

Supplementary Document 5

This supplementary document is for: Reproducibility Probability Score: Incorporating Measurement Variability across Laboratories for Gene Selection (2006). Lin G, He X, Ji H, Shi L, Davis R and Zhong S, Nature Biotechnology.

Discussion on the RPS

We developed a statistical metric and procedure that takes into account inter-laboratory measurement variability for gene selection. This procedure allows the rich MAQC dataset to be applied to analyze other microarray data generated by a single laboratory. To our knowledge, no other gene selection metrics explicitly takes into account the inter-laboratory variation. Without a reference study such as the MAQC, the task of incorporating the inter-laboratory variability would not have been practical when an individual study is carried out in only one site. We demonstrated the merits of the RPS using validations on the MAQC data and on an independent study for colorectal adenocarcinoma. Both validations showed that the genes selected by the RPS had much higher reproducibility in the genes selected by other methods.

A basic assumption underlying the RPS procedure is a fraction of the gene-specific inter-laboratory variation obtained from reference data will persist. This assumption is based on the fact that consistent gene expression measurements are partly attributable to the biophysical properties of a given microarray format and its probesets. So long as the same microarray is used for a given gene expression study, the biophysical properties of a probeset are likely to persist. This assumption is better satisfied when the amount of reference data increases. We note that the model training step, i.e. the learning of the inter-laboratory correlation, can always be improved with more reference data. As a result, the power of the RPS metric increases as the reference dataset expands. Future multi-laboratory studies can contribute new reference data to the RPS procedure.

In the RPS package, we provide a function “RPS.learn()” to read future multi-laboratory data as the reference data and to update the inter-laboratory correlations.

If a gene satisfies a gene selection criterion ($\Theta > \theta$) based on the data from the real laboratory L_0 , it is easy to understand the “reproducibility” of this gene. However, by definition, every gene has an RPS regardless of whether it satisfies $\Theta > \theta$. In general, RPS is the estimated probability of a gene satisfying $\Theta > \theta$ if data were to be generated from a randomly selected laboratory, and therefore every gene has an RPS. This definition avoids the problem of a “hard” cutoff in gene selection. Instead, the RPS provides an educated gene selection procedure. For example, if the “hard” gene selection cutoff is $FC > 2$, a gene with $FC = 1.99$ will not be selected. However, if this gene has consistent measurements across the reference laboratories, it could have a higher RPS than another gene with a higher FC in L_0 but have inconsistent measurements across the reference laboratories.

The mixed-effects model (Model 6) resembles the formulation by Kerr and Churchill [1-3], although it is applied for a different purpose. It does not aim to identify differentially expressed genes in the reference data. Instead, it is used to estimate the gene-specific interlaboratory correlation ρ . We note that the ρ computed from Equation 10 is not the raw correlation between the gene expression values in two laboratories. The former is not affected by the average expression value in any sample, while the latter is. The sample average expression value is represented by the μ in Model 6. It is independent to ρ , because ρ only depends on L_i , S_j , and ε_{ijr} , which are all independent to μ . Figure S8 gives the scatter plots between ρ and the average expression value in each sample. Ideologically we are learning the part of the cross-laboratory variation that is independent to the average expression values in any reference samples. This is the only useful information we could get from the reference data,

and the average expression values in the reference samples do not contain any useful information for the analysis of the new samples. Equation 14 shows that the inter-laboratory variation of the simulated data for a new sample is a product of two components. The first component is a sample independent variation, represented by ρ and can be learnt from the reference data. The second component, σ^2 , depends on the particular biological sample to be analyzed, and it is estimated from the data in the real laboratory L_0 .

We also note that previous variations ANOVA methods [1-5] cannot be directly extended to perform similar analysis as the RPS analysis. This is because the reference data are usually generated from different biological samples as the biological samples in a new study.

A natural extension of our work is to model cross-platform correlations as well as the cross-laboratory correlations. This might be achieved by modifying Model (6) into:

$$Y_{ijt} = L_i + S_j + P_t + \varepsilon_{ijt}$$

where Y_{ijt} is the expressions on sample j in laboratory i with microarray platform t , and L_i , S_j , and P_t are laboratory, sample, and platform effects, respectively. Here, P_t will be treated as a fixed effect, and the platform effect might be nested within the laboratory effect, depending on how the multiple laboratory study is designed. The normality and independence assumptions similar to Equations (7-9) can be made. This mixed-effects model could be used to estimate the laboratory-to-laboratory correlations regardless of microarray platforms. Future versions of the RPS built upon this modified model might be more powerful, because it accounts for both the inter-laboratory and inter-platform variabilities. A higher RPS provides higher confidence in that the differential expression of a gene is more likely to be valid regardless of which platform and which laboratory conducted the study. This assumes that the gene expression measurements from different

microarray platforms and different laboratories are unbiased and non-negatively correlated.

Reference:

1. Kerr, M.K. and G.A. Churchill, *Experimental design for gene expression microarrays*. Biostatistics, 2001. **2**(2): p. 183-201.
2. Kerr, M.K. and G.A. Churchill, *Statistical design and the analysis of gene expression microarray data*. Genet Res, 2001. **77**(2): p. 123-8.
3. Kerr, M.K., M. Martin, and G.A. Churchill, *Analysis of variance for gene expression microarray data*. J Comput Biol, 2000. **7**(6): p. 819-37.
4. Wolfinger, R.D., et al., *Assessing gene significance from cDNA microarray expression data via mixed models*. J Comput Biol, 2001. **8**(6): p. 625-37.
5. Baldi, P. and A.D. Long, *A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes*. Bioinformatics, 2001. **17**(6): p. 509-19.

Figure S8: (Left) Scatter plot between the interlaboratory correlation ρ and the average expression value in sample A. (Right) Scatter plot between ρ and the average expression value in sample B. These scatter plots show that the ρ computed from Equation 10 is not correlated to the average gene expression value in either reference sample.

