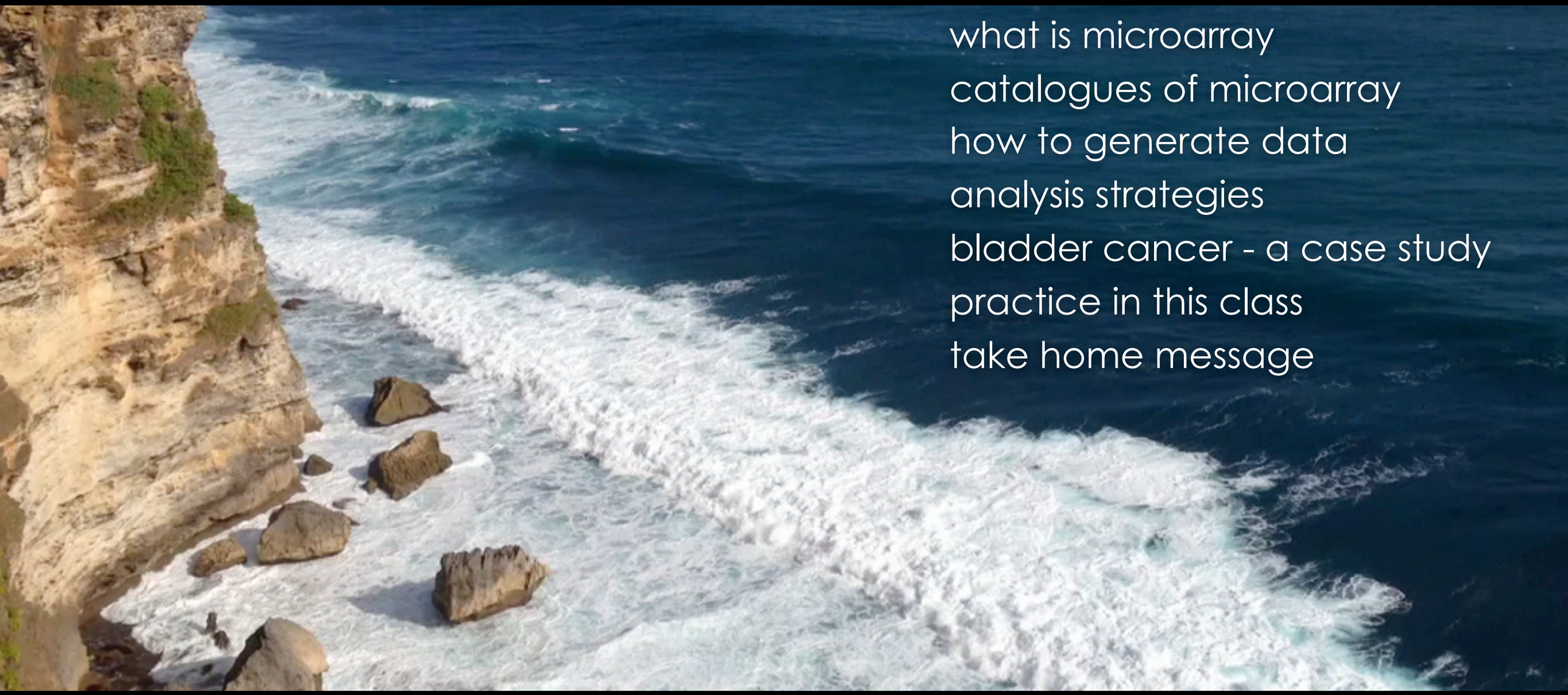
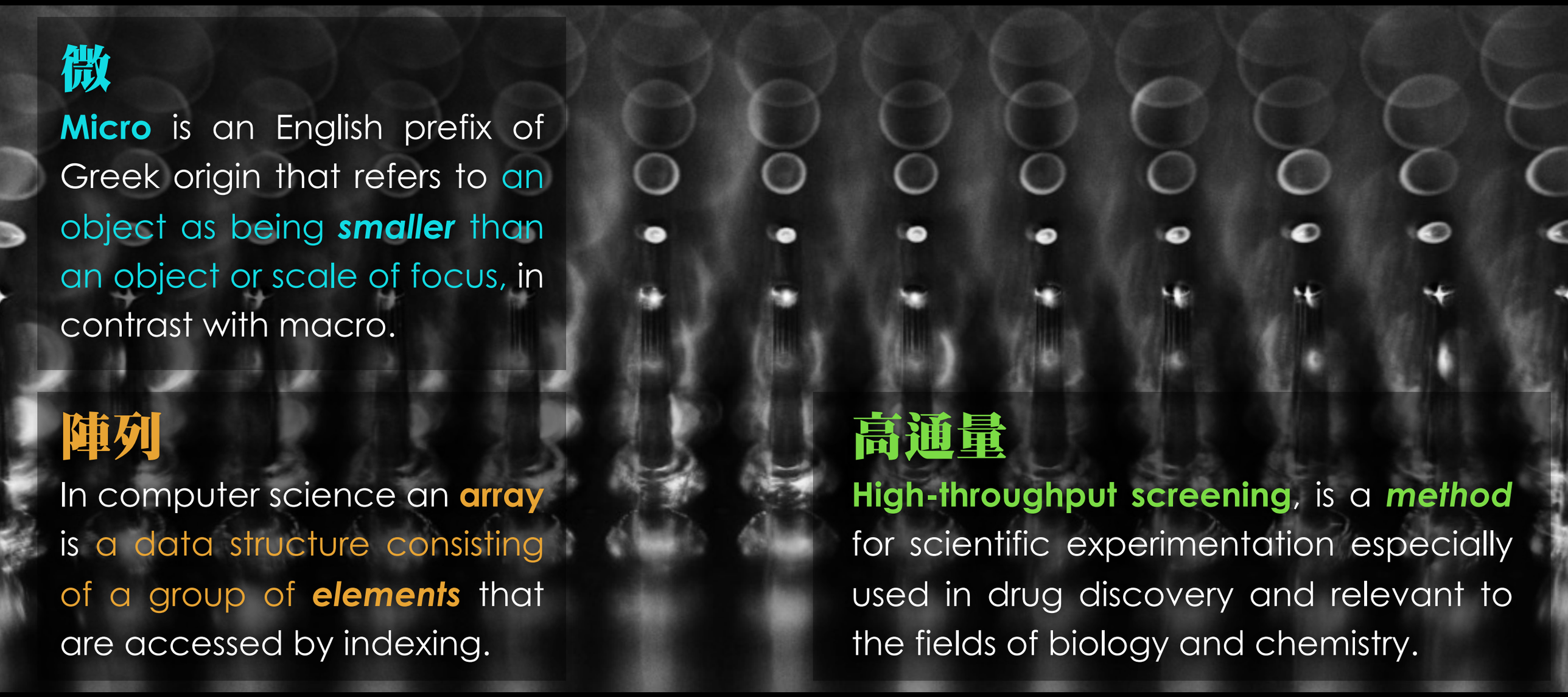
Microarray - Analysis and Its Strategies

Shin, Jyh Wei Ph.D.



what is microarray
catalogues of microarray
how to generate data
analysis strategies
bladder cancer - a case study
practice in this class
take home message

Pura Luhur Uluwatu , Bali, Indonesia



微

Micro is an English prefix of Greek origin that refers to an object as being ***smaller*** than an object or scale of focus, in contrast with macro.

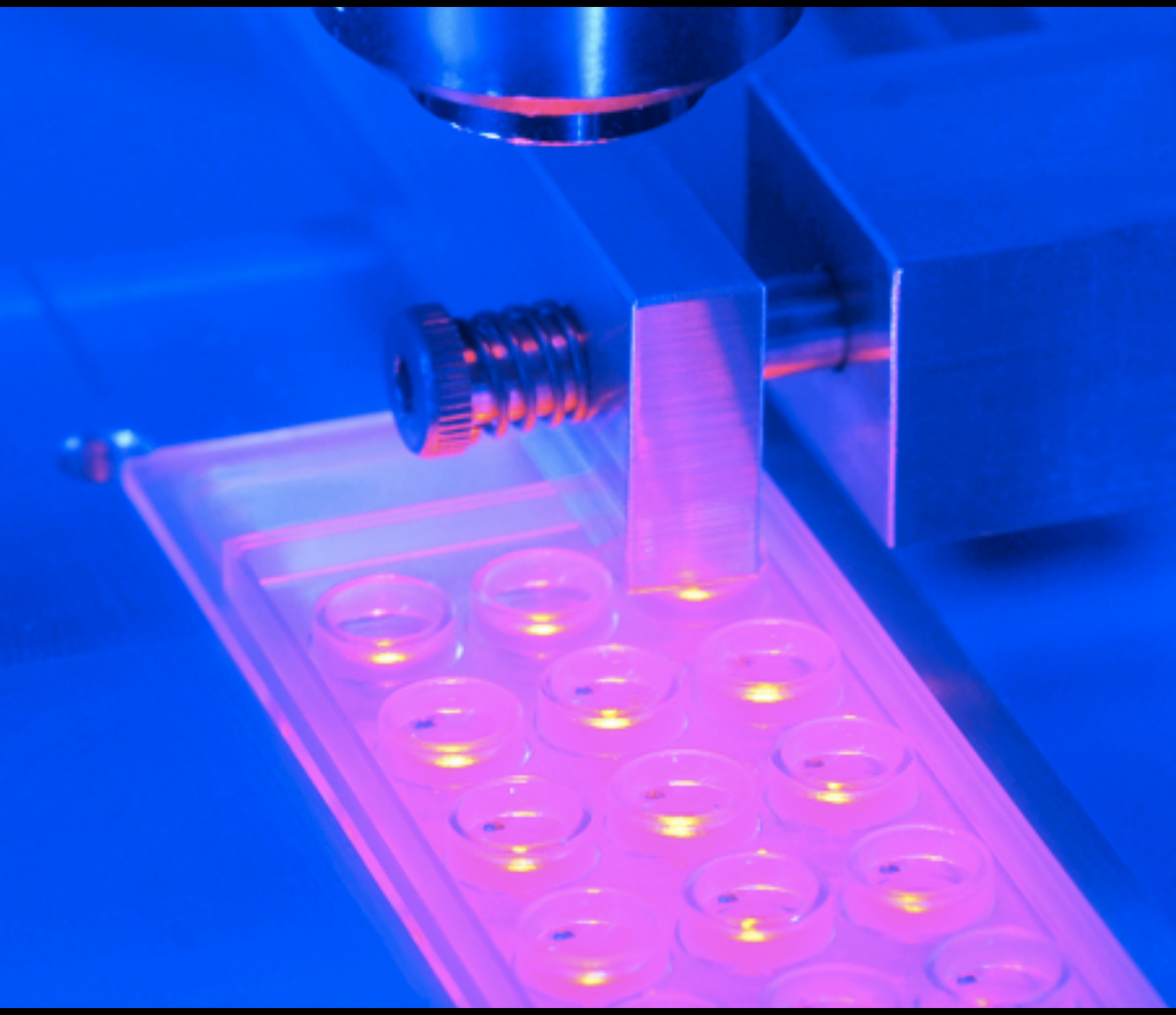
陣列

In computer science an **array** is a data structure consisting of a group of ***elements*** that are accessed by indexing.

高通量

High-throughput screening, is a ***method*** for scientific experimentation especially used in drug discovery and relevant to the fields of biology and chemistry.

500px.com

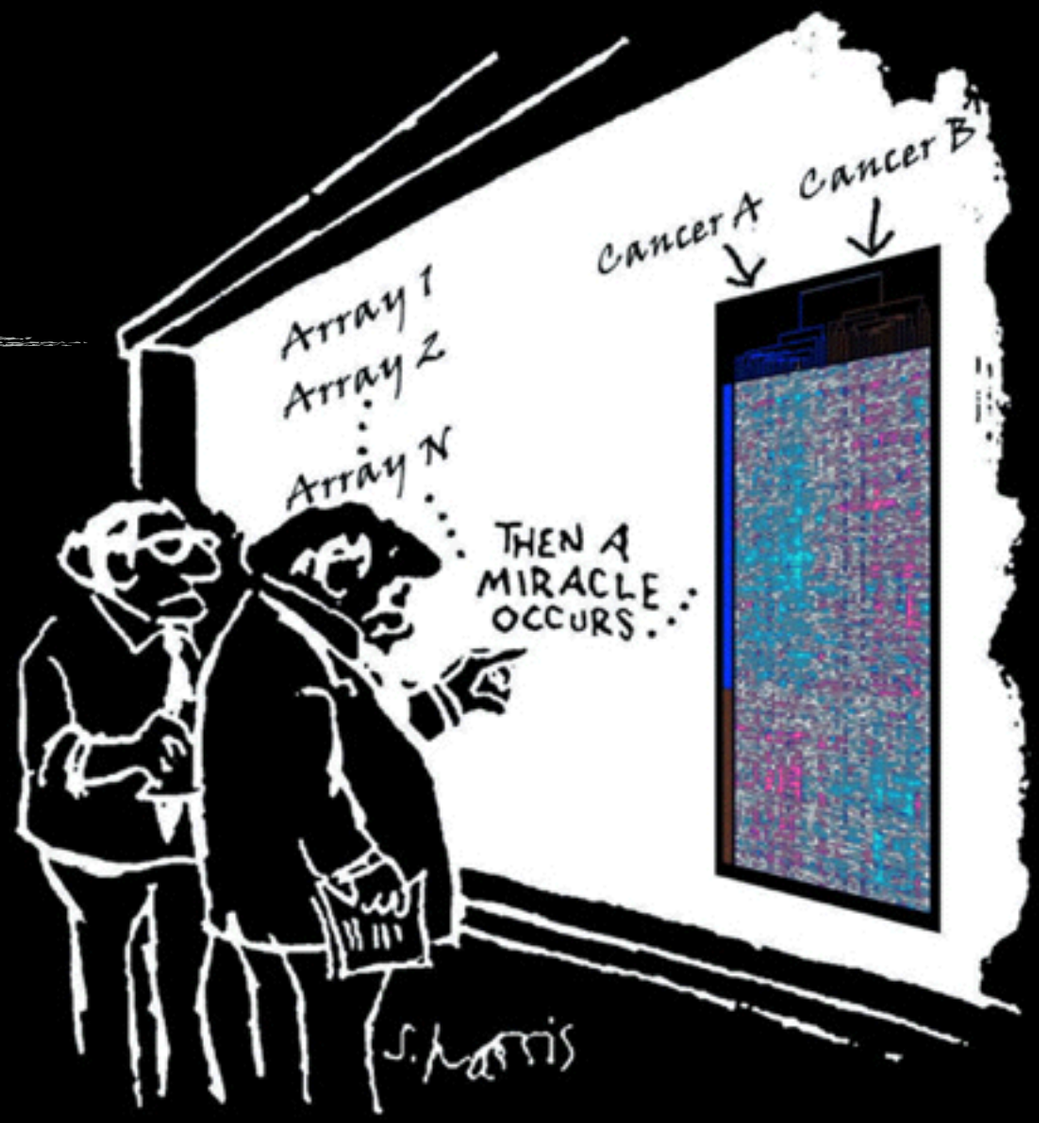


Microarrays are miniaturized biological devices consisting in molecules, for example DNA or protein, named the "**probes**", that are **orderly arranged** at a **microscopic scale onto a solid support** such as a membrane or a glass microscope slide.

http://www.ilt.fraunhofer.de/en/publication-and-press/press-release/press-release-2013/press-release-2013-07-26.html

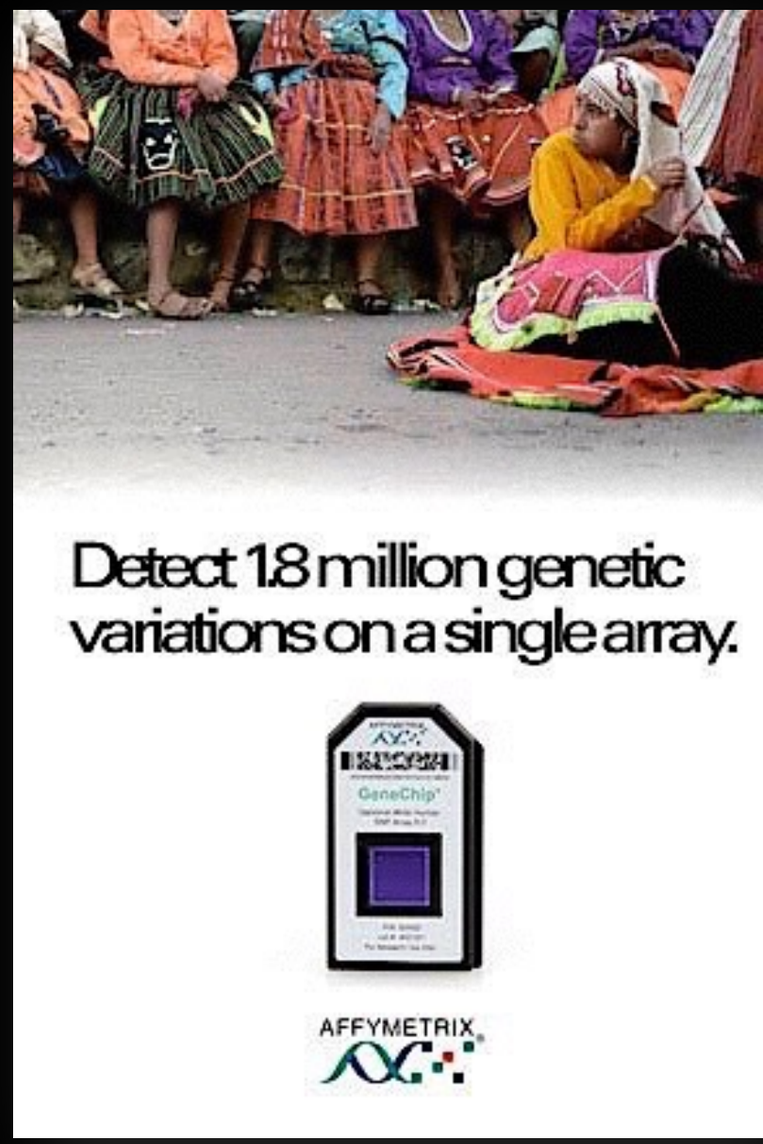
Microarray Features

- ***Parallelism***: Thousands of genes simultaneously
- ***Miniaturization***: Small chip size
- ***Multiplexing***: Multiple samples at the same time
- ***Automation***: Chip manufacturing and Reagents

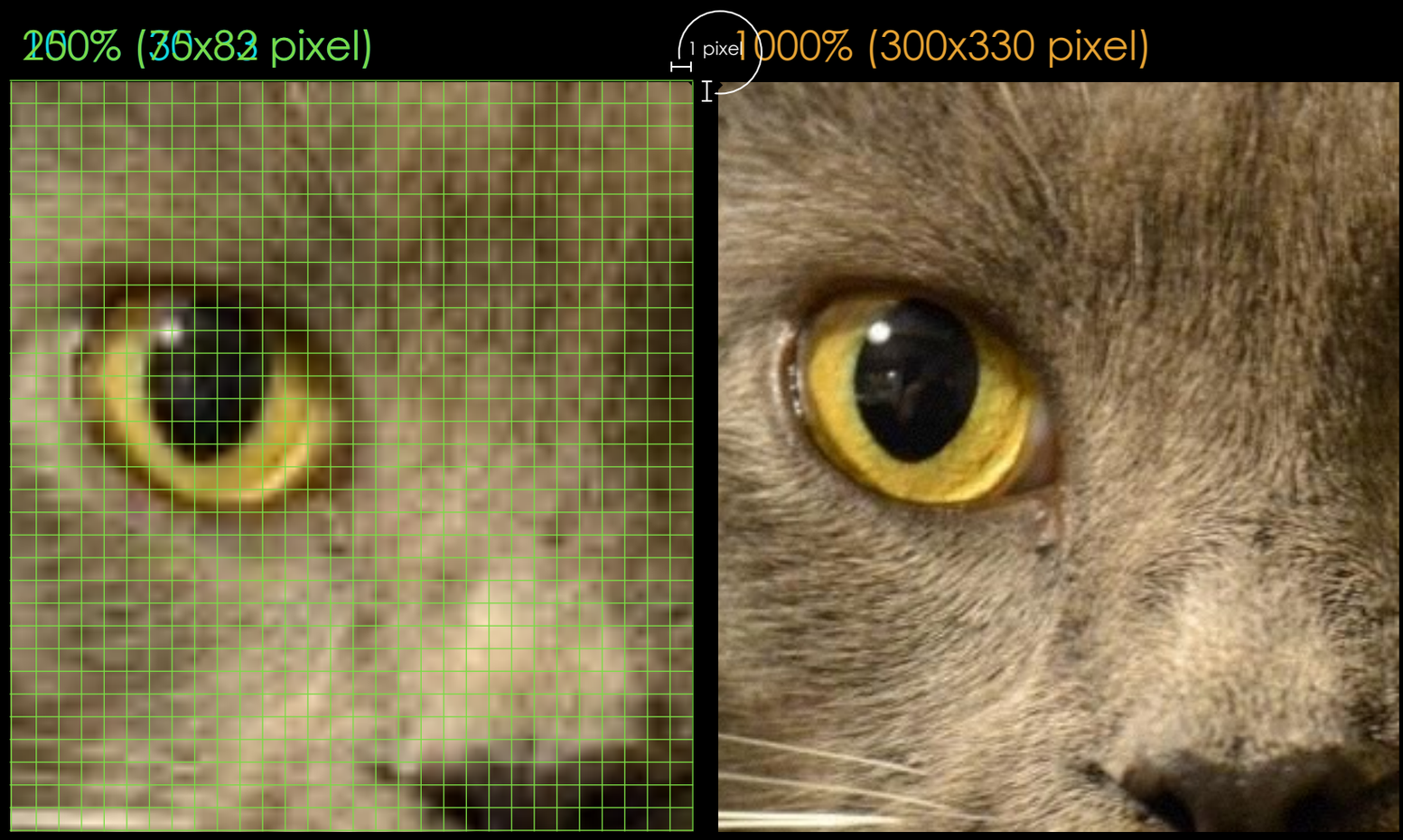


What circumstances brought out this technology?

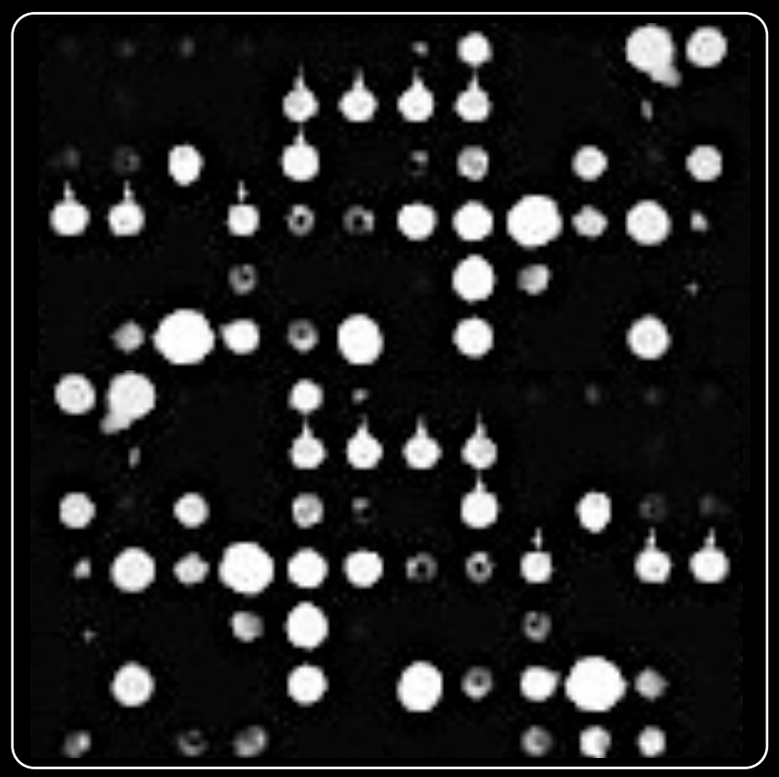
- Fact: biological processes are complicated with many molecules ***working together***.
- Biologists are eager to obtain the "***whole picture***".
- ***Genome sequences availability***.
- ***Computer aids.***



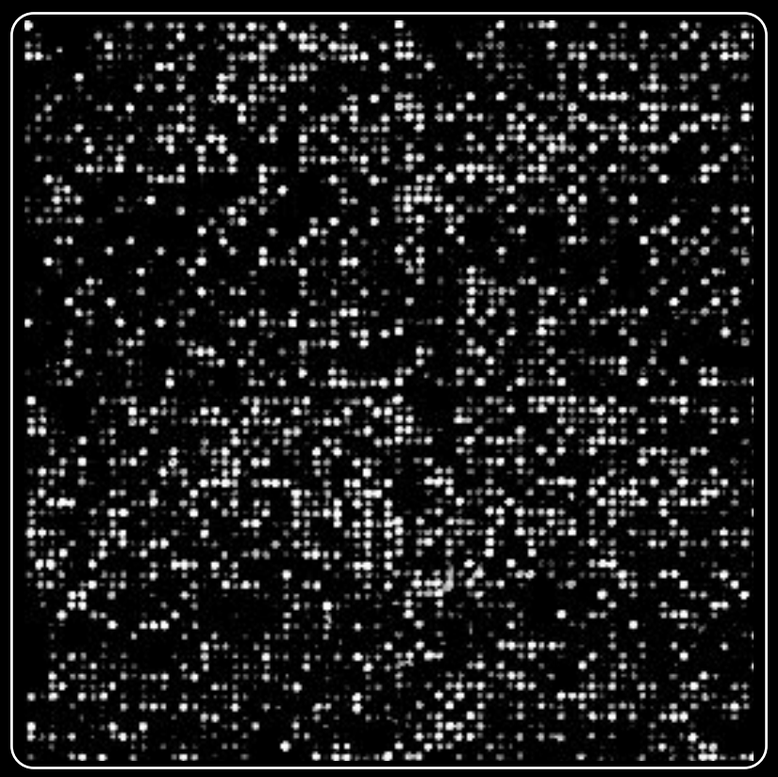
Density of microarray



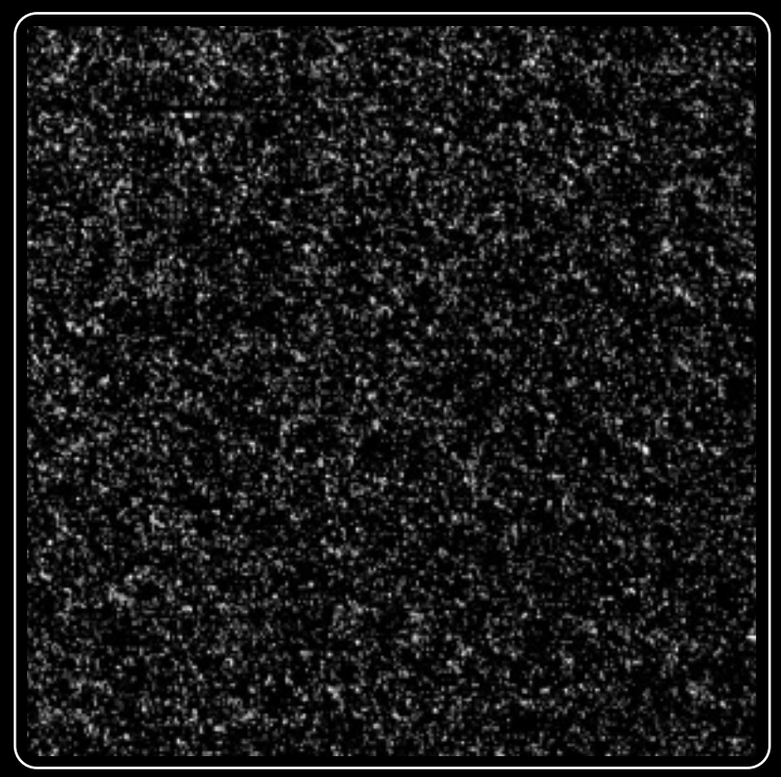
Density of microarray



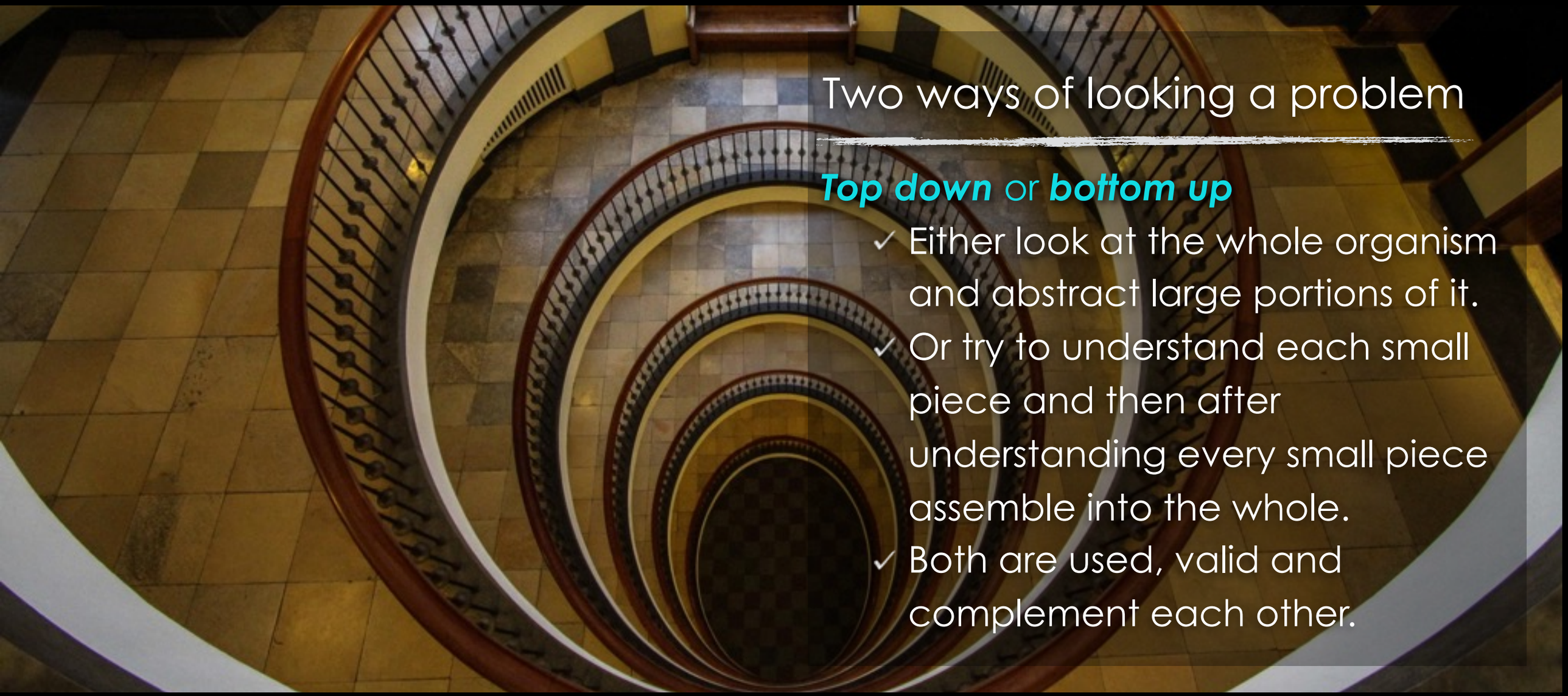
Isotope
Nylon – cDNA (300-900 nt)



Two-channel
cDNA or Oligo (80 nt)
500 – 11,000 elements



Affymetrix
Silicone – oligo (20 nt)
22 ,000 elements



Two ways of looking a problem

Top down or ***bottom up***

- ✓ Either look at the whole organism and abstract large portions of it.
- ✓ Or try to understand each small piece and then after understanding every small piece assemble into the whole.
- ✓ Both are used, valid and complement each other.

500px.com

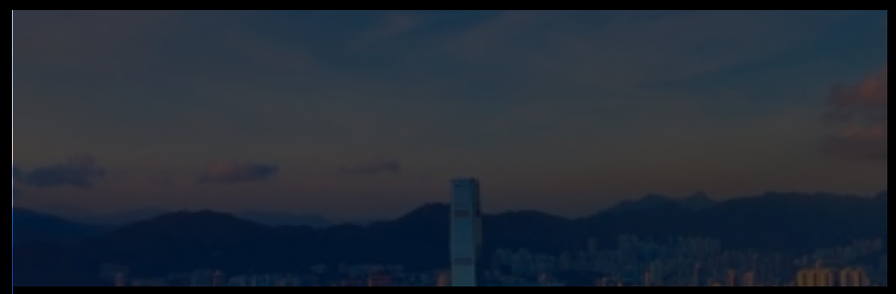
Combination of microarray

Northern blot

microarray analysis

mRNA expression →





CGH →

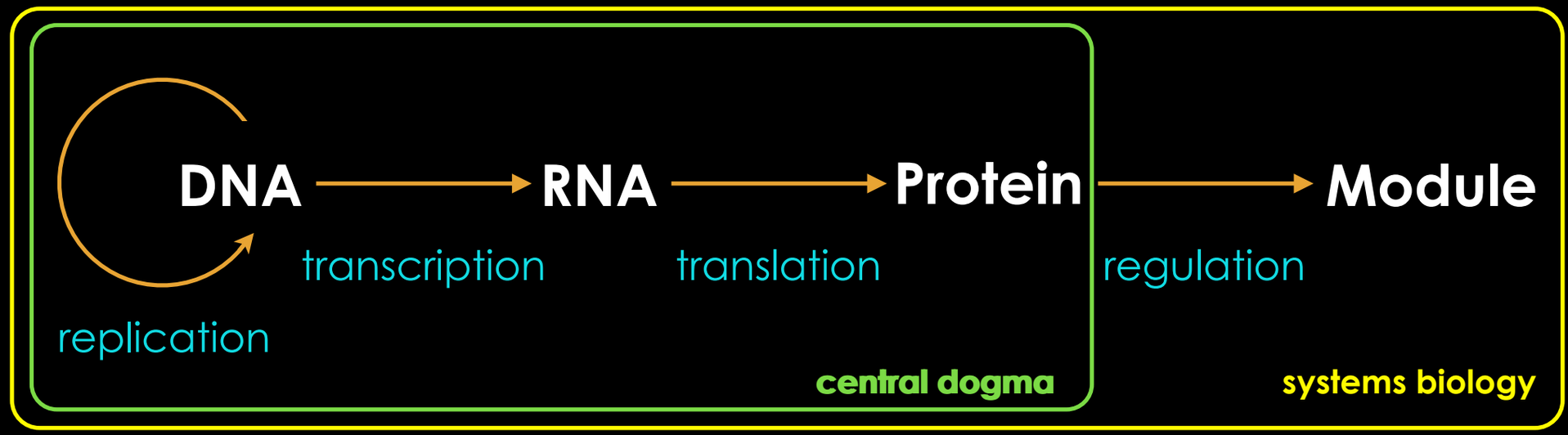


CHIP →



specificity, density and content of probes

low → high


DNA
RNA
Protein
Module
transcription
translation
regulation
replication
central dogma
systems biology

Catalogues of microarray

molecule as probe

aCGH
cDNA array
CHIP on chip
SNP array
miRNA array
ELISA array
Tilling array
Protein array

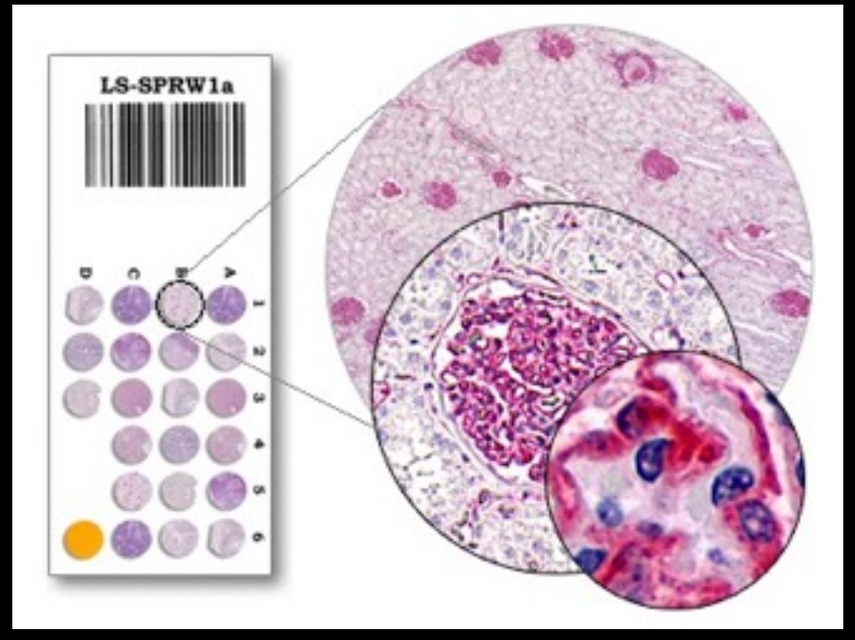
cell as probe

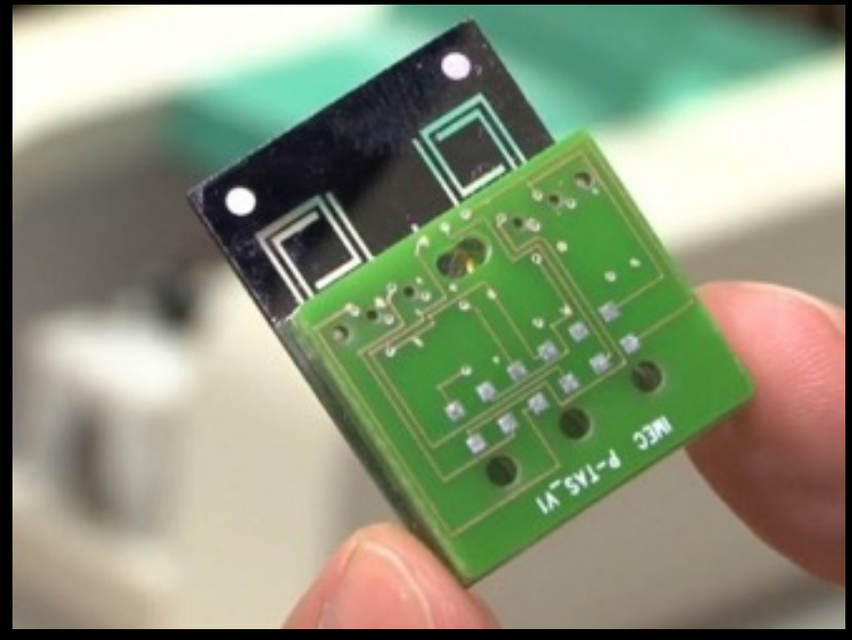
Cell array
Tissue array

Arranged device

qPCR array
Lab-on-a-chip

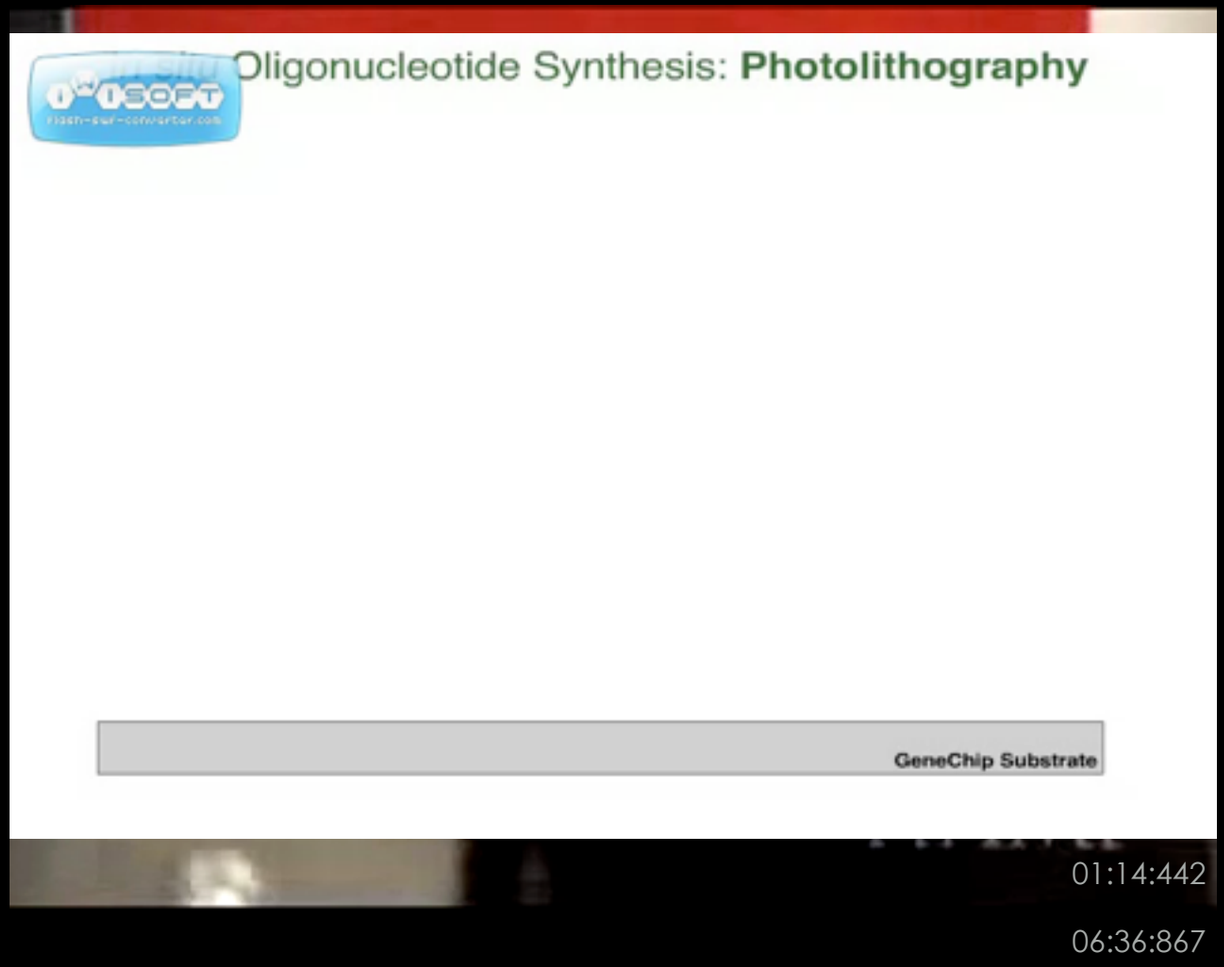






Molecule as probe

A **DNA microarray** (also commonly known as gene or genome chip, DNA chip, or gene array) is a collection of microscopic DNA spots, commonly representing single genes, arrayed on a solid surface by covalent attachment to a chemical matrix. DNA arrays are different from other types of microarray only in that they either **measure DNA or use DNA as part of its detection system**. Qualitative or quantitative measurements with DNA microarrays utilize the selective nature of DNA-DNA or DNA-RNA hybridization under high-stringency conditions and **fluorophore-based detection**. DNA arrays are commonly used for expression profiling, i.e., monitoring expression levels of thousands of genes simultaneously, or for comparative genomic hybridization.



Molecule as probe

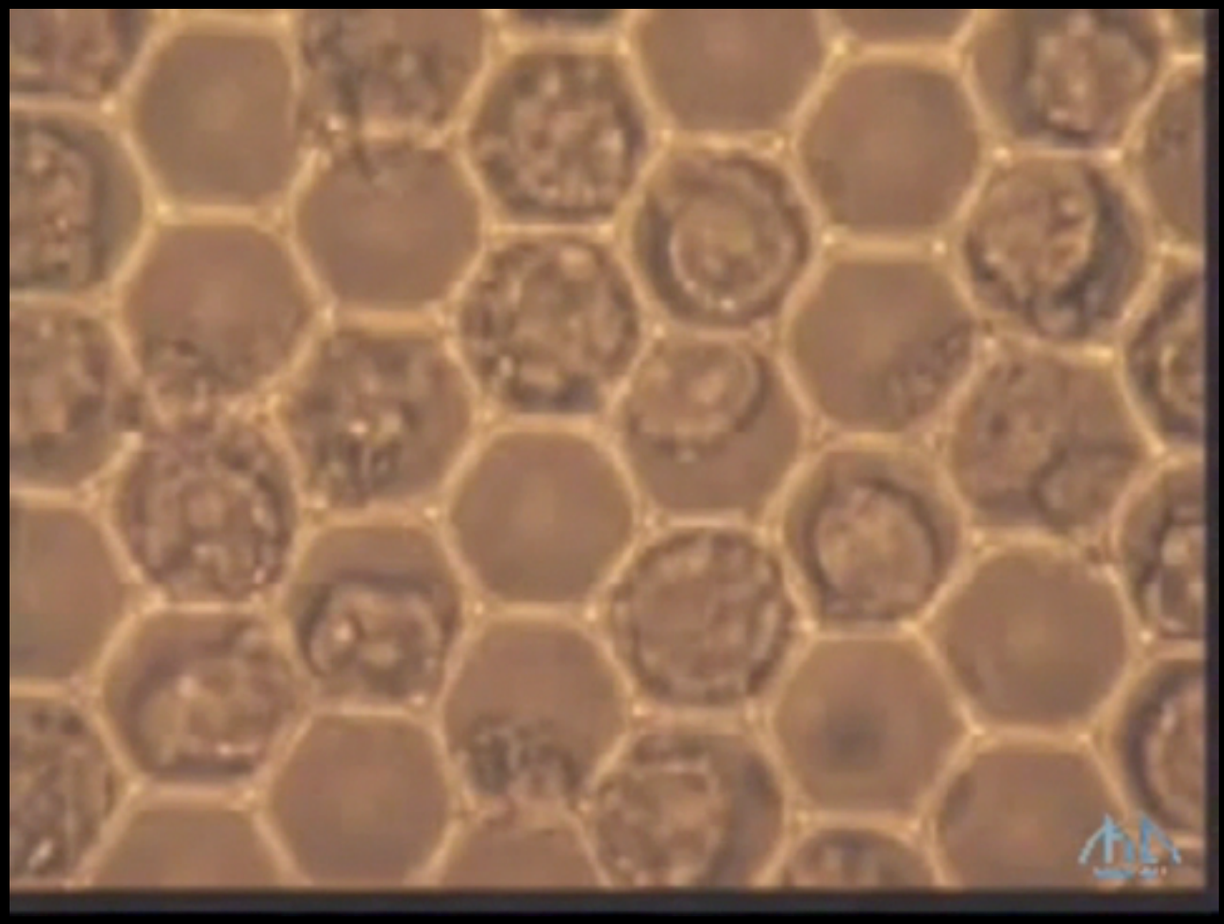
Enzyme-Linked ImmunoSorbent Assay, or **ELISA**, is a biochemical technique used mainly in immunology to detect the presence of an antibody or an antigen in a sample. The ELISA has been used as a diagnostic tool in medicine and plant pathology, as well as a quality control check in various industries. ELISA array (or array ELISA) is a new technology capable of simultaneously **identifying multiple biomarkers to generate a biochemical profile.** Once the ELISA protocol is completed the array is imaged using either a CCD imaging system or X-ray film to capture the **chemiluminescent signal**. The pixel intensity of each spot is directly correlated to the concentration.



01:13:713

Cell as probe

Cell array is a common term for different techniques, which are used in genomic-level cell biological testing. The live cell array, a microscope slide-based high content analysis tool, enables **multi-parametric imaging-based assays on thousands of intact individual cells**, including non-adhering blood and bone marrow cells. Cells can be observed in their own identified location, tracking individual, real-time responses to intervention. Multiple functional assays can be performed on a **living cell**, followed by post-fixation studies on the same cell to maximize usage of cell samples.



02:04:040

Cell as probe

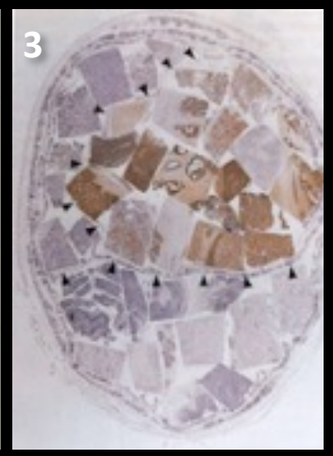
Tissue microarrays are produced by a method of re-locating tissue from conventional histologic paraffin blocks so that tissue from **multiple patients or blocks can be seen on the same slide**. This is done by using a needle to biopsy a standard histologic sections and placing the core into an array on a recipient paraffin block.



whole tissue



tissue cocktail



tissue sausage



tissue array

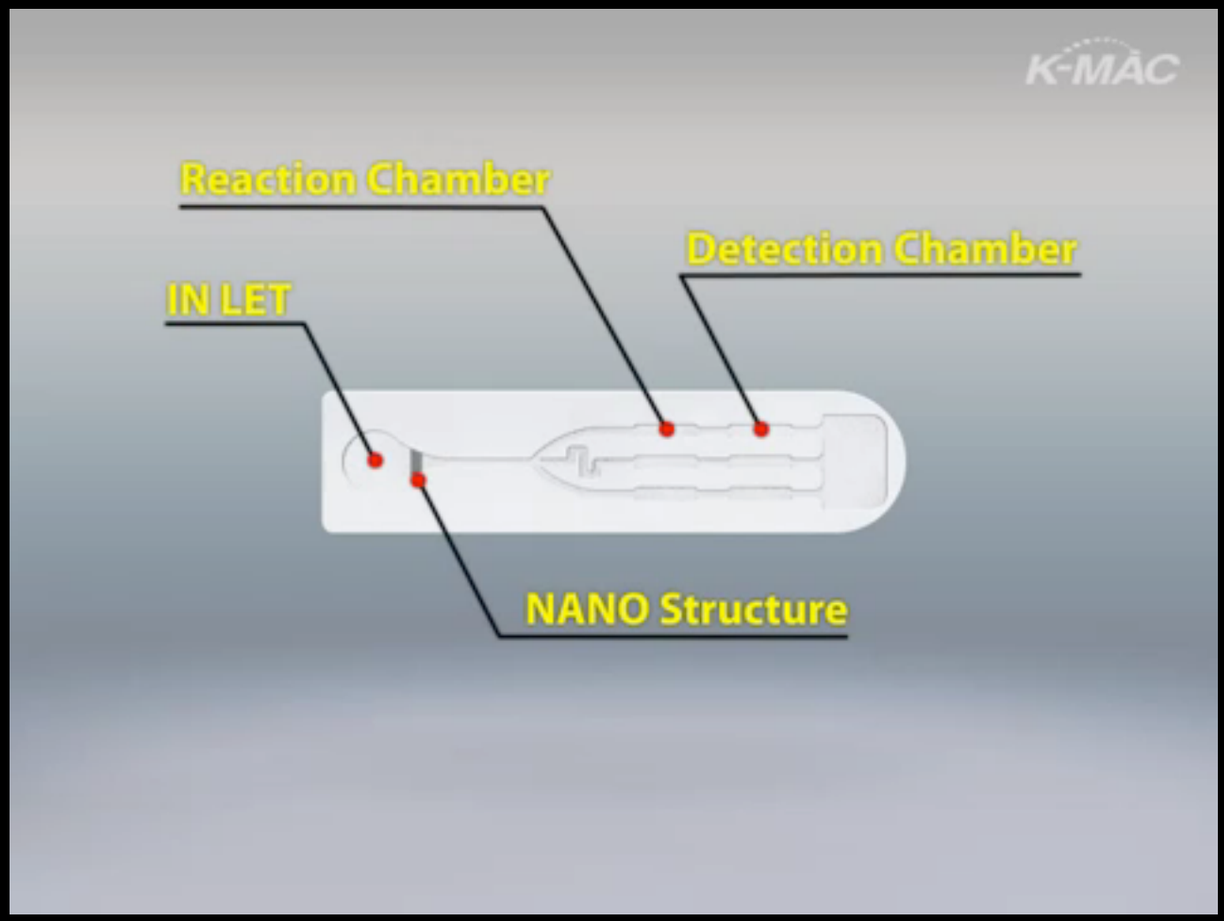


03:39:422

Arranged device

Lab-on-a-chip (LOC) is a term for **devices** that integrate (multiple) laboratory functions on a single chip of only millimeters to a few square centimeters in size and that are capable of handling extremely **small fluid volumes** down to less than pico liters.





00:59:325



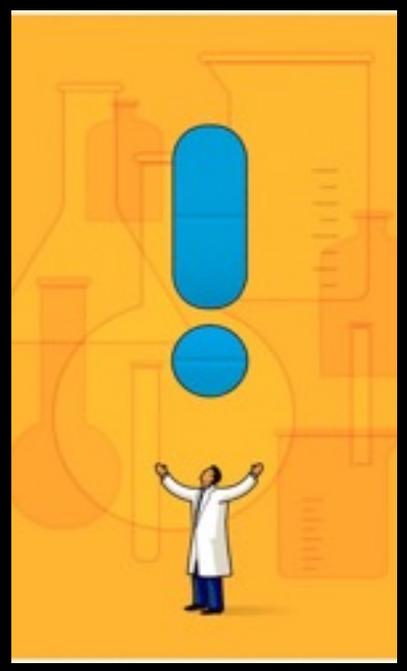
The 6 steps of a microarray experiment

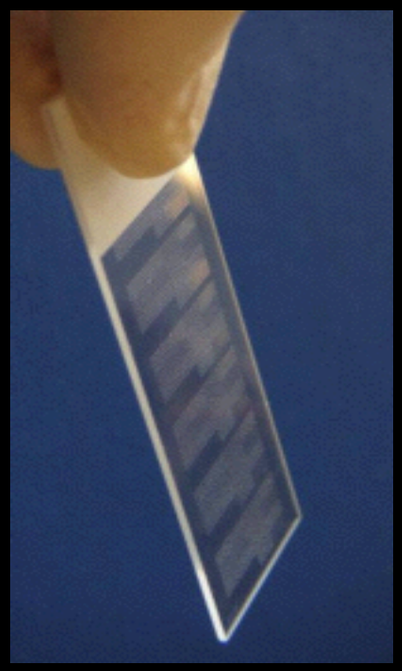
- Manufacturing of the microarray
- **Experimental design and choice of reference**: what to compare to what?
- Target preparation (labeling) and hybridization
- Image acquisition (scanning) and quantification (signal intensity to numbers)
- Database building, filtering and normalization
- **Statistical analysis, data mining and modulation**

500px.com

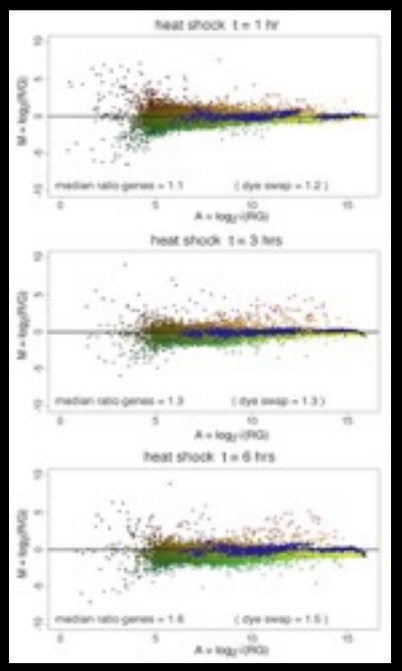
Experimental methodology for microarray

1. *Experimental Design*
2. *Microarray Experiments*
3. *Image Processing*
4. *Data Normalization*
5. *Data Analysis*











Data report format of microarray

Probe location and gene annotation file

xxx.gal

Color image file (8 bit)

xxx.jpg

Reconstructed image file (16 bit)

xxx.tif

Scan result file

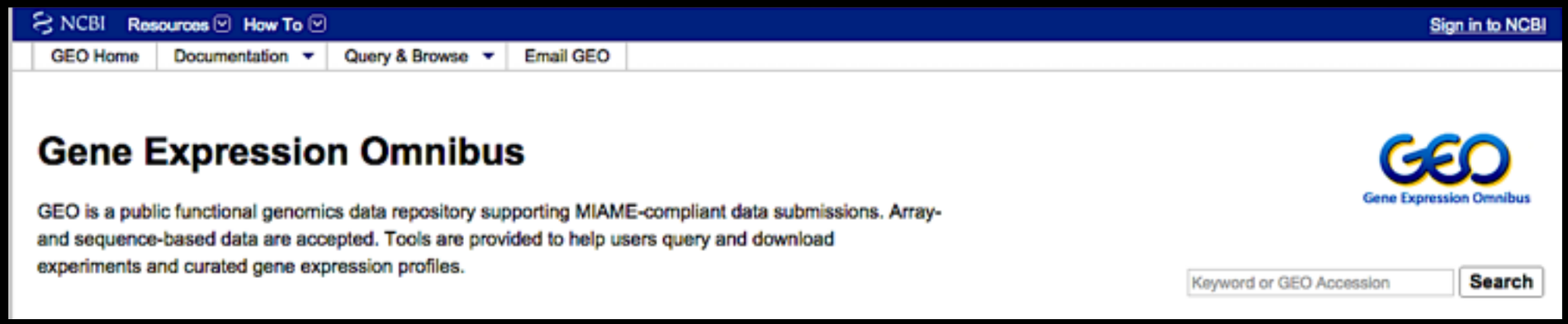
xxx.gpr

	A	B	C	D	E	F	G	H	I
1	ATF	1							
2	29	48							
3	Type=GenePix Results 3								
4	DateTime=2006/06/15 09:19:02								
5	Settings=								
6	GalFile=C:\Axon\Params\chip6old.gal								
7	PixelSize=10								
8	Wavelengths=635 532								
9	ImageFiles=C:\Documents and Settings\rc\桌面\20060411_LiuHS\20060411_LiuHS_7-11 cell_IPTG_0hr.tif								
10	NormalizationMethod=None								
11	NormalizationFactors=1 1								
12	JpegImage=C:\Documents and Settings\rc\桌面\20060411_LiuHS\20060411_LiuHS_7-11 cell_IPTG_0hr.jpg								
13	StdDev=Type 1								
14	RatioFormulations=W1/W2 (635/532)								
15	Barcode=								
16	BackgroundSubtraction=LocalFeature								
17	ImageOrigin=5840, 4300								
18	JpegOrigin=6580, 5120								
19	Creator=GenePix Pro 4.1.1.42								
20	Scanner=GenePix 4000B [92487]								
21	FocusPosition=0								
22	Temperature=28.34								
23	LinesAveraged=1								
24	Comment=								
25	PMTGain=700 640								
26	ScanPower=100 100								
27	LaserPower=1.73 3.24								
28	LaserOnTime=8118 8169								
29	Filters=<Empty> <Empty>								
30	ScanRegion=584,430,1999,2027								
31	Supplier=								
32	Block	Column	Row	Name	ID	X	Y	Dia.	F635 Media
33	1	1	1	SIAT1	AI286209.1	6860	5490	150	356
34	1	2	1	RBL2	AI459215	7090	5520	150	195
35	1	3	1	PTGS2	R80322.1	7310	5510	150	424

Minimum **I**nformation **A**bout a **M**icroarray **E**xperiment - **MIAME**

MIAME describes the Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment **unambiguously and potentially to reproduce the experiment**. [Brazma et al., Nature Genetics, PMID: 11726920]

Many journals and funding agencies require microarray data to comply with MIAME. **GEO** deposit procedures enable and encourage submitters to supply MIAME compliant data. [Edgar R, Barrett T. Nature Biotechnology, PMID: 17160034]



The 6 most critical elements contributing towards MIAME are:

1. The raw data for each hybridization (e.g., CEL or **GPR** files)
2. The final processed (**normalized**) data for the set of hybridizations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential **sample annotation** including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The **experimental design** including sample data relationships (e.g., which raw data file relates to which sample, which hybridizations are technical, which are biological replicates)
5. **Sufficient annotation** of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and **data processing protocols** (e.g., what normalization method has been used to obtain the final processed data)



500px.com

When you want to determine whether some program or treatment ***causes*** some outcome or outcomes to occur, then you are interested in having strong internal validity. Essentially, you want to assess the proposition:

If X, then Y

or, in more colloquial terms:

If the program is given, then the outcome occurs

Unfortunately, ***it's not enough*** just to show that when the program or treatment occurs the expected outcome also happens. That's because there may be lots of reasons, other than the program, for why you observed the outcome. To really show that there is a causal relationship, you have to simultaneously address the two propositions:

If X, then Y

and

If *not* X, then *not* Y

Or, once again more colloquially:

If the program is given, then the outcome occurs

and

If the program is *not* given, then the outcome does *not* occur

$$p \longrightarrow q$$

$$\sim p \longrightarrow \sim q$$

how **MANY** replicates for microarray is necessary ?

It depends on

- ☑ the quality of the sample
- ☑ the magnitude of the expected effect
- ☑ the method of analysis
- ☑ the experimental design

- smaller samples (single cells) are more noisy than large samples (tissue homogenates)
- some problematic cases
- cell cultures are less noisy than patient biopsies
- sample pooling can decrease noise – if individual variation is not of interest
- if you knock-out organisms (even the same tissue can be a different)
- Microarrays are very sensitive
- to keep costs small:
- study transfected cell lines (are they still the same cells?)
- knock-out organisms (even the same tissue can be different)
- local changes may be diluted of cell isolation will increase noise
- dogs and donuts
- experiment failed!

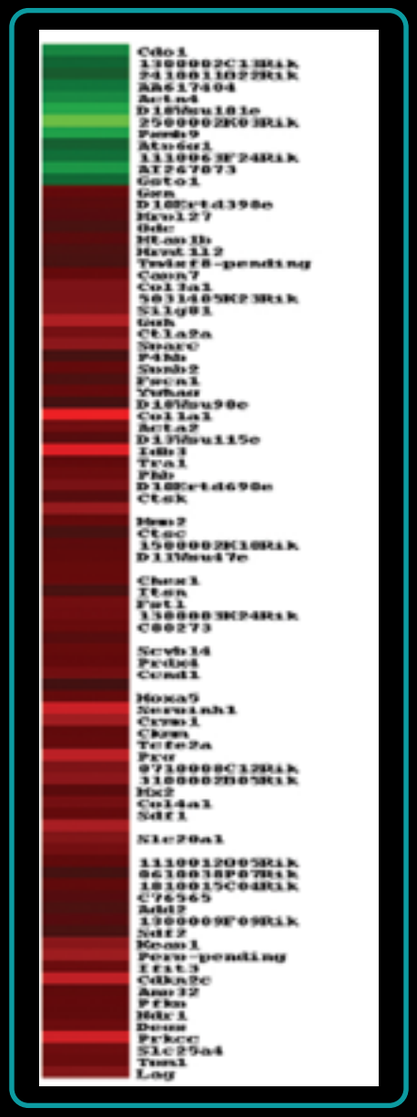
Experiment design for microarray

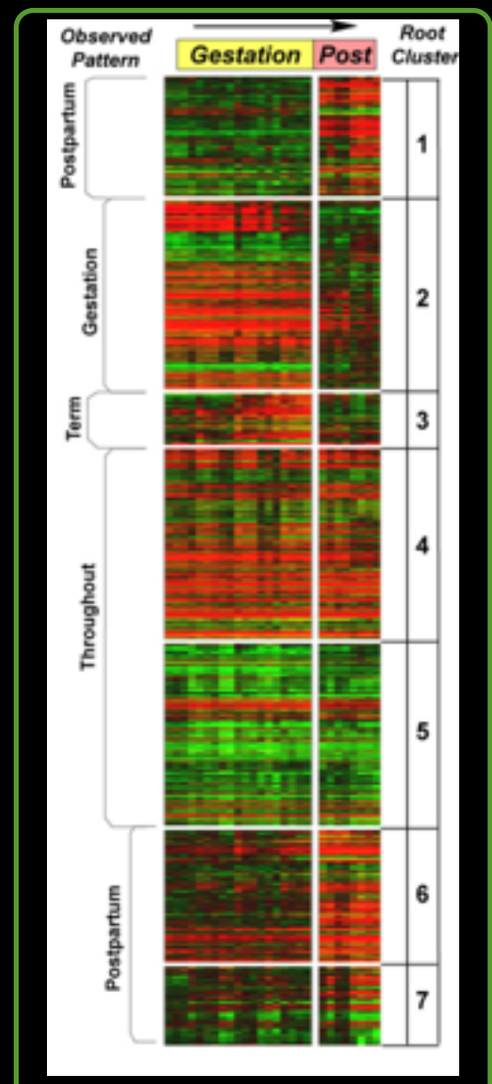
Type I: (n = 2) Differential Expression

- How is this gene expressed in target 1 as compared to target 2?
- Which genes show ***up/down regulation*** between the two targets?

Type II: (n > 2) Data Partitioning

- How does the expression of gene A ***vary over time, tissues, or treatments***?
- Do any of the expression profiles exhibit similar patterns of expression?

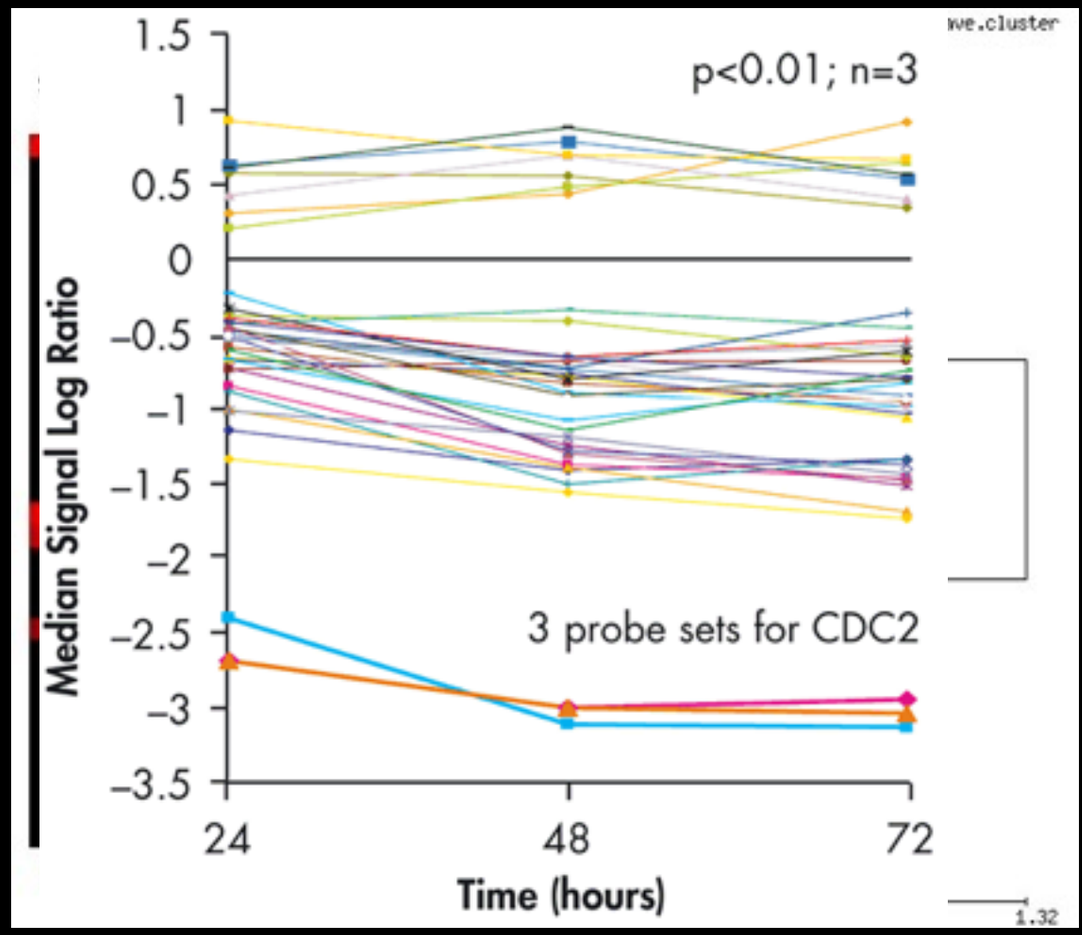




Reference and balanced block design

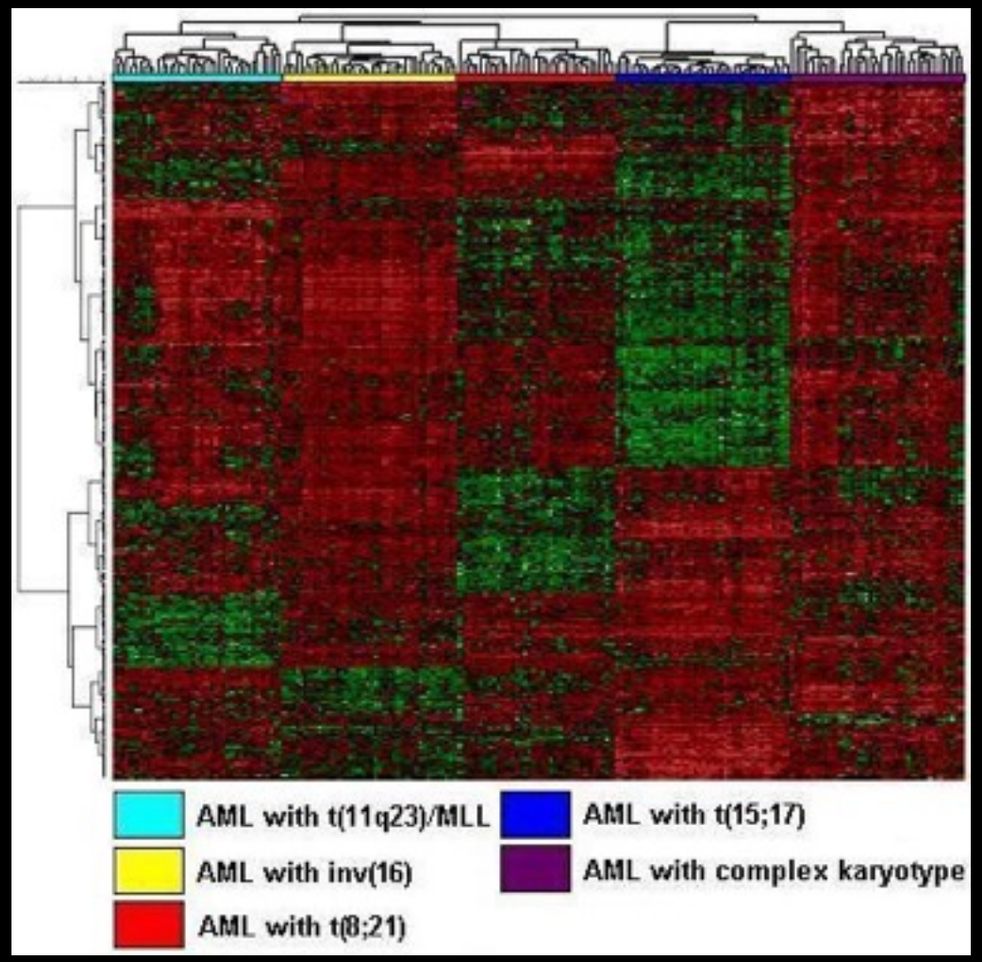
Reference design (flexible)
is far superior for more complex class comparison as well as for class prediction or discovery where ***clustering analysis is required.***

Balanced block design (efficient)
can only be used only for ***simple class comparison type*** of analysis (looking for genes that are **up or down** regulated between two samples).



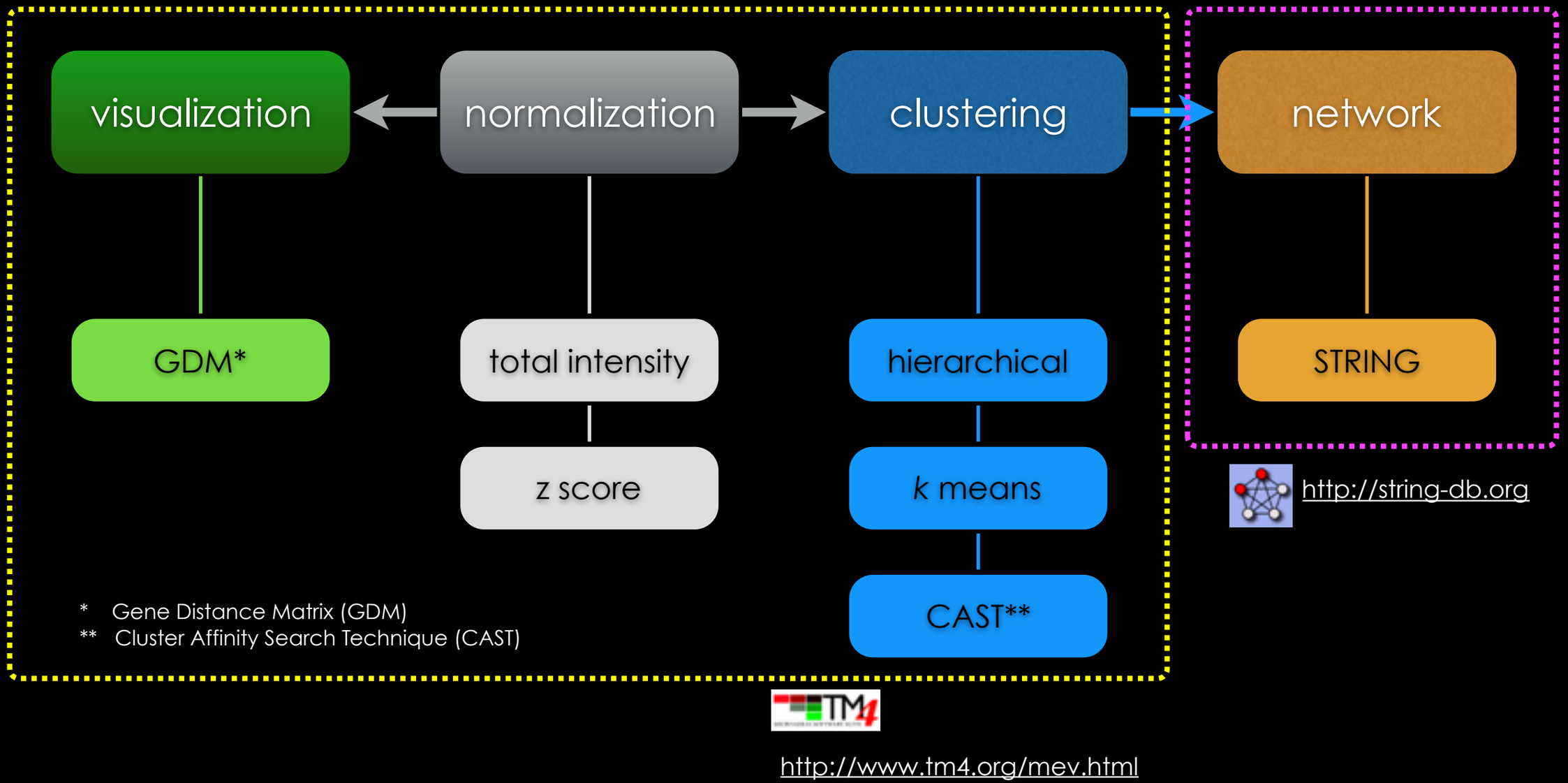
Loop design (for microarray center)

Reference RNA for reference design experiments used does ***not necessarily*** need to be physiologically relevant to your RNA. However, reference design should have at least 80% coverage for the particular slide (i.e. should contain **mRNA representing at least 80% of genes on the slide**). RNA isolated from a cells or tissues usually has 50-60% coverage. Reference RNA is, most often, a pool of RNA from several different cell lines or tissues. Human, mouse, and rat universal reference **RNAs are available from several vendors**.




data process and analysis
500px.com

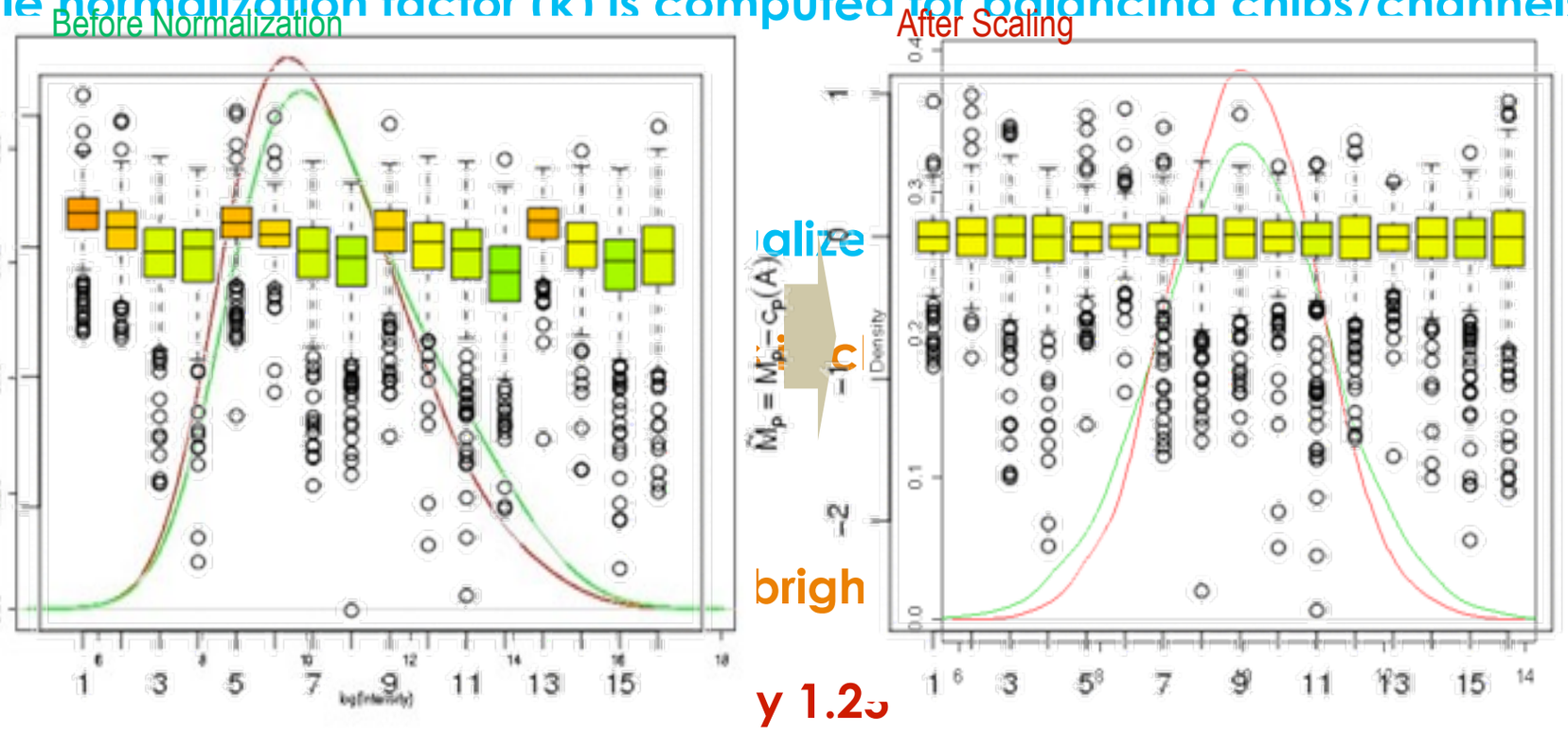
Laboratory of microarray analysis



Normalization (1): Total Intensity Normalization

Total intensity normalization data relies on the assumption that ***the quantity of initial mRNA is the same for both labelled samples***. Furthermore, one assumes that some genes are up-regulated in the query sample relative to the control and that others are down-regulated.

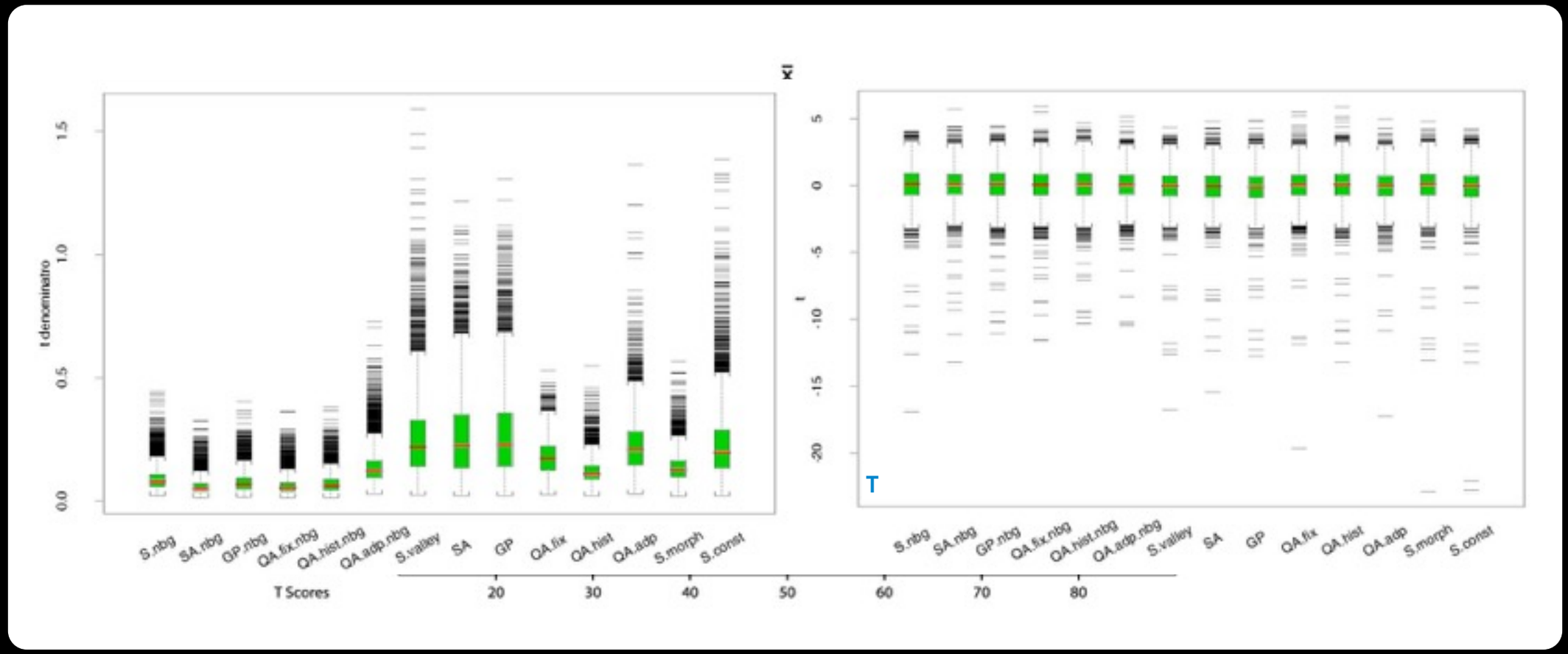
- **A single normalization factor (k) is computed for balancing chips/channels:**



- **Mul alize**
  - **2**
  - **brigh**
  - **1 y 1.2**

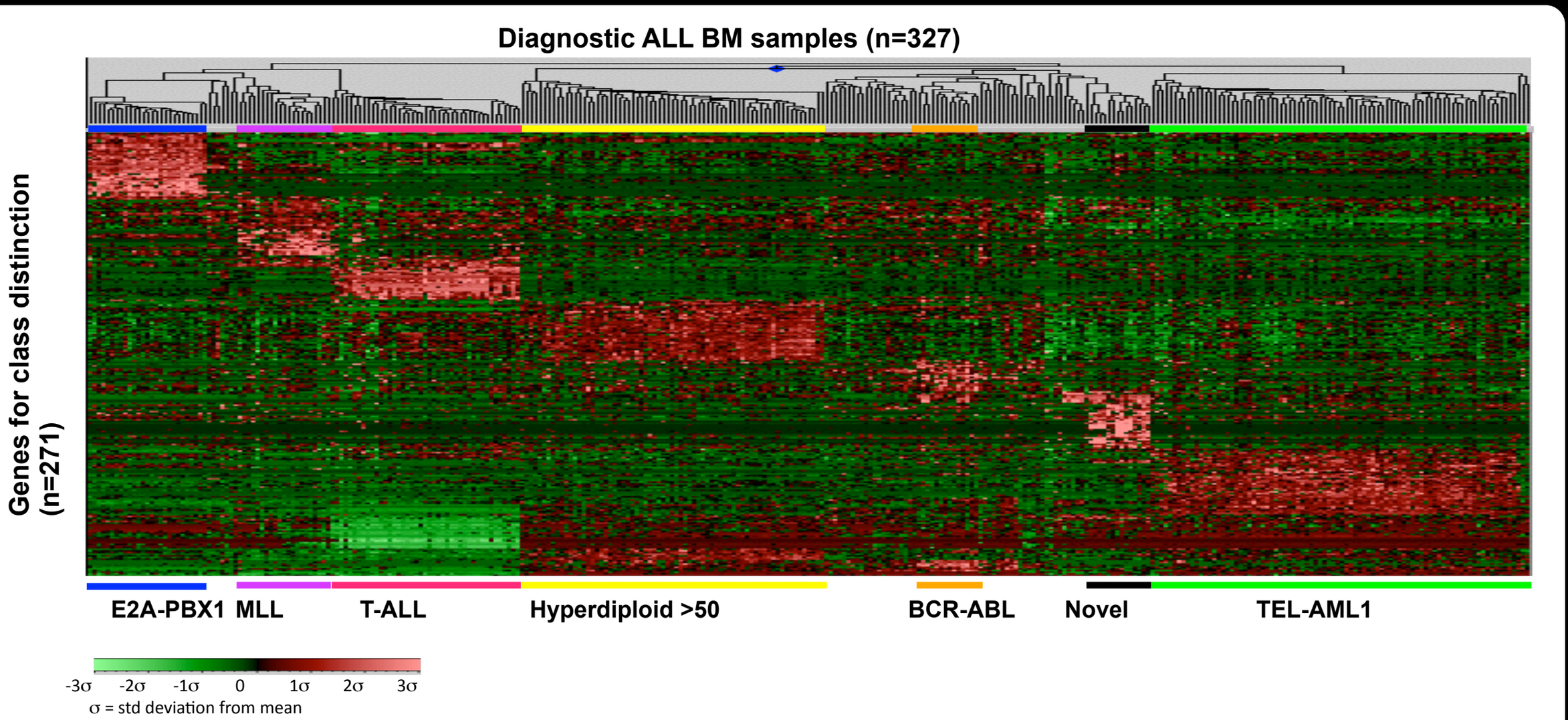
Normalization (2): Standard Score

Standard scores are also called z-values, **z-scores**, normal scores, and standardized variables; the use of "Z" is because the normal distribution is also known as the "Z distribution". They are most frequently used to ***compare a sample to a standard normal deviate***, though they can be defined without assumptions of normality.



Clustering (1): Hierarchical Clustering

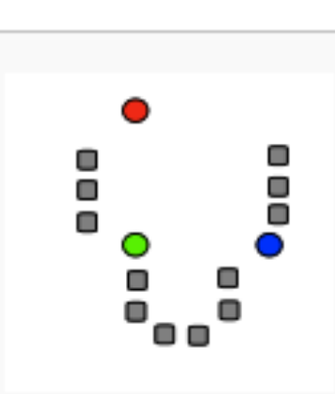
Hierarchical Clustering is an agglomerative clustering method which ***joins similar genes into groups***. The iterative process continues with the joining of ***resulting groups based on their similarity*** until all groups are connected in a hierarchical tree.



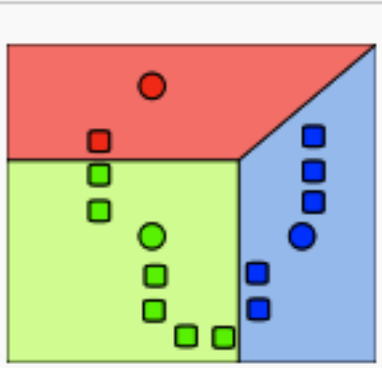
Clustering (2): *k*-means clustering

k*-means clustering** aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the ***nearest mean, serving as a prototype of the cluster.

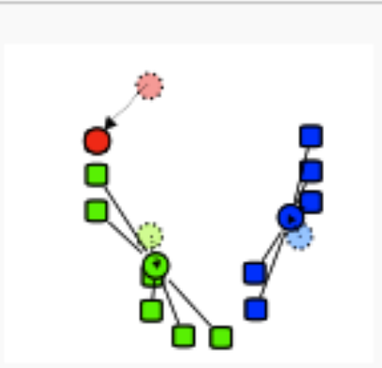
Demonstration of the standard algorithm



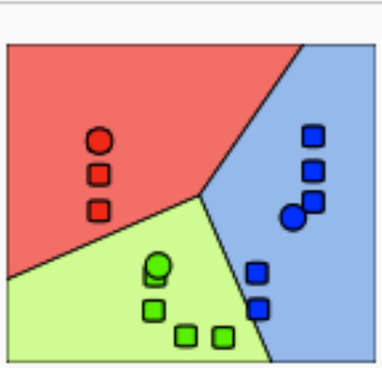
1) *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



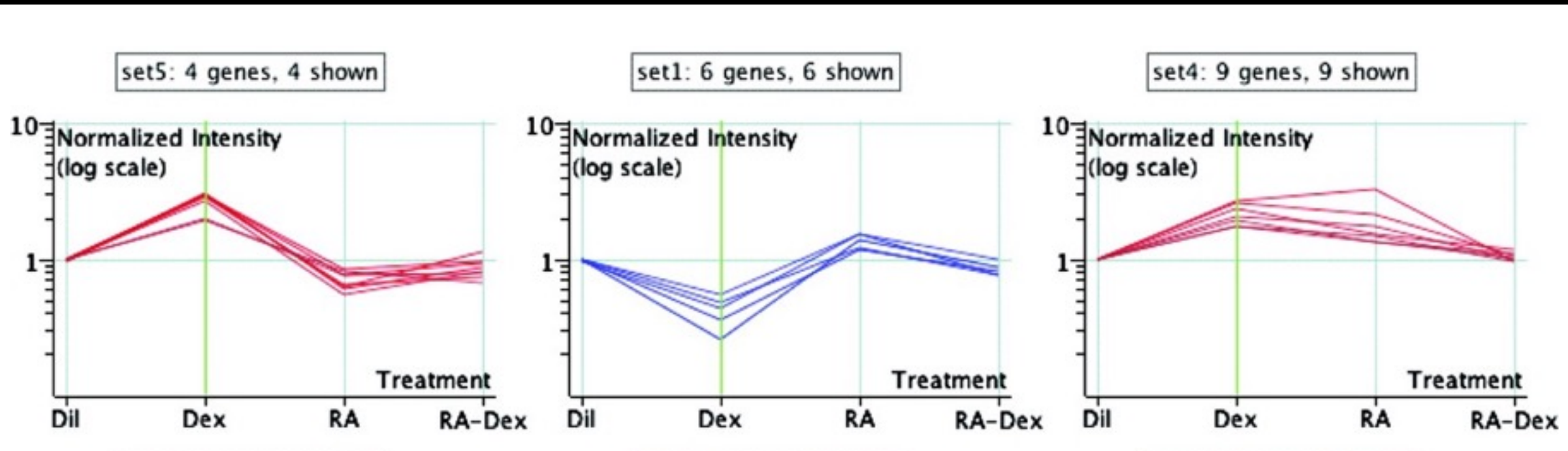
3) The centroid of each of the *k* clusters becomes the new mean.



4) Steps 2 and 3 are repeated until convergence has been reached.

Clustering (3): Cluster Affinity Search Technique

Cluster Affinity Search Technique uses an iterative approach to segregate elements with 'high affinity' into a cluster. The process iterates through two phases, (1) ***addition of high affinity elements to the cluster*** being created and (2) ***removal or clean-up of low affinity elements from the cluster*** being created.



Network: Protein-protein interactions

Protein–protein interactions refer to intentional physical contacts established between two or more proteins as a result of ***biochemical events and/or electrostatic forces***.





Why do we need protein interaction information

- **Learning protein functions:**

  If two proteins interact, there is a very high possibility that their ***functions*** are related as well
- Cellular operations are largely endured by interactions among proteins
- From protein pathways to understanding cells, tissues, … to life and evolution

Interaction → Networks → Pathways

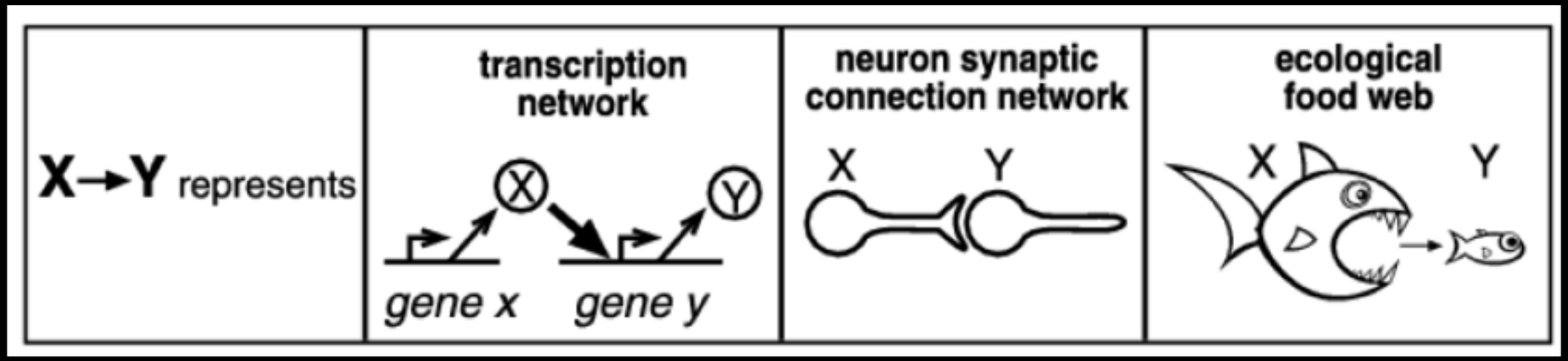
- A network is the ***backbone*** of a complex system.
- A group or system of (electric) components and *connecting circuitry* designed to function in a specific manner.
- Such motifs are found in networks from:

Engineering : Electronic circuits, World Wide Web *et al.*,

Biochemistry : Transcriptional regulation networks

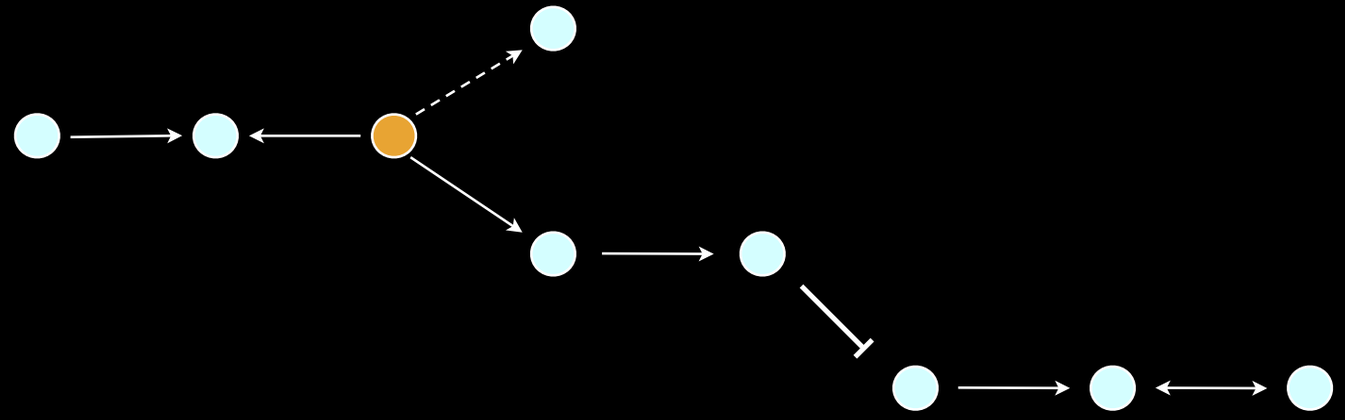
Neurobiology : Neuron connectivity

Ecology : Food webs



Interaction ⟶ Networks ⟶ Pathways

- It can be defined a modular unit of interacting molecules to fulfill a cellular function.
- It is usually represented by a ***2-D diagram with characteristic symbols*** linking the protein and non-protein entities.



A circle indicates a protein or a non-protein biomolecule.
An symbol in between indicates the nature of molecule-molecule interaction.

Gene annotation

Level 1: component, such as nuclear acid, amino acid, small molecule[1] …….

Level 2: component interactions, such as STRING[2] ……

Level 3: genome spatial orientation, such as Chromosome map[3] ……

Level 4: evolutionary change: such as PHYLIP[4] ……

1: NCBI Gene: http://www.ncbi.nlm.nih.gov/gene/
2: STRING: http://string-db.org
3: Ensembl: http://asia.ensembl.org/Homo_sapiens/Location/Genome
4: PHYLIP: http://evolution.genetics.washington.edu/phylip.html



500px.com

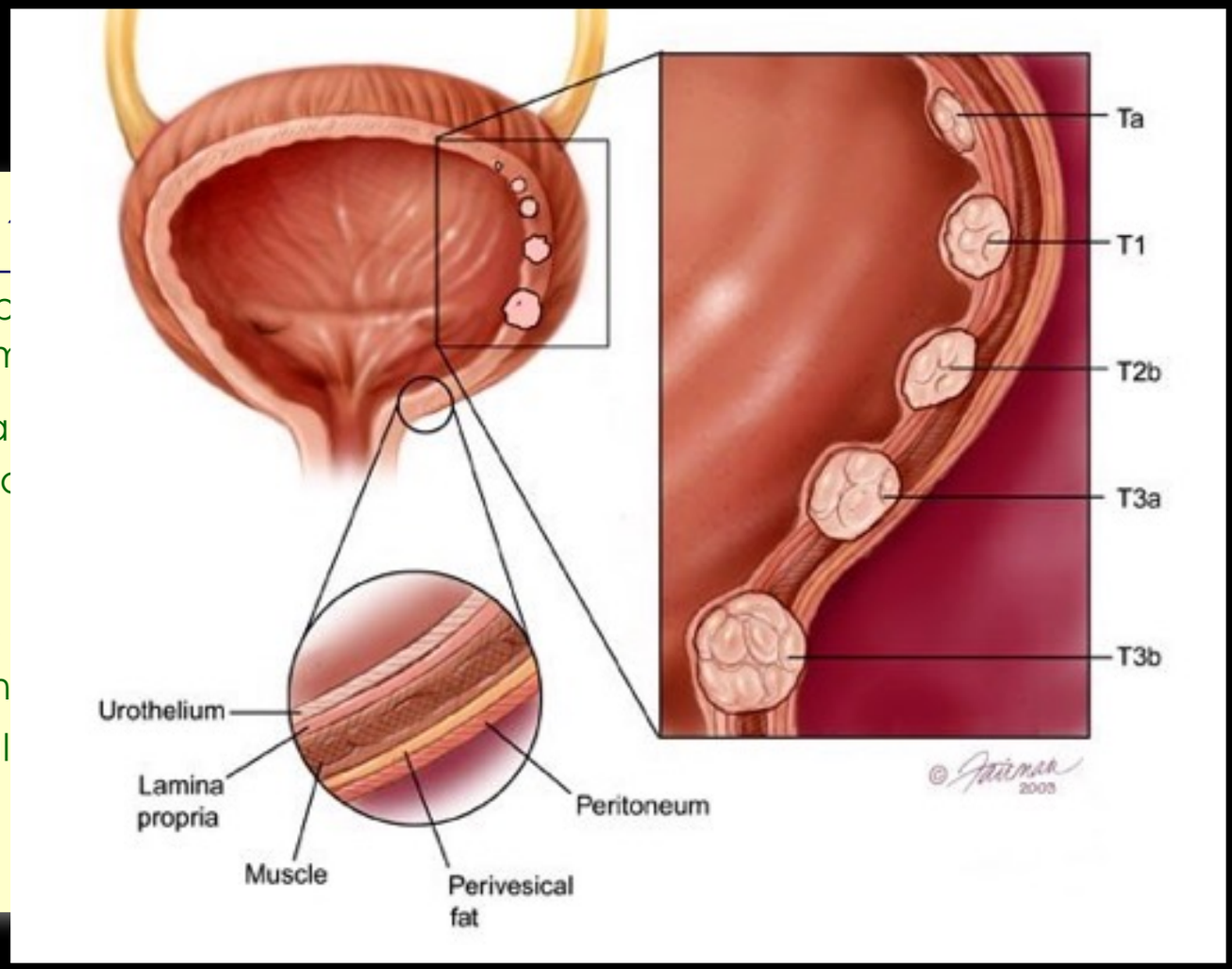
Towards the Systems Biology of Bladder Carcinoma

Jyh-wei Shin[1], Petrus Tang[2],
Jung-hsien Chiang[3], Shin-mu Vincent Tseng[3],
Yat-ming Yung[4], Phei-lang Chang[5]

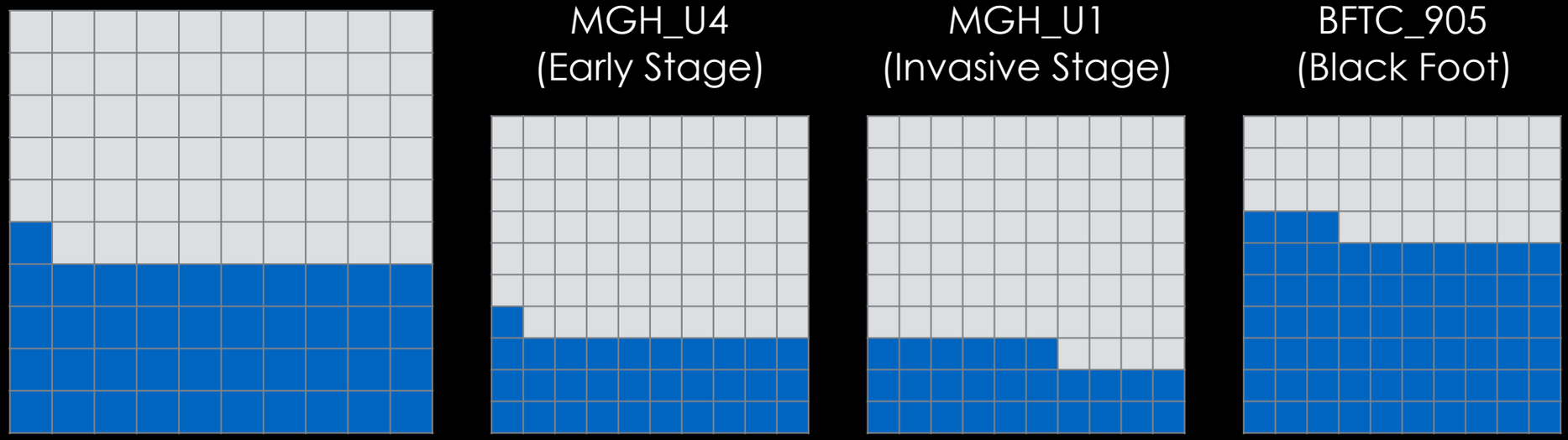
[1]Department of Parasitology, [3]Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan, R.O.C.
[2]Department of Parasitology, [4]Department of Pharmacology, [5]Department of Medicine, College of Medicine, Chang Gung University, Taiwan, R.O.C.

Bladder Carcinoma Cell Lines

	MGH_U4 (Early Stage)	MGH_U
	Human bladder carcinoma	Human b carcinom
Species	Homo sapiens	Homo sa
Strain/Ethnicity		Caucasi
Age/Stage		81 years
Gender		Female
	Bladder; transitional cell carcinoma	Bladder; transition
Morphology	Epithelial	Epithelial
Stage		B
Grade	I	III

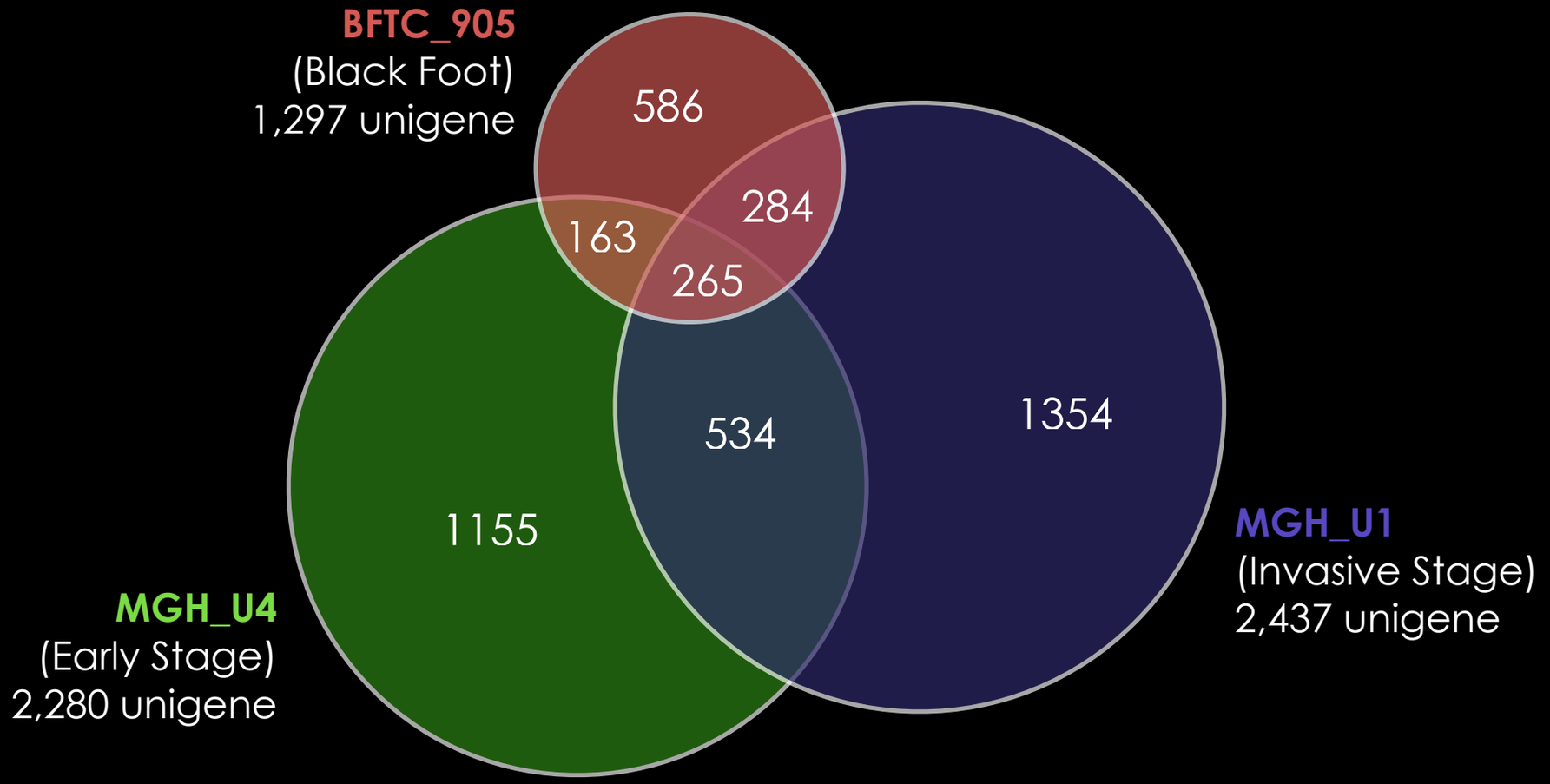


Clustering of 9,649 ESTs and the distribution of transcripts

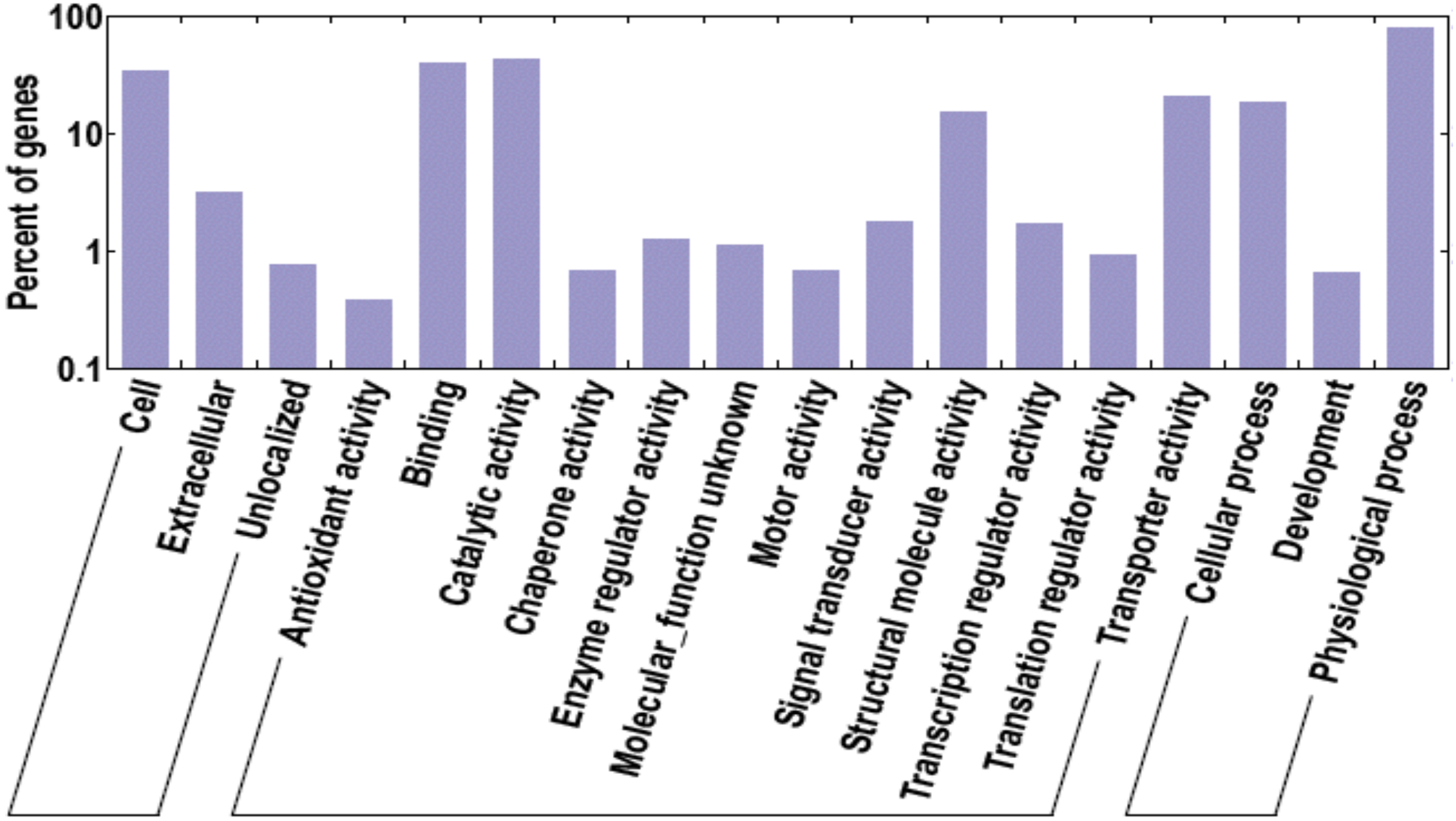


Clusters: 1,725	642	568	749
Singletons: 2,508	1,457	1,581	449
ESTs: 9,649	3,680	3,517	2,452

Comparative transcriptomics of 4,341 unigenes



Gene Ontology Classification



Cellular location

subcellular structures, locations, and macromolecular complexes; examples include nucleus, telomere, and origin recognition complex .

Molecular function

the tasks performed by individual gene products; examples are transcription factor and DNA helicase.

Biological process

broad biological goals, such as mitosis or purine metabolism, that are accomplished by ordered assemblies of molecular functions.

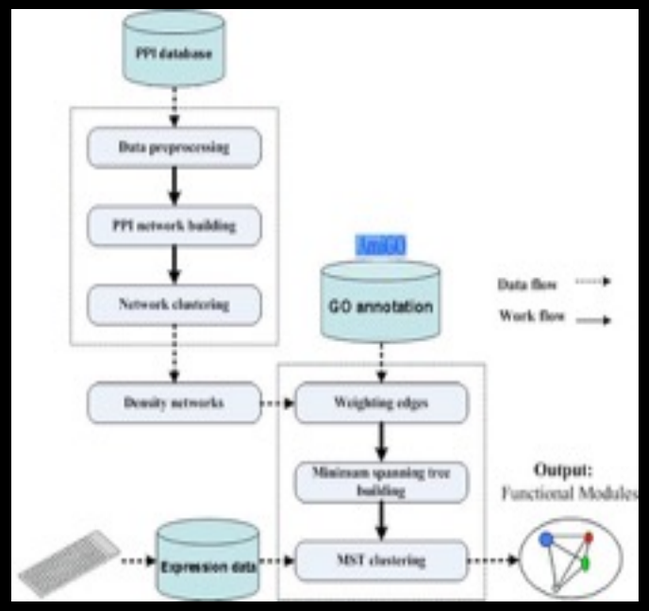
Protein Protein Interaction Functional Module Network

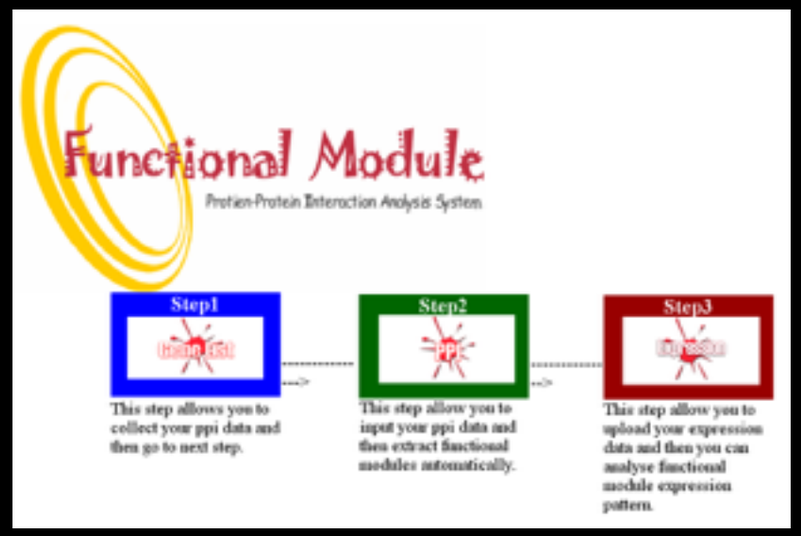
Unigenes

Scaffold

Subtraction

mapping





EST

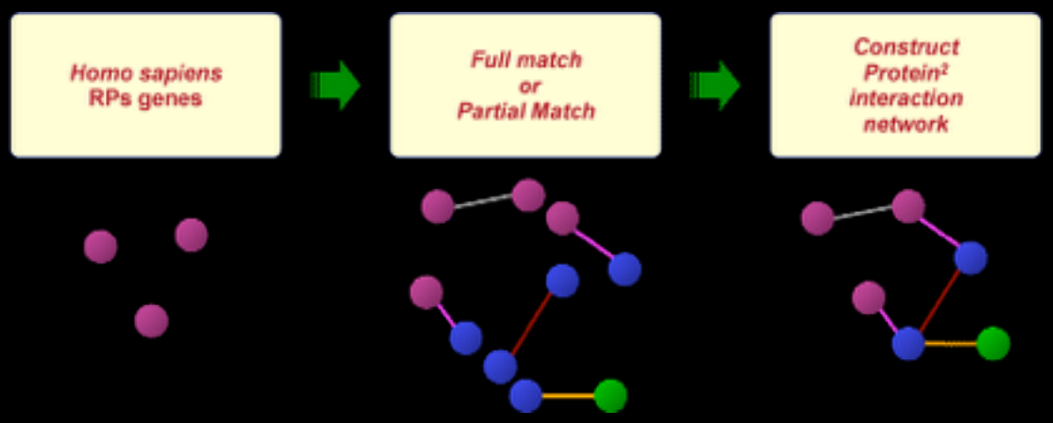
SAGE

Proteomics

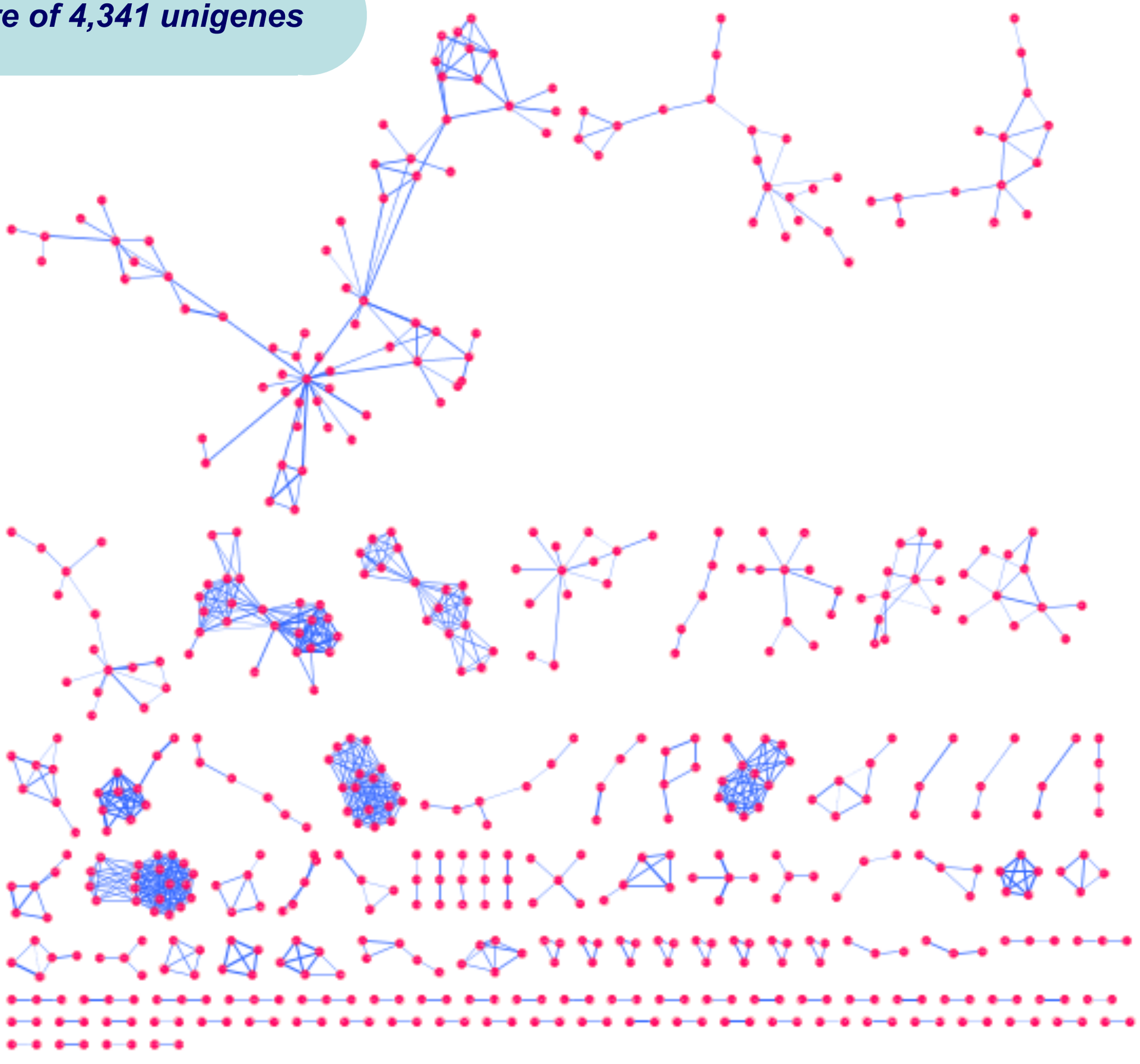
Microarray

Text mining

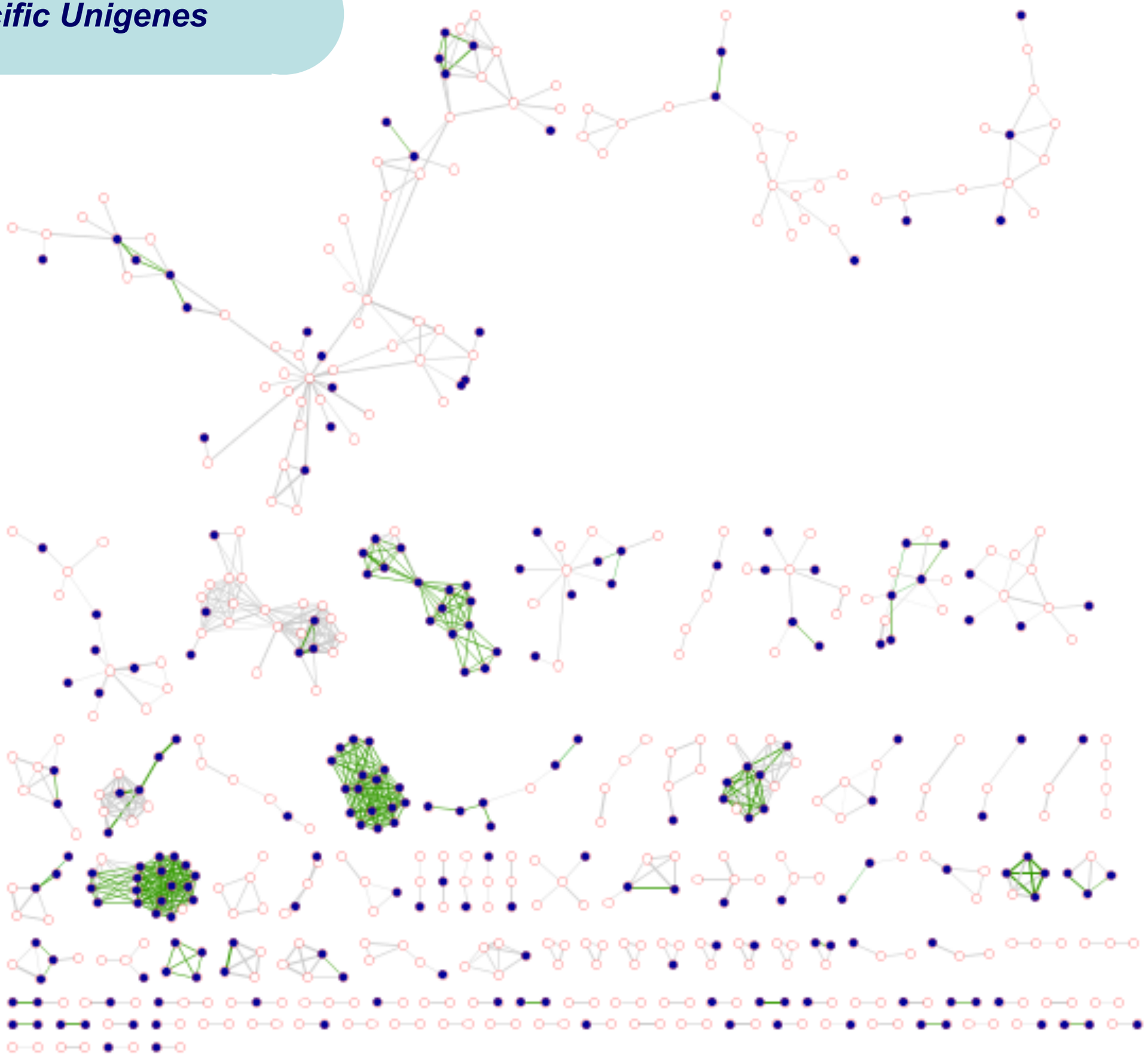
Functional Module



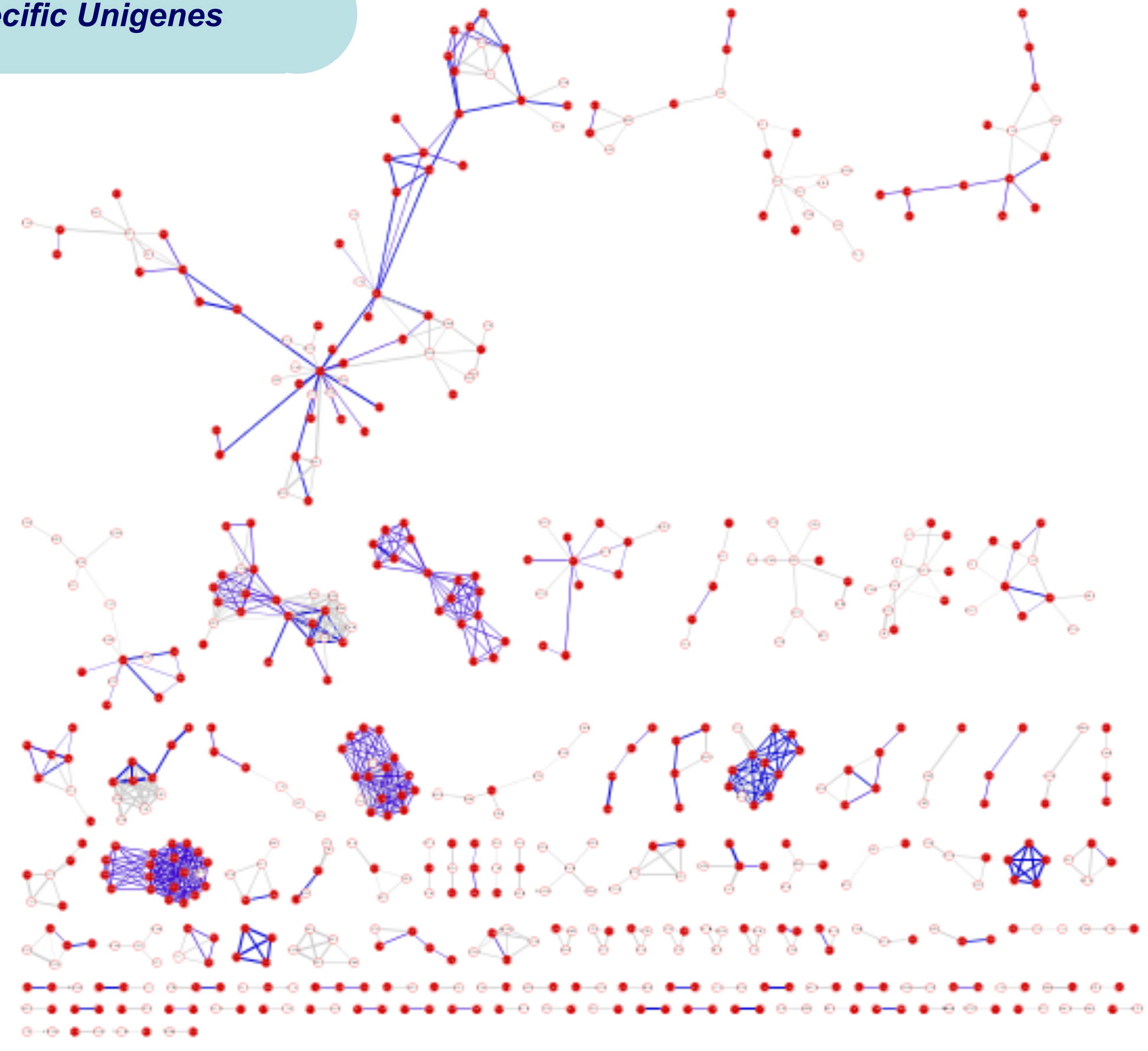
Scaffold Figure of 4,341 unigenes



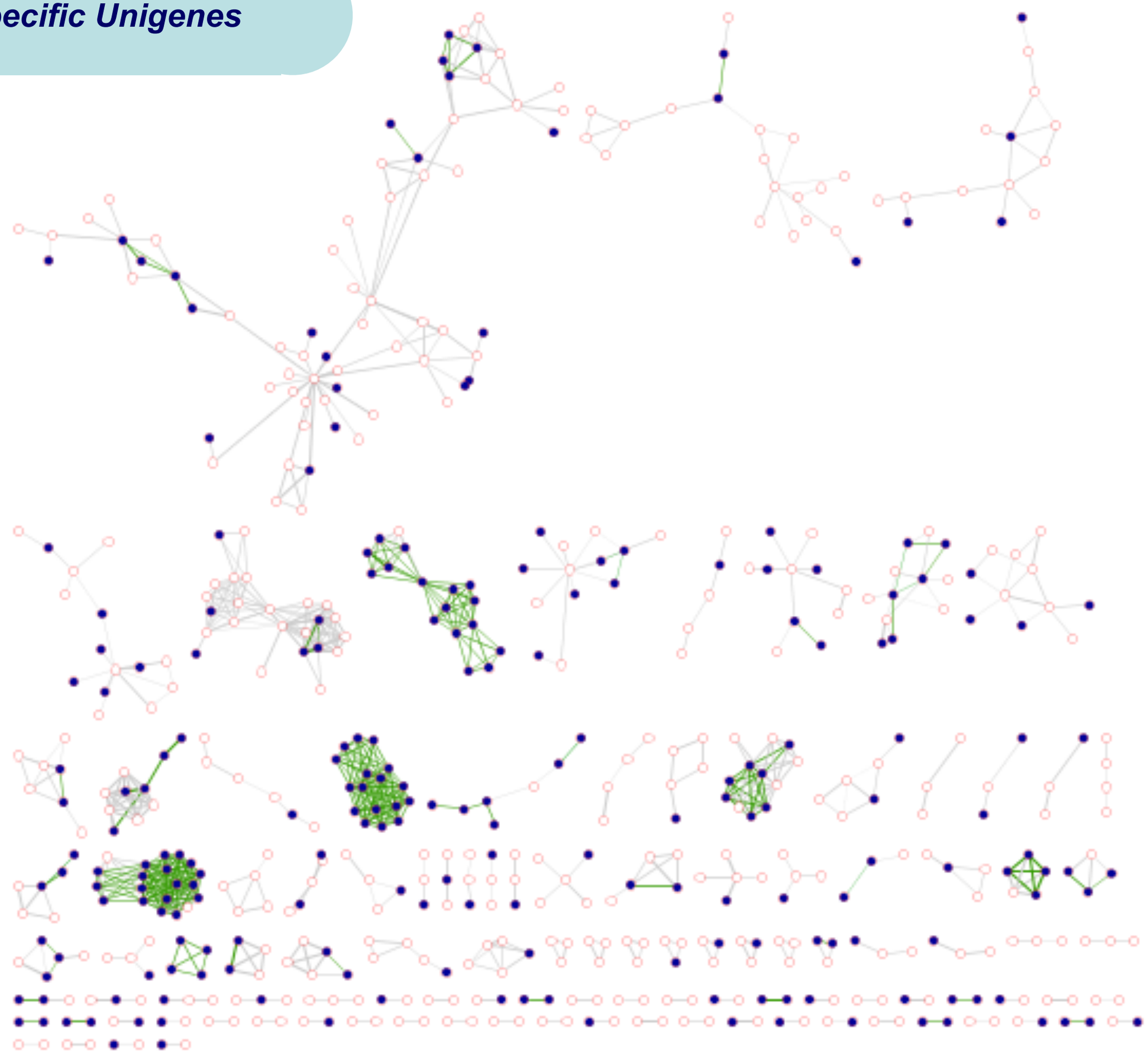
MGH_U4 Specific Unigenes



MGH_U1 Specific Unigenes



BFTC_905 Specific Unigenes



MGH_U4, MGH_U1 and BFTC_905 expression

Oncogene Candidate

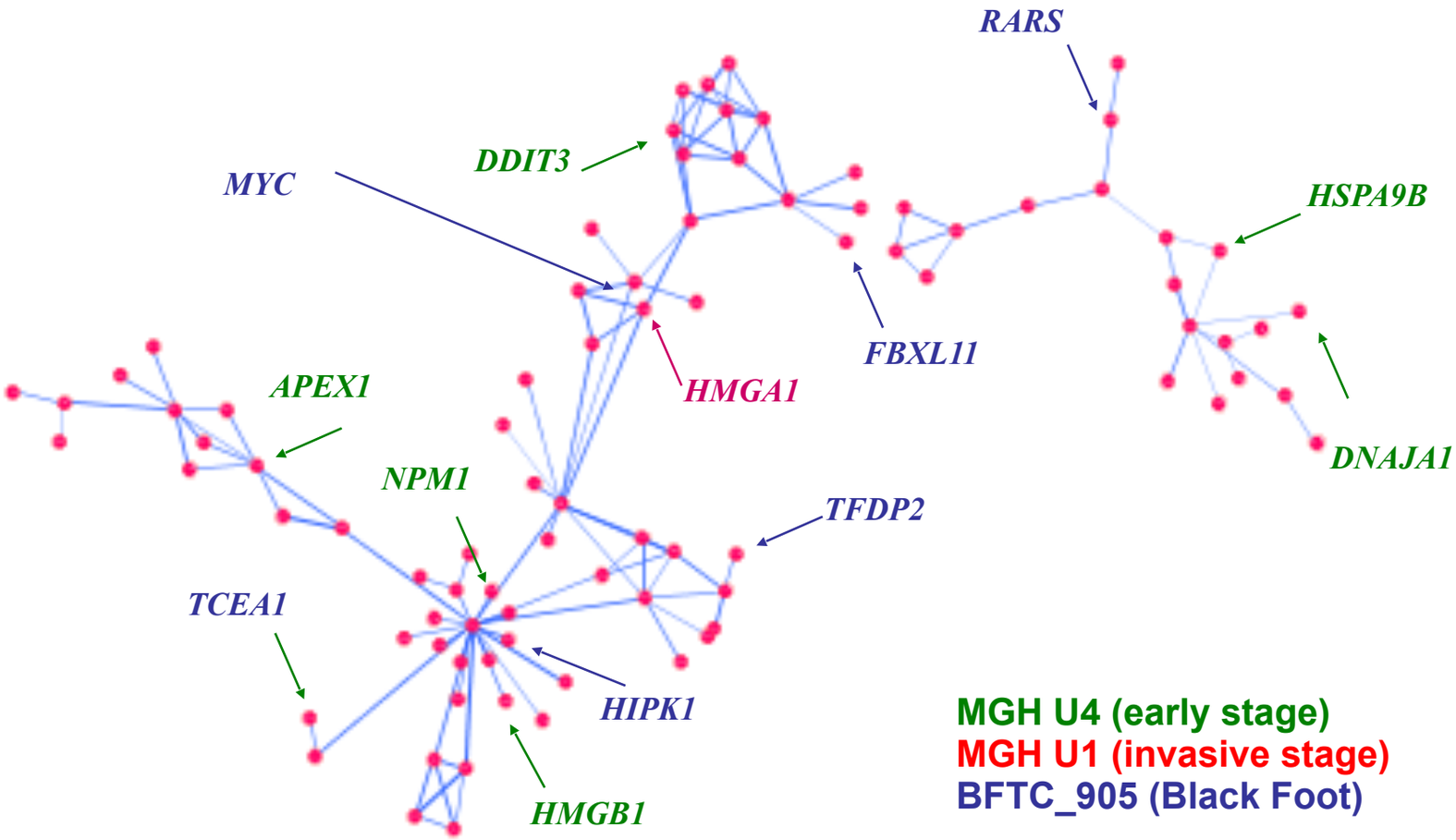
HMGA1

Tumor Suppressor Gene Candidate

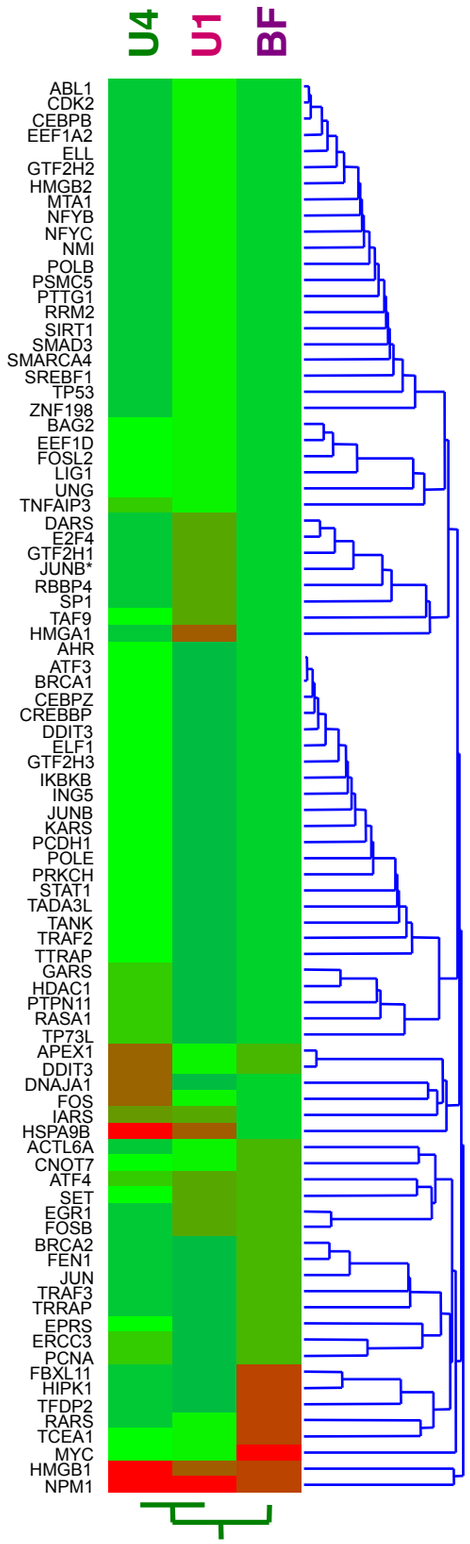
APEX1, DDIT3, DNAJA1, HMGB1, HSPA9B, NPM1

Black Foot Related Gene

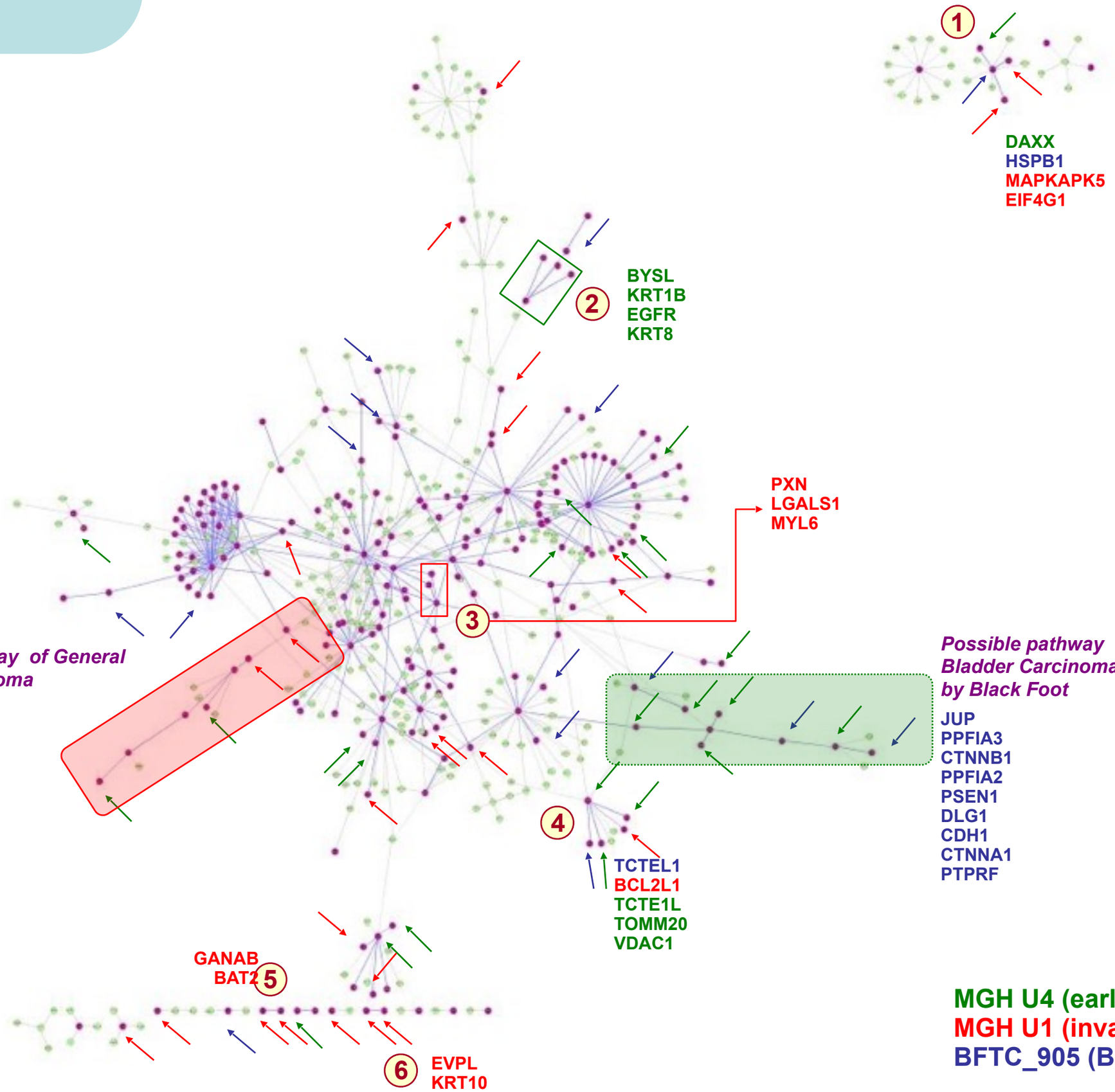
FBXL11, HIPK1, MYC, RARS, TFDP2



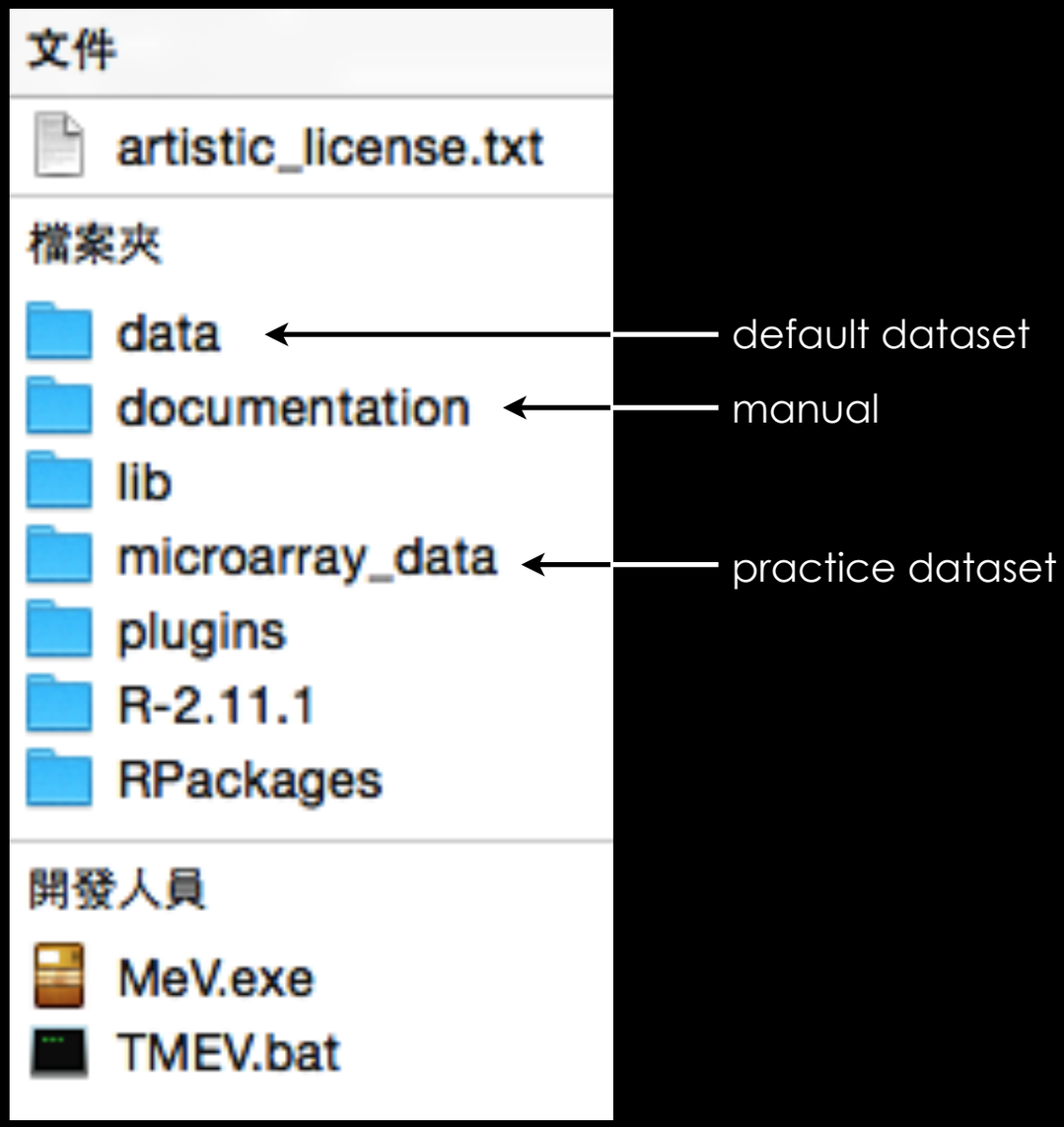
MGH U4 (early stage)
MGH U1 (invasive stage)
BFTC_905 (Black Foot)



ESTs mapping



Practice and Take home message





Jerry Jyh-wei Shin

hippotntw@gmail.com
hippotntw@163.com

212258307