生物信息学分析

章节	源数据	结果知识	种类
四、序列分析 *	DNA序列	基因等特征序列	Seq.
	蛋白质序列	特征域、特性	
	EST	表达基因（mRNA）	Expr.
五、系统发育分析	DNA/蛋白质序列	进化历史	Evol.
六、基因组分析	基因组序列	基因位置、功能、 物种进化历史	Seq. Evol.
（转录组分析）	Microarray	表达基因（mRNA）	Expr.
	RNA-seq		
七、蛋白质组分析	2D-Page	表达基因（蛋白质）	Expr.
	Y2-hybrid ...	蛋白质相互作用…	Net.
八、结构分析	蛋白质序列	蛋白质结构	Struct.
	RNA序列	RNA结构	

Microarrays (1)

——本质是生物信息的集成性平行分析：利用核酸分子杂交（蛋白质分子亲和）原理，通过荧光标记可视化，借助计算机分析处理，可迅速获取大量生物信息，效率是传统手段的成百上千倍。

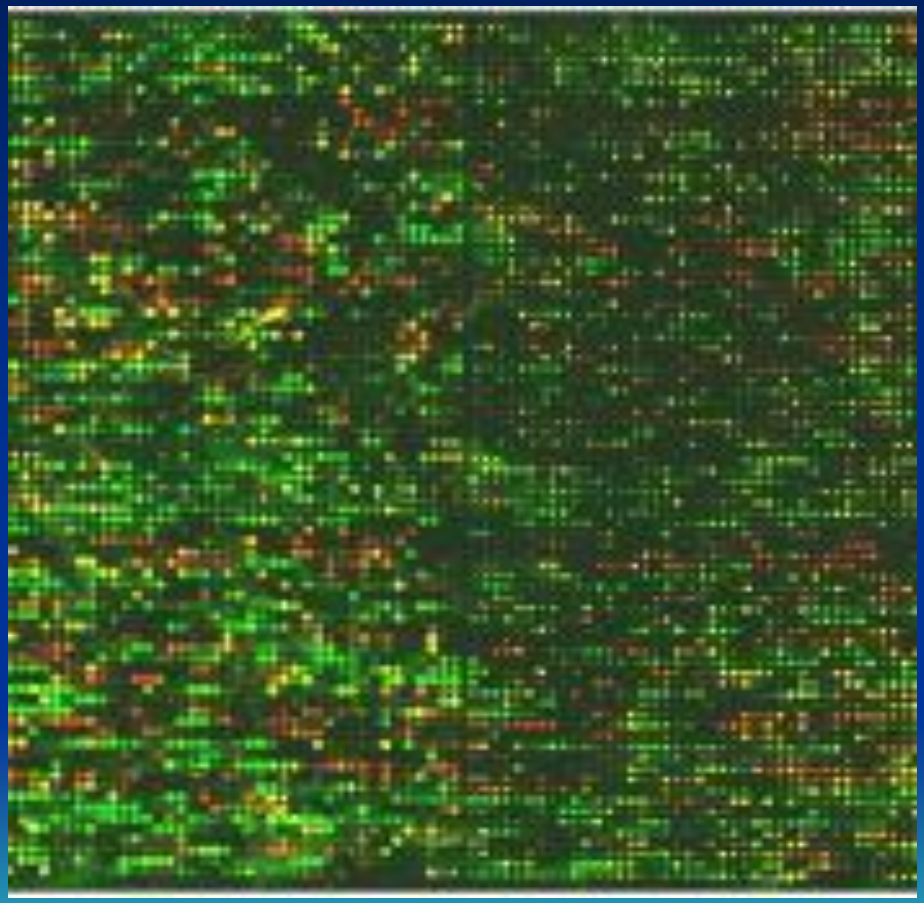
•There are several names for this technology – biochips, microchips, DNA microarrays, DNA arrays, DNA chips, gene chips, others. Sometimes a distinction is made between them but in fact they are all synonyms as there are no standard definitions for each name —— EBI.

•Two major technologies ----

Spotted DNA microarrays

Oligonucleotide GeneChips （ US company Affymetrix Inc.)

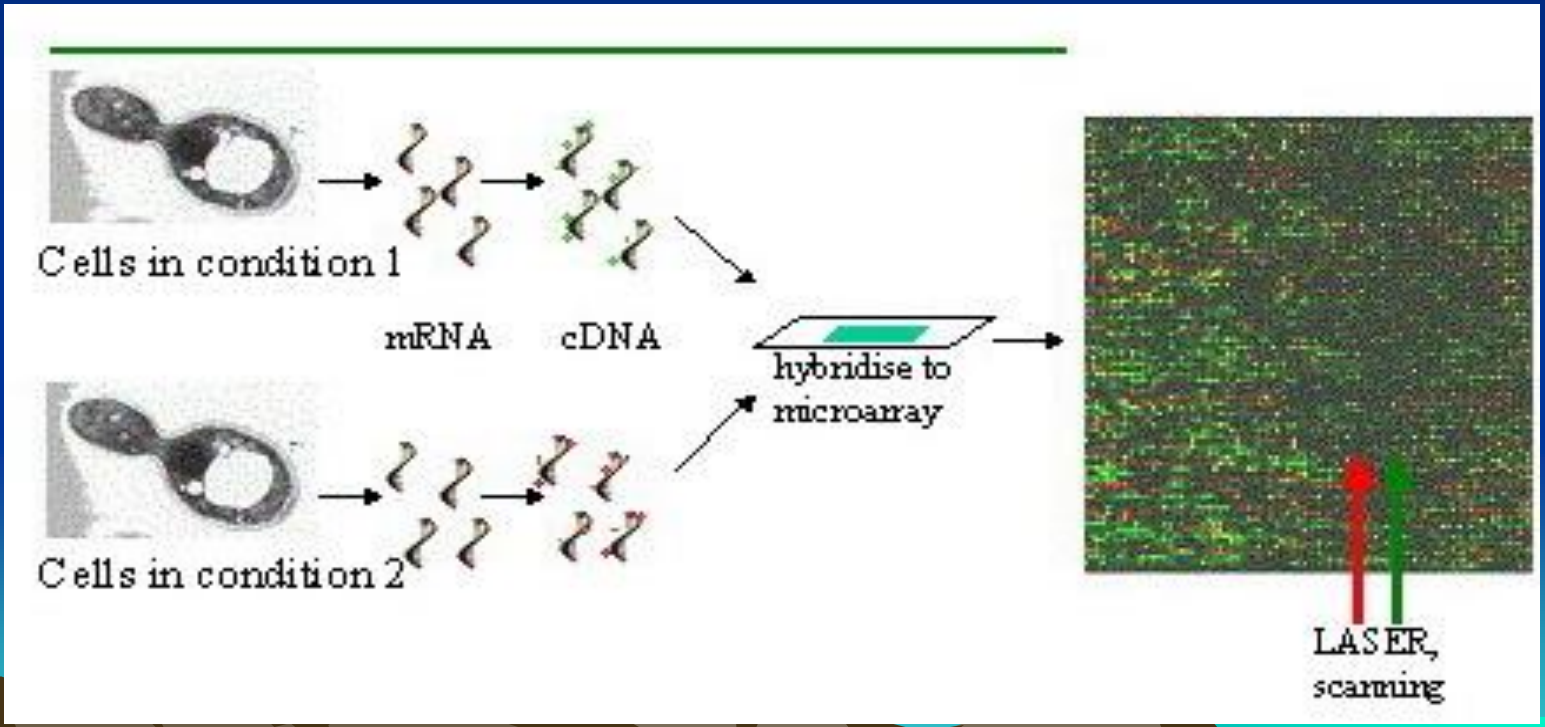
Microarrays (2)



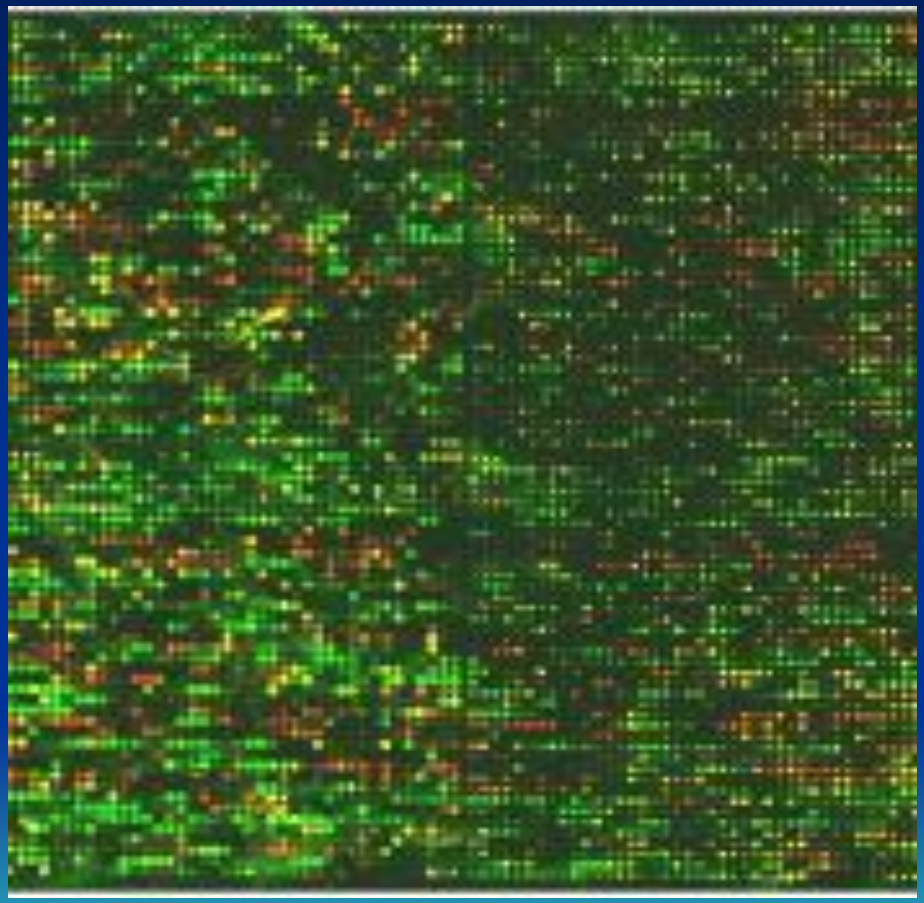
- DNA molecules are attached at fixed locations (5,000 spots/cm^2, diameter 0.1 mm).
- Each point contains a huge number (10^7-10^8) of identical DNA molecules (clone).
- Each clone ideally should identify one gene or one exon in the genome.

Microarrays (3)

- Example --- comparing gene expression levels in a healthy cell and a diseased cell.



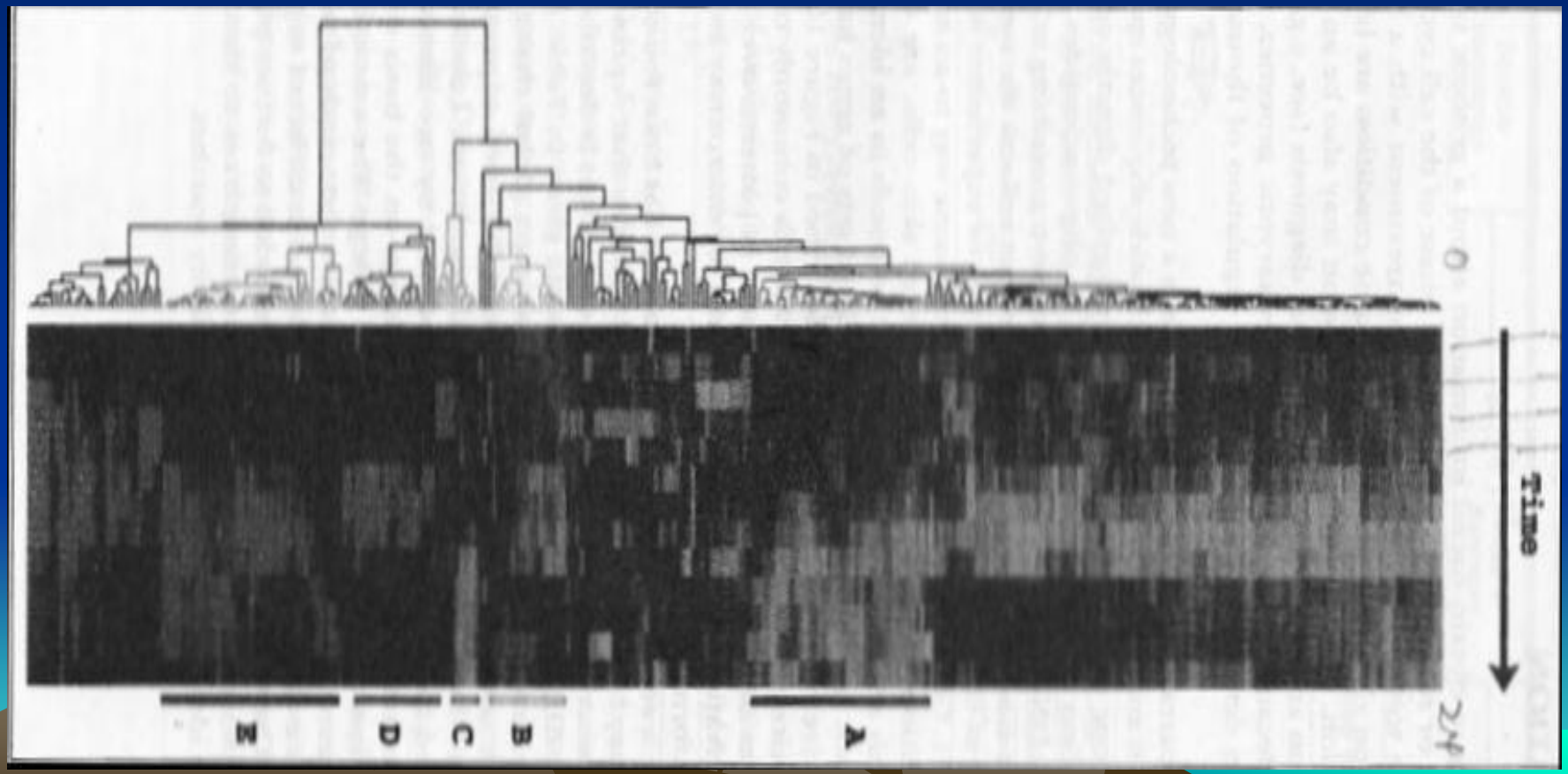
Microarrays (2)



- DNA molecules are attached at fixed locations (5,000 spots/cm^2, diameter 0.1 mm).
- Each point contains a huge number (10^7-10^8) of identical DNA molecules (clone).
- Each clone ideally should identify one gene or one exon in the genome.

Microarrays (4)

- Microarray分析：图像分析（去噪音和信号数据化）、标准化（重复实验的可比性）、Ratio分析（两色荧光的比值）、基因聚类分析（寻找同类基因）。



RNA-seq

- RNA-seq describes experimental and computational methods to determine the identity and abundance of RNA sequence in biological samples.
- RNA-seq methods are derived from generational changes in sequencing technology.

从EST 到 RNA-seq

	EST	RNA-seq
cDNA文库的建立	克隆	PCR
测序方法	第一代	第二代
读段量	**	******
辅助基因发现	***	*****
表达差异分析	*	*****

新兴测序技术

- 第二代测序（SGS/NGS）：通量高

  Illumina: 600G nt / run (Sanger: 0.1M nt / run) ;

  Roche 454: 300~400nt / read (似Sanger法).

  SOLID: sequencing by oligonucleotide ligation – accuracy 99.99% (Sanger：99%) ;

- **第三代测序（TGS）：单分子测序**

  Pacific Biosciences: 5000 nt or longer / read ;

  Nanopore technologies: 微电流测序 – …RNA/P

GENCODE

Human Mouse How to access data FAQ Documentation Abo

Long-read RGASP

The Long-read RNA-seq Genome Annotation Assessment Project (LRGASP) Consortium is organizing a systematic evaluation of different methods for transcript computational identification and quantification using long-read sequencing technologies such as PacBio and Oxford Nanopore. We are interested in characterizing the strengths and potential remaining challenges in using these technologies to annotate and quantify the transcriptomes of both model and non-model organisms.

从EST 到 RNA-seq

	EST	RNA-seq
cDNA文库的建立	克隆	PCR
测序方法	第一代	第二代
读段量	**	******
辅助基因发现	***	*****
表达差异分析	*	*****

Microarray *vs.* RNA-seq

表 6-7　RNA-seq 的技术优势

	基因芯片	RNA-seq
参考序列	需要	不需要
动态范围	小	大
背景噪声	大	小
受降解影响	大	小
序列变异	无法检测	可以检测
转录组方向	不能确定	能确定
可重复性	一般	高

相关计算机分析

- **RNA-seq：**质量控制、读段清理、表达定量、序列拼接、相关系数分析、表达差异分析……
- Microarray：图像分析、标准化、Ratio分析、表达差异分析或基因聚类分析。
- EST：预处理（去除载体、接头以及引物等“污染物”）、聚类、序列拼接…

基因组、转录基因组和蛋白质组

- 染色体基因组，或简称基因组，即生物体内所有细胞中的遗传信息。—>DNA
- 转录基因组，即细胞某个特定生长阶段中的表达部分。—>mRNA （EST, microarray……）
- 蛋白质组，反映细胞特性和功能的所有蛋白质分子。 —>蛋白质

Omics

后基因组时代“组学”研究的三个层次

RNA-seq --- 真正的转录组分析

- RNA-seq describes experimental and computational methods to determine the identity and abundance of RNA sequence in biological samples.
- RNA-seq 分析的靶标并不局限于mRNA，还包括noncoding RNA，如tRNA、 rRNA、miRNA和lncRNA等，是真正的、全面的转录组分析技术。