orthology mapping from InParanoid[2] to translate gene information between species. In the case of mapping to multiple orthologs, the ortholog with the closest value to the input gene is used to determine the distance for the comparison (**Supplementary Methods**).

As an example of ExpressionBlast usage, we studied an expression data set of male mice overexpressing *Sirt6*. Although phenotypic effects of *Sirt6* have been documented (including increased lifespan[3] and reduced levels of diet-induced obesity[4]), little is known about the specific mechanisms by which it operates or the pathways it regulates. ExpressionBlast gives us an opportunity to look for other mice experiments whose results have significant similarity to or are negatively correlated with the *Sirt6* overexpression profile. Such matching can allow users to determine the functional categories and pathways activated by *Sirt6* and may lead to novel hypotheses and follow-up experiments regarding the role of *Sirt6* in regulating cellular activity. We also envisage using ExpressionBlast to study expression data from patients[5] to identify relevant studies that can be used to explain underlying causes of human diseases.

By allowing users to mine large, unstructured expression databases, ExpressionBlast can become a useful tool leading to important new hypotheses and findings.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nmeth.2630).

ACKNOWLEDGMENTS

This work was supported in part by US National Institutes of Health grant 1R01 GM085022 and US National Science Foundation awards DBI-0965316 and I-Corps 1242525 to Z.B.-J.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Guy E Zinman[1,3], Shoshana Naiman[2,3], Yariv Kanfi[2], Haim Cohen[2] & Ziv Bar-Joseph[1]

[1]Lane Center for Computational Biology, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA. [2]The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel. [3]These authors contributed equally to this work.
e-mail: zivbj@cs.cmu.edu

1. Barrett, T. *et al. Nucleic Acids Res.* **35**, D760–D765 (2007).
2. Ostlund, G. *et al. Nucleic Acids Res.* **38**, D196–D203 (2010).
3. Kanfi, Y. *et al. Nature* **483**, 218–221 (2012).
4. Kanfi, Y. *et al. Aging Cell* **9**, 162–173 (2010).
5. Rajkumar, R. *et al. Am. J. Physiol. Heart Circ. Physiol.* **298**, H1235–H1248 (2010).

Building an ENCODE-style data compendium on a shoestring

To the Editor: One perhaps unintended consequence of the success of the human genome project has been a shift in the biomedical research funding landscape toward large-scale programs, commonly involving several hundred scientists and budgets of hundreds of millions of dollars. However, this emphasis on large-scale projects has been questioned, as illustrated by recent debates following last year's publications from the Encyclopedia of DNA Elements (ENCODE) project[1,2]. Rather than making decisions ahead of time about what data sets should be generated for a given research community, as large-scale projects must do, we have explored an alternative approach, compiling all data sets produced by one such community as soon as they have been deposited in public databases. We demonstrate that the compendium size resulting from such real-time curation can exceed that of large-consortium efforts, thereby providing a highly topical contribution to the ongoing 'small science versus big science' debate.

We created HAEMCODE, a repository for transcription factor (TF)-binding maps in mouse blood cells; the maps are generated from chromatin immunoprecipitation followed by sequencing (ChIP-seq) data. Using a standardized analysis pipeline, we manually curated more than 300 TF ChIP-seq studies from a wide range of primary mouse hematopoietic cells and major cell line models. As of September 2013, the HAEMCODE compendium covered 84 TFs across 24 major blood cell types. Hemopoiesis is also a major focus of ENCODE, yet the currently available mouse ENCODE data (36 TFs; May 2013) cover less than half the HAEMCODE contents, with only 9 TFs investigated by ENCODE not available elsewhere.

We developed a Web interface (http://haemcode.stemcells.cam.ac.uk/) to provide data access as well as a range of online analysis tools that we designed to be useful to both experimentalists and computational biologists. In the classical use case, a user selects experiments within HAEMCODE before being directed to a workspace that offers precomputed options to inspect and/or download selected ChIP-seq data sets. Additional online tools can compute global similarity between selected experiments, investigate overrepresentation of a user-submitted gene list in any subset of ChIP-seq experiments[3], inspect precomputed results from *de novo* motif discovery and output all ChIP-seq experiments with binding peaks for a user-supplied gene locus.

Integration of publicly available data represents a powerful approach to make novel discoveries across diseases, species and platforms that would be impossible to achieve from single projects[4]. Successful completion of the HAEMCODE project on a small budget highlights this approach as a potentially widely applicable complement to multimillion-dollar research initiatives.

ACKNOWLEDGMENTS

This work was funded by the Biotechnology and Biological Sciences Research Council, Leukaemia and Lymphoma Research, the Medical Research Council (MRC), Cancer Research UK, the Cambridge National Institute for Health Research (NIHR) Biomedical Research Center and core support grants from the Wellcome Trust–MRC Cambridge Stem Cell Institute. F.S.L.N. is supported by a Yousef Jameel scholarship.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

David Ruau[1,2], Felicia S L Ng[1,2], Nicola K Wilson[1,2], Rebecca Hannah[1,2], Evangelia Diamanti[1,2], Patrick Lombard[1,2], Steven Woodhouse[1,2] & Berthold Göttgens[1,2]

[1]Department of Hematology, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [2]Wellcome Trust–Medical Research Council (MRC) Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK.
e-mail: djr62@cam.ac.uk or bg200@cam.ac.uk

1. Alberts, B. *Science* **337**, 1583 (2012).
2. The ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
3. Joshi, A., Hannah, R., Diamanti, E. & Göttgens, B. *Exp. Hematol.* **41**, 354–366 (2013).
4. Butte, A.J. & Kohane, I.S. *Nat. Biotechnol.* **24**, 55–62 (2006).