

BINF 732 Genomics

Dr. Saleet Jafri

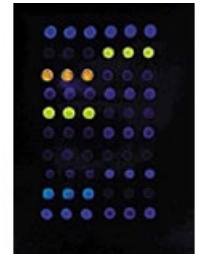
Genome Expression and Microarrays

Sept. 19, 2007

Kevin Thompson

Slides provided by Dr. Jennifer Weller

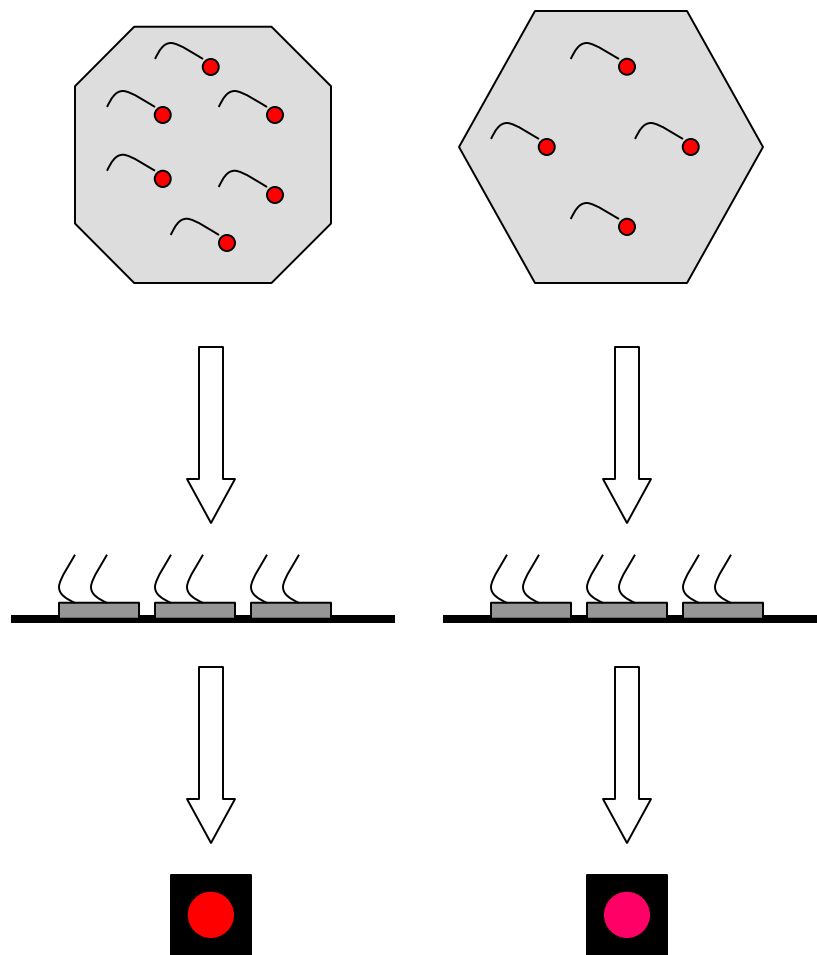
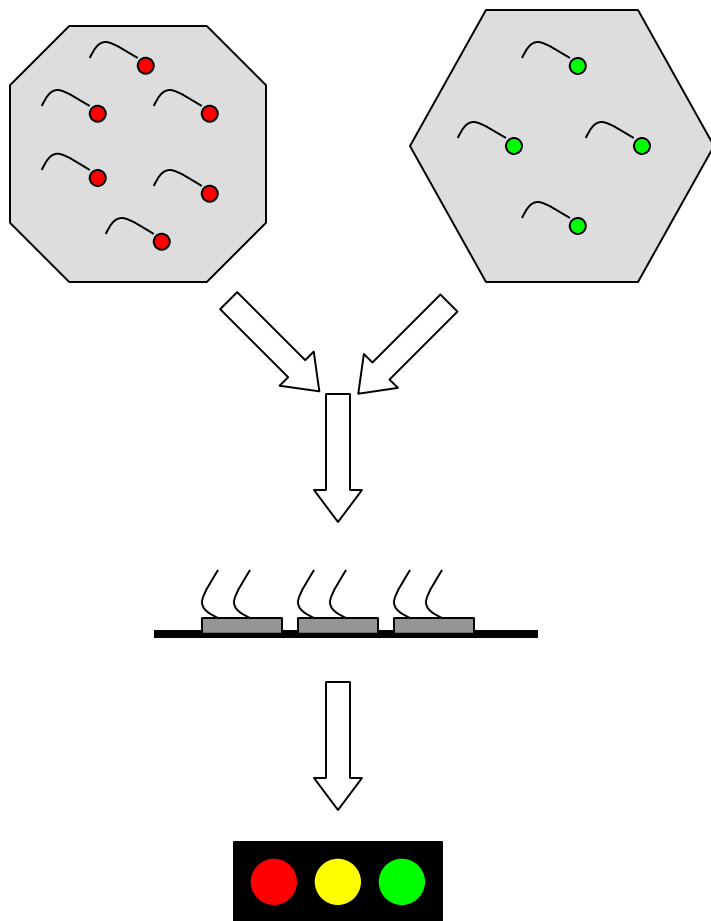
Microarray Experiments



- At their most general a microarray is a flat surface on which one molecule interacts with another, and the location and type of signal produced by the interaction are used to infer important characteristics about the interacting partners.
 - The *probe* is the characterized entity that monitors that behavior of the *target*.
 - Either the probe or target may be the entity fixed to the surface. In gene expression microarrays the probe is a (usually relatively short) sequence of DNA that is fixed to the surface.
 - Either the probe or the target may carry the signaling molecule. Usually the strength of the interaction determines whether the signal is stable enough to be detected.

Two Dye

Single Dye

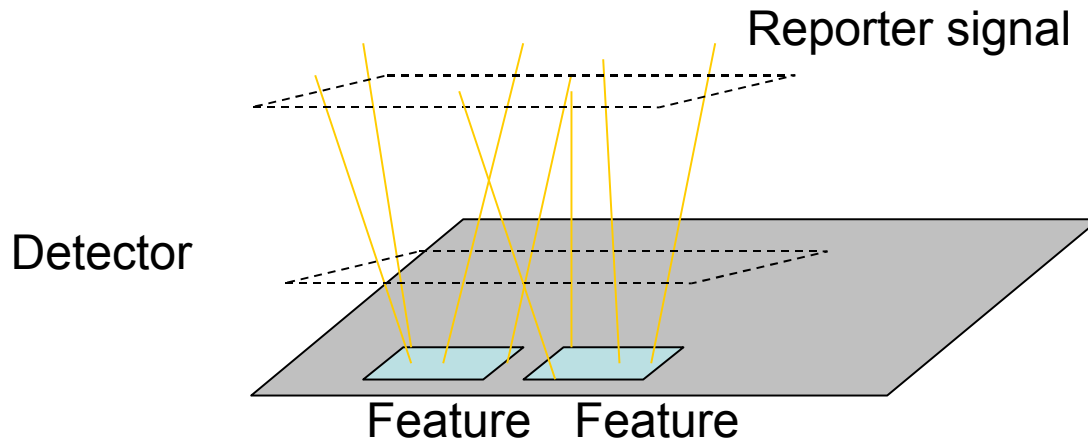


Common terminology

- From the Microarray Gene Expression Databases (MGED) group and MAGE model definitions (<http://www.mged.org/>) there is a specified vocabulary:
- Probe: the known entity attached to the array surface
- Target: the labeled unknown entities
- Feature: an (x,y) location
- Reporter: the probe sequence (may report on multiple targets)
- Composite Feature: a set of reporters (if for example more than one probe binds to a single target)
- *Note: Schena, for historical reasons (Northern blot nomenclature) reverses probe and target terms.

General Array attributes

- Physical dimensions of the array
- Number of probes per array
- Surface and attachment chemistries
- Probe density: how close together (on average) the identical sequences are.
 - Allow optimal target interaction
 - Fit the optical detection system -> to signal spread and boundary overlap



Biophysical processes underlying Nucleic acid- based Microarrays

- Nucleic acids have a particular structure and mode of interaction with one another
- On microarrays, the complementarity of base-pairing of single-stranded nucleic acids is the interaction that we intend to monitor.
 - The more stable the duplex the greater sensitivity; the duplex must withstand the wash conditions
 - the *specificity* of the assay depends on this interaction
 - The probe is always DNA, but the target may be RNA or DNA

Solution Diffusing-Fixed Partner Hybridization

- Microarrays fix one binding partner to the surface and wash the other across in solution.
 - The probe has
 - a locally high concentration
 - is not homogeneously distributed in solution where random processes govern the likelihood of probe-target interactions.
 - The target is
 - homogeneously distributed in solution
 - except where it has formed a duplex with probe – then it is also locally higher in concentration when it has dissociated from the probe.
 - The original model for this platform was the Southern blots
 - Genomic DNA is restricted using endonucleases
 - Separate fragments through gel electrophoresis
 - Fixed to nitrocellulose membranes
 - Radio-labeled DNA probes of known sequence are hybridized to the attached unknowns
 - Non-binding material is washed away
 - Duplex product is visualized using film (autoradiography).

Outcomes

- Properly processed data has the characteristic that the modified value of the signal intensity correlates in the same way for all individual measurements with the amount of mRNA in the original sample.
- The *result* is a set of observations of the concentration of all represented transcripts that were present *when* the sample was collected.
 - Observing how transcripts respond to stimuli can give an understanding of the transcriptional control and regulation in a cell under particular stresses

Why?

- F. Crick, Nature, 1970
 - Unidirectional transfer of sequence information
 - Withstands challenges by prions, mature mRNA, methylation, etc
- D. Thielfry, TiBS 23, Aug 1998

Central Dogma of Molecular Biology

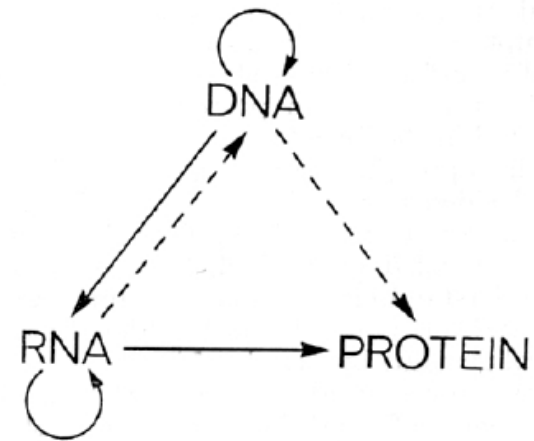


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

High Throughput Replacement of Traditional Analysis

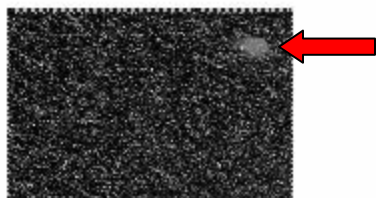
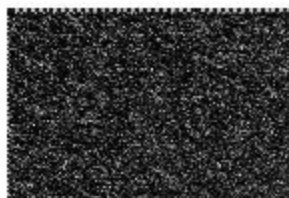
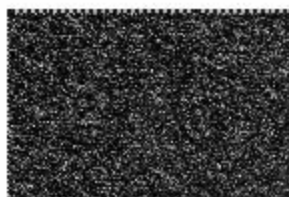
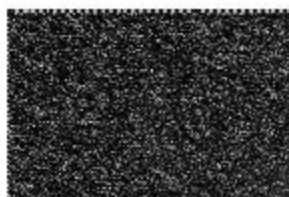
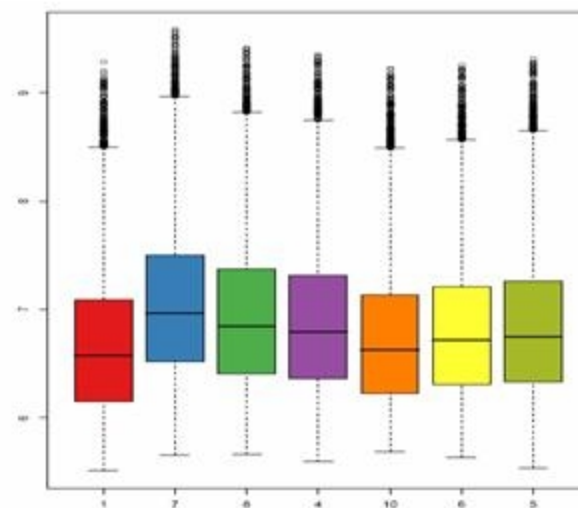
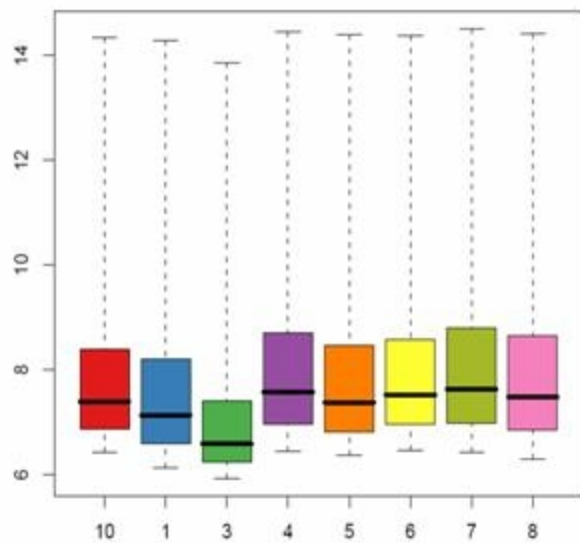
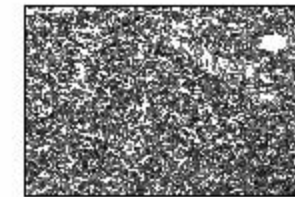
- Single parameter (protein) experiments
 - Study changes in protein kinetics, interactions, functionality (localization), concentrations
 - Quantitative measurements
 - Westerns, ELISA's, HPLC, etc
- Multivariate analysis of mRNA levels
 - ✓ Qualitative analysis
 - ✓ One step removed from the cellular protein activities

ELISA vs Microarray

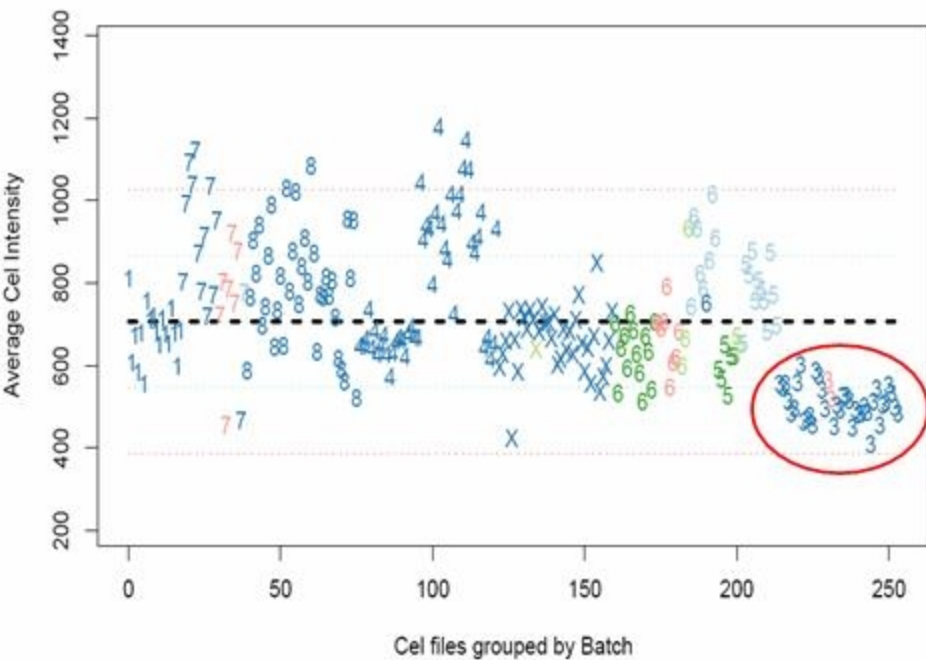
- Antibody Specificity
 - Increased through blocking, washings, and antibody scaffolding
- Sample Replication
- Background Controls
- Internal Standard
 - w/ corr coeff $>.98$ (low)
- $0.0 < \text{O.D.} < 2.0$
- Sequence Specificity?
 - Washings
- Sample Replication?
- Background?
 - Chip, block, spot
- No standard
 - Spiked-in experiments
- Scanner limitations

Other Microarray Issues

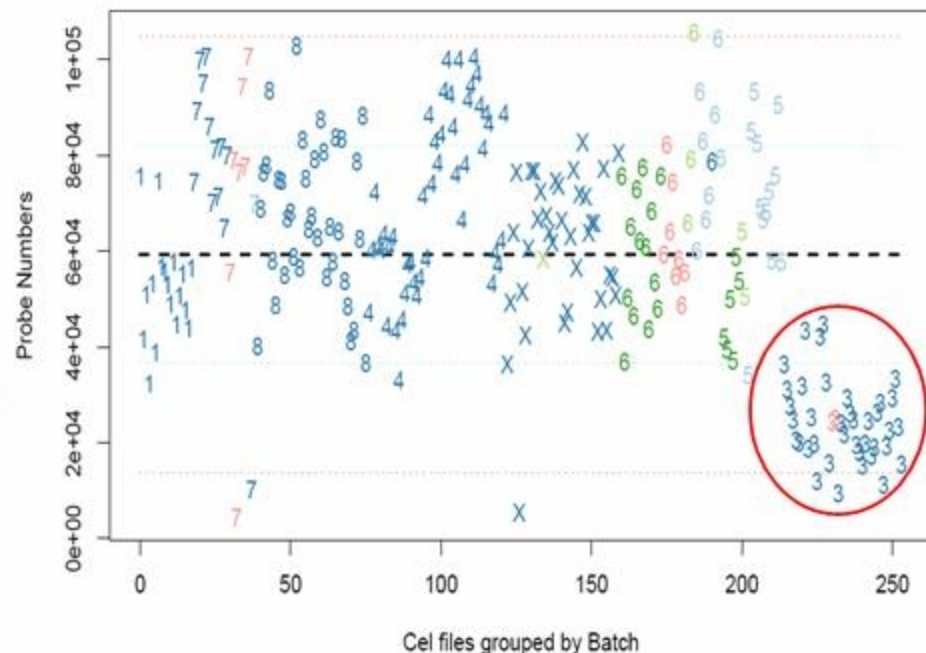
- Biological/Sample Variation
 - SNP's, x-hybridizations, variance
- Lab Variation
 - Bubbles, scrapings, degradation
- Dye incorporation
 - 3' bias, Cy3 better than Cy5
- Heterogeneity of samples
 - Mixture of cells
- Ploidy
- Splicing Events
- Annotations (gene models, exons, introns, GO, gene function)
- Statistically
 - $n \ll P$
 - Independence

Batch10**Batch1****Batch10****Batch1****Batch4****Batch5****Batch4****Batch5**

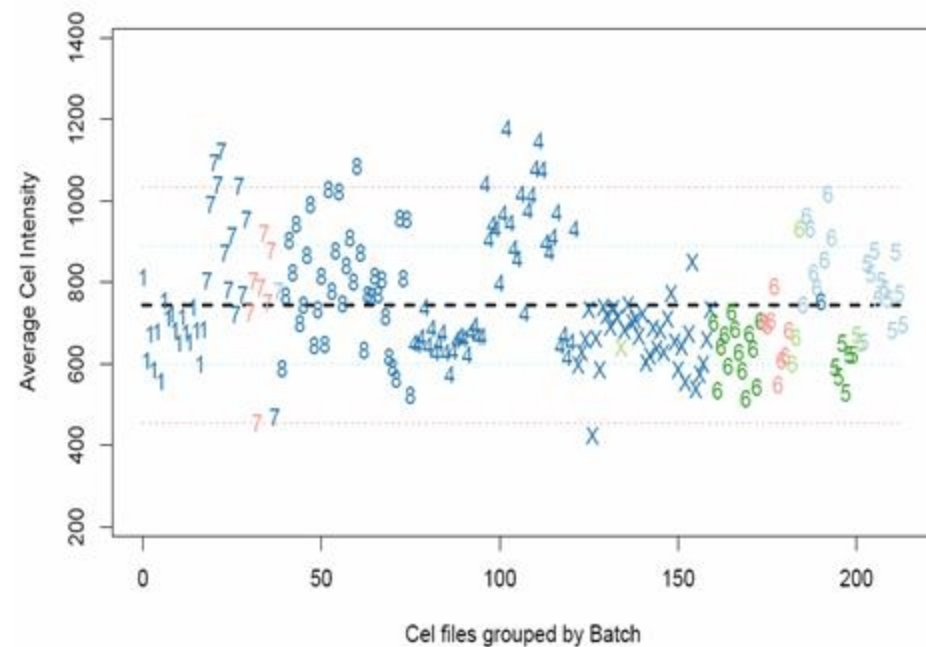
Sample Analysis of Cleansed Probes



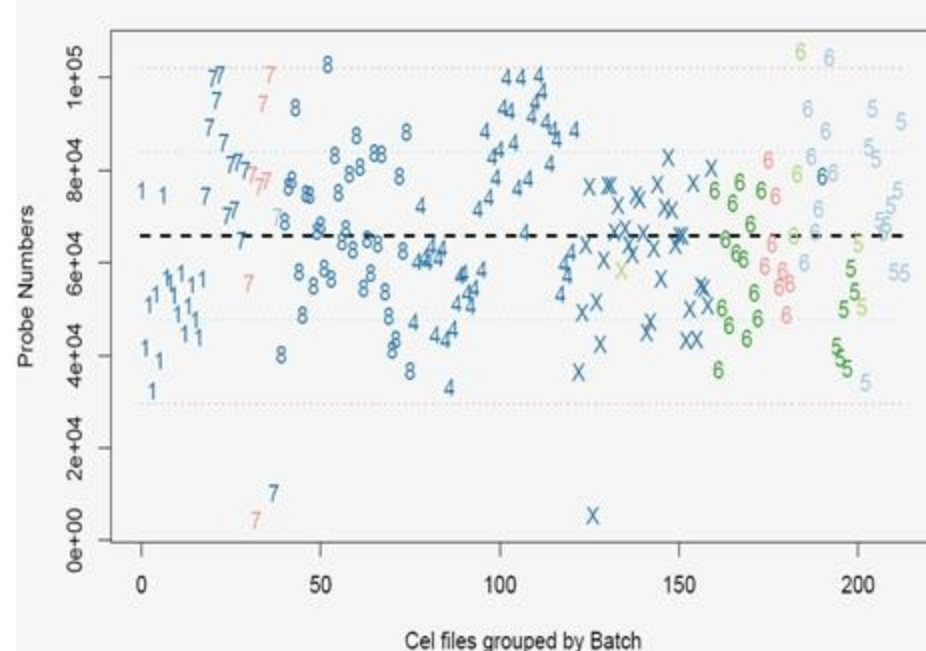
Sample Analysis of Cleansed Probes



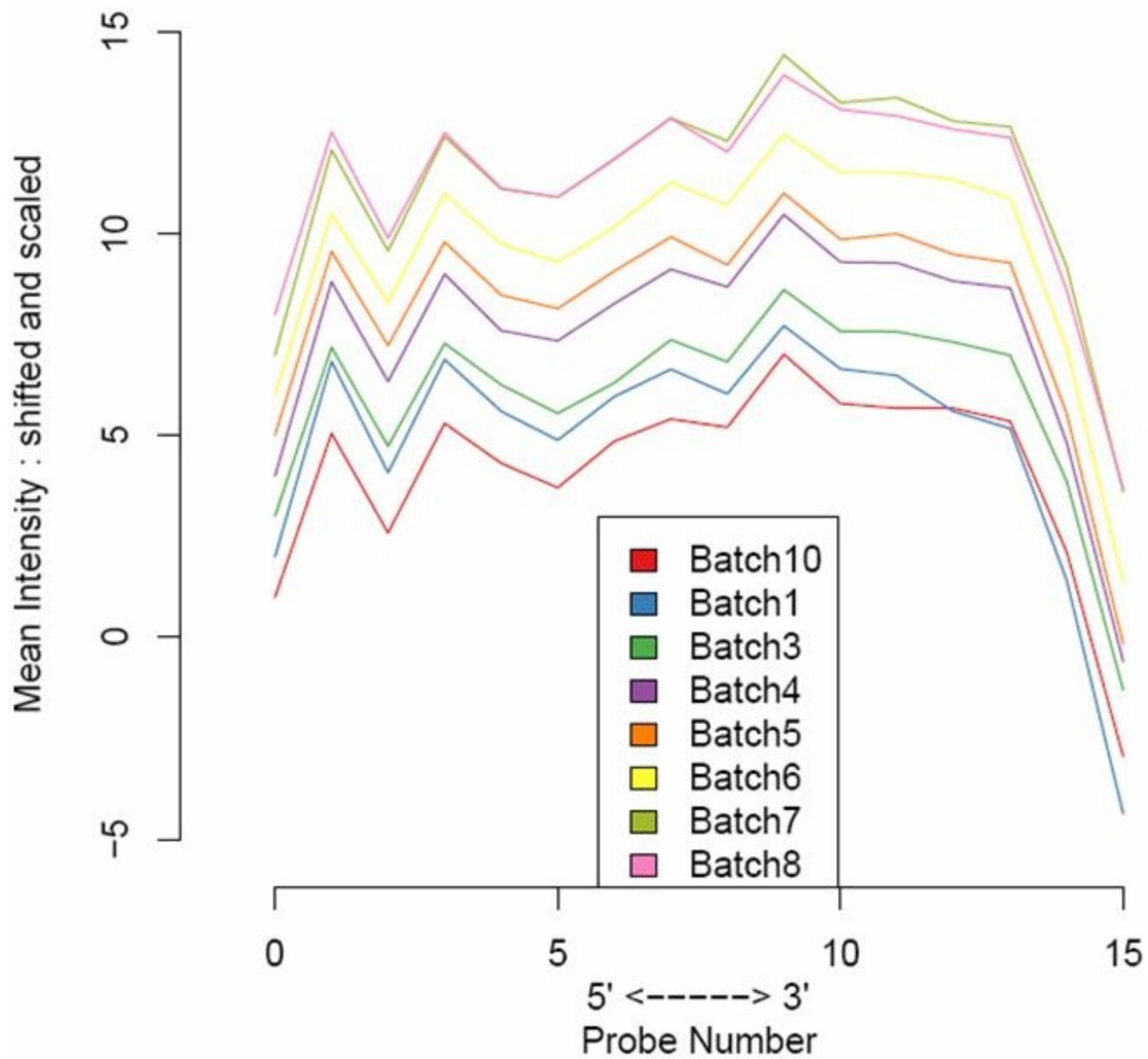
Cleansed Probes, Batch 3 Removed

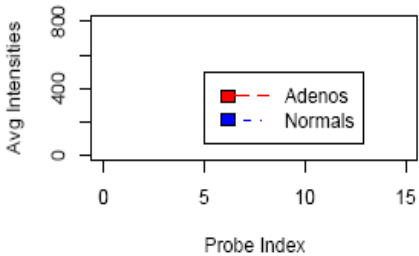
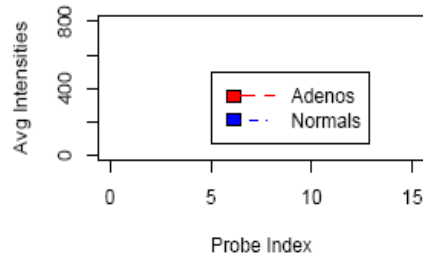
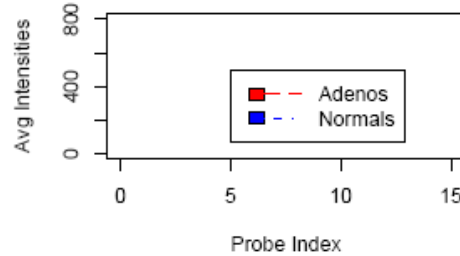
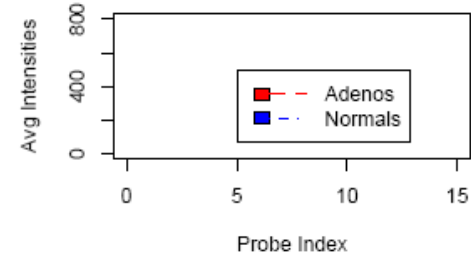
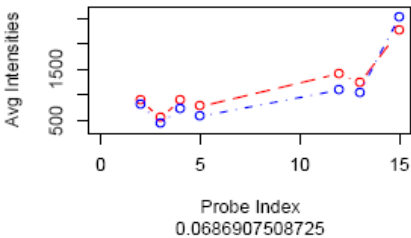
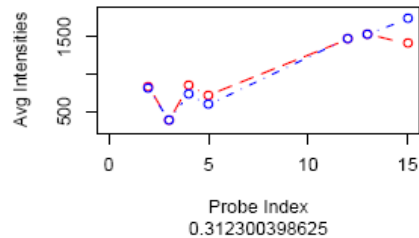
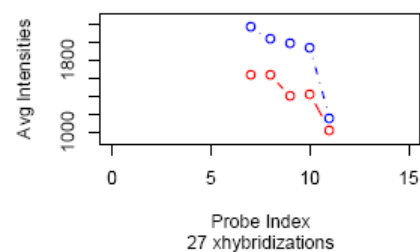
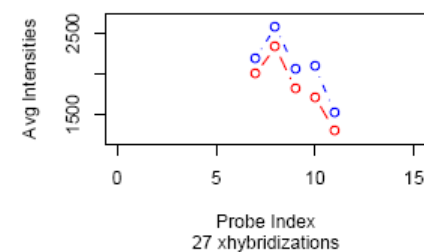
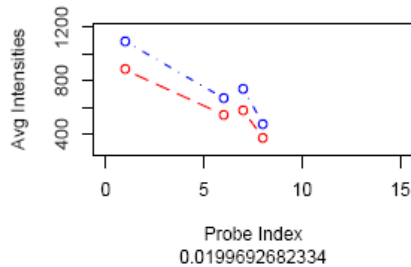
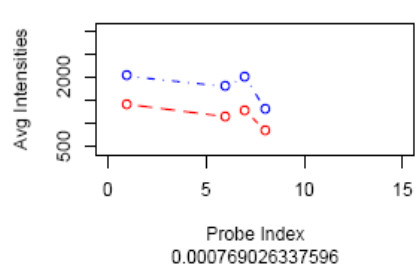
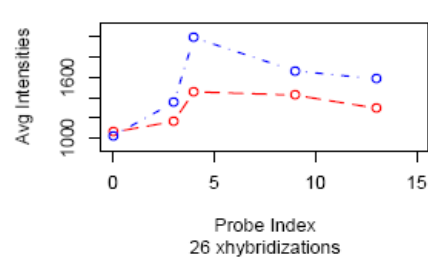
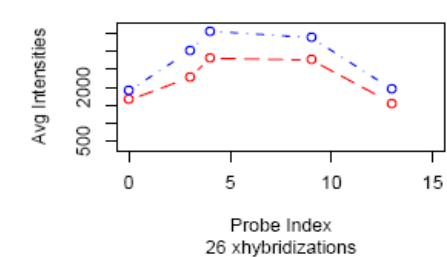


Cleansed Probes, Batch 3 Removed



RNA digestion plot

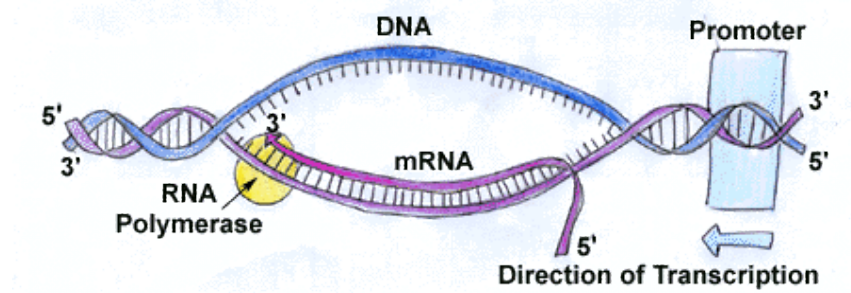


Bhattacharjee Data**Stearman Data****Bhattacharjee Xhybrids****Stearman Xhybrids****Bhattacharjee 32629_f_at****Stearman 32629_f_at****Bhattacharjee 32629_f_at****Stearman 32629_f_at****Bhattacharjee 39867_at****Stearman 39867_at****Bhattacharjee 39867_at****Stearman 39867_at**

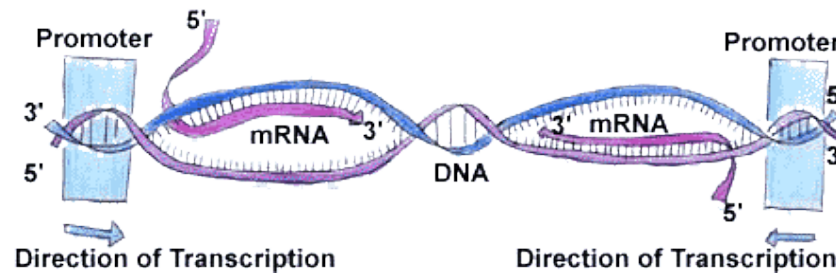
Genes, Targets and Probes

- Gene Expression microarrays bring together two nucleic acid reactants, probes and targets, to get the duplex products.
 - Probes must correspond to regions of genes that will appear in the mRNA
 - Design depends on prior knowledge of sequence and recognition of those regions that will appear in the mRNA
 - Targets must be complementary to probes, and must be quantitatively purified and intact, and must be homogeneously labeled

General Process of Transcription



Off one strand

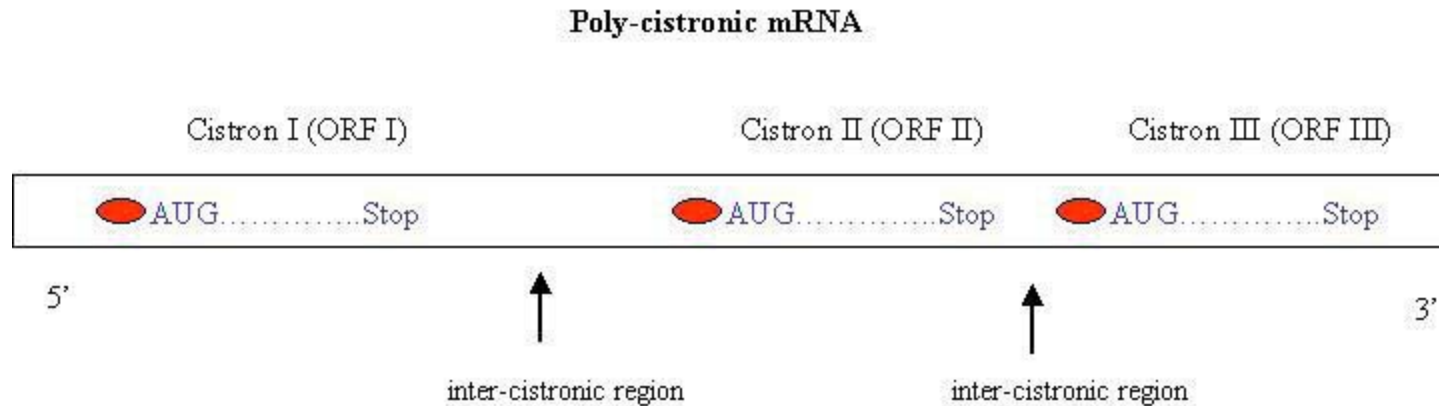


Off both strands


Gene Structure

- Units: genomic DNA occurs as a double-stranded helix and the unit of measure for genes is basepairs
 - Written in the 5' to 3' orientation (usually implied)
 - the molecule type is inferred by whether T or U is used
 - ss vs ds is implied from molecule type.
- Prokaryotic gene structure components
 - Polycistronic mRNAs
 - Translation signals (Shine-Dalgarno sequence)
 - Overlapping start positions
- Eukaryotic gene structure components
 - Introns
 - Primary vs mature transcript
 - Alleles or sequence variants
 - Determine whether they are real or from a sequencing artifact
 - Alternatively spliced forms

Prokaryotic mRNA structure



Prokaryotic mRNA **does not** have cap or polyA tail

 Shine-Dalgarno consensus sequence: **AAGGAG**

Deviations from this sequence occur, but it always is 6-7 purines. This sequence is recognized by the small subunit of the ribosome. Always resides 5-10 bases upstream to AUG

GENERAL STRUCTURE OF PROKARYOTIC mRNA

Eukaryotic Gene Structure

- Human genes on average have 6– 8 introns and exons each.
 - Exons average 100-200 nt apiece.
 - Introns average 1000 kb apiece.
 - An average gene has 10kb, but the Duchenne MD gene has 75 exons and covers 2.4 million bases.
 - GeneScan is a well-regarded tool for the analysis of eukaryotic genomes.

Eukaryotic UTR mRNA structure

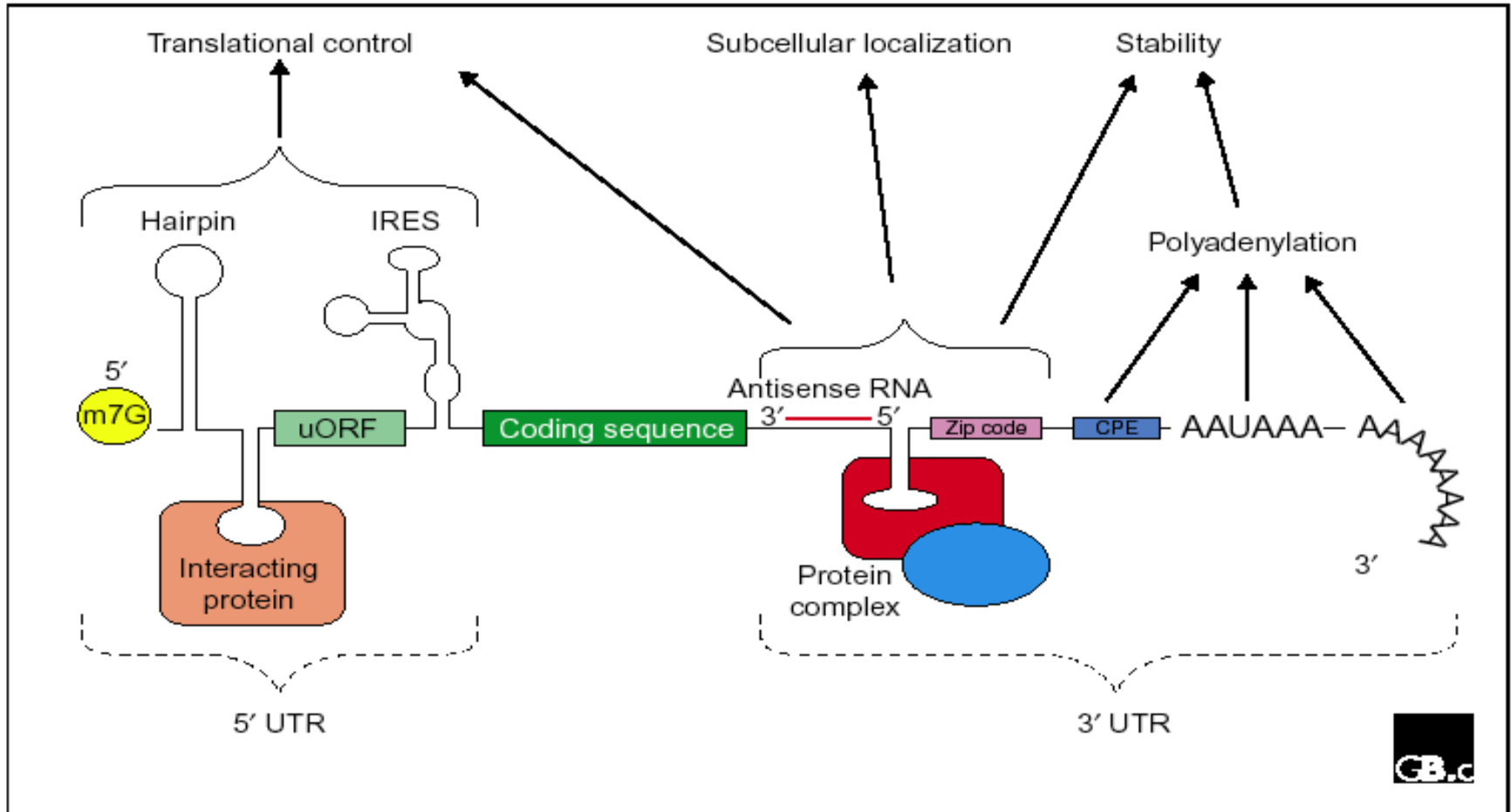


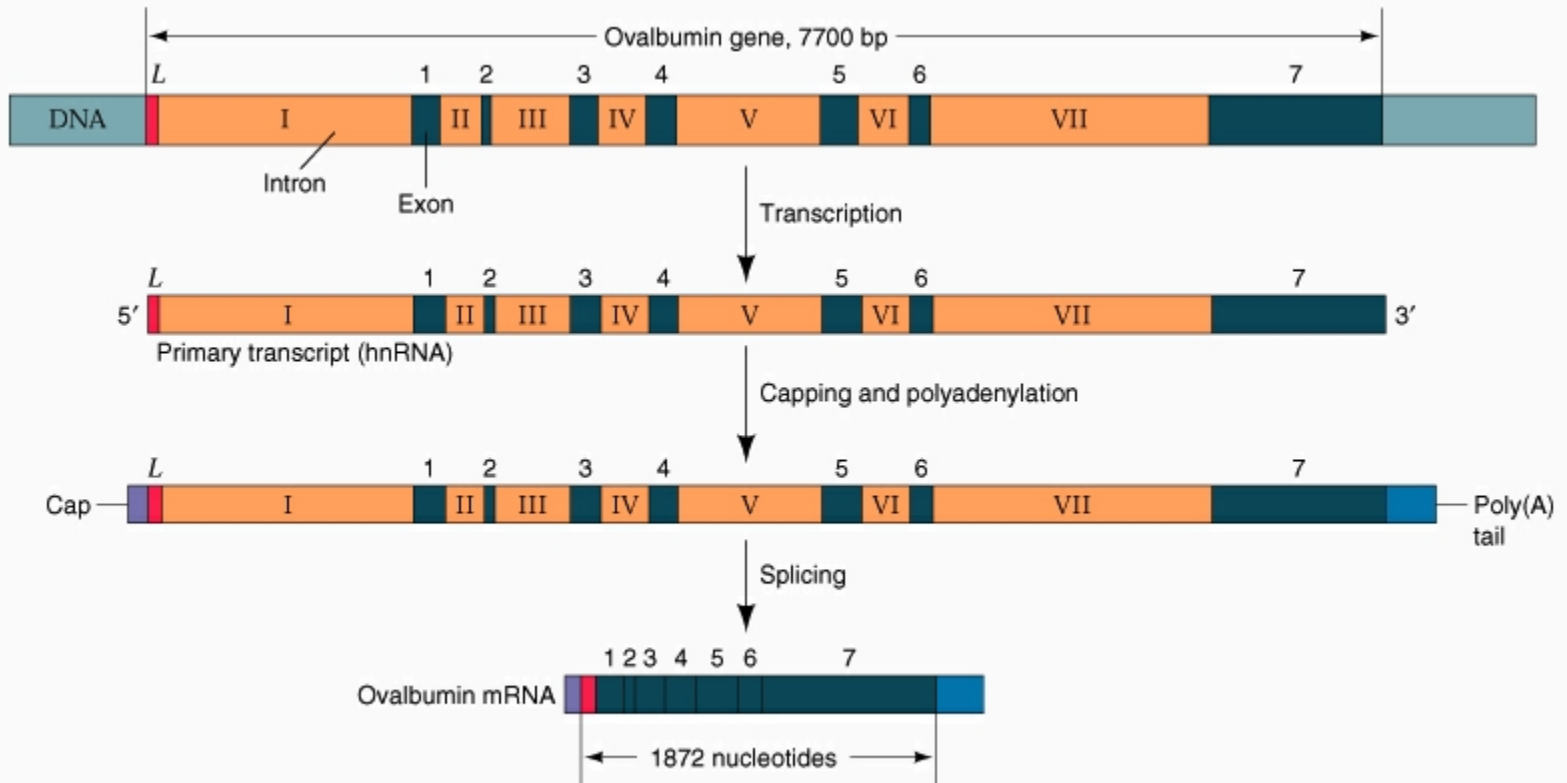
Figure 1

The generic structure of a eukaryotic mRNA, illustrating some post-transcriptional regulatory elements that affect gene expression. Abbreviations (from 5' to 3'): UTR, untranslated region; m7G, 7-methyl-guanosine cap; hairpin, hairpin-like secondary structures; uORF, upstream open reading frame; IRES, internal ribosome entry site; CPE, cytoplasmic polyadenylation element; AAUAAA, polyadenylation signal.

Eukaryotic mRNA processing

- In eukaryotes the *primary mRNA transcript* undergoes many processing steps.
 - The primary transcript has
 - untranslated regions (UTRs at both the 5' and 3' termini) that are important for regulation and correct processing
 - intergenic sequences (introns) between the segments that code for protein (exons) and must be removed by the splicing apparatus.
 - Signals for addition of polyA
- The mature message has
 - A 'cap' structure, a modified G nucleotide is added to the 5' end
 - A polyA tail is added to the 3' end (this increased the stability, or half-life, of the mRNA)
 - The introns are spliced out

Introns and Exons

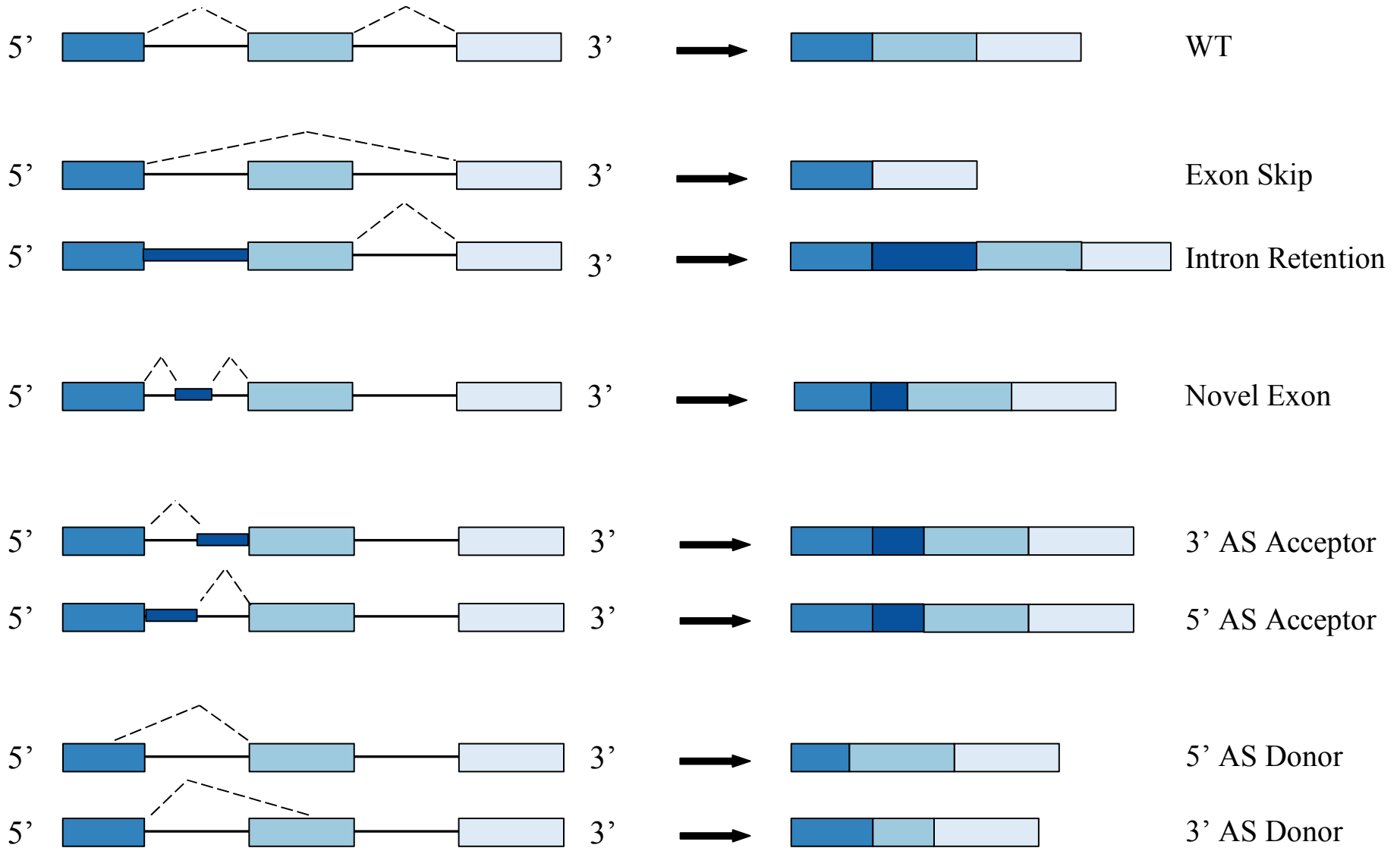


Splice variants

- *Splice variants*: Alternative splicing allows exons from the same nuclear mRNA to be combined in different ways to give different proteins.
 - Alternative splicing often occurs when a protein specific for a type of cell is made.
 - The sequence at the boundary-junctions between introns and exons provide the information about the possibility for forming splice forms.
 - Bioinformatics tools that can predict splice variants include Gene Finder and HMMgene.

Nucleic mRNA

Cytosolic mRNA



Microarray Grand Tour

- The organismal biology and the trait explored impose restrictions on the design of probes:
 - Gene expression vs SNP analysis
 - Which exons should be monitored? Which variants are present at SNP loci and at what frequency? What if multiple SNPs are present in the 25-mer?
 - The type of sample population
 - Does out-crossing occur? Are you using a highly inbred cell line?
 - The amount of material likely to be obtained from an average sample
 - A biopsy gives a small amount of heterogenous tissue that must be shared with the histopathologist
 - Variability arises from both biological and technical sources – the technical sources you can (perhaps) control, the biological you cannot.

Generic Features of a Microarray Experiment

- Select or design an appropriate microarray
- Set up and perform experiments on organisms / cells
 - If this is a gene expression experiment, decide on the conditions that are going to be tested and whether secondary assays will be used to verify results for particular genes of interest
- Collect the samples, process to purify the target, which is mRNA for gene expression arrays
 - If using mRNA, decide if you want to transform it to cDNA, and whether you want to perform other procedures such as cloning, amplification, etc.
 - Quantify the final material, QC for integrity.
- Label the sample material, standardize the targets as much as possible (parameters such as average length affect the hybridization time) to control variation.
- Hybridize the sample to the array.
- Remove unreacted material (whatever does not find a binding partner stable enough to resist the wash conditions).
- Develop the signal
- Collect the signal, then process and proceed to the analysis.

Experimental Design

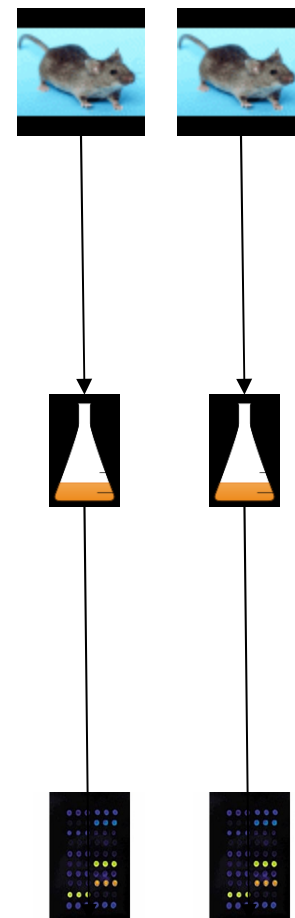
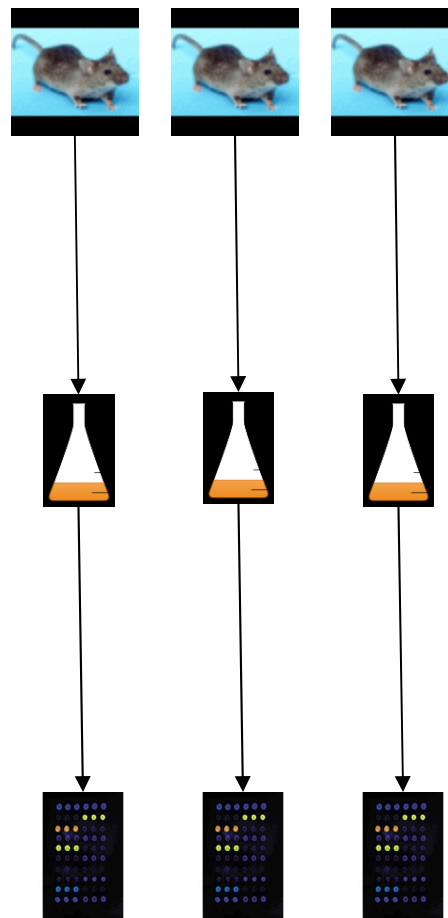
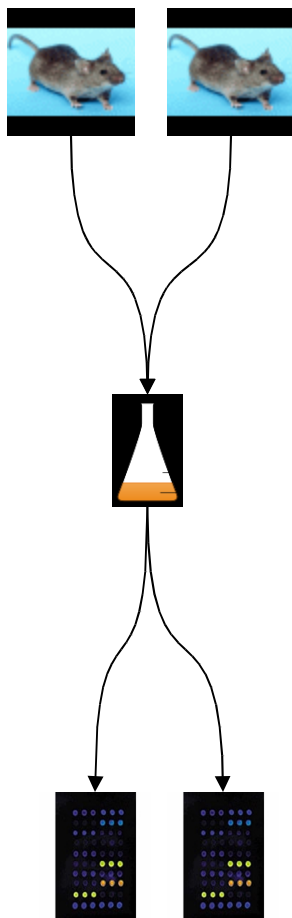
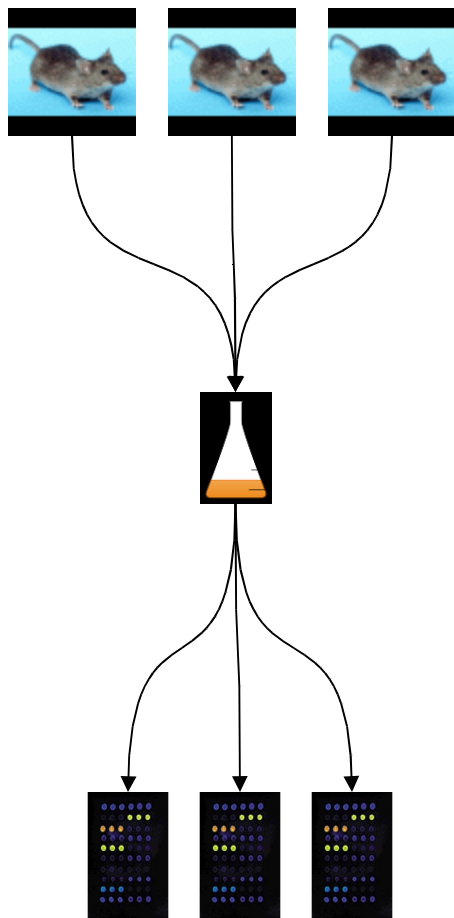
- Step 1: Collecting sequence information requires access to public sequence repositories
 - Does a completed reference genome exist? If so what else is known?
 - Annotation level
 - Are you starting from data from one or more EST libraries?
 - Constructing UniGene sets or some other normalization method.
 - Are you starting with a random collection of sequences from an incomplete genome?
 - Should you use a comparative genomics approach?
- Step 2: Constructing a target list – several categories are common
 - Experiments involving single organisms
 - Experiments involving mixtures of organisms
 - Using pre-designed target lists

Group A

Group B

Group A

Group B



Technology Issues

- Step 3: Selecting the type of platform and linker and attachment chemistry
 - The available platforms and their limitations (cDNA, long oligo, short oligo)
 - Surface chemistries differ depending on the type of platform used
- Probe attachment methods affect the sensitivity of the outcome
 - Covalent linker chemistries vs probe deposition and cross-linking
- Printing and deposition methods/ QC
 - Photolithography, ink-jet, pins
- Biophysical limitations imposed on response range for hybridization (temperature, solvent, voltage, concentration, time)

Assay Considerations: Hybridization & Scanning

- Hybridization
 - Methods and shortcomings
 - Mixing in thin layers
 - Time to equilibrium
 - Important chemical reaction parameters
 - Kinetics and thermodynamics
 - Hybridization controls
- Data Acquisition: Scanning
 - Excitation/detection platforms
 - Selection criteria
 - Common scanner errors
 - Scanner controls (two laser settings)

Assay Considerations

- Form of the target molecule
 - mRNA, cRNA, aRNA, cDNA
 - Intact molecule or fragmented
- Type, position and density of signal molecules
 - Radioactive, fluorescent (one-color or two-color), other
 - Direct or indirect
 - End-label or incorporated

Probe Composition

- The goal is to have a probe that unambiguously identifies exactly one target.
 - Assessment is based first upon sequence similarity
 - Optimization of hybridization (composition) and consideration of internal secondary structure (availability)
 - The source of sequence information is databases of public sequence, genomic and expressed
 - Design tools are diverse; those for sequence similarity are better developed than those that optimize hybridization – this is discussed in detail in future lectures

Probe length, Specificity and Sensitivity

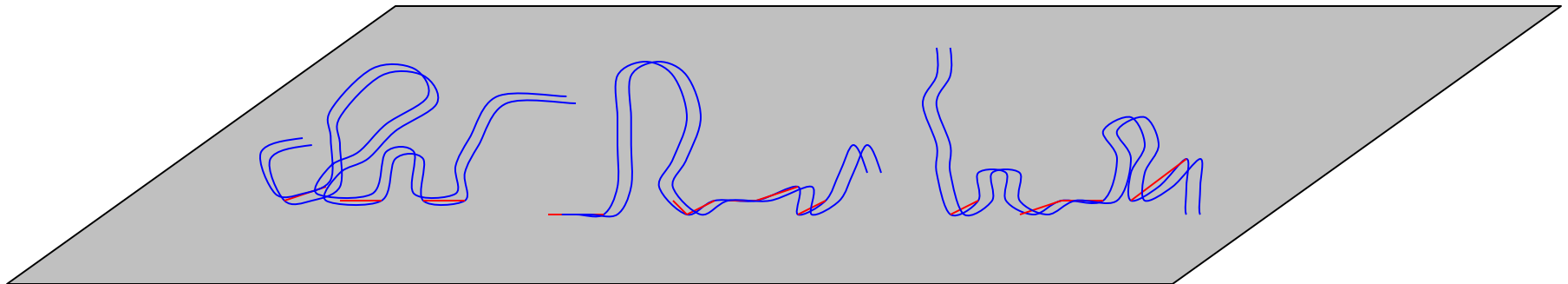
- Nucleic acid hybridization gives rise to double-stranded molecules stabilized with hydrogen bonds. The stability of the duplex depends partly on the length and partly on the composition of the duplex partners.
 - If the composition/complexity is about the same, longer makes for a more stable duplex.
 - A more stable duplex is more able to *tolerate* small inaccuracies,
 - what it reports is less accurate for a sequence ‘match’ than a short duplex will report (less specific).
 - A more stable duplex that does have the correct match will be more accurate in reporting the concentration of that target (more sensitive).

Probe to Surface Attachment

- For cDNA arrays, a solution is usually picked up with a fine metal pin and the pin touches the array surface, depositing some of the probe-containing solution.
- The attachment may be very specific, or very non-specific.
 - With cDNA arrays, since there is no unique chemical feature to use, the attachment is non-specific.

Reproducibility of cDNA deposition

- How the PCR product is ‘attached’ to the array surface is very inconsistent – the actual concentration of a particular subset of the sequence varies
- The PCR product is double stranded, so there is competition from re-annealing to itself after the pre-hybridization melt step.



Informatics Issues

- Step 4. Design of controls
 - Signal controls (detection of a given fluor)
 - Response Range controls (response range of the scanner)
 - Technical replicates (spots)
 - Hybridization controls (spiked in targets)
 - Slide variability controls (labeled probes at a variety of locations and concentrations)
 - Randomization
- Step 5. Design of probes
 - Criteria to consider
 - Sequence
 - Composition
 - Biophysical structure and accessibility
 - Software tools
 - Independent evaluation for significant responses (qRT-PCR)
 - Validating existing designs (spike in different alleles)
 - Iterated design methods

Signal Conversion & storage

- Data conversion: Image analysis
 - Extracting the image
 - Formats
 - Algorithms
 - Data
 - Software
 - Capabilities and weaknesses
- Data Storage
 - Data standards
 - MIAME
 - Additional (protocols and procedures)
 - Common data formats
 - Software
 - Capabilities and weaknesses, public repositories

Measuring the Performance of Technical Parameters

- Experimental Design
 - Slide design
 - Randomization or related genes
 - Pseudo-Random dispersion of controls
 - Determining the appropriate number of replicates
 - Applying appropriate statistical tests
 - Achieving economies of scale
 - Storage of oligos and slides
 - Design informatics
 - MIAME checklist for reference parameters
 - » Manufacturing parameters
 - » Probes sequence + location
 - » Standard nomenclature for array layout and related annotation
 - Databases

Extracting information

- Data Preprocessing and Data cleansing
 - Using signal controls
 - Filtering for linear range and confounding factors (such as cross-hybridization)
 - Normalization/ standardization
- Data Analysis and Visualization
 - Classifiers
 - Standard methods
 - Clustering, PCA, SOMs, K-means
 - Handling time series data (linear and multifactorial)
 - Statistical validation
 - Software tools

Caveats to Expression Monitoring

- The assumption in many gene expression microarray experiments is that observations of the modulation of all the mRNAs in response to a treatment can lead to an understanding of transcriptional control and regulation.
- At the cellular level, most of the phenotype (morphology and behavior) is due to the action of proteins and metabolites.
 - There is a strong correlation between the concentration of mRNA and its protein for many genes but a good number of exceptions have also been shown.
 - The half-lives of a mRNA and its protein are not always proportional (the proteins of the cellular matrix turn over at a far slower rate than do the mRNAs)
 - The half-life of different mRNA species is quite different (the length of the polyA tail is known to correlate well with this) – you would need to monitor at the appropriate intervals to capture this behavior.
 - This is not trivial: a change in the half-life of dystrophin mRNA is part of the cause of MD. However a steady-state and constant across all species is an implicit assumption of many analyses.
 - Metabolites may cause cellular responses without a transcriptional event being required (nerve excitation, hormonal release, muscle contraction)

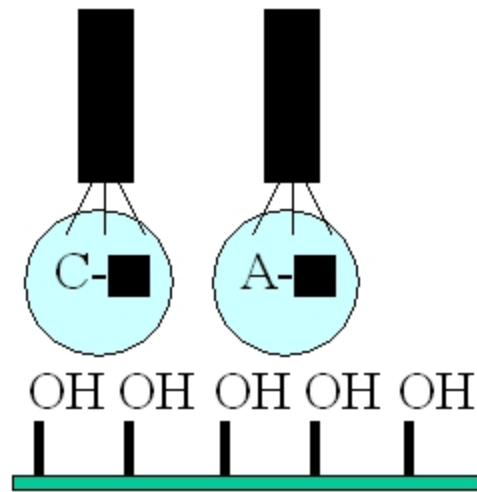
Completion of Grand Tour

- Nature BioTechnology has had two specialized issues covering many of these topics, called the Chipping Forecasts.
 - Available on-line. The article by Ed Southern is a good summary of many of the principles discussed next.

in situ Probe Synthesis

- Short oligonucleotide probes may be built by synthetic organic chemistry.
 - The chemistries available for surface reaction are rather limited
 - The efficiency of the reaction is such that after ~25 bases failure sequences start to become a significant fraction of the total
 - the actual concentration of the true probe is a low

Ink Jet Synthesis



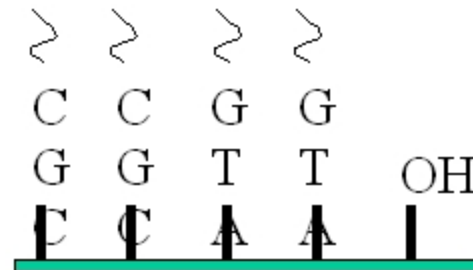
1) Deposit phosphoramidite



2) After coupling

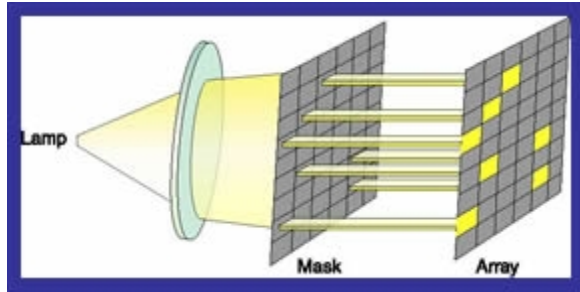


3) After deprotection

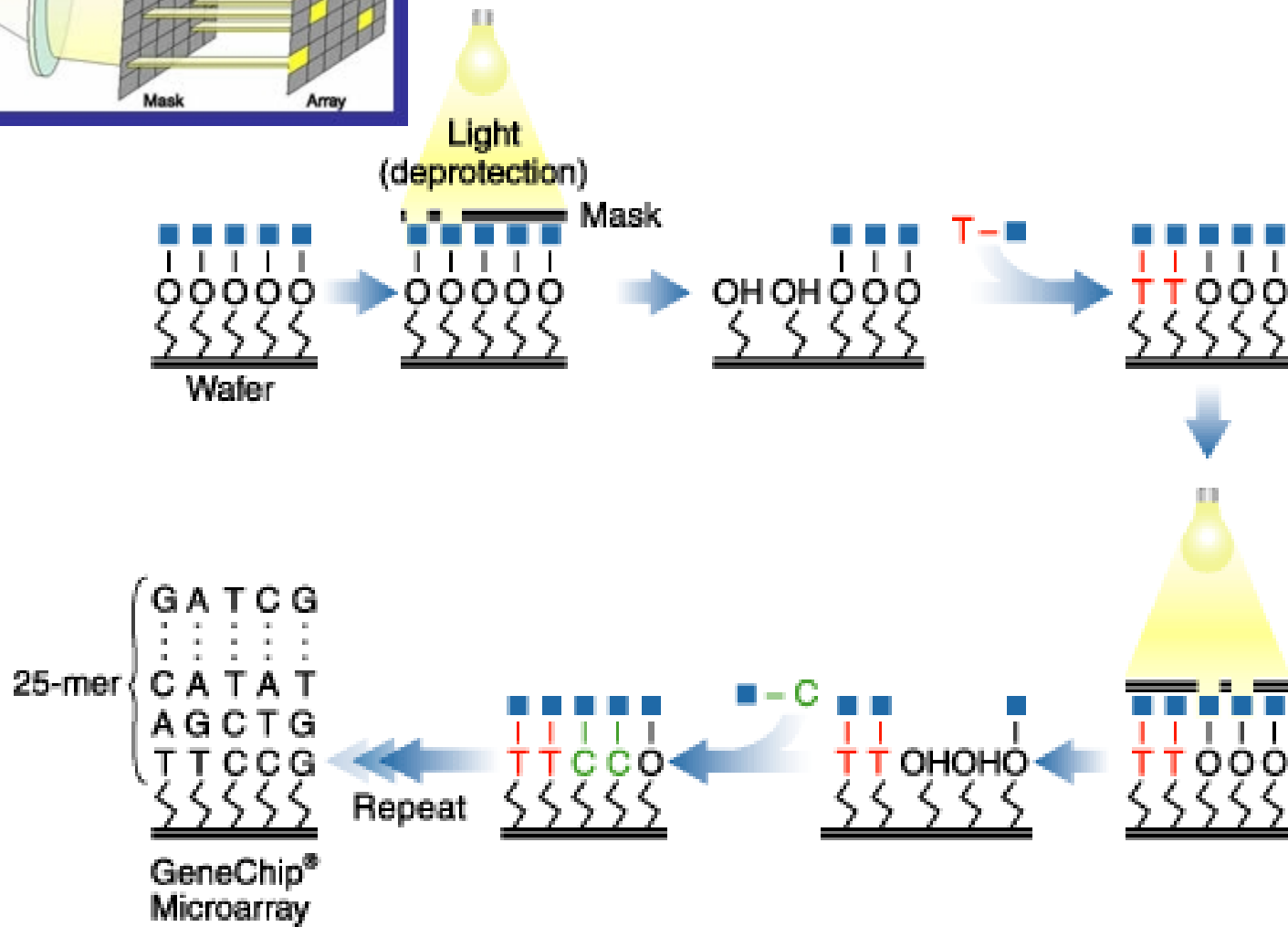


4) Repeat

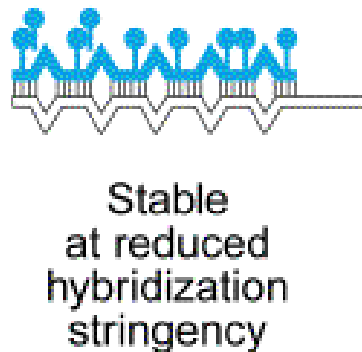
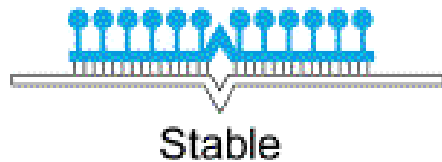
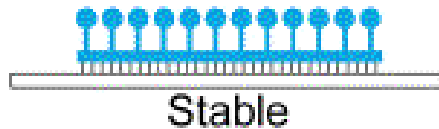
Affymetrix: photolithography



<http://www.bcm.edu/mcfweb/Images/new%20images/affy3.jpg>



Conventional
DNA probe

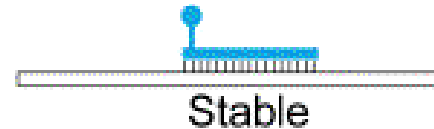


Perfect
match

Single
mismatch
(e.g. allelic)

20% mismatch
(e.g. coding
sequences of
human and
mouse genes)

Oligonucleotide
probe



Microarray Types

- Gene Expression Microarrays
 - For one target genome, regions that are expressed as transcripts
- SNP microarrays (SBH)
 - For one genome, tiling across segments known to have a single nucleotide polymorphism
- Exon Hit microarrays
 - For one genome, designed to have at least one probe per exon and one per intron-exon junction
- Comparative Genome Hybridization Microarrays (CGH)
 - Have large immobilized regions of target organism chromosome, used to identify regions of mutation (especially deletion and rearrangement)
- Chromatin ImmunoPrecipitation Microarrays (ChIP on chip)
 - Protein and DNA are cross-linked, immunoprecipitated samples are compared to control samples by hybridization to DNA probes on an array
- Protein interaction microarrays (e.g. SELDI)
 - Either a protein or chemical group that a protein is known to react to is attached as bait for proteins in the experimental mixture. Identification may be by an antibody or mass spec
- Diagnostic microarrays
 - For a number of genomes, usually designed against a small number of unique features within each genome.

Non Gene-Expression Experiments

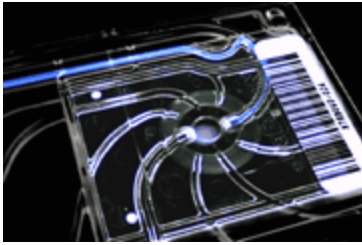
- Gene variant monitoring
 - Alternatively spliced transcripts or gene family assays
- Mutation detection
 - SNPs, small indels
- Genomic structure
 - Monitoring large changes and rearrangements in a genome, also loss of heterozygosity
- Gene or organism monitoring
 - Put sequences unique to an organism or class of organisms (ribosomal genes are often used)
 - Put multiple genes from a given organism on the chip, as well as common variants
- Re-sequencing and sequencing by hybridization
 - For regions of interest you can walk down the sequence a base at a time, including all four possible changes at the test location
- DNA-protein interactions
 - Attach DNA, put on a protein solution, identify the attached protein with Mass Spec, and antibody stain or some other method
- DNA-small molecule interactions
 - Allow small molecules to bind, identify with Mass Spec.

Commercial Suppliers

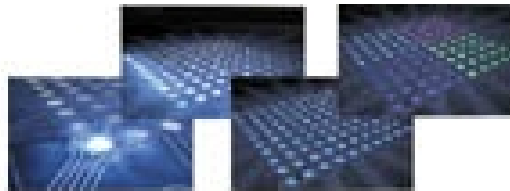
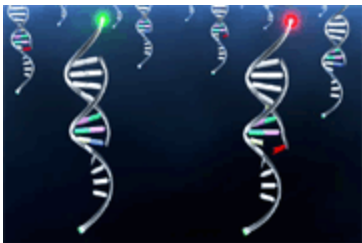
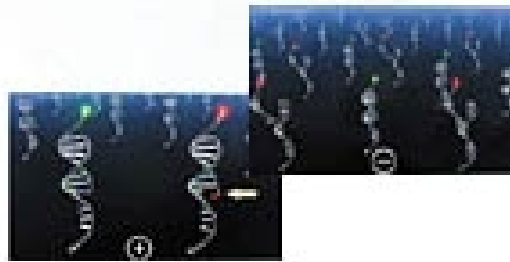
- Affymetrix
 - Nimblegen
 - Agilent
 - Nanogen
 - Illumina
 - Codelink
 - Counting methods (SAGE, MPSS, AFLP...)
 - Many others.....
-
- Multiple-laboratory comparison of microarray platforms, Irizarry, *et al*, Nature Methods, 2005

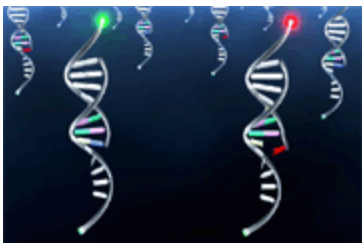
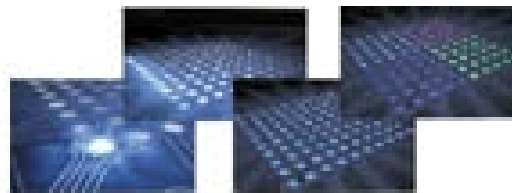
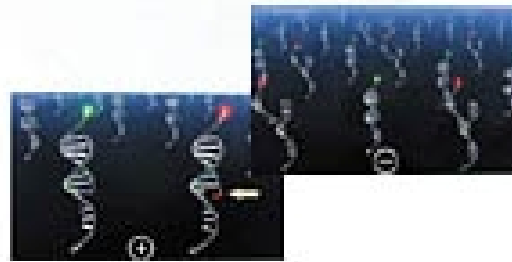
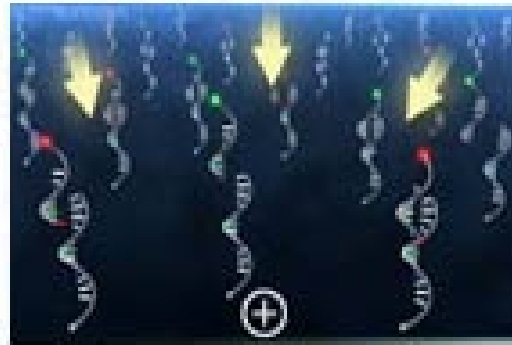
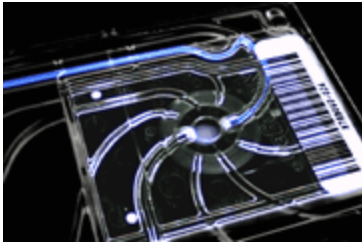
Nanogen

- Electronic positioning of target molecules
 - A cDNA molecule has a negative charge, so it will move to an area of net positive charge
 - Rather than sloshing a solution around and waiting for molecules to move down to the probe interface you can use electronic charge to pull them into the interaction region.
 - You can also use charge to alter the stringency of the probe-target interaction.
 - molecular binding onto the NanoChip® microarray is accelerated up to 1000 times
 - From the Nanogen Web site
http://www.nanogen.com/technology/core_technology.



“Applying an electric current to individual test sites on the NanoChip® microarray enables rapid movement and concentration of the molecules.



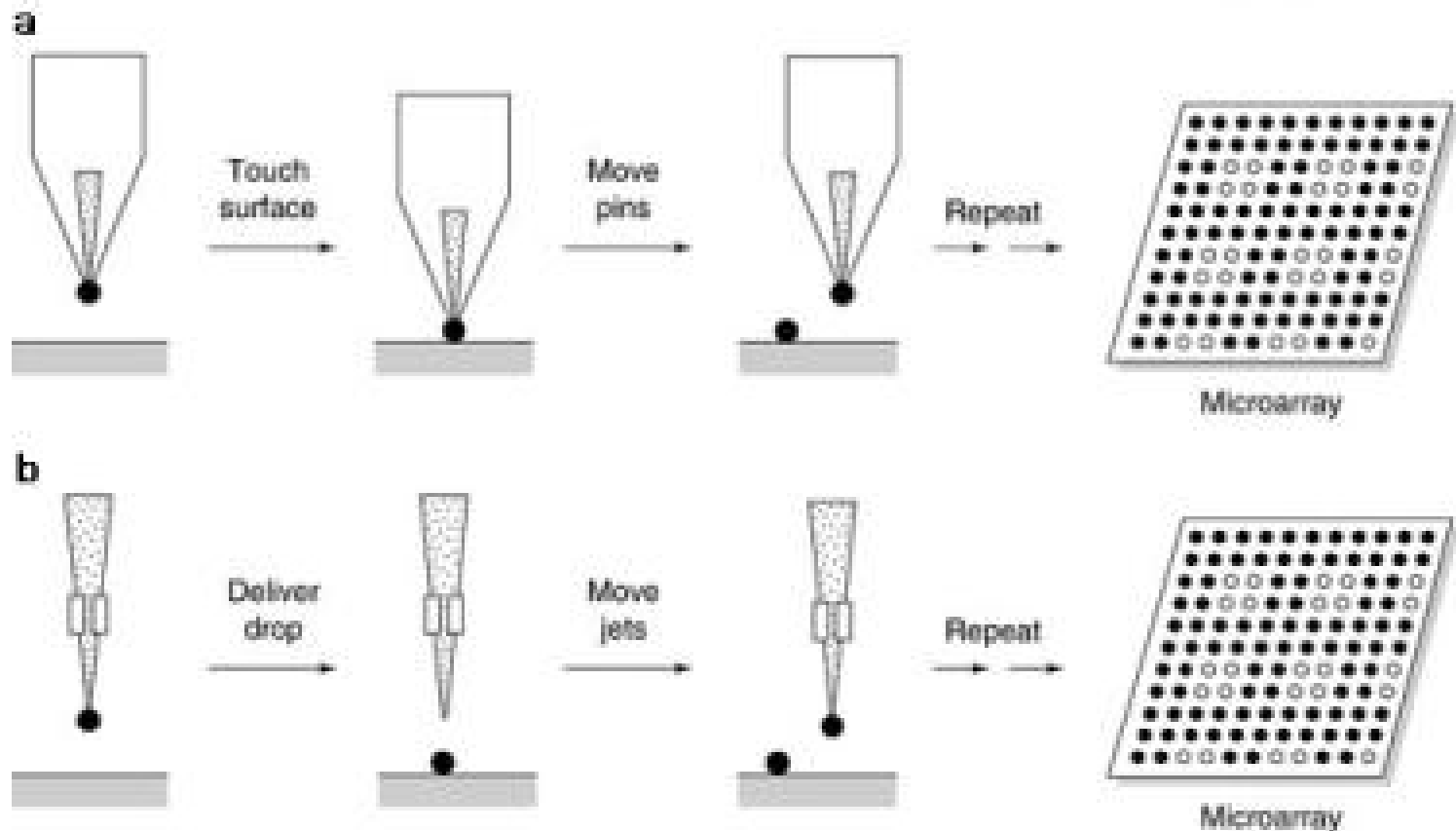


Applying an electric current to individual test sites on the NanoChip® microarray enables rapid movement and concentration of the molecules. Through electronics, molecular binding onto the NanoChip® microarray is accelerated up to 1000 times faster than traditional passive methods. Nanogen's technology involves electronically addressing biotinylated DNA samples, hybridizing complementary DNA reporter probes and applying stringency to remove unbound and nonspecifically bound DNA after hybridization.

Agilent

- Ink-jet spotted arrays can be made in two ways:
 - Deliver reagents to the spot using an ink-jet device, spraying them onto the target area in a very directed fashion
 - Deliver oligonucleotides to the location using an ink-jet device
 - In this case, the oligonucleotides are pre-synthesized (with organic chemistry methods) with correct attachment modifications, then purified, and sequence-validated as needed.

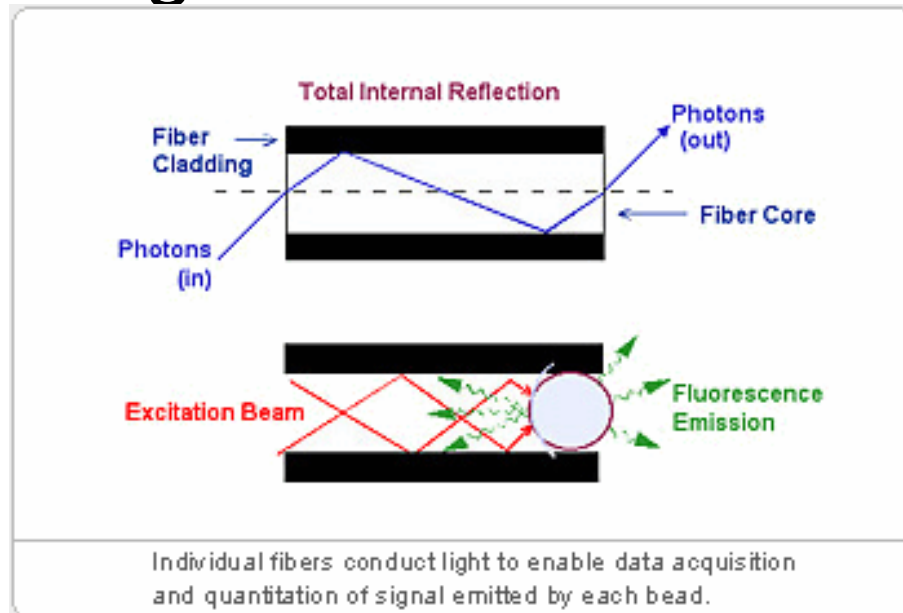
Targeted Reagent Delivery: Drop vs ink-jet



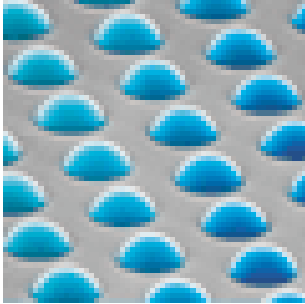
http://bric.postech.ac.kr/bbs/biostat/2001/images/nondan2_3.jpg

Illumina

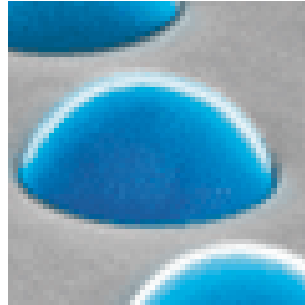
- Oligonucleotides are linked to beads, each bead is color-coded. The beads are bound in the tips of optical fibers, which are then bundled together.



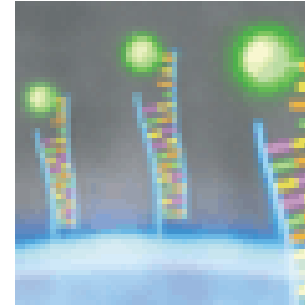
Illumina Bead Array



Ten of thousands of wells with hundreds to thousands of bead types can be assembled into each array bundle



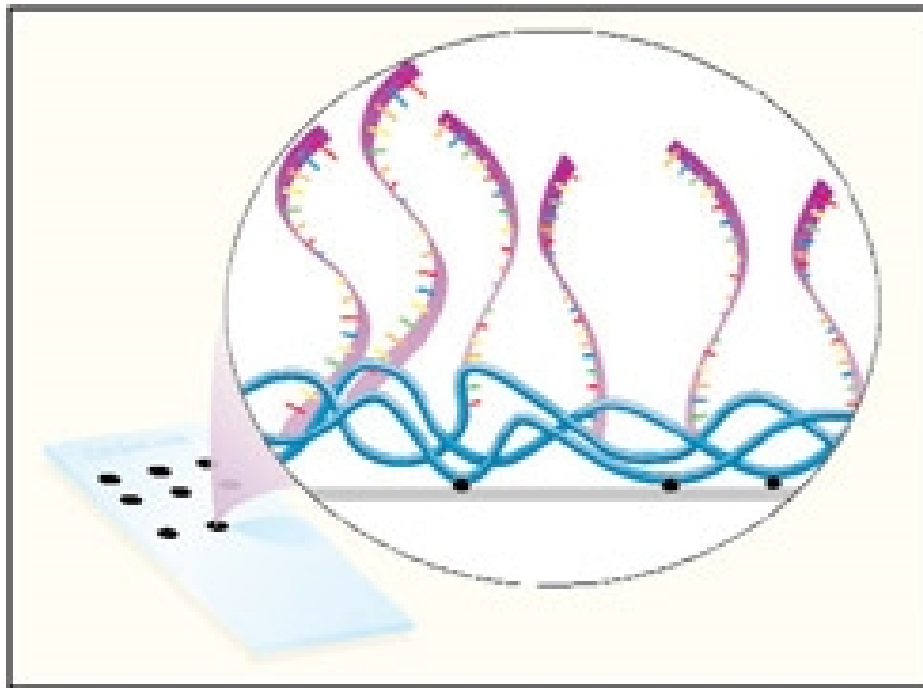
Determine which bead type occupies which well using a proprietary decoding process



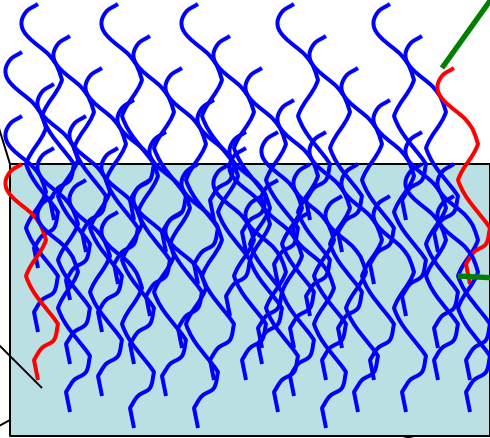
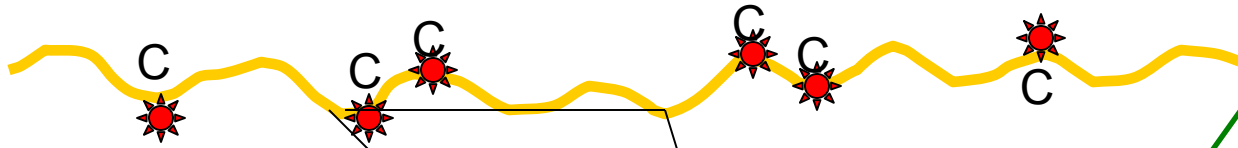
The molecules in the sample bind to their matching molecules on the coated bead

CodeLink

- Motorola/Amersham



Gel pad technology: using an acrylamide gel 3-D pad you can imbed oligonucleotides – the gel matrix does not seem to inhibit diffusion and you can get a high density of probes without limiting access.



(X_4, Y_1)

X_1 →

Y_1
↓
