Stephen J Chapman[1,2], Chiea C Khor[1], Fredrik O Vannberg[1], Nicholas A Maskell[2], Christopher WH Davies[3], Emma L Hedley[2], Shelley Segal[4], Catrin E Moore[4], Kyle Knox[5], Nicholas P Day[6], Stephen H Gillespie[7], Derrick W Crook[5], Robert JO Davies[2] & Adrian VS Hill[1]

[1]The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK. [2]Oxford Centre for Respiratory Medicine, Churchill Hospital Site, Oxford Radcliffe Hospital, Oxford OX3 7LJ, UK. [3]Department of Respiratory Medicine, Royal Berkshire Hospital, Reading RG1 5AN, UK. [4]Department of Paediatrics, John Radcliffe Hospital, Oxford OX3 9DU, UK. [5]Department of Microbiology, John Radcliffe Hospital, Oxford OX3 9DU, UK. [6]Centre for Clinical Vaccinology and Tropical Medicine, Oxford OX3 9DU, UK. [7]Centre for Medical Microbiology, Department of Infection, University College London, London NW1 2BU, UK.
e-mail: adrian.hill@well.ox.ac.uk

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

ACKNOWLEDGMENTS
S.J.C. is a Wellcome Trust Clinical Research Fellow, and A.V.S.H. is a Wellcome Trust Principal Fellow. C.C.K. is a scholar of the Agency for Science, Technology and Research (A-STAR) of Singapore and a member of the Master of Medicine and Surgery (MBBS) PhD program of the Faculty of Medicine, National University of Singapore.

1. Vang, T. *et al. Nat. Genet.* **37**, 1317–1319 (2005).
2. Bottini, N. *et al. Nat. Genet.* **36**, 337–338 (2004).
3. Begovich, A.B. *et al. Am. J. Hum. Genet.* **75**, 330–337 (2004).
4. Kyogoku, C. *et al. Am. J. Hum. Genet.* **75**, 504–507 (2004).
5. Smyth, D. *et al. Diabetes* **53**, 3020–3023 (2004).
6. Siminovitch, K.A. *Nat. Genet.* **36**, 1248–1249 (2004).
7. Roy, S. *et al. Lancet* **359**, 1569–1573 (2002).
8. Maskell, N.A. *et al. N. Engl. J. Med.* **352**, 865–874 (2005).
9. Kadioglu, A. & Andrew, P.W. *Trends Immunol.* **25**, 143–149 (2004).
10. Epplen, J.T. *Hum. Genet.* **90**, 331–341 (1992).

GenePattern 2.0

To the Editor:

Whole-genome expression profiling has created a revolution in the way we study disease and basic biology. Since 1997, the number of published results based on an analysis of gene expression microarray data has grown from 30 to over 5,000 publications per year. Sophisticated mathematical methods have been developed for use in patient diagnosis and prognosis, identification of new drug targets and understanding biological mechanisms. However, these tools are often out of the direct reach of the biomedical researchers who can so critically benefit from them because they can be difficult to understand and use correctly. This challenge is even more relevant in the context of 'integrative' approaches, where a multitude of data sources and methods are combined in the analysis of a single problem.

To address this challenge in genomics research we have developed a software package called GenePattern (see the **Supplementary Tutorial** and list of frequently asked questions in the **Supplementary Note** online), which provides a comprehensive environment that can support (i) a broad community of users at all levels of computational experience and sophistication, (ii) access to a repository of analytic and visualization tools and easy creation of complex analytic methods from them and (iii) the rapid development and dissemination of new methods. **Supplementary Figures 1** and **2** show the functional and software architecture of the system. Perhaps the most important feature of GenePattern is that it supports a mechanism to guarantee the capture and independent replication of published computational methods and *in silico* results. Although there are many packages available for microarray analysis, few currently provide all the features we describe above (see **Supplementary Table 1** online).

The best way to illustrate the utility of the GenePattern approach is to give an example that motivated its development. In 1999 we published a methodology to classify leukemia patient samples. The method builds predictive models using marker genes that are significantly differentially expressed between two subtypes of leukemia, acute lymphoblastic (ALL) and acute myelogenous (AML)[1]. The steps of the method (**Fig. 1**, left) were originally implemented manually by running a sequence of required software tools to identify the statistically significant genes to be used as features in the predictive model, visualize their expression profiles, build the predictor and set its parameters via cross-validation runs on the training data set, test the final model on a separate test data set, and import the data into Excel to visualize the results. This cumbersome process required programming skill as well as access to the appropriate software tools. In addition, there was no good, automated way to ensure the accurate capture of this process, including the order of tool invocation and parameter settings.

Although the tools and the data sets used to obtain these results were made publicly available as part of their publication, this was not sufficient for independent replication. We received hundreds of e-mail messages with questions on how to reproduce the results and requests for analysis details that were not included in the paper: for instance, the exact way genes were filtered or the exact parameters settings for the classification model. What was missing was a detailed, self-contained, executable description of the method that allowed its reproduction by others in an automated way.

With GenePattern, we have been able to make a 'reproducible research'[2] version of the method by capturing the entire set of steps (along with the parameter settings) using a simple form-based environment (**Fig. 1**, right). The resulting 'pipeline' makes all the necessary calls—with no manual intervention—to the required analysis and visualization tools, all of which are available in the GenePattern module repository. Moreover, this and similar pipelines can be created by a nonprogramming user. The methodology can easily be executed on different data sets or modified to perform variations of the method without programming. The ALL-AML analysis pipeline is now part of the core GenePattern module library distribution.

A distinguishing feature of GenePattern is the inclusion of multiple user environments including a point-and-click graphical user interface, an analytic method pipeline builder and a programming language interface. For nonprogramming users, the interface provides prepackaged analysis modules of both individual algorithms and more complex methodologies encapsulated as pipelines. Modules and pipelines are available to the user via pull-down menu task lists, and data sets are accessible through a file browser. The repository provides a core library of approximately 60 analysis modules, including most of the common functions of gene expression analysis, such as preprocessing, clustering, prediction and

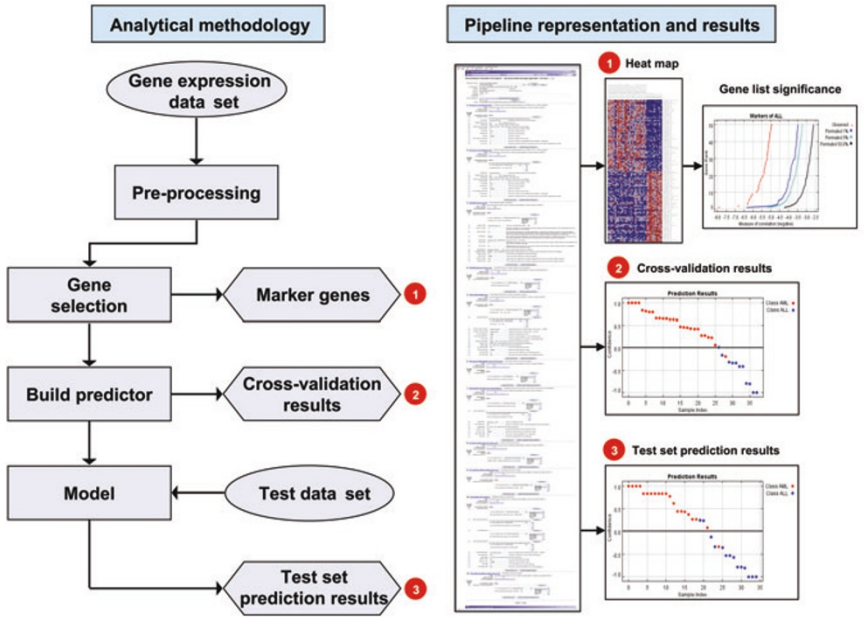

Figure 1 GenePattern pipelines. At left is a diagrammatic representation of the computational method used in ref. 1 for classifying leukemia patient samples. Generally, a researcher has to perform each step of this analysis manually, including invoking parameter settings, matching data objects and even moving files. In GenePattern, she can encapsulate this set of steps in a single, reproducible pipeline (at right) that choreographs the entire process. The pipeline is then available for modification, with each revision preserved for reproducibility and to share with colleagues and the community.

marker selection, as well as more complex methods previously published by our group. While many of the modules were originally developed for use with Affymetrix GeneChip data, they can also be applied to two-channel data (see the frequently asked questions in the **Supplementary Note**). A complete list of the repository contents can be found in **Supplementary Table 2**. Each module in the repository comes with extensive documentation (http://www.broad.mit.edu/genepattern/doc/modules) about the specifics of the algorithm and its use, including references for further study. In this way, we hope to minimize the danger of misuse or erroneous conclusions. For users with programming experience, GenePattern provides libraries that allow transparent access to GenePattern modules from the R, MATLAB and Java programming languages.

In an active research environment, the ability to quickly integrate new methods with existing tools is essential to advance research goals associated with large-scale genomic projects. Software engineering teams can rarely keep up with the rate at which new prototypes and methods are created. GenePattern provides a rapid, language-independent mechanism to deploy new tools via a simple form-based process. Thus, the system enables the creation, evaluation, and adoption of new methods in almost real-time, and it allows users to easily supplement the contents of the module repository with their own or other tools; for example, from the Bioconductor project (http://www.bioconductor.org).

A typical analysis of genomic data involves a complex multistep sequence of computation and data access across a multitude of tools, platforms and data sets. Writing one-off programs to join these together can be tiresome and out of the reach of many laboratory scientists, and the end product may not be accessible for another researcher's use. The GenePattern pipeline builder provides an easy-to-use, form-based method for stringing data processing, analytic and visualization modules together into methodologies. The methods can then easily be integrated back into the module repository—without additional engineering as described above—and can be made available to other researchers for their use or modification. The GenePattern interface provides the capability to export a method, including all modules, data and associated parameters, packaged as a single file that can be imported and run by any other GenePattern user (**Supplementary Tutorial**). Thus, GenePattern provides a painless way to create and distribute an entire computational analysis methodology in a unified and executable form.

GenePattern is freely available at the Broad Institute website (http://www.broad.mit.edu/genepattern). The software was originally released in March 2004 and currently has over 2,200 registered users. GenePattern received an Editors' Choice award in the 2005 Bio-IT World Best Practices competition, which recognizes technologies that further the goal of bridging information technology and biomedical research and development.

GenePattern version 2.0 was recently released, adding a new suite of modules for the processing, analysis and visualization of proteomic data and incorporating additional support for reproducible research. This includes recording user actions so that analyses can be re-run with different data or parameters and automatically creating the corresponding analysis pipeline. Further planned enhancements to the software include addition of modules for the analysis of SNP and genotype data, templates for creating text documents with embedded data and analytic pipelines and, as part of the National Cancer Institute's caBIG (Cancer Biomedical Informatics Grid) project, infrastructure to support access to remote resources across the web.

Michael Reich, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo & Jill P Mesirov

Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.
e-mail: mesirov@broad.mit.edu

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

1. Golub, T.R. *et al. Science* **286**, 531–537 (1999).
2. Gentleman, R. *Stat. Appl. Genet. Mol. Biol. [online]* **4**, 2 (2005) <http://www.bepress.com/sagmb/vol4/iss1/art2>.