These background are needed:

1. **- Basic Molecular Biology & Genetics**

DNA replication
Transcription
Post-transcriptional RNA processing
Translation
Post-translational protein modification
Gene expression regulation (basic mechanisms)

Basics of protein structure and molecular representations

Example:
Chapters 1 through 10 from "Essential Cell Biology" 2° or 3° Edition – Alberts et al., Garland, 2004 (2°), 2009 (3°)
Italian version – Zanichelli (2005)

Basic recombinant DNA methodology:

1. DNA replication (in vivo and in vitro)
2. PCR, rt-PCR and real-time PCR
3. Basic DNA cloning in plasmids and other vectors
4. Libraries, clones, colonies, storage, propagation, analysis.
5. DNA sequencing, restriction, Southern blot
6. RNA analysis, Northern blot, Rnase protection

Basic bioinformatics:

1. Database organization
2. Finding gene and protein sequences
3. Basic alignment protocols

Dale & von Schantz - Dai Geni ai Genomi – Edises 2008 (19 €)
Reece – Analisi dei geni e genomi – Edises 2006 (30€)
Watson Caudy Myers Witkowski – DNA ricombinante – Zanichelli 2009

Nice book for advanced students: T.A. Brown - Genomi 3 - Edises 2008

Riassunto delle lezioni precedenti

Analisi di espressione genica su vasta scala:

- Microarrays
  1. Spotted – Probes: PCR products or long oligos — 2-colors relative meas.
  2. In situ synthesized:
     1. short – Affymetrix (photolitography, probeset) — 1 color absolute meas.
     2. longer (ink-jet technology) — 1 or 2 – colors
     3. arrayed beads (Illumina Bead-arrays) — 1 color absolute meas.

- Serial sequencing
  1. EST (cloning and serial sequencing of cDNA libraries)
  2. SAGE (short 3' tags concatamers from cDNA)
  3. Deep-sequencing

What we see:

Microarrays: since we use "probes" we obviously must know the sequences we are looking at !

Sequencing: theoretically, all the sequences that are represented in RNA are read, but:

EST – which primer in cDNA synthesis?

SAGE – short tags are identified (mapped to genome sequence) only if we consider 3'UTR only

1st problem: **Sensitivity:**

mRNA is 2-4 % of total RNA

1μg Tot RNA → 20-40 ng mRNA

Assuming that 10,000 genes are expressed, on average each mRNA species is 2-4 pg

The number of mRNA molecules/cell of individual genes ranges from 0 to some thousands.

i.e. , for some genes, we are measuring a **very low** number of mRNA molecules.

Sensitivity:

Microarrays

Problem: low-expressing genes are hardly seen, since we must subtract from the individual probe signal a median background fluorescence value.

Solution: amplifying the complex probe

Sequencing

Theoretically all expressed sequences (tags) can be identified, independently of their relative abundance. Practically, this is limited by the number of "reads" we make.

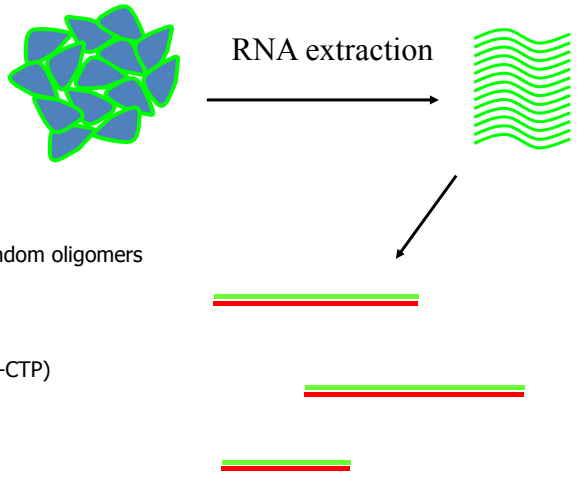
Problem: DNA fragments may contaminate mRNA preparations: increased sensitivity means also increased detection of (false) positives (FDR).

Microarrays:

Sample preparation

(sometimes the RNA sample is also called “complex probe”….not to be confused with the “probes” that are oligos or cDNA fixed to the chip surface…)

Direct labelling, no amplification

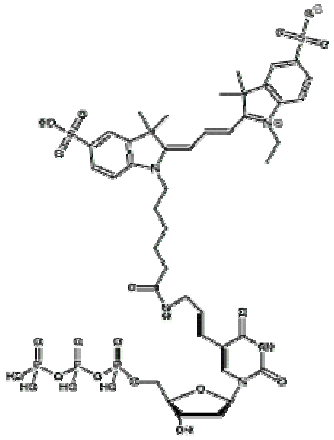


Primers: oligo(dT) or random oligomers

Reverse transcriptase

dNTPs

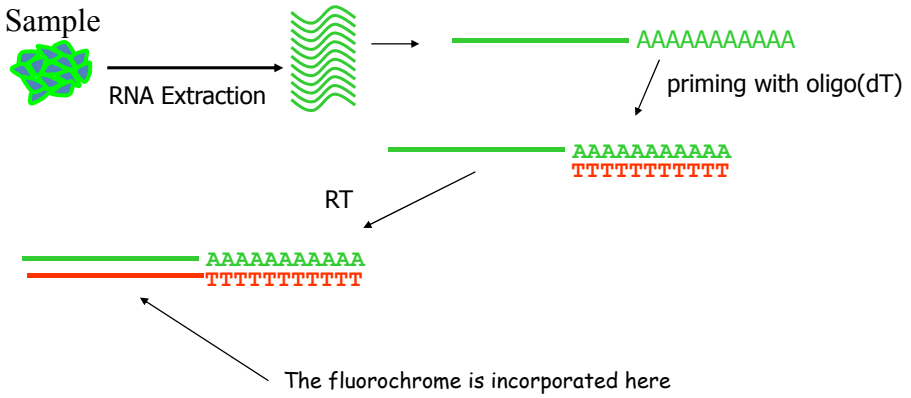
1 labelled NTP (e.g. Cy3-CTP)



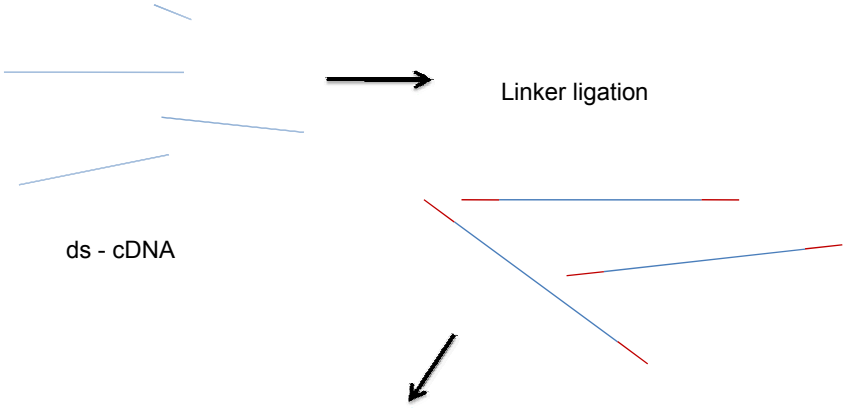
Cy3-dUTP

Nonamplified complex probe preparation:

RNAs are labelled using fluorochrome-conjugated ribonucleotides (NTP) directly by reverse transcription (RT), priming the synthesis either with oligo-dT or with random primers.

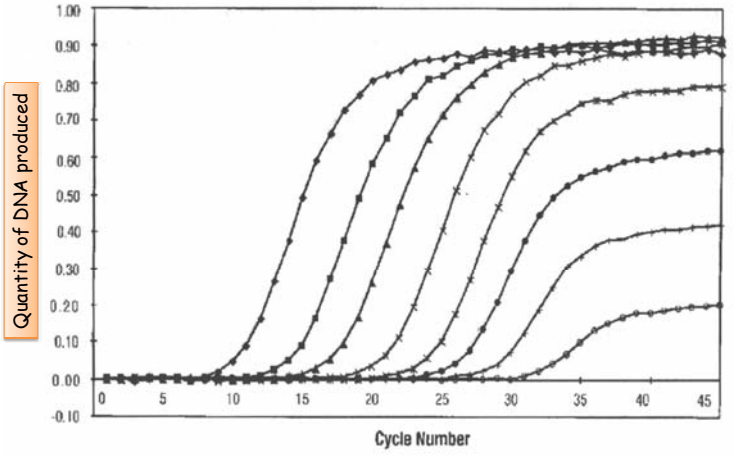


PCR amplification ?



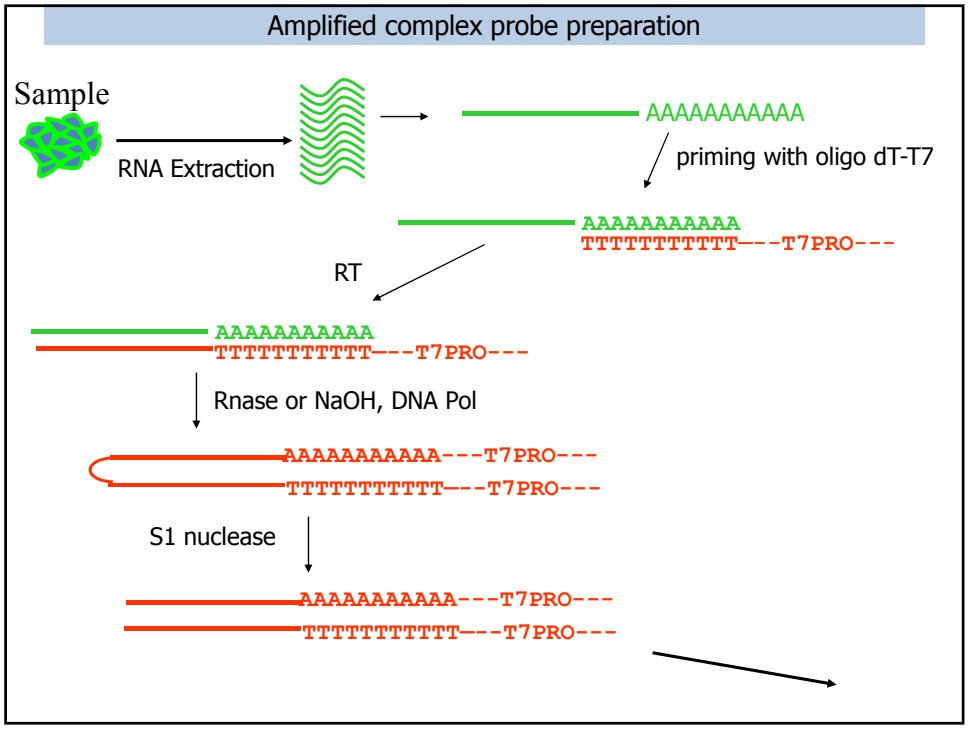
PCR amplification

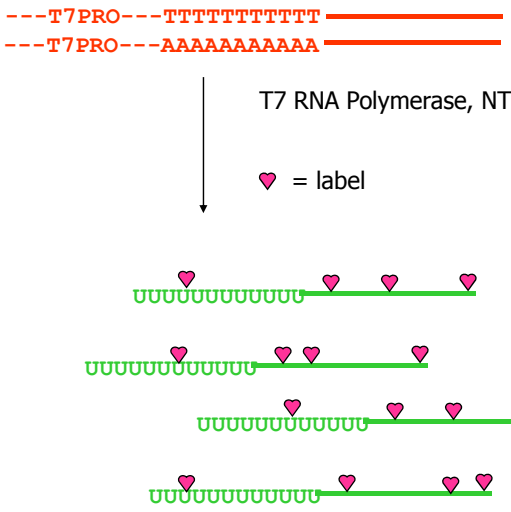
The exponential nature of PCR amplification, however, tends to alter the relative ratios of mRNA in a sample.



Much better a **linear** amplification protocol !

……synthesis of cRNA





Label may be a fluorochrome or a detectable modification, like biotin or digoxigenin, or a chemical group that can be conjugated with fluorochrome after transcription (e.g. allyl-UTP)

TRANSCRIPTION
Linear Amplification of each sequence that was originally present in starting RNA, but **complementary**
= "cRNA"

The problem of identifying genes in the genome

Using known cDNA → quite easy problem → BLAST

In 2001 there were relatively "few" genes known (10K); the number is incread today, but we still do not have a complete catalogue.

Bioinformatic approach → look for gene "features" in the genomic sequence.

EST approach → BLAST, then look around for gene "features"

a) Known genes
b) Predicted genes
c) Ab initio predicted genes

articles

Experimental annotation of the human genome using microarray technology

D. D. Shoemaker*, E. E. Schadt*, C. D. Armour, Y. D. He, P. Garrett-Engele, P. D. McDonagh, P. M. Loerch, A. Leonardson, P. Y. Lum, G. Cavet, L. F. Wu, S. J. Altschuler, S. Edwards, J. King, J. S. Tsang, G. Schimmack, J. M. Schelter, J. Koch, M. Ziman, M. J. Marton, B. Li, P. Cundiff, T. Ward, J. Castle, M. Krolewski, M. R. Meyer, M. Mao, J. Burchard, M. J. Kidd, H. Dai, J. W. Phillips, P. S. Linsley, R. Stoughton, S. Scherer & M. S. Boguski

Rosetta Inpharmatics, Inc., 12040 115th Avenue N.E., Kirkland, Washington 98034, USA
* *These authors contributed equally to this work*

The most important product of the sequencing of a genome is a complete, accurate catalogue of genes and their products, primarily messenger RNA transcripts and their cognate proteins. Such a catalogue cannot be constructed by computational annotation alone; it requires experimental validation on a genome scale. Using 'exon' and 'tiling' arrays fabricated by ink-jet oligonucleotide synthesis, we devised an experimental approach to validate and refine computational gene predictions and define full-length transcripts on the basis of co-regulated expression of their exons. These methods can provide more accurate gene numbers and allow the detection of mRNA splice variants and identification of the tissue- and disease-specific conditions under which genes are expressed. We apply our technique to chromosome 22q under 69 experimental condition pairs, and to the entire human genome under two experimental conditions. We discuss implications for more comprehensive, consistent and reliable genome annotation, more efficient, full-length complementary DNA cloning strategies and application to complex diseases.

Exonic arrays

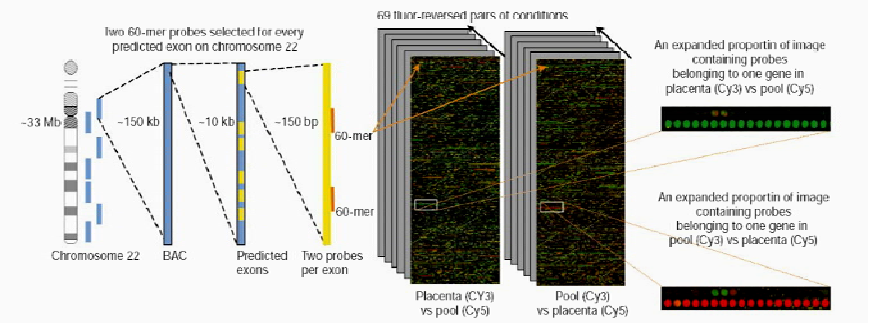


Figure 1 Design and fabrication of exon arrays for the predicted exons on human chromosome 22. Two 60-mers were selected from each of 8,183 predicted exons on human chromosome 22q and printed on a single 1 x 3 inch array (~25,000 60-mers). This array was hybridized with 69 pairs of RNA samples using a two-colour hybridization technique. Each experiment was performed in duplicate with a fluor reversal to minimize possible bias caused by the molecular structure of the Cy3 and Cy5 dyes (138 arrays in total). Red and green spots, as shown in the expanded panels on the right, are probes representing experimentally verified genes (groups of differentially expressed exons that are located next to each other in the genome).

To solve these problems, several approaches were developed:

1. Tiling arrays
2. CAGE (5'-CAP-linked serial Gene Expression)
3. Deep-sequencing (re-sequencing)

Review TRENDS in Genetics Vol.21 No.2 February 2005

Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments

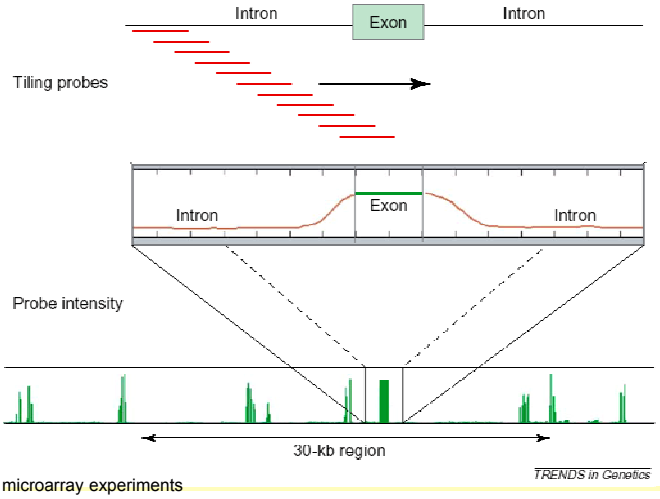
Jason M. Johnson[1], Stephen Edwards[1], Daniel Shoemaker[2] and Eric E. Schadt[1]

[1]Rosetta Inpharmatics LLC*, 401 Terry Avenue North, Seattle, WA 98109, USA
[2]GHC Technologies, 505 Coast Boulevard South, Suite 309, La Jolla, CA 92037, USA

Microarrays provide the opportunity to measure transcription from regions of the genome without bias towards the location of known genes. This technology thus offers an important source of genomic sequence annotation that is complementary to cDNA sequencing and computational gene-finding methods. Recent 'tiling' microarray experiments that assay transcription at regular intervals throughout the genome have shown evidence of large amounts of transcription outside the boundaries of known genes. This transcription is observed in polyadenylated RNA samples and appears to be derived from intergenic regions, from introns of known genes and from sequences antisense to known transcripts. In this article, we discuss different explanations for this phenomenon.

review

This paper is part of the course



Box 1. Tiling microarray experiments

Tiling microarrays are designed to assay transcription at regular intervals of the genome using regularly spaced probes (horizontal red lines) that can be overlapping (Figure I) or separated. The distance between the centers of successive probes is the 'step' size and probes can be selected to be complementary to one strand (as shown) or both strands. Probes can be synthesized directly onto or spotted onto glass slides, and can be synthesized oligonucleotides or PCR products. They are hybridized with fluorescently labeled cRNA or cDNA prepared from cell samples. Regions of greater fluorescent intensity (green peaks in lower panel) can reveal transcription within a large genomic region. In addition, the correlation of probe intensities in several different tissues (co-expression analysis) can be used to identify probes that are detecting exons of the same transcript. The lower panel shows the extent of a hypothetical transcript within the genome. The middle panel is a schematic, magnified view of the hybridization of a genomic region containing an exon.

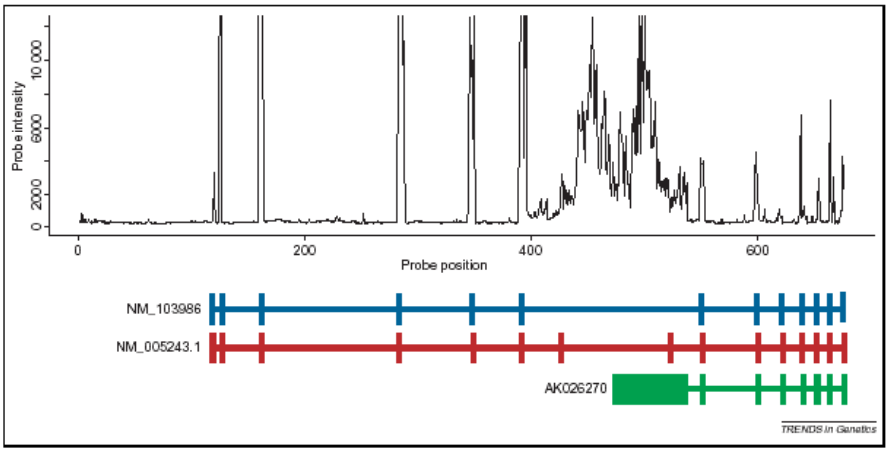


Figure 1. Microarray intensity profile for human thymus poly(A)$^+$ cDNA profiled on a 60mer ink-jet tiling array representing the genomic locus of the Ewing sarcoma breakpoint region 1 gene (*EWSR1*) in 30-nt steps, with probe index as the x-axis. No probes are shown for repeat-masked regions. The tiling data for this locus are shown in relation to the exon positions (below the plot) of three EWSR1 cDNAs (Genbank accession numbers: NM_013986, NM_005243.1 and AK026270). The 5′-most exon lies in a repeat-masked region and is not shown. A few peaks with the highest intensity have been truncated in Figures 1–3.

Tiling microarray analysis of RNA from thalamus, testes and uterus (mouse)

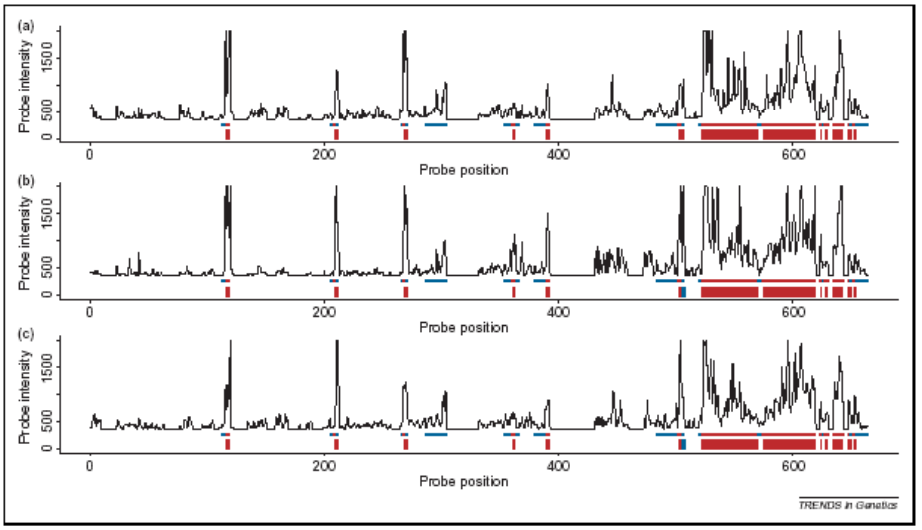


Figure 2. Microarray tiling confirmation of a predicted gene. Microarray tiling probe intensities for a region of human chromosome 20 that contains an *ab initio* gene prediction (made by the program GENSCAN [48]). Predicted exons for this gene are shown with blue lines. The transcription detected by microarrays has been grouped into a single transcriptional unit (dark red) by the correlated behavior of these probes across different human samples [11]. The conditions displayed are **(a)** thalamus, **(b)** testes and **(c)** uterus.

Perspective

What is a gene, post-ENCODE? History and updated definition

Mark B. Gerstein,[1,2,3,9] Can Bruce,[2,4] Joel S. Rozowsky,[2] Deyou Zheng,[2] Jiang Du,[3] Jan O. Korbel,[2,5] Olof Emanuelsson,[6] Zhengdong D. Zhang,[2] Sherman Weissman,[7] and Michael Snyder[2,8]

[1]Program in Computational Biology & Bioinformatics, Yale University, New Haven, Connecticut 06511, USA; [2]Molecular Biophysics & Biochemistry Department, Yale University, New Haven, Connecticut 06511, USA; [3]Computer Science Department, Yale University, New Haven, Connecticut 06511, USA; [4]Center for Medical Informatics, Yale University, New Haven, Connecticut 06511, USA; [5]European Molecular Biology Laboratory, 69117 Heidelberg, Germany; [6]Stockholm Bioinformatics Center, Albanova University Center, Stockholm University, SE-10691 Stockholm, Sweden; [7]Genetics Department, Yale University, New Haven, Connecticut 06511, USA; [8]Molecular, Cellular, & Developmental Biology Department, Yale University, New Haven, Connecticut 06511, USA

While sequencing of the human genome surprised us with how many protein-coding genes there are, it did not fundamentally change our perspective on what a gene is. In contrast, the complex patterns of dispersed regulation and pervasive transcription uncovered by the ENCODE project, together with non-genic conservation and the abundance of noncoding RNA genes, have challenged the notion of the gene. To illustrate this, we review the evolution of operational definitions of a gene over the past century—from the abstract elements of heredity of Mendel and Morgan to the present-day ORFs enumerated in the sequence databanks. We then summarize the current ENCODE findings and provide a computational metaphor for the complexity. Finally, we propose a tentative update to the definition of a gene: A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products. Our definition sidesteps the complexities of regulation and transcription by removing the former altogether from the definition and arguing that final, functional gene products (rather than intermediate transcripts) should be used to group together entities associated with a single gene. It also manifests how integral the concept of biological function is in defining genes.

This paper is part of the course

Deep-sequencing

or

mass sequencing

Developed for DNA resequencing, but applied to RNA analysis

"RNA-Seq"

Gene expression by deep-sequencing

All genome can be re-sequenced starting from known primers by Solexa® - Illumina technology or other competitor technology.

The platform uses Solid-phase sequencing (see http://www.genetic-inference.co.uk/blog/?p=413)

RNA application (RNA-Seq)

1. RNA extraction from cells or tissues
2. cDNA synthesis from oligo(dT) primers
3. ds cDNA tags ligated to A and B primers
4. all cDNA re-sequenced on a chip (microarray) containing millions of spot with complementary aA and bB primers
5. sequence is read → table of frequency of each match

Ilumina Solexa deep sequencer

http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf