DNA Chips and Microarray Analysis

– An Overview

Sangdun Choi

Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125, USA

Tel: (626) 395-8732

Fax: (626) 796-7066

E-mail: **schoi@caltech.edu**

Running title: DNA microarrays and their analyses

1. INTRODUCTION

DNA microarray technology has given rise to the study of functional genomics (Brown and Botstein 1999; Lockhart and Winzeler 2000). The entire set of genes of an organism can be microarrayed on an area as small as a fingernail and the expression levels of thousands of genes are simultaneously studied in a single experiment. DNA microarray technology allows comparisons of gene expression levels on a genomic scale in all kinds of combinations of samples derived from normal and diseased tissues, treated and nontreated time courses, and different stages of differentiation or development. Further computational analysis of microarray data allows the classification of known or unknown genes by their mRNA expression patterns. Global gene expression profiles in cells or tissues will provide us with a better understanding of the molecular basis of phenotype, pathology, or treatment. This review article presents an introduction to the DNA microarray technology and its analysis.

2. MICROARRAY FABRICATION

On a glass surface, complementary DNAs (cDNAs) or gene-specific oligonucleotides can be spotted. Oligonucleotides can also be synthesized *in situ*.

2.1. cDNA or pre-synthesized oligonucleotide deposited arrays

In deposited arrays, which were developed by Pat Brown lab in Stanford University (Schena *et al*. 1995; the general guide is well described in the web site of http://cmgm.stanford.edu/pbrown/mguide/index.html), a high-speed spotting robot is used to affix PCR-amplified cDNA or synthesized oligonucleotides onto a chemically

modified glass slide. Polylysine or aminosilane is commonly used as a substrate for coating the surface of these glass slides. The DNA arrayed slides are then hybridized with fluorescently labeled cDNAs reverse-transcribed from mRNA populations. During this process, the slide is hybridized with two different cDNA samples labeled separately with two distinct fluorescent dyes, such as Cy3 and Cy5 (two-color hybridization). The relative intensities of the two fluorescent dyes within a spot represent the relative mRNA expression levels of the gene. For example, if fluorescent labels Cy3 (green) and Cy5 (red) are used to make each sample's cDNA probe, the expression level of a gene will be displayed as green or red when the gene is differentially expressed, or yellow when the level is the same in the two samples (Fig.17.1).

Mechanical microspotting allows transfer of premade substances onto a solid surface. The recently developed inkjet printing method (Agilent, Palo Alto, CA) produces more uniform spots (Fig.17.1A), which had been difficult to achieve with available pin spotting techniques.

2.2. *in situ* synthesized oligonucleotide arrays

This method was developed by Fodor *et al*. (1991) and it is now referred to as the Affymetrix GeneChip system (Affymetrix Inc., Santa Clara, CA). In these oligonucleotide arrays, DNA oligonucleotides are synthesized *in situ* onto the DNA chip using photolabile protecting groups and photolithographic masks to add the selective sequences of nucleotides (Lipshutz *et al*. 1999). The recent version of GeneChip contains 400,000 of 25-base-pair (bp) oligonucleotides where twenty different oligonucleotide pairs represent one gene. A pair consists of a 'perfect-match' and a 'mismatch'

oligonucleotide, in which the 13th nucleotide is mutated. The ‘perfect match’ signals are subtracted by 'mismatch' signals, and the net values are used for the comparisons. In the Affymetrix GeneChip system, in contrast to the cDNA or pre-synthesized oligonucleotide deposited arrays, only one sample is hybridized on to one array and comparisons can be made among multiple arrays (one-color hybridization) (Fig.17.2).

2.3. cDNA vs. oligonucleotide as a target DNA

For reference, the term "probe" refers to the fluorescently labeled DNAs or RNAs and the term "target" refers to the DNAs on the slide. The definitions are reversed in some articles, especially when the Affymetrix GeneChip system is mentioned.

The oligonucleotide targets require only the sequence information of genes, and thereby can maximally exploit the genome sequences of the organism. The precise fold changes from cDNA or oligonucleotide microarrays are not identical, but the general trends of changes and the fold differences are similar to each other when the same genes are compared in both array experiments. In oligonucleotide arrays, the design of oligonucleotides is very critical to the results and they should be designed very carefully. In cDNA arrays, there may be potentially confounding effects of cross-hybridization due to sequence homologies among members of a gene family. When oligonucleotides are used, they can be specifically designed to differentiate between highly homologous members of a gene family as well as alternatively spliced forms of the same gene (exon-specific). Oligonucleotide arrays can also be designed to allow detection of mutations and single nucleotide polymorphisms.

3. HYBRIDIZATION

In two-color hybridization, fluorescently labeled cDNA probes, directly reverse-transcribed from two separate mRNA populations, are hybridized onto the target DNA fixed on the slide. One cDNA sample is labeled with Cy3 fluorescent dye, and the other sample with Cy5. After hybridization and washing on a single array, the relative level in mRNA transcript for each gene is measured by the Cy5/Cy3 color ratio using a confocal laser scanner.

3.1. Preparing probes

3.1.1. Labeling from total RNA or poly(A) mRNA

Total RNA or poly(A) mRNA can be used as a template for microarray analysis. An array generally requires 2-200 μg of total RNA or 0.2-2 μg of poly(A) mRNA as a starting material. These values are adjusted depending on the number of elements, area of hybridization, method of labeling, etc. When poly(A) RNAs are reverse-transcribed, Cy3- or Cy5-linked dNTP is added to generate the final cDNA probe.

A minimum of 0.2 μg poly(A) mRNA or 5 μg total RNA is recommended to start the Affymetrix cDNA synthesis protocol. Basically, the Affymetrix system uses *in vitro* transcription after making double stranded cDNA from poly(A) RNA. During *in vitro* transcription of the complimentary RNA (or linear amplification: details in 3.1.2.), biotin-labeled NTPs are incorporated, probes are hybridized onto the GeneChip, and Streptavidin-Phycoerythrin is used to stain the biotin-bound targets. Phycoerythrin is detected in the wavelength of 570 nm by an appropriate scanner.

3.1.2. Amplification

Direct labeling of cDNA microarray usually requires about 100 μg of total RNA or 1 μg of mRNA to determine an expression profile. This becomes a problem when only a small amount of RNA is available for analysis. In this case, optimized linear amplification methods can be performed to increase the number of transcript messages (Fig.17.3). One example of this method is using a primer containing the T7 promoter sequence fused to an oligo-dT sequence to generate double strand cDNAs from the template mRNA. Antisense RNAs (aRNA) are then linearly amplified from cDNAs using T7 RNA polymerase. A portion of the amplified aRNAs is now reverse transcribed in the presence of Cy-3-dNTP or Cy-5-dNTP to generate fluorescently labeled cDNAs. The expression results obtained from linear amplified aRNA are comparable to those from the original mRNA.

3.2. Laser capture microdissection

Integration of laser capture microdissection to microarray technology (Luo *et al*. 1999) can potentially provide substantially higher homogeneity in samples by determining with microscopic visualization. It is very useful to define the gene expression profile within a small group of a specific cell population. A cell typically produces 10-30 pg of total RNA. An amplification technique is usually needed to generate a sufficient amount of probe from small but homogeneous cell populations for microarray hybridizations. Faithfully amplified cellular RNAs from a few (or even single) cells isolated by laser capture microdissection will reach the ultimate sensitivity of

the smallest functional tissue unit. This will eventually lead to a cellular-level molecular understanding of the functional processes.

3.3. Scanning

Hybridized slides are scanned with a confocal laser scanner, such as GenePix 4000A scanner (Axon Instruments, Union City, CA), ScanArray (GSI Lumonics, Farmington Hills, MI), Agilent Scanner (Agilent, Palo Alto, CA), *etc*. The majority of the confocal laser microarray scanners use photomultiplier tubes (PMTs) as detectors. PMTs can detect fluorescence intensities in the visible wavelength range. By varying the control voltage, the sensitivity of the detector can be modified easily. The acquired images are analyzed with one of the available softwares, such as GenePix Pro3.0 (Axon), Image Analysis (Agilent), *etc*.

4. DATA ANALYSIS

4.1. Normalization

4.1.1. Total intensity normalization

The total intensity normalization method stands on the assumption that the total quantities of messages from both samples are the same. Under this assumption, a normalization factor can be calculated from the total integrated intensity (in one color hybridization, for example, in Affymetrix GeneChip system) or from the total average fold difference of the Cy3 and Cy5 channels (in two color hybridization, for example, in deposited cDNA arrays) for all the elements in one array. This normalization factor is then used to adjust the scale or fold for each gene in the array.

4.1.2. Regression normalization

In a scatter plot of both channels in the two-color hybridization, the genes would scatter along a diagonal straight line when closely related two samples are compared. Normalization of this data can be performed by calculating the best-fit slope and by applying the regression to adjust the levels of all the genes. Adjustment using local regression is more suitable in cases where the fold differences are nonlinear.

4.1.3. Normalization using ratio statistics

This method assumes that a subset of genes, referred to as 'housekeeping genes', do not change their profiles throughout the experiments. The normalization factor calculated from this subset of housekeeping genes is used to adjust experimental variability in the samples being compared. Alternatively, a set of exogenous controls can be spiked onto the arrays and mRNAs from the set are equally added into the initial RNA samples before labeling. The average expression ratio from these controls should be equal to one and this factor is used to normalize the data to identify differentially expressed genes.

4.1.4. Universal standard

Great advantages arise in the scientific community when research data is shared between laboratories and experiments. Unfortunately, microarray data is not easily shared due to the variation of standards among experiments. The need for researchers to agree on one particular standard, referred to as the universal standard, is very difficult to

achieve. Thus, ongoing efforts to find a common standard sample for all experiments are in progress to facilitate widespread data sharing.

There are also some critical caveats in using of a universal standard: 1) When everyone attempts to prepare a large batch of the reference sufficient for the entire project, the reproducibility among the standards cannot be assured. 2) Standards as denominators below the threshold of accurate measurement must be avoided. 3) An alternative, much discussed in the array community, is a pool of about 10 cell lines. The universal human or mouse reference RNA isolated from 10 (human) or 11 (mouse) cell lines representing different tissues is available from Stratagene (La Jolla, CA). The idea is that this will cover most genes, and that with fresh culture, although there might be differences in some cell lines, the effects would be smoothed out over the entire group.

Still there may be some problems associated with universal standards. For example, you cannot hope to obtain better data than the system will produce under idealized conditions. Whenever possible, the use of a direct comparison - the same cells split into two different conditions and then co-hybridized - will yield the most accurate results, particularly for small induction/repression levels. However, a universal standard, if possible, is the best way overall to ensure consistency for large data sets.

4.2. Clustering

DNA microarray experiments generate unprecedented quantities of genome-wide data which can greatly overwhelm biologists. To extract useful information from expression profiles, computational tools that cluster and display data can be used. Although there are many ways to analyze gene expression data, hierarchical clustering

(Eisen *et al*. 1998) and self-organizing map clustering (Tamayo *et al*. 1999) have been widely used to display the data.

Hierarchical clustering is simple and the results are easily visualized (Fig.17.4). In hierarchical clustering, the distances between genes are calculated for all of the genes based on their expression pattern and the closer genes are merged to produce a cluster. The distances between these small clusters are calculated to produce a new cluster.

Self-organizing map (SOM) clustering assigns genes to a series of groups on the basis of expression pattern similarities. Random vectors are constructed for each group and a gene is assigned to the closest vector.

5. APPLICATIONS OF MICROARRAYS

Traditional molecular research tools for gene expression study are limited to a small group of genes at a time. Recent advances in microarray field have enabled the study of large numbers of genes in a single experiment. DNA microarrays not only detect global changes of gene expression, but also have many other potential applications including the identification of gene copies in a genome (Pinkel *et al*. 1998), mutation and polymorphism detection (Wang *et al*. 1998), sequencing, diagnostic tools for diseases, and drug discovery.

6. BEYOND THE DNA MICROARRAYS

The question in dealing with the current humongous amount of microarray data is whether we are ready to decode these messages. Very often, the array data might not apply well to our previously accumulated knowledge. This is due to the fact that data

obtained from expression profiling are too complex to interpret and for a long time we have been used to studying one gene or one system at a time. Although the array data on a genomic scale is significantly insightful in understanding the mechanisms in biological systems, mRNA profiling provides us with only the levels of mRNA messages. The problem of biological interpretation of gene expression data occurs when cellular events are mediated in protein levels. In addition, the current array data include the transcriptional behaviors of large portion of yet-uncharacterized genes.

Recently NCBI has launched the Gene Expression Omnibus (GEO) site (http://www.ncbi.nlm.nih.gov/geo/), allowing public access to a broad range of gene expression data. GEO is not only a gene expression data repository but also an online resource for the retrieval of gene expression data from any organism and among laboratories. The meaningful use of these and other array data will require systematic computational analyses and cross-comparisons with other data sets obtained from different experimental systems, such as two-dimensional gel electrophoresis of proteins, yeast two-hybrid system, and other conventional molecular biological tools.

7. CONCLUSIONS

By the time of publication of this article, it is expected that the whole set of human or mouse genes will be available in a chip. Since DNA microarrays have the possibility of incorporating thousands of genes, it would not be impossible to scan the whole genome of a particular organism in one experiment. Thus, these techniques will allow the complete comparison of almost all transcribed genes' expression levels.

The influence of microarray technology has been powerful in both basic and applied biology. Unmanageable amounts of microarray data have out-paced and even stimulated the development of a new science area, bioinformatics. One important goal of computational analysis is to extract clues from microarray data and translate the information into biological understanding. Systematic analysis of microarray data will yield insight into molecular biological processes and functions of thousands of gene products in parallel. This approach allows for improved understandings in cellular signaling, disease classification, diagnosis, prognosis, and drug design. To pinpoint specific genes as research targets, such as drug targets, traditional methods are still required. What microarrays do best is high-throughput screening.

The future of plant microarrays will diverge from *Arabidopsis* (Finkelstei *et al*. 2002) to agriculturally important plants, such as maize, rice, soybean, *etc*. When the genome sequencing is completed, full genome microarrays can also be constructed. By using microarrays, disease resistance, cold or drought response, and product yield can be explored on a genomic scale and comprehensive dynamic view of what happens in the molecular level.

Proteome technologies for monitoring changes in protein abundance and protein modification are important because the correlation between gene and protein expression is variable, and the post-translational protein modifications are responsible for realizing the signaling and information processing. Tissue microarrays (Kononen *et al*. 1998) and protein microarrays (MacBeath and Schreiber 2000; Zhu *et al*, 2001) have been developed where samples from up to hundreds of tissues or proteins are analyzed simultaneously on one glass slide.

As with DNA microarray, orchestrated and genome-wide gene expression studies can be efficiently carried out, and these explorations will provide clues for new hypotheses. Microarray technologies will accelerate scientific discoveries and become one of the key technologies in biology.

ACKNOWLEDGMENTS

I thank Anna Cao and Becky Hart for editing the manuscript and Alex Hoffman for expertise and advice on arrays.

REFERENCES

Brown PO and Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nature Genet 21; 33-37.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863-14868.

Finkelstei D, Ewing R, Gollub J, Sterky F, Cherry JM, and Somerville S (2002) Microarray data quality analysis: lessons from the AFGC project. Arabidopsis Functional Genomics Consortium. Plant Mol Biol 48:119-131.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D (1991). Light-directed, spatially addressable parallel chemical synthesis. Science 251:767-773.

Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med. 4:844-847.

Lipshutz RJ, Fodor SP, Gingeras TR, and Lockhart DJ (1999). High density synthetic oligonucleotide arrays. Nat Genet 21:20-24.

Lockhart DJ and Winzeler EA (2000). Genomics, gene expression and DNA arrays. Nature 405:827-836.

Luo L, Salunga RC, Guo H, Bittner A, Joy KC, Galindo JE, Xiao H, Rogers KE, Wan JS, Jackson MR, and Erlander MG (1999). Gene expression profiles of laser-captured adjacent neuronal subtypes. Nat Med. 5:117-122.

MacBeath G and Schreiber S (2000). Printing proteins as microarrays for high-throughput function determination. Science 289:1760-1763

Pinkel D, Segraves R, Sudar S, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, and Zhai Z (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 20:207-211.

Schena M, Shalon D, Davis RW, and Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470.

Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907-2912.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E,

Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, and Lander ES (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280:1077-1082.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, and Snyder M (2001) Global analysis of protein activities using proteome chips. Science 293:2101-2105.

Legends

Fig.17.1. Deposited cDNA microarrays cohybridized with fluorescently labeled reference (Cy3) and test (Cy5) samples. The hybridization output is measured by the relative intensities of the two colors, red (Cy5) and green (Cy3). Yellow spots indicate equal amounts of expressed mRNA from each sample, red indicating higher expression in the test sample and green lower expression. A. Inkjet spotted cDNA array. B. Pin spotted cDNA array.

Fig.17.2. Affymetrix GeneChip image.

Fig.17.3. Linear amplification method.

Fig.17.4. Hierarchical clustering of gene expression profiles: The column lists the time point studied. The row corresponds to individual gene surveyed. Upregulated genes appear in red, and downregulated genes appear in green, with the relative $\log_2$ ratio reflected by the intensity of the color.

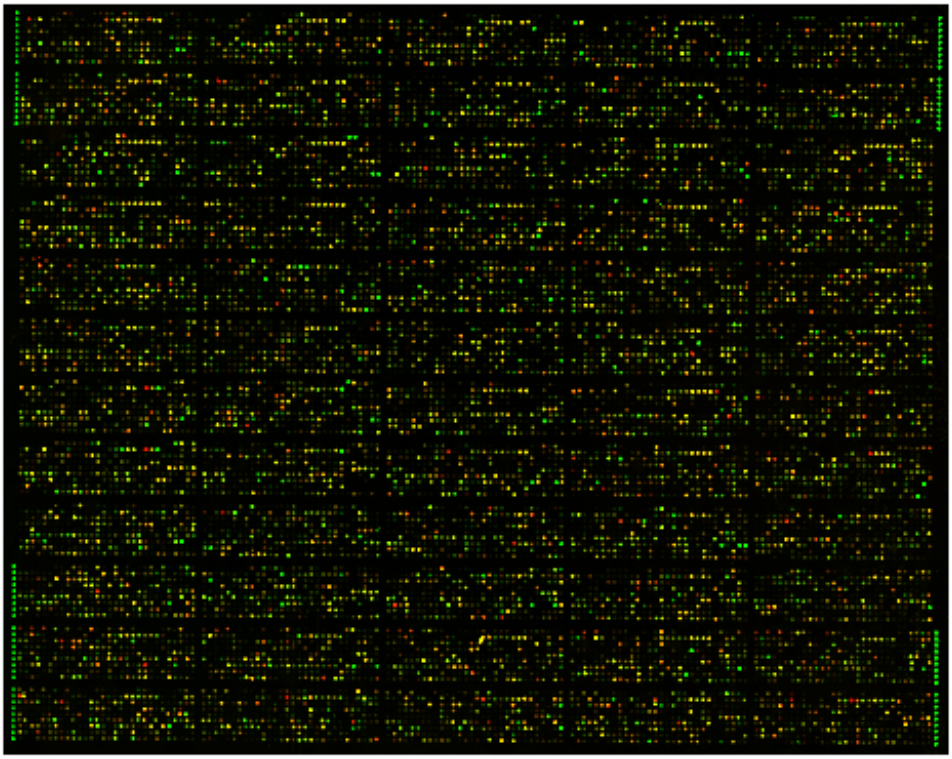


Fig.17.1A.

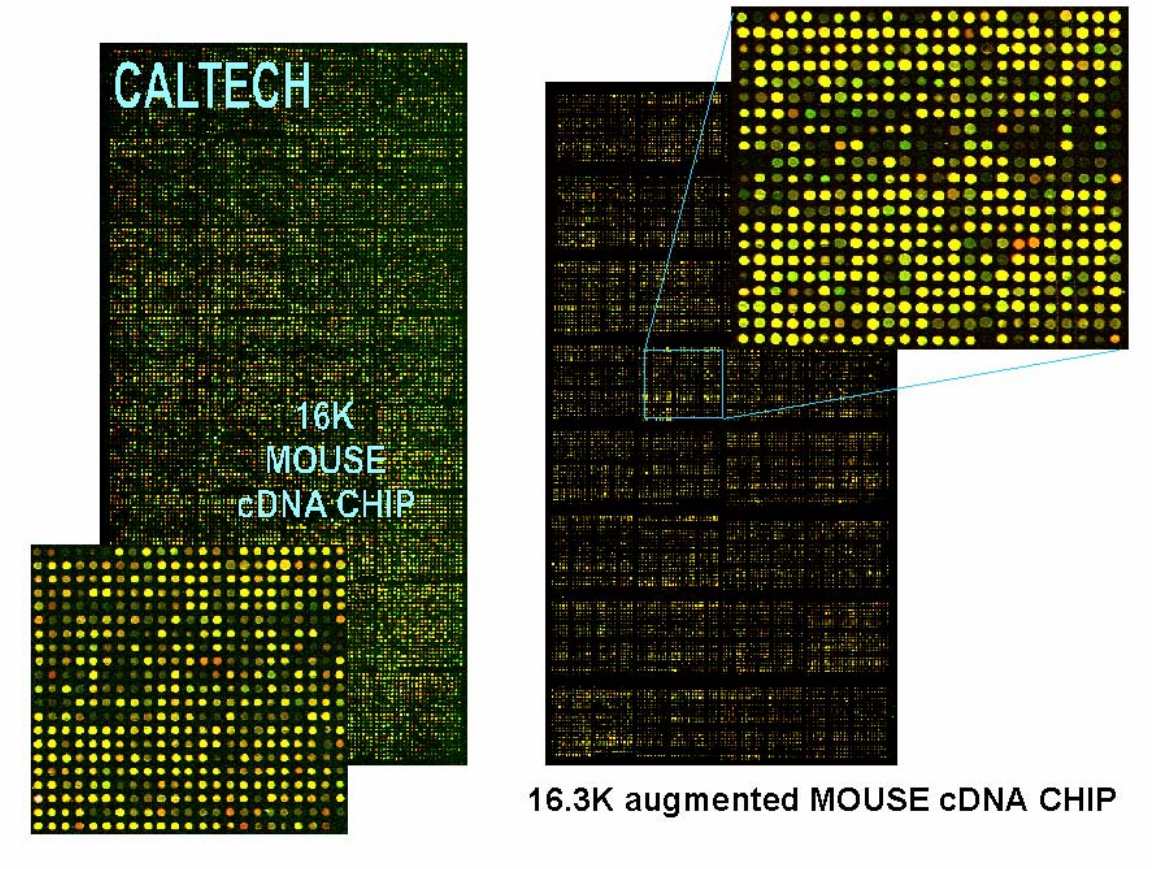


Fig.17.1B.

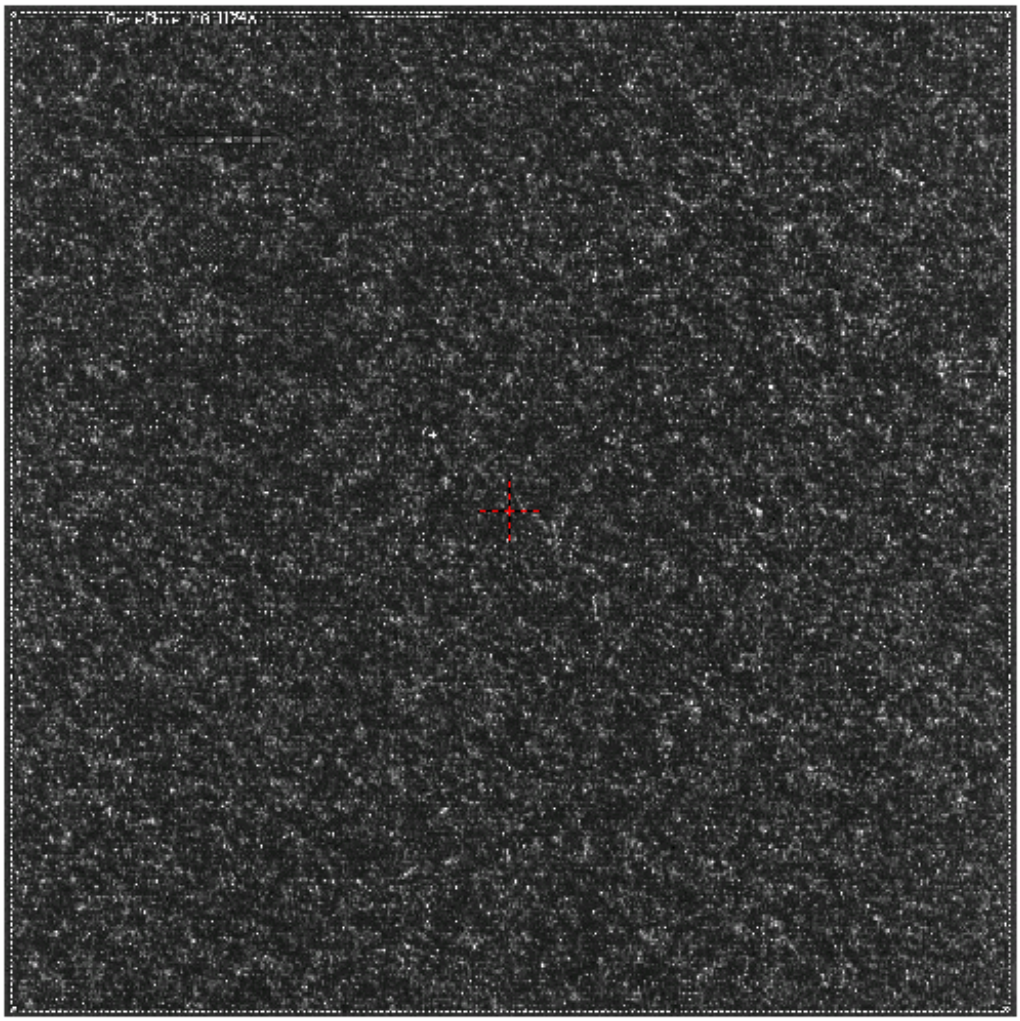
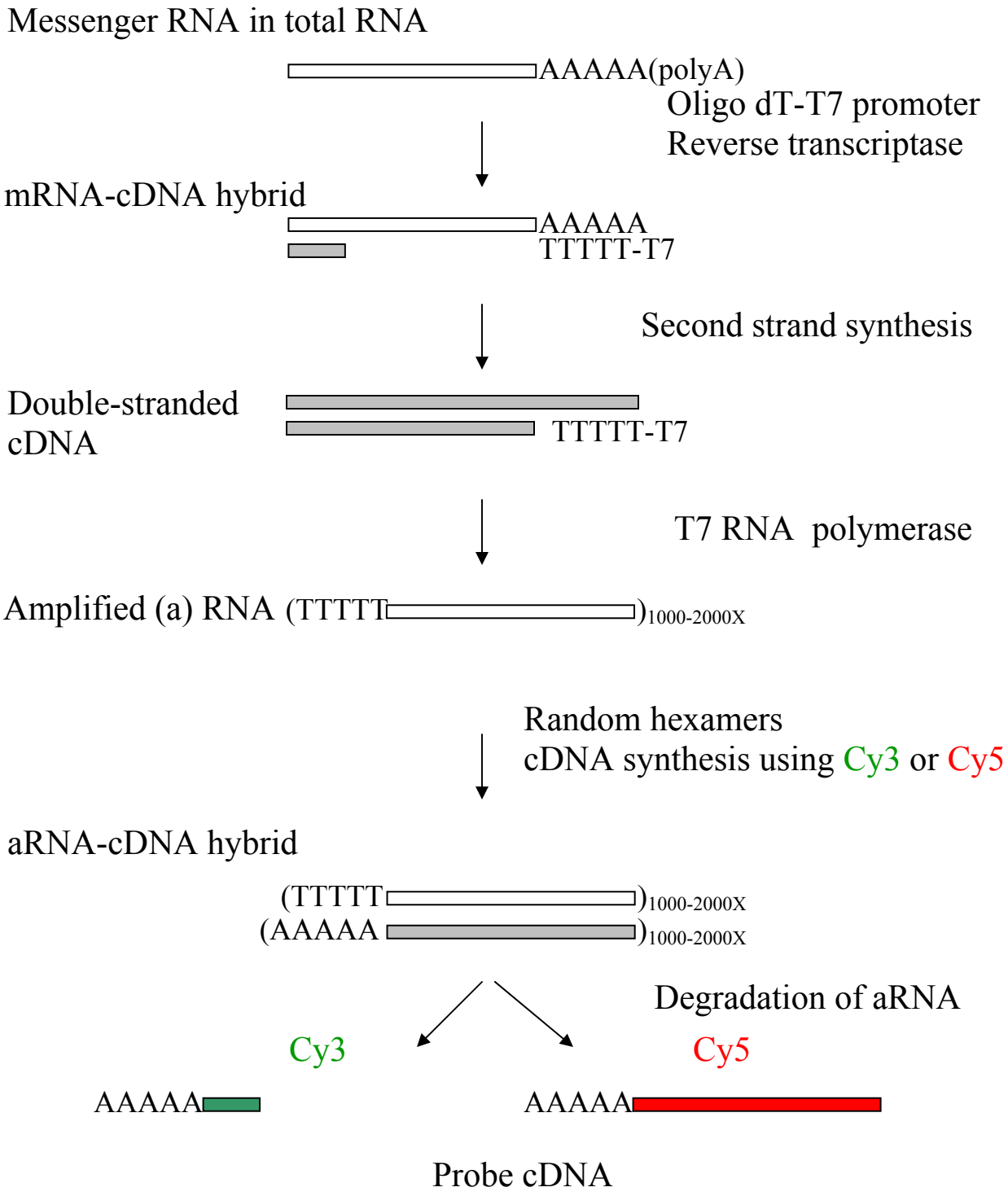


Fig.17.2.

Fig.17.3.


Messenger RNA in total RNA
AAAAA(polyA)
Oligo dT-T7 promoter
Reverse transcriptase
mRNA-cDNA hybrid
AAAAA
TTTTT-T7
Second strand synthesis
Double-stranded cDNA
TTTTT-T7
T7 RNA polymerase
Amplified (a) RNA (TTTTT)1000-2000X
Random hexamers
cDNA synthesis using Cy3 or Cy5
aRNA-cDNA hybrid
(TTTTT)1000-2000X
(AAAAA)1000-2000X
Degradation of aRNA
Cy3
Cy5
AAAAA
AAAAA
Probe cDNA

Apoptosis

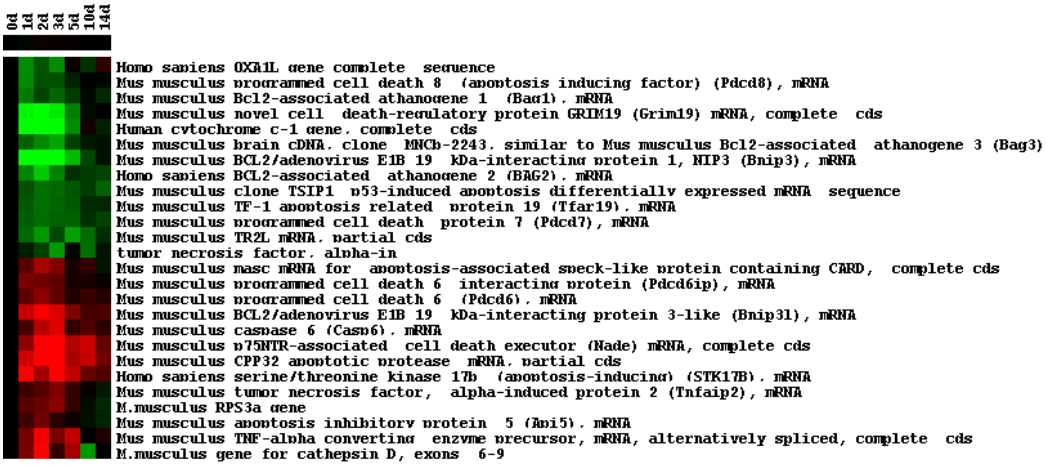


Fig.17.4.