Biological data sciences in genome research

Michael C. Schatz

Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

The last 20 years have been a remarkable era for biology and medicine. One of the most significant achievements has been the sequencing of the first human genomes, which has laid the foundation for profound insights into human genetics, the intricacies of regulation and development, and the forces of evolution. Incredibly, as we look into the future over the next 20 years, we see the very real potential for sequencing more than 1 billion genomes, bringing even deeper insight into human genetics as well as the genetics of millions of other species on the planet. Realizing this great potential for medicine and biology, though, will only be achieved through the integration and development of highly scalable computational and quantitative approaches that can keep pace with the rapid improvements to biotechnology. In this perspective, I aim to chart out these future technologies, anticipate the major themes of research, and call out the challenges ahead. One of the largest shifts will be in the training used to prepare the class of 2035 for their highly interdisciplinary world.

Modern quantitative biology is in some regards no different than previous eras. At its core, it follows the same principles as ever by asking questions of a biological system, and then recording and integrating observations made under varying conditions until some conclusion can be made. Consider, for example, how the very foundation of modern genetics, Mendel's principles of inheritance, was established through an analysis of some 30,000 pea plants and in recognizing the inheritance of certain traits could be explained by a few simple mathematical rules (Mendel 1866). This accomplishment demonstrates the power of quantitative biology, especially considering that these laws were established in advance of modern molecular biology, including deciphering how genes encode for proteins or even that DNA is the modality of inheritance.

What is new in the modern era is, unlike Mendel and his successors that collected data by hand, the methods for collecting observations are now largely automated digital sensors. This includes the rise of DNA sequencing instruments, super-resolution digital microscopy, mass spectrometry, magnetic resonance imagery, or even satellite imagery used for studying biological systems with greater throughput and resolution than ever before. Furthermore, although the instruments are now providing great quantities of data, they do not by themselves do any meaningful interpretation of them. Consequently, the second major advance in the field has been the increased importance of computational and analytical techniques used to study biological data. These approaches, including large-scale multicore computing systems, advanced search and indexing algorithms, numerical optimization and modeling techniques, and many others, have primarily originated outside of biology in other quantitative disciplines. This new paradigm, often called "Biological Data Science" acknowledges that computer science, mathematics, physics, statistics, and other quantitative fields have developed advanced techniques that can be applied toward understanding biological data (https://datascience.nih.gov/).

The power of this approach comes from its ability to find relationships over very large numbers of observations, commonly stored in terabytes or petabytes of data (Marx 2013). However, given the size and complexities of the relationships, this pursuit requires an end-to-end integration of approaches, forming an analysis stack starting with data collection and continuing through computational and statistical evaluations toward higher-level biological interpretations and insights (Fig. 1). Neglecting any layer of this stack will limit progress toward the higher goals: The instruments do not interpret data, the data will be inaccessible without high performance computing systems, and abstract quantitative approaches can be misled by technical artifacts or spurious correlations without a deep understanding of the underlying biology. Consequently, many of the future advances in biology and medicine will come from the integration of biotechnologies, computational technologies, and quantitative reasoning, all designed by scientists with broad training.

With this powerful combination of rapidly advancing biotechnology and rapidly advancing data science, quantitative biology has a very bright future over the next 20 years. Many fundamental questions in genomics and biology will start to be addressed regarding the structure and function of genomes, the molecular interplay within cells and organs, and the properties of entire species and ecosystems. One likely outcome from this research within human genetics will be an extended family tree linking together much of the world's population, and with it, unprecedented power to study the forces of inheritance as well as our own origins (Ledford 2013). As we become more capable at interpreting genomes and monitoring molecular changes, we will also see profound advances in the medical community toward recognizing genetic risk factors and treating diseases based on one's personal genomic makeup (Collins and Varmus 2015). Outside of human genetics, we will see major efforts to use quantitative biology to enhance agriculture (McCouch et al. 2013), monitor the microbiome (The Human Microbiome Project Consortium 2012), or understand the brain (Insel et al. 2013), among many other projects.

In this perspective, I aim to chart out the major advances and challenges in quantitative biology in genome research over the next 20 years, including an analysis of what new biotechnologies and analytical tools are expected to develop, along with how those developments will lead to advances in biology and medicine. We end with a look into what technological and societal challenges

Corresponding author: mschatz@cshl.edu

Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.191684.115. Freely available online through the *Genome Research* Open Access option.

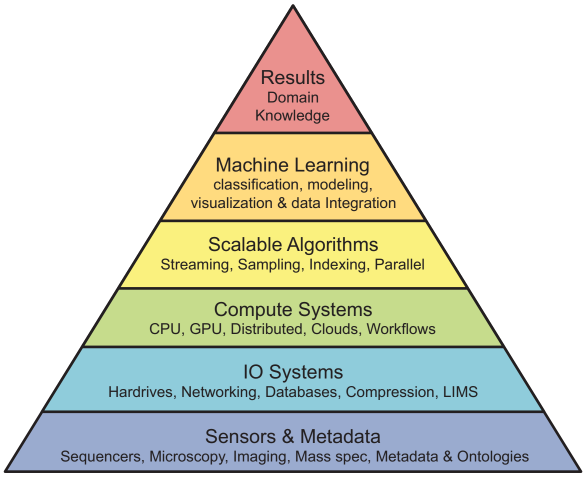
Results
Domain Knowledge
Machine Learning
classification, modeling, visualization & data Integration
Scalable Algorithms
Streaming, Sampling, Indexing, Parallel
Compute Systems
CPU, GPU, Distributed, Clouds, Workflows
IO Systems
Hardrives, Networking, Databases, Compression, LIMS
Sensors & Metadata
Sequencers, Microscopy, Imaging, Mass spec, Metadata & Ontologies

Figure 1. Data science analysis stack. Large-scale projects in quantitative biology must address a multilayer stack of approaches moving toward increasing levels of abstraction. At its base, the experiments begin with the technologies for collecting data and metadata from various biological sensors. The processing then proceeds upward through the input/output (IO) and Compute layers that can support large-scale data processing, statistical and analysis software layers that can summarize and identify trends in the data, until finally biological results can be achieved at the top, leveraging the domain knowledge of the problem.

remain ahead, especially the shift in education that is needed to support the next generation of quantitative biologists.

Advances in biotechnology

Quantitative biology is becoming an increasingly data-rich discipline, especially in the fields of genomics, systems biology, and computational neuroscience. Each of these fields has benefited tremendously from recent improvements to sensor technology to probe into the inner workings of cells with increasingly greater resolution. Interestingly, many of these advances have benefited from the core improvements to CCD technology, leading to higher resolution and more affordable digital photography. These have naturally translated into improved microscopy or cytometry but have had far-reaching benefits in other areas. For example, within DNA sequencing, improved CCD technology has enabled sequencing greater numbers and densities of molecules with fluorescently tagged nucleotides (Ansorge 2009).

A premier example of how improved biotechnology has accelerated biology has been the development of high-throughput DNA sequencing. Originally developed in the 1970s, the initial protocols for sequencing DNA were slow and laborious with hazardous reagents (Sanger and Coulson 1975). These protocols were revolutionary at the time, but could only reliably sequence a few hundred or a few thousand base pairs per week per person, with substantial labor and reagent costs involved. In the past 20 years, however, several fully automated digital instruments have become commercially available that have dramatically accelerated the pace of sequencing (Reuter et al. 2015). During this time, the worldwide capacity for sequencing DNA has doubled approximately every nine to 12 months; remarkably, the trillionfold improvements to throughput have been matched by similar reductions in costs. Today, the most powerful sequencing instrument available, the Illumina X10, can sequence the equivalent of one haploid human genome every minute (3 Gbp/min) and has capacity to sequence 18,000 whole human genomes per year to deep coverage for about $1000 each (Illumina 2015).

Furthermore, the technologies used for sequencing DNA have been cleverly repurposed for several other applications. This includes protocols for measuring the levels of mRNA transcription or translation inside of cells (RNA-seq, Ribo-seq), the presence of methylation (Methyl-seq), or the location and frequency of proteins binding to DNA (ChIP-seq), among dozens of other “omics” assays (Soon et al. 2013). Significant developments have also pushed these technologies into ever more minute samples, including techniques to measure the genomes, transcriptomes, and epigenomes of individual cells, especially to probe the heterogeneity of gametes, cancer, the brain, and other complex samples (Wigler 2012).

Consequently, genomics is now a data-rich and diverse subfield within modern quantitative biology, with studies exploring nearly every branch of the tree of life. There are sequencing instruments in more than 60 different countries on nearly every continent (Stephens et al. 2015), and the worldwide sequencing capacity currently exceeds 35 petabases a year (35 million billion bases a year), enough capacity to sequence approximately 250,000 human genomes per year (Regalado 2014). Much of that capacity is concentrated at research institutions, hospitals, and agricultural companies and is used to study the genetics of humans and other species, especially those species with medical, agricultural, or bioenergy importance. The sequencing company, Illumina, projects sequencing capacities will continue to double year over year and projects that by the end of 2017 more than 1.6 million human genomes will be sequenced (Regalado 2014). Remarkably, at that rate of growth, over the next 20 years the worldwide sequencing capacity will grow to reach more than 1 billion human genomes per year (Stephens et al. 2015).

It is most likely that human and medical genomics, along with the genomics of agriculture and energy production, will come to dominate the sequencing field as they have the largest economic incentives. In addition to widespread DNA sequencing, the various “omics” assays will become increasingly important for monitoring health and disease over time, perhaps even daily measurements to one’s own molecular profile as the technologies become cheaper and more accessible (Chen et al. 2012). This will be used, for example, to monitor diseases and recommend treatments based on changes to gene expression patterns before any macroscopic symptoms can be seen (Simon and Roychowdhury 2013). Single-cell approaches will also become increasingly important, especially for profiling the heterogeneity in tumors before and during chemotherapy (Wigler 2012). Electronic medical records and personalized activity monitors, including Fitbits and mobile phones, will be used to continuously refine and update our understanding of our health and behavior (Gottesman et al. 2013).

Outside of medical genomics, many novel applications will develop as the instrumentation becomes smaller and less expensive (Rusk 2015). One of the most important applications is the formation of a large distributed network of sequencers that can monitor for potential pathogens around the world. This sensor network will form the basis for a “digital immune system,” analogous to the worldwide weather network that can recognize changes to the composition of the viruses, microbes, and other agents in the environment (Schatz and Phillippy 2012). Biological sensors are already at high profile public sites, such as major sports arenas and transportation centers, used to monitor for pathogens passing in the air. Just as weather prediction becomes more informed and more accurate by worldwide monitoring, establishing a worldwide sequencing net would help us to monitor and contain epidemics before they can spread to larger populations.

Complementary to the technologies for measuring biological systems, new technologies for manipulating cells and synthesizing molecules will also become extremely important. Existing techniques for manipulating DNA or expression levels, such as restriction digests or RNAi (Hannon 2002), will be enhanced with new higher precision and more flexible technologies. Already, CRISPR/Cas9 systems are emerging as powerful techniques for editing genomes (Jinek et al. 2012), and optogenetics can be used for real-time activation or repression of targeted cells (Deisseroth 2011). Complementary advances in synthetic genomics are starting to yield entirely artificial genomes with genes to produce compounds of our own design (Gibson et al. 2010).

Altogether, the next 20 years holds enormous potential for sensing and manipulating cells and molecules, creating an efficient feedback cycle between reading a genome or molecular activity, modeling its function, and measuring the effect of changing it.

Computational and quantitative advances

Most immediately impacted by the massive growth to sequencing and sensor technologies will be the computational systems used for storing and transferring biological data. For more than 20 years, NCBI and its international counterparts at the EBI and DDBJ have served as the central clearinghouse for genomic data (NCBI Resource Coordinators 2015). Over the next 20 years, these resources will continue to steadily grow, although as the sequencing facilities grow from petabyte to exabyte scale, it will become less and less practical to transfer data into these archives as they exist today. Furthermore, as sequencing shifts from research purposes and into more direct medical applications, the incentive for making the data publicly available in a centralized archive will be reduced or perhaps even legally restricted.

In its place, we will see the rise of federated approaches for exchanging biological data, especially computing centers dedicated to large sequencing facilities. Already this trend is beginning, and the NCBI Sequence Read Archive (SRA) currently only stores approximately one-tenth of the worldwide sequence production, ~3.8 Pbp of the >35 Pbp sequenced so far (http://www.ncbi.nlm.nih.gov/Traces/sra/). Fortunately, the rest of the data are not completely lost, and we are beginning to see the emergence of new exchange systems outside of traditional archives. These systems often consolidate regional and/or topical interests inside a dedicated cloud-based portal, such as CGHub (Wilks et al. 2014) or ICGC (The International Cancer Genome 2010) for consolidating cancer genomic data, or the recently launched BGI-Cloud to provide access to the great resources available there (http://bgiamericas.com/data-analysis/bgi-cloud/). Illumina BaseSpace (https://basespace.illumina.com), DNAnexus (https://www.dnanexus.com/), Google Genomics (https://cloud.google.com/genomics/), and other commercial vendors are also emerging to help manage the deluge of data using commercial cloud platforms and are likely to play an increasingly important role in genomics in the future.

The major technical reason this model will become more widespread is that at large scales, it is overwhelmingly more efficient to upload code segments, measured in kilobytes to megabytes, rather than to download entire large collections, measured in petabytes or beyond (Schatz et al. 2010). Economies of scale are also made possible through consolidating purchasing and time-sharing of equipment, especially for the many thousands of cores and petabytes of storage that will need to be purchased and maintained. It also allows for the different cloud resources to specialize their services for different biological systems or political requirements; despite using common sequencing technologies, the tools and data sets needed for studying cancer are very different than those needed for studying crop development, as are the legal requirements in some countries compared to others.

Although this model of federated cloud-based data warehousing offers many advantages, it also presents significant new challenges. Foremost, genomic data are most useful when they can be aggregated and combined in very large numbers. Otherwise, subtle genetic signals may be lost if the data are partitioned or if the measurements are recorded with incompatible formats. Therefore, the resources will have to establish common application programming interfaces (APIs) to enable remote access to their data, along with strong encryption and authentication safeguards to protect from theft or abuse. Today, the Global Alliance for Genomics and Health is one of the leading efforts to define such standards for genomics and other personal biomedical data. Although even the most basic of federated tasks, so-called “Beacons” that identify if a resource has any individual with a particular mutation, are proving to be difficult to execute mostly for nontechnical reasons (http://ga4gh.org/#/beacon). These include the challenges of agreeing on an API for all institutions to use, privacy concerns over releasing these data, and a general reluctance to share unpublished data.

Within each data warehouse there will be major systems engineering challenges to consider: Instead of a program crashing on one server, a code failure could disrupt thousands of cores at once leading to years or centuries of wasted computing effort. Most of the currently available genomics applications are not designed for this level of parallelism, needing new higher-level and easy-to-use workflow systems to orchestrate the scheduling and management of resources (Afgan et al. 2011), as well as improved software engineering practices to build more robust software (Wilson et al. 2014). Unlike current research practices that often store all data in a uniformly accessible way, the massive data warehouses must be built with tiered storage systems to prioritize access to the data within while keeping storage costs manageable (Haussler et al. 2012). Summary statistics, variant lists, expression profiles, and other highly processed data can be kept in active memory, while raw reads and measurements can be archived to slower and less expensive medium. Specialized data compression algorithms (Hsi-Yang Fritz et al. 2011), while extremely important to make the best use of every available byte, are unlikely to completely solve the storage needs. In an effort to control costs and complexity, there will be growing urgency to delete data as soon as possible. This will mark a radical shift in how data are currently viewed, and requires careful consideration of the “preciousness” of a sample: A research project exploring a rare cancer or ancient DNA sample will likely choose to archive everything, whereas studies of more abundant samples may elect to only store the processed results.

The applications and algorithms built for these warehouses will be focused on integrating the analysis over very large populations. Certain topics that are widely studied today, such as short-read mapping or de novo assembly of individual genomes, will fade away as the algorithms mature and new sequencing technologies producing longer reads take over, leading to extremely high quality genomes (Berlin et al. 2015). In its place will be the need for algorithms and systems for studying the genomes, transcriptomes, and epigenomes of millions of individuals at once, especially systems that can do so within a graph of sequence variability (Church et al. 2015). As sequencing becomes more widely used for real-time applications, such as a real-time readout of the transcription levels in a blood sample, the interpretation can be done on the fly without the raw data recorded at all. Already several "streaming" methods for doing this are becoming available for inferring expression levels from RNA-seq reads that are just as effective as their more traditional counterparts (Patro et al. 2014), and other omics data types inferring quantification from sequences are likely to follow similar developments.

Altogether, over the next 20 years, we will see the development of major institutional and governmental data warehousing for millions to billions of genomes and other biomedical data. It will require maturing of the algorithms and formats used today into more scalable and interoperable systems as well as designing new systems to solve problems that are not even considered today. A growing need will be for streaming algorithms and other approaches that can make inferences over diverse data types as soon as they are produced, so that data storage needs will be as limited as possible.

Insights of genome research

If one of the greatest accomplishments from the last 20 years has been sequencing the first human genomes, one of the greatest pursuits for the next 20 years will be trying to understand what it all means. Although we have been quite successful deciphering the genetic code of how genes code for proteins, the grander challenges of understanding how genome sequences ultimately code for traits and relate to disease has remained largely a mystery. For example, the current best estimates suggest several thousands of variants are related to human height but only a few hundred candidates are currently known. Even less is known of the genetic factors involved in major diseases such as coronary artery disease (Ozaki and Tanaka 2015) or cerebrovascular disease (Markus 2010) despite strong evidence for a genetic component. Determining the molecular basis, be it genetic, epigenetic, or even metagenomic, for these and other human conditions is one of the central pursuits of quantitative biology and one of the great unanswered questions in all of science.

The rationale for sequencing one million or one billion genomes using this powerful stack of technologies is not to produce one million or one billion separate lists of variants. Rather, the hope is that the whole will be greater than the sum of its parts, and something new will emerge that sheds light on to what those sequences mean. Like Mendel, the hope is that bringing together these data will lead to discovering the underlying patterns and rules of genetics and organism biology: who has which variants, how they are inherited or evolve, and how variants are related to diseases and other traits. The insights we will gain will come in phases depending on several factors, including how many samples are available, what other measurements are available, and crucially, how complex the traits are, ranging from simple monogenic conditions to highly complex polygenic traits or those with subtle environmental components (McCarthy et al. 2008). In some cases, the relationships will be very clear, allowing us to leap almost immediately from genetic variants to their associated traits. More often than not, however, this analysis will require a series of stepping stones to connect how a variant can alter the sequence or expression of a single gene, which in turn influences a pathway of interconnected genes, which then influences the overall cell or individual, all within a particular environment (Wang et al. 2010a).

Considering the multitude of environmental factors, cell-type specificities, and diversity of genetic backgrounds, understanding what each base of a genome means may be a never-ending pursuit. Nevertheless, over the next 20 years, great strides will be made in raising our level of understanding. Today, we are most successful at interpreting major alterations within gene sequences (Wang et al. 2010b), but we are substantially less informed about interpreting the relative importance and mechanisms for noncoding mutations (Ward and Kellis 2012). Over the next 20 years, however, our power for doing so will greatly improve building on the pioneering work of ENCODE (The ENCODE Project Consortium 2012), the Roadmap Epigenomics Project (Romanoski et al. 2015), and similar projects that are starting to provide detailed annotations as to the roles and evolution of sequences all throughout the genome. The community is also currently largely focused on single nucleotide and other small variants, but as the sequencing technologies improve, it is likely that the widespread nature and significance of structural variations will become even more pronounced (Chaisson et al. 2015). Indeed, although a typical person has more SNPs than any other class of mutations, the total number of bases that are mutated, and hence the largest aspect of genetic diversity in a person, is mostly due to copy number and other structural variations (Sebat et al. 2004; Zarrei et al. 2015). Finally, the widespread deployment of diverse sensors will allow us to more carefully consider environmental factors, such as Fitbit-type data of location and activities integrated with detailed molecular and microbiome profiles.

Although extremely powerful, care must be taken when using these data and techniques to avoid unfounded generalizations or mistaking spurious correlations as causal relationships. Quantitative biology needs integration of quantitative expertise with domain expertise to design the proper experiments, to fight technical artifacts, and to recognize relationships outside of the expected. As recently highlighted, although much scrutiny has been given to the importance of *P*-values, these evaluations come at the top of a complex analysis stack that requires careful scrutiny at every stage (Leek and Peng 2015). Whenever the result is most surprising or unexpected, that is exactly when one should be most critical of the methodology. Caution is also needed to consider the social changes and privacy risks that will come from building these massive data warehouses of personal information, especially when sequencing a genome becomes as cheap and easy as taking someone's photograph. It is surprisingly easy to identify someone from their genome today and will become even easier and more accurate as these massive collections are established (Erlich and Narayanan 2014). As genome sequences are connected to other personal data, the threat of genetic discrimination and abuse becomes real and everlasting.

Nevertheless, the potential benefits for building and mining these massive collections are overall much larger than the potential risks. One of the major lessons learned from machine learning has been, when coupled with the right algorithms and systems, increasing the amount of data often leads to greatly increased performance and interpretation, including chess programs (Campbell et al. 2002), translation systems (Koehn et al. 2007), and even virtual Jeopardy contestants (Ferrucci 2012) that can outperform any human. Within genomics and quantitative biology, we hope and anticipate this will prove true as well, leading to many breakthroughs of profound significance over the next 20 years.

Recommendations for the class of 2035

Children born this summer will grow up in a drastically different world than those from 20 years ago. They will grow up with unprecedented access to information, quantitative techniques, and biotechnologies that will be used to manipulate biological systems in currently unimaginable detail. Although the foundations of biology will continue to be observation, experimentation, and interpretation, the technologies and approaches used will become ever more powerful and quantitative. More so than ever, we need to revise the curriculum to integrate computational and quantitative analysis as early as possible into their training so they are ready for the world ahead (Schatz 2012).

My recommendation to the next generation of scientists is to embrace the integration of fields that is forming modern biology. To be competitive, you will need to establish a broad interdisciplinary foundation of math and science as well as strong communication skills. One of the most important skills you can develop early is computer programming. Much like learning to speak a new language is often easier the younger you begin, learning to "speak" to a computer seems to follow a similar path. Although sequencing technologies and other instrumentation will come and go over the next 20 years, biology will only continue to grow its dependency on computational analysis as data sets grow ever larger. Mathematics may be the language of nature, but code is the language of data.

Students should be prepared for the technologies and systems for coding and data analysis to change over the next two decades, just as C++ and Python have largely displaced Machine Language and Fortran in the previous decades. However, as has been broadly documented in software engineering, while higher-level programming languages, expert systems, and other advanced human-computing interfaces do simplify and accelerate software development, they do not solve the underlying intrinsic complexities of developing or deciphering large abstract systems (Brooks 1995). As Fred Brooks so elegantly described it, there is inherently no "silver bullet" that will make building software easy. All signs indicate developing software and performing analysis for quantitative biology will be at least as difficult, as the number of possible states and interactions that are potentially relevant to a biological system are astoundingly large, and we cannot expect any semblance of a designed architecture as we would in an engineered system.

Most of all, I encourage you to follow your sense of curiosity. The most brilliant advances in science all start with asking the right questions, especially when encountering something unexpected or unexplained. To do so, I also encourage you to learn the experimental side of biology as early as possible so that you can pursue your own experiments as you notice something interesting about yourself or the world around you. Many of the most profound advances have occurred when a new biotechnology or analysis technique have been applied to a deep biological question. When you are first to generate a novel data type you can also be first to unravel a mystery of how the world operates. Finally, always remember to keep focused on the most important problems that you can hope to address.

Acknowledgments

I dedicate this article to my wife, Kelly, and our incredible son, James, born July 29, 2015. I thank all of the participants of the Biological Data Sciences conference, co-organized with Anne Carpenter and Matt Wood, held last fall at Cold Spring Harbor Laboratory that was inspirational to this work. Finally, I also thank all of my great colleagues, friends, and family that have contributed to the ideas presented here. The project was supported, in part, by National Science Foundation award DBI-1350041.

References

Afgan E, Baker D, Coraor N, Goto H, Paul IM, Makova KD, Nekrutenko A, Taylor J. 2011. Harnessing cloud computing with Galaxy Cloud. *Nat Biotechnol* **29:** 972–974.

Ansorge WJ. 2009. Next-generation DNA sequencing techniques. *N Biotechnol* **25:** 195–203.

Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* **33:** 623–630.

Brooks FP. 1995. *The mythical man-month: essays on software engineering*. Addison-Wesley Publishing Co., Reading, MA.

Campbell M, Hoane AJ, Hsu FH. 2002. Deep blue. *Artif Intell* **134:** 57–83.

Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al. 2015. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517:** 608–611.

Chen R, Mias GI, Li-Pook-Than J, Jiang L, Lam HY, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, et al. 2012. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148:** 1293–1307.

Church DM, Schneider VA, Steinberg KM, Schatz MC, Quinlan AR, Chin CS, Kitts PA, Aken B, Marth GT, Hoffman MM, et al. 2015. Extending reference assembly models. *Genome Biol* **16:** 13.

Collins FS, Varmus H. 2015. A new initiative on precision medicine. *N Engl J Med* **372:** 793–795.

Deisseroth K. 2011. Optogenetics. *Nat Methods* **8:** 26–29.

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489:** 57–74.

Erlich Y, Narayanan A. 2014. Routes for breaching and protecting genetic privacy. *Nat Rev Genet* **15:** 409–421.

Ferrucci DA. 2012. Introduction to "This is Watson". *IBM J Res Dev* **56:** 1:1–1:15.

Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, et al. 2010. Creation of a bacterial cell controlled by a chemically synthesized genome. *Science* **329:** 52–56.

Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, Sanderson SC, Kannry J, Zinberg R, Basford MA, et al. 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med* **15:** 761–771.

Hannon GJ. 2002. RNA interference. *Nature* **418:** 244–251.

Haussler D, Patterson DA, Diekhans M, Fox A, Jordan M, Joseph AD, Ma S, Paten B, Shenker S, Sittler T, et al. 2012. *A million cancer genome warehouse*. EECS Department, University of California, Berkeley, CA.

Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E. 2011. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res* **21:** 734–740.

The Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. *Nature* **486:** 207–214.

Illumina. 2015. *HiSeq X series of sequencing systems*. Vol. 2015. Illumina, Inc., San Diego.

Insel TR, Landis SC, Collins FS. 2013. Research priorities. The NIH BRAIN Initiative. *Science* **340:** 687–688.

The International Cancer Genome Consortium. 2010. International network of cancer genome projects. *Nature* **464:** 993–998.

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

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, et al. 2007. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 177–180. Association for Computational Linguistics, Prague, Czech Republic.

Ledford H. 2013. Genome hacker uncovers largest-ever family tree. *Nature* doi: 10.1038/nature.2013.14037.

Leek JT, Peng RD. 2015. Statistics: *P* values are just the tip of the iceberg. *Nature* **520:** 612.

Markus HS. 2010. Unravelling the genetics of ischaemic stroke. *PLoS Med* **7:** e1000225.

Marx V. 2013. Biology: the big challenges of big data. *Nature* **498:** 255–260.

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9:** 356–369.

McCouch S, Baute GJ, Bradeen J, Bramel P, Bretting PK, Buckler E, Burke JM, Charest D, Cloutier S, Cole G, et al. 2013. Agriculture: feeding the future. *Nature* **499:** 23–24.

Mendel G. 1866. Versuche über Plflanzen-hybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd IV für das Jahr, 1865 Abhandlungen*, pp. 3–47. [For the English translation, see Druery CT, Bateson W. 1901. Experiments in plant hybridization. *J Royal Hort Soc* **26:** 1–32.]

NCBI Resource Coordinators. 2015. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **43:** D6–D17.

Ozaki K, Tanaka T. 2015. Molecular genetics of coronary artery disease. *J Hum Genet* doi: 10.1038/jhg.2015.70.

Patro R, Mount SM, Kingsford C. 2014. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol* **32:** 462–464.

Regalado A. 2014. EmTech: Illumina says 228,000 human genomes will be sequenced this year. In *MIT technology review* (ed. Pontin J), Vol. 2015. MIT, Cambridge, MA.

Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. *Mol Cell* **58:** 586–597.

Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. 2015. Epigenomics: roadmap for regulation. *Nature* **518:** 314–316.

Rusk N. 2015. MinION takes center stage. *Nat Methods* **12:** 12–13.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94:** 441–448.

Schatz MC. 2012. Computational thinking in the era of big data biology. *Genome Biol* **13:** 177.

Schatz MC, Phillippy AM. 2012. The rise of a digital immune system. *Gigascience* **1:** 4.

Schatz MC, Langmead B, Salzberg SL. 2010. Cloud computing and the DNA data race. *Nat Biotechnol* **28:** 691–693.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Simon R, Roychowdhury S. 2013. Implementing personalized cancer genomics in clinical trials. *Nat Rev Drug Discov* **12:** 358–369.

Soon WW, Hariharan M, Snyder MP. 2013. High-throughput sequencing for biology and medicine. *Mol Syst Biol* **9:** 640.

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. 2015. Big data: astronomical or genomical? *PLoS Biol* **13:** e1002195.

Wang K, Li M, Hakonarson H. 2010a. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* **11:** 843–854.

Wang K, Li M, Hakonarson H. 2010b. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38:** e164.

Ward LD, Kellis M. 2012. Interpreting noncoding genetic variation in complex traits and human disease. *Nat Biotechnol* **30:** 1095–1106.

Wigler M. 2012. Broad applications of single-cell nucleic acid analysis in biomedical research. *Genome Med* **4:** 79.

Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, Murphy D, Pierce H, Black J, Nelson D, et al. 2014. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database (Oxford)* **2014**. doi: 10.1093/database/bau093.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SH, Huff KD, Mitchell IM, Plumbley MD, et al. 2014. Best practices for scientific computing. *PLoS Biol* **12:** e1001745.

Zarrei M, MacDonald JR, Merico D, Scherer SW. 2015. A copy number variation map of the human genome. *Nat Rev Genet* **16:** 172–183.

Biological data sciences in genome research

Michael C. Schatz

Genome Res. 2015 25: 1417-1422
Access the most recent version at doi:10.1101/gr.191684.115