Implications of Big Data for cell biology

Kara Dolinski[a,*] and Olga G. Troyanskaya[a,b,c,*]
[a]Lewis-Sigler Institute for Integrative Genomics and [b]Department of Computer Science, Princeton University, Princeton, NJ 08540; [c]Simons Center for Data Analysis, Simons Foundation, New York, NY 10010

ABSTRACT **"Big Data" has surpassed "systems biology" and "omics" as the hottest buzzword in the biological sciences, but is there any substance behind the hype? Certainly, we have learned about various aspects of cell and molecular biology from the many individual high-throughput data sets that have been published in the past 15–20 years. These data, although useful as individual data sets, can provide much more knowledge when interrogated with Big Data approaches, such as applying integrative methods that leverage the heterogeneous data compendia in their entirety. Here we discuss the benefits and challenges of such Big Data approaches in biology and how cell and molecular biologists can best take advantage of them.**

Monitoring Editor
Keith G. Kozminski
University of Virginia

Received: Feb 26, 2015
Revised: May 15, 2015
Accepted: May 18, 2015

What is "Big Data," and what, if anything, can it do for cell biologists? The definition of Big Data is changing as rapidly as genomics data are being generated. All biologists are faced with data growing at a rate that could not have been imagined just 20 years ago, when most labs were still running polyacrylamide gels to sequence individual genes over the course of a couple of days. Soon after, results from a single microarray were intimidating enough to most biologists to be considered Big Data. Now, it is routine to analyze the entire compendia of expression and protein–protein interaction data. A similar "data avalanche" is happening in DNA sequencing, in which thousands of genomes are being analyzed in concert, and in imaging, in which cellular and organismal phenotypes can be systematically assessed in high-throughput format. Rather than setting a size threshold to define it (lest we fall into a "620K is enough memory for anyone" trap), Big Data is a moving bar that is set just beyond what we can, at a particular time, routinely annotate, analyze, and visualize—that is, Big Data is positioned where the challenges are in interpreting the wealth (and noisiness) of data now readily available. In other words, Big Data can be characterized by the three Vs: volume, variety, and velocity. The key question here is whether these virtual mountains of expression, sequence, proteomics, imaging, and other data can be transformed into biological knowledge in such way that it is both trusted and useful to cell biologists.

It is interesting that the two most-cited articles in this journal are among the very first Big Data papers (Spellman *et al.*, 1998; Gasch *et al.*, 2000). David Botstein showed, in a retrospective, that indeed it was the data that were important: roughly half of the citations for these articles came from computational biologists and statisticians (Botstein, 2010). Thus, these articles not only defined for the first time a genome-wide set of genes regulated by the cell cycle and stress response, respectively, but they also provided data for follow-up analyses, both experimental and computational, that enabled systems-level understanding of these processes and how they work in concert with other pathways. For example, a subsequent article used data from these two studies combined with growth rate under different limiting conditions to characterize the coordination of cell cycle, stress response, and growth rate in *Saccharomyces cerevisiae* (Brauer *et al.*, 2008). Since those articles, Big Data has grown, not simply in size—more than 1.3 million samples are now available from the Gene Expression Omnibus database alone—but also in diversity. Nowadays, no area of molecular biology or genetics is insulated from high-throughput data—whether it is exploring genomic diversity in the context of evolution or human disease, considering epigenetic changes in development, or understanding transcriptional regulation of genes or posttranslational protein modifications. These are produced not only by individual laboratories, but also by large consortia, such as the Encyclopedia of DNA Elements (ENCODE) project to identify all the functional elements in the human genome (www.encodeproject.org/), the GTeX project to generate expression data across different human tissues

DOI:10.1091/mbc.E13-12-0756

*The authors contributed equally to this work.

Address correspondence to: Kara Dolinski (dolinski@princeton.edu); Olga G. Troyanskaya (ogt@genomics.princeton.edu).

Abbreviations used: GTeX, Genotype-Tissue Expression; LINCS, Library of Integrated Network-based Cellular Signatures.

(www.gtexportal.org/), and the LINCS program to produce a large-scale set of cellular signatures in response to different molecular or genetic perturbations (www.lincsproject.org/).

These data sets can already be helpful to the everyday bench biologist, where one might find some new information about particular genes of interest that are discussed in the text or pop up in a reanalysis of an individual data set. However, these data, when analyzed in concert both within and across data types have the potential to provide substantial biological knowledge, generating hypotheses that cannot be readily extracted from either the literature or any individual, even genome-scale, data set. For example, combining information in the Gasch *et al.* (2000) yeast stress response data set and the Spellman *et al.* (1998) yeast cell cycle data set can enable us to identify cell cycle–regulated genes that respond to DNA damage. This can then be extended with additional data to include identification of regulatory networks (e.g., by including chromatin immunoprecipitation followed by high-throughput sequencing [ChIP-seq] and nucleosome occupancy data). One such method, BETA, integrates ChIP-seq data sets that provide information about transcription factor binding with gene expression data informative of differential expression to predict direct target genes of transcription factors (Wang *et al.*, 2013). Emerging Big Data analysis methods and prediction platforms can generate hypotheses about the function, interactions, and regulation of proteins, RNAs, and other biomolecules, their behavior in biological pathways, and relationships between various molecular entities and phenotypes. For example, Inferelator 2.0 uses a combination of Markov chain Monte Carlo and ordinary differential equations to learn both the topology and dynamics of regulatory networks in Halobacterium (Madar *et al.*, 2009). Furthermore, many of these approaches can, through integrative analysis of Big Data collections, provide insight into biological contexts (such as specific tissues, developmental stages, and perturbations) that are challenging to directly assay experimentally. For example, in humans, the Genome-scale Integrated Analysis of gene Networks in Tissues (GIANT) webserver predicts functional maps of protein–protein relationships for 144 tissues through Bayesian integration of thousands of expression, sequence, and protein–protein interaction experiments (Greene *et al.*, 2015); experimental biologists can explore these functional maps to better understand cell-specific processes in human disease.

In addition, Big Data, while naturally subject to signal-to-noise challenges, can compensate for the noisiness of each individual data set precisely because of its scale. Intuitively, signals that occur independently in multiple data sets are more likely to be "real"; for example, genes identified as cell-cycle regulated in multiple genome-scale studies are more likely to be truly cell-cycle regulated. Of course, simply identifying repeating signals can also zero in on common technical and biological artifacts or very broad (and thus often less interesting) biological signals, such as the general stress response that *S. cerevisiae* exhibit across essentially all treatments or broad growth regulators in human cell culture data. Sophisticated computational approaches based on Big Data collections, such as those described earlier, can specifically focus on the biologically informative signals relevant to a specific biological question, including those hard or impossible to detect by simple analyses.

Furthermore, as each individual experiment is inherently assaying only specific aspects of cellular complexity, combining data sets from different experimental conditions, platforms, and experimental approaches provides insights into molecular pathways that cannot be realized from individual studies done in isolation. Thus, analyses based on large collections of data hold the promise of systematic insights into how pathways function across mechanistic interaction and regulation modes, spanning, for example, transcriptional regulation by histone proteins and transcription factors (from ENCODE data), RNA stability effects, and protein transport, interactions, and posttranscriptional modifications from imaging and proteomics studies.

In the past decade, many systematic approaches have been developed for integrating across genomic Big Data collections to provide novel biological insights—for example, to identify biological processes in which genes with unknown function participate (Pena-Castillo *et al.*, 2008) and to predict physical and regulatory protein interactions and posttranslational modification in a large-scale, automated way (Vaske *et al.*, 2010; Zhong *et al.*, 2014; Park *et al.*, 2015). These approaches use statistical techniques that aim to isolate signal from noise in these diverse and heterogeneous data collections, often relying on examples of known biological associations (e.g., genes with previously discovered biological function) to identify informative signals and make new discoveries. For example, Integrative Multi-species Prediction (IMP; imp.princeton.edu) probabilistically combines a large collection of expression, sequence, and protein interaction data to provide functional network and function predictions for any protein in the human and major model organism genome or proteome. Although most of these methods analyze data in preset ways, a recent trend includes development of approaches that enable the user to focus analysis on a specific biological area or question, essentially directing the analysis of Big Data without having to do any programming. Many Big Data integrative methods now provide highly targeted analysis, such as in the tissue-specific functional networks provided by GIANT (giant.princeton.edu) or Th17-focused prediction of TH17 regulatory networks (Ciofani *et al.*, 2012).

In fact, integrative analysis of functional genomics data coupled with computational modeling has effectively directed laboratory experiments and given rise to novel experimental discoveries in multiple model organisms and humans (Hess *et al.*, 2009; Yan *et al.*, 2010; Guan *et al.*, 2012; Wong *et al.*, 2012). For example, Doherty *et al.* (2012) showed that the BLM10-20S proteasome activator mediates DNA damage and other cellular stresses, in part by examining the predicted functional networks of the genes that were induced in blm10 mutants using the BioPIXIE tool (Myers *et al.*, 2005). Similarly, Sanchez-Garcia *et al.* (2014) integrated expression data from primary breast tumors with data from RNA interference screens using their Helios tool to identify candidate cancer-driver genes. They went on to experimentally characterize one of their novel predictions, RSF-1, showing that when amplified, RSF-1 increased both tumorigenesis and metastasis in mouse models of breast cancer.

Such *data-driven* approaches provide an important complement to the highly curated, aggregate databases that provide access to valuable information such as comprehensively curated physical and genetic interaction data (e.g., BioGRID; Chatr-Aryamontri *et al.*, 2015) or phenotype information for model organisms through the model organism databases (e.g., Engel *et al.*, 2010; Bult *et al.*, 2013; Deans *et al.*, 2015). Because they are based on high-throughput data, not literature-based curation or collections of specific experiments, the Big Data–based resources tend to be less biased toward prior knowledge and are able to make predictions even in areas in which prior knowledge may be very sparse or nonexistent. The price for this is of course the higher potential for errors due to noise levels in the data, although these can be mitigated by careful analysis, making genomic data collection a great source of hypotheses that can drive traditional experiments. Together the Big Data–driven methods and the curated databases are powerful tools for the cell biologist. The curated data within the databases can serve as

important gold standards for evaluating computational predictions, and the predictions can be used to guide and refine the annotation provided by the curated databases.

Big Data also has the potential of revolutionizing our use of model organisms, enabling accurate, less-biased, molecular-level identification of the most informative model for genes and diseases in the least expensive and most tractable experimental system. The key advantage is the ability to go beyond sequence-based orthology to systematically assess functional conservation, promising a functional mapping of proteins, pathways, and phenotypes across organisms. For example, biologists can use a method based on probabilistically mapping protein networks from a large compendium of high-throughput expression data across organisms to systematically predict which genes are most likely to participate in the same biological process and thus have analogous function in different organisms (Singh *et al.*, 2008; Chikina and Troyanskaya, 2011; Park *et al.*, 2013). Such approaches can succeed where sequence-based methods often fail, such as resolving paralogues based on tissue expression and correctly identifying functional divergence when orthology is predicted based on sequence and evolutionary relationships. The growing Big Data compendia in model organisms and humans, combined with sophisticated computational approaches, are bringing in an era in which we will be able to quantitatively and systematically identify the best experimental model (or models) for a given disease or process, pinpoint specific aspects of relevant biology that are or are not conserved across organisms, and generally be able to effectively and accurately integrate our knowledge across organisms.

What does this mean for a cell biologist? This means that all biologists contemplating their next study should consider using Big Data—based tools to inform their hypotheses, whether to identify additional proteins that may be relevant to the process they are studying, examine predicted molecular functions of a protein of interest, or consider pathways that may be relevant to the experimental treatment or genetic modification they are considering. If they are interested in a specific cell type or developmental stage, they can use Big Data–based resources to identify proteins expressed in this cell type, tissue-specific interactions and functions, and perhaps even cell-type specific predictions of perturbation effects on phenotypes. Many of the prediction systems and algorithms necessary for these analyses are available publicly, often in a user-friendly form aimed at biomedical researchers with no or limited computational training (Table 1). A more in-depth analysis, especially involving unpublished data from a cell biologist's laboratory, can be undertaken either in their own lab (if they have the requisite computational skills or through the newly emergent systems that enable sophisticated computational analysis by nonspecialists; e.g., Greene and Troyanskaya, 2011) or with a computational collaborator. Most institutions now include "card-carrying" computational biologists or bioinformaticians, as well as many experimentalists with substantial computational expertise, although looking for the best collaborative fit may require crossing departmental boundaries into computational biology, computer science, or similarly focused departments.

Does this mean that experimental cell biologists should look for alternative careers? Absolutely not! Computational approaches based on Big Data generate *hypotheses*, not experimentally verified *biological knowledge*. In addition, the broader, less-biased, Big Data–driven information can be a powerful guide for cell biology studies. In the past, cell biologists would read articles and look at their last gel to inform the next set of experiments. As we continue to make progress in harnessing Big Data, cell biologists can obtain a new, valuable tool from its the broader, less-biased information. Cell biologists who do not use Big Data to inform their experiments are squandering a valuable resource. It is analogous to a biologist doing DNA amplification manually in water baths when PCR machines are available. The wealth that Big Data brings will enable cell biologists to better design and focus their experimental programs with the expectation that biological insights will come faster and more efficiently. We are not even close to replacing individual experiments (and the cell biologists who do them!) with computers, but instead are in the midst of an exciting time when we are just beginning to tap the major effect of Big Data on the world of cell biology.

galaxyproject.org	Platform for genome-scale biomedical research
imp.princeton.edu	Functional networks in model organisms and humans
giant.princeton.edu	Tissue-specific networks and genome-wide association studies in humans
thebiogrid.org	Database of protein and genetic interactions
seek.princeton.edu	Cross-platform search engine for expression data
genomespace.org	Framework for integrative genomics analysis
cbioportal.org	Visualization and analysis of cancer genomic data

TABLE 1: Examples of user-friendly systems for Big Data analysis in biology.

REFERENCES

Botstein D (2010). It's the data! Mol Biol Cell 21, 4–6.

Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D (2008). Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. Mol Biol Cell 19, 352–367.

Bult CJ, Eppig JT, Blake JA, Kadin JA, Richardson JE, Mouse Genome Database Group (2013). The mouse genome database: genotypes, phenotypes, and models of human disease. Nucleic Acids Res 41, D885–D891.

Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, Stark C, Breitkreutz A, Kolas N, O'Donnell L, *et al.* (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res 43, D470–D478.

Chikina MD, Troyanskaya OG (2011). Accurate quantification of functional analogy among close homologs. PLoS Comput Biol 7, e1001074.

Ciofani M, Madar A, Galan C, Sellars M, Mace K, Pauli F, Agarwal A, Huang W, Parkurst CN, Muratet M, *et al.* (2012). A validated regulatory network for Th17 cell specification. Cell 151, 289–303.

Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, *et al.* (2015). Finding our way through phenotypes. PLoS Biol 13, e1002033.

Doherty KM, Pride LD, Lukose J, Snydsman BE, Charles R, Pramanik A, Muller EG, Botstein D, Moore CW (2012). Loss of a 20S proteasome activator in Saccharomyces cerevisiae downregulates genes important for genomic integrity, increases DNA damage, and selectively sensitizes cells to agents with diverse mechanisms of action. G3 (Bethesda) 2, 943–959.

Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL, *et al.* (2010). Saccharomyces Genome Database provides mutant phenotype data. Nucleic Acids Res 38, D433–D436.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241–4257.

Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, *et al*. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat Genet, doi: 10.1038/ng.3259.

Greene CS, Troyanskaya OG (2011). PILGRM: an interactive data-driven discovery platform for expert biologists. Nucleic Acids Res 39, W368–W374.

Guan Y, Gorenshteyn D, Burmeister M, Wong AK, Schimenti JC, Handel MA, Bult CJ, Hibbs MA, Troyanskaya OG (2012). Tissue-specific functional networks for prioritizing phenotype and disease genes. PLoS Comput Biol 8, e1002694.

Hess DC, Myers CL, Huttenhower C, Hibbs MA, Hayes AP, Paw J, Clore JJ, Mendoza RM, Luis BS, Nislow C, *et al*. (2009). Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLoS Genet 5, e1000407.

Madar A, Greenfield A, Ostrer H, Vanden-Eijnden E, Bonneau R (2009). The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. Conf Proc IEEE Eng Med Biol Soc 2009, 5448–5451.

Myers CL, Robson D, Wible A, Hibbs MA, Chiriac C, Theesfeld CL, Dolinski K, Troyanskaya OG (2005). Discovery of biological networks from diverse functional genomic data. Genome Biol 6, R114.

Park CY, Krishnan A, Zhu Q, Wong AK, Lee YS, Troyanskaya OG (2015). Tissue-aware data integration approach for the inference of pathway interactions in metazoan organisms. Bioinformatics 31, 1093–1101.

Park CY, Wong AK, Greene CS, Rowland J, Guan Y, Bongo LA, Burdine RD, Troyanskaya OG (2013). Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. PLoS Comput Biol 9, e1002957.

Pena-Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, Guan Y, Leone M, Pagnani A, Kim WK, *et al*. (2008). A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome Biol 9(Suppl 1), S2.

Sanchez-Garcia F, Villagrasa P, Matsui J, Kotliar D, Castro V, Akavia UD, Chen BJ, Saucedo-Cuevas L, Rodriguez Barrueco R, Llobet-Navas D, *et al*. (2014). Integration of genomic data enables selective discovery of breast cancer drivers. Cell 159, 1461–1475.

Singh R, Xu J, Berger B (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc Natl Acad Sci USA 105, 12763–12768.

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9, 3273–3297.

Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237–i245.

Wang S, Sun H, Ma J, Zang C, Wang C, Wang J, Tang Q, Meyer CA, Zhang Y, Liu XS (2013). Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protocols 8, 2502–2515.

Wong AK, Park CY, Greene CS, Bongo LA, Guan Y, Troyanskaya OG (2012). IMP: a multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. Nucleic Acids Res 40, W484–W490.

Yan H, Venkatesan K, Beaver JE, Klitgord N, Yildirim MA, Hao T, Hill DE, Cusick ME, Perrimon N, Roth FP, Vidal M (2010). A genome-wide gene function prediction resource for Drosophila melanogaster. PLoS One 5, e12139.

Zhong J, Wasson T, Hartemink AJ (2014). Learning protein-DNA interaction landscapes by integrating experimental data through computational models. Bioinformatics 30, 2868–2874.