stress through its kinase domain. But how this relates to disease resistance remains a mystery. Several other QDR-associated genes encode proteins with kinase domains[9]: do these genes confer resistance through similar mechanisms?

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Saikkonen, K., Gundel, P. & Helander, M. *J. Chem. Ecol.* **39**, 962–968 (2013).
2. Nagabhyru, P., Dinkins, R., Wood, C., Bacon, C. & Schardl, C. *BMC Plant Biol.* **13**, 127 (2013).
3. Zuo, W. *et al. Nat. Genet.* **47**, 151–157 (2015).
4. Matyac, C. *Phytopathology* **75**, 924–929 (1985).
5. Martinez, C., Jauneau, A., Roux, C., Savy, C. & Dargent, R. *Protoplasma* **213**, 83–92 (2000).
6. Jones, J.D.G. & Dangl, J.L. *Nature* **444**, 323–329 (2006).
7. Bent, A.F. & Mackey, D. *Annu. Rev. Phytopathol.* **45**, 399–436 (2007).
8. Poland, J.A., Balint-Kurti, P.J., Wisser, R.J., Pratt, R.C. & Nelson, R.J. *Trends Plant Sci.* **14**, 21–29 (2009).
9. Roux, F. *et al. Mol. Plant Pathol.* **15**, 427–432 (2014).
10. Studer, A.J. & Doebley, J.F. *Genetics* **188**, 673–681 (2011).
11. Fu, H. & Dooner, H. *Proc. Natl. Acad. Sci. USA* **99**, 9573–9578 (2002).
12. Kohorn, B.D. *et al. Plant J.* **60**, 974–982 (2009).
13. Kohorn, B.D. *et al. Plant J.* **46**, 307–316 (2006).

Big data mining yields novel insights on cancer

Peng Jiang & X Shirley Liu

Recent years have seen the rapid growth of large-scale biological data, but the effective mining and modeling of ‘big data’ for new biological discoveries remains a significant challenge. A new study reanalyzes expression profiles from the Gene Expression Omnibus to make novel discoveries about genes involved in DNA damage repair and genome instability in cancer.

Since the invention of gene expression microarray technology almost 20 years ago, numerous mRNA profiling data sets have been generated for diverse biological processes in many organisms. Currently, there are over 30,000 series and 1 million samples of array-based gene expression data deposited in the NCBI Gene Expression Omnibus (GEO) database. In this issue, Rudolf Fehrmann and colleagues comprehensively reanalyzed the expression profiles of 77,840 Affymetrix gene expression data sets from GEO, using principal-components analysis (PCA) to identify ‘transcriptional components’, which each capture a part of the variance seen in gene expression across samples[1]. Using this test set of samples, the authors developed a method for extracting biological information about the regulatory program of the samples. They then used this method to analyze expression data from 16,172 tumor samples for cancer biology discovery.

The vast amounts of biological big data—genomic, transcriptomic, proteomic and epigenomic—available through public repositories are a potential source for novel biological discoveries. To make these discoveries, however, bioinformatic tools are needed to integrate the different data types and platforms. There have been efforts to create processed public data resources for the scientific community[2–4], which require extensive investment in data collection, curation and processing. There are also studies integrating expression data sets from GEO to make new discoveries. For example, expression compendia integration identified the conditional activity of expression modules in cancer[5], expression outlier analysis predicted the frequent fusion of the *TMPRSS2* and *ETS* transcription factor genes in prostate cancer[6] and mutual information has been used to infer post-translational modulators of transcription factor activity[7]. The current study by Fehrmann *et al.* represents a fresh angle for big data integration and novel discovery[1].

Peng Jiang and X. Shirley Liu are in the Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard School of Public Health, Boston, Massachusetts, USA.
e-mail: xsliu@jimmy.harvard.edu

Landscape of mRNA profiles

Using PCA, Fehrmann *et al.* identified principal components (PCs), which they refer to as transcriptional components, from public gene expression profiles (**Fig. 1a**). Each PC explained a portion of the total variation in gene expression across samples. Understandably, some of the PCs reflect technical artifacts, and these components can be used to remove batch effects. However, if some of the PCs contain high-coefficient genes that are known to be associated with a certain biological process, then other genes with similarly high PC coefficients might also be involved in this process. The authors used their PCA approach, combined with gene set enrichment analysis, to build a model of the regulatory network of 19,997 genes, which they used to predict the biological function of some genes within the network.

Through painstaking comparison with other methods, the authors demonstrate the superior performance of their PCA approach and make some new discoveries about the gene regulatory network. For example, they found that *FEN1* had coefficients similar to those of *BRCA1* and *BRCA2* across PCs (**Fig. 1b**). *FEN1* was previously known to be involved in DNA repair, and *BRCA1* and *BRCA2* are well known as mediators of homologous recombination in DNA damage repair. Through guilt-by-association analysis, Fehrmann and colleagues predicted that, like *BRCA1* and *BRCA2*, *FEN1* had a function in homologous recombination–mediated repair. The authors experimentally validated this prediction, showing that *FEN1* inactivation impaired homologous recombination–mediated repair in human cells.

The authors also demonstrate the use of their PCA approach to identify somatic copy number alterations (SCNAs) by locating neighboring genes on a chromosome with consistently higher or lower coefficients in one PC (**Fig. 1c**). This approach is based on the finding that coordinated aberrations in expression for nearby genes suggest the presence of SCNAs[8]. The association of PCs with SCNAs was only observed in human samples derived from cancer tissues or cell lines; non-tumor samples and samples from rodents did not show this association. On the basis of these observations, the authors developed a computational method, termed ‘functional genomic mRNA’ (FGM) profiling that uses non-genetic transcriptional components to correct raw expression data, and they used this method to determine the landscape of genome-wide SCNAs in cancer samples. The authors also derived a genome instability value for each sample, which was used to measure the overall degree of genome-wide SCNA (or total functional aneuploidy). In comparison to a previous study[8], Fehrmann *et al.* had improved power to detect associations with genomic instability, likely owing to

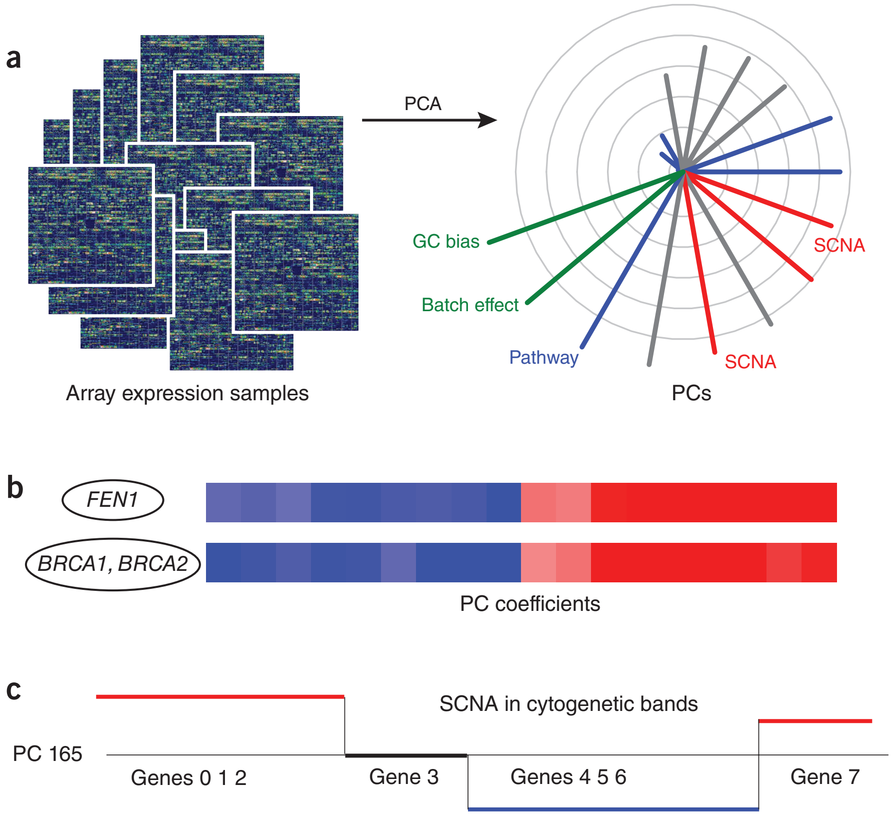

Figure 1 PCA of gene expression. (**a**) PCA is applied to array profiles to decompose them into different PCs. (**b**) Pearson correlation of PC coefficients helps infer genes with similar functions. (**c**) Consecutive genes on a chromosome with high or low PC coefficients reflect SCNAs in cancer samples.

their use of the corrected expression levels (FGM profiles) to better reflect copy number variations.

Finally, the authors show that a higher degree of genome instability is correlated with progression-free survival in ovarian cancer. The genes they identify as being associated with genome instability might be used in future studies to help predict tumor sensitivity to DNA-damaging chemotherapies and to eventually develop new therapeutics (or repurpose existing ones).

Perspectives on integrating big data

As a general limitation of studies with large-scale public data sets, labor-intensive work is required to annotate sample metadata and to process and normalize different data types. Organizing the vast number of samples, derived from diverse biological conditions, will be extremely challenging for any single laboratory. The current study by Fehrmann *et al.*, despite its impressive scale, uses less than 10% of all GEO expression profiles[1]. The study was designed to use data obtained from Affymetrix chip expression profiling experiments, although the majority of GEO expression data were not obtained with Affymetrix chips. Furthermore, only human, mouse and rat data were considered in the current study. In the future, bioinformatic tools allowing for integration across multiple sample types, along with better database organization, ontology structures and crowdsourcing, will likely facilitate easier knowledge mining and yield more biological discoveries.

The current study focuses on integrating public gene expression data from microarrays. Many other public data resources are or are becoming available. These include the 1000 Genomes Project[9], the Encyclopedia of DNA Elements (ENCODE) Project[10], Roadmap Epigenomics[11], Genotype-Tissue Expression (GTEx), the Library of Integrated Network-Based Cellular Signatures (LINCS) and The Cancer Genome Atlas (TCGA)[12]. There are also tools developed to conduct integrative data modeling (Paradigm[13], Cistrome[14] and Oncomine[3]) and visualization (UCSC Genome Browser, WashU Epigenome Browser[15] and cBioPortal[4]). All of these resources can uncover different aspects of biological systems. As any biological process is a cooperative progression of the whole cell system, systematically integrating different types of public data sets will yield a deeper understanding of both normal biology and disease.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Fehrmann, R.S. *et al. Nat. Genet.* **47**, 115–125 (2015).
2. McCall, M.N. *et al. Nucleic Acids Res.* **42**, D938–D943 (2014).
3. Rhodes, D.R. *et al. Neoplasia* **6**, 1–6 (2004).
4. Cerami, E. *et al. Cancer Discov.* **2**, 401–404 (2012).
5. Segal, E., Friedman, N., Koller, D. & Regev, A. *Nat. Genet.* **36**, 1090–1098 (2004).
6. Tomlins, S.A. *et al. Science* **310**, 644–648 (2005).
7. Wang, K. *et al. Nat. Biotechnol.* **27**, 829–839 (2009).
8. Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N. & Szallasi, Z. *Nat. Genet.* **38**, 1043–1048 (2006).
9. 1000 Genomes Project Consortium. *Nature* **491**, 56–65 (2012).
10. ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
11. Bernstein, B.E. *et al. Nat. Biotechnol.* **28**, 1045–1048 (2010).
12. Cancer Genome Atlas Research Network. *Nat. Genet.* **45**, 1113–1120 (2013).
13. Vaske, C.J. *et al. Bioinformatics* **26**, i237–i245 (2010).
14. Liu, T. *et al. Genome Biol.* **12**, R83 (2011).
15. Zhou, X. *et al. Nat. Methods* **8**, 989–990 (2011).