Power and minimal sample size for multivariate analysis of microarrays

Maarten van Iterson[1,*], **José A. Ferreira**[2], **Judith M. Boer**[1,3,4] **and Renée X. Menezes**[5]

1. Center for Human and Clinical Genetics, LUMC, Leiden
2. RIVM, Bilthoven
3. Department of Pediatrics Oncology and Hematology, Erasmus MC Sophia Childrens Hospital, Rotterdam
4. Netherlands Bioinformatics Center
5. Department of Epidemiology and Biostatistics, VUmc, Amsterdam.

*Contact author: M.van_iterson.HG@lumc.nl

Keywords: average power, sample size determination, high-throughput data, kernel deconvolution estimator, moderated t test statistics.

Choosing the appropriate sample size for high-throughput experiments, such as those involving microarrays and next-generation sequencing is complicated. Traditional univariate sample size determinations relate power and significance level to sample size, effect size and sample variability. However, for high-dimensional data these quantities need to be redefined: average power instead of power, significance level needs to take multiple testing into account, and both effect sizes and variances have many values.

Some authors (see Ferreira and Zwinderman, 2006 and Dobbin and Simon, 2005) have proposed such methods for two-group comparisons of high-dimensional data. The most general of those, by Ferreira and Zwinderman, uses the entire set of test statistics from pilot data to estimate the effect size distribution, power and minimal sample size, as opposed to most other published methods that either fix an effect size of interest, or assume (partial) homoscedasticity. Ferreira and Zwinderman assume that the test statistics follow a normal distribution, which is unlikely to hold in practice as many comparisons involve a Student-t test statistic, and the performance of the method in such cases was not evaluated.

We aimed at a generalization of power and sample size estimation more applicable to high-throughput genomics data. First, we extended Ferreira and Zwindermans method to the case of a Student-t test. Second, we considered t-test statistics generated by testing if a coefficient of a general linear model is equal to zero. Furthermore, we considered Student-t tests that use a shrunken variance estimator, such as those produced by empirical Bayes linear models as implemented in the BioConductor package **limma** (Smyth, 2005). These extensions represent a considerable improvement on the power and sample size estimation compared to when the normal assumption is used, which we illustrate via a simulation study. Finally, we will extend the method to generalized linear models aimed at power and sample size estimation for RNA-seq data. The extensions will be implemented as part of our BioConductor package **SSPA** (van Iterson *et al.*, 2009), forming a valuable tool for experimental design of microarray experiments.

References

Dobbin K, Simon R. (2005). Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*. 6(1):27-38.

Ferreira JA, Zwinderman A. (2006). Approximate sample size calculations with microarray data: an illustration. *Statistical Applications in Genetics and Molecular Biology*. 5(25).

Smyth GK (2005). Limma: linear models for microarray data, Bioinformatics and Computational Biology Solutions using R and Bioconductor, R. Gentleman and V. Carey and S. Dudoit and R. Irizarry and W. Huber, Springer, New York, 397–420.

van Iterson *et al.* (2009). van Iterson M, 't Hoen PA, Pedotti P, Hooiveld GJ, den Dunnen JT, van Ommen GJ, Boer JM, Menezes RX. Relative power and sample size analysis on gene expression profiling data. *BMC Genomics*. 10:439.