COXPRESdb

gene coexpression database for Human, mouse and rat

Kengo Kinoshita

kinosita@hgc.jp

Human Genome Center
Institute of Medical Science
University of Tokyo

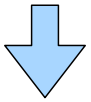
2009/07/29
@IUPS2009

Today's topic

- Brief introduction of COXPRESdb by KK
  - What is gene coexpression?
    - Biological meaning of coexpression
  - What is the characteristics of COXPRESdb?
- Tutorial by Takeshi Obayashi
  - How to use in practical scene
- Open for discussions; questions & comments

Gene function identification in silico

Many uncharacterized genes on human genome

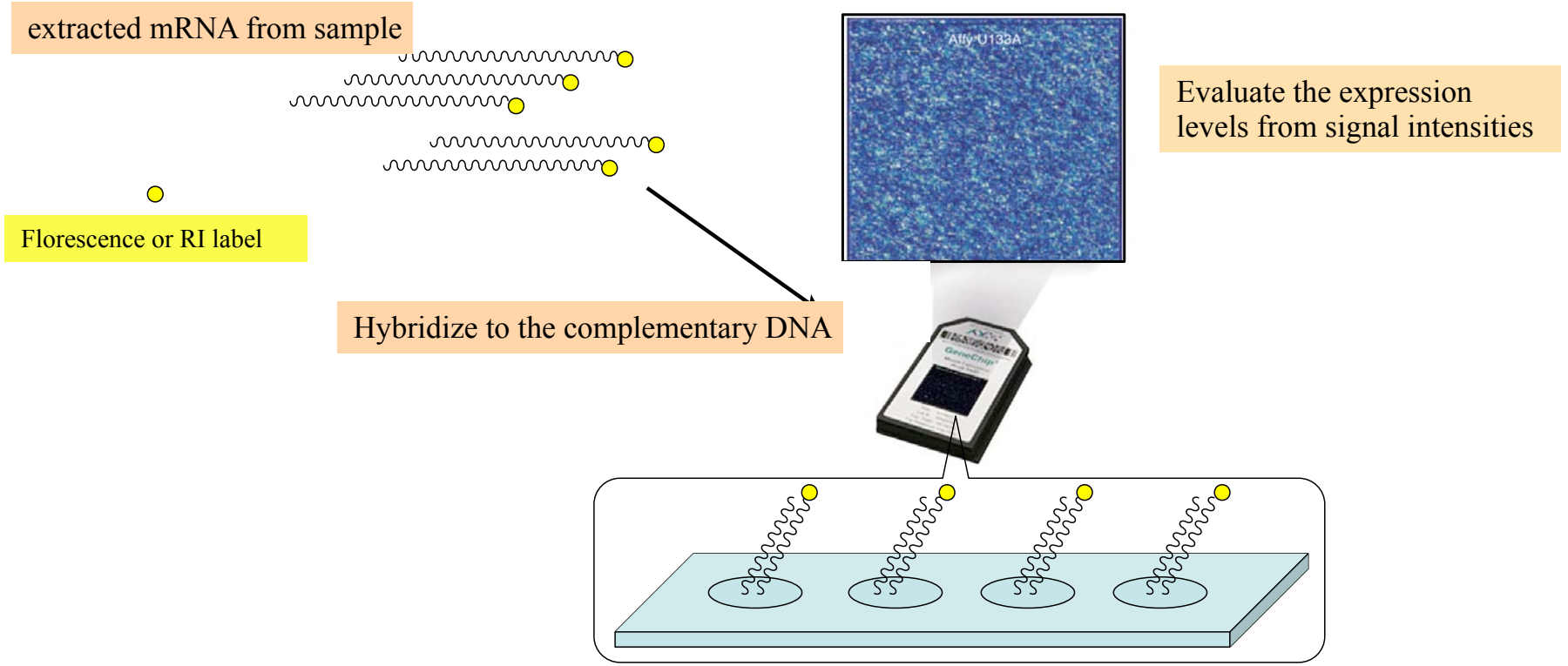


How to identify the function of proteins?

- **Sequence similarity search**
  - BLAST, PSI-BLAST ... many methods
  - Weak for paralogous proteins and orphan proteins
- **Structure similarity search**
  - SeSAW, GIRAF, eF-site ... some methods
  - Suitable for molecular function
  - Weak for the identification of cellular function
    (∵ CF is determined by interaction network)
- **Expression pattern similarity (today's topic)**
  - Gene coexpression
  - For cellular function

Expression pattern by DNA micro array

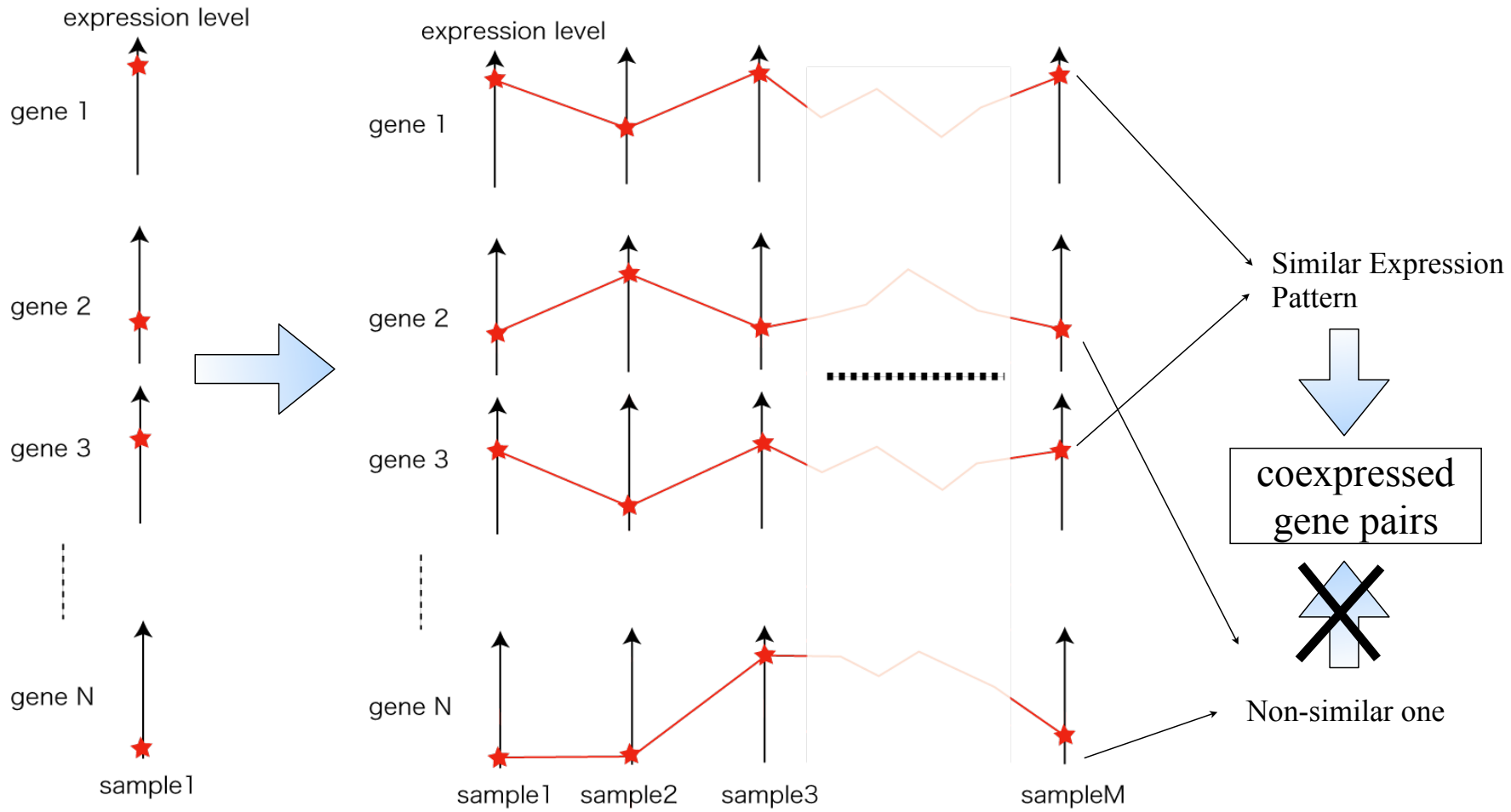
- Expression levels of genes are measured by DNA micro array
  - High density oligo nucleotide array
  - Measure about 30-50 thousands of mRNA levels at a time



Similarity of the expression patterns
Gene co-expression

Single Array Data

Multiple Array Data

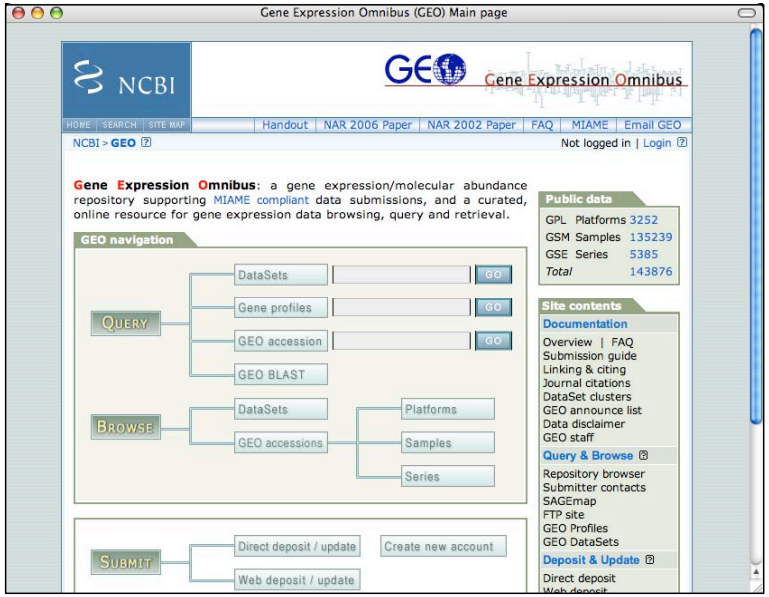


Usual measure of expression pattern is Pearson correlation coefficient (PCC).

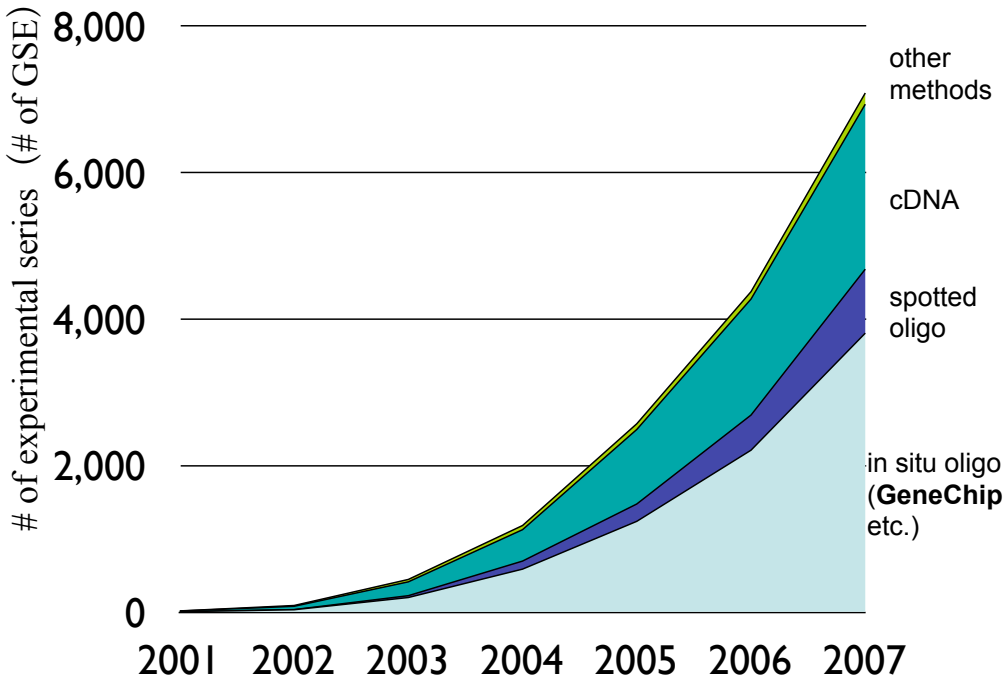
Accumulation of Expression Data

- Increase of expression data in public database

NCBI GEO：repository



# of GeneChip samples@NCBI GEO



- Quality enhancement

1994: 100μ
2006: 11μ：55k-probe, 11k locus
in development: 2μ：25 times density
→ almost all genes in higher organisms are covered!

Now, it is a good timing for us to get reliable gene co-expressions

What is the meaning of co-expression?

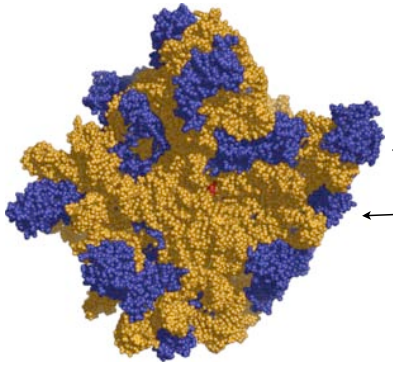
Cellular function requires coexistence of proteins!

Co-expressed gene pairs have some functional relationships

Various functional relationships were known

co-expressed gene pairs

Supulamolecule



(ex) ribosome

metabolic pathway

Aspartate

Ass1

co-expression

arginosuccinate

Asl

Arginine

- signal transduction
- expressed in the same tissue
- …

Datasets of COXPRESdb

- Raw data obtained from NCBI/GEO
- RMA summarization
- Expression level normalization by gene centering
  - Normalized data are available in COXPRESdb
- One platform was selected for each species
  - GeneChip (affymetrix) data in the current version

Current version:

	n. samples	n. probes	n. genes	Platform
human	4401	56163	19777	human genome U133 Plus 2.0
mouse	2226	45037	21036	mouse genome 430-2.0
rat	632	31099	11912	rat genome 230-2.0

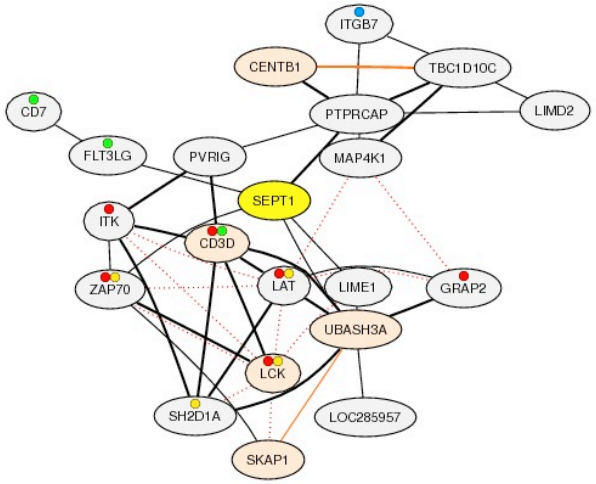
Next version:
(Sep, 2009 ?)

	n. samples
human	24266
mouse	14077
rat	4440

Characteristics of COXPRESdb

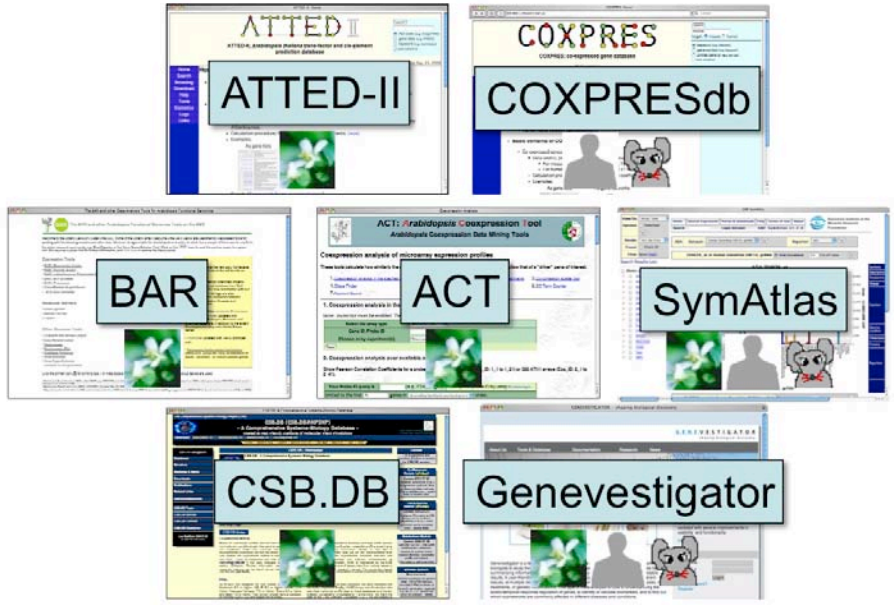
- Optimized for **function identification** (next slide)
- Coexpression measured by **large scale data**
- **Network** drawing including KEGG annotation
  - conserved coexpression (orange lines)
  - PPI data is also shown (red dotted lines) (from HPRD & IntAct)
- **Rat** data are available

An example network



KEGG ID	Title	#genes	Link to the KEGG map (multiple genes)
hsa04660	T cell receptor signaling pathway	6	
hsa04650	Natural killer cell mediated cytotoxicity	4	
hsa04640	Hematopoietic cell lineage	3	
hsa04510	Focal adhesion	1	

coexpression DB in the world



Several options to get coexpressed genes

GeneChip data

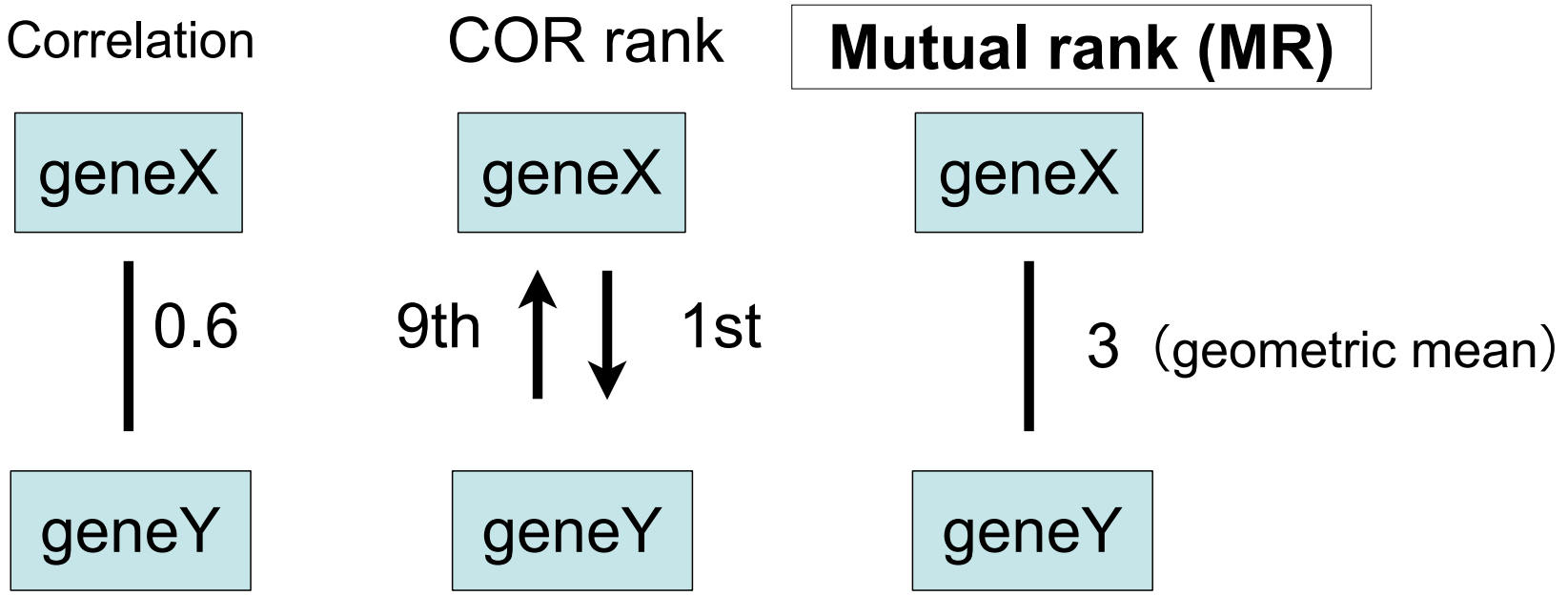
calculation

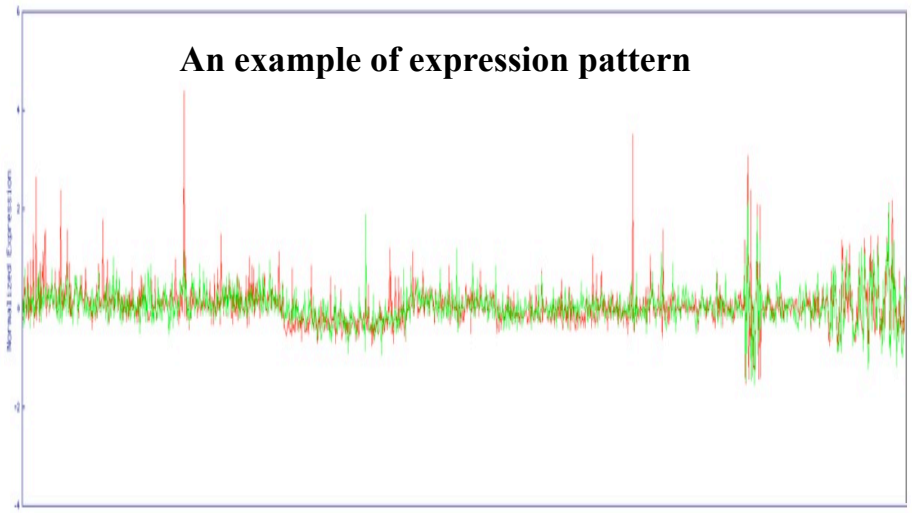
→→→ Gene coexpression list

1. GeneChip sample selection
   - **All**, subgroup
2. GeneChip summarization
   - **RMA**, GCRMA, MAS5, Plier
3. Sample redundancy treatment
   - **weight**, no weight, average
4. Measure of coexpression
   - correlation, cor. rank, **mutual rank**

Combinations were optimized for the function prediction

Mutual Rank





		MR*	COR*	symbol	function	coexpression detail
0	☑			Sept4	septin 4	
1	☐	3.2	0.25	Higd1b	HIG1 domain family, member 1B	[detail]
2	☐	4.6	0.24	Car2	carbonic anhydrase 2	[detail]
3	☐	7.3	0.26	Frmd4b	FERM domain containing 4B	[detail]
4	☐	17.2	0.24	6230424C14Rik	RIKEN cDNA 6230424C14 gene	[detail]
5	☐	17.3	0.26	Agpat3	1-acylglycerol-3-phosphate O-acyltransferase 3	[detail]
6	☐	25.3	0.24	Tmem204	transmembrane protein 204	[detail]
7	☐	28.6	0.26	Nme3	expressed in non-metastatic cells 3	[detail]
8	☐	41.2	0.22	Fn3k	fructosamine 3 kinase	[detail]
9	☐	47.8	0.24	Pfkm	phosphofructokinase, muscle	[detail]
10	☐	48.6	0.24	Myo6	myosin VI	[detail]
11	☐	52.2	0.20	5430421B17	hypothetical protein 5430421B17	[detail]
12	☐	55.3	0.22	Ptprb	protein tyrosine phosphatase, receptor type, B	[detail]
13	☐	56.0	0.25	Ndufs3	NADH dehydrogenase (ubiquinone) Fe-S protein 3	[detail]
14	☐	61.8	0.23	Cryab	crystallin, alpha B	[detail]
15	☐	62.0	0.24	Gkap1	G kinase anchoring protein 1	[detail]
16	☐	64.7	0.23	Iqsec1	IQ motif and Sec7 domain 1	[detail]
17	☐	66.9	0.22	Lamb2	laminin, beta 2	[detail]
18	☐	66.9	0.21	Ushbp1	Usher syndrome 1C binding protein 1	[detail]
19	☐	77.1	0.22	C530044N13Rik	RIKEN cDNA C530044N13 gene	[detail]
20	☐	80.4	0.20	Afap1l1	actin filament associated protein 1-like 1	[detail]

Summary

- Increase of the quality and quantity of DNA micro array data
- Similarity of gene expression pattern (coexpression) can be a good measure of the functional relation

- COXPRESdb provides
  - gene coexpression table sorted by MR (mutual rank)
    - Tables are pre-calculated (quick response)
    - A tool to pick up the coexpressed genes from multiple query genes is available
  - gene coexpression network with functional annotation
  - Human, Mouse and Rat coexpression
    - Homologous coexression
  - PPI data taken from HPRD and IntAct
- Usually, we will update COXPRESdb twice in a year
  - Next update will be Sep 2009 with a few new features

FAQ

- Negative correlation
  - negative correlations were used as is. In other words, negatively correlated gene pairs will have large MR.
- How to treat the multiple probes for a single gene?
  - First, calculate all probe-vs-probe correlation, then the maximum value for a gene pair was used as the PCC value of the gene pair.
- promiscuous probe (1 probe → multiple genes)
  - All promiscuous probes were eliminated
- How often updated?
  - Usually, twice in a year
- Further questions are welcome at **coxpresdb@hgc.jp**