



Opinion

Data Incompleteness May form a Hard-to-Overcome Barrier to Decoding Life's Mechanism

Liya Kondratyeva ^{1,*}, Irina Alekseenko ^{1,2}, Igor Chernov ¹ and Eugene Sverdlov ^{2,3,*}

¹ Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry of the Russian Academy of Sciences, Moscow 117997, Russia

² Institute of Molecular Genetics of National Research Centre “Kurchatov Institute”, Moscow 123182, Russia

³ Kurchatov Center for Genome Research, National Research Center “Kurchatov Institute”, Moscow 123182, Russia

* Correspondence: liakondratyeva@yandex.ru (L.K.); edsverd@gmail.com (E.S.)

Simple Summary: The influence of data incompleteness on the correctness of conclusions about the structure and functions of the objects under study is widely discussed in the literature. It was noted that even a small percentage of missing data can lead to incorrect conclusions and imperfect knowledge. In particular, incompleteness can lead to critical errors in the qualitative and quantitative assessments of interactions in biological systems and a distorted understanding of the functioning mechanisms of living systems. In this brief review, we attempt to demonstrate the extent of this incompleteness in functional information about living systems using the best-studied examples. We suggest that this incompleteness may form seemingly insurmountable barriers in deciphering the mechanisms of the functioning of complex systems with unpredictable properties arising from the interaction of the system components.

Abstract: In this brief review, we attempt to demonstrate that the incompleteness of data, as well as the intrinsic heterogeneity of biological systems, may form very strong and possibly insurmountable barriers for researchers trying to decipher the mechanisms of the functioning of live systems. We illustrate this challenge using the two most studied organisms: *E. coli*, with 34.6% genes lacking experimental evidence of function, and *C. elegans*, with identified proteins for approximately 50% of its genes. Another striking example is an artificial unicellular entity named JCVI-syn3.0, with a minimal set of genes. A total of 31.5% of the genes of JCVI-syn3.0 cannot be ascribed a specific biological function. The human interactome mapping project identified only 5–10% of all protein interactions in humans. In addition, most of the available data are static snapshots, and it is barely possible to generate realistic models of the dynamic processes within cells. Moreover, the existing interactomes reflect the de facto interaction but not its functional result, which is an unpredictable emerging property. Perhaps the completeness of molecular data on any living organism is beyond our reach and represents an unsolvable problem in biology.

Keywords: bioinformatics; big data; genome; systems biology; complexity



Citation: Kondratyeva, L.; Alekseenko, I.; Chernov, I.; Sverdlov, E. Data Incompleteness May form a Hard-to-Overcome Barrier to Decoding Life's Mechanism. *Biology* **2022**, *11*, 1208. <https://doi.org/10.3390/biology11081208>

Academic Editor: Tianwei Yu

Received: 19 June 2022

Accepted: 10 August 2022

Published: 12 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In December 2021, a series of papers was published by a research group involved in an eight-year project to reproduce the results of cancer preclinical trials described in more than 50 papers between 2010 and 2012 [1–4]. The shockingly high irreproducibility revealed by this group has recently been discussed in various scientific journals. “One of my biggest frustrations as a scientist is that it is so hard to know which exciting results are sturdy enough to build on”; this is how Dr. Yusuf A. Hannun, director of the Stony Brook University Cancer Center in New York, reacted to the results in his comment in *Nature* [5]. The estimates published in other fields are quite similar. This irreproducibility calls into

question the reliability of the conclusions regarding the mechanisms of disease development and, consequently, the development of methods for their treatment. Irreproducibility is not limited to cancer research. The problem was reviewed in detail by Begley and Ioannidis [6]. These authors underlined an important problem: “The variability of biological systems means that we should not expect an obligatory reproduction of the results to the smallest detail”.

Leaving aside numerous other shortcomings of the available data that have been noted, especially since the new generation of sequencing (NGS) has ushered life sciences into the era of “big data” (defined in [7–10]), the main challenge of biological big data is the low quality of the data themselves and their annotations. Genome-wide analyses give the impression of a huge amount of information, which nevertheless contains a large number of false positives and false negatives that mislead researchers [11–14].

In this review, we focus on important and seemingly insurmountable barriers faced by researchers who attempt to obtain correct conclusions about the mechanisms of life. This is data incompleteness, $n \neq \text{all}$ (Figure 1). To address the phenomenon of missing data, we searched the Google scholar database for review articles that include “missing data” in their title published since 2021. The search revealed 58 relevant articles. Almost all articles related to biological and medical research highlight the problem of unavoidable missing data. The problem of missing data is also widely observed in real-life databases, which often are imperfect, so that only incomplete, undefined, and invalid data sets are available. Missing data can be caused by a lot of factors, such as human errors in the experimental process, technical errors in the software or corrupt records in the databases, etc. [15].

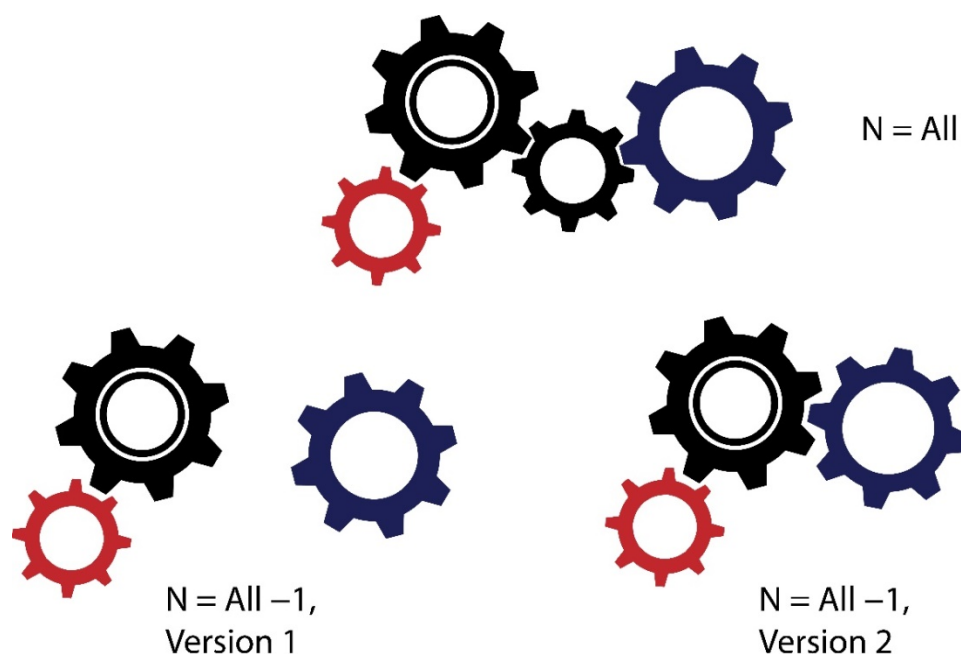


Figure 1. A simple illustration of the data incompleteness problem. Achieving “ $n = \text{all}$ ” for many biological data may be an unreachable or unrealistic goal. With the same set of incomplete data, it is possible to arrive at different versions of the organization of any biological process.

We use for our demonstration the two most studied organisms: *E. coli*, with 600 genes—that is, 34.6% of 4623 unique genes—lacking experimental evidence of function, and *C. elegans*, with identified proteins for approximately 50% of its genes. In addition, for *C. elegans*, 23% of the protein-coding genes have no phenotypic or cell-specific expression data, and 96% of protein–protein interactions are yet to be documented. Another striking example is an artificial unicellular bacteria-like entity named JCVI-syn3.0, which has in its genome a minimal set of genes—473. In this reduced set, function could not be defined for 149 genes

(31.5% of the genome). Only approximately 5–10% of all protein–protein interactions in the human interactome could be identified in recent studies [10,16,17].

These figures reflect the fundamental difficulty of deciphering the mechanisms of the functions of complex systems with unpredictable properties arising from interactions among system components.

2. Incompleteness of Genomic Data

Although research is currently developing in different areas within the multi-omics paradigm, the most advanced and most information-rich research is on genomes and transcriptomes and, to a much lesser extent, on proteomes. Therefore, to demonstrate the incompleteness of the available information, we mainly focus on the most advanced areas, using high-throughput NGS technologies. Interested readers can find information on multi-omics in the latest reviews [18–21]. We use two of the best-studied organisms, *E. coli* and *C. elegans*, to provide illustrative insight into the degree of data incompleteness.

3. *E. coli* Data Incompleteness

The first genome sequence of *Escherichia coli* was established in 1997. Meanwhile, the molecular and physiological functions of many genes are still unknown [22].

Studies on *Escherichia coli* K-12 have often been carried out with poorly annotated genes [23]. Using various databases (EcoCyc, EcoGene, UniProt, and RegulonDB), the authors identified 34.6% (1600 out of 4623) of the genes for which no function was experimentally detected. These unannotated *E. coli* genes have a name beginning with “y”, known as “y-genes”. Moreover, for 111 of the genes, the authors [23] found no information concerning their functions in the available knowledge bases such as UniProt, EcoCyc, and others. Unannotated genes in *E. coli* and other model organisms still play an important role in determining cell phenotypes [24–26]. The situation of the number, location, and strength of *E. coli* promoters is also incomplete. Recently (24 November 2021), 4042 promoters from the *Escherichia coli* K-12 strain MG1655 were reported ([27], EcoCyc base, <https://ecocyc.org/> (accessed on 10 December 2021)).

Only 2228 active promoters have been precisely mapped. The predicted promoter activity deduced from the genome sequence is highly unreliable [28], and the number of promoters functioning in *E. coli* is unknown, as is the extent to which the level of proteins in the cell is determined by regulation at the level of the promoter and, eventually, whether we can predict from the sequence if a promoter is contained within it, let alone the strength of the promoter and the mechanism of its regulation.

It is evident that predictions based on models of the genotype–phenotype intercommunication for whole *E. coli* cells may be highly unrealistic. In conclusion, it can be noted that in a recent study, the authors [29] compared the potential of various metrics to measure the similarity of phenotypic patterns to deduce gene function. Comparable results were obtained for most gene pairs for the three tested metrics. The conclusion was that, currently, there is no clearly preferred method of comparison.

4. One of the Best-Studied Multicellular Models, *C. elegans*, Is Still Very Far from “n = all”

In the late 1960s, Nobel Prize winner Sydney Brenner realized that a suitable organism for investigations into the development of the nervous system should contain a rather small number of cells, “so that exhaustive studies of lineage and patterns can be made” (quoted in [10]). The small roundworm (nematode) *Caenorhabditis elegans* was chosen for this purpose. Its relative simplicity allowed for the accumulation of data before the advent of next-generation sequencing. In an attempt to achieve “n = all” [30], numerous pieces of data were collected; the neuronal connections of all neurons were constructed from electron microscopy (302 neurons were from hermaphrodites, 385 were from males, and 294 neurons were common among them [31]), identifying the synaptic connections

“connectome”). The complete sequence of the genome, consisting of 100,291,840 base pairs, was determined [32].

The *C. elegans* knockout project has provided loss-of-function mutations in more than 14,000 of 20,000 protein-coding genes. It is incomplete, and it does not contain microRNAs or other RNA-modulating processes, nor control regions [10]. Recently, Li-Leger et al. [33] reported an estimation of the number of essential genes in *C. elegans* to be approximately 15%–30% of the 20,000 genes (for an essentiality definition see [34]). The authors identified 58 putative essential genes involved, in particular, in cell division and morphogenesis, and male-expressed genes required for fertilization and embryonic development. Another group [35] using genetic balancer systems, allowing the effective capture and maintenance of lethal mutations [36], identified 104 essential genes. This last group also reported that 604 essential genes were previously identified by other authors. One can see that the number of identified essential genes is much less than the number of previously estimated ones. More than 900 different families of transcripts have been predicted (see review in [31]). It was reported in 2015 [10] that about 10% of all TFs have been identified.

Only about 50% of the genes in *C. elegans* have been identified [10,37,38] (for comparison, about 86% of genes in humans have been associated with proteins [39], though in this case a complete standardized catalogue of protein-coding genes is also unavailable [40]). The *C. elegans* Deletion Mutant Consortium reported in 2012 the phenotypic identification of 6841 mutations in 6013 protein-coding genes [37,41]. Using Mos1 transposon insertions, 10,858 mutants in 4700 genes were exposed [42] (for a recent review, see [41]).

Identifying peptides for all exons and splice variants is still far from completion. It was assessed that 96% of the protein–protein interactions in *C. elegans* remain to be documented [43]. Such a shortage of data is also typical for other organisms. One should also keep in mind the nonlinear growth of the number of interactions with the number of proteins or genes. Therefore, “n = all” in these cases will be more difficult to achieve [44].

It was also pointed out that 23% of the protein-coding genes in WormBase release WS238 do not possess data concerning their cell-specific expression or phenotypic manifestation. There is also no GO annotation for them; therefore, they remain totally uncharacterized [10].

This lack of completeness of big data datasets poses a major challenge for data integration and the extraction of biological knowledge [10]. In addition, as we indicated before, big data collections are “noisy” and contain unreliable or even false data [10].

5. The Unsolved Mysteries of the Fully Synthetic JCVI-syn3 Genome

Recently, Venter and colleagues synthesized an artificial genome called JCVI-syn3.0, containing a minimal necessary set of DNA containing 473 genes (438 protein-coding genes and 35 genes for RNAs). The JCVI-syn3.0 genome is smaller than that of known natural independently replicating cells [45]. For 31.5% of the genes—149—of the minimized obligate genome, a specific biological function could not be ascribed [46]. Some of these important but “nonfunctional” genes appeared to be conserved across other species, including *Homo sapiens*. The elimination of all the “non-essential” genes resulted in a nonviable genome (reviewed in [47]). Moreover, 79 genes could not be assigned to even a broad functional category. The authors of [46] used *in silico* methods and assigned presumable functions to 66 of the 149 proteins. These proteins lack orthologues, lack protein domains, and/or are characteristic of membrane proteins. Among them, 24 are possible transporter proteins.

Several drawbacks, including extensive filamentation and vesicle formation, were revealed during JCVI-syn3.0 growth [48]; therefore, the new design JCVI-syn3A was constructed. It incorporated 19 additional genes from the very first version JCVI-syn1.0 that were not present in JCVI-syn3.0. With a 543 kbp genome and 492 genes (452 protein-encoding and 38 RNA-encoding), JCVI-syn3A still has a smaller genome than any natural autonomously replicating organism. However, the problems of unknown functions remained unchanged. Further information beyond the scope of this review, but which may be of interest to readers, can be found in recent articles [49,50].

In conclusion, it should be noted that missing functional annotation is quite common. It has been reported, in particular, that UniProt5 contains less than 1% of 148 million protein sequences with experimentally validated functions in gene ontology (GO) (April 2019) [46].

6. Complete Interactomes—An (Unreachable?) Dream of Systems Biologists

Two decades after the draft sequence of the human genome was published, the “entire” sequence was finally deciphered, with all gaps filled and errors corrected from previous versions ([51] and five accompanying papers in the same source). The complete sequence contains 3.055 billion base pairs (bps) in 22 autosomes plus Chromosome X. Nearly 200 million bps of the novel sequence were added, including 2226 paralogous gene copies; among these, 115 are probably protein-coding.

The complete reference genome sequence is very important; however, all genomic and cellular functions can be implemented only through the spatio-temporal interaction networks of cellular components [52–54].

Despite the advances in genome sequencing, transcriptome sequencing, and proteome sequencing, we are still far from understanding the cellular mechanisms responsible for organism function.

The entire complements of functional molecular associations that may occur in a cell are known as the “interactome”, and this encompasses a range of cellular functional networks. Complex interactome networks form through interactions among DNA and RNA molecules, also involving proteins, lipids, and small metabolites [55,56].

As a reference genome sequence is important for human genetics, reference interactome networks are critical for the full appreciation of genotype–phenotype relationships [57–59]. Currently, the existing information is highly fractional and represents only a small proteome fraction. Therefore, the construction and verification of the cellular interactome are two of the most important goals of genome functional analyses [60].

A variety of biochemical, genetic, and cellular methods have been developed for mapping interactomes [10,44,58], which are often represented as networks (graphs), allowing for both visual and computational analyses of their structure and connectivity ([61–66] with references from [67]). Figure 2 schematically illustrates the information that emerges from the research on interactomes.

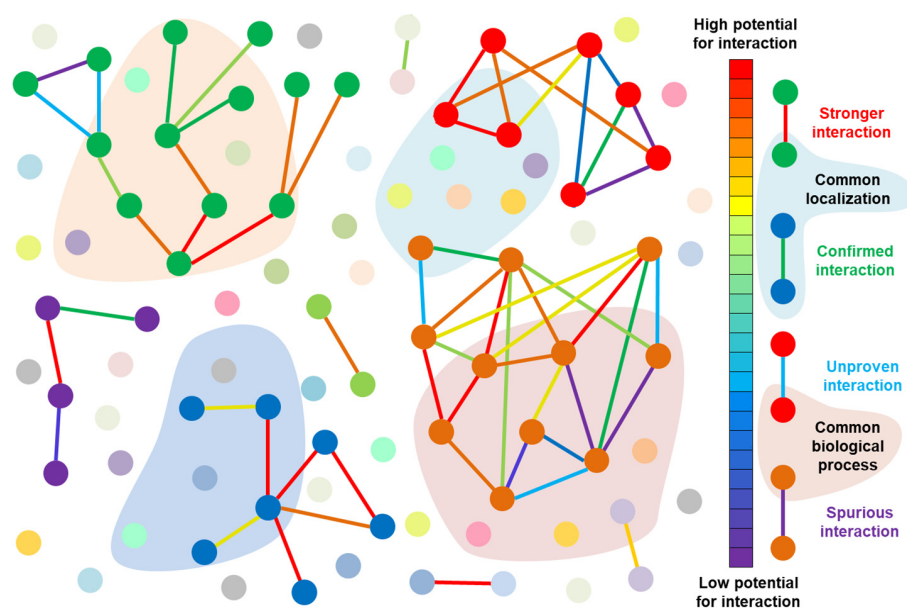


Figure 2. Illustration of a protein–protein interaction network: colored lines indicate an interaction between a pair of proteins, some of which are possibly spurious. The strength of the interaction between a pair of proteins is indicated by its color and is stronger when the color is closer to red. Translucent-colored clouds uniting proteins in a cluster symbolize common localization or function.

Recently, “a reference map of the human binary protein interactome” was published [17]. The authors [17] indicated that it is still not possible to generate a reference map of the interactome by systematically identifying protein–protein interactions (PPIs) in thousands of cellular variations. However, various technologies [16,68] have allowed for creating maps of the human protein interactome with high functional significance. In particular, using yeast two-hybrid (Y2H) assays, the authors previously generated an interactome consisting of ~14,000 PPIs [16] and presented a reference interactome map of human binary protein interactions—the Human Reference Interactome (HuRI), containing ~53,000 high-quality PPIs, which is significantly higher than previous data. Together with genome [69], transcriptome [70], and proteome [71] data, HuRI has allowed the obtainment of more information about cellular functions. The high-quality PPI data collection can be found at <http://interactome-atlas.org> (accessed on 10 March 2022).

An interesting comparison demonstrating how far we are from a complete human interactome recently appeared [72]. The authors noted that since PPIs are made up of two or more proteins, the total number of PPIs is much greater than the number of human protein-coding genes (about 19000, see above), and they provided the following data on the number of PPIs reported by different sources: The Center for Cancer Systems Biology (CCSB) Interactome Database—13,993 human PPIs [16]; HuRI (see above [17])—64,006 binary interactions; the STRING database [73]—505,116 high-confidence, experimental, or predicted human PPIs. Venkatesan et al. [74] compared the PPIs found with the yeast two-hybrid method and known human PPIs, and evaluated the size of the human interactome to be ~130,000. Stumpf et al. [75] estimated the size of the interactome as 650,000. One can see significant variability in the estimation, and an understandable question appeared recently as the heading of a paper: “How Far Are We from the Completion of the Human Protein Interactome Reconstruction?” [76].

The authors [76] wrote: “After more than fifteen years from the first high-throughput experiments for human protein–protein interaction (PPI) detection, we are still wondering how close the completion of the genome-scale human PPI network reconstruction is, what needs to be further explored, and whether the biological insights gained from the holistic investigation of the current network are valid and useful”. They suggest that “an almost complete picture of a structurally defined network has been reached”.

However, practically all major processes in a cell are implemented not by binary PPI interactions but by complexes consisting of several participants, and the cell can be considered an interactome of protein machines [77,78].

As indicated by Alberts [77], the cell can be thought of as a factory containing a complex network of interconnected assembly lines of large molecular machines. In addition, these machines are in permanent motion, whereas our data on interactomes are inevitably static [77]. Alberts also indicated that although we have undoubtedly made great progress in deciphering the structure of protein assemblies, we still have an enormous amount left to learn. Understanding a protein machine function will require, along with knowledge of its static structure, knowledge of the kinetics and energetics of all the intermediate products that are formed during the reaction. This remains a problem that will require methodologies that do not yet exist. The question is whether it will ever be possible to develop, for example, a technique that allows researchers “to follow the kinetics and structure of each of the intermediates involved in the many fascinating transport reactions that occur deep within the lipid bilayer membrane”. Alberts wrote about this in 1998; almost a quarter of a century has passed, but the situation has not changed considerably (see, for example [79–81]).

In the meantime, intensive work continues to attempt to use the capabilities of modern computer tools for the analysis of available interactomes, especially bacterial ones. As illustrative examples, one can cite the studies of Wuchty et al. [82] and Dilucca et al. [83].

In the first study [82], the authors attempted to summarize and analyze protein interactions using data from different sources. Two main observations stand out: (i) the evolutionary conservation of some interactions and the complete absence of others. For example, 80 protein complexes of *H. pylori* were also observed in *Escherichia coli*, whereas

120 complexes were not found in *E. coli*; (ii) the comparison of various species allowed for the obtainment of the putative functions of approximately 300 poorly characterized or previously uncharacterized proteins.

In the second study [83], the authors related the evolutionary conservation, essentiality, and functional repertoire of a gene to the connectivity k (the number of interprotein links) in the PPI network of 42 bacteria with genomes of different sizes, and reasonably separated evolutionary trajectories. In particular, they demonstrated that highly connected proteins (with connectivity a $k \geq 40$) were encoded by genes that were conserved and essential among the species considered [84]. Despite the undoubtedly interesting results, they should be treated with reasonable caution because of the incomplete nature of the interactomes used for the analyses. Huxley's (Thomas H. Huxley, the famous English biologist and anthropologist) warning always stands: "Mathematics may be compared to a mill of exquisite workmanship, which grinds you stuff of any degree of fineness; but, nevertheless, what you get out depends upon what you put in; and as the grandest mill in the world will not extract wheat flour from peas cods, so pages of formulae will not get a definite result out of loose data" [85].

7. Bio-Databases and Ontologies for Biomedical Literature: The Inherent Incompleteness of Gene Ontology

Thousands of global biodatabases currently exist (reviewed in [86]). UniProtKB now includes more than nine million entries. Swiss-Prot annotates UniProtKB records, and it is expected that it will soon contain 500,000 protein sequences. The research team manually annotated approximately half of these entries by analyzing thousands of articles and data from hundreds of other databases. The team members have processed an enormous amount of information: an estimated 25,000 peer-reviewed journals publishing about 2.5 million articles per year. For life sciences alone, this amounts to two new articles published in MEDLINE each minute [86].

We will use just one of the databases, gene ontology (GO), as an illustrative example. GO is a universal portal for operations with high-capacity biological datasets [87]. It is a formidable resource and it is relatively easy to use without a deep understanding of its structure. GO is very useful when dealing with large databases and data mining; however, as we explain later, we have to treat this information with caution. A very widespread type of analysis involves comparing gene sets using their functional annotations to detect the enrichment or depletion of functional groups in a given sample of genes. It can also be used to determine the relationship of certain functions to regulatory networks, sequence convergence or divergence, and other aspects of gene activities and evolution. Since the appearance of its introductory article [88], about 280,000 papers with the words "gene ontology" have been published (Google Scholar).

The following can be found in the current release (1 July 2022): 43,558 GO terms, 7,483,496 annotations, 1,480,259 gene products, and 5213 species (<http://geneontology.org/>, accessed on 10 July 2022). However, most of its gene annotations are scanty and incomplete [89,90] (also see below). As a result, it is rather difficult to identify associations among genes and a huge number of terms. The ontology is very dynamic and constantly improving in order to better represent the evolutionary and functional relationships among organisms. However, numerous aspects of the gene ontology database are still poorly comprehended. One should keep in mind that the information contained in the GO database is necessarily incomplete (see below for more details), and the absence of functional evidence does not mean that there is no function [87]. This is described as the open world assumption. Disregarding the open world assumption can cause inflated large numbers of false positive rates when using gene function prediction tools. Numerous pitfalls, depending on the structure of GO, the methods of compiling annotations, and their variability based on newly emerging data (dictating the need to use the latest version of GO), are given in two brilliant reviews [88,89].

As noted by Gaudet regarding GO, the central database for functional genetics, there are “misconceptions and misleading assumptions commonly made about GO, including the effect of data incompleteness, the importance of annotation qualifiers, and the transitivity or lack thereof associated with different ontology relations” [87].

While great efforts are being made to increase the coverage of annotated gene products, it should not be expected that eventually every gene product will be annotated.

Another problem is that the incompleteness is very uneven. In addition, even rather well-annotated GO parts can also cause trouble, providing users with seemingly contradictory results.

8. Conclusions: Is Data Completeness One More Unsolvable Problem of Biology?

The famous scientist Jan Baptista van Helmont (1580–1644) undertook an experiment to demonstrate where trees obtain material from for growth. He grew a willow with a pre-weighed volume of soil. After five years, he found that the willow weighed about 74 kg more than at the beginning. Since the weight of the soil did not change much, van Helmont concluded that the additional plant material only came from the water. Van Helmont was unaware of photosynthesis. This incomplete knowledge led to incorrect conclusions. This is a useful lesson for both data producers and analysts.

The incomplete nature of the available data is exacerbated by the fact that we do not really know how far we are from completeness. Although attempts are being made with the help of computer analysis to predict completeness (see, for example, [79–81]), it is not clear how realistic these predictions are given the poor quality of the data on which they are based.

In addition, there is a strongly uneven distribution among organisms, among their tissues and cells, and among their compartments [91]. Some of the reasons for this unevenness are objective, for example, the difficult accessibility of certain cells and tissues, while others are often of an organizational nature. This last problem was discussed by Alberts [91], who named it the “Canalization of Research Areas on the Principle of Training Inertia”. He noted that while many important areas of cell biology are little explored, there are overcrowded areas of science where many scientists conduct almost identical experiments. This is most likely because numerous students of each major scientist, after receiving a doctoral degree, create their own laboratories where they continue research in the same area. This may be because researchers are concerned about the risk of entering a new area, whereas the completeness of scientific knowledge requires many people to work in unexplored areas. Alberts gives the example of *Escherichia coli*, for which a little more than 4000 different proteins have been detected. However, even today, we have no idea what ~1000 of them do (see above). This is an unexplored area in which almost no one works today. This situation also occurs in more complex organisms. However, if we have such gaps in our knowledge, we cannot hope to understand the mechanisms of functioning of even the simplest living cells, not to mention human cells [92]. The modern system of rewards in science (the publication of articles and distribution of grants) supports this fear. Therefore, we must consider ways to encourage research in risky areas. Interestingly, when Alberts and others were making their fundamental discoveries, this fear was not so prevalent. An outstanding example is the work of Brenner on the development of the nematode *Caenorhabditis elegans* as an object of study in molecular genetics. Brenner decided to study this transparent worm to solve an important new problem in molecular biology, and he started this around 1966. The first paper on genetics and a number of *C. elegans* mutants, written exclusively by Brenner, appeared in 1974 [93]. For eight years, he worked without a single publication. Today, it is impossible to imagine the effectiveness of a scientist not being evaluated by the number of articles published.

Among the objective reasons for data incompleteness, the two most important are (i) the complex nature of the interactions among biological systems components, and (ii) that the majority of processes in cells are dynamic and occur over a large range of time scales. Both of these are discussed above.

In summary, every major process in a cell is materialized by complexes consisting of tens of protein molecules, each of which interacts with several other large complexes of proteins often connected to membranes or nucleic acids [77,78]. Combining many components with nonlinear interactions leads to new emergent properties that can only be understood in the context of the whole system. The emergent properties are nonlinear and cannot be described by the linear superposition of the experimental data [94,95].

Most of the available data are static snapshots, and it is barely possible to generate realistic models of the dynamic processes within cells occurring in times from milliseconds to years [94].

As a result, as formulated by Sydney Brenner: “We are drowning in a sea of data and starving for knowledge. Biological sciences have exploded largely through our unprecedented power to accumulate descriptive facts. How to understand genomes and how to use them is going to be a central task of our research in the future” [96]. In other places, Brenner asked: “Is there some other approach? If I knew it, I would be doing it, and not writing about the problem” [92].

We believe that no one knows. Some time ago, one of us indicated three unsolvable problems in biology: “It is impossible to create two identical organisms, to defeat cancer, or to map organisms onto their genomes” [95]. Perhaps the unattainability of the completeness of molecular data for any living organism is the fourth unsolvable problem.

Finally, it should be noted that old discussions of whether a “hypothesis first” [97] or “data first” [98] acceptance of the data incompleteness throw additional weight on the “hypothesis first” side, provided, of course, that the completeness and consistency of the data underlying the hypothesis are good enough (for a more current discussion, see: [99,100]). Completeness (the proportion of stored data against the potential of “ $n = \text{all}$ ”) is one of the most important characteristics of big data quality [101,102]. Another important characteristic is “consistency”; consistent representation is the degree, in particular, to which data are compatible with previous data [103]—that is, reproducible.

We have attempted to demonstrate that biological data (especially in the case of cancer) are incomplete and may be poorly reproducible. It seems axiomatic that incomplete data cannot be the basis for constructing a correct theory or for designing an effective treatment system in medicine. This has recently been dramatically demonstrated by a series of hasty and incorrect conclusions drawn from incomplete data on the effectiveness of anti-infection drugs against SARS-CoV-2 [104]. However, as the history of science shows, even incomplete data can serve as a basis for proposing and testing hypotheses, which must be consistent with the fundamental principle of falsifying (discrediting), that is, containing a system of its own refutation. The principle was put forward in 1934 by Sir Karl Popper [105], and since then has proven its significance time and again. This was once again confirmed by the previously mentioned Taran et al. [104] on the example of SARS-CoV-2 drug searches.

Unfortunately, we do not know what we do not know. This important problem was discussed in a comprehensive review by Hutter and Moerman [10]. The authors noted that it is difficult to even define completeness, that achieving $n = \text{all}$ for many biological data may be an unreachable or unrealistic goal, and that now the problem is whether there is a point at which data collection is “good enough,” even if it is not comprehensive. Depending on the scientific question, it might be possible to obtain a sufficiently “complete” understanding of a complex system functioning with a limited data set. They believe that “not being able to achieve $n = \text{all}$ does not necessarily mean that a scientific question is unsolvable”. We completely agree with such an optimistic point of view.

The general strategies for data completeness improvement and minimization of the incompleteness effect were reviewed in detail by Begley and Ioannidis [6] and Danchin et al. [14]. Are there any strategies for the improvement of data completeness? At the dawn of genomics in 2000, Nature Genetics published a commentary titled “Grass-roots genomics” where the authors gave the following recipe for deciphering the organisms: “Genomic technologies can give hints about the functions of genes, but the information they get is usually too limited to draw any conclusions from it. These hints can only point the

way to discovery if they are used by someone with experience and intuition to see the direction they point. There is only one way—gene by gene, process by process, researcher by researcher—by which we can “decipher the organism” [106].

Author Contributions: Conceptualization, E.S.; writing—original draft preparation, E.S. and L.K.; writing—review and editing, I.A., E.S. and I.C.; supervision—I.A., E.S.; funding acquisition, E.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Russian Science Foundation (project no. 22-14-00308).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Errington, T.M.; Mathur, M.; Soderberg, C.K.; Denis, A.; Perfito, N.; Iorns, E.; Nosek, B.A. Investigating the replicability of preclinical cancer biology. *eLife* **2021**, *10*, e71601. [[CrossRef](#)]
2. Errington, T.M.; Denis, A.; Perfito, N.; Iorns, E.; Nosek, B.A. Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology. *eLife* **2021**, *10*, e67995. [[CrossRef](#)] [[PubMed](#)]
3. Errington, T.M.; Denis, A.; Allison, A.B.; Araiza, R.; Aza-Blanc, P.; Bower, L.R.; Campos, J.; Chu, H.; Denson, S.; Donham, C. Experiments from unfinished Registered Reports in the Reproducibility Project: Cancer Biology. *eLife* **2021**, *10*, e73430. [[CrossRef](#)]
4. Rodgers, P.; Collings, A. Reproducibility in Cancer Biology: What have we learned? *eLife* **2021**, *10*, e75830. [[CrossRef](#)]
5. Hannun, Y.A. Build a registry of results that students can replicate. *Nature* **2021**, *600*, 571. [[CrossRef](#)]
6. Begley, C.G.; Ioannidis, J.P. Reproducibility in science: Improving the standard for basic and preclinical research. *Circ. Res.* **2015**, *116*, 116–126. [[CrossRef](#)] [[PubMed](#)]
7. Helzlsouer, K.; Meerzaman, D.; Taplin, S.; Dunn, B.K. Humanizing Big Data: Recognizing the Human Aspect of Big Data. *Front. Oncol.* **2020**, *10*, 186. [[CrossRef](#)]
8. Stevens, M.; Wehrens, R.; de Bont, A. Conceptualizations of Big Data and their epistemological claims in healthcare: A discourse analysis. *Big Data Soc.* **2018**, *5*, 2053951718816727. [[CrossRef](#)]
9. Mayer-Schonberger, V.; Cukier, K. *Big Data: A Revolution that will Transform How We Live, Work, and Think*; Mariner Books; Houghton Mifflin Harcourt: Boston, MA, USA, 2014.
10. Hutter, H.; Moerman, D. Big Data in *Caenorhabditis elegans*: Quo vadis? *Mol. Biol. Cell* **2015**, *26*, 3909–3914. [[CrossRef](#)]
11. Aggarwal, S.; Raj, A.; Kumar, D.; Dash, D.; Yadav, A.K. False discovery rate: The Achilles’ heel of proteogenomics. *Brief. Bioinform.* **2022**, bbac163. [[CrossRef](#)] [[PubMed](#)]
12. Elouataoui, W.; Alaoui, I.E.; Gahi, Y. Data Quality in the Era of Big Data: A Global Review. *Big Data Intell. Smart Appl.* **2022**, *994*, 1–25. [[CrossRef](#)]
13. Kasif, S.; Roberts, R.J. We need to keep a reproducible trace of facts, predictions, and hypotheses from gene to function in the era of big data. *PLoS Biol.* **2020**, *18*, e3000999. [[CrossRef](#)]
14. Danchin, A.; Ouzounis, C.; Tokuyasu, T.; Zucker, J.D. No wisdom in the crowd: Genome annotation in the era of big data—Current status and future prospects. *Microb. Biotechnol.* **2018**, *11*, 588–605. [[CrossRef](#)] [[PubMed](#)]
15. Nijman, S.; Leeuwenberg, A.M.; Beekers, I.; Verkouter, I.; Jacobs, J.; Bots, M.L.; Asselbergs, F.W.; Moons, K.; Debray, T. Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *J. Clin. Epidemiol.* **2022**, *142*, 218–229. [[CrossRef](#)] [[PubMed](#)]
16. Rolland, T.; Tasan, M.; Charlotiaux, B.; Pevzner, S.J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.; et al. A proteome-scale map of the human interactome network. *Cell* **2014**, *159*, 1212–1226. [[CrossRef](#)] [[PubMed](#)]
17. Luck, K.; Kim, D.K.; Lambourne, L.; Spirohn, K.; Begg, B.E.; Bian, W.; Brignall, R.; Cafarelli, T.; Campos-Laborie, F.J.; Charlotiaux, B.; et al. A reference map of the human binary protein interactome. *Nature* **2020**, *580*, 402–408. [[CrossRef](#)] [[PubMed](#)]
18. Tarazona, S.; Arzalluz-Luque, A.; Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat. Comput. Sci.* **2021**, *1*, 395–402. [[CrossRef](#)]
19. Miao, Z.; Humphreys, B.D.; McMahon, A.P.; Kim, J. Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **2021**, *17*, 710–724. [[CrossRef](#)]
20. Wu, S.; Chen, D.; Snyder, M.P. Network biology bridges the gaps between quantitative genetics and multi-omics to map complex diseases. *Curr. Opin. Chem. Biol.* **2022**, *66*, 102101. [[CrossRef](#)] [[PubMed](#)]
21. Vahabi, N.; Michailidis, G. Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review. *Front. Genet.* **2022**, *13*, 854752. [[CrossRef](#)]
22. Kurokawa, M.; Ying, B.W. Experimental Challenges for Reduced Genomes: The Cell Model *Escherichia coli*. *Microorganisms* **2019**, *8*, 3. [[CrossRef](#)] [[PubMed](#)]

23. Ghatak, S.; King, Z.A.; Sastry, A.; Palsson, B.O. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Res.* **2019**, *47*, 2446–2454. [[CrossRef](#)] [[PubMed](#)]
24. Dellomonaco, C.; Clomburg, J.M.; Miller, E.N.; Gonzalez, R. Engineered reversal of the beta-oxidation cycle for the synthesis of fuels and chemicals. *Nature* **2011**, *476*, 355–359. [[CrossRef](#)] [[PubMed](#)]
25. Sandberg, T.E.; Pedersen, M.; LaCroix, R.A.; Ebrahim, A.; Bonde, M.; Herrgard, M.J.; Palsson, B.O.; Sommer, M.; Feist, A.M. Evolution of Escherichia coli to 42 degrees C and subsequent genetic engineering reveals adaptive mechanisms and novel mutations. *Mol. Biol. Evol.* **2014**, *31*, 2647–2662. [[CrossRef](#)] [[PubMed](#)]
26. Hufnagel, D.A.; DePas, W.H.; Chapman, M.R. The disulfide bonding system suppresses CsgD-independent cellulose production in Escherichia coli. *J. Bacteriol.* **2014**, *196*, 3690–3699. [[CrossRef](#)] [[PubMed](#)]
27. Keseler, I.M.; Gama-Castro, S.; Mackie, A.; Billington, R.; Bonavides-Martinez, C.; Caspi, R.; Kothari, A.; Krummenacker, M.; Midford, P.E.; Muniz-Rascado, L.; et al. The EcoCyc Database in 2021. *Front. Microbiol.* **2021**, *12*, 711077. [[CrossRef](#)] [[PubMed](#)]
28. Urtecho, G.; Tripp, A.D.; Insigne, K.D.; Kim, H.; Kosuri, S. Systematic Dissection of Sequence Elements Controlling sigma70 Promoters Using a Genomically Encoded Multiplexed Reporter Assay in Escherichia coli. *Biochemistry* **2019**, *58*, 1539–1551. [[CrossRef](#)] [[PubMed](#)]
29. Wu, P.I.; Ross, C.; Siegele, D.A.; Hu, J.C. Insights from the reanalysis of high-throughput chemical genomics data for Escherichia coli K-12. *G3* **2021**, *11*, jkaa035. [[CrossRef](#)] [[PubMed](#)]
30. Glenwinkel, L.; Taylor, S.R.; Langebeck-Jensen, K.; Pereira, L.; Reilly, M.B.; Basavaraju, M.; Rafi, I.; Yemini, E.; Pocock, R.; Sestan, N.; et al. In silico analysis of the transcriptional regulatory logic of neuronal identity specification throughout the C. elegans nervous system. *eLife* **2021**, *10*, e64906. [[CrossRef](#)]
31. Godini, R.; Handley, A.; Pocock, R. Transcription Factors That Control Behavior-Lessons From C. elegans. *Front. Neurosci.* **2021**, *15*, 745376. [[CrossRef](#)]
32. Hillier, L.W.; Coulson, A.; Murray, J.I.; Bao, Z.; Sulston, J.E.; Waterston, R.H. Genomics in C. elegans: So many genes, such a little worm. *Genome Res.* **2005**, *15*, 1651–1660. [[CrossRef](#)]
33. Li-Leger, E.; Feichtinger, R.; Flibotte, S.; Holzkamp, H.; Schnabel, R.; Moerman, D.G. Identification of essential genes in Caenorhabditis elegans through whole-genome sequencing of legacy mutant collections. *G3* **2021**, *11*, jkab328. [[CrossRef](#)] [[PubMed](#)]
34. Rancati, G.; Moffat, J.; Typas, A.; Pavelka, N. Emerging and evolving concepts in gene essentiality. *Nat. Rev. Genet.* **2018**, *19*, 34–49. [[CrossRef](#)]
35. Yu, S.; Zheng, C.; Chu, J.S. Identification of Essential Genes in Caenorhabditis elegans with Lethal Mutations Maintained by Genetic Balancers. *Methods Mol. Biol.* **2022**, *2377*, 345–362. [[CrossRef](#)]
36. Edgley, M.; Baillie, D.; Riddle, D.; Rose, A. Genetic Balancers. *WormBook: The Online Review of C. elegans Biology.* *Nucleic Acids Res.* **2007**, *35*, D472-5. [[CrossRef](#)]
37. Consortium, C.e.D.M. large-scale screening for targeted knockouts in the Caenorhabditis elegans genome. *G3* **2012**, *2*, 1415–1425. [[CrossRef](#)]
38. Walther, D.M.; Kasturi, P.; Zheng, M.; Pinkert, S.; Vecchi, G.; Ciryam, P.; Morimoto, R.I.; Dobson, C.M.; Vendruscolo, M.; Mann, M.; et al. Widespread Proteome Remodeling and Aggregation in Aging C. elegans. *Cell* **2015**, *161*, 919–932. [[CrossRef](#)] [[PubMed](#)]
39. Hatje, K.; Muhlhause, S.; Simm, D.; Kollmar, M. The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *Bioessays* **2019**, *41*, e1900066. [[CrossRef](#)] [[PubMed](#)]
40. Wilhelm, M.; Schlegl, J.; Hahne, H.; Gholami, A.M.; Lieberenz, M.; Savitski, M.M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582–587. [[CrossRef](#)] [[PubMed](#)]
41. Campos, T.L.; Korhonen, P.K.; Hofmann, A.; Gasser, R.B.; Young, N.D. Harnessing model organism genomics to underpin the machine learning-based prediction of essential genes in eukaryotes—Biotechnological implications. *Biotechnol. Adv.* **2021**, *54*, 107822. [[CrossRef](#)] [[PubMed](#)]
42. Vallin, E.; Gallagher, J.; Granger, L.; Martin, E.; Belougne, J.; Maurizio, J.; Duverger, Y.; Scaglione, S.; Borrel, C.; Cortier, E.; et al. A genome-wide collection of Mos1 transposon insertion mutants for the C. elegans research community. *PLoS ONE* **2012**, *7*, e30482. [[CrossRef](#)] [[PubMed](#)]
43. Simonis, N.; Rual, J.F.; Carvunis, A.R.; Tasan, M.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Sahalie, J.M.; Venkatesan, K.; Gebreab, F.; et al. Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network. *Nat. Methods* **2009**, *6*, 47–54. [[CrossRef](#)] [[PubMed](#)]
44. Rimmelzwaal, S.; Boxem, M. Protein interactome mapping in Caenorhabditis elegans. *Curr. Opin. Syst. Biol.* **2019**, *13*, 1–9. [[CrossRef](#)] [[PubMed](#)]
45. Hutchison, C.A., 3rd; Chuang, R.Y.; Noskov, V.N.; Assad-Garcia, N.; Deerinck, T.J.; Ellisman, M.H.; Gill, J.; Kannan, K.; Karas, B.J.; Ma, L.; et al. Design and synthesis of a minimal bacterial genome. *Science* **2016**, *351*, aad6253. [[CrossRef](#)]
46. Antczak, M.; Michaelis, M.; Wass, M.N. Environmental conditions shape the nature of a minimal bacterial genome. *Nat. Commun.* **2019**, *10*, 3100. [[CrossRef](#)]
47. Coyle, M.; Hu, J.; Gartner, Z. Mysteries in a Minimal Genome. *ACS Cent. Sci.* **2016**, *2*, 274–277. [[CrossRef](#)] [[PubMed](#)]
48. Breuer, M.; Earnest, T.M.; Merryman, C.; Wise, K.S.; Sun, L.; Lynott, M.R.; Hutchison, C.A.; Smith, H.O.; Lapek, J.D.; Gonzalez, D.J.; et al. Essential metabolism for a minimal cell. *eLife* **2019**, *8*, e36842. [[CrossRef](#)]

49. Pelletier, J.F.; Glass, J.I.; Strychalski, E.A. Cellular mechanics during division of a genomically minimal cell. *Trends Cell Biol.* **2022**, preprint. [[CrossRef](#)]
50. Zhang, C.; Zheng, W.; Cheng, M.; Omenn, G.S.; Freddolino, P.L.; Zhang, Y. Functions of Essential Genes and a Scale-Free Protein Interaction Network Revealed by Structure-Based Function and Interaction Prediction for a Minimal Genome. *J. Proteome Res.* **2021**, *20*, 1178–1189. [[CrossRef](#)]
51. Nurk, S.; Koren, S.; Rhie, A.; Rautiainen, M.; Bzikadze, A.V.; Mikheenko, A.; Vollger, M.R.; Altemose, N.; Uralsky, L.; Gershman, A.; et al. The complete sequence of a human genome. *bioRxiv* **2021**. [[CrossRef](#)]
52. Hartwell, L.H.; Hopfield, J.J.; Leibler, S.; Murray, A.W. From molecular to modular cell biology. *Nature* **1999**, *402*, C47–C52. [[CrossRef](#)] [[PubMed](#)]
53. Eisenberg, D.; Marcotte, E.M.; Xenarios, I.; Yeates, T.O. Protein function in the post-genomic era. *Nature* **2000**, *405*, 823–826. [[CrossRef](#)]
54. Brehme, M.; Vidal, M. A global protein-lipid interactome map. *Mol. Syst. Biol.* **2010**, *6*, 443. [[CrossRef](#)]
55. Kunowska, N.; Stelzl, U. Decoding the cellular effects of genetic variation through interaction proteomics. *Curr. Opin. Chem. Biol.* **2021**, *66*, 102100. [[CrossRef](#)] [[PubMed](#)]
56. Luck, K.; Sheynkman, G.M.; Zhang, I.; Vidal, M. Proteome-Scale Human Interactomics. *Trends Biochem. Sci.* **2017**, *42*, 342–354. [[CrossRef](#)] [[PubMed](#)]
57. Yook, S.H.; Oltvai, Z.N.; Barabasi, A.L. Functional and topological characterization of protein interaction networks. *Proteomics* **2004**, *4*, 928–942. [[CrossRef](#)]
58. Snider, J.; Kotlyar, M.; Saraon, P.; Yao, Z.; Jurisica, I.; Stagljar, I. Fundamentals of protein interaction network mapping. *Mol. Syst. Biol.* **2015**, *11*, 848. [[CrossRef](#)]
59. Vidal, M.; Cusick, M.E.; Barabasi, A.L. Interactome networks and human disease. *Cell* **2011**, *144*, 986–998. [[CrossRef](#)]
60. Cusick, M.E.; Klitgord, N.; Vidal, M.; Hill, D.E. Interactome: Gateway into systems biology. *Hum. Mol. Genet.* **2005**, *14*, R171–R181. [[CrossRef](#)]
61. Huber, W.; Carey, V.J.; Long, L.; Falcon, S.; Gentleman, R. Graphs in molecular biology. *BMC Bioinform.* **2007**, *8* (Suppl. 6), S8. [[CrossRef](#)]
62. Koh, G.C.; Porras, P.; Aranda, B.; Hermjakob, H.; Orchard, S.E. Analyzing protein-protein interaction networks. *J. Proteome Res.* **2012**, *11*, 2014–2031. [[CrossRef](#)]
63. Mason, O.; Verwoerd, M. Graph theory and networks in Biology. *IET Syst. Biol.* **2007**, *1*, 89–119. [[CrossRef](#)] [[PubMed](#)]
64. Bu, D.; Zhao, Y.; Cai, L.; Xue, H.; Zhu, X.; Lu, H.; Zhang, J.; Sun, S.; Ling, L.; Zhang, N.; et al. Topological structure analysis of the protein-protein interaction network in budding yeast. *Nucleic Acids Res.* **2003**, *31*, 2443–2450. [[CrossRef](#)] [[PubMed](#)]
65. Jeong, H.; Mason, S.P.; Barabasi, A.L.; Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **2001**, *411*, 41–42. [[CrossRef](#)]
66. Wuchty, S.; Oltvai, Z.N.; Barabasi, A.L. Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat. Genet.* **2003**, *35*, 176–179. [[CrossRef](#)]
67. James, K.; Wipat, A.; Cockell, S. Expanding Interactome Analyses beyond Model Eukaryotes. *Brief Funct. Genom.* **2021**, *21*, 243–269. [[CrossRef](#)]
68. Rual, J.F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G.F.; Gibbons, F.D.; Dreze, M.; Ayivi-Guedehoussou, N.; et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **2005**, *437*, 1173–1178. [[CrossRef](#)] [[PubMed](#)]
69. Amberger, J.S.; Bocchini, C.A.; Schiettecatte, F.; Scott, A.F.; Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **2015**, *43*, D789. [[CrossRef](#)] [[PubMed](#)]
70. Mele, M.; Ferreira, P.G.; Reverter, F.; DeLuca, D.S.; Monlong, J.; Sammeth, M.; Young, T.R.; Goldmann, J.M.; Pervouchine, D.D.; Sullivan, T.J.; et al. Human genomics. The human transcriptome across tissues and individuals. *Science* **2015**, *348*, 660–665. [[CrossRef](#)] [[PubMed](#)]
71. Thul, P.J.; Akesson, L.; Wiking, M.; Mahdessian, D.; Geladaki, A.; Ait Blal, H.; Alm, T.; Asplund, A.; Bjork, L.; Breckels, L.M.; et al. A subcellular map of the human proteome. *Science* **2017**, *356*, eaal3321. [[CrossRef](#)] [[PubMed](#)]
72. Shin, W.H.; Kumazawa, K.; Imai, K.; Hirokawa, T.; Kihara, D. Current Challenges and Opportunities in Designing Protein-Protein Interaction Targeted Drugs. *Adv. Appl. Bioinform. Chem. AABC* **2020**, *13*, 11–25. [[CrossRef](#)] [[PubMed](#)]
73. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [[CrossRef](#)] [[PubMed](#)]
74. Venkatesan, K.; Rual, J.F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.L.; et al. An empirical framework for binary interactome mapping. *Nat. Methods* **2009**, *6*, 83–90. [[CrossRef](#)] [[PubMed](#)]
75. Stumpf, M.P.; Thorne, T.; de Silva, E.; Stewart, R.; An, H.J.; Lappe, M.; Wiuf, C. Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 6959–6964. [[CrossRef](#)] [[PubMed](#)]
76. Dimitrakopoulos, G.N.; Klapa, M.I.; Moschonas, N.K. How Far Are We from the Completion of the Human Protein Interactome Reconstruction? *Biomolecules* **2022**, *12*, 140. [[CrossRef](#)]
77. Alberts, B. The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists. *Cell* **1998**, *92*, 291–294. [[CrossRef](#)]

78. von Hippel, P.H. From “simple” DNA-protein interactions to the macromolecular machines of gene expression. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 79–105. [[CrossRef](#)] [[PubMed](#)]
79. Plewczynski, D.; Ginalski, K. The interactome: Predicting the protein-protein interactions in cells. *Cell. Mol. Biol. Lett.* **2009**, *14*, 1–22. [[CrossRef](#)] [[PubMed](#)]
80. Kovacs, I.A.; Luck, K.; Spirohn, K.; Wang, Y.; Pollis, C.; Schlabach, S.; Bian, W.; Kim, D.K.; Kishore, N.; Hao, T.; et al. Network-based prediction of protein interactions. *Nat. Commun.* **2019**, *10*, 1240. [[CrossRef](#)]
81. Johnson, K.L.; Qi, Z.; Yan, Z.; Wen, X.; Nguyen, T.C.; Zaleta-Rivera, K.; Chen, C.J.; Fan, X.; Sriram, K.; Wan, X.; et al. Revealing protein-protein interactions at the transcriptome scale by sequencing. *Mol. Cell* **2021**, *81*, 3877. [[CrossRef](#)]
82. Wuchty, S.; Muller, S.A.; Caufield, J.H.; Hauser, R.; Aloy, P.; Kalkhof, S.; Uetz, P. Proteome Data Improves Protein Function Prediction in the Interactome of *Helicobacter pylori*. *Mol. Cell. Proteom. MCP* **2018**, *17*, 961–973. [[CrossRef](#)]
83. Dilucca, M.; Cimini, G.; Giansanti, A. Bacterial Protein Interaction Networks: Connectivity is Ruled by Gene Conservation, Essentiality and Function. *Curr. Genom.* **2021**, *22*, 111–121. [[CrossRef](#)] [[PubMed](#)]
84. Luo, H.; Gao, F.; Lin, Y. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Sci. Rep.* **2015**, *5*, 13210. [[CrossRef](#)] [[PubMed](#)]
85. Huxley, T. Thomas Huxley Quotes. Available online: <https://www.quotes.net/quote/56043> (accessed on 10 March 2022).
86. Attwood, T.K.; Kell, D.B.; McDermott, P.; Marsh, J.; Pettifer, S.R.; Thorne, D. Calling International Rescue: Knowledge lost in literature and data landslide! *Biochem. J.* **2009**, *424*, 317–333. [[CrossRef](#)]
87. Gaudet, P.; Dessimoz, C. Gene Ontology: Pitfalls, Biases, and Remedies. In *The Gene Ontology Handbook, Methods in Molecular Biology*; Dessimoz, C., Škunca, N., Eds.; Springer Open Humana Press: Berlin/Heidelberg, Germany, 2017; Volume 1446.
88. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [[CrossRef](#)]
89. Zhao, Y.; Wang, J.; Chen, J.; Zhang, X.; Guo, M.; Yu, G. A Literature Review of Gene Function Prediction by Modeling Gene Ontology. *Front. Genet.* **2020**, *11*, 400. [[CrossRef](#)]
90. Zhang, D.; Guelfi, S.; Garcia-Ruiz, S.; Costa, B.; Reynolds, R.H.; D’Sa, K.; Liu, W.; Courtin, T.; Peterson, A.; Jaffe, A.E.; et al. Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders. *Sci. Adv.* **2020**, *6*, eaay8299. [[CrossRef](#)] [[PubMed](#)]
91. Alberts, B. Biology Past and Biology Future: Where have we been and where are we going. *Neural Regen. Res.* **2013**, *8*, 2309–2316.
92. Brenner, S. Loose ends. *Curr. Biol.* **1995**, *5*, 1328. [[CrossRef](#)]
93. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **1974**, *77*, 71–94. [[CrossRef](#)]
94. Brenner, S. Sequences and consequences. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **2010**, *365*, 207–212. [[CrossRef](#)] [[PubMed](#)]
95. Sverdlov, E.D. Unsolvable Problems of Biology: It Is Impossible to Create Two Identical Organisms, to Defeat Cancer, or to Map Organisms onto Their Genomes. *Biochemistry* **2018**, *83*, 370–380. [[CrossRef](#)] [[PubMed](#)]
96. Brenner, S. Nobel lecture: Nature’s gift to science. *Biosci. Rep.* **2003**, *23*, 225–237. [[CrossRef](#)]
97. Weinberg, R. Point: Hypotheses first. *Nature* **2010**, *464*, 678. [[CrossRef](#)] [[PubMed](#)]
98. Golub, T. Counterpoint: Data first. *Nature* **2010**, *464*, 679. [[CrossRef](#)] [[PubMed](#)]
99. Hulsen, T.; Jamuar, S.S.; Moody, A.R.; Karnes, J.H.; Varga, O.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E.F. From Big Data to Precision Medicine. *Front. Med.* **2019**, *6*, 34. [[CrossRef](#)] [[PubMed](#)]
100. Voit, E.O. Perspective: Dimensions of the scientific method. *PLoS Comput. Biol.* **2019**, *15*, e1007279. [[CrossRef](#)] [[PubMed](#)]
101. Ramasamy, A.; Chowdhury, S. Big data quality dimensions: A systematic literature review. *JISTEM-J. Inf. Syst. Technol. Manag.* **2020**, *17*, e202017003. [[CrossRef](#)]
102. Hassenstein, M.J.; Vanella, P. Data Quality—Concepts and Problems. *Encyclopedia* **2022**, *2*, 498–510. [[CrossRef](#)]
103. Wang, R.Y.; Strong, D.M. Beyond accuracy: What data quality means to data consumers. *J. Manag. Inf. Syst.* **1996**, *12*, 5–33. [[CrossRef](#)]
104. Taran, S.; Adhikari, N.K.; Fan, E. Falsifiability in medicine: What clinicians can learn from Karl Popper. *Intensive Care Med.* **2021**, *47*, 1054–1056. [[CrossRef](#)] [[PubMed](#)]
105. Popper, K. *The Logic of Scientific Discovery*; Routledge: Oxfordshire, UK, 2005.
106. Johnston, M.; Fields, S. Grass-roots genomics. *Nat. Genet.* **2000**, *24*, 5–6. [[CrossRef](#)] [[PubMed](#)]