expected. However, the identity of expression QTLs (eQTLs) should remain unaffected if they genuinely influence transcript expression. Contrary to this prediction, in the three recombinant inbred line Affymetrix data sets[4–6] we found an overall concordance of 39% to 70% of eQTLs between the Microarray Analysis Suite 5 (MAS5) and Robust Multiarray Average (RMA) analyses at a genome-wide significance level of 0.05; the highest concordance was observed when comparing RMA with gene chip RMA analyses (45%–85%). At the same significance level, the 31-BxD liver data (two-color glass) showed similar patterning of concordance: the highest concordance (76%) was observed when comparing print-tip *loess* normalization with print-tip *loess* that was followed by interarray quantile normalization.

Loci with association to many transcripts might seem to be strong candidates for harboring a *trans*-acting transcriptional effector and are likely to be prioritized in downstream analyses—particularly since such loci could imply a functional relationship between the coregulated genes, or 'regulons'[9]. We hypothesized that reproducibility of linkage from multitranscript linkage peaks (**Supplementary Methods**) across different normalizations would be higher than the overall degree of overlap, but in practice this varied greatly, both within and between different experiments (**Supplementary Fig. 3** online).

In order to gain further insight into the degree of between-array normalization that is appropriate for detecting plausible genetic signal, we undertook a simulation study of 32 RI strains, with 1.5% of simulated transcripts being subject to a genetic influence from a single locus (**Supplementary Methods**). These simulated data were then further modified to introduce systematic biases (**Supplementary Methods**), including per-array changes in mean and variance, in order to model systematic bias consistent with that observed in empirical data. From this modified data set, the best recovery of modeled genetic signal was made from quantile-normalized data (**Supplementary Table 1** online), indicating that conservative adjustment of differences in interarray data structure permits recovery of an embedded genetic signal relative to less conservative removal of interarray structure.

Collectively, these observations demonstrate that differences in normalization of raw microarray data can have a profound influence on the ability of genetical genomics experiments to identify transcripts demonstrating genetic linkage and their cognate eQTLs. We cannot easily determine the nature of the artifactual signals that are propagated through normalization, in part because it is likely to vary from experiment to experiment. We would suggest that the following considerations should apply: (i) following initial quality control (and, if appropriate for the platform, intra-array normalization), remove as much interarray data structure as possible using platform-specific normalizations (*i.e.*, those based on scale or quantile normalization); (ii) exclude genes with low-intensity signal levels that are likely to arise from nonexpressed transcripts, as such data lacks biological plausibility; (iii) perform linkage analysis on the same data normalized by different approaches; and (iv) treat linkage that is robust to independent normalization techniques as more reliable than linkage that is not. Ultimately, the interpretation of linkage data will be resolved by experimental validation of large numbers of detected linkage signals in genetical-genomics experiments; these are presently unavailable in any study published thus far.

Rohan B H Williams[1,2,4], Chris J Cotsapas[1,4], Mark J Cowley[1], Eva Chan[1], David J Nott[3] & Peter F R Little[1]

[1]School of Biotechnology and Biomolecular Sciences, [2]The Clive and Vera Ramaciotti Centre for Gene Function Analysis and [3]School of Mathematics, The University New South Wales, Sydney 2052, Australia. [4]These authors contributed equally to this work.
e-mail: p.little@unsw.edu.au

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank J. Schimenti (formerly of the Jackson Laboratory) and M. Bucan (University of Pennsylvania) for assistance in obtaining mouse strains. Supported by an Australian Research Council Discovery Grant award (P.F.R.L.), a National Health and Medical Research Council Peter Doherty Fellowship (R.B.H.W.), the University of New South Wales Genome Informatics Scholarship (C.J.C.) and Australian Postgraduate Awards (E.C. and M.J.C.) and a grant-in-aid from the Australian Centre for Advanced Computing and Communications (P.F.R.L.).

1. Schadt, E.E. *et al. Nature* **422**, 297–302 (2003).
2. Yvert, G. *et al. Nat. Genet.* **35**, 57–64 (2003).
3. Morley, M. *et al. Nature* **430**, 743–747 (2004).
4. Chesler, E.J. *et al. Nat. Genet.* **37**, 233–242 (2005).
5. Bystrykh, L. *et al. Nat. Genet.* **37**, 225–232 (2005).
6. Hubner, N. *et al. Nat. Genet.* **37**, 243–253 (2005).
7. Smyth, G.K. & Speed, T. *Methods* **31**, 265–273 (2003).
8. Irizarry, R.A. *et al. Nat. Methods* **2**, 345–350 (2005).
9. Cotsapas, C.J. *et al. Cold Spring Harb. Symp. Quant. Biol.* **68**, 109–114 (2003).

Chesler *et al.* reply:

Williams *et al.* raise an important point concerning normalization that is not unique to 'genetical genomics' but is generic to microarray experiments. We address their comments based on (i) the conceptual and methodological grounds on which conclusions were drawn, (ii) the observed patterns of QTL concordance, (iii) the degree to which we have covered these issues in our work and (iv) a biological explanation of the apparently widespread lack of concordance among QTLs despite highly consistent patterns of expression regulation.

Genetical genomics is subject to challenges of both microarray and QTL analysis, including appropriateness of normalization and mapping methods and assessment of statistical significance in a context of massive correlation among statistical tests. The difficulty in microarray analysis comes from determining which method best approximates 'truth' in a situation in which truth is typically unknown. Seeking consensus among multiple methods, some of which are notoriously poor, does not improve one's ability to identify true positives. Such an approach is likely to miss many real loci because each method yields both false positives and false negatives, and the set of results concordant for all methods will amplify false negatives because failure to detect a QTL with any single method will result in its exclusion. This is particularly so in the methods of Williams *et al.*, because they used model-based nominal *P* values and Bonferroni adjustment across markers to establish linkage. The accuracy of these values depends on the degree to which the distribution of each transcript satisfies model assumptions. Bonferroni correction is a poor substitute for genome-wide *P* values[1] because the dense marker map used by Williams *et al.* renders tests of linkage to a transcript non-independent. Excessively conservative significance thresholds such as these are biased against detection of concordance because they increase the number of false negatives. A comparison of the RMA and MAS5 algorithms uncovers a predominant pattern of *trans*-regulatory band concordance (see Fig. 2 of our central nervous system study[2]).

The challenge is to choose an approach that best approximates biological reality. In genetic association studies, there is a method of applying this criterion for success. *Cis*-regulatory QTLs are transcripts for which the top locus controlling expression of a given gene maps back to the gene itself.

The *cis*-regulatory QTLs provide a measure of 'correct' QTL targeting. Their enrichment gives an approximation of a method's ability to detect true positives. Given a genome size of 3,000 Mb and markers every 10 Mb, this should occur by chance only 0.33% of the time (41 out of 12,422 results). *Cis*-regulatory QTLs are unlikely mapping artifacts. (They can be hybridization artifacts, a problem addressed in our[2] and others' work[3]). Using *cis*-regulatory QTL enrichment, we have compared normalizations and report that RMA[4] and PDNN[5] (Paired Difference Nearest Neighbors) methods perform best. These methods also perform best on measures of biological consistency, determined using strain intraclass correlation for each transcript[6]. This is the amount of trait variance among strains relative to total transcript variance. The normalization that produces the highest median genetic variance over the array maximizes between-strain signal and reduces within-strain noise. Again, RMA and PDNN are best (Fig. 1a). Although it may seem that many more QTLs are called using noisy condensation algorithms, a greater proportion of these are for traits with low heritability—chance associations of noise to genotypes. We have made all of this data publicly available at http://www.genenetwork.org so that users can compare the diverse set of normalization methods. In a second confirmatory approach, we report consistency of genetic regulation of gene expression across tissues in our original papers[2,7]. However, the most important test of reliability is concordance on fully independent biological replication; Peirce *et al.*[8] deal with independent replication of QTLs using two pairs of BXDs and F2 data sets available in GeneNetwork: BXD F2 and BXD RI.

We concur that *trans*-regulatory QTLs detected using single-QTL models vary in their location owing to vagaries of normalization. The explanation for this is as much biological as it is statistical. Most continuously distributed phenotypes, including gene expression[9], are regulated by several genetic loci. Variations in normalization may influence which of several loci are detected when a single locus model is applied. We have characterized the relationships among *trans*-regulatory bands, and we show that correlated gene expression levels are actually regulated by combinations of loci at the *trans*-regulatory band locations[10]. Slight relative changes introduced by various normalizations may enhance association to one genetic locus over another. Therefore, the maximum LRS location may move from one *trans*-regulatory band to another depending on the method used, but the *trans*-regulatory bands remain consistent. When the locations of single-locus QTL peaks are compared across normalizations, most non-concordant transcript regulatory loci map to the locations of other *trans*-regulatory bands (Fig. 1b). Furthermore, because the single-locus model is inappropriate, there will be many false-negative QTLs. Each locus by itself does not account for enough of the trait variance to be detected without fitting additional loci. Excessive thresholding will increase false negatives. In Figure 3 of Chesler *et al.*[2], the heritability filtering and false discovery rate (FDR) shading in the plot of transcriptome QTLs for RMA- versus MAS5-normalized data show this.

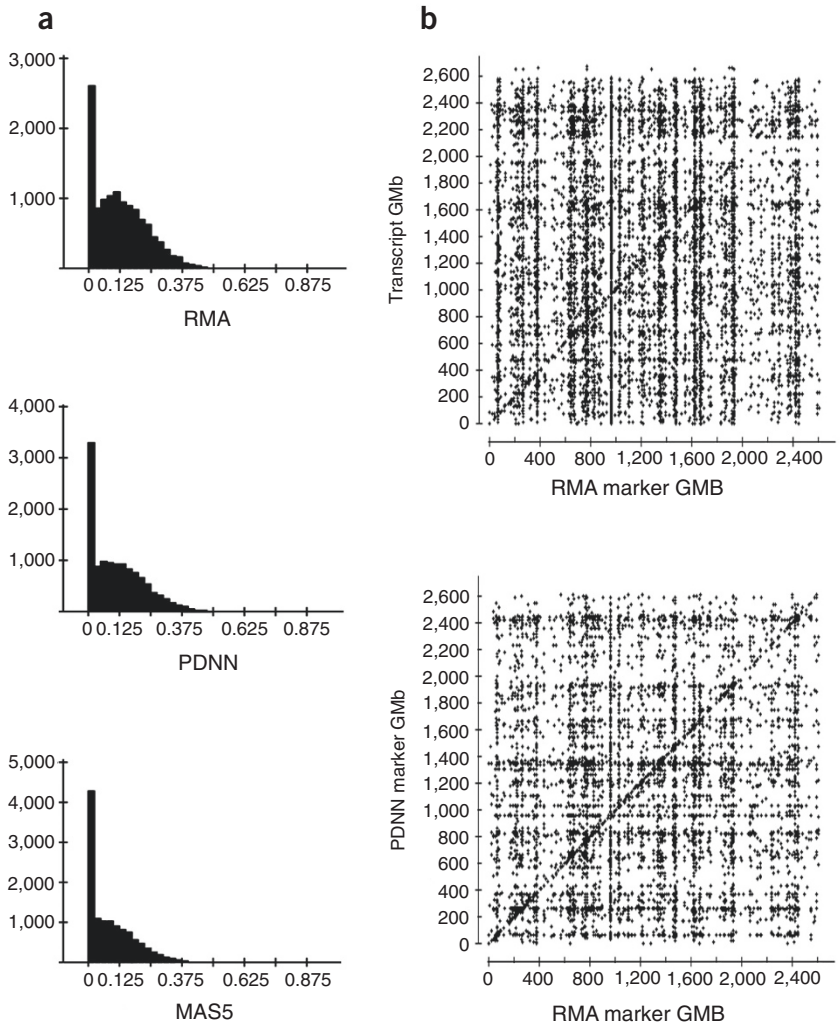

Figure 1 Effect of normalization on genetic analysis of gene expression. (**a**) Histograms of strain intraclass correlation for expression of each of the 12,422 transcripts analyzed in Chesler *et al.*[2] using three different normalization methods. Median intraclass correlation is highest using RMA, indicating that this method has generated the highest signal-to-noise ratio in this data set. (**b**) Upper panel: the RMA transcriptome map shows a plot of peak QTL location for each transcript versus each transcript location. Lower panel: a strikingly similar plot of peak RMA QTL location versus peak PDNN QTL location shows that when there is a lack of concordance, QTL peaks are often found on a different *trans*-regulatory band. The plot indicates the presence of polygenic regulatory mechanisms. GMb: genome megabase (genome location sequentially from proximal chromosome 1 to distal chromosome 2).

Williams *et al.* recommend a four-step process similar to that described in our previous work[2], in which we report that (i) the data were scaled, (ii) strain intraclass correlation was used to select an optimal normalization and identify transcripts for which genetic linkage was a plausible explanation for variability, (iii) mapping was evaluated using several normalization approaches and (iv) normalizations were compared and illustrated, revealing the same overall pattern of *trans*-regulatory loci

across normalizations. A wealth of analytic results went into the development of this process, and we presented a select few in our reports. We carefully evaluated different normalizations and heritabilities and chose to concentrate on specific normalizations that minimize the number of QTLs called for non-heritable traits. The answer to the question of which method is best will await large-scale validation of the gene expression data by independent methods and, more importantly, determination of the number of biological insights gained from these approaches.

Elissa J Chesler

Mammalian Genetics and Genomics Group, Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. e-mail: cheslerej@ornl.gov

Leonid Bystrykh & Gerald de Haan

Department of Cell Biology, Stem Cell Biology, University Medical Center Groningen, 9713 AV Groningen, The Netherlands.

Michael P Cooke & Andrew Su

Genomics Institute of the Novartis Research Foundation, 10675 John J. Hopkins Drive, San Diego, California 92121, USA.

Kenneth F Manly & Robert W Williams

Center for Genomics and Bioinformatics, University of Tennessee Health Science Center, 855 Monroe Ave., Memphis, Tennessee 38163, USA.

1. Churchill, G.A. & Doerge, R.W. *Genetics* **138**, 963–971 (1994).
2. Chesler, E.J. *et al. Nat. Genet.* **37**, 233–242 (2005).
3. Alberts, R., Terpstra, P., Bystrykh, L.V., de Haan, G. & Jansen, R.C. *Genetics* **171**, 1437–1439 (2005).
4. Irizarry, R.A. *et al. Biostatistics* **4**, 249–264 (2003).
5. Zhang, L., Miles, M.F. & Aldape, K.D. *Nat. Biotechnol.* **21**, 818–821 (2003).
6. Carlborg, O. *et al. Bioinformatics* **21**, 2383–2393 (2005).
7. Bystrykh, L. *et al. Nat. Genet.* **37**, 225–232 (2005).
8. Peirce, J.L. *et al. Mamm. Genome* **17**, 643–656 (2006).
9. Brem, R.B. & Kruglyak, L. *Proc. Natl. Acad. Sci. USA* **102**, 1572–1577 (2005).
10. Chesler, E.J. & Langston, M.A. *Lect. Notes Bioinformatics* (in the press).

Petretto *et al.* reply:

R.B.H. Williams and colleagues discuss the robustness and consistency of linkage analysis of microarray-based 'genetical-genomics' experiments. The challenges of data handling and statistical analysis of microarray experiments have been acknowledged, but we believe that standards have now emerged that, in contrast to the viewpoint of Williams *et al.*, give credibility and rigor to microarray research in general and to the genetical-genomics design in particular. We find that the Williams *et al.* correspondence contains several methodological weaknesses that undermine the authors' main conclusions.

We also believe that most of the issues raised have been researched, discussed and resolved in previous publications.

The central proposition of Williams *et al.* is a lack of agreement between gene lists for expression quantitative trait locus (eQTL) linkages derived using different normalization methods, such as MAS5 and RMA (Fig. 1 in Williams *et al.*). However, the difficulty of comparing lists of genes has long been recognized[1]. Williams *et al.* employed the 'correspondence at the top' (CAT) plot[2] to quantify the concordance of linkage results in RMA, gcRMA and MAS5 data sets but fail to recognize that this method introduces sensitivity to small changes in gene order by arbitrarily dichotomizing the continuous distribution of *P* values. For instance, the top 20 significant transcripts identified as dysregulated by normalization method A may not necessarily be within the top 20 transcripts identified by method B, despite all transcripts achieving statistical significance with both methods. To illustrate this point, we used the expression data in our SHR and BN parental strains[3] to compare the RMA and MAS5 normalization methods. The proportion of differentially expressed genes in common between the RMA and MAS5 data sets depends on the ranks that are compared (Fig. 1a) and

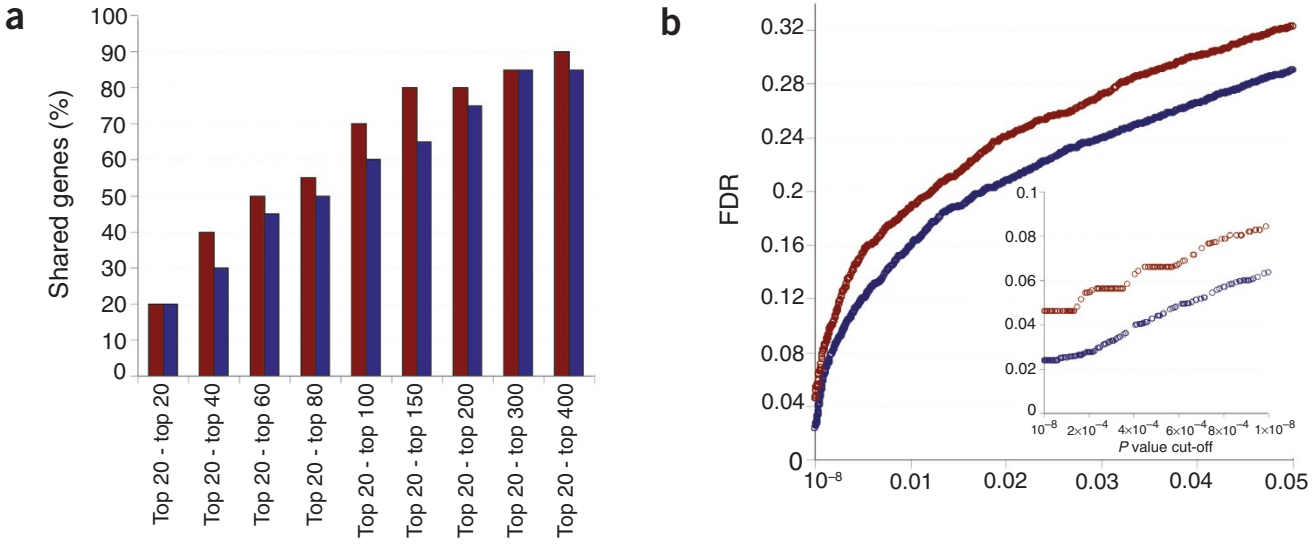

Figure 1 Comparison of the RMA and MAS5 normalization methods. (**a**) We compared the top 20 differentially expressed genes ($P \leq 5 \times 10^{-5}$) detected using MAS5 (red) or RMA (blue) with the top 20, 40, 60, 80, 100, 150, 200, 300 and 400 differentially expressed genes detected using RMA or MAS5 in fat tissue. The proportion of genes that are shared between the top ranks is reported for each comparison. For both RMA and MAS5 data sets, the top 400 genes all show differential expression with $P \leq 5 \times 10^{-3}$. Differential expression analysis in kidney tissue showed similar results. (**b**) False discovery rate (FDR) for different *P* value thresholds for the genes showing differential expression ($P \leq 0.05$) in fat tissue between SHR and BN parental strains. Two normalization procedures are compared: MAS5 (red) and RMA (blue). Inset: FDR for different *P* value thresholds in the range 10^{-8} to 10^{-3}.