Genet. Res., Camb. (2007), **89**, *pp.* 505–507. © 2008 Cambridge University Press
doi:10.1017/S0016672308009725 Printed in the United Kingdom

Proper experimental design and sound statistical inference win every time: a commentary on 'Statistical design and the analysis of gene expression microarray data' by M. Kathleen Kerr and Gary A. Churchill

R. W. DOERGE*
Departments of Statistics and Agronomy, Purdue University, 150 N. University Street, West Lafayette, IN 47907, USA

'The designs that are most suitable for any particular experiment depend on the questions of interest and the hypotheses to be investigated.' Kerr & Churchill (2001).

The three basic experimental design concepts that every statistician attributes to R. A. Fisher are randomization, replication and blocking. Given the historic bonds between statistics, genetics and agriculture, it seems fitting that microarray technology (Brown & Botstein, 1999) was the vehicle that brought history full circle. In 2001, Kerr and Churchill provided what some may call a review, but really their paper was one of the first that brought experimental design concepts to microarray experiments by reminding readers of the intimate connection between statistics and agriculture. Using analogies to agricultural (i.e. field trials) experiments, they carefully laid out the fundamental issues of experimental design as applied to microarray technology. The relevance of the statistical issues associated with the analysis of microarray data as identified by Kerr and Churchill (2001) has stood the test of time while providing the much needed guidance to an expanding community.

Microarray technology has enabled the quantification of expression for a large number of genes, or even all genes in a genome, simultaneously through the exploitation of messenger RNA (mRNA). The Central Dogma (Crick, 1970) of molecular biology dictates that genes consist of DNA, and when genes are transcribed or read by an enzyme (i.e. RNA polymerase), a complementary strand of RNA (i.e. mRNA) is produced and is referred to as a transcript. In short, these transcripts are then translated by other enzymes to produce protein sequences. If the mRNA can be captured (i.e. sampled) and quantified, then the amount of mRNA present in the sample is an indication of how much a gene has been transcribed or expressed. Since different genes participate in different biological processes (e.g. development and disease), it is of interest to evaluate differences in a gene's expression for different conditions, treatments or states.

In 2001, the potential of microarray technology was seemingly vast and the excitement from biologists, who were suddenly able to simultaneously quantify the expression of every gene in a genome, was unmistakable. Thrilled statisticians were equally excited about relying on rather simple statistical models to partition the sources of variation as easily as they could for any experiment. However, as quickly as the excitement entered, it waned (on both sides). Biologists were unable to confirm or reproduce known results (i.e. genes that were known to be differentially expressed under certain conditions), and statisticians were suddenly aware that they had their work cut out for them as the data were limited in sample size, of high dimension and, in the early days, very noisy. Kerr & Churchill's (2001) review was timely in that it focused on the basic concepts that were important and accessible to both biologists and statisticians, namely issues about defining an experimental unit, technical and biological replication, experimental design, identifying and partitioning sources of variation, and testing correct hypotheses.

Kerr and Churchill set the stage for future experimental designs and analyses by discussing as examples both the dye-swap design (Fig. 1; Churchill, 2002) and the reference design in the context of an analysis of variance (ANOVA) model. In doing so, discussions about normalization of data, balanced versus unbalanced designs, replication, estimation of variation and statistical inference arose naturally. While the discussion of these issues was certainly important and necessary, the simple manner in which the points were made was probably most relevant at the time. For example, Kerr and Churchill (2001) pointed out that when mRNA samples are obtained from two different individuals under different conditions, the experimental unit is the 'spot' on the array that represents a gene, and therefore comparisons between samples are made 'within' genes. With an

* e-mail: doerge@purdue.edu

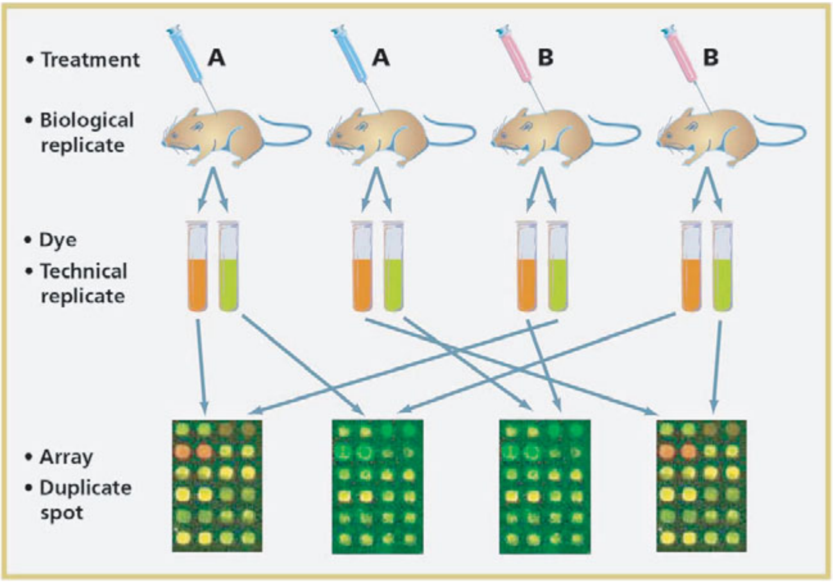

Fig. 1. Dye-swap experimental design taken from Churchill (2002). Two of four biological replicates receive treatment A, while the other two biological replicates receive treatment B. mRNA is extracted from each biological replicate. The samples are then split into subsamples, or technical replicates, and labelled with one of two dyes (i.e. red or green). There are four possible comparisons of a sample from mouse treated with A with a mouse treated with B, and as such, the alternatively labelled samples from mouse A and mouse B are combined and hybridized to an array that comprises duplicate spots for every gene. Notice that in two of the four comparisons, of mouse A with mouse B, the colour of the labelled samples are exchanged, or swapped. After hybridization, the predominant 'spot' colour (red or green) on the array represents more expression/transcripts of a gene in the respectively labelled sample than its counterpart. If the 'spot' is yellow, this indicates equal amounts of both samples or no differential expression between samples for that gene. The arrays are then scanned, and an intensity signal for each 'spot' on the array obtained as a continuous data point. Reproduced with permission of Nature Publishing Group.

appreciation of the experimental unit in hand, Kerr and Churchill's discussions and statements (e.g. 'Without the ability to estimate error there is no basis for statistical inference.') about data normalization, when and where to replicate and how this affects estimation provided a foundation for future statistical models, approaches and theories.

The impact of the work by Kerr and Churchill in the early 2000s was twofold. Firstly, through their clever analogies and simple examples, they were able to attach statistical concepts to biological phenomena, and thus enabled many members of the statistical community to become more actively involved in genomics. Secondly, they greatly influenced our current understanding of the important statistical issues that are associated with the analysis of microarray data. While most of these same statistical issues remain of importance today, it is interesting to realize that the challenges Kerr and Churchill predicted for the future have indeed happened: namely, that mixed models and random effects (Wolfinger *et al.*, 2001) do have a role in the analysis of microarray data; that genes with small, but reproducible, changes in expression are of biological interest; that without replication, biologists are unable to assess which features in the data arose by chance; and that discovering that a model is not adequate often assists the modelling process by identifying sources of variation and bias that were either missed or not understood.

So, how have microarrays changed over the last 10 years, and is the information provided by Kerr and Churchill still relevant? The quantification of data are better and the dimensionality higher, but as is said in Kerr & Churchill (2001) 'collecting data and acquiring data are not the same thing'. The statistical issues that were introduced as being of greatest importance in the early days of microarrays remain, and it turns out that Kerr and Churchill, like many others, were correct in their opinions that sound statistical inference was/is indeed the crucial factor that fulfilled the potential of microarray technology to impact science. Will microarray technology continue to impact the future of science? Microarrays will have their place in science for a while, but most certainly will be replaced by the next latest, greatest technology (i.e. next-generation sequencing). More, better, faster data and a seemingly new set of statistical challenges will arise. However, if we look very closely and remember the lessons learned from Kerr and Churchill, we will realize that the three important concepts to be applied to every experiment are the same (i.e. randomization, replication and blocking), and that regardless of the technology a solid

https://doi.org/10.1017/S0016672308009725 Published online by Cambridge University Press

design and sound statistical inference will win out every time.

References

Brown, P. O. & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics* **21**, 33–37.

Churchill, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490–495.

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature* **227**, 561–563.

Kerr, M. K. & Churchill, G. A. (2001). Statistical design and the analysis of gene expression microarray data. *Genetical Research* **77**, 123–128.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. & Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* **8**, 625–637.

https://doi.org/10.1017/S0016672308009725 Published online by Cambridge University Press