Gene sets expression

Vincent Detours

IRIBHM, Université Libre de Bruxelles (U.L.B.)
vdetours@ulb.ac.be

Limits of the gene-wise view of gene expression data

- Pleiotropy is ubiquous: over-expression of a given gene provides ambiguous information about the underlying biological process
- Biological systems are highly redundant, non expression of a gene may be complemented by another

Limits of the gene-wise view of gene expression data

- The gene-wise view focus on large regulation amplitudes but will miss biological processes that involve groups of genes *coordinately* regulated at low levels
- Higher-level biological processes rely on the activity of many genes, not single genes
- Nearly all microarray-derived markers rely on multi-genes signatures

Limits of the gene-wise view of gene expression data

- The overlap of gene lists from replicated microarray studies has been notoriously low
- Interpreting lists of 100-1,000 of regulated genes is difficult in practice...
- ...or too easy: one can always make a biological argument by focusing on a subset of the list

Limits of the gene-wise view of gene expression data

It make sense to study the collective behavior of sets of functionally related genes, but

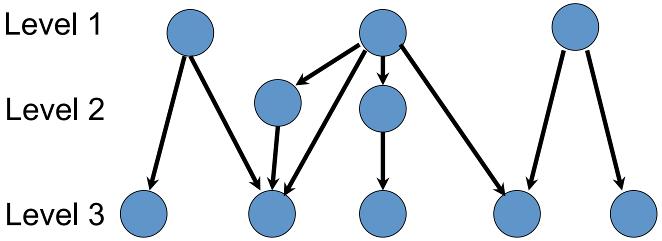
- What genes sets?
- What statistical procedures?

The Gene Ontology

- "The Gene Ontology project provides a *controlled vocabulary* to describe gene and gene product attributes in any organism"
- It is a gene-centered database of gene function curated by human beings
- See www.geneontology.org

The Gene Ontology: structure

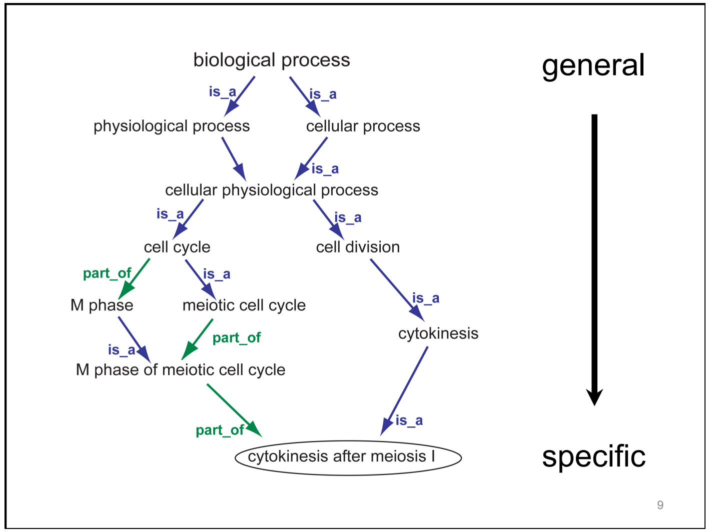
- GO is a *directed acyclic graph*, i.e. a hierarchical structure. It goes from general to specific concepts

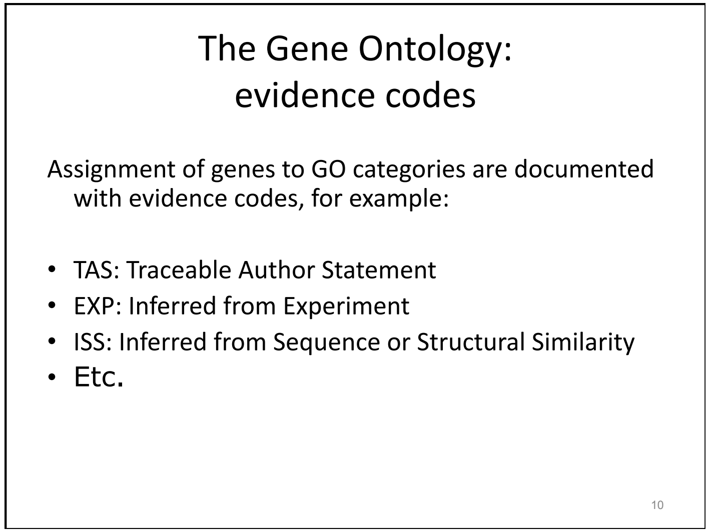


The Gene Ontology: main branches

There are three ontologies

- Biological process: cell cycle, apoptosis, oxydative phosphorylation,...
- Molecular function: catalytic activity, transporter,...
- Cellular compartment: nucleus, membrane,...

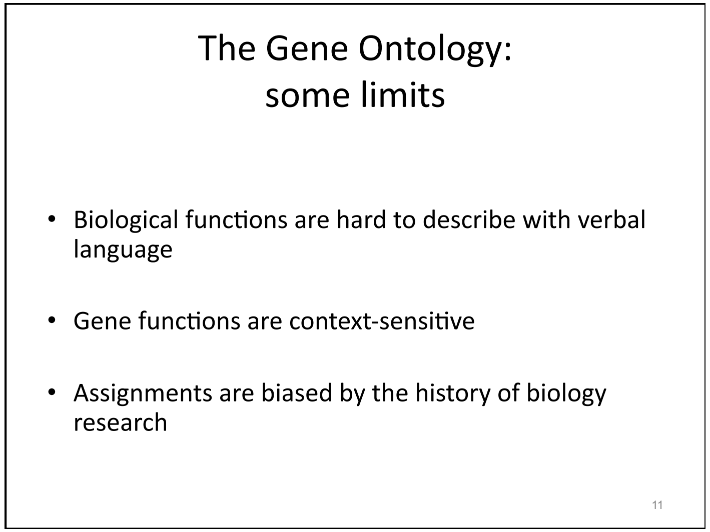




The Gene Ontology: evidence codes

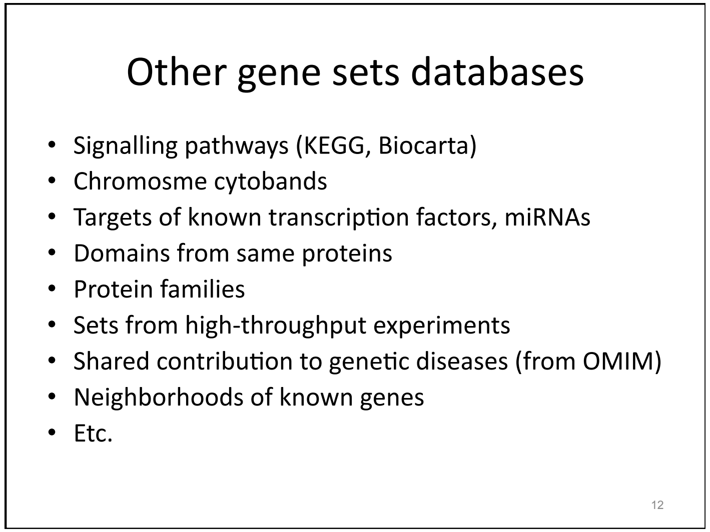
Assignment of genes to GO categories are documented with evidence codes, for example:

- TAS: Traceable Author Statement
- EXP: Inferred from Experiment
- ISS: Inferred from Sequence or Structural Similarity
- Etc.



The Gene Ontology: some limits

- Biological functions are hard to describe with verbal language
- Gene functions are context-sensitive
- Assignments are biased by the history of biology research



Other gene sets databases

- Signalling pathways (KEGG, Biocarta)
- Chromosme cytobands
- Targets of known transcription factors, miRNAs
- Domains from same proteins
- Protein families
- Sets from high-throughput experiments
- Shared contribution to genetic diseases (from OMIM)
- Neighborhoods of known genes
- Etc.

Most available tools operate downstream of gene selection

The user

1. provides a list of differentially expressed genes from his/her microarray experiments, e.g. the output of SAM
2. (provides a background gene list, e.g. genes present on the microarrays; it's the negative control)
3. selects a collection of functionally characterized gene sets (e.g. some level of the Gene Ontology)

The statistical procedure

1. measures how each functional gene set overlaps with the user-supplied list of step #1.
2. corrects for multiple testing (many gene sets tested)

The significance of overlap is estimated with a test of independence in a 2x2 contigency table

	Differentially expressed	Not differentially expressed
In gene set	18	31
Not in gene set	486	11,800

May use

- χ^2
- hypergeometric test
- etc.,...

This very popular approach is flawed and anticonservative

- Biological samples, not genes are replicated in actual microarrays experiments: the power of the test should increase with the number of samples, it is not
- Genes expression levels are not independent variables

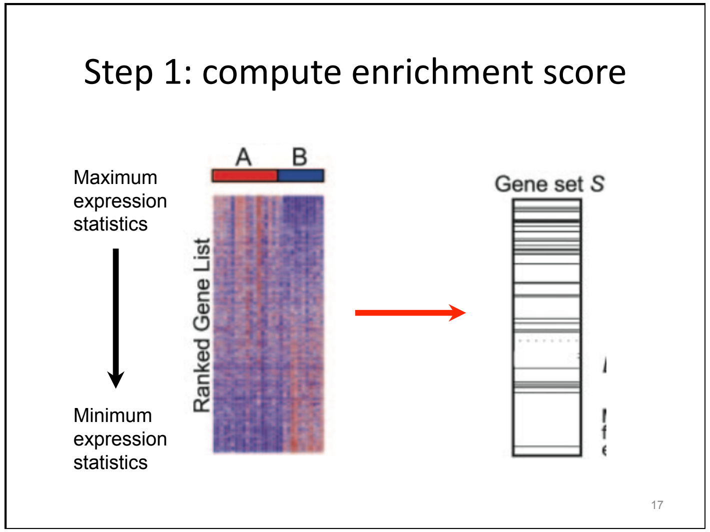
› Artificially high power due to (inapropriate) gene sampling---there are >20,000 genes

› Overstated p-values due to departure from independence

Gene Set Enrichment Analysis

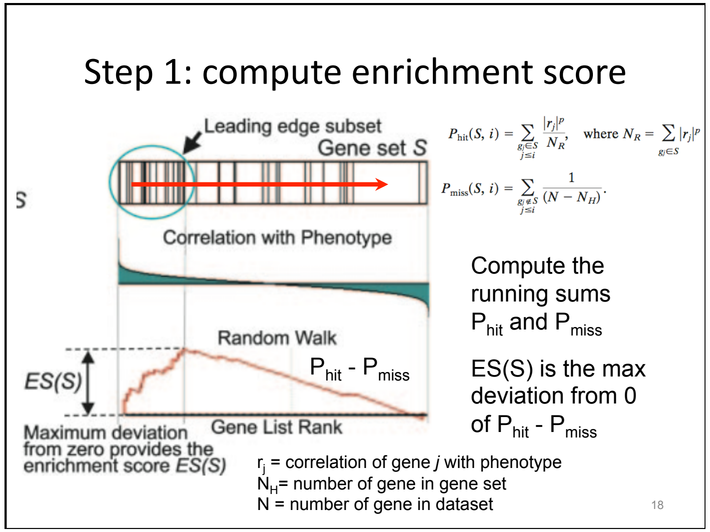
Subramanian et al. (PNAS, 2005) proposed a method, GSEA, that

- uses patient sampling
- does not require gene expression threshold
- comes with a large collection of experimentally derived gene sets



Step 1: compute enrichment score

17



Step 1: compute enrichment score

$$P_{\text{hit}}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

$$P_{\text{miss}}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

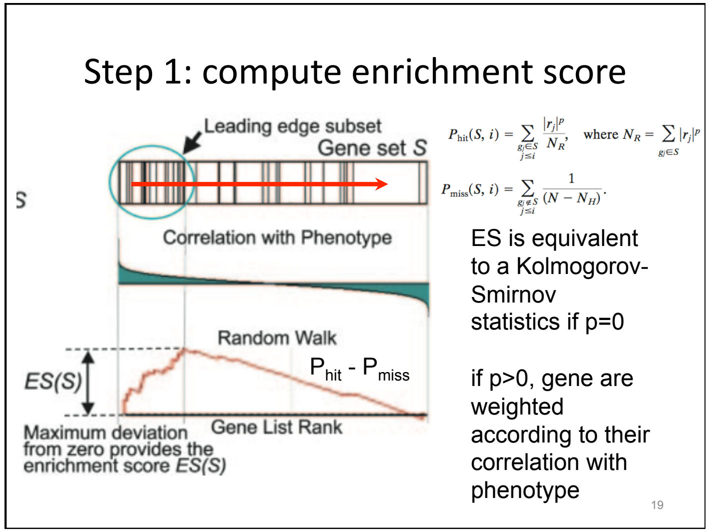
Compute the running sums P_{hit} and P_{miss}

ES(S) is the max deviation from 0 of P_{hit} - P_{miss}

r_j = correlation of gene *j* with phenotype
N_H= number of gene in gene set
N = number of gene in dataset

Leading edge subset
Gene set *S*
S
Correlation with Phenotype
Random Walk
P_{hit} - P_{miss}
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

18



Step 1: compute enrichment score

$$P_{\text{hit}}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p$$

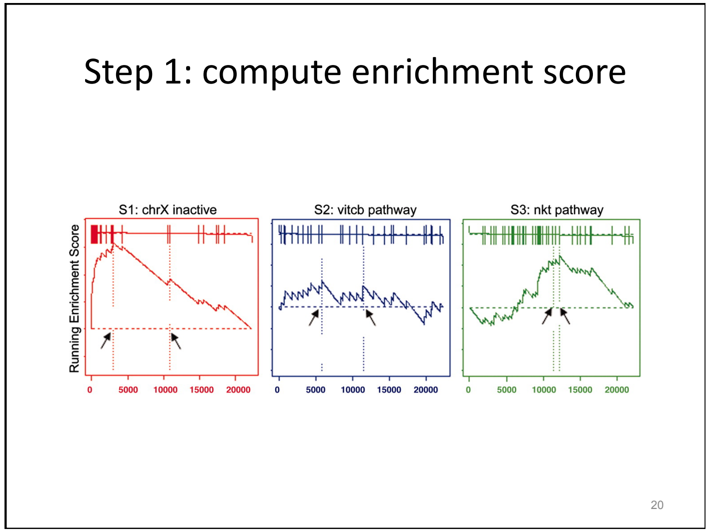
$$P_{\text{miss}}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}.$$

ES is equivalent to a Kolmogorov-Smirnov statistics if p=0

if p>0, gene are weighted according to their correlation with phenotype

Leading edge subset
Gene set *S*
S
Correlation with Phenotype
Random Walk
P_{hit} - P_{miss}
ES(S)
Gene List Rank
Maximum deviation from zero provides the enrichment score *ES(S)*

19



Step 1: compute enrichment score

20

Step 2: estimate the significance of ES

Do many times:

1. permute sample labels at random
2. recompute ES

Compute p-values from permutation ES distribution

› Power does depend on sample size
› Gene correlation structure is preserved

Step 3: Adjust for multiple testing

Once steps 1 and 2 have been run for all gene sets

1. For all gene set S, adjust its ES for gene set size, by dividing ES(S, original) and ES(S, perm) by the mean of ES(S, perm)

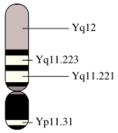
2. Compute FDRs from normalized original and permutation ES

MSigDB

Gene sets from the Broad Institute (check updates)

- C1: cytogenetics
- C2: Functional sets (from experiments)
- C3: Regulatory motifs
- C4: Neighborhood sets
- C5: Gene Ontology
- C6: Oncogenic signatures
- C7: Immunological signatures

Control: Male vs. Female Lymphoblastoid cells lines



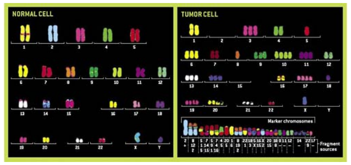
mRNA profiles from 15 male and 17 female lymphoblastoid cell lines

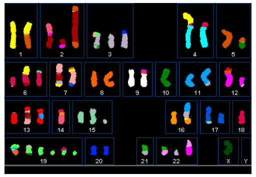
- testing C1 in the male>female comparizon detects chromosome Y, and cytobands Yp11 and Yq11 (the most gene-dense)

- surprisingly, testing C2 yields genes enriched in reproductive organs, testis and uterus, even when analysis is restricted to autosomal genes

Application: cytogenetic abnormalities in acute leukemias

- cancer cells often have an abnormal karyotype

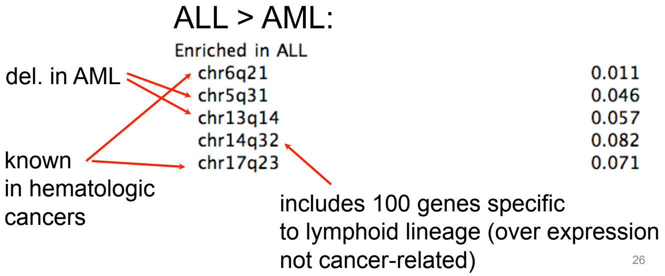




Human chromosomes,
metaphase state

Application: cytogenetic abnormalities in acute leukemias

- Run GSEA with C1 on 24 acute lymphoid leukemias vs. 24 acute myeloid leukemias



ALL > AML:

Enriched in ALL	
chr6q21	0.011
chr5q31	0.046
chr13q14	0.057
chr14q32	0.082
chr17q23	0.071

del. in AML

known
in hematologic
cancers

includes 100 genes specific
to lymphoid lineage (over expression
not cancer-related)

Application: p53 status in cell lines

- p53 is mutated in a much as 50% of all human cancers
- it is a transcription factor that regulates genes in response to various cellular stresses, including DNA damage

Idea

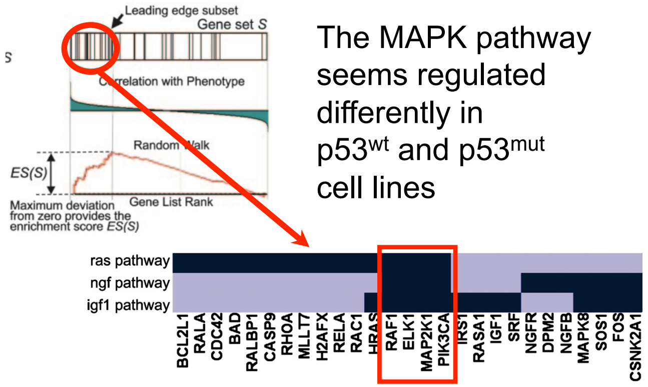
use published NCI-60 cell lines expression profiles (the p53 mutational status of 50 of them has been determined, 17 p53^{wt}, 33 p53^{mut})

Application: p53 status in cell lines

Run C2 gene sets

Enriched in p53 mutant	
Ras signaling pathway	0.171
Enriched in p53 wild type	
Hypoxia and p53 in the cardiovascular system	0.001
Stress induction of HSP regulation	0.001
p53 signaling pathway	0.001
p53 up-regulated genes	0.013
Radiation sensitivity genes	0.078

Application: p53 status in cells lines



The MAPK pathway seems regulated differently in $p53^{wt}$ and $p53^{mut}$ cell lines

Application: outcome in two lung cancer studies

- Two lung cancer data sets of 62 and 86 primary tumors mRNA profiles
- Patients stratified in 'good' and 'bad' outcome

› Gene-wise analysis detects no outcome-related genes at q<0.05, in either study

› The two top-100 genes lists seems very discordant (intersect of 12 genes)

Application: outcome in two lung cancer studies

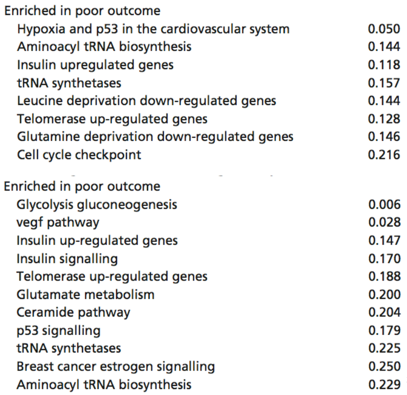
- S1 = top 100 outcome-related genes in study 1
- S2 = top 100 outcome-related genes in study 2

GSEA finds that

- NES for S1 in study 2 is 1.9, p<0.001
- NES for S2 in study 1 is 2.1, p<0.001

Gene sets are more robust than individual genes

Application: outcome in two lung cancer studies

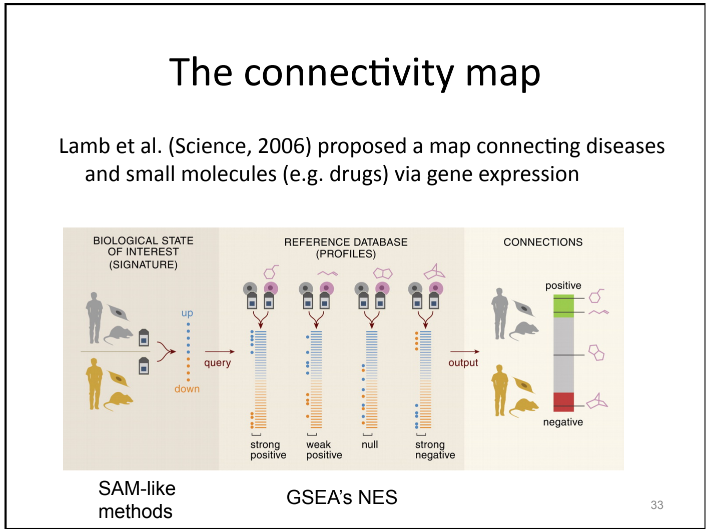


study 1

Enriched in poor outcome	
Hypoxia and p53 in the cardiovascular system	0.050
Aminoacyl tRNA biosynthesis	0.144
Insulin upregulated genes	0.118
tRNA synthetases	0.157
Leucine deprivation down-regulated genes	0.144
Telomerase up-regulated genes	0.128
Glutamine deprivation down-regulated genes	0.146
Cell cycle checkpoint	0.216

study 2

Enriched in poor outcome	
Glycolysis gluconeogenesis	0.006
vegf pathway	0.028
Insulin up-regulated genes	0.147
Insulin signalling	0.170
Telomerase up-regulated genes	0.188
Glutamate metabolism	0.200
Ceramide pathway	0.204
p53 signalling	0.179
tRNA synthetases	0.225
Breast cancer estrogen signalling	0.250
Aminoacyl tRNA biosynthesis	0.229

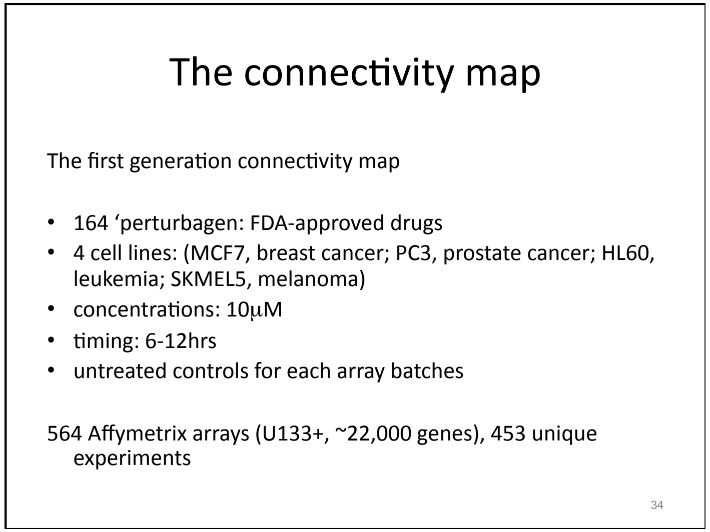


The connectivity map

Lamb et al. (Science, 2006) proposed a map connecting diseases and small molecules (e.g. drugs) via gene expression

SAM-like methods

GSEA's NES

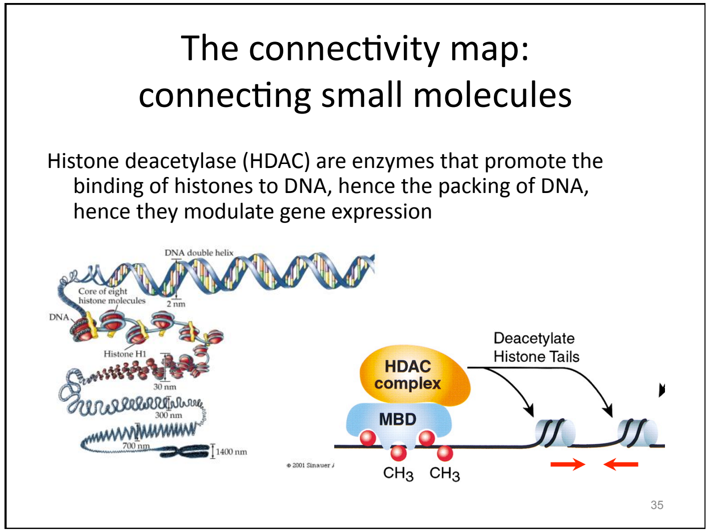


The connectivity map

The first generation connectivity map

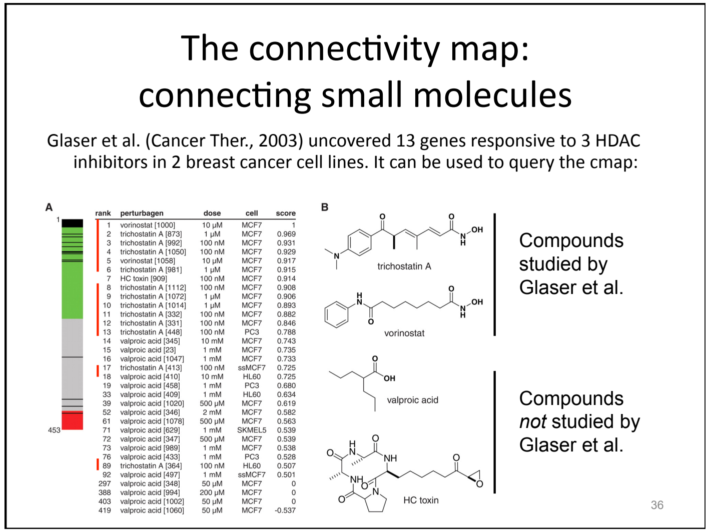
- 164 ‘perturbagen: FDA-approved drugs
- 4 cell lines: (MCF7, breast cancer; PC3, prostate cancer; HL60, leukemia; SKMEL5, melanoma)
- concentrations: 10μM
- timing: 6-12hrs
- untreated controls for each array batches

564 Affymetrix arrays (U133+, ~22,000 genes), 453 unique experiments



The connectivity map: connecting small molecules

Histone deacetylase (HDAC) are enzymes that promote the binding of histones to DNA, hence the packing of DNA, hence they modulate gene expression



The connectivity map: connecting small molecules

Glaser et al. (Cancer Ther., 2003) uncovered 13 genes responsive to 3 HDAC inhibitors in 2 breast cancer cell lines. It can be used to query the cmap:

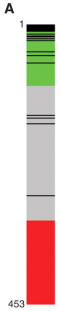
rank	perturbagen	dose	cell	score
1	vorinostat [1000]	10 μM	MCF7	1
2	trichostatin A [873]	1 μM	MCF7	0.969
3	trichostatin A [992]	100 nM	MCF7	0.931
4	trichostatin A [1050]	100 nM	MCF7	0.929
5	vorinostat [1058]	10 μM	MCF7	0.917
6	trichostatin A [981]	1 μM	MCF7	0.915
7	HC toxin [909]	100 nM	MCF7	0.914
8	trichostatin A [1112]	100 nM	MCF7	0.908
9	trichostatin A [1072]	1 μM	MCF7	0.906
10	trichostatin A [1014]	1 μM	MCF7	0.893
11	trichostatin A [332]	100 nM	MCF7	0.882
12	trichostatin A [331]	100 nM	MCF7	0.846
13	trichostatin A [448]	100 nM	PC3	0.788
14	valproic acid [345]	10 mM	MCF7	0.743
15	valproic acid [23]	1 mM	MCF7	0.735
16	valproic acid [1047]	1 mM	MCF7	0.733
17	trichostatin A [413]	100 nM	ssMCF7	0.725
18	valproic acid [410]	10 mM	HL60	0.725
19	valproic acid [458]	1 mM	PC3	0.680
33	valproic acid [409]	1 mM	HL60	0.634
39	valproic acid [1020]	500 μM	MCF7	0.619
52	valproic acid [346]	2 mM	MCF7	0.582
61	valproic acid [1078]	500 μM	MCF7	0.563
71	valproic acid [629]	1 mM	SKMEL5	0.539
72	valproic acid [347]	500 μM	MCF7	0.539
73	valproic acid [989]	1 mM	MCF7	0.538
76	valproic acid [433]	1 mM	PC3	0.528
89	trichostatin A [364]	100 nM	HL60	0.507
92	valproic acid [497]	1 mM	ssMCF7	0.501
297	valproic acid [348]	50 μM	MCF7	0
388	valproic acid [994]	200 μM	MCF7	0
403	valproic acid [1002]	50 μM	MCF7	0
419	valproic acid [1060]	50 μM	MCF7	-0.537

Compounds studied by Glaser et al.

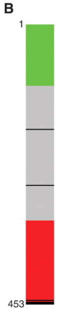
Compounds *not* studied by Glaser et al.

The connectivity map: connecting small molecules

Frasor et al. (Cancer Res., 2004) treated MCF7 breast cancer cell lines with 17β-estradiol, the natural eostrogen receptor ligand. Their 129 genes signature can be used to query the cmap:



rank	perturbagen	dose	cell	score
2	estradiol [988]	100 nM	MCF7	0.936
3	estradiol [373]	10 nM	ssMCF7	0.918
4	genistein [1015]	10 µM	MCF7	0.913
5	estradiol [1079]	10 nM	MCF7	0.899
6	estradiol [1021]	10 nM	MCF7	0.813
8	alpha-estradiol [990]	10 nM	MCF7	0.809
9	alpha-estradiol [403]	10 nM	ssMCF7	0.807
10	estradiol [414]	10 nM	ssMCF7	0.794
11	estradiol [121]	10 nM	MCF7	0.758
12	genistein [1073]	10 µM	MCF7	0.753
13	genistein [638]	10 µM	MCF7	0.730
17	alpha-estradiol [1048]	10 nM	MCF7	0.646
20	genistein [268]	1 µM	MCF7	0.619
21	estradiol [365]	100 nM	MCF7	0.610
25	genistein [382]	10 µM	MCF7	0.561
27	genistein [267]	1 µM	MCF7	0.552
46	alpha-estradiol [122]	10 nM	MCF7	0.435
51	estradiol [387]	10 nM	HL60	0.421
64	estradiol [782]	10 nM	HL60	0.376
148	alpha-estradiol [702]	10 nM	PC3	0
152	genistein [703]	10 µM	PC3	0
162	alpha-estradiol [762]	10 nM	MCF7	0
278	estradiol [665]	10 nM	PC3	0



rank	perturbagen	dose	cell	score
171	fulvestrant [704]	1 µM	PC3	0
261	fulvestrant [523]	1 µM	ssMCF7	0
447	fulvestrant [367]	1 µM	MCF7	-0.749
450	fulvestrant [310]	10 nM	MCF7	-0.843
451	fulvestrant [985]	1 µM	MCF7	-0.961
452	fulvestrant [1076]	10 nM	MCF7	-0.989
453	fulvestrant [1043]	1 µM	MCF7	-1

The connectivity map: connecting small molecules

Connecting molecules is of upmost relevance to toxicology

Major drugs have been recently withdrawn frommarket because of unexpected toxicities (e.g. Merk's painkiller, COX-2 inhibitor, Vioxx)

The European Union REACH regulations require

- proving the innocuity of tens thousands chemicals
- minimize animal testing

The connectivity map: connections with diseases

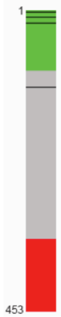
Lopez et al. (Obes. Res., 2003) derived a signature from a rat model of diet-induced obesity (expresson in adipose tissues)

Differences between the two systems:

•tissues vs. cell lines

•65 days vs. 6hrs

•rat vs. human



rank	perturbagen	dose	cell	score
3	indometacin [452]	100 µM	PC3	0.874
4	rosiglitazone [430]	10 µM	PC3	0.838
11	troglitazone [462]	10 µM	PC3	0.737
20	troglitazone [431]	10 µM	PC3	0.696
116	15-delta prostaglandin J2 [446]	10 µM	PC3	0

peroxisome proliferation-activated receptor γ agonists

only PC3 expresses PPARγ

Other PPARγ agonists were not detected

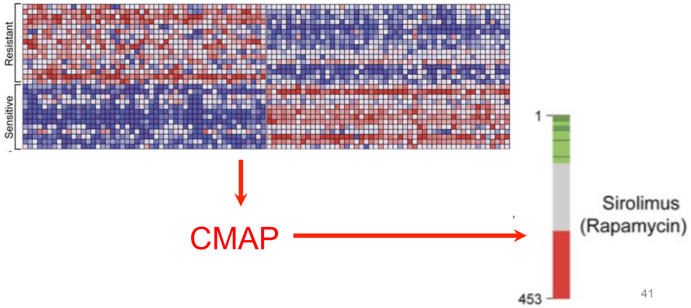
The connectivity map: connections with diseases

In another publication (Wei et al., Cancer Cell 2006), the cmap was used to investigate the resistance of acute lymphoblastic leukemia to glucocorticoids

- GC is a class of steroid hormones that suppress the immune system
- GC resistance is a marker of bad prognosis in ALL
- Resistance mechanisms are unknown
- Wei et al. investigate apoptosis as a possible mechanism

The connectivity map: connections with diseases

- Expression analysis of pre-treatment ALL samples
- GC resistance determined according to apoptosis (sensitive IC50 < 150mg/ml prednisolone, resistant otherwise)



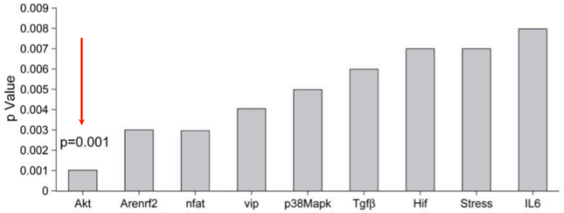
CMAP

Sirolimus (Rapamycin)

The connectivity map: connections with diseases

- Rapamicyn is an immunosuppressant
- It inhibits mTOR, which is activated by the PI3K/Akt pathway

GSEA on the sensitive/resistant ALL samples detect the following Biocarta pathways



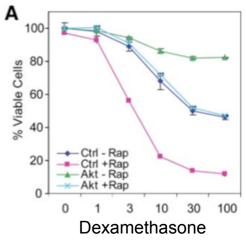
The connectivity map: connections with diseases

- mTOR is downstream of Akt in some cell types
- Does activation of Akt induce GC resistance?
- If yes, can rapamycin reverse resistance?

T cells were infected with a virus containing a constitutively active form of Akt

Apoptosis was assayed with and without rapamycn

(further molecular biology dissected the details GC and rapamycin action)



Strenghts of GSEA

Flexible set up

- Choice of expression metrics
- Choice of gene sets with a virtually infinite range of biological significations

> Really takes advantage of the universality microarray-derived mRNA phenotypes

Statistically sound

- sample permutation
- preserve genes correlation sctructure
- robust investigation of data at metagenes level

Criticisms of GSEA

- Kolmogorov-Smirnov-like statistics lacks statistical power, Efron & Tibshirani proposed an alternative (Ann. Appl. Stat., 2007)
- Not clear how it bahaves with respect to gene set size
- Crude gene set definition, no sophisticated gene expression directionality, for example

CMAP development

- CMAP 2.0 contains profiles of 1,309 FDA-approved compounds
- Library of Integrated Network-based Cellular Signatures (LINCS):
  - 1 million profiles!
  - 1000 genes capturing 80% expression variance
  - 24,413 compounds, 22,119 gene KO and over-expression assays
  - 18 cell types
  - https://clue.io

LINCS

- L1000: 1000 genes that capture 80% of global gene expression variance across
- So far, L1000 profiles available for
  - ~20,500 compounds
  - ~18,500 gene shRNA inhibitions assays
  - ~3,500 gene overexpression assays
  - In 59 cell lines, 10 primary cultures types,…
  - About 10^6 LINCS profiles, overall