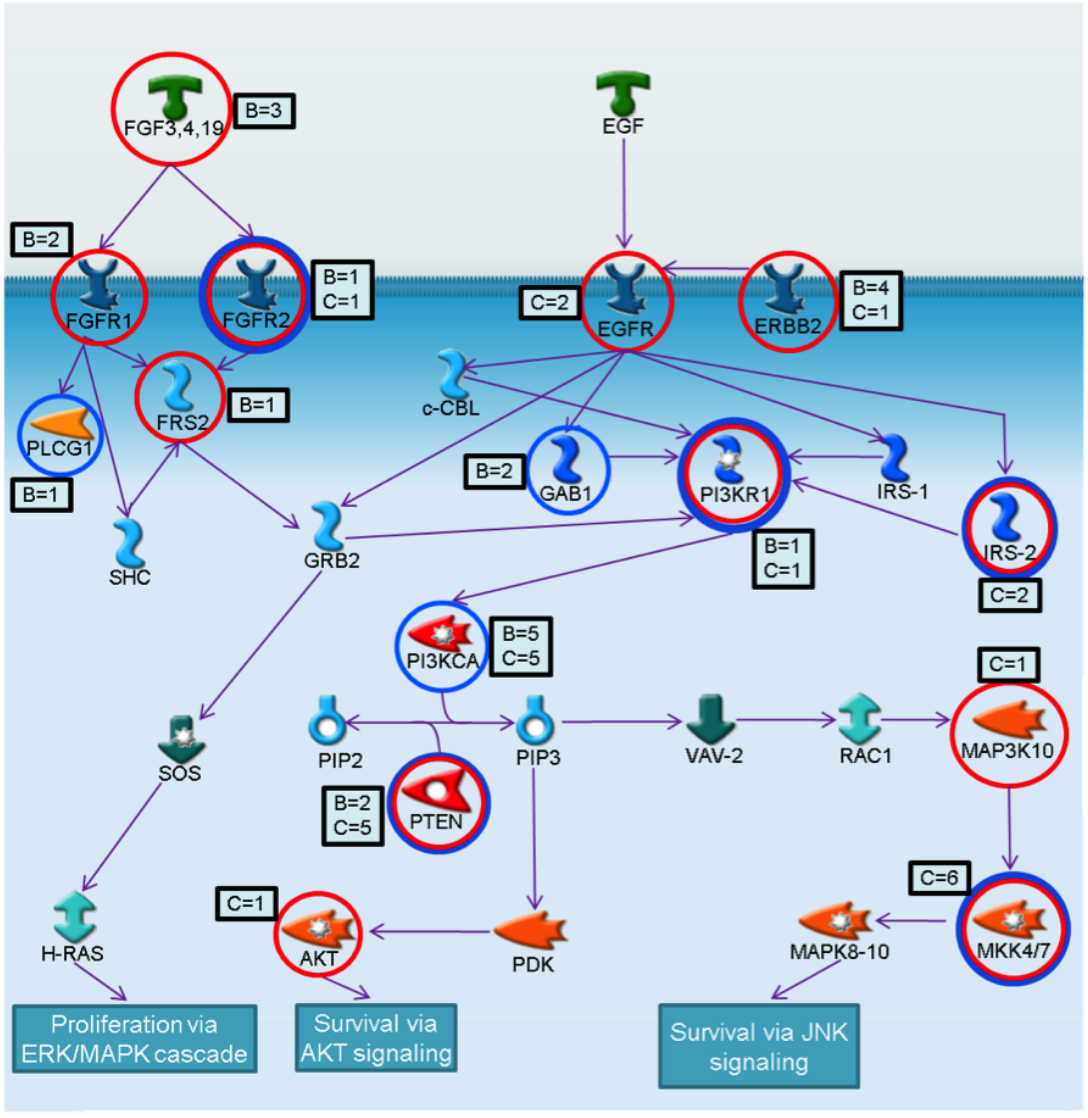
Interpreting Diverse Genomic Data Using Gene Sets

giovanni_parmigiani@dfci.harvard.edu

FHCRC February 2012

Gene Sets

Leary 2008 PNAS



Alterations in the combined FGF, EGFR, ERBB2 and PIK3 pathways.

Red: Copy number alterations; Blue: Point mutations.

Why gene-set analysis?

Improvements in interpretability of experimental results.

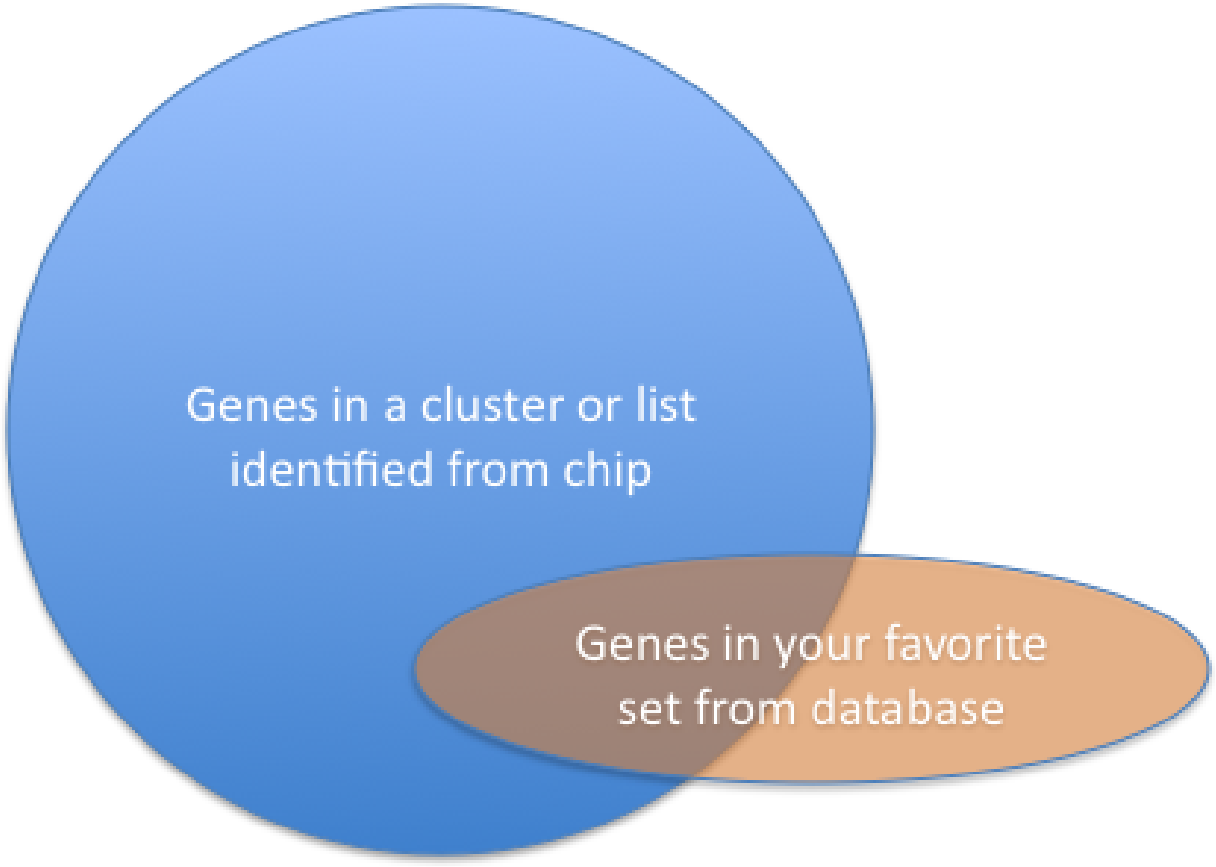
Detection of subtle correlated changes in sets.

Detection of set-level biological signals.

Integration of diverse data sources.

The birthplaces of gene set analysis: I

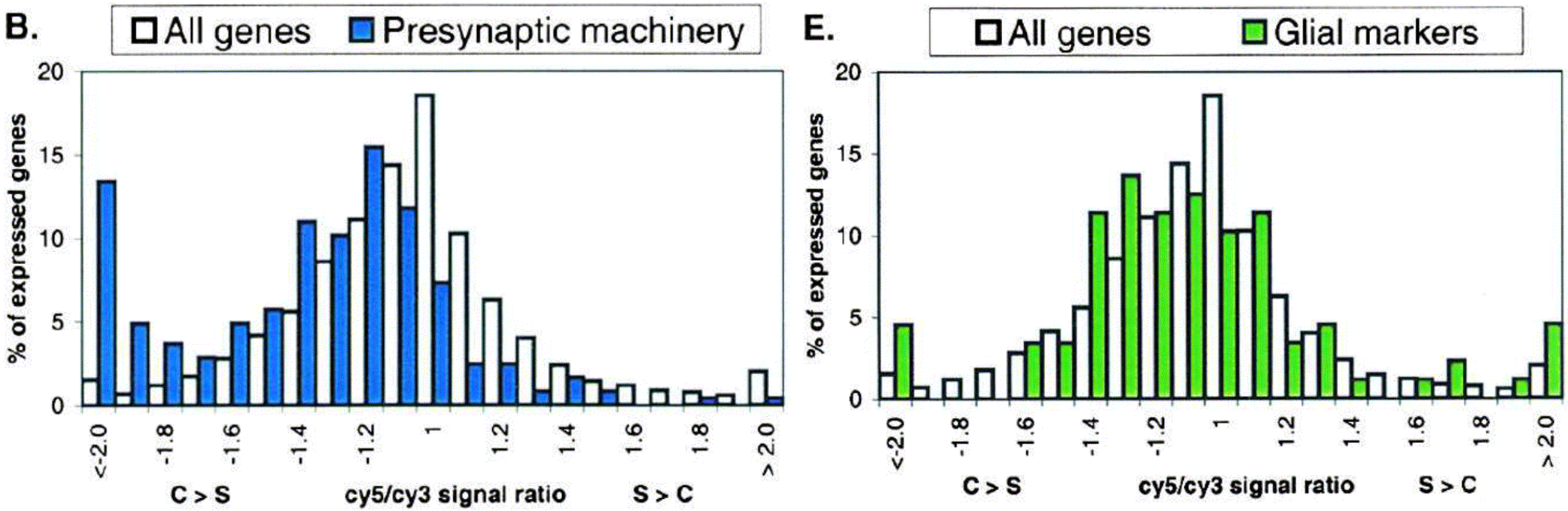
Tavazoie etal Nature Genetics 2000



Hypergeometric p-value.

The birthplaces of gene set analysis: II

Mirnics et al Neuron 2000



Molecular Characterization of Schizophrenia Viewed by Microarray Analysis of Gene Expression in Prefrontal Cortex.

A Formalism for Two-Stage Gene Set Analysis

Binary response vector Y (phenotype, class label, case-control...)
one for each of N samples

$G \times N$ matrix X of genetic information on samples

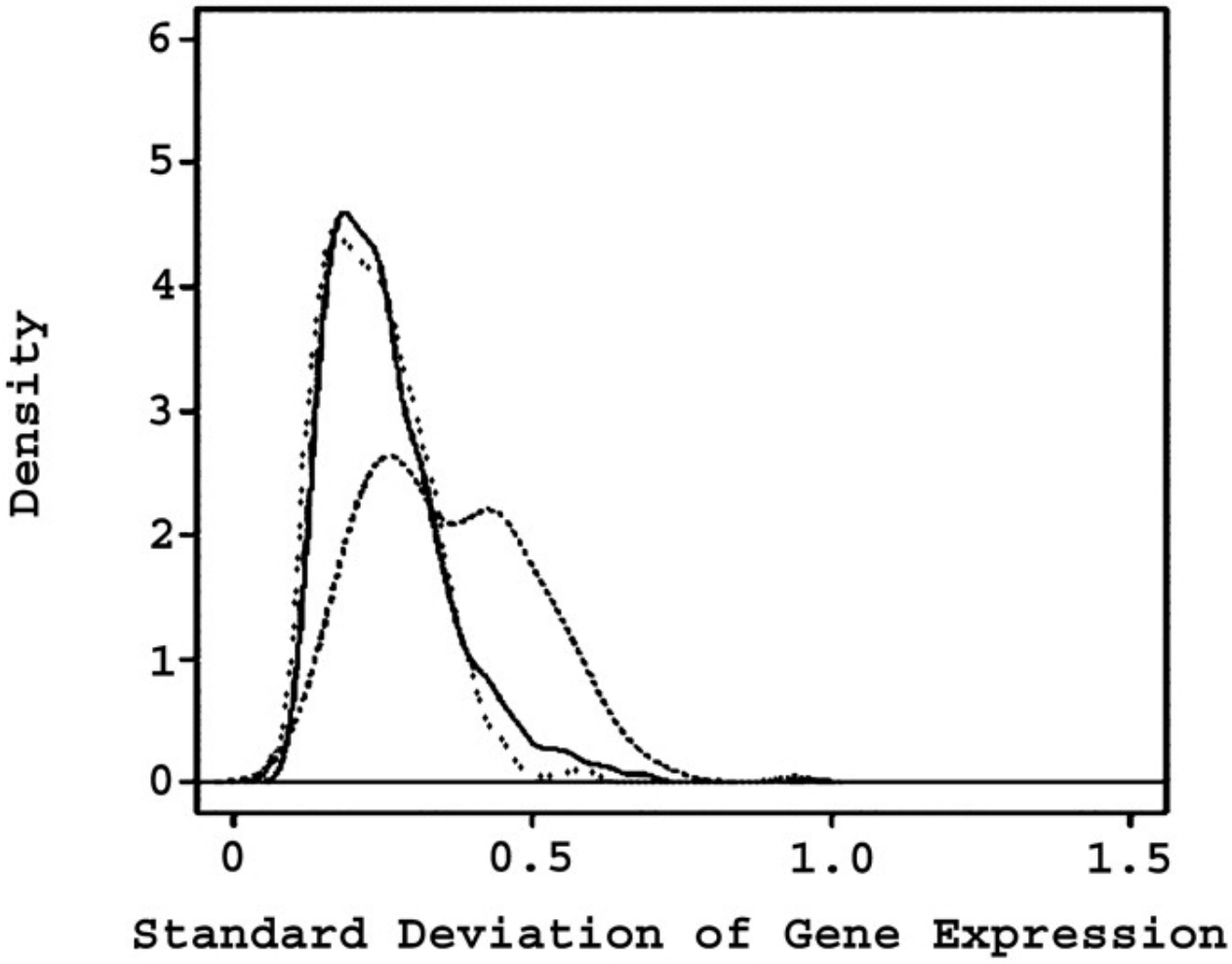
$G \times S$ binary membership matrix M

Stage I *Testing of differences between groups for each gene.* Compute for each gene g a score $s_g(X, Y)$, capturing the relationship between the genomic measurements and a phenotype of interest.

Stage II *Testing of differences in scores between sets.* Take the scores computed in Stage I as data, and look for association between the scores and the columns of M.

An Early Gene Set Analysis

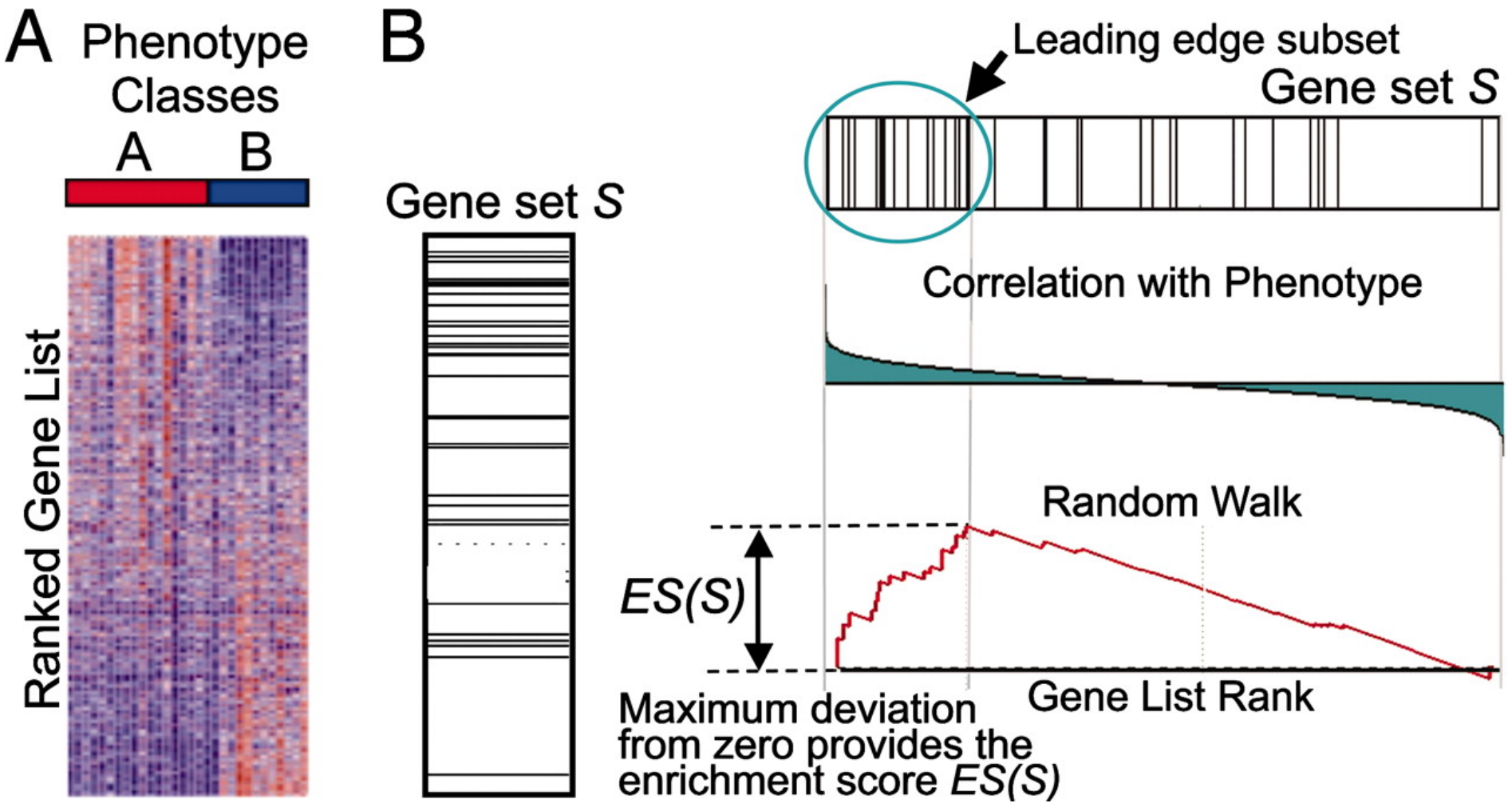
Chowers etal Human Molecular Genetics, 2003



Distribution of standard deviations for expression ratios of all genes of known function on the array (solid line), photoreceptor genes (dashed line), and genes involved in cell proliferation (dotted line).

Gene Set Enrichment Analysis

Mootha et al Nature Genetics, 2003; Subramanian PNAS 2005



Caveats I: Biology

- SET QUALITY
- SET OVERLAP
- TISSUE SPECIFICITY
- PATHWAY TOPOLOGY

Caveats II: Statistics

- GENES ARE NOT INDEPENDENT
- WHAT IS THE NULL HYPOTHESIS?
- BIG SET BIAS

Outline, References and Acknowledgments

MANY TECHNOLOGIES

S. Tyekucheva, L. Marchionni, R. Karchin and G. Parmigiani Integrating diverse genomic data using gene sets . *Genome Biol.*, 12: R105, 2011.

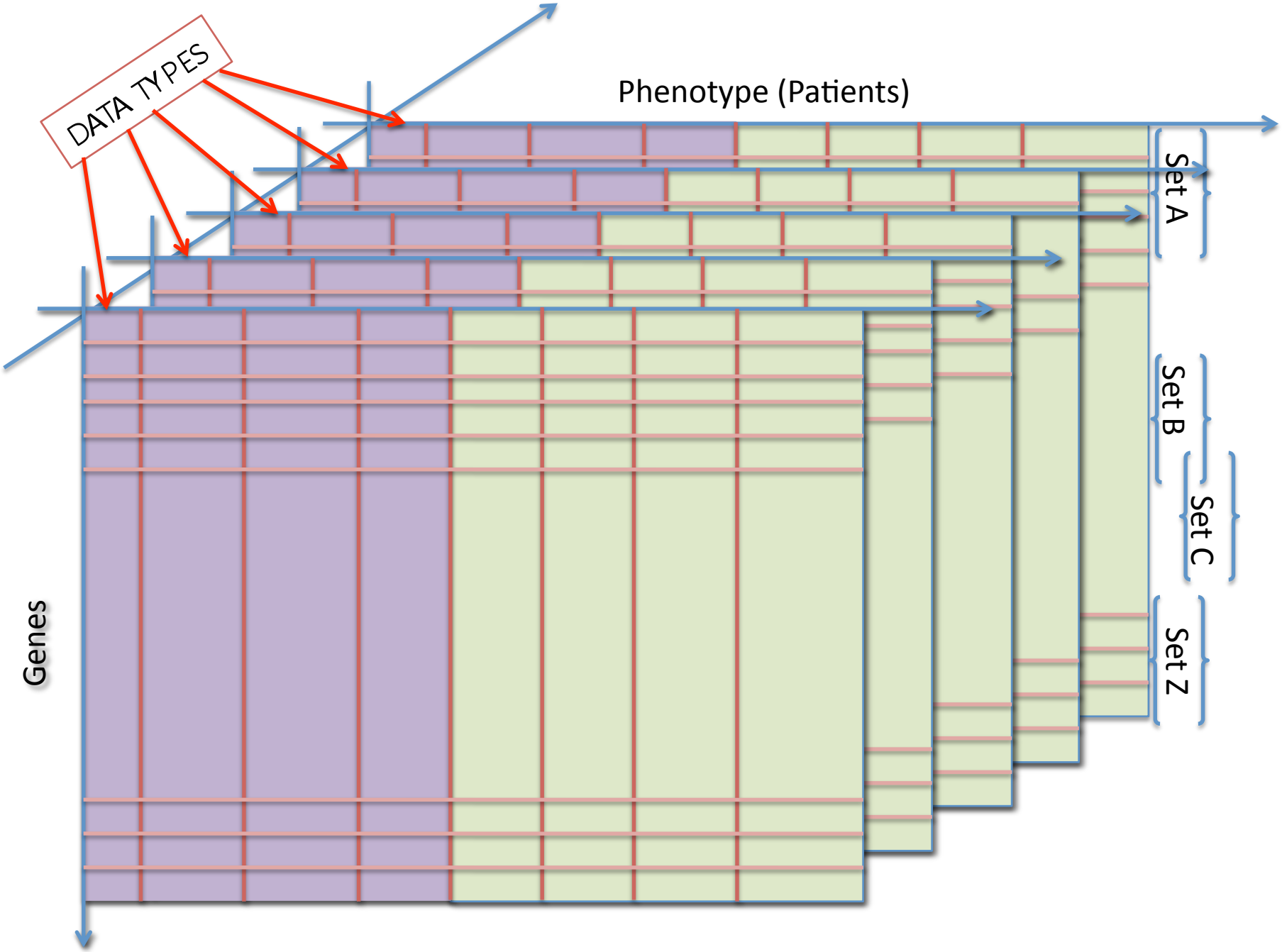
ATOMS

S.M. Boca, H. Corrada Bravo, B. Caffo, J.T. Leek and G. Parmigiani. A decision-theory approach to interpretable set analysis for high-dimensional data. *JHU Biostat Working Paper 211*, 2010.

PATIENTS

S.M. Boca, K.W. Kinzler, V.E. Velculescu, B. Vogelstein and G. Parmigiani. Patient oriented gene-set analysis for cancer mutation data. *Genome Biol.*, 11: R112, 2010.

Multiple data types



Gene-centric approaches for multiple data types

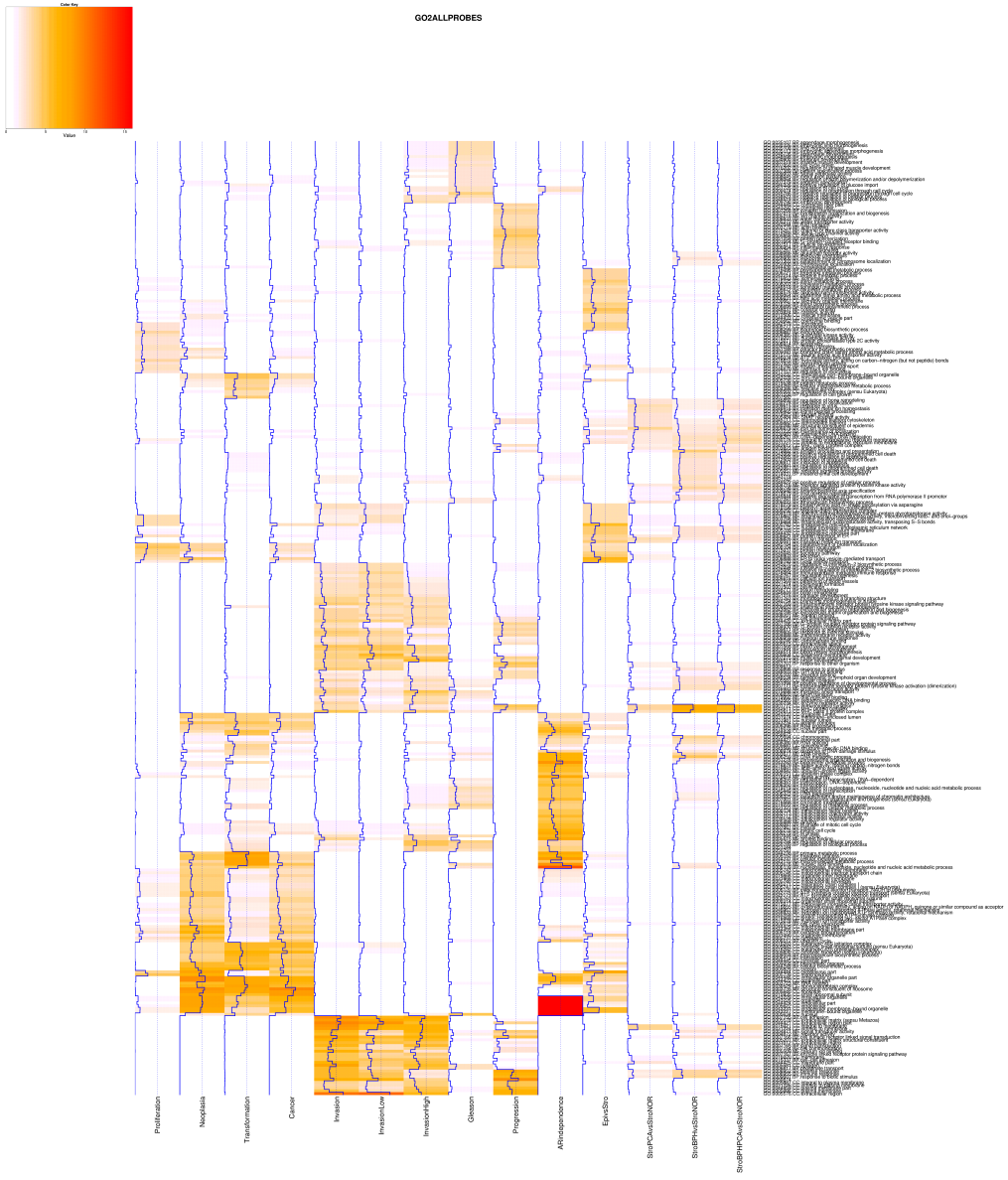
Binary response vector Y (phenotype, class label, case-control...)

$G \times N$ matrix X of genetic information on samples

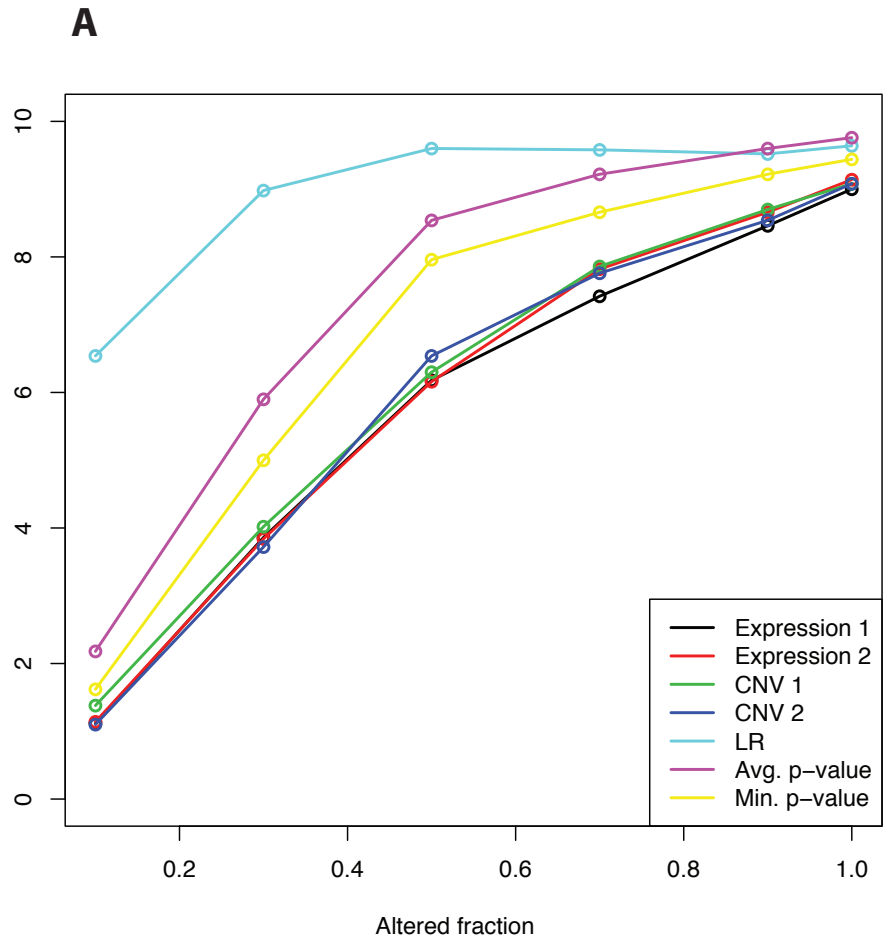
$G \times S$ binary membership matrix M

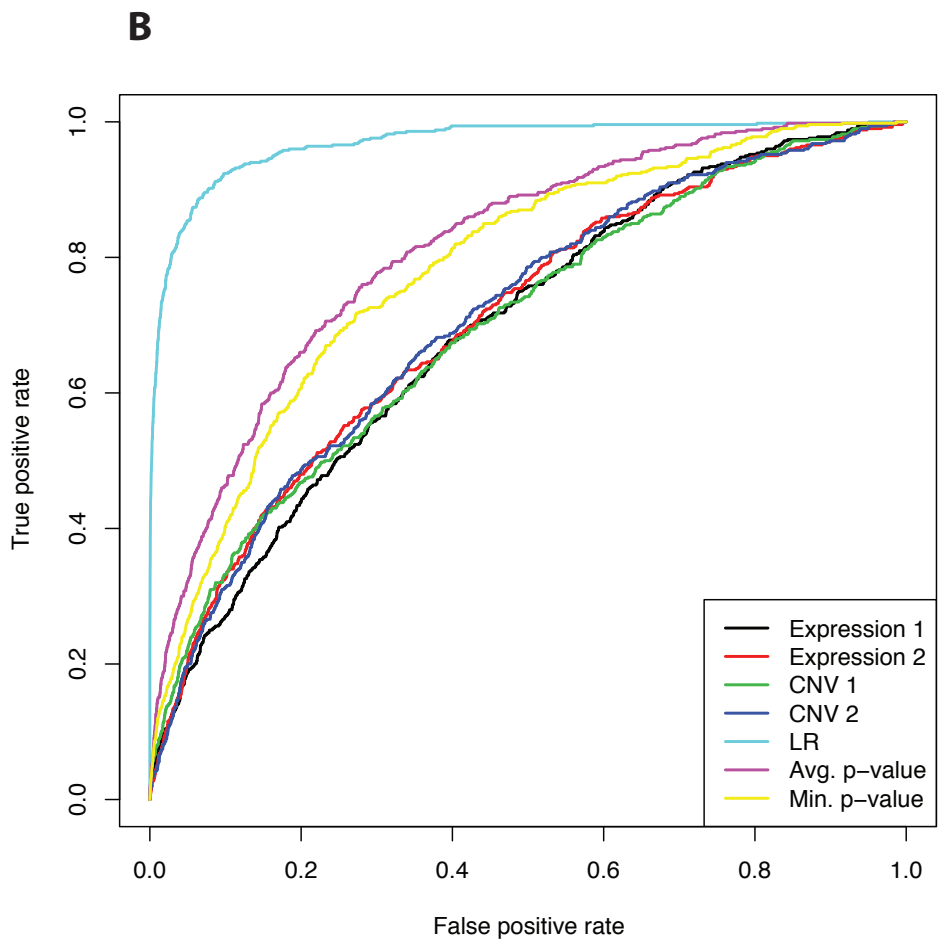
Stage I	Stage II	
$s_g(X^1, \ldots, X^D, Y)$	$t_s(\mathbf{s}, M_s)$	Integration
$s_g^1(X^d, Y) \ldots s_g^D(X^d, Y)$	$t_s(\mathbf{s}^1 \ldots \mathbf{s}^D, M_s)$	Meta-analysis
$s_g^1(X^d, Y) \ldots s_g^D(X^d, Y)$	$t_s(\mathbf{s}^1, M_s) \ldots t_s(\mathbf{s}^D, M_s)$	Visualization

Clustering Sets to Compare Experiments



Integrative more powerful than Meta-analytic

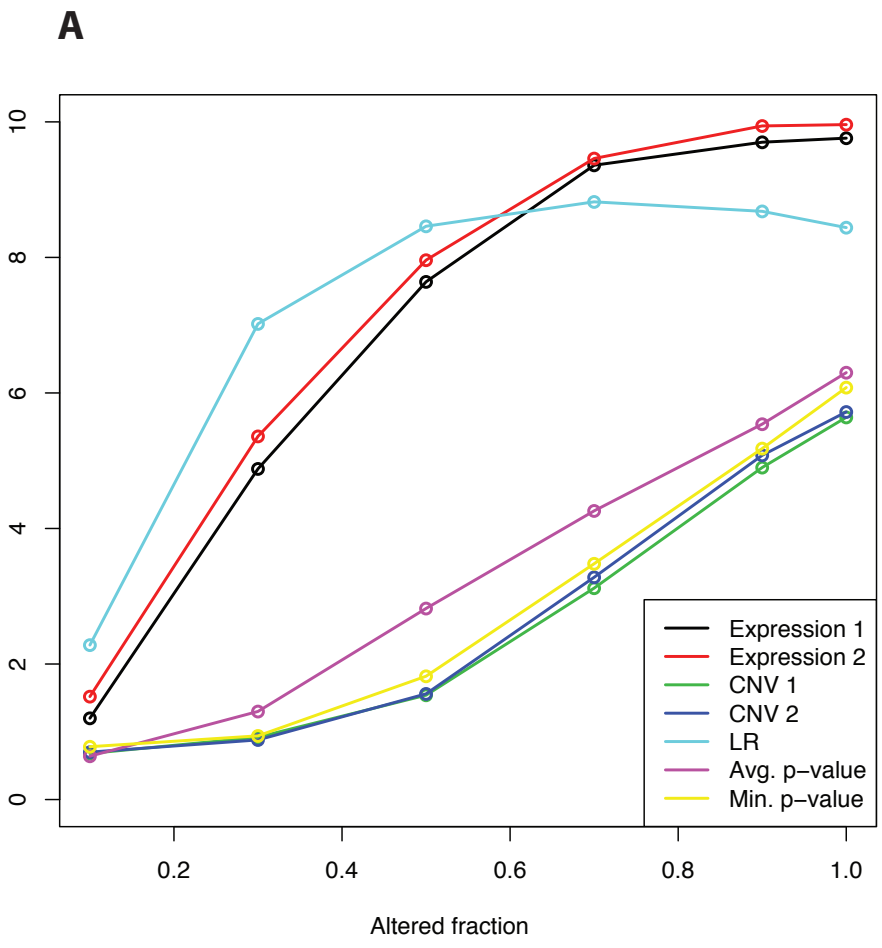


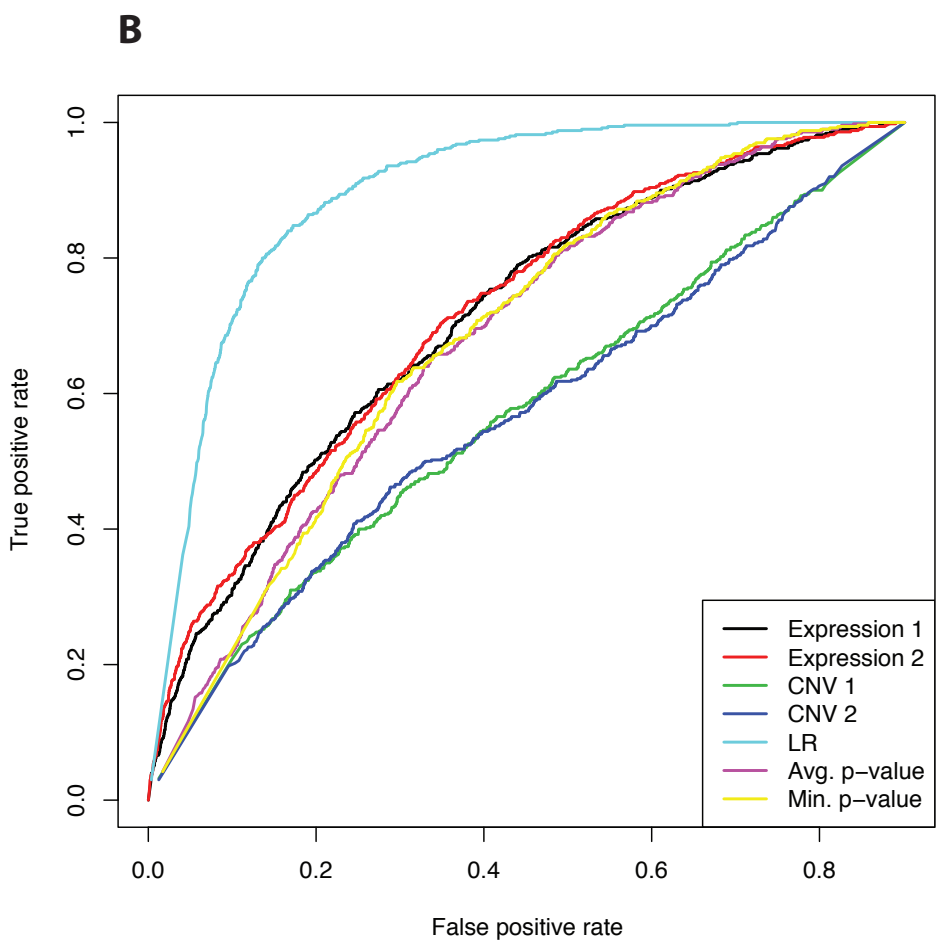


Independent Sets

ROC for classification of spiked-in sets

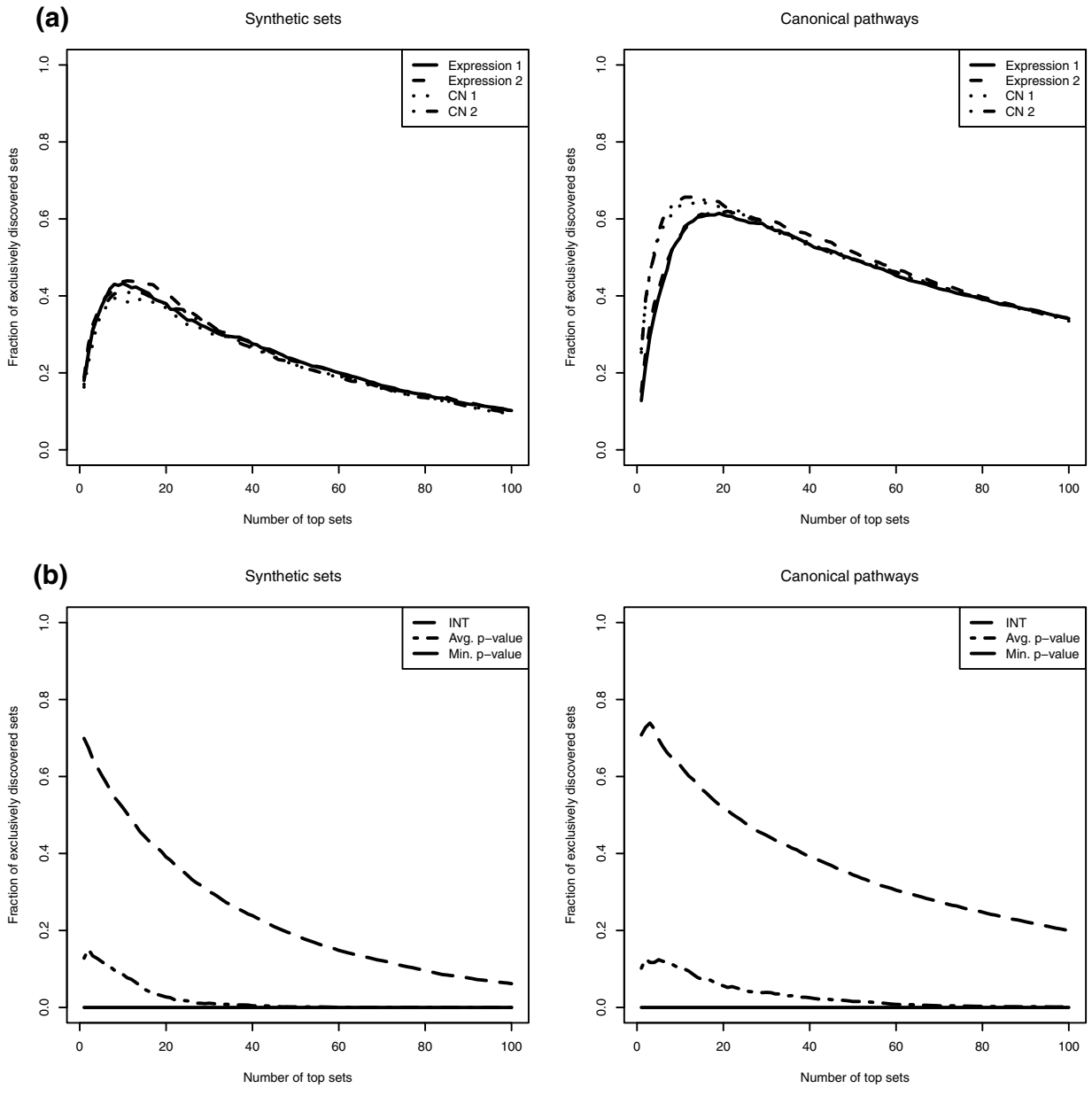
Integrative more robust than Meta-analytic



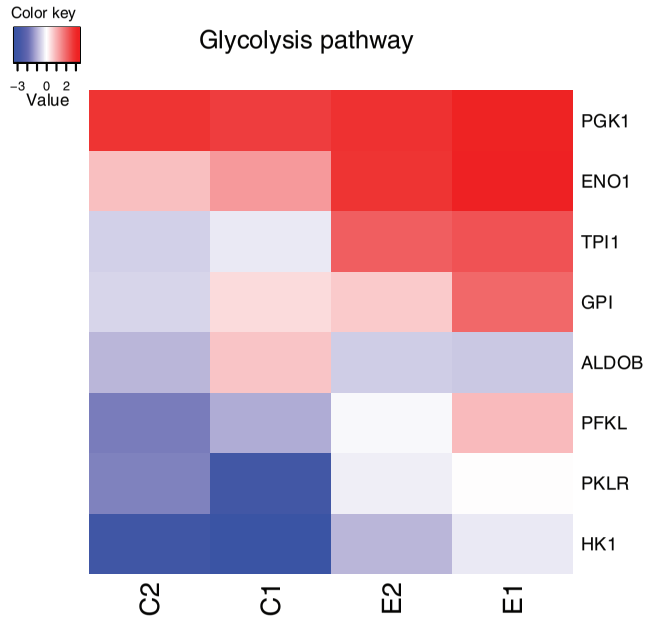


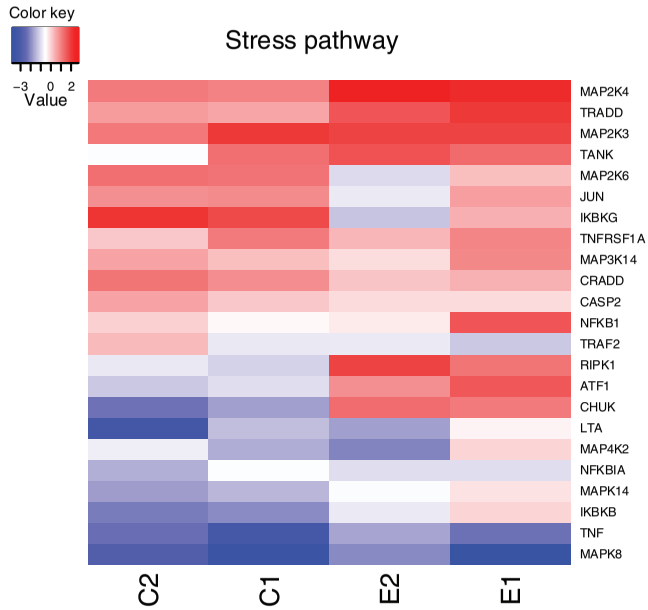
Chromosomal Segments
ROC for classification of spiked-in sets

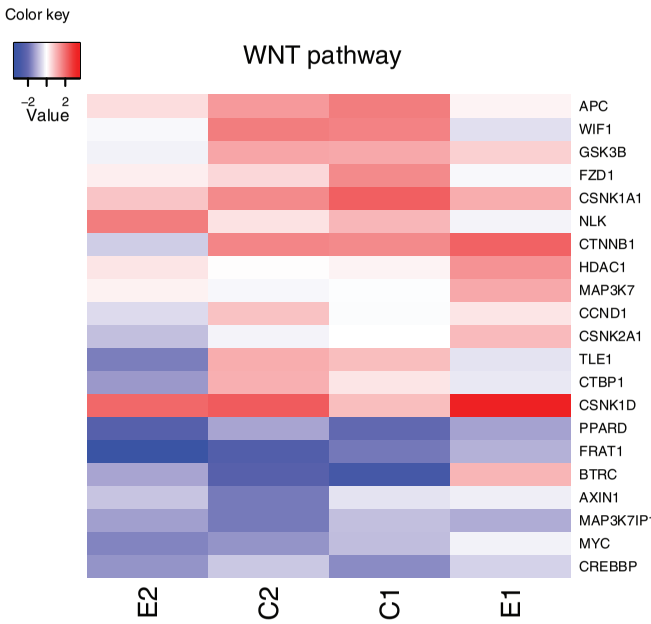
Integrative discovers novel sets



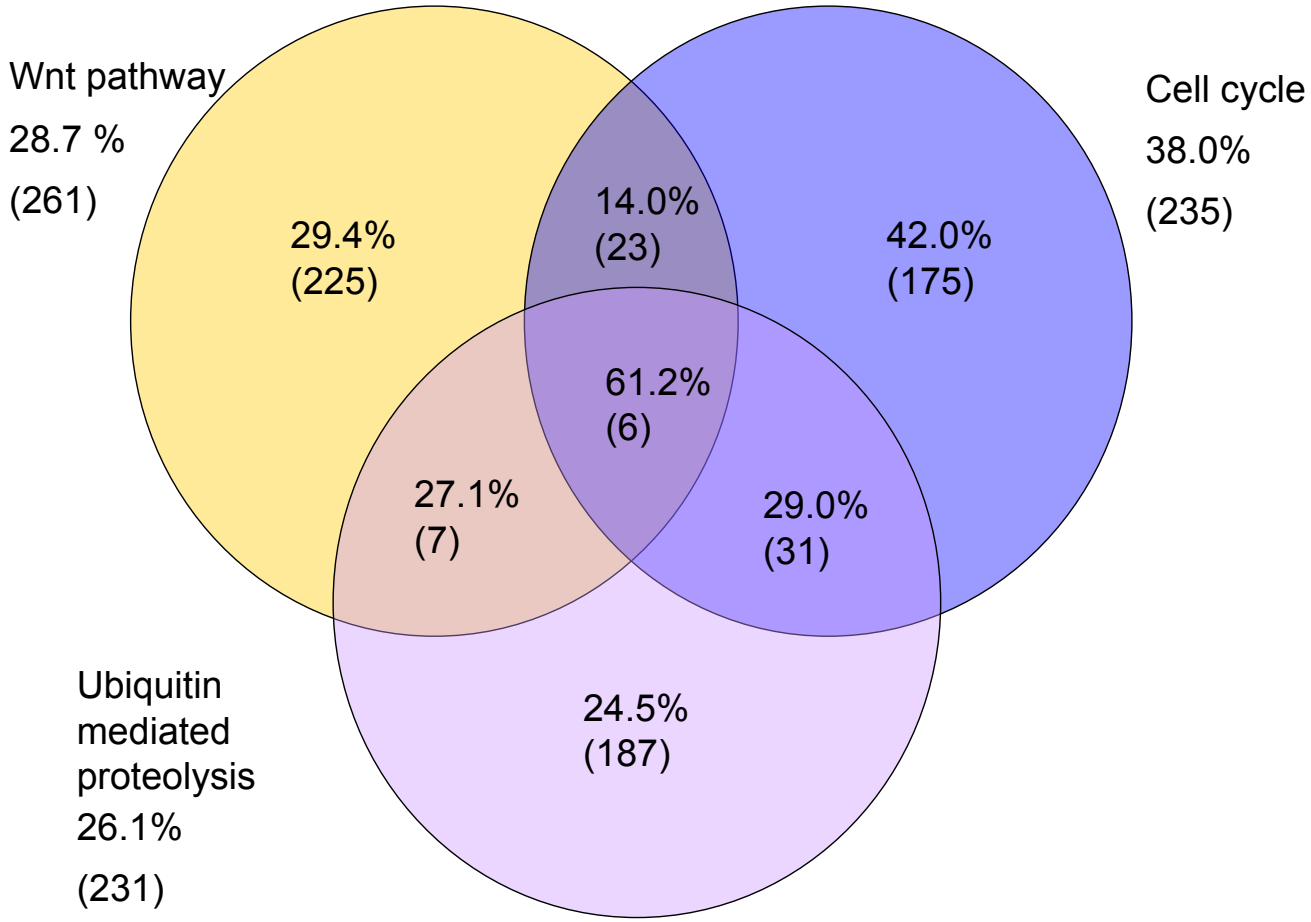
GBM


Color key
−3 0 2
Value
Glycolysis pathway
PGK1
ENO1
TPI1
GPI
ALDOB
PFKL
PKLR
HK1
C2
C1
E2
E1


Color key
−3 0 2
Value
Stress pathway
MAP2K4
TRADD
MAP2K3
TANK
MAP2K6
JUN
IKBKG
TNFRSF1A
MAP3K14
CRADD
CASP2
NFKB1
TRAF2
RIPK1
ATF1
CHUK
LTA
MAP4K2
NFKBIA
MAPK14
IKBKB
TNF
MAPK8
C2
C1
E2
E1


Color key
−2 2
Value
WNT pathway
APC
WIF1
GSK3B
FZD1
CSNK1A1
NLK
CTNNB1
HDAC1
MAP3K7
CCND1
CSNK2A1
TLE1
CTBP1
CSNK1D
PPARD
FRAT1
BTRC
AXIN1
MAP3K7IP1
MYC
CREBBP
E2
C2
C1
E1

Enrichment of 3 intersecting pathways for ER+ BC



X% (Y): X% out of Y genes are estimated to have densities from the alternative distribution.

Decision Theoretic Angle

- Divide genes into atoms based on sets.
- Truth is the list of alternatives.
- We search for estimators among the unions of atoms.
- The estimators are based on the loss function:

$$(1-w)\times \text{ \# of FD } + w \times \text{ \# of MD.}$$

- The posterior expected loss is:

$$(1-w)\times \text{EFD} + w \times \text{EMD.}$$

Atomic False Discovery Rate

- We define the *atomic false discovery rate* for atom A as:

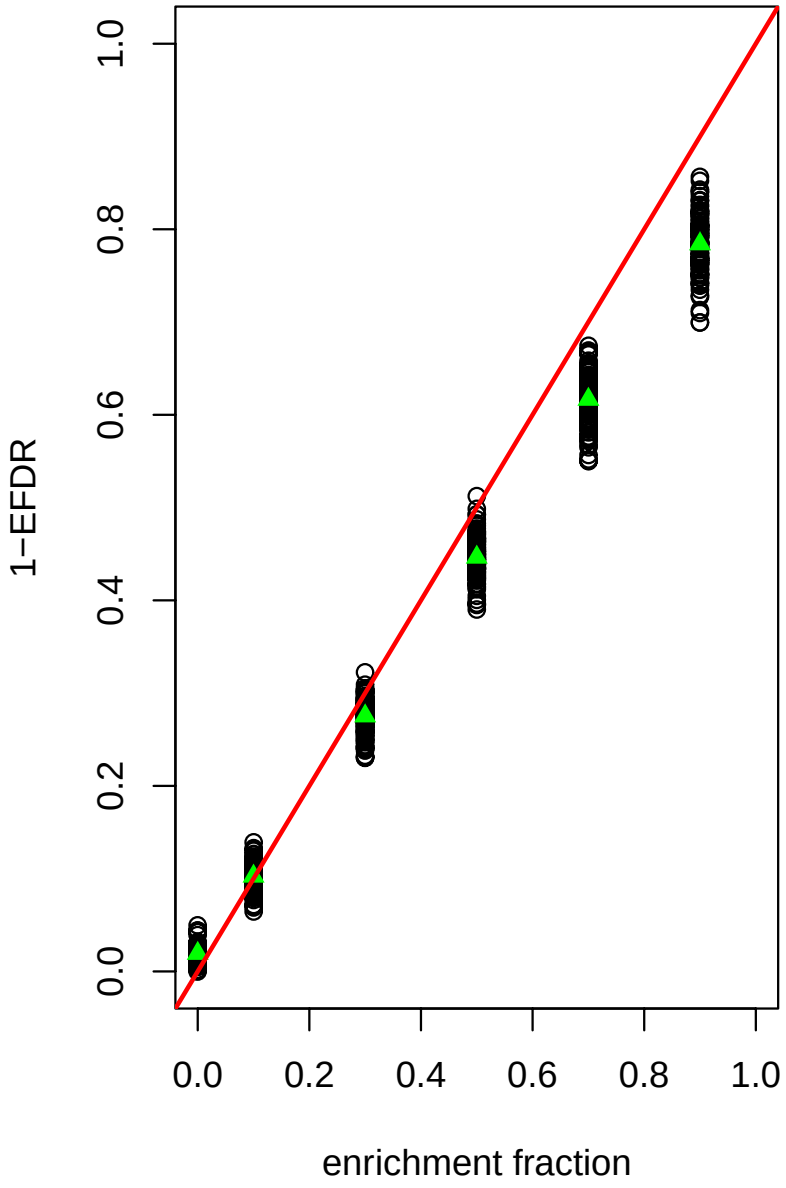
$$\mathrm{AFDR}(A) = \mathrm{FD}(A)/n_A.$$

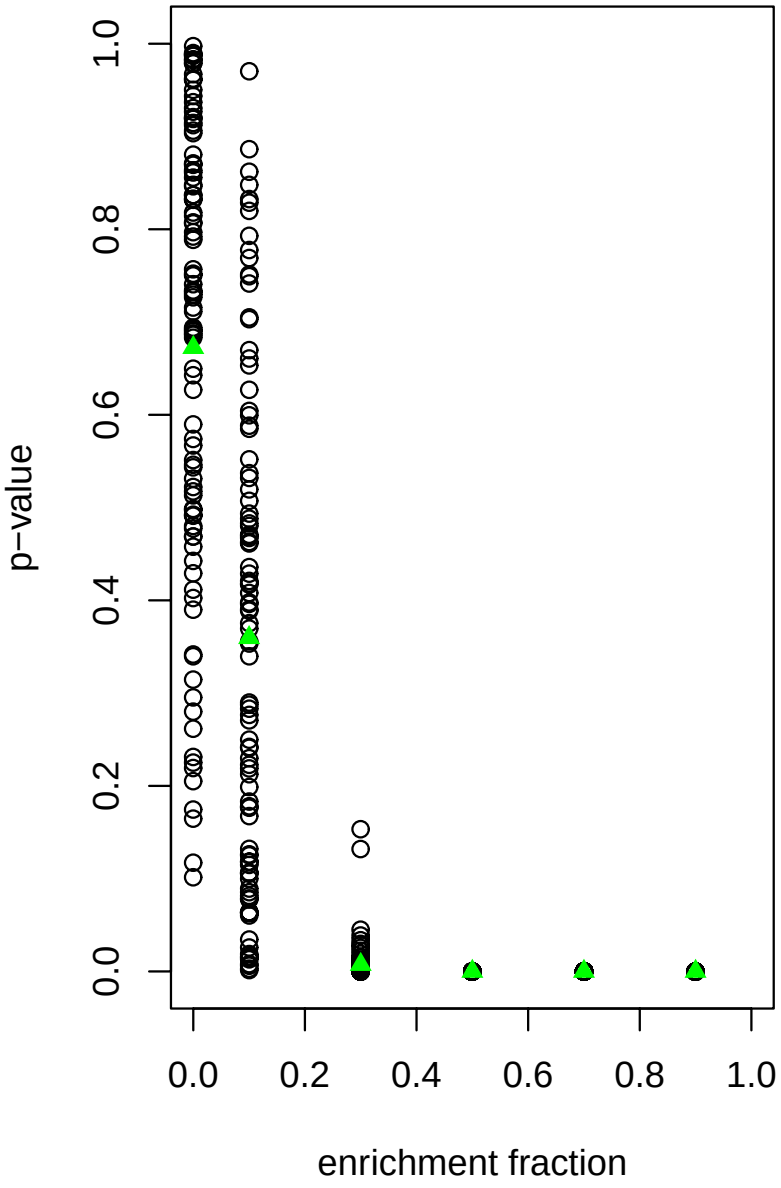
- **Theorem (Boca et al., 2010)** Atom A is included in the Bayes estimator if and only if the atomic FDR is thresholded by w:

$$\widehat{\mathrm{AFDR}}(A) \leq w.$$

- $1 - \widehat{\mathrm{AFDR}}$ estimates the fraction of alternatives in an atom.

Atomic FDR measures enrichment





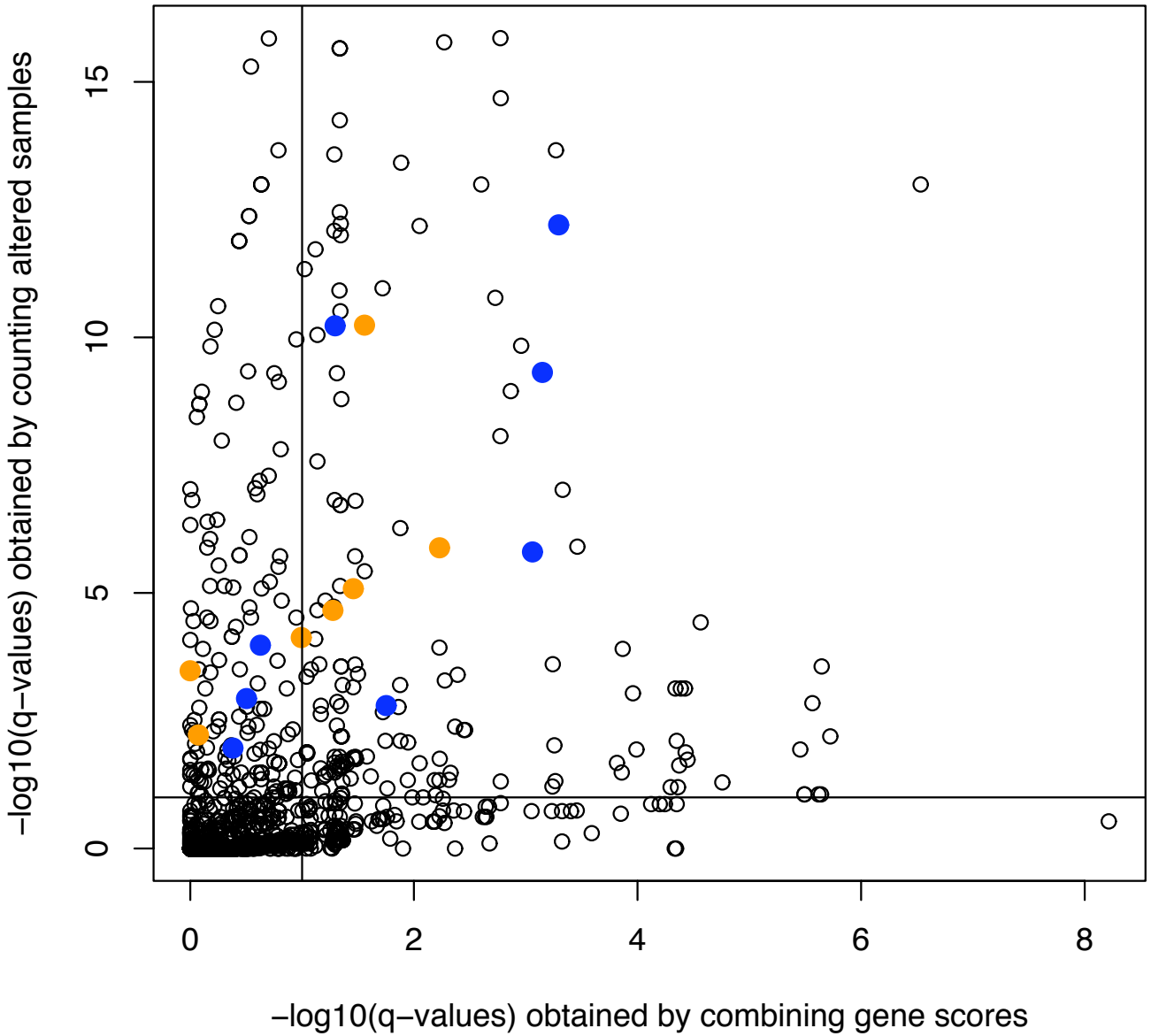
Altered Pathways in Glioblastoma

Parsons 2008

	TP53 pathway				PI3K Pathway					RB1 pathway			
Tumor sample	TP53	MDM2	MDM4	All genes	PTEN	PIK3CA	PIK3R1	IRS1	All genes	RB1	CDK4	CDKN2A	All genes
Br02X	Del			Alt				Mut	Alt			Del	Alt
Br03X	Mut			Alt	Mut				Alt				
Br04X	Mut			Alt	Mut				Alt	Mut			Alt
Br05X			Amp	Alt		Mut			Alt			Del	Alt
Br06X												Del	Alt
Br07X	Mut			Alt	Mut				Alt	Del			Alt
Br08X												Del	Alt
Br09P	Mut			Alt							Amp		Alt
Br10P	Mut			Alt									
Br11P	Mut			Alt									
Br12P	Mut			Alt			Mut		Alt				
Br13X	Mut			Alt								Del	Alt
Br14X							Mut		Alt			Del	Alt
Br15X										Mut		Del	Alt
Br16X		Amp		Alt							Amp		Alt
Br17X					Mut				Alt			Del	Alt
Br20P													
Br23X	Mut			Alt	Del				Alt				
Br25X					Mut				Alt			Del	Alt
Br26X						Mut			Alt			Del	Alt
Br27P	Mut			Alt							Amp		Alt
Br29P	Mut			Alt									
Fraction of tumors with altered gene/pathway*	0.55	0.05	0.05	0.64	0.27	0.09	0.09	0.05	0.50	0.14	0.14	0.45	0.68

* Mut, mutated; Amp, amplified; Del, deleted; Alt, altered *Fraction of affected tumors in 22 Discovery Screen samples

Gene-centric vs Patient Centric Scores



Outline, References and Acknowledgments

MANY TECHNOLOGIES

S. Tyekucheva, L. Marchionni, R. Karchin and G. Parmigiani Integrating diverse genomic data using gene sets . *Genome Biol.*, 12: R105, 2011.

ATOMS

S.M. Boca, H. Corrada Bravo, B. Caffo, J.T. Leek and G. Parmigiani. A decision-theory approach to interpretable set analysis for high-dimensional data. *JHU Biostat Working Paper 211*, 2010.

PATIENTS

S.M. Boca, K.W. Kinzler, V.E. Velculescu, B. Vogelstein and G. Parmigiani. Patient oriented gene-set analysis for cancer mutation data. *Genome Biol.*, 11: R112, 2010.