Microarray pipeline

Introduction to Microarray Technology course
May 2011
Solveig Mjelstad Olafsrud
solvegi@microarray.no
The microarray pipeline

The goal for this lecture: Make this figure understandable
Workflow microarray experiment

1. Problem-driven experimental design
   Wet-lab experiments
   Quality control
   RNA labelling
   Microarrays
      hybridization
      washing
      scanning
   Image analysis
      gridding
      feature extraction
   Raw data

2. Data pre-processing
   filtering
   normalization
   transformation
   missing values, ..... 
   Gene expression table

3. Secondary data analysis
   differential expression
   pattern recognition
   functional characterisation
   knowledge integration
   Biological interpretation or discovery
Experimental design (1)

• What is your **biological question**?
  - Important to make effective study design and appropriate analysis strategy!!!

• Typical types of study objectives:
  - Class **comparison** (finding genes)
  - Class **prediction** (prediction of clinical outcome)
  - Class **discovery** (gouping samples or genes with similar expression profiles)
Experimental design (2)

• Good questions to ask:
  ✓ What are you comparing? Groups, time series..
  ✓ How many samples?
  ✓ How many biological replicates?
  ✓ Do you expect big changes in gene expression?
  ✓ Budget?
  ✓ What technology to use? Commercial/in house, one/two channel

• Make experimental plan
  ✓ Avoid technical batch effects in the experiment
Wet lab

• Sampling
  ✓ Prepare tissue/cells/animals for extraction
  ✓ Proper storage important!!

• Extraction
  ✓ Use microarray compatible method
    • Column based recommended
  ✓ DNase treatment?
  ✓ Follow experimental plan!
Quality control

- RNA quality measures
  - 260/280 ratio
    - NanoDrop value +/- 2
  - 230/260 ratio
    - NanoDrop value normally 1.8-2.2
  - Bioanalyzer (Agilent Bioanalyzer 2100) profile
    - **RIN-number** (RNA Integrity Number)
      - 18S and 28S peaks (ribosomal RNA), baseline
      - Rule of thumb: RIN $\geq 7.5$
- DNA contamination?
- Globin peak?
Characteristic of intact eucaryotic total RNA

- No small, well defined peaks between ribosomal peaks
- Distinct 28S Ribosomal Subunit (usually ~2X 18S)
- Distinct 18S Ribosomal Subunit
- (5s Subunit) Prep Dependant
- Flat Baseline throughout electropherogram
Partially digested total RNA

Baseline between and to the left of the ribosomal peaks becomes jagged.

Intensities of the smaller degraded RNA increases.

The peaks begin to shift toward the left of the electropherogram.

Intensities of the peaks decrease.

In general, the 28S peak begins to degrade first.
totRNA contaminated with genomic DNA

- Genomic DNA skewing 28S peak
- Nano Peak
- Sharper 28S Peak
- Flat baseline indicates Genomic DNA digestion.
Labelling and hybridization

RNA → Reverse Transcription → cDNA-mRNA hybrid → 2nd strand synthesis → cDNA → In vitro Transcription and Labelling → Labelled cRNA → Fragmenting → Raw Data → Wash, Scan, Quantitate → Hybridisation
Scanning

- Excitation of dyes
- Emission of fluorescence
- Signal detection, amplification, digitisation
- Data storage, analysis

Photomultiplier tube (PMT)
Image analysis

Gridding / addressing:
   Add coordinates to each of the spots

Feature extraction:
   Measure foreground & background signals
   Quality measures
Data Pre-processing

• Filtering:
  ✓ Low intensity spots (signal to noise)
  ✓ Bad quality spots

• Normalization:
  ✓ Arrays needs to have a similar signal distribution to be comparable

• Missing values:
  ✓ Filtering leaves missing values,
  ✓ these has to be replaced

• Log transformation:
  ✓ To variansstable the microarray data
Tools for data analysis

J-express
Tools for data analysis

TMeV

R/Bioconductor
Validating

- Some of the genes found differentially expressed should be validated using Real-Time PCR.

- Do not only validate genes on the top of the list but some further down as well.
Many journals demand that microarray data should be stored in public repositories such as ArrayExpress or GEO.