QLUCORE

RNAseq or microarrays – which one should I choose?

INTRODUCTION

For the past two decades, microarrays have been used extensively for high-throughput quantification of mRNA abundance. In the last few years, transcriptomic profiling via next generation sequencing (RNAseq) has emerged as a powerful competitor. In this document we briefly discuss and compare the characteristics of the two methods.

THE TECHNIQUES

Modern microarrays typically consist of a large number of short oligonucleotide probes, representing genomic regions of interest, attached to a chip. Transcripts are fluorescently labeled and allowed to hybridize to the probes, and the signal intensity from each probe is used as a measure of RNA abundance. RNAseq, in contrast, uses next generation sequencing methods to directly determine the nucleotide sequence of millions of short pieces of RNA (called reads). Typically, the short reads are then mapped to a reference genome and the number of reads that map within a given region is used as a measure of the abundance of that region. If no reference genome exists for the studied organism, it is possible to assemble it using the sequenced reads.

PROS AND CONS

One of the most prominent advantages of RNAseq compared to array-based techniques is that RNAseq can be applied without extensive knowledge of the genomic sequence and the location of genes or other features of interest. This makes it possible, for example, to detect previously unknown transcripts, isoforms and splice junctions and also means that RNAseq can be used for organisms where no microarray or reference genome is available [1, 2]. Arrays, in contrast, rely on hybridization to pre-defined probes and allow detection only of the features encoded in these probes. Different types of arrays, such as gene expression arrays, exon arrays and splicing arrays, have therefore been developed to address different questions. Most similar to RNAseq are the tiling arrays, where the probes are constructed to overlap each other and are distributed along the entire genome. Comparisons between exon arrays and RNAseq have shown that RNAseq is more precise in estimating exon boundaries, which has been attributed to the higher resolution provided by the RNAseq technique [1].

The reliance on hybridization makes arrays susceptible to cross-hybridization, that is, hybridization of RNA that is similar but not identical to the probe target. This can have a measurable effect on expression levels particularly for genes with low expression [1]. The use of hybridization also imposes a limitation on the dynamical range of expression levels, and a saturation effect can be seen for highly expressed probes. RNAseq does not suffer from the

COPYRIGHT 2015 QLUCORE AB

cross-hybridization problem, and has a considerably larger dynamical range than microarrays, both in terms of expression levels and in terms of fold changes [1, 3]. However, RNAseq is based on sampling (i.e the reads to be sequenced are sampled from the pool of available short segments) and for weakly expressed features the sampling variation dominates the biological variation, and weakly expressed features may not be detected at all. Except for the very highly or weakly expressed features, several studies have reported a high correlation between expression measurements from RNAseq and different types of microarrays, and there is usually a large overlap between differentially expressed genes found by using the two techniques [1, 2, 3].

One of the biggest hurdles for RNAseq to overcome is still the higher cost compared to microarrays, even though the gap is narrowing rapidly. The higher cost of RNAseq may lead researchers to reduce the number of biological replicates, which would then make it harder to perform reliable statistical analyses. Another advantage, stemming from the extensive use of microarrays over many years, is that their biases are well known and understood, and that there are well developed analysis pipelines [2]. These aspects are also currently being intensely studied for RNAseq. Finally, there is a considerable difference in the amount of data generated from the two technologies. While the data files obtained from microarray analyses are typically some tens of MB, the sequence files from RNAseq experiments are usually several GB, which drastically increases the need for storage space and computational resources [2].

ANALYSIS TECHNIQUES

Microarrays have been used extensively for transcription profiling for many years, and the computational techniques used to find differentially expressed genes are well developed. Since the (log-transformed) microarray signal intensity can be assumed to follow a normal distribution, the most common approaches are regular or moderated t-tests or more general methods based on linear models. The number of methods designed for differential expression analysis of RNAseq data increases rapidly, and many of them are explicitly modeling the observed read counts using, e.g., Poisson or Negative Binomial distributions. However, recent publications have shown that methods that apply a data transformation, followed by typical "microarray" differential expression analysis methods, perform well in many situations and seem to be more resistant to outliers [4].

Both microarray data and RNAseq data can be analyzed with Qlucore Omics Explorer (OE). Microarray data obtained by Affymetrix or Agilent arrays can be automatically normalized when imported into OE. For RNAseq data (gene expression) aligned BAM files can be directly imported.

TO READ MORE

[1] Agarwal et al., Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics 11:383 (2010)

[2] Malone and Oliver, Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biology 9:34 (2011)

COPYRIGHT 2015 QLUCORE AB

[3] Raghavachari et al., A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. BMC Medical Genomics 5:28 (2012)

[4] Soneson and Delorenzi, A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91 (2013)

DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.

Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.

Qlucore Omics Explorer is only intended for research purposes.

COPYRIGHT 2015 QLUCORE AB

COPYRIGHT 2015 QLUCORE AB