nature genetics

Towards sound epistemological foundations of statistical methods for high-dimensional biology

Tapan Mehta[1,2], Murat Tanik[1,2] & David B Allison[1,3]

A sound epistemological foundation for biological inquiry comes, in part, from application of valid statistical procedures. This tenet is widely appreciated by scientists studying the new realm of high-dimensional biology, or 'omic' research, which involves multiplicity at unprecedented scales. Many papers aimed at the high-dimensional biology community describe the development or application of statistical techniques. The validity of many of these is questionable, and a shared understanding about the epistemological foundations of the statistical methods themselves seems to be lacking. Here we offer a framework in which the epistemological foundation of proposed statistical methods can be evaluated.

The challenge we face

High-dimensional biology (HDB) encompasses the 'omic' technologies[1] and can involve thousands of genetic polymorphisms, sequences, expression levels, protein measurements or combination thereof. How do we derive knowledge about the validity of statistical methods for HDB? A shared understanding regarding this second-order epistemological question seems to be lacking in the HDB community. Although our comments are applicable to HDB overall, we emphasize microarrays, where the need is acute. "The field of expression data analysis is particularly active with novel analysis strategies and tools being published weekly" (ref. 2; **Fig. 1**), and the value of many of these methods is questionable[3]. Some results produced by using these methods are so anomalous that a breed of 'forensic' statisticians[4,5], who doggedly detect and correct other HDB investigators' prominent mistakes, has been created.

Here we offer a 'meta-methodology' and framework in which to evaluate epistemological foundations of proposed statistical methods. On the basis of this framework, we consider that many statistical methods offered to the HDB community do not have an adequate epistemological foundation. We hope the framework will help methodologists to develop robust methods and help applied investigators to evaluate whether statistical methods are valid.

Our vantage point: from samples to populations

We study samples and data to understand populations and nature. From this perspective (**Table 1**), the sampling units are cases (*e.g.*, mice) and not genes. Although this may seem obvious, methods in which inferences about differences in gene expression between populations are made by comparing observed sample differences with an estimated null distribution of differences based on technical rather than biological replicates have been proposed[6]. Measurement error should not be confused with true biological variability among cases in a population. This conflates the standard error of measurement with the standard error of the sample statistic; it takes observations from Level I (**Table 1**), makes an inference to Level II and conflates this inference with the desired inference to Level III. This is one example of a common class of mistakes that can be avoided by considering the sample-to-population perspective.

What is validity?

Assessing validity requires explicit standards for evaluating methods. This requires an explanation of what a method is supposed to do or what properties it is supposed to have. A full description of various properties that a statistical procedure should have is beyond our scope. There is inherent subjectivity in choosing which properties are of interest or desired, but once criteria are chosen, methods can and should be evaluated objectively. Validity can be relative and situation-specific. This is noteworthy in considering the merit of a newly proposed procedure when one or more procedures already exist for similar purposes. In such cases, it may be important to ask not only whether the new method is valid in an absolute sense, but whether and under what circumstances it confers any relative advantage with respect to the chosen properties. **Table 2** outlines four common statistical activities in HDB, how validity might be defined in each and special issues with their application to HDB.

The search for proof: deduction

A proof is a logical argument proceeding from axioms to eventual conclusion through an ordered deductive process. Its certainty stems from the deductive nature by which each step follows from an earlier step. As things proven and methods of their proof have become more complex, certainty is not always easy to achieve and what is obvious to one person may not be to another[7]. The key structure that we should seek in a proof that a method has a certain property has three parts: precise formulation of axioms, statement of the method's purported property and logical steps connecting the two.

[1]Department of Biostatistics, Section on Statistical Genetics, Ryals Public Health Building, Suite 327; [2]Department of Electrical and Computer Engineering; [3]Clinical Nutrition Research Center; University of Alabama at Birmingham, 1665 University Boulevard, Birmingham, Alabama 35294, USA.
Correspondence should be addressed to D.B.A. (dallison@uab.edu).

Published online 30 August 2004; doi:10.1038/ng1422

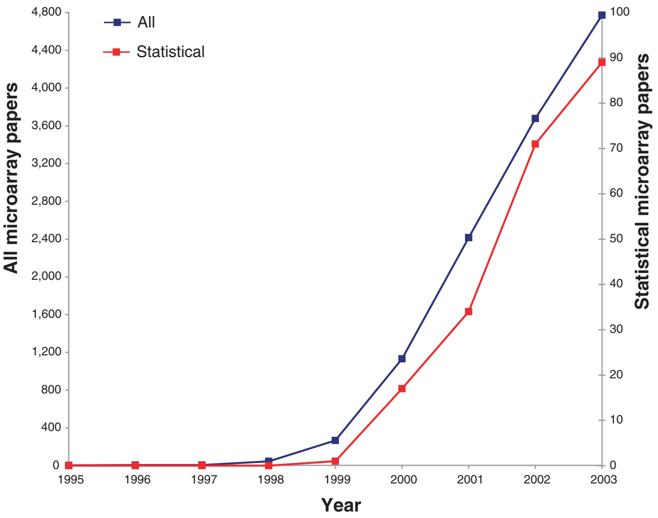

Figure 1 Growth of microarray and microarray methodology literature listed in PubMed from 1995 to 2003. The category 'all microarray papers' includes those found by searching PubMed for microarray* OR 'gene expression profiling'. The category 'statistical microarray papers' includes those found by searching PubMed for 'statistical method*' OR 'statistical techniq*' OR 'statistical approach*" AND microarray* OR 'gene expression profiling'.

Proofs begin with axioms or postulates (*i.e.*, assumptions) and are valid only when the assumptions hold. The proof's practical conclusions may hold across broader circumstances, but additional evidence is required to support this. Therefore, it is important to state and appreciate the assumptions underlying any method's validity. This allows assessment of whether those assumptions are plausible and, if not, what the effect of violations might be.

Many methods assume that residuals from some fitted model are normally distributed. It is unclear, however, whether transcriptomic or proteomic data are normally distributed even after the familiar log transformation. For least squares–based procedures, the central limit theorem guarantees robustness with large sample sizes. But HDB sample sizes are typically small. Some analyses allow the enormous numbers of measurements to compensate for the few cases[8], but the extent to which such procedures compensate for robustness to departures from distributional assumptions is unclear.

An equally important Gauss-Markov assumption[9], homoscedasticity (homogeneity of variance), is crucial for most least squares–based tests. Violation can greatly affect power and type 1 error levels. Here, it is important to highlight a common misconception about nonparametric statistics. Nonparametric statistics, including permutation tests, are distribution-free. Their validity does not depend on any particular data distribution. But distribution-free is not assumption-free. Many HDB methodologists use nonparametric, particularly permutation or bootstrap, testing as though it eliminates all assumptions and is universally valid. This is not so[10,11]. For example, conventional permutation tests assume homoscedasticity and can be invalidated by outliers[10]. Moreover, conducting inference for one's method by permutation, even if this yields correct type 1 error rates, may not be optimal for all purposes. For example, in some transcriptomic studies, investigators may primarily wish to rank genes by their 'importance' or the magnitude of their effect. In such cases, permutation tests may yield valid type 1 error rates but may be outperformed by parametric tests in terms of ranking genes by magnitude of effect[12].

Another common assumption about statistical techniques is that certain elements of the data are independent[9], and violations can markedly invalidate tests. This includes permutation and bootstrap tests, unless the dependency is built into the resampling process, as some have done[13]. Thus, we should ask whether dependency is accommodated in our methods. A popular approach in microarray data is to calculate a test statistic for each gene and then permute the data multiple times, each time recalculating and recording the test statistics, thereby creating a pseudonull distribution against which observed test statistics can be compared for statistical significance. If one uses only the distribution of test statistics in each gene, then, given the typically small samples, there are insufficient possible permutations and the distribution is coarse and minimally useful[14,15]. Some investigators[16] pool the permutation-based test statistics across all genes to create a pseudonull distribution with lesser coarseness. But this approach treats all genes as independent, which is not the case. Therefore, *P* values derived from such permutations may not be strictly valid[17].

Statements about proposed approaches can be supported by referring to proofs already published. For example, those proposing a particular mixed model approach[18] correctly realized that they did not need to prove that (under certain conditions) this model is asymptotically valid for frequentist testing, because this has already been shown. They needed only to cite those references. Recognizing the limits of what has been previously shown is important, and mixed models exemplify an acute concern in HDB. Certain mixed model tests are asymptotically valid but can be invalid under some circumstances with samples as small as 20 per group[19], far larger than those typically used in HDB. Thus, validating methods with small samples when their validity relies on asymptotic approximations is vital.

Finally, we note that mathematical description of some process is not equivalent to proof that the result of the process has any particular properties. Methodological papers in HDB often present new algorithms with exquisite mathematical precision. Those who are less comfortable with mathematics may mistake this for proof. Writing an equation may define something, but it does not prove anything.

The proof of the pudding is in the eating: induction

In induction, there is no proof that a method has certain properties. Instead we rely on extra-logical information[20,21]. If a method performs in a particular manner across many instances, we assume it will probably do so in the future. We therefore seek to implement methods in situations that can provide feedback about their performance[22]. Simulation and plasmode studies (below) are two such methods.

Many methodologists use simulation to examine methods for HDB[8,14]. Because the data are simulated, one knows the right answers and can unequivocally evaluate the correspondence between the underlying 'truth' and estimates, conclusions or predictions derived with the method. Moreover, once a simulation is programmed, one can generate and analyze many data sets and, thereby, observe expected performance across many studies. Furthermore, one can manipulate many factors in the experiment (*e.g.*, sample size, measurement reliability, effect magnitude) and observe performance as a function. There are two key challenges to HDB simulation: computational demand and representativeness.

Regarding computational demand, consider that we need to analyze many variables (*e.g.*, genes) and may use permutation tests that necessitate repeating analyses many times per data set. This demand is compounded when we assess method performance across many conditions and wish to work at α levels around 10^{-4} or less, necessitating on the order of 10^6 simulations per condition to accurately estimate (*i.e.*, with 95% confidence to be within 20% of the expected

Table 1 Iconic representation of levels of observation and inference

	Level I	Level II	Level III
Aspects of variables studied	Measurements of specific variables on cases	True values of specific variables on cases	Population distribution of variables
Units of observation	Cases/samples	Cases/samples	Population/universe
Subject of study	Data	Data	Nature
Conceptual description	Second-order shadows	Shadows	Reality

We observe imperfect measurements of variables from specific objects drawn from larger populations, and these observations form our base data. Using these data, we wish to make inferences from these imperfect measurements to the real variables they represent and then from the sample cases to the population they were sampled from. Borrowing the language of Plato's *Allegory of the Cave*, we can view the data at level II as manifestations of a random sampling process and 'shadows' of the population-level reality we are trying to model. Similarly, we can view the data at level III as manifestations of our inevitably error-prone measurement processes and 'second-order shadows' of the actual sample values. An unfortunately common mistake in HDB is to confuse the variability among measurements of the same object with variability among objects sampled from a population and, thereby, mistake a procedure that can make valid inferences from level I to level II for one that makes inferences to level III.

value) type 1 error rates. Simulating at such low α levels is important, because a method based on asymptotic approximations may perform well at higher α levels but have inflated type 1 error rates at lower α levels. In such situations, even a quick analysis for an individual variable becomes a computational behemoth at the level of the simulation study. Good programming, ever-increasing computational power and advances in simulation methodology (*e.g.*, importance sampling)[23] are, therefore, essential.

The second challenge entails simulating data that reasonably represent actual HDB data, despite limited knowledge about the distribution of individual mRNA or protein levels and the transcriptome- or proteome-wide covariance structure. Consequently, some investigators believe that HDB simulation studies are not worthwhile. This extreme and dismissive skepticism is ill-founded.

First, although we have limited knowledge of the key variables' distributions, this is not unique to HDB[24], and we can learn about such distributions by observing real data. We rarely know unequivocally the distribution of biological variables, yet we are able to develop and evaluate statistical tests for these. One can simulate data from an extraordinarily broad variety of distributions[25]. If tests perform well across this variety, we can be relatively confident of their validity. Moreover, if we identify specific 'pathological' distributions for which our statistical procedures perform poorly, then by using them in practice, we can attempt to ascertain whether the data have such distributions.

Table 2 Four common statistical activities in HDB and issues in the validation of their methods

Statistical activity	Description	Criteria for evaluating validity	Comments	Special issues in HDB
Inference	The process of making decisions or drawing conclusions about the truth or falsity of hypotheses by applying frequentist, Bayesian or other paradigms[43,44].	Family-wise error rate control; false discovery rate control; proportion of false positive control; power (type II error minimization); total error minimization[45,46].	Distributional and other assumptions about the nature of the data are usually paramount.	Small sample sizes offer little power, especially if corrected for multiple testing. Greater power and flexibility might be obtained by borrowing information across genes[15], but doing so often makes assumptions about exchangeability across or independence of genes that may not be explicit or valid[17].
Estimation	Computing sample statistics (*e.g.*, relative change in gene expression) on observed data as estimates of corresponding unobserved population parameters (*e.g.*, the relative change in gene expression in the population), called estimands.	Unbiasedness; minimum mean square error; efficiency; sufficiency; consistency.	HDB offers especially rich ground for modern 'information-borrowing' techniques that can radically improve estimation when many estimates are simultaneously made on small samples[15,44].	Estimators that are ordinarily unbiased can be markedly biased when used only to estimate significant effects in genome-wide contexts[47].
Prediction (unsupervised classification)	In HDB, typically entails predicting status of some categorical variable (*e.g.*, malignant versus benign) on the basis of observed variables (*e.g.*, proteomic measurements).	Minimizing and accurately estimating expected prediction error probabilities; positive predictive value; area under receiver operating curve.	The predictive validity of a prediction rule developed in a single data set should not be confused with the validity of a general method for developing prediction rules on data sets.	The seemingly simple problem of estimating predictive accuracy on a single data set is not so simple in HDB. HDB methodologists have markedly overestimated predictive accuracy by using invalid methods for estimating predictive accuracy of derived predictive rules[4].
Classification (supervised classification)	The construction of classification schemes and assignment of objects to classes within schemes. In transcriptomics, classification often occurs through some form of cluster analysis and is applied to the genes (variables) as opposed to the cases.	The extent to which they yield classifications that are replicable across multiple samplings from a population beyond chance levels[48]. Although a replicable classification is not necessarily useful, a useful classification that characterizes some aspect of the population must be replicable[35].	Classifications do not exist; we create them. Thus, there is no null hypothesis to test, no independent reality to compare with a derived classification —no right answer. Some[49] disagree on this point and state that "the null hypothesis that is being tested here is that of no structure in the data." Exactly what 'no structure' means is not clear, nor can it be taken to be equivalent to 'no classification.'	Some authors have, in our opinion, mistakenly resampled across genes when trying to assess stability of cluster solutions [48], which makes little sense from the samples-to-population perspective shown in **Table1**.

Regarding correlation among genes, it is easy to simulate a few, even non-normal, correlated variables[26]. In HDB, the challenge is simulating many correlated variables. Using block diagonal correlation matrices[14] oversimplifies the situation. 'Random' correlation matrices[27] are unlikely to reflect reality. Alternatively, one can use real data to identify a correlation structure from which to simulate. This can be done by using the observed expression values and simulating other values (*e.g.*, group assignments, quantitative outcomes) in hypothetical experiments or by generating simulated expression values from a correlation matrix that is based in some way on the observed matrix[28] using factoring procedures. Exactly how to do this remains to be elucidated, but the challenge seems to be surmountable. Investigators are addressing this challenge[29–31], and several microarray data simulators exist (refs. 32–34 and the gene expression data simulator at http://bioinformatics.upmc.edu/GE2/index.html).

Another challenge in simulation is to make the covariance structure 'gridable', meaning that the theoretically possible space of a parameter set can be divided into a reasonably small set of mutually exclusive and exhaustive adjacent regions. Typically, simulation is used when we are unable to derive a method's properties analytically. Therefore, it is usually desirable to evaluate performance across the plausible range of a key factor. If that factor is the correlation between two variables, one can easily simulate along the possible range (–1,1) at suitably small adjacent intervals (a grid). With multiple variables under study, the infinite number of possible correlation matrices is not obviously represented by a simple continuum, and it is not obvious how to establish a reasonably sized grid. But if one could extract the important information from a matrix in a few summary metrics, such as some function of eigenvalues, it might be possible to reduce the dimensionality of the problem and make it 'gridable'. This is an important topic for future research.

A plasmode is a real data set whose true structure is known[35]. As in simulations, the right answer is known *a priori*, allowing the inductive process to proceed. Plasmodes may represent actual experimental data sets better than simulations do. In transcriptomics, the most common type of plasmode is the 'spike-in' study. For example, real cases from one population are randomly assigned to two groups and then known quantities of mRNA for specific genes (different known quantities for each group) are added to the mRNA samples. In this situation, the null hypothesis of no differential expression is known to be true for all genes except those that were spiked, and the null hypothesis is known to be false for all those that are spiked. One can then evaluate a method's ability to recover the truth.

Plasmode studies have great merit and are being used[15,36], but there is a need for greater plurality. Because statistical science deals with random variables, we cannot be certain that a method's performance in one data set will carry over to another. We can only make statements about expected performance, and estimating expected or average performance well requires multiple realizations. Analysis of a single plasmode is minimally compelling. Because plasmode creation can be expensive and laborious, it is difficult for investigators to create many. Additionally, although plasmodes might offer better representations of experimental data sets, there is no guarantee. For example, in spike-in studies, it is unclear how many genes should be spiked or what the distribution of spike-created effects should be to reflect reality.

Combined modes

One can also combine the approaches above[15]. When two or more modes yield consistent conclusions, confidence is strengthened. One could also creatively combine deduction and induction. For example, suppose there were two alternative inferential tests, A and B, which could be proven deductively to have the correct type 1 error rate under the circumstances of interest. If one applied the tests to multiple real data sets and consistently found that test A rejected more null hypotheses than did test B, one could reasonably conclude that test A was more powerful than test B. This makes sense only if both tests have correct type 1 error rates.

Data sets of unknown nature: circular reasoning

Authors often purport to demonstrate a new method's validity in HDB by applying it to one real data set of unknown nature. A new method is applied to a data set, and a new interesting finding is reported; for example, a gene previously not known to be involved in disease X is found to be related to the disease, and the authors believe that the finding shows their method's value. The catch is this: if the gene was previously not known to be involved in disease X, how do the authors know that they got the right answer? If they do not know that the answer is right, how do they know that this validates their method? If they do not know that their method is valid, how do they know that they got the right answer? We are in a loop (circular argument). Illustration of a method's use is not demonstration of its value. Illustration with single data sets of unknown nature, though interesting, is not a sound epistemological foundation for method development.

Where to from here?

We offer four suggestions for progress:

(i) Vigorous solicitation of rigorous substantiation. Guidelines have been offered or requested for genome scan inference[37], transcriptomic data storage[38], specimen preparation and data collection[39], and result confirmation[40]. We agree that these should remain guidelines and not rules[41]. Such guidelines help evaluate evidential strength of claims. But there are no guidelines for presentation and evaluation of methodological developments[22]. Thus, we offer the guidelines in **Box 1** to be used in evaluating proffered methods.

(ii) 'Meta-methods'. For methodologists to strive for high standards of rigor, they must have the tools to do so. An important area for new research is HDB 'meta-methodology', methodological research about how to do methodological research. Such second-order methodological research could address how to simulate realistic data and how to meet computational demands. Public plasmode database archives would also be valuable.

(iii) Qualified claims? A risk in requesting more rigorous evidential support for new HDB statistical techniques is that if such requests became inflexible demands, progress might be slowed. 'Omic' sciences

Box 1 Suggested guidelines for promoting a sound epistemological foundation for new statistical methodology in HDB

1. State exactly what the method is intended to do or what properties it is intended to have in objectively testable terms.
2. State the assumptions under which these properties or expected outcomes should occur.
3. Provide evidence that the method has the claimed performance or properties from simulation studies, analytic proofs and/or multiple plasmode data analyses.
4. In the absence of compelling evidence as described in point 3, state clearly that the claimed properties are conjectured and await substantiation.
5. Where an alternative method already exists, compare the properties of the new method with those of the existing method, or, at minimum, note that an alternative exists, conjecture why the new method may be superior in some situations and suggest future testing of the conjecture.

move fast, and investigators need new methodology. Therefore, although we hope methodologists publish new methods with the most rigorous validation possible, public scientific conjecture has an illustrious history, and it is in the interests of scientific progress and intellectual freedom that compelling methods, though merely conjectured to be useful, be published. But, as Bernoulli wrote, “In our judgments we must beware lest we attribute to things more than is fitting to attribute…and lest we foist this more probable thing upon other people as something absolutely certain”[42]. Thus, it is reasonable to publish methods without complete evidence regarding their properties, provided we follow Bernoulli: state the claims we are making for our proffered methods and whether such claims are supported by simulations, proofs, plasmode analyses or merely conjecture.

(iv) *Caveat emptor*. Ultimately, we offer the ancient wisdom, “*caveat emptor*”. Statistical methods are, by definition, probabilistic, and in using them, we will err at times. But we should have the opportunity to proceed knowing how error-prone we will be, and we appeal to methodologists to provide that knowledge.

ACKNOWLEDGEMENTS
We thank J. Dickson for editorial assistance and S. Vollmer, S. Barnes and E. K. Allison for discussions. This work was supported by grants from the US National Institutes of Health and from the US National Science Foundation.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 5 August; accepted 9 August 2004
Published online at http://www.nature.com/naturegenetics/

1. Evans, G.A. Designer science and the “omic” revolution. *Nat. Biotechnol.* **18**, 127 (2000).
2. Gracey, A.Y. & Cossins, A.R. Application of microarray technology in environmental and comparative physiology. *Annu. Rev. Physiol.* **65**, 231–259 (2003).
3. Tilstone, C. DNA microarrays: vital statistics. *Nature* **424**, 610–612 (2003).
4. Ambroise, C. & McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **99**, 6562–6566 (2002).
5. Baggerly, K.A. *et al.* A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3**, 1667–1672 (2003).
6. Toda, K. *et al.* Test of significant differences with a priori probability in microarray experiments. *Anal. Sci.* **19**, 1529–1535 (2003).
7. Lakatos, I. Proofs and refutations: I. *Br. J. Philos. Sci.* **14**, 1–25 (1963).
8. Baldi, P. & Long, A.D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519 (2001).
9. Berry, W.D. A formal presentation of the regression assumptions. in *Understanding Regression Assumptions* (ed. Lewis-Beck, M.S.) 3–11 (Sage University Publications, Thousand Oaks, 1993).
10. Roy, T. The effect of heteroscedasticity and outliers on the permutation t-test. *J. Stat. Comput. Simul.* **72**, 23–26 (2002).
11. Hall, P. & Wilson, S.R. Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762 (1991).
12. Xu, R.H. & Li, X.C. A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics* **19**, 1284–1289 (2003).
13. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375 (2003).
14. Gadbury, G.L., Page, G.P., Heo, M., Mountz, J.D. & Allison, D.B. Randomization tests for small samples: an application for genetic expression data. *J. R. Stat. Soc. Ser. C Appl. Stat.* **52**, 365–376 (2003).
15. Newton, M.A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176 (2004).
16. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
17. Kowalski, J., Drake, C., Schwartz, R.H. & Powell, J. Non-parametric, hypothesis-based analysis of microarrays for comparison of several phenotypes. *Bioinformatics* **20**, 364–373 (2004).
18. Wolfinger, R.D. *et al.* Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **8**, 625–637 (2001).
19. Catellier, D.J. & Muller, K.E. Tests for Gaussian repeated measures with missing data in small samples. *Stat. Med.* **19**, 1101–1114 (2000).
20. Russell, B. On induction. in *Basic Writings* 149–155 (Touchstone-Simon and Schuster, London, 1961).
21. Ertas, A., Maxwell, T., Rainey, V. & Tanik, M.M. Transformation of higher education: the transdisciplinary approach in engineering. *IEEE Trans. Education* **46**, 289–295 (2003).
22. Spence, M.A., Greenberg, D.A., Hodge, S.E. & Vieland, V.J. The emperor’s new methods. *Am. J. Hum. Genet.* **72**, 1084–1087 (2003).
23. Malley, J.D., Naiman, D.Q. & Bailey-Wilson, J.E. A comprehensive method for genome scans. *Hum. Hered.* **54**, 174–185 (2002).
24. Miccerri, T. The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* **105**, 156–166 (1989).
25. Karian, Z.A. & Dudewicz, E.J. Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods 1–38 (CRC, New York, 2000).
26. Headrick, T.C. & Sawilowsky, S.S. Simulating correlated multivariate non-normal distributions – Extending the Fleishman power method. *Psychometrika* **64**, 25–35 (1999).
27. Davies, P.I. & Higham, N.J. Numerically stable generation of correlation matrices and their factors. *BIT Num. Math.* **40**, 640–651 (2000).
28. Cherepinsky, V., Feng, J., Rejali, M. & Mishra, B. Shrinkage-based similarity metric for cluster analysis of microarray data. *Proc. Natl. Acad. Sci. USA* **100**, 9668–9673 (2003).
29. Bailey, L.R. & Moore, J.H. Simulation of gene expression patterns in cDNA microarray data. *Am. J. Hum. Genet.* **65**, 473 (1999).
30. Balagurunathan, Y., Dougherty, E.R., Che,n Y., Bittner, M.L. & Trent, J.M. Simulation of cDNA microarrays via a parameterized random signal model. *J. Biomed. Opt.* **7**, 507–523 (2002).
31. Perez-Enciso, M., Toro, M.A., Tenenhaus, M. & Gianola, D. Combining gene expression and molecular marker information for mapping complex trait genes: A simulation study. *Genetics* **164**, 1597–1606 (2003).
32. Mendes, P., Sha, W. & Ye, K. Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics* **19**, II122–II129 (2003).
33. Michaud, D.J., Marsh, A.G. & Dhurjati, P.S. eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods. *Bioinformatics* **19**, 1140–1146 (2003).
34. Singhal, S. *et al.* Microarray data simulator for improved selection of differentially expressed genes. *Cancer. Biol. Ther.* **2**, 383–391 (2003).
35. Blashfield, R.K. & Aldenderfer, M.S. The methods and problems of cluster analysis. in *Handbook of Multivariate Experimental Psychology* 2nd edn. (eds. Nesselroade, J.R., & Cattell, R.B.) 447–473 (Plenum, New York, 1988).
36. Irizarry, R.A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15 (2003).
37. Lander, E. & Kruglyak, L. Genetic dissection of complex traits - guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).
38. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
39. Benes, V. & Muckenthaler, M. Standardization of protocols in cDNA microarray analysis. *Trends Biochem. Sci.* **28**, 244–249 (2003).
40. Rockett, J.C. & Hellmann, G.M. Confirming microarray data-is it really necessary? *Genomics.* **83**, 541–549 (2004).
41. Witte, J.S., Elston, R.C. & Schork, N.J. Genetic dissection of complex traits. *Nat. Genet.* **12**, 355–356 (1996).
42. Bernoulli, J. *Ars Conjectandi* (1713).
43. Edwards, A.W. Statistical methods in scientific inference. *Nature* **222**, 1233–1237 (1969).
44. Yang, D. *et al.* Applications of Bayesian statistical methods in microarray data analysis. *Am. J. Pharmacogenomics* **4**, 53–62 (2004).
45. Gadbury, G.L. *et al.* Power and sample size estimation in high dimensional biology. *Stat. Methods Med. Res.* (in the press).
46. van den Oord, E.J. & Sullivan, P.F. False discoveries and models for gene discovery. *Trends Genet.* **19**, 537–542 (2003).
47. Allison, D.B. *et al.* Bias in estimates of quantitative-trait-locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *Am. J. Hum. Genet.* **70**, 575–585 (2002).
48. Famili, A.F., Liu, G. & Liu, Z. Evaluation and optimization of clustering in gene expression data analysis. *Bioinformatics* **20**, 1535–1545 (2004).
49. Smolkin, M. & Ghosh, D. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics* **4**, 36 (2003).