WEEK 4 ASSIGNMENT

Data Systems in the Life Sciences – BIOL 51000 | Fall 2 2020

Christina Morgenstern

MICROARRAY DATA

1. Why microarrays are considered multiplexed experiments

In general, performing many experiments of the same type in parallel on a small scale are considered to be multiplexed. So are microarrays that perform hundreds to millions of individual experiments with the same reagents on a chip-based substrate.
These parallel experiments of high-throughput character can be performed to detect potential drug compounds, RNA molecules, antibodies, protein crystal structures as well as changes in gene expression amongst others. The parallelization of these experiments has the advantage of making the individual experiment better comparable as well as help with normalizing and organizing the data.

2. How to adjust and normalize microarray data

When dealing with gene expression microarrays, we are assuming that each measured intensity corresponds to the relative expression level of each gene. Comparing the expression levels between two states i.e., healthy and disease, we are looking for relevant patterns differentiating the two conditions. Before comparing the different levels, the data needs to be processed in order to eliminate low-quality measurements, to adjust measurements for better comparison and to select genes that are significantly differentially expressed between samples [1].
When comparing genes, the $\log_2$ ratios between the red and the green colors are taken as expression levels. Normalization is the first transformation applied to the expression data and adjusts the individual intensities to balance them and to make them comparable with each other. The data needs to be normalized because of unequal quantities in starting material, differences in labeling or detection efficiencies between dyes used and systemic biases in measured expression levels [1].
Depending on the experimental design and the biological question, a number of normalization methods can be used: Standard deviation, mean, centered mean, Z-score and the maximum value can all be applied to each value in the matrix or certain (rows or columns) array data. Normalization can also be performed against an internal control value a so-called reference. The intensity of the fluorescent signal can also be used for normalization by taking the log of the mean. Lastly, quantiles can also be used to normalize array data albeit this is often considered a less scientific way.

3. Compare and contrast microarrays and RNA sequencing - how to choose which to use?

The technology of microarrays was developed in 1999 and evolved from previous gene expression techniques like Southern and Northern blotting. Microarrays were one of the first multiplexed methods that allowed the analysis of millions of individual analyses in one experiment. The steps involved in a traditional gene expression microarray experiment are as follows: RNA is isolated from control and test cells and reverse transcribed into cDNA. The samples are further labeled with fluorescent dyes, green (Cy 3) and red (Cy 5), for control and test sample, respectively. The two labeled samples are then mixed together and hybridized on to the microarray that has an array of DNA samples spotted. Using a laser, the fluorescence of the bound quantities is detected and measured.

RNA sequencing (RNA-Seq) evolved in 2008 with the rise of next-generation sequencing technologies (NGS). In order to determine gene expression with RNA-Seq, total RNA is extracted from the samples of interest. The RNA is reverse transcribed to yield cDNA. After the attachment of linkers, the cDNA libraries are sequenced generating 100s of millions of reads. The reads are further either mapped to a reference genome or assembled to generate expression matrices with gene/transcript counts.

When comparing microarrays with RNA-Seq, we can see that the principle of the technology is quite different. While microarrays use hybridization, RNA-seq uses cloning and sequencing steps. The amount of RNA needed, is for microarray experiments substantially higher than for RNA-Seq experiments. As for resolution, microarrays are able to provide information on several and up to 100 bp whereas RNA-Seq is capable of single base resolution. Differences in allelic expression is limited in microarrays while RNA-Seq is able to pick up differences in alleles. Novel genes can be uncovered using RNA-Seq while microarrays are dependent on previous knowledge [2].

Advantages of microarrays are the well-defined protocols for hybridization and analysis pipelines as well as the standardized approaches for data submission and the relatively low cost of materials and equipment. Disadvantages of the microarray technology are the analysis of only pre-defined sequences, the reliance on hybridization which might not be specific, the high variance for low expressed genes and the inability of identifying splice variants [3].

Advantages of RNA-Seq are the fact that no prior knowledge on sequence information is needed, the ability to detect alternative splicing if aligned to the genome, the possibility to detect SNPs and paralogous genes. The high cost attributes for a disadvantage of RNA-Seq as well as the need for dedicated computing facilities and the complexity of the analysis [3].

When cost is an issue, microarrays (currently) almost always outperform RNA-Seq experiments. However, the benefits of RNA-Seq can outweigh the cost issues.

Depending on the experimental question, it might still make sense to use a microarray over an RNA-Seq experiment for example in diagnostic tests with proven clinical utility [4].

4. Discuss what types of experimental questions can be answered using microarrays

Microarrays or arrays in general can be used to answer a variety of biological questions. In fact everything that involves some kind of interaction between DNA, proteins or peptides, lipids, carbohydrates or small molecules can be addressed using the platform of arrays.

Traditionally, microarrays evolved to analyze the expression of thousands of genes simultaneously. These types of arrays are called Gene expression profiling arrays. Comparative Genomic Hybridization arrays are used to assess the genome content in different cells or related organisms. The identification of single base differences, single nucleotide polymorphisms (SNPs) is possible with SNP arrays. ChIP-on-chip (Chromatin Immunoprecipitation) arrays aide in determining protein binding sites within a genome. Epigenetic markers like methylated DNA can be assessed using methylation arrays.

Researchers can address questions like the amount of mRNA expressed by a gene by using gene expression, exon or tiling arrays. The latter two microarray technologies are also used when determining the amount of mRNA expressed by a specific exon. Exon and tiling arrays can further evaluate which strand of DNA is expressed.

Protein arrays can analyze a variety of interactions: screening for antibodies that bind certain proteins, performing enzymatic assays, determine protein-DNA interaction or protein-small molecule, protein-lipid or protein-protein interactions.

References:

[1] J. Quackenbush, ‘Microarray data normalization and transformation’, *Nature Genetics*, vol. 32, Dec. 2002, doi: doi:10.1038/ng1032.

[2] S. Chavan, M. Bauer, E. Peterson, C. Heuck, and D. Jr, ‘Towards the integration, annotation and association of historical microarray experiments with RNA-seq’, *BMC Bioinformatics*, vol. 14, Oct. 2013, doi: 10.1186/1471-2105-14-S14-S4.

[3] S. Martin, C. Dehler, and E. Krol, ‘Transcriptomic responses in the fish intestine’, *Developmental & Comparative Immunology*, vol. 64, Mar. 2016, doi: 10.1016/j.dci.2016.03.014.

[4] ‘RNA-sequencing vs microarrays’, *Genevia Technologies*. https://geneviatechnologies.com/blog/rnasequencing/ (accessed Nov. 21, 2020).