9. MICROARRAYS - 3

BIOINFORMATICS COURSE
MTAT.03.239

07.11.2012
MICROARRAYS

Slides adapted from Konstantin Tretyakov’s 2011/2012
MICROARRAYS

RNA

http://www.sciencephoto.com
HOW TO DESIGN AN EXPERIMENT?

• 16 rats 🐭 🐭
• 2 array batches (print runs)
• two scientists: Alison and Brain
EXPERIMENTAL DESIGN

Experimental design 1

Alison chooses 8 rats and treats them with benzopyrene. She prepares liver samples from the rats and hybridizes them to the 8 arrays from the first print run. Brian takes the remaining rats and treats them with control substance; he prepares samples and hybridizes them to the 8 arrays from the second print run.
EXPERIMENTAL DESIGN

Experimental design 2

Alison chooses eight rats, and treats four with benzopyrene and four with control substance. She chooses four arrays from each of the print runs and hybridizes samples from two treated rats and two control rats to each of the batches of four arrays. Brian does likewise with the remaining eight rats and eight arrays.
EXPERIMENTAL DESIGN

Experimental design 3

Eight rats are randomly allocated to Alison; similarly, four arrays from each of the two print runs are also randomly allocated to Alison. Four preparations of benzopyrene and four preparations of control compound are given to Alison in a way that she does not know the identity of any of the preparations. The arrays are prearranged for Alison so that she will hybridize two treated and two control rats to four arrays from each batch, with random allocation. Brian does likewise.
EXPERIMENTAL DESIGN

Experimental design 1

Alison and Brian might handle the samples differently, which might lead to technical and systematic difference in gene expression.

Experimental design 2

A bias problem. Alison might choose rats that are in some way similar. They might be the healthiest looking or the most docile, or something else. Also one might treat the poisoned rats with greater care than the control rats.

Experimental design 3

Uses randomization to overcome the bias problem. WINNER!
MICROARRAYS

GENE EXPRESSION

• What is it?
• How it is measured?
MICROARRAYS

**GENE EXPRESSION**
- What is it?
- How it...

**EXPRESSION DATA**
- Data formats and databases
- Normalization
- Differential expression analysis
MICROARRAYS

**GENE EXPRESSION**
- What is it?
- How it is measured?

**EXPRESSION DATA**
- Data formats and databases
- Normalization
- Differential expression analysis

**VISUALIZATION**
- Line plots
- Heatmaps
- Scatterplots (PCA)
MICROARRAYS

GENE EXPRESSION
• What is it?
• How it is measured?

EXPRESSION DATA
• Data formats and databases

VISUALIZATION
• Line plots
• Heatmaps
• Scatterplots (PCA)

CLUSTERING
• Distances
• Hierarchical
• K-means

“Microarrays 3" Bioinformatics Course
MICROARRAYS

Gene expression
- What is it?
- How it is measured?

Expression data
- Data formats and databases

Visualization
- Line plots
- Heatmaps
- Scatterplots (PCA)

Clustering
- Distances
- Hierarchical
- K-means

Enrichment analysis
- Gene ontology
- Hypergeometric test
- Multiple testing correction
- Tools like g:profiler, MEM
APPLICATION OF MICROARRAYS

- **Gene discovery**
  Identification of new genes and their function and level in different conditions.

- **Molecular classification of complex diseases**
  To classify the different cancer types based on the patterns of gene activity in the tumor cells, to develop drugs specifically designed for these cancers.

- **Drug discovery**
  Comparative analysis of the genes from a diseased and a normal cell help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. This information can be used to synthesize drugs that combat with these proteins and reduce their effect.

- **Toxicological research**
  Microarray technology provides a robust platform to study the effect of different toxins on the cells and its progenitors.

http://estbioinfo.stat.ub.es/docs/jlmosquera/teaching/BIOINF/MDALab/Lecture1-IntroductionToMicrarrayDataAnalysis.pdf

“Microarrays 3" Bioinformatics Course
APPLICATION OF MICROARRAYS

- Gene discovery
  Identification of new genes and their function and level in different conditions.

- Molecular classification of complex diseases
  To classify the different cancer types based on the patterns of gene activity in the tumor cells, to develop drugs specifically designed for these cancers.

- Drug discovery
  Comparative analysis of the genes from a diseased and a normal cell help the identification of the biochemical constitution of the proteins synthesized by the diseased genes. This information can be used to synthesize drugs that combat with these proteins and reduce their effect.

- Toxicological research
  Microarray technology provides a robust platform to study the effect of different toxins on the cells and it’s progenitors.

What else can we study with microarrays besides gene expression?

http://estbioinfo.stat.ub.es/docs/jmosquera/teaching/BIOINF/MDALab/Lecture1-IntroductionToMicrarrayDataAnalysis.pdf
# TYPES OF MICROARRAYS

<table>
<thead>
<tr>
<th>Array</th>
<th>Probes</th>
<th>Targets</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene expression arrays</td>
<td>Gene representatives (cDNA; oligos, ...)</td>
<td>mRNA/cDNA</td>
<td>Gene expression</td>
</tr>
<tr>
<td>Comparative Genomic Hybridization (CGH) array</td>
<td>DNA</td>
<td>DNA</td>
<td>Genomic changes in cancer</td>
</tr>
<tr>
<td>SNP arrays</td>
<td>DNA (oligos)</td>
<td>DNA</td>
<td>Genotyping; genomic changes</td>
</tr>
<tr>
<td>Methylation arrays</td>
<td>DNA (CpG islands)</td>
<td>DNA (bisulfite-treated)</td>
<td>Methylation-status in genes</td>
</tr>
<tr>
<td>Promoter arrays</td>
<td>DNA (promoter 1kb)</td>
<td>DNA (ChIP-enriched)</td>
<td>Transcription factor binding sites</td>
</tr>
<tr>
<td>Tiling arrays</td>
<td>DNA</td>
<td>DNA/mRNA</td>
<td>All of the above; sequencing; gene annotation</td>
</tr>
<tr>
<td>Protein arrays</td>
<td>Antibody</td>
<td>Protein</td>
<td>Protein expression (ELISA)</td>
</tr>
<tr>
<td>Tissue arrays</td>
<td>Tissues</td>
<td>Proteins</td>
<td>Histology; protein expression; Immunohistochemistry; in situ hybridization</td>
</tr>
</tbody>
</table>
## TYPES OF MICROARRAYS

<table>
<thead>
<tr>
<th>Array</th>
<th>Probes</th>
<th>Targets</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene expression arrays</td>
<td>Gene representatives (cDNA; oligos, ...)</td>
<td>mRNA/cDNA</td>
<td>Gene expression</td>
</tr>
<tr>
<td><strong>Comparative Genomic Hybridization (CGH) array</strong></td>
<td>DNA</td>
<td>DNA</td>
<td>Genomic changes in cancer</td>
</tr>
<tr>
<td>SNP arrays</td>
<td>DNA (oligos)</td>
<td>DNA</td>
<td>Genotyping; genomic changes</td>
</tr>
<tr>
<td>Methylation arrays</td>
<td>DNA (CpG islands)</td>
<td>DNA (bisulfite-treated)</td>
<td>Methylation-status in genes</td>
</tr>
<tr>
<td>Promoter arrays</td>
<td>DNA (promoter 1kb)</td>
<td>DNA (ChIP-enriched)</td>
<td>Transcription factor binding sites</td>
</tr>
<tr>
<td>Tiling arrays</td>
<td>DNA</td>
<td>DNA/mRNA</td>
<td>All of the above; sequencing; gene annotation</td>
</tr>
<tr>
<td>Protein arrays</td>
<td>Antibody</td>
<td>Protein</td>
<td>Protein expression (ELISA)</td>
</tr>
<tr>
<td>Tissue arrays</td>
<td>Tissues</td>
<td>Proteins</td>
<td>Histology; protein expression; Immunohistochemistry; in situ hybridization</td>
</tr>
</tbody>
</table>

http://www.epibiostat.ucsf.edu/biostat/cbmb/bmi209.fall05/lectures/bmi209fall05-lecture1.pdf
CHG

CHG was developed to survey DNA copy-number variations across a whole genome. With CGH, differentially labeled test (i.e. tumor) and reference (i.e. normal individual) genomic DNA are cohybridized to normal metaphase chromosomes, and fluorescence ratios along the length of chromosomes provide a cytogenetic representation of the relative DNA copy-number variation. The resolution is limited to 10-20 Mb - therefore anything smaller than that will not be detected. Researchers can very precisely determine the chromosomal regions and genes that are amplified or missing.
SINGLE NUCLEOTIDE POLYMORPHISMS (SNP)

GOOD MATCH STICKS, BAD ONE DOESN’T

http://www.csus.edu/indiv/r/rogersa/Bio181/GenechipHO.pdf
SNP’S CAN AFFECT

- Protein sequence if located in the exon and a nonsense or missense mutation occurs.
- Alternative splicing if at critical location in the intron.
- The binding affinity of a transcription factor if located in the DNA binding site.
SNP

- Two unrelated people share about 99.5% of their DNA sequence.
- The most common difference is in single nucleotide positions where a subpopulation can (for instance) have a G and others might have an A.
- Each variant of a base that can occur in the SNP throughout the population is called an allele.
- For each SNP, an individual’s genotype is the specific combination of alleles that it possesses.
- The map of SNPs serves as an excellent genotypic marker for research.
HAPMAP PROJECT - SNP

rs290485: Allele Frequencies in HapMap Populations

<table>
<thead>
<tr>
<th>Panel</th>
<th>Description</th>
<th>Frequency of A (ref)</th>
<th>Frequency of T</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASW(A)</td>
<td>African ancestry in Southwest USA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CEU(C)</td>
<td>Utah residents with Northern and Western European ancestry from the CEPH collection</td>
<td>82%</td>
<td>18%</td>
</tr>
<tr>
<td>CHB(H)</td>
<td>Han Chinese in Beijing, China</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>CHD(D)</td>
<td>Chinese in Metropolitan Denver, Colorado</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>GIH(G)</td>
<td>Gujarati Indians in Houston, Texas</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>JPT(J)</td>
<td>Japanese in Tokyo, Japan</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>LWK(L)</td>
<td>Luhya in Webuye, Kenya</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>MEX(M)</td>
<td>Mexican ancestry in Los Angeles, California</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>MKK(K)</td>
<td>Maasai in Kinyawa, Kenya</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>TSI(T)</td>
<td>Tuscan in Italy</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>YRI(Y)</td>
<td>Yoruban in Ibadan, Nigeria</td>
<td>97%</td>
<td>3%</td>
</tr>
</tbody>
</table>


"Microarrays 3" Bioinformatics Course
What are the possible genotypes of CEU?
What are the possible genotypes of CEU?

AA, AT, TT


"Microarrays 3" Bioinformatics Course
MUTATIONS / SUBSTITUTIONS

- Adenine (A)
- Guanine (G)
- Cytosine (C)
- Thymine (T)

Transitions:
- A → G
- C → T

Transversions:
- A → T
- C → G

Purines: Adenine and Guanine
Pyrimidines: Cytosine and Thymine

http://en.wikipedia.org/wiki/Transition_(genetics)

“Microarrays 3" Bioinformatics Course
At the top is the fragment of DNA harboring an A/C SNP to be interrogated by the probes shown. (a) In the Affymetrix assay, there are 25-mer probes for both alleles, and the location of the SNP locus varies from probe to probe. The DNA binds to both probes regardless of the allele it carries, but it does so more efficiently when it is complementary to all 25 bases (bright yellow) rather than mismatching the SNP site (dimmer yellow). This impeded binding manifests itself in a dimmer signal. (b) Attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site. The single-base extension (T or G) that is complementary to the allele carried by the DNA (A or C, respectively) then binds and results in the appropriately-colored signal (red or green, respectively). For both platforms, the computational algorithms convert the raw signals into inferences regarding the presence or absence of each of the two alleles.

LaFramboise T Nucl. Acids Res. 2009;37:4181-4193

“Microarrays 3” Bioinformatics Course
SNP ARRAYS

- Used to detect polymorphism within a population.
- Ables to determine the genotypes.
- Each SNP on the array is interrogated with different probes – one for each allele.
- SNPs can be used as markers for genetic diseases that have complex traits.
- Help to design drugs that act on a group of individuals.
- SNP arrays can be used to study loss of heterozygosity (LOH).
DNA METHYLATION

- A chemical modification that predominantly affects the cytosine base of CG nucleotides (CpG).
- A methyl group is added to the 5th carbon position of the cytosine base.
- CpGs are statistically underrepresented in the human genome.
- Methylation in promoter region can prevent transcription.
- CpG-rich sequences in **actively transcribed regions** are mostly **unmethylated** (CpG islands – often associated with gene promoters and regulatory regions).

http://www.cmb.usc.edu/people/stavare/STpapers-pdf/Tetal07.pdf
http://extremelongevity.net/2012/06/12/dna-switches-discovered-to-decline-significantly-with-age/
FIGURE 13.1 Illustration of (a) a methylated CpG dinucleotide. The cytosine and guanine bases are joined by a phosphodiester bond and a methyl group has been added to the cytosine. (b) gives a detailed illustration of double-stranded DNA with methylated CpGs in positive and negative strands.
FIGURE 13.2 Two double-stranded DNA fragments that are both (a) methylated and (b) unmethylated. Filled circles represent methylated CpGs, whereas open circles represent unmethylated CpG sites. Two hemimethylated CpG sites are shown in the lower fragment in (b).
METHYLATION ARRAY PRINCIPLES

- **Enrichment**
  
  Fragments are separated either by being methylated or not being methylated. A methylation-sensitive restriction enzyme digestion can be used.

- **Bisulfite treatment**
  
  Similar to SNP detection. After bisulfite treatment the unmethylated cytosine is converted to uracil.

---

Allele 1 (methylated)

\[ \text{---ACTCCACGG---TCCATCGCT---} \]

\[ \text{---TGAGGTGCCC---AGGTACCGA---} \]

\[ \text{---AUTUUAUGG---TUUATCGUT---} \]

\[ \text{---TGAGGTCGUU---AGGTAGCGA---} \]

Allele 2 (unmethylated)

\[ \text{---ACTCCACGG---TCCATCGCT---} \]

\[ \text{---TGAGGTGCCC---AGGTACCGA---} \]

\[ \text{---AUTUUAUGG---TUUATCGUT---} \]

\[ \text{---TGAGGTCGUU---AGGTAGUGA---} \]
Through a bisulfite conversion step, unmethylated cytosines are converted to uracils, while methylated cytosines remain unchanged. Each CpG locus is represented by two bead types. The U bead type matches the unmethylated CpG site; the M bead type matches the methylated state. In the top figure, the unmethylated CpG target site matches perfectly with the U probe, enabling single-base extension and detection. It has a single-base mismatch to the M probe which inhibits extension. If the CpG locus of interest is methylated (bottom figure), the reverse occurs. When the CpG locus has an intermediate methylation state, both probes match the target site and are extended. Methylation status of a CpG site is determined by the beta-value calculation, which is the ratio of fluorescent signals from the methylated probes to total locus intensity.
ChIP – chromatin immunoprecipitation
CHIP – microarray
POI – protein of interest
TECHNIQUES – LARGE SCALE

- Polymerase chain reaction (PCR)
- Real - Time PCR (qPCR)
- Serial Analysis of Gene Expression (SAGE)
- Northern Blot
- Microarrays
- RNA - seq
TECHNIQUES – SINGLE GENE

- Polymerase chain reaction (PCR)
- Real - Time PCR (qPCR)
- Serial Analysis of Gene Expression (SAGE)
- Northern Blot
- Microarrays
- RNA - seq
PCR

PCR is a process for amplification of specific fragments of DNA.

David A. Palmer,
pathmicro.med.sc.edu/rtpcr/rt-pcr.ppt

“Microarrays 3" Bioinformatics Course
PCR is a process for amplification of specific fragments of DNA.

David A. Palmer,
pathmicro.med.sc.edu/rtpcr/rt-PCR.ppt

“Microarrays 3" Bioinformatics Course
PCR

PCR is a process for amplification of specific fragments of DNA.
REAL-TIME PCR

Real-Time PCR requires a specific probe with a fluorescent reporter.

Fluorescence dye
Reports what is going on in real-time by emitting fluorescent signal.

Quencher
Absorbes the fluorescence signal from the fluorescence reporter
REAL-TIME PCR

Quencher
Absorbes the fluorescence signal from the fluorescence reporter

Fluorescence dye
Reports what is going on in real-time by emitting fluorescent signal.

David A. Palmer, pathmicro.med.sc.edu/rtpcr/rt-PCR.ppt

“Microarrays 3" Bioinformatics Course
REAL-TIME PCR

- Allows us to measure minute amounts of DNA in the sample.
- Amplified DNA is detected as the reaction progresses in real-time.
- If there was four times as much gene 1 DNA template as gene 2 DNA template then gene 1 would reach 1,000,000 copies two cycles earlier than gene 2.

<table>
<thead>
<tr>
<th>Cycle</th>
<th>Gene 1</th>
<th>Gene 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>250,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>24</td>
<td>500,000</td>
<td>2,000,000</td>
</tr>
<tr>
<td>25</td>
<td>1,000,000</td>
<td>4,000,000</td>
</tr>
</tbody>
</table>

David A. Palmer,
pathmicro.med.sc.edu/rtpcr/rt-pcr.ppt
HOUSEKEEPING GENE

- The total amount of mRNA in our sample can vary.
- It is possible, though very unlikely, that expression level of all genes has increased.
- We have to compare the expression of our gene of interest to a gene that is expressed at a constant level, a housekeeping gene.
AREA UNDER THE ROC CURVE (AUC)

- Receiver Operating Characteristic (ROC) is a graphical plot that illustrates the performance of a binary classifier system.
- ROC plots the fraction of true positives out of the positives (TPR) vs. the fraction of false positives out of the negatives (FPR), at various threshold settings.
- AUC is the area under the ROC curve.

$$TPR = \frac{TP}{(TP+FN)}$$
$$FPR = \frac{FP}{(FP+TN)}$$

R package "ROCR"
MATTHEW’S CORRELATION COEFFICIENT (MCC)

- Used in machine learning as a measure of quality of binary (two-class) classification.
- Has a value between -1 and +1.
- A coefficient of +1 represents a perfect prediction, 0 is no better then random prediction and -1 indicates total disagreement between prediction and observation.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

TP – true positive
TN – true negative
FP – false positive
FN – false negative
ARRAYS CAN BECOME OUTDATED

- Gene definitions change
- The reference genome sequence gets finished
- Novel splice variants are found
- Errors are made in the initial design and remain present in all arrays made
CUSTOM CDF

- Chip Definition File (CDF)
- This file maps probes into probe-sets.
- One can update those mappings:
  - Ignore deprecated or cross-hybridizing probes.
  - Merge multiple probes that recognize the same gene.
  - Account for entirely new genes that were not known at the time of array design.

CLOSING SUMMARY

- Pre-processing is very hard
  Nobody knows exactly how to do it.
  Worth careful time & consideration.

- Pay attention to statistics
  Start off with good experimental design.
  Carefully ponder all statistical tests.

- Microarray analysis Involves a pipeline
  Complete each step before moving to the next.

- Not all results make sense in the beginning
  Try to understand the initial problem and the experimental design.


“Microarrays 3" Bioinformatics Course