Analysis of differential expression in microarrays where either only up- or down-regulated genes are relevant or expected by application of extreme value theory

R. Ivanek[1,*], Y.T. Gröhn[1], M.T. Wells[2], S. Raengpradub[2], M.J. Kazmierczak[3] & M. Wiedmann[2]

[1]Texas A&M University, College Station, TX, USA; [2]Cornell University, Ithaca, NY, USA; [3]Harvard Medical School, Boston, MA, USA.

1 BACKGROUND

Microarrays - Measure mRNA transcript levels to characterize gene expression.
Two-color (spotted) array - Measure expression in two biological samples, X and Y.

$$\frac{X}{Y} = \text{FoldChange} = FC$$

FC: symmetric around 1; log(FC): symmetric around 0 (Fig 1)

A new method needed to analyze spotted microarrays where the interest (e.g., to identify up-regulated gene markers of a disease) or the design of the experiment (e.g., some "wild-type v. mutant" experiments) limits identification of differentially expressed genes to those regulated in a single direction (either up or down; Fig 1, panel b).

Objective: Develop a new approach for analysis of differential expression in experiments interested in or expecting either up- or down-regulated genes.

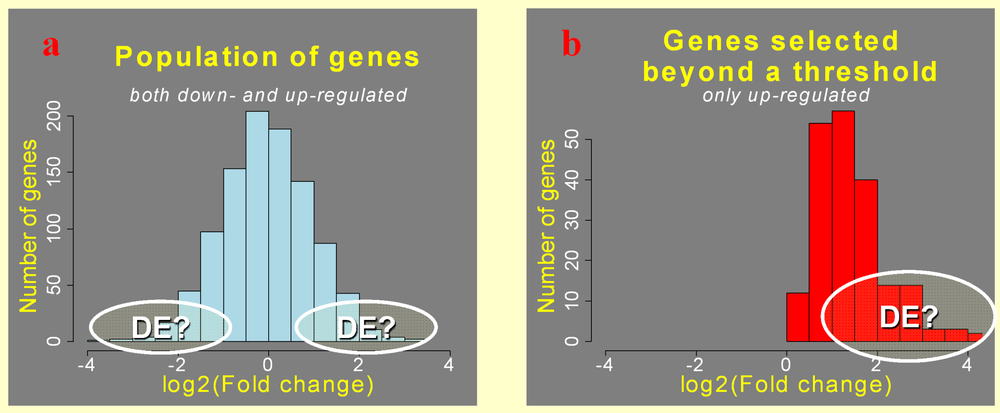

Fig 1. Comparison of histograms of gene expression data from (a) large scale microarrays and (b) microarrays limited to a single direction of expression. "DE"=differentially expressed.

2 METHODS

Develop a new Empirical Bayes approach based on the Extreme Value Theory, and compare its performance with two existing empirical Bayes methods (Fig 2): Limma [2,4] (BN) and Lapmix[1] (BL) on a real dataset and in a simulation study.

Empirical Bayes: (1) Guess the probability distribution of the parameter in question (prior; estimate parameters of the prior based on the data). (2) Observe data. (3) Use Bayes' theorem and modify the prior based on the data to get a posterior distribution.

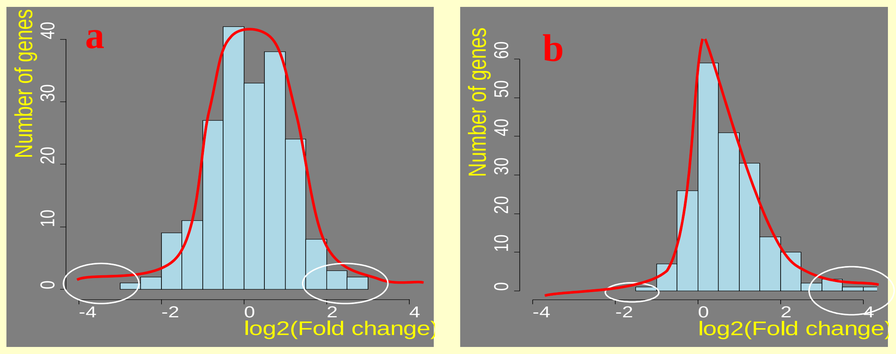

Fig 2. Assumed prior distributions in the existing empirical Bayes methods: (a) Normal prior in the BN, and (b) Asymmetric Laplace in the BL method.

3 RESULTS

(i) New approach developed:

Empirical Bayes Extreme Value Distribution mixture model (BE).

BE = log(Posterior Odds Ratio) = log(O_{ij})

$$\log(O_{ij}) = \log\frac{p(M_1 \mid data)}{p(M_0 \mid data)} = \log\frac{p(M_1)\,p(data \mid M_1)}{p(M_0)\,p(data \mid M_0)}$$

M_1= the gene is differentially expressed
M_0= the gene is not differentially expressed

- p(M_1) = *pDE* (guessed % of differentially expressed genes)
- p(M_0) = 1-*pDE*

μ_g ~ **Extreme Value Distribution**

$$p(\hat{\alpha}_g, s_g^2 \mid M_1) = \iint f(\hat{\alpha}_g \mid \mu_g, \sigma_g^2) f(s_g^2 \mid \sigma_g^2)\, f(\mu_g)\, f(\sigma_g^2) d\mu_g d\sigma_g^2$$

$$p(\hat{\alpha}_g, s_g^2 \mid M_0) = \int f(\hat{\alpha}_g \mid \mu_g = 0, \sigma_g^2) f(s_g^2 \mid \sigma_g^2)\, f(\sigma_g^2) d\sigma_g^2$$

Data likelihoods — Parameter priors

Monte Carlo integration method.

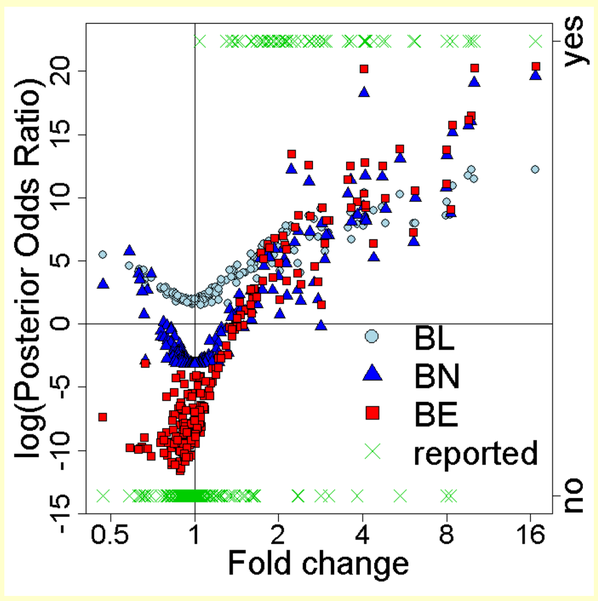

Fig 3. The BN, BL and BE statistics plotted against the Fold changes. "reported" and the associated right y-axis indicate whether a gene has been previously reported as differentially expressed ("yes") or not ("no").

(ii) New approach fits better to the real data[3] (Fig 3):

Listeria monocytogenes grown under osmotic stress. Mutant: deleted *sigB* gene (lacks σ^B protein). Wild type: parent strain - intact *sigB* gene. σB protein=positive regulator → expected up-regulated genes.

(iii) New approach showed better accuracy and precision in a simulation study (Fig 4):

False discovery rate (FDR) = expected proportion of errors among the genes selected to be differentially expressed. False negative rate (FNR) = the (1-sensitivity) or (1-power).

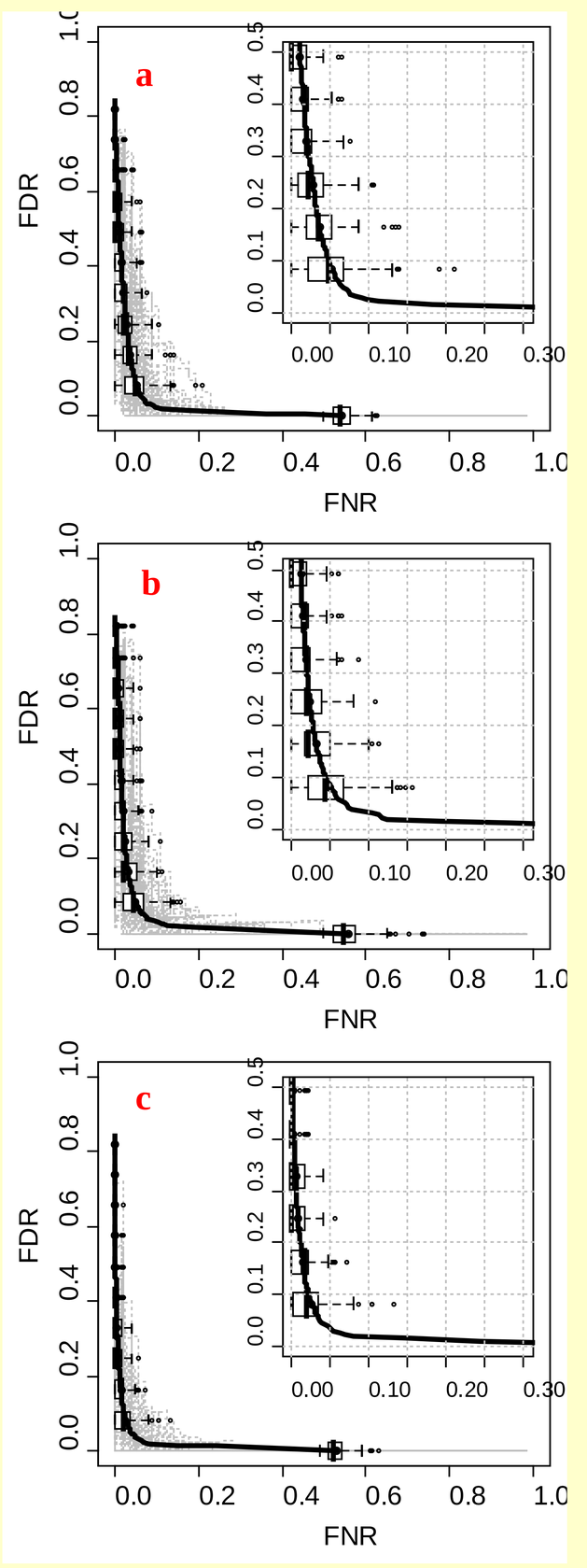

Fig 4. False discovery rate (FDR) v. false negative rate (FNR) plots of 100 simulated datasets overlaid by the horizontal average curve and box plots showing horizontal spread of the performance (achieved FNR for a selected FDR) of the (a) BN, (b) BL and (c) BE statistics.

4 CONCLUSIONS

Compared with the BN and BL, the new method (BE) fits better to the real data. In the analysis of simulated data, the BE method showed better accuracy and precision. The BE method, therefore, seems promising and useful for inference about differential expression in custom, restricted coverage microarray experiments where either only up- or down-regulated genes are relevant or expected.

REFERENCES:
1.Bhowmick, D., Davison, A.C., Goldstein, D.R. and Ruffieux, Y. (2006) A Laplace mixture model for identification of differential expression in microarray experiments. Biostatistics. 7(4):630-641.
2.Lonnstedt, I., and Speed, T.P. (2002) Replicated microarray data. Statistica Sinica. 12:31–46.
3.Kazmierczak, M.J., Mithoe, S.C., Boor, K.J., and Wiedmann, M. (2003) *Listeria monocytogenes* σ^B regulates stress response and virulence functions. Journal of Bacteriology. 185: 5722-5734.
4.Smyth, G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. 3(1), Article 3.