Quantitative Biology 2021, 9(4): 366–370
https://doi.org/10.15302/J-QB-021-0273

FEATURE

# A wonderful time – exciting progress made in the past 20 years in genetics powered by the Human Genome Project

**Zhaohui S. Qin***

Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA
* Correspondence: zhaohui.qin@emory.edu

Received July 28, 2021; Revised August 4, 2021; Accepted August 6, 2021

This year marks the 20-year anniversary of the publication of the draft human genome. In this article, I review the fundamental impact the Human Genome Project (HGP) has made in biomedical research, especially in the genetics field. I also offer some perspective from the viewpoint of a bioinformatics and quantitative genetics researcher, what the future looks like for the next 20 years and beyond.

## INTRODUCTION

It is hard to believe 20 years have passed since the publication of the draft human genome [1,2]. Two thousand and one—the year after the turn of the century, is destined to go down the history as a special moment in humanity. For me, the announcement made by the HGP at that time also had a profound impact on me personally, helped to convert me from a statistician to a bioinformatics researcher.

Looking back 20 years later, I believe the excitement among the biomedical research community when the announcement is made has been justified. Indeed, research progress in genetics and genomics has accelerated massively since the completion of the HGP. Not only its precious product—the near-complete reference human genome sequence, but the technologies incubated by the HGP, the vast amount of data produced by the HGP and follow-up projects, as well as the new ways of conducting biomedical team research by international consortia. It is safe to say that, the HGP ushered in a new era of biomedical research. Many wonderful review papers have been published about the HGP, offer insights on its impact and vision for the future [3]. In this article, I try to offer some personal opinion from the perspective of a bioinformatics and statistical genetics researcher.

## THE LEGACY OF THE HGP

The HGP has been compared to the Apollo moon landing, foreseeing the dawn of a new era of genomics and transformed biology and medicine. Like the Apollo project, the legacy of the HGP is far-reaching and long-lasting. From the long list, I think the following three are perhaps the most impactful.

### Technology development

Robert Oppenheimer, one of the greatest physicists in the 20th century, once famously said, “It is a profound and necessary truth that the deep things in science are not found because they are useful; they are found because it was possible to find them.” Indeed, many of the ground-making discoveries in the scientific history are made using the latest technologies that are simply not available before. On the other hand, important and acute scientific questions, such as “what is our genome?” push and even force technologies moving forward. At the beginning of HGP, the technology is not advanced enough to meet the demand of the grand project. Thus many in the research community were skeptical whether the ambitious goals of the project

could be achieved on time. But since the question is so important, the international Human Genome Sequencing Consortium played a leadership role to challenge the entire scientific community to come up with innovative new sequencing technologies which push the HGP to complete on time and under budget. And after HGP, the need to lowering the formidable sequencing cost, again drives scientists and engineers to keep inventing, which result in the development of the next generation sequencing technologies, dramatically lowered the sequencing cost. Today, the "$1,000 genome" dreamed by genetics merely ten years ago has become a reality. The newly developed technologies, in turn, gave rise to many innovative applications, resulted in exciting new scientific discoveries.

The innovation on sequencing technology did not stop at "$1,000 genome". In recent years, the third-generation sequencing, such as nanopore sequencing from Oxford Nanopore and single molecule real time (SMRT) sequencing technologies from Pacific BioSciences, have emerged. The ability to generate super long reads make these technologies very attractive since they complement the dominating next generation sequencing technologies led by Illumina.

Another contribution of HGP on technology development is the genotyping array technologies. Using the genetic variant information discovered and cataloged by the international HapMap and 1000 Genomes projects, commercial array-based genotyping technologies have been developed and improved over the years, which made the genome-wide association study (GWAS) possible. Today, thousands of GWASs have been conducted across the globe which resulted in tens of thousands of significant associations identified. None of these findings would be possible without the array-based genotyping technologies.

### Sequencing-based applications

When electronic computer was invented more than 70 years ago, the intention is to relieve engineers from tedious computation tasks and greatly speed up processes such as rocket design. However, in today's world, scientific computing account for only a small proportion of all computer applications. Similarly, so many applications of sequencing technologies have been invented, sequencing the genome account for only a small proportion of all usages in the lab. Many sequencing-based applications have been invented over the years, including TF-binding profiling (ChIP-seq) [4], transcriptome profiling [5], chromatin accessibility (ATAC-seq) [6], 3D chromosomal organization (Hi-C) [7], just to name a few. Hundreds of sequencing-based applications have been developed and reported in the literature and the number is increasing steadily.

### From genome to multi-ome

In the pre-HGP era, the scale of biomedical research is rather limited, many researchers spend their whole career working on one or few genes. One of the greatest legacies of the HGP is the promotion of high throughput technologies, which enables scientists to study the whole organism in a holistic manner. The hallmark of this technology revolution is the birth of many "-ome" in the biomedical literature, from genome to transcriptome, epigenome, proteome, metabolome, microbiome, interactome, phenome and exposome.... Even new research discipline—systems biology was initiated to study those "omes". The latest trend is multi-omics, which advocate for investigating a biological phenomenon from multiple angles to get a bird's eye view on the subject or entity being studied.

### Team science

A huge effort like the HGP requires the entire global biomedical research community to come together and collaborate. The HGP is such a success story that its operation model was quickly cloned to many other efforts. Indeed, multiple international consortia are formed to tackle various fundamental biological questions soon after the completion of HGP including the Encyclopedia of DNA Elements (ENCODE) project [8], the international HapMap project [9,10], and the 1000 Genomes project. The ENCODE project conducted comprehensive survey of important transcription factors and histone marks on a diverse set of cell lines. The high-quality data produced by the ENCODE project gave us a detailed catalog of regulatory elements throughout the genome. The International HapMap and 1000 Genomes projects gave us comprehensive catalogs of common and rare genetic mutations in multiple populations. All of these projects are highly successful. The data generated by these projects have become indispensable resources for the biomedical research community. Many ground-break discoveries would not have been possible without them.

## WHAT'S NEXT?

Looking forward, emerging technologies trace back to the HGP have created exciting new opportunities for researchers across biomedical research fields. On the other hand, challenges remain. Here I present a brief summary.

### Many more genomes will be sequenced

With the sequencing cost falls steadily, more and more whole genome sequencing is poised to be conducted, especially in the clinics, and the number is expected to climb exponentially for the next decade. It is not unreasonable to assume that a significant portion of the entire population will undergo whole genome sequencing (WGS) or other similar alternative such as whole exome sequencing (WES). At the meantime, WGS or WES data in large data repositories such as various biobanks are expected to grow substantially. The dramatic increasing sample size of WGS and WES will give researchers better opportunities to investigate genotype-phenotype association. How to utilize WGS or WES information in the clinic, the central theme of precision medicine, is a major challenge going forward. One of the current focus is to use WGS or WES data to inform disease risk. Polygenic risk score (PRS) complied based on GWAS findings has long been introduced.

### Better reference genome

Perhaps the most important product from the HGP is the reference human genome. Reference genome is required for almost all sequencing-based studies. Improvement such as filled gaps have been made continuously over the many iterations of the reference genomes maintained by the Genome Reference Consortium (https://www.ncbi.nlm.nih.gov/grc). The latest version of the human reference genome is the build 13 of *GRch38* (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.28_GRCh38.p13/), released in March 1, 2019. Scientists are working on the next release aided by the latest sequencing technologies including the third sequencing. Additionally, new strategies are being considered. The next release, GRCg39, are expected to be drastically different from its predecessors, called the pangenome or graph genome, where diversity will be highlighted. As a result, the next reference genome will no longer be a consensus, but rather a vast collection of genomes representing all possible variants at any given locus of the genome. To address diversity issues in the reference genome, perhaps alternative routes such as ethnicity-specific reference genomes [11], or even personalized reference genome [12] may also be considered and utilized.

### Better understanding of the coding and non-coding variants

GWAS have identified tens of thousands of genetic variants that show robust statistical associations with various disease phenotypes [13]. However, the majority of the identified variants fall outside of the protein-coding regions [14]. How to discover cryptic link between non-coding genetic variants and pathophysiology of the disease has emerged as the main challenge in the post-GWAS era. Researchers have developed multiple computational approaches to identify non-coding risk variations [15–19]. Please see review paper [20] for more details. However, there are still much room left for improvement. Newly generated genome-wide profiling data, especially those from single cell experiments provide new opportunities as informative features. Additionally, with the ever-increasing computing power, machine learning and artificial intelligence have advanced rapidly, the development and application of deep learning techniques have dramatically improved the performance of classical machine learning tasks such as image recognition. Using the latest machine learning technologies may greatly improve our ability to better understand the mechanisms behind the most of the coding and non-coding variants.

### Other types of mutations

Thanks to the Herculean effort made by the HGP, the International HapMap and the 1000 Genomes projects, a rather comprehensive catalog of single nucleotide variants (SNVs), short indel, and copy number variants have been established. However, there are other types of genetic variants still unaccounted for. In particular, short tandem repeats (STR), a sequence of two or more nucleotides that is repeated multiple times continuously, have been shown to be associated with multiple neuropathological disorders including Huntington's disease, Kennedy's disease, myotonic dystrophy, Fragile X syndrome and several spinocerebellar ataxias [21,22]. Due to the nature of STR, short read-based sequencing reads that contain STR variants will have difficulty to map to the reference gnome, hence their presence is difficult to detect using current technologies. But the situation is likely to change when the third-generation sequencing technologies becomes widely-available. We expect more disease-associated STRs to be discovered in the near future.

### Challenges in ethical, legal, and social implications (ELSI)

A unique and urgent challenge emerged with the HGP is the unforeseen ways that genomics interacts with our daily life. Such interactions give rise to many ELSI issues that call for serious attention from the scientific community. Here are some examples.

Privacy

Genetic mutation profiles derived from high-throughput genotyping, WGS or WES contains more than enough information to ID the person. Although a blessing for the law-enforcement agencies if used properly, this could be a double-edge sword and if abused, may have a potential devastating adverse impact on the society. Furthermore, the health-related information contained in such data may be exploited to put individual citizen at a disadvantaged position.

Health disparity

A signature new research paradigm originated from the HGP is precision medicine, in which treatment design and delivery is tailored toward individual patients. The success of precision medicine relies on our knowledge of genotype-phenotype association. It has long been recognized that there is much diversity among different populations. Unfortunately, minority populations are underrepresented in most of the genomic databases. Combating health disparity requires resources, expertise and training devoted to underrepresented and underserved communities.

Ethics

Rapidly developing genetic editing technologies such as CRISPR/Cas9 [23] will soon find their way to the clinics. While powerful and potentially life-saving, such technologies have the potential to be used to create heritable genetic alterations of different organisms including plants, animals and human. These changes have the potential to affect physical traits, disease risks and survival. Application of such technologies will create many ethical, social and legal issues that the scientific community must work with other community stakeholders to address. Currently, research in these areas are lagging behind the rapid technology development.

## CONCLUDING REMARKS

There is no doubt that the HGP will be recognized as a major scientific break-through of the 21st century. Its long-lasting impact will be felt in many years to come. I felt extremely lucky to start my career soon after the completion of the HGP 20 years ago. Many of the research problems I worked on today are unthinkable without the HGP. Although just a light-weight researcher, I do feel the same way as the titan—Sir Issac Newton once said in 1675 “If I have seen further it is by standing on the shoulders of Giants.”



## REFERENCES

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., 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
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. Collins, F. S., Green, E. D., Guttmacher, A. E. and Guyer, M. S., and the US National Human Genome Research Institute. (2003) A vision for the future of genomics research. Nature, 422, 835−847
4. Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007) High-resolution profiling of histone methylations in the human genome. Cell, 129, 823−837
5. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods, 5, 621−628
6. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods, 10, 1213−1218
7. Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., *et al*. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science, 326, 289−293
8. Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., *et al*. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447, 799−816
9. International HapMap Consortium. (2005) A haplotype map of the human genome. Nature, 437, 1299−1320

10. Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., *et al*. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature, 449, 851−861
11. Dewey, F. E., Chen, R., Cordero, S. P., Ormond, K. E., Caleshu, C., Karczewski, K. J., Whirl-Carrillo, M., Wheeler, M. T., Dudley, J. T., Byrnes, J. K., *et al*. (2011) Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet., 7, e1002280
12. Yuan, S., Johnston, H. R., Zhang, G., Li, Y., Hu, Y. J. and Qin, Z. S. (2015) One size doesn't fit all-refeditor: Building personalized diploid reference genome to improve read mapping and genotype calling in next generation sequencing studies. PLOS Comput. Biol., 11, e1004448
13. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., *et al*. (2014) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. Nucleic Acids Res., 42, D1001−D1006
14. Zhu, Y., Tazearslan, C. and Suh, Y. (2017) Challenges and progress in interpretation of non-coding genetic variants associated with human disease. Exp. Biol. Med. (Maywood), 242, 1325−1334
15. Ritchie, G. R., Dunham, I., Zeggini, E. and Flicek, P. (2014) Functional annotation of noncoding sequence variants. Nat. Methods, 11, 294−296
16. Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M. and Shendure, J. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet., 46, 310−315
17. Ionita-Laza, I., McCallum, K., Xu, B. and Buxbaum, J. D. (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet., 48, 214−220
18. Chen, L., Jin, P. and Qin, Z. S. (2016) DIVAN: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. Genome Biol., 17, 252
19. Zhou, L. and Zhao, F. (2018) Prioritization and functional assessment of noncoding variants associated with complex diseases. Genome Med., 10, 53
20. Rojano, E., Seoane, P., Ranea, J. A. G. and Perkins, J. R. (2019) Regulatory variants: from detection to predicting impact. Brief. Bioinform., 20, 1639−1654
21. Ellegren, H. (2004) Microsatellites: simple sequences with complex evolution. Nat. Rev. Genet., 5, 435−445
22. Ryan, C. P. (2019) Tandem repeat disorders. Evol. Med. Public Health, 2019, 17
23. Hsu, P.D., Lander, E.S. and Zhang, F. (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell. 157, 1262−1278