Quantitative Biology 2017, 5(3): 272–275
DOI 10.1007/s40484-017-0114-5

NEWS

Strategic planning for national biomedical big data infrastructure in China

Zhen Wang[1], Zefeng Wang[1,*] and Yixue Li[1,2,*]

[1] Key Lab of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] Shanghai Center for Bioinformation Technology, Shanghai Industrial Technology Institute, Shanghai 201206, China
* Correspondence: wangzefeng@picb.ac.cn, yxli@sibs.ac.cn

Received May 19, 2017; Revised June 16, 2017; Accepted June 16, 2017

The promise that big data will revolutionize scientific discovery and technology innovation is now being widely recognized. With the explosive growth of biomedical data, life science is being transformed into a digital science in which novel insights are gained from in-depth data analysis and modeling. Extensive and innovative utilization of biomedical big data is a key to the success of precision medicine. Therefore, constructing a centralized national-level biomedical big data infrastructure becomes crucial and urgent for China. Such infrastructure should achieve superb capacity of safe data storage, standardized data processing and quality control, systematic data integration across multiple types, and in-depth data mining and effective data sharing. Full data chain service including information retrieval, knowledge discovery and technology support can be provided to data centers, research institutes and healthcare industries. Relying on Shanghai Institutes for Biological Sciences, agreements have been signed that a main node of the infrastructure will be located in Shanghai, and a backup node will be set up in Guizhou Province. After a construction period of five years, the infrastructure should greatly enhance China's core competence in collection, interpretation and application of biomedical big data.

Keywords: biomedical big data; national infrastructure; precision medicine

SCIENTIFIC SIGNIFICANCE OF BIOMEDICAL BIG DATA

The big data revolution will transform our life, business and technology [1]. In the thematic studies in 2011 and 2014, McKinsey Global Institute brought forward that the application of big data will become the basis of future competition, "underpinning new waves of productivity growth, innovation, and consumer surplus". It emphasizes that big data is the critical infrastructure to support all 12 disruptive technologies on the horizon. Big data provide fascinating possibility to create advanced artificial intelligence (AI) with faster knowledge extraction and decision-making. For example, the landmark victory of AlphaGO in 2016 was a powerful demonstration of the potential of AI that can learn from huge amounts of data [2].

Many disciplines of modern sciences, including life science, astronomy and high-energy physics, are transformed by the broad application of big data. A vision of data-intensive scientific research was declared by A. M. Turing Award winner Jim Gray in 2007 [3]. In 2009, Microsoft Research introduced the term "*Fourth Paradigm*", pointing out that in addition to theory, experiment, and simulation, data-intensive discovery will bring revolutionary change for scientific research [4]. In fact, the completion of Human Genome Project in the early 21st century indicates that life science had entered the era of big data and a new research paradigm of life science appears [5]. With the declined cost of next-generation sequencing since 2007, the output of biological data has grown explosively. In 2015, the scale of genomic data have well been beyond petabyte (10^{15} bytes, PB) and will soon exceed exabyte (10^{18} bytes, EB), which is either on par with or exceed the capacity of other highly demanding data domains such as astronomical data [6]. Besides, biomedical big data have the characteristics of being all-

round, multi-dimensional, dynamic and of high content. Even for a single individual, various levels of omics data and clinical/medical records could be used to decipher the biological system at different levels [7].

Biomedical big data is closely embedded with a series of national life science projects, and is also a curial concept underlying precision medicine [8,9], a initiative advocated by US President Obama in 2015 to develop tailor-made treatments based on an individual's genetic content or other molecular or cellular analysis. Under this initiative, healthcare will be profoundly affected by digital management and interventions resulted from the wide and deep application of biomedical big data.

NECESSITY AND URGENCY OF NATIONAL INFRASTRUCTURE OF BIOMEDICAL BIG DATA

We are living in the era of big data, which has become another fundamental resource of a nation like population, land and minerals. Big data infrastructure, which provides the capacities of large-scale data collecting and storage, as well as processing and analysis, is the important part of the country's core competitiveness. Infrastructure for biomedical big data with outstanding scale and function is the key platform to support both basic research in life sciences and technology innovation in biomedical industry. In fact, constructing a centralized national infrastructure of biomedical big data is the main strategy embraced by developed countries like United States, European Union and Japan. Founded in late 1980s and early 1990s, data centers in the National Center of Biotechnology Information (NCBI) of US, the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ) have dominated the international life science research up to now [10]. Especially, NCBI provides the most comprehensive collection of online resources for biological data in the world, covering sequences, genes, genomes, proteins, chemicals, literature and healthcare [11].

Although significant fractions of the genomic data deposited in NCBI is generated from China, there is no dependable domestic infrastructure to conduct comprehensive management of these data and to provide effective services over the entire country. As a result, China has been forced to become the biggest data exporting country. Taking the genomic data as an example, although national projects provide the majority of funding to acquire a large number of data, China is heavily dependent on foreign countries for management of database and knowledge base system. Furthermore, many genomic data are scattered in individual laboratories and institutions with unreliable data quality, fragmented storage and severe loss, making the development of effective knowledge a very difficult task and resulting in an insufficient use.

Just one year after former US President Obama announced the Precision Medicine Initiative, China has also incorporated its own version of precision medicine into the 13th Five-Year National Science and Technology Innovation Plan, as well as a 15-year outline through 2030 [12]. This series of projects will produce a huge amount of biomedical data through basic researches using various omic technologies, data repository of cohort studies, and clinical studies using clinical trails or real world evidences. The abilities of large-scale data storage, standardization, analysis and integration, together with the policies for data security and data sharing, play a vital role in these projects. A consensus has been reached among the research institutes, healthcare institutions and biomedical enterprises that the goals of precision medicine cannot be achieved without a national infrastructure of biomedical big data, which should provide dependable support for data archiving and management. It is critical not only to serve the demand of national strategic plans for science, but also to guarantee the security of national data resources.

To meet the urgent demand of national strategy for scientific, technological and social development, we propose to establish a comprehensive national infrastructure for biomedical big data in China as soon as possible. The infrastructure should rely on a third party national scientific institute, and should be a high-level, non-profit, well-managed entity operated by personnel with proven experiences to implement national scientific infrastructure. The infrastructure should be able to integrate various data types, accomplish data standardization and quality control, and provide services for data analysis and sharing. Using such infrastructure as a launch pad, we will be able to continuously elevate the national ability for safe data storage, information sharing, technological innovation, standardization system improvement, Intellectual property appreciation and efficient utilization of biomedical data. In response to this proposal, the Chinese government has announced a guideline at the end of 2016 to set up a national infrastructure of biomedical big data in the next five years with an initiative budget around 200 million USD. In March 2017, the president of Chinese Academy of Sciences and the mayor of Shanghai Municipality have signed an agreement to promote this national infrastructure to be built at Shanghai Zhangjiang high-tech park as part of the international innovation center.

FRAMEWORK FUNCTIONS AND KEY FEATURES FOR THE INFRASTRUCTURE

The infrastructure we seek to build will have dual

functions to support both scientific research and new technology development. It will enhance life science researches like big data processing related to precision medicine, and strengthen the integration between omics data and clinical information with emphasis on complete data connection. Data generated from national projects in life science and precision medicine, as well as the international life science databases will be incorporated. The relevant policies dealing with personal privacy, data security and data sharing will also be taken into consideration as the "soft" components of the infrastructure.

We seek to achieve four capacities in biomedical big data: i) the capacity of safe data storage; ii) the capacity of standardized data processing and quality control; iii) the capacity of multi-dimensional data integration; iv) the capacity of in-depth data mining/analysis and effective data/knowledge sharing. These four capabilities can be achieved by four major functional modules, which are named as software and hardware supporting module, data specification and standardization module, multi-level database module as well as biomedical big data mining module, respectively. These systems and modules are established based on three sub-facilities: i) Sub-facility of storage, disaster recovery, computing and network, which provides the hardware and software to support data storage and ensure data security; ii) Sub-facility of big data standard research and development, which standardizes various types of biomedical data and related analyses and applications; iii) Sub-facility of big data management and application, which provides various databases, searching engines and visualization interfaces. In addition to standard pipelines, a workflow engine will be implemented to support an open platform for user-defined data analyses.

The most important feature of the infrastructure is to generate a chain of scientific and technological innovation through integrating data chain. The biomedical big data are structured as a data chain flowing from data to information, scientific knowledge, novel technology and engineering. The huge amount of biomedical big data are in rapid growth and from multiple sources, which should firstly be standardized and aggregated into high-quality data sets. Subsequently, comprehensive and integrated analyses are applied to extract biomedical information for various applications. Finally, novel knowledge on biology and medicine are acquired through data mining and analysis. All related participants, including sequencing centers, research and clinical institutions and healthcare industries could be integrated into the data chain, which will release the full potentials of big data.

Overall, the infrastructure will be implemented as a center for biomedical big data standardization and safe storage, a center for data analysis service, and a center for technology development. It also serves as a platform open to the public for scientific research as well as a base for international data exchange and cooperation. The comprehensive extension of core competence will not only support user-oriented data chain service, but also guarantee national co-ordinate activities.

CONSTRUCTIONS PLAN AND FEASIBILITY

The main node of the biomedical big data infrastructure will be set up in Zhangjiang Hi-Tech Park, Shanghai. With well-established top colleges, research institutes and hospitals, Shanghai takes a leading position in scientific resources, talents and research demands. Furthermore, many national scientific facilities and bio-pharmaceutical industries are clustered in Zhangjiang Hi-Tech Park, reaching the critical mass for that promotes novel scientific discoveries. Shanghai Institutes for Biological Sciences (SIBS) is seeking to host the project and Shanghai government will provide financial support. A land scale of 20,500 square meters is planned for the first phase and work site construction, including 10,000 square meters reserved for the second phase. A server room of 8,500 square meters will be equipped with 200 server cabinets, providing an initial data storage capacity of 100 PB to 1 EB. The computing capability will be implemented by 100 high-performance blade servers, 200 parallel computational nodes, 50 fat nodes and 100 GPU computing servers.

In June 2016, SIBS signed a contract with Gui'an government to build a data storage/backup node in the national big data industrial park in Gui'an, GuiZhou Province. Financial support will be provided by the local government, and 50–70 acres of land will be allocated in the first phase. The operation of the facility can save energy consumption by 80% per year. Also, the long-term plan will build additional nodes of the infrastructure to form a network over the entire country to provide regional service.

As bases of the infrastructure construction, we have mirrored over 30 international databases, including the 1000 Genome Project, UCSC Genome Browser and Ensembl. We have deployed the Bio-Med Big-data Center (http://www.biosino.org/bigbim/index), which catalogs series of databases we have built including genomics, transcriptomics, proteomics, epigenomics, variations and phenotypes/diseases. Especially, we have created the National Omics Data Encyclopedia (NODE, http://www.biosino.org/node/index), which is a database for submitting and publishing raw high-throughput sequencing data like NCBI SRA. A collection of database systems for microbiome, precision medicine and large-scale visualization are also being built. To encourage data produced

from projects funded by national agencies to be deposited to the infrastructure, we have established collaborations with the most influential journals of biology and medicine in China, including *Cell Research*, *Molecular Plant*, *Journal of Molecular Cell Biology* and *Acta Biochimica et Biophysica Sinica*.

After three years' construction, the infrastructure is expected to achieve a data storage volume of 100 PB or more, a daily data entry volume of 15 TB and a daily access of 1.2 million (equivalent to 1/3 of NCBI). When the infrastructure construction is completed after five years, China will have the full capability to manage and utilize biomedical big data.

SUMMARY AND OUTLOOK

China is facing an aging population in the 21st century. Constructing a centralized and non-profit national infrastructure for biomedical big data will enable an effective management and utilization of the precious data resource in health science, which is vital and urgent for the competitiveness of this nation. The large scale infrastructure with diverse functions will not only consolidate the basis of biomedical "Big Country of Data", but also realize the dream of biomedical "Powerful Country of Data". The next five years will be a critical window to build such a national center, because related technologies in both life science and data science are becoming mature so that the academic institutions, biomedical industry and healthcare providers can all benefit greatly.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program on Precision Medicine (Nos. 2016YFC0901704, 2016YFC0901900 and 2016YFC0901600), the National Grand Program on Key Infectious Diseases (No. 2015ZX10004801-005), and the National High Technology Research and Development Program (Nos. 2015AA020104 and 2015AA020108).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Zhen Wang, Zefeng Wang, and Yixue Li declare that they have no conflict of interests.

REFERENCES

1. Mayer-Schönberger, V. and Cukier, K. (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Boston: Houghton Mifflin Harcourt
2. Chouard, T. (2016) The Go Files: AI computer wraps up 4-1 victory against human champion. Nature, doi: 10.1038/nature.2016.19575.
3. Gray, J. (2009) Jim Gray on eScience: A Transformed Scientific Method. Hey, T., Tansley, S., and Tolle, K. M. eds. In The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, WA: Microsoft Research, xix
4. Hey, T., Tansley, S. and Tolle, K. M. (2009) The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, WA: Microsoft Research
5. Hood, L. and Rowen, L. (2013) The Human Genome Project: big science transforms biology and medicine. Genome Med., 5, 79
6. Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S. and Robinson, G. E. (2015) Big data: astronomical or genomical? PLoS Biol., 13, e1002195.
7. Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. and Tegnér, J. (2014) Data integration in the era of omics: current and future challenges. BMC Syst. Biol., 8, I1.
8. Ashley, E. A. (2016) Towards precision medicine. Nat. Rev. Genet., 17, 507–522.
9. Gligorijević, V., Malod-Dognin, N. and Pržulj, N. (2016) Integrative methods for analyzing big data in precision medicine. Proteomics, 16, 741–758.
10. Cochrane, G., Karsch-Mizrachi, I., Takagi, T. and the International Nucleotide Sequence Database Collaboration. (2016) The International Nucleotide Sequence Database Collaboration. Nucleic Acids Res., 44, D48–D50.
11. Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Edgar, R., Federhen, S. (2016) Database resources of the national center for biotechnology information. Nucleic Acids Res., 44, D7–D19.
12. Zhan, Q. and Qian, H. (2016) Opportunities and Advantages for The Development of Precision Medicine in China. In Precision Medicine in China. Sanders, S. and Oberst, J. eds., pp. 6–9. Washington, DC: Science/AAAS

