Science
AAAS

A Postgenomic Visual Icon

John N. Weinstein

Science **319**, 1772 (2008);
DOI: 10.1126/science.1151888

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

discovered slowing of the rate of CO_2 uptake by the Southern Ocean appears to hinge on questionable CO_2 measurements made at Ascension Island in the 1980s (*7*, *8*). A duplicate record would have settled the issue.

The Scripps CO_2 program is now a component of a multinational collaboration aimed at tracking changes in greenhouse gases and related species, coordinated by the World Meteorological Organization. A major justification of this effort is the promise of quantifying the sources and sinks of greenhouse gases at Earth's surface. Soon, the compliance of international treaties to curb greenhouse gas emissions may be assessed using these capabilities.

If long-term observations are fundamental to understanding global change, why have they proved so hard to support? The costs of sustained measurements can be high, so prioritization is clearly an issue. The Scripps program has proved, however, that a long-term observational program is not necessarily incompatible with the normal peer review system. The Scripps program continues to be funded—if perilously—one grant at a time. Even within agencies committed to long-term observations, such as the National Oceanographic and Atmospheric Administration, funding is tight and a hiatus may be only one political wind shift or economic downturn away. A diversity of funding sources supporting a heterogeneous mixture of overlapping programs is probably the best formula for long-term stability.

A continuing challenge to long-term Earth observations is the prejudice against science that is not directly aimed at hypothesis testing. At a time when the planet is being propelled by human action into another climate regime with incalculable social and environmental costs, we cannot afford such a rigid view of the scientific enterprise. The only way to figure out what is happening to our planet is to measure it, and this means tracking changes decade after decade and poring over the records. A point of diminishing scientific returns has never been realized in what is now known as the "Keeling Curve," the Mauna Loa CO_2 record.

References

1. C. D. Keeling, *Annu. Rev. Energy Environ.* **23**, 25 (1998).
2. C. D. Keeling, in *Proceedings of the Conference on Recent Research in Climatology*, H. Craig, Ed. (Committee on Research in Water Resources and University of California, Scripps Institution of Oceanography, La Jolla, CA, 1957), pp. 43–49.
3. C. D. Keeling, *Tellus* **12**, 200 (1960).
4. H. L. Bryden, H. R. Longworth, S. A. Cunningham, *Nature* **438**, 655 (2005).
5. R. B. Bacastow, *Nature* **261**, 116 (1976).
6. R. B. Bacastow, C. D. Keeling, T. P. Whorf, *J. Geophys. Res.* **90**, 10529 (1985).
7. R. M. Law, R. J. Matear, R. J. Francey, *Science* **319**, 570a (2008).
8. C. Le Quéré *et al.*, *Science* **319**, 570c (2008).

10.1126/science.1156761

BIOCHEMISTRY

A Postgenomic Visual Icon

A decade of experience in visualizing large-scale genotypic and phenotypic data as heat maps has illuminated the strengths and limitations of the approach.

John N. Weinstein

The "postgenomic era" in biology may be hard to define, and when it actually began is subject to debate. But its most characteristic feature is clearly the accumulation of massive amounts of genotypic and phenotypic data that must be organized, analyzed, visualized, and interpreted. That series of challenges has been central to recent bioinformatics. For visualization, by far the most popular graphical representation has been the "clustered heat map," which compacts large amounts of information into a small space to bring out coherent patterns in the data. Despite its popularity, however, are such maps optimal for visually integrating information to extract valuable insights and generate fresh hypotheses? That question can be addressed through understanding the strengths and limitations of heat map visualization.

Since their debut over 10 years ago (*1*) (see the figure), clustered heat maps have appeared in well over 4000 biological or biomedical publications. They have been used for two-dimensional display of patterns in all types of molecular data, including messenger RNA (mRNA) and microRNA expression, protein expression, DNA copy number, DNA methylation, metabolite concentration, and drug activity (*1*–*8*). They have proved useful for microarray data (*2*) and have sometimes been engineered for "integromic" merging (*1*, *9*, *10*) of different types of molecular information. The figure, for example, combines data on mRNA expression, protein expression, mutations, cell cycle properties, stress responses, a yeast-based functional assay, and drug activity in cancer cells. Organisms analyzed have spanned the phylogenetic tree from the plant *Arabidopsis thaliana* to rainbow trout to suicidal crickets (*11*–*13*). Diseases analyzed have ranged from AIDS to cancer to bubonic plague (*1*, *14*, *15*).

In the case of gene expression data, the color assigned to a point in the heat map grid indicates how much of a particular RNA or protein is expressed in a given sample. The gene expression level is generally indicated by red for high expression and either green or blue for low expression. Coherent patterns (patches) of color are generated by hierarchical clustering on both horizontal and vertical axes to bring like together with like. Cluster relationships are indicated by tree-like structures adjacent to the heat map, and the patches of color may indicate functional relationships among genes and samples. Occasionally, a source of order other than clustering (for example, time in a series of measurements) is used on one or both axes. Without some basis for functional ordering on both axes, however, there would be no coherent patterns of color.

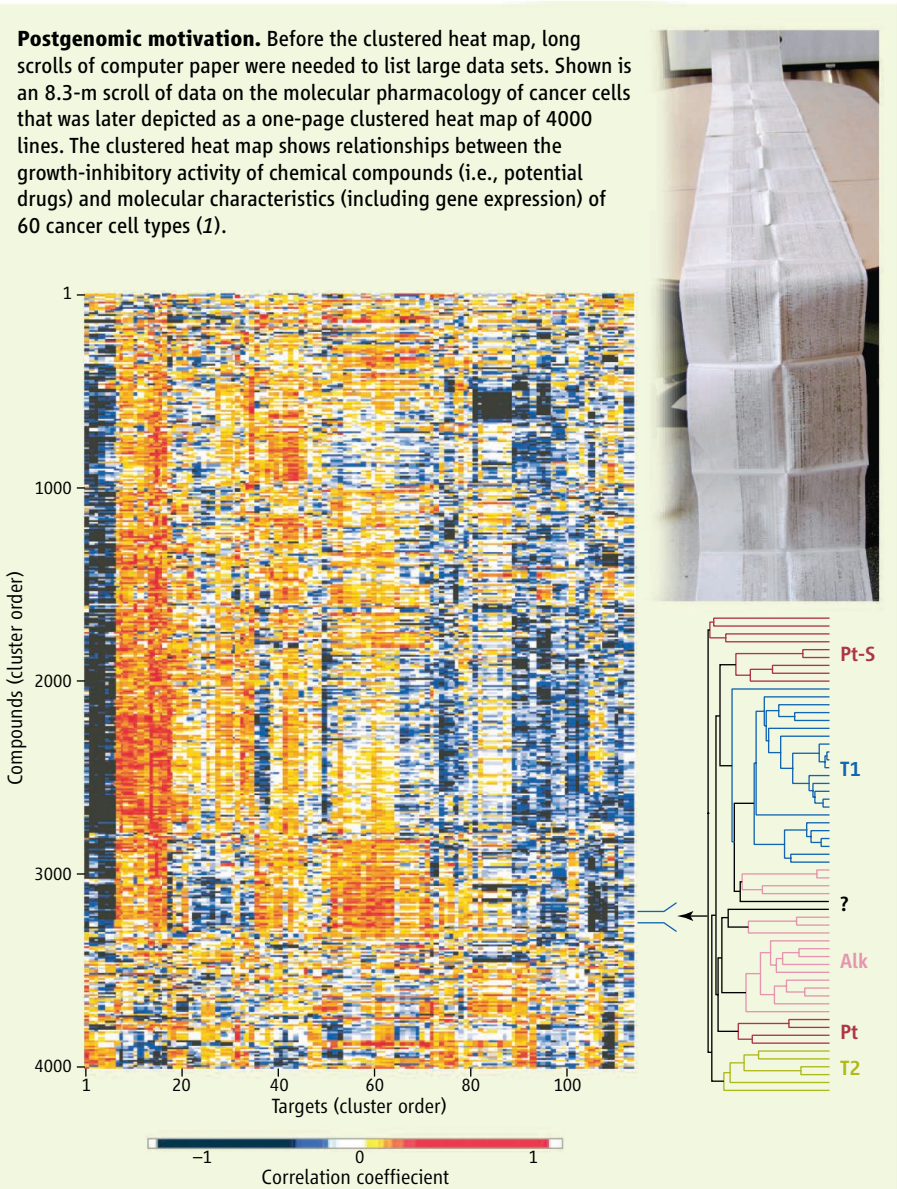
Seductive though it may be, the clustered heat map has its limitations and potential for misinterpretation or misuse. Most prominently among the limitations, it provides only first-order insight into the data; complex patterns of nonlinear relationship among only a few of the samples are unlikely to show up. A computer-intensive variant based on "biclustering" has been developed to reveal such relationships (*16*). A second problem is that, in hierarchical clustering, each bifurcation of the cluster tree can be "swung" in either direction at each fork in the tree, so some objective (but, to a degree, arbitrary) rule must be invoked to decide which way each branch will, in fact, swing. There is also the temptation to select a small subset of the variables (for example, genes in a microarray study), and represent them in a clustered heat map. That is common (and appropriate) practice in the discovery of new biomarkers and gene expression signatures for discriminating subtypes of a disease such as cancer (*17*). However, if one picks a signature consisting of only a few dozen genes out of a set of more than 10,000, then even randomized

Department of Bioinformatics and Computational Biology, M. D. Anderson Cancer Center, 1515 Holcombe Boulevard, Houston, TX 77030, USA. E-mail: jweinste@mdanderson.org

data can produce clustered heat maps that appear spuriously to show good distinction of two subclasses. Typically, in such small-subset "cherry-picking" analysis, the upper left and lower right quadrants of the figure will be predominantly one color (for example, red), whereas the lower left and upper right quadrants will be the other color (for example, green or blue). It may be legitimate to show such maps for gene signatures simply because the entire set of genes cannot be annotated on a single map. But the lack of statistically robust meaning to the pattern should be clearly stated, and that is rarely done.

Even beyond those limitations and concerns, the generation of clustered heat maps is a surprisingly subtle process. It requires that a large number of choices be made, and those choices dictate the type and meaning of pattern that emerges. The required decisions include (i) the preprocessing algorithm (e.g., type of background subtraction, data normalization, and data filtering), which hopefully minimizes noise in the system while keeping the meaningful signal; (ii) the clustering algorithm (e.g., average linkage, complete linkage, or centroid-based), which determines how the data will be grouped; (iii) the distance metric (e.g., Euclidean or correlation), which defines what is meant by similarity of genes or samples to each other; and (iv) the color scheme (linear, logarithmic, quantile, two-color, or three-color), which determines what patterns in the data will be emphasized visually. The decision to use relative or absolute data must also be made, and one data set can be subtracted from another to create a "difference" heat map (e.g., of gene expression before and after treatment of cells with a drug). [See supporting online material and (*1*, *18*) for more detailed explanations of those choices and their implications.] The point is that many different heat maps, each with its own visual meaning, can be generated from the same experiment. Unless the details of the parameter choices are specified, the analysis is incomplete at best and subject to misinterpretation at worst.

Postgenomic data sets are becoming larger and larger. A decade ago, microarrays produced thousands of numbers; now they often produce millions. Hardware, software, mathematical algorithms, and the human mind are all being stressed to the limit by the flood of data. Every biomedical research institution is finding its human and computational resources inadequate to the task. Our need for graphical representations that illuminate patterns in the data and evoke new hypotheses will only increase over the next few years, particularly in the drive to identify biomarkers useful for "personalizing" treatment of diseases such as cancer. The ability of the human eye to recognize patterns, coupled with modern data analysis, "can stimulate excitement, awe, new ways of looking at things, and, above all, a broad appreciation of even the most esoteric scientific information" (*19*). For the last decade, the ubiquitous clustered heat map has served those purposes, even if imperfectly, for postgenomic biology.



Postgenomic motivation. Before the clustered heat map, long scrolls of computer paper were needed to list large data sets. Shown is an 8.3-m scroll of data on the molecular pharmacology of cancer cells that was later depicted as a one-page clustered heat map of 4000 lines. The clustered heat map shows relationships between the growth-inhibitory activity of chemical compounds (i.e., potential drugs) and molecular characteristics (including gene expression) of 60 cancer cell types (*1*).

PHOTO CREDIT: GENOMICS & BIOINFORMATICS GROUP

References

1. J. N. Weinstein *et al.*, *Science* **275**, 343 (1997).
2. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
3. D. T. Ross *et al.*, *Nat. Genet.* **24**, 227 (2000).
4. U. Scherf *et al.*, *Nat. Genet.* **24**, 236 (2000).
5. G. Szakacs *et al.*, *Cancer Cell* **6**, 129 (2004).
6. B. R. Zeeberg *et al.*, *BMC Bioinformatics* **6**, 168 (2005).
7. S. Nishizuka *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 14229 (2003).
8. M. J. Brauer *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 19302 (2006).
9. T. G. Myers *et al.*, *Electrophoresis* **18**, 467 (1997).
10. J. N. Weinstein, Y. Pommier, *C. R. Biol.* **326**, 909 (2003).
11. J. Bobe, J. Montfort, T. Nguyen, A. Fostier, *Reprod. Biol. Endocrinol.* **4**, 39 (2006).
12. A. C. Doxey *et al.*, *Mol. Biol. Evol.* **24**, 1045 (2007).
13. D. G. Biron *et al.*, *Insect Mol. Biol.* **15**, 731 (2006).
14. Y. Le Priol *et al.*, *J. Immunol.* **177**, 5145 (2006).
15. X. Y. Wang *et al.*, *Arch. Microbiol.* **186**, 151 (2006).
16. Y. Kluger, R. Basri, J. T. Chang, M. Gerstein, *Genome Res.* **13**, 703 (2003).
17. T. R. Golub *et al.*, *Science* **286**, 531 (1999).
18. http://discover.nci.nih.gov/cimminer
19. J. Nesbit, M. Bradford, *Science* **317**, 1857 (2007).

Supporting Online Material
www.sciencemag.org/cgi/content/full/319/5871/1772/DC1

10.1126/science.1151888

Published by AAAS