Gene expression and regulation

Subha Narayan Rath

Proteomics and Transcriptomics

- Proteomics is the study of the distribution and interactions of proteins in time and space in a cell or organism
- Transcriptomics is the study of all RNA molecules in a tissue
- High throughput methods of data analysis: microarray analysis and mass spectrometry> gives large-scale picture of proteins in living tissues
- They are active in controlling: transcription and translation
- Goal of systems biology is the synthesis of genomic, transcriptomics, proteomics and other data into an integrated picture of structure, dynamics, and logic of living tissues
- 2 techniques show the distribution of RNA showing indirectly proteins in cells: Microarray technique and RNAseq (high throughput sequencing of RNAs in a sample)
- Mass spectrometry for proteomics: Not discussed!!

Different tests to detect gene of interest

Type of tests	Probe (synthesized and immobilized material in the chip)	Target (the sample which is extracted, labelled & then tested)	Comments
One-to-one test	Oligo with a complementary sequence	One oligonucleotide with a known sequence	Hybridization
Many-to-one test	One probe with complementary sequence	To find the query oligo in a mixture, spread the mixture out and test each component of the mixture	Northern or southern blot
Many-to-many test	A set of oligos are synthesized one complementary to each sequence of query	To detect many oligos in a mixture, they are prepared with different colored fluorescent tags for different samples	Microarrays where DNA oligomers are affixed to known locations on a rigid support in a regular 2D array

DNA microarrays

- DNA microarrays analyze
  - 1. the mRNAs in a cell to reveal the expression patterns of proteins; or
  - 2. genomic DNA to reveal absent or mutated genes

- The following answers can be found by DNA microarrays
- A. Integrated characterization of cellular activity: what proteins are present at what amounts and where exactly?
- B. Measuring expressed genes help to identify which genes are causative for diseases
- DNA microarrays or DNA chips are devices for checking a sample simultaneously for the presence of many sequences

DNA microarrays

- Distributed on small wafer of glass or nylon typically 2cm square
- Spot size ~ 150 um in diameters
- DNA chip: 400 000 probe oligomers (larger than total genes)
- The data is scanned to get it computer readable forms
- Affymetrix and Illumina: 25-mer probes synthesized in situ vs. multiple copies of single 50-mer probes attached to a microbead
- 1. Expression chips: immobilized oligos are cDNA samples (20-80bp) from mRNAs of known genes; Target samples are mixture of mRNAs of normal or diseased tissue
- 2. Genomic hybridization: gains or losses of genes or changes in copy number. Target sequences fixed on chips are large pieces of genomic DNA, 500-5000bp long; probe mixture contain genomic DNA from normal or diseased states
- 3. Mutation microarray analysis: one looks for SNPs

Microarray data are quantitative but imprecise

- Precision is low and hence it is semi-quantitative
- mRNA levels detected by the array, do not reflect protein level
- Even yield in RT to make cDNA may be non-uniform
- MIAME: Minimum information about a Microarray Experiment: describe the contents and formats of the information to be recorded in the experiment
- European bioinformatics institute: array express: https://www.ebi.ac.uk/biostudies/arrayexpress/studies
- US NCBI hosts GENE EXPRESSION OMNIBUS: https://www.ncbi.nlm.nih.gov/geo/geo2r/
- Princeton university microarray database: https://puma.princeton.edu/
- Microarray database of plants:

Analysis of microarray data

- Steps are:
- 1. collection of samples and isolation of mRNA
- 2. Labelled Cdna
- 3. Hybridization
- 4. Scanning and analysis
- Color and intensity of the fluorescence reflect the extent of hybridization
- One gene may correspond to 30-40 spots and highly redundant

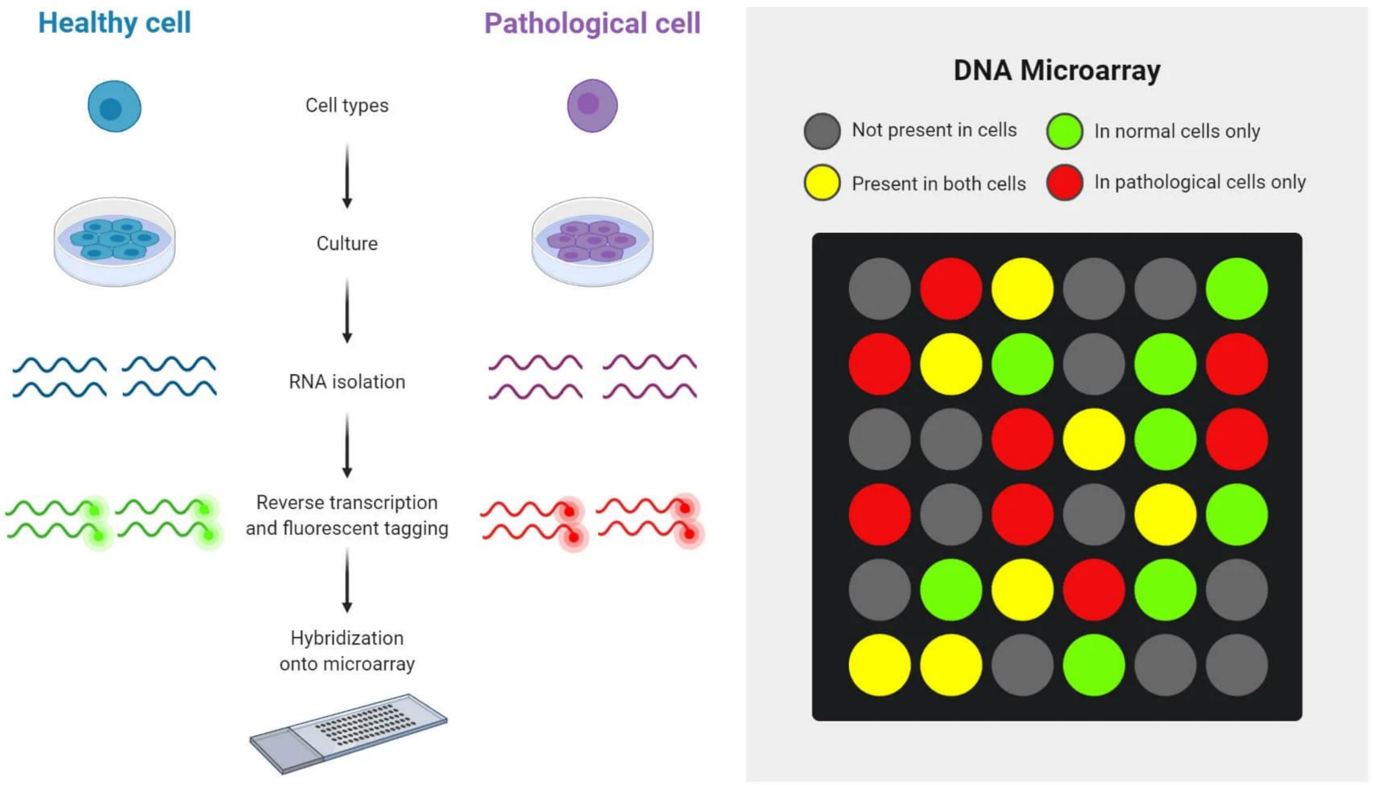


Image By Sagar Aryal, created using biorender.com

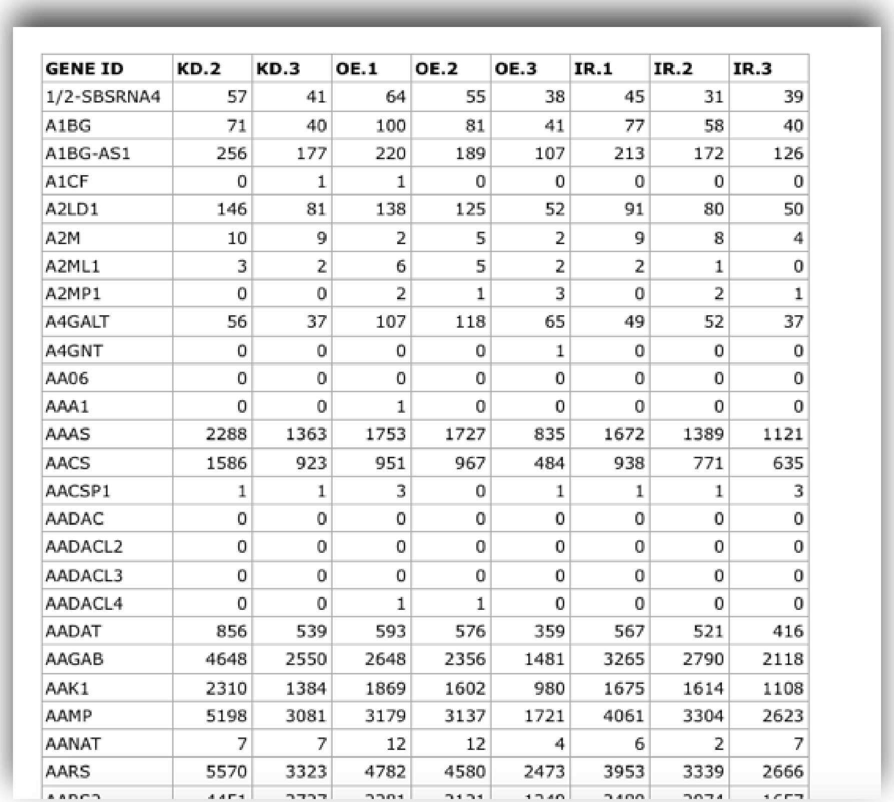
https://microbenotes.com/dna-microarray/

Data processing generated gene expression matrix table

- By image processing, checking internal controls, dealing with missing data, selecting reliable measurements, putting the results in consistent scales
- Change by 1.5-2 is considered significant in each row or column, considered as vector
- Two approaches for analysis:
- 1. comparisons focused on genes by comparing rows
- 2. comparison focused on different samples by comparing columns

Each column is a sample

Each row is a gene



GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/05_counting_reads.html

Case study: Mechanotransduction

- About GEO2R: https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html
- geo2R: https://www.ncbi.nlm.nih.gov/geo/geo2r/
- Volcano plot: to see differentially expressed plots by plotting statistically significant changes vs differentially expressed plots
- Mean difference plot: see differentially expressed plots