

# **FUNCTIONAL GENOMICS**

**PRESENTED BY:**

**AJAY**

**M.Sc. Biotechnology (P)**

**Roll no. 4**

# INTRODUCTION

- **Genomics** - It is the study of genomes.
- The field of genomics comprises of two main areas:
  1. Structural genomics
  2. Functional genomics
- **Structural genomics** - deals with genome structures with a focus on the study of genome mapping and assembly as well as genome annotation and comparison.

# Functional genomics

- It is largely experiment based with a focus on gene functions at the whole genome level using high throughput approaches.
- The high throughput analysis of all expressed genes is also termed **transcriptome analysis**
- Transcriptome analysis can be conducted by two approaches:
  - 1) sequence based approaches
  - 2) microarray based approaches

# Sequence based approaches

- **Expressed sequence tags** :- ESTs are short sequences of cDNA typically 200-400 nucleotides in length.
- Obtained from either 5' end or 3' end of cDNA inserts of cDNA library.

# ADVANTAGES OF E.S.T

- **Provide a rough estimate of genes** that are actively expressed in a genome under a particular physiological condition.
- Help in **discovering new genes**, due to random sequencing of cDNA clones.
- EST libraries can be **easily generated**.

# DRAWBACKS OF USING E.S.Ts

- Automatically generated without verification thus contain **high error rates**.
- There is often **contamination** by vector sequence , introns, ribosomal RNA, mitochondrial RNA.
- **Weakly expressed genes** are hardly found in EST sequencing survey.
- ESTs represent only **partial sequences** of genes.

# E.S.Ts DATABASE

- **dbEST** is a EST database provided by GenBank. That contains EST collections for a large number of organisms.
- The database is **regularly updated**.

# EST INDEX CONSTRUCTION

- Goal of EST database – organize and combine the EST data to improve the quality of sequence information.

**“A collection of nonredundant and annotated EST sequence is known as gene index construction”**

- main advantage – **reduces redundancy.**



# STEPS INVOLVED

Remove vector contaminants and masks repeats using **Vecscreen**

**Clustering** – associates EST sequences with unique genes

Derive **consensus sequences** by fusing redundant, overlapping ESTs to correct errors

Results in **longer contigs**

Coding region is defined by excluding introns and 3'- untranslated sequences.

# Contd...

coding sequence **translated** into **protein** sequence for db similarity searching.

Alignment of these compiled ESTs with genomic sequence, performed by **program SIM4**

➔ **Major EST index sequence databases are:**

- **Unigene** – is an **NCBI** EST cluster database.
- **TIGR Gene indices**

# SAGE

- **Serial analysis of gene expression**
- Another high throughput, Sequence-based approach for gene expression profile analysis.
- **More quantitative** in determining mRNA expression in cell.
- Short fragments (**taqs**) of DNA, excised from cDNA sequences , act as unique markers of gene transcript
- SAGE invented at **Johns Hopkins University** in USA (Oncology Center) by **Dr. Victor** in 1995.

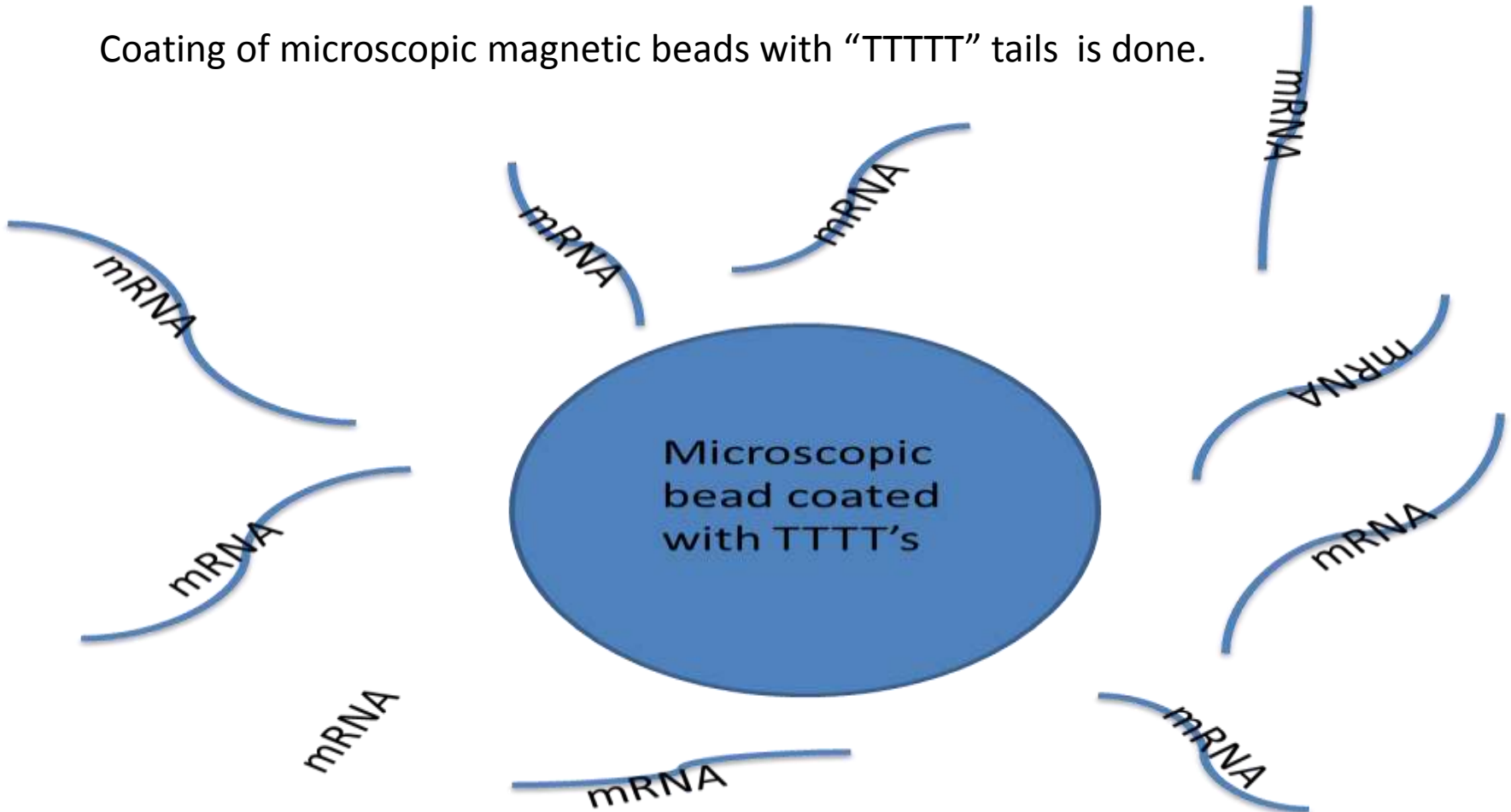
# Principle Underlining SAGE methodology

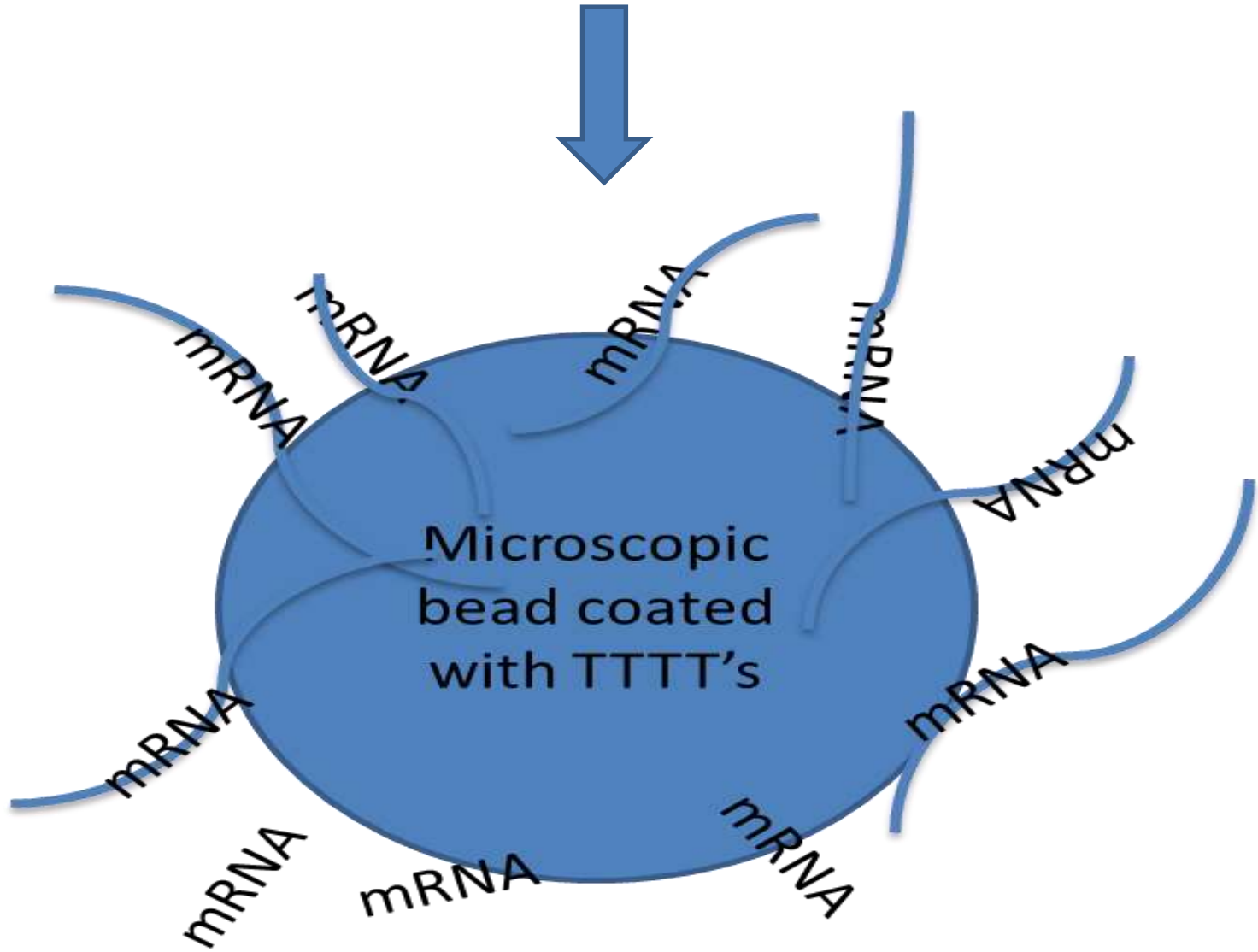
- A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript.
- Sequence tag can be **linked together** to form long serial molecules that can be cloned and sequenced.
- the number of times a particular tag is observed provides the expression level of the corresponding transcript.

# Steps In Brief...

## 1. Trapping of RNA with beads

Coating of microscopic magnetic beads with “TTTTT” tails is done.



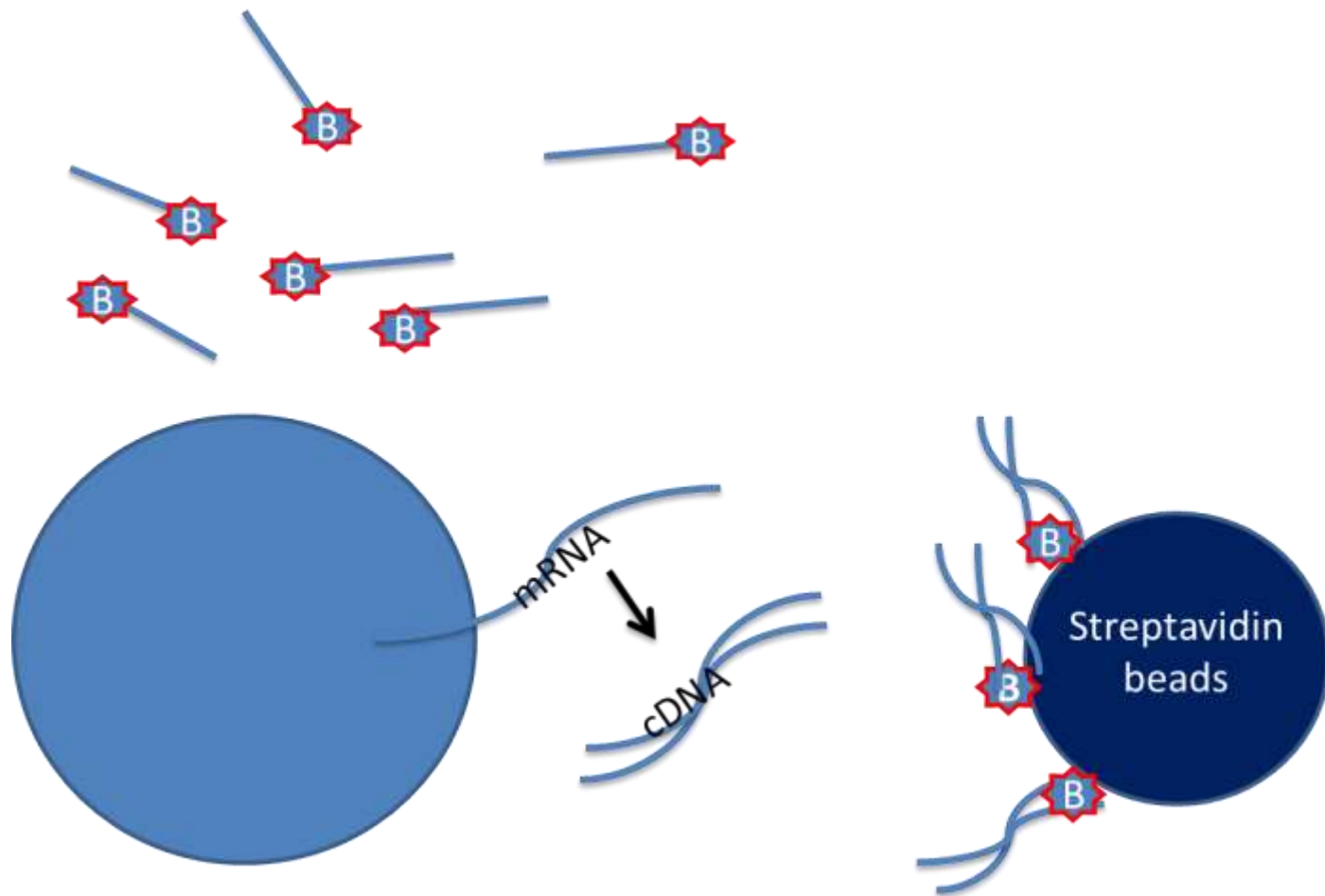


## 2. cDNA synthesis

ds cDNA is synthesized from the extracted mRNA by means of **biotinylated oligo (dT) primer**.

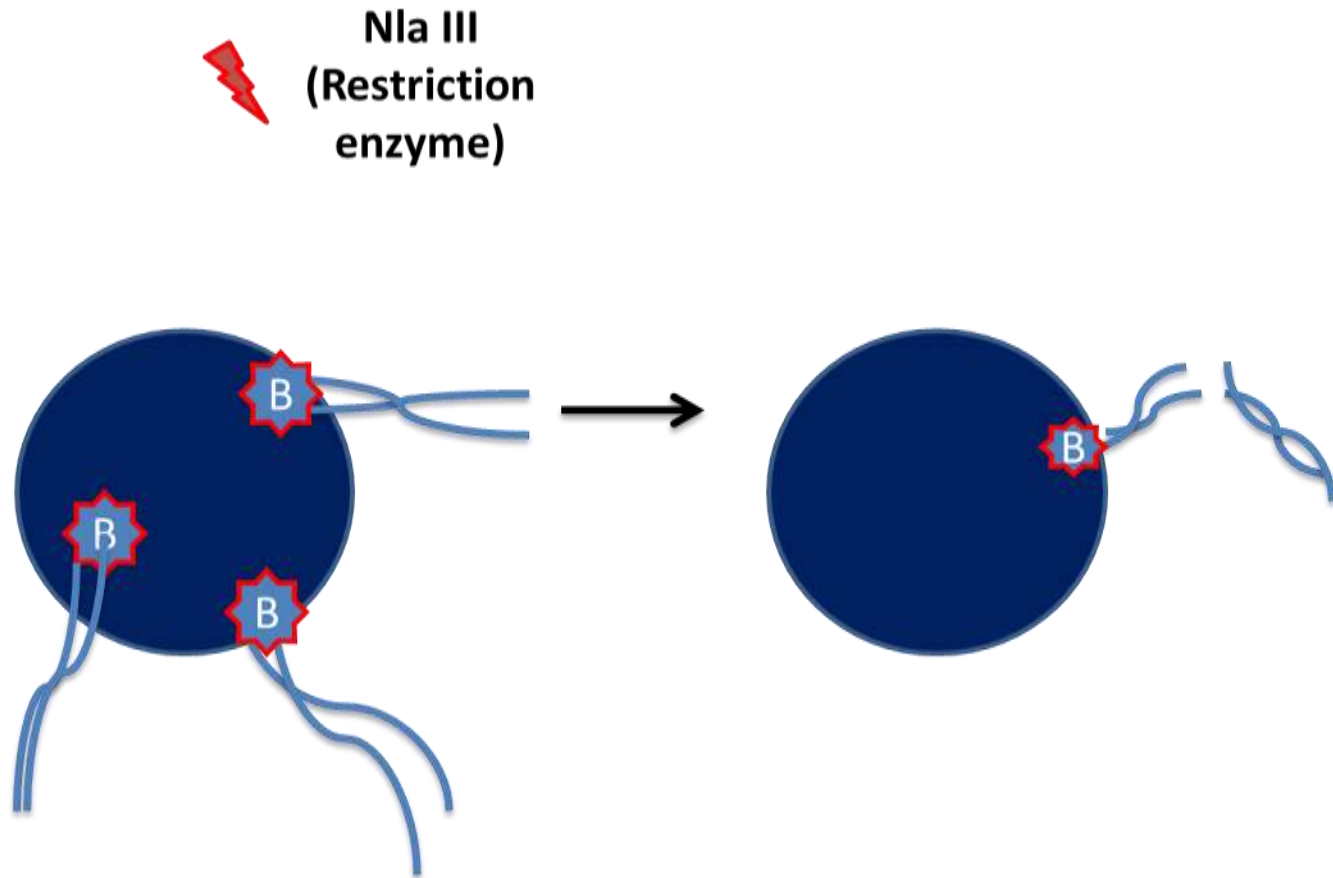
- cDNA synthesis is immobilized to **streptavidin beads**.





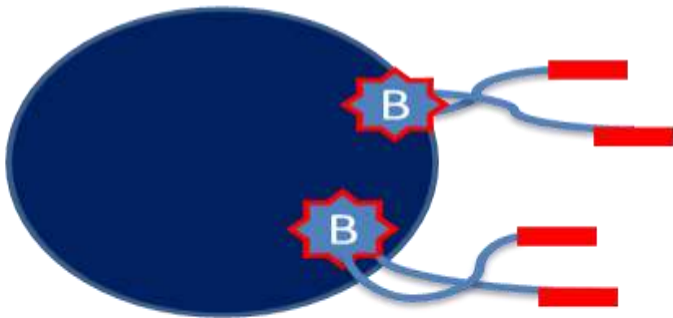
# 3. Enzymatic cleavage of cDNA

- Type II restriction enzyme used (E.g. NlaIII.)

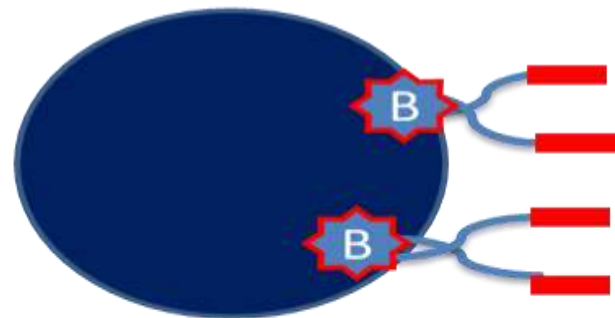


# 4. Ligation of Linkers to bound cDNA

- Captured cDNA are then ligated to linkers at their ends.



Pool A



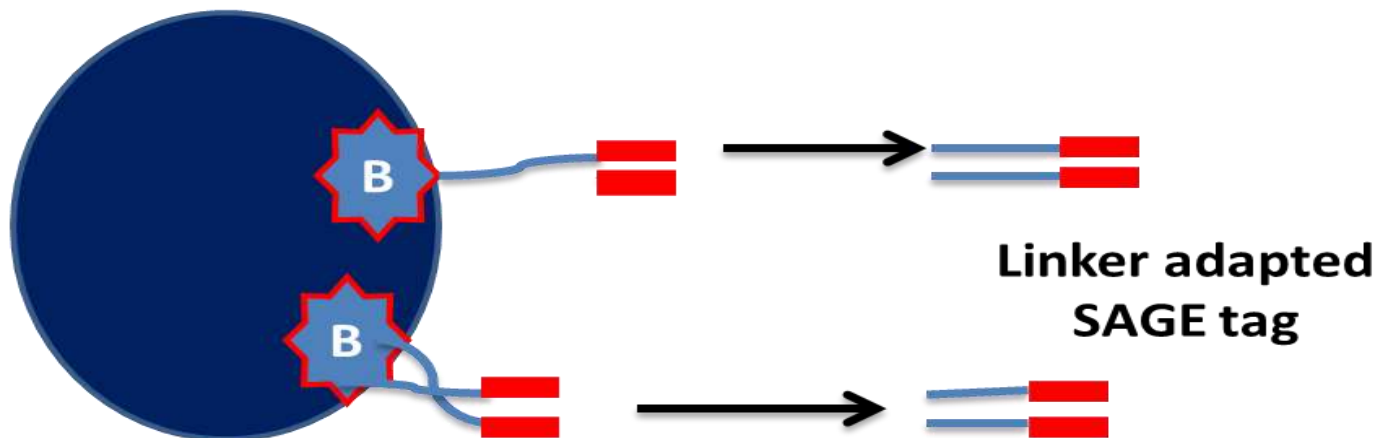
Pool B

# 5. Cleaving with tagging enzyme

- Tagging enzyme, (usually BsmF1) cleave DNA, releasing the **linker-adapted SAGE tag** from each cDNA.

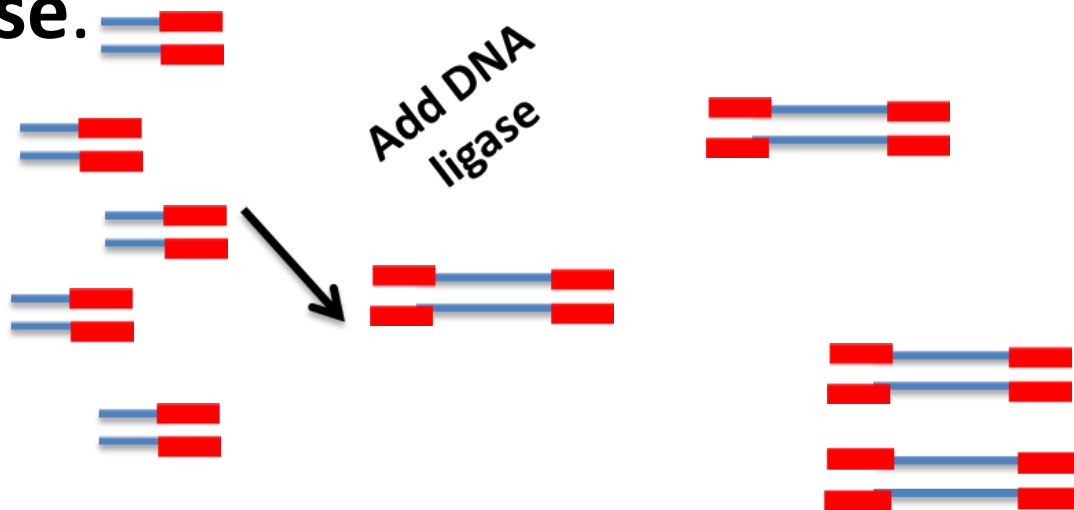


**Bsm FI**  
(tagging Enzyme)



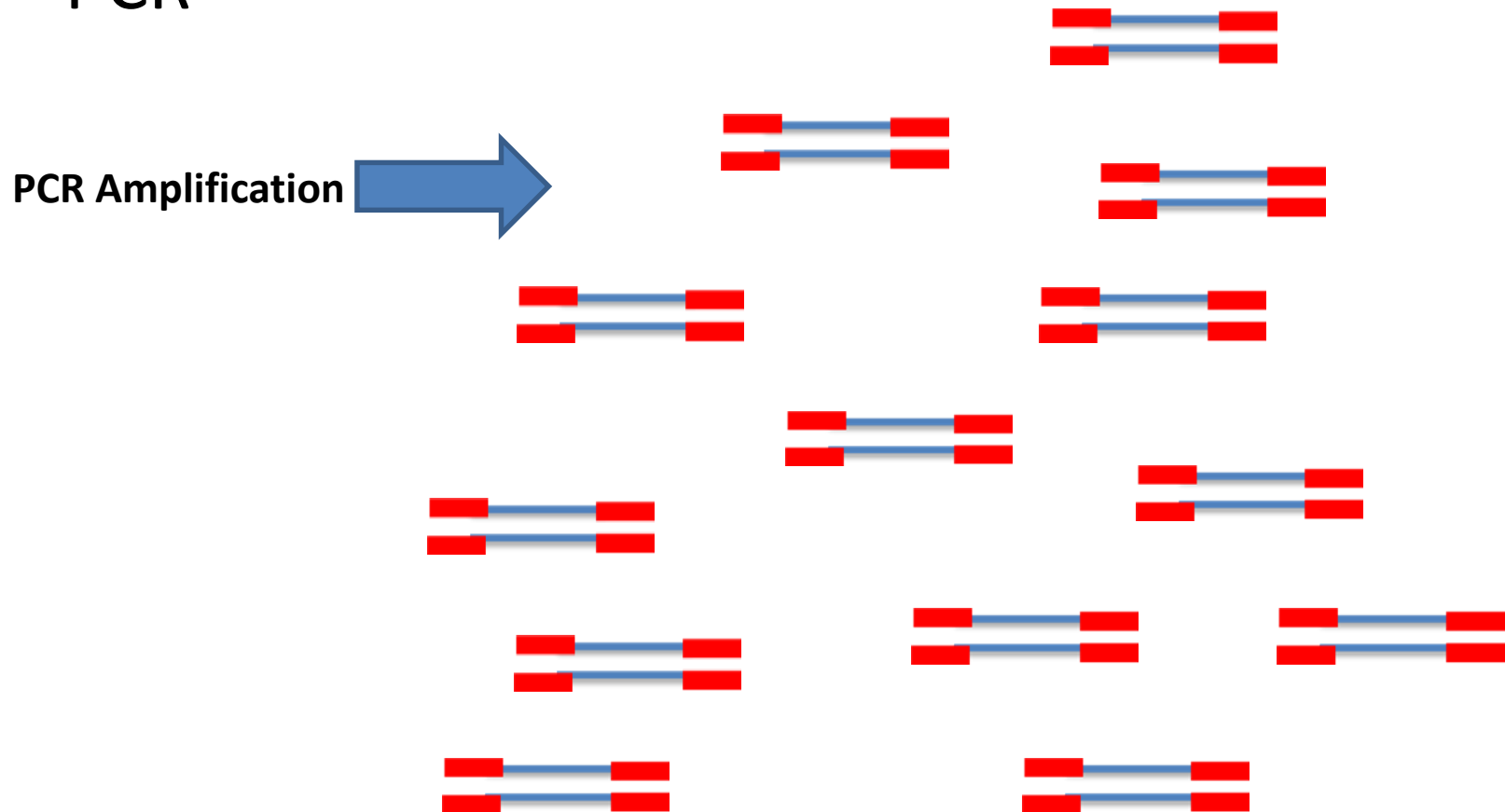
## 6. Formation of Ditags

- Two groups of cDNAs are ligated to each other, to create a “ditag” with linkers on either end.
- Two tags are linked together using **T4 DNA ligase**.



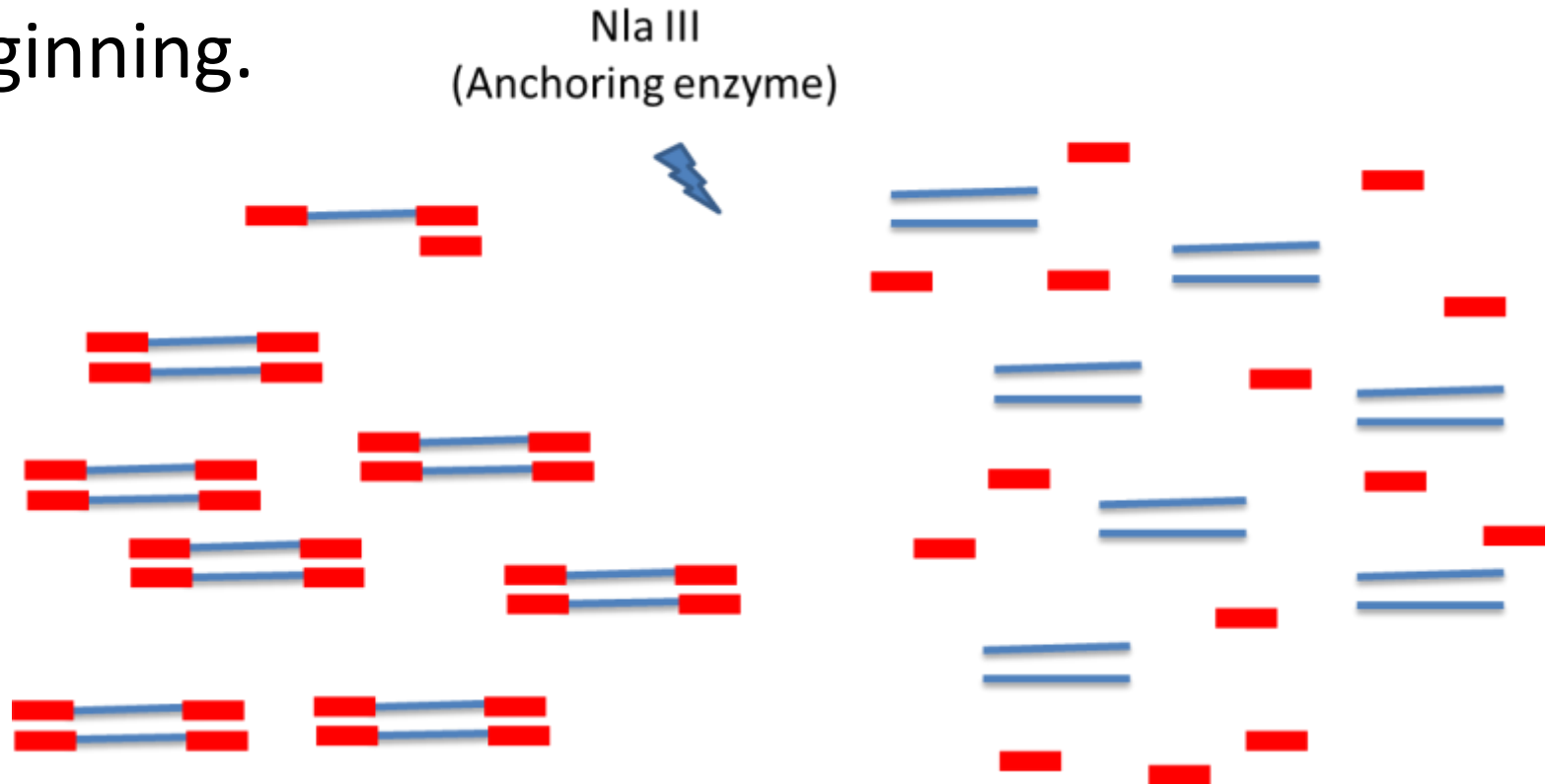
# 7. PCR amplification of Ditags

- linker-ditag-linker constructs are amplified by PCR



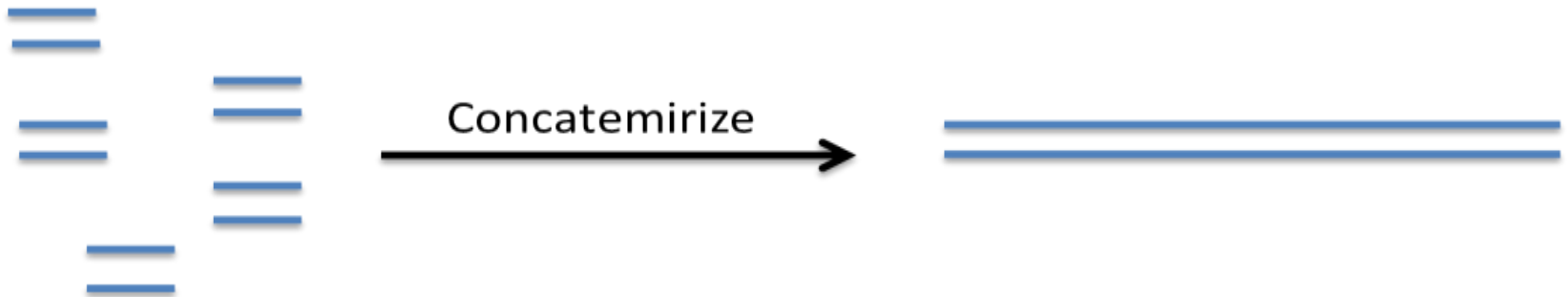
# 8. Isolation of Ditags

- The cDNA is again digested by the Anchoring enzyme (AE)
- Breaking the linker off right where it was added in beginning.



# 9. Concatamerization of Ditags

- Tags are combined into much longer molecules, called concatamers





# Cloning Concatamers and Sequencing...

- Concatamers are inserted into bacteria, which act like living “copy machines” to **create millions of copies** from original.
- Copies are then **sequenced**, using machines that can read the nucleotides in DNA.

- **Software tools for SAGE analysis:**

  - SAGE map

  - SAGE xprofiler

  - SAGE Genie

- **Advantages over EST analysis:**

  - a. Detect **weakly expressed** genes

  - b. It uses a short nucleotide tag and allows **sequencing of multiple tags in a single clone**

# MICROARRAY-BASED APPROACHES

A microarray is a pattern of ssDNA probes which are immobilized on a surface called a **chip** or a **slide**.

- Microarrays use hybridization to detect a specific DNA or RNA in a sample.
- DNA microarray uses a million different probes, fixed on a solid surface.
- Microarray technology **evolved from Southern blotting**.

# WHY?

- To analyze the expression of thousands of genes in **single reaction**, very quickly and in an efficient manner.
- To understand the genetic causes for the abnormal functioning of the human body.
- To understand which genes are active and which genes are inactive in different cell types.

# OLIGONUCLEOTIDE DESIGN

- Fixed oligonucleotide onto a solid support such as glass slide
- Length of oligonucleotides is in range of **25-70 bases** long
- Oligonucleotides are called **probes** (should be specific to minimize cross-hybridization)
- Oligonucleotide should not form stable internal secondary structure, A program **M fold** helps to detect secondary str.
- All probes should have approx. equal **T<sub>m</sub>**.
- **OligoWiz & OligoArray** are 2 programs used in designing probe for microarray spotting.

# STEPS

Sample  
preparation  
and  
labeling

Hybridisation

Washing

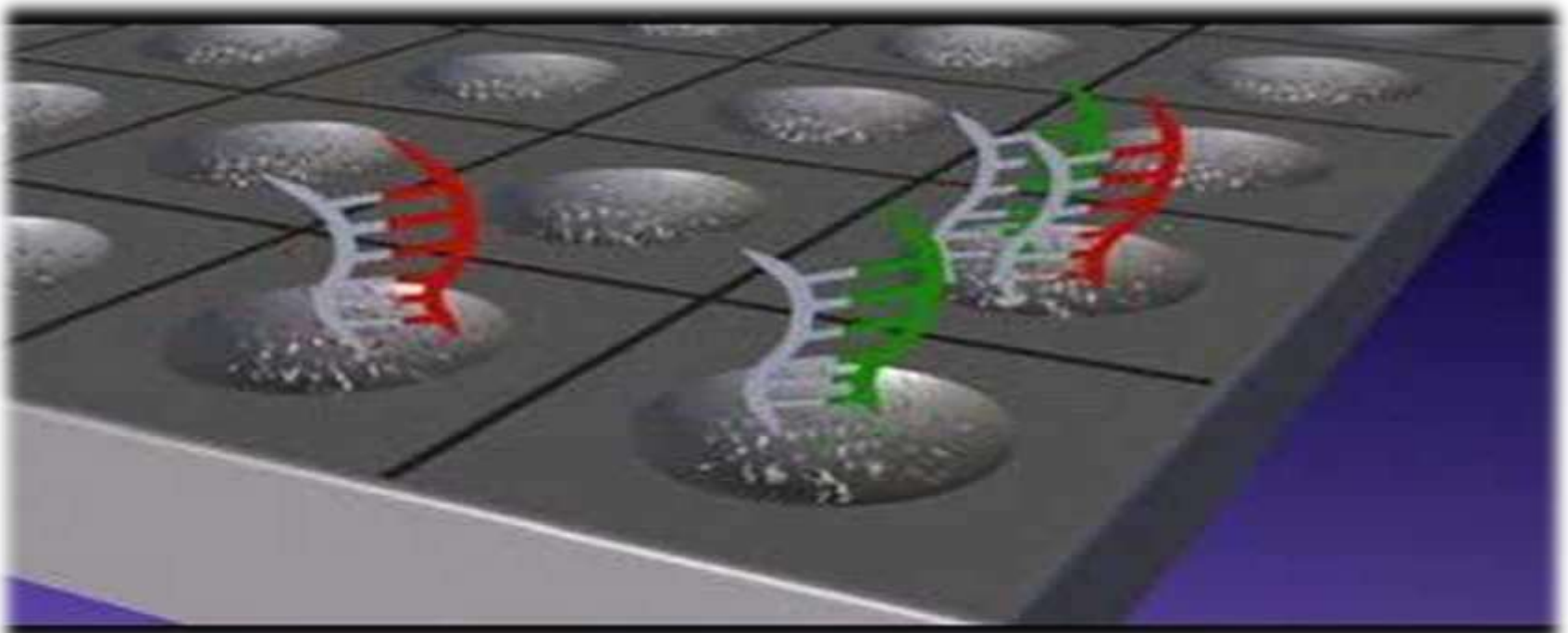
Image processing  
and  
Data analysis

# 1 .Sample preparation and labeling

- **Isolate a total RNA containing mRNA** that ideally represents genes, that are expressed at the time of sample collection.
- **Preparation of cDNA from mRNA** using a reverse-transcriptase enzyme.
- Short primer is required to initiate cDNA synthesis.
- Each cDNA (Sample and Control) is **labelled with fluorescent cyanine dyes** (i.e. Cy3 and Cy5).

## 2. Hybridization

- labelled cDNA is competitively hybridized against cDNA molecules spotted on a glass slide.





# **3. Image acquisition and data analysis**

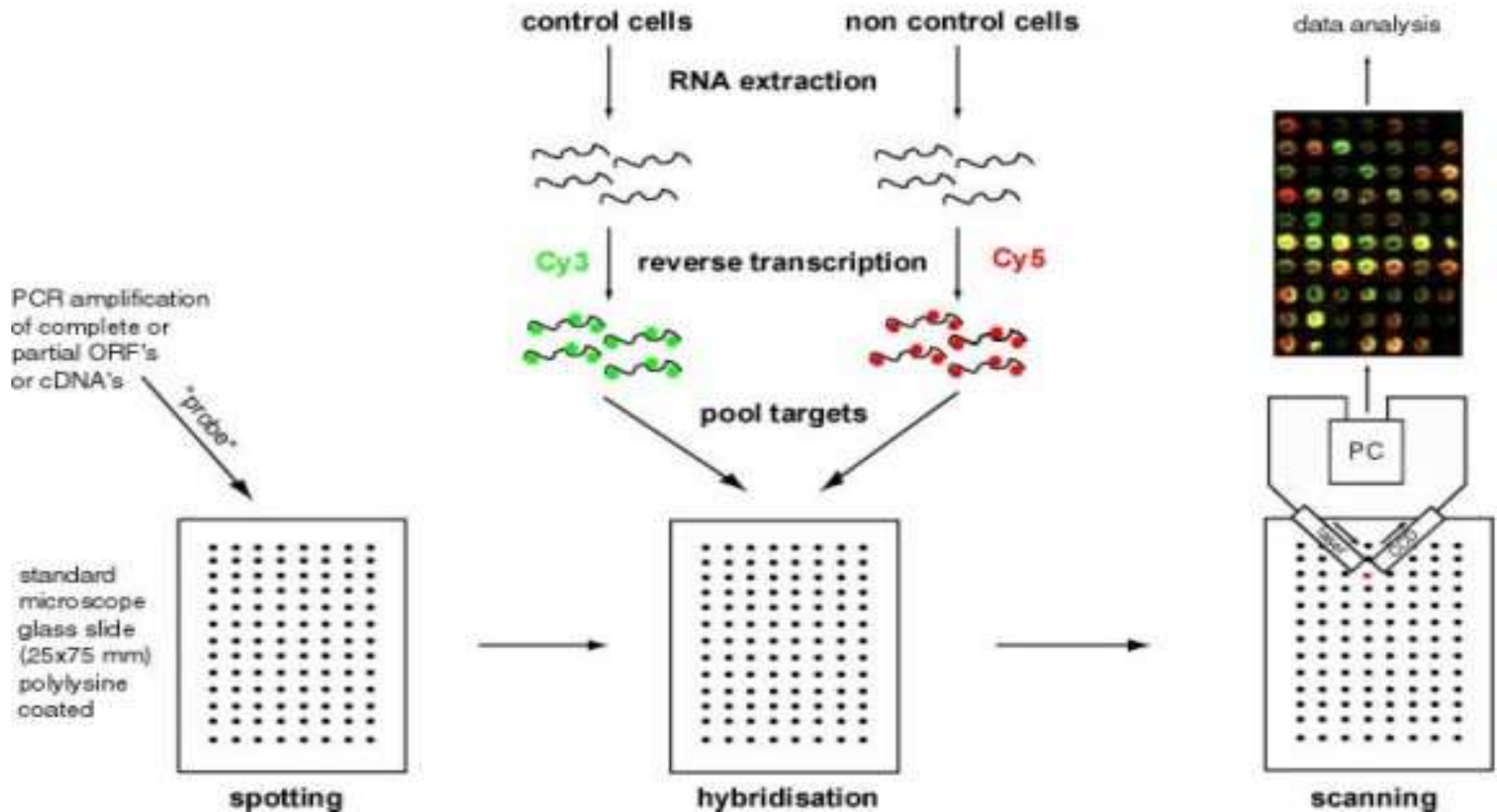
Scanning hybridization signals

Image analysis

Transformation and normalization of data

Analyzing data

# Diagram



# Image processing

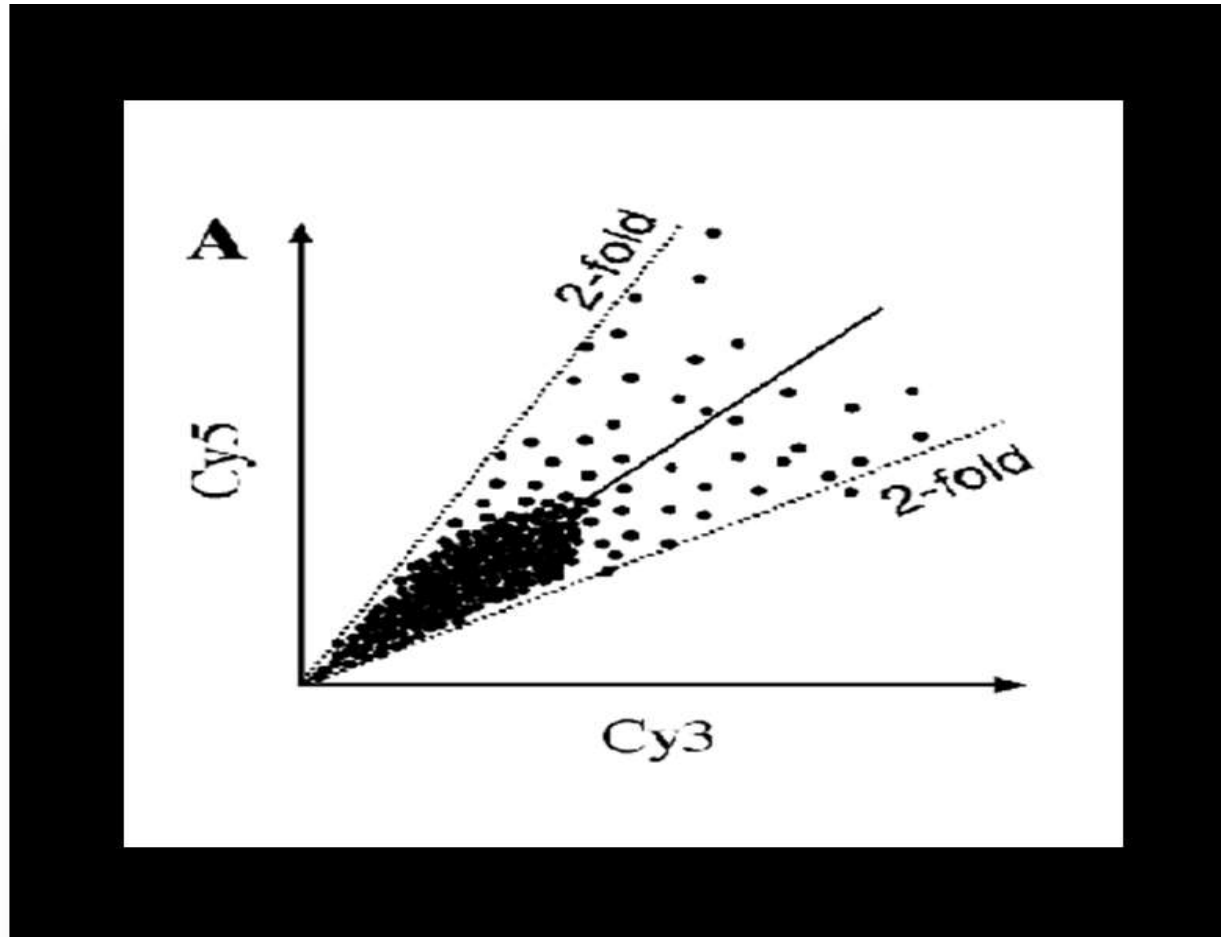
- To **locate and quantify hybridization spots** and to separate true hybridization signals from background noise ( may be **non specific hybridization, unevenness of the slide** surface, presence of contaminants such as **dust** on surface of slide).
  1. Computer programs are used to correctly locate the boundaries of spots.
  2. Arrayed signals are converted into numbers and observed as ratio between cy5 and cy3 for each spot.
  3. data are often presented as false colour.
  4. If ratios value are above 1 the false colour appears as red if it is below 1 then appear as green.

- Software programs to perform microarray image analysis:
- ArrayDB
- TIGR Spotfinder

# Data transformation and normalization

- Digitized gene expression data **need to be further processed** before differentially expressed gene can be identified.
- When **raw fluorescence intensity** cy5 is plotted against cy3, most of data are clustered near bottom left of plot.
- -It shows **non- normal distribution of raw data.**
- -It represents **imbalance of red & green intensities.**

# Plot of raw fluorescence signal intensities of Cy5 versus Cy3



# Improvement in data quality

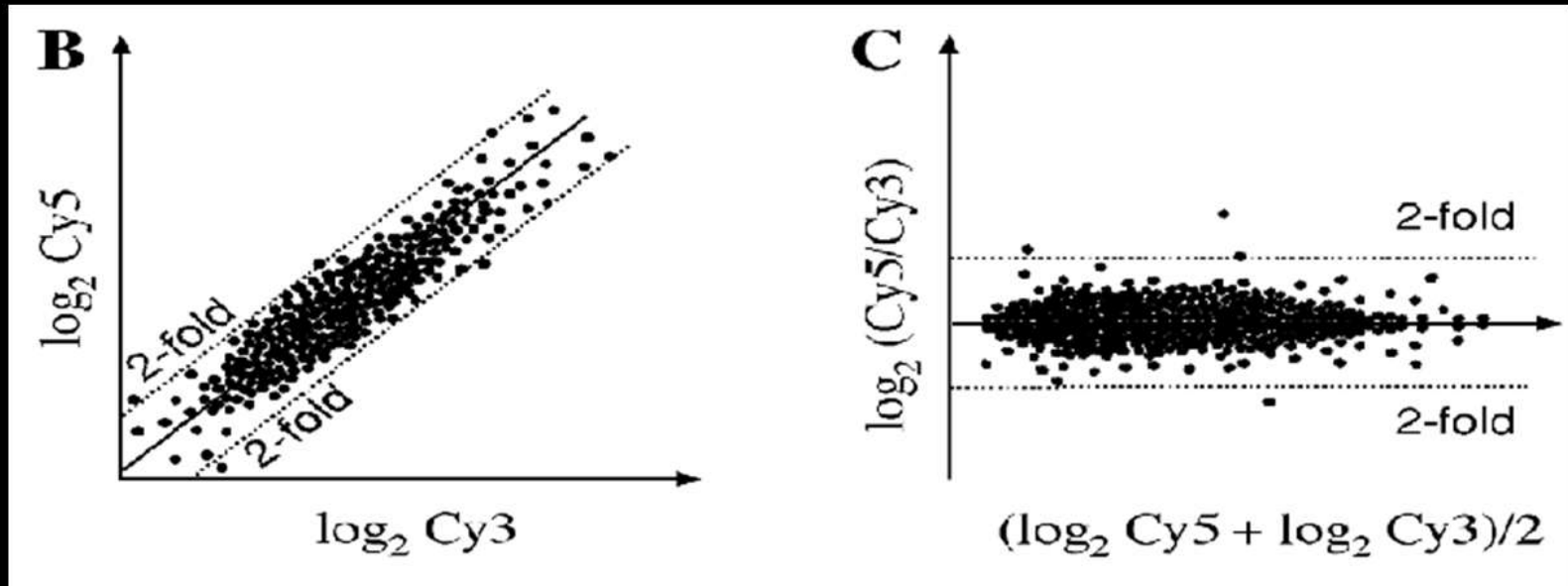
1). Transform raw cy5 & cy3 values **by taking log to the base of 2.**

It produces **uniform distribution** of data.

2). Further normalization of data is done by plotting the data points horizontally

plot the log **ratios(cy5/cy3)** against the average log intensities  **$(\log_2 \text{cy5} + \log_2 \text{cy3})/2$**

- This form is known as “**intensity- ratio plot**” In which differentially expressed genes are more easily visible about a horizontal axis.
- In all these types **linear regression** is used.



B. Plot of data after log transformation

C. Plot of mean log intensity v/s log ratio of two fluorescence intensities



- **Software programs for microarray data normalization:**

**Arrayplot**

**SNOMAD**

# Statistical Analysis to identify Differentially Expressed Genes

- Differentially expressed genes are separated, using **normalization cutoff** of generally two fold.
- But it is arbitrary cutoff value depending on **data variability**.
- Ways to ensure that gene is truly differentially expressed :
  - 1) **Multiple replicate experiments**
  - 2) **Statistical testing**

## 1. Multiple replicate experiment:

- Repeat experiments provide replicate data points that provide information about variability of expression data.
- **Helps to identify differentially expressed genes.**

## 2. Statistical Tests:

- t-Test
- &
- ANOVA (Analysis of variance).

- **Other Statistical Programs**

- 1) MA-ANOVA

- 2) Cyber-T

- Advantage of microarray based approach:

**To study the expression of many genes in parallel & identify groups of genes that exhibit similar expression patterns.**

# Comparison of SAGE & Microarray based approach

## SAGE

- 1) Doesn't require prior knowledge of transcript sequence.
- 2) SAGE measures absolute mRNA expression level.
- 3) A minute quantity of sample mRNA is required.
- 4) It is expensive & gene identification is difficult.

## DNA Microarray

- 1) Detect only the genes spotted on microarray.
- 2) It indicates the relative expression level.
- 3) It requires larger quantity of mRNA sample.
- 4) It is less expensive & identity of probes are already known.



**THANK YOU**