Quantification of reproducibility of microarray experiments by semi-parametric mixture models applied to the detection of differentially expressed genes in B-cell subpopulations

Bilgrau, Anders Ellern; Bergkvist, Kim Steve; Kjeldsen, Malene Krag; Larsen, Steffen Falgreen; Rodrigo, Maria; Schmitz, Alexander; Bødker, Julie Støve; Nyegaard, Mette; Johnsen, Hans Erik; Sørensen, Karen Dybkær; Rasmussen, Jakob Gulddahl; Bøgsted, Martin

Publication date:
2012

Document Version
Early version, also known as pre-print

NORTH DENMARK REGION

Quantification of reproducibility of microarray experiments by semi-parametric mixture models applied to the detection of differentially expressed genes in B-cell subpopulations

Anders Bilgrau[1,2,3], Kim Steve Bergkvist[1], Malene Krag Kjeldsen[1], Steffen Falgreen[1], Maria Rodrigo-Domingo[1,2], Alexander Schmitz[1], Julie Støve Bødker[1], Mette Nyegaard[1], Hans Erik Johnsen[1], Karen Dybkær[1], Jakob Guld-dahl Rasmussen[2], and Martin Bøgsted[1]

Background

Detection of differences in expression of thousands of genes using microarrays between malignant and non-malignant tissues with a limited number of samples, due to heavy costs, is a typical task in cancer research. There is, however, substantial variability in the lists of differentially expressed genes produced by multiple studies and experimental platforms, as high signal-to-noise ratios and false discoveries due to multiple testing are intrinsic to such experiments. The scientific principle of reproducibility is highly relevant in these studies but has been somewhat overlooked. A tool capable of quantifying the amount of reproducibility based on statistical modelling [1] was used to quantify the reproducibility of genes identified as differentially expressed between the Naïve (N), Germinal Center (GC), Memory (M), Centroblasts (CB), and Centrocytes (CC) B-cell subpopulations across two different microarray platforms.

Results

First an exploratory statistical analysis (Figure 1) showed evidence for a small reproducible set of genes in each comparison. Next, the model based method (Figure 2) was able to quantify 334, 1815, and 199 genes in comparisons of N vs. M (A), N vs. GC (B), and CC vs. CB (C), respectively, as reproduced between the microarrays at an irreducible discovery rate (IDR) of 0.05.

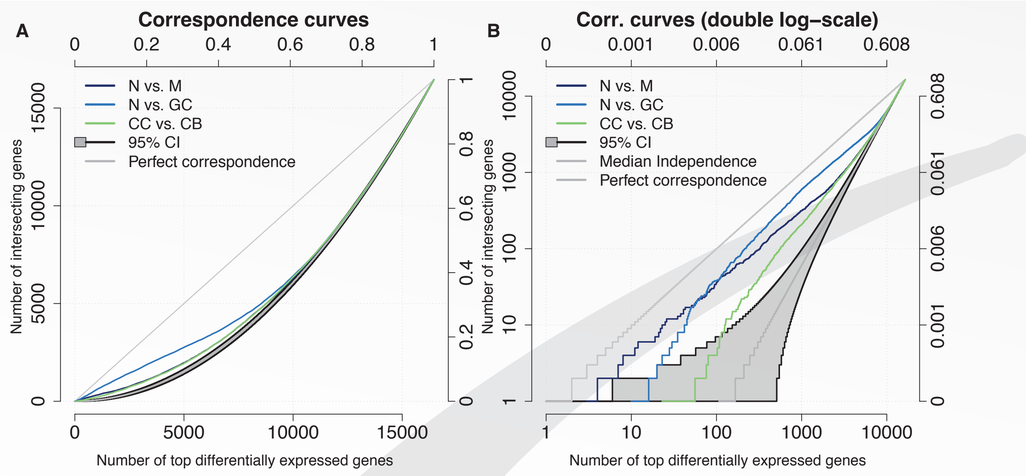

Figure 1: Correspondence Curves counts the number *y* of intersecting genes between the top *x* ranked genes in each experiments. The alternative axes show the counts relative to the total number of genes. Panel **B** is simply a log-log scaled version of panel **A**.

Model

The model assumes an observed signal depending on an unobservable latent process given by a Gaussian mixture model of a reproducible and non-reproducible group; i.e.

$$K \sim \text{Bernoulli}(\pi_0), \qquad \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \Big| K{=}k \sim N_2\left(\begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{bmatrix} \sigma_k^2 & \rho_k\sigma_k^2 \\ \rho_k\sigma_k^2 & \sigma_k^2 \end{bmatrix} \right),$$

under the assumption that

$$\mu_0 = 0, \quad \sigma_0 = 1, \quad \rho_0 = 0.$$

Thus there are 4 parameters to be estimated:

$$\theta = (\pi_1, \mu_1, \sigma_1, \rho_1).$$

The observed quantities are given by

$$x_1 = F_1^{-1}\big(\Phi_\theta(z_1)\big), \quad x_2 = F_2^{-1}\big(\Phi_\theta(z_2)\big),$$

where none assumptions are made on the marginal distribution functions F_1 and F_2.

Materials

Affymetrix GeneChips *HG-U133 Plus 2.0* and *HuEx 1.0 ST v2* microarrays f samples of removed non-malignant tonsils (n = 6) sorted with flow cytometry into 5 B-cell subpopulations N, CC, CB, M, and Plasmablast (PC) cells. The CB and CC were categorized as GC cells.

Method

Li et al [1] use a Gaussian mixture copula (see "**Model**") to model the correlation structures needed to do inter-study comparisons and meta-analyses with differences in experimental settings. The model makes it possible to obtain the a posteriori probability that each gene belongs to irreproducible component given the observed data, called the IDR. For each gene on each microarray a P-value originating from t-tests of no significant difference in gene expression in the comparisons of N vs. M, N vs GC, and CC vs CB was computed. Next, for each comparison, the IDRs across the platforms was estimated. A refined version of Li et al's algorithm was used to compute the IDRs.

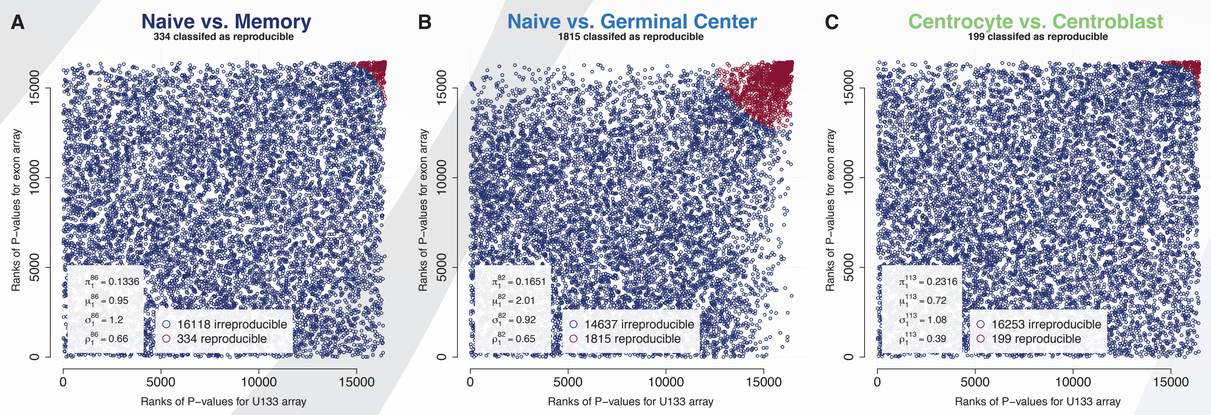

Figure 2: Scatter plot of ranks of P-values from each comparison showing the genes classified as reproducible **(red)** and irreproducible **(blue)** and the fitted parameters of the model. The model identifies the concentration of point in the upper right (i.e. genes ranking high in *both* experiments) as reproducible. Genes in the upper left or lower right are likely false positives or genes failed to be identified as differentially expressed due to low power.

Conclusion & Perspectives

The Gaussian copula mixture model was able to identify reproducible genes and assess the reproducibility of the experiment as a whole. It is, however, expected the approximative algorithm [1] can be extended along the lines of Tewari et al. [2] and thereby add to the specificity of the reproduced genes. The statistical framework is widely applicable and capable of quantifying the reproducibility between interplatform, interpopulation, and experimentally different studies to achieve greater specificity in conclusions and ensure reliable cancer research.

References

[1] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, "*Measuring reproducibility of high-throughput experiments*". The Annals of Applied Statistics, vol. 5, no. 3, pp. 1752-1779, Sep. 2011.

[2] A. Tewari, M. Giering, and A. Raghunathan, "*Parametric Characterization of Multimodal Distributions with Non-gaussian Modes*". IEEE 11th International Conference on Data Mining Workshops, 2011 286-292 (2011).

[1]Department of Haematology, Aalborg Hospital, Aarhus University Hospital
[2]Department of Mathematical Sciences, Aalborg University
[3]To whom correspondence should be adressed: **a.bilgrau@rn.dk**

Forskningens Dag, Aalborg Sygehus, 2012-04-19

AALBORG UNIVERSITET

AALBORG HOSPITAL
Aarhus University Hospital