Measures of Gene Expression for Affymetrix High Density Oligonucleotide Arrays

Rafael A. Irizarry

Abstract

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. In Affymetrix GeneChip array technology, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, separately referred to as probes, arrayed on a silicon chip. After chip measurements are preprocessed, a fluorescence intensity value for each probe is obtained. A necessary step for defining a measure of expression (*ME*) is to summarize the probe intensities for a given gene. In this paper, we review the ideas that motivate a summary statistic, referred to as the robust multi-array average (*RMA*), that improves the default Affymetrix approach and provides substantial benefits to users of the GeneChip technology.

Keywords: Affymetrix GeneChip arrays; background correction; gene expression; normalization; summary measure

1 Introduction

High density oligonucleotide expression array technology is widely used in many areas of biomedical research for quantitative and highly parallel measurements of gene expression. Affymetrix GeneChip arrays use oligonucleotides of length 25 base pairs to probe genes. In this technology, each gene is typically represented by a set of 11-20 pairs of oligonucleotides, separately referred to as *probes*, arrayed on a silicon chip. Details of this array technology are described by [1] and [10]. Briefly, though, RNA samples are prepared according to a specific protocol. A fluorescently labeled RNA sample is hybridized to probes on the chip. After some processing steps, the array is scanned with a laser. This scan produces an image that is analyzed to produce an intensity value for each probe (see [9] for more details). These intensities quantify the extent of the hybridization between the labeled target sample and the oligonucleotide probe. A final step to obtain a measure of gene expression (*ME*) is to summarize the intensities for a given gene in order to quantify the amount of corresponding mRNA species in the sample. The intensities obtained for each probe are denoted by PM_{ijn} and MM_{ijn}, $i = 1,\ldots,I$, $j = 1,\ldots,J_n$, and $n = 1,\ldots,N$, with i representing different RNA samples, j representing the probe pair number (this number is related to the physical position of