Introduction to Gene Sets Analysis

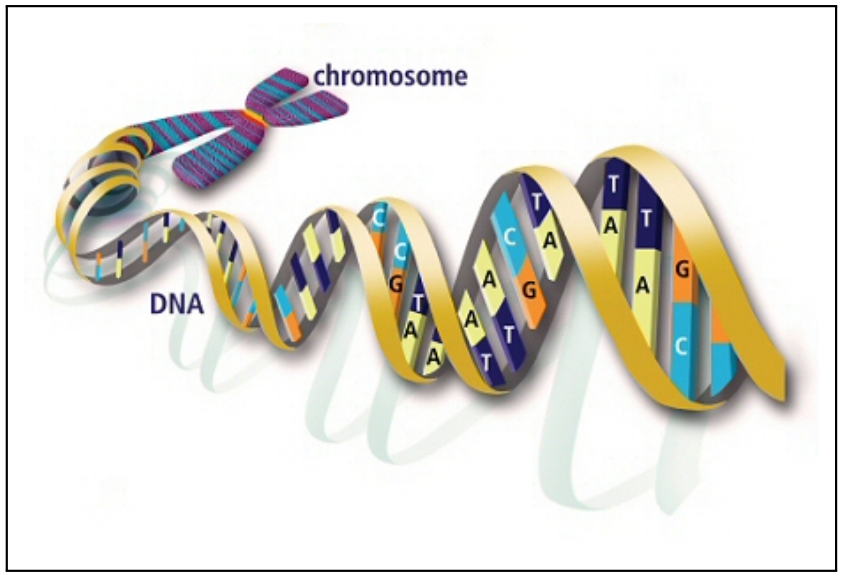
Svitlana Tyekucheva

Dana-Farber Cancer Institute

May 15, 2012

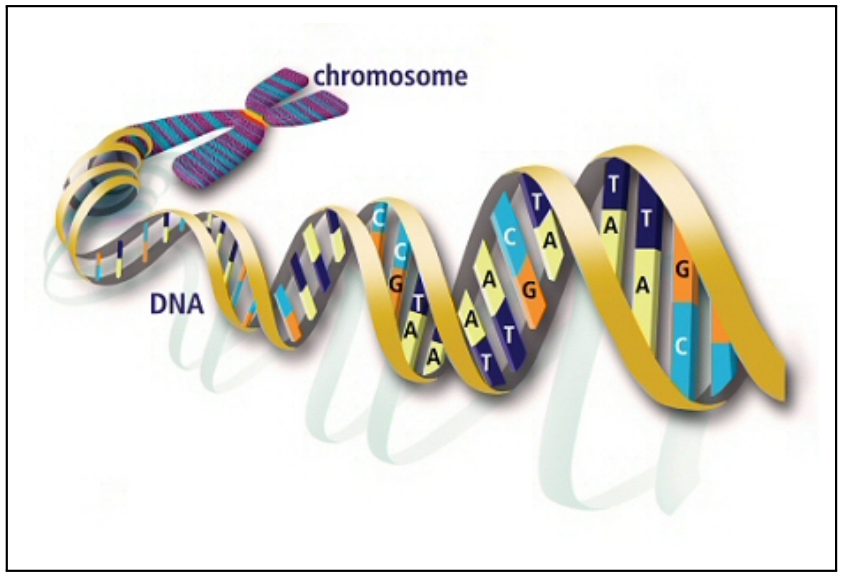
Introduction

- Various measurements: gene expression, copy number variation, methylation status, mutation profile, etc.
- Main question: what makes two (or more) phenotypes different (tumor vs normal, short vs long survival, molecular subtypes, etc.)
- Major problems: data are noisy, the signal is subtle, results are hard to interpret in a biologically meaningful way



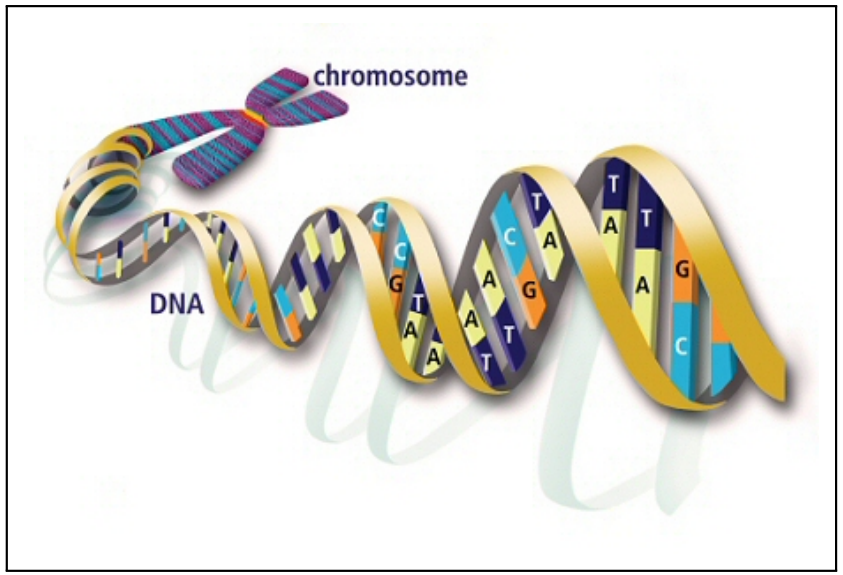
Introduction

- Various measurements: gene expression, copy number variation, methylation status, mutation profile, etc.
- Main question: what makes two (or more) phenotypes different (tumor vs normal, short vs long survival, molecular subtypes, etc.)
- Major problems: data are noisy, the signal is subtle, results are hard to interpret in a biologically meaningful way

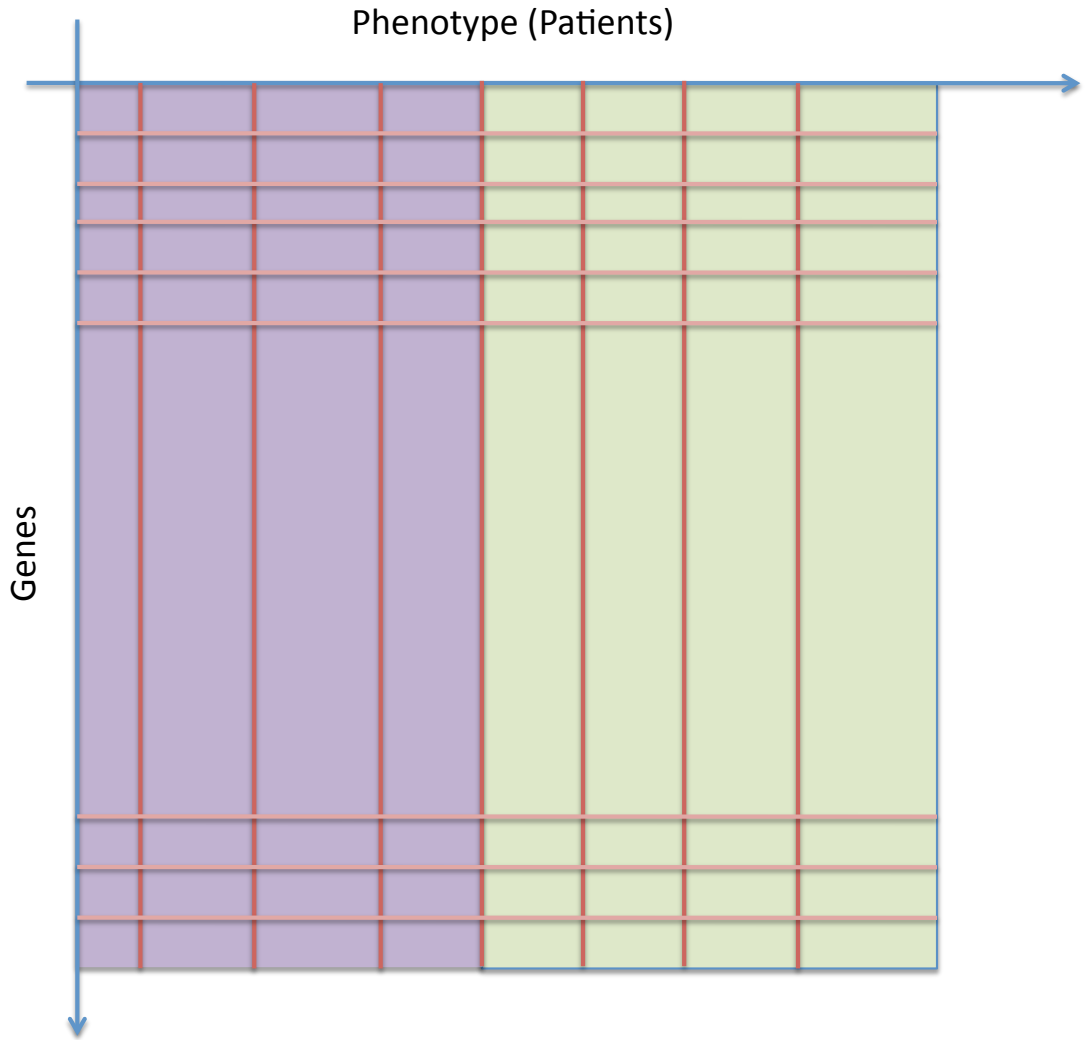


Introduction

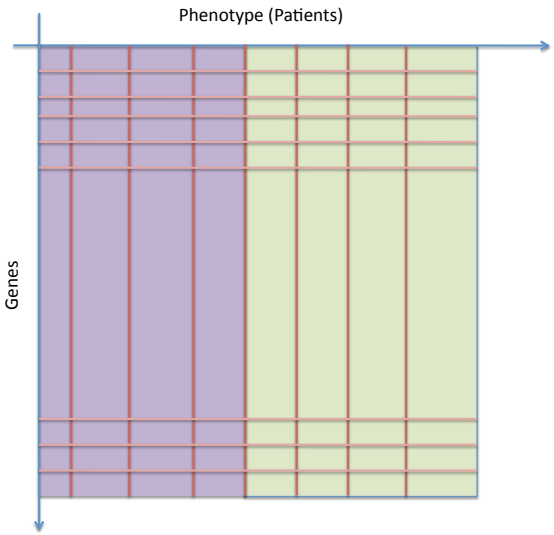
- Various measurements: gene expression, copy number variation, methylation status, mutation profile, etc.
- Main question: what makes two (or more) phenotypes different (tumor vs normal, short vs long survival, molecular subtypes, etc.)
- Major problems: data are noisy, the signal is subtle, results are hard to interpret in a biologically meaningful way



Basic setup



Traditional approach



- Find some gene-to-phenotype association score (say, using t-test, ANOVA, etc.)
- Rank genes according to the score and take top-'your favorite number' or assess statistical significance of the discriminating power of each gene

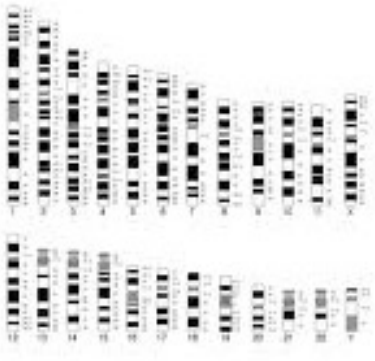
* Multiple testing

* It is not uncommon that similar studies report nonintersecting lists of "top genes"

* Individual genes might not contribute too much to the difference between phenotypes, together, though, they might!

Why gene sets?

- Instead of 'list of genes' - think about 'list of gene sets'



MSigDB
Molecular Signatures Database

Why gene sets?

- Gene sets encompass larger amount of biological information, this helps to make results more interpretable.
- Information on the gene set level is comparable across different types of measurements (different platforms)
- Multiple testing issue: we will usually (not always...) have less sets than individual genes
- Same biological mechanisms can manifest in different parts of the pathway and via different alterations in different subjects (!!!)

Glioblastoma multiforme

- Glioblastoma multiforme (GBM) - very aggressive and the most common primary brain tumor
- primary (90%-95%), secondary (5%-10%)
- median survival 15 month

Pathways altered in GBM

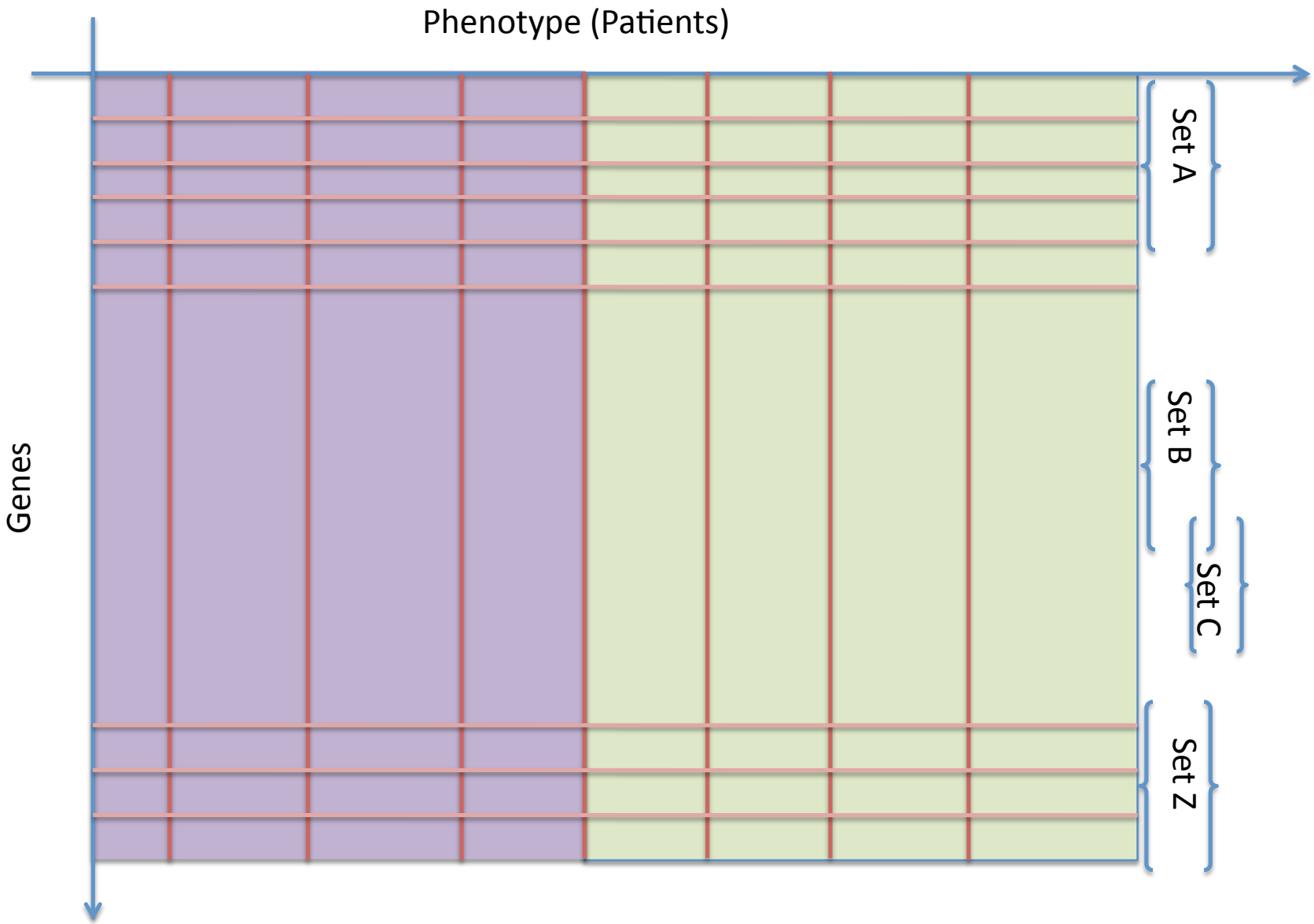
Table 3. Mutations of the TP53, PI3K, and RB1 pathways in GBM samples. Mut, mutated; Amp, amplified; Del, deleted; Alt, altered.

	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

*Fraction of affected tumors in 22 Discovery Screen samples.

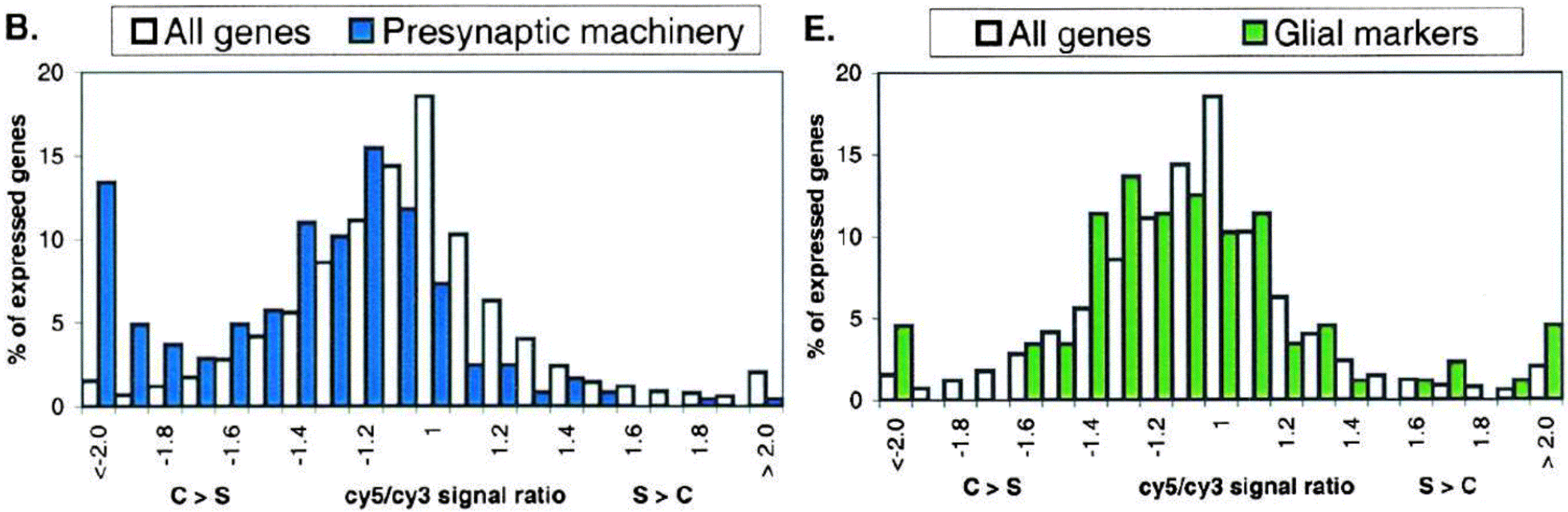
(Parsons, 2008)

Gene sets approach



The birthplace of gene set analysis

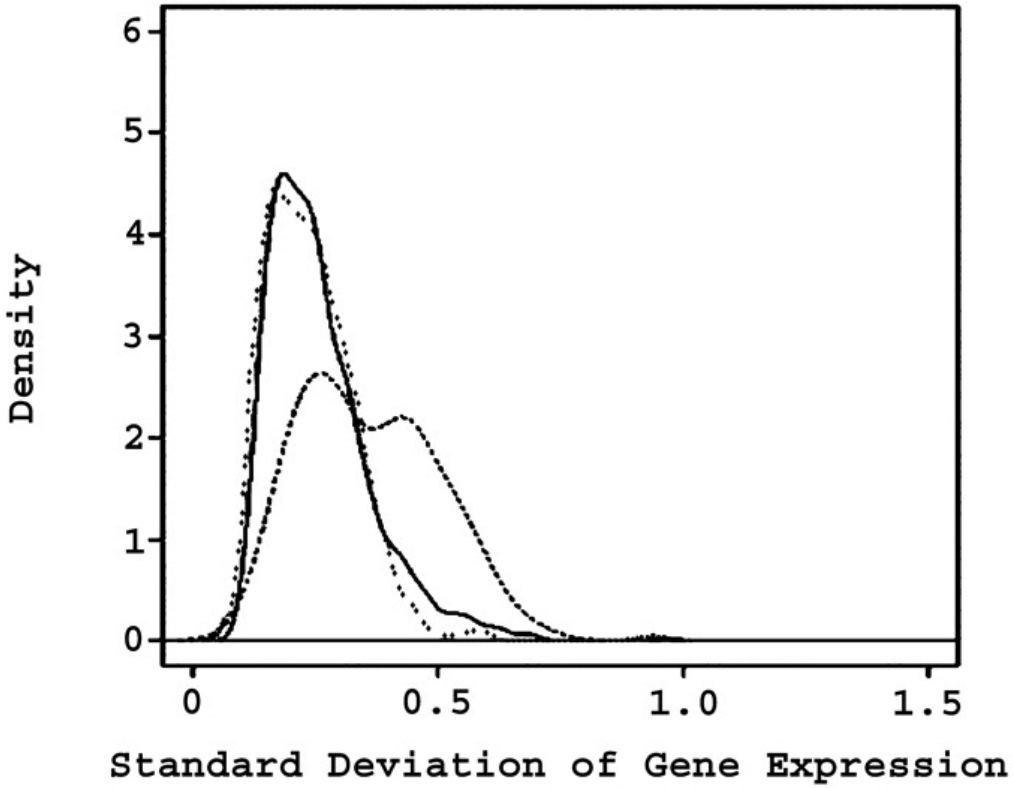
Mirnics et al Neuron 2000



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

An Early Gene Set Analysis

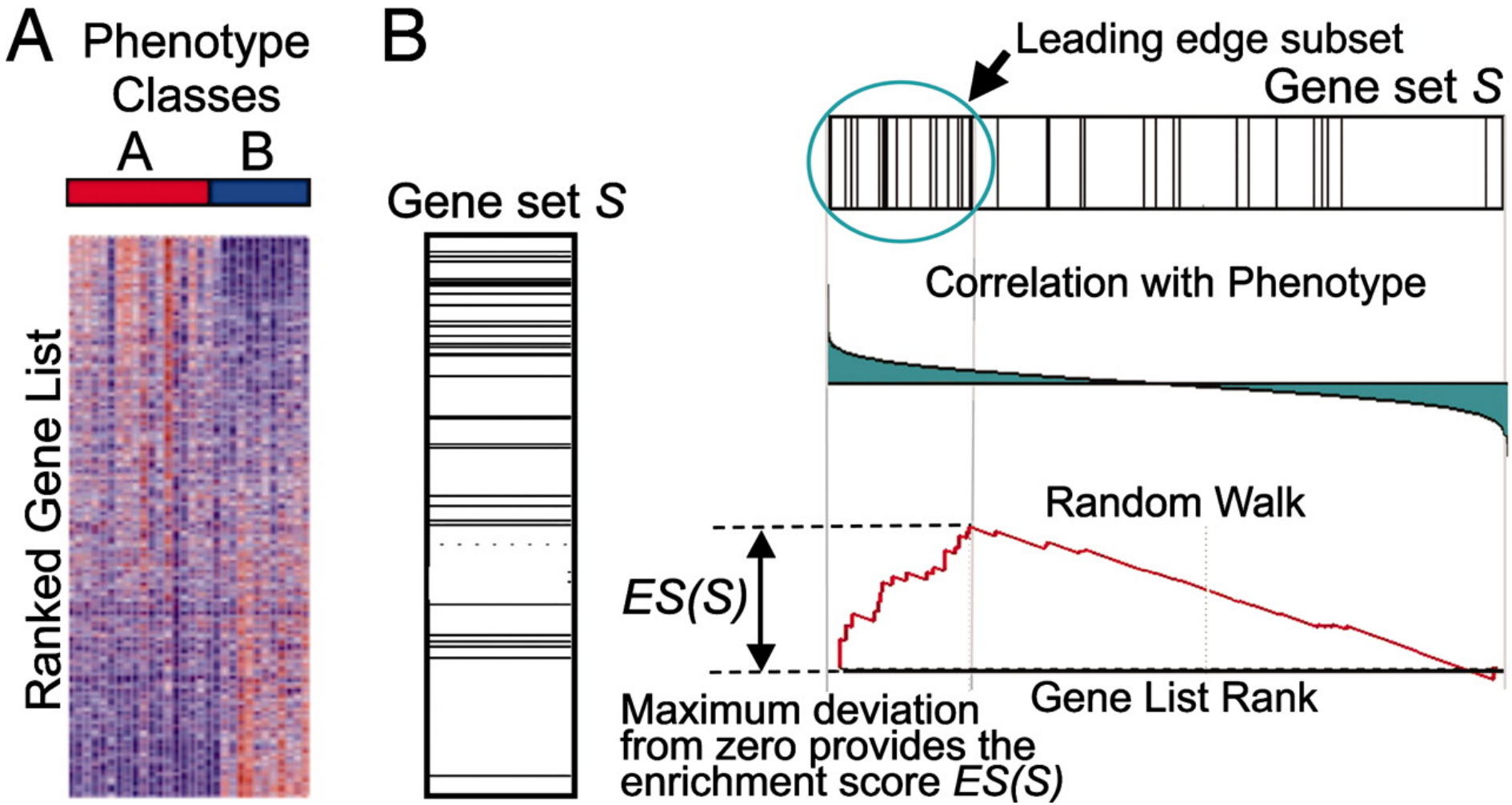
Chowers et al. Human Molecular Genetics, 2003



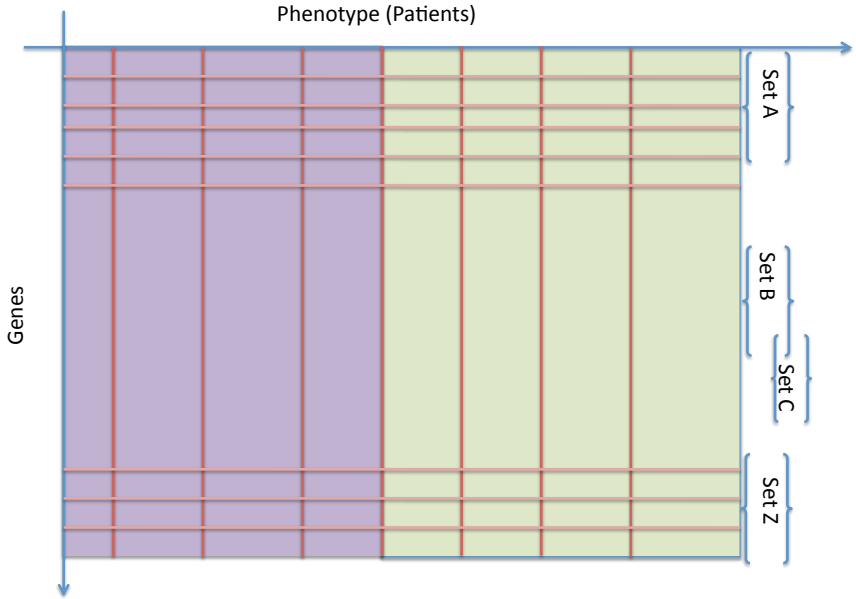
Distribution of standard deviations for expression ratios of all genes of known function on the array (solid line), photoreceptor genes (dashed line), and genes involved in cell proliferation (dotted line).

Gene Set Enrichment Analysis

Mootha et al Nature Genetics, 2003; Subramanian PNAS 2005

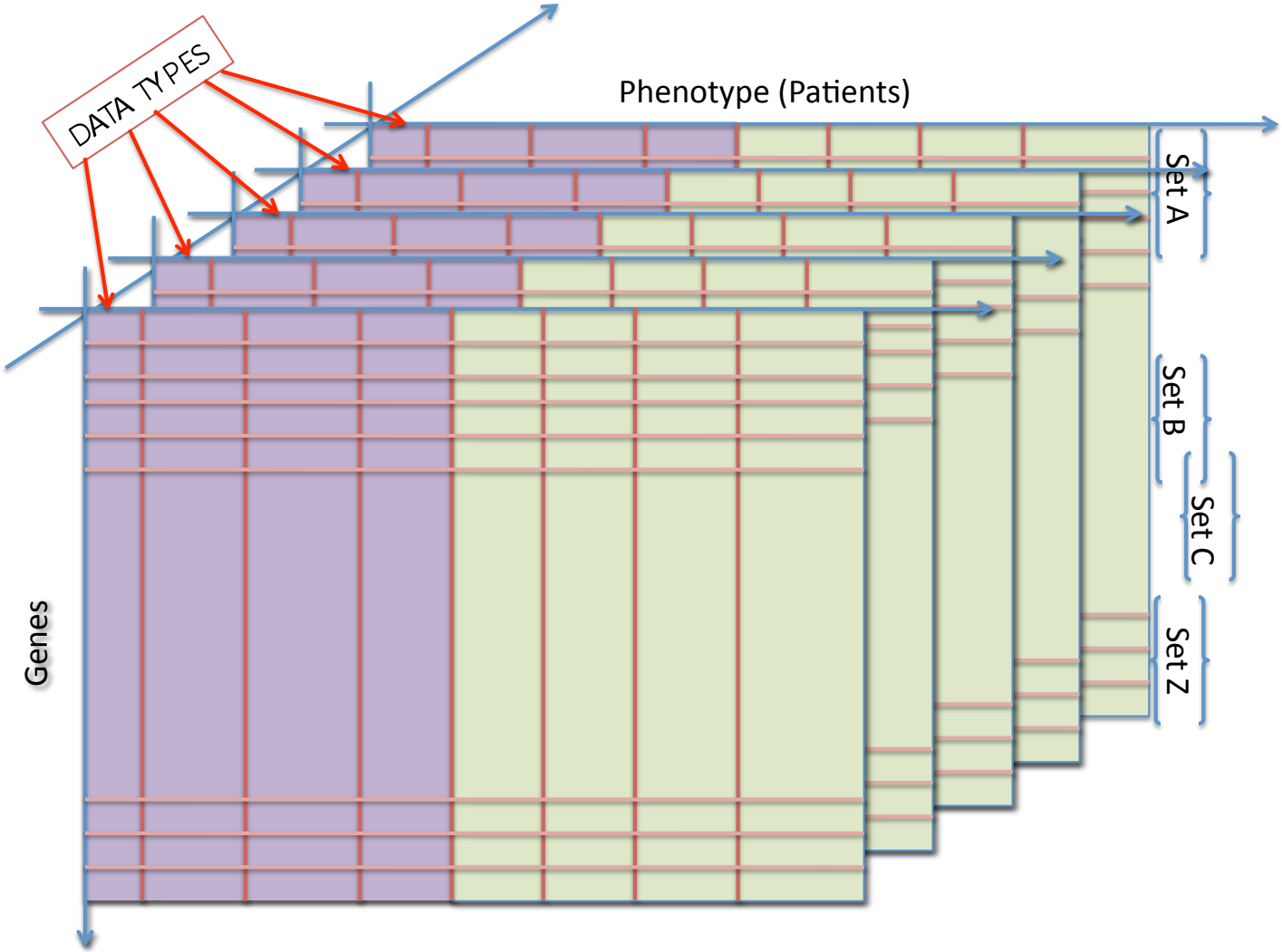


Two stage approach

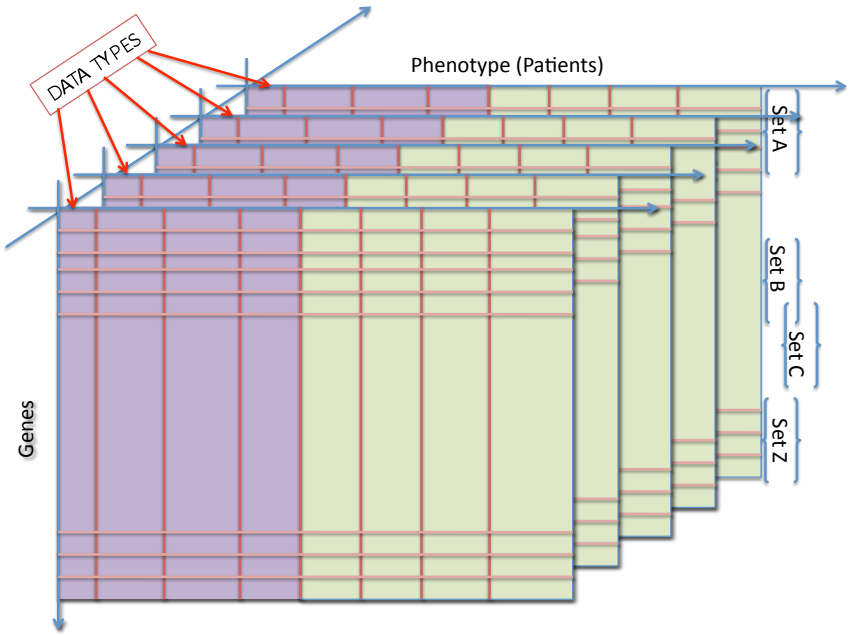


- Stage I: compute some gene-to-phenotype association scores (say, t-values) and rank genes according to these values
- Stage II : check whether the distribution of the ranks is different in a given set vs the rest of the genes (sets) – competitive null, or vs the distribution of the ranks in the same set when there is no association with the phenotype – self-contained null
- Infer enriched sets, say by ranking sets according to the outcome of the Mann-Whitney or Wilcoxon test.
- Supposedly, this approach allows to tease out subtle signals.

Now suppose we have a lot of data types



Suggested integration approaches



Integrate on

- Stage I (Integration+GSA) Compute gene-to-phenotype association scores using all available data types (say, using logistic regression or other linear model)

OR

- Stage II (GSA+Integration) Use, say, Wilcoxon p-values and take their geometric average, or take the smallest one across all data types (some consensus measurement)

Suggested integration approaches

Stage I integration: Integration+GSA. Heterogeneous data is integrated into a single gene-specific score $s_g(X^1, \ldots, X^D, Y)$ that draws from all the measurements available from gene g across all the dimensions studied, followed by one-dimensional GSA. *E.g.*:

$$\phi\big(E(Y_i|X^1_{gi}, \ldots, X^d_{gi})\big) = \sum_{d\in\{1,\ldots,D\}} X^d_{gi}\beta^d_g$$

where ϕ is a link function and i the biological sample. For each gene, the Stage I score can be provided by a measure of the overall fit of the model, say, a likelihood ratio for comparing this model to the "null" model in which all the β^d_g coefficients are zero. In Stage II these scores can then be analyzed using traditional methods, finally giving set-specific scores $t_s(s, M_s)$.

Suggested integration approaches

Stage II integration: GSA+Integration. This approach starts as a standard one-dimensional GSA: we determine a gene-to-phenotype association scores separately for each dimension $s_g^d(X^d, Y)$, and in Stage II we compute set-specific scores $t_s^d(s, M_s)$, $d \in 1, \ldots, D$, for each dimension. Next these scores (e.g. p-values) can be integrated, say, by averaging:

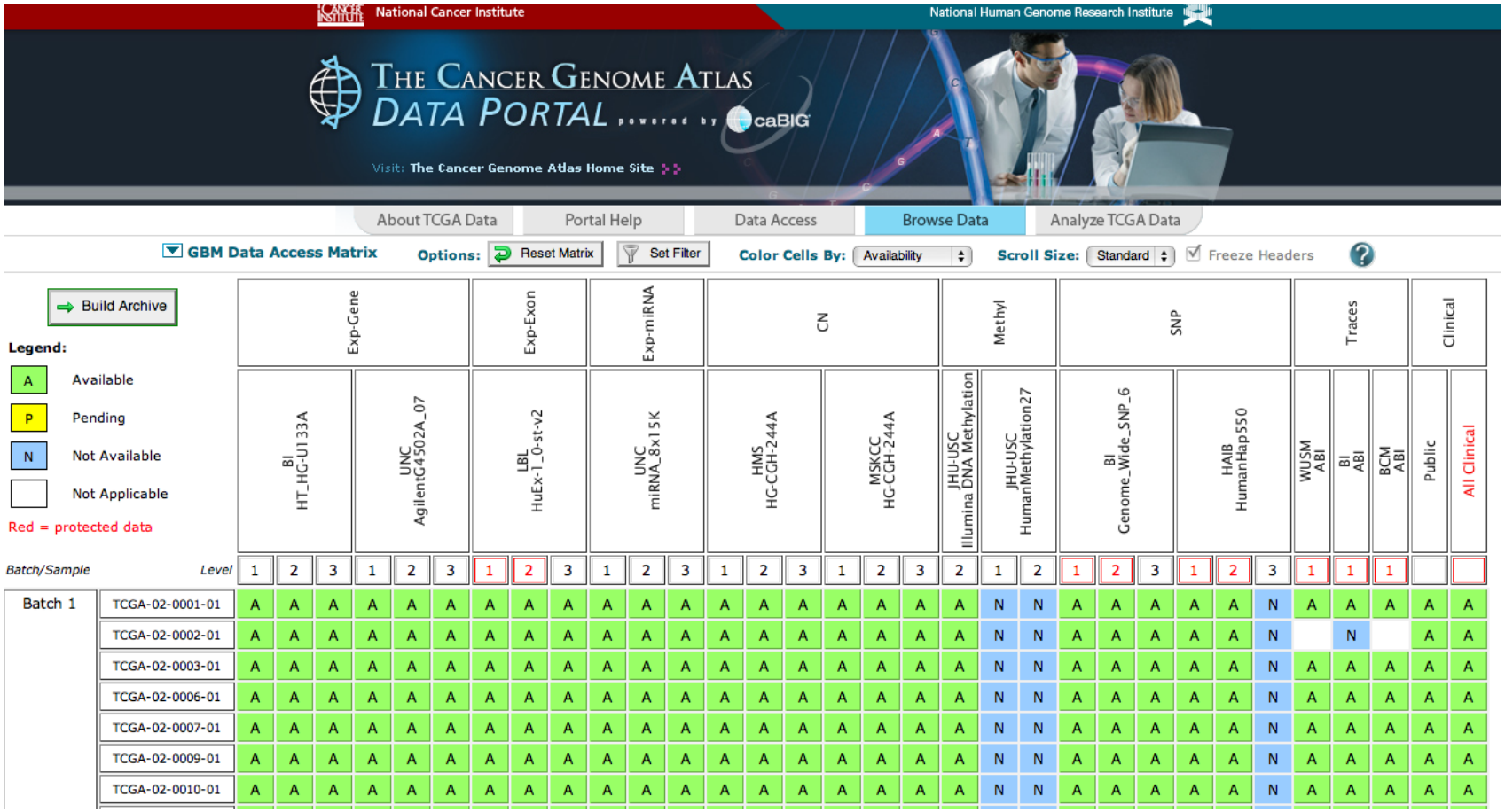
$$t_s(s, M_s) = \operatorname*{avg}_{d \in \{1,\ldots,D\}} t_s^d(s, M_s),$$

when evidence of significance from several data types is needed, or by taking the extremum score:

$$t_s(s, M_s) = \operatorname*{extremum}_{d \in \{1,\ldots,D\}} t_s^d(s, M_s),$$

when strong evidence from a single dimension seems to be sufficient.

Data: The Cancer Genome Atlas



TCGA GBM Data We Will Use

- Gene expression: affymetrix and agilent arrays from two centers
- Copy number variation: agilent (same platform, but two different centers)
- We have one observation from each type for each gene, for each sample (patient)
- Phenotype (response variable): dichotomized survival ($\leq 25\%$ – 190 days, $\geq 75\%$ – 594 days)
- Collections of gene sets from Broad MSigDb