Adaptive informatics for multifactorial and high-content biological data

Bjorn L Millard[1], Mario Niepel[1], Michael P Menden[1,3], Jeremy L Muhlich[1] & Peter K Sorger[1,2]

Whereas genomic data are universally machine-readable, data from imaging, multiplex biochemistry, flow cytometry and other cell- and tissue-based assays usually reside in loosely organized files of poorly documented provenance. This arises because the relational databases used in genomic research are difficult to adapt to rapidly evolving experimental designs, data formats and analytic algorithms. Here we describe an adaptive approach to managing experimental data based on semantically typed data hypercubes (SDCubes) that combine hierarchical data format 5 (HDF5) and extensible markup language (XML) file types. We demonstrate the application of SDCube-based storage using ImageRail, a software package for high-throughput microscopy. Experimental design and its day-to-day evolution, not rigid standards, determine how ImageRail data are organized in SDCubes. We applied ImageRail to collect and analyze drug dose-response landscapes in human cell lines at single-cell resolution.

It is widely accepted that biomedical data should be machine-readable and web-accessible. Relational database management systems[1,2] have proven highly effective with sequence data that are string-based, invariant in organization and interpretable without knowledge of the experiments, instruments or algorithms used to gather them. It has proven more difficult to manage data arising from complex biochemical measurements, imaging, flow cytometry and phenotypic assays of cells and tissues. The interpretation of these data, which are often unstructured (for example, images), is critically dependent on experimental context, and this context changes frequently. The difficulty in developing satisfactory database solutions for 'high-content' data is widely ascribed to insufficient standardization or poor implementation[3], but we believe the problem is more fundamental: it reflects the impossibility of fully specifying a priori complex experimental designs. Flexible and creative design is the essence of good experimental science, and because design determines data structure (the number of time points, repeats, conditions and others), structures frequently change (**Fig. 1a**). To accommodate these changes, database schema must be reconfigured frequently, a complex and time-consuming task. Thus, most experimental data reside in unlinked, loosely annotated spreadsheets that are easily fragmented or lost[4,5]. When data scope and complexity demand a more capable repository, a new database is often created *ad hoc*.

As an illustrative problem in biological data management, we focus here on high-throughput, high-content microscopy[6,7]. Microscopy presents two distinct data management challenges. One is the sheer size of the data, which can exceed many terabytes per month. The second involves the difficulties of working with numerical data extracted by image analysis, which can include a large number of data types that have complex relationships to each other (for example, the boundaries and intensities of cells or compartments and computed features such as nuclear translocation; **Fig. 1b**)[8]. For example, a typical genome-wide RNAi screen might generate ~7×10^5 images (~1.3 terabytes of data); analysis would increase the size only modestly (by ~100 megabytes), but the number of data entries would increase from ~10^6 images to >10^9 features (**Supplementary Fig. 1**). Conventional spreadsheets and comma-separated value (CSV) files perform poorly with 10^9 data entities, and relational databases impose the organizational costs described above.

Here we propose a potential solution to the challenge of managing high-dimensionality biomedical data based on the use of semantically typed data hypercubes (SDCubes) in which binary data are stored in hierarchical data format 5 (HDF5; http://www.hdfgroup.org/HDF5/), and metadata and data ontologies are stored in extensible markup language (XML; http://www.w3.org/standards/xml). We created a new open-source Java library, the SDCube programming library (**Supplementary Software 1**; http://www.semanticbiology.com/software/sdcube/) that can create SDCubes with appropriate dimensionality, encode the data model in a machine-readable XML ontology and reformat SDCubes as needed when experiments change (**Fig. 2a**). To illustrate the use of SDCubes, we created a second program, ImageRail (**Supplementary Software 2**; http://www.semanticbiology.com/software/imagerail/), for high-content microscopy that (i) segments images of cells grown in 96- and 384-well plates to extract features such as cell shape and size, (ii) stores experimental metadata and results of image analysis in SDCubes, (iii) computes sets of cellular features from the image (for example, fluorescence and localization metrics), and (iv) displays metadata,

[1]Center for Cell Decision Processes, Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Present address: Department of Biotechnology and Bioinformatics, University of Applied Sciences Weihenstephan-Triesdorf, Freising, Germany. Correspondence should be addressed to P.K.S. (peter_sorger@hms.harvard.edu).

RECEIVED 1 FEBRUARY; ACCEPTED 22 MARCH; PUBLISHED ONLINE 24 APRIL 2011; DOI:10.1038/NMETH.1600

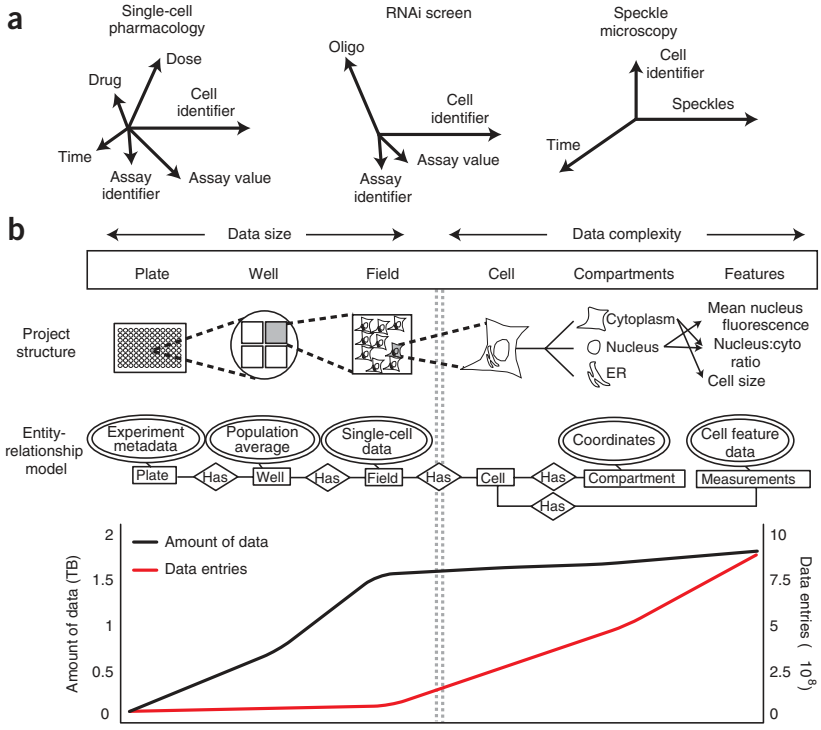

Figure 1 | Challenges in management of multidimensional data. (**a**) Schematic representation of various experimental protocols to illustrate that experimental design is the key determinant of data dimensionality (numbers and types of dimensions, which are represented by axes whose length represents magnitude). In the case of single-cell pharmacology, this includes the selection of drug, dose, time and type of assay and number of cells analyzed. (**b**) Schematic of high-throughput immunofluorescence experiments managed by ImageRail, which have a hierarchy describable in an entity-relationship model. The graph shows the amount of data in terabytes during the illustrated steps in the image processing workflow and the number of individual data entries.

images and analysis in various formats[9]. By using SDCubes, ImageRail can organize data according to the design of an experiment and its day-to-day evolution rather than an inflexible, predetermined schema. We used these tools to characterize the responses of tumor cells to therapeutic small molecules and show that the apparent half-maximal inhibitory concentration (IC_{50}) for receptor inhibitors varies with ligand dose, that cell-to-cell variability is maximal as ligands and drugs approach concentrations likely to be encountered *in vivo*, and that variance impacts the shape of dose-response curves. Our results suggest that monitoring variance will be broadly useful in preclinical pharmacology. Moreover, because flow cytometry and multiplex biochemistry have similar workflows to imaging[4,10], ImageRail and the SDCube programming library are starting points for managing diverse experimental data.

RESULTS

Managing complex and heterogeneous data using SDCubes

HDF5 files can contain both structured and unstructured data, can encode data hierarchically using 'groups' (analogous to file system folders), are unlimited in size and can be opened progressively using software libraries that read and write selected slices of data. The latter feature is critical for files that exceed the size of physical memory. To date, HDF5 has been used primarily in observational sciences (particularly remote Earth sensing) involving highly standardized data collection and little or no directed perturbation of the system under study. It has been suggested that HDF5 might be applied to biological imaging[11], but no practical implementations exist, and HDF5 alone appears to be insufficient to meet the challenges of biological experiments involving complex perturbations such as gene knockdown, drug and ligand dose response, pulse-chase and others. SDCubes address this challenge by encoding the design of perturbation-rich experiments in XML and using the design to create HDF5 files of appropriate dimensionality. A two-format solution is needed because XML is ill-suited for storage of large numerical datasets, and HDF5 lacks easy integration with 'minimum information' standards such as minimum information for biological and biomedical investigations (MIBBI)[12] and other web-based ontologies.

The HDF5 component of an SDCube is composed of basic data modules, each of which contains the HDF5 groups 'data', 'meta', 'raw' and 'children' (**Fig. 2b**). The data group contains measured or computed data stored in *N*-dimensional arrays; the meta group contains metadata such as plate address, sample identifiers and the SDCube XML file; and the raw group contains original CSV, TIFF, FCS and other primary data as byte arrays. The children group allows creation of nested data modules, each containing progressively more detailed information (**Fig. 2c**). The top-level children group is special in that it is always organized by 'sample', a label identical to 'experiment' in the minimum information about a cellular assay standard[12].

The XML component of SDCubes contains four types of information: (i) standard metadata (for example, investigator and research group); (ii) experimental protocol (for example, information on cell lines and reagents, in formats conforming to MIBBI standards when possible); (iii) experimental design (for example, species and other variables in the protocol, such as time or perturbation, that are applicable to each sample); and (iv) the identities of algorithms and free parameters used during conversion of raw data into useful experimental measurements (Online Methods). Using methods in the SDCube programming library (**Supplementary Note 1**), new samples, dimensions or assays can be appended to or inserted into an existing SDCube simply by modifying the XML file and adding to the children group at the top level of the HDF5 hierarchy (**Fig. 2d**). SDCubes are adaptable to a variety of biological assays (**Supplementary Fig. 2**) and can be combined to aggregate data from other SDCubes or divided up to create data subsets.

Implementing the HDF5-XML SDCube standard in ImageRail

ImageRail is a standalone program for high-throughput image analysis that creates and manipulates SDCubes and serves as a test of the concepts outlined above. ImageRail has four software components. First, formatting tools create and modify SDCubes so that the children group is formatted to create a five-level

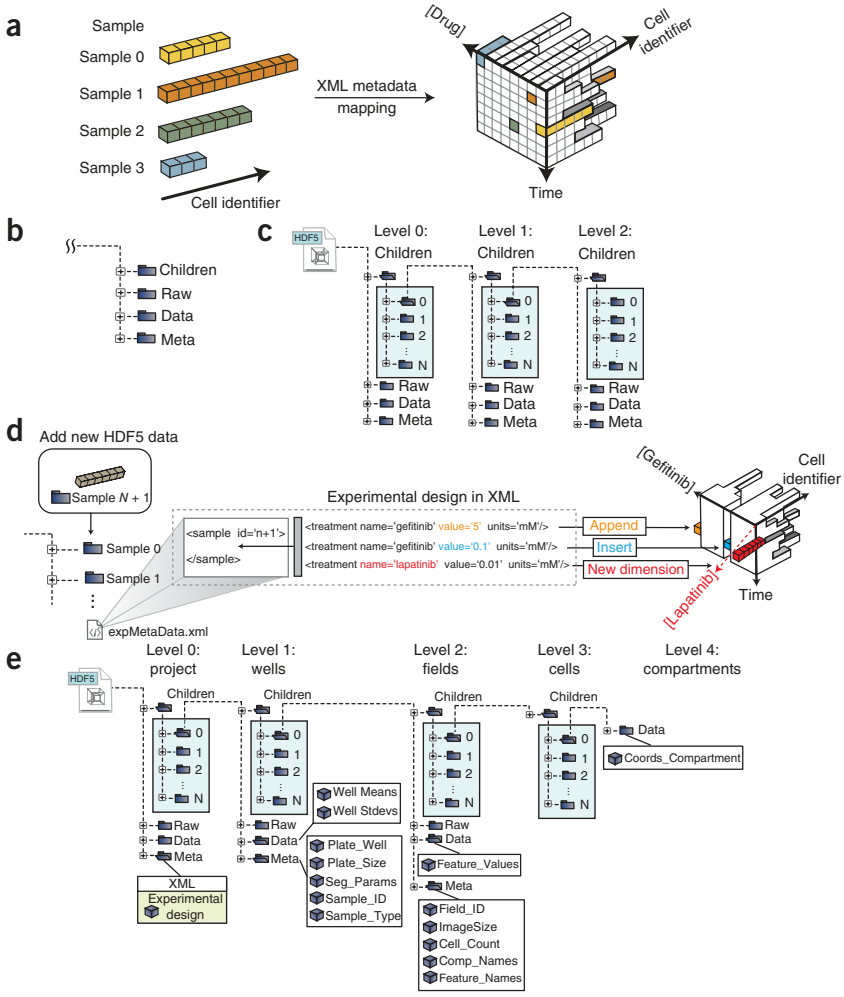

Figure 2 | SDCubes are built from a collection of linked data modules that can encode diverse experimental data with varying requirements. (**a**) Schematic of XML metadata mapping the HDF5 encoded data samples onto a biological data space. Data from each cell are represented by colored boxes, and different numbers of cells are collected for each condition. (**b**) The SDCube data module is composed of four HDF5 groups, each storing a different type of data. (**c**) The 'children' group in each module can contain additional data modules, generating an arbitrarily complex data tree. (**d**) A previously defined SDCube can be modified to append a new data sample to the end of an existing series, insert data into the middle of a series or add a new type of data that requires addition of a new dimension (in this case, use of lapatinib). All three operations are performed by modifying the XML file while recording the data in the appropriate place in the HDF5 file hierarchy. (**e**) ImageRail uses a five-level SDCube encoding high-throughput fixed-cell imaging data and progressively increasing detail (project, well, field, cell and compartment).

data hierarchy comprising project, plate, well, (image) field, cell and (cellular) compartment (conforming to the entity-relationship model in **Figs. 1b** and **2e**). Dropdown lists and a graphical user interface for highlighting wells make it possible to specify which experimental conditions map to which wells, thereby specifying the experimental design and SDCube dimensionality, and creating XML annotation (**Fig. 3a**). Second, image analysis tools create and store segmentation masks based on standard algorithms for cell monolayers, which can be extended using existing software such as ImageJ[13] (**Fig. 3b**). Third, data viewers display raw data and computed features as images, line plots, histograms, scatter plots and multiwell plate views. Scatter plotting includes multidimensional gating similar to that used for analysis of flow cytometry data (**Fig. 3c**). Finally, embedded routines enable dynamic linking of data points to specific image features. Dynamic linking allows users to highlight cells in an image that correspond to selected data points in a scatter plot (**Fig. 3b**,**c**), facilitating the identification of outliers and experimental artifacts such as bubbles, tissue-culture debris or edge effects (**Supplementary Fig. 3**). Users choose the level of detail at which to store the link between segmentation and data; at one extreme, pixel-by-pixel information can be stored, but we generally find it more useful to store either the centroid of each cell or a bounding box (**Fig. 3d**). Although the SDCube data group 'raw' can store image data, we are in the process of integrating ImageRail with the open microscopy environment remote objects (OMERO) image server[14]. Thus, ImageRail currently stores TIFF files alongside SDCubes and not within them. OMERO provides powerful tools for processing and organizing images, is used widely in open-source and commercial image management applications[14] and OME-TIFF has found wide acceptance as a file standard for biological microscopy[15].

Monitoring cell-to-cell variability in drug responses

It is widely hypothesized that variability in cellular responses to drugs and the presence of drug-resistant cell subpopulations can impact cancer therapy[16]. One application of ImageRail is to systematize single-cell drug-response studies and uncover the origins and importance of variability. Our proof-of-principle studies focused on the effect of changes in the concentration of epidermal growth factor (EGF) on the IC_{50} of the ATP-competitive EGF receptor (EGFR) inhibitor gefitinib[17]. We assayed inhibition by immunofluorescence microscopy, using antibodies to the downstream kinase ERK1/2 phosphorylated at Thr202 and Tyr204 (henceforth called ppERK). EGFR mutation and overexpression are implicated in a wide range of tumors[18] and gefitinib is used clinically to treat lung, colorectal and other cancers[19,20].

Here we exposed cells to EGF at ten doses over a 10^4 concentration range in combination with gefitinib at eight doses over a 10^3 range using a simple adaptive design in which each 96-well plate was subjected to a different and changeable set of treatments and measurements. To enable image segmentation with a standard watershed algorithm, we treated cells with nuclear and cytoplasmic stains (**Supplementary Fig. 4**). The dataset comprised 160 conditions, 1.4×10^6 individual cells and an SDCube with 2.8×10^6 entries (a tenfold larger dataset is shown in **Supplementary Fig. 5**). By accessing different slices of the cube, we can view data as IC_{50} curves at various EGF

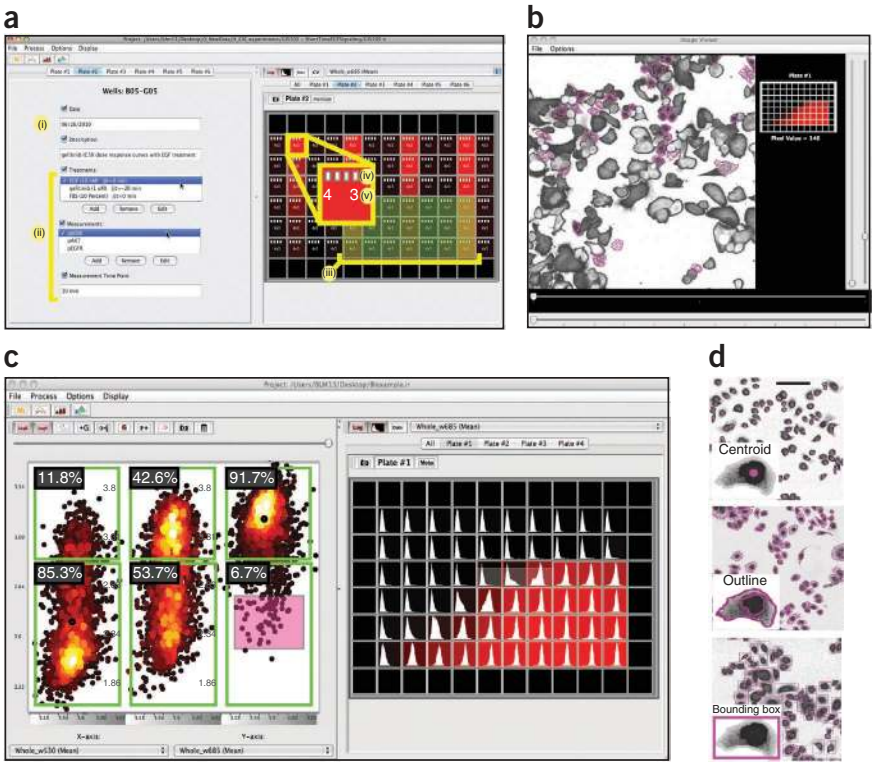

Figure 3 | Annotated and simplified screen shots from ImageRail software. (**a**) Screenshot showing general experiment metadata (i) and computable information derived from image analysis across perturbations and measurements (ii) associated with selected wells of a microtiter plate (iii). White document icons represent the number of image fields that have single-cell data stored in the HDF5 file available for analysis (iv), and numbers represent imaged fields and wavelengths (v). (**b**) Screenshot illustrating dynamic linking of extracted data to the source images that shows which cells gave rise to which measurements and is implemented using an image viewer and scatter plot (red box in **c**). (**c**) Screenshot illustrating data visualization, which includes single-cell scatter plots with flow cytometry–style gating (left) and plate heatmaps of population averages along with a representation of the underlying single-cell distributions (right). (**d**) Results of image segmentation can be stored in different ways, including centroid, outline and bounding box. Scale bar, 100 µm.

concentrations or as EGF dose-response curves at various drug concentrations; cell-to-cell variability can also be visualized at any point (**Fig. 4a**). We observed that average amounts of ppERK increased with increasing [EGF] and decreased with increasing [gefitinib], and that the apparent IC_{50} was sensitive to EGF concentration, varying ~20-fold as exogenous EGF varied from 0 ng ml^{-1} to 100 ng ml^{-1} (**Fig. 4b**). Well-average data computed from images closely matched dose-response data obtained using conventional biochemical assays (**Supplementary Fig. 6**). The relationship between IC_{50} and [EGF] varied substantially with cell type (**Fig. 4c**): whereas IC_{50} was strongly sensitive to [EGF] in SKBR3 and T47D cells, it was less so in MCF7 cells (**Supplementary Fig. 7**). Data exploration of this type is intuitively simple but involves the manipulation of many data entries; because HDF5 successively loads data, there is no limit a priori to the number of entries, and we validated ImageRail with ~10^8–10^9 data points.

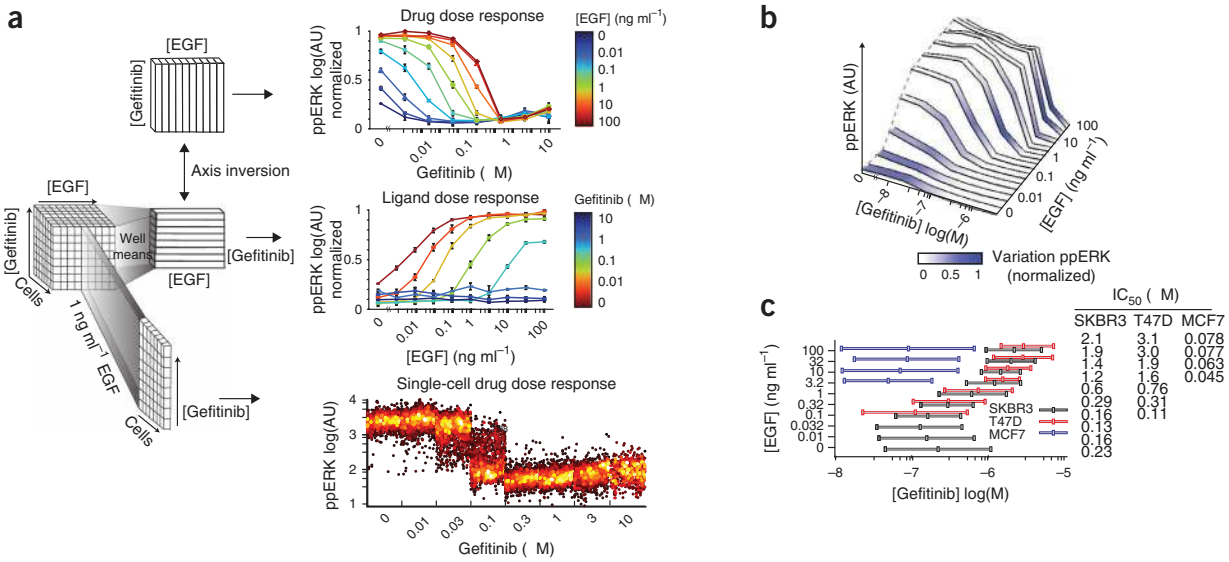

Figure 4 | Exploring different dimensions of a multivariate drug and ligand dose-response series using SDCubes. (**a**) Well-mean values are computed from single-cell data recorded from cultured SKBR3 cells exposed to exogenous EGF for 10 min over a range of concentrations and then stained with antibodies to ppERK. Error bars, s.e.m. of biological triplicates. Plotted data show conventional drug dose-response relationships at different ligand concentrations (top). Inverting the axes allows the same data to be plotted as a ligand dose-response curve at different drug doses (middle). For each mean value in either plot, the underlying single-cell distribution can be visualized as scatter-plots (bottom; gefitinib dose-response at 1 ng ml^{-1} EGF). (**b**) ppERK signal in SKBR3 cells treated as in **a** and colored according to the degree of cell-to-cell variation, with 1 being a high coefficient of variation. (**c**) Whisker plots of gefitinib IC_{10}, IC_{50} and IC_{90} values for the inhibition of ERK phosphorylation by gefitinib in SKBR3, T47D and MCF7 cells treated for 10 min with a range of EGF concentrations.

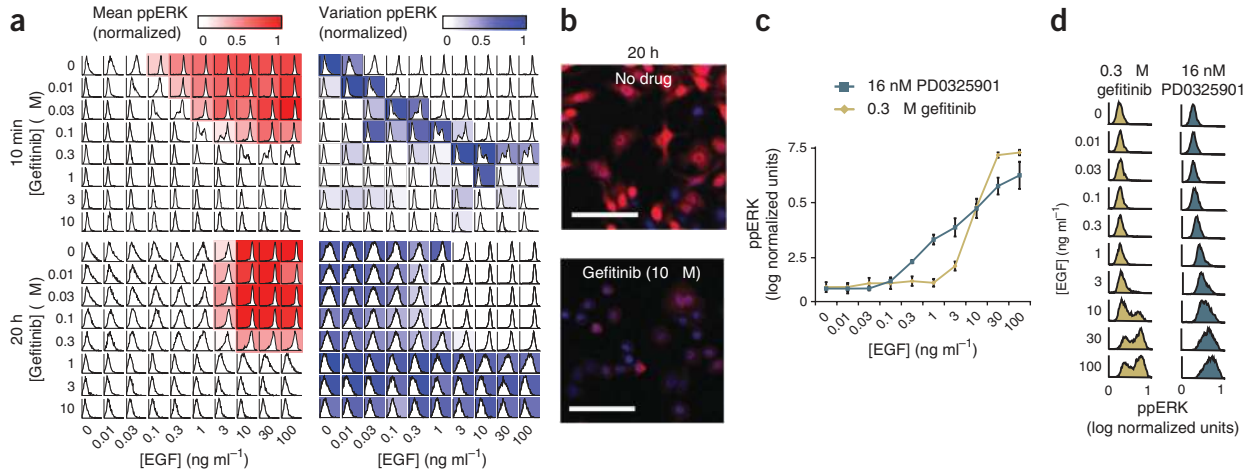

Figure 5 | Single-cell analysis of drug-ligand dose responses uncovers cell-to-cell heterogeneity. (**a**) ppERK signal was measured at 10 min and 20 h in SKBR3 cells treated with indicated doses of EGF and gefitinib. Heat maps of the mean values and coefficients of variation of the underlying cell population histograms are overlaid on a representation of a standard 96-well microtiter plate. (**b**) Selected immunofluorescence images of ppERK staining (red) and Hoechst staining (blue) of cells 20 h after exposure first to 10 μM gefitinib and then to 100 ng ml^{-1} EGF. Scale bars, 100 μm. (**c**) EGF-induced ppERK dose-response curves in SKBR3 cells pretreated with subsaturating doses of gefitinib or the MEK inhibitor PD0325901. Error bars represent the s.e.m. of biological triplicates. (**d**) Single-cell distributions for the population mean data shown in **c**.

On comparing mean ppERK levels with cell-to-cell variance using plate maps (**Fig. 5a**), we observed maximum variability at physiologically relevant doses of drug and ligand (estimated to be 0.1–1.5 ng ml^{-1} for EGF and 0.4–50 μM for gefitinib[21,22]). Mean value and variance in response changed over time, such that 20 h after EGF and gefitinib treatment, IC_{50} was less dependent on [EGF] but the variance increased. By linking back to the underlying images, we observed that even in cells exposed to saturating doses of gefitinib (10 μM) for 20 h, a subpopulation of cells (~1%) had elevated ppERK levels. This implies not only that these cells were drug-resistant but also that ERK signaling could be sustained in the absence of exogenous ligand (a behavior different from that of cells that are simply gefitinib-insensitive; **Fig. 5a**,**b**). Thus, single-cell data revealed three interesting features of cellular responses to gefitinib and EGF. First, IC_{50} varied with the concentration of extracellular ligand, particularly at early time points. Second, the extent of cell-to-cell variability was maximal near intermediate, physiologically relevant concentrations; conversely, it was masked when drug or ligand were added at high levels. Third, cell-to-cell heterogeneity changed over time, being dominated initially by broad distributions and subsequently by rare cells with sustained signaling. Whether the differences we observed are genetic[23], epigenetic[24] or stochastic[25] in origin is not yet clear, but reversibility implies that some are indeed stochastic, as we have previously demonstrated for TNF-responsive apoptosis-inducing ligand (TRAIL)[25].

The shape of dose-response curves for drugs and ligands often depends on the agent and cell type (**Fig. 4c** and **Supplementary Fig. 7**). Gefitinib dose response of cells exposed to EGF conformed to a sigmoidal shape as expected for simple ligand-receptor binding, but the dose-response for an inhibitor of MEK kinase (PD0325901)[26], an enzyme acting immediately upstream of ERK1/2 kinases in EGFR signaling, was nearly linear over a 10^3 [EGF] range (**Fig. 5c**). At the single-cell level, responses to gefitinib were bimodal, with low ppERK levels in some cells and 100-fold higher levels in other cells, but responses to PD0325901 were continuous, with cells exhibiting a wide range of activities (**Fig. 5d**). We conclude that the mean-value dose-response curves for PD0325901 and gefitinib differed because of variability at the single-cell level and speculate that this might be a general explanation for nonsigmoidal dose-response relationships.

DISCUSSION

By creating a lightweight data repository customized to the design of a specific experiment and then storing the design in a machine-readable XML format, the SDCube programming library places experimental design foremost in organizing data for storage. The use of XML to encode ontologies simplifies harmonization with existing web-based standards[12], and the use of HDF5 allows progressive access to even very large files. As the design changes or expands, the dimensionality of SDCubes changes as do the metadata tags that point to specific data elements. The result is an approach to data and metadata storage that aims to address the competing demands of data integrity and flexibility. Little attention has been paid to computer-readable experimental designs, and only one public specification exists (minimum information for data analysis in systems biology; MIDAS)[27]. It is possible, however, to document the format of any hypothesis-driven or systematic experiment in XML, making it straightforward to use resource description framework (RDF) and web ontology language (OWL) to share and analyze experimental designs, a critical step in making the results from complex experiments machine-interpretable in light of their purpose and context.

Although we applied SDCubes to microscopy data using ImageRail, the SDCube format is in principle adaptable to any type of high-dimensional data, and we created preliminary schemata for multicolor flow cytometry[10] and multiplex or array-based biochemical assays[28] (**Supplementary Fig. 2**). Matlab users will recognize that some SDCube functionality is already present in Matlab, which makes extensive internal use of multidimensional data arrays (indeed, Matlab can read HDF5 files). However, Matlab files cannot duplicate key features of SDCubes: they cannot be read incrementally, their data models cannot be referenced to external ontologies or parsed using web-based tools, and Matlab is not open-source software, an important consideration for an open data standard.

We designed ImageRail to be interoperable with existing open-source image analysis software, including ImageJ, CellProfiler and OME[13–15,29]. Interoperability is important to avoid duplication of effort, but ImageRail also needs to function as a stand-alone application; hence we included common segmentation and visualization routines.

The ability of SDCubes and ImageRail to systematize data from complex dose-response experiments has made it possible to implement an efficient scheme for single-cell pharmacology. Exposing tumor cells to growth factors and kinase inhibitors in combination reveals many examples of cell-to-cell variability; some of these are likely to have nongenetic origins, by direct analogy to the variability observed in cellular responses to TRAIL[25], T-cell receptor agonists[30] and other ligands[31]. Variability is maximal at doses close to the IC_{50} of gefitinib or the half-maximal activation (EC_{50}) by EGF, precisely the doses likely to prevail in humans. It therefore seems reasonable that application of single-cell pharmacology will help to uncover the basis of fractional killing by anticancer drugs and assist in dissecting the origins of intrinsic and acquired drug resistance[32].

In many exploratory biological experiments, data collection and analysis are iterative processes undertaken by a limited number of people. In this environment, the high-integrity, multiuser, read-write operations enabled by conventional databases are unnecessary overhead, and SDCubes offer an effective alternative. But as data become more mature or an experiment nears completion, it will often be advantageous to move key results to a relational database. One way to accomplish this is to create a specialized summary view of an SDCube and then import the summary data into a database. Only data conforming to a pre-existing standard would be accessible in the database, but an SDCube containing all primary data could easily be called using a uniform resource identifier (URI, akin to a web URL). It is possible that new types of databases will be developed with science in mind (for example, SciDB), but we predict that lightweight, adaptable, file-based data storage will always coexist with server-based data management and that sophisticated file formats such as SDCubes will provide a missing link between creative experimentation and machine-interpretable data.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

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

ACKNOWLEDGMENTS

This work was supported by US National Institutes of Health grants HG006097, HG005693 and GM68762. We thank G. Danuser, T. Mitchison and M. Eisenstein for help with the manuscript; Applied Precision Inc., C. Brown and K. Teplitz for help with instrumentation; and G. Odell and J. Baker for inspiration.

AUTHOR CONTRIBUTIONS

B.L.M., M.P.M. and J.L.M. programmed the software. B.L.M., M.N., J.L.M. and P.K.S. developed the method and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
2. Maheswari, U. *et al.* The Diatom EST database. *Nucleic Acids Res.* **33**, D344–D347 (2005).
3. Pawley, J.B. *Handbook of Biological Confocal Microscopy*. 3rd edition. (Springer Science + Business Media, 2006).
4. Gaudet, S. *et al.* A compendium of signals and responses triggered by prodeath and prosurvival cytokines. *Mol. Cell. Proteomics* **4**, 1569–1590 (2005).
5. Neve, R.M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**, 515–527 (2006).
6. Conrad, C. & Gerlich, D.W. Automated microscopy for high-content RNAi screening. *J. Cell Biol.* **188**, 453–461 (2010).
7. Loo, L.H., Wu, L.F. & Altschuler, S.J. Image-based multivariate profiling of drug responses from single cells. *Nat. Methods* **4**, 445–453 (2007).
8. Snijder, B. *et al.* Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature* **461**, 520–523 (2009).
9. Gehlenborg, N. *et al.* Visualization of omics data for systems biology. *Nat. Methods* **7**, S56–S68 (2010).
10. Krutzik, P.O., Crane, J.M., Clutter, M.R. & Nolan, G.P. High-content single-cell drug screening with phosphospecific flow cytometry. *Nat. Chem. Biol.* **4**, 132–142 (2008).
11. Dougherty, M.T. *et al.* Unifying biological image formats with HDF5. *ACM Queue* **52**, 42–47 (2009).
12. Taylor, C.F. *et al.* Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.* **26**, 889–896 (2008).
13. Abramoff, M.D., Magelhaes, P.J. & Ram, S.J. Image processing with ImageJ. *Biophotonics International* **11**, 36–42 (2004).
14. Moore, J. *et al.* Open tools for storage and management of quantitative image data. *Methods Cell Biol.* **85**, 555–570 (2008).
15. Goldberg, I.G. *et al.* The Open Microscopy Environment (OME) data model and XML file: open tools for informatics and quantitative analysis in biological imaging. *Genome Biol.* **6**, R47 (2005).
16. Gupta, P.B., Chaffer, C.L. & Weinberg, R.A. Cancer stem cells: mirage or reality? *Nat. Med.* **15**, 1010–1012 (2009).
17. Ciardiello, F. *et al.* Antitumor effect and potentiation of cytotoxic drugs activity in human cancer cells by ZD-1839 (Iressa), an epidermal growth factor receptor-selective tyrosine kinase inhibitor. *Clin. Cancer Res.* **6**, 2053–2063 (2000).
18. Yarden, Y. & Sliwkowski, M.X. Untangling the ErbB signalling network. *Nat. Rev. Mol. Cell Biol.* **2**, 127–137 (2001).
19. Ciardiello, F. & Tortora, G. EGFR antagonists in cancer treatment. *N. Engl. J. Med.* **358**, 1160–1174 (2008).
20. Paez, J.G. *et al.* EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **304**, 1497–1500 (2004).
21. Blaimauer, K. *et al.* Effects of epidermal growth factor and keratinocyte growth factor on the growth of oropharyngeal keratinocytes in coculture with autologous fibroblasts in a three-dimensional matrix. *Cells Tissues Organs* **182**, 98–105 (2006).
22. McKillop, D. *et al.* Tumor penetration of gefitinib (Iressa), an epidermal growth factor receptor tyrosine kinase inhibitor. *Mol. Cancer Ther.* **4**, 641–649 (2005).
23. Turke, A.B. *et al.* Preexistence and clonal selection of MET amplification in EGFR mutant NSCLC. *Cancer Cell* **17**, 77–88 (2010).
24. Sharma, S.V. *et al.* A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell* **141**, 69–80 (2010).
25. Spencer, S.L., Gaudet, S., Albeck, J.G., Burke, J.M. & Sorger, P.K. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature* **459**, 428–432 (2009).
26. Brown, A., Carlson, T., Loi, C.-M. & Graziano, M. Pharmacodynamic and toxicokinetic evaluation of the novel MEK inhibitor, PD0325901, in the rat following oral and intravenous administration. *Cancer Chemother. Pharmacol.* **59**, 671–679 (2007).
27. Saez-Rodriguez, J. *et al.* Flexible informatics for linking experimental data to mathematical models via DataRail. *Bioinformatics* **24**, 840–847 (2008).
28. Albeck, J.G. *et al.* Collecting and organizing systematic sets of protein data. *Nat. Rev. Mol. Cell Biol.* **7**, 803–812 (2006).
29. Lamprecht, M.R., Sabatini, D.M. & Carpenter, A.E. CellProfiler: free, versatile software for automated biological image analysis. *Biotechniques* **42**, 71–75 (2007).
30. Feinerman, O., Veiga, J., Dorfman, J.R., Germain, R.N. & Altan-Bonnet, G. Variability and robustness in T cell activation from regulated heterogeneity in protein levels. *Science* **321**, 1081–1084 (2008).
31. Niepel, M., Spencer, S.L. & Sorger, P.K. Non-genetic cell-to-cell variability and the consequences for pharmacology. *Curr. Opin. Chem. Biol.* **13**, 556–561 (2009).
32. Yang, R., Niepel, M., Mitchison, T.K. & Sorger, P.K. Dissecting variability in responses to cancer chemotherapy through systems pharmacology. *Clin. Pharmacol. Ther.* **88**, 34–38 (2010).

ONLINE METHODS

Software availability. Software, user manuals and periodic updates are available at http://www.semanticbiology.com/software/. Software is also available as **Supplementary Software 1** and **2**.

Details of SDCubes. SDCubes store experimental data from various bioassays in HDF5 file format with an XML file describing how each sample was treated experimentally. Four types of information are recorded in the XML component of an SDCube, conforming to an XML schema specification. The first consists of standard information such as date, experimenter, research group and others. The second describes the experimental protocol, including information on cell lines, reagents, drugs and environmental factors. Existing minimum information standards such as minimum information about a microarray experiment and minimum information about a cellular assay (components of MIBBI) are used where possible[12], and integration of other relevant XML formats such as CML (chemical markup language)[33] will be straightforward. The third type of information relates to experimental design, that is, how variables in the protocol such as time, perturbation or assay are applied to each sample in the experiment. Design is the primary determinant of data cube dimensionality, but relatively little effort has been devoted thus far to making experimental design computer-readable. All SDCubes currently conform to our MIDAS standard for experimental design[27], but we anticipate development of other machine-readable experimental designs in the future. The fourth block of XML-encoded metadata specifies the identities of all algorithms and free parameters used during conversion of raw data into useful experimental measurements; these include algorithms for cell tracking, background subtraction, intensity normalization and descriptor calculation. Given the potential complexity and heterogeneity of such algorithms, we intend to index them in XML using URIs or persistent unique digital identifiers (for example, Digital Object Identifiers as applied to journal articles), but it is also possible simply to embed equations or software code in HDF5 files.

ImageRail software. Many image analysis software packages are already available in both the commercial and academic domains, and we designed ImageRail to be interoperable with key open-source applications (**Supplementary Table 1**). ImageRail follows an 'overlapping modular design' to create an application with new capabilities but sufficient functionality in common tasks, such as metadata entry, image segmentation, feature extraction and image and data visualization, to function as a stand-alone application. Additional functionality is acquired through the use of existing software (**Supplementary Fig. 8**). For example, ImageRail imports standard TIFF images from microscope-control software, uses Java-encoded analysis algorithms and exports slices of HDF5-XML data as CSV files for analysis by software such as Excel (Microsoft) or Spotfire (TIBCO). We also expect to build a link to CellProfiler[29] to enable use of its image-processing engine and storage of the resulting data in SDCubes. ImageRail can also export data slices in CSV-MIDAS format to be interoperable with DataRail[27], a software package we had developed previously to manipulate multidimensional biochemical data and construct regression models. DataRail currently does not use SDCubes, but we are writing a new version that will. Integration of ImageRail with other software packages requires Java programming through the use of the provided application programming interface. The overall goal of the integration effort is to leverage a rich set of existing software and to allow ImageRail to fit into existing data workflows. Conversely, interested developers can use the provided SDCube programming library to create new software for flow cytometry[34], protein arrays[35,36] and multiplex immunoassays[37] or even for nonbiological data. Refer to the ImageRail user manual (**Supplementary Note 2**) for specific software instructions.

Cell treatment and immunofluorescence staining. Cells were plated at 7,500 cells per well in 96-well microscopy plates (Corning) in recommended medium for 24 h and then starved in medium lacking serum for 16 h. Cells were pretreated for 10 min with tenfold stock solutions of gefitinib (LC Laboratories) or MEK inhibitor PD0325901 (Selleck Chemicals) and treated with tenfold stock solutions of EGF (PeproTech) for the indicated amounts of time. Cells were fixed in 2% paraformaldehyde for 10 min at room temperature (20–25 °C) and washed in PBS with 0.1% Tween 20 (Sigma-Aldrich) (PBS-T). Cells were permeabilized in methanol for 10 min at room temperature, washed with PBS-T and blocked in Odyssey Blocking Buffer (LI-COR Biosciences) for 1 h at room temperature. Cells were incubated overnight at 4 °C with antibodies to ppERK, Akt phosphorylated on Ser 473 (pSer473Akt) or cJUN phosphorylated on Ser 73 (pS73cJUN) (Cell Signaling Technology) diluted 1:400 in Odyssey Blocking Buffer. Cells were washed three times in PBS-T and incubated with rabbit-specific secondary antibody labeled with Alexa Fluor 647 (Invitrogen) diluted 1:2,000 in Odyssey Blocking Buffer. Cells were washed once in PBS-T, once in PBS and incubated in 250 ng ml^{-1} Hoechst 33342 (Invitrogen) and 1:1,000 Whole Cell Stain (blue; Thermo Scientific) solution. Cells were washed two times with PBS and imaged in an imageWoRx high-throughput microscope (Applied Precision). The microscope had a 10× objective and 12-bit camera sensor under 2 × 2 binning giving 1,024 × 1,024 pixels per image with final spatial resolution of 1.48 μm per pixel. Microscopy exposure times were 1 s for the 647 nm fluorophore (far-red) and 0.012 s for the Hoechst (blue) channel. ImageJ software was used to compute the linear color scaling of [10, 100] pixel intensity units in the red channel and [10, 1,500] pixel intensity units in the blue channel, and enable creation of the multichannel pseudo-colored images shown **Figure 5b**. Bioplex assays were performed as previously described[38]. Data for **Supplementary Figure 5** was plotted using DataPflex[39].

33. Murray-Rust, P. & Rzepa, H.S. Chemical markup, XML and the world wide web. 4. CML schema. *J. Chem. Inf. Comput. Sci.* **43**, 757–772 (2003).
34. Krutzik, P.O. & Nolan, G.P. Fluorescent cell barcoding in flow cytometry allows high-throughput drug screening and signaling profiling. *Nat. Methods* **3**, 361–368 (2006).
35. Sevecka, M. & MacBeath, G. State-based discovery: a multidimensional screen for small-molecule modulators of EGF signaling. *Nat. Methods* **3**, 825–831 (2006).
36. Wolf-Yadlin, A., Sevecka, M. & MacBeath, G. Dissecting protein function and signaling using protein microarrays. *Curr. Opin. Chem. Biol.* **13**, 398–405 (2009).
37. Alexopoulos, L.G., Saez-Rodriguez, J., Cosgrove, B.D., Lauffenburger, D.A. & Sorger, P.K. Networks inferred from biochemical data reveal profound differences in toll-like receptor and inflammatory signaling between normal and transformed hepatocytes. *Mol. Cell. Proteomics* **9**, 1849–1865 (2010).
38. Chen, W.W. *et al.* Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.* **5**, 239 (2009).
39. Hendriks, B.S. & Espelin, C.W. DataPflex: a MATLAB-based tool for the manipulation and visualization of multidimensional datasets. *Bioinformatics* **26**, 432–433 (2010).