Quantitative Biology 2022, 10(1): 6–16
https://doi.org/10.15302/J-QB-021-0286

FEATURE

Looking back at the first twenty years of genomics

John Quackenbush[1,2,3,*]

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[2] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02215, USA
[3] Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02115, USA
* Correspondence: johnq@hsph.harvard.edu

Received October 21, 2021; Revised November 2, 2021; Accepted November 2, 2021

Earlier this year, we celebrated the twentieth anniversary of the publication of the first draft human genome sequence — or more correctly, the nearly simultaneous publication of draft versions of the public human genome sequence and the private (Celera Genomics, Inc.) human genome sequence in *Nature* [1] and *Science* [2], respectively. Like many of the achievements of the Human Genome Project (HGP), the accomplishment was part scientific milestone, and part political compromise — in this case a "co-victory" brokered in hopes of ending the increasingly vitriolic exchanges in the press between the public and private projects[①]. Not surprisingly, the publications generated their own controversies, with the groups arguing over which version of the genome sequence was "better." In some sense, the argument was ridiculous — Celera had its data as well as all the public data, which gave them a tremendous advantage in terms of assembling the sequences and resolving ambiguities. However, the public genome was freely available, could be downloaded and reanalyzed, and was obviously going to evolve and improve over time as more data became available.

The concept of a "draft" genome was an odd one and it was not at all clear what would constitute a finished genome or what the timeline was. There were obviously gaps in the sequence to fill, and genes and other features to be found and annotated, but most of the genome seemed to be there. However, when the completed mouse genome was published in 2002 without a draft [3], there was pressure to finish the human genome, resulting in its announced completion in 2003 [4] and a single publication from the public HGP in 2004 [5] — despite the fact that there were chromosomes continued to be "finished" and published over the ensuing years. Again, some of this was political compromise. The public genome project had divided chromosomes among various sequencing groups all of whom wanted their own signature paper; the largest chromosome, Chromosome 1 was not announced complete until 2006 [6]! And even in the completed genome and its complete chromosomes, there were still lots of holes and gaps — so much so that a completed genome was published in bioRxiv in 2021 [7]. More surprisingly, the sex chromosomes, X and Y, had been more or less neglected, with X only being announced completed in 2020 [8] and the Y chromosome still lost somewhere in the land of myths and legends [9].

It is also somewhat astonishing that many questions central to sequencing the genome remain unanswered. For example, there is still no clear consensus as to how

This article is dedicated to the feature in 20th anniversary of Human Genome (Eds. Michael Q. Zhang and Xuegong Zhang).

[①] Ari Patrinos of the US Department of Energy, a visionary who also played a role in conceiving of the genome project itself, largely deserves credit for getting the Francis Collins and Craig Venter to declare mutual victory. The publications were preceded by June 26, 2000, ceremony held at the White House (with satellite ceremonies held in other countries) in which President Bill Clinton, J. Craig Venter of Celera Genomics, and Francis Collins of the National Institutes of Health came together to announce draft genome's completion. The seven-month delay between announcement and publication is a sign of how much was left to do in mid-2000 and echoes the multiple announcements of the genome being finished over the coming years.

© The author(s) 2022. Published by Higher Education Press

many genes there are in the human genome; the major data resources provide different answers, sometimes even from different groups within their organizations. There isn't even a clear consensus of what a gene is. We now have many fully assembled genomes, all of which differ to some extent in their sequence, structure, and gene content, leading us without a definitive reference genome[②] . And, of course, there are other aspects of the genome that are still unexplored. But fretting over what remains unanswered misses the point. In the early days of the HGP, we talked about mapping the genome, and whatever version we have is, at its core, a map — imperfect, imprecise, missing features — but a beautiful creation that is useful for navigation and discovery.

Ultimately, the sequencing the genome was a triumph of science and technology that ushered in today's biotech revolution and allowed us to embark on a journey of unparalleled biological discovery. The twentieth anniversary of the creation of what remains an invaluable resource provides an opportunity for us to look back at the important lessons that we learned both in the process of sequencing the genome and over the two decades as we've struggled with the to realize the promise of a nearly complete catalog of human genes. And in my opinion, far more than the sequence, the changes in how we do science that were driven by the rise of genomics are in many ways far more significant than the production of the sequence itself. In this retrospective, I will explore four unexpected lessons that emerged from the genome project and our ongoing analysis of the genome that have had a profound influence on how modern biomedical research is conducted.

THE PRIMACY OF TECHNOLOGY AND DATA

When the Human Genome Project was launched, everyone recognized that the existing technologies were not sufficient to sequence the genome. The success of the project ultimately relied on improvements in automated Sanger sequencing technology to generate the data, and the development of new computational techniques to assemble the final sequence and to identify and annotate the genes encoded within. And in that sense, the success of the project was fundamentally driven by technology and data.

However, to truly understand what the genes are and what they do — to link genotype to phenotype — it is obvious that you need more than one genome. After all, it is hard to understand what makes one genome unique if we don't have others to which to compare it. But we also had to recognize that no matter how many genomes we sequenced, sequence alone simply could never be sufficient. To begin to understand the genome, we needed to collect genomic data on large populations on which we have extensive data about the state of health and physical traits of those being assayed[③] . But, fundamentally, how many genomes you can sequence is limited by how much sequencing a genome costs[④] .

The cost of sequencing the first genome is often estimated to be almost anything from $300,000,000 to $3,000,000,000, but it is almost impossible to actually calculate what the true cost was. There were many technologies that had to be developed, sequencing and mapping strategies that were developed and abandoned, and inefficiencies and redundancies that were the result of doing something that had never been done before[⑤] . However, shortly after the draft genome was published, it was estimated that the cost of sequencing the next genome would be about $100,000,000. Over the next few years, that cost fell steadily, dropping by about a factor of two every eighteen to twenty-four months[⑥] . But the cost of sequencing a genome was still far above what could conceivably support any sort of population-based studies.

Fortunately, in 2007, when the cost to sequence a genome had fallen to about $10,000,000, a new generation of sequencing instruments were introduced that dramatically increased the rate at which data could be generated. The cost of sequencing also started to plummet, dropping by about one-third every three months. In 2009 my colleagues and I sequenced an ovarian cancer genome, the project took about two months. We almost immediately started a second ovarian cancer genome, and that took us only two weeks. Although the pace of data generation and the costs have since leveled off, today one can sequence a genome essentially overnight for less than $1,000 (Fig. 1).

As a result, genome sequencing has become a commodity that is incorporated into a growing number of studies. Even more exciting projects, such as the UK

[②]I fully recognize that the major genome databases now have a unified reference genome, but the truth is that your genome and mine likely differ to an extent in gene content and certainly differ in sequence and there is no guarantee that the reference genome doesn't encode potentially deleterious variants.

[③]As I have said many times, the next large cohort study really needs to be everyone.

[④]The most important -omic science is econ-omics.

[⑤]It is also difficult to know how to reconcile the sequencing costs of the public human genome project and what was done at Celera, in part because Celera was launched with money from the company selling sequencers to everyone.

[⑥]People often compare this trend to Moore's law and argue that the drop is consistent with the falling cost of analyzing the data, which could account for some, but not all of the cost.

© The author(s) 2022. Published by Higher Education Press

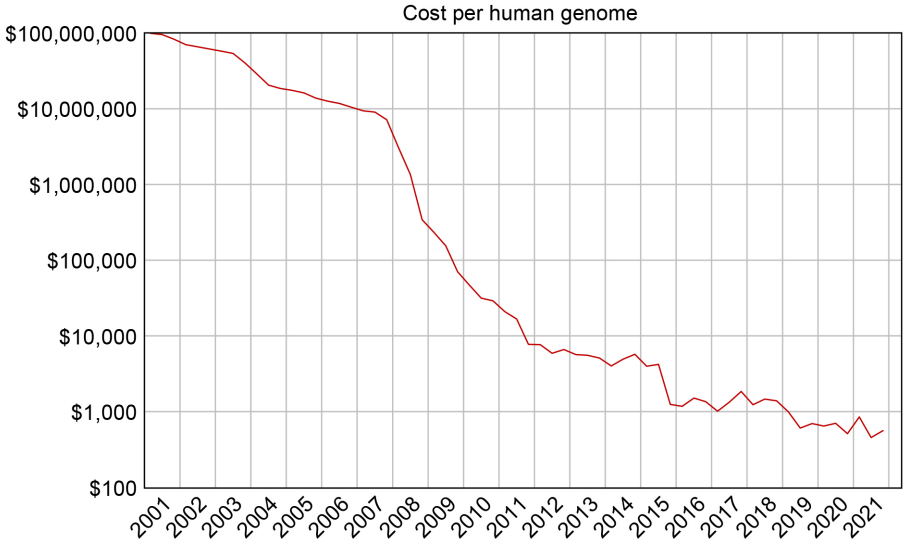

Figure 1. The estimated cost of sequencing a human genome has fallen dramatically since the first genome was sequenced in 2001. Initially, the cost was dropping by about a factor of two every eighteen months, but the introduction of new "next generation" sequencing technologies in 2007 caused costs to plummet until they leveled off at about $1000 per genome in 2016. As costs fell, so did the time required to sequence, assemble, and annotate a new genome, resulting in the dramatic increase in the number of genomes available today. Adapted from the website of National Human Genome Research Insititute in NIH.

Biobank, the Million Veterans Project, TOPMed, and All of US, have given us nearly 1,000,000 fully sequenced human genomes, and genome-wide association studies (GWASes) now regularly genotype as many as 10,000,000 single nucleotide polymorphisms (SNPs) in populations of 1,000,000 or more with GWAS populations projected to grow to 5,000,000 for many applications [10]. Handling these data requires increasingly complex data analysis methods that build on advances in data management methods, ever more powerful computing technologies, and increasingly sophisticated analytical methods. Not surprisingly, analyzing the data that we have often reveals the need for even more data — generally meaning that we need to sample larger, well-annotated populations.

As scientists, we all understand the importance of data. Data are the raw materials out of which we create our models. They are the primary resource we use to falsify or validate our models. And they are then the material we use to build new, better models. The wonderful thing about genomics is that technological advances have allowed us to begin collecting data of a scope and scale necessary to address fundamental questions that would otherwise be out of our grasp. As a result, we've seen the evolution of health and biomedical research into what is increasingly a technology-driven data and information science in which those best positioned to make progress are those with the ability to collect, manage, and analyze large, complex data.

THE COMPLEXITY OF BIOLOGICAL SYSTEMS

One of the greatest surprises from sequencing the draft human genome was just how few (protein-coding) genes were encoded within. The genomes of free-living bacteria typically encode a few thousand genes. Common brewer's yeast, *Saccharomyces cerevisiae*, has about 6,600 genes. The 1 mm long nematode, *Caenorhabditis elegans*, has nearly 20,000 genes. And, despite the fact that the number is still disputed [11], a good estimate is that the human genome encodes about 25,000 genes — far fewer than many plants. Nevertheless, when the draft genome was published, many people claimed that having a nearly complete catalog of all human genes would allow us to rapidly discover the genetic roots of most human disease and identify the drivers of many of our traits.

As comparative analyses of human genome sequence data identified approximately 6,000,000 common[⑦] single nucleotide polymorphisms, many scientists began to believe that common traits, including genetic diseases, could be associated with common variants through genome-wide association studies (GWAS), which looked for variants that were significantly overrepresented in disease populations compared to controls. The development of low-cost, high-throughput SNP genotyping arrays, which provided reasonably

[⑦]Meaning that they appear in more than 5% of the population.

© The author(s) 2022. Published by Higher Education Press

comprehensive coverage of common variants in the human genome and made GWAS and quantitative trait locus (QTL) analyses economically feasible. And these studies have been effective. Since the first GWAS study was published in 2005, more than 50,000 associations have been reported between SNPs and complex phenotypes, and such studies have helped identify disease-causing genes [12].

Despite these successes, the odds ratios and effect sizes for the identified genetic determinants have been surprisingly low [12], the percentage of phenotypic variation explained by GWAS signals has generally been modest, and most heritability for complex traits remains unexplained [13]. For example a comprehensive meta-analysis of GWAS data found that 697 genetic variants could explain about 20% of human height, but nearly 2000 were necessary to raise the explanatory power to 21%, and 9500 to reach 29% explanatory power [14]. In examining body mass index (BMI), this same group found that 97 SNPs could explain 2.7% of BMI, but they estimated that all common variants would account for a bit more than 20% of the trait. Rare genetic variants, occurring spontaneously and appearing in limited family groups or individuals, have been suggested as possible drivers of disease, but these too generally been found to have little explanatory power [15].

Expression quantitative trait locus (eQTL) analysis treats the expression of each gene as a quantitative trait and tests each SNP position in the genome to search for linear relationships between a given variant (0, 1, or 2 copies) and the level of expression of each gene. Although there have been tens of thousands of variants found to influence expression, few eQTLs have been directly linked in a causal way to disease or phenotype [16]. And further, when looking at gene or protein expression levels as biomarkers to distinguish between biological or disease subtypes, many fail to validate between studies, have modest predictive power, and do not fully explain the biological basis for the phenotypic subgroups with which they are associated.

Instead, what these failures indicate is that biological states are defined by complex interacting networks of cellular elements that work together to alter cellular processes and, ultimately, phenotypic states. Indeed, if one represents eQTL associations using bipartite networks that link "regulatory" SNPs to expressed genes, one finds that these networks are organized into highly modular "communities" in which the genes often represent coherent biological functions. Surprisingly, the SNPs found through disease studies do not map to the global "hubs" in the network, but instead appear as local hubs, or "core SNPs," in their functional communities. This network analysis also identified the fact that core SNPs have different regulatory potential and epigenetic states than other SNPs in the network, and that tissue-specific functional communities appear in the networks and are associated with SNPs that appear in tissue-specific open chromatin states [17,18].

Most importantly, eQTL networks help us understand that it is not individual SNPs that determine phenotypes, but families of variants that work in a nonlinear fashion to alter biological function. This concept is similar to the "omnigenic" that was later proposed and that suggests that the entirety of an individual's genetic background contributes to subtly alter regulatory pathways and gene expression in individual cells, ultimately contributing to phenotypes [19].

Central to these models is the idea that gene expression (and downstream protein expression) is essential for determining phenotype. Indeed, methods that use gene-gene correlation measures to construct expression networks have proven their value in identifying co-expressed modules of genes representing functional groups that differ between phenotypes [20]. However, correlation in gene expression, while consistent with co-regulation, is not necessarily indicative of it. Methods such as PANDA explicitly model the gene regulatory processes as one involving interactions between transcription factors and their targets. PANDA takes as input a "prior" regulatory network constructed by mapping transcription factor motifs to the genome. PANDA uses message passing to optimize the structure of this seed network by searching for consistency between its structure, condition-specific gene expression data, and protein-protein interactions, thus accounting for co-regulation and the formation of regulatory complexes. The outputs from PANDA are condition-specific regulatory networks that can be compared to understand condition-specific differences in regulatory processes. PANDA has been extended to include additional regulatory factors such as miRNAs [21] and epigenetic factors [22], and a linear interpolation method, LIONESS, allows gene regulatory network models to be inferred for each individual sample analyzed in a study. Gene regulatory network modeling has been used to identify potential drug targets in ovarian cancer subtypes [23], identify regulatory changes as tissues are converted to cell lines [24], find network structures defining tissue-specific functions in thirty-eight different tissues [25], and comparing regulatory network drivers of sexual differences in twenty-nine tissues [26] and in colon cancer [27]. Not surprisingly, these complex networks have found biologically relevant changes in regulatory processes that distinguish phenotypes even when there are few significant differences in gene expression [28]. Most importantly, the recognition that it is networks that drive

© The author(s) 2022. Published by Higher Education Press

even the expression we see has opened up new ways of analyzing and modeling biological processes, provided ways of estimating and comparing regulatory networks between conditions, and even the recognition that differences in individual networks may provide more informative biomarkers than the -omic features from which they are derived (Fig. 2).

Of course, none of this would have been possible without the Human Genome Project. The genes themselves, and the regulatory elements within the genome, are really the raw material upon which such models are built. The beauty of these complex networks we are discovering is that not only do they help explain the complexity of human phenotypes given the relatively small number of genes, but these complex adaptive networks provide a means of understanding the stability of cellular processes to perturbation and the orchestrated changes in regulation through both evolutionary and developmental time.

THE IMPORTANCE OF OPEN SCIENCE

When the sequencing of the human genome began, the project was far too large and complex for any single academic institution to tackle. Even though the National Institutes of Health and other agencies around the world funded the establishment of large-scale sequencing centers, the project was clearly going to require the cooperative efforts of many groups working together. The public genome project was largely built around a "divide and conquer" approach in which large fragments of DNA (Bacterial Artificial Chromosomes or "BACs") were mapped to the genome, a minimally overlapping set was selected, and these were then subcloned and sequenced. Because of this, the project evolved to allot various chromosomes or chromosomal regions to individual groups or consortia so that they could generate and analyze contiguous sequencing data and publish their analyses — largely describing the gene content of the chromosomes and possible disease associations based on genetic mapping studies.

But the funders and organizers also recognized that the most useful genome sequence would be one that met minimal quality standards and one that would provide free and open access to the underlying data. This latter requirement reflected a number of realities, including an interest in assessing genetic variation, in setting confidence on the base calls and the overall consensus sequence, and the nascent state of large-scale sequence assembly and annotation tools (which tacitly acknowledged that the genome sequence might be reassembled and refined through the addition of more

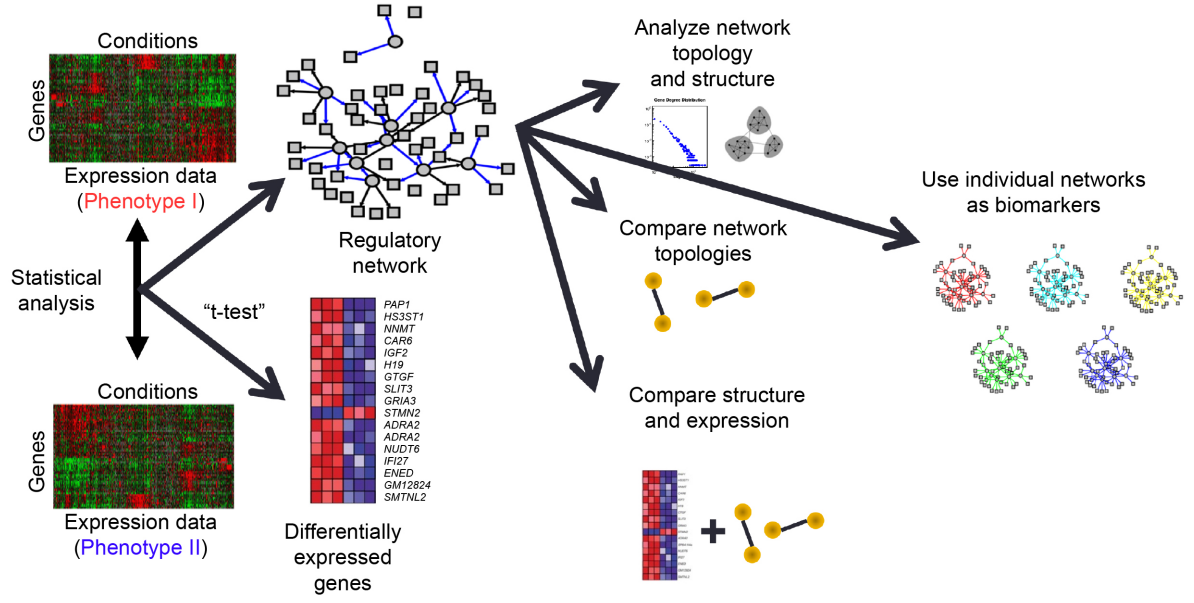

Figure 2. Much of -omic data analysis has relied on comparison of the characters of -omic features (such as gene mutational status, gene expression levels, or epigenetic modifications) between phenotypes. However, because single elements in cells do not operate in isolation, methods for inferring gene regulatory and other networks have become more important in delving into the processes driving the phenotypes we observe. These networks change in their overall topology between biological states, often in subtle but meaningful ways, which can include simple changes in regulatory "edges" between elements or larger structural changes that can create or destroy network communities that may control specific disease-associated processes. Differences in regulatory processes between individuals have led to the recognition that networks themselves can serve as biomarkers to distinguish disease states or predict clinical outcomes such as responses to therapy.

© The author(s) 2022. Published by Higher Education Press

data). The project leaders also wanted to avoid creating rifts in the scientific community between the genome data "haves" and "have nots". In addition to randomly assigning BACs to competing labs to check the consistency of sequence data, the leaders and funders gathered at a meeting in Bermuda in 1996 and arrived at an agreement unprecedented in biomedical research, requiring that all DNA sequence data be released in publicly accessible databases within twenty-four hours after generation rather than awaiting publication. These "Bermuda Principles" not only set a precedent in genomics and other fields for the rapid release of both raw and "finished" data, but it also reaffirmed a fundamental tenet in research — that results of any scientific inquiry should be open to reanalysis and falsification by other scientists. And with a project requiring what was, at the time, an irreplicable data generation effort, the only way to preserve such inquiry was to require that the data be available.

The importance of this commitment became very clear after the publication of the draft genome sequences. The genome sequencing "war" over who would finish first devolved into a sort of "cold war" over which version of the genome was best. When a number of papers pointed out shortcomings in the public genome sequence [29–32], many questioned the value of the public project. To me, however, these re-analyses were a sign of the superiority of the public effort because data availability meant that one had the opportunity to find errors. This is something that simply would not be possible with the Celera genome at the time [33]; Celera had reached an agreement with *Science* to allow data access from its servers, but limited downloads to one megabase per group per week, meaning downloading their genome would take an individual scientist about fifty-eight years[⑧] .

This commitment to basic scientific principles and open science reverberated throughout the genomics community. Many of the pioneers in large-scale gene expression analysis founded the Microarray Gene Expression Data society and proactively creating the open-data Minimal Information about a Microarray Experiment (MIAME) standard [34] and then pushed journals, funders, a scientists relentlessly to adopt useful data sharing as a requirement for publication and funding [35–38]. The success of this movement is reflected in the more than thirty number minimum information standards for data sharing that have been established over the past twenty years in fields ranging from genome sequencing to single cell analysis, metabolomics, proteomics, *in vivo* research, and even medical AI reporting. And the original proponents of the Bermuda Principles gathered in 2003, along with other open science advocates, funders, and journal publishers to affirm a broader commitment to the free and open use of genomic data pre-publication.

Despite the enthusiasm for data sharing, there have been notable failures, and successes, that themselves have further advanced the commitment to open science. After a team of scientists led by Anil Potti published a series of papers claiming to have discovered gene expression signatures that could predict a patient's response to chemotherapy — resulting in the launch of three clinical trials and the founding of a company. But a number of scientists, including Keith Baggerly of the MD Anderson Cancer Center, were unable to independently reproduce Potti's results and found contradictions between the published data and what was represented as the study data. They raised the alarm. Although the funders that had supported him and the journals that had published his work were slow to take these claims seriously, eventually the house of cards collapsed, the trials stopped, the papers were retracted, and Potti left academic medicine [39]. An Institute of Medicine committee at the National Academies was charged with looking into what led to the failures at Duke. While the committee declined to point fingers, it issued findings that called on funders and publishers to be more open to exploring questions raised about research results, called on the scientific community to less hierarchical in considering questions about research results (some junior scientists at Duke had raised questions, but were not taken seriously), and, most importantly, drew a bright line around any research destined for clinical application, calling for release of not only data, but methods and open-source software so that, at a minimum, others could reproduce the published findings if not replicate them in other data (Fig. 3) [40].

And open sharing of data has led to better science. In 2012, two large, well-funded studies simultaneously published (in *Nature*) analyses of gene expression profiles from cell lines treated with a wide range of pharmaceutical compounds, claiming that the data could be used to predict drug response [41,42]. To their credit, both groups made much of their data available and provided descriptions of their methods, but when an independent group attempted to use the data to develop predictors of drug response, they found inconsistencies between the published drug sensitivities and the overall results [43], as well as missing raw data that could possibly have been used to resolve some of these issues. The journal was at first hesitant to publish the critique, but eventually relented. And despite claims that there

[⑧]The truth is that I and many others hacked the Celera website and wrote bots that scraped the genome over a few days—before Celera plugged the holes.

© The author(s) 2022. Published by Higher Education Press

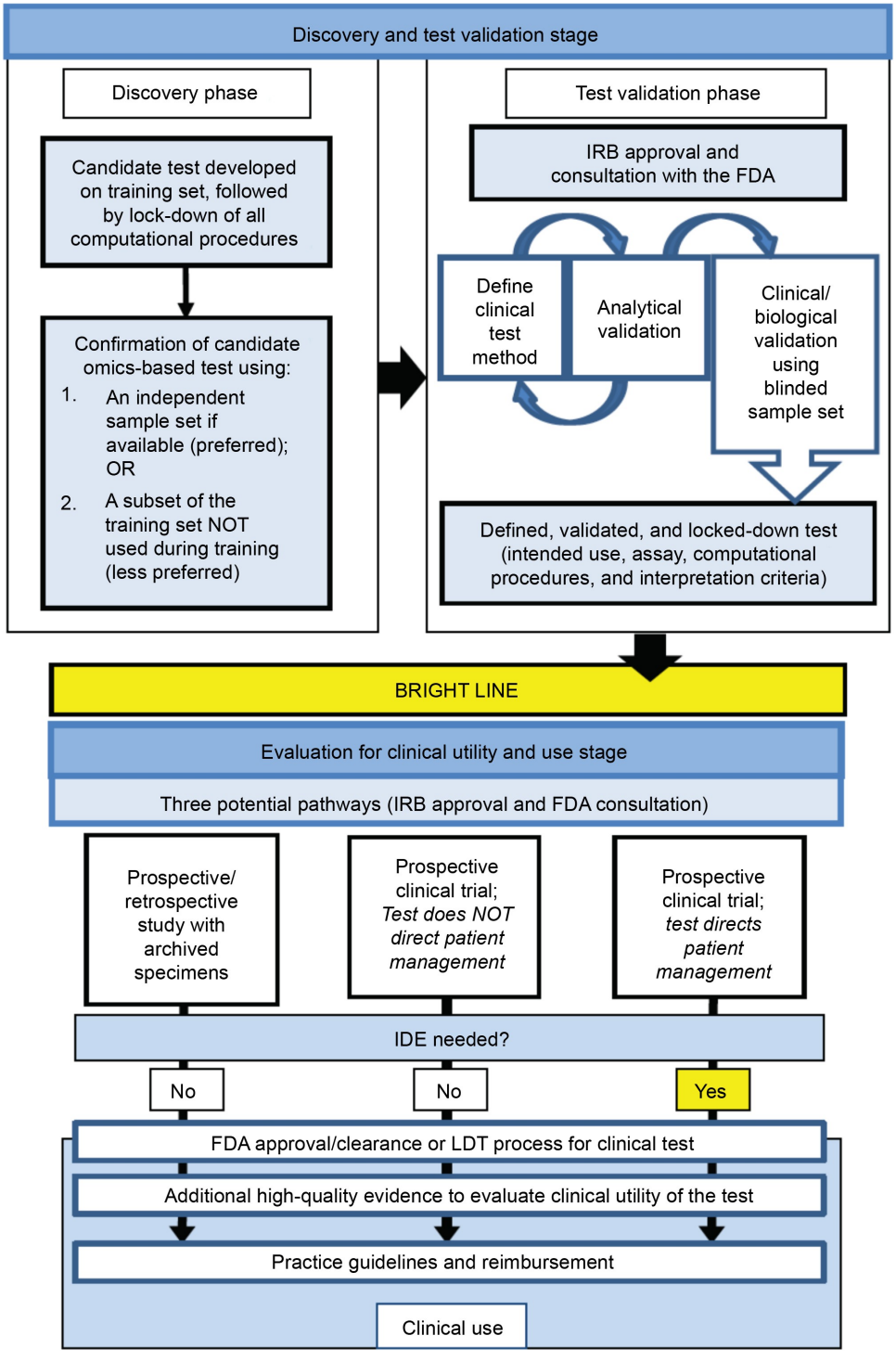

Figure 3. In its 2012 report, *Evolution of Translational Omics: Lessons Learned and the Path Forward*, the National Academies recognized the need for reproducibility in biomarker research, particularly in the use of complex, multifactorial biomarkers such as those based on algorithmic analysis of -omic features for classification, with a "bright line" between any exploratory analysis and any potential clinical application. The report called for availability of both measurement data and the software code used for prediction or classification (and overall data analysis) and sufficient demonstration of reproducibility of the proposed biomarker. Today we understand such standards must extend to other types of complex biomarkers and we recognize that reproducibility overall requires access to data and analytical methods and models.

© The author(s) 2022. Published by Higher Education Press

were no problems, the groups eventually revised their analytical methods [44] and their recent pre-publication data, in which both groups use the same viability assay, show much better consistency.

More importantly, the hard-leaned lessons about the importance of open science, including shared data, models, and software are being recognized beyond the genomics community as essential, particularly in assays being developed for clinical applications. When a team led by data scientists at Google published a breast cancer radiographic classification system claiming that it was robust enough to begin clinical trials [45], leading scientists who had been involved in ensuring genomic biomarker reproducibility quickly responded, identifying the need for transparency and laying out the tools available to achieve it [46]. Although the sequencing of the genome remains a scientific tour de force, the open science movement in biology that it inspired is arguably the Human Genome Project's contribution to the science itself.

THE VALUE OF ASKING "UNANSWERABLE" QUESTIONS

Many people view hypothesis testing as the defining element of the scientific method. And many have criticized genomics-based inquiries, including the sequencing of the genome, as "fishing expeditions"[⑨] or "hypothesis-free research." But the truth is that you generally need some data or observations to begin to formulate a hypothesis. The type of exploratory research enabled by genomics has allowed us to pose questions that could not be conceived of without genome-scale data. But this approach is not unprecedented and often arises when new technologies open new ways of looking at the world.

When Galileo Galilei built a telescope and turned it toward Jupiter, there was no way he could have hypothesized it had moons, yet the observation that Jupiter possessed four moons that clearly orbited the planet eventually led him to conclude that Nicolaus Copernicus's heliocentric model was correct — and that all of the planets, including the earth, orbited the sun. This helped drive acceptance of the model and catalyzed the development of much of modern physics and our understanding of the universe. In the same way, there was no way to hypothesize what the 25,000 genes encoded in the genome are, or that breast cancer has a set of subtypes defined by patterns of expression and that correlate with clinical subtypes [47] — although both of these technology-driven observations created new ways of understanding human health and disease. Armed with breast cancer molecular subtypes, one could then hypothesize about what biological processes drive and distinguish different subtypes and, more importantly, how one might target specific drugs to treat them.

One of the greatest joys — and most daunting aspects — of being a scientist is taking on problems for which no one knows the answer. I remember the day I "became" a scientist. It was a cool, rainy January day and I was sitting in my graduate student office, working on a problem that I had been wrestling with for months when I finally arrived at an answer. I had checked it multiple times and I knew it was right. And an immense joy washed over me because, for that brief moment, I was the only person in the world who knew the solution to the problem I was wrestling with. But this was a problem I knew about when I started, I just didn't know the answer. Doing Genomics has opened up a world of new research questions.

If we think about the universe of facts, we can divide them into subsets depending on what we know about them[⑩] . There are known knowns, which are the things we think we know and understand. Then there are the known unknowns, which in science are the problems that are most amenable to hypothesis-driven research. Then there is the class of unknown unknowns — the things that we do not yet know we even need to know. When the genome was sequenced the concept of precision medicine, in which we match patients to therapies based on their germline or somatic genome, became a priority, particularly in diseases like cancer (Fig. 4). Finding targets, developing and testing therapeutics, and understanding why those therapies don't work in everyone carrying a particular target genetic variant, requires a mix of exploratory and hypothesis-driven research — together with an informed understanding of the biology associated with the disease processes.

And while doing exploratory research can be a data driven exercise, it does not mean that the approaches we take can ignore the basics of experimental design and analysis. Do we have sufficiently large sample size to draw a meaningful conclusion? Can we develop a meaningful validation strategy? Do we have potential confounders in our assays that will bias the outcome in some way? And if we find something, will we question the results and look for potential biases and spurious correlations[⑪] ? Rather than a lazy approach to research, discovery-driven science done well actually requires a

[⑨] As Roger Bumgarner of University of Washington likes to say, "Fishing is the right choice if your goal is to catch fish."

[⑩] This is an approach that is widely used in project management, particularly for large, complex tasks, although it has roots in cognitive psychology with a technique known as the Johari Window.

[⑪] The tylervigen website.

© The author(s) 2022. Published by Higher Education Press

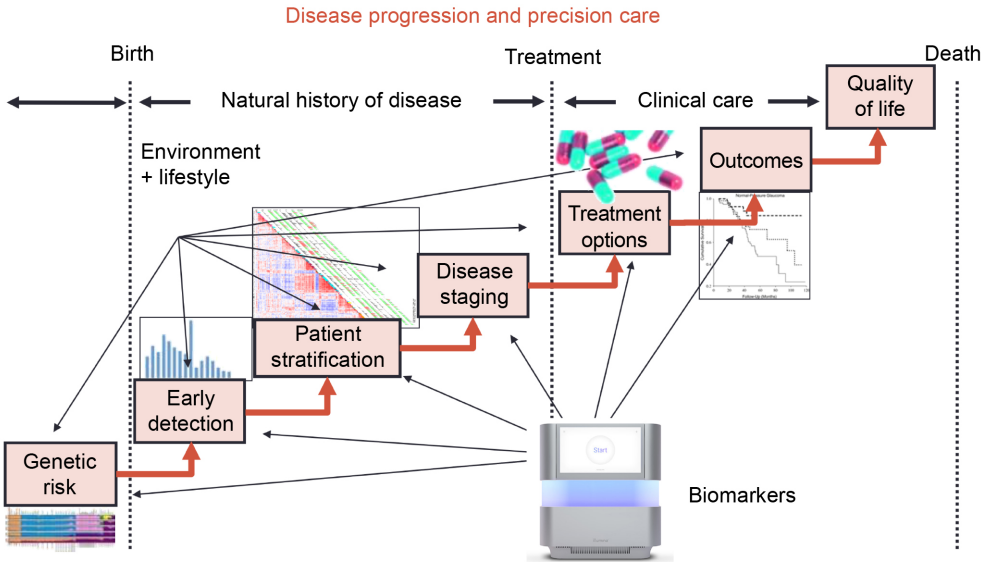

Figure 4. Precision medicine is often seen as the ultimate end goal of the Human Genome Project, with genomic data providing insight into everything from disease risk to optimal preventative interventions. But seeing this vision realized will require that we understand far more about the genome and cellular processes than we know today and, further, that we develop models that account for environmental and lifestyle effects as well as chance. All of this must be driven by a combination of exploratory and hypothesis driven research, as well as an embrace of the complexity of biological systems. Adapted from a presentation by Peter van der Spek, Erasmus University.

level of rigor in analysis that can sometimes be overlooked in the race to prove a hypothesis.

The success of genomics has really been driven by the things we didn't expect to find. That joy of discovery, of finding things that nobody else knew before because nobody had the opportunity to ask, is what makes being a scientist worthwhile. Over the years, I have gotten increasingly interested in understanding what drives phenotypic differences in health and disease, in development, between different cell types, and between individuals. I have a whole list of questions that are unanswerable simply because they are beyond the scope of available data. But if the first twenty years of the genome have taught me anything, it's that technology will likely give me data to explore, to develop hypotheses, and to test those hypotheses. Rather than focusing on narrowly defined questions, genomics has opened up the world of unexplored phenomenon and given us the opportunity to be explorers of the unknown. As a scientist, there is little that gives me greater pleasure than investigating phenomena about which we know very little or questions whose answers have long eluded us.

CLOSING THOUGHTS

The past twenty years of health and biomedical research have largely been defined by the availability of the genome sequence. Whether it is direct use of the genome sequence itself, the catalog of genes, or the technologies that have been spawned by the genome project, nearly every study in biology today relies on genomics in some way. We've mapped countless diseases, begun to unravel regulatory processes in the cell, discovered new therapies, sampled the microbiome, traced human evolution to Africa and migration across the globe, and started along the path of sequencing a substantial fraction of the human population. Genomics has enabled non-invasive prenatal testing, been used to monitor the COVID-19 pandemic, and is slowly making its way into routine healthcare testing. Sequencing isn't limited to humans either — from ecology to plant science, genomic studies are essential tools for advancing the field. And genomics is even working its way into our broader culture — it is almost impossible to turn on the radio without hearing ads for genomic ancestry testing or genome-wide precision cancer screening. It is simply what we use to make advances.

When the first genome was sequenced in 2001, it would have been almost to imagine genomics as something other than powerful approach to investigation, but one of limited potential application, but that was changing rapidly. By 2009 when I spoke about genomics, I would say that if my wife or son had a rare tumor, I would mortgage our house and sequence their genome because even then we knew enough about therapies and genomic variants that it would have been worthwhile. Today my house is safe — I can pay for a genome sequence with a credit card, store the genome

© The author(s) 2022. Published by Higher Education Press

on the cloud, and analyze it with any number of open-source tools.

Despite the progress we've made in genomic science, much remains to be done. We are still trying to understand how our genome makes us who we are and, more importantly, to model disease risk and the trajectory of its progression, as well as predicting response to therapy. As a scientist, I have been privileged to work on these fundamental problems and honored to work as part of a talented community of scientists from around the world. While I sometimes lament the fact that we will not answer all of our questions in my lifetime, I revel in the deeper understanding of who and what we are that scientists develop every day. To snowclone writer Peter De Vries description of the universe, "The genome is like a safe to which there is a combination. But the combination is locked up in the safe." For me, I am happy to finish this essay and get back to trying to unlock that safe.

COMPLIANCE WITH ETHICS GUIDELINES

The author John Quackenbush declares he has no conflict of interests.

OPEN ACCESS

This article is licensed by the CC By under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

REFERENCES

1. Rheinbay, E., Parasuraman, P., Grimsby, J., Tiao, G., Engreitz, J. M., Kim, J., Lawrence, M. S., Taylor-Weiner, A., Rodriguez-Cuevas, S., Rosenberg, M., *et al*. (2017) Recurrent and functional regulatory mutations in breast cancer. Nature, 547, 55−60
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001) The sequence of the human genome. Science, 291, 1304−1351
3. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al*. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature, 420, 520−562
4. NHGRI Press Release. "International Consortium Completes Human Genome Project". Available from the website of NHGRI in NIH
5. International Human Genome Sequencing Consortium. (2004) Finishing the euchromatic sequence of the human genome. Nature, 431, 931−945
6. Gregory, S. G., Barlow, K. F., McLay, K. E., Kaul, R., Swarbreck, D., Dunham, A., Scott, C. E., Howe, K. L., Woodfine, K., Spencer, C. C., *et al*. (2006) The DNA sequence and biological annotation of human chromosome 1. Nature, 441, 315−321
7. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., *et al*. (2021) The complete sequence of a human genome. bioRxiv, 2021.05.26.445798
8. Miga, K. H., Koren, S., Rhie, A., Vollger, M. R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G. A., *et al*. (2020) Telomere-to-telomere assembly of a complete human X chromosome. Nature, 585, 79−84
9. Reardon, S. (2021) A complete human genome sequence is close: how scientists filled in the gaps. Nature, 594, 158−159
10. Loos, R. J. F. (2020) 15 years of genome-wide association studies and no signs of slowing down. Nat. Commun., 11, 5900
11. Salzberg, S. L. (2018) Open questions: How many genes do we have? BMC Biol., 16, 94
12. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G. and Meyre, D. (2019) Benefits and limitations of genome-wide association studies. Nat. Rev. Genet., 20, 467−484
13. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., *et al*. (2009) Finding the missing heritability of complex diseases. Nature. 461, 747−53
14. Wood, A. R., Esko, T., Yang, J., Vedantam, S., Pers, T. H., Gustafsson, S., Chu, A. Y., Estrada, K., Luan, J., Kutalik, Z., *et al*. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat. Genet., 46, 1173−1186
15. Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., *et al*. (2016) The genetic architecture of type 2 diabetes. Nature, 536, 41−47
16. GTEx Consortium; Laboratory, Data Analysis & Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; *et al*. (2017) Genetic effects on gene expression across human tissues. Nature, 550, 204−213
17. Platig, J., Castaldi, P. J., DeMeo, D. and Quackenbush, J. (2016) Bipartite community structure of eQTLs. PLOS Comput. Biol., 12, e1005033
18. Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C. Y., Lopes-Ramos, C. M., Glass, K., Quackenbush, J. and Platig, J. (2017) Exploring regulation in tissues with eQTL networks. Proc. Natl. Acad. Sci. USA, 114, E7841−E7850
19. Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell, 169,

1177−1186
20. Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics, 9, 559
21. Kuijjer, M. L., Fagny, M., Marin, A., Quackenbush, J. and Glass, K. PUMA: PANDA using microRNA associations. (2019) bioRxiv, 2019.12.18.874065
22. Sonawane, A. R., DeMeo, D. L., Quackenbush, J. and Glass, K. (2020) Constructing gene regulatory networks using epigenetic data. bioRxiv. 2020.10.19.345827
23. Glass, K., Quackenbush, J., Spentzos, D., Haibe-Kains, B. and Yuan, G. C. (2015) A network model for angiogenesis in ovarian cancer. BMC Bioinformatics, 16, 115
24. Lopes-Ramos, C. M., Paulson, J. N., Chen, C. Y., Kuijjer, M. L., Fagny, M., Platig, J., Sonawane, A. R., DeMeo, D. L., Quackenbush, J. and Glass, K. (2017) Regulatory network changes between cell lines and their tissues of origin. BMC Genomics, 18, 723
25. Sonawane, A. R., Platig, J., Fagny, M., Chen, C. Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K. and Kuijjer, M. L. (2017) Understanding tissue-specific gene regulation. Cell Rep., 21, 1077−1088
26. Lopes-Ramos, C. M., Chen, C. Y., Kuijjer, M. L., Paulson, J. N., Sonawane, A. R., Fagny, M., Platig, J., Glass, K., Quackenbush, J. and DeMeo, D. L. (2020) Sex differences in gene expression and regulatory networks across 29 human tissues. Cell Rep., 31, 107795
27. Lopes-Ramos, C. M., Kuijjer, M. L., Ogino, S., Fuchs, C. S., DeMeo, D. L., Glass, K. and Quackenbush, J. (2018) Gene regulatory network analysis identifies sex-linked differences in colon cancer drug metabolism. Cancer Res., 78, 5538−5547
28. Weighill, D., Ben Guebila, M., Glass, K., Platig, J., Yeh, J. J. and Quackenbush, J. (2021) Gene targeting in disease networks. Front. Genet., 12, 649942
29. Katsanis, N., Worley, K. C. and Lupski, J. R. (2001) An evaluation of the draft human genome sequence. Nat. Genet., 29, 88−91
30. Salzberg, S. L., White, O., Peterson, J. and Eisen, J. A. (2001) Microbial genes in the human genome: lateral transfer or gene loss? Science, 292, 1903–1906
31. Andersson, J. O., Doolittle, W. F. and Nesbø, C. L. (2001) Genomics. Are there bugs in our genome? Science, 292, 1848−1850
32. Stanhope, M. J., Lupas, A., Italia, M. J., Koretke, K. K., Volker, C. and Brown, J. R. (2001) Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. Nature, 411, 940−944
33. Quackenbush, J. (2001) The power of public access: the Human Genome Project and the scientific process. Nat. Genet., 29, 4−6
34. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., *et al*. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat. Genet., 29, 365−371
35. Ball, C. A., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., *et al*. (2004) Submission of microarray data to public repositories. PLoS Biol., 2, e317
36. Ball, C., Brazma, A., Causton, H., Chervitz, S., Edgar, R., Hingamp, P., Matese, J. C., Parkinson, H., Quackenbush, J., Ringwald, M., *et al*. (2004) Standards for microarray data: an open letter. Environ. Health Perspect., 112, A666−A667
37. Ball, C. A., Sherlock, G., Parkinson, H., Rocca-Sera, P., Brooksbank, C., Causton, H. C., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F., *et al*. (2002) Standards for microarray data. Science, 298, 539
38. Ball, C. A., Sherlock, G., Parkinson, H., Rocca-Sera, P., Brooksbank, C., Causton, H. C., Cavalieri, D., Gaasterland, T., Hingamp, P., Holstege, F., *et al*. (2002) The underlying principles of scientific publication. Bioinformatics, 18, 1409
39. Kaiser, J. (2015) Potti found guilty of research misconduct. Available from the website of Science
40. Institute of Medicine; Board on Health Care Services; Board on Health Sciences Policy; Committee on the Review of Omics-Based Tests for Predicting Patient Outcomes in Clinical Trials. (2012) Evolution of Translational Omics: Lessons Learned and the Path Forward. Micheel, C. M., Nass, S. J. and Omenn, G. S., (Eds.). Washington, DC: The National Academies Press
41. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., *et al*. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 483, 603−607
42. Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X., Soares, J., *et al*. (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature, 483, 570−575
43. Haibe-Kains, B., El-Hachem, N., Birkbak, N. J., Jin, A. C., Beck, A. H., Aerts, H. J. and Quackenbush, J. (2013) Inconsistency in large pharmacogenomic studies. Nature, 504, 389−393
44. Cancer Cell Line Encyclopedia Consortium and Genomics of Drug Sensitivity in Cancer Consortium. (2015) Pharmacogenomic agreement between two cancer cell line data sets. Nature, 528, 84−87
45. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., *et al*. (2020) International evaluation of an AI system for breast cancer screening. Nature, 577, 89−94
46. Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Shraddha, T., Kusko, R., Sansone, S.-A., Tong, W., Wolfinger, R. D., Mason, C. E., *et al*. (2020) Transparency and reproducibility in artificial intelligence. Nature, 586, E14−E16
47. Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. and Brown, P. O. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat. Genet., 23, 41−46

© The author(s) 2022. Published by Higher Education Press