Liverome: a curated database of liver cancer-related gene signatures with self-contained context information

➢ Motivation of the study

- Our group has been performing microarray experiments on a large cohort of liver cancer patients (~300 patients).

- Needed to compare our own data with publicly available liver cancer data for prioritization of genes for further follow-up studies.

- The result from the public molecular profiling data most often comes in the form of a list of genes, also called a gene signature, reported in articles as a table

- These signatures are scattered in individual articles, buried in main or supplementary tables, thus all the valuable information is underused.

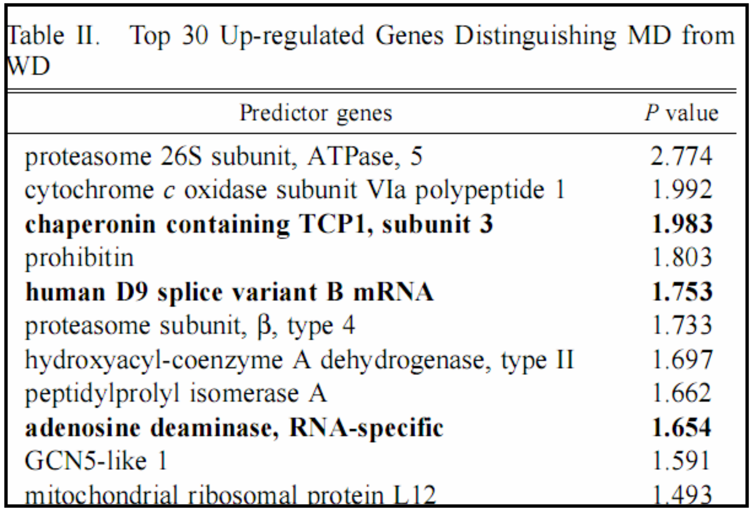


Table II. Top 30 Up-regulated Genes Distinguishing MD from WD

Predictor genes	*P* value
proteasome 26S subunit, ATPase, 5	2.774
cytochrome *c* oxidase subunit VIa polypeptide 1	1.992
chaperonin containing TCP1, subunit 3	**1.983**
prohibitin	1.803
human D9 splice variant B mRNA	**1.753**
proteasome subunit, β, type 4	1.733
hydroxyacyl-coenzyme A dehydrogenase, type II	1.697
peptidylprolyl isomerase A	1.662
adenosine deaminase, RNA-specific	**1.654**
GCN5-like 1	1.591
mitochondrial ribosomal protein L12	1.493

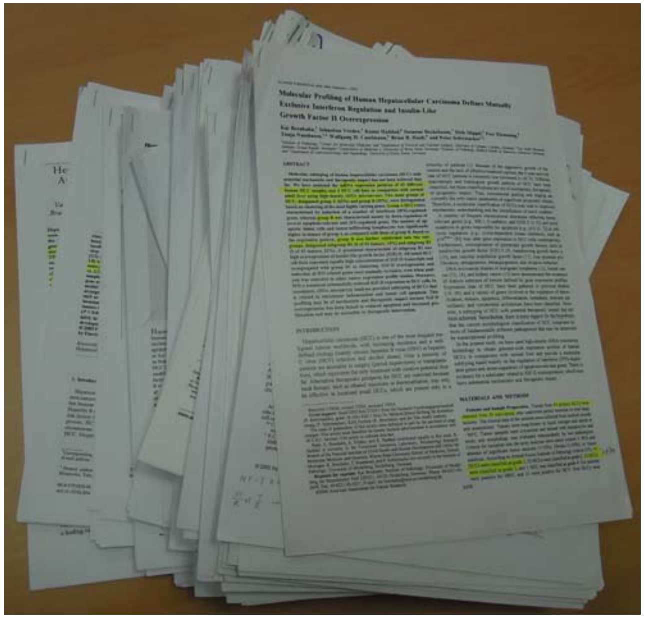
- To address this need, several signature databases have been constructed to serve as a repository of signatures. But several limitations were observed.

Seungwoo Hwang
Korean Bioinformation Center (KOBIC)
Korea Research Institute of Bioscience and Biotechnology (KRIBB)

2011/12/01

Database construction in a nutshell

~100 articles on liver cancer microarray and proteome studies



~150 gene signatures that appeared as tables

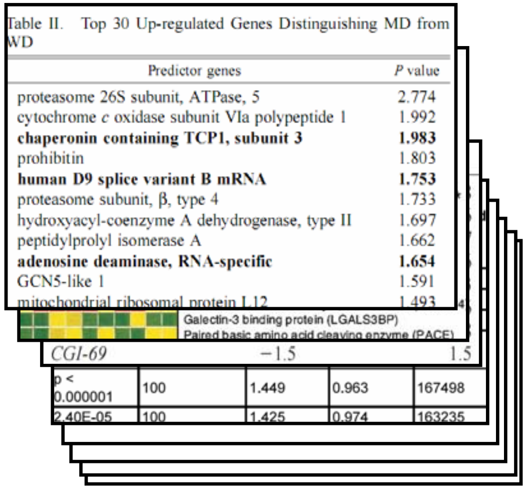


Table II. Top 30 Up-regulated Genes Distinguishing MD from WD

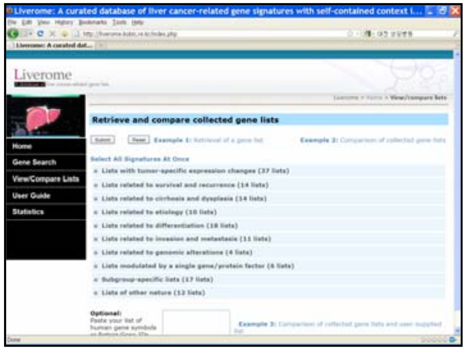
Predictor genes	*P* value
proteasome 26S subunit, ATPase, 5	2.774
cytochrome *c* oxidase subunit VIa polypeptide 1	1.992
chaperonin containing TCP1, subunit 3	**1.983**
prohibitin	1.803
human D9 splice variant B mRNA	**1.753**
proteasome subunit, β, type 4	1.733
hydroxyacyl-coenzyme A dehydrogenase, type II	1.697
peptidylprolyl isomerase A	1.662
adenosine deaminase, RNA-specific	**1.654**
GCN5-like 1	1.591
mitochondrial ribosomal protein L12	1.493

Galectin-3 binding protein (LGALS3BP)
Paired basic amino acid cleaving enzyme (PACE)

CGI-69	−1.5	1.5

p < 0.000001	100	1.449	0.963	167498
2.40E-05	100	1.425	0.974	163235

Liverome DB



Liverome: A curated database of liver cancer-related gene signatures with self-contained context i...

File Edit View History Bookmarks Tools Help
http://liverome.kobic.re.kr/index.php
Liverome: A curated dat...
Liverome
Liverome > Home > View/compare lists
Home
Gene Search
View/Compare Lists
User Guide
Statistics
Submit
Reset
Done

Retrieve and compare collected gene lists

Example 1: Retrieval of a gene list
Example 2: Comparison of collected gene lists

Select All Signatures At Once

- Lists with tumor-specific expression changes (37 lists)
- Lists related to survival and recurrence (14 lists)
- Lists related to cirrhosis and dysplasia (14 lists)
- Lists related to etiology (10 lists)
- Lists related to differentiation (18 lists)
- Lists related to invasion and metastasis (11 lists)
- Lists related to genomic alterations (4 lists)
- Lists modulated by a single gene/protein factor (6 lists)
- Subgroup-specific lists (17 lists)
- Lists of other nature (12 lists)

Optional:
Paste your list of human gene symbols

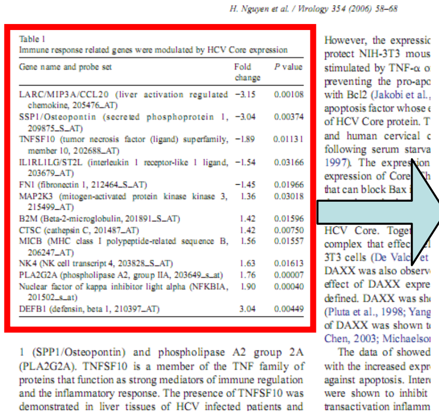
Example 3: Comparison of collected gene lists and user-supplied list

- ➢ Thorough manual annotation of database content
- ➢ Comprehensive coverage
- ➢ Straightforward web interface designed for liver cancer biologists

What is a signature database

➢ With respect to **data source**

Publication-derived signatures



H. Nguyen et al. / Virology 354 (2006) 58–68

Table 1
Immune response related genes were modulated by HCV Core expression

Gene name and probe set	Fold change	P value
LARC/MIP3A/CCL20 (liver activation regulated chemokine, 205476_AT)	−3.15	0.00108
SSP1/Osteopontin (secreted phosphoprotein 1, 209875_S_AT)	−3.04	0.00374
TNFSF10 (tumor necrosis factor (ligand) superfamily, member 10, 202688_AT)	−1.89	0.01131
IL1RL1LG/ST2L (interleukin 1 receptor-like 1 ligand, 203679_AT)	−1.54	0.03166
FN1 (fibronectin 1, 212464_S_AT)	−1.45	0.01966
MAP2K3 (mitogen-activated protein kinase kinase 3, 215499_AT)	1.36	0.03018
B2M (Beta-2-microglobulin, 201891_S_AT)	1.42	0.01596
CTSC (cathepsin C, 201487_AT)	1.42	0.00750
MICB (MHC class I polypeptide-related sequence B, 206247_AT)	1.56	0.01557
NK4 (NK cell transcript 4, 203828_S_AT)	1.63	0.01613
PLA2G2A (phospholipase A2, group IIA, 203649_s_at)	1.76	0.00007
Nuclear factor of kappa inhibitor light alpha (NFKBIA, 201502_s_at)	1.90	0.00040
DEFB1 (defensin, beta 1, 210397_AT)	3.04	0.00449

1 (SPP1/Osteopontin) and phospholipase A2 group 2A (PLA2G2A). TNFSF10 is a member of the TNF family of proteins that function as strong mediators of immune regulation and the inflammatory response. The presence of TNFSF10 was demonstrated in liver tissues of HCV infected patients and

However, the expressic
protect NIH-3T3 mous
stimulated by TNF-α o
preventing the pro-apo
with Bcl2 (Jakobi et al.,
apoptosis factor whose e
of HCV Core protein. T
and human cervical c
following serum starva
1997). The expression
expression of Core. Th
that can block Bax i

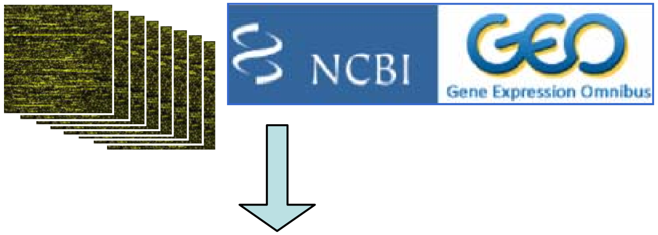
HCV Core. Toget
complex that effec
3T3 cells (De Vald et
DAXX was also observ
effect of DAXX expre
defined. DAXX was sh
(Pluta et al., 1998; Yang
of DAXX was shown t
Chen, 2003; Michaelsor
The data of showed
with the increased expr
against apoptosis. Inter
were shown to inhibit
transactivation inflamm

Signature tables in articles

Strength

- Utilize the end results from expert analysis of individual studies
- Can always obtain signatures from articles

Raw data-derived signatures



Signatures are generated by a re-analysis of expression profile data from public repositories

Strength

- Consistent data processing scheme may generate signatures that are more reproducible across datasets
- Full list of genes are generated

➢ With respect to **phenotype coverage**

	Specialized	All-inclusive
	for example, liver cancer	for example, all types of cancer all phenotypes
Strength	▪ Phenotype-specific coverage is generally high	▪ Enables inter-phenotype comparison

Signature databases (a partial list)

	Publication-derived	Raw data-derived
Specialized	❑ **Liverome** ❑ EHCO (*BMC Bioinfo* 2007)	❑ Pancreatic Expression DB (*BMC Genomics* 2007; *NAR* 2011)
All-inclusive	❑ CCancer (*NAR* 2010) ❑ dbDEPC (*NAR* 2010) ❑ GeneSigDB (*NAR* 2010) ❑ MSigDB (*Bioinformatics* 2011) ▪ Employed in GSEA ▪ A subset “cgp” contains publication-derived signatures	❑ Oncomine (*Neoplasia* 2004, 2007) ❑ GeneChaser (*BMC Bioinfo* 2008)

Liver cancer-specific data coverage

	Publication-derived
Specialized	□ **Liverome** □ EHCO (*BMC Bioinfo* 2007)
All-inclusive	□ CCancer (*NAR* 2010) □ dbDEPC (*NAR* 2010) □ GeneSigDB (*NAR* 2010)

- ➢ Our data coverage is >3 times larger

- ➢ Even in large databases whose overall coverage is much higher than ours, their liver cancer-specific coverage was much lower than ours

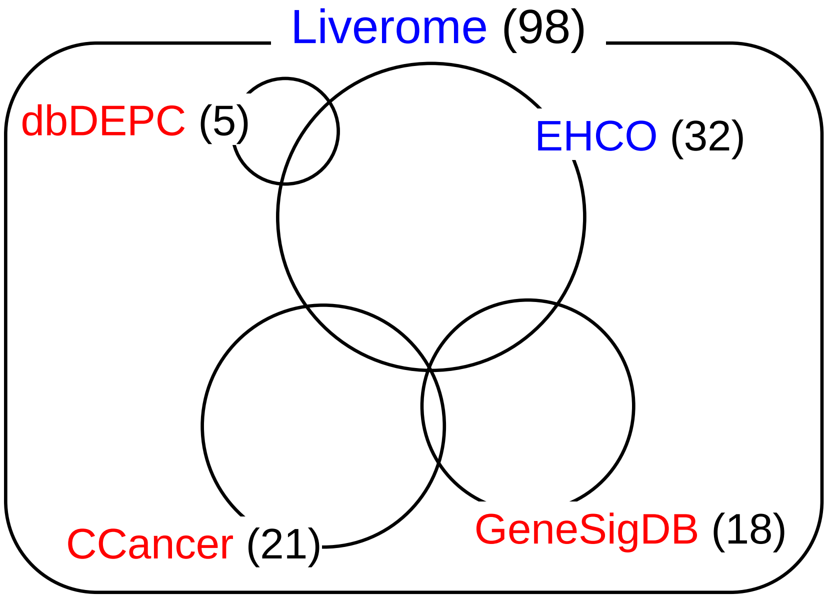
	GeneSigDB	CCancer
Overall coverage	~10 times	~26 times
Liver cancer-specific coverage	1/5	1/5

⇨ Points to the need for specialized database

Red: All-inclusive DB

Blue: Liver cancer DB

(Number): Liver cancer-related articles



Database statistics of Liverome

- With respect to **clinicopathological category**

Category	# Signatures	# Articles
1) Tumor vs Normal comparison	37	32
2) Survival & Recurrence	14	14
3) Cirrhosis & Dysplasia	14	9
4) Etiology	10	9
5) Differentiation	18	13
6) Invasion & Metastasis	11	10
7) Genomic alterations	4	4
8) Modulated by a single gene/protein factor	6	6
9) Subgroup-specific	17	3
10) Other	12	11
Total	**143 signatures**	**98 articles**

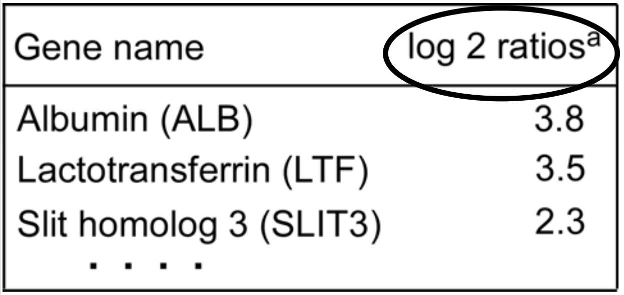
- With respect to **type of experiment**

Type of experiment	# Signatures	# Articles
Transcriptomics	124	83
Proteomics	16	12
Others	3	3
Total	**143 signatures**	**98 articles**

Well-annotated version of signatures: Main strength of Liverome

Signature table **as appeared in an article**

Table 2 of Lau *et al* (2006) *Oncogene*



Gene name	log 2 ratios[a]
Albumin (ALB)	3.8
Lactotransferrin (LTF)	3.5
Slit homolog 3 (SLIT3)	2.3
. . . .	

Compared sample groups?

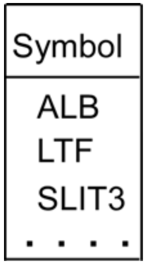
- Tumor vs Normal?
- Between subtypes?
- Something else?

Info is missing from table

Merely extracted gene IDs

Uninformative form of signature **from other DB** (CCancer)

Table 2

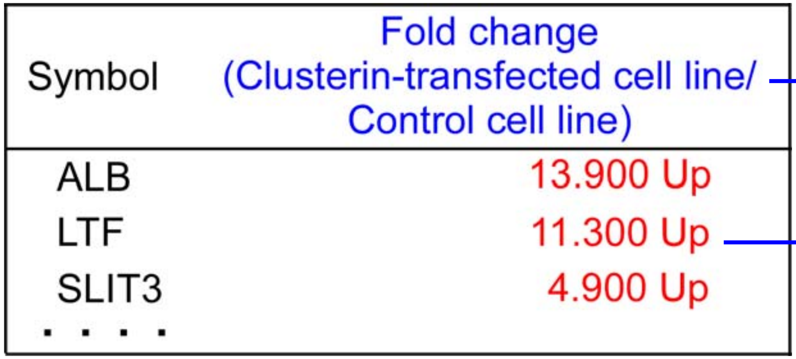


Symbol
ALB
LTF
SLIT3
. . . .

Read each article and derived an informative form

Self-contained form of signature **from Liverome**

Lau (2006) Oncogene [Genes regulated by clusterin]



Symbol	Fold change (Clusterin-transfected cell line/ Control cell line)
ALB	13.900 Up
LTF	11.300 Up
SLIT3	4.900 Up
. . . .	

Informatively named the signature

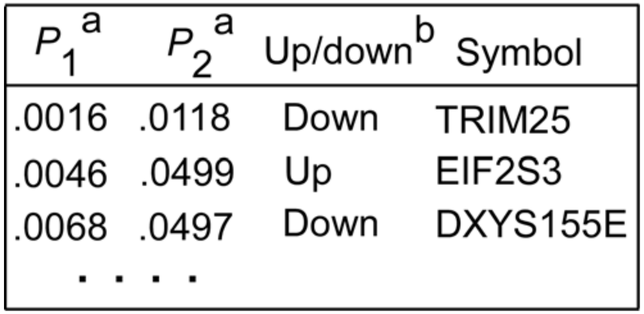
Specifies the compared groups

Represents fold change values in a scale that is more recognizable

Well-annotated version of signatures: Main strength of Liverome (cont'd)

Signature table **as appeared in an article**

Table 2 of Okamoto *et al* (2006)

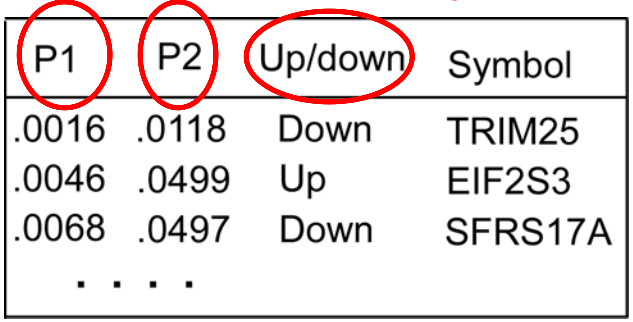


P_1[a]	P_2[a]	Up/down[b]	Symbol
.0016	.0118	Down	TRIM25
.0046	.0499	Up	EIF2S3
.0068	.0497	Down	DXYS155E
. . . .			

Extracted the signature table as-is

Uninformative form of signature **from other DB** (GeneSigDB)

Viral_Okamoto06_36genes



P1	P2	Up/down	Symbol
.0016	.0118	Down	TRIM25
.0046	.0499	Up	EIF2S3
.0068	.0497	Down	SFRS17A
. . . .			

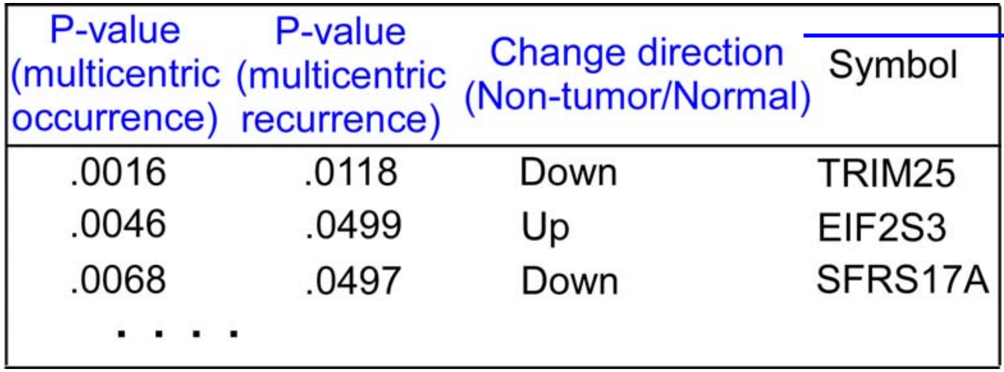
Manual annotated all the information

Self-contained form of signature **from Liverome**

Okamoto (2006) Ann Sur Oncol

[Markers for multicentric hepatocarcinogenesis]

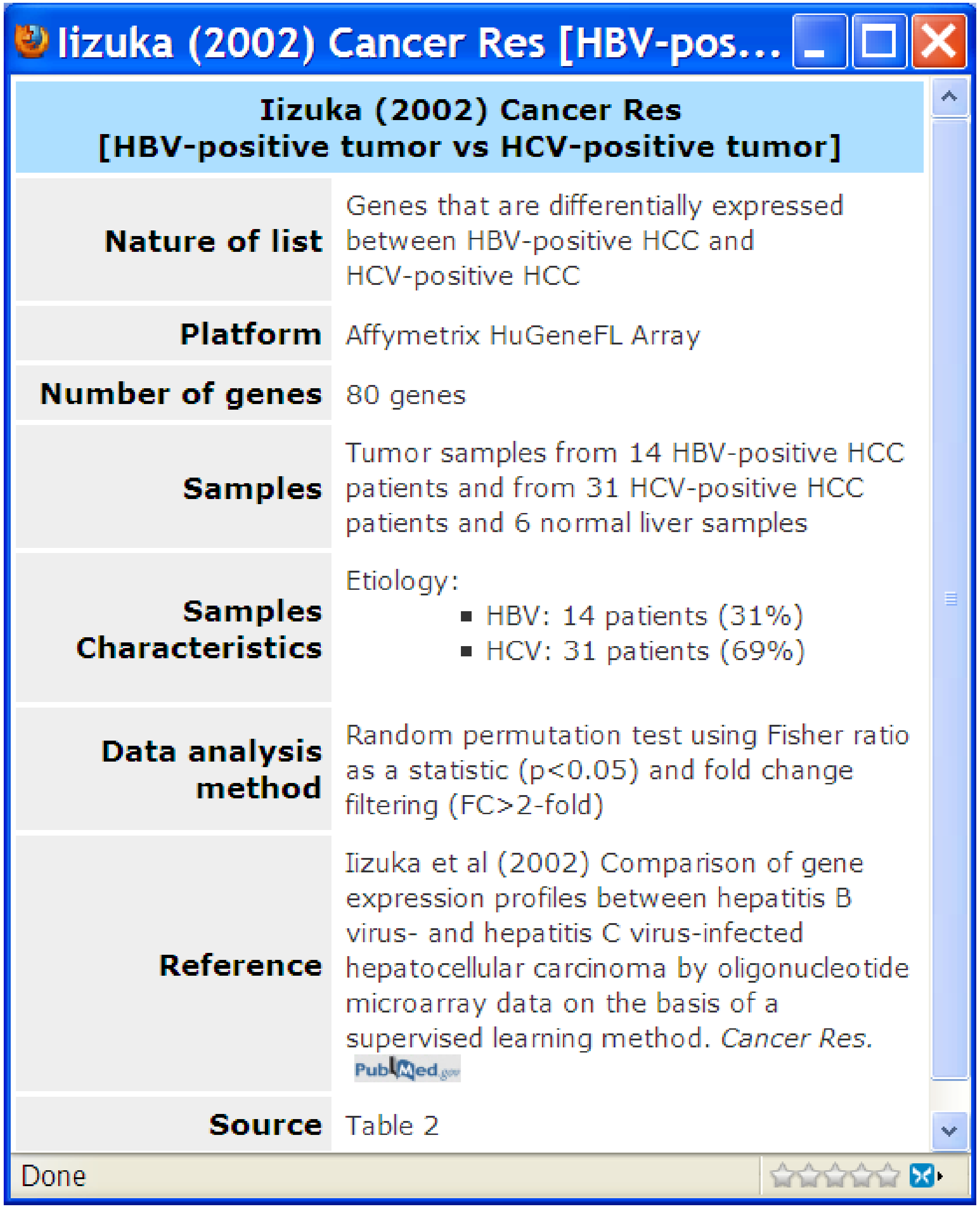
Informatively named the signature



P-value (multicentric occurrence)	P-value (multicentric recurrence)	Change direction (Non-tumor/Normal)	Symbol
.0016	.0118	Down	TRIM25
.0046	.0499	Up	EIF2S3
.0068	.0497	Down	SFRS17A
. . . .			

Specified the compared groups

Summarized all the essential information underlying the signature



Iizuka (2002) Cancer Res
[HBV-positive tumor vs HCV-positive tumor]

Nature of list	Genes that are differentially expressed between HBV-positive HCC and HCV-positive HCC
Platform	Affymetrix HuGeneFL Array
Number of genes	80 genes
Samples	Tumor samples from 14 HBV-positive HCC patients and from 31 HCV-positive HCC patients and 6 normal liver samples
Samples Characteristics	Etiology: ▪ HBV: 14 patients (31%) ▪ HCV: 31 patients (69%)
Data analysis method	Random permutation test using Fisher ratio as a statistic ($p<0.05$) and fold change filtering (FC>2-fold)
Reference	Iizuka et al (2002) Comparison of gene expression profiles between hepatitis B virus- and hepatitis C virus-infected hepatocellular carcinoma by oligonucleotide microarray data on the basis of a supervised learning method. *Cancer Res.* PubMed.gov
Source	Table 2

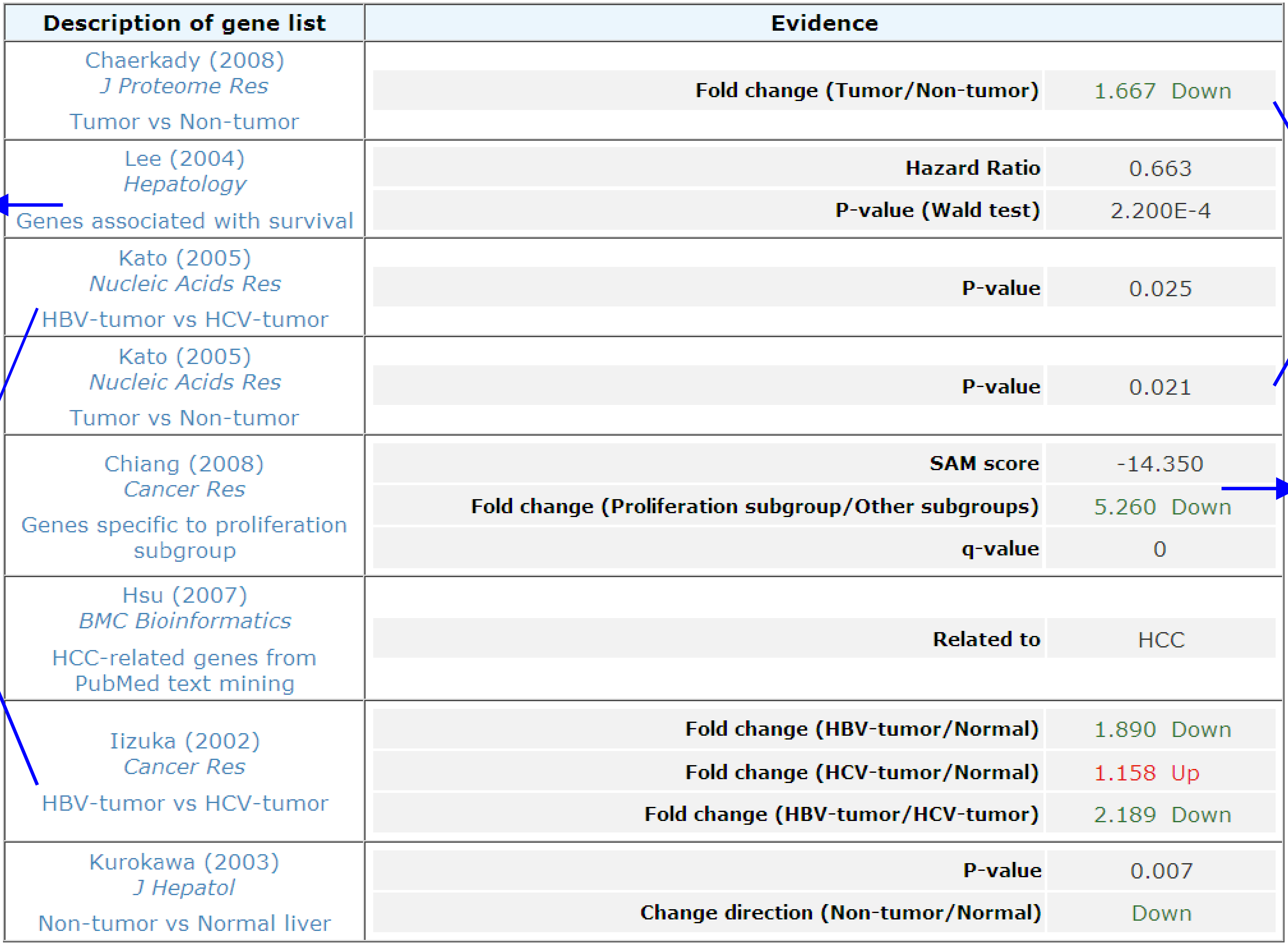
Main point:

➢ Made extensive manual annotation efforts to contain all context information within the database

➢ Should enable easier database browsing without a need to refer to the original publication

Straightforward web interface: Gene search

- A gene search result for “CES2 (carboxylesterase 2)



Description of gene list	Evidence	
Chaerkady (2008) *J Proteome Res* Tumor vs Non-tumor	**Fold change (Tumor/Non-tumor)**	1.667 Down
Lee (2004) *Hepatology* Genes associated with survival	**Hazard Ratio**	0.663
	P-value (Wald test)	2.200E-4
Kato (2005) *Nucleic Acids Res* HBV-tumor vs HCV-tumor	**P-value**	0.025
Kato (2005) *Nucleic Acids Res* Tumor vs Non-tumor	**P-value**	0.021
Chiang (2008) *Cancer Res* Genes specific to proliferation subgroup	**SAM score**	-14.350
	Fold change (Proliferation subgroup/Other subgroups)	5.260 Down
	q-value	0
Hsu (2007) *BMC Bioinformatics* HCC-related genes from PubMed text mining	**Related to**	HCC
Iizuka (2002) *Cancer Res* HBV-tumor vs HCV-tumor	**Fold change (HBV-tumor/Normal)**	1.890 Down
	Fold change (HCV-tumor/Normal)	1.158 Up
	Fold change (HBV-tumor/HCV-tumor)	2.189 Down
Kurokawa (2003) *J Hepatol* Non-tumor vs Normal liver	**P-value**	0.007
	Change direction (Non-tumor/Normal)	Down

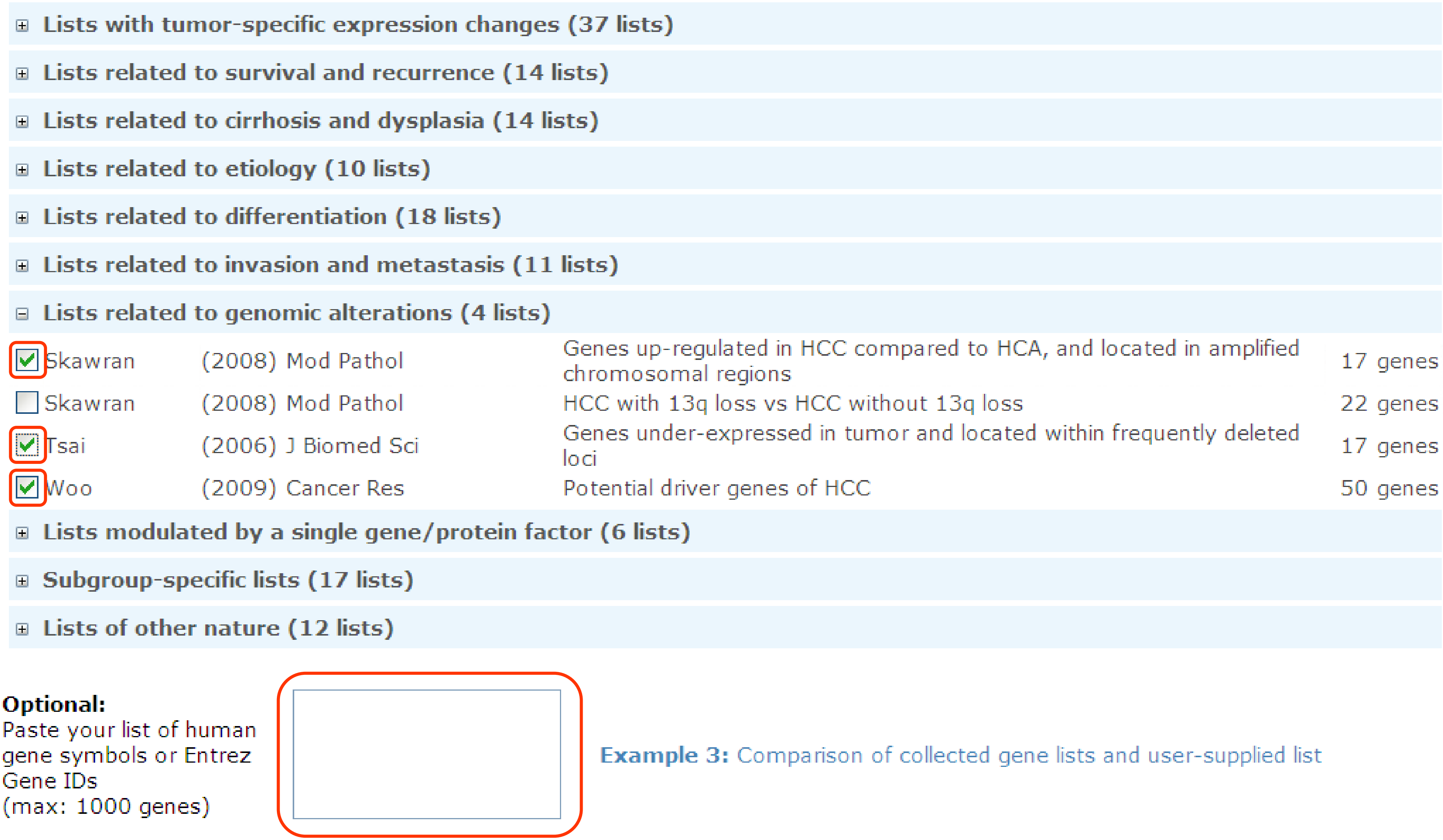
Survival

Viral infection status

DE in T vs NT

Subtype -specific

Straightforward web interface: Signature comparison



⊞ **Lists with tumor-specific expression changes (37 lists)**

⊞ **Lists related to survival and recurrence (14 lists)**

⊞ **Lists related to cirrhosis and dysplasia (14 lists)**

⊞ **Lists related to etiology (10 lists)**

⊞ **Lists related to differentiation (18 lists)**

⊞ **Lists related to invasion and metastasis (11 lists)**

⊟ **Lists related to genomic alterations (4 lists)**

☑	Skawran	(2008) Mod Pathol	Genes up-regulated in HCC compared to HCA, and located in amplified chromosomal regions	17 genes
□	Skawran	(2008) Mod Pathol	HCC with 13q loss vs HCC without 13q loss	22 genes
☑	Tsai	(2006) J Biomed Sci	Genes under-expressed in tumor and located within frequently deleted loci	17 genes
☑	Woo	(2009) Cancer Res	Potential driver genes of HCC	50 genes

⊞ **Lists modulated by a single gene/protein factor (6 lists)**

⊞ **Subgroup-specific lists (17 lists)**

⊞ **Lists of other nature (12 lists)**

Optional:
Paste your list of human gene symbols or Entrez Gene IDs
(max: 1000 genes)

Example 3: Comparison of collected gene lists and user-supplied list

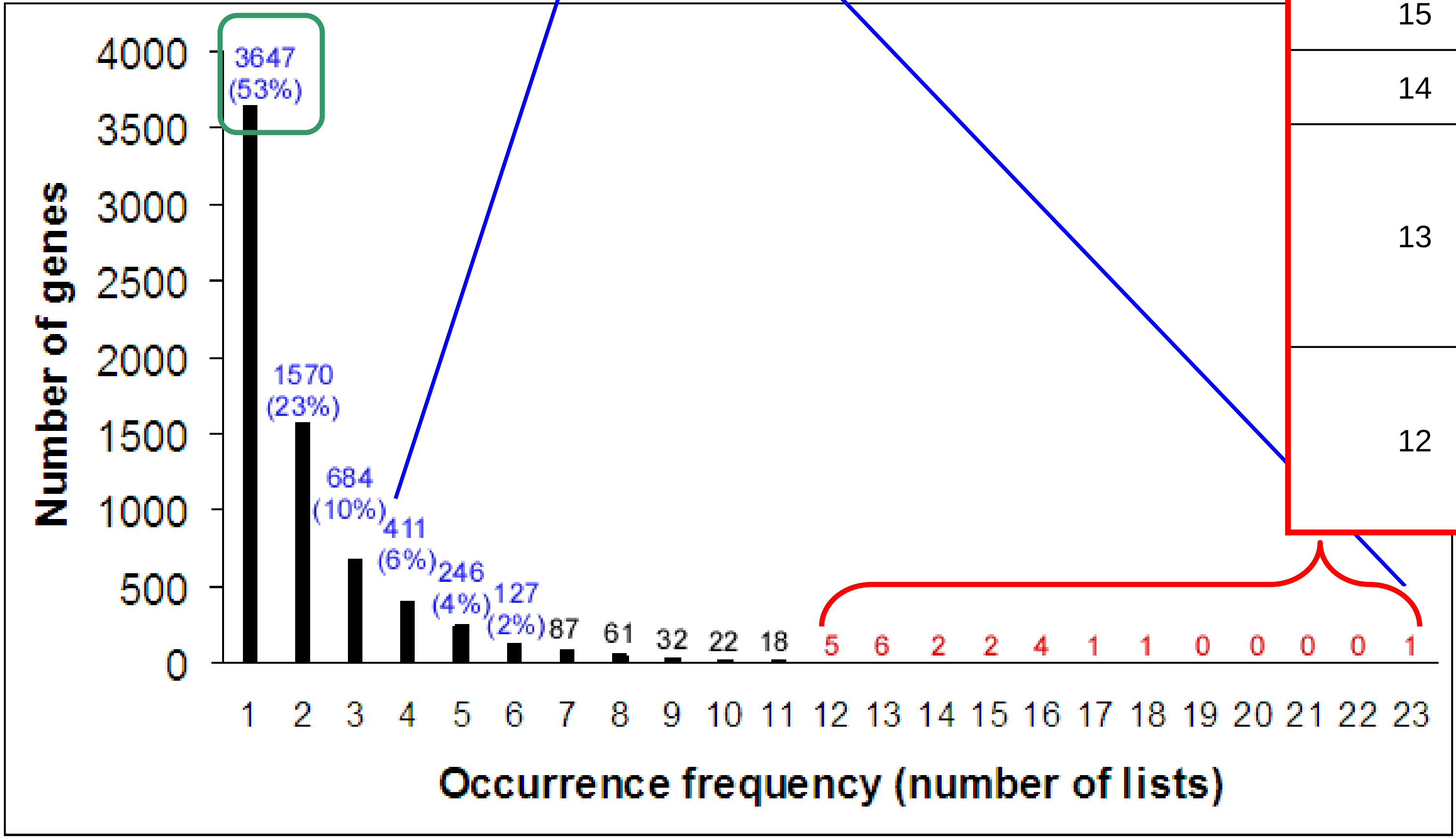
Prioritization of genes according to occurrence frequency

~20 genes occur **very frequently** in ≥ **12 signatures**

Occurrence	Symbol
23	ECHS1
18	ADH1B
17	GPC3
16	ALB
	BHMT
	PLG
	VIM
15	RGN
	TF
14	FABP1
	HPD
13	ACADSB
	CAT
	MTHFD1
	RPSA
	SLC22A1
	TDO2
12	ADH4
	CP
	CYP2E1
	PCK1
	SPARC

~1,000 genes occur **frequently** in ≥ **4 signatures**

~ A half of the genes occur in only one signature



Construction of co-occurrence network of genes

A network analysis using Liverome-collected signatures

Method

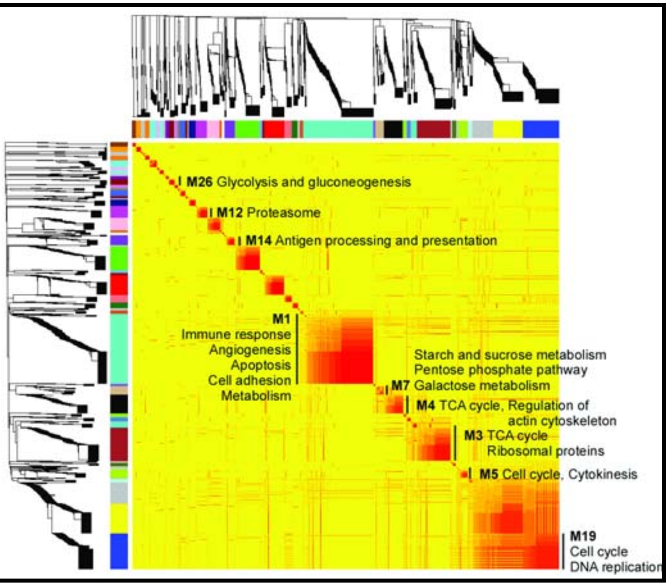
- Used WGCNA (Weighted Gene Co-expression Network Analysis) Langfelder & Horvath, 2007, BMC Bioinformatics
- Usually used to construct co-expression network from expression profile data
- Here, used it to construct co-occurrence network from signature data
- A pair of genes are considered as similar if they co-occur in many of the signatures

	Signature1	Signature2	Signature3	Signature4	Signature5	Signature6	. . .	Signature *N*
Gene A	🟦		🟦					
Gene B			🟥	🟥				

$$\text{Co-occurrence (Jaccard similarity coefficient)} = \frac{\text{\# Signatures containing both gene A and gene B}}{\text{\# All other signatures}}$$

Result: a co-occurrence network

- Genes are shown on rows and columns
- Color coding represents the similarity measure
- Each block represents a module which consists of a set of genes that have similar liver cancer signature membership
- Cancer-related pathways are enriched in the modules (glycolysis, cell cycle, apoptosis, etc.)
- Co-occurrence network constructed from liver cancer signature data alone recapitulates known liver cancer biology



Summary

- Comprehensive collection of liver cancer-related gene signatures
- All the database content was made into a self-contained form by extensive manual annotation
- Limitations: All limitations inherent to publication-based signature database
  - Each signature contains only a few selected genes above the significance cutoff
  - Each signature was derived from its own data processing scheme, which may decrease reproducibility across datasets
- Usefulness
  - Most useful to retrieve known differential expression information of a gene
  - To compare your own gene list with previously reported gene lists
  - An interesting bioinformatics analysis may be possible using the DB contents

Main contributors

Korea Research Institute of Bioscience & Biotechnology	Hyang-Sook Yoo	Project conception & supervision
	Langho Lee	Database and web programming
Pfizer	Kai Wang	Co-occurrence network analysis
	Gang Li	Biological discussion
Personal Genome Institute	Jong Bhak	Project initiation

Also thank to the authors of the articles that we collected and included in the database