Large-scale gene expression analysis

Qualitative and quantitative information on all transcribed sequences ("Transcriptome")

- Changes during developmental processes and cell differentiation
- Cell-type and tissue specific gene expression profiles
- Genes induced and repressed in response to drugs or environmental hits
- Genes regulated in response to cellular communication (signal transduction)
- Genes regulated in response to hormones
- Individual specificity of gene expression
- Changes during neoplastic transformation or other disease

How to measure the activity of all genes (genome-wide) in cells/tissues (mRNA).

Sequencing methods

(EST)

SAGE (LongSAGE, CAGE)

direct re-sequencing (deep sequencing) — Sequence approach

DNA microarrays, oligonucleotide microarrays.

Spotted arrays

In situ synthesized oligo arrays

Bead-arrays® — Hybridization approach

DNA chips: there are currently several types available:

Spotted

cDNA microarrays DNA microarrays	10,000 - 40,000 probes / cm^2 are spotted on glass slides using an automated microarrayer. Probes are cDNA or PCR products representing known genes or simply EST Chemically synthesized long oligonucleotides (50 – 70 nt) representing cds of known genes/EST

DNA chips: there are currently several types available:

2. In situ synthesized

Oligonucleotide arrays (Affychip®) Affymetrix	Up to 450,000 20-25nt long oligonucleotides / cm^2 are sythesized directly on the chip surface, using a photolitographic technique. Each gene is represented by a "probeset" of 12-13 probes.
Long oligonucleotide arrays	Up to 250,000 different 30-60 nt long oligonucleotides / cm^2 are sythesized directly on the chip surface, using an ink-jet technique.
Bead Arrays® Illumina	Oligonucleotide probes (30-50 nt) are synthesized on beads, with a identification address. Beads are randomly arrayed on surfaces and position of each oligonucleotide determined using addresses.

Spotted (pinspotted) DNA microarrays

A **1 - 40,000 probes / cm² are spotted on glass slides using an automated microarrayer.**

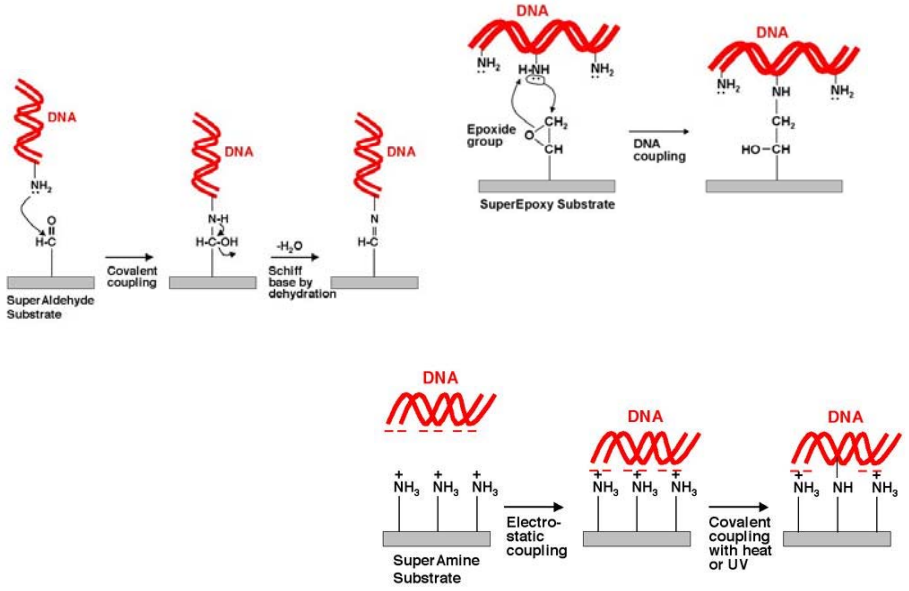
Probes are cDNA fragments or PCR products representing known genes or simply EST

Probe size: 200-1,000 nt

B **Same as above, but with:**

Oligonucleotide probes (30 – 70 nt in lenght)

Single-strand



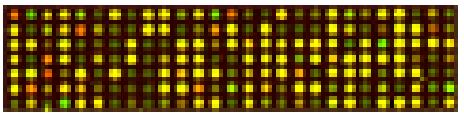
Long Oligonucleotide arrays

Up to 250,000 oligonucleotides / cm^2 are sythesized directly on the chip surface, using an ink-jet technique.

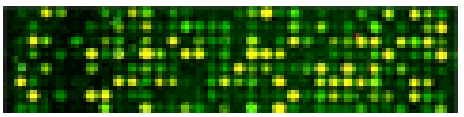
These 60 nt long oligonucleotides represent sequences of known genes or EST

Chemical synthesis of oligonucleotides

In situ



spotted



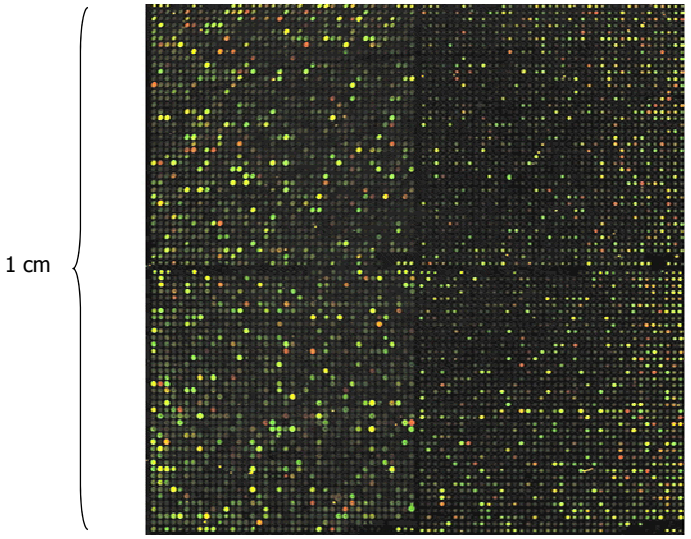
Spotted arrays, or ink-jet oligo arrays are commonly used for **relative** measurements, i.e. to compare gene expression between two biological samples.

RNA from **sample** and from **reference** are labeled by introducing two different fluorochromes.

This allows co-hybridization of the two samples to the same chip, providing direct comparison by two-color analysis

The most common fluorochromes are the cyanines Cy3 (red) and Cy5 (green)

How a spotted microarrays hybridized with two-colors probes looks like



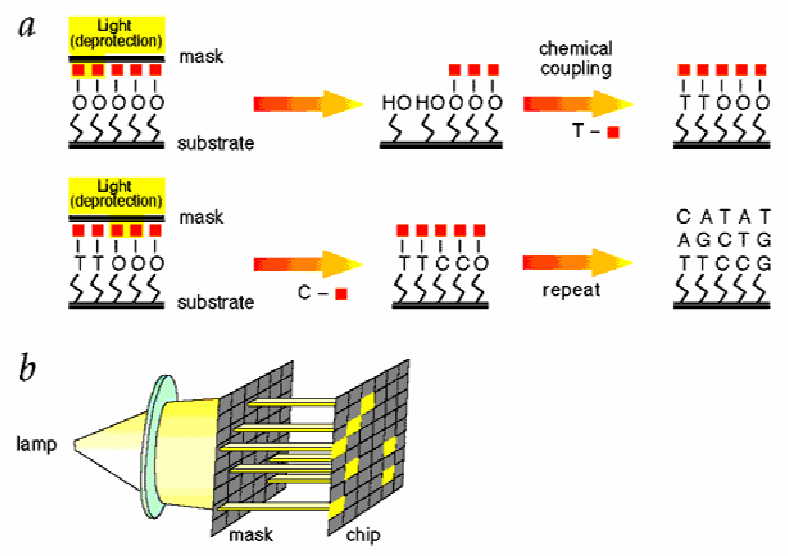
Short Oligonucleotide arrays (Affymetrix)

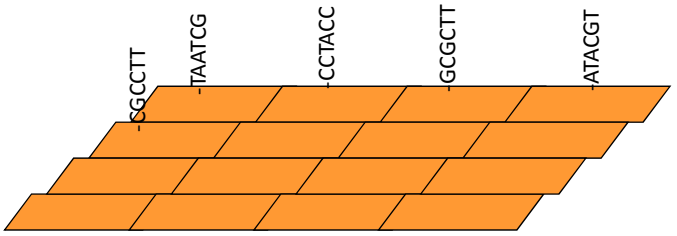
Up to millions of oligonucleotides / cm^2 are synthesized directly on the chip surface, using a photolithographic technique.

These 20-25 nt long oligonucleotides represent sequences of known genes or EST

Chemical synthesis of oligonucleotides

Affymetrix technology allows direct synthesis of oligonucleotides in micro-zones on the glass slide by photolithograpy





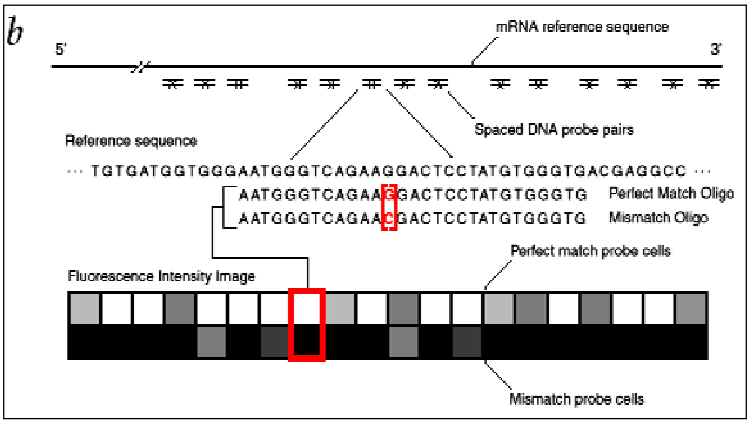
(Ovviamente, in ogni quadratino ci sono migliaia di oligonucleotidi uguali)

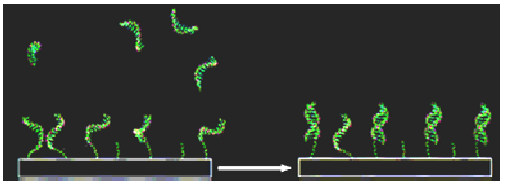
One limitation: the lenght of oligonucleotides (20-25 nt)

This can give problems of aspecific hybridization due to:
-Similar sequences can be present in different mRNAs
-Hybridization is done at the same temperature that is average of the optimal temperature for individual probe-target pairs

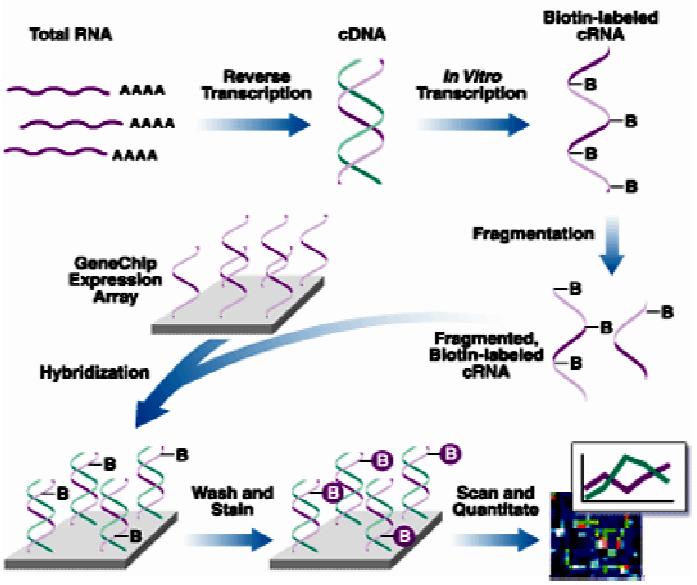
Affymetrix gives one solution to these problems:

Affymetrix GeneChip probe set

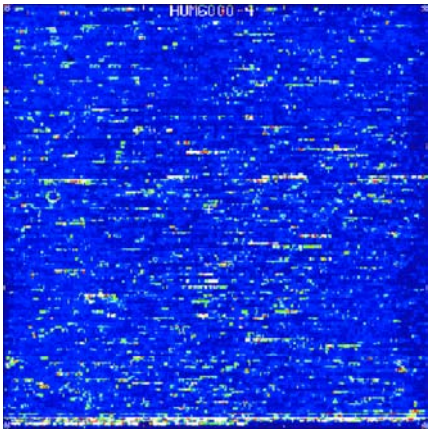




How an Affychip is used



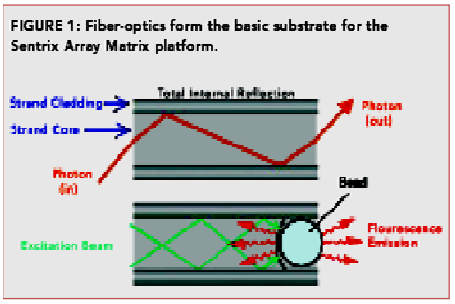
How an Affychip result looks like

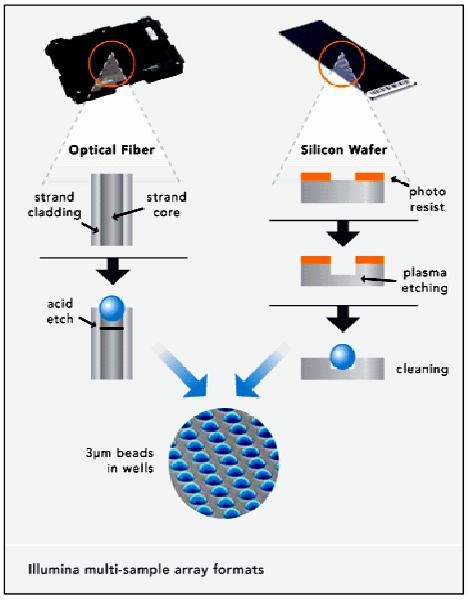


An oligonucleotide array (Affychip®) hybridized to biotin-labelled cRNA and revealed with fluorochrome-conjugated avidin

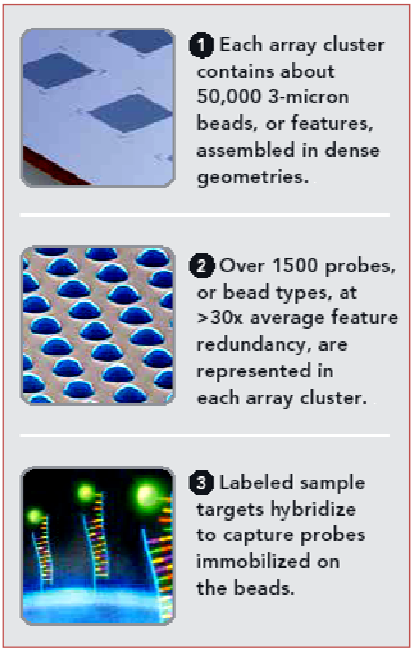
Last generation:
Bead-arrays®

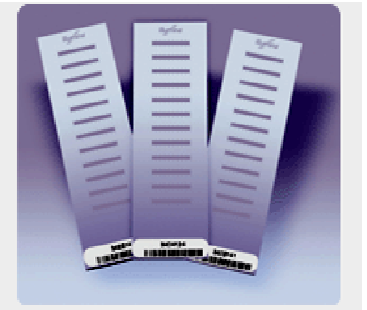
This kind of array was patented by Illumina (San Diego, CA, USA

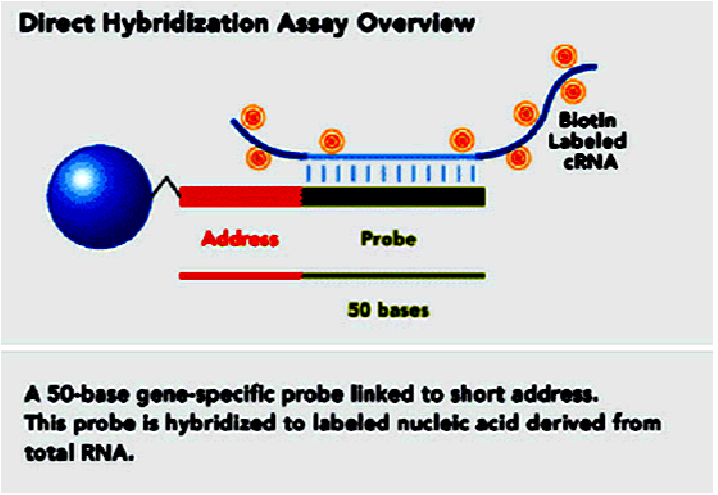




illumina.







Reading of microarrays is performed with laser scanners, which allow a quantitation of fluorescence in different channels

Scanners produce a **table of values** that are intensities at each spot. If using double colors, relative fluorescence intensities in two channels are read.

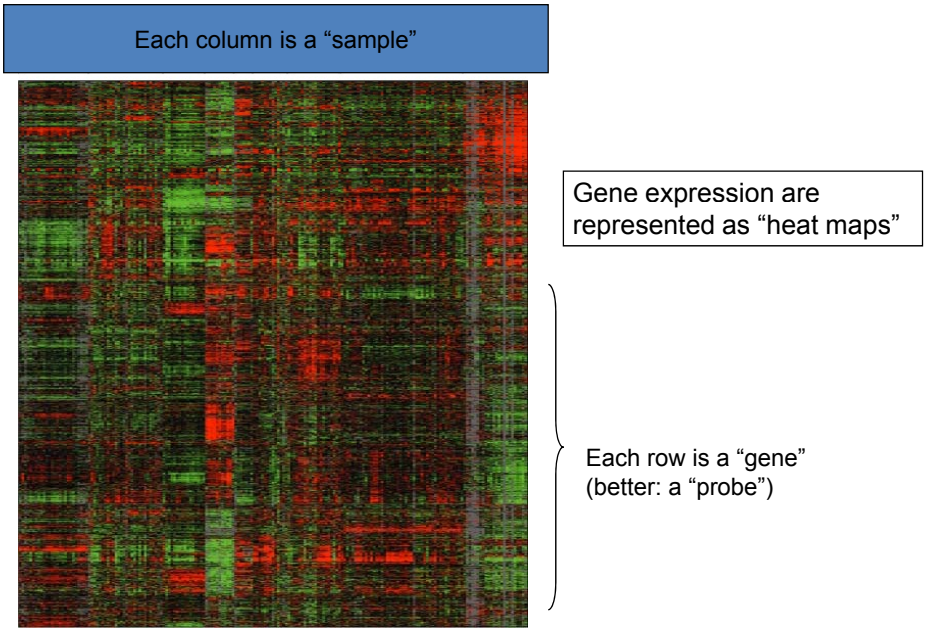
First important difference to know is therefore:

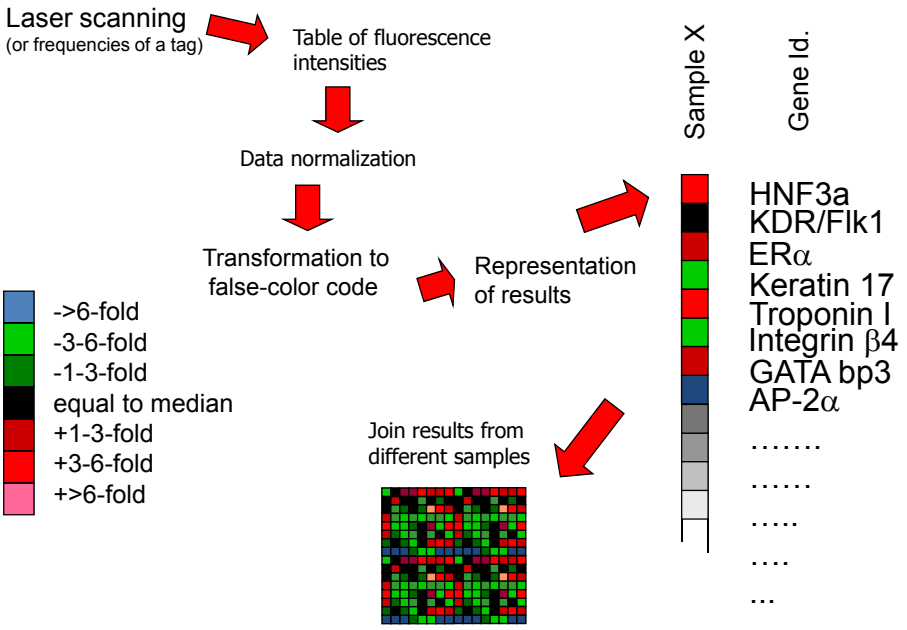
Absolute *versus* relative measurement

Using hybridization reaction as an absolute measurement of RNA requires that the amount of probe on each spot be uniform and reproducible.

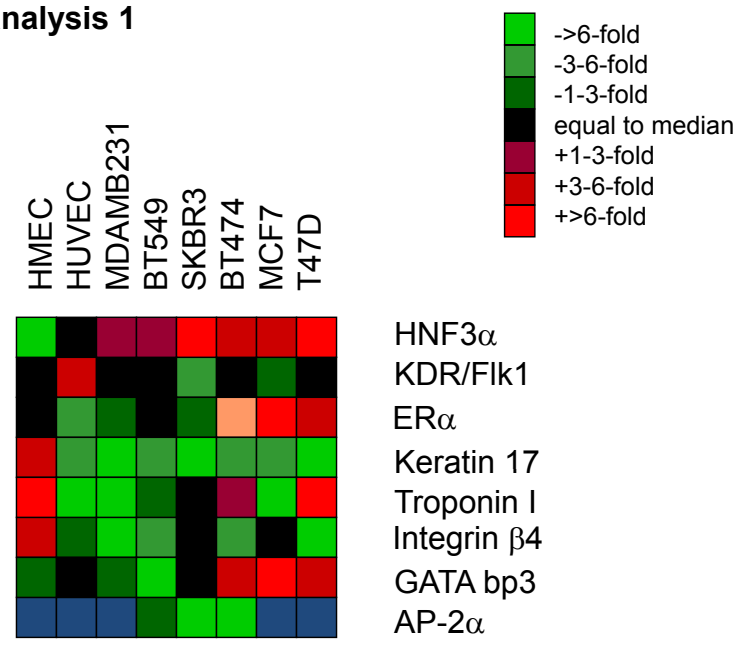
This is a requirement fullfilled by Affymetrix arrays and by the latest generation of in situ synthesized long-oligo arrays and bead arrays.

In all the other cases, the amount of probe / spot is variable and unassessable, so that **relative** measurements are necessary.

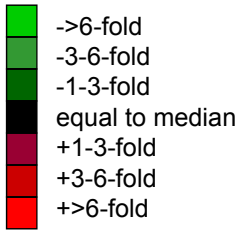


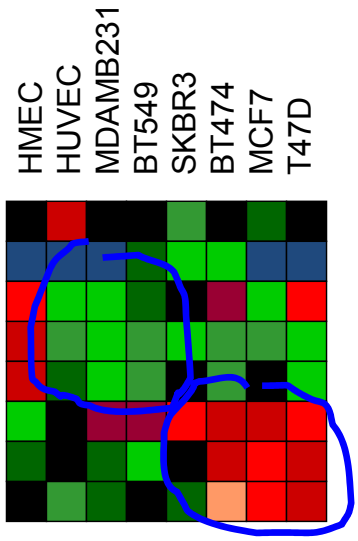


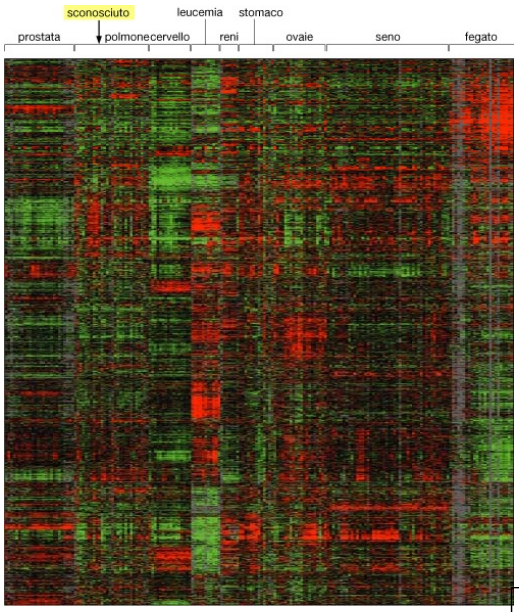
Cluster analysis 1



Cluster analysis 2







Gene expression are represented as “heat maps”

Different expression profiles in human cells of different tissues: 1800 genes probes

The Transcriptional Program of Sporulation in Budding Yeast

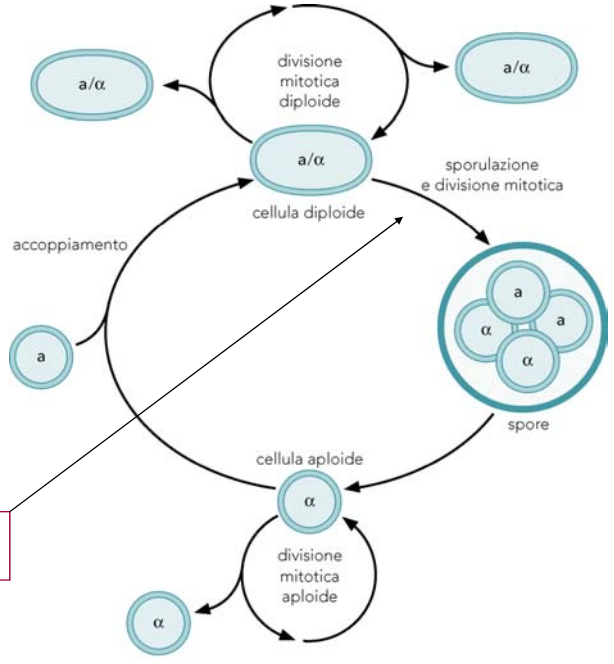
S. Chu,* J. DeRisi,* M. Eisen, J. Mulholland, D. Botstein,
P. O. Brown,† I. Herskowitz†

Diploid cells of budding yeast produce haploid cells through the developmental program of sporulation, which consists of meiosis and spore morphogenesis. DNA microarrays containing nearly every yeast gene were used to assay changes in gene expression during sporulation. At least seven distinct temporal patterns of induction were observed. The transcription factor Ndt80 appeared to be important for induction of a large group of genes at the end of meiotic prophase. Consensus sequences known or proposed to be responsible for temporal regulation could be identified solely from analysis of sequences of coordinately expressed genes. The temporal expression pattern provided clues to potential functions of hundreds of previously uncharacterized genes, some of which have vertebrate homologs that may function during gametogenesis.

SCIENCE VOL 282 23 OCTOBER 1998 699

This paper is included as a resource in the Moodle site.
It makes part of the course and students should read and study it

Saccaromyces cerevisiae



nitrogen-deficient medium induces sporulation

How to measure the activity of all genes (genome-wide) in cells/tissues (mRNA).

Sequencing methods

(EST)	Sequence approach
SAGE (LongSAGE, CAGE)	
direct re-sequencing (deep sequencing)	
DNA microarrays, oligonucleotide microarrays.	
Spotted arrays	Hybridization approach
In situ synthesized oligo arrays	
Bead-arrays®	

how to create a cDNA library

EST = expressed sequence tags

1. mRNA extraction from cells or tissues
2. cDNA synthesis (oligo-dT or random-primed)
3. cloning into plasmid vectors
4. sequencing from vector primers (200-300nt)
5. Estimate expression from frequency

automated DNA sequencing

[Current EST Databases contain millions of EST]

on line sequence databases, BLAST

SAGE

SCIENCE • VOL. 270 • 20 OCTOBER 1995

Serial Analysis of Gene Expression

Victor E. Velculescu, Lin Zhang, Bert Vogelstein,
Kenneth W. Kinzler*

The characteristics of an organism are determined by the genes expressed within it. A method was developed, called serial analysis of gene expression (SAGE), that allows the quantitative and simultaneous analysis of a large number of transcripts. To demonstrate this strategy, short diagnostic sequence tags were isolated from pancreas, concatenated, and cloned. Manual sequencing of 1000 tags revealed a gene expression pattern characteristic of pancreatic function. New pancreatic transcripts corresponding to novel tags were identified. SAGE should provide a broadly applicable means for the quantitative cataloging and comparison of expressed genes in a variety of normal, developmental, and disease states.

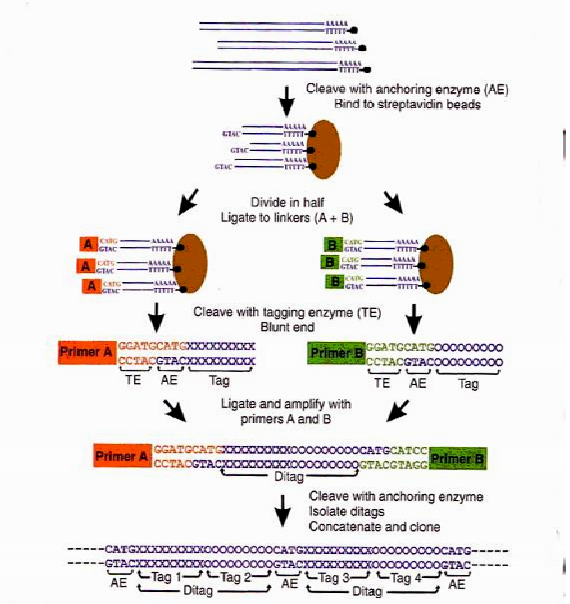
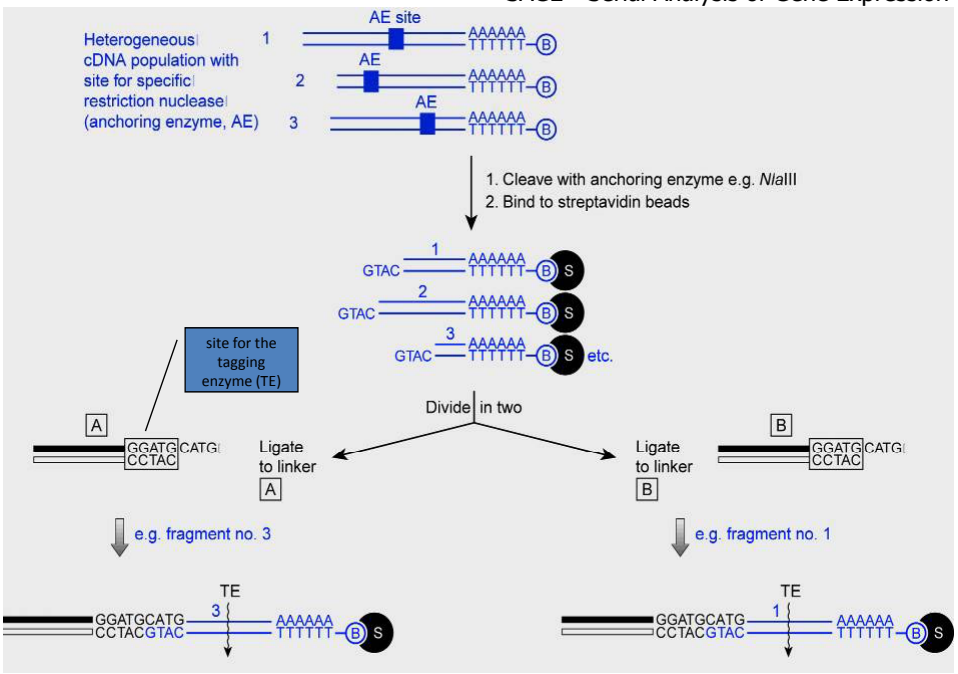
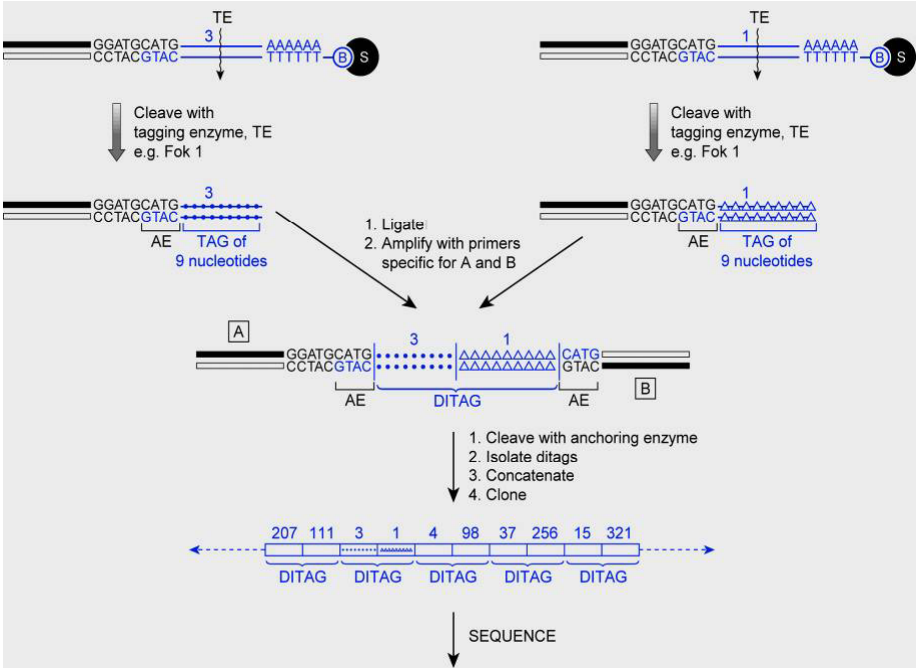


Fig. 1. Schematic of SAGE. The anchoring enzyme is Nla III and the tagging enzyme is Fok I. Sequences colored red and green represent primer-derived sequences, whereas blue represents transcript-derived sequences, with X and O indicating nucleotides of different tags. See text for further explanation.

SAGE= Serial Analysis of Gene Expression





SAGE analysis produces a table of frequencies for each tag

Frequencies are proportional to representation of that tag in the sample, i.e. to the level of expression of that specific mRNA

Table 1. Pancreatic SAGE tags. Tag indicates the 9-bp sequence identifying each tag, adjacent to the 4-bp anchoring Nla III site. *n* and Percent indicate the number of times the tag was identified and its frequency, respectively. Gene indicates the description and accession number of the GenBank release 87 entry found to exactly match the indicated tag when the SAGE software group was used, with the following exceptions. When multiple entries were identified because of duplicated entries (*7*), only one entry is listed. For chymotrypsinogen and trypsinogen 1, other genes (adenosine triphosphatase and myosin alkali light chain, respectively) were identified that were predicted to contain the same tags, but subsequent hybridization and sequence analysis identified the listed genes as the source of the tags. Alu entry indicates a match with a GenBank entry for a transcript that contained at least one copy of the Alu consensus sequence (*15*).

Tag	Gene	*n*	Percent
GAGCACACC	Procarboxypeptidase A1 (X67318)	64	7.6
TTCTGTGTG	Pancreatic trypsinogen 2 (M27602)	46	5.5
GAACACAAA	Chymotrypsinogen (M24400)	37	4.4
TCAGGGTGA	Pancreatic trypsin 1 (M22612)	31	3.7
GCGTGACCA	Elastase IIIB (M18892)	20	2.4
GTGTGTGCT	Protease E (D00300)	16	1.9
TCATTGCCC	Pancreatic lipase (M20285)	16	1.9
CCAGAGAGT	Procarboxypeptidase B (M81057)	14	1.7
TCCTCAAAA	No match (see Table, 2, P1)	14	1.7
AGCCTTGGT	Bile salt stimulated lipase (X54457)	12	1.4
GTGTGCGCT	No match	11	1.3
TGCGAGACC	No match (see Table 2 P2)	9	1.1
GTGAAACCC	21 Alu entries	8	1.0
GGTGACTCT	No match	8	1.0
AAGGTAACA	Secretory trypsin inhibitor (M11949)	6	0.7
TCCCCTGTG	No match	5	0.6
GTGACCAGG	No match	5	0.6
CCTGTAATC	M91159, M29366, 11 Alu entries	5	0.6
CAGGTTGGA	No match	5	0.6
AGCCCTACA	No match	5	0.6
AGCACCTCC	Elongation factor 2 (Z11692)	5	0.6
ACGCAGGGA	No match (see Table 2, P3)	5	0.6
AATTGAAGA	No match (see Table 2, P4)	5	0.6
TTCTGTGGG	No match	4	0.5
TTCATAGAG	No match	4	0.5
GTGGCAGGC	NF-κB (X61499), Alu entry (S04541)	4	0.5
GTAAAACCC	TNF receptor II (M55994), Alu entry (X01448)	4	0.5
GAACACACA	No match	4	0.5
CCTGGGAAG	Pancreatic mucin (J05582)	4	0.5
CCCATCGTC	Mitochondrial CytC oxidase (X15759)	4	0.5
SAGE tags occurring:	Greater than three times	380	45.2
	Three times (15 × 3 =)	45	5.4
	Two times (32 × 2 =)	64	7.6
	One time	351	41.8
	Total SAGE tags	840	100.0

Students are expected to explore and use resources that are allocated in the Moodle website.

Some interactive problems and exercise are also present: I strongly invite students to perform these activities.