BSC 4934: Q'BIC Capstone Workshop

Giri Narasimhan

ECS 254A; Phone: x3748

giri@cs.fiu.edu

http://www.cs.fiu.edu/~giri/teach/BSC4934_Su10.html

July 2010

Sources of Variations & Experimental Errors

- Variations in cells/individuals
- Variations in mRNA extraction, isolation, introduction of dye, variation in dye incorporation, dye interference
- Variations in probe concentration, probe amounts, substrate surface characteristics
- Variations in hybridization conditions and kinetics
- Variations in optical measurements, spot misalignments, discretization effects, noise due to scanner lens and laser irregularities
- Cross-hybridization of sequences with high sequence identity
- Limit of factor 2 in precision of results
- Variation changes with intensity: larger variation at low or high expression levels

Need to Normalize data

Early Molecular Biology Contributions

- Prostate cancer: prostate-specific antigen screening
- Protein kinase inhibitors as cancer drugs
  - Gleevec: some forms of Leukemia
  - Monocloan antibody Herceptin: some forms of Breast cancer

Analyzing Microarray Data

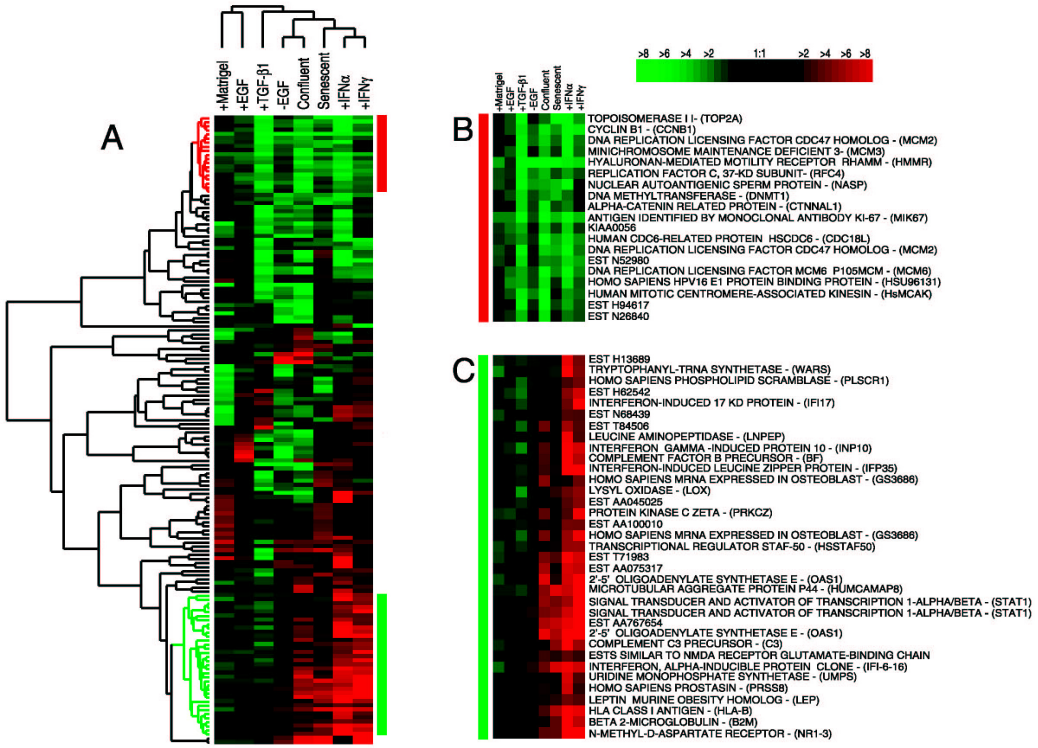
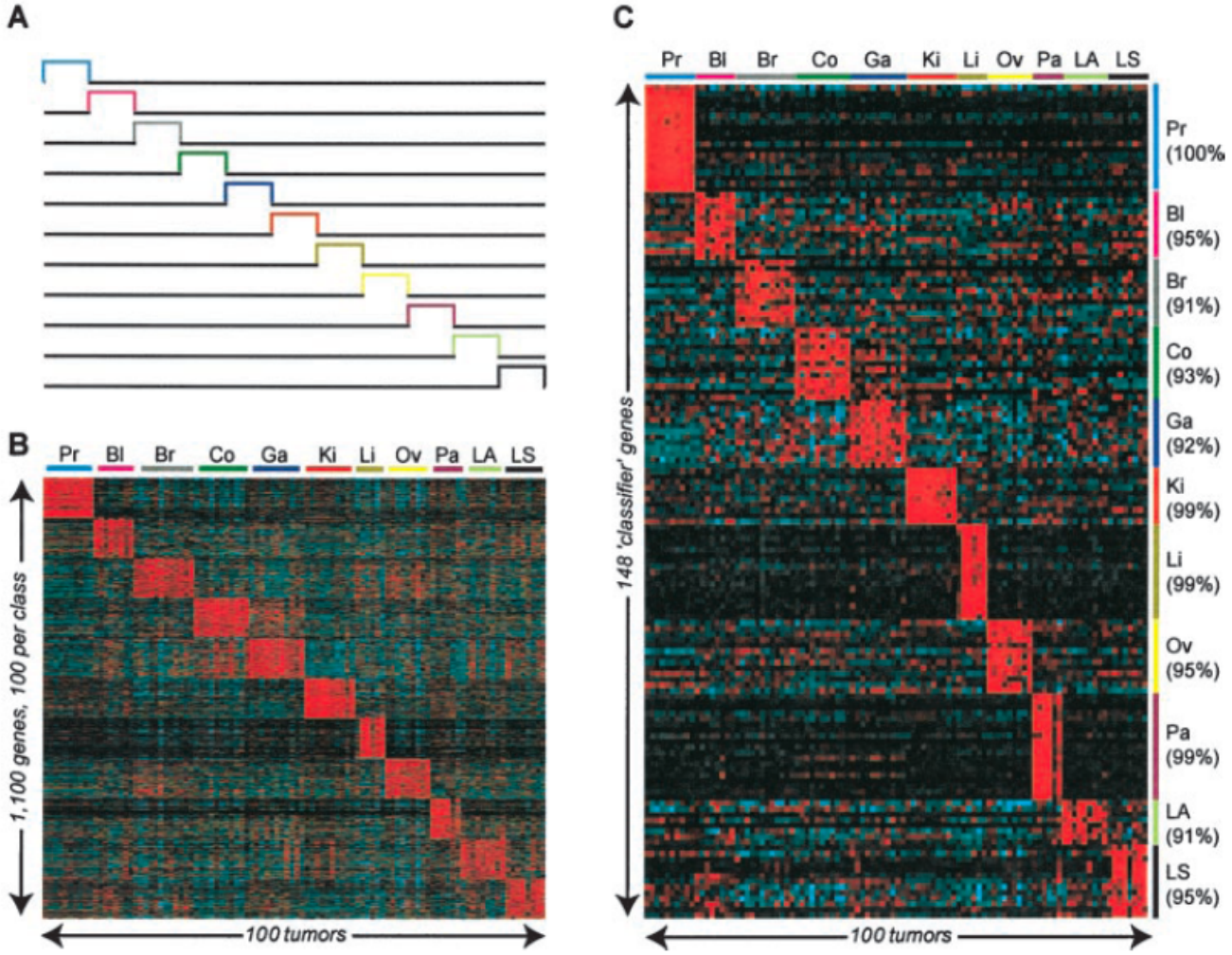


FIG. 1. (*A*) Cluster diagram of HMEC *in vitro* experiments. Each column represents a single experiment, and each row represents a single gene. Ratios of gene expression relative to HMEC control samples grown under standard conditions are shown. Green squares represent lower than control levels of gene expression in the experimental samples (ratios less than 1); black squares represent genes equally expressed (ratios near 1); red squares represent higher than control levels of gene expression (ratios greater than 1); gray squares indicate insufficient or missing data. The color saturation reflects the magnitude of the log/ratio [see scale at top right and Fig. 5 (see Supplemental data at www.pnas.org) for the full cluster diagram with all gene names]. (*B*) Expanded view of the subset of genes whose expression was decreased in association with reduced HMEC proliferation. (*C*) Expanded view of the IFN-regulated gene cluster. In many instances, multiple independent clones/cDNA representing the same gene were spotted on different locations on these microarrays, and in most cases, these copies usually clustered together, either very near each other or immediately adjacent to each other.

Microarray Data Analysis: Subtyping

Fig. 1. Selection of tumor-specific genes for cancer class prediction. *A*, schematic diagram depicting the idealized expression profile of tumor-specific genes that the method selects as classifiers. The shape of each profile represents genes that are highly expressed in each cancer type relative to all other tumors in the training set. *B*, 100 genes per tumor class (total, 1100) with the most significant scores in a Wilcoxon rank-sum test for equality were selected as likely candidates for tumor classifiers. *Pr*, prostate; *Bl*, bladder/ureter; *Br*, breast; *Co*, colorectal; *Ga*, gastroesophagus; *Ki*, kidney; *Li*, liver; *Ov*, ovary; *Pa*, pancreas; *LA*, lung adenocarcinomas; *LS*, lung squamous cell carcinoma. *C*, the final refined set of gene classifiers was generated after the genes in *B* were ranked by SVM/LOOCV accuracy. Annotations of the genes from which 110 "predictor" genes were bootstrapped are provided on our website.[4] For clarity, only 8 of 76 predictor genes for lung adenocarcinomas are depicted here. Levels of gene expression (depicted in each *row*) across all samples (*columns*) were median-centered and normalized by "Cluster" and output in "Treeview" (12). *Red*, increased gene expression; *blue*, decreased expression; *black*, median level of gene expression. The color intensity is proportional to the hybridization intensity of a gene from its median level across all samples.



Differential Analysis

- Determine differentially expressed genes
  - Need for Replication and Normalization
  - Differential Analysis: test statistics
    - Fold-change (Sample vs Control)
    - t-test
    - F-statistic
    - Other Non-parametric rank-based statistics
  - Significance of observed statistic (Permutation test)
  - False Discovery Rate
    - Multiple test corrections
  - Pattern Discovery

Pattern Discovery

- Dimensionality reduction
  - Principal Component Analysis
  - Multidimensional scaling
  - Singular-value decomposition
- Visualization methods

Pattern Discovery

Principal Component Analysis

Clustering

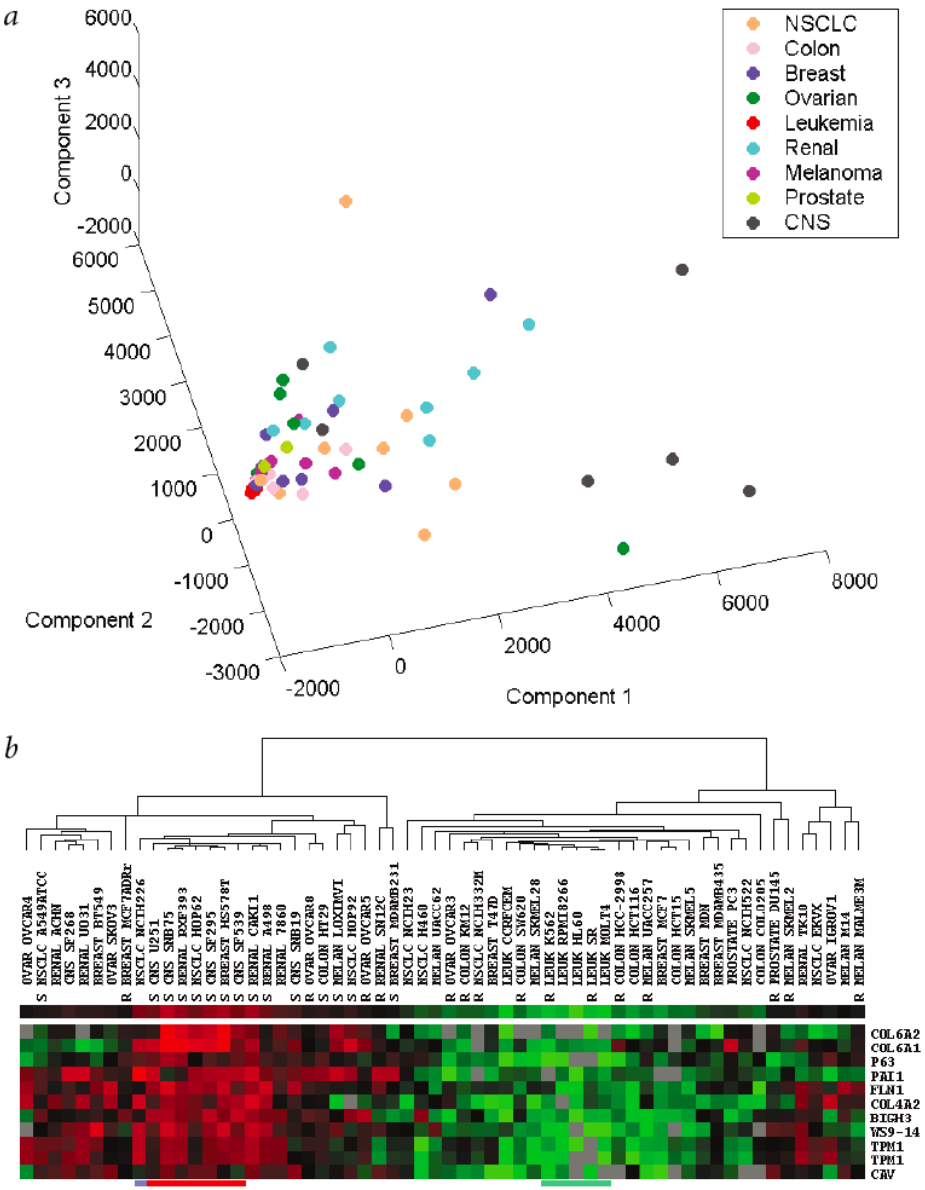


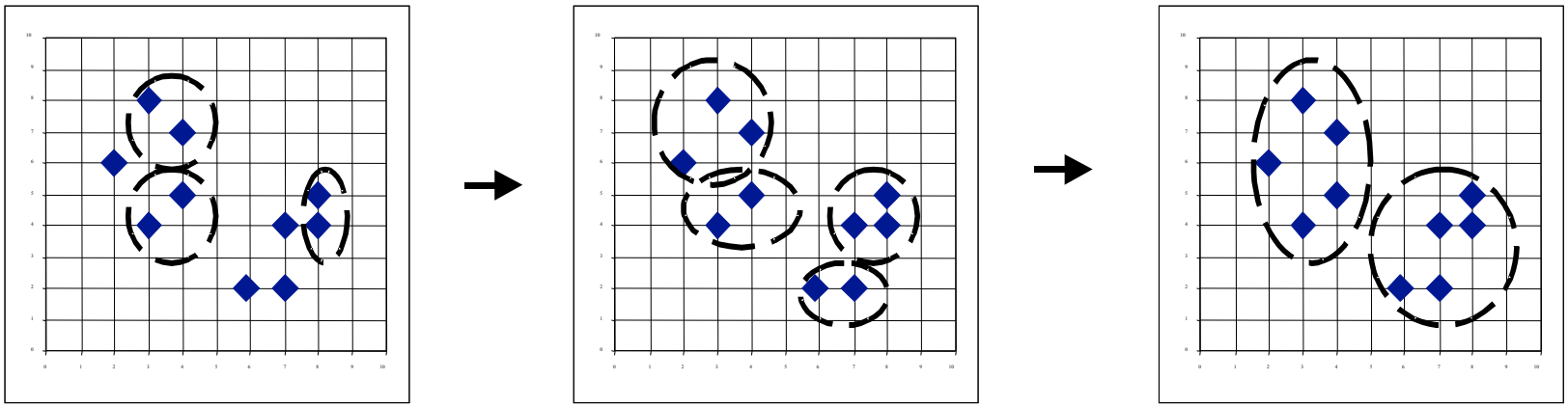
Fig. 2 Two pattern-discovery techniques. Data for both figures measure expression for 11 genes characterizing sensitivity to compound cytochalasin D in 60 cancer cell lines[97]. ***a***, The first three principal components, plotted using Matlab software (Mathworks). Apparent features include a tight cluster of leukemia samples (red dots, nearly superimposed) and the more scattered outlying cluster of CNS tumors (black dots). A single lung cancer sample (NSCLC-NCIH226) also appears as an outlier — the solitary orange dot at the top. ***b***, Hierarchical clustering of the same data, using Cluster/TreeView (http://rana.lbl.gov/EisenSoftware.htm). Names of samples extremely sensitive or resistant to cytochalasin D (see Supplementary information) are prefixed 'S' and 'R' respectively. The samples fall into two main clusters, roughly, but not perfectly, separating the sensitive and resistant samples. As in ***a***, fine structure shows a tight leukemia cluster (underlined in green) and a tight CNS cluster (underlined in red), but does not suggest that the CNS cluster or NSCLC-NCIH226 (underlined in blue) are outliers. Apparent in both ***a*** and ***b*** is the relative heterogeneity of the breast cancer cell lines.

merging the two closest clusters is repeated until a single cluster remains. This arranges the data into a tree structure that can be broken into the desired number of clusters by cutting across the tree at a particular height. Tree structures are easily viewed and understood (Fig. 2*b*), and the hierarchical structure provides potentially useful information about the relationships between clusters. Trees are known to reveal close relationships very well. However, as

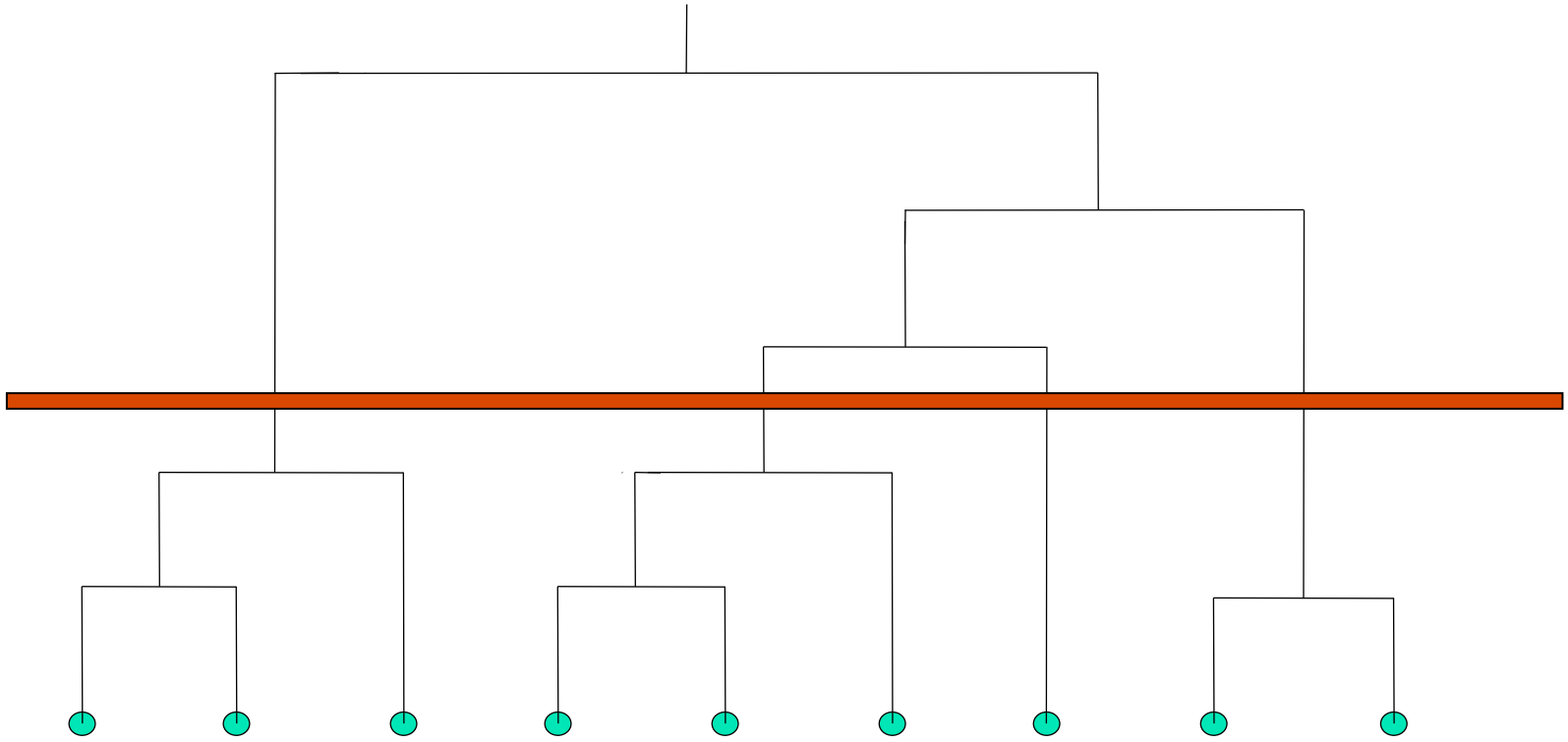
Clustering

- Clustering is a general method to study patterns in gene expressions.
- Several known methods:
  - Hierarchical Clustering (Bottom-Up Approach)
  - K-means Clustering (Top-Down Approach)
  - Self-Organizing Maps (SOM)

Hierarchical Clustering: Example



A Dendrogram



Hierarchical Clustering [Johnson, SC, 1967]

- Given n points in R^d, compute the distance between every pair of points
- While (not done)
  - Pick closest pair of points s_i and s_j and make them part of the same cluster.
  - Replace the pair by an average of the two s_{ij}

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

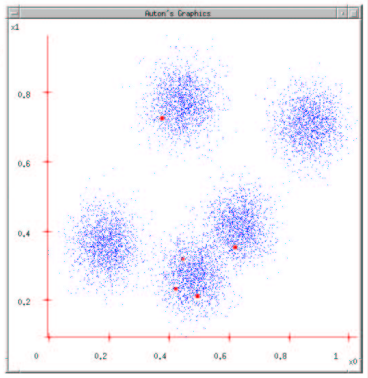
K-Means Clustering: Example

Example from Andrew Moore's tutorial on Clustering.

Start

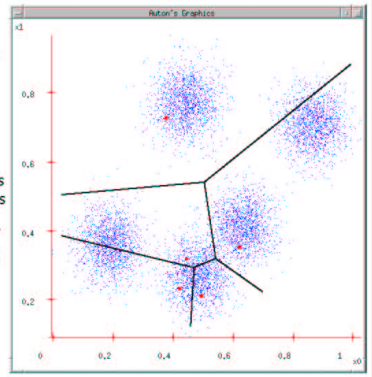
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



K-means

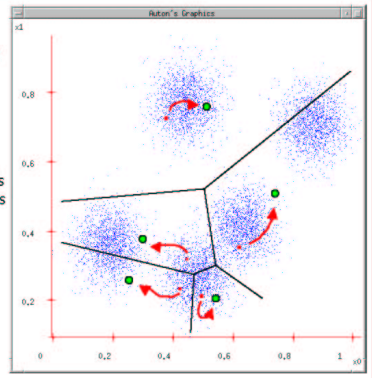
1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



4

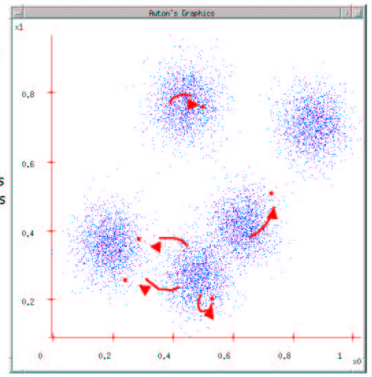
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns...
5. ...and jumps there
6. ...Repeat until terminated!



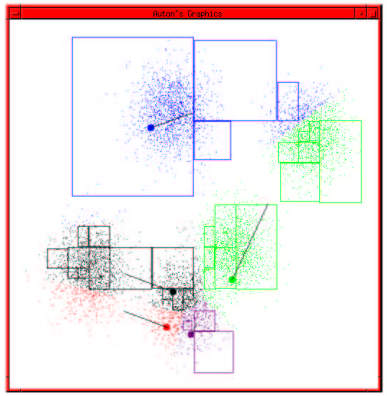
5

K-means Start

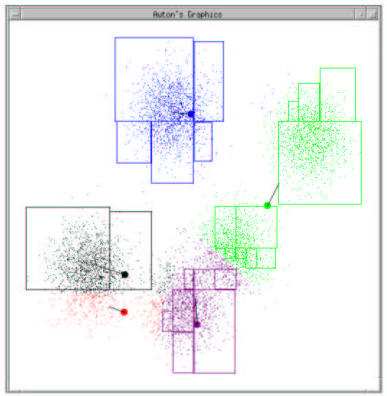
Advance apologies: in Black and White this example will deteriorate

Example generated by Dan Pelleg's super-duper fast K-means system:

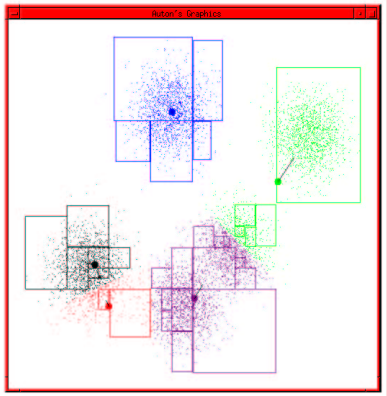
Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html*)*



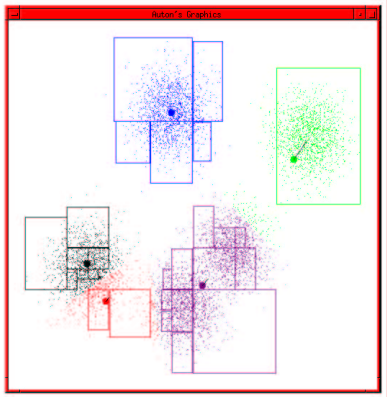
K-means continues ...

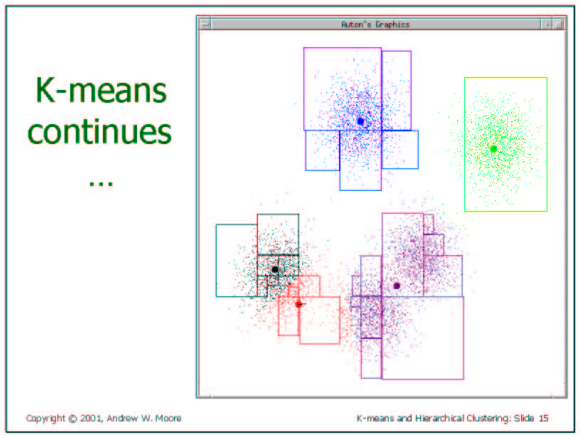


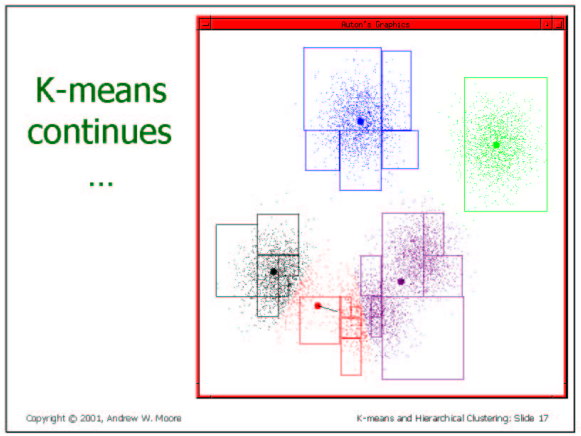
K-means continues ...

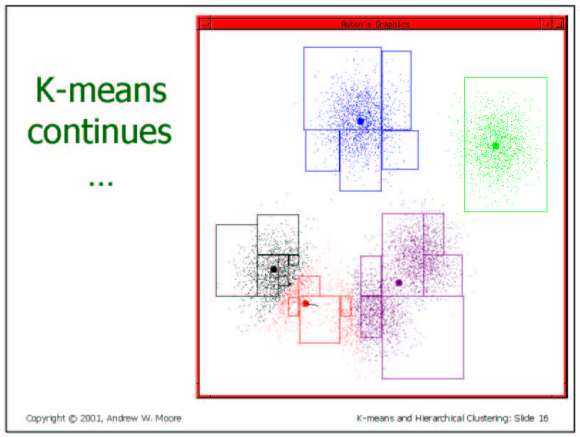


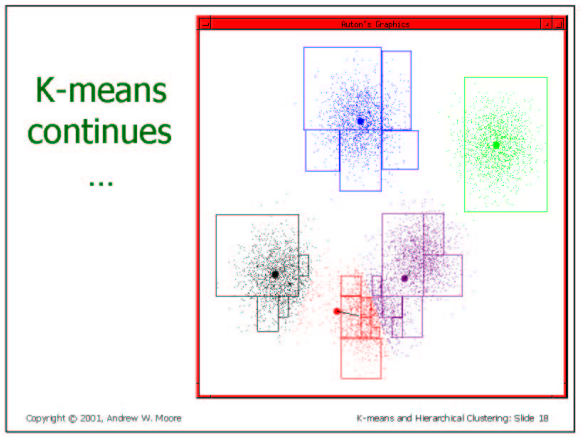
K-means continues ...




K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 15


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 17


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 16


K-means
continues
...
Copyright © 2001, Andrew W. Moore
K-means and Hierarchical Clustering: Slide 18

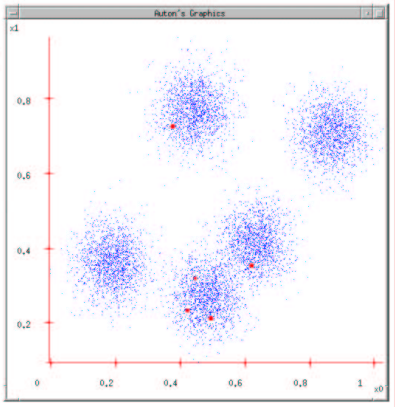
8

9

Start

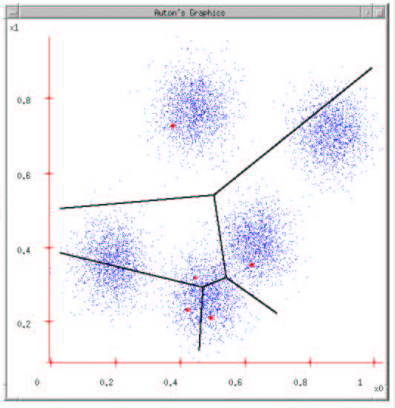
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



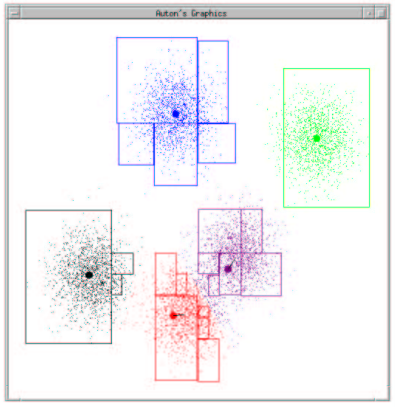
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)

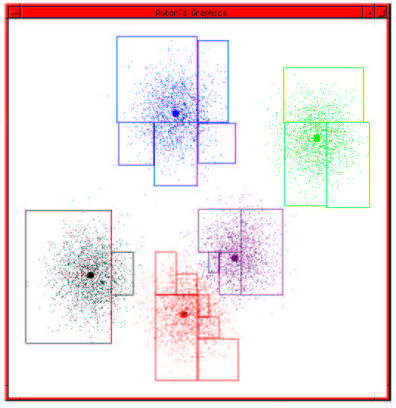


4

K-means continues ...



K-means terminates



10

End

K-Means Clustering [McQueen '67]

Repeat

- Start with randomly chosen cluster centers
- Assign points to give greatest increase in score
- Recompute cluster centers
- Reassign points

until (no changes)

Try the applet at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

Comparisons

- Hierarchical clustering
  - Number of clusters not preset.
  - Complete hierarchy of clusters
  - Not very robust, not very efficient.
- K-Means
  - Need definition of a mean. Categorical data?
  - More efficient and often finds optimum clustering.

Class Prediction

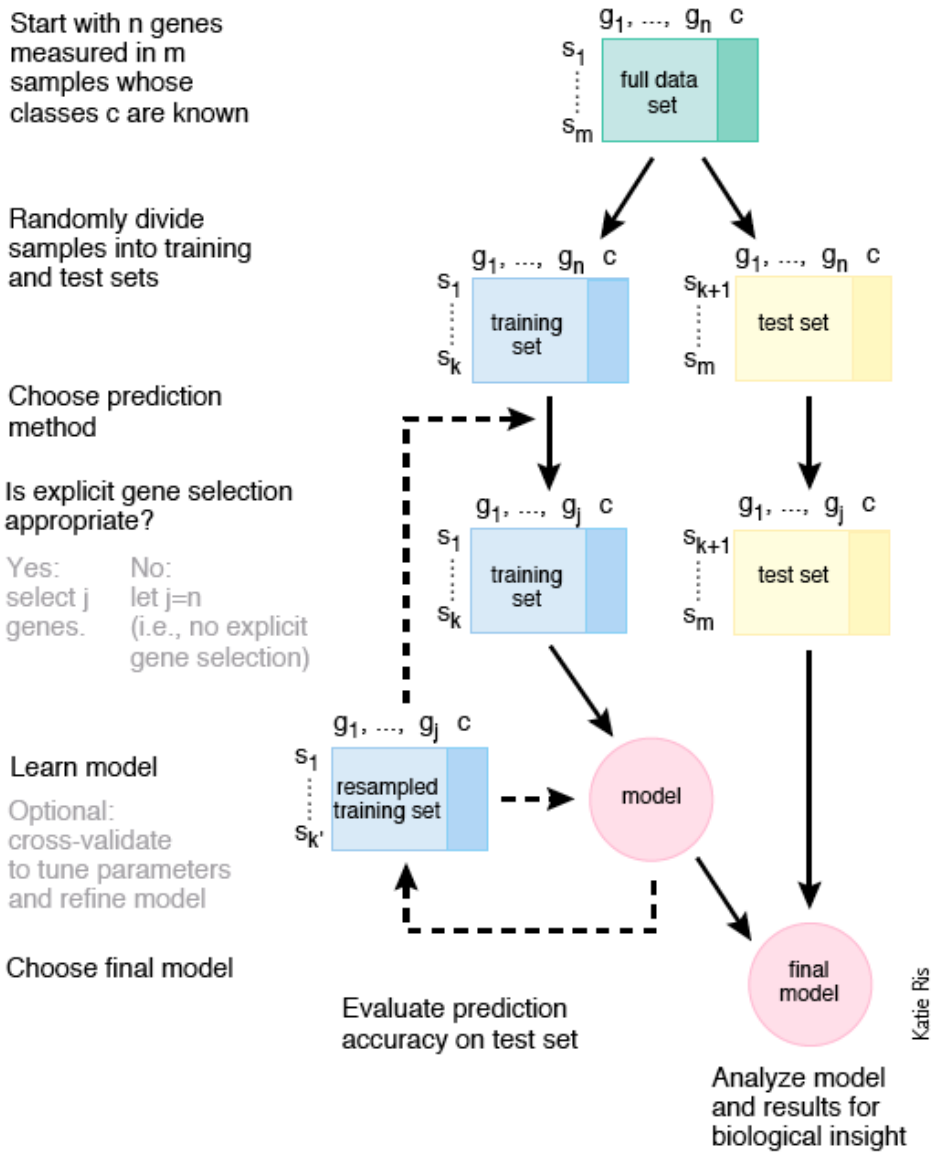


Fig. 3 An overview of the process for building a prediction model to classify samples. The partition into training and test data is ideally chosen at random across the entire set of samples. Many prediction methods require tuning some parameter (such as the number of genes, the number of nearest-neighbors to consider, or the number of decision trees built). This choice is often evaluated by cross-validation — the process of repeatedly removing smaller test sets from the training set, building new models (starting with the gene selection process) with the remaining data, and evaluating performance across all the different models built. For example, "leave-one-out cross validation" (also called "*n*-way") builds *n* models, each using *n*–1 training examples and evaluated on the remaining one; the accuracy for predicting all *n* samples is reported. Observing that predictors may succeed by chance even in cross-validation, Radmacher *et al.* suggest using permutation testing to determine the significance of the observed results[98]. Ultimately the final model, perhaps chosen during the cross-validation process, is then tested on entirely new data not used in the model generation process. The model itself, as well as the prediction results and the influential genes, may yield new biological insights.

Katie Ris

Class Prediction Methods

- Decision Trees
- Support Vector Machines (SVM)
- k-NN or k-nearest neighbor method
- Fisher's linear discriminant method
- Neural Networks
- Self-Organizing Maps
- Ensemble methods
  - Boosting
  - Bagging

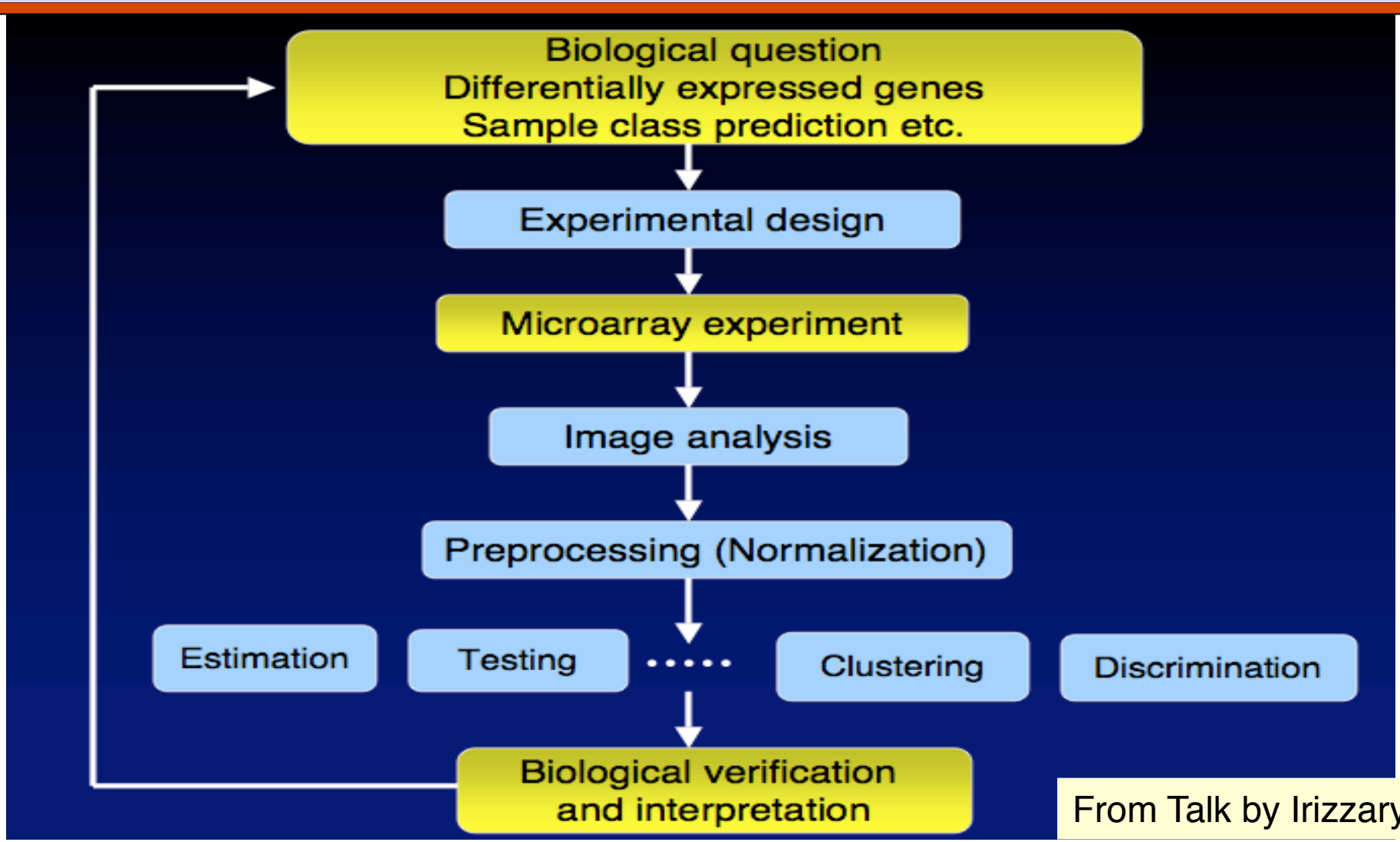
Functional Biases, Pathways & Networks

- Over/Under-representation of functional groups of genes
- Over/Under-representation of genes involved in functional pathways
- Inferring of regulatory relationships
- Inferring of protein-protein interactions

Reading

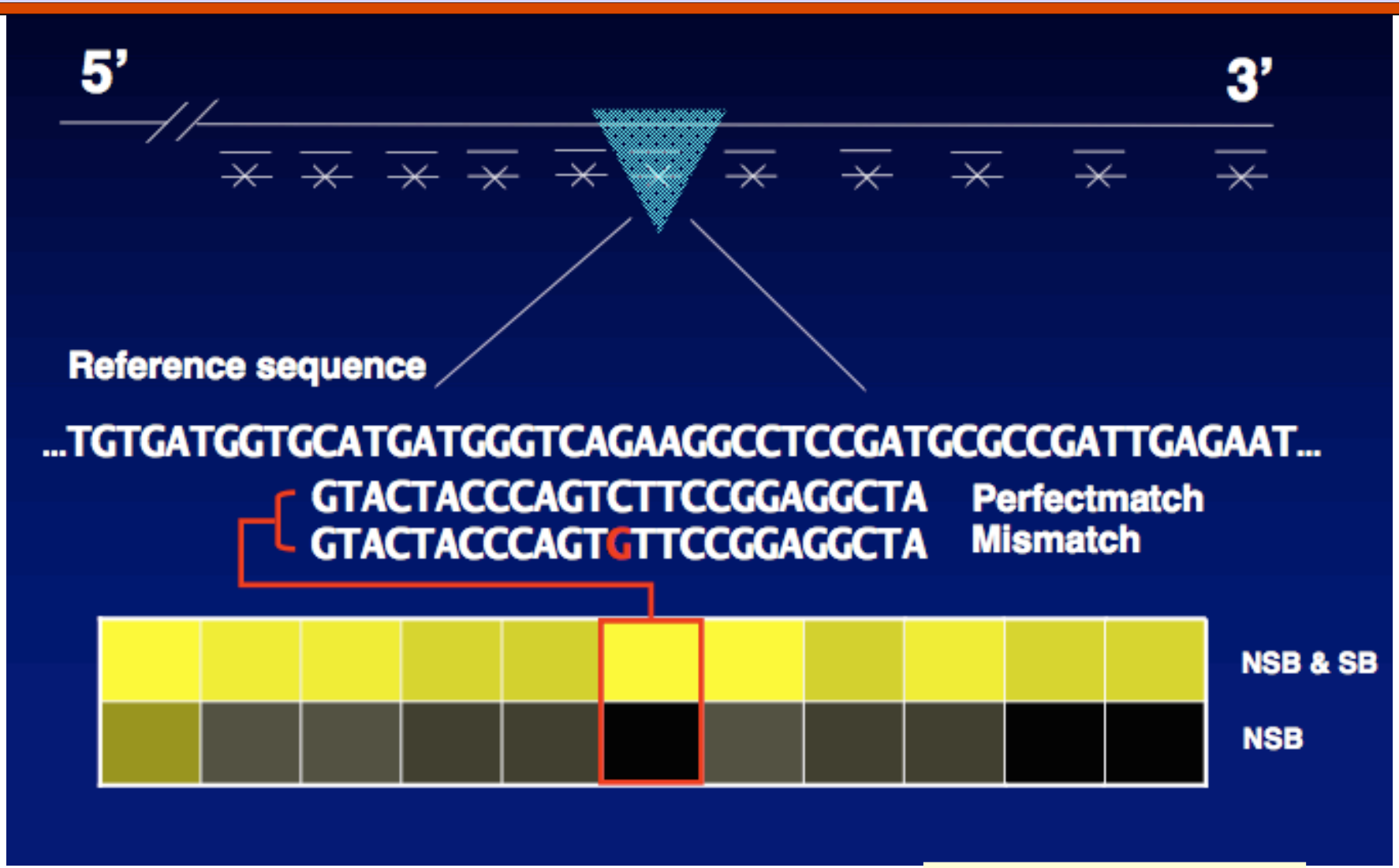
- The following slides come from a series of talks by Rafael Irizzary from Johns Hopkins
- Much of the material can be found in detail in the following papers from [http://www.biostat.jhsph.edu/~ririzarr/papers/]
  - Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics. Vol. 4, Number 2: 249-264.
  - Bolstad, B.M., Irizarry RA, Astrand, M, and Speed, TP (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics. 19(2):185-193.

Inference Process



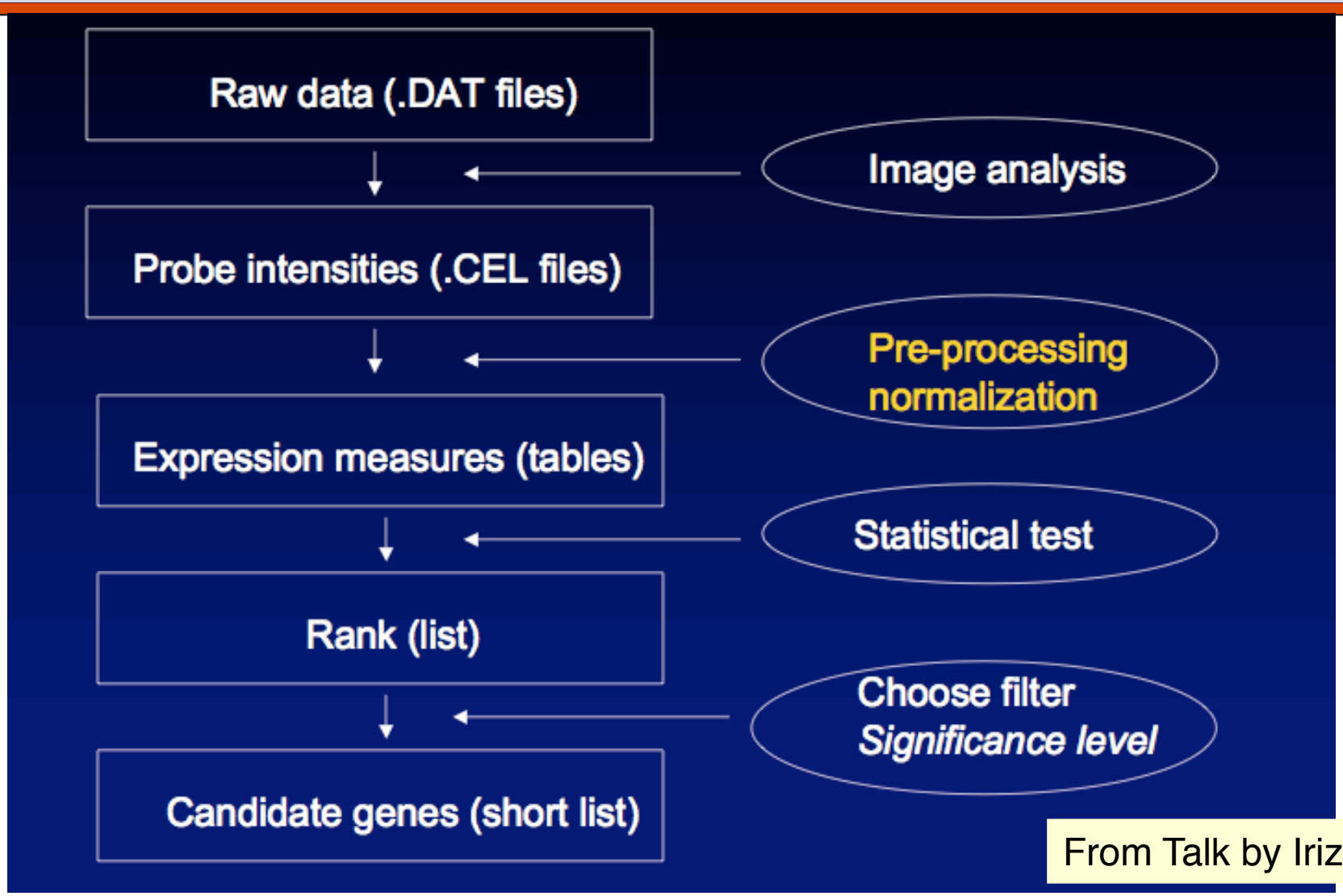
From Talk by Irizzary

Affymetrix Genechip Design



From Talk by Irizzary

Workflow: Analyzing Affy data



From Talk by Irizzary

Affy Files

- **DAT** file: image file, about 10 million pixels, 30-50 MB
- **CEL** file: cell intensity file with probe level PM and MM values
- **CDF** file: chip description file describing which probes go in which probe sets and the location of probe-pair sets (genes, gene fragments, ESTs)

From Talk by Irizzary

Image analysis & Background Correction

- Each probe cell: 10 X 10 pixels
- Gridding estimates location of probe cell centers
- Signal is computed by
  - Ignoring outer 36 pixels leaving a 8 X 8 pixel area
  - Taking the 75 percentile of the signal from the 8 X 8 pixel area
- Background signal is computed as the average of the lowest 2% probe cell values, which is then subtracted from the individual signals

From Talk by Irizzary

Standard Normalization Procedure

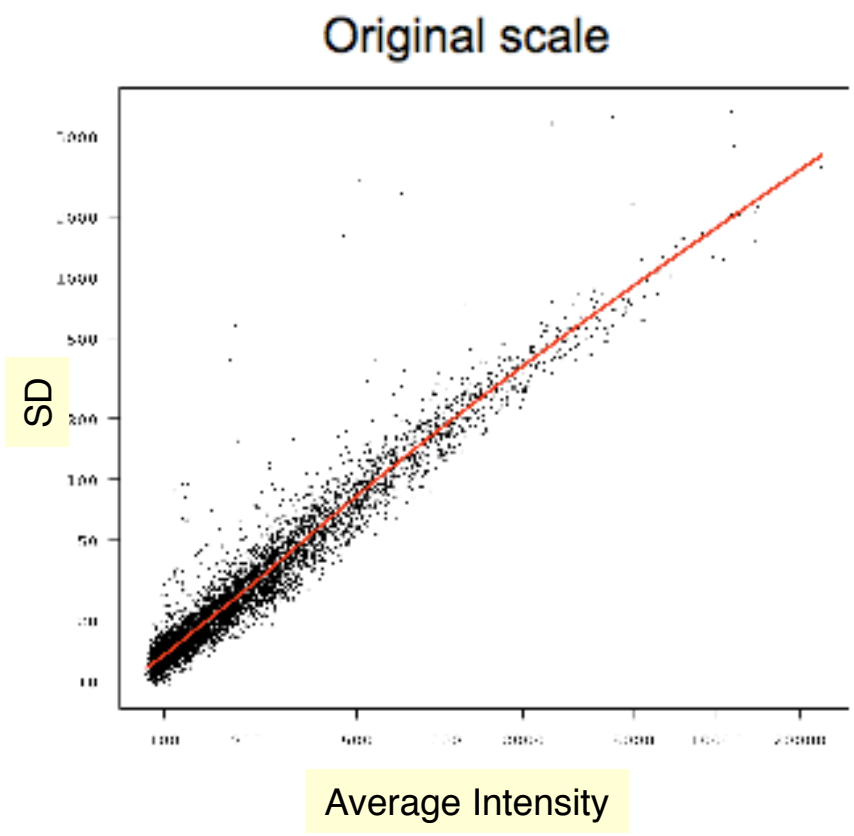
- Log-transform the data
- Ensure that the average intensity and the standard deviation are the same across all arrays.
- This requires the choice of a baseline array, which may or may not be obvious.

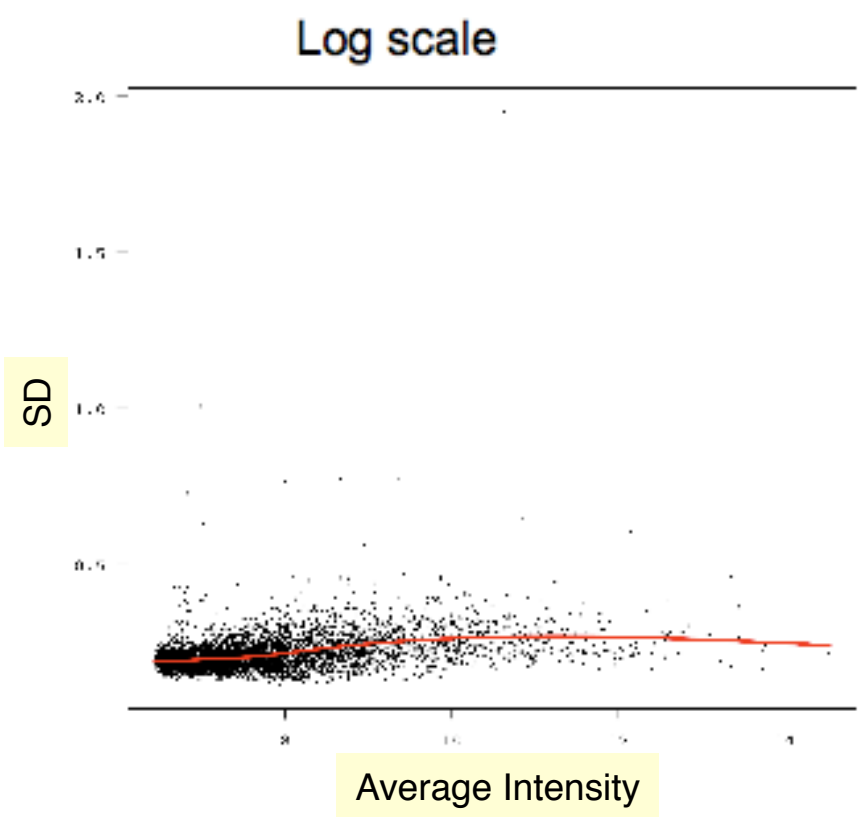
Analyzing Affy data

- MAS 4.0
  - Works with PM-MM
  - Negative values result very often
  - Very noisy for low expressed genes
  - Averages without log-transformation
- dChip [Li & Wong, PNAS **98**(1):31-36]
  - Accounts for probe effect
  - Uses non-linear normalization
  - Multi-chip analysis reveals outliers
- MAS 5.0
  - Improves on problems with MAS 4.0

From Talk by Irizzary

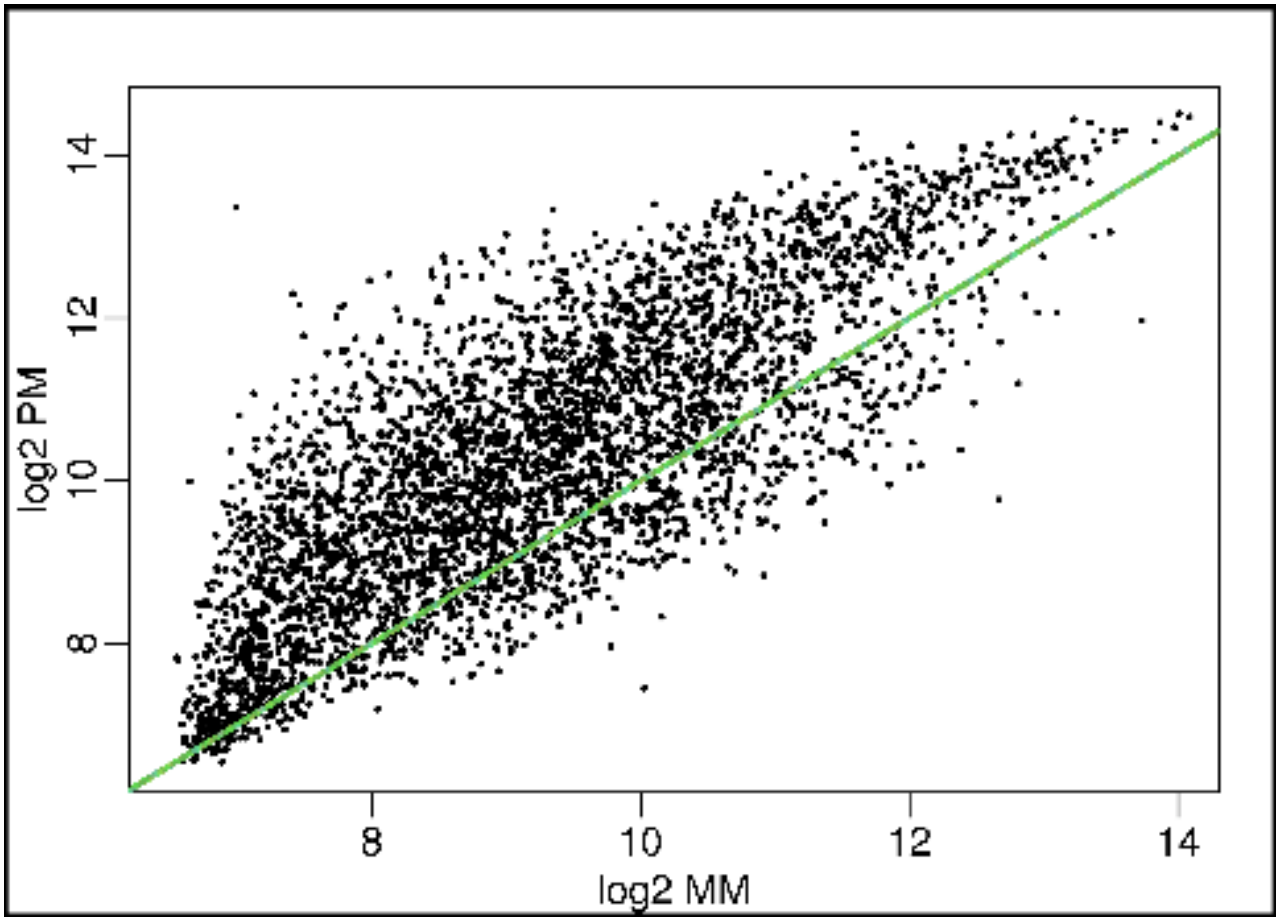
Why you use log-transforms?





From Talk by Irizzary

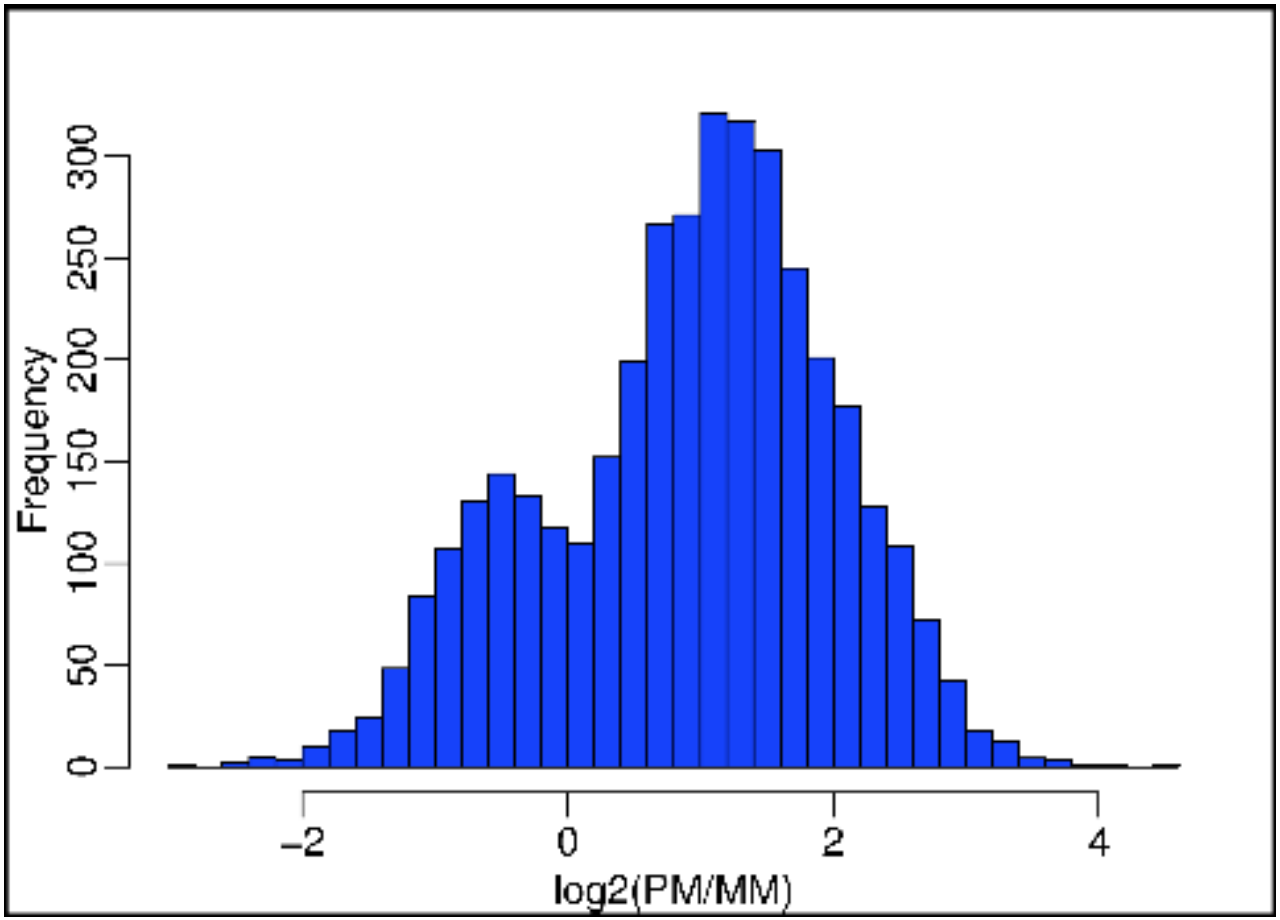
Problem with using (transformed) PM-MM



Sometimes MM is larger than PM!

From Talk by Irizzary

Bimodality for large expression values



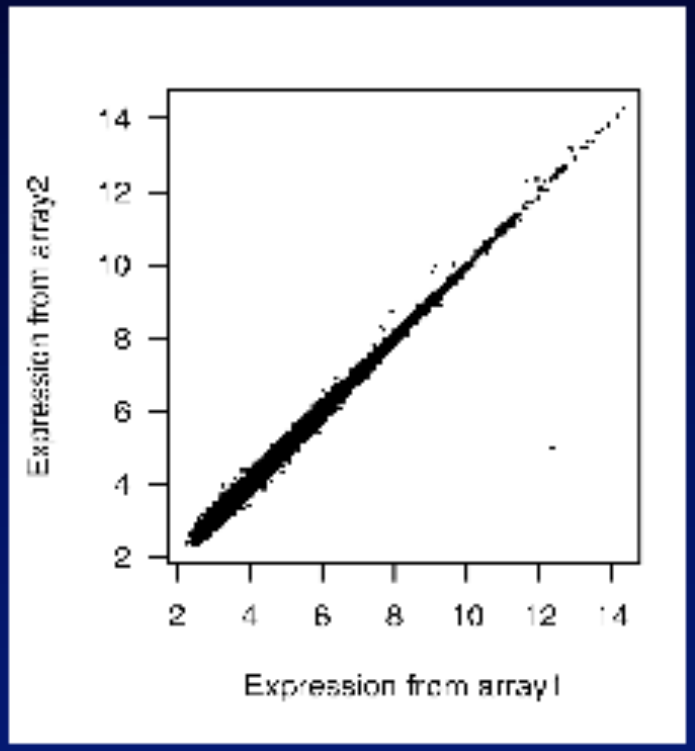
From Talk by Irizzary

MAS 5.0

- **MAS 5.0** is Affymetrix software for microarray data analysis.
- Ad hoc background procedure used
- Summarization: Averaging over multiple probes
- For summarization, MAS 5.0 uses:
  - **Signal = TukeyBiweight{log($PM_j - MM_j^*$)}**
  - Tukey Biweight: $B(x) = (1 - (x/c)^2)^2$, if $x<c$
    $= 0$ otherwise
- Ad hoc scale normalization used

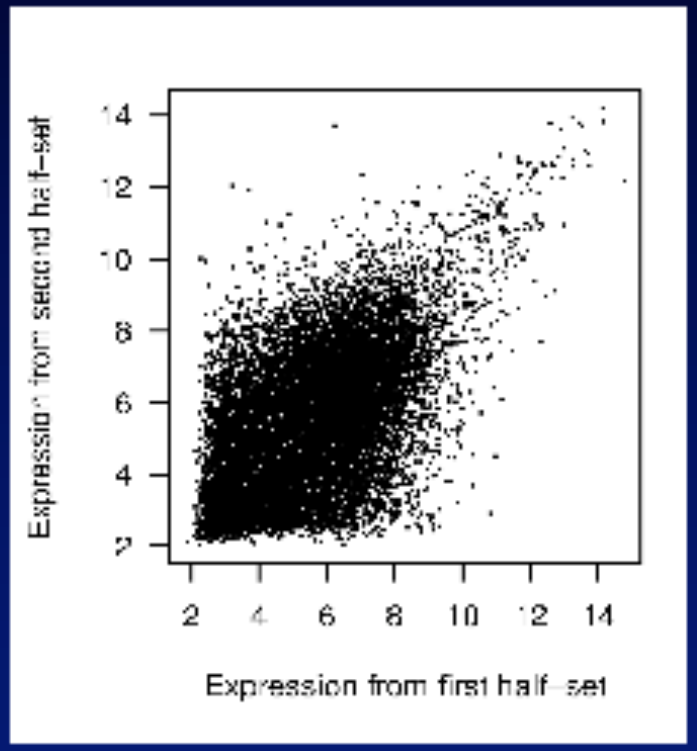
From Talk by Irizzary & PhD thesis by Astrand

2 replicate arrays



Expression from corresponding probes are highly correlated

Correlation is higher than 0.99

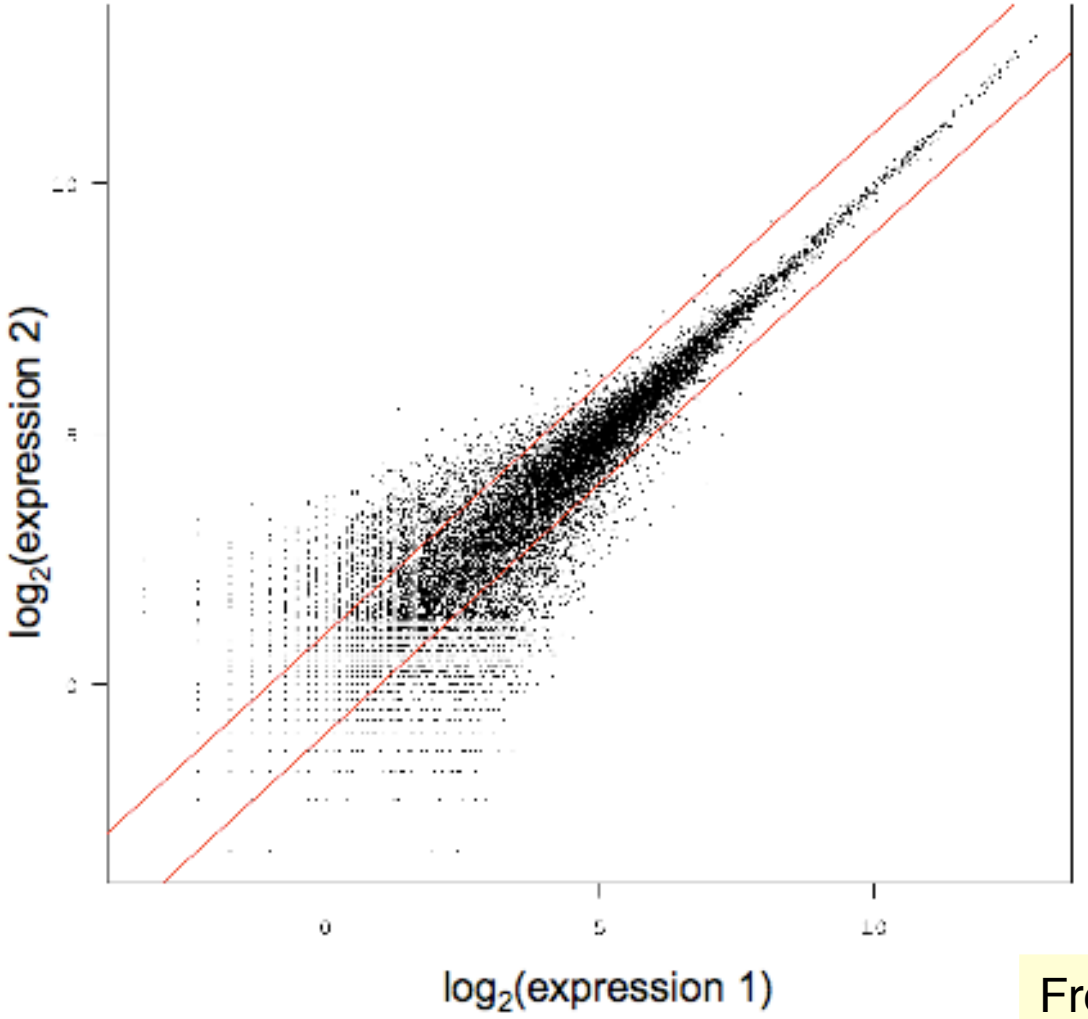


Expression not correlated when probes randomly partitioned

Correlation drops to 0.55

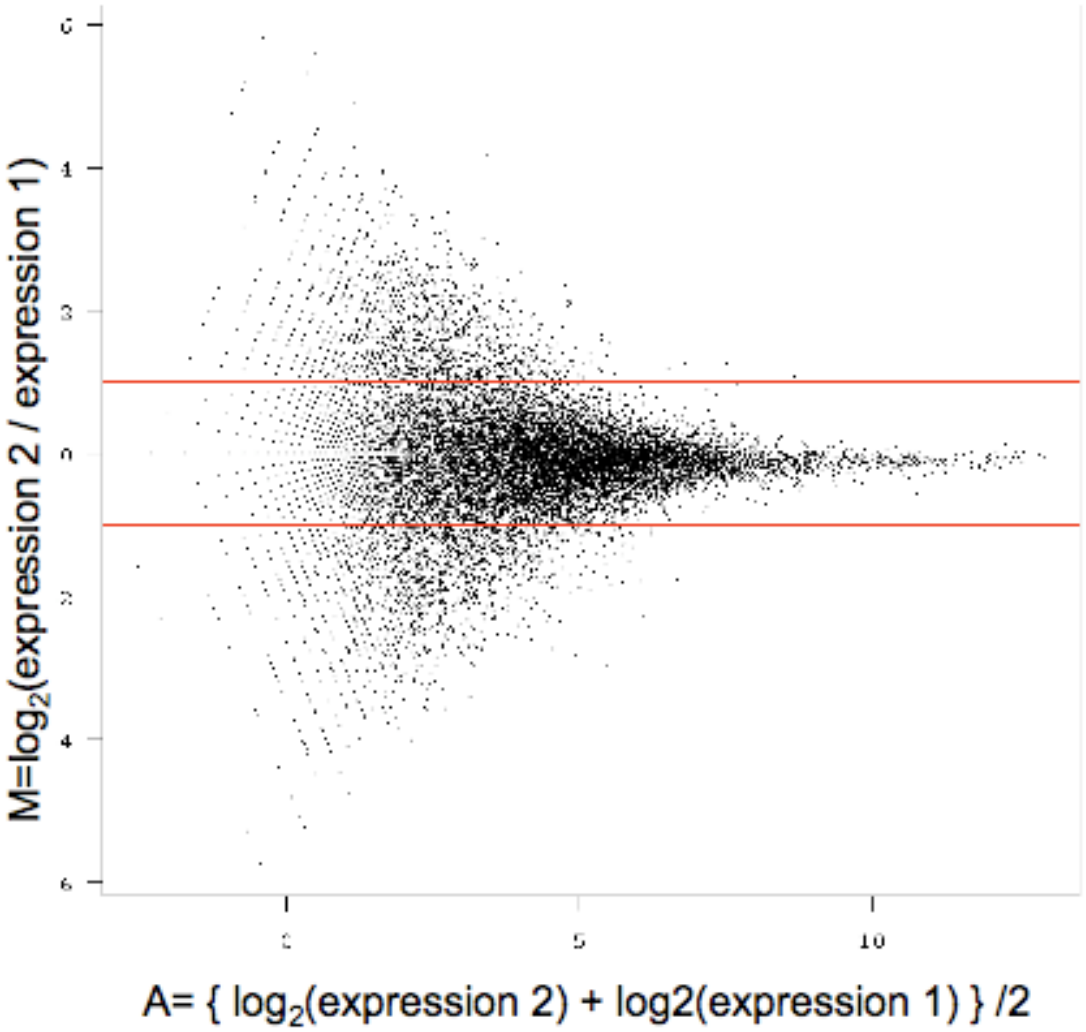
From Talk by Irizzary

We have to deal with **variations**!



From Talk by Irizzary

MvA Plots



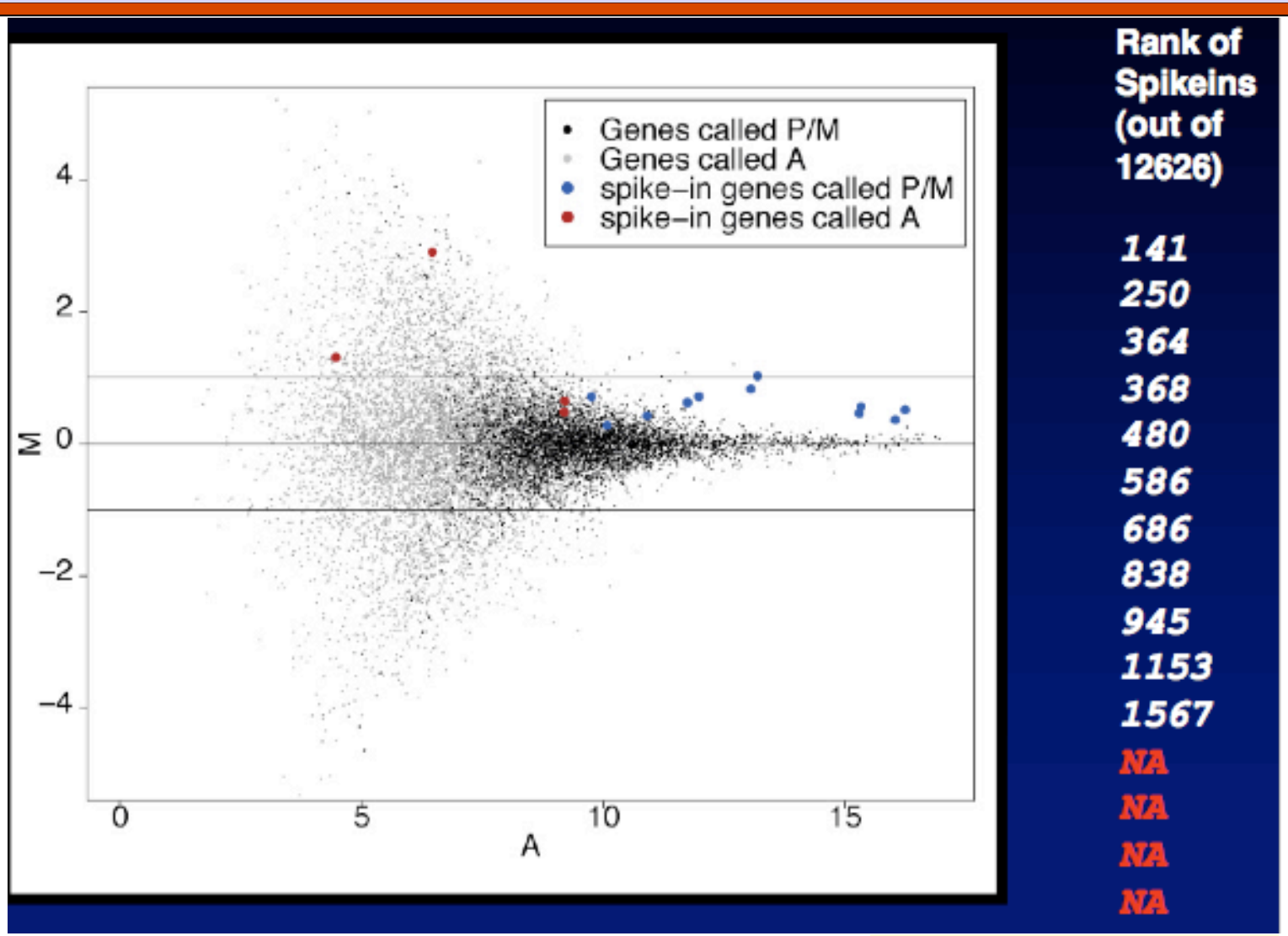
From Talk by Irizzary

Spike-in Experiment

- Replicate RNA samples were hybridized to various arrays
- Some probe sets were spiked in at different concentrations across the different arrays
- Goal was to see if these spiked probe sets "stood out" as differentially expressed

From Talk by Irizzary

Analyzing Spike-in data with MAS 5.0



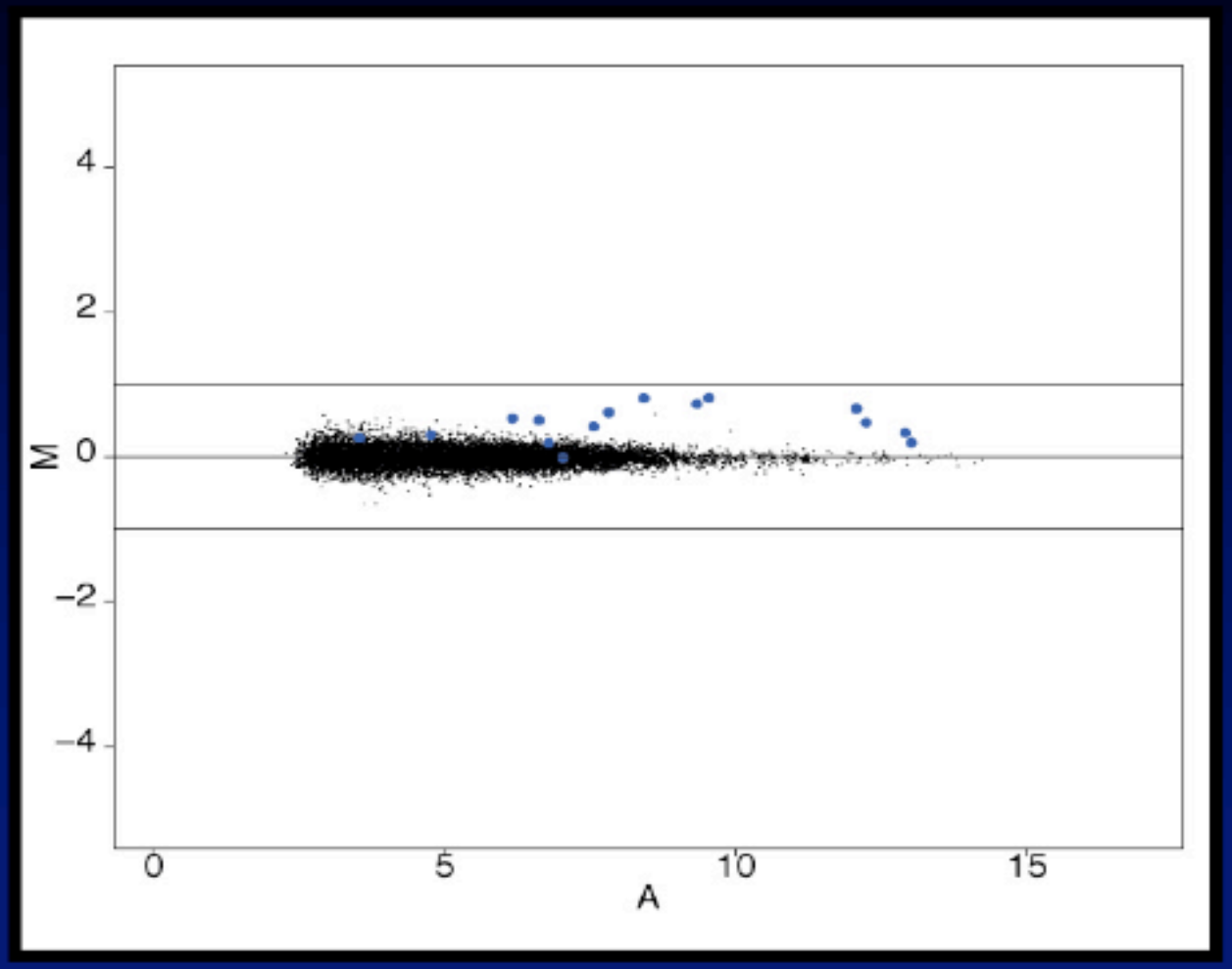
From Talk by Irizzary

Robust Multiarray normalization (RMA)

- Background correction separately for each array
  - Find E{Sig | Sig+Bgd = PM}
  - Bgd is normal and Sig is exponential
- Uses quantile normalization to achieve "identical empirical distributions of intensities" on all arrays
- Summarization: Performed separately for each probe set by fitting probe level additive model
- Uses median polish algorithm to robustly estimate expression on a specific chip
- Also see GCRMA [Wu, Irizzary et al., 2004]

From Talk by Irizzary & PhD thesis by Astrand

Analyzing Spike-in data with RMA



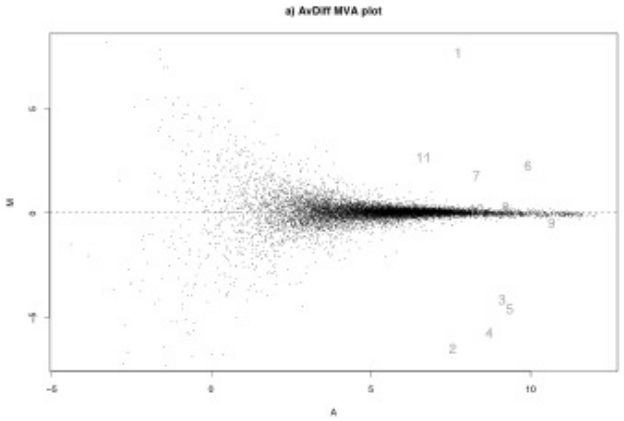
Rank of Spikeins (out of 12626)

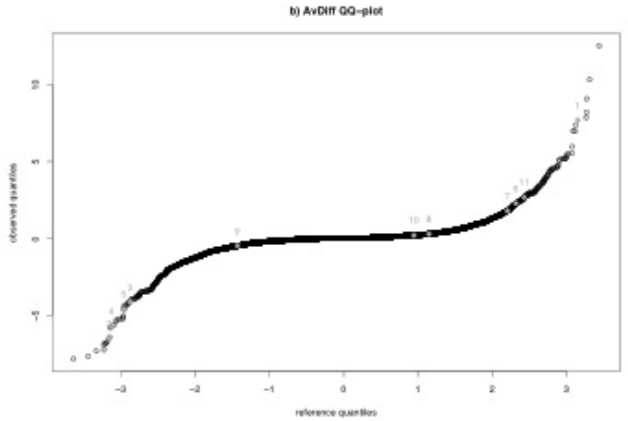
1
2
3
4
7
11
15
21
35
122
1182
230
450
1380
11700

Irizarry et al. (2003) *NAR* 31:e15

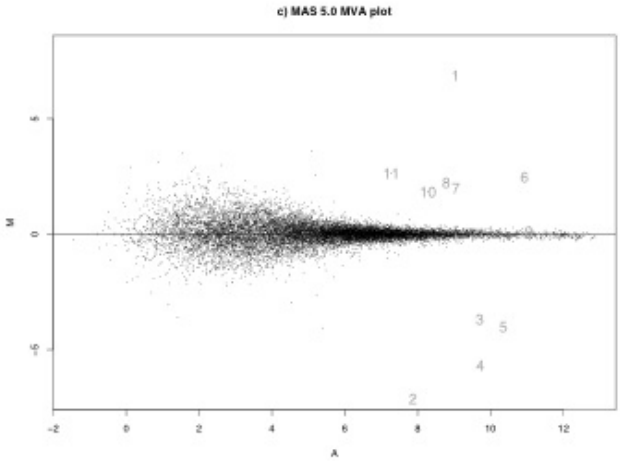
From Talk by Irizzary

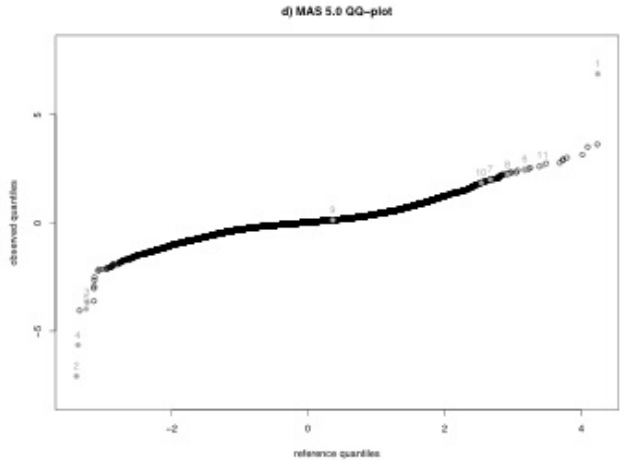
MvA and q-q plots





MAS 4.0

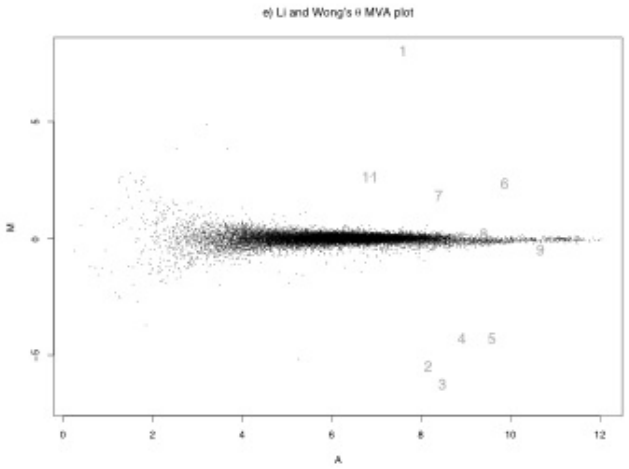


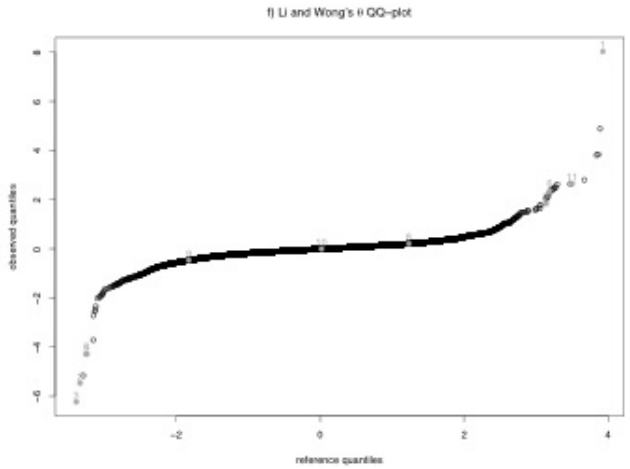


MAS 5.0

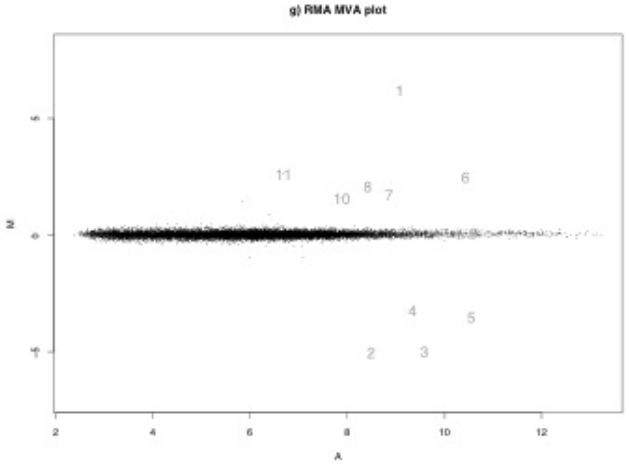
From Talk by Irizzary

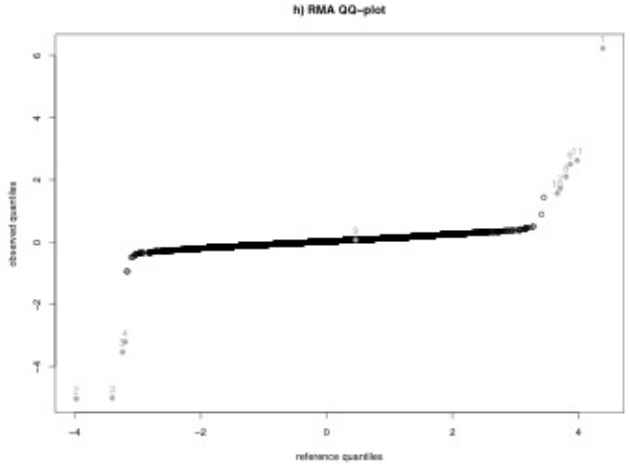
MvA and q-q Plots





MBEI





RMA

From Talk by Irizzary

Before and after quantile normalization

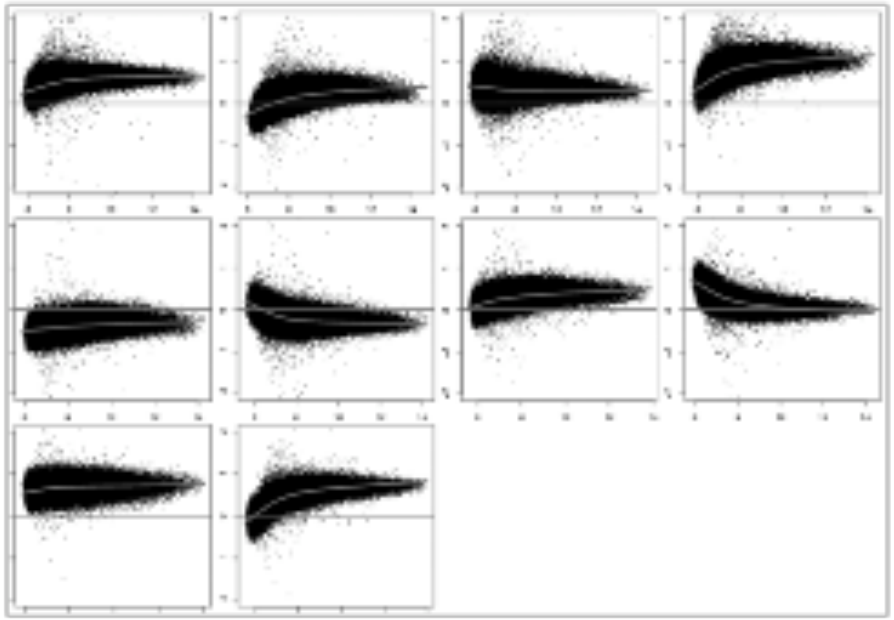


Fig. 2. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data for unadjusted data.

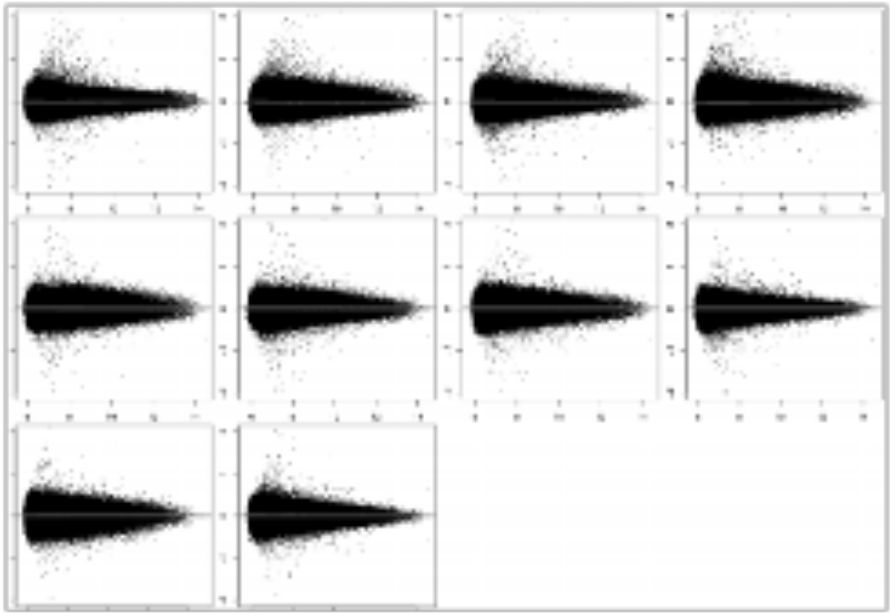


Fig. 3. 10 pairwise M versus A plots using liver (at concentration 10) dilution series data after quantile normalization.

From Talk by Irizzary

Bioconductor

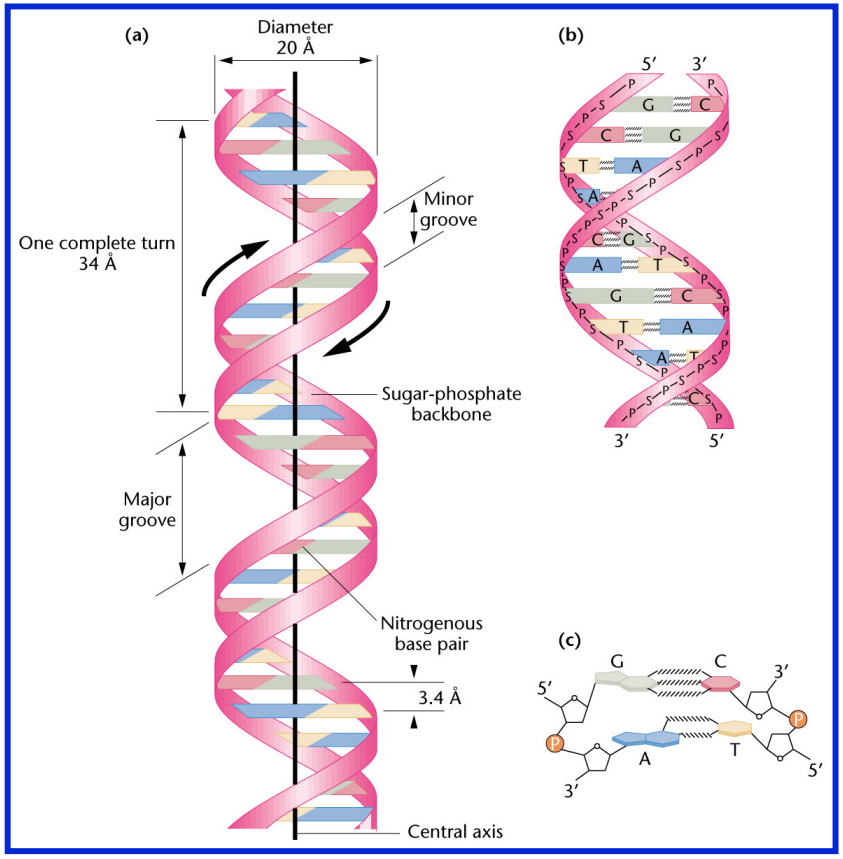
- **Bioconductor** is an **open source** and open development software project for the analysis of biomedical and genomic data.
- World-wide project started in 2001
- **R** and the **R package system** are used to design and distribute software
- Commercial version of Bioconductor software called **ArrayAnalyzer**

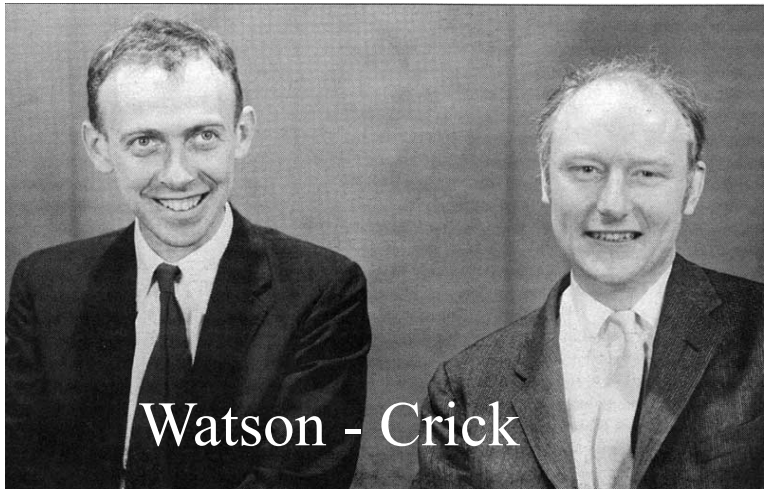
From Talk by Irizzary

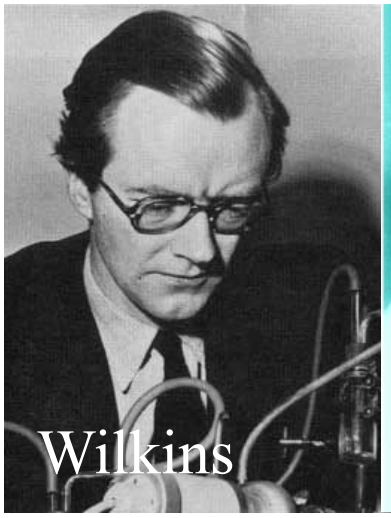
R: A Statistical Programming Language

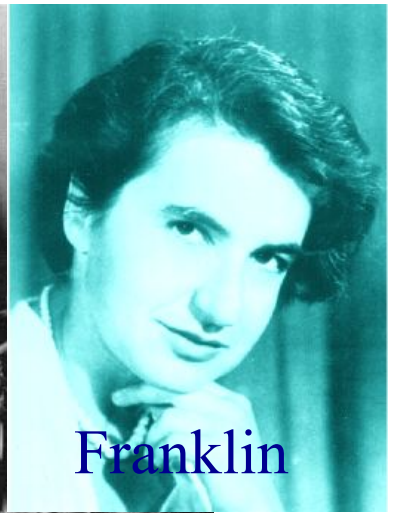
- Try the tutorial at: [http://www.cyclismo.org/tutorial/R/]
- Also at: [http://www.math.ilstu.edu/dhkim/Rstuff/Rtutor.html]

DNA Structure - 1953









DNA Controversy

1. Double Helix by Jim Watson - Personal Account (1968)
2. Rosalind Franklin by Ann Sayre (1975)
3. The Path to the Double Helix by Robert Olby (1974)
4. Rerelease of Double Helix by Jim Watson with Franklin's paper
5. Rosalind Franklin: The Dark Lady of DNA by Brenda Maddox (2003)
6. Secret of Photo 51 - 2003 NOVA Series

What are the next big Qs?

1. What is order of DNA sequence in a chromosome?
2. How does the DNA replicate?
3. How does the mRNA transcribe?
4. How is the protein gets translated?

Etc

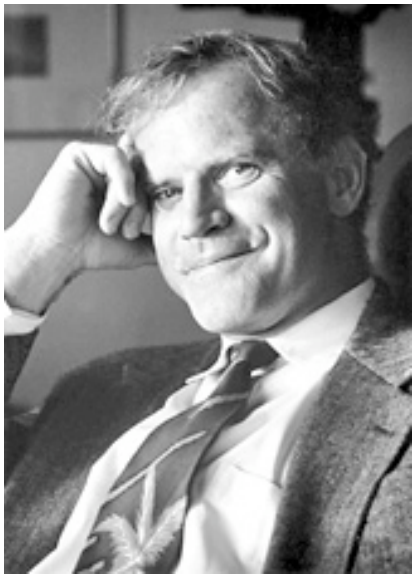
One of the tool that made a difference

Polymerase Chain Reaction

Polymerase Chain Reaction

1983 - technique was developed by Kary Mullis & others (1944-)

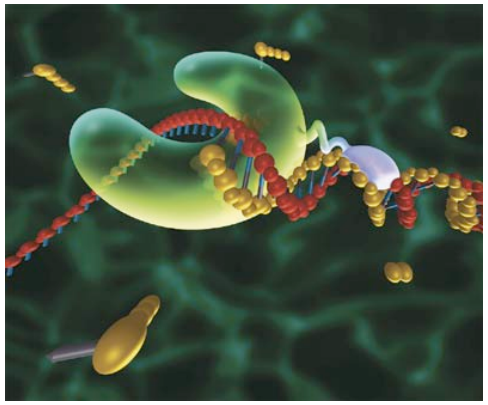
1993 Nobel prize for Chemistry



Controversy: Kjell Kleppe, a Norwegian scientist in 1971, published paper describing the principles of PCR

Stuart Linn, professor at University of California, Berkeley, used Kleppe's papers in his own classes, in which Kary Mullis was a student at the time

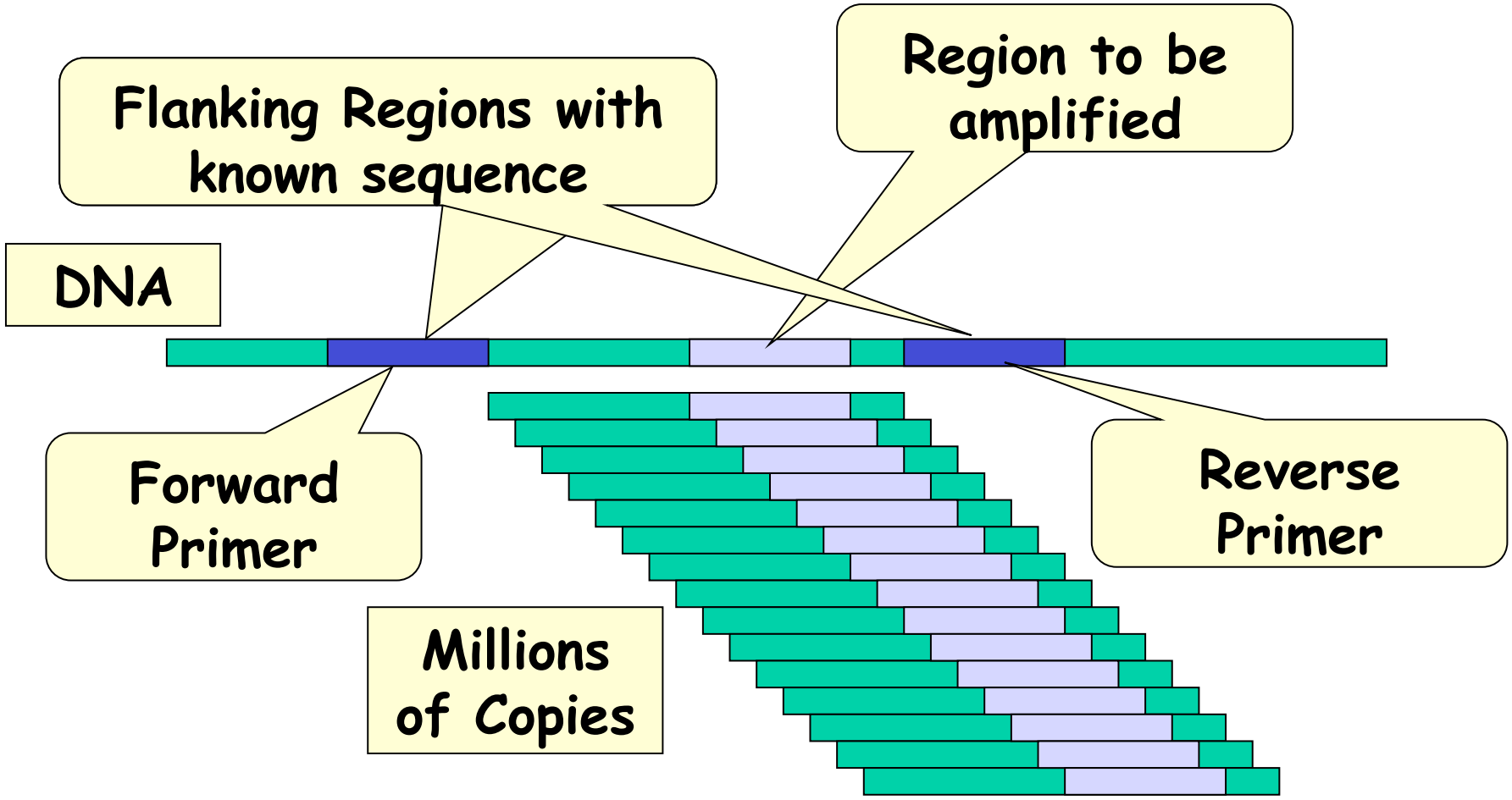
DNA Replication & Polymerase



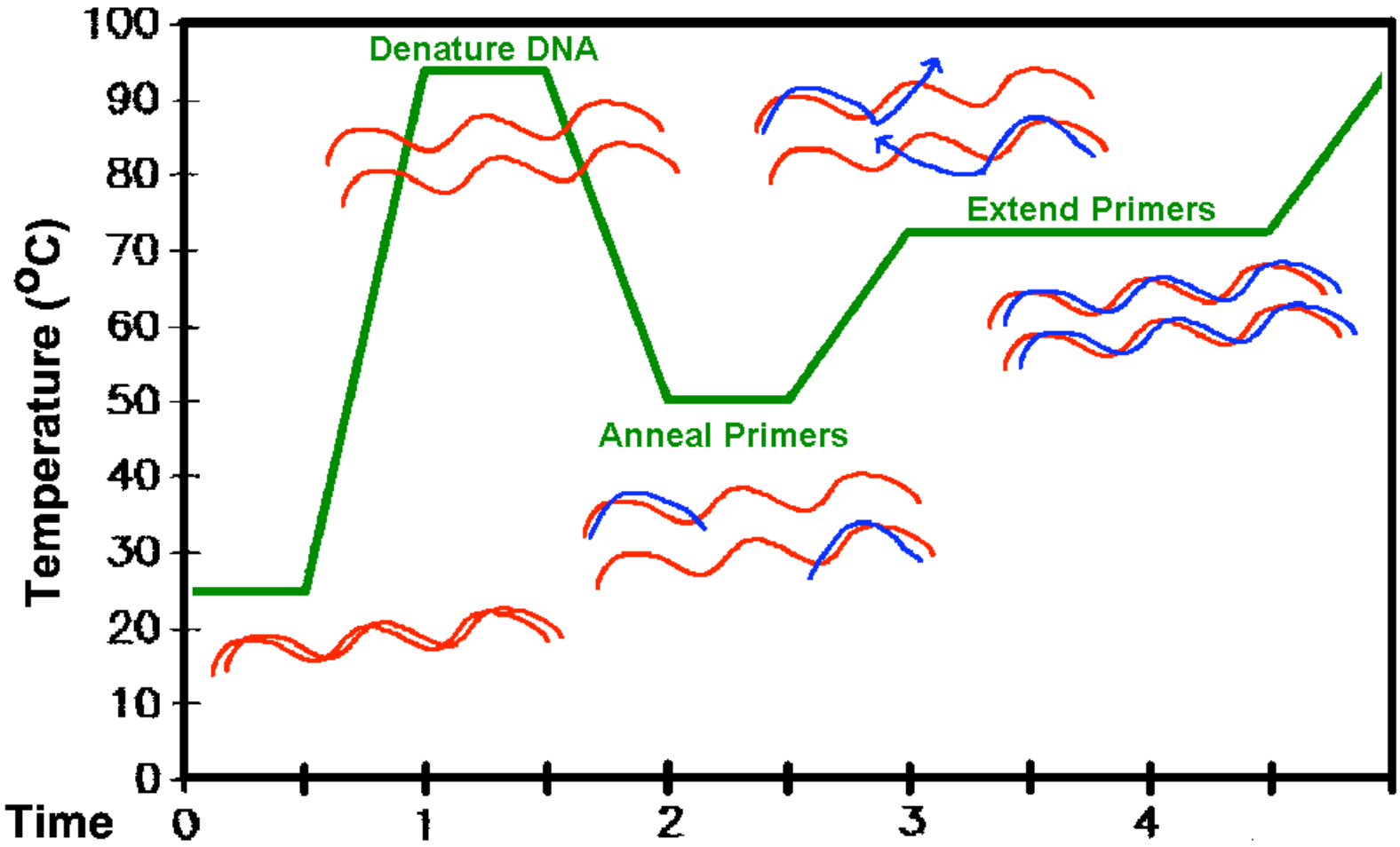
Polymerase Chain Reaction (PCR)

- PCR is a technique to amplify the number of copies of a specific region of DNA.
- Useful when exact DNA sequence is unknown
- Need to know "flanking" sequences
- Primers designed from "flanking" sequences
- If no info known, one can add adapters (short known sequence) then use a primer that recognizes the adaptor

PCR

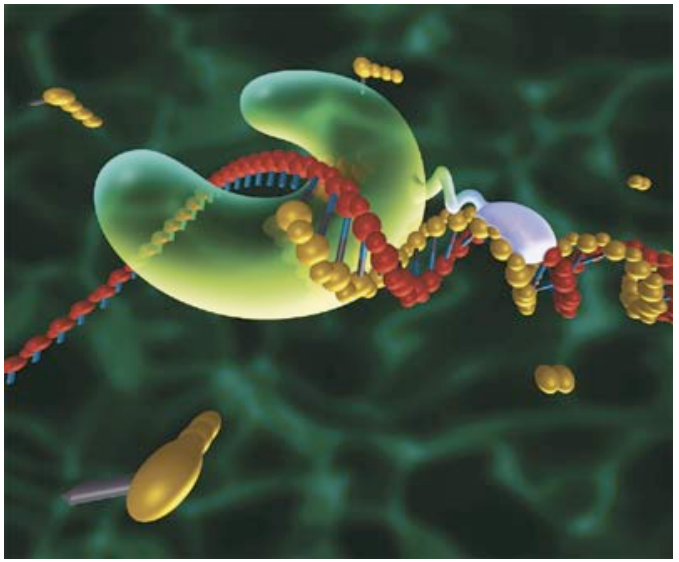

Flanking Regions with known sequence
Region to be amplified
DNA
Forward Primer
Reverse Primer
Millions of Copies

PCR

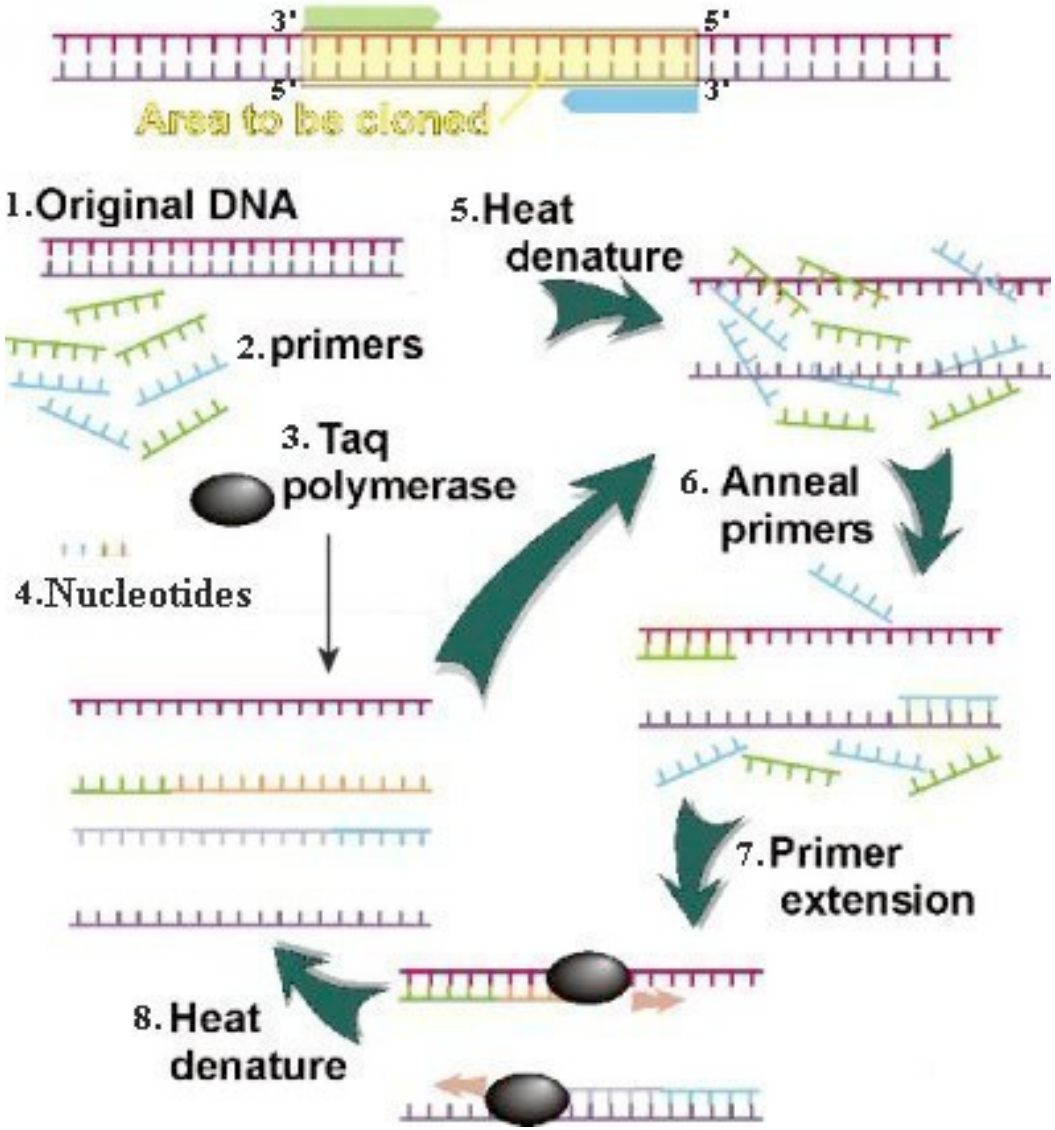

Denature DNA
Extend Primers
Anneal Primers
Temperature (°C)
100
90
80
70
60
50
40
30
20
10
0
Time
0
1
2
3
4

Taq polymerase

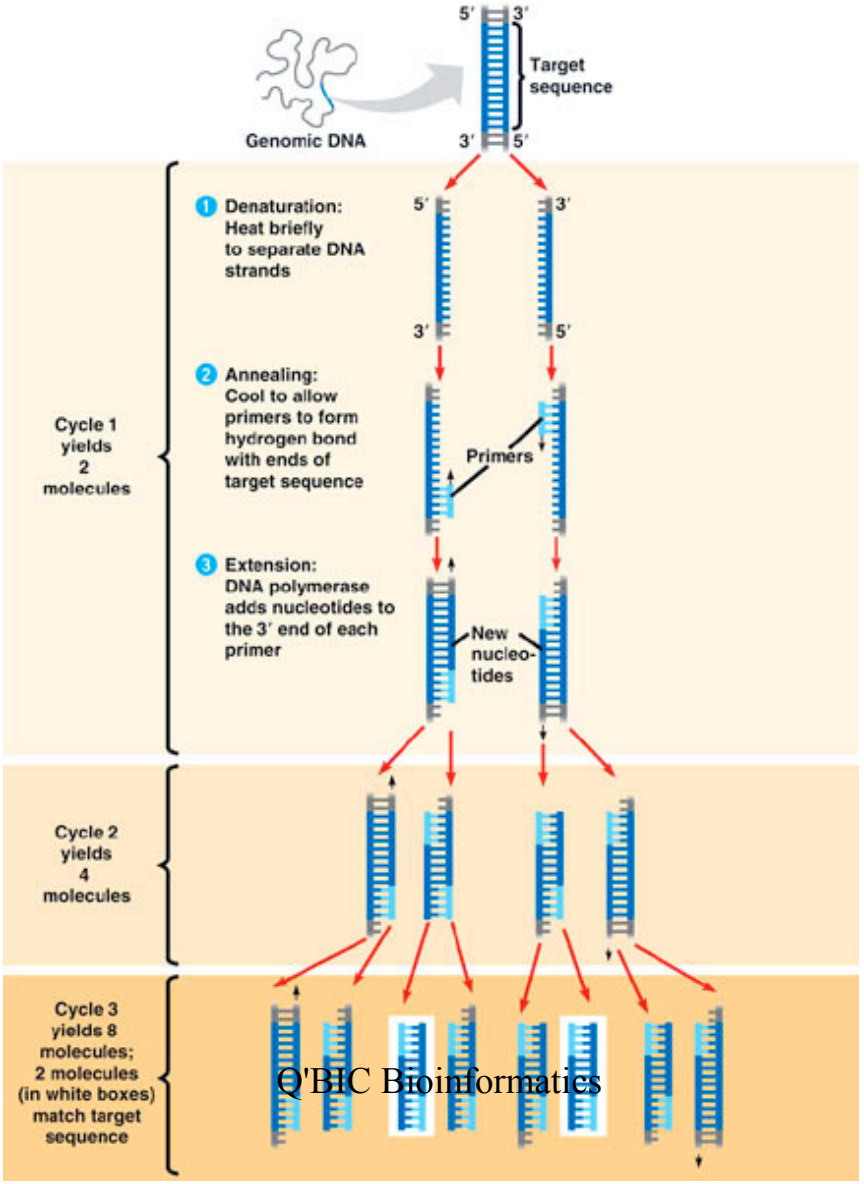
- Thermostable DNA polymerase named after the thermophilic bacterium *Thermus aquaticus*
- Originally isolated by Thomas D. Brock in 1965
- Molecule of the 80s
- Many versions of these polymerases are available
- Modified for increased fidelity



Schematic outline of a typical PCR cycle

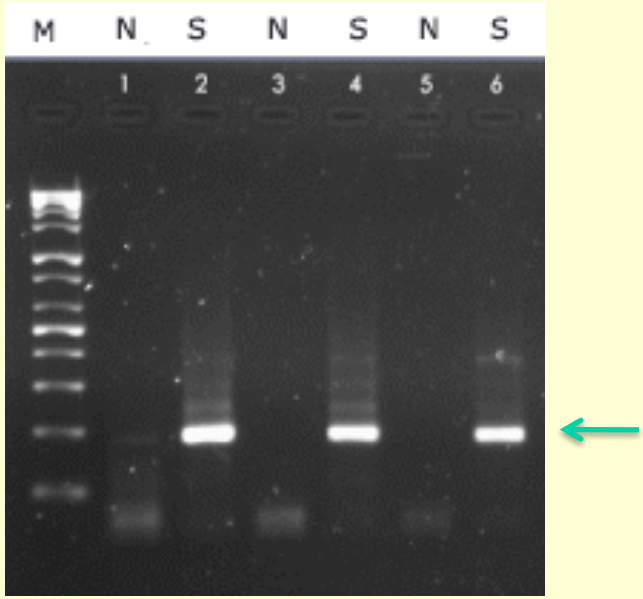


PCR


5′
3′
Target sequence
Genomic DNA
3′
5′
1 Denaturation: Heat briefly to separate DNA strands
5′
3′
3′
5′
2 Annealing: Cool to allow primers to form hydrogen bond with ends of target sequence
Primers
3 Extension: DNA polymerase adds nucleotides to the 3′ end of each primer
New nucleo-tides
Cycle 1 yields 2 molecules
Cycle 2 yields 4 molecules
Cycle 3 yields 8 molecules; 2 molecules (in white boxes) match target sequence

Gel Electrophoresis

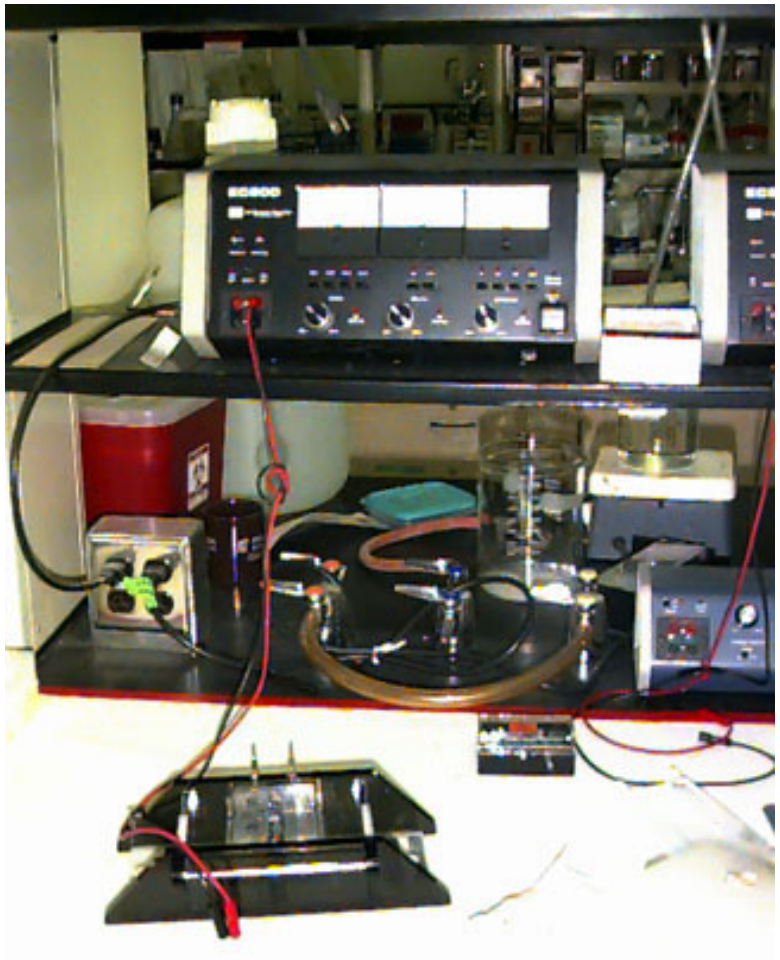
- Used to measure the size of DNA fragments.
- When voltage is applied to DNA, different size fragments migrate to different distances (smaller ones travel farther).

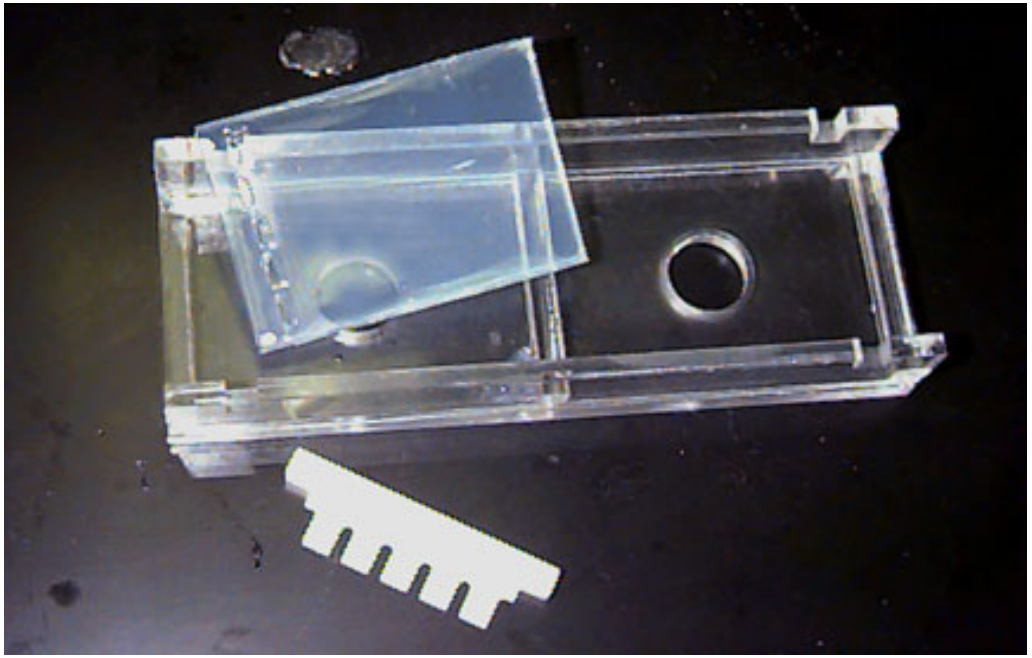


Gel Electrophoresis for DNA

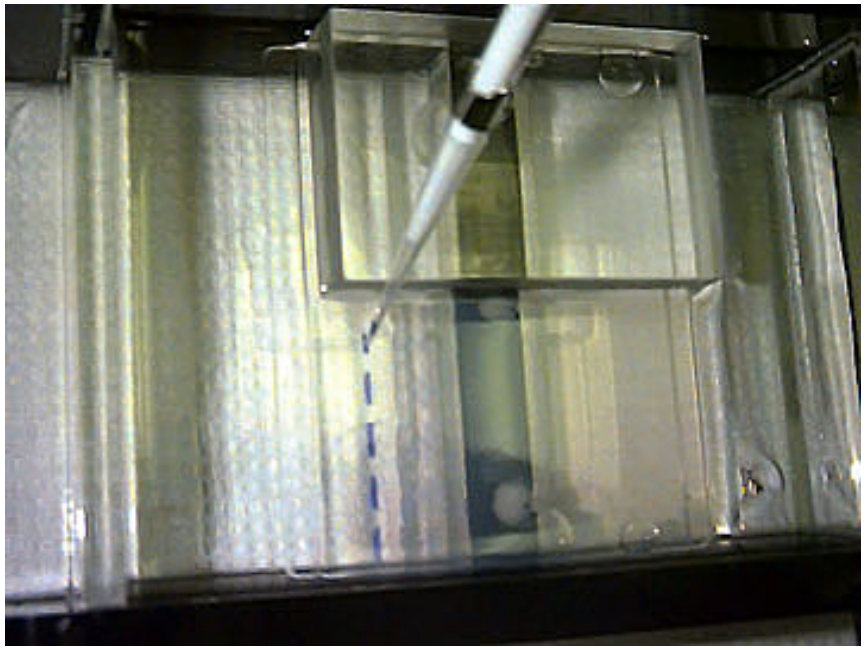
- DNA is negatively charged – WHY?
- DNA can be separated according to its size
- Use a molecular sieve – Gel
- Varying concentration of agarose makes different pore sizes & results
- Boil agarose to cool and solidify/polymerize
- Add DNA sample to wells at the top of a gel
- Add DNA loading dye (color to assess the speed and make it denser than running buffer)
- Apply voltage
- Larger fragments migrate through the pores slower
- Stain the DNA – EtBr, SyberSafe, etc

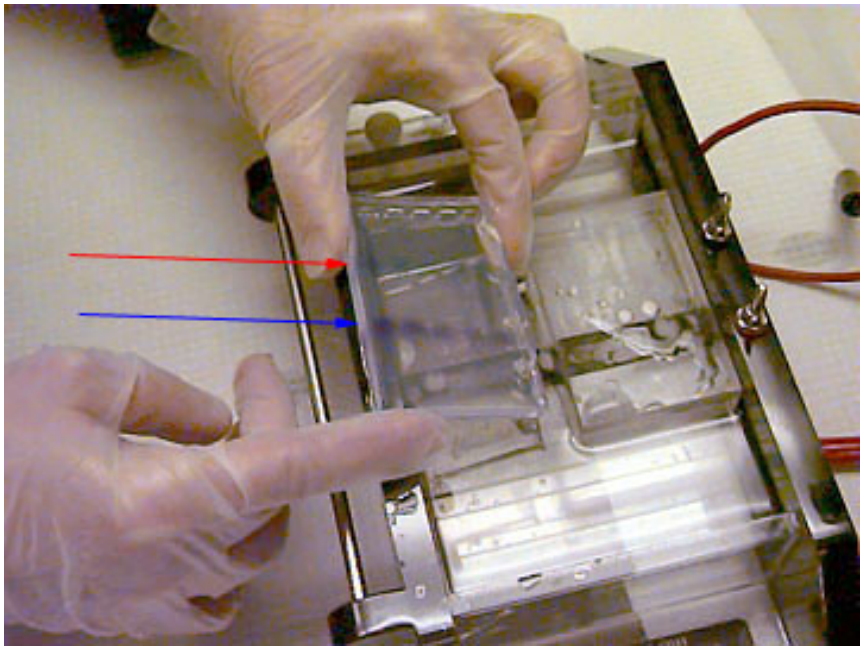
Gel Electrophoresis





Gel Electrophoresis





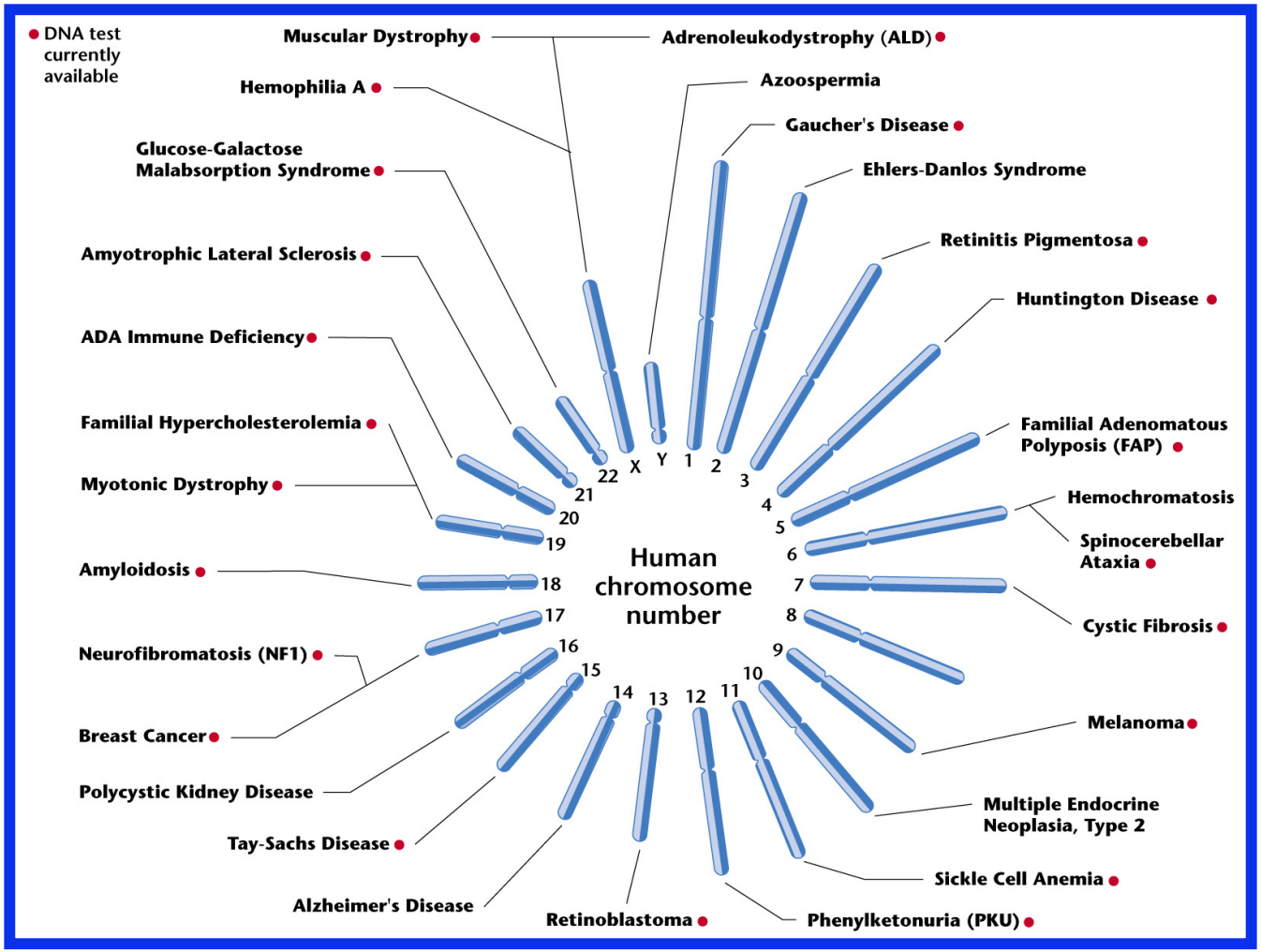
Sequencing

Why sequencing?

- Useful for further study:
  - Locate gene sequences, regulatory elements
  - Compare sequences to find similarities
  - Identify mutations – genetic disorders
  - Use it as a basis for further experiments
  - Better understand the organism
  - Forensics

Next 4 slides contains material prepared by Dr. Stan Metzenberg. Also see:
http://stat-www.berkeley.edu/users/terry/Classes/s260.1998/Week8b/week8b/node9.html

Human Hereditary Diseases

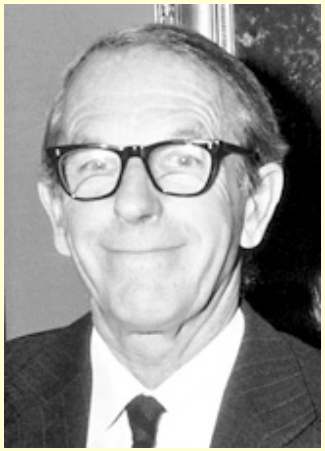


Those inherited conditions that can be diagnosed using DNA analysis are indicated by a (•)

History

- Two methods independently developed in 1974
  - Maxam & Gilbert method
  - Sanger method: became the standard
- Nobel Prize in 1980

Insulin; Sanger, 1958



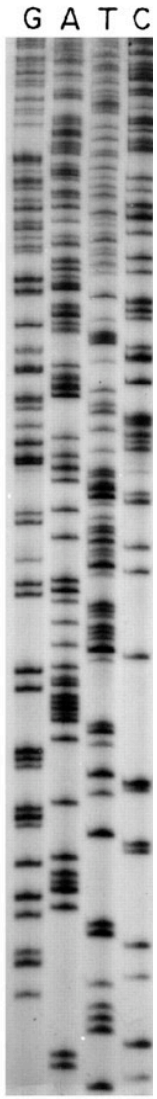
Sanger

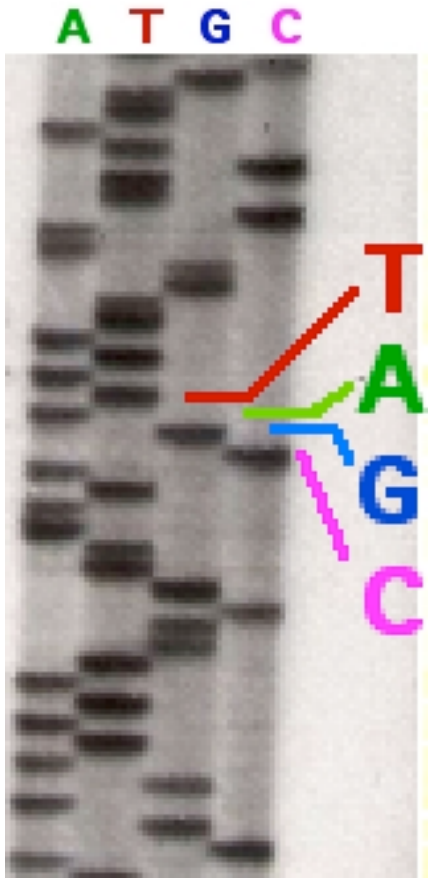
Gilbert

Original Sanger Method

- (Labeled) Primer is annealed to template strand of denatured DNA. This primer is specifically constructed so that its 3' end is located next to the DNA sequence of interest. Once the primer is attached to the DNA, the solution is divided into four tubes labeled "G", "A", "T" and "C". Then reagents are added to these samples as follows:
  - "G" tube: ddGTP, DNA polymerase, and all 4 dNTPs
  - "A" tube: ddATP, DNA polymerase, and all 4 dNTPs
  - "T" tube: ddTTP, DNA polymerase, and all 4 dNTPs
  - "C" tube: ddCTP, DNA polymerase, and all 4 dNTPs
- DNA is synthesized, & nucleotides are added to growing chain by the DNA polymerase. Occasionally, a ddNTP is incorporated in place of a dNTP, and the chain is terminated. Then run a gel.
- All sequences in a tube have same prefix and same last nucleotide.

Sequencing Gel



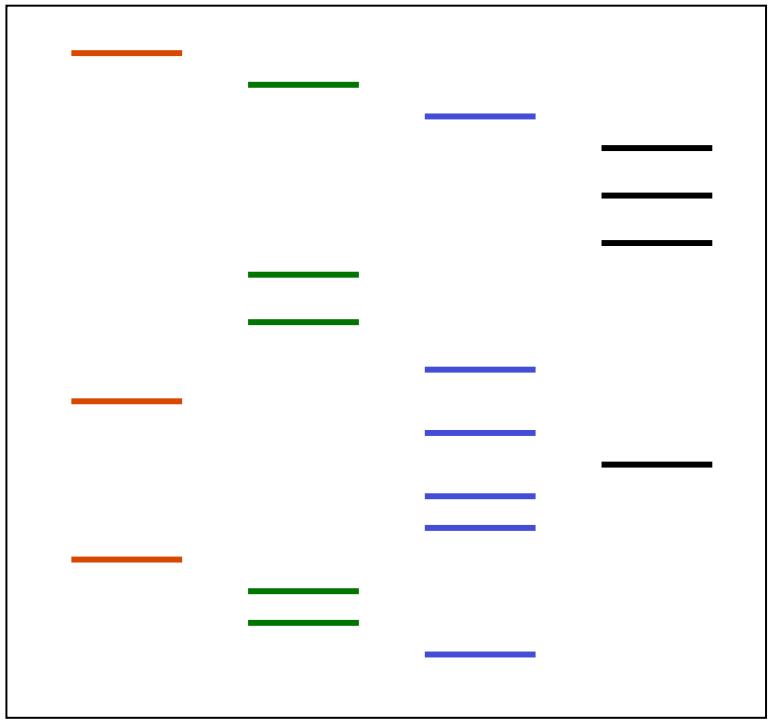


Modified Sanger

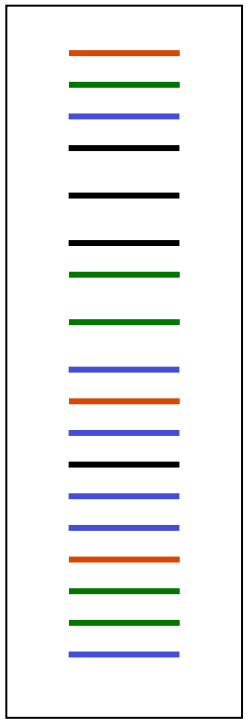
- Reactions performed in a single tube containing all four ddNTP's, each labeled with a different color fluorescent dye

Sequencing Gels: Separate vs Single Lanes

GCCAGGTGAGCCTTTGCA



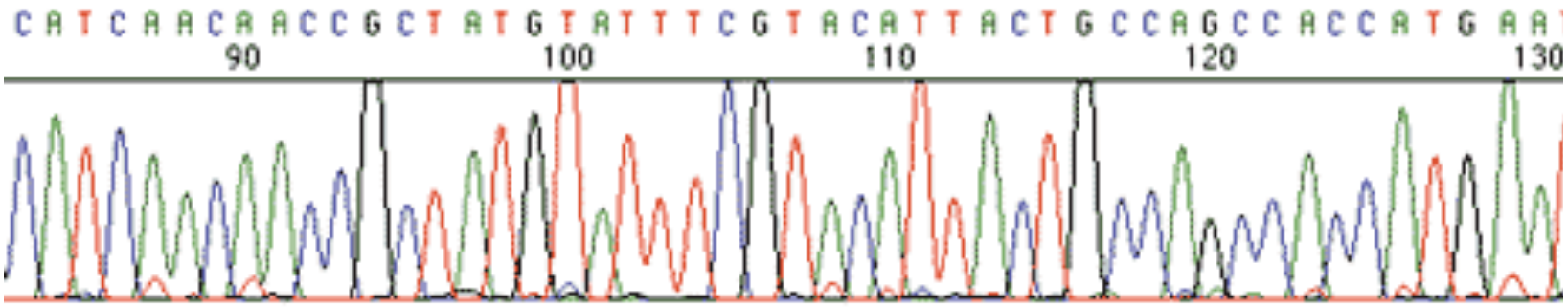
A C G T



Automated Sequencing Instruments

Sequencing

- Flourescence sequencer
- Computer detects specific dye
- Peak is formed
- Base is detected
- Computerized



Maxam-Gilbert Sequencing

- Not popular
- Involves putting copies of the nucleic acid into separate test tubes
- Each of which contains a chemical that will cleave the molecule at a different base (either adenine, guanine, cytosine, or thymine)
- Each of the test tubes contains fragments of the nucleic acid that all end at the same base, but at different points on the molecule where the base occurs.
- The contents of the test tubes are then separated by size with gel electrophoresis (one gel well per test tube, four total wells), the smallest fragments will travel the farthest and the largest will travel the least far from the well.
- The sequence can then be determined from the picture of the finished gel by noting the sequence of the marks on the gel and from which well they came from.

Human Genome Project

Play the Sequencing Video:

- Download Windows file from http://www.cs.fiu.edu/~giri/teach/6936/Papers/Sequence.exe
- Then run it on your PC.

Human Genome Project

1980 The sequencing methods were sufficiently developed

International collaboration was formed: International Human Genome Consortium of 20 groups - a Public Effort (James Watson as the chair!)

Estimated expense: $3 billion dollars and 15 years

Part of this project is to sequence: *E. coli, Sacchromyces cerevisiae, Drosophila melanogaster, Arabidopsis thaliana, Caenorhabdidtis elegans*

- Allow development of the sequencing methods

Got underway in October 1990

Automated sequencing and computerized analysis

Public effort: 150,000 bp fragments into artificial chromosomes (unstable - but progressed)

In three years large scale physical maps were available

Venter vs Collins

National Human Genome Research Institute

Venter's lab in NIH (joined NIH in 1984) is the first test site for ABI automated sequences; he developed strategies (Expressed Sequence Tags - ESTs)

1992 - decided to patent the genes expressed in brain - "Outcry"

Resistance to his idea

Watson publicly made the comment that Venter's technique during senate hearing - "wasn't science - it could be run by monkeys"

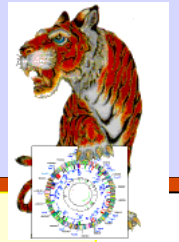
In April 1992 Watson resigned from the HGP

Craig Venter and his wife Claire Fraser left the NIH to set up two companies

- the not-for-profit TIGR The Institute for Genomic Research, Rockville, Md
- A sister company FOR-profit with William Hazeltine - HGSI - Human Genome Sciences Inc., which would commercialize the work of TIGR
- Financed by Smith-Kline Beecham ($125 million) and venture capitalist Wallace Steinberg.

Francis Collins of the University of Michigan replaced Watson as head of NHGRI.

Venter vs Collins



HGSI promised to fund TIGR with $70 million over ten years in exchange for marketing rights TIGR's discoveries

PE developed the automated sequencer & Venter - Whole-genome short-gun approach

"While the NIH is not very good at funding new ideas, once an idea is established they are extremely good," Venter

In May 1998, Venter, in collaboration with Michael Hunkapiller at PE Biosystems (aka Perkin Elmer / Applied Biosystems / Applera), formed Celera Genomics

Goal: sequence the entire human genome by December 31, 2001 - 2 years before the completion by the HGP, and for a mere $300 million

April 6, 2000 - Celera announces the completion "Cracks the human code"

Agrees to wait for HGP

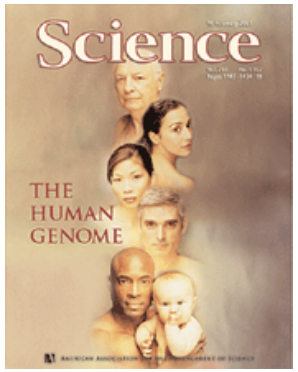
Summer 2000 - both groups announced the rough draft is ready

Human Genome Sequence

6 months later it was published - 5 years ahead of schedule with $ 3 billion dollars

50 years after the discovery of DNA structure

Human Genome Project was completed - 3.1 billion basepairs



Pros: No guessing of where the genes are

Study individual genes and their contribution

Understand molecular evolution

Risk prediction and diagnosis

Con: Future Health Diary --> physical and mental

Who should be entrusted? Future Partners, Agencies, Government

Right to "Genetic Privacy"