BLOOD RESEARCH
Volume 48 · Number 3 · September 2013
http://dx.doi.org/10.5045/br.2013.48.3.167

Perspective

Surviving in the era of "Big Data"

Seog-Woon Kwon, M.D., Ph.D.

Department of Laboratory Medicine, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

In the Stone Age, livelihood required only a few tools, including stone axes, stone blades, stone mortars, and earthenware. These simple tools may have been adequate for survival in that period. However, human lifestyle has dramatically changed over a long history of "ups and downs." There have been many remarkable advances in the fields of knowledge, education, and science. These advances have brought us many benefits and led us to marvelous scientific breakthroughs, truly a great leap for mankind.

The invention of computers is one of the greatest masterworks. According to the records, the world's first programmable automatic computing machine was invented in early 1940s. In 1981, International Business Machines Corporation (IBM) introduced its first personal computer (PC) equipped with 16 kilobytes of memory (expandable to 256 kilobytes) and using the Microsoft Disk Operating System (MS-DOS). Although Bill Gates once said (and later denied) that "640 kilobytes ought to be enough for anybody" with respect to the IBM PC at that time, we now use PCs with more than 4 gigabytes of memory.

The invention of the Internet along with the search engines, including Google (which derives its name from a misspelling of "Googol," which means 10^{100}), is another triumph for humanity. These powerful tools enable us to surf worldwide information networks, including GenBank (the NIH genetic sequence database), and provide us a mechanism for dissemination of information and a forum for collaboration and interaction between scientists.

In 1665, an English scientist, Robert Hooke, discovered the cell. He saw the "small rooms" through his microscope and coined the word "cell" simply because they looked like small rooms where the monks lived in. Now we can see the "cosmos" in the cell, where an overwhelming number of functional molecules are living. In 1953, James Watson and Francis Crick saw the double helix structure of DNA [1]. We can now see the complete DNA sequence or genomes of numerous species, including human genome. By virtue of the Human Genome Project, which was started in 1990 and completed in 2004, approximately 3.3 billion base-pairs of human genome have been fully sequenced and identified, and we became to know that there are approximately 20,500 genes in human beings [2], roughly the same range as in mice.

Having a great diversity, the genes are expressed by specific signals in the microenvironments, and sometimes, are actively involved in disease development and progress. Scientists are now confronted by a vast number of tasks required for understanding the expression of these genes and their association with various diseases. The Sanger sequencing method definitely contributed to the completion of the Human Genome Project, and ushered in the age of genomics. However, this generated a requirement for a new tool for performing high-throughput sequencing.

The advent of next-generation sequencing (NGS) technology has enabled rapid sequencing of large stretches of DNA base-pairs [3]. Up to one terabase of data can be obtained in a single sequencing run by using NGS technology. It is really breathtaking to realize that researchers are now able to sequence more than five human genomes in a single run, producing data in a week by computer at a low cost of less than $5,000 per genome. This revolutionary advance in genomic science will surely help scientists extract genetic information from biological systems as much as they need and provide a powerful tool to explore cells and diseases.

Scientists frequently encounter limitations because of large data sets in not only meteorology (the study of atmosphere) and connectomics (the study of the connectivity of synapses in the brain) but also genomics. In being committed to our work as hematologists, we have to look into the fine details of blood cells (including immune cells), both healthy and malignant, at the genetic and molecular levels. This generates data about genomic sequences, protein sequences, protein structure and function, molecular interactions, signaling and metabolic pathways, regulatory motifs, etc. [4]. We are inevitably facing these "big data" challenges as well as those from laboratory and clinical findings. Maybe we have to learn how to manage these "big data" for surviving.

REFERENCES

1. Watson JD, Crick FH. The structure of DNA. Cold Spring Harb Symp Quant Biol 1953;18:123-31.
2. An overview of the Human Genome Project. National Human Genome Research Institute, National Institutes of Health. (Accessed September 13, 2013, at http://www.genome.gov/12011238.)
3. Church GM. Genomes for all. Sci Am 2006;294:46-54.
4. Editorial. Community cleverness required. Nature 2008;455:1.