PBMs are an important technological development, especially in the latest implementations that include all possible 10-mer binding sites. They provide an inexpensive and high-throughput method for determining binding specificities of transcription factors and are rapidly increasing the database of characterized transcription factors. To maximize the information obtained from this technique, it is critical to employ optimized analysis methods. The success of the BEEML-PBM method is mainly due to the power of regression analysis and demonstrates that quantitative PBM data can be analyzed in the traditional biochemical framework of equilibrium binding to obtain accurate binding energies.

With a few exceptions, the simple PWM model performs very well, supporting the hypothesis that the energetics of transcription factor–DNA recognition is generally simple. This simplicity has considerable practical implications. The main difficulty in the study of transcription factor specificity is one of scale. Unlike protein–protein interactions, a single affinity is not sufficient to parameterize transcription factor specificity. For example, there are more than a million possible sequences for a 10-nt-long binding site. Even with high-throughput techniques, direct measurement of affinity for all sites is not practical. However, if the bases contribute to the total binding free energy independently, then a model with only 31 parameters can give accurate predictions of the million binding energies. Even if neighboring dinucleotide interactions are important, only 112 parameters are necessary[9]. Furthermore, this simplicity can be exploited in the design of promoters with tunable induction or transcription factors with custom specificity.

We conclude that the widespread phenomenon of secondary binding preference identified by Badis *et al.*[6] from PBM data is not supported by the data. The suboptimal estimation of the PWMs in previous studies can be accounted for by the lack of a biophysical model for transcription factor binding and the use of summary statistics, such as *E*-scores and *Z*-scores. This can be corrected by taking into account the specific characteristics of PBM data and maximizing the fit to the intensity data directly.

A support vector regression (SVR) method has also been used to improve PWM predictions compared with UniPROBE PWMs, yielding superior results in most, but not every, case[11]. In comparison, BEEML-PBM improved the predictions in every case (compared with UniPROBE PWMs), the resulting model has many fewer parameters than the SVR model, and each parameter has a specific biophysical interpretation (e.g., a binding energy contribution of a specific base-pair to the transcription factor–DNA interaction). The software code for BEEML-PBM is available in **Supplementary Note** and at http://ural.wustl.edu/~zhaoy/beeml/.

ACKNOWLEDGMENTS

We thank T. Hughes, M. Bulyk and Q. Morris for very helpful comments on the manuscript. We also thank members of the Stormo laboratory for their comments and suggestions throughout the course of this work. This work was supported by National Institutes of Health (NIH) grant R01 HG00249 to G.D.S. and NIH training grant T32 HG00045 to Y.Z.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Yue Zhao & Gary D Stormo

Washington University Medical School, St. Louis, Missouri, USA.
e-mail: stormo@wustl.edu

1. Stormo, G.D. & Zhao, Y. *Nat. Rev. Genet.* **11**, 751–760 (2010).
2. Luscombe, N.M. & Thornton, J.M. *J. Mol. Biol.* **320**, 991–1009 (2002).
3. Stormo, G.D. *Bioinformatics* **16**, 16–23 (2000).
4. Bulyk, M.L., Huang, X., Choo, Y. & Church, G.M. *Proc. Natl. Acad. Sci. USA* **98**, 7158–7163 (2001).
5. Berger, M.F. *et al. Nat. Biotechnol.* **24**, 1429–1435 (2006).
6. Badis, G. *et al. Science* **324**, 1720–1723 (2009).
7. Zhao, Y. Granas, D. & Stormo, G.D. *PLoS Comput. Biol.* **5**, e1000590 (2009).
8. Benos, P.V., Bulyk, M.L. & Stormo, G.D. *Nucl. Acids Res.* **30**, 4442–4451 (2002).
9. Stormo, G.D. *Genetics* published online, doi:10.1534/genetics.110.126052 (4 February 2011).
10. Newburger, D.E. & Bulyk, M.L. *Nucl. Acids Res.* **37**, D77–82 (2009).
11. Agius, P., Arvey, A., Chang, W., Nobel, W.S. & Leslie, C. *PLoS Comput. Biol.* **6**, e1000916 (2010).

Jury remains out on simple models of transcription factor specificity

To the Editor:

Zhao and Stormo[1] introduce a new method for deriving position weight matrices (PWMs) from protein binding microarrays (BEEML-PBM). Using this method, they challenge a central claim of our 2009 paper[2] and conclude "that the widespread phenomenon of secondary binding preference identified by Badis *et al.* is not supported by our data" and that the PWMs were suboptimally estimated.

BEEML-PBM is simple, elegant and corrects for a pronounced positional effect of transcription factor (TF) binding in the PBM assay; however, we do not agree with their overall conclusion and believe that it is based on incomplete and biased analysis of our data. The conclusions of Zhao and Stormo[1] are based on comparing the performance of BEEML-PBM PWMs and our methods on held-out data. However, they overestimate the performance of their PWMs and underestimate the performance of our methods.

First, their claims of suboptimality of our PWMs are based on results from only one of the three motif finders that we employed, Seed-and-Wobble (SnW). SnW was not developed to predict probe intensities and does not attempt to produce a summary PWM that optimizes performance over all probes in predicting probe intensities. Instead, it was developed for the purpose of summarizing the 8-mer data, seeding with the highest scoring 8-mer, in a compact way for use in visual depiction as sequence logos. In contrast, another of the methods we employed, RankMotif++, is designed to produce summary PWMs and we have previously reported[3] that it, like BEEML-PBM, better predicts probe intensities than SnW. So we suspect its performance would be much more competitive. In fact, RankMotif++ is very similar to the BEEML base method[4]; it fits a PWM model using a regression-like procedure to optimally predict PBM intensity data. RankMotif++ differs from BEEML primarily in that it regresses on a partial preference ordering of probes inferred from their PBM intensities rather than on their actual intensities themselves. We acknowledge that comparisons with RankMotif++ PWMs would have been difficult because, although the source code for RankMotif++ has been available for 3 years, the PWMs we learned for Badis *et al.*[2] were until recently only available as sequence logos. However, we made the motifs available to Zhao and Stormo[1] when we were notified of this oversight and before the final submission of their paper. The motifs are available here: http://the_brain.bwh.harvard.edu/suppl105/.

Second, we note that Zhao and Stormo[1] use a positional effect model when training their PWMs but do not allow the methods that they are comparing against the same opportunity to correct this bias during training. We propose that this correction is a major cause of BEEML-PBM's success and that both the

multiple PWM methods and the 8-mer affinity estimates we employed would greatly benefit from a similar correction, thus restoring our reported gain in performance. For example, the 8-mer median intensities used in their Figure 2a are not corrected for positional biases and this leads to the counterintuitive claim that for the 15–20 (of 41) data points that lie above the diagonal, BEEML-PBM PWMs capture >100% of the replicate reproducibility. A more appropriate comparison would either employ PWMs uncorrected for positional bias (as we did in our original paper) or to compare against similarly corrected 8-mer median intensities. Zhao and Stormo[1] do neither and, as such, we believe that their prediction accuracy estimates are inflated.

Finally, we note that explaining 90% of the reproducible binding signal is not the same as explaining 100%, and proteins that we and others have confirmed have multiple binding modes do not satisfy Zhao and Stormo's 90% cut-off. For example, we reported that Jundm2 (Jdp2) binds two half-sites with variable spacing between them; this is clearly observed in the top-scoring 8-mers[2]. This mode of binding is common among other bZIP proteins. Furthermore, Zhao and Stormo[1] do not consider the PBM data for Bcl6b, a C2H2 zinc finger for which we obtained two very different PWMs; these are also clearly observed in the top-scoring 8-mers, and, moreover, enrichment for motif matches can be observed in associated ChIP-chip data[2]. In general, variable spacing in long C2H2 zinc-finger array seems to be common; for example, ChIP-seq for RE1-silencing transcription factor also supports use of partial versus full sites and different spacings[5]. Single, summary PWMs cannot capture these binding modes, and it is important to do so, as C2H2 zinc fingers are the most common domain in metazoa, and long arrays of these domains are common in human and mouse genomes.

We agree that simple and accurate representation of transcription factor sequence specificity on the basis of PBM data is an important problem. We ourselves have been working on extensions to our algorithms to capture the PBM positional and orientation effects (which we have previously reported[6]). We also have recently conducted a DREAM (Dialogue for Reverse Engineering Assessments and Methods) competition in which the goal was to predict PBM probe intensities using a two-array framework and evaluation criteria similar to those used in reference 2 and Zhao *et al.*[4]. A manuscript describing these new data, the competition, the methods of ~20 groups, their evaluation and a web site that allows benchmarking any method to the DREAM results is in preparation (M. Weirauch *et al.*, unpublished data). We have now obtained the BEEML-PBM code, and we look forward to comparing it to alternatives.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Quaid Morris[1–3], Martha L Bulyk[4–7] & Timothy R Hughes[1,2]

[1]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [3]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [4]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [5]Department of Pathology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [6]Harvard-MIT Division of Health Sciences and Technology (HST) Harvard Medical School, Boston, Massachusetts, USA. [7]Committee on Higher Degrees in Biophysics, Harvard University, Cambridge, Massachusetts, USA.
e-mail: quaid.morris@utoronto.ca

1. Zhao, Y. & Stormo, G.D. *Nat. Biotechnol.* **29**, 480–483 (2011).
2. Badis, G. *et al. Science* **324**, 1720–1723 (2009).
3. Chen, X., Hughes, T.R. & Morris, Q. *Bioinformatics* **23**, i72–i79 (2007).
4. Zhao, Y., Granas, D. & Stormo, G.D. *PLoS Comput. Biol.* **5**, e1000590 (2009).
5. Johnson, D.S., Mortazavi, A., Myers, R.M. & Wold, B. *Science* **316**, 1497–1502 (2007).
6. Berger, M.F. *et al. Nat. Biotechnol.* **24**, 1429–1435 (2006).

US attitudes toward human embryonic stem cell research

To the Editor:
Although scientifically promising, research using human embryonic stem cells (hESCs) has roused political controversy for nearly two decades, with sharp differences between policies in different nations and, in the United States, repeated changes in policy—the latest of which is the lifting of the ban on US federal funding of hESC work[1]. US federal funding for research on stem cells derived from nuclear transfer of a patient's own genes, a promising approach sometimes called therapeutic cloning, remains banned[2]. Arguments for

Box 1 Questions on reproductive cloning

Scientists can now make clones—baby animals that are exact genetic copies of an adult. This is how it works:

- BEGIN with a fertilized egg. For humans this is likely to be a spare embryo from an IVF program which would otherwise be thrown away.
- REMOVE the original genes.
- REPLACE them with genes from the person to be cloned.
- GROW the fertilized egg in the lab for a few days into an early embryo, a little ball of cells.
- IMPLANT the embryo in the womb of a surrogate mother where it develops into a baby. The baby is the donor's identical twin, except for the difference in age.

Do you approve of...

Cloning endangered animals?
Cloning the best farm animals to improve breeding stock—for example, cloning a superb dairy bull?
Cloning a child killed in a traffic accident?
Cloning a child that is a copy of its father or mother?

The answer options for these questions were:

Definitely yes
Yes
Undecided, mixed feelings
No
Definitely not

We scored these options in equal intervals: Definitely not = 0; No = 25; Undecided, mixed feelings = 50; Yes = 75; Definitely yes = 100.