Title

Future Promises and Concerns of Ubiquitous Next-Generation Sequencing

Permalink

https://escholarship.org/uc/item/0dq7986g

Journal

Cold Spring Harbor Perspectives in Medicine, 9(9)

ISSN

0079-1024

Authors

McCombie, W Richard
McPherson, John D

Publication Date

2019-09-01

DOI

10.1101/cshperspect.a025783

Peer reviewed

Future Promises and Concerns of Ubiquitous Next-Generation Sequencing

W. Richard McCombie[1] and John D. McPherson[2]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724

[2]University of California Davis Comprehensive Cancer Center, Sacramento, California 95817

Correspondence: jdmcpherson@ucdavis.edu

Since the first draft of the human genome was completed, next-generation DNA sequencing technology has dramatically reduced the cost of sequencing a genome. Computational analysis has not advanced as fast as the instruments that generate the data, and storing all the data remains a challenge. Nevertheless, personal genomics has arrived and is already being used in the clinic. Significant privacy issues remain, however, and these are not widely understood. The Genetic Information Non-Discrimination Act (GINA) needs to be extended and the probabilistic nature of genetic predisposition must be better explained to both the public and physicians. We must also be wary that this promising new technology and its applications do not amplify existing healthcare disparities.

The first human genome draft was completed in 2001 (Lander et al. 2001) with a more complete version released in 2003. This sequence is often priced at $3 billion, although the Human Genome Project encompassed much more than just that one haploid sequence. Regardless of the exact cost of that first genome, it represented a milestone in possibilities and paved the way for the now feverish pace of genome sequencing. The tremendous increase in capacity has been driven by rapidly decreasing cost, making genome analysis on a large scale tractable, therefore fueling demand (www.genome.gov/sequencingcostsdata). Moore's Law is often invoked in discussing either sequencing cost decline or capacity increase. Moore's Law came about from Gordon Moore's (Fairchild Semiconductor) initial 1965 prediction of a trend for the doubling of the density of transistors in an integrated circuit about every year (Moore 1965). This was later revised to a doubling every 2 years by Moore and later 18 months by David House (Intel) referring more to the overall computer performance (en.wikipedia.org/wiki/Moore's_law). This comparison was apt in the early years of sequencing until a few years after the reference human genome release (2003–2007). But both sequencing costs and capacity increase have greatly exceeded Moore's prediction since about 2007, perhaps begging for a new term to describe this trajectory. Stephens et al. (2015) proposed that big data may appropriately be described with the adjective "genomical" and that this term could supplant "astronomical" for referring to very large entities. Sequencing costs have been plateauing for the past few years but

Editors: W. Richard McCombie, Elaine R. Mardis, James A. Knowles, and John D. McPherson
Additional Perspectives on Next-Generation Sequencing in Medicine available at www.perspectivesinmedicine.org

doi: 10.1101/cshperspect.a025783
Cite this article as *Cold Spring Harb Perspect Med* 2019;9:a025783

capacity continues to grow, likely currently approaching an exa-basepair per year (Stephens et al. 2015). The long-awaited $1000 genome has been touted as possible with Illumina's latest instrument (www.illumina.com/company/news-center/feature-articles/the-1000-dollar-genome.html) but the reality of this cost point is yet largely elusive. Regardless, genome sequencing at a personal level has arrived and has reached a reasonable diagnostic return on investment in some cases. As sequencing costs continue to decrease and available capacity increases, it is hard to imagine that genome sequence will not become part of a medical record.

But is this all there is to the equation, which is that cheap sequence equals sequence utility and adoption? Not exactly. The semiconductor analogy above is apropos as the availability of cheap computing power is an essential component of genome sequencing. Raw genome sequence data, the captured image data from base addition cycles is readily changed to a FASTQ format, representing the sequence reads and the associated quality value of each base. These fragmented data are then aligned to a reference genome and the differences between the collective sequence reads and the reference, variants, are noted. The variants are then annotated to put them in context of known data such as genes and regulatory elements and further characterized for potential functional impact and possible disease association.

However, the fact that genome sequencing improvements have surpassed those predicted by Moore's Law begs a difficult question. How are we to analyze data when the computers required to do it are not advancing as quickly as the instruments that generated it? Throughout the genome project, it has probably been the case that the combination of software and hardware (usually the former) have always been a bit behind the ability to generate data. The ability to analyze data has become much more sophisticated in recent years; however, the ability to generate it has advanced extremely rapidly. As a result, it is the case now that even a cursory analysis of the data takes more time than is required to generate it. A more sophisticated data analysis—one hesitates to use the word insightful but it is probably appropriate—is much more difficult and much more time-consuming if it can be done at all.

Storing genome data is another issue. It is often said that you only need your genome sequenced once for life. While largely true for generating a snapshot of your genetic makeup, notwithstanding sequencing of microbiome, tumors, and pathogens as needed, the longevity and cost of the storage media may complicate this view. As the electronic footprint of an individual genome requires only two bits per base—less than 2 GB total storage—a common thumb drive is sufficient. To get the genome sequenced, however, requires more data be generated as a genome is sequenced multiple times to ensure accuracy with a human whole genome sequence binary file containing these sequences; their base qualities and alignment of the sequences to a reference genome being about 50–60 GB. It could be argued that only the differences need be stored further, compressing this to less than 10 MB or approximately 10 photos taken on your phone. Unfortunately, the human genome reference is still evolving, albeit in small increments today, which may require retaining the larger sequencing data set for reanalysis to keep information current. For an individual, even that amount of data is manageable, but the collective genomic data sets for a caregiving facility loom large. This just refers to the base data but its utility comes in the information derived from the genome of an individual. These metadata add to the overall data footprint increasing the problem. Computational power is such, however, that regenerating the metadata could be done on the fly. With sequencing capacity and costs as envisioned in the future, it may become much easier, more reliable, and cheaper to utilize the extraordinary data compression of DNA and simply regenerate the data as needed from a new patient sample. This is an attractive possibility invoking images of the Star Trek tricorder but it does not solve the problem of maintaining genetic lineage for use of family genomic history in patient care.

Another issue that is problematic is simply the transfer of data itself. There are probably 50–100 centers in the world that are doing population-level sequencing. This means tens of thousands of whole genomes sequenced per year. There are hundreds of sites doing much less than that per year but still sequencing at a level that far exceeds what the entire world was capable of doing 20 years ago. This is generating vast amounts of data that are difficult to transfer to central data repositories. The reverse is true as well, that it is very difficult to transfer these large amounts of data to a local site for analysis by researchers. To some extent, the community is struggling with this issue continually. The question of whether to maintain large databases at the local site for more rapid access by researchers at that site is counterbalanced by the difficulty of downloading, storing, and frequently updating such large data sets. The alternative to this is carrying out the computational analyses at the sites where the data is stored but that has difficulties in that it may require supporting many thousands of researchers and their computational needs at a few central repositories. In reality, right now it is a mix of both of these options that are happening in the community. It is not clear how this will evolve as larger data sets become available.

The amazing expansion of sequence capacity has revolutionized research in biology and medicine with considerable advances happening every year. One area that has the potential to still open up new possibilities is in the application of long-read sequencing. The price of that still exceeds the industry standard of short-read sequencing; however, the difference between the two is shrinking rapidly. Even as little as 5 years ago, the price difference between short-read and long-read sequencing was probably in the range of 50- to 100-fold. Now, however, it is getting to be closer to five- or 10-fold higher to do long-read sequencing over short-read sequencing. As advances continue, it is likely that this cost gap will continue to decrease. Cost is not the only advantage of short versus long-read sequencing, but it is certainly a major one. In addition to cost, short-read sequencing requires far less DNA than long-read sequencing platforms but this gap too is closing. In some types of sequencing, DNA can readily be obtained for either, but for others, such as sequencing from tumors, it is difficult to obtain enough DNA in many cases for generating the longer sequence reads. With all of these gaps between short-read sequencing and long-read platforms narrowing it does raise the interesting prospect of another fundamental shift in sequencing as long reads become possible and cost-effective in more applications. Being able to do long-read sequencing in a number of different applications is perhaps a holy grail of sequencing. De novo assembly of all samples would be the norm. Right now, short-read sequencing is optimized for mapping back to a reference genome and those references are only an approximation of the new genome being sequenced. Thus, differences within individuals as well as any errors in the reference genome are not correctly reflected in the sample under study. Using individual assemblies produced by long reads could eliminate this limitation of short-read sequencing and shift the balance among platforms used in the future.

PRIVACY AND ETHICS

One area that has received constant and recently increasing attention is the issue of genomic privacy. The recent case of the capture of a suspect in the Golden State killer case made many headlines and of course no one could be against bringing justice to a serial killer. However, apart from this particular case, the issue of genomic privacy and the implications of very large databases generated, currently largely by consumer genomic companies, has to be addressed. At this point we do not really have any firm answers. There has been a general willingness to sacrifice privacy whether it be for free email or for genealogy purposes and, at some point, all of us make these compromises. But we have a concern that people do not understand the sacrifices they are making and hence cannot fully weigh the cost–benefit equation as it applies to them in a particular instance. Most of us are more than willing to accept someone knowing what we might buy on Amazon for access to free email. But are we being naïve? In the example given, probably not. But perhaps time will prove otherwise. It seems that the value of genetic information is potentially subject to considerably more misuse than information about shopping habits. Perhaps though, that is not the case simply because we are frankly pretty weak at analyzing whole-genome sequences with the exception of a few well-known traits or diseases that we may look for. But this is bound to change as we have more genomes to analyze from a research standpoint. And it seems very likely that we cannot put the genie back in the bottle as the saying goes. Once your shopping habits or your genome sequence are out there on the Internet, it is likely they can never be made to disappear. And once our data are out there, it does not result in a potential loss of just our privacy. Our genome sequence implies at a level of probability based on the degree of relatedness the sequence of our relatives. Our parents, our siblings, and our children have a partial reflection of our genome in their own. For instance, our children or our siblings would have roughly half of their genome in common with us. Do we have the right to make privacy decisions for them?

It seems though that a large number of people want to make their genomes available for purposes of genealogy or, in a more restricted availability, to answer health questions. As long as that trend continues, and it is very unlikely that it will not, it will have to be addressed at the legislative level to minimize the possibility of abuse. The 2008 passage in the U.S. Genetic Information Nondiscrimination Act (GINA) bill was a milestone in the process of protecting us from genetic discrimination through the misuse of our genomic data. Genetic discrimination occurs when an individual is treated differently by an employer (GINA Title I) or health insurance company (GINA Title II) because they have a gene alteration that causes or increases the risk of an inherited disorder. Genetic information here includes family history of disease as it often more accurately reflects the impact of your genome on your health. However, as with any law, it is not perfect in that it cannot cover every eventuality. Notably, GINA does not protect against genetic discrimination in forms of insurance other than health, such as life, disability, or long-term care insurance. While we have some ability to predict future uses that may be an abuse of genomic data, our view forward must be, by definition, imperfect. Therefore, it is probably the case that we will have to be continually vigilant to identify and legislatively address potential sources of abuse as they are detected.

The same is true regarding the privacy concerns of clinical sequence data, although the way those concerns are balanced is likely different. The interpretation of clinical sequence data, which is now the rate-limiting aspect of genomic medicine both in terms of time and precision, will be greatly benefited by the inclusion and aggregation of as many data sets as possible. There are nuances of how these data sets need to be combined based on population structure and other factors but the general premise is true. However, by combining these data sets from a large number of individuals at one level each of them is losing their privacy. It would be extremely valuable in interpreting genetic variance if patients with known phenotypes in adequate numbers could be examined to see whether there is a link between the phenotype and the genetic variant. But this raises serious concerns about privacy and protecting the individual patients from biases because of their genetic makeup. It is becoming clear that protecting privacy by methods that unlink the data from the patient identity can limit its utility. In reality, it is possible to identify a patient based on their sequence and sequences from their relatives, which may be publicly available from such sites as those used in genealogy identification. It would of course certainly be possible to identify the subject by resequencing candidate subjects and comparing the sequences or relatives of candidates. This, coupled with the above-mentioned significant benefit of pooling large numbers of patient samples to be able to associate genetic variance with disease, again leads us back to the view that the best way to balance an individual's right to privacy versus society's benefit from pooling large numbers of patient sequence data files is to stringently limit how such data could be used. As mentioned, GINA was a significant piece of legislation but it should only be viewed as the beginning. The type of protections afforded to patients in GINA should be extended to prevent DNA sequence data from being used in a variety of other ways that would be unfairly detrimental to the patients from whom the sequence was derived.

Whereas physicians and their patients are already dealing with this in some very critical interactions such as whether to get a prophylactic mastectomy in response to a genetic predisposition to increased likelihood of getting breast cancer, this is really only the beginning of such interactions and we as a society are probably not very well prepared to deal with the information. First of all, society seems ill-prepared in general to deal with probabilities. One frequently hears the response when talking about the risk of smoking that "my (fill in the relative or friend) lived to be (fill in a large number) years old and was a heavy smoker." There are no doubt cases of this but equally without doubt is the probability that a heavy smoker will not live as long as a nonsmoker. It is the probabilistic nature that is not well understood by society in general. This will be even more problematic in cases such as a genetic predisposition that makes someone more or less likely to be addicted to drugs or to commit a crime. What is often viewed as the presence or absence of willpower is at its core an individual example of an event that is probabilistic on the population and that that probability is more or less dependent upon genetic variants in individuals.

Another area of genetic privacy concerns that may have broad impact is thus far primarily been the stuff of TV shows. This topic relates to what extent a person's genetic makeup, or their genetic makeup and their various environmental interactions, affects their behavior. At this point, there are not a lot of clear examples of this. Indeed, most of these, and to some extent most genotype–phenotype relationships in general, are inherently basically probabilistic; in other words, a given genotype or genotype plus phenotype means that a person has an increased or decreased probability of having a certain condition or exhibiting a certain behavior. But the fact that such a relationship may be probabilistic does not make it any less real than a deterministic relationship.

The broad question is how knowledge of our genetic makeup and how it interacts with all aspects of our existence will impact both the perception of ourselves and how others in the overall society perceive us. At this point, examples of this may be few and somewhat vague. However, as this collection shows, the one thing that is incredibly clear is the phenomenal pace at which our knowledge of that genetic basis and its effect on us is progressing. Conditions like addiction and mental illness, which in some areas of society have considerable social stigmas attached to them, are in some cases related to a certain degree to the genetic hand that we were dealt. The same is true of other characteristics of our behavior, many of which are less well understood than addiction and mental illness. In the case of mental illness, it seems that one of the real benefits that could accrue from a better understanding of its genetic underpinnings is a removal of at least some of the stigma attached to it. It was true of cancer probably 50 years ago that there was a considerable social stigma attached to it and it was not openly discussed. That is clearly much less the case now as evidenced by the reasonably successful efforts such as Stand Up To Cancer to really draw attention to the need for a better understanding of the disease, and this has reflected very favorably in the removal of any stigma attached to sufferers of the disease.

Likewise, the same will probably happen as it becomes more ingrained into the cultural psyche that mental illness is at least largely a result of, in most cases, the individual's genetic makeup. Recent advances in the genetics of psychiatric disease, particularly of autism, will clearly go a long way toward eradicating culturally unenlightened views of what leads to these illnesses. It is far less likely to have a stigma attached to a condition if it is viewed as a medical condition rather than something that "the parents did" or resulting from a lack of willpower or some other readily addressable issue in the sufferer themselves.

Some of these issues, however, will be more complex than others. As mentioned, much of what we are learning references a probabilistic relationship between a person's genotype and various aspects of their physical and mental being. In other words, we are not our genetic makeup but we are influenced by it to varying degrees, depending upon the aspects of our physical and mental being that are being discussed. How is society to deal with a person that is perhaps twice as likely than another person to have some negative facet in their existence? What about a person that is 10 times or a thousand times more likely to have this characteristic that is perceived as negative by society? As a concrete example, how does a jury react to a person that is 50% more likely to commit a crime based on the population analysis of a particular genetic variant or combination of variants? What about 10 times more likely than other people? These complex issues, where our rapidly advancing genetic understanding interacts with other aspects of society, be they public perception or the very specific requirements of the law, are issues that will need to be grappled with increasingly in the future until society comes to an understanding of how to incorporate them into our thinking and ultimately other institutions such as the law.

The rate at which sequencing technology has been advancing is showing that considerable gaps are present in our ability to use this potential very powerful knowledge infrastructure. Many of these have to do with training of appropriate personnel. Much of this is now manifesting itself in the burgeoning fields built around analyzing sequence data. This is by nature almost exclusively a computational endeavor requiring at least some degree of knowledge in computational methods, statistical methods, and what are the most significant biological questions to ask. While it is true that many more people are being trained in this area versus 20 or 25 years ago, the demand is still being outstripped by the supply of trained investigators in this area. Generating and analyzing the data from the first human genome project took heroic efforts from computational biologists to cobble together the software to manage the data acquisition, and then to assemble, analyze, and visualize the first sequence. However, when you consider that in a period of 10 to 15 years we have gone from getting one genome done to doing tens of thousands of genomes per year, it is not possible for the analysis of that data to be scaled. Certainly, the very economies of scale are helping with this. New software is being developed to more rapidly carry out all the steps ranging from laboratory information-management systems to annotation of a final sequence of the genome. Continued efforts by the bioinformatics community has made a real impact on our ability to generate and use these data. It is clear that the system is strained as the sequencing capacity climbs upward at a relatively dizzying rate. Eventually, of course, this will reach equilibrium as new and better tools are developed, but it is always a bit of an arms race between the bioinformatics community analyzing the data and the combination of wet lab and bioinformatics generating the data.

This problem is perhaps even more challenging in a clinical setting. A very large percentage of the physicians practicing medicine today probably began their practices, and hence their education, prior to the completion of the Human Genome Project. An even larger percentage began prior to the advent of NexGen sequencing. As a result, most of these physicians did not receive much, if any, training in genomic or precision medicine. This of course can be addressed by continuing education but it still remains a problem. Some medical specialties are clearly embracing precision medicine at a faster rate than others. For instance, oncology is probably leading the pack among medical specialties adapting the new genomic information into their practice. This problem of education resulting from the newness of the capabilities is reflected through the entire medical infrastructure. For instance, genetic counselors are really at the forefront of the interface between the medical community and the patient population. However, many of these counselors, like the physicians they work with, are coming from an area in time previous to when the human genome was sequenced and previous to the past 10 years wherein next-generation sequencing has been making a growing impact. It will take time for all of these professionals to learn how to incorporate genomic information into their practices especially as it is still a topic of research and how to manage the plethora of information brought to them by their patients from increasing direct-to-consumer mechanisms. There will continue to be a growing demand for genetic counselors moving forward and it is essential that programs are available to meet this demand.

This touches on another area that we feel is an unrecognized ethical consideration. Most ethical concerns, at least by those in the bioethics profession, have to do with inappropriate application of genomic information to patients or inappropriate handling of patient genomic data. Both of these are of course considerable areas of concern. However, another important area of concern is that genomic technology and its applications to medicine clearly magnify the disparities that exist between different patients based on socioeconomic and other factors. To put it simply, a patient in an area that is, for whatever reason, medically isolated, either socioeconomically or geographically, and not served by state-of-the-art medical care using genomic information, is possibly at a considerable disadvantage from a health standpoint than those who do have access to the latest genomic information. It has always been the case of course that some patients have access to higher quality healthcare than others, but the rapidly growing genomic information is going to exacerbate those disparity issues. In most cases, a patient giving a family history to a doctor and the doctor using that information in diagnostics can happen anywhere. However, the newest genomic analysis, whether it be in oncology or the diagnosis of a disease in a newborn child that is not readily recognized from the symptoms alone, provides a potentially significant advance in possible treatment, diagnosis, or both. This disparity is present within the confines of most technologically advanced countries. It is vastly greater when comparing those technologically developed countries with the developing world. Addressing this broadening gulf between those with and without access to the latest in healthcare will be a constant and growing challenge for years to come.

But it would be inappropriate to end this article in the collection on a negative note. There clearly are issues that need to be resolved by the scientific and medical communities, the individual citizens, and society as a whole. However, hundreds, perhaps thousands, of scientists did not work for years to cause problems for people (with the possible exception of to each other—the topic of other research). In fact, there was a concerted, funded effort made from the beginning of the Human Genome Project to deal with the ethical issues that emerged. And beyond that, there are staggering potential benefits from the advances in genetics over the past 25 years, or, depending on how you would like to count it, maybe even 65 or 70 years.

One of the areas that has great potential for benefit both for the individual and for the society as a whole in terms of tangible items like driving down healthcare costs, is our ability to more accurately carry out risk management. Many diseases that have a genetic component are subject even now to such things as lifestyle interventions. One can argue, and often does sometimes over a refreshing beverage, why people do not do more in terms of intervention in their lifestyle to maximize their health. There are clearly many reasons for this and they vary to some extent from individual to individual. However, it is clear that one of the reasons is probably a lack of a precise understanding of what the risks are that one is taking. It is one thing to get into your car and go for a drive knowing that there are some risks, probably relatively few and difficult to define, associated with that and yet another to drive your car into a wall. In the latter case, the risks are much higher and much easier to understand. More and more we are learning where the walls are and this will affect behavior.

In addition, in many cases, we do not manage risk because we are totally unaware of it. If we can predict for instance that we have the likelihood of a bad response to a drug, it is unlikely that a doctor would prescribe it for us or that we would take it. Examples of this currently exist such as allergies to penicillin and others, but our knowledge of pharmacogenetics will expand the known risks that we have and help us minimize them. Another area where this new genetic knowledge is likely to have big payoffs in people's lives is in early diagnosis. Like many areas of applying genomic information to health, oncology is leading the way in this. Numerous efforts are underway to diagnose cancer at a very early stage when it can be much more readily treated using either circulating tumor DNA or enrichment analysis of cancer cells from the blood. It is probably safe to say that blood-based tests that will predict the presence of a tumor at a very early stage is the holy grail of cancer medicine. We are not there yet, but the advances that we are observing are substantial and very encouraging. Even if the test has a certain false-positive rate, it would allow further follow-up, be it radiological or by continued monitoring at decreased time intervals to allow early detection of the disease. Whereas cancer is leading the way in this, there is no reason that many diseases cannot be ultimately diagnosed at a much earlier stage, or even the risk of them diagnosed, so that the patient could be monitored more closely thus reducing morbidity and mortality from that disease. This will of course lead to improvements in interventional medicine. Oncology is leading the way here as well where individual patient tumors can be analyzed to provide improved prognostic and treatment options. Being able to target a tumor specifically based on genetic makeup is in its relative infancy but progressing very rapidly. Other areas of medicine will follow in oncology's footprint in this regard. There will be of course some diseases that are refractory to this, whereas in others it will be successful at a very early stage.

So, we have fully entered the genomic era. In both research and medicine and in our understanding of ourselves, genomics will have a role. As with anything with such powerful positive potential, there are clearly potholes in the road in terms of ethical issues, privacy concerns, and potential misuse. But there is also a tremendous benefit that is waiting for us as we use this new and powerful information to improve our ability to learn about us, the world around us, and to better understand and treat the diseases that have plagued us since the beginning of our species' time on earth.

REFERENCES

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062

Moore GE. 1965. Cramming more components onto integrated circuits. *Electronics* **38:** 114–117.

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. doi:10.1371/journal.pbio.1002195