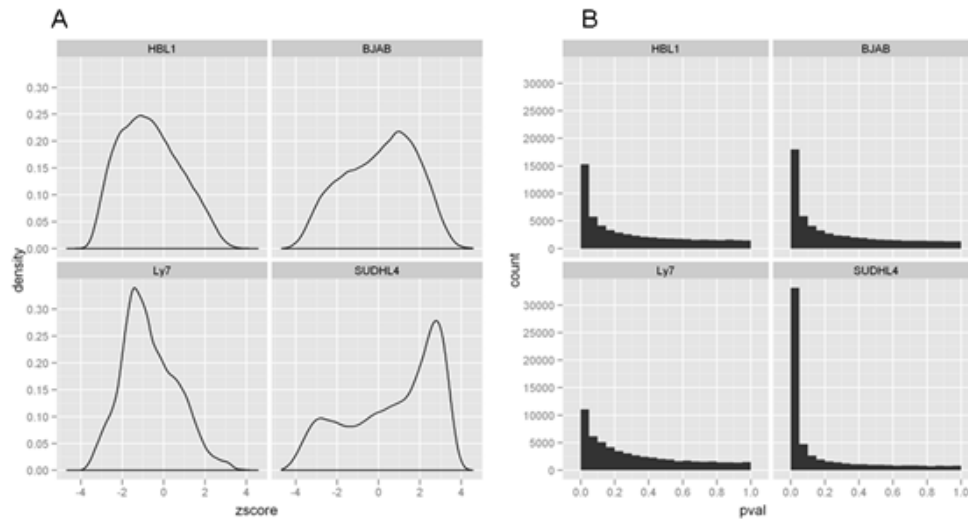


Figure 6

Figure 6: Density plots of Z score and histograms of p values for differential-representation results. Negative Z score means depletion of shRNA in experiment of study.

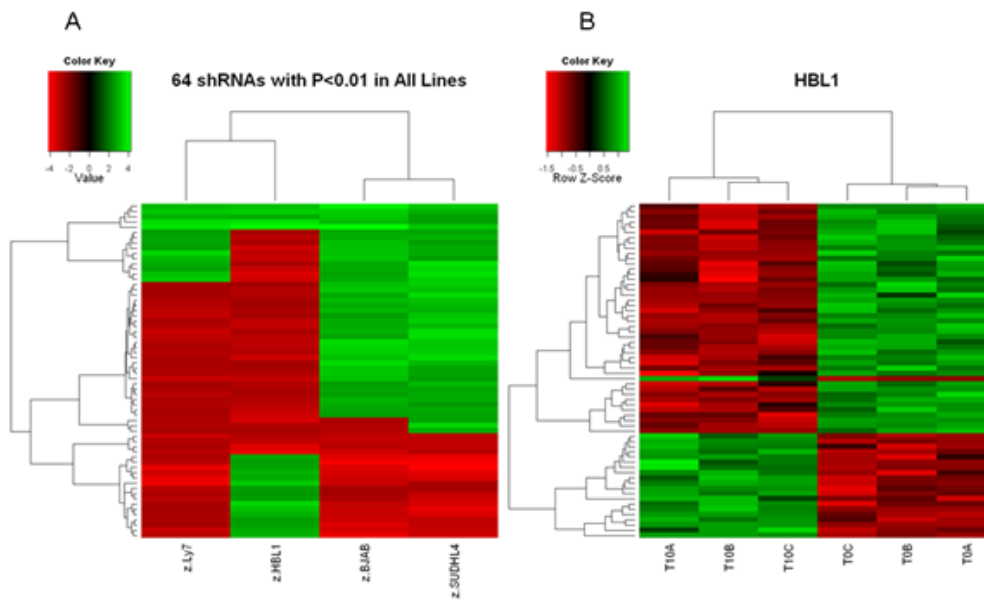


Figure 7

Figure 7: Heatmap of (A) Z scores and (B) normalized data of shRNAs showing significantly ($P < 0.01$) differential-representation in all screened cell lines.

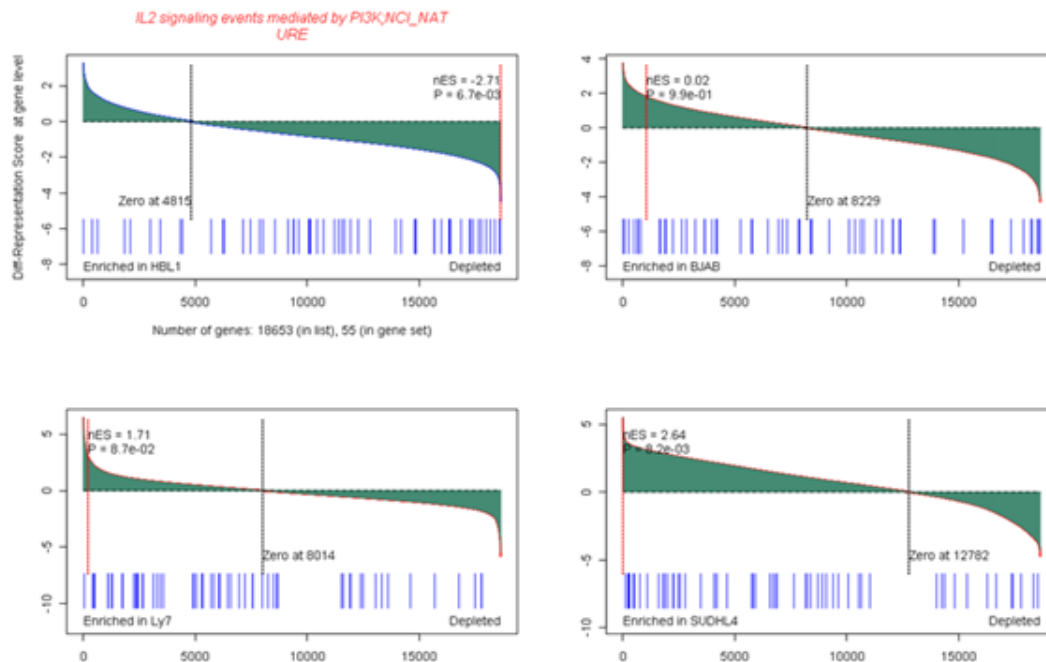


Figure 8

Figure 8: An example GSEA plot of pathway or GO gene sets in differentially-represented shRNAs. Y axis shows the z score of differential representation at shRNA level or gene level. The red dashed lines indicate normalized Enrichment Score (nES) and P value.