Genome Biology

Genomics and data science: an application within an umbrella

Fábio C. P. Navarro[1,2], Hussein Mohsen[1,2], Chengfei Yan[1,2], Shantao Li[5,6], Mengting Gu[1,2], William Meyerson[1,2] and Mark Gerstein[1,2,3,4*] ⓘ

Abstract

Data science allows the extraction of practical insights from large-scale data. Here, we contextualize it as an umbrella term, encompassing several disparate subdomains. We focus on how genomics fits as a specific application subdomain, in terms of well-known 3 V data and 4 M process frameworks (volume-velocity-variety and measurement-mining-modeling-manipulation, respectively). We further analyze the technical and cultural "exports" and "imports" between genomics and other data-science subdomains (e.g., astronomy). Finally, we discuss how data value, privacy, and ownership are pressing issues for data science applications, in general, and are especially relevant to genomics, due to the persistent nature of DNA.

Introduction

Data science as a formal discipline is currently popular because of its tremendous commercial utility. Large companies have used several well-established computational and statistical techniques to mine high volumes of commercial and social data [1]. The broad interest across many applications stirred the birth of data science as a field that acts as an umbrella, uniting a number of disparate disciplines using a common set of computational approaches and techniques [2]. In some cases, these techniques were created, developed, or established in other data-driven fields (e.g., astronomy and earth science). In fact, some of these disciplines significantly predate the formal foundation of data science and have contributed to several techniques to cope with knowledge extraction from large amounts of data.

Many scholars have probed the origins of data science. For example, in 1960 Tukey described a new discipline called data analysis, which some consider being a forerunner of data science. He defined data analysis as the interplay between statistics, computer science, and mathematics [3]. Jim Gray also introduced the concept of data-intensive science in his book *The Fourth Paradigm* [4], and discussed how the developments in computer science would shape and transform segments of science to a data-driven exercise. More practically, the maturation of modern data science from an amorphous discipline can be tracked to the expansion of the technology industry and its adoption of several concepts at the confluence of statistics and algorithmic computer science, such as machine learning [5]. Somewhat less explored is the fact that several applied disciplines have contributed to a collection of techniques and cultural practices that today comprise data science.

Contextualizing natural science within the data science umbrella

Long before the development of formal data science, and even computer science or statistics, traditional fields of natural sciences established an extensive culture around data management and analytics. For instance, physics has a long history of contributions of several concepts that are now at the foundation of data science. In particular, physicists such as Laplace, Gauss, Poisson, and Dirichlet have led the way for the development of hypothesis testing, least squares fits, and Gaussian, Poisson, and Dirichlet distributions [6].

More recently, physics also has contributed new data techniques and data infrastructure. For example, Ulam originally invented the Monte Carlo sampling method while he was working on the hydrogen bomb [7] and

* Correspondence: mark@gersteinlab.org
[1]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA
[2]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

Berners-Lee, from the CERN (European Organization for Nuclear Research), developed the World Wide Web [8] to enable distributed collaboration in particle physics. While most disciplines are now experiencing issues with rapid data growth [9, 10], we find it interesting that physics had issues with data management long before most disciplines. As early as the 1970s, for example, Jashcek introduced the term "information explosion" to describe the rapid data growth in astrophysics [11].

Fundamental contributions to data management and analytics have not been exclusive to physics. The biological sciences, perhaps most prominently genetics, also have significantly influenced data science. For instance, many of the founders of modern statistics, including Galton, Pearson, and Fisher, pioneered principal component analysis, linear regression, and linear discriminant analysis while they were also preoccupied with analyzing large amounts of biological data [6]. More recently, methods such as logistic regression [12], clustering [13], decision trees [14], and neural networks [15] were either conceptualized or developed by researchers focused on biological questions. Even Shannon, a central figure in information theory, completed a short PhD in population genetics [16].

Genomics and data science

More recent biological disciplines such as macromolecular structure and genomics have inherited many of these data analytics features from genetics and other natural sciences. Genomics, for example, emerged in the 1980s at the confluence of genetics, statistics, and large-scale datasets [17]. The tremendous advancements in nucleic acid sequencing allowed the discipline to swiftly assume one of the most prominent positions in terms of raw data scale across all the sciences [18]. This pre-eminent role of genomics also inspired the emergence of many "-omics" terms inside and outside academia [19, 20]. Although today genomics is pre-eminent in terms of data scale, this may change over time due to technological developments in other areas, such as cryo-electron microscopy [21] and personal wearable devices [22]. Moreover, it is important to realize that many other existing data-rich areas in the biological sciences are also rapidly expanding, including image processing (including neuroimaging), macromolecular structure, health records analysis, proteomics, and the inter-relation of these large data sets, in turn, is giving rise to a new subfield termed biomedical data science (Fig. 1a).

Here, we explore how genomics has been, and probably will continue to be, a pre-eminent data science subdiscipline in terms of data growth and availability. We first explore how genomics data can be framed in terms of the 3Vs (data volume, velocity, and variety) to contextualize

the discipline in the "big-data world". We also explore how genomics processes can be framed in terms of the 4Ms (measurement, mining, modeling, and manipulating) to discuss how physical and biological modeling can be leveraged to generate better predictive models. Genomics researchers have been exchanging ideas with those from other data science subfields; we review some of these "imports" and "exports" in a third section. Finally, we explore issues related to data availability in relation to data ownership and privacy. Altogether, this perspective discusses the past, present, and future of genomics as a subfield of data science.
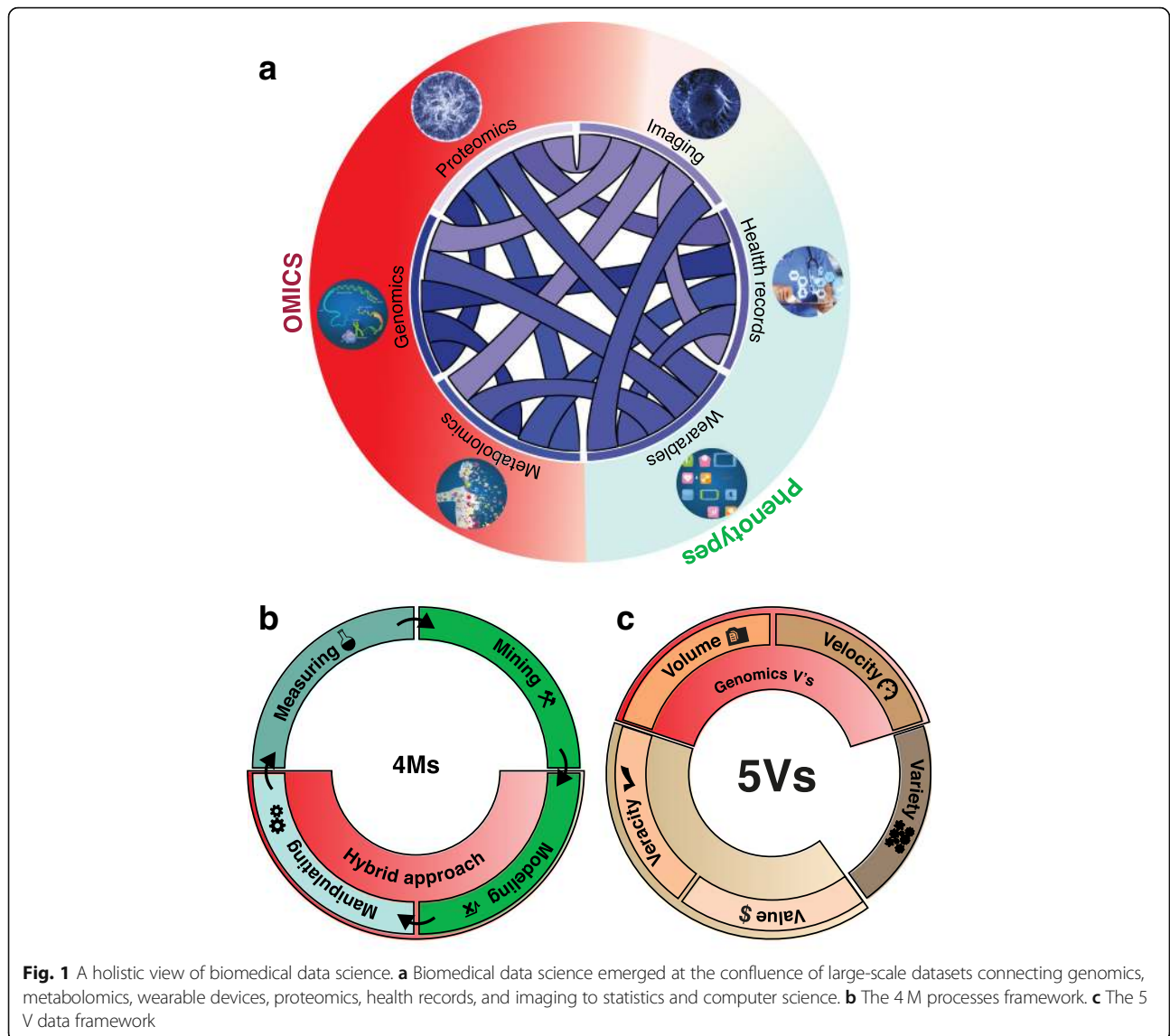
Genomics versus other data science applications in terms of the V framework

One way of categorizing the data in data science disciplines is in terms of its volume, velocity, and variety. Within data science, this is broadly referred to as the V framework [23]. Over the years, the V framework has been expanded from its original 3Vs [24] (volume, velocity, and variety) to the most recent versions with four and five Vs ($3V + value$ and veracity; Fig. 1c) [25]. In general, the distinct V frameworks use certain data-related parameters to recognize issues and bottlenecks that might require a new set of tools and techniques to cope with unstructured and high-volume data. Here, we explore how we can use the original 3 V framework to evaluate the current state of data in genomics in relation to other applications in data sciences.

Volume

One of the key aspects of genomics as a data science is the sheer amount of data being generated by sequencers. As shown in Fig. 2, we tried to put this data volume into context by comparing genomics datasets with other data-intensive disciplines. Figure 2a shows that the total volume of data in genomics is considerably smaller than the data generated by earth science [26], but orders of magnitude larger than the social sciences. The data growth trend in genomics, however, is greater than in other disciplines. In fact, some researchers have suggested that if the genomics data generation growth trend remains constant, genomics will soon generate more data than applications such as social media, earth sciences, and astronomy [27].

Many strategies have been used to address the increase in data volume in genomics. For example, researchers are now tending to discard primary data (e.g., FASTQ) and prioritizing the storage of secondary data such as compressed mapped reads (BAMs), variant calls (VCFs), or even only quantifications such as gene expression [28].

Navarro *et al. Genome Biology* (2019) 20:109

Page 3 of 11

Fig. 1 A holistic view of biomedical data science. **a** Biomedical data science emerged at the confluence of large-scale datasets connecting genomics, metabolomics, wearable devices, proteomics, health records, and imaging to statistics and computer science. **b** The 4 M processes framework. **c** The 5 V data framework

In Fig. 2b, we compare genomics to other data-driven disciplines in the biological sciences. This analysis clearly shows that the large amount of early biological data was not in genomics, but rather in macromolecular structure. Only in 2001, for example, did the number of datasets in genomics finally surpass protein-structure data. More recently, new trends have emerged with the rapidly increasing amount of electron microscopy data, due to the advent of cryo-electron microscopy, and of mass-spectrometry-based proteomics data. Perhaps these trends will shift the balance of biomedical data science in the future.
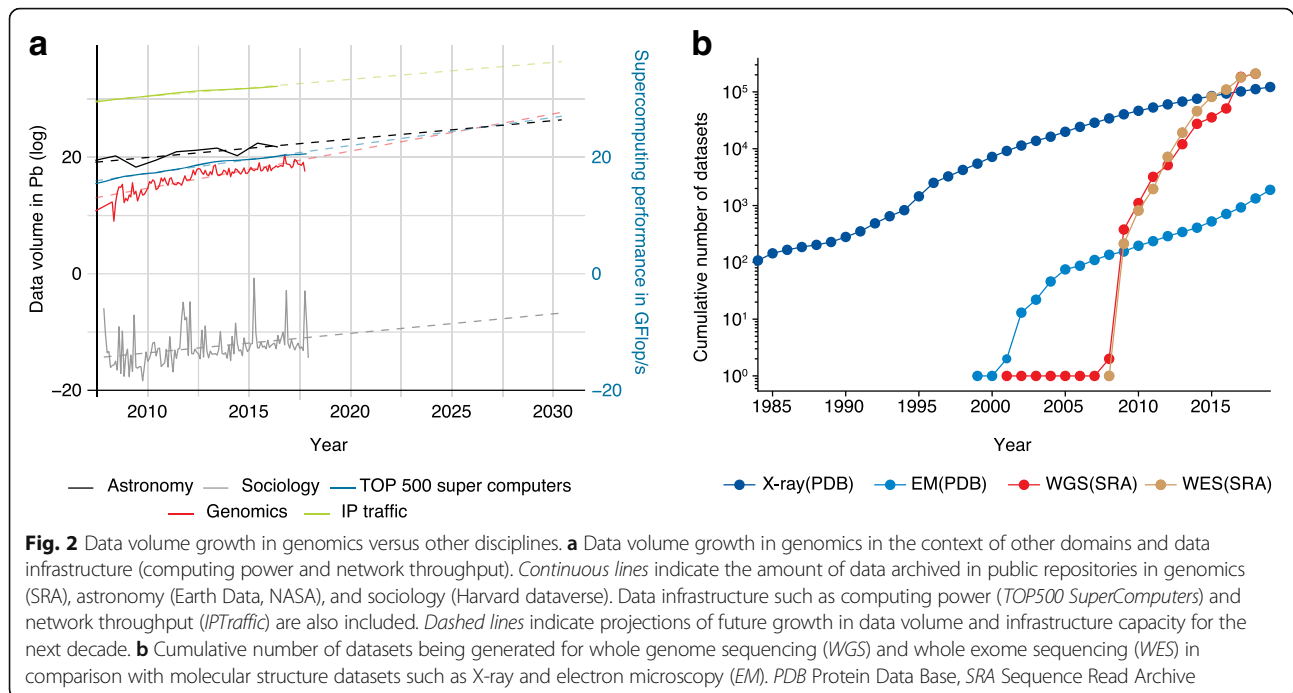
Velocity

There are two widely accepted interpretations of data velocity: (i) the speed of data generation (Fig. 2) and

(ii) the speed at which data are processed and made available [29].

We explored the growth of data generation in the previous section in relation to genomics. The sequencing of a human genome could soon take less than 24 h, down from 2 to 8 weeks by currently popular technologies and 13 years of uninterrupted sequencing work by the Human Genome Project (HGP) [30]. Other technologies, such as diagnostic imaging and microarrays, have also experienced remarkable drops in cost and complexity and, therefore, resulting data are much quicker to generate.

The second definition of data velocity speaks to the speed at which data are processed. A remarkable example is the speed of fraud detection during a credit card transaction or some types of high-frequency trading

Fig. 2 Data volume growth in genomics versus other disciplines. **a** Data volume growth in genomics in the context of other domains and data infrastructure (computing power and network throughput). *Continuous lines* indicate the amount of data archived in public repositories in genomics (SRA), astronomy (Earth Data, NASA), and sociology (Harvard dataverse). Data infrastructure such as computing power (*TOP500 SuperComputers*) and network throughput (*IPTraffic*) are also included. *Dashed lines* indicate projections of future growth in data volume and infrastructure capacity for the next decade. **b** Cumulative number of datasets being generated for whole genome sequencing (*WGS*) and whole exome sequencing (*WES*) in comparison with molecular structure datasets such as X-ray and electron microscopy (*EM*). *PDB* Protein Data Base, *SRA* Sequence Read Archive

in finance [31]. In contrast, genomics data and data processing have been traditionally static, relying on fixed snapshots of genomes or transcriptomes. However, new fields leveraging rapid sequencing technologies, such as rapid diagnosis, epidemiology, and microbiome research, are beginning to use nucleic acid sequences for fast, dynamic tracking of diseases [32] and pathogens [33]. For these and other near-future technologies, we envision that fast, real-time processing might be necessary.

The description of the volume and velocity of genomics data has great implications for what types of computations are possible. For instance, when looking at the increase of genomics and other types of data relative to network traffic and bandwidth, one must decide whether to store, compute, or transfer datasets. This decision-making process can also be informed by the 3 V framework. In Fig. 2, we show that the computing power deployed for research and development (using the top 500 supercomputers as a proxy) is growing at a slower pace than genomic data growth. Additionally, while the global web traffic throughput has no foreseeable bottlenecks (Fig. 2a) [34], for researchers the costs of transferring such large-scale datasets might hinder data sharing and processing of large-scale genomics projects. Cloud computing is one way of addressing this bottleneck. Large consortia already tend to process and store most of their datasets on the cloud [35–37]. We believe genomics should consider the viability of public repositories that leverage cloud computing more broadly. At the current rate, the field will soon reach a critical point
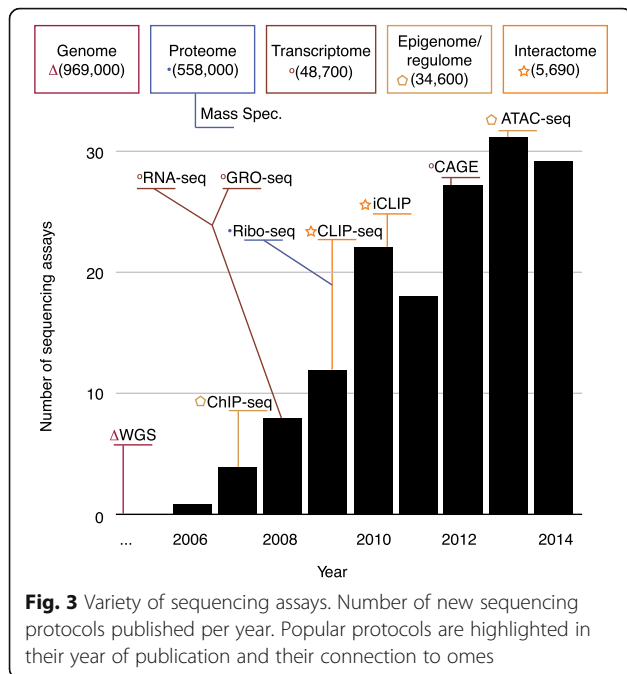
at which cloud solutions might be indispensable for large-scale analysis.

Variety

Genomics data have a two-sided aspect. On one side is the monolithic sequencing data, ordered lists of nucleotides. In human genomics, traditionally these are mapped to the genome and are used to generate coverage or variation data. The monolithic nature of sequencing output, however, hides a much more varied set of assays that are used to measure many aspects of genomes. In Fig. 3 we illustrate this issue by showing the growth in the diversity of sequencing assays over time and displaying a few examples. We also display how different sequencing methods are connected to different omes [19]. The other side of genomics data is the complex phenotypic data with which the nucleotides are being correlated. Phenotypic data can consist of such diverse entities as simple and unstructured text descriptions from electronic health records, quantitative measurements from laboratories, sensors, and electronic trackers, and imaging data. The varied nature of the phenotypic data is more complicated; as the scale and diversity of sequencing data grow larger, more attention is being paid to the importance of standardizing and scaling the phenotypic data in a complementary fashion. For example, mobile devices can be used to harness large-scale consistent digital phenotypes [38].

Genomics and the 4 M framework

Two aspects distinguish data science in the natural sciences from social science context. First, in the

Fig. 3 Variety of sequencing assays. Number of new sequencing protocols published per year. Popular protocols are highlighted in their year of publication and their connection to omes

natural sciences much of the data are quantitative and structured; they often derive from sensor readings from experimental systems and observations under well-controlled conditions. In contrast, data in the social sciences are more frequently unstructured and derived from more subjective observations (e.g., interviews and surveys). Second, the natural sciences also have underlying chemical, physical, and biological models that are often highly mathematized and predictive.

Consequently, data science mining in the natural sciences is intimately associated with mathematical modeling. One succinct way of understanding this relationship is the 4 M framework, developed by Lauffenburger [39]. This concept describes the overall process in systems biology, closely related to genomics, in terms of (i) Measuring the quantity, (ii) large-scale Mining, which is what we often think of as data science, (3) Modeling the mined observations, and finally (4) Manipulating or testing this model to ensure it is accurate.

The hybrid approach of combining data mining and biophysical modeling is a reasonable way forward for genomics (Fig. 1b). Integrating physical–chemical mechanisms into machine learning provides valuable interpretability, boosts the data-efficiency in learning (e.g., through training-set augmentation and informative priors) and allows data extrapolation when observations are expensive or impossible [40]. On the other hand, data mining is able to accurately estimate model parameters, replace some complex parts of the models where theories are weak, and emulate some physical models for computational efficiency [41].

Short-term weather forecasting as an exemplar of this hybrid approach is perhaps what genomics is striving for. For this discipline, predictions are based on sensor data from around the globe and then fused with physical models. Weather forecasting was, in fact, one of the first applications of large-scale computing in the 1950s [42, 43]. However, it was an abject flop, trying to predict the weather solely based on physical models. Predictions were quickly found to only be correct for a short time, mostly because of the importance of the initial conditions. That imperfect attempt contributed to the development of the fields of nonlinear dynamics and chaos, and to the coining of the term "butterfly effect" [43]. However, subsequent years dramatically transformed weather prediction into a great success story, thanks to integrating physically based models with large datasets measured by satellites, weather balloons, and other sensors [43]. Moreover, the public's appreciation for the probabilistic aspects of a weather forecast (i.e., people readily dress appropriately based on a chance of rain) foreshadows how it might respond to probabilistic "health forecasts" based on genomics.

Imports and exports

Thus far, we have analyzed how genomics sits with other data-rich subfields in terms of data (volume, velocity, and variety) and processes. We argue that another aspect of genomics as an applied data science subfield is the frequent exchange of techniques and cultural practices. Over the years, genomics has imported and exported several concepts, practices, and techniques from other applied data science fields. While listing all of the movements is impossible in this piece, we will highlight a few key examples.

Technical imports

A central aspect of genomics—the process of mapping reads to the human reference genome—relies on a foundational technique within data science: fast and memory-efficient string-processing algorithms. Protein pairwise alignment predates DNA sequence alignment. One of the first successful implementations of sequence alignment was based on Smith–Waterman [44] and dynamic programming [45, 46]. These methods were highly reliant on computing power and required substantial memory. With advances in other string-alignment techniques and the explosion of sequencing throughput, the field of genomics saw a surge in the performance of sequence alignment. As most sequencing technologies produce short reads, researchers generated several new methods using index techniques, starting around 2010. Several methods are now based on the Burrows–Wheeler transformation (BWA, bowtie) [47, 48], De Bruijn

graphs (Kallisto, Salmon) [49, 50], and the Maximal Mappable Prefix (STAR) [51].

Hidden Markov models (HMMs) are well-known algorithms used for modeling the sequential or time-series correlations between symbols or events. HMMs have been widely adopted in fields such as speech recognition and digital communication [52]. Data scientists also have long used HMMs to smooth a series of events in a varied number of datasets, such as the stock market, text suggestions, and in silico diagnosis [53]. The field of genomics has applied HMMs to predict chromatin states, annotate genomes, and study ancestry/population genetics [54]. Figure 4a displays the adoption of HMMs in genomics compared with other disciplines. It shows that the fraction of HMM papers related to genomics has been growing over time and today it corresponds to more than a quarter of the scientific publications related to the topic.
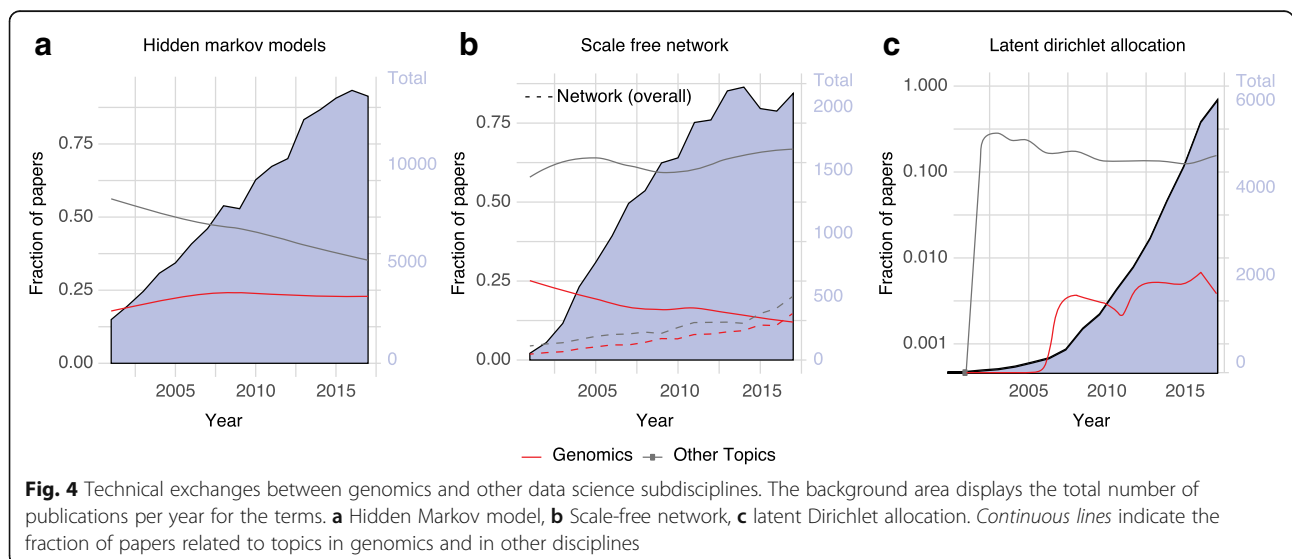
Another major import into genomics has been network science and, more broadly, graphs. Other subfields have been using networks for many tasks, including algorithm development [55], social network research [56], and modeling transportation systems [57]. Many subfields of genomics rely heavily on networks to model different aspects of the genome and subsequently generate new insights [58]. One of the first applications of networks within genomics and proteomics was protein–protein interaction networks [59]. These networks are used to describe the interaction between several protein(s) and protein domains within a genome to ultimately infer functional pathways [60]. After the development of large-scale transcriptome quantification and chromatin immunoprecipitation sequencing (ChIP-Seq), researchers built regulatory

networks to describe co-regulated genes and learn more about pathways and hub genes [61]. Figure 4b shows the usage of "scale-free networks" and "networks" as a whole. While the overall use of networks has continued to grow in popularity in genomics after their introduction, the specific usage of scale-free has been falling, reflecting the brief moment of popularity of this concept.

Given the abundance of protein structures and DNA sequences, there has been an influx of deep-learning solutions imported from machine learning [62]. Many neural network architectures can be transferred to biological research. For example, the convolutional neural network (CNN) is widely applied in computer vision to detect objects in a positional invariant fashion. Similarly, convolution kernels in CNN are able to scan biological sequences and detect motifs, resembling position weight matrices (PWMs). Researchers are developing intriguing implementations of deep-learning networks to integrate large datasets, for instance, to detect gene homology [63], annotate and predict regulatory regions in the genome [64], predict polymer folding [65], predict protein binding [66], and predict the probability of a patient developing certain diseases from genetic variants [67]. While neural networks offer a highly flexible and powerful tool for data mining and machine learning, they are usually "black box" models and often very difficult to interpret.

Cultural imports

The exchanges between genomics and other disciplines are not limited to methods and techniques, but also include cultural practices. As a discipline, protein-structure prediction pioneered concepts such as the Critical Assessment of protein Structure Prediction (CASP) competition

Fig. 4 Technical exchanges between genomics and other data science subdisciplines. The background area displays the total number of publications per year for the terms. **a** Hidden Markov model, **b** Scale-free network, **c** latent Dirichlet allocation. *Continuous lines* indicate the fraction of papers related to topics in genomics and in other disciplines

format. CASP is a community-wide effort to evaluate predictions. Every 2 years since 1994, a committee of researchers has selected a group of proteins for which hundreds of research groups around the world will (i) experimentally describe and (ii) predict in silico its structure. CASP aims to determine the state of the art in modeling protein structure from amino acid sequences [68]. After research groups submit their predictions, independent assessors compare the models with the experiments and rank methods. In the most recent instantiation of CASP, over 100 groups submitted over 50,000 models for 82 targets. The success of the CASP competition has inspired more competitions in the biological community, including genomics. DREAM Challenges, for example, have played a leading role in organizing and catalyzing data-driven competitions to evaluate the performance of predictive models in genomics. Challenge themes have included "Genome-Scale Network Inference", "Gene Expression Prediction", "Alternative Splicing", and "in vivo Transcription Factor Binding Site Prediction" [69]. DREAM Challenges was initiated in 2006, shortly before the well-known Netflix Challenge and the Kaggle platform, which were instrumental in advancing machine-learning research [70].

Technical exports
A few methods exported from genomics to other fields were initially developed to address specific biological problems. However, these methods were later generalized for a broader set of applications. A notable example of such an export is the latent Dirichlet allocation (LDA) model. Pritchard et al. [71] initially proposed this unsupervised generative model to find a group of latent processes that, in combination, can be used to infer and predict individuals' population ancestry based on single nucleotide variants. Blei, Ng, and Jordan [72] independently proposed the same model to learn the latent topics in natural language processing (NLP). Today, LDA and its countless variants have been widely adapted in, for example, text mining and political science. In fact, when we compare genomics with other topics such as text mining we observe that genomics currently accounts for a very small percentage of work related to LDA (Fig. 4c).

Genomics has also contributed to new methods of data visualization. One of the best examples is the Circos plot [73], which is related to the import above of network science. Circos was initially conceptualized as a circular representation of linear genomes. In its conception, this method displayed chromosomal translocations or large syntenic regions. As this visualization tool evolved to display more generic networks, it was also used to display highly connected datasets. In particular, the media has used Circos to display and track customer behavior, political citations, and migration patterns [73].

In genomics, networks and graphs are also being used in order to represent the human genome. For instance, researchers are attempting to represent the reference genome and its variants as a graph [74].

Another prominent idea exported from genomics is the notion of family classification based on large-scale datasets. This derives from the biological taxonomies dating back to Linnaeus, but also impacts the generation of protein and gene family databases [75, 76]. Other disciplines, for example, linguistics and neuroimaging, have also addressed similar issues by constructing semantic and brain region taxonomies [77, 78]. This concept has even made its way into pop culture; for example, Pandora initially described itself as the music genome project [79]. Another example is the art genome project [80], which maps characteristics (referred to as "genes") that connect artists, artworks, architecture, and design objects across history.

Cultural exports
Genomics has also tested and exported several cultural practices that can serve as a model for other data-rich disciplines [81]. On a fundamental level, these practices promote data openness and re-use, which are central issues to data science disciplines.

Most genomics datasets, and most prominently datasets derived from sequencing, are frequently openly accessible to the public. This practice is evidenced by the fact that most genomics journals require a public accession identifier for any dataset associated with a publication. This broad adoption of data openness is perhaps a reflection of how genomics evolved as a discipline. Genomics mainly emerged after the conclusion of HGP—a public initiative that, at its core, was dedicated to release a draft of the human genome that was not owned or patented by a company. It is also notable that the public effort was in direct competition with a private effort by Celera Genomics, which aimed to privatize and patent sections of the genome. Thus, during the development of the HGP, researchers elaborated the Bermuda principles, a set of rules that called for public releases of all data produced by HGP within 24 h of generation [82]. The adoption of the Bermuda principles had two main benefits for genomics. First, it facilitated the exchange of data between many of the dispersed researchers involved in the HGP. Second, perhaps due to the central role of the HGP, it spurred the adoption of open-data frameworks more broadly. In fact, today most large projects in genomics adopt Bermuda-like standards. For example, the 1000 Genomes [83] and the ENCODE [35] projects release their datasets openly before publication to allow other researchers to use their datasets [84]. Other subfields such as neuroscience (e.g.,

Navarro *et al. Genome Biology* (2019) 20:109

Page 8 of 11

the human connectome) were also inspired by the openness and setup of the genomics community [81].

In order to attain a broad distribution of open datasets, genomics has also adopted the usage of central, large-scale public dataset repositories. Unlike several other applied fields, genomics data are frequently hosted on free and public platforms. The early adoption of these central dataset resources, such as the Sequence Read Archive (SRA), European Nucleotide Archive (ENA), GenBank, and Protein Data Base (PDB), to host large amounts of all sorts of genetics data, including microarray and sequencing data, has allowed researchers to easily query and promote reuse datasets produced by others [85].

The second effect of these large-scale central dataset repositories, such as the National Center for Biotechnology Information (NCBI) and ENA, is the incentive for early adoption of a small set of standard data formats. This uniformity of file formats encouraged standardized and facilitated access to genomics datasets. Most computations in genomics data are hosted as FASTA/FASTQ, BED, BAM, VCF, or bigwig files, which respectively represent sequences, coordinates, alignments, variants, and coverage of DNA or amino acid sequences. Furthermore, as previously discussed, the monolithic nature of genomic sequences also contributes to the standardization of pipelines and allows researchers to quickly test, adapt, and switch to other methods using the same input format [86].

The open-data nature of many large-scale genomics projects may also have spurred the adoption of open-source software within genomics. For example, most genomics journals require public links to source codes to publish in silico results or computational methods. To evaluate the adoption of open source in genomics, we used the growth of GitHub repositories and activity (commits) over time (Fig. 5). Compared with many fields of similar scale (e.g., astronomy and ecology) genomics has a particularly large representation on GitHub and this is growing rapidly.

Data science issues with which genomics is grappling
Privacy
In closing, we consider the issues that genomics and, more broadly, data science face both now and in the future. One of the major issues related to data science is privacy. Indeed, the current privacy concerns related to email, financial transactions, and surveillance cameras are critically important to the public [87]. The potential to cross-reference large datasets (e.g., via quasi-identifiers) can make privacy leaks non-intuitive [70]. Although genomics-related privacy overlaps with data science-related privacy, the former has some unique aspects given that the genome is passed down through generations and

is fundamentally important to the public [88]. Leaking genomic information might be considered more damaging than leaking other types of information. Although we may not know everything about the genome today, we will know much more in 50 years. At that time, a person would not be able to take their or their children's variants back after they have been released or leaked [88]. Finally, genomic data are considerably larger in scale than many other bits of individual information; that is, the genome carries much more individual data than a credit card or social security number. Taken together, these issues make genomic privacy particularly problematic.

However, in order to carry out several types of genomic calculations, particularly for phenotypic associations like genome-wide association studies, researchers can get better power and a stronger signal by using larger numbers of data points (i.e., genomes). Therefore, sharing and aggregating large amounts of information can result in net benefits to the group even if the individual's privacy is slightly compromised. The Global Alliance for Genomics and Health (GA4GH) has made strides in developing technical ways to balance the concerns of individual privacy and social benefits of data sharing [89]. This group has discussed the notion of standardized consents associated with different datasets. The fields of security and privacy are undertaking projects like homomorphic encryption, where one can make certain calculations on an encrypted dataset without accessing its underlying contents [90].

Data ownership
Privacy is an aspect of a larger issue of data ownership and control. Although the individual or patient typically is thought to own their personal data, a countervailing trend in biomedical research is the idea that the researcher who generates a dataset owns it. There is a longstanding tradition among researchers who have generated large datasets to progressively analyze their data over the course of several papers, even a career, to extract interesting stories and discoveries [91]. There is also the notion that human data, particularly health data, have obvious medical and commercial value, and thus companies and nations often seek ownership and control over large datasets.

From the data miner's perspective, all information should be free and open, since such a practice would lead to the easy aggregation of a large amount of information, the best statistical power, and optimally mined results. Intuitively, aggregating larger datasets will, most frequently, give progressively better genotypes being associated to phenotypes.

Furthermore, even in an ideal scenario in which individuals consent to free access and the resulting dataset is completely open and freely shared by users, we imagine
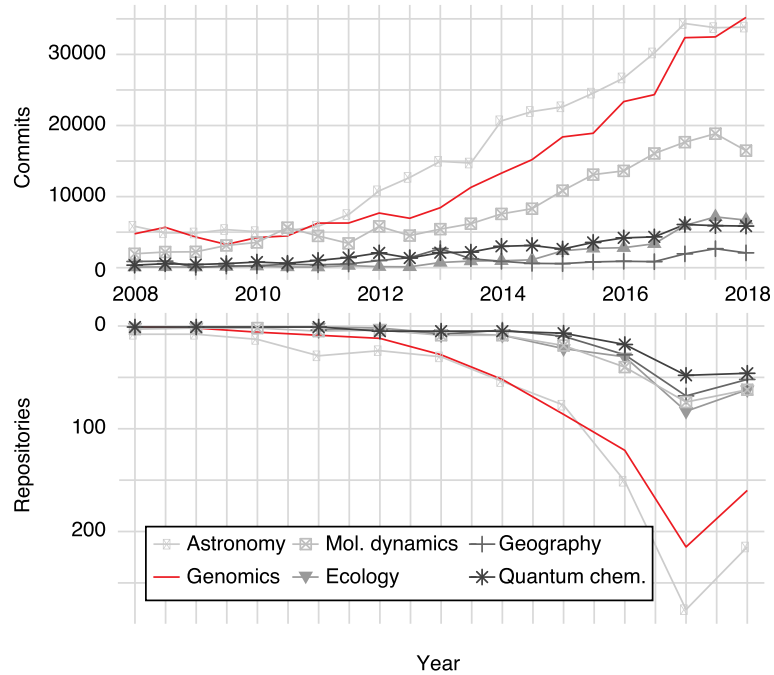
Fig. 5 Open source adoption in genomics and other data science subdisciplines. The number of GitHub commits (*upper panel*) and new GitHub repositories (*lower panel*) per year for a variety of subfields. Subfield repositories were selected by GitHub topics such as genomics, astronomy, geography, molecular dynamics (*Mol. Dynamics*), quantum chemistry (*Quantum Chem.*), and ecology

complications will arise from collection and sharing biases such as particular cohort ethnicity, diseases, and phenotypes being more open to share their genetic data. Socioeconomic status, education, and access to health-care can all possibly cause skew in datasets, which would further bias mining efforts such as machine learning algorithms and knowledge extraction. For example, ImageNet, a heavily used dataset in image classification, has nearly half of the images coming from the USA. Similarly, about 80% of genome-wide association study catalog participants are of European descent, a group which only makes up 16% of the world population [92].

For this reason, completely open data sharing will probably not be reasonable for the best future genomic association studies. One possible technical solution for sharing genomics data might be the creation of a massive private enclave. This is very different from the World Wide Web, which is fundamentally a public entity. A massive private enclave would be licensed only to certified biomedical researchers to enable data sharing and provide a way to centralize the storage and computation of large datasets for maximum efficiency. We believe this is the most practical viewpoint going forward.

On the other hand, the positive externality of data sharing behaviors will become more significant as genomic science develops and becomes more powerful in aggregating and analyzing data. We believe that, in the future, introducing data property rights, Pigouvian

subsidies, and regulation may be necessary to encourage a fair and efficient data trading and use environment. Furthermore, we imagine a future where people will grapple with complex data science issues such as sharing limited forms of data within certain contexts and pricing of data accordingly.

Lastly, data ownership is also associated with extracting profit and credit from the data. Companies and the public are realizing that the value of data does not only come from generating it per se, but also from analyzing the data in meaningful and innovative new ways. We need to recognize the appropriate approaches to not only recognize the generation of the data but also to value the analysis of large amounts of data and appropriately reward analysts as well as data generators.

Conclusion

In this piece, we have described how genomics fits into the emergence of modern data science. We have characterized data science as an umbrella term that is increasingly connecting disparate application subdisciplines. We argue that several applied subdisciplines considerably predate formal data science and, in fact, were doing large-scale data analysis before it was "cool". We explore how genomics is perhaps the most prominent biological science discipline to connect to data science. We investigate how genomics fits in with many of the other areas of data science, in terms of its data volume, velocity, and

Navarro *et al. Genome Biology* (2019) 20:109

Page 10 of 11

variety. Furthermore, we discuss how genomics may be able to leverage modeling (both physical and biological) to enhance predictive power, similar in a sense to what has been achieved in weather forecasting. Finally, we discuss how many data science ideas have been both imported to and exported from genomics. In particular, we explore how the HGP might have inspired many cultural practices that led to large-scale adoption of open-data standards.

We conclude by exploring some of the more urgent issues related to data, and how they are impacting data in genomics and other disciplines. Several of these issues do not relate to data analytics per se but are associated with the flow of data. In particular, we discuss how individual privacy concerns, more specifically data ownership, are central issues in many data-rich fields, and especially in genomics. We think grappling with several of these issues of data ownership and privacy will be central to scaling genomics to an even greater size in the future.

Abbreviations
CASP: Critical Assessment of Protein Structure Prediction; CNN: Convolutional Neural Network; ENA: European Nucleotide Archive; HGP: Human Genome Project; HMM: Hidden Markov model; LDA: Latent Dirichlet allocation

Authors' contributions
FCPN and MBG conceived and planned the study, prepared the figures, and wrote the manuscript. HM prepared the figures and wrote the manuscript. CY, SL, MG, and WM collected data and wrote the manuscript. All authors discussed the results and commented on the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA. [2]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA. [3]Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA. [4]Department of Statistics and Data Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520, USA. [5]Department of Computer Science, Stanford University, Stanford, CA 94305, USA. [6]Department of Biomedical Data Sciences, Stanford University, Stanford, CA 94305, USA.

References
1. Davenport TH, Patil DJ. Data scientist: the sexiest job of the 21st century. Harv Bus Rev. 2012;90:70–6.
2. Provost F, Fawcett T. Data science and its relationship to big data and data-driven decision making. Big Data. 2013;1:51–9.
3. Tukey JW. The future of data analysis. Ann Math Stat. 1962;33:1–67.
4. Tansley S, Tolle KM. The fourth paradigm: Microsoft Press; 2009.
5. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349:255–60.
6. Fienberg SE. A brief history of statistics in three and one-half chapters: a review essay. Stat Sci. 1992;7:208–25.
7. Robert C, Casella G. A short history of Markov chain Monte Carlo: subjective recollections from incomplete data. Stat Sci. 2011;26:102–15.
8. Lee TB, Cailliau R, Groff JF, Pollermann B. World-wide web: the information universe. Internet Res. 2013;2:52–8.
9. Kodama Y, Shumway M, Leinonen R. International nucleotide sequence database collaboration. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res. 2012;40:D54–6.
10. Hey T, Trefethen A. The data deluge: an e-science perspective. In: Berman F, Fox G, Hey T, editors. Grid computing: making the global infrastructure a reality. Chichester: Wiley-Blackwell; 2003. p. 809–24.
11. Jaschek C. Data in astronomy. Cambridge: Cambridge University Press; 1989.
12. Cox DR. Analysis of binary data. New York: Routledge; 1970.
13. Blashfield RK, Aldenderfer MS. The methods and problems of cluster analysis. In: Nesselroade JR, Cattell RB, editors. Handbook of multivariate experimental psychology. Boston: Springer; 1988. p. 447–73.
14. Belson WA. Matching and prediction on the principle of biological classification. App Stat. 1959;8:65.
15. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biol. 1943:99–115 discussion 73–97.
16. Shannon CE. An algebra for theoretical genetics. PhD thesis. Cambridge: Massachusetts Institute of Technology; 1940.
17. Kuska B. Beer, Bethesda, and biology: how "genomics" came into being. J Natl Cancer Inst. 1998;90:93.
18. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17:333–51.
19. Greenbaum D, Luscombe NM, Jansen R, Qian J, Gerstein M. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. Genome Res. 2001;11:1463–8.
20. Eisen JA. Badomics words and the power and peril of the ome-meme. Gigascience. 2012;1:6.
21. Cheng Y. Single-particle cryo-EM – how did it get here and where will it go. Science. 2018;361:876–80.
22. Althoff T, Sosič R, Hicks JL, King AC, Delp SL, Leskovec J. Large-scale physical activity data reveal worldwide activity inequality. Nature. 2017;547:336–9.
23. Wamba SF, Akter S, Edwards A, Chopin G, Gnanzou D. How "big data" can make big impact: findings from a systematic review and a longitudinal case study. Int J Prod Econ. 2015;165:234–46.
24. McAfee A, Brynjolfsson E. Big data: the management revolution. Harv Bus Rev. 2012;90:61–7.
25. White M. Digital workplaces: vision and reality. Bus Inf Rev. 2012;29:205–14.
26. NASA. https://earthdata.nasa.gov. Accessed 10 May 2019.
27. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: astronomical or genomical? PLoS Biol. 2015;13:e1002195.
28. Marx V. Biology: The big challenges of big data. Nature. 2013;498:255–60.
29. Zikopoulos P, Eaton C. IBM. Understanding big data: analytics for enterprise class hadoop and streaming data. India: McGraw-Hill; 2011.
30. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
31. Gandomi A, Haider M. 2015. Beyond the hype: big data concepts, methods, and analytics. Int J Inf. 2015;35:137–44.
32. Saunders CJ, Miller NA, Soden SE, Dinwiddie DL, Noll A, Alnadi NA, et al. Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med. 2012;4:154ra135.
33. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530:228–32.
34. Cisco Visual Networking Index: forecast and trends, 2017–2022 White Paper. 2018. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html. Accessed 10 May 2019.
35. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.
36. Campbell PJ, Getz G, Stuart JM, Korbel JO, Stein LD. ICGC/TCGA Pan-Cancer analysis of whole genomes net. Pan-cancer analysis of whole genomes. BioRxiv. 2018:1–29.
37. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010;467:1061–73.
38. Onnela J-P, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. 2016;41:1691–6.
39. Ideker T, Winslow LR, Lauffenburger DA. Bioengineering and systems biology. Ann Biomed Eng. 2006;34:1226–33.

40. Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, et al. Deep learning and process understanding for data-driven earth system science. Nature. 2019;566:195–204.

41. Artificial intelligence alone won't solve the complexity of Earth sciences [Comment]. Nature. 2019;566:153.

42. Murphy AH. The early history of probability forecasts: some extensions and clarifications. Wea Forecasting. 1998;13:5–15.

43. Bauer P, Thorpe A, Brunet G. The quiet revolution of numerical weather prediction. Nature. 2015;525:47–55.

44. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195–7.

45. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227:1435–41.

46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

47. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754–60.

48. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nature. 2012;9:357–9.

49. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.

50. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.

51. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

52. Gales M, Young S. The application of hidden Markov models in speech recognition. FNT in Signal Processing. 2007;1:195–304.

53. Gagniuc PA. Markov chains. Hoboken: John Wiley; 2017.

54. Eddy SR. Profile hidden Markov models. Bioinformatics. 1998;14:755–63.

55. Mealy GH. A method for synthesizing sequential circuits. Bell Syst Tech J. 1955;34:1045–79.

56. Ediger D, Jiang K, Riedy J, Bader DA, Corley C. Massive social network analysis: mining twitter for social good. 2010. 39th International Conference on Parallel Processing (ICPP) IEEE; p 583–593.

57. Guimera R, Mossa S, Turtschi A, Amaral LA. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci U S A. 2005;102:7794–9.

58. McGillivray P, Clarke D, Meyerson W, Zhang J, Lee D, Gu M, et al. Network analysis as a grand unifier in biomedical data science. Annu Rev Biomed Data Sci. 2018;1:153–80.

59. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402:C47–52.

60. Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796–804.

61. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302:249–55.

62. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nature. 2018;12:878.

63. Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. Bioinformatics. 2007;23:1728–36.

64. Jia C, He W. EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. Sci Rep. 2016;6:38741.

65. Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep. 2015;5:11476.

66. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33:831–8.

67. Wang D, Liu S, Warrell J, Won H, Shi X, Navarro FCP, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018;362:eaat8464.

68. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins. 1995;23:ii–v.

69. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, et al. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PLoS One. 2010;5:e9202.

70. Narayanan A, Shi E, Rubinstein BIP. Link prediction by de-anonymization: how we won the Kaggle Social Network Challenge. 2011 International Joint Conference on Neural Networks (IJCNN 2011, San Jose). IEEE; p. 1825–34.

71. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.

72. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.

73. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.

74. Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017;27:665–76.

75. Schreiber F, Patricio M, Muffato M, Pignatelli M, Bateman A. TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. 2014;42:D922–5.

76. Lam HYK, Khurana E, Fang G, Cayting P, Carriero N, Cheung K-H, et al. Pseudofam: the pseudogene families database. Nucleic Acids Res. 2009; 37:D738–43.

77. Panagiotaki E, Schneider T, Siow B, Hall MG, Lythgoe MF, Alexander DC. Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison. Neuroimage. 2012;59:2241–54.

78. Ponzetto SP, Strube M. Deriving a large-scale taxonomy from Wikipedia. Proceedings of the National Conference on Artificial Intelligence, 2007. Palo Alto: Association for the Advancement of Artificial Intelligence; 2007. p. 440–5.

79. Prockup M, Ehmann AF, Gouyon F, Schmidt EM, Kim YE. Modeling musical rhythmatscale with the music genome project. 2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA). Piscataway: IEEE; 2015. p. 1–5.

80. Artsy. www.artsy.net. Accessed 10 May 2019.

81. Choudhury S, Fishman JR, McGowan ML, Juengst ET. Big data, open science and the brain: lessons learned from genomics. Front Hum Neurosci. 2014;8:239.

82. Cook-Deegan R, Ankeny RA, Maxson Jones K. Sharing data to build a medical information commons: from Bermuda to the global alliance. Annu Rev Genomics Hum Genet. 2017;18:389–415.

83. 1000 Genomes Project Consortium, Auton A, Brooks LD, Garrison EP, Kang HM, Marchini JL, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

84. Wang D, Yan K-K, Rozowsky J, Pan E, Gerstein M. Temporal dynamics of collaborative networks in large scientific consortia. Trends Genet. 2016;32:251–3.

85. Rung J, Brazma A. Reuse of public genome-wide gene expression data. Nat Rev Genet. 2013;14:89–99.

86. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8.

87. Acquisti A, Gross R. Imagined communities: awareness, information sharing, and privacy on the Facebook. In: Danezis G, Golle P, editors. Privacy enhancing technologies. PET 2006. Lecture notes in computer science, vol 4258. Berlin: Springer; 2006. p. 36–58.

88. Greenbaum D, Sboner A, Mu XJ, Gerstein M. Genomics and privacy: implications of the new reality of closed data for the field. PLoS Comput Biol. 2011;7:e1002278.

89. Knoppers BM. International ethics harmonization and the global alliance for genomics and health. Genome Med. 2014;6:13.

90. Erlich Y, Narayanan A. Routes for breaching and protecting genetic privacy. Nat Rev Genet. 2014;15:409–21.

91. Longo DL, Drazen JM. Data sharing. N Engl J Med. 2016;374:276–7.

92. Zou J, Schiebinger L. AI can be sexist and racist – it's time to make it fair. Nature. 2018;559:324–6.

Publisher's Note