Review

Cell PRESS

Behavioral barcoding in the cloud: embracing data-intensive digital phenotyping in neuropharmacology

David Kokel[1,2], Andrew J. Rennekamp[1,2], Asmi H. Shah[3], Urban Liebel[4] and Randall T. Peterson[1,2]

[1] Cardiovascular Research Center and Division of Cardiology, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, 149 13th Street, Charlestown, MA 02129, USA
[2] Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA
[3] Institute of Toxicology and Genetics (ITG), KIT Karlsruhe Institute of Technology, Karlsruhe, Germany
[4] Institute of Applied Informatics (IAI), KIT Karlsruhe Institute of Technology, Karlsruhe, Germany

For decades, studying the behavioral effects of individual drugs and genetic mutations has been at the heart of efforts to understand and treat nervous system disorders. High-throughput technologies adapted from other disciplines (e.g., high-throughput chemical screening, genomics) are changing the scale of data acquisition in behavioral neuroscience. Massive behavioral datasets are beginning to emerge, particularly from zebrafish labs, where behavioral assays can be performed rapidly and reproducibly in 96-well, high-throughput format. Mining these datasets and making comparisons across different assays are major challenges for the field. Here, we review behavioral barcoding, a process by which complex behavioral assays are reduced to a string of numeric features, facilitating analysis and comparison within and across datasets.

Behavior-based drug discovery

Neuroactive drugs are among the most powerful tools available for neuroscience research. Most neuroactive drugs, including the prototypes of most modern psychiatric medicines, were discovered based on their behavioral phenotypes. These discoveries were largely due to serendipitous observations, suggesting that systematic behavior-based chemical screening is likely to identify additional neuroactive drugs [1]. As more researchers combine high-throughput phenotyping with behavior-based chemical screening, a new influx of chemobehavioral data is flooding the neuropharmacology field. New data sharing and analytical tools will be needed to leverage this data deluge [2].

Historically, many behavioral phenotypes were manually analyzed and scored with textual descriptions. This 'sensor bottleneck' focused phenotypic analyses on qualitative measurements and limited the scale and utility of phenotypic datasets. Text searching is a valuable way to identify and retrieve information, but is also extremely limited in terms of accuracy, specificity, and quantitative data analysis. For example, the book 'Phenethylamines I Have Known and Loved – A chemical love story' [3], records a dedicated attempt to document the psychoactive effects of 179 phenethylamines using the Shulgin Rating scale, a semi-quantitative but subjective five-point scale ranging from (+/–), a false positive, to (++++) 'a rare and precious transcendental state'. Such ratings are enlightening, but have the drawbacks of being subjective, low throughput, and difficult to analyze mathematically.

Behavioral phenomics beyond functional genomics

Current genomic and chemical databases are excellent tools for organizing relatively small-scale phenotyping studies. But they are fundamentally limited for analyzing large-scale behavioral datasets. Searching for behavioral ontology terms in genomic databases typically retrieves tabular results with behavioral phenotypes presented as links associated with specific genes (Figure 1a,b). For example, a search for 'behavioral defective' on the *Drosophila* genomic database, FlyBase, returns 2396 alleles with links to original papers and detailed text-based phenotypic descriptions [4]. This is an excellent way to organize genomic information when the goal is to identify records that annotate a specific gene. However, large-scale systematic phenotyping generates more than a collection of individual gene annotations – it generates coherent sets of data that can be analyzed quantitatively. Most genomic and chemical databases are simply not designed for this type of behavioral analysis.

The future of digital phenotyping will invert the standard genome–browser paradigm, by focusing on digital phenotypes themselves while using genes and drugs as descriptors (Figure 1c). For example, the Mouse Phenome Database (MPD) currently catalogs about 1000 measurements relating to a wide variety of phenotypes [5]. Behavioral data is grouped into seven different types, 22 apparatuses, and several different mouse strains. Beyond the standard tabular gene-behavior format, the MPD also provides access to a tool for exploring phenotype-to-phenotype and phenotype-to-genotype correlations. The ability to easily compare quantitative relationships between two different phenotypes is an exciting application of behavioral analytics that

Corresponding author: Peterson, R.T. (peterson@cvrc.mgh.harvard.edu).
Keywords: chemobehavioral data; data-intensive digital phenotyping; dataset mining; neuroscience; pharmacology; high-thoughput screen.

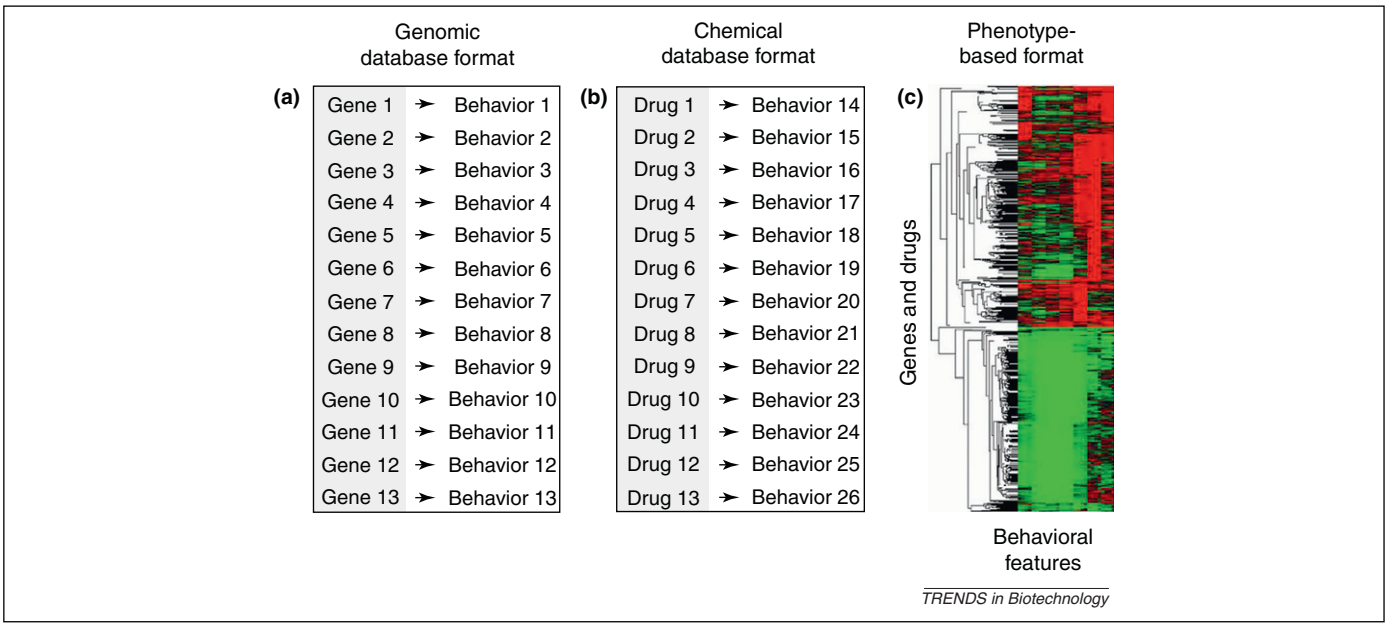


Figure 1. Behavioral phenotypes are more than just annotations. Genomic **(a)** and chemical **(b)** databases typically present behavioral data as simple annotations linked to single genes or chemicals. By contrast, a phenotype-based approach to the analysis of large behavioral datasets helps researchers to understand patterns between genes, drugs, and their behavioral effects. **(c)** Dendrogram showing genes and drugs (*y*-axis) clustered based on similarity of their behavioral features. The intensity of red and green bars in the heatmap indicates the magnitude of deviation for each behavioral feature (*x*-axis) relative to the control set.

takes behavioral data beyond annotating the genome. These types of research tools need to be developed and expanded.

Chemical biology is particularly well poised to benefit from the integration of phenomic information. High-throughput chemical screening datasets are data-rich but phenotypically poor compared to other fields of biology [6]. Large chemical databases, such as Pubchem, provide structural and activity searching for >2.5 million chemicals, including tools for structure–activity analyses and chemical structure clustering [7]. Some compounds are annotated with bioactivity links [8]. However, comprehensive behavioral phenomics remains an unrealized goal.

Data-intensive behavioral analytics

Until recently, there has been a relative scarcity of large behavioral datasets, but high-throughput technologies are beginning to generate an abundance of behavioral data. Increasingly accessible digital phenotyping systems are available for organisms ranging from humans to fruit flies [9–11]. Automated video tracking systems coupled to machine vision algorithms can monitor behavior 24 h a day. Such systems have arrived at a time when data-intensive, discovery-based research is becoming increasingly common in academia and in industry.

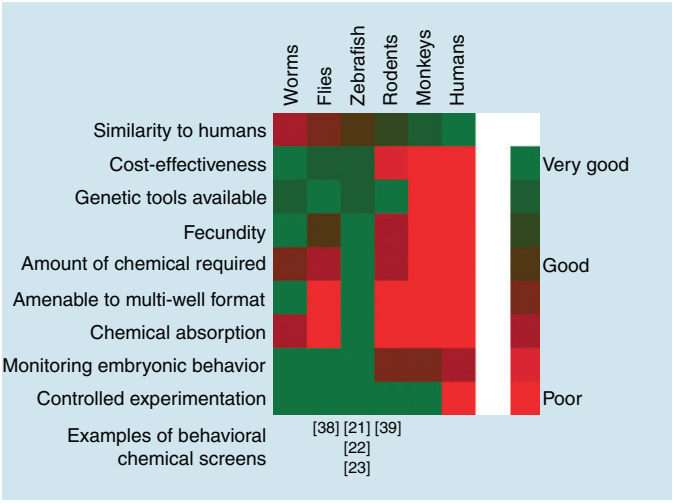
'Systems biology', 'the fourth paradigm of science', and 'big data' are terms used to describe efforts to systematically integrate technology, biology, and computation [12]. The question 'Why not measure it all?' has been applied to genomes [13]. Perhaps the time has come to take a similar approach to large-scale phenotyping. It may not be possible to comprehensively quantify all of an animal's behaviors, but researchers can collect large amounts of data describing how certain responses vary across millions of individuals and different conditions. The challenge will be devising systems to curate and analyze this behavioral data deluge.

Today, the largest behavioral databases are owned and operated by internet technology companies such as Google, Facebook, and online gaming communities. The sheer magnitude of behavioral analytics at these companies is enviable for basic science researchers in neuropharmacology. Still, complimentary efforts are being developed in research and industry to analyze animal and human behavior. In both situations, researchers are interested in categorizing a large number of responders into meaningful groups based on an even larger number of behavioral features. Fortunately, these large datasets are driving new approaches to behavioral analytics [14]. For example, Google Flu Trends correlates internet search behavior with geographical data to predict flu outbreaks [15]. Google Maps uses real-time cell phone location to analyze traffic patterns and suggest travel directions [16]. Analysis of Twitter feeds has been used to predict national moods and interests [17]. These examples provide lessons and inspiration for scaling neuropharmacology informatics in the laboratory.

Zebrafish and the chemobehavioral data deluge

The most popular models in behavioral neuroscience include humans, monkeys, rodents, *Drosophila*, and *Caenorhabditis elegans*. However, behavior-based drug discovery is inefficient using these models. Humans, rodents, and other large vertebrate animals are simply too large for cost-effective and systematic chemical screening (Table 1). By contrast, invertebrate model organisms, including *C. elegans* and *Drosophila*, are powerful models in developmental biology and genetics. Despite their high fecundity and small size, these invertebrates are difficult to dose with small molecules (Table 1). *Drosophila* can be exposed by mixing compounds in with their food, however this approach requires relatively large amounts of stock compounds [18]. *C. elegans* can be grown in liquid culture,

Table 1. Comparison of different models for behavior-based chemical screening [21–23,38,39]



but have a tough cuticle that blocks the absorption of many compounds. By comparison, zebrafish are excellent models for behavior-based chemical screening (Table 1).

Zebrafish are relative newcomers to neuropharmacology. However, they are poised to generate an increasingly large share of the data in the field. The reason is that zebrafish are an excellent model for chemical biology [19,20]. At the embryonic and larval stages, they can be easily soaked in chemical solutions and are small enough for automated behavioral phenotyping in 96-well plate format. In a typical screen, animals are arrayed in multi-well plates along with chemicals from a chemical library. Automated systems systematically deliver stimuli and capture digital video that can be efficiently analyzed by machine vision algorithms. Such systems generate large amounts of raw image data and quantitative behavioral readouts [21–23].

Behavioral assays in zebrafish present opportunities for remarkable gains in throughput compared to other model organisms. For example, using a 96-well plate loaded with ten zebrafish per well, it should be possible to assay 960 animals and 96 chemical treatments in only a few minutes (Figure 2). By contrast, most behavioral assays in rodents

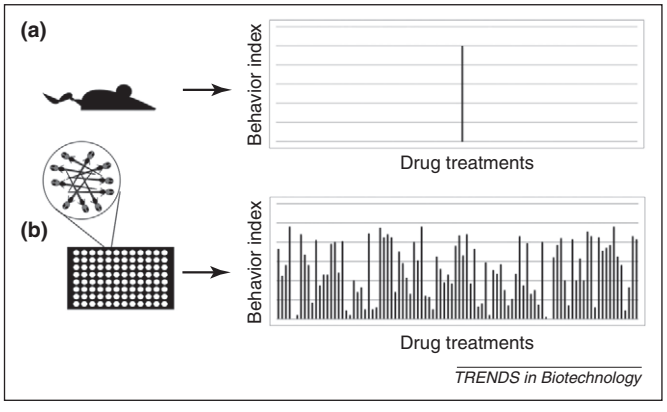


Figure 2. Zebrafish digital phenotyping and the behavioral data deluge. **(a)** Many behavioral models in neuropharmacology are based on relatively low-throughput assays in mice and other rodents leading to a relative scarcity of behavioral data. **(b)** High-throughput phenotyping technologies in the zebrafish are generating an unprecedented amount of behavioral data from large-scale chemical genetic screening.

are designed to score one animal at a time [24]. As a result, an unprecedented amount of phenotypic data can be expected from behavior-based chemical screens in the zebrafish.

Scalable zebrafish behavioral assays include the photomotor response, rest/wake behaviors, acoustic startle, and habituation [21,23,25–28]. Many of these assays have already been used to screen thousands of chemical treatments. As researchers work to understand how chemicals affect behaviors, it will be useful to combine the large-scale phenotyping efforts of different researchers and laboratories.

The behavioral barcoding analytical standard

Behavioral barcoding is a statistical profiling approach for analyzing large behavioral datasets [21]. The term 'behavioral barcoding' refers to a systematic analysis of behavioral features for classifying and organizing phenotypes. Barcodes are simple and familiar tools for tracking and organizing all kinds of complex information from consumer products to geographical locations. Like other kinds of multidimensional profiling, behavioral barcoding is a standardized and scalable approach for systematically classifying information into hierarchical classes.

Recently, the behavioral barcoding approach was used to systematically analyze the zebrafish photomotor response (PMR), a robust pattern of motor activity in response to bright light stimuli for >20 000 small molecule treatments [21]. To generate PMR barcodes for 20 000 chemical treatments, 4 terabytes (TB) of raw image data was processed to generate 7 gigabytes (GB) of raw quantitative data. This quantitative data was processed to generate 1 GB of graphical plots (one plot per response) and then further processed to generate 5 MB of behavioral barcodes. The first step in the barcoding analysis is to subdivide the response into a set of quantitative features that summarize salient features of the behavioral response. These features, or their combinations, represent the 'bars' in the barcode. For example, features may include different phases of the response such as background, latency, and excitation phases that occur in almost all behavioral phenotypes. When combined, these summary features together comprise the barcode, providing a concise quantitative summary of the complex behavior under study (Figure 3).

The behavioral barcoding approach can be generalized to almost any quantitative assay. Behavioral profiling of zebrafish rest/wake behavior has been used in a screen of 3968 compounds to link drugs to their biological targets and rest/wake regulation [23]. Automated phenotyping of non-associative learning behaviors has been used in a screen of 1760 compounds to identify chemicals with unknown roles in learning [22]. The approach could also be applied to many other behavioral assays. Because many phenotypes are screened against common chemical libraries, chemical treatments can be used to link behavioral barcodes from different assays and laboratories.

The purpose of creating behavioral barcodes and sorting them into particular orders is to understand the functional relationships between genes, chemical treatments, and their phenotypes. Different neuroactive chemicals affect behavioral barcodes in specific ways.

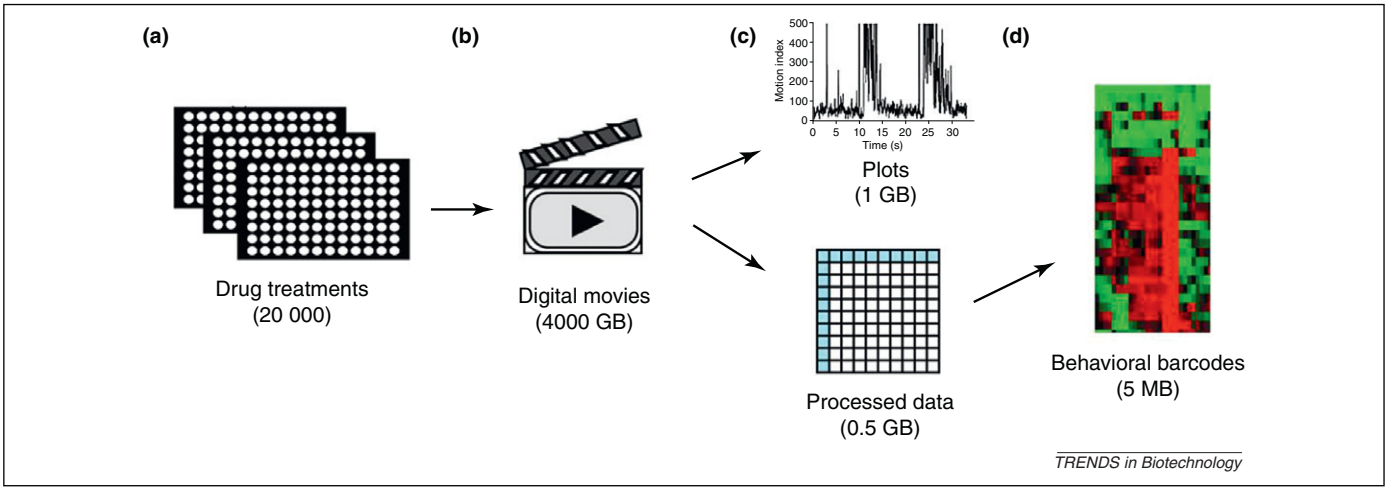


Figure 3. Behavioral barcoding is a systematic approach to analyzing large behavioral datasets. The schematic workflow depicts information processing steps in a large-scale behavior-based chemical screen, from many drug treatment groups **(a)**, to a large number of raw digital recordings **(b)**, to a processed dataset consisting of each drug's behavioral features **(c)**, and finally a concise collection of behavioral barcoding data **(d)**. Additional processing steps are necessary to link barcodes to the original data and outside information.

Hierarchical clustering of behavioral barcodes groups compounds into mechanistically related phenoclusters. Recent studies have visualized the clustering of behavioral barcodes as a 2D heat map [21,23]; for example, with compounds along the *y*-axis, behavioral features on the *x*-axis, and signal strength as colors on the heat map. These clusters can then be used to predict the molecular mechanisms of novel neuroactive drugs, identify novel neuroactive compounds, and describe new behavioral phenotypes. In theory, the predictive resolution of behavioral barcodes could be improved by adding additional features from different assays and genetic backgrounds. Reaching this goal will require the efforts of many laboratories and will be aided by computational tools for analyzing and sharing behavior barcodes. For example, cloud computing services could host large datasets and provide a pipeline of standardized analytical tools. Here, we use the term 'cloud computing' to refer to a wide range of opportunities facilitated by internet applications and data service centers.

The 'matrix tool' scenario

We envision developing an internet-based 'matrix tool' for mining behavioral barcodes and other phenotypic data from large-scale chemical screens (Figure 4). Like a standard database, the matrix tool could easily link each barcode to its respective digital movie file, graphical plot, and chemical structure information. However, unlike a standard database, it could also analyze all the phenotypic data in the matrix for comprehensive phenotypic correlations. Many labs have developed valuable behavioral models and would be willing to contribute their data to larger analyses. But the field is lacking standardized tools for helping researchers to get their data into the cloud. Fortunately, lessons from more established '-omic' sciences provide some rough ideas for how to produce a matrix tool for chemobehavioral phenomics.

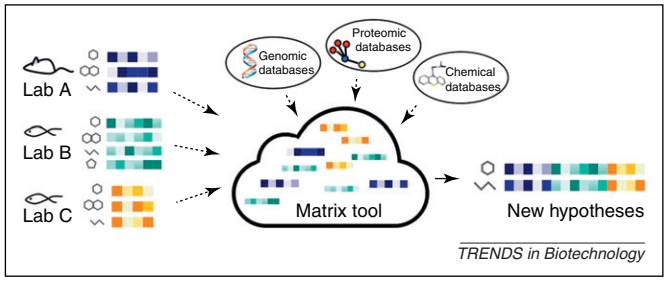


Figure 4. A cloud-based 'matrix tool' is envisioned for mining behavioral barcodes and other phenotypic data from large-scale chemical screens. The matrix tool is a proposed cloud-based toolbox to standardize, integrate, and share large-scale behavioral datasets. Laboratories can contribute behavioral barcodes and other matrix compliant datasets to the matrix tool, which will facilitate analysis across all stored datasets and crosslink information from additional bioinformatic tools to improve understanding, generate new hypotheses, and share data with the wider community.

One of the first steps is to develop guidelines for the minimum information necessary to describe a behavioral phenomics experiment. There are several precedents in the biomedical sciences for such guidelines including the 'minimum information about a…' microarray experiment (MIAME), proteomics experiment (MIAPE), and bioactive entity experiment (MAIBE) [29–31]. For behavioral phenomics experiments, the matrix tool could provide templates for organizing screening data in to standardized machine-searchable spreadsheet formats based on standardized chemical, genetic, and behavioral ontologies. Producing high quality matrix-compatible data is a win-win situation for individual laboratories and the neuropharmacology community. For participating laboratories, standardized workflows will facilitate data analysis and make large and complex behavioral phenomic datasets more manageable. For the community, standard guidelines will make behavioral data more accessible and useful.

The matrix tool would be a very small part of a large and rapidly expanding bioinformatic universe. Beyond cross-linking screening results, the matrix tool can integrate with the entire scientific knowledgebase to improve understanding and generate new hypotheses (Figure 4). Scientific search engines and aggregators such as sciencenet, and the 'Bioinformatic Harvester' could be used to cross-link additional databases and search tools like Pubchem, the similarity ensemble approach (SEA), Stitch, and others to predict chemical-target interactions [32–36]. Similarly, augmented-reality browsing tools can help researchers

understand complex information by automatically overlaying genes and chemical names with descriptive pop-up windows [37]. Integrating phenomic information with the larger spectrum of bioinformatic tools remains a key future goal.

'A precious transcendental state'

Behavioral phenomics is catching up to other '-omic' sciences with automated phenotyping, online repositories, behavioral barcoding, and other standardized workflows. Building a matrix tool will accelerate the pace of neuroscience research while transcending the boundaries of life science research laboratories. For example, the matrix would help to reach dedicated computer graphics and data mining labs (which often have no direct access to biologists) to work and collaborate with behavioral phenomics projects. It would also enable crowd-sourcing tools for citizen scientists and other interested parties to comment on and complement machine-generated information. Although no such matrix tool currently exists for zebrafish chemobehavioral data, the first versions cannot be too far away.

Acknowledgments

This work was supported by National Institutes of Health grants K01MH091449 (D.K.), T32HL07208 (A.J.R.), MH086867 and MH085205 (R.T.P.), the BOLD Marie Curie Initial Training Network grant 238821 (A.H.S.), the KIT Biointerfaces program (U.L.), and by the Charles and Ann MGH Research Scholars Award (R.T.P.).

References

1 Kokel, D. and Peterson, R.T. (2008) Chemobehavioural phenomics and behaviour-based psychiatric drug discovery in the zebrafish. *Brief. Funct. Genomics Proteomics* 7, 483
2 Schadt, E.E. *et al.* (2010) Computational solutions to large-scale data management and analysis. *Nat. Rev. Genet.* 11, 647–657
3 Shulgin, A. and Shulgin, A. (1991) *Pihkal: A Chemical Love Story*, Transform Press
4 Drysdale, R. (2008) FlyBase: a database for the *Drosophila* research community. *Methods Mol. Biol.* 420, 45–59
5 Maddatu, T.P. *et al.* (2011) Mouse Phenome Database (MPD). *Nucleic Acids Res.* 40, D887–D894
6 Peterson, R.T. (2008) Chemical biology and the limits of reductionism. *Nat. Chem. Biol.* 4, 635–638
7 Wang, Y. *et al.* (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* 37, W623–W633
8 Wang, Y. *et al.* (2011) PubChem's BioAssay Database. *Nucleic Acids Res.* 40, D400–D412
9 Dankert, H. *et al.* (2009) Automated monitoring and analysis of social behavior in *Drosophila*. *Nat. Methods* 6, 297–303
10 Hoyer, S.C. *et al.* (2008) Octopamine in male aggression of *Drosophila*. *Curr. Biol.* 18, 159–167
11 Steele, A.D. *et al.* (2007) The power of automated high-resolution behavior analysis revealed by its application to mouse models of Huntington's and prion diseases. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1983–1988
12 Hey, A. *et al.* (2009) *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research
13 Houle, D. *et al.* (2010) Phenomics: the next challenge. *Nat. Rev. Genet.* 11, 855–866
14 Fox, P. and Hendler, J. (2011) Changing the equation on scientific data visualization. *Science* 331, 705–708
15 Carneiro, H. (2009) Google Trends: a web-based tool for real-time surveillance of disease outbreaks. *Clin. Infect. Dis.* 49, 1557–1564
16 Mitchell, T.M. (2009) Computer science. Mining our reality. *Science* 326, 1644–1645
17 Bollen, J. *et al.* (2011) Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pp. 450–453
18 Chang, S. *et al.* (2008) Identification of small molecules rescuing fragile X syndrome phenotypes in *Drosophila*. *Nat. Chem. Biol.* 4, 256–263
19 MacRae, C.A. and Peterson, R.T. (2003) Zebrafish-based small molecule discovery. *Chem. Biol.* 10, 901–908
20 McKinley, E.T. *et al.* (2005) Neuroprotection of MPTP-induced toxicity in zebrafish dopaminergic neurons. *Brain Res. Mol. Brain Res.* 141, 128–137
21 Kokel, D. *et al.* (2010) Rapid behavior-based identification of neuroactive small molecules in the zebrafish. *Nat. Chem. Biol.* 6, 231–237
22 Wolman, M. *et al.* (2011) Chemical modulation of memory formation in larval zebrafish. *Proc. Natl. Acad. Sci. U.S.A.* 180, 15468–15473
23 Rihel, J. *et al.* (2010) Zebrafish behavioral profiling links drugs to biological targets and rest/wake regulation. *Science* 327, 348–351
24 Jhuang, H. *et al.* (2010) Automated home-cage behavioural phenotyping of mice. *Nat. Commun.* 1, 68
25 Baraban, S.C. *et al.* (2007) A large-scale mutagenesis screen to identify seizure-resistant zebrafish. *Epilepsia* 48, 1151–1157
26 Burgess, H.A. and Granato, M. (2007) Modulation of locomotor activity in larval zebrafish during light adaptation. *J. Exp. Biol.* 210, 2526–2539
27 Burgess, H.A. and Granato, M. (2007) Sensorimotor gating in larval zebrafish. *J. Neurosci.* 27, 4984–4994
28 Burgess, H.A. *et al.* (2010) Distinct retinal pathways drive spatial orientation behaviors in zebrafish navigation. *Curr. Biol.* 20, 381–386
29 Brazma, A. *et al.* (2001) Minimum information about a microarray experiment (MIAME) -toward standards for microarray data. *Nat. Genet.* 29, 365–371
30 Orchard, S. *et al.* (2011) Minimum information about a bioactive entity (MIABE). *Nat. Rev. Drug Discov.* 10, 661–669
31 Taylor, C.F. *et al.* (2007) The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* 25, 887–893
32 Hert, J. *et al.* (2008) Quantifying the relationships among drug classes. *J. Chem. Inform. Model.* 48, 755–765
33 Kuhn, M. *et al.* (2009) STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.* 38, D552–D556
34 Laggner, C. *et al.* (2011) Chemical informatics and target identification in a zebrafish phenotypic screen. *Nat. Chem. Biol.* 8, 144–146
35 Liebel, U. *et al.* (2004) Harvester: a fast meta search engine of human protein resources. *Bioinformatics* 20, 1962–1963
36 Lutjohann, D.S. *et al.* (2011) Sciencenet – towards a global search and share engine for all scientific knowledge. *Bioinformatics* 27, 1734–1735
37 Pafilis, E. *et al.* (2009) Reflect: augmented browsing for the life scientist. *Nat. Biotechnol.* 27, 508–510
38 Stilwell, G.E. *et al.* (2006) Development of a Drosophila seizure model for in vivo high-throughput drug screening. *Eur. J. Neurosci.* 24, 2211–2222
39 Roberds, S.L. *et al.* (2011) Rapid, computer vision-enabled murine screening system identifies neuropharmacological potential of two new mechanisms. *Front. Neurosci.* 5, 103