# Big Data in *Caenorhabditis elegans*: *quo vadis?*

Harald Hutter[a] and Donald Moerman[b]

[a]Department of Biological Sciences, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; [b]Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

**ABSTRACT A clear definition of what constitutes "Big Data" is difficult to identify, but we find it most useful to define Big Data as a data collection that is complete. By this criterion, researchers on *Caenorhabditis elegans* have a long history of collecting Big Data, since the organism was selected with the idea of obtaining a complete biological description and understanding of development. The complete wiring diagram of the nervous system, the complete cell lineage, and the complete genome sequence provide a framework to phrase and test hypotheses. Given this history, it might be surprising that the number of "complete" data sets for this organism is actually rather small—not because of lack of effort, but because most types of biological experiments are not currently amenable to complete large-scale data collection. Many are also not inherently limited, so that it becomes difficult to even define completeness. At present, we only have partial data on mutated genes and their phenotypes, gene expression, and protein–protein interaction—important data for many biological questions. Big Data can point toward unexpected correlations, and these unexpected correlations can lead to novel investigations; however, Big Data cannot establish causation. As a result, there is much excitement about Big Data, but there is also a discussion on just what Big Data contributes to solving a biological problem. Because of its relative simplicity, *C. elegans* is an ideal test bed to explore this issue and at the same time determine what is necessary to build a multicellular organism from a single cell.**

**Monitoring Editor**
Jeffrey D. Hardin
University of Wisconsin

Received: Jul 13, 2015
Revised: Aug 20, 2015
Accepted: Aug 26, 2015

> As was predicted at the beginning of the Human Genome Project, getting the sequence will be the easy part as only technical issues are involved. The hard part will be finding out what it means, because this poses intellectual problems of how to understand the participation of the genes in the functions of living cells.
>
> *Sydney Brenner (1995)*

## INTRODUCTION: WHAT IS "BIG DATA"?

Collecting "Big Data" is currently seen as the panacea for many research problems in contemporary science, as well as for diverse fields outside of science. What is most curious about this state of affairs is that hardly anybody defines it, and when you start to dig, you realize that the term "Big Data" is used in many different ways. Compare, for example, the use of this term in biology, astrophysics, and finance and by the U.S. National Security Agency or in social media analysis. The size scales are very different in each of these examples, often by orders of magnitude. Broadly speaking, Big Data are large amounts of data, possibly heterogeneous in nature, that are difficult to analyze by current data-processing tools. As data-processing methods change over time, mainly due to increasing computing power, the definition of Big Data also changes. Another way to look at Big Data is in the context of the overall information available. With this in mind, Big Data can be usefully defined as, not just the collection of lots of samples, but of "all" data, which has been described simply as $n = \text{all}$ (Mayer-Schönberger and Cukier, 2013). A key feature of Big Data in this sense is the option to reuse data for a different, even completely unanticipated purpose without the need to collect data again. For example, whole-genome sequence data originally collected to identify mutations in protein-coding genes could be used to identify changes in other parts of the genome or even to identify commensal organisms or parasites of the host whose genome might have been sequenced together with the host genome. Because of this added utility, we would argue that completeness should be an important aspect of large-scale studies in biology.

DOI:10.1091/mbc.E15-05-0312

Address correspondence to: Donald Moerman (moerman@zoology.ub.ca).

Abbreviations used: CRISPR, clustered regularly interspaced short palindromic repeats; MS, mass spectrometry; RNAi, RNA interference; RNA-seq, RNA sequencing; SAGE, Serial Analysis of Gene Expression; smFISH, single-molecule fluorescent in situ hybridization; Y2H, yeast two-hybrid.

## BIG DATA IN *CAENORHABDITIS ELEGANS*

Along with the question of what Big Data are, we need to ask what Big Data are for. In biology today, collections of large data sets often ultimately relate to human health, but Big Data can and are being used to address all kinds of questions. What if the object is to understand life—for example, how to build an organism. Sydney Brenner once said he wanted to build a Gryphon, a mythological creature with the head, wings, and talons of an eagle and the body of a lion, as only by building such an organism could he confirm that he understood how development works down to the smallest detail. When he was looking for a suitable organism to study the development of the nervous system in the late 1960s, "Big Data" in the sense of $n$ = all was already in his mind: "It [the organism] should have relatively few cells, so that exhaustive studies of lineage and patterns can be made" (quoted in Wood, 1988). His search led him to the small roundworm (nematode) *Caenorhabditis elegans*. Indeed, Big Data were collected in *C. elegans* almost from the beginning. As a harbinger of things to come, John Sulston measured the total genomic DNA content (Sulston and Brenner, 1974). John White and his colleagues traced all of the neuronal connections from electron microscopic images of serially sectioned animals (White *et al.*, 1986). The reconstruction of the synaptic connections, now dubbed the "connectome," was done manually, as attempts to automate the task by digitizing the images failed due to a lack of computing power at the time, so this was truly "Big Data" in many ways. Additional "Big Data" sets in *C. elegans* now include the complete cell lineage (Sulston and Horvitz, 1977; Sulston *et al.*, 1983) and the sequence of the genome, which indeed is complete, with no gaps in any of the six chromosomes and totaling 100,291,840 base pairs (Hillier *et al.*, 2005). These data sets are important as a basis for further experiments and formulating hypotheses, but they do not solve a scientific question in and of themselves, that is, the lineage does not "explain" how development works, and the connectome does not "explain" how the brain works. As these illustrate, Big Data alone do not provide answers to many critical questions in biology, leading detractors of Big Data collections often to refer to them as "fishing expeditions" or use the pejorative "hypothesis-free" (supporters of this approach prefer the term "hypothesis generating"). Nonetheless, it would be a mistake to dismiss these collections as unworthy projects, as is amply demonstrated by the value of the connectome, the lineage, and the complete genome sequence to subsequent studies of the worm. In addition, Big Data projects like these make two underappreciated contributions: they provide a driving force to innovate the technology required for more efficient data acquisition and, when $n$ = all, they help to define the limits of the problem and provide a framework for developing future experimental ideas.

Somewhat surprisingly, there are no other complete data sets for *C. elegans*. This absence has not been for lack of investment of both money and labor but is mainly due to the fact that, apart from the examples just given and possibly RNA sequences (see later discussion), the desirable data sets today either are simply too large to be readily obtained with current methods or inherently open ended. At the DNA level, a reasonable goal would be to create deletion or other loss-of-function alleles for every gene. The *C. elegans* knockout project, after 15 years of systematically generating mutations, has generated bona fide loss-of-function mutations in >14,000 of 20,000 protein-coding genes, but it will likely take at least another 5 years to complete the collection (*C. elegans* Deletion Mutant Consortium, 2012; Thompson *et al.*, 2013). This target does not even include microRNAs or other RNAs modulating cellular processes, and we dare not even mention control regions. Even with new technological developments such as clustered regularly interspaced short palindromic repeats (CRISPR; Jinek *et al.*, 2012), finding loss-of-function mutations for all functional genetic elements remains a monumental undertaking. Reducing gene activity by RNA interference (RNAi) offers a way to scale things up, that is, to do genome-wide screens (Fraser *et al.*, 2000; Gönczy *et al.*, 2000; Kamath *et al.*, 2003). However, RNAi does not work for every gene (or every phenotype), the collection of RNAi clones does not cover the entire genome, and the actual screening effort for many phenotypes is prohibitive.

At the RNA level, determining gene expression, although posing some problems in collection and analysis, might be close to achieving $n$ = all, at least at the whole-animal level (but see later discussion for limitations). Over the past two decades, there have been several large-scale projects to monitor mRNA levels in wild-type and mutant animals, various developmental stages, and specific tissues and cells. These studies have used the technology of the period, whether it be in situ hybridization, microarrays, Serial Analysis of Gene Expression (SAGE), or, more recently, RNA sequencing (RNA-seq). There is a large body of data here to mine (see, e.g., modSEEK; Zhu *et al.*, 2015; seek.princeton.edu/modSeek/worm/index.jsp), but the different techniques used pose a serious challenge for the integration of data. For example, how does a twofold microarray signal compare to five SAGE tags or 200 RNA-seq reads for the same transcript? A single gold standard set of expression data would in principle eliminate this problem. A subset of RNA-seq identified transcripts validated using single-molecule fluorescent in situ hybridization (smFISH) is one possible approach to establishing a gold standard set of expression data (Raj *et al.*, 2008). Whereas high-throughput approaches give us a comparative look at multiple transcripts simultaneously, and, in addition, with RNA-seq detail the level of alternative splicing, we still miss much of the nuance of gene regulation and expression. This is because most large-scale expression studies measure transcript levels of the entire organism or at the level of a single tissue, and none has yet given us a glimpse at the breadth of transcripts of a single cell in *C. elegans*.

Although there is now a burgeoning field examining RNA regulation, most large-scale RNA-seq studies use mRNA as a proxy for what is occurring at the protein level. There are other possible approaches to perform this type of study. For example, a more informative approach, albeit with much lower throughput, is to tag newly synthesized proteins and follow their expression through development. Following this theme, the Waterston group at the University of Washington (Seattle, WA) is attempting to map the expression pattern of the >900 transcription factors (TFs; Reece-Hoyes *et al.*, 2005; Weirauch and Hughes, 2011) in the nematode. Using four-dimensional (4D) microscopy and green fluorescent protein–tagged genes, they can monitor changes of protein levels over time with single-cell resolution during embryonic development (Murray *et al.*, 2008). So far, they have done ~10% of all TFs (Robert Waterston, personal communication), enough to demonstrate that this is a feasible approach. Multiple laboratories working together could give us an expression map for all TFs at the cellular and temporal detail required for a detailed understanding of cell specification. Although several labs across the world have the capacity to collect 4D expression data in the embryo, to our knowledge, a coordinated approach across multiple laboratories is not being pursued, but we see no technological barriers to this happening. An additional contribution of this approach is that tagged proteins allow one to determine subcellular localization. Use of tagged proteins for this purpose is already well established in yeast and has the added benefit that one can follow protein localization under different conditions (Tkach *et al.*, 2012; Chong *et al.*, 2015).

Although there have been several informative small-scale studies directed at specific protein complexes, in general, protein data lag far behind DNA or RNA data. To date, there are only a few published large-scale proteomic studies in *C. elegans*, and these have identified proteins for ~50% of the genes in this organism (Merrihew *et al.*, 2008; Schrimpf *et al.*, 2009; Walther *et al.*, 2015). By comparison, proteins have been associated with ~84% of genes in humans (Kim *et al.*, 2014; Wilhelm *et al.*, 2014). The lower percentage of coverage for the worm reflects that these studies were done on whole animals, whereas the human studies were done on a wide variety of different tissues. For neither organism are we close to identifying peptides for all exons and splice variants. It is now possible to do cell- and tissue-specific proteomic analysis in the worm, which should allow the generation of a more detailed proteome (Yuet *et al.*, 2015). The largest compendia of protein interaction data for *C. elegans* are yeast two-hybrid (Y2H) data sets. These data sets are no doubt useful as a guide, but they suffer from lack of context, validation, and completeness. One of the leading laboratories conducting high- throughput Y2H screening estimates that 96% of the protein–protein interactions in *C. elegans* have yet to be documented and that their high-throughput system will ultimately be able to detect ~16% of these interactions (Simonis *et al.*, 2009). Lest this be viewed as an anomaly specific to the worm, a recent human interactome–mapping project identified 14,000 Y2H interactions for 13,000 genes; this is estimated to be between 5 and 10% of all protein interactions for humans (Rolland *et al.*, 2014). An additional problem for gene or protein interaction data sets is that the number of possible interactions grows nonlinearly with the number of proteins or genes, so that *n* = all might become substantially more difficult to achieve. Mass spectrometry–based (MS) shotgun proteomics coupled with immunoprecipitations (reviewed in Fonslow *et al.*, 2014) might be a way to identify biologically meaningful interactions more directly and accelerate the identification of such interactions. There are clearly limitations to this approach, as it requires good antibodies and stable complexes of interacting proteins. As with expression profiling, mapping proteins and protein complexes in the worm could benefit from an organized, sustained, distributed, and cooperative endeavor. Current MS technology applied as a cooperative effort could approach *n* = all. Creating a map of interactions that actually occur would reduce the search space, which is an important starting point. However, to understand how a cell, tissue, or organism functions, a dynamic view of protein interactions is ultimately needed.

This is where we reach an impasse. At this juncture in biology, there is no easy way to collect *n* = all for this type of protein data, nor is it possible to identify all possible mutations or all variant transcripts and their cell and temporal expression in a genome of this size. It is not just the problem that *n* is too large, but also the problem of defining what *n* should include in each of these instances. A more feasible approach may be to identify a specific scientific question, such as control of gene expression, so that only a small subset of all the genes need be investigated and *n* = all can amount to a small and manageable number—for example, examining the expression pattern of all transcription factors as described earlier. A divide-and-conquer strategy could also be applied to the issue of proteomics, with a division of labor among cooperating laboratories along the lines of protein families or by tissue. That scale is everything in trying to attain completeness of this sort of data has been amply demonstrated in prokaryotes, where researchers using the bacteria *Bartonella henselae* were able to identify the complete expressed proteome of 1250 proteins using shotgun proteomics and saturating RNA-seq (Omasits *et al.*, 2013). The alternative to defining smaller problems is to develop technologies that allow us to increase the scale at which we can collect these types of data.

## THE HORROR! THE HORROR! WORKING WITH BIG DATA

WormBase is the main repository for all data pertaining to *C. elegans* and many of its relatives (Harris *et al.*, 2014; www.wormbase.org/). Although there is a great deal of gene annotation for some genes, many genes still have only sparse annotation or none at all. This is sobering, considering this nematode is probably among the most thoroughly studied of metazoans. Regarding gene function, data on gene expression and phenotype are most relevant. Phenotypic descriptions are difficult to analyze on a large scale when written in plain language. This has led to the proposal of the adoption of a Worm Phenotype Ontology (WPO; Schindelman *et al.*, 2011), and phenotypic data in this standardized vocabulary can now be extracted from WormBase, which greatly simplifies large-scale phenotypic comparisons. Some aspects of phenotype are also captured in Gene Ontology (GO) terms associated with the gene (Ashburner *et al.*, 2000). In addition, GO terms include information about putative biochemical function due to sequence similarity. However, 23% of the protein-coding genes in WormBase release WS238 have neither phenotypic nor cell-specific expression data and no GO annotation, that is, these are completely uncharacterized genes for which we cannot even predict a putative function based on sequence similarity. Eighteen percent of the genes have cellular expression data, mainly based on reporter gene expression, and 35% have either RNAi or mutant phenotypes (32% have RNAi phenotypes, 11% have mutant phenotypes, and some genes have both). Curiously, we are in a situation in which 70% of genes have mutations, but only 11% have any phenotypic description based on mutations. Obviously, the bottleneck is not the generation of mutants, but the characterization of mutants, which seems a peculiar state of affairs for such a pivotal genetic model organism. Unlike yeast, for which high-throughput phenotypic screening is prominent, high-throughput phenotypic screening has yet to take hold in the worm community even with the availability of high-throughput technology in the form of the Union Biometrica COPAS BioSorter, worm trackers (Husson *et al.*, 2012), and microfluidic devices.

The lack of completeness or paucity of Big Data sets (in the sense of *n* = all) obviously leads to major challenges for data integration. The dearth of phenotypic data exacerbates this problem because phenotype can integrate many seemingly unrelated data sets. Big Data sets tend to be "noisy" and inevitably contain imperfect or false data. Not every Y2H interaction represents a real interaction (let alone a biologically meaningful one). Combining Big Data sets is one way to improve the quality of the observations. For a protein interaction to be "real," the two proteins have to be expressed in the same cell at the same time, so that expression data combined with Y2H data can improve data quality. Obviously this approach is more effective when all relevant data sets are truly "big" in the sense of *n* = all. Because most current data sets are "small," that is, effectively just samples of the "expressome," "phenome," or "interactome," our ability to use data integration to improve quality is severely compromised. Consequently, we can use these integrated data sets mainly to add a few additional uncharacterized genes to existing networks, but we are not yet at the point where novel insights at the "systems" level might emerge. High-throughput sequencing is by far the main avenue to generate Big Data, although various time-lapse imaging strategies and electron microscopic reconstructions are beginning to generate substantial amounts of image data, which provide challenges for data analysis and presentation (e.g., Jarrell *et al.*, 2012; also see WormAtlas, www.wormatlas.org; Hall and Altun, 2008). Although sequencing has opened up a substantial number of research areas, many scientific questions cannot be addressed with the limited set of Big Data currently available. The value of the genome sequence will increase significantly when we can effectively integrate it with expression phenotype and protein interaction data. For this reason, we think that the near future for Big Data collection in *C. elegans* should include single-cell expression studies, more emphasis on large-scale protein studies, and certainly more emphasis on phenotypic analysis at both the whole-organism and the tissue level.

As mentioned earlier, Big Data sets inevitably contain errors due to experimental constraints and the large scale of data collection. Error rates of individual large-scale studies can vary substantially and are often unknown. Although validation will always be a problem when dealing with Big Data, since traditional types of validation experiments cannot be scaled up easily, and even performing replicate experiments might not be possible due to costs, we need to define the "known unknowns" whenever possible. Data annotation, integration, and conversion can further compromise data quality. Particularly critical here is GO annotation, as it is widely used across many organisms (see, e.g., geneontology.org/page/download-annotations; see Rhee *et al.* (2008) for a review of use and misuse of GO). GO annotations integrate a variety of different data. They are associated with an evidence code covering different types (and quality) of evidence, ranging from Inferred from Electronic Annotation (IEA) to Inferred by Curator. By far the largest number of annotations comes from IEA, not necessarily the most reliable source of annotation. Frequently, the quality of the underlying data used for the annotation is unclear. GO annotations are an integral part of many bioinformatics studies, so that lack of a sense of quality of the underlying data is a bit disturbing. We might be at risk of "garbage in, garbage out" or, even worse, "garbage in, problem solved" after drawing conclusions from large integrated data sets.

Data visualization tools and user-friendly interfaces for queries are essential for handling Big Data, and a major emphasis now is on the development and improvement of such tools. See, for example, Cytoscape (Shannon *et al.*, 2003; www.cytoscape.org). Gene and protein interactions are typically shown as networks with genes/proteins as nodes and interactions as lines connecting nodes. GeneMANIA (Montojo *et al.*, 2014; www.genemania.org/) is a program that provides a flexible, interactive user interface with access to information about which data sets are used to infer functional connections. We think this is one of its most powerful features. A related strength of this interface is that it allows the user to exclude certain data sets as a way to identify where the evidence for the interactions comes from and to estimate robustness of the network. Often it reveals a surprisingly small number of data sets contributing to a particular interaction. The program mentioned earlier, modSEEK, which allows comparisons of RNA profiles, also has these features. The evidence for interactions will become much stronger once we have more data sets and independent types of data, but in many instances, these validating data are still to come.

## MORE HORROR! INTERPRETING BIG DATA

For DNA sequences, we can obtain raw data with high accuracy, limited only by instrument quality and the resources to generate them. Interpreting the primary consequences of a detected sequence change is straightforward if it involves the amino acid encoding sequence of a protein but less straightforward when intergenic regulatory elements are affected. Secondary consequences—for example, consequences of an amino acid change for protein function—are frequently impossible to predict in the absence of further information on the protein in question. This challenge is exemplified by the results of the Million Mutation Project in *C. elegans*, which has identified >180,000 missense mutations with largely unknown consequences regarding protein function (Thompson *et al.*, 2013; genome.sfu.ca/mmp/about.html). Whereas the resource does provide an allelic series for most genes, it will be up to the individual investigator to tease apart which protein changes have a phenotypic effect. For any other kind of molecular data, even interpreting the primary data is a challenge. What does it mean when expression of gene X is three times higher than the expression of gene Y? In comparative expression studies, it is impossible to define a threshold for changes in gene expression that is significant in a biological sense. In fact, it is likely that the threshold will be different for different genes. Thresholds for "significant" change in expression levels in comparative expression studies are arbitrary from a biological perspective (thresholds, of course, can be and are specified on purely statistical grounds). Because posttranscriptional regulation of mRNA levels does not necessarily correlate with protein levels, changes in protein concentration cannot be predicted unambiguously from changes in mRNA levels. Limitations in data interpretation pose substantial challenges for a meaningful use of Big Data in biological sciences.

## SUMMARY: *QUO VADIS*?

> The future is already here—it's just not very evenly distributed.
>
> *William Gibson (1999)*

Big Data can lead to new insights by establishing unexpected correlations that allow for new predictions and hypotheses. Big Data alone, however, cannot establish causality (Leonelli, 2014). Correlations are sufficient in certain circumstances—for example, if a combination of genetic markers predicts that a drug will likely work for a particular patient, there is no need to establish causality in order to help the patient. However, not being able to establish causality is an unsatisfactory situation in basic research, such as research done with model organisms such as *C. elegans*. "Big Data" does not tell us how to build a worm, but it does provide us with a blueprint and a parts list. Therefore, "Big Data" in basic research is a tool and the starting point for further experimentation, not the endpoint. In one form or another, "Big Data" has been driving scientific progress in *C. elegans* since Brenner chose it for his studies. The complete lineage, connectome, and genome sequence provided a foundation for testing hypothesis and accelerated discoveries. Because of its relative simplicity compared with other metazoans, collecting Big Data in *C. elegans* historically has not depended on the development of revolutionary technologies. The lineage was worked out by simple observation without the sophisticated time-lapse recordings available today, the wiring diagram of all 302 neurons was established by manual tracing of neuronal processes on printouts of 7000 electron micrograph images without the aid of computers, and the genome was sequenced using standard Sanger sequencing techniques with incremental improvements long before next-generation sequencing tools became available. Each of these achievements took years of focused, dedicated effort to achieve $n$ = all, and building the connectome and sequencing the genome were viewed as high throughput for their time. Incremental changes to technology, as well as all new technologies, allow these types of data to be obtained faster and at much lower cost today.

Achieving *n* = all for many biological data may be an unreachable or unrealistic goal. A question to ask is whether there is a point at which data collecting is "good enough," even if it is not comprehensive. Depending on the scientific question, it might be possible to get a sufficiently "complete" answer with a limited data set. For example, one can get a good understanding of the principles of pattern formation in early embryogenesis from a selected set of informative mutants without having to analyze every single gene involved in the process, so not being able to achieve *n* = all does not necessarily mean that a scientific question is unsolvable.

The *C. elegans* community has a long history of sharing and pooling data. One of our colleagues referred to organizing large collaborative research efforts as akin to "herding cats," and so perhaps we are being Pollyannaish, but we believe that Big Data for this organism can be generated in a distributed manner, as we suggested for expression profiling of genes, protein–protein interaction maps, and large-scale studies of mutant phenotypes. "Big Data" has played a big part in *C. elegans* research in the past and will continue to accelerate scientific discoveries for the foreseeable future. It would certainly help if granting agencies appreciate this and commit to supporting these types of projects over the long term and to their completion. Obtaining *n* = all might not be sexy, but for carefully selected data sets, it should help to guide further experiments to understand basic questions in biology. As the following quote implies, data alone will not solve any problem:

> Actually, the orgy of fact extraction in which everybody is currently engaged has, like most consumer economies, accumulated a vast debt. This is a debt of theory, and some of us are soon going to have an exciting time paying it back—with interest, I hope.
>
> *Sydney Brenner (1997)*

## ACKNOWLEDGMENTS

Several colleagues made useful comments in an attempt to keep us from completely embarrassing ourselves, including Jack Chen, Mark Edgley, Linda Matsuuchi, Corey Nislow, Ralf Schnabel, Karina Simonsen, Tiffany Timbers, Ryan Viveiros, and Robert Waterston. The reader can decide whether their efforts were in vain. We also thank two anonymous reviewers for their comments. Research in the laboratories of H.H. and D.M. is sponsored by the Canadian Institute for Health Research and the Natural Sciences and Engineering Research Council of Canada. D.M. is a Senior Fellow of the Genetics Network node of the Canadian Institute for Advanced Research.

## REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al.* (2000). Gene ontology: tool for the unification of biology. Gene Ontology Consortium Nat Genet 25, 25–29.

Brenner S (1995). Loose ends. Curr Biol 5, 332.

Brenner S (1997). In theory. Curr Biol 7, R202.

*C. elegans* Deletion Mutant Consortium (2012). Large-scale screening for targeted knockouts in the Caenorhabditis elegans genome. G3 (Bethesda) 2, 1415–1425.

Chong YT, Koh JL, Friesen H, Duffy SK, Cox MJ, Moses A, Moffat J, Boone C, Andrews BJ (2015). Yeast proteome dynamics from single cell imaging and automated analysis. Cell 161, 1413–1424.

Fonslow BR, Moresco JJ, Tu PG, Aalto AP, Pasquinelli AE, Dillin AG, Yates JR (2014). Mass spectrometry-based shotgun proteomic analysis of *C. elegans* protein complexes. WormBook 2014(Jun 2), 1–18.

Fraser AG, Kamath RS, Zipperlen P, Martinez-Campos M, Sohrmann M, Ahringer J (2000). Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. Nature 408, 325–330.

Gibson W (1999). The science in science fiction [Interview]. National Public Radio. Talk of the Nation 1999 November 30. Available at www.npr.org/templates/story/story.php?storyId=1067220 (accessed 20 September 2015).

Gönczy P, Echeverri C, Oegema K, Coulson A, Jones SJ, Copley RR, Duperon J, Oegema J, Brehm M, Cassin E, *et al.* (2000). Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. Nature 408, 331–336.

Hall D, Altun Z (2008). *C. elegans* Atlas, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Harris TW, Baran J, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, Done J, Grove C, Howe K, *et al.* (2014). WormBase 2014: new views of curated biology. Nucleic Acids Res 42 (Database Issue), D789–D793.

Hillier LW, Coulson A, Murray JI, Bao Z, Sulston JE, Waterston RH (2005). Genomics in C. elegans: so many genes, such a little worm. Genome Res 15, 1651–1660.

Husson SJ, Costa WS, Schmitt C, Gottschalk A (2012). Keeping track of worm trackers. WormBook 2013(Feb 22), 1–17.

Jarrell TA, Wang Y, Bloniarz AE, Brittin CA, Xu M, Thomson JN, Albertson DG, Hall DH, Emmons SW (2012). The connectome of a decision-making neural network. Science 337, 437–444.

Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337, 816–821.

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, *et al.* (2003). Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421, 231–237.

Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, Chaerkady R, Madugundu AK, Kelkar DS, Isserlin R, Jain S, *et al.* (2014). A draft map of the human proteome. Nature 509, 575–581.

Leonelli S (2014). What difference does quantity make? On the epistemology of big data in biology. Big Data Soc 1, doi: 10.1177/2053951714534395.

Mayer-Schönberger M, Cukier K (2013). Big Data: A Revolution That Will Transform How We Live, Work, and Think, London: John Murray.

Merrihew GE, Davis C, Ewing B, Williams G, Kall L, Frewen BE, Noble WS, Green P, Thomas JH, MacCoss MJ (2008). Use of shotgun proteomics for the identification, confirmation, and correction of *C. elegans* gene annotations. Genome Res 18, 1660–1669.

Montojo J, Zuberi K, Rodriguez H, Bader GD, Morris Q (2014). GeneMANIA: fast gene network construction and function prediction for Cytoscape. F1000Res 3, 153.

Murray JI, Bao Z, Boyle TJ, Boeck ME, Mericle BL, Nicholas TJ, Zhao Z, Sandel MJ, Waterston RH (2008). Automated analysis of embryonic gene expression with cellular resolution in C. elegans. Nat Methods 5, 703–709.

Omasits U, Quebatte M, Stekhoven DJ, Fortes C, Roschitzki B, Robinson MD, Dehio C, Ahrens CH (2013). Directed shotgun proteomics guided by saturated RNA-seq identifies a complete expressed prokaryotic proteome. Genome Res 11, 1916–1927.

Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008). Imaging individual mRNA molecules using multiple singly labeled probes. Nat Methods 5, 877–879.

Reece-Hoyes JS, Deplancke B, Shingles J, Grove CA, Hope IA, Walhout A (2005). A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. Genome Biol 6, R110.

Rhee SY, Wood V, Dolinski K, Draghici S (2008). Use and misuse of the gene ontology annotations. Nat Rev Genet 9, 509–515.

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, *et al.* (2014). A proteome-scale map of the human interactome network. Cell 159, 1212–1226.

Schindelman G, Fernandes JS, Bastiani CA, Yook K, Sternberg PW (2011). Worm Phenotype Ontology: integrating phenotype data within and beyond the C. elegans community. BMC Bioinformatics 12, 32.

Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, Malmström J, Brunner E, Mohanty S, Lercher MJ, Hunziker PE, *et al.* (2009). Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. PLoS Biol 7, e48.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498–2504.

Simonis N, Rual JF, Carvunis AR, Tasan M, Lemmens I, Hirozane-Kishikawa T, Hao T, Sahalie JM, Venkatesan K, Gebreab F, *et al.* (2009). Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat Methods 6, 47–54.

Sulston JE, Brenner S (1974). The DNA of Caenorhabditis elegans. Genetics 77, 95–104.

Sulston JE, Horvitz HR (1977). Post-embryonic cell lineages of the nematode Caenorhabditis elegans. Dev Biol 56, 11–56.

Sulston JE, Schierenberg E, White JG, Thomson JN (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Dev Biol 100, 64–119.

Thompson O, Edgley M, Strasbourger P, Flibotte S, Ewing B, Adair R, Au V, Chaudhry I, Fernando L, Hutter H, *et al.* (2013). The Million Mutation Project: a new approach to genetics in Caenorhabditis elegans. Genome Res 23, 1749–1762.

Tkach JM, Yimit A, Lee AY, Riffle M, Costanzo M, Jaschob D, Hendry JA, Ou J, Moffat J, Boone C, *et al.* (2012). Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress. Nat Cell Biol 9, 966–976.

Walther DM, Kasturi P, Zheng M, Pinkert S, Vecchi G, Ciryam P, Morimoto RI, Dobson CM, Vendruscolo M, Mann M, Hartl FU (2015). Widespread proteome remodeling and aggregation in aging C. elegans. Cell 161, 919–932.

Weirauch MT, Hughes TR (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Subcell Biochem 52, 25–73.

White JG, Southgate E, Thomson JN, Brenner S (1986). The structure of the nervous system of the nematode *Caenorhabditis elegans*. Philos Trans R Soc Lond B Biol Sci 314, 1–340.

Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, Savitski MM, Ziegler E, Butzmann L, Gessulat S, Marx H, *et al.* (2014). Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587.

Wood WB, ed. (2015). The Nematode *Caenorhabditis elegans*, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Yuet KP, Doma MK, Ngo JT, Sweredoski MJ, Graham RL, Moradian A, Hess S, Schuman EM, Sternberg PW, Tirrell DA (2015). Cell-specific proteomic analysis in *Caenorhabditis elegans*. Proc Natl Acad Sci USA 112, 2705–2710.

Zhu Q, Wong AK, Krishnan A, Aure MR, Tadych A, Zhang R, Corney DC, Greene CS, Bongo LA, Kristensen VN, *et al.* (2015). Targeted exploration and analysis of large cross-platform human transcriptomic compendia. Nat Methods 12, 211–214.