On the ultimate finishing line of the Human Genome Project

Jun Yu[1,*] **and Songnian Hu**[1,2]

[1]University of Chinese Academy of Sciences, Beijing 100190, China

[2]The Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

*Correspondence: junyu@big.ac.cn

Received: April 10, 2021; Accepted: June 8, 2021; Published Online: June 12, 2021; https://doi.org/10.1016/j.xinn.2021.100133

© 2021 This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Citation: Yu J. and Hu S. (2021). On the ultimate finishing line of the Human Genome Project. The Innovation **2(3)**, 100133.

The Human Genome Project (HGP) has paved the way for the Digital Personal Genomes (DPG), whereby a person's complete genome sequence serves as the primary entry for the Digital Healthcare Systems (DHS). If the goal is to deliver PDG on demand, affordability, mainly the cost, becomes one of the primary issues. However, as a once-in-a-lifetime event, a total cost of US $100 is close to $1 per year contribution. Therefore, what remains is still largely an engineering challenge, since a multi-fold reduction of the current per-genome sequencing cost may be achievable by increasing the scale of operation and the degree of automation over 5–10 years. The second issue is to differentiate scientific achievements from what is applicable to the healthcare systems and human well-being in general. The original thought of HGP is to understand the genetics of cancers—an idea or a proposal was made and debated in the early 1980s, and it is still clear that we need both high-quality genome sequences and time to fully understand their encoded biological information; the success of HGP lies on the separation of these two goals. Over the past 40 years or so, not only has the urge to use PDG for healthcare purposes reached a new height as the two National Institutes of Health-led projects, HGP and PMP (Precision Medicine Project),[1,2] have been paving the way toward this finishing line and the "All-of-Us" effort has delivered its sequence data approaching millions, but also there have been increasing demands on genome sequences of entire populations. It becomes more apparent that, at the finishing line, fully digitalized genomic data will serve as a permanent ID in DHS and Digital Life System (DLS; a more detailed description about the relationship between PDG, DHS, and DLS is illustrated in Figure 1). Even if we assume the existence of such DHS and DLS, there is still an enormous number of foreseeable challenges, both scientifically and organizationally, some of which we would like to share here.

The scope of PDG data

A current view of PDG data (all data deposited to DHS are preferably longitudinal over a person's lifespan) include at least four basic categories: (1) the genomes (both the mitochondrial and the nuclear, together with their heteroplasmic and non-germline variants, respectively); (2) cell- and organ-based ribogenomes; (3) epigenomes at single-cell and single-molecule levels; and (4) condition-defined microbiomes of the human body.

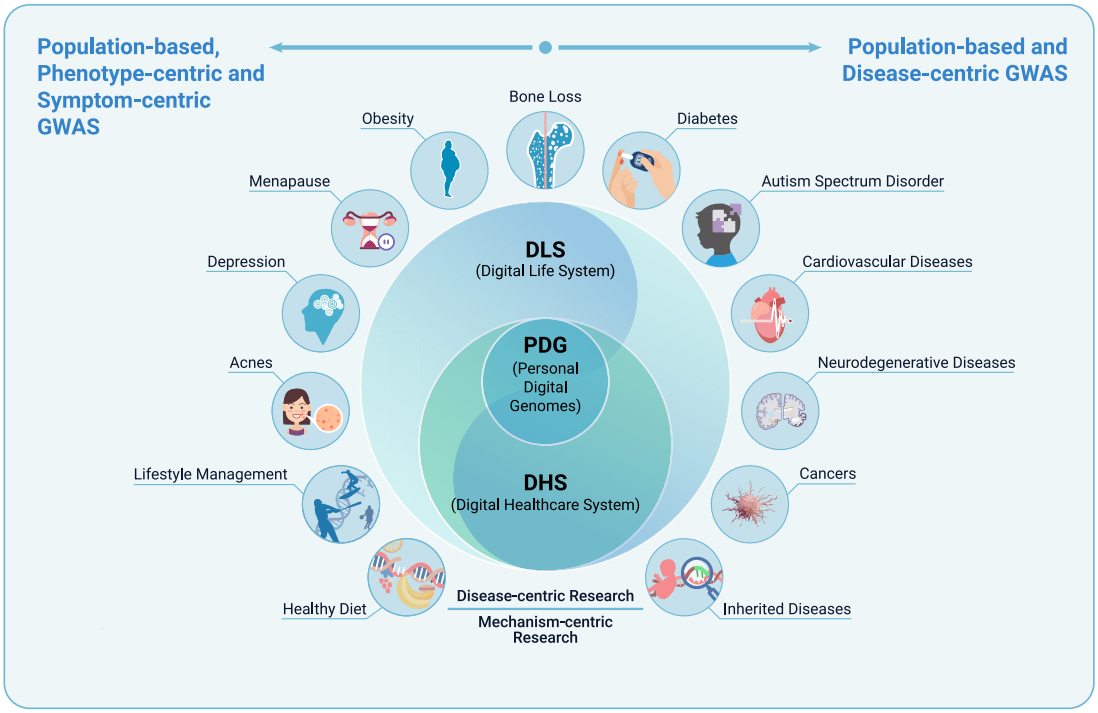

Figure 1. An illustration of how the three digital systems, PDG (Personal Digital Genomes), DHS (Digital Healthcare System), and DLS (Digital Life System), are unified
DHS starts with PDG but DLS contains all and covers the entire lifespan. Association studies are thus expanded to link genotype and phenotype as symptom-associated (even trait-associated) in addition to disease-associated investigations. The Daoist Yin-Yang background indicates that physiological and pathological elements, including diet, lifestyle, lifespan-related preclinical conditions (such as *Propionibacterium acnes* infection and menopause) and early phase signatures of common diseases, disease stage-associated symptoms (such as obesity and bone loss), and disease complexity (inherited and complex diseases), are all intertwined as gradual processes from small to large or from loose to tight within a person's lifespan at population levels.

Commentary

The breadth of phenotyping

When a research project morphs into an infrastructure-creation project, disease-based cohort studies expand into population-based symptom phenotyping and the magnitude of such effort changes considerably; billions in currencies have to be invested and nationwide efforts have to be orchestrated. There are at least three tracks to be considered: (1) the biomedical track, designed to look for inheritable factors and including diseases of all categories, rare or common, simple or complex, transmittable or age-related, long-term or casual; (2) the environmental track, looking for geographical, vocational, and hazardous health-risk factors; and (3) the lifespan track, defining and collecting age-associated preclinical data for various chronic diseases. It is undoubted that an all-out effort, track-specific study designs, and all-purpose data management platforms are all necessary.

In the Core Facility and Information Hub for Genotype-to-Phenotype (G2P) Data Generation and Management, data generation ensures quality and efficiency of the data production and houses and partitions data, while management converts data into information and knowledge and connects it to relevant systems of various stakeholders including, but not limited to, the healthcare, public health, drug discovery, infant care, early education, and aged care systems. The two entities are designed to be permanently operational, and the digital G2P records accompanied by PDG and scientific literature are intended to be stored as public archives. Population-based and disease-centric genome-wide association studies (GWAS) are to be extended to phenotype-centric and symptom-centric GWAS.

The timely development of phenotyping assays and instruments

In addition to best-formulated questionnaires, accurate measurements to generate high-quality data are of essence for phenotyping, such as those for heart rate, blood sugar and pressure, and urine tests. There are two alternatives for sampling and data collection: the use of either a direct-to-consumer device (such as wearable device) or a POCT (point-of-care-test) instrument. For instance, to collect data for a population-based study on sleeping disorders, a wearable or positioned surveillance device that traces and quantifies the subject's movement becomes prerequisite.

A watchdog system for causal agents of global pandemics

SARS-CoV-2, the causal agent of COVID-19, has led to an estimated economic loss of many trillions of US dollars in 2020 alone.[3,4] Billions of US dollars should have been invested on a per annum basis to study and survey the origin and transmission routes of coronaviruses via natural hosts and to establish a global surveillance system for this virus in particular as well as others, such as the flu viruses and Ebola, in general. The HIV/AIDS studies and intervention efforts in the past 35 years have provided a positive example of disease prevention but a final solution, such as a vaccine for HIV, has not been achieved.

International collaboration and data sharing

The research communities, especially the genome research community, hold the tradition to maintain public databases for best sharing. The new challenges are the enormous amount of data and sharing mechanisms beyond research communities. Nevertheless, discussions among the stakeholders are of essence for the formulation of guidelines and standards for data quality. For PDG data, we have to be extremely careful when differentiating private genetic variations (that are limited in number and unique to a person) and privacy protection for individual PDG data; the former is less informative for population-based studies and the latter has to be shared after desensitization.

Plans for all-life-form genome sequencing

These plans are not going to be random but prioritized based on the need of biomedical, agricultural, and environmental studies, as well as biodiversity conservation. For instance, our own microbiomes are related to and frequently exchange genetic materials with those of our living environments, and it would be very dangerous to ignore the fact that these environmental microbiomes are also variable according to seasons and climate variables. Since whole genome sequencing for population-based studies of any wild animals and plants are rare, affordable alternative genotyping techniques should be developed.

New data types and an outlook for future technology platforms

The ultimate demand from PDG is faster and cheaper sequencing. DNA sequencing, especially the next-generation sequencing platforms, are adequate for the current need in terms of both cost and throughput, albeit exhibiting a few pitfalls (such as short read length and long turnaround time). Two features for future sequencing technology development are of essence: long read length for better sequence assembly and direct sequencing of RNA. Two long-read platforms have been provided by Oxford Nanopores and Pacific Biosciences, but the cost and throughput of their platforms have to be comparable when applied as complementary or alternative data, accompanying those of the short-read platforms. The current ribogenomics relies on reverse transcription and RNA-specific characteristics, such as hundreds of covalently modified nucleotides and their synthetic pathways, are lost completely in the process. A meaningful analysis of an RNA molecule has to meet three simple parameters: position (presence or absence of a modified nucleotide with single-molecule resolution), identity (molecular structure of the modified nucleotide at the position), and degree (the actual count of modified sites over the expected). Therefore, convergence of single-cell, single-molecule technologies and platforms represent the future direction of sequencing technology development.

Integration and extension of data beyond genomics

Two lines of research activities and their data integration have to go forward: one concerns functional interpretations of the genomic data at large multi-dimensionality and the other is the extension of other omics data generation based on improved technologies. The scale and resolution of such omics data types, such as proteomics, glycomics, interactomics, and metabolomics, should match those of genomics, preferably at single-cell and single-molecule levels. Therefore, microelectronics and microfluidics technologies are the choices for highly parallel and high-precision operations, such as lab-on-a-chip and organ-on-a-chip. Of course, aside from computing power and proficient platforms, visualization technology and system animation at cellular level are also highly beneficial for data integration and comprehension.

The success of HGP and its sequels suggests that it is time to sequence everyone's genome by building an infrastructure to sequence better and sequence cheaper. HGP is undoubtedly an excellent investment, as its economic impact has been reported to have a return on investment of over 250-fold.[5] A step into the PDG era marks another massive investment, 20- to 40-fold more than that of HGP as a decade-long project, and once such an infrastructure is ready for operation the annual budget for maintaining it should be a trivial fraction of the annual healthcare expenditure over a lifespan. What we are promoting here is a sharp transition of genomic information acquisition from the coupling of basic and clinical research projects and patients to a full call for concerted efforts of all members of Village Earth for their health and well-being.

REFERENCES

1. NHGRI History and Timeline of Events. https://www.genome.gov/about-nhgri/Brief-History-Timeline.
2. Yu, J. (2016). Precision medicine: what do we expect in the scope of basic biomedical sciences? Genomics Proteomics Bioinformatics ***14***, 1–3.
3. Cutler, D.M., and Summers, L.H. (2020). The COVID-19 pandemic and the $16 trillion virus. JAMA ***324***, 1495–1496.
4. Global economy gets COVID-19 shot from US stimulus, but pre-existing conditions worsen. (2021) https://unctad.org/news/global-economy-gets-covid-19-shot-us-stimulus-pre-existing-conditions-worsen.
5. Battelle Technology Partnership Project for United for Medical Research (UMR) (2013). The Impact of Genomics on the U.S. Economy. https://web.ornl.gov/sci/techresources/Human_Genome/publicat/2013BattelleReportImpact-of-Genomics-on-the-US-Economy.pdf.

DECLARATION OF INTERESTS

The authors declare no competing interests.