Reliability and Reproducibility in Microarray measurements

By Jake Sauter

- Introduction
- Sensitivity
- Accuracy
- Reproducibility
- Cross - Platform Consistency
- MAQC Project

Microarray Results

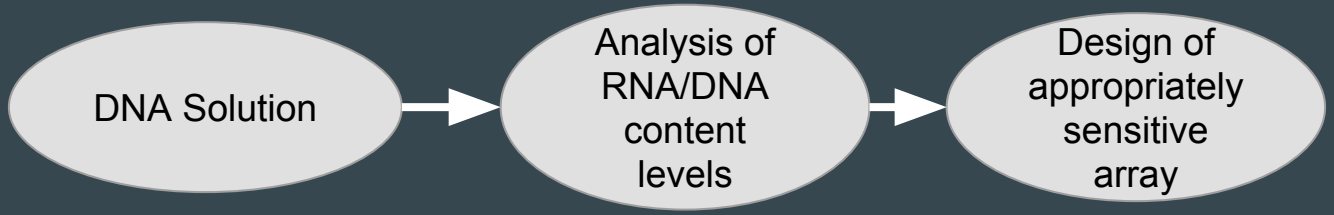
- Results are still marred by several technical issues that are often neglected
- Each manufacturing technique has its own weaknesses
- Different arrays have different results concerning accuracy, sensitivity, specificity and robustness
    - Claims have been made of reduced reproducibility of results from one platform to another
- These issues became much more important when microarrays were proposed as a diagnostic tool in molecular disease classification
    - Regulatory agencies such as the FDA require solid, empirically supported data about the accuracy, sensitivity, specificity, reproducibility and reliability of microarrays for clinical use

Fabrication Issues

- Nucleotide probes synthesized on the chip are not 100% accurate due to base-skipping
- Deposition based techniques can lead to incorrect cDNA intermixed on the probe
- Analysis is based on the assumption that most microarray probes produce specific signals under a single, rather lenient hybridization process
    - Probably not true as testified by widespread cross-hybridization of transcripts on microarrays

Microarray Sensitivity

- Determining the sensitivity of microarray measurements is essential in order to define the concentration range in which accurate measurements can be made



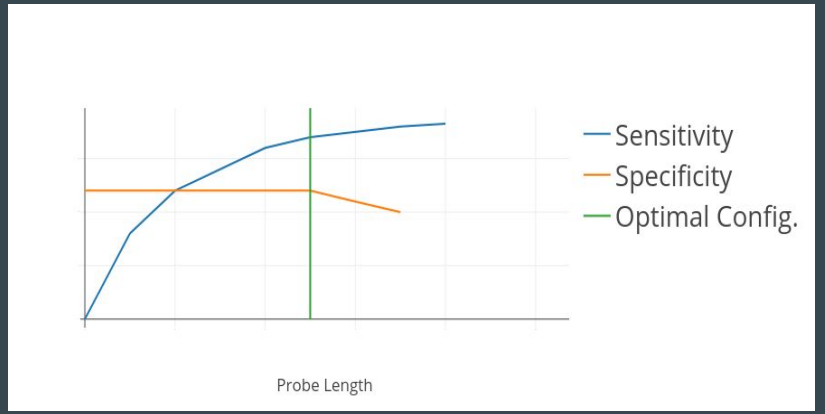
- Results from using microarrays with suggested solution concentrations were verified for many manufacturers, however any lower than the recommended concentrations caused non-meaningful measurements

Sensitivity Issues

- 40-50% of present transcripts from RNA samples are estimated to be below sensitivity threshold levels
- Failure to detect a highly relevant gene EGFR (important for cancer diagnosis) encourages reflection of using microarrays in the diagnosis pipeline

Probe Length and Detection Limits

- Probe length can provide trade-offs in sensitivity and specificity
- Sensitivity increases with probe length, however after ~30-mer probes specificity starts to decline for less of an increase in sensitivity
- Detection limit of current microarray technology appears to be between 1 and 10 copies of mRNA per cell.
    - Might still be insufficient for detection of relevant changes in low abundance genes (e.g. transcription factors that control cell replication)



Microarray Accuracy

- Microarrays can measure
  - Absolute transcript concentration
  - Relative transcript concentration (expression ratios, comparing two samples)
- Expression ratios can be measured with a high accuracy and are usually favored
- Probes in a given Affymetrix probe set (probes to detect the same gene transcript) while producing significantly different intensities may still produce consistent ratio values across the very same probe set when two RNA samples are compared

Validation of Accuracy

- Assessing the accuracy of microarray measurements requires that the true concentrations or concentration ratios are known for a large amount of transcripts
- True concentrations can be obtained by
  - Spike-in or dilution experiments (small amounts of genes per test)
  - Independent means such as quantitative RT-PCR or Northern Blots (costly process so done for a small number of genes)
- At most 42 genes can be assessed per spike-in study, much less than the 10-30k possibly informative genes
- Usually only 3-10 genes are usually verified using independent quantitative techniques per study

Key Points for Accuracy of Microarrays

1. In their appropriate dynamic range, microarray measurements accurately reflect the existence and direction of expression changes in approximately 70-90% of genes.
2. Microarrays measure expression ratios more accurately than absolute expression levels
    a. High correlation in expression ratio measurements has been seen with more robust experimental methods
3. This relatively good correlation is not perfect due to compressed (underestimated) microarray expression ratios

Reproducibility

- A platform can have excellent reproducibility without producing any accurate or cross-platform consistent measurements
- "Good" reproducibility requires that a given probe bind the same number of labelled transcripts in repeated measurements of the same sample
- Badly designed probes that perhaps cross-hybridize with a number of other transcripts besides the target transcript can easily provide highly reproducible yet useless data

Cross-Platform Consistency

- If cross-platform consistency and reproducibility were high, one could use appropriately normalized data regardless of the platform it was collected on
  - This would also reduce the need to replicate experiments and allow researcher to build a universal gene expression database
- Cross-platform consistency has become a top required characteristic of most platforms for reliability
  - Though this is not a sound technique as it is required but not sufficient to validate microarray technology
- Incorrect probe matching can contribute to low cross-platform consistency
  - Usually during analysis there are only 1-2 thousand highly cross-platform consistent genes
  - Up to 50% of contradictory cDNA to oligo arrays can be explained by incorrect clones on cDNA probe

Causes of Inaccuracy and Inconsistencies in Microarrays

- The transcript hybridization signal is composed of 3 signals that can be difficult to isolate
  - Specific signal from target
  - cross-hybridization
  - background (present when absence of sequence similarity)
- The relationships between probe sequences, target concentration and probe intensity is poorly understood
  - Some probes require more energy to bind then others, meaning less target sequences would bind than other target in the solution even if there is the same quantity
- Splice variants are problematic
  - More than 50% of human genes are alternatively spliced
  - Probes must be designed to bind to all variants of a desired target

Microarray Quality Control Project (MAQC)

- FDA spawned project initiated September 2006 with the goals of
  - Provide quality control, tools to the microarray community to avoid procedural failures
  - Develop guidelines for microarray data analysis by providing the public with larger reference data sets along with readily accessible reference RNA samples
  - Establish QC metrics and thresholds for objectively assessing the performance achievable by various microarray platforms
  - Evaluate the advantages and disadvantages of various data analysis methods
- Goals of this project update periodically once the direction wished has been sufficiently explored

References

[1] Drăghici Sorin. Statistics and Data Analysis for Microarrays: Using R and Bioconductor. Chapman and Hall, 2012.