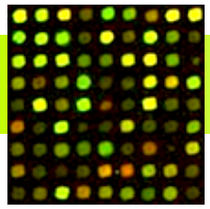
Reliability and Reproducibility Issues in DNA Microarray Measurements

Soran Draghici et al., Februar 2006

Special Topics on Bioinformatics
SS 2010

J. Palme

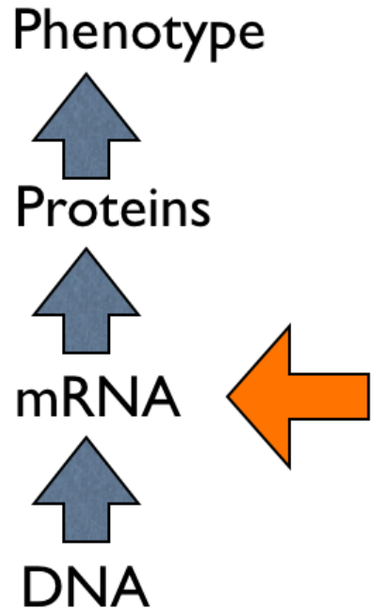


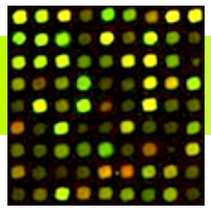
Microarrays

- Allow to monitor the expression of thousands of genes simultaneously
- Measurement of absolute levels and differential gene expression analysis
- Measurement of mRNA concentration values for single tissue or mRNA concentration differences for two tissues (measured and reference tissue)
- Search for determining factors of a specific phenotype using gene expression levels

High Level Model

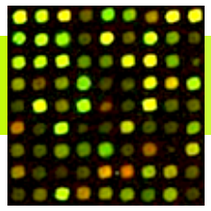
- Nucleus as controlling entity
  Cell as executing entity
- mRNA as one-way signalling from controlling to executing entity, each mRNA as separate signal
- Control signal: mRNA concentration of different mRNAs in tissue => signal strength per signal
- Capturing this control signalling gives insight into the "driving forces" for phenotype creation





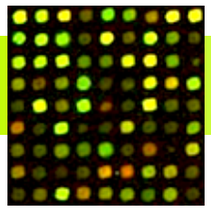
Assumption in Model I

- No interaction between signals -> clean signal separation
- Linear relation between signalling and phenotype
- Phenotype mainly driven by proteins
- Concentration of mRNA correlates to generated protein amount
- More specifically - linear relation between proteins and signals (mRNAs)



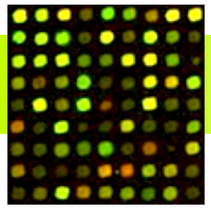
Assumption in Model II

- mRNA pretranslation processing / controls negligible (maybe except for alternate splicing)
- Posttranslation protein-preprocessing negligible
- Deterministic behavior of transcription and translation
- Identical effectiveness of proteins on phenotype creation (same concentration leads to same effect)
- Other or additional assumptions ?



Model Documentation

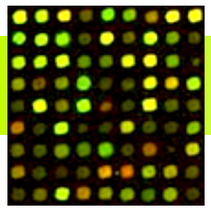
- Theoretical model of observed phenomenon
- Theoretical model of measurement system
- Shared knowledge in heads of research community
- Everybody has slightly different picture / infos
- Formal model documentation seems to be helpful
- Mutually agreed model describing “one reality”



Model Documentation

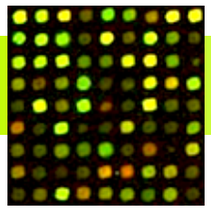
Formal model documentation helpful for

- documentation of current state of knowledge
- shared mutual understanding
- document assumptions, boundaries, exceptions ...
- give direction to model verification / falsification
- give direction to model improvements
- traceability of changes / model evolution



Array Technology

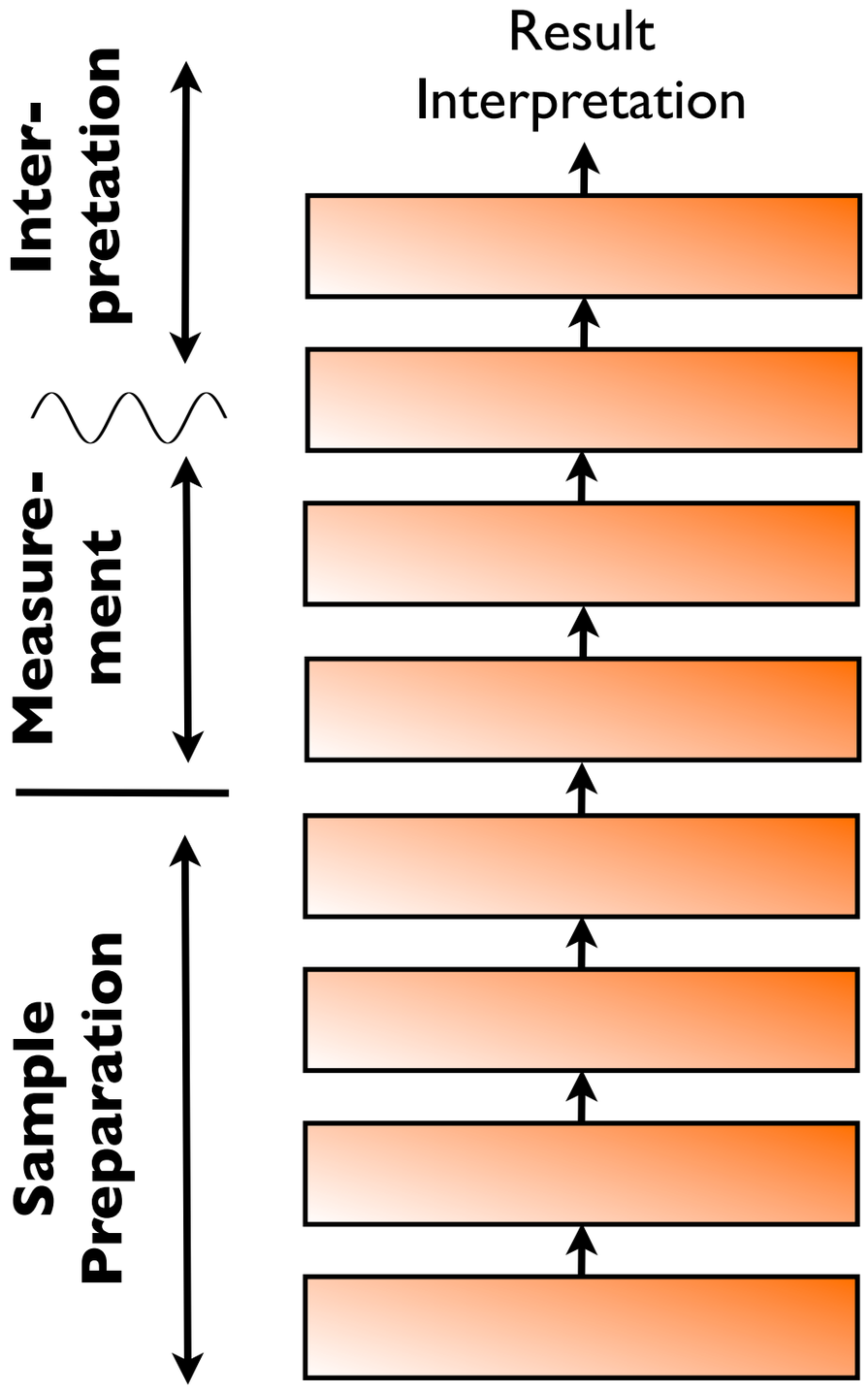
- Technology introduced in 1995
- Multiple array technologies developed: mainly cDNA array, oligonucleotide array
- Main question: measurement quality? (especially for use in medical diagnostics)
- FDA: MicroArray Quality Control Project to define quality metrics



FDA - MAQC

- Multiphase project that started in 2005
  http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/default.htm
- Phase I focus: QC guidelines, metrics, tools
- 51 organizations, 137 scientists in phase I
- 4 well defined samples prepared in one company measured at 28 locations with 7 diff. technologies
- Serious attempt to set up a process for the project detailed descriptions for parts of the process

Microarray Process



Process Phases

Data Analysis

Data Cleaning / Preprocessing

Measurement

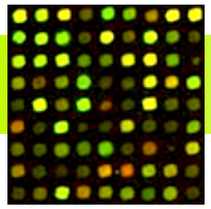
Premeasurement Treatment

Store/Transport Sample

Prepare Sample from Tissue

Store/Transport Tissue

Take Tissue Sample



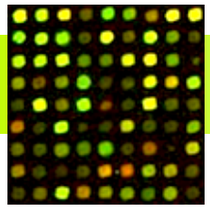
Microarray Process

Process Parts

- defined phases with clear boundaries
- defined activities per phase
- defined input and output per phase
- quality control in each phase
- common and variant parts of process

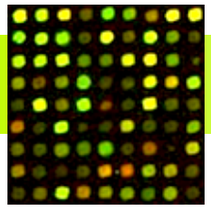
Process Variation

- sample specific / domain spec. parts
- sample preparation specific parts
- parts related to specific storage / transport needs
- technology related parts
- platform / vendor specific parts



Diagnostic Aspects

- Medical use for reliable diagnosis of diseases on molecular level
- Expression levels are very precise indicators of modification of genetic activity - e.g. in case of cancer or genetic disorders
- Creation of new disease classifiers based on gene expression levels
- Requires reliable measurement of expression levels - both absolute and differential

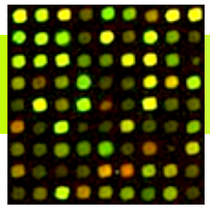


Microarray Quality

- Accuracy
  conformity of measured value with the actual value
  (mRNA concentration in tissue)

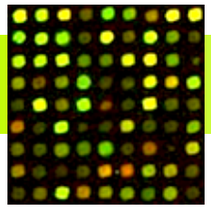
- Precision (Reproducibility)
  degree of identical results for repeated measure-
  ments with identical conditions

- Specificity
  ability of probe to bind unique target sequence
  => value represents single mRNA concentration

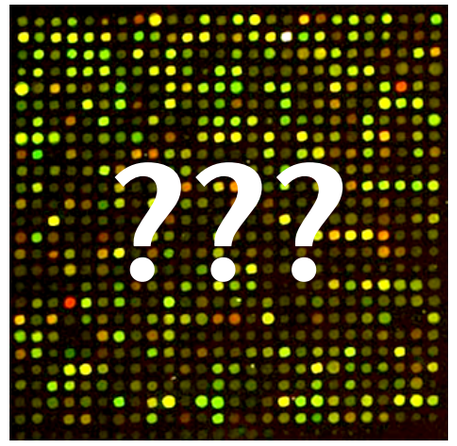


Measurement Types

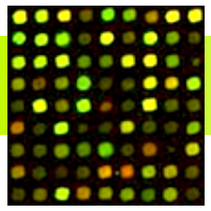
- Absolute value of mRNA concentration
- Differential expression analysis
  comparison of sample with reference sample
  only value difference between the two samples
  errors even out - as long as they are identical in both cases
- often direction of change more important than absolute value
- differential measurem. more robust against noise



However

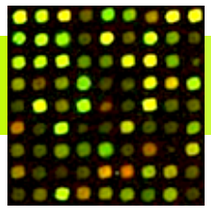


- Research in recent years has shown many effects that impact measurement quality
- Platform specific effects
- Sample specific effects
- Sample preparation specific effects
- “Microarray - the unknown creature?”



Accuracy

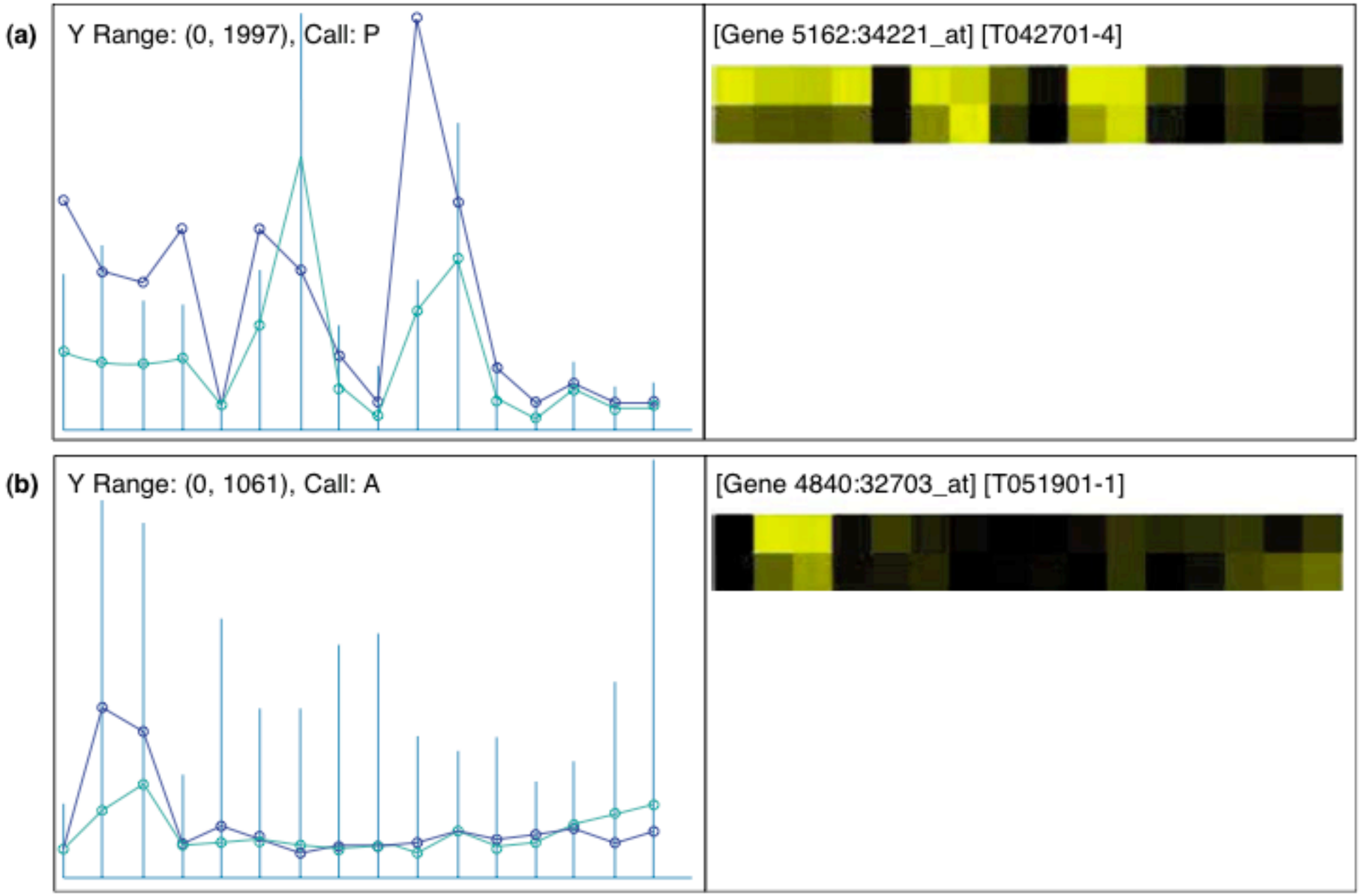
- Differential measurement requires linear relation between signal intensity and concentration of specific transcript
- Absolute values not important - as long as the same linear relation applies for both compared tissue samples
- Reliable absolute values require careful calibration with known transcripts/mixtures

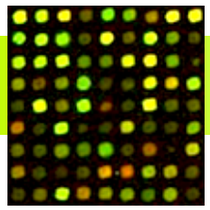


Accuracy-Problems

- Often different values for probes in probe-set - sometimes by orders of magnitude
- Three possibilities for binding variation:
  - variation through seq. dep. binding affinity
  - variation through different probe quality
  - cross-hybridization from other mRNAs
- Indicates, that measured values often do not directly correspond to concentrations!

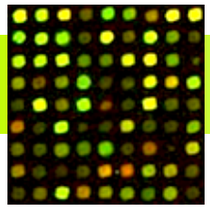
Probe Dependency





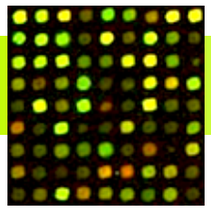
Spike-In Datasets

- A few available from Affymetrix / partners
- Contain only a small amount of genes (42 genes at most!) - unbiased selection?
- Show much smaller intensity variation than real-life data with strong intensity/sequence dependency
- Extrapolation from 42 to 10000 genes?



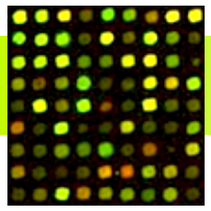
Spike-In Datasets

- Also provided by academic labs
- Drosophila RNA samples with ~1300 spiked-in genes (background of 2500 genes)
- Significant agreement between observed and actual fold changes (R^2 = 0,86)
- Probesets in the lowest quartile were filtered out



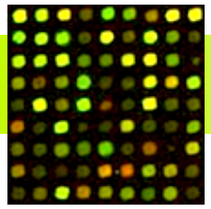
Verification via RT-PCR

- High costs => only limited no of transcripts typically < 20
- Usually verification for genes with well agreed upon sequences
- in 85-90 % verifiable differential expression results on widely used platforms
- BUT only for stronger expressed genes



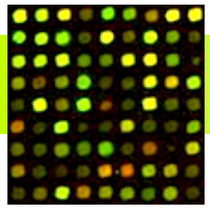
Verification via RT-PCR

- typically 40 - 50 % of transcripts in a RNA sample are not in the sensitivity range of the platform
- this part of expression information is lost
- sometimes the correct fold changes of highly relevant genes (e.g. epidermal growth factor receptor - EGFR) for cancer diagnostics are not recognized



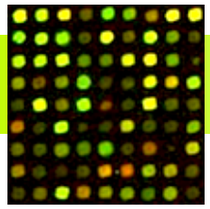
RT-PCR Studies Results

- Above sensitivity threshold in ~ 70-90 % of genes direction of expression changes o.k.
- Magnitude of changes different between microarray and values from RT-PCR
- Both dual & single channel arrays measure ratios more accurately then absolute levels
- M.Array expression ratios are compressed



Precision

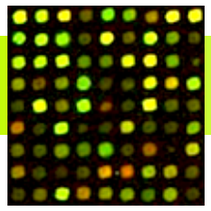
- Reproducable results for defined conditions
- Independence from laboratory personal
- Cross platform consistency
- Reproducable results do not necessarily mean accurate results - effects can be reproduceable as well
  =>Reproducable+cross-platform consistent



Reproducability

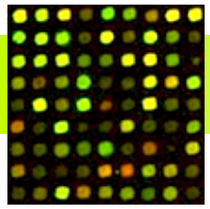
- Reproducability within sensitivity range
- Oligonucleotide Arrays (Affy, Agilent, Codelink): correlation coefficient > 0,9
- cDNA arrays or Mergen platform correlation coefficient: 0,5 up to 0,95

  => impact to cross platform consistency



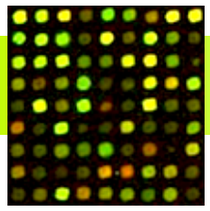
Cross-PLF Consistency

- What do we wish:
  - highly consistent results across platforms
  - identical and correct absolute values
- Platform dependency would be removed
- No replicated experiments with different platforms / technologies
- Basis for universal gene-expression DBs



Cross-PLF consistency

- Consistency still no proof of accuracy
- Lack of consistency could be produced by single platform - unclear which one is bad
- Cross-hybridization could be consistent between platforms
- Necessary but insufficient condition!

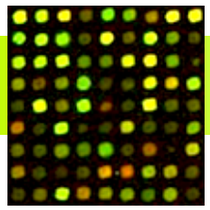


Cross-PLF Results

- Data available for NCI60 cell line
- profiled on cDNA- and Affymetrix array
- reanalyzed multiple times with improving results
- result highly dependent on probe matching strategy between platforms!

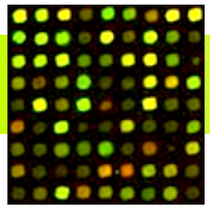
Cross-PLF Results

- Initially no sequence info released => probe matching based on Unigene ID (produces significant no of incorr. pairings
- Pearson correlation < 0.34
- Less cross PLF correlation for probes with greater no of cross matches to other genes
- Low correlation for low intensity genes



Cross-PLF Results

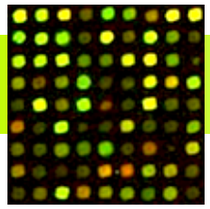
- later probe sequence data published
  => better matching strategies
- filtering out incorrect sequences matching to other transcripts
  => for same NCI60 data: Pears.corr. ~ 0.6
- Using only probes targeting same region of transcript, NCI60 data: Pearson corr. ~ 0.7



Cross-PLF Results

One comparison from 2003:

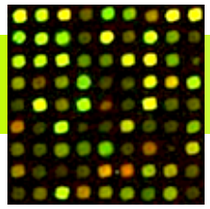
- Highest correlation: 0.59 between oligonucleotide arrays (Affy and Codelink)
  Correlation: 0.49 between oligo- and cDNA arrays
- Measurement without sequence verification and filtering of low expression levels
- Very different genes differentially expressed on differnt PLFs but gene ontology mapping shows consistent biological processes for all PLFs !



Cross-PLF Results

Further Results (from different studies)

- Oligonucleotid arrays - higher correlation (>0.7) than cDNA arrays and Mergen platform (≤0.5)
- In studies of cross-PLF reproducability: cross-PLF consistency between 0.11 and 0.76
- Affymetrix produced highest correlation when same PLF was used by different laboratories: 0.91

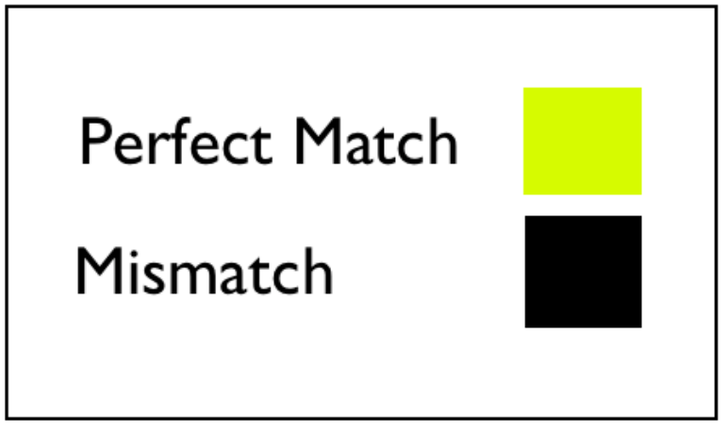


Cross-PLF Results

High correlation results only with:

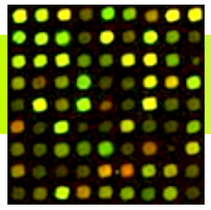
- Strict sequence matching strategy
- Check of sequence data against high quality sequence DBs - throw out uncertain seq.
- Filtering out of low intensity genes
- Filter/analyse data based on technological background - not only raw data into DB

Specificity



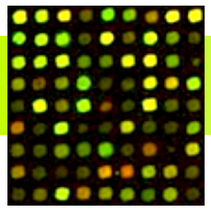
- Wrong probe sequences
- Binding affinity varies with sequence,
  e.g mismatch with higher affinity as target seq.

- alternative splicing - probe can target all variants or specific splice variant => contributes to cross-PLF discrepancies
- Folding of target transcripts impacts binding
- Cross-hybridization



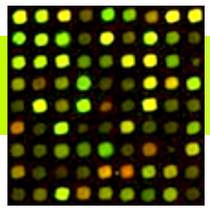
Problem Areas

- Probe design and technological inaccuracies
- Homogeneity of hybridization process and lack of understanding of hybridization kinetics
- Cross-hybridization signal from transcripts with sufficient similarity
- Alternate splicing impacts
- Nucleotide insertion during labelling



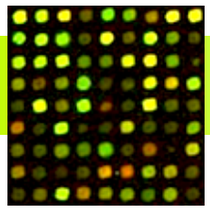
Recommendations

- Calibrate whenever possible
- Use most recent annotation from vendor
- Verify sequences against high quality DBs
- Remove erroneous sequences before probe data mixing
- Filter out low level expression genes



SUMMARY I

- First decade of microarray technology produced rather limited data
- Current microarrays suited for differential expression analysis
- Absolute values and detection of low abundance genes currently beyond reach
- Consider best practices



SUMMARY II

- With high quality a big step in diagnostics
- FDA: Micro Array Quality Control Project was a first step in the right direction!
- Further steps needed - e.g. cross-validation with NGS and better defined process
- Attention: Microarray -> FRAGILE
  HANDLE WITH CARE