Analyzing ’omics data using hierarchical models

Hongkai Ji & X Shirley Liu

Hierarchical models provide reliable statistical estimates for data sets from high-throughput experiments where measurements vastly outnumber experimental samples.

Interpreting ’omics data often involves statistical analysis of large numbers of loci such as genes, binding sites or single-nucleotide polymorphisms (SNPs). Although the data set as a whole may be rich in information, each individual locus is typically only associated with a limited amount of data. Statistical inference in this context is challenging. A hierarchical model is a useful statistical tool to more efficiently analyze the data, and it is increasingly being used in computational genomics.

A motivating example

Consider a hypothetical microarray experiment with ten genes. For each gene, $\log_2$ expression fold-changes (hereafter referred to as simply ‘expression’) are observed between tumor and normal tissues in three biological replicates (**Table 1**). To select a gene for follow-up study that is differentially expressed in tumor compared with normal cells, which gene should be the top candidate?

A simple solution is to rank the genes by *t*-statistics

$$t_i = \bar{x}_i / \sqrt{s_i^2 / n}$$

Here n (= 3) is the number of replicates, $\bar{x}_i$ is the average expression of gene i, and s_i^2 is the sample variance. Based on the absolute values of *t*-statistics, gene 2 is the top candidate (**Table 1**).

The data in this example, however, are simulated, with each gene having a ‘true’ expression μ_i whose measurement is confounded by experimental or biological variability represented by the parameter σ_i^2. (In fact, each expression measurement was randomly drawn from a bell curve–shaped normal distribution with a mean μ_i and variance σ_i^2). The true values of μ_i and σ_i^2, which are unknown to you, are shown in **Table 1**. It turns out the only truly differentially expressed gene is gene 10, which has a nonzero μ_i. Gene 2 thereby represents a false-positive call.

What causes this mistake? Small sample size and the multiplicity of the problem are the reasons. To understand why, it may be helpful to briefly review the key ideas behind statistical inference. The first concept to understand is that of the ‘distribution’. Briefly, in the presence of biological or experimental noise and variability, repeated biological measurements are unlikely to be identical, giving rise to a collection, or distribution, of data values. This distribution can be characterized by parameters, such as its mean (or average value) and variance (which quantifies how far the measurements are expected to be from the mean). The parameters are properties associated with infinitely many measurements. In a real scenario, when only a finite number of measurements are available, the true parameter value cannot be obtained. Statistical inference seeks to make statements about the true, also referred to as ‘unobserved’, parameter value based on the observed data which are called by statisticians as ‘samples’ drawn from the distribution.

In a *t*-statistic, the sample mean $\bar{x}_i$ represents an estimate of the true mean μ_i of the distribution from which gene *i*’s data are sampled, and the sample variance s_i^2 represents an estimate of the true variance σ_i^2. If the true mean is zero (that is, gene *i* is not differentially expressed), it is unlikely to obtain a *t*-statistic with a large magnitude.

When the sample size is small, however, the observed sample variance is an unreliable estimate of the true variance of the system. To see why, imagine randomly selecting three data points from a normal distribution with mean 0 and variance 1, which results in the values 0.1, 0.09 and 0.11 (**Fig. 1a**, blue dots). As a result, the observed variance is 0.0001 (or approximately 0) even though the true variance is 1 (that is, much bigger than 0). Another random draw of three data points from the same distribution may give you –1.1, –0.2 and 0.7 (**Fig. 1a**, orange dots) and a totally different observed variance of 0.81. Although the probability that the observed variance significantly deviates from the true variance is small for each individual gene, in a genomic study with many genes, the chance to encounter such deviants for some genes is high.

Small sample variances obtained by chance give rise to large *t*-statistics, which can incorrectly rank nondifferentially expressed genes at the top. This is what happened in our example. The true variance of gene 2 is 1, but the sample variance is 0.005 (**Table 1**); as a result, the *t*-statistic incorrectly ranked gene 2 ($t_2 = 17.5$) on top of the truly differentially expressed gene 10 ($t_{10} = 3.42$). In general, when data analysis involves estimating many parameters or testing many hypotheses but the sample size is small, it is difficult to reliably estimate all parameters or to make correct decisions for all tests simultaneously. This problem is less serious if more samples are available, as more reliable estimates of parameters can be obtained for each gene.

Real gene expression microarray experiments with tens of thousands of genes are examples

Hongkai Ji is in the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA, and X. Shirley Liu is in the Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts, USA. e-mail: hji@jhsph.edu or xsliu@jimmy.harvard.edu

Table 1 Statistical analysis of example data using either *t*-statistics or a hierarchical model

	Gene, *i*	1	2	3	4	5	6	7	8	9	10
Unobserved parameters	Mean, μ_i	0	0	0	0	0	0	0	0	0	2
	Variance, σ_i^2	2.0	1.0	1.5	0.5	0.7	1.1	1.3	0.9	1.2	1.0
Observed expression data ($\log_2$ fold change)	x_{i1}	0.97	0.73	0.63	1.20	–0.57	3.68	–0.45	1.14	0.34	1.31
	x_{i2}	–0.47	0.78	–0.41	1.48	0.33	–0.68	–0.06	0.40	–0.08	2.59
	x_{i3}	–0.19	0.64	1.93	–0.02	0.26	2.08	–0.74	0.30	1.74	1.03
Gene selection by *t*-statistics	Mean, $\bar{x}_i$	0.10	0.72	0.72	0.89	0.01	1.69	–0.42	0.61	0.67	1.64
	Sample variance s_i^2	0.58	0.005	1.37	0.64	0.25	4.86	0.12	0.21	0.91	0.69
	t_i	0.23	***17.50***	1.06	1.93	0.02	1.33	–2.12	2.32	1.21	3.42
Gene selection by hierarchical model	Adjusted variance, $\hat{\sigma}_i^2$	0.84	0.65	1.09	0.85	0.73	2.22	0.69	0.72	0.95	0.88
	New t_i	0.20	1.53	1.19	1.66	0.01	1.97	–0.87	1.25	1.19	***3.04***

[a]Numerals in bold italics indicate the gene for which the absolute value of the *t*-statistic (t_i) is the largest.

of a 'large *p*, small *n*' problem, where *p* refers to the number of genes and *n* refers to the number of samples. In addition to the multiplicity issue mentioned before, another potential problem is that if the data are not normally distributed, applying a *t*-test can be invalid when the sample size is small[1]. However, this problem is not the focus of the current primer, in which the data in our example are assumed to be normally distributed.

What is a hierarchical model?

One statistical tool for handling large-*p*, small-*n* problems is a hierarchical model. Such a model describes hierarchical relationships between various sources of data variation. The model structure effectively makes it possible to 'borrow' information from all genes to make more reliable statistical inferences about a particular gene. Hierarchical models are conceptually related to regularization techniques, which include Lasso and ridge regression and represent a broad class of methods for handling large-*p*, small-*n* problems (reviewed in refs. 2,3).

In our example, a hierarchical model can be built by assuming that the unobserved mean and variance parameters (that is, μ_i and σ_i^2) of different genes are also sampled from a distribution (denoted as F_0). The distribution is characterized by parameters, such as mean and variance of infinitely many μ_i and σ_i^2 hypothetically collected from different genes. Accordingly, one can imagine that the observed expression data are generated hierarchically by first drawing the mean and variance parameters for each gene from F_0, and then drawing expression measurements for each gene from a gene-specific distribution (that is, a normal distribution with mean μ_i and variance σ_i^2) (**Fig. 1b**).

Naturally, this model describes two sources of variation in the observed expression data. At the top of the hierarchy, the intrinsic similarities and differences between the expression of different genes is mathematically modeled using a distribution (that is, F_0) of the unobserved gene-specific parameters. At the bottom of the hierarchy, the cross-sample variability within a single gene is modeled using a gene-specific distribution with parameters generated from the top-level distribution (**Fig. 1c**). In effect, the top-level distribution describes which gene-specific parameter values are common and which are unusual. The data contain information about the distributions at both levels because there are several replicates for each gene over many different genes.

Although the top-level distribution is usually unknown, it can be estimated using data from the thousands of genes available. Then, using this distribution, the hierarchical model allows one to 'borrow' information across genes to facilitate inference. How much information to borrow is determined by how similar the genes are relative to the cross-sample variability. The intuition is that if the heterogeneity across genes is small, then data from all genes could be informative about the parameters of a particular gene (**Fig. 1c**). Borrowing information across genes essentially increases the effective sample size for making inferences about individual genes[4]. In contrast, the *t*-statistic approach only uses information from a single gene to estimate the mean and variance of the bottom-level distribution for that gene.

Inference using the hierarchical model

The first step in using the hierarchical model is to find a top-level distribution that fits the data (**Fig. 1d**). This process can be intuitively interpreted as learning the cross-gene heterogeneity from the data. The top-level distribution is usually assumed to be a member of a broad family of distributions. In other words, a large number of candidate distributions with the same mathematical form but different parameter values are considered. By varying the parameter values, members in the family are able to describe a variety of distribution patterns of the gene-specific parameters. The analysis starts by finding the distribution (through identifying the parameter value) in the family that fits the data well, and then using the identified distribution to help infer the gene-specific parameters. Commonly used top-level distribution families include 'conjugate priors' and mixtures of simple distributions (e.g., mixture of normal distributions)[5]. The former is typically used if developing a simple computational algorithm is required, and the latter is used if one needs flexibility to describe very complex cross-gene variation patterns.

Next, the top-level distribution is used to adjust the parameter estimate of every gene (**Fig. 1e**). If cross-gene heterogeneity is small, the adjustment will make the parameter estimates of different genes more similar to each other. Here, the hierarchical model borrows from Bayesian inference, a general approach to make statistical inference by combining prior information with observed data[5,6], with the top-level distribution being treated as the prior knowledge about the unobserved mean and variance parameters of individual genes.

Algorithmically, finding the top-level distribution and inferring gene-specific parameters can be implemented simultaneously using standard Bayesian or empirical Bayes techniques[5,6], which sometimes requires advanced and computation-intensive techniques such as Markov chain Monte Carlo[5].

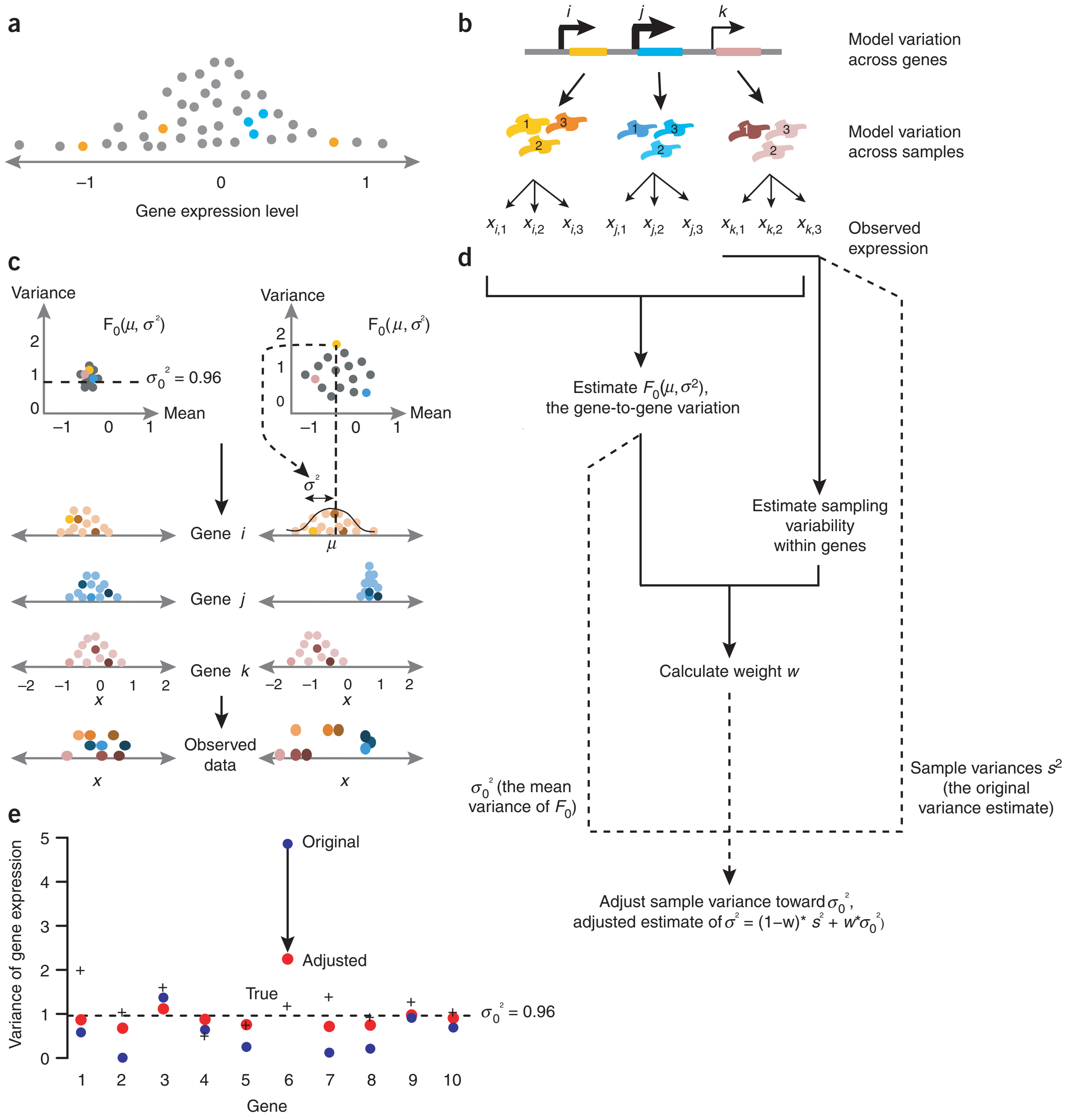
a
–1
0
1
Gene expression level
b
i
j
k
Model variation across genes
1
2
3
Model variation across samples
$x_{i,1}$ $x_{i,2}$ $x_{i,3}$ $x_{j,1}$ $x_{j,2}$ $x_{j,3}$ $x_{k,1}$ $x_{k,2}$ $x_{k,3}$
Observed expression
c
Variance
$F_0(\mu, \sigma^2)$
2
1
0
σ_0^2 = 0.96
Mean
–1 0 1
σ^2
μ
Gene i
Gene j
Gene k
–2 –1 0 1 2
x
Observed data
d
Estimate $F_0(\mu,\sigma^2)$, the gene-to-gene variation
Estimate sampling variability within genes
Calculate weight w
σ_0^2 (the mean variance of F_0)
Sample variances s^2 (the original variance estimate)
Adjust sample variance toward σ_0^2, adjusted estimate of σ^2 = (1–w)* s^2 + $w^*\sigma_0^2$)
e
Variance of gene expression
5
4
3
2
1
0
Original
Adjusted
True
σ_0^2 = 0.96
1 2 3 4 5 6 7 8 9 10
Gene

Figure 1 Hierarchical modeling. (**a**) Many analysis techniques, such as *t*-statistics, consider each gene separately. Owing to different sources of biological and experimental variation, if triplicate measurements of the expression of the same gene are collected twice (blue dots and orange dots), the measurements may yield different estimates of the mean and variance of the true distribution that describes the gene's expression (gray). (**b**) A hierarchical model helps produce more reliable estimates of the mean and variance by considering all genes together. It models different sources of biological variation hierarchically. A top-level distribution (F_0) models variation across genes and a bottom-level distribution models variation of the same gene between samples. Data are described by first drawing μ and σ^2 from F_0 for each gene and then drawing expression fold-changes for each gene. (**c**) If different genes have similar mean and variance, data from one gene are informative about the mean and variance of another gene. It is not known a priori whether genes are similar (left, F_0 is tightly clustered) or not (right, F_0 is more spread out). However, this can be learned by looking at the data of many different genes. If genes are similar, the observed gene-to-gene differences can be largely explained by the sampling variability within a gene (bottom, left); otherwise genes are heterogeneous (bottom, right). (**d**) The hierarchical model is applied by first using the observed data to estimate cross-gene variation (that is, F_0), then comparing it to within-gene sampling variability to determine a rule to combine the characteristics shared by all genes with the data specific to each gene for estimating μ and σ^2 (solid lines). In our example, this yields an adjusted variance estimate in the form of a weighted average between the sample variance and the mean of variances σ^2 in F_0 (that is, σ_0^2) (dashed lines). The model was not applied to estimate the gene-specific mean μ. (**e**) The genes' true variances in our example are similar (as in the left side of **c**), which is perceived by the model. As a result, the adjusted variance estimates (red) are closer than the original variance estimates (blue) to the mean variance σ_0^2 (dotted line), which incorporates data from all genes. Overall, the adjusted variance estimates are also closer to the unobserved true variances listed in **Table 1** (black '+').

In our example, applying the hierarchical model yields a new estimate of the variance parameter of a gene. The new estimate of σ_i^2 is a weighted average between the sample variance s_i^2 and the estimated mean variance of all genes (that is, the mean of all variances σ^2 in the estimated F_0, also denoted as σ_0^2) (ref. 7). The sample variance is an estimate of σ_i^2 based on gene *i*'s data, and σ_0^2 represents a shared property of all genes. These two pieces of information are combined using a weight determined automatically by comparing the magnitude of cross-gene variation (with respect to σ_i^2) with that of the within-gene sampling variability (with respect to s_i^2). If the variability among genes is low relative to the sampling variability within a gene, the mean variance σ_0^2 will receive a high weight. On the other hand, if the cross-gene variation is high compared to the within-gene sampling variability, more weight will be given to s_i^2.

The new estimates shift the sample variances toward the common population mean of σ_i^2, and pulls small variances by chance away from zero. Compared with the old estimates s_i^2, the sum of squared error of the new estimates $\hat{\sigma}_i^2$ from the true values is much smaller (3.50 versus 19.46). When the sample variances s_i^2 in the *t*-statistics are replaced by the new estimates, the new *t*-statistics correctly rank gene 10 before gene 2 (**Table 1**). This weighted average technique to estimate the variance is called variance stabilization. It is widely used in analyses of gene expression microarrays[4,8] and chromatin immunoprecipitation on tiling microarrays (ChIP-chip)[7] to detect differentially expressed genes and protein-DNA binding sites, respectively. Naturally, real microarray experiments are more complicated and contain more sources of variation than our example; thus, they can benefit from more sophisticated hierarchical models that capture those types of variation.

The validity of model assumptions, such as those on the hierarchical structure and the distributions at the top and bottom levels, is crucial for the successful application of hierarchical models. When the assumptions hold true, the model brings additional power. Otherwise, the model may not use the information optimally, or may introduce bias that leads to misleading results. Therefore, it is always wise to check the model assumptions by exploring characteristics of the raw data and testing the analysis results using independent information or cross-validation[2].

Other applications

Hierarchical models can be applied to many other problems besides gene expression microarrays and ChIP-chip. For example, in genome-wide association studies, hundreds of thousands of SNPs are tested for association with a phenotype. In a simple scenario, the association can be studied in a linear regression "phenotype = $\alpha_i + \beta_i$ *genotype + noise," where a nonzero coefficient β_i (*i* indexes SNPs) indicates association. With a limited number of samples and many SNPs to evaluate, this approach often lacks the power to distinguish relevant SNPs from random associations. Because SNPs with similar characteristics, such as those that reside in genes in the same pathway or that show a similar degree of evolutionary conservation, have similar potentials to be associated with the phenotype, one can build a hierarchical model to borrow information from similar SNPs to increase the statistical power of association studies[9]. To use this information, one can assume that β_i from different SNPs follow a top-level distribution $N(\mu + \eta^* z_i, \tau^2)$, where z_i is an observed characteristic of SNP *i*, such as conservation score. Here, $\mu + \eta^* z_i$ describes the relationship between a SNP's characteristic and its potential association with the phenotype, and τ^2 describes the heterogeneity among SNPs with the same characteristic. The model can be generalized to incorporate multiple characteristics. One can use data from all SNPs to estimate this top-level distribution (that is, μ, η, τ^2), and make an inference based on new estimates of β_i that combines the top-level distribution with the SNP-specific data.

Application of hierarchical models is not limited to large-*p*-small-*n* data. The models are useful in a broad spectrum of large-*p* problems where the amount of information per locus is limited, with small sample size being a special case. For example, predicting transcription factor binding sites in DNA sequences can be viewed as a problem that probabilistically assigns a 0–1 label to each locus by matching the sequence to a motif model as opposed to a background model. If the sequences are long, there could be random matches to the motif, which leads to false-positive predictions. However, functional transcription factor binding sites tend to cluster in the genome to form regulatory modules. One can build a hierarchical model by assuming that the input sequences consist of background and modules, and the modules in turn consist of background and binding sites, hence binding sites only occur within modules; given the binding site locations, nucleotides are generated according to either the motif or background probability models. Using this hierarchical model, one can first infer the top-level module status by checking sequences from nearby genomic loci, and then combine the module status as prior and the DNA sequence at each locus to infer its binding status. The module status estimated using information across loci helps eliminate many false-positive binding site predictions. In ref. 10, it was shown that the improved estimates of binding site locations increase the power of *de novo* motif discovery.

We conclude by providing two other examples where hierarchical models might be useful yet have not been fully explored. First, if you want to estimate the fold enrichment at ChIP-seq binding loci, but each ChIP and control library has only one replicate sequenced not so deeply, you may estimate a more robust background read count at one locus by borrowing information from other loci. Second, if you want to estimate the binding motif matrices for several transcription factors in the same protein family, but have only a handful of known binding sites for each factor, you can estimate more robust motif matrices by borrowing information across the family. What are other examples? Looking at your own data might reveal the answer.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ramsey, F.L. & Schafer, D.W. *The Statistical Sleuth: A Course in Methods of Data Analysis* (Duxbury/Thomson Learning; 2002).
2. Hastie, T., Tibshirani, R. & Friedman, J.H. *The Elements of Statistical Learning*, edn. 2 (Springer; 2009).
3. Tibshirani, R. *J. Roy Stat. Soc. B* **58**, 267–288 (1996).
4. Smyth, G.K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
5. Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. *Bayesian Data Analysis* edn. 2 (Chapman & Hall/CRC; 2004).
6. Beaumont, M.A. & Rannala, B. *Nat. Rev. Genet.* **5**, 251–261 (2004).
7. Ji, H. & Wong, W.H. *Bioinformatics* **21**, 3629–3636 (2005).
8. Sartor, M.A. *et al.* BMC *Bioinformatics* **7**, 538 (2006).
9. Chen, G.K. & Witte, J.S. *Am. J. Hum. Genet.* **81**, 397–404 (2007).
10. Zhou, Q. & Wong, W.H. *Proc. Natl. Acad. Sci. USA* **101**, 12114–12119 (2004).