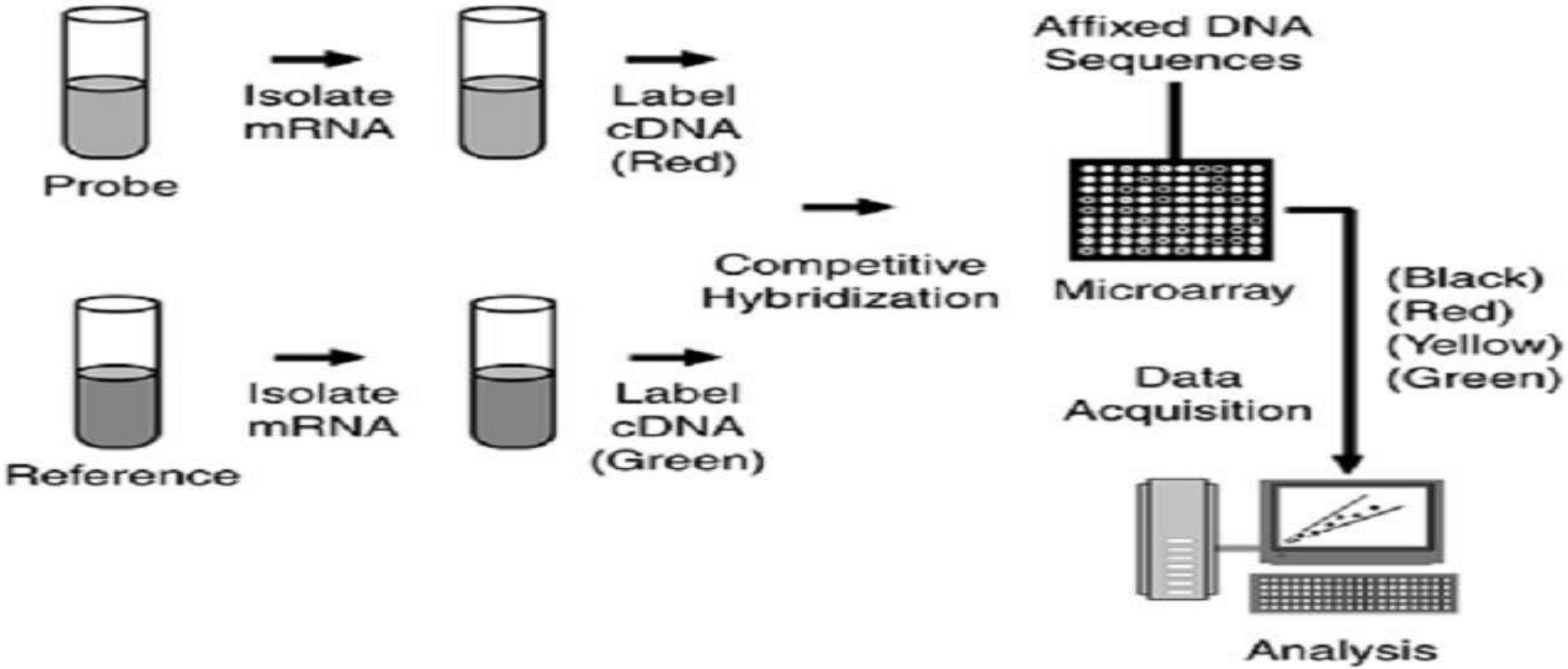
Understanding Microarray Data

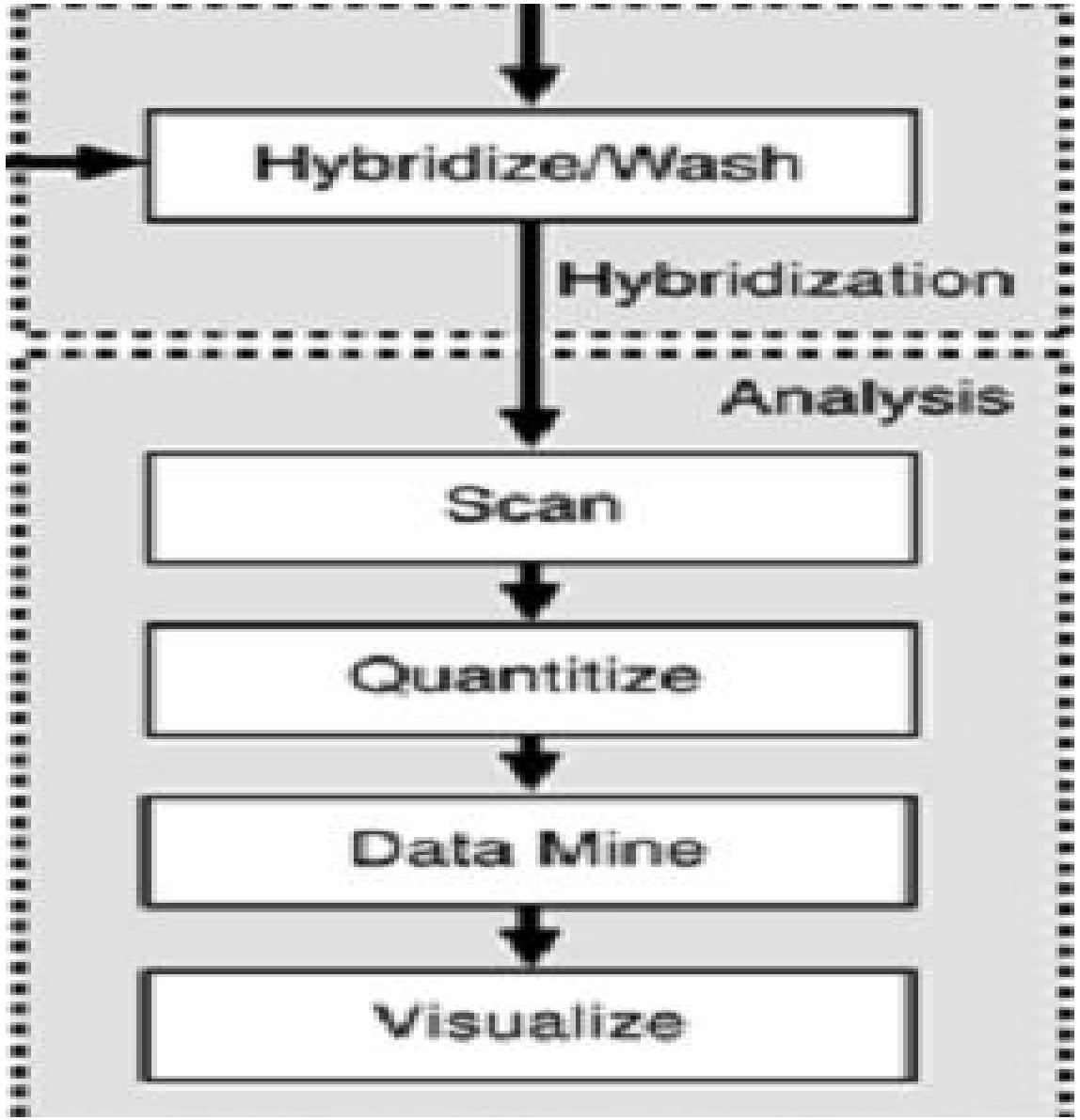
MBI401-High throughput Data Generation & analysis

Mamta Sagar
Department of Bioinformatics
University Institute of Engineering & Technology, CSJM University, Kanpur

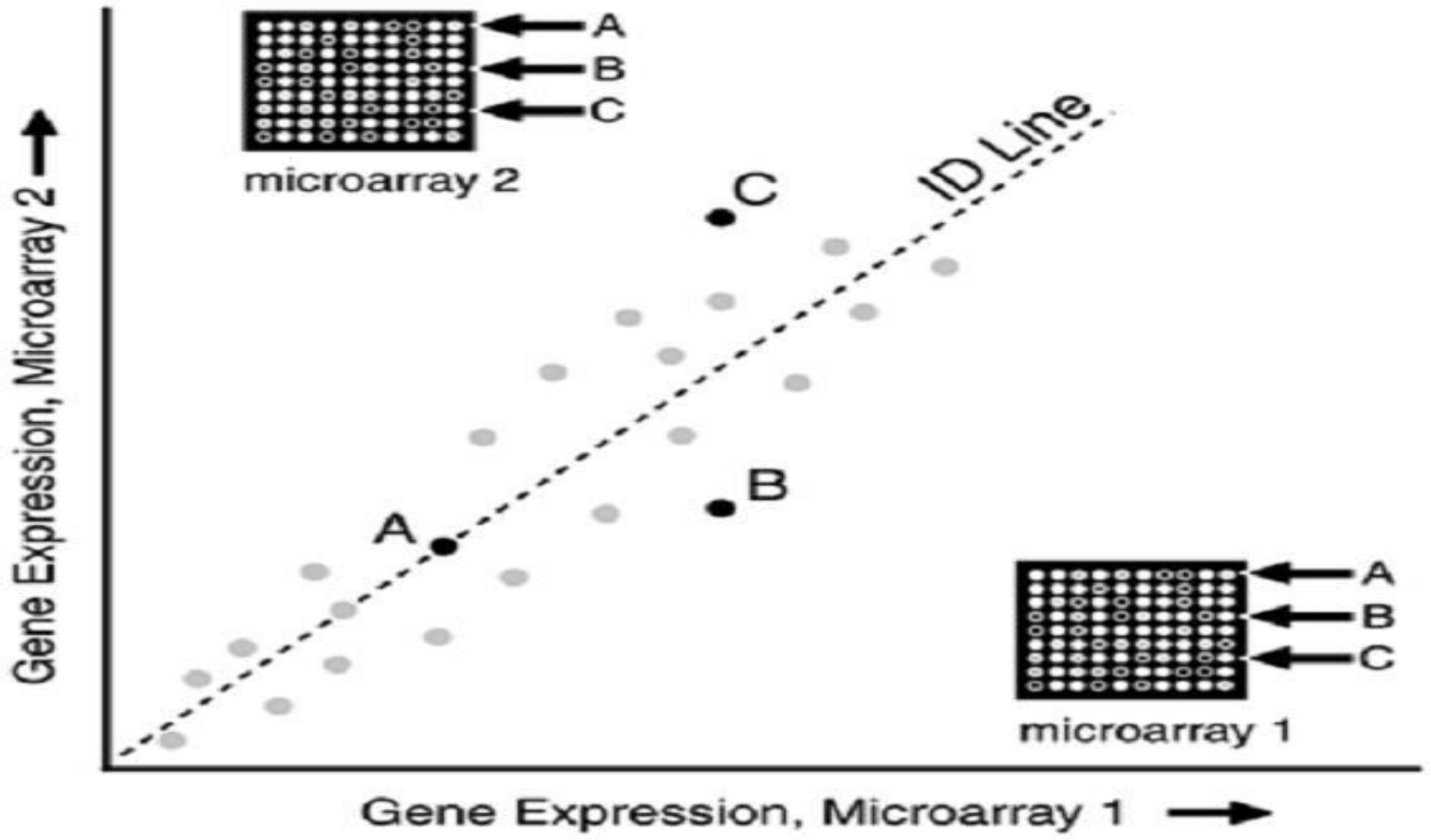
Schematic Presentation of Microarray Technology



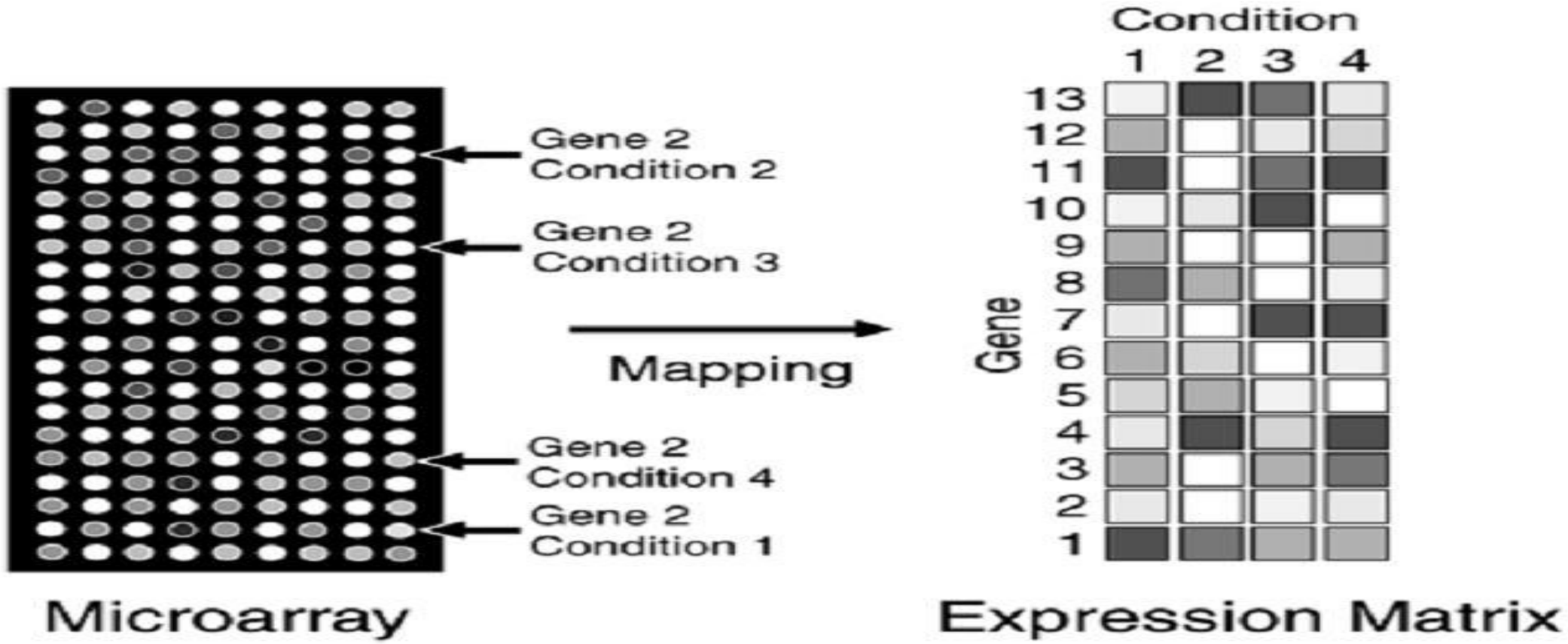
Analysis involve data mining



Microarray Results Analysis. Scatter plot illustrating inter-microarray variability in two identically treated microarrays, Microarray 1 and Microarray 2. Ideally, all data points fall on the ID line, as illustrated by data point (A).

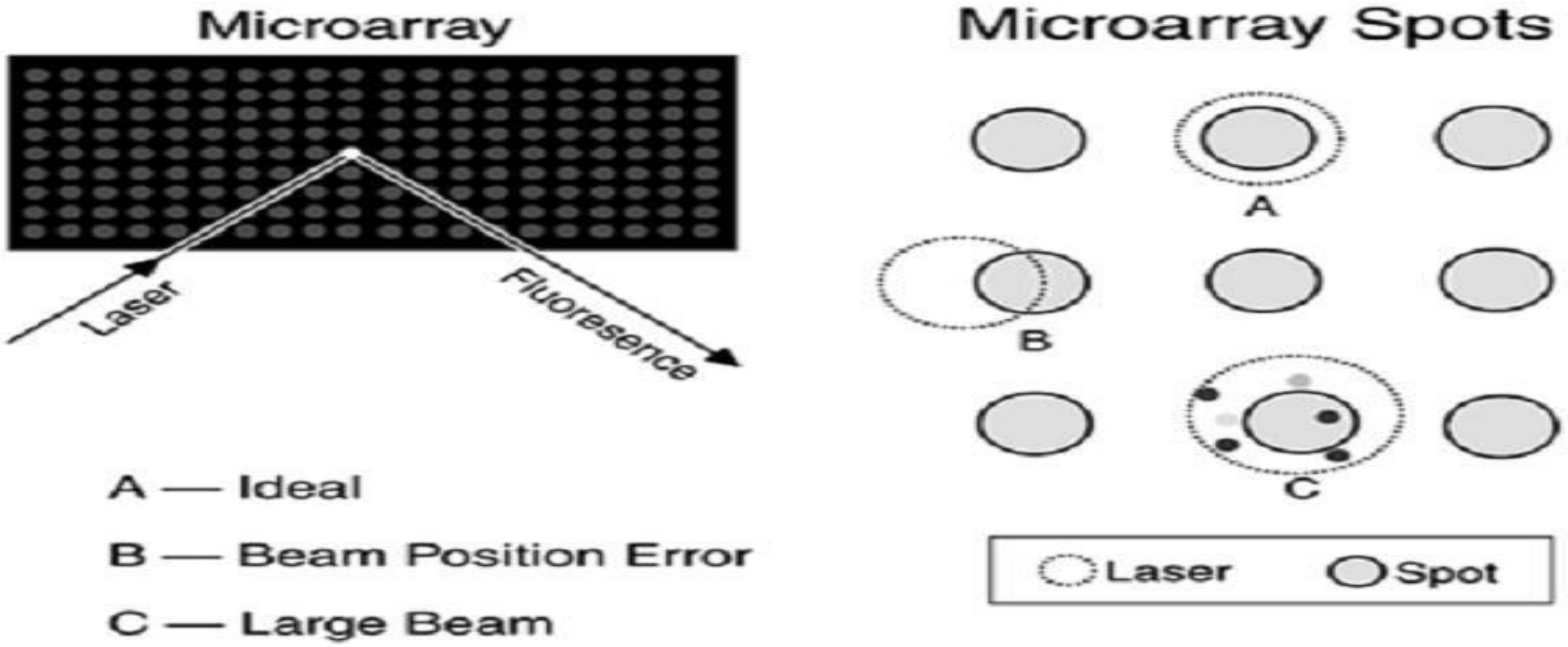


Mapping Microarray Data to an Expression Matrix. Note the lack of correlation between physical experimental position on the microarray and the mapping of data in the expression matrix. Although shown here in grayscale, the individual squares in the gene expression matrix are normally represented by the fluorescence color of the corresponding microarray spot.

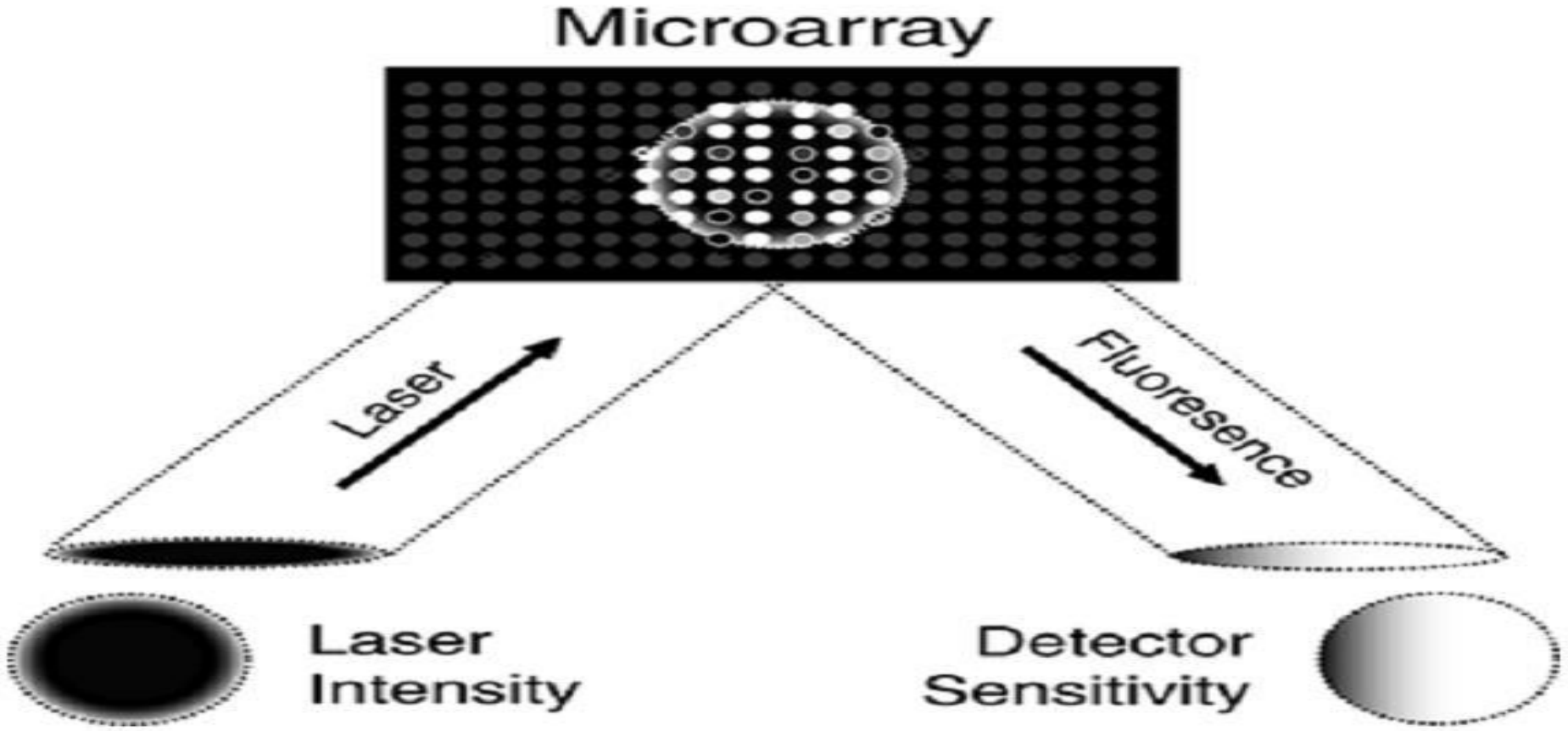


Source of Variablity

ideal.



Sources of Variability in the Starring Method of Reading a Microarray. Not only may the laser intensity be nonlinear across the area of the microarray that is excited by the swath of laser light, but the photodetector may exhibit variations in sensitivity across the detector aperture as well.



Common Sources of Variability Associated with Microarray Preparation. These sources of error affect both the spotting and starring methods of microarray reading.

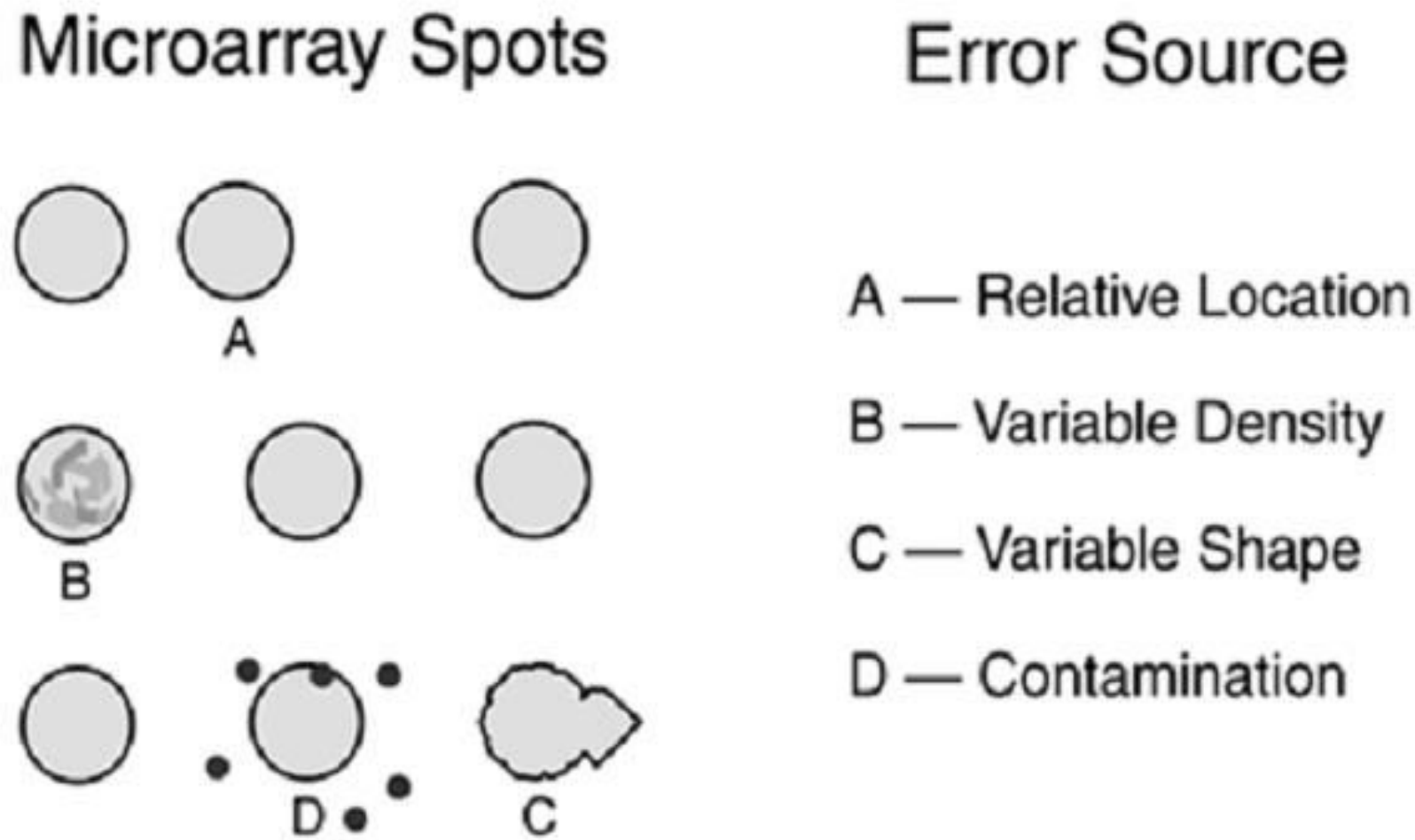


Figure 6-11. The Z-Distribution. This distribution is a special case of the Normal distribution, with mean of zero and a standard deviation of one.

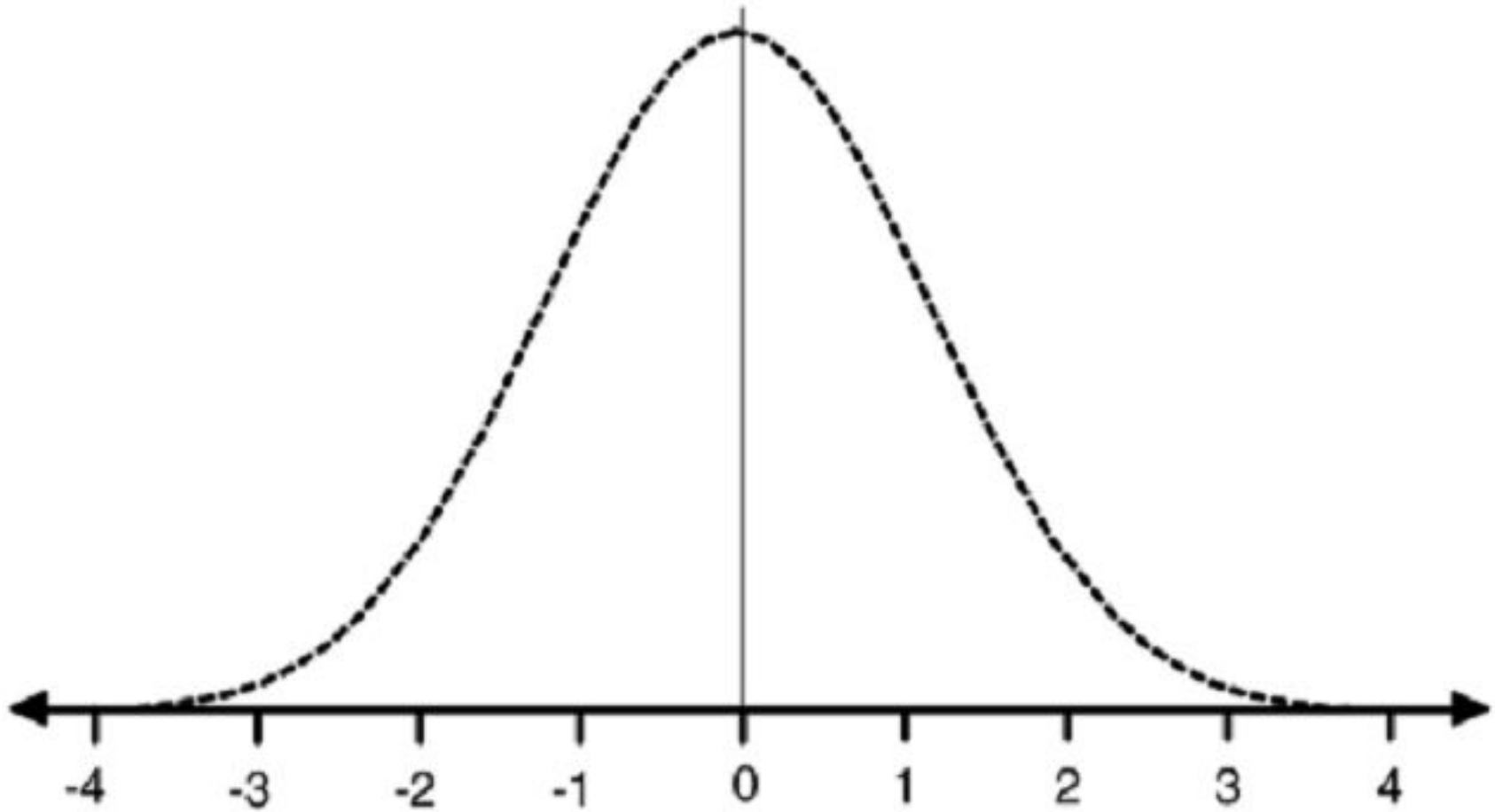


Figure 6-12. Deviations from the Normal Distribution. Although statistical analysis of continuous random variables assumes a normal distribution, many distributions are not normal, as illustrated by the skewed and expected distributions.

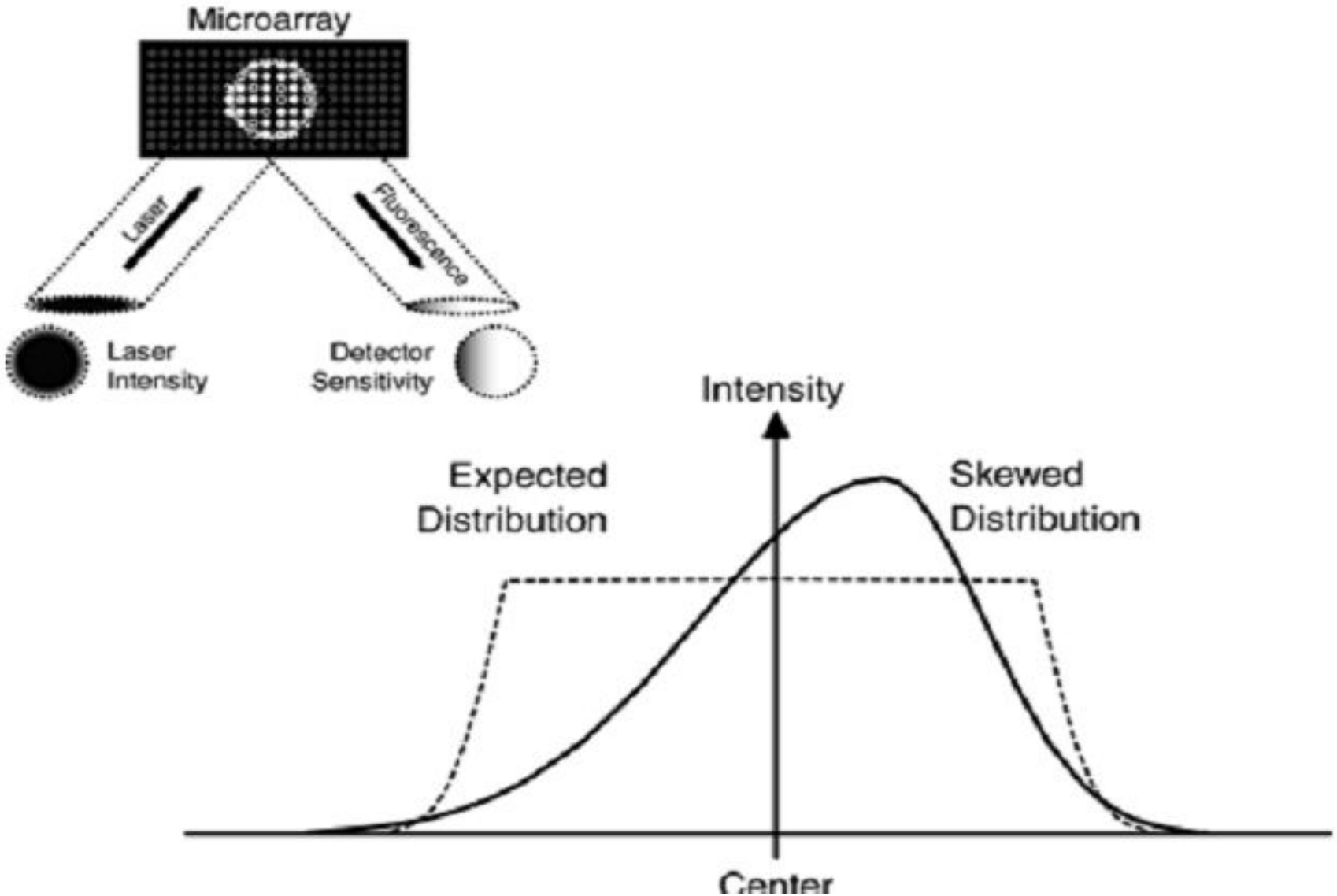


Figure 6-14. Analog-to-Digital Conversion. The dynamic range of the microarray experiment is limited by the resolution or bit depth of the A-to-D conversion process, as illustrated by the magnified view of the digital signal.

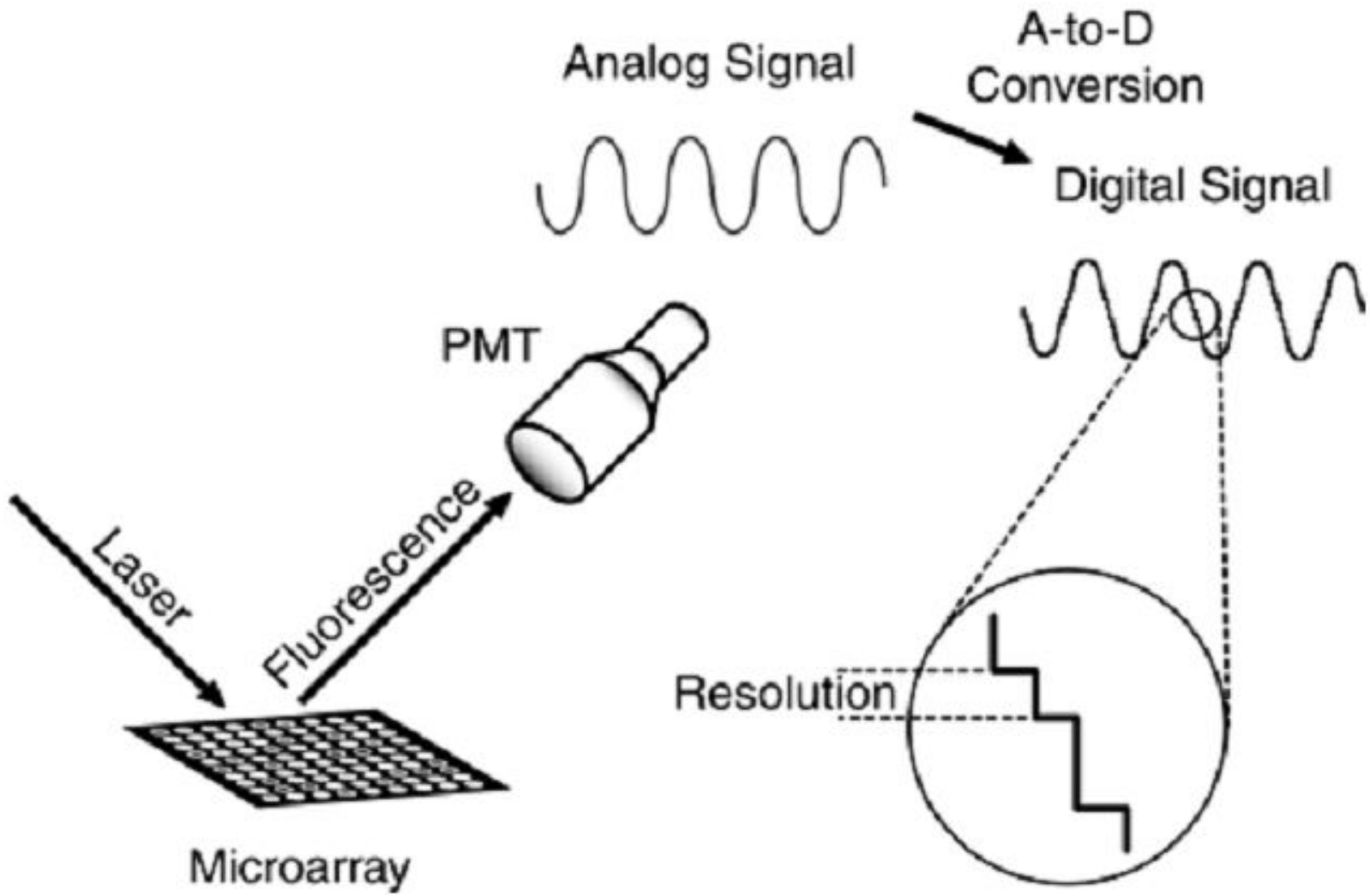
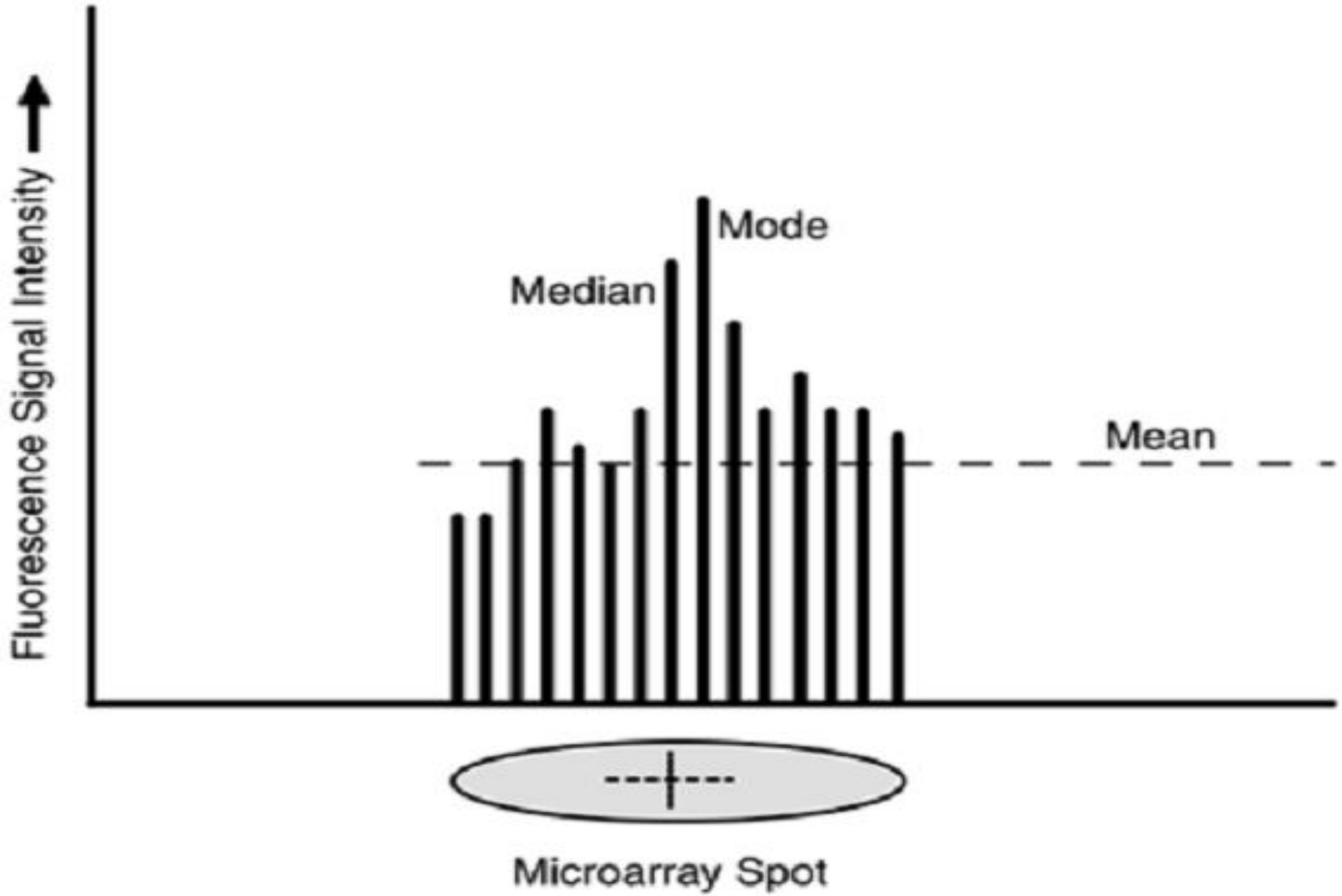


Figure 6-15. Microarray Fluorescence Statistical Analysis.



Microarray spot intensity distriution

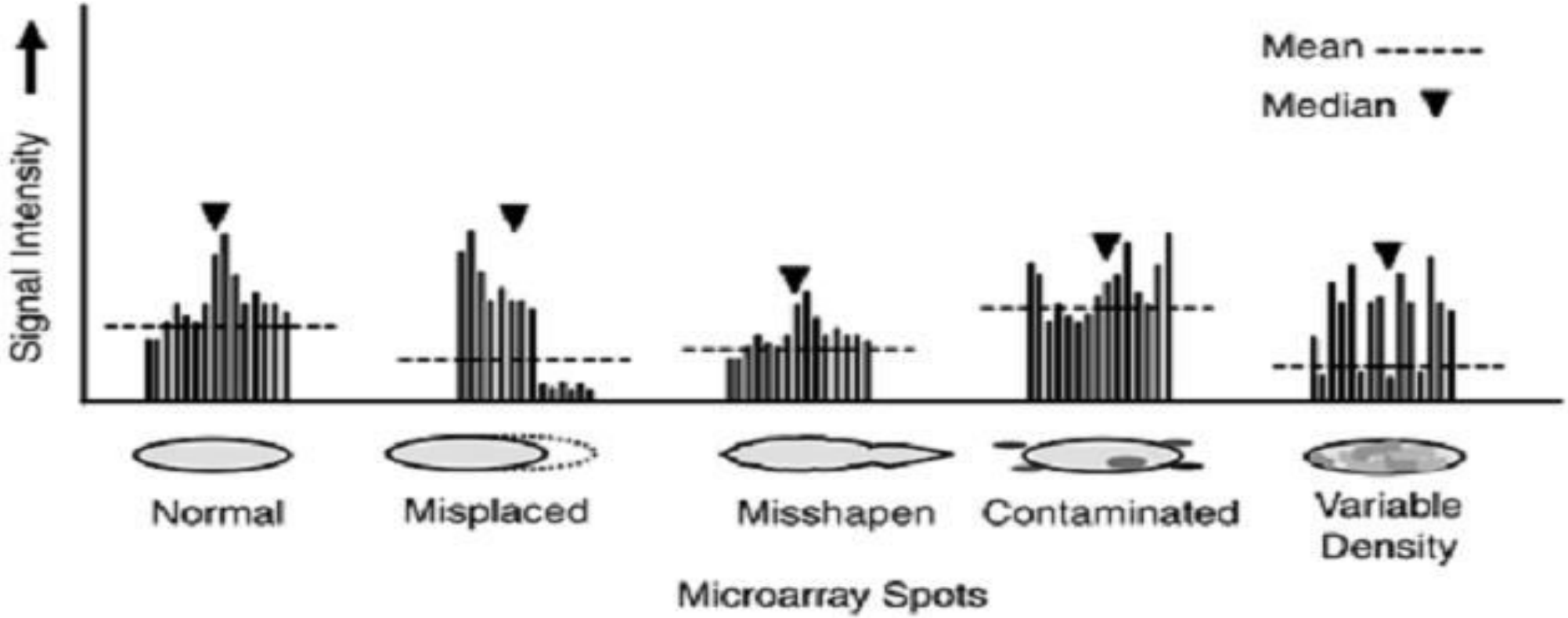


Figure 6-17. Intra-Microarray Intensity Comparisons. Statistical analysis of the means of relative fluorescence intensity can be used to programmatically identify a contaminated sample (far right) that can be discarded from the final gene expression analysis, thereby reducing variability in the experiment.

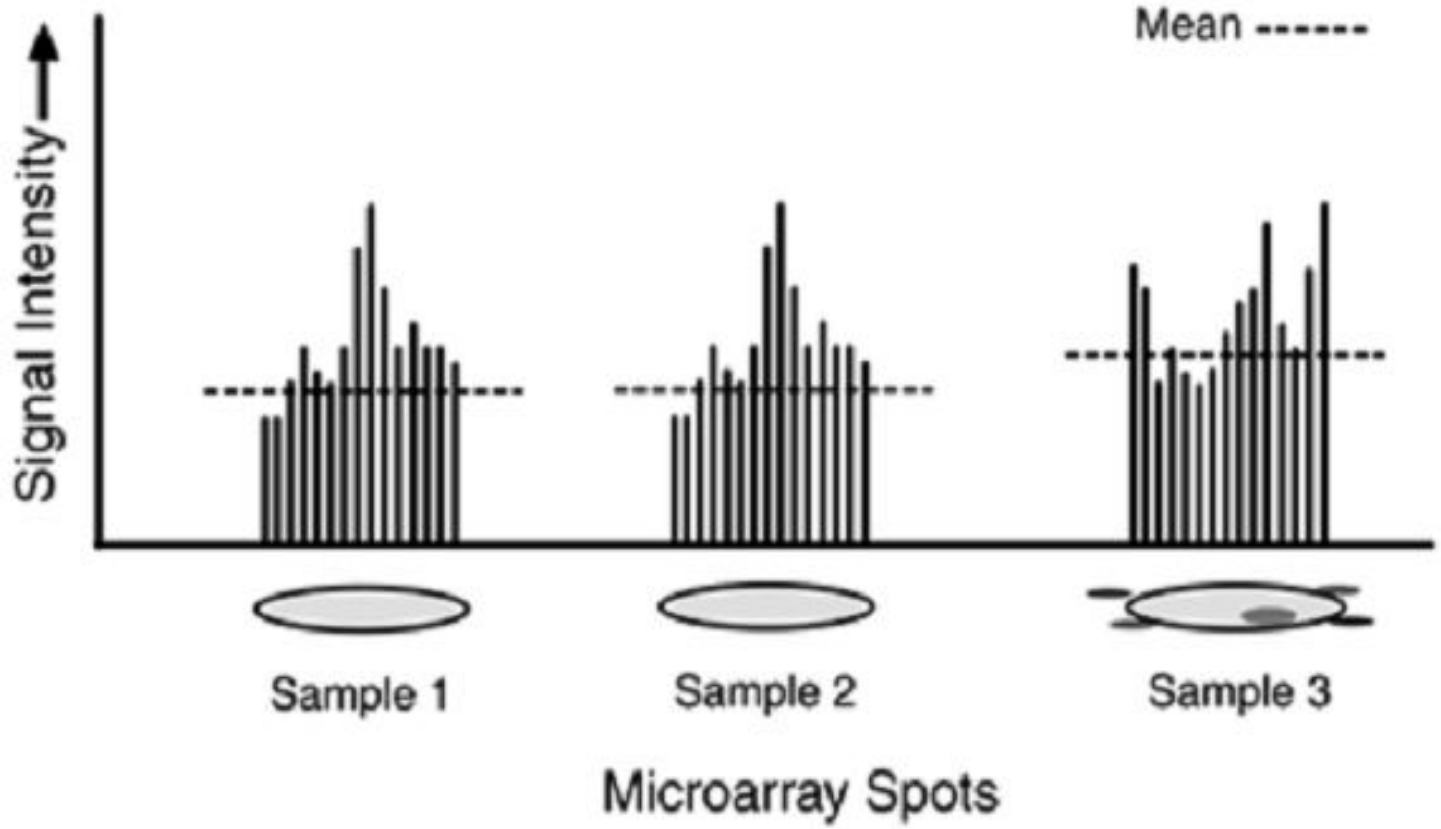
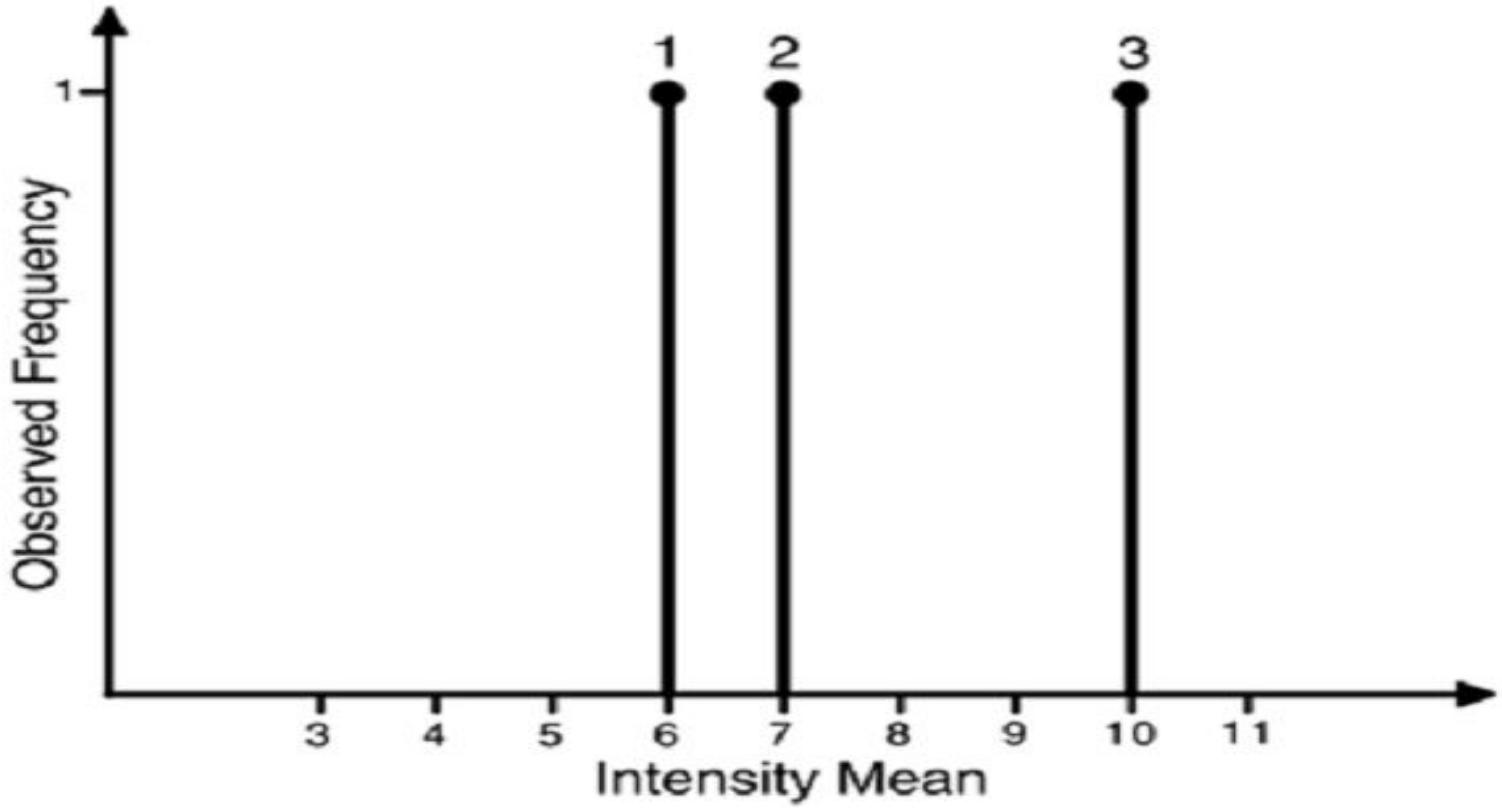


Figure 6-18. Observed Frequency of Differences Between Means. The intensity values associated with sample 3 appears to be different from the values derived from samples 1 and 2. The scale of intensity mean values is arbitrary.



Mathematically, the mean intensity value is computed as:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{i} = \frac{6+7+10}{3} = 7.67$$

Figure 6-19. Z-Scores of Mean Intensity Values. All values are within one z-score (one standard deviation from the mean).

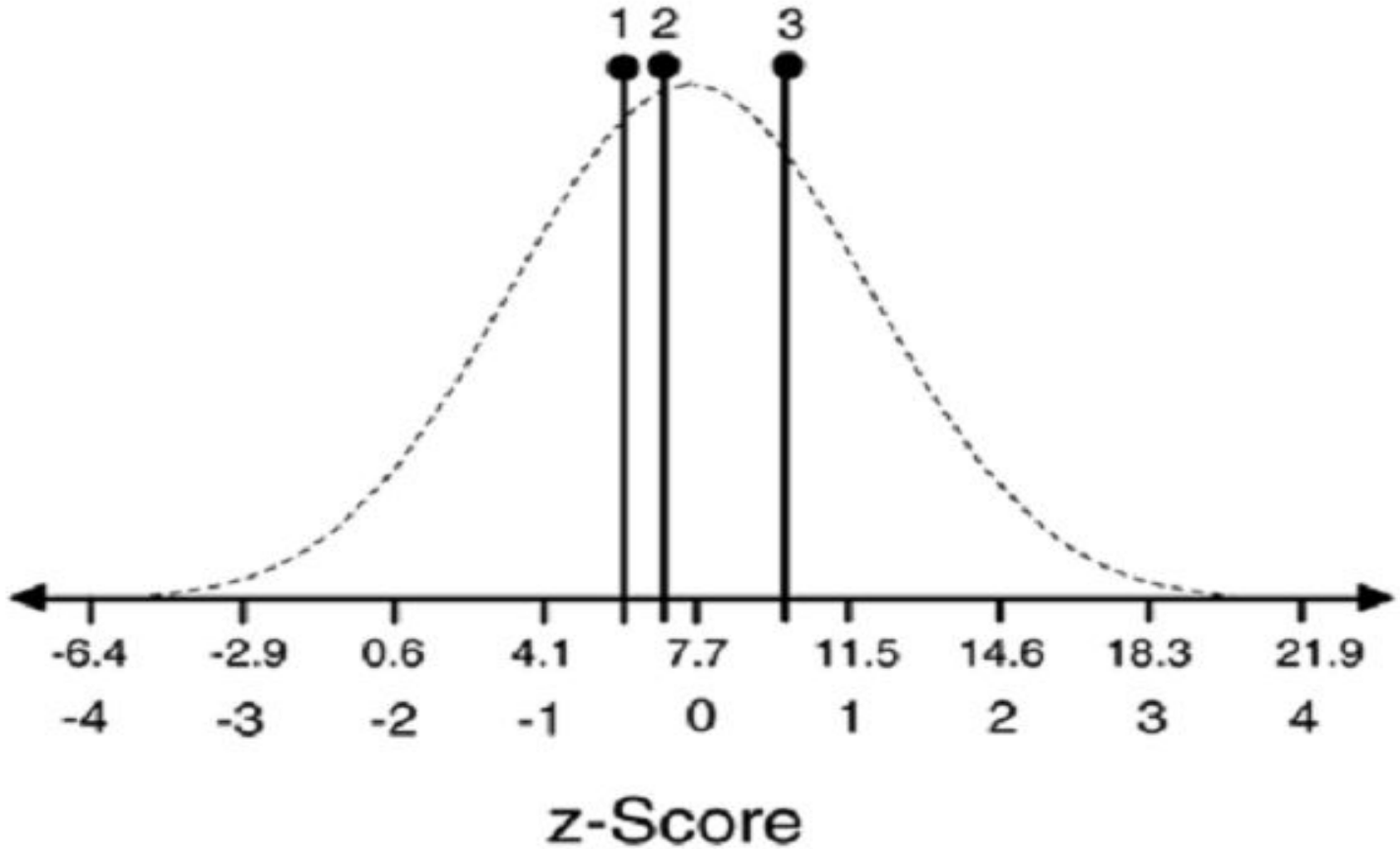
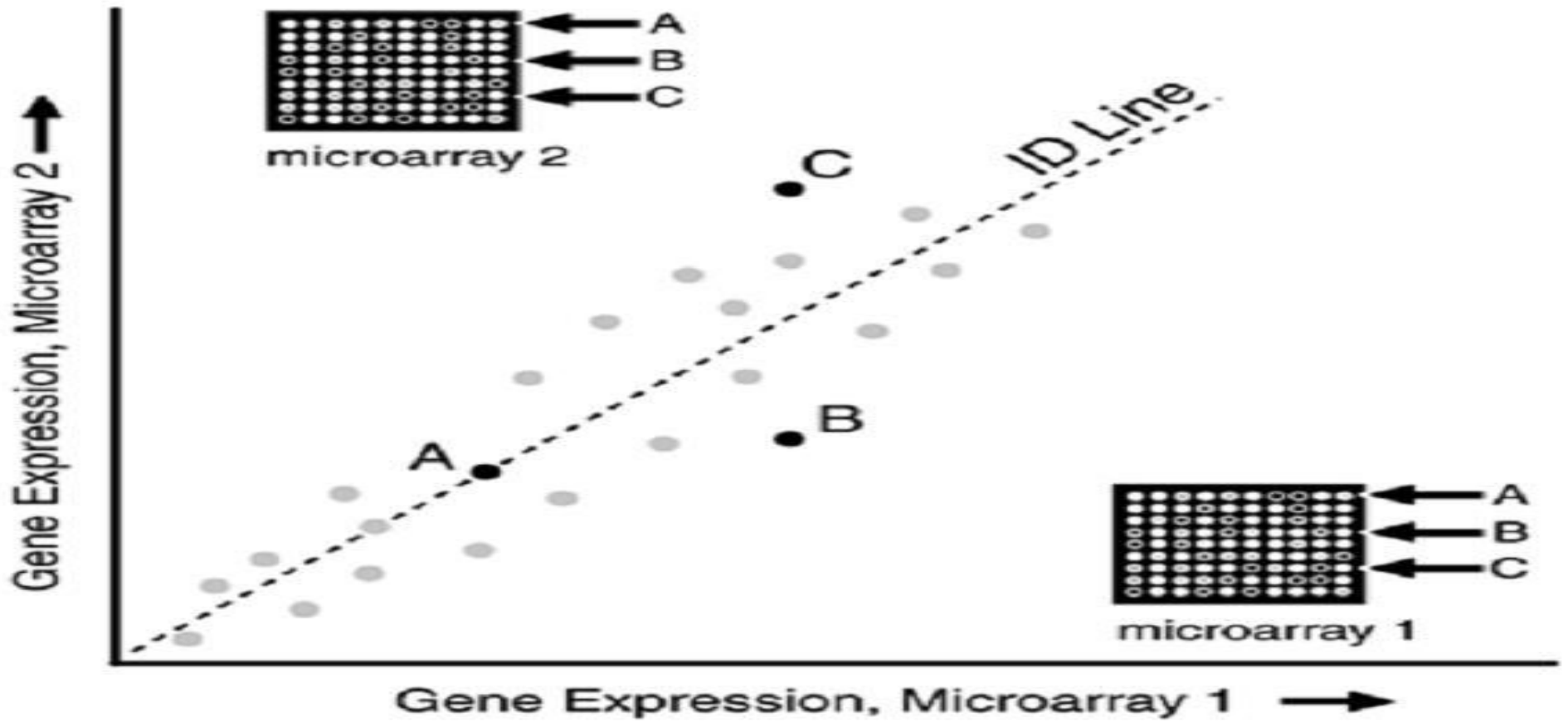
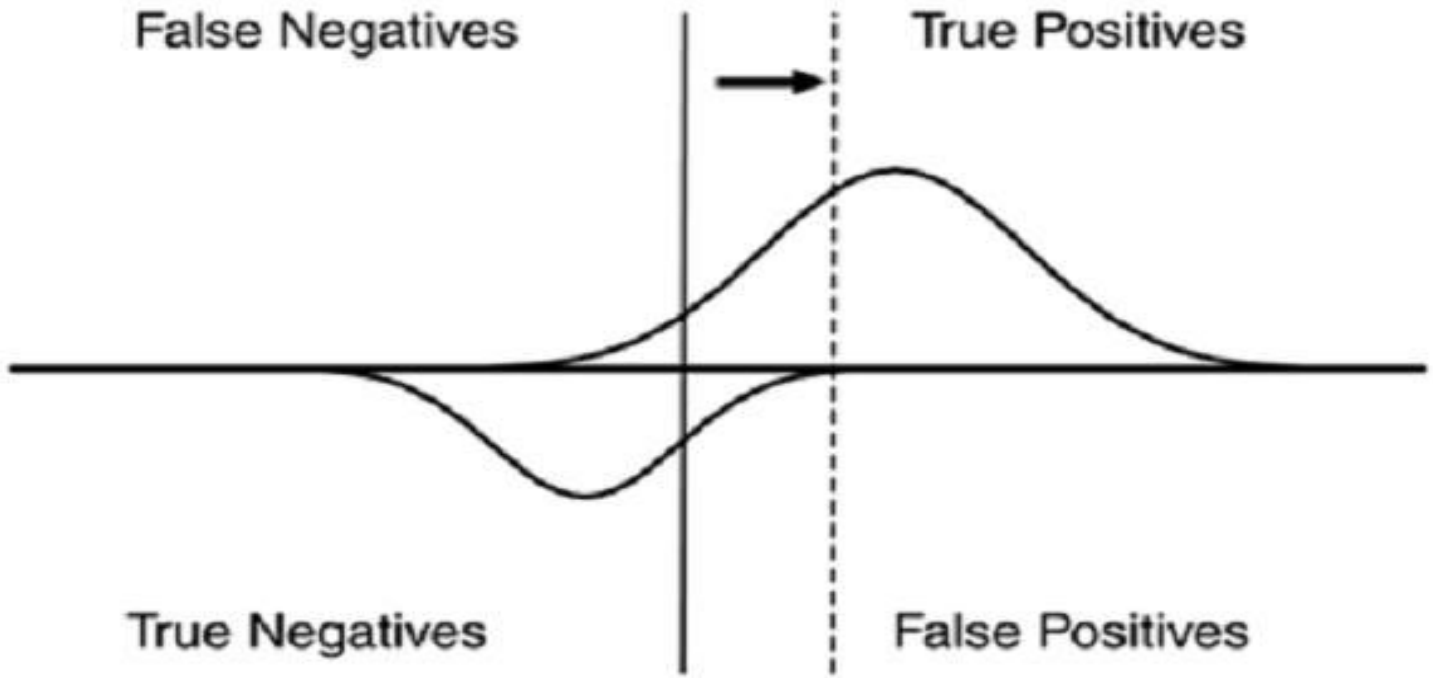


Figure 6-4. Microarray Results Analysis. Scatter plot illustrating inter-microarray variability in two identically treated microarrays, Microarray1 and Microarray 2. Ideally, all data points fall on the ID line, as illustrated by data point (A).





$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Figure 6-21. Receiver Operating Characteristic (ROC) Curves For the twotests shown here, Test A provides superior discrimination over Test B.

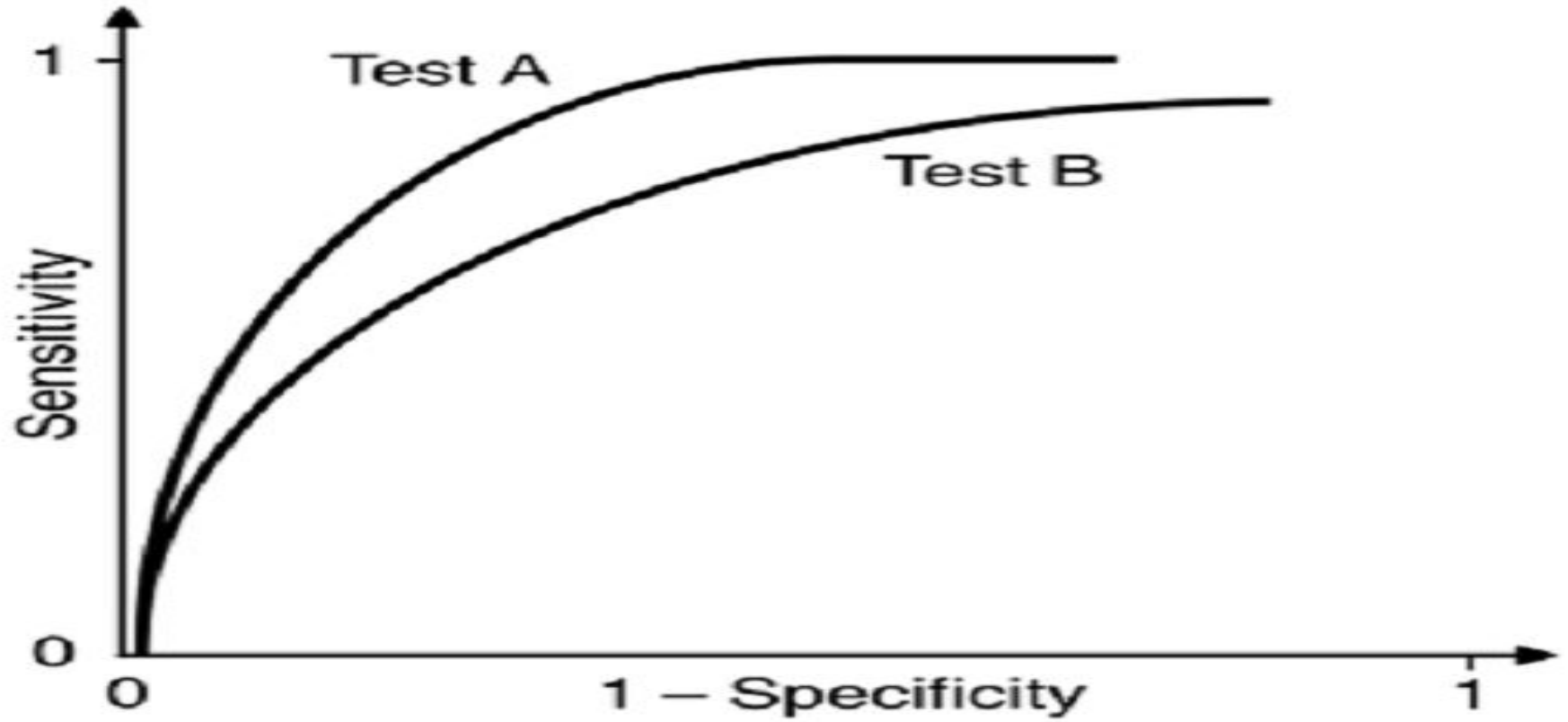


Figure 6-22. Hierarchical Clustering. Data in the expression matrix can be clustered to an arbitrary depth.

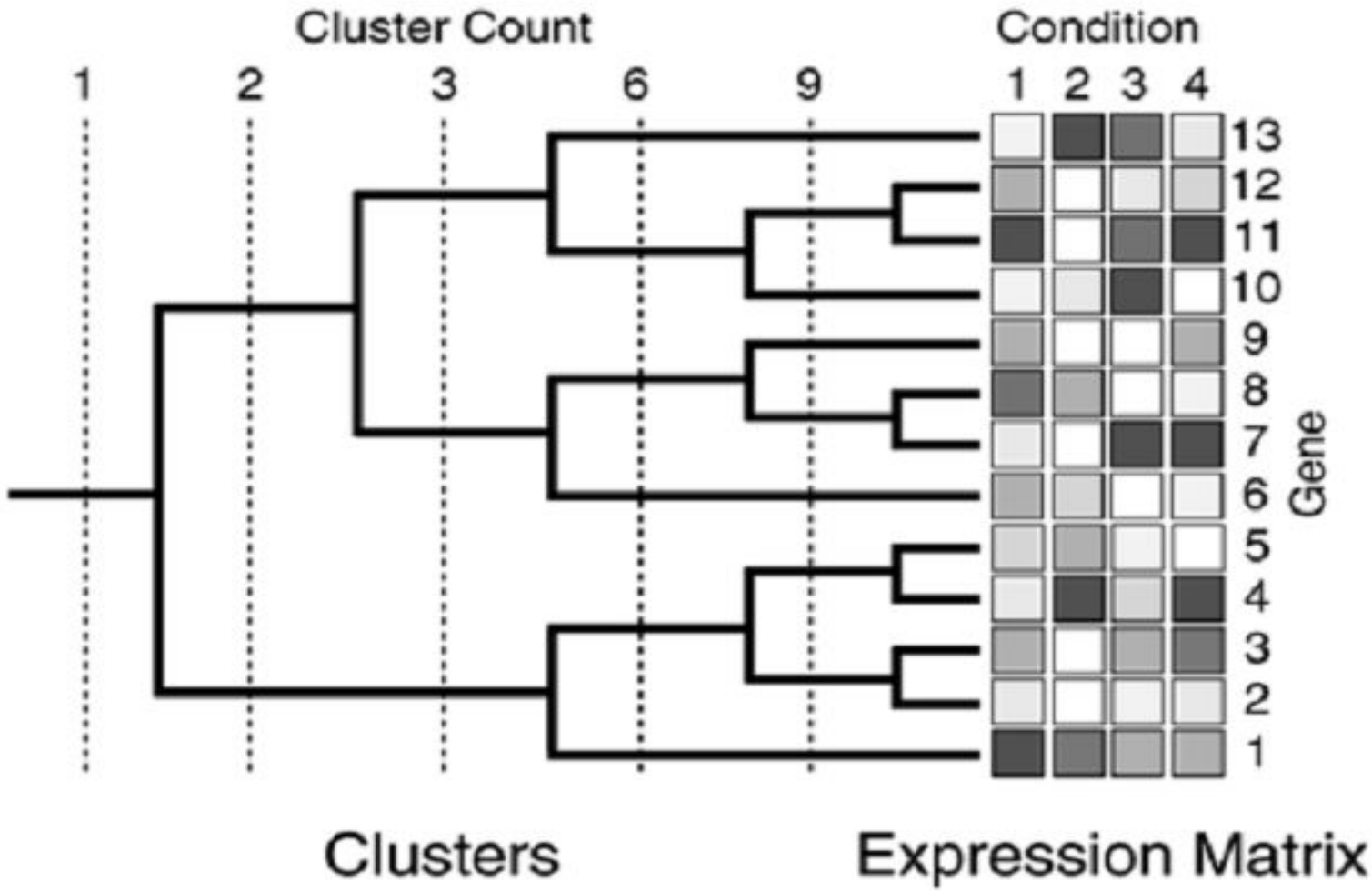


Figure 6-23. K-Means Clustering. Items are assigned to the nearest cluster and the cluster centers (squares) are recalculated. This process is repeated until the cluster centers don't change significantly. In the end, there are two clusters, one with filled circles and one with empty circles.

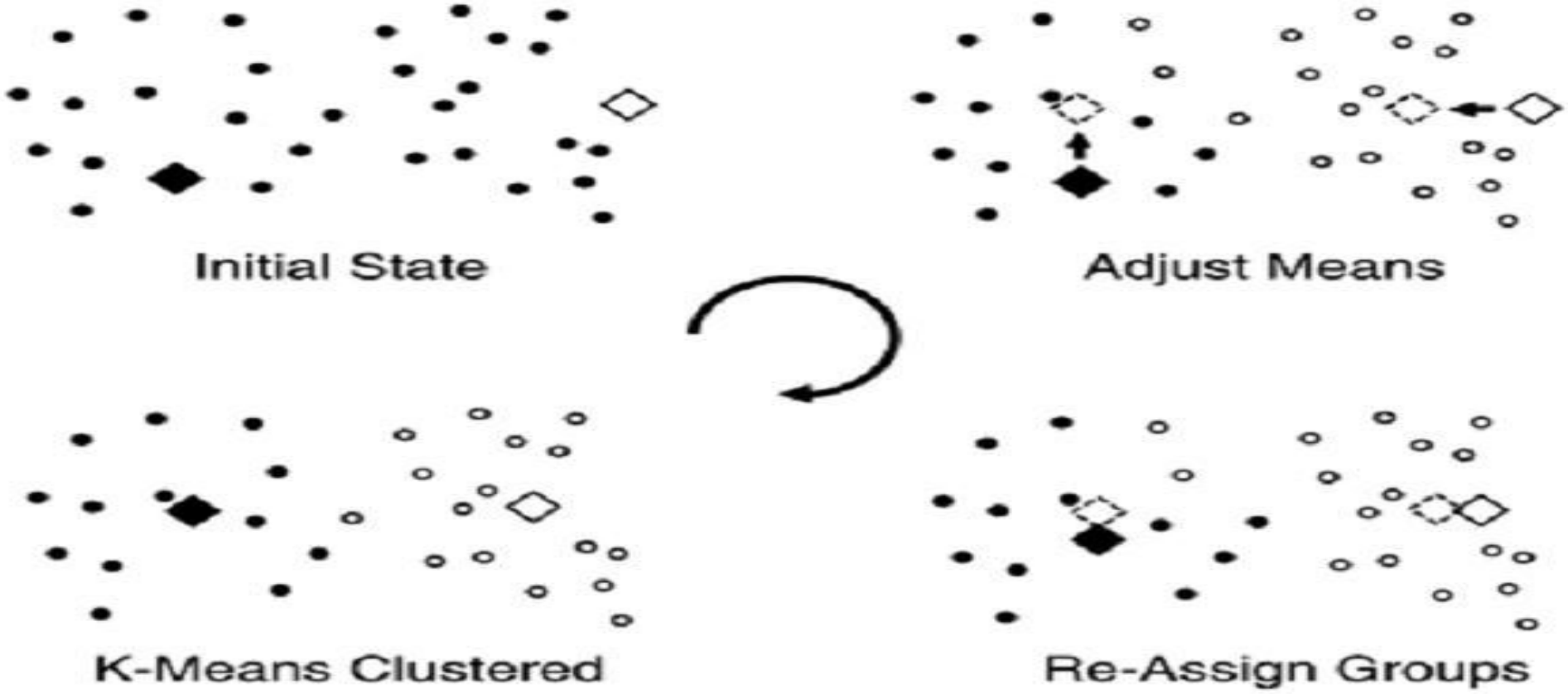


Figure 6-24. Bayes' Theorem Example. The data points A, B, and can be classified using Bayes' Theorem.

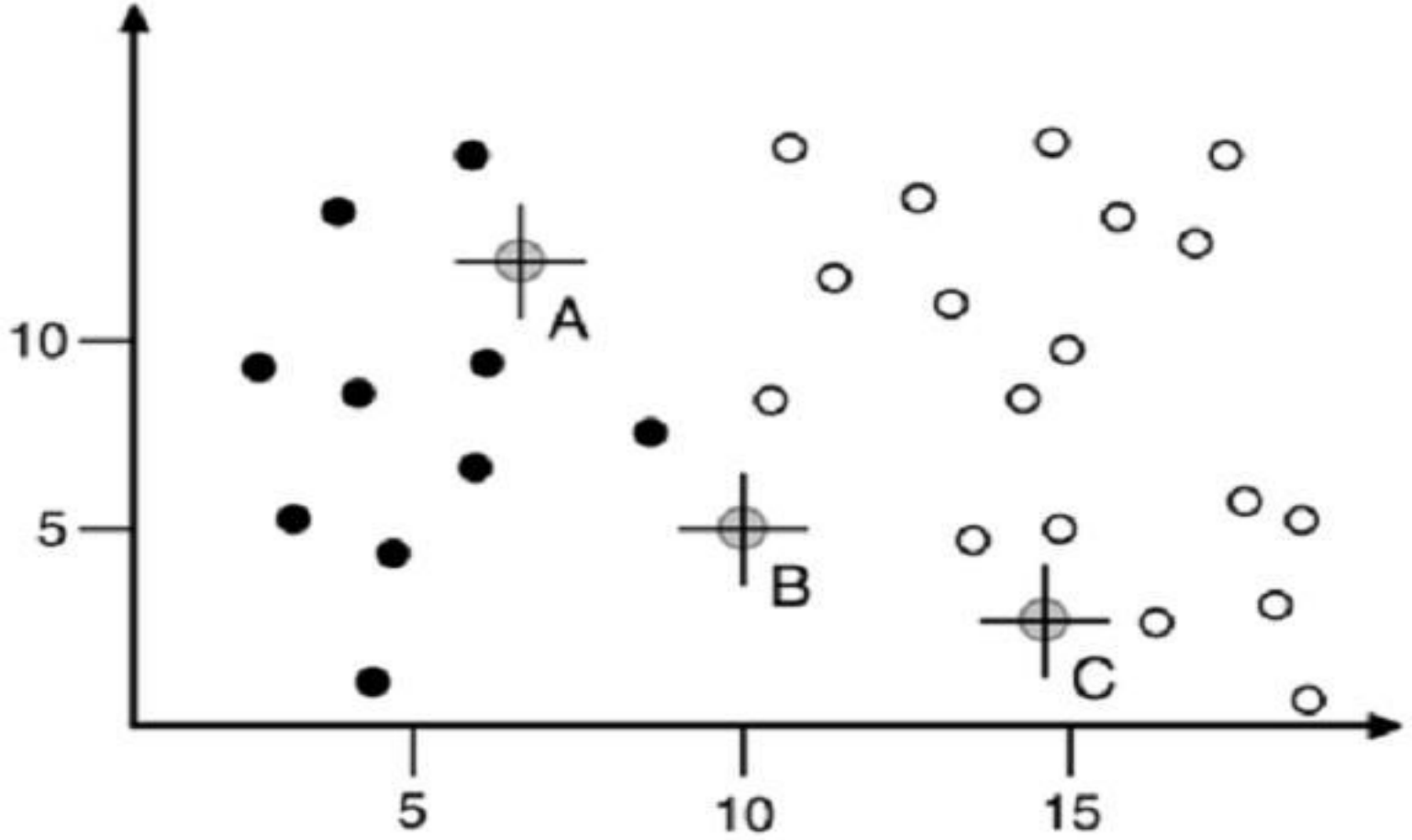
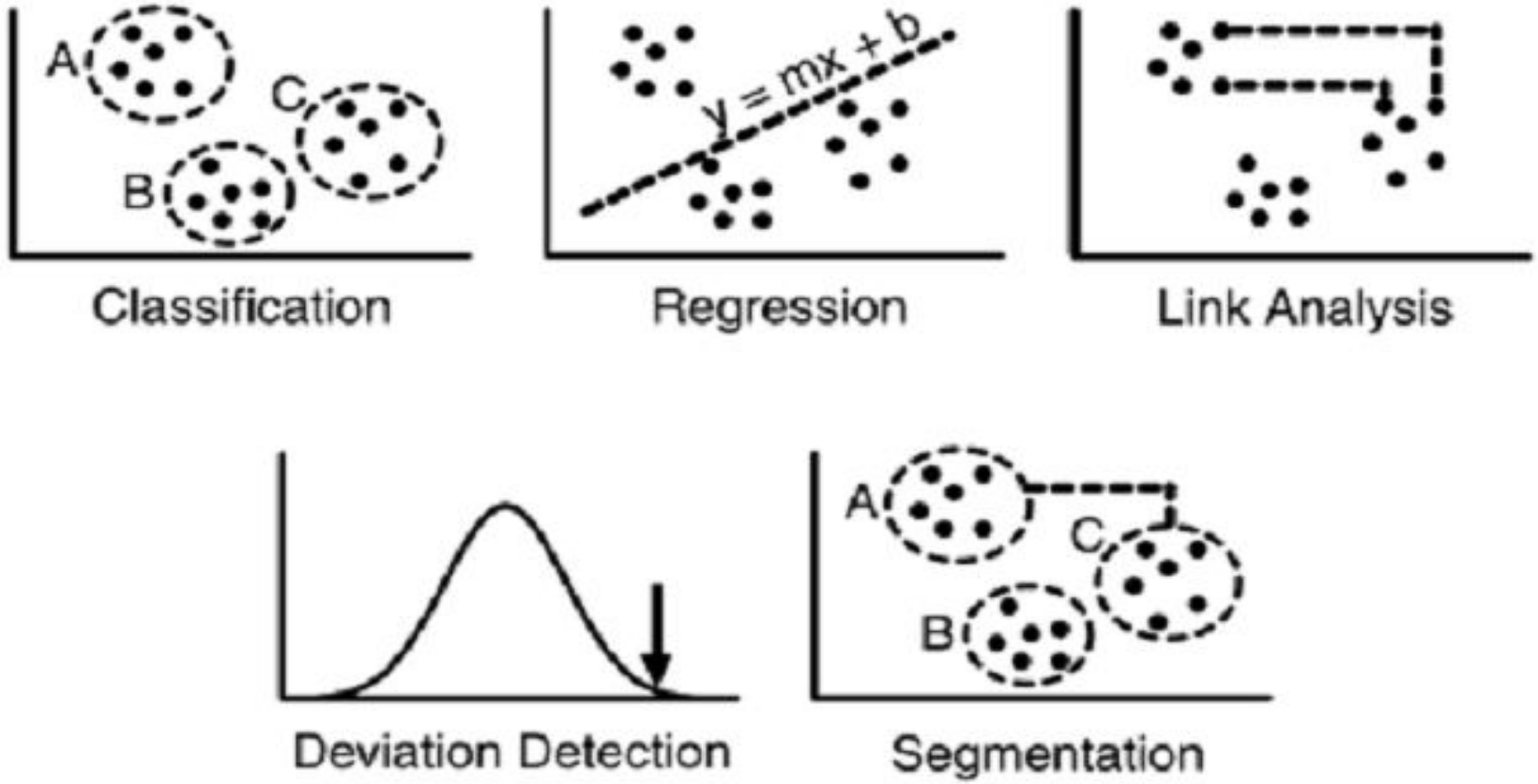


Figure 7-2. Data Mining Methods. Classification—Mapping to a class or group. Regression—Statistical analysis. Link Analysis—Correlation of data. Deviation Detection—Difference from the norm. Segmentation—Similarity function.



OVERVIEW OF IMAGE PROCESSING, TRANSFORMATION AND NORMALIZATION

Image processing involves the following steps:

- 1. *Identification of the spots and distinguishing them from spurious signals.*
- 2. *Determination of the spot area to be surveyed, determination of the local region to estimate background hybridization.*
- 3. *Reporting summary statistics and assigning spot intensity after subtracting for background intensity.*

Intensity Signal

- Most approaches use the spot median value, with the background median value subtracted from it, as the metric to represent spot intensity.
- The median intensity is a value where half the measured pixels have intensities greater than this value and the other half of the measured pixels have intensities less than this value
- The other method is to use total intensity values, which has an advantage of being insensitive to misidentification of spots (as few more pixels with zero value in the background will not affect the total intensity), but has a disadvantage of being prone to be skewed by a few pixels with extreme intensity values

- Another consideration in image processing is the number of pixels to be included for measurement in the spot image. For many scanners, the default pixel size is 10µm.
- This means that an average spot of diameter of 200µm will have ~314 pixels.
- However, for a smaller spot diameter, it is better to use a smaller pixel size to ensure enough pixels are sampled. Most scanners now allow the pixel size of 5µm.

2.2 Expression ratios: the primary comparison

- We saw that the relative expression level for a gene can be measured as the amount of red or green light emitted after excitation. The most common metric used to relate this information is called expression ratio. It is denoted here as *Tk and defined as:*

$$T_k = \frac{R_k}{G_k}$$

- For each gene *k on the array, where Rk represents the spot intensity metric for the test* sample and *Gk represents the spot intensity metric for the reference sample. As mentioned* above, the spot intensity metric for each gene can be represented as a total intensity value or a background subtracted median value. If we choose the median pixel value, then the median expression ratio for a given spot is:

$$T_{median} = \frac{R_{median}^{spot} - R_{median}^{background}}{G_{median}^{spot} - G_{median}^{background}}$$

- where R_{median}^{spot} *and* $R_{median}^{background}$
- *are the median intensity values for the spot and background* respectively, for the test sample.

References

Lecture was prepared using following study material

- Roger Bumgarner, DNA microarrays: Types, Applications and their future Curr Protoc Mol Biol. 2013 January ; 0 22: Unit–22.1.. doi:10.1002/0471142727.mb2201s101.
- **Madan Babu, M., 2015. *An Introduction to Microarray Data Analysis*. [online] Mrc-lmb.cam.ac.uk. Available at: <http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/> [Accessed 3 December 2015].**