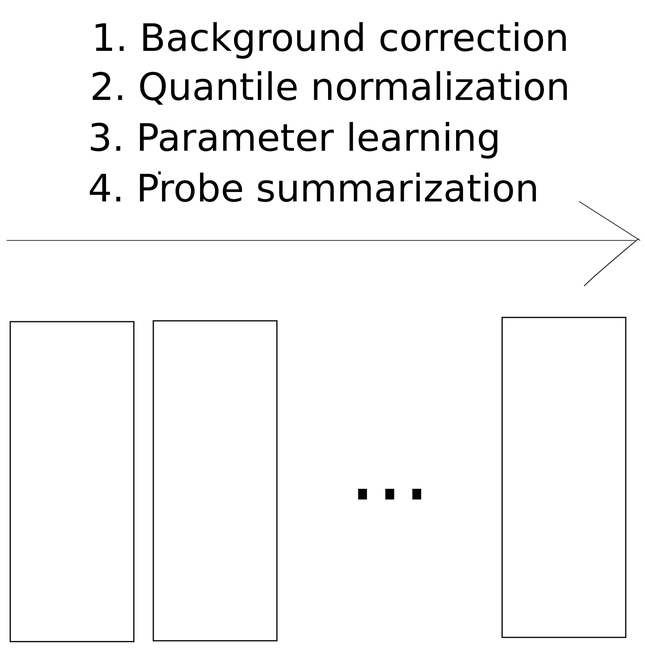
Fully scalable online-preprocessing approach for large-scale gene expression atlases

Leo Lahti (1,2), Aurora Torrente (3), Laura L. Elo (4), Johan Rung (3)
(1) Wageningen University, Laboratory of Microbiology, Netherlands (2) University of Helsinki, Department of Veterinary Bioscience, Finland (3) European Bioinformatics Institute EBI, Hinxton, UK (4) University of Turku, Department of Mathematics, Finland

Standardization of measurement platforms and accumulation of research data in public repositories has opened up novel opportunities for genome- and organism-wide investigations of gene activity via integrative analysis of large-scale data collections. Short oligonucleotide microarrays present a major source of genome-wide profiling data, but the limited scalability of the current preprocessing techniques has formed a bottleneck for comprehensive meta-analyses of contemporary microarray data collections. Scalable multi-array preprocessing techniques have been available only for particular microarray platforms based on pre-calculated probe effect terms estimated from restricted reference training sets.

We introduce a fully scalable online-learning algorithm that overcomes these limitations by providing the tools to extract probe-level information across large collections of short oligonucleotide microarray data based on sequential hyperparameter updates, allowing the preprocessing of large-scale microarray atlases in small consecutive batches in linear time with respect to sample size. In contrast to alternatives, the method is fully scalable and readily applicable to all measurement platforms. Incorporating prior information of the probes in the analysis can be used to further improve preprocessing performance, and estimates of probe affinity and variance terms based on the most comprehensive collections of short oligonucleotide array data can potentially guide array design and quality control. This is opening up novel opportunities to take full advantage of contemporary microarray data collections based on a fully scalable extension to the probabilistic model originally proposed for probe reliability analysis in Lahti et al. (2011).



Overview of the online-learning algorithm. The algorithm performs four consecutive sweeps over the data collection to perform (i) background correction; (ii) quantile normalization; (iii) hyperparameter estimation; and (iv) probe summarization. At each sweep, the data is read and processed in moderately sized batches that fit in computer memory. Online-update of the hyperparameters allows scalable preprocessing of the data collection.

The probe-level online-learning model

The probe-level signal s is modeled as a sum of underlying gene expression signal, probe affinity term and probe-specific stochastic noise (with variance denoted by tau below):

$$s_{ij} = a_i + \mu_j + \varepsilon_{ij}.$$

Calculating the differential probe-level signal m of the array collection s with respect to an arbitrary reference array r leads to cancellation of the unidentifiable probe affinity term. The probabilistic formulation allows incorporation of prior information of the model parameters:

$$P(\mathbf{d}, \boldsymbol{\tau}^2|\mathbf{m}) \sim P(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2)P(\mathbf{d}, \boldsymbol{\tau}^2).$$

Point estimates of the posterior hyperparameters provide priors for the next batch, allowing estimation of the model parameters in a sequential manner:

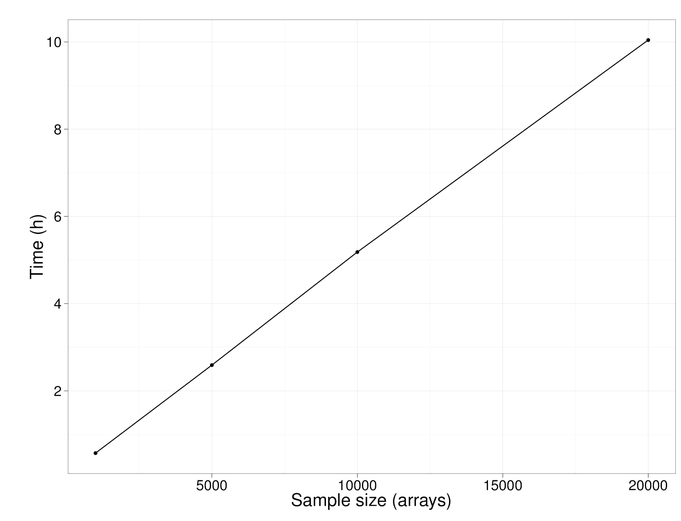
$$P(\mathbf{d}, \boldsymbol{\tau}^2) = P(\mathbf{d})P(\boldsymbol{\tau}^2) \sim \prod \Gamma^{-1}(\tau_j^2; \alpha_j, \beta_j).$$

$$P(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2) = \prod_{tj} \int N(m_{tj}|d_t - \varepsilon_{rj}, \tau_j^2) N(\varepsilon_{rj}|0, \tau_j^2) d\varepsilon_{rj}$$

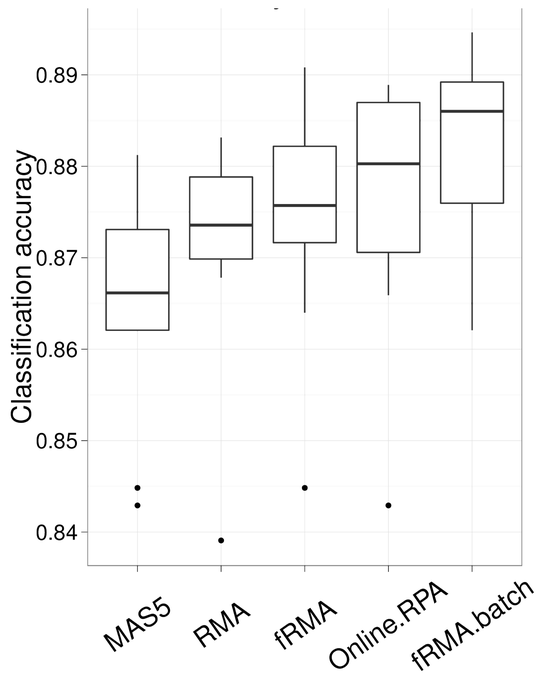
$$\sim \prod_j (2\pi\tau_j^2)^{-\frac{T}{2}} exp(-\frac{\sum_t (m_{tj} - d_t)^2 - \frac{[\sum_t (m_{tj} - d_t)]^2}{T+1}}{2\tau_j^2}).$$

The probeset level summaries of gene expression are obtained after learning the final hyperparameters based on the complete data collection. For details, see the RPA manual.

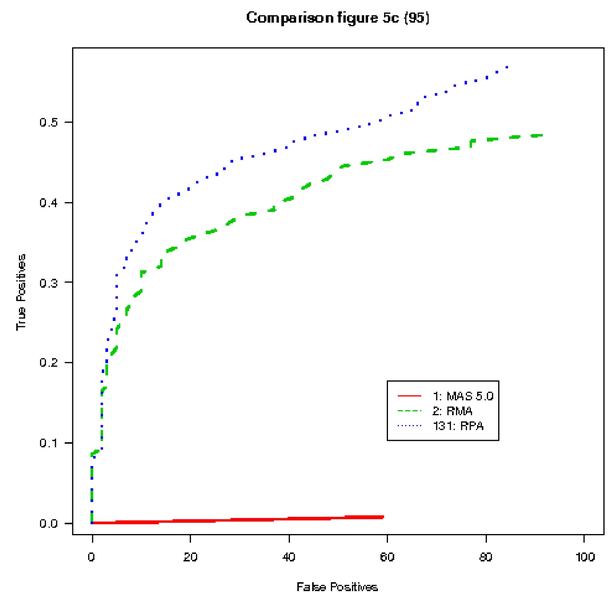
The implementation is available through the RPA package in R/Bioconductor.



The execution time scales linearly with respect to sample size, allowing scalable preprocessing of very large gene expression atlases. The calculations for this example were performed on a Z400 Desktop using four 3.06 GHz processor cores.



Sample classification performance. Online-RPA out-performs RMA and MAS ($p < 0.05$); differences between RPA-online and fRMA are not significant; fRMA-batch outperforms the other methods ($p < 0.05$) as it utilizes additional, experiment-specific information.



Average ROC curves for low-abundance concentrations with nominal concentrations at most 4 picoMolar and nominal fold changes at most 2 in the AffyCompIII spike-in data for MAS, RMA, and RPA. The fRMA is not included in the comparisons as the required probe effect terms were not available for the HG-U95Av2 microarray platform. For details, see Cope et al., 2004.

References

Cope et al. A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 20:323-331, 2004.

Lahti et al. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. TCBB/IEEE 8(1):217-25, 2011.

McCall et al. Frozen robust multiarray analysis (fRMA). Biostatistics 11:242--53, 2010.

LL has been supported by the Finnish Alfred Cordelin foundation and the Academy of Finland (decision 256950).