clinical data, providing a platform to present these data in an integrated system for cancer researchers and the broader scientific community. The utility of this browser is not limited to cancer-related data. As genome-wide high-throughput data become more available, we expect such tools to be increasingly important in disease research.

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

Jingchun Zhu[1,5], J Zachary Sanborn[1,5], Stephen Benz[1,5], Christopher Szeto[1], Fan Hsu[1], Robert M Kuhn[1], Donna Karolchik[1], John Archie[1], Marc E Lenburg[2], Laura J Esserman[3], W James Kent[1], David Haussler[1,4] & Ting Wang[1]

[1]Center for Biomolecular Science and Engineering, University of California at Santa Cruz, Santa Cruz, California, USA. [2]Department of Pathology and Laboratory Medicine, Boston University Medical School, Boston, Massachusetts, USA. [3]Department of Surgery, University of California at San Francisco, San Francisco, California, USA. [4]Howard Hughes Medical Institute, University of California at Santa Cruz, Santa Cruz, California, USA. [5]These authors contributed equally to this work. e-mail: tingwang@soe.ucsc.edu or haussler@soe.ucsc.edu

1. Karolchik, D. *et al. Nucleic Acids Res.* **36**, D773–D779 (2008).
2. Kent, W.J. *et al. Genome Res.* **12**, 996–1006 (2002).
3. Chin, K. *et al. Cancer Cell* **10**, 529–541 (2006).
4. Chin, S.F. *et al. Genome Biol.* **8**, R215:1–R215:17 (2007).
5. TCGA Research Network *Nature* **455**, 1061–1068 (2008).

mzAPI: a new strategy for efficiently sharing mass spectrometry data

To the Editor: The call for data-access standards in mass spectrometry–based proteomics has led to proposals focused on the extraction of native data to XML-based formats[1,2]. Although self-describing and human-readable formats are laudable goals, particularly for archival purposes, they are not well-suited to large numeric datasets. Consequently, whereas metadata in mzML (ref. 2) remain human-readable, the vast majority of the file is devoted to a hexadecimal representation of the mass spectra. Moreover, the transition from mzXML to a true XML format (mzML) eliminates embedded indexing schemes; consequently, extracted files are compromised in both content and access efficiency[1,3].

Based on similarities in data structure and access patterns, we suggest that fields such as astronomy are better models for proteomics data analysis (**Fig. 1**). These fields also rely on common file formats but typically use binary standards such as HDF (http://www.hdfgroup.org/) or netCDF[4]. In contrast, the commercial nature of mass spectrometry has led to the evolution of proprietary binary file formats. In light of these observations, we propose that a common and redistributable application programming interface (API) is a more viable approach to data access in mass spectrometry. In effect, we propose to shift the burden of standards compliance to the manufacturers' existing data-access libraries.

Although our proposal for abstraction through a common API is a clear departure from current attempts to define a universal file format, it is by no means unique within the broader scientific community. For example, standardized APIs have enabled the development of portable applications in such diverse areas as computer graphics (OpenGL; http://www.opengl.org/) and parallel processing (message passing interface, MPI; http://www-unix.mcs.anl.gov/mpi/). More importantly, we believe that a common API will benefit all stakeholders. For example, free redistribution of compiled, vendor-supplied dynamically linked libraries will protect the proprietary layout of native files and provide users with direct and flexible access to data system–specific and instrument-specific functionality, which are typically ignored by lowest common denominator export routines. In addition, mzAPI naturally supports US Food and Drug Administration's regulatory requirements for electronic records (http://www.fda.gov/ora/compliance_ref/Part11/). Finally, a community-driven API standard will facilitate manufacturer support of UNIX, in addition to Windows, by highlighting the subset of procedures, from each data system (Thermo Fisher's Xcalibur, Applied Biosystems-SCIEX's Analyst and others), which need to be ported.

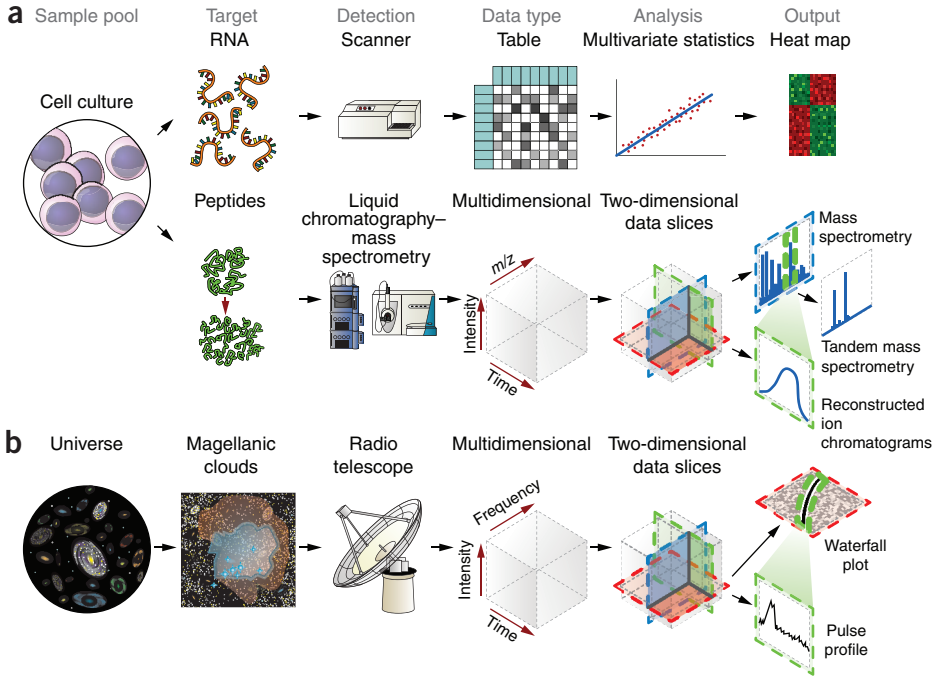

Figure 1 | Array scanners, telescopes and mass spectrometers: XML, HDF or API? (**a**) Although both DNA and protein can be extracted from biological samples, subsequent large-scale analyses (microarray or proteomics) yield data structures that diverge with respect to typical access patterns. For example, features detected by array scanners can be exported to a tabular format, immediately suited for clustering (or other multivariate analysis; top). In contrast, liquid chromatography–mass spectrometry experiments for proteomics generate complex multidimensional data, in which feature characterization is itself still an active area of research. The underlying raw data are repeatedly accessed as two-dimensional slices, reconstructed ion chromatograms, for example, and hence requires inclusion of indices in file formats designed to accommodate large-scale experimental results (bottom). (**b**) A similar situation exists in fields outside biomedical research, where, for example, slices taken through radiofrequency data yield waterfall plots and pulse profiles, which are used to characterize signals of astrophysical origin[5].

Motivated originally by our desire to provide a more intimate environment for flexible and in-depth exploration of mass spectrometry data, particularly from low-throughput experiments, we developed a preliminary common API (mzAPI) consisting of just five procedures (http://blais.dfci.harvard.edu/mzapi/). To maximize accessibility we exposed mzAPI in the form of a Python library within a flexible, mass-informatics desktop framework called multiplierz (http://blais.dfci.harvard.edu/multiplierz/). We are encouraged by results from this test harness, in particular how well mzAPI and our desktop environment support a variety of data-analytic operations. Equally impressive is how quickly nonprogrammers can customize scripts for their specific tasks. Despite success to date in our laboratory, we recognize that mzAPI will benefit from additional refinement and stress testing. Accordingly, we are actively seeking input from the research community with respect to both concept and implementation of a comprehensive API-based standard for mass spectrometry data access and analysis.

Manor Askenazi[1–3], Jignesh R Parikh[1,4] & Jarrod A Marto[1,2]

[1]Department of Cancer Biology and Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Biological Chemistry, The Hebrew University of Jerusalem, Jerusalem, Israel. [4]Bioinformatics Program, Boston University, Boston, Massachusetts, USA.
e-mail: jarrod_marto@dfci.harvard.edu

ACKNOWLEDGMENTS

We thank Y. Zhang and S. Ficarro for valuable discussion and input, and E. Smith for preparing the figure. This work was supported by the Dana-Farber Cancer Institute and the National Human Genome Research Institute (P50HG004233).

1. Pedrioli, P.G. *et al. Nat. Biotechnol.* **22**, 1459–1466 (2004).
2. Deutsch, E. *Proteomics* **8**, 2776–2777 (2008).
3. Lin, S.M. *et al. Expert Rev. Proteomics* **2**, 839–845 (2005).
4. Rew, R. & Davis, G. *Computer Graphics and Applications, IEEE* **10**, 76–82 (1990).
5. Lorimer, D.R. *et al. Science* **318**, 777–780 (2007).