puma: a Bioconductor package for Propagating Uncertainty in Microarray Analysis

The analysis of microarray experiments typically involves a number of stages. The first stage for analysis of Affymetrix GeneChip arrays is usually the application of a summarisation method such as MAS5.0 or RMA in order to obtain an expression level for each probeset on each array. Subsequent analyses then use these expression levels, for example to determine differentially expressed (DE) genes, or to find clusters of genes and/or conditions. Although there are a number of summarisation methods which can give accurate point estimates of expression levels, few can also provide any information about uncertainty in expression levels (such as standard errors). Even for methods than can provide uncertainty information, this is rarely used in subsequent analyses due to the lack of available methods for dealing with such information. A large amount of potentially valuable information is therefore lost.

The multi-mgMOS method of Liu *et al.* (2005) uses Bayesian methods to associate credibility intervals with expression levels. Sanguinetti *et al.* (2005) describe the noise-propagation in principal components analysis (NPPCA) method which can propagate the expression level uncertainty to improve the results of PCA. The Probability of Positive Log Ratio (PPLR) method of Liu *et al.* (2006) can combine uncertainty information from replicated experiments in order to obtain point estimates and standard errors of the expression levels within each condition. These point estimates and standard errors can then be used to obtain a PPLR score for each probeset, which can then be used to rank probesets by probability of differential expression between two conditions. The PUMA-CLUST method of Liu *et al.* (2007) uses uncertainty propagation to improve results of a typical clustering analysis.

The *puma* package combines the various methods described in the previous paragraph in a single, easy-to-use package, and overcomes some of the shortcomings of these methods. *puma* offers the following contributions:

- pumaDE - an extension of the PPLR method to the multi-factorial case
- The automated creation of design and contrast matrices for typical experimental designs
- pumaComb - an implementation of the method of combining information from replicates from Liu *et al.* (2006) that is significantly speeded up through the use of parallel processing
- pumaPCA - an R implementation of the NPPCA algorithm, with much improved execution speed over the previous matlab version.
- Bringing together for the first time in a single package a suite of algorithms for propagating uncertainty in microarray analysis, together with tools for plotting, data manipulation, and comparison to other methods
- Demonstration of uncertainty propagation methods on Affymetrix GeneChip and Illumina BeadChip data.

Liu,X., Milo,M., Lawrence,N.D. and Rattray,M. (2005) A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. Bioinformatics, 21:3637-3644.

Sanguinetti,G., Milo,M., Rattray,M. and Lawrence, N.D. (2005) Accounting for probe-level noise in principal component analysis of microarray data. Bioinformatics, 21:3748-3754.

Liu,X., Milo,M., Lawrence,N.D. and Rattray,M. (2006) Probe-level measurement error improves accuracy in detecting differential gene expression. Bioinformatics, 22:2107-2113.

Liu,X., Lin,K.K., Andersen,B., and Rattray,M. (2007) Including probe-level uncertainty in model-based gene expression clustering. BMC Bioinformatics, 8(98).