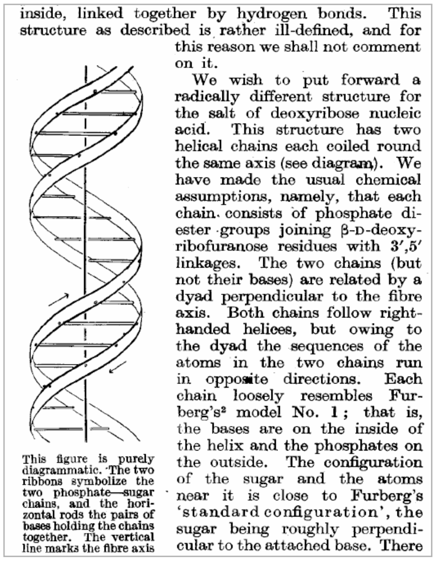
Functional genomics using DNA microarrays

CH370
Jan 31 2006

Vishy Iyer
Molecular Genetics and Microbiology
vishy@mail.utexas.edu



inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

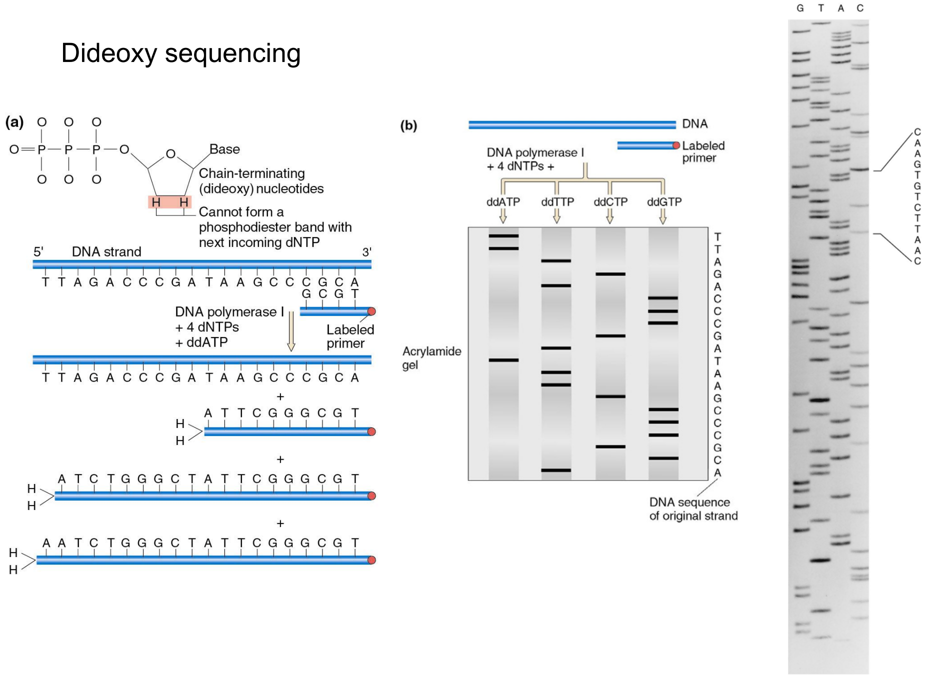
We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate—sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Nature – 1953

Nature – 2001

Dideoxy sequencing


(a)
O=P—P—P—O
Base
Chain-terminating (dideoxy) nucleotides
H H
Cannot form a phosphodiester band with next incoming dNTP
5' DNA strand 3'
T T A G A C C C G A T A A G C C C G C A
G C G T
DNA polymerase I + 4 dNTPs + ddATP
Labeled primer
T T A G A C C C G A T A A G C C C G C A
+
A T T C G G G C G T
+
A T C T G G G C T A T T C G G G C G T
+
A A T C T G G G C T A T T C G G G C G T
(b)
DNA
Labeled primer
DNA polymerase I + 4 dNTPs +
ddATP
ddTTP
ddCTP
ddGTP
Acrylamide gel
T T A G A C C C G A T A A G C C C G C A
DNA sequence of original strand
G T A C
C A A G T G T C T T A C

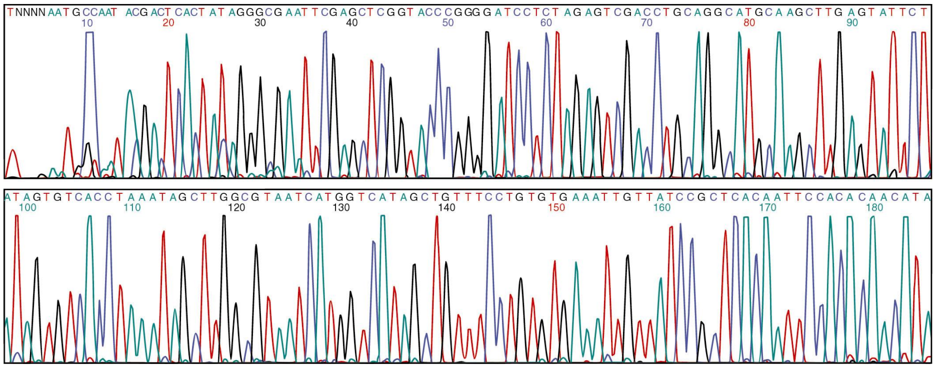
Automated dye-terminator sequencing

4-fluorescently labelled dideoxy dye terminators

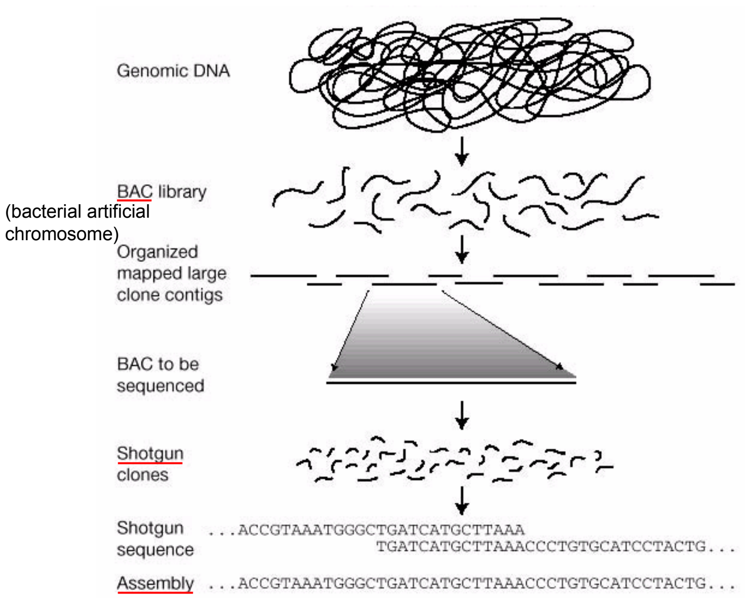
ddATP
ddGTP
ddCTP
ddTTP

} pool and load in a single well or capillary

- scan with laser + detector specific for each dye
- automated base calling
- very long reads (~ 1000 bases)/run



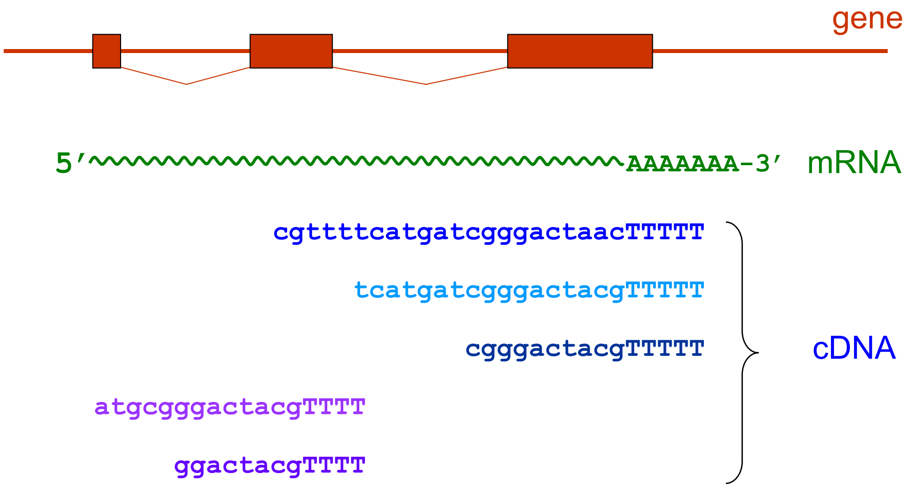
Physical mapping and sequencing of the human genome



```
Shotgun sequence  ...ACCGTAAATGGGCTGATCATGCTTAAA
                                       TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly          ...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```

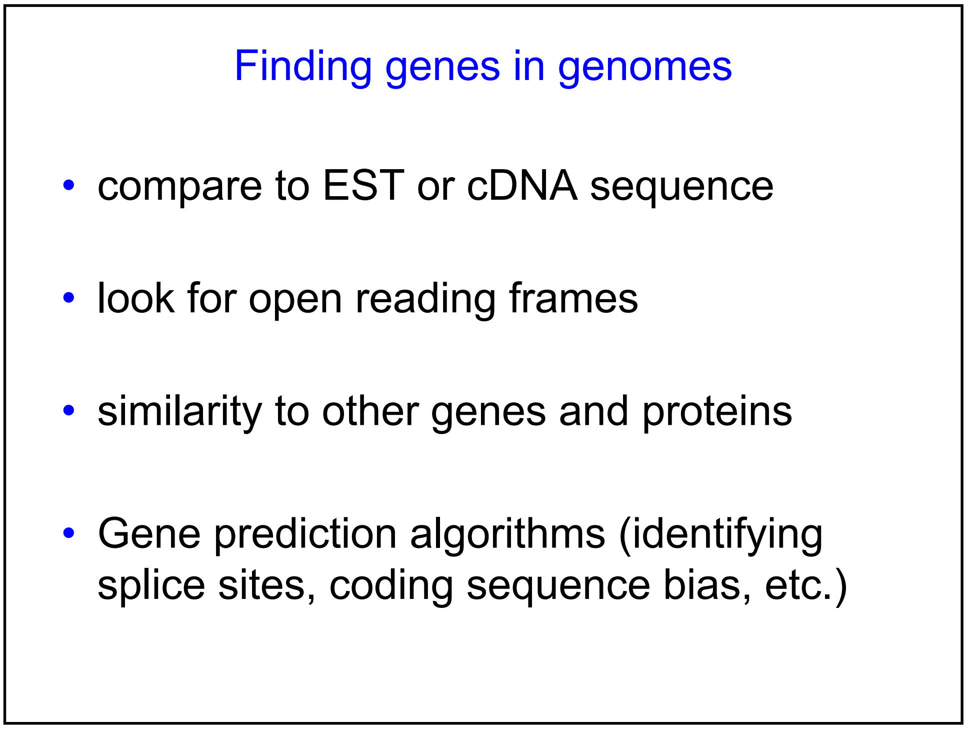
Nature (2001) **409** p. 860-921

Genes can also be identified by sequencing cDNAs at random. The sequenced cDNAs are called ESTs (expressed sequence tags)



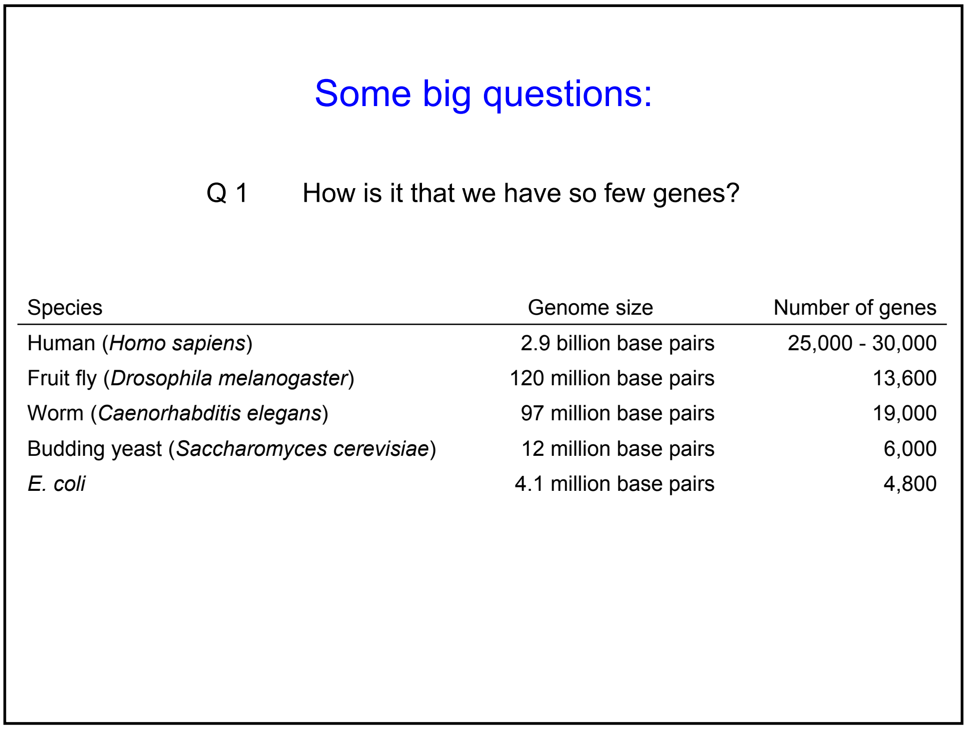
```
5′                                   AAAAAAA-3′

              cgttttcatgatcgggactaacTTTTT
                  tcatgatcgggactacgTTTTT
                       cgggactacgTTTTT
atgcgggactacgTTTT
    ggactacgTTTT
```



Finding genes in genomes

- compare to EST or cDNA sequence
- look for open reading frames
- similarity to other genes and proteins
- Gene prediction algorithms (identifying splice sites, coding sequence bias, etc.)



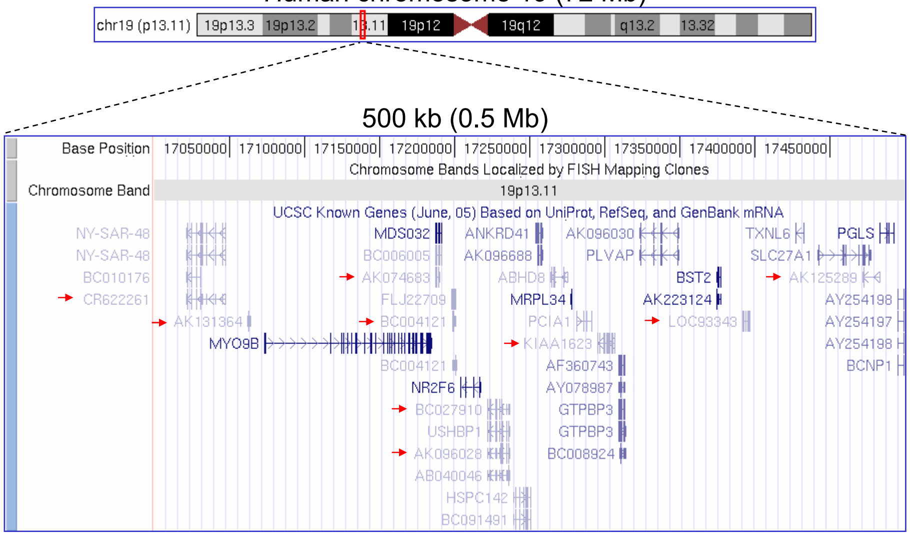
Some big questions:

Q 1 How is it that we have so few genes?

Species	Genome size	Number of genes
Human (*Homo sapiens*)	2.9 billion base pairs	25,000 - 30,000
Fruit fly (*Drosophila melanogaster*)	120 million base pairs	13,600
Worm (*Caenorhabditis elegans*)	97 million base pairs	19,000
Budding yeast (*Saccharomyces cerevisiae*)	12 million base pairs	6,000
E. coli	4.1 million base pairs	4,800

Q 2 What are the functions of all the unknown genes?

Human chromosome 19 (72 Mb)



500 kb (0.5 Mb)

Base Position 17050000 17100000 17150000 17200000 17250000 17300000 17350000 17400000 17450000

Chromosome Bands Localized by FISH Mapping Clones

Chromosome Band 19p13.11

UCSC Known Genes (June, 05) Based on UniProt, RefSeq, and GenBank mRNA

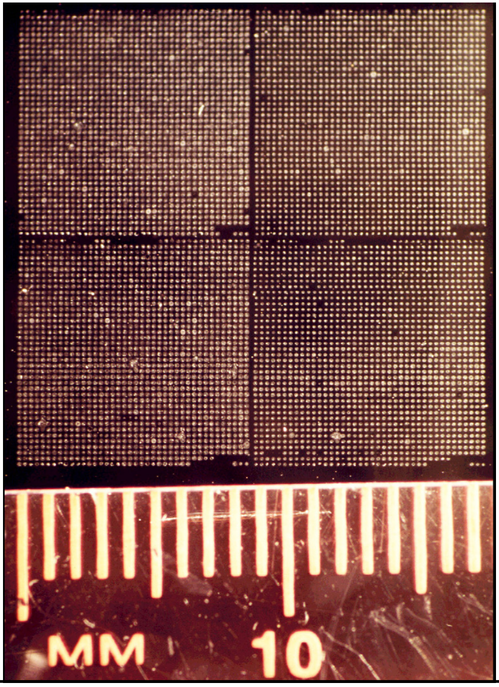
http://genome.ucsc.edu/

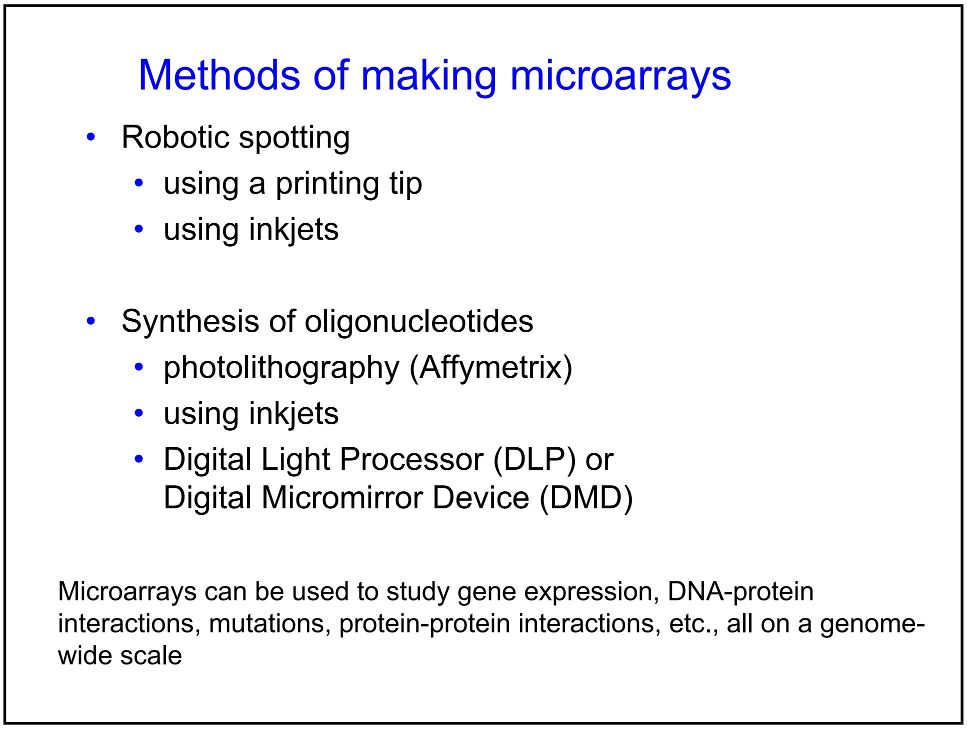
Drew Sheneman -*The Newark Star Ledger*

Functional genomics and proteomics

- Identify genes and proteins encoded in the genome (Gene finding)
- Measure gene expression on a genome-wide scale (microarrays)
- Identify protein function
  30-50% of the genes in a genome are of unknown function
- Identify protein interactions, biochemical pathways, gene interaction networks inside cells

DNA microarray (chip)



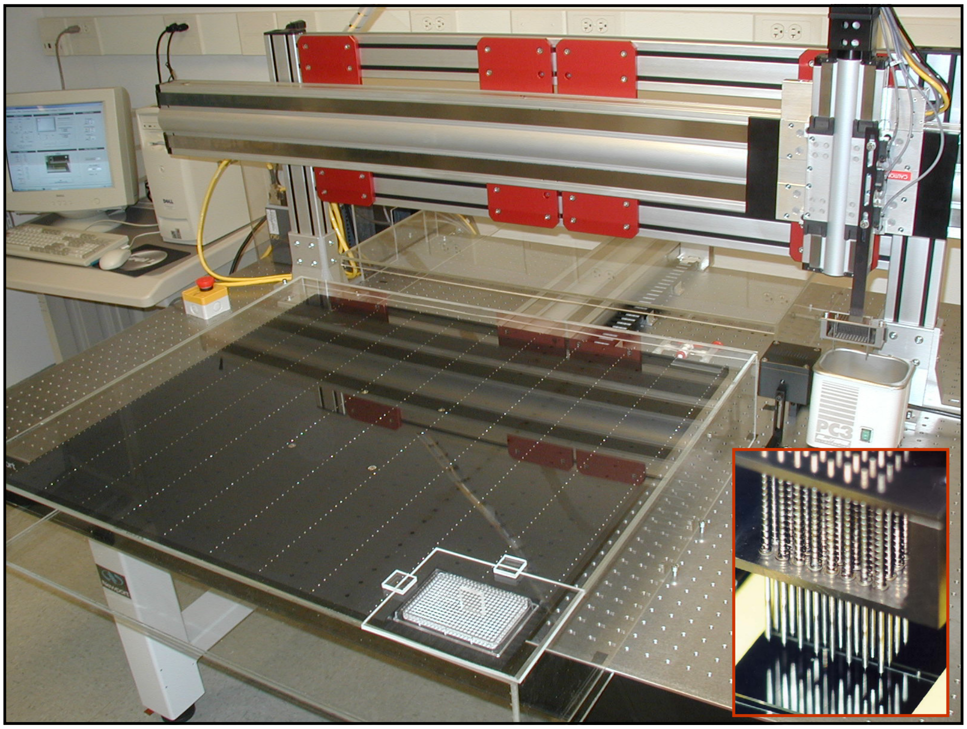


Methods of making microarrays

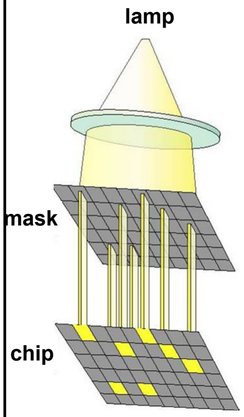
- Robotic spotting
  - using a printing tip
  - using inkjets

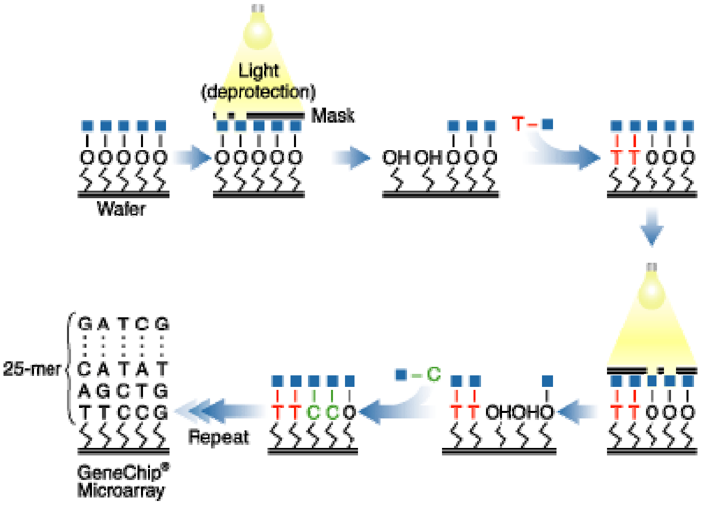
- Synthesis of oligonucleotides
  - photolithography (Affymetrix)
  - using inkjets
  - Digital Light Processor (DLP) or Digital Micromirror Device (DMD)

Microarrays can be used to study gene expression, DNA-protein interactions, mutations, protein-protein interactions, etc., all on a genome-wide scale



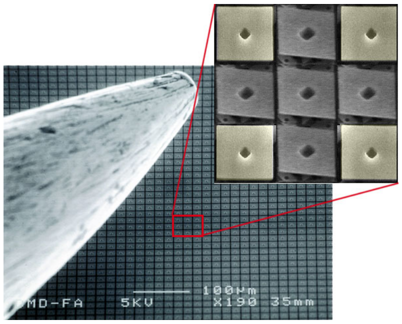
Affymetrix GeneChip

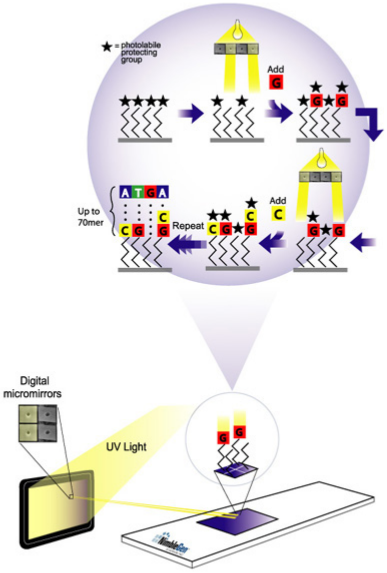




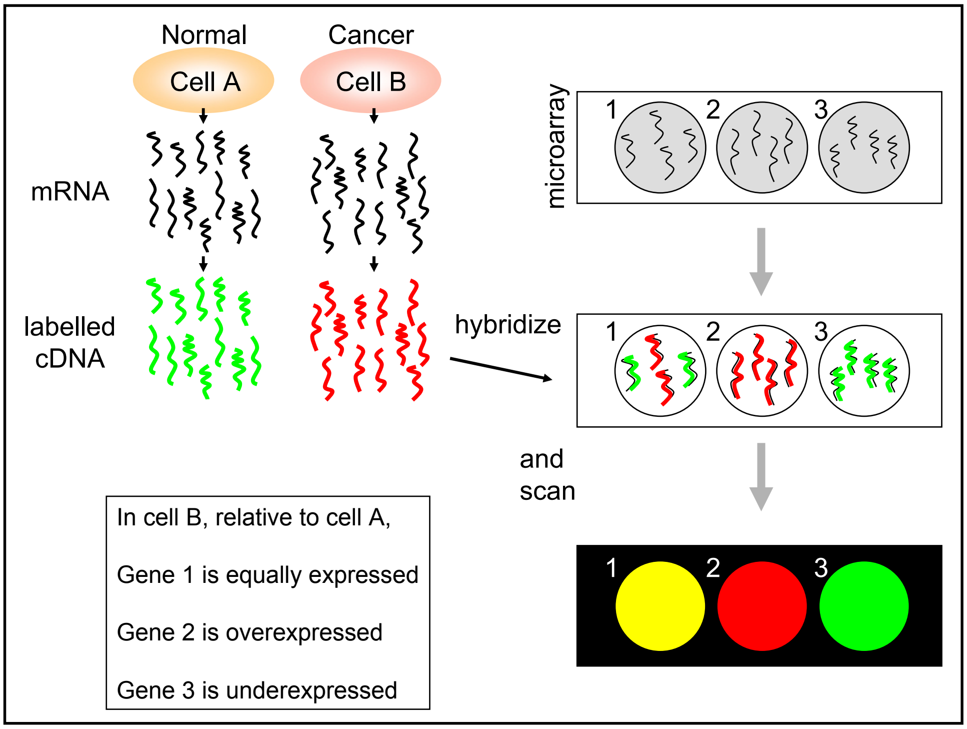
courtesy: *www.affymetrix.com*

Digital Micromirror Device





courtesy: *www.nimblegen.com*



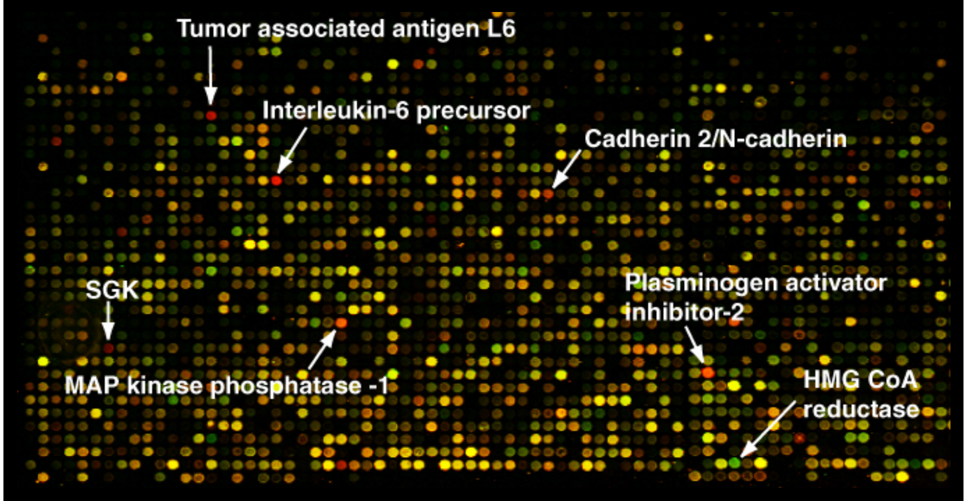
In cell B, relative to cell A,

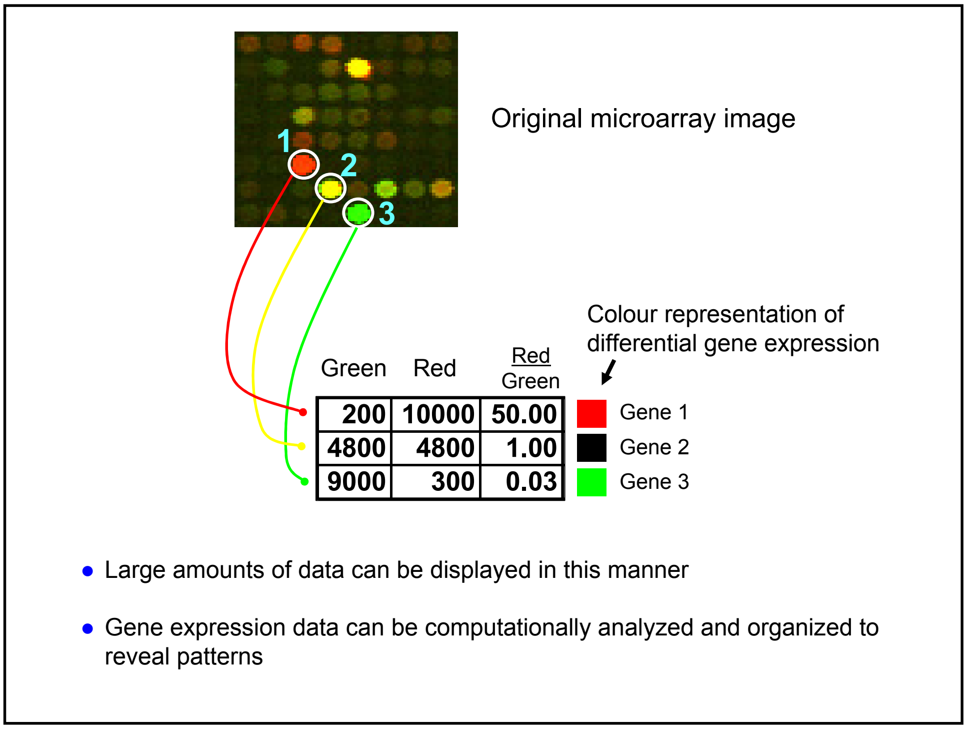
Gene 1 is equally expressed

Gene 2 is overexpressed

Gene 3 is underexpressed

DNA microarray after hybridization of fluorescent probes



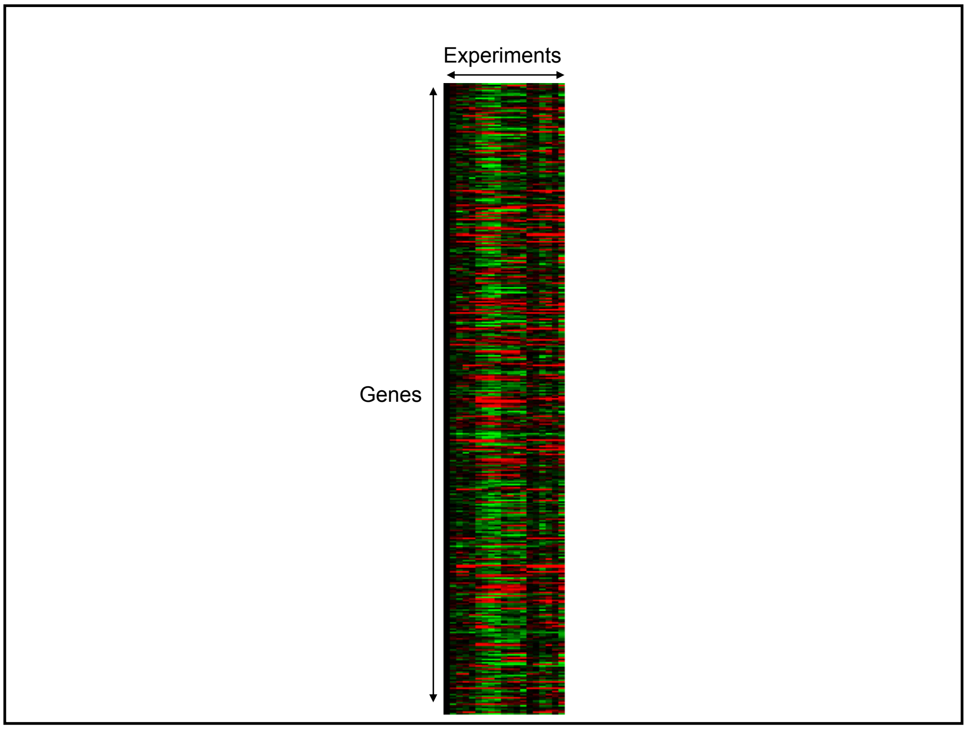


Original microarray image

Colour representation of differential gene expression

Green	Red	$\frac{\text{Red}}{\text{Green}}$		
200	10000	50.00	🟥	Gene 1
4800	4800	1.00	⬛	Gene 2
9000	300	0.03	🟩	Gene 3

- Large amounts of data can be displayed in this manner
- Gene expression data can be computationally analyzed and organized to reveal patterns



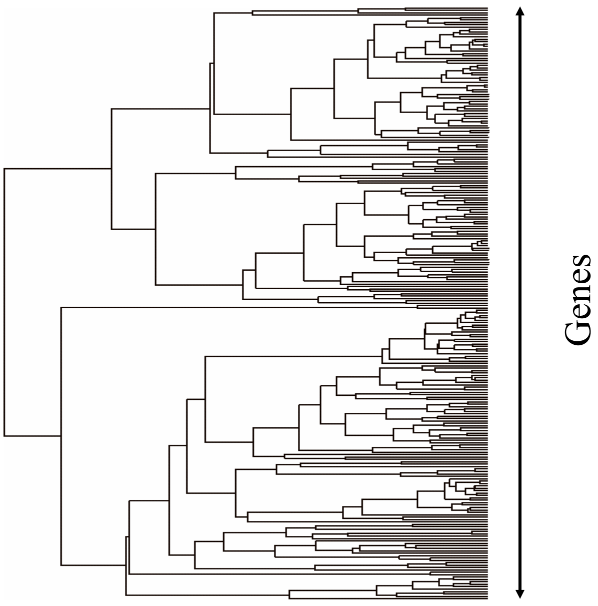
	Expt. 1	Expt. 2	Expt. 3	Expt. 4	Expt. 5	Expt. 6	Expt. 7	Expt. 8	Expt. 9	Expt. 10	Expt. 11	Expt. 12
Gene a	1.27	2.28	2.46	0.01	-0.54	-1.03	-0.94	-1.12	-0.29	-0.38	-0.15	0.03
Gene b	-0.45	1.62	1.83	0.03	0.33	0.25	-0.07	0.23	-0.4	-0.1	-0.36	-0.32
Gene c	1.42	3.03	3.67	0.58	0.66	0.78	0.3	-0.38	0.19	-0.01	-0.17	0.11
Gene d	0.56	2.05	2.43	0	1.36	0.06	-0.58	-0.04	-0.76	0.16	0.21	0.07
Gene e	0.01	2.24	3.41	1.58	1.86	0.69	0.08	-0.22	0.74	0.61	-0.32	-0.23
Gene f	0.58	0.59	1.31	0.75	2.58	1.22	1.08	0.93	0.38	-0.04	-0.09	-0.01
Gene g	-0.76	0.01	1.15	0.77	1.74	0.72	-0.36	-1.18	-0.15	-0.58	-0.45	-0.51
Gene h	-0.54	-0.38	-0.1	1.29	1.95	1.63	1.07	-0.86	-0.56	-0.64	-0.3	-0.42
Gene i	0.07	-0.67	0.94	0.4	1.81	1.64	1.1	-0.01	0.18	0.18	-0.07	0.1
Gene j	-0.42	-0.92	0.45	1.45	1.49	0.73	0.97	0.24	0.04	-0.14	-0.23	0.16
Gene k	0.37	0.07	-0.45	-0.47	2.49	1.81	0.96	-0.09	0.41	0.76	0.91	0.1
Gene l	-0.07	-0.14	0.01	0.1	2.8	1.34	0.56	0.55	0.48	0.18	0.33	-0.3
Gene m	-0.54	-0.27	-1.06	0.43	1.66	1.7	1.52	0.64	0.21	0.2	-0.12	0.23
Gene n	0.07	0.5	-0.09	0.01	1.57	1.71	1.54	0.86	-0.09	-0.49	-0.64	0.71
Gene o	0.25	0.82	0.78	0.61	2.26	2.61	1.77	1.17	0.66	-0.18	-0.29	1.14
Gene p	-0.07	0.56	0.93	0.28	1.37	2.85	2.21	0.84	0.37	0.29	-0.23	0.68
Gene q	0.23	0.56	0.39	0.23	1.64	3.16	2.89	0.28	-0.04	-0.36	-0.45	-0.29
Gene r	1.42	1.27	1.91	2.63	5.28	6.44	4.68	3.89	2.75	1.44	1.28	0.53
Gene s	-0.27	0.74	1.43	0.63	2.34	1.63	1.24	0.78	0.68	0.5	0.82	1.04
Gene t	0.1	0.55	0.71	0.59	2.37	1.59	1.12	0.63	-0.29	-0.17	-0.23	0.04

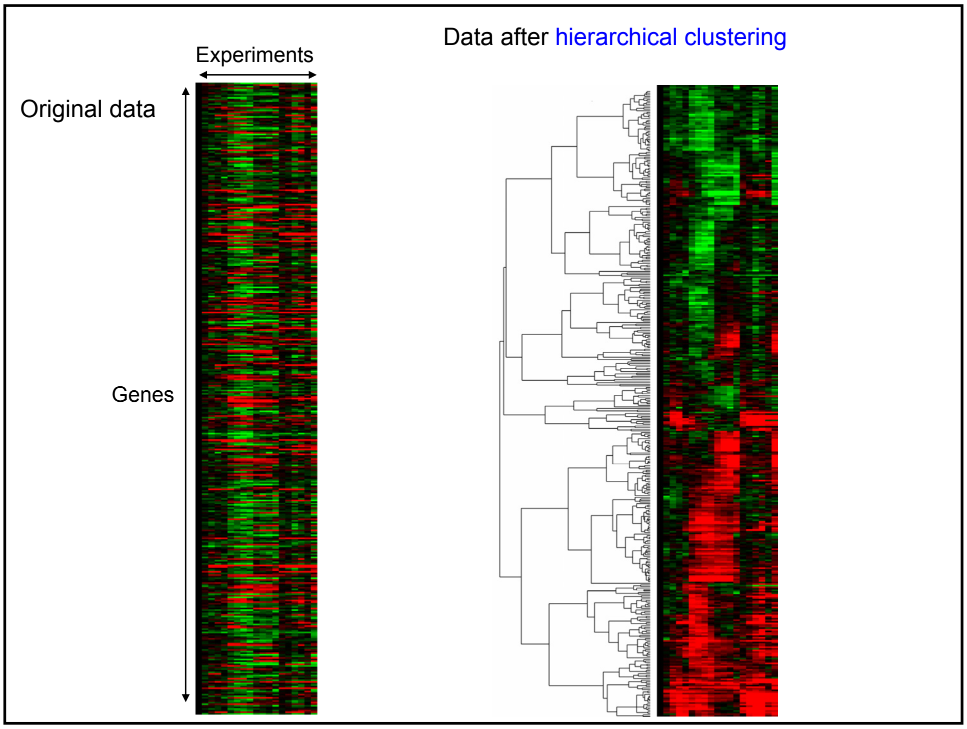
Expression vector

The Pearson correlation coefficient ***r,*** between two number series

$$X = \{X_1, X_2, \ldots X_N\} \text{ and } Y = \{Y_1, Y_2, \ldots Y_N\}$$

is given by $r = \frac{1}{N} \sum_{i=1,N} \left(\frac{X_i - \overline{X}}{\sigma_X} \right) \left(\frac{Y_i - \overline{Y}}{\sigma_Y} \right)$



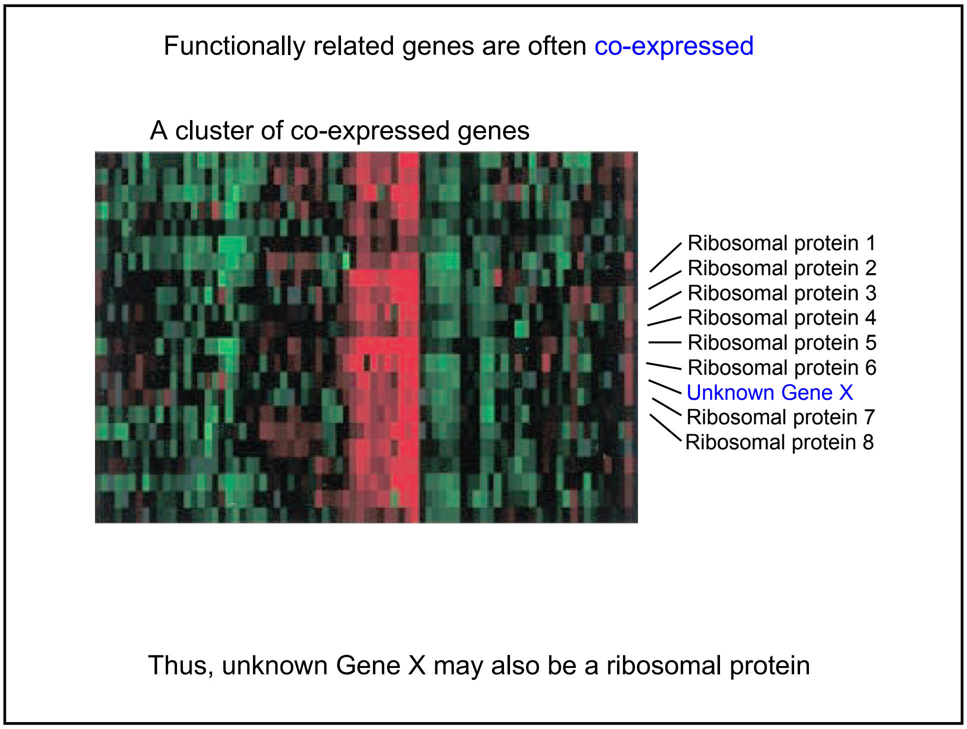


Data after hierarchical clustering

Experiments

Original data

Genes



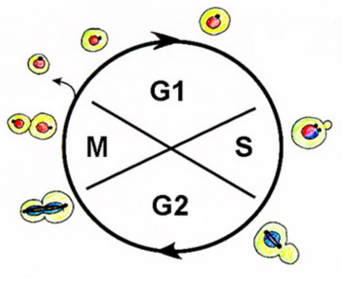
Functionally related genes are often co-expressed

A cluster of co-expressed genes

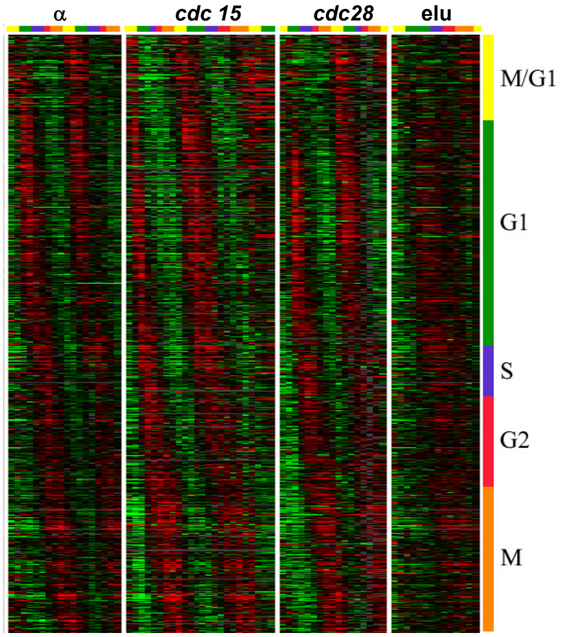
Ribosomal protein 1
Ribosomal protein 2
Ribosomal protein 3
Ribosomal protein 4
Ribosomal protein 5
Ribosomal protein 6
Unknown Gene X
Ribosomal protein 7
Ribosomal protein 8

Thus, unknown Gene X may also be a ribosomal protein

S. cerevisiae mitotic cell-cycle

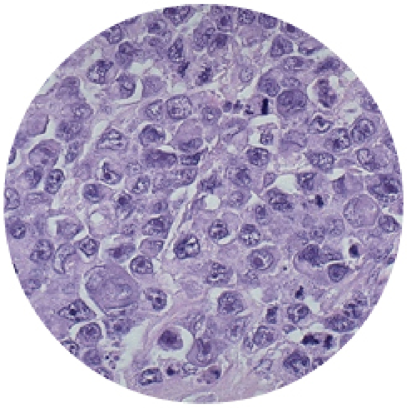


Brenda Andrews lab, University of Toronto



Spellman *et al*, (1998) *Mol. Biol. Cell*

The challenge of cancer diagnosis

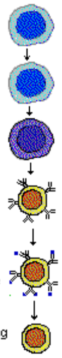


What type of cancer?

What is the underlying molecular basis?

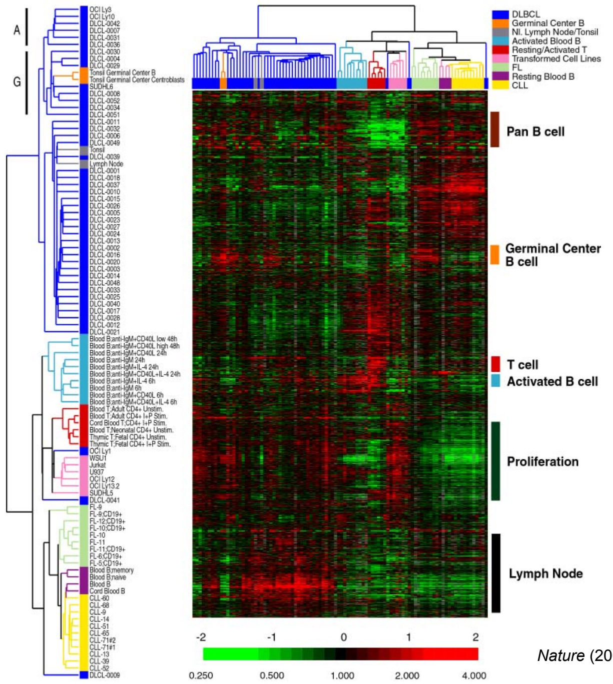
What is the optimal treatment?

B Cell Cancers by Cell Development

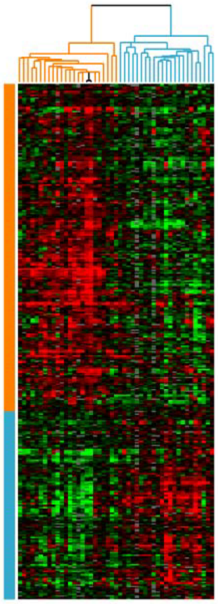


Cell Stage	Malignancy
Stem Cell	acute lymphocytic leukemia (ALL) lymphoblastic lymphoma/leukemia
Precursor B-Cell	Precursor B-lymphoblastic lymphoma / leukemia
Immature B-Cell	Small Non-Cleaved Cell Lymphoma (Burkitt's/Non-Burkitt's) (Possibly)
B-Cell Before Antigen Exposure	Chronic Lymphcytic Leukemia (CLL) Small Lymphocytic Lymphoma
B-Cell After Antigen Exposure	Follicular Lymphomas Large Cell Lymphoma Immunoblastic Lymphoma
Antibody Producing B-Cell	

Lymphoma Information Network
`http://www.lymphomainfo.net/`

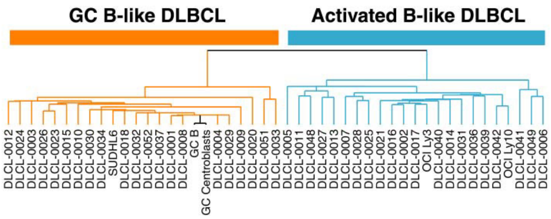


Nature (2000) **403**: 503

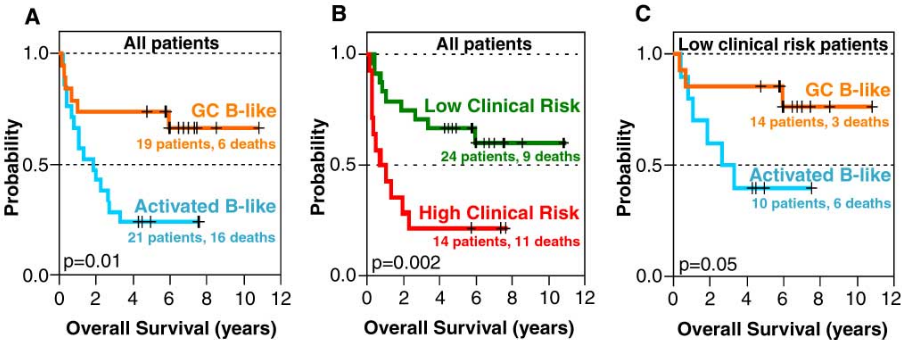


Clustering of tumour samples from cancer patients can be used for molecular classification of cancers. This may be useful for diagnosis and treatment

Subtypes of Diffuse Large B-Cell Lymphoma (DLBCL)



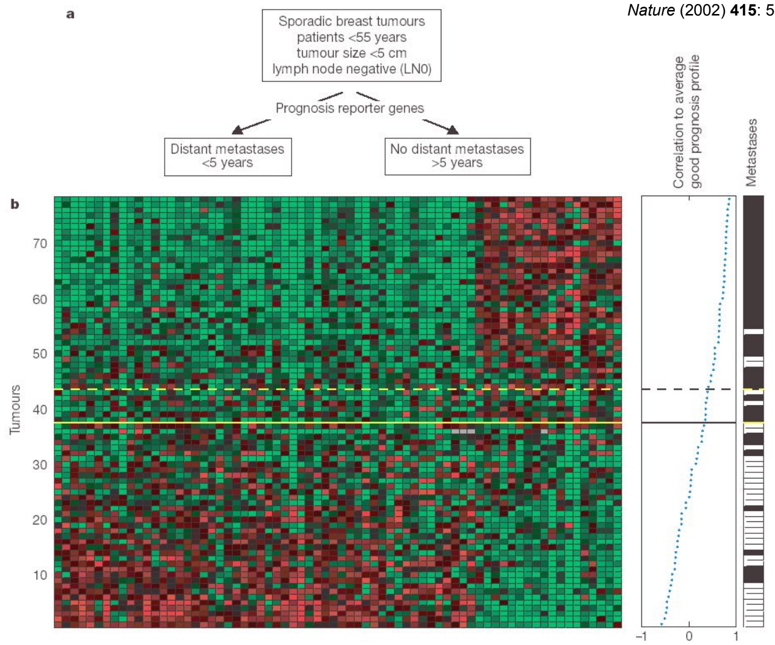
Nature (2000) **403**: 503



Nature (2000) **403**: 503

Identifying a gene expression signature for breast cancer metastasis

Nature (2002) **415**: 530



Nature (2002) **415**: 530

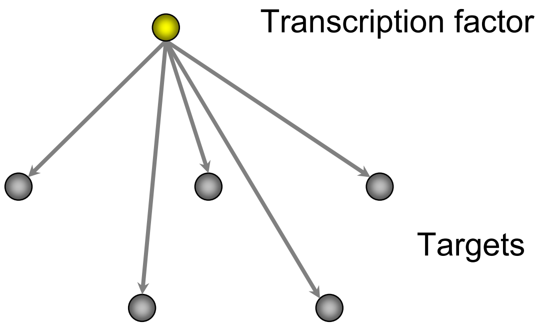
Bioinformatics & computational biology

- Databases
  - Genbank, SwissProt (DNA and protein sequence)
  - functional genomics and proteomics data (gene expression, protein profiles, drug data)
  - protein structure data (crystallography, NMR)
  - biomedical literature (PubMed)
- Analysis algorithms
  - finding patterns in expression data (clustering)
  - gene and protein interaction networks
  - data mining
  - regulatory elements, novel genes etc.
  - visualization

Other applications of microarrays

- Genomic amplifications and deletions
  Comparative Genomic Hybridization
- DNA-protein interactions
  mapping genome-wide distribution of proteins that interact with DNA
- RNA and protein localization
  analysis of RNAs associated with membrane-associated ribosomes, polysomes, different sub-cellular fractions
- Polymorphisms
  oligonucleotide (Affymetrix) arrays used for analyzing single nucleotide polymorphisms (SNPs) for linkage mapping and association studies
- Protein microarrays
  detecting proteins in complex mixtures
- Tissue microarrays
  high-throughput pathology

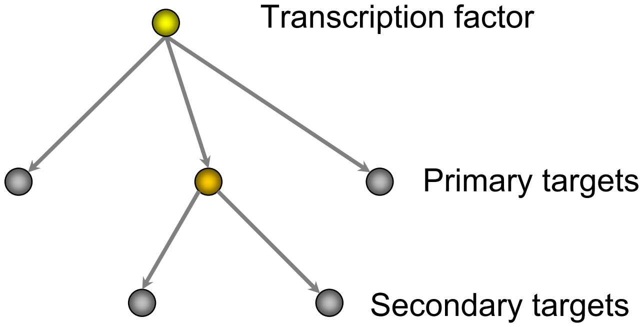
Transcription factor targets



Transcription factor

Targets

Transcriptional regulatory network



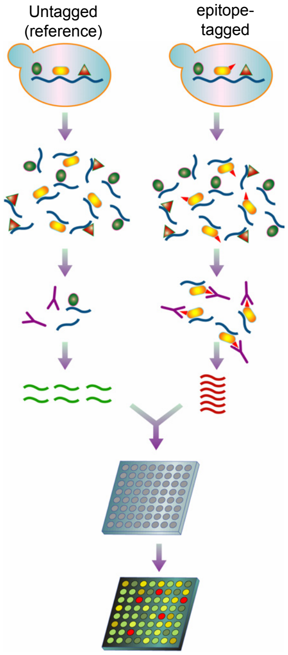
Transcription factor

Primary targets

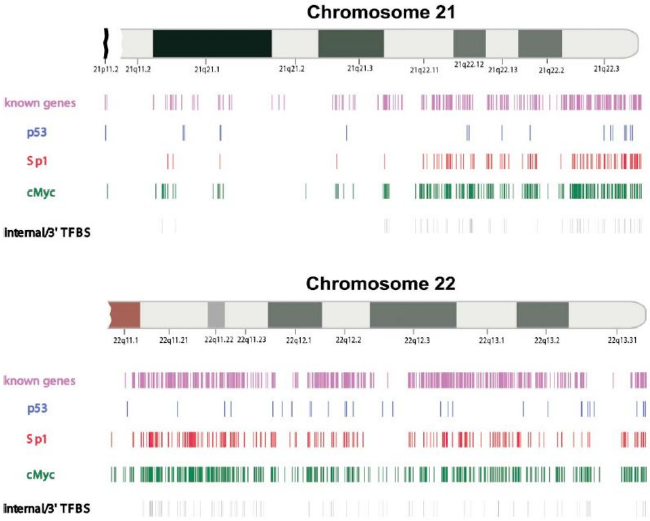
Secondary targets

Mapping the binding distribution of proteins on the genome

1. Crosslink proteins to DNA
2. Extract and shear crosslinked DNA
3. Immunoprecipitate with specific antibody
4. Reverse crosslinks, amplify and label DNA
5. Hybridize to intergenic microarray
6. Fluorescence ratio at each element is proportional to extent of binding



Protein binding sites on human chromosomes



	Chr 21 + 22	Whole human genome (?)
p53	48	1,600
Sp1	353	12,000
c-myc	756	25,000

Cell (2004) **116**: 499

Single Nucleotide Polymorphisms (SNPs)

SNPs are the main kind of measurable human genetic variation

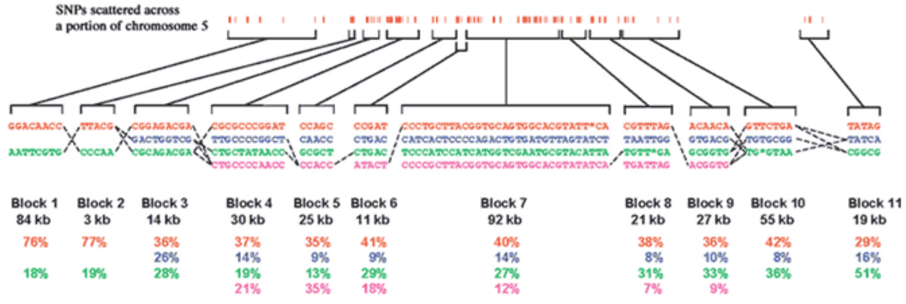
Allele 1 . . .A C T A A . . . . G G T A G . . . . A G C A. . .

Allele 2 . . .A C C A A . . . . G G T G G . . . . A G A A. . .

Allele 3 . . .A C T A A . . . . G G T A G . . . . A G C A. . .

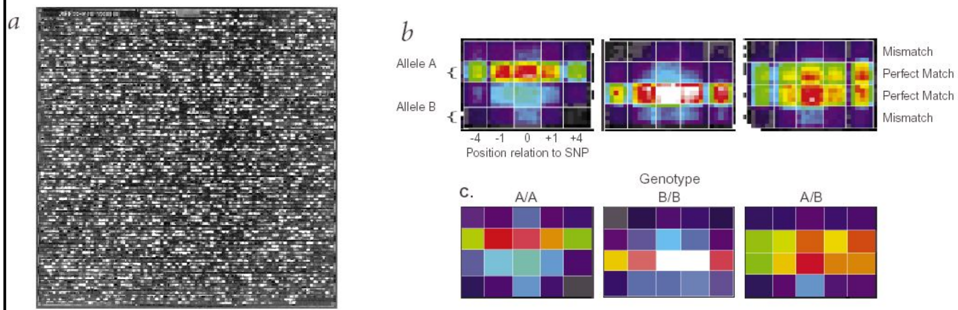
SNPs are inherited as blocks of associated SNPs = haplotype

Haplotypes in Crohn’s disease



Mark Daly, Whitehead Insititute - http://www.wi.mit.edu/nap/2002/nap_feature_hapmap-illus1.html

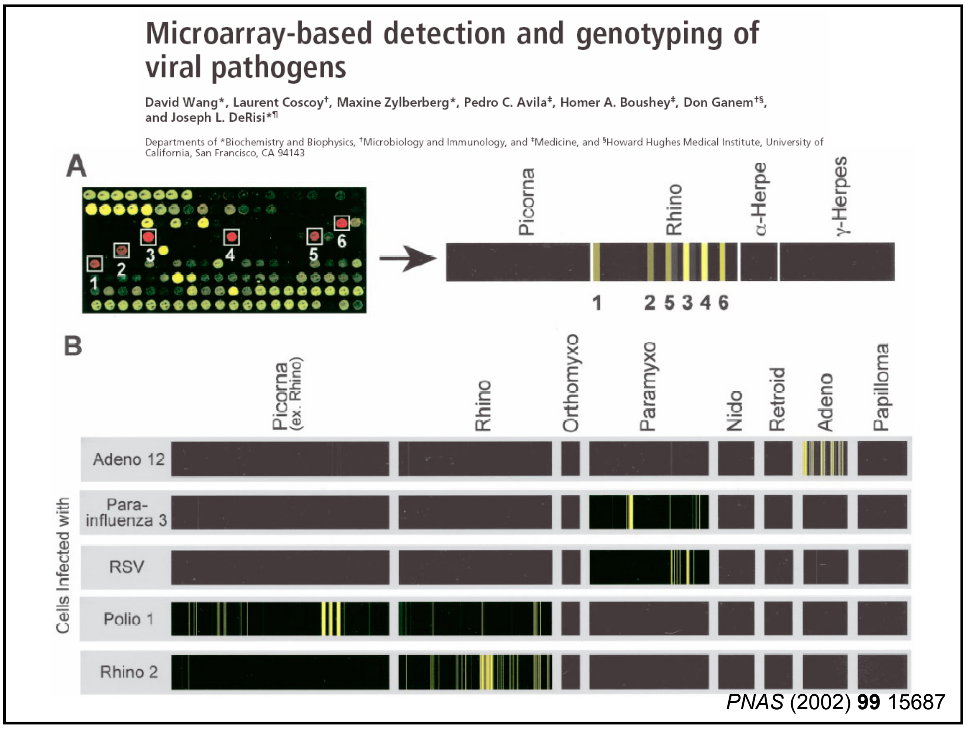
SNP genotyping with oligonucleotide arrays



Nature Genetics Suppl. (1999) **21**: 20

- HapMap Project – International project to map all human haplotypes
- The HapMap will be very useful for association studies for complex traits (Linking genotypes to inherited disease traits)

A haplotype map of the human genome (2005) *Nature* **437:** 1299



Microarray-based detection and genotyping of viral pathogens

David Wang*, Laurent Coscoy†, Maxine Zylberberg*, Pedro C. Avila‡, Homer A. Boushey‡, Don Ganem†§, and Joseph L. DeRisi*¶

Departments of *Biochemistry and Biophysics, †Microbiology and Immunology, and ‡Medicine, and §Howard Hughes Medical Institute, University of California, San Francisco, CA 94143

PNAS (2002) **99** 15687