MINI REVIEW Genomics & Informatics Vol. 6(4) 161-165, December 2008

Personal Genomics, Bioinformatics, and Variomics

Jong Bhak[1], Ho Ghang[1], Rohit Reja[1] and Sangsoo Kim[2]*

[1]KOBIC (Korean Bioinformation Center), KRIBB, Daejeon 305-806, Korea, [2]Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea

Abstract

In 2008 at least five complete genome sequences are available. It is known that there are over 15,000,000 genetic variants, called SNPs, in the dbSNP database. The cost of full genome sequencing in 2009 is claimed to be less than $5000 USD. The genomics era has arrived in 2008. This review introduces technologies, bioinformatics, genomics visions, and variomics projects. Variomics is the study of the total genetic variation in an individual and populations. Research on genetic variation is the most valuable among many genomics research branches. Genomics and variomics projects will change biology and the society so dramatically that biology will become an everyday technology like personal computers and the internet. 'BioRevolution' is the term that can adequately describe this change.

Keywords: personal genomics, bioinformatics, variomics

Introduction

Since the launch of the Human Genome Project (HGP) in 1990 by NIH of USA, researchers have been developing faster DNA sequencers (Chan, 2005; Gupta, 2008; Mardis, 2008; Metzker, 2005; Shendure *et al.*, 2004). HGP has been said to be led by James Watson who modeled DNA in Cambridge, UK in 1953. In 2003, the International Human Genome Sequencing Consortium held a press conference to announce the completion of the human genome (IHGSC, 2004). In 2008, after 55 years, Watson's complete genome sequence was publicized by using 454 DNA sequencers developed by a company rather than a research institute (Wheeler *et al.*, 2008). In 2007, Craig Venter, a former Celera founder, published his own personal genome in PLoS Biology (Levy *et al.*, 2007). We are entering the personalized biology era with the advent of next generation sequencing technologies.

*Corresponding author: E-mail sskimb@ssu.ac.kr
Tel +82-2-820-0457, Fax +82-2-824-4383
Accepted 29 November 2008

DNA Sequencing

The first breakthrough in genome sequencing came from Watson's colleague, Fred Sanger, in Cambridge, UK. In 1977, Sanger and his team produced the first useful DNA sequencing method and publicized the first complete genome (Sanger *et al.*, 1977). It was a tiny virus genome known as phi X 174. Soon after phi X 174, he published the first complete organelle genome which was a mitochondrion (Anderson *et al.*, 1981). By 1998, researchers in the US evaluated multiplex genome sequencing technologies and were aware that one person's whole genome could be sequenced in a day using contemporary technologies. George Church was a Ph.D. student of Walter Gilbert who received a Nobel Prize with Sanger for developing a sequencing method. Gilbert's method was not widely used. However, his colleague Church continued to develop sequencing methods. One of them is based on the Polony idea (Porreca *et al.*, 2006). This technology is used by KNOME Inc., a full genome sequencing company. Along with KNOME, other companies, such as Complete Genomics, are now producing DNA sequences cheaply and in an unprecedented capacity. The speed of sequencing is advancing many folds per year, much faster than the cycle of semiconductor chips in computer industries. Also, genome sequencing technology is becoming an everyday technology at the level as computer CPUs are universally used. In five years' time, experts predict that everyone in developed nations will be able to have his or her own genome information. Due to its far reaching consequences in medicine, health, biology, nanotechnology, and information technology, DNA sequencing will become the most important industrial technology ever developed during the next decades.

Personal Genomics

In 2009, genome sequencing technologies will achieve one person's whole genome per day in terms of DNA fragments sequenced. Personal genomics is a new term that utilizes such fast sequencers. In 2008, the cost for one personal genome is less than $350,000 USD. If the cost goes down below $1,000 USD, the impact of personal genomics is predicted to be the largest ever in biology in common people's lives. Reflecting this technological advancement to society is the PGP (Personal Genome Project), a project to sequence as many people as possible with lowest possible cost (Church, 2005). At

present, Google, Inc. and the Church group are working together to sequence 100,000 people's genetic regions of DNA. In Saudi Arabia, the government is planning to sequence 100 Arabic people's genome. In Europe, there are various groups of people and nations who have been genotyping those populations. Iceland has been especially successful in that effort by utilizing their well-kept genealogical data encompassing hundreds of thousands of people. In Asia, Jeongsun Seo of Seoul National University has been working on the East Asia Genome Project during the past several years. His group has collected thousands of samples from Mongolian tribes with a extremely large genealogical tree among them (Park *et al*., 2008; Sung *et al*., 2008). Seo is said to be sequencing at least 100 Korean genomes in collaboration with Church and Green Cross, Inc. of Korea. The aim of Seo's genome project is to produce a resource for East Asians. He is presently sequencing at least two Korean people. In China, Beijing Genome Institute has been successful in terms of sequencing. Their first achievement came from a plant genome, rice. After rice, they launched a 100 Han Chinese genome sequencing project. In Nov. 2008, they published their first Chinese genome in a journal, Nature. In Dec. 2008, another Korean group, Lee Gilyeo Cancer and Diabetes Institute (LCDI) and Korean Bioinformation Center (KOBIC) made a Korean genome sequence public. The genome was sequenced by Solexa paired-end sequencer, and comparative genomics analyses and SNP data were uploaded as a public resource. It took only one week to analyze the 7.8x Korean genome using 150 computer CPUs to produce mapping DNA fragments to a reference genome, generate new SNP information, compare that with other individual genomes, and map it with 1600 already known phenotype information from the public literature.

Genome Revolution

These public genome data alongside previously known Craig Venter's and James Watson's mark that full genome sequences are not soley in academic domain anymore. Anyone who has money and the will can sequence human genomes. This 'genomic revolution' will eventually lead to the 'BioRevolution' in terms of making the most essential human information completely mapped and publically available. This is revolutionary, because humans can now engineer themselves with a map or a blue print not directly relying on trial and error style conventional evolutionary methods. This indicates that evolution has moved to a conscious level driving evolution. We are in effect designing evolution using computers.

Genomes and Personalized Medicine

The consequences of 'BioRevolution' where genomic information is utilized by scientists to engineers all kinds of biological processes, including evolution itself, will bring us personalized medicine. The essence of personalized medicine is that enzymes in our tissues, such as cytochrome P450, have distinct differences among individuals and populations. Certain drugs produce different responses in individuals.

Cytochrome p450 Family Example

The cytochrome P450 (CYP) family of liver enzymes is responsible for breaking down more than 30 different classes of drugs during Phase I of drug metabolism. Structural and SNP variations of the genes that code for these enzymes can influence their ability to metabolize certain drugs. Based upon this, a population can be categorized into four major types of drug metabolizers:

- Extensive metabolizers: Individuals that can be administered with normal drug dosage
- Intermediate metabolizers: Individuals that metabolize drugs with a slower than normal rate.
- Poor metabolizers: Individuals with poor metabolizing rates. Drugs may accumulate and cause serious adverse effects.
- Ultra metabolizers: Individuals with metabolizing rates even faster than extensive metabolizers. They may experience no effect of drug activity.

In early 2005, the US FDA cleared the AmpliChip® CYP450 Test, which measures variations in two genes of the CYP450 enzyme system: CYP2D6 and CYP2C19. The Roche AmpliChip CYP450 Test is intended to identify a patient's CYP2D6 and CYP2C19 genotype from genomic DNA extracted from a whole blood sample. Information about CYP2D6 and CYP2C19 genotype may be used as an aid to clinicians in determining therapeutic strategy and treatment dose for therapeutics that are metabolized by the CYP2D6 or CYP2C19 gene product.

Variomics

The most important scientific data out of personal genomes are the precise sequence differences among individuals. Such differences have many types. There are structural differences among chromosomes. There can be insertions and deletions of DNA segments. There are certain fragments that appear as repeats in genomes. Mapping all these structural genetic variations can be briefly termed 'variomics'. A variome is the totality of genetic variation found in an individual, a population,

and a species. Among all the variations we know, the most common is the single nucleotide polymorphisms (SNP). In Korea, mapping the variome has been pursued relatively early, and there are several groups who are mapping the genetic variations. KOBIC has several very early stage, if not the earliest in the world, variome servers: http://variome.net and http://variomics.net. Along with SNP variation, the copy number variation (CNV) is also important. Some recent reports tell us that CNVs can be as variable as or even more variable than SNPs that are simple DNA base changes in populations. Yeun-Jun Chung of the Catholic University of Korea has been mapping CNVs among Korean people (Kim *et al*., 2008).

Human Variome Project (HVP)

As an international collaboration, headed by Richard Cotton, HVP was launched in 2006 (http://humanvariomeproject.org) (Ring *et al*., 2006). HVP aims to make clinicians who have been working on rare diseases, to work together with molecular biologists and bioinformaticians. Their goal is to link medical information with genotype information. Succinctly, this process is called genotype to phenotype mapping. As several full human genome sequences are already available, mapping phenotypes to full genomes will be the major challenge of biology in the next 20 years.

Asian Variome Project (AVP)

Alongside and with the associations of eIMBL, A-IMBN, and HVP, a variome project that is working to map the Asian population variome was launched in 2008. This was a group effort by Korean researchers who have been interested in genome sequences, SNPs, and CNVs. They have formed the KOrean VAriome Consortium (KOVAC: http://variome.kr) and support AVP as one of the first projects. eIMBL, the virtual laboratory network of Asia linking key biology groups modeled after EMBL, has acquired $80,000 USD in 2008 to support AVP. eIMBL aims to establish a virtual bioinformatics center in the Asia Pacific region that will link many bioinformation processing scientists in Asia.

Construction of Reference Genomes for the World

Sanger Center, EBI, NCBI, and the University of Washington Genome Center have formed a consortium to produce a reference genome (http://referencegenome.org). A reference standard is the most important standard among all the standards. Providing an accurate reference genome to biologists is an important task. The first reference genome by the above consortium is based on Caucasian genomes. Due to the extent of SNPs and CNVs, it is necessary to construct reference genomes for diverse ethnic groups. In Korea, since 2006, the reference standard genome project began and produced the first draft for Koreans in November, 2008, using a male donor. Through the bioinformatic analysis, the Korean researchers in LCDI and KOBIC found that there was a good justification for any nation to launch large scale genome projects to map population diversities. Even such close populations as Korean and the Chinese showed a large quantity of SNP differences.

Bioinformatics for Personal Genomes and Variomes

Bioinformatics is the key in personal genome projects and variome projects. Bioinformatics is not merely a set of tools but a scientific discipline. It regards life as a gigantic information processing phenomenon and works to map its components and to model the emerging networks of the components. Bioinformatics in 2008 is driving biology into an information science. Most biology research projects produce massive amounts of data that cannot be processed by hand. Nearly all biological research outcomes in the next five years will have some form of high throughput data such as genome sequences, microarray data, proteome analyses, SNPs, epigenome chips, and large scale phenotype mapping. Bioinformatics tools in genomics and variomics can be found from various internet resources. There are several bioinformatics hubs such as NCBI (National Center for Biotechnology Information), EBI (European Bioinformatics Institute), DDBJ (Databank of Japan), and KOBIC. Some others are: Bioinformatics Organization (http://Bioinformatics.Org), EMBnet (http://www.embnet.org/), and The International Society for Computational Biology (http://iscb.org).

The following are major bioinformatics journals:

- Algorithms in Molecular Biology (http://www.almob.org/)
- Bioinformatics (http://bioinformatics.oxfordjournals.org/)
- BMC Bioinformatics (http://www.biomedcentral.com/bmcbioinformatics)
- Briefings in Bioinformatics (http://bib.oxfordjournals.org/)
- Genome Research (http://genome.cshlp.org/)
- Genomics and Informatics (http://www.genominfo.org)
- The International Journal of Biostatistics (http://www.bepress.com/ijb/)
- Journal of Computational Biology (http://www.

liebertpub.com/Products/Product.aspx?pid=31&AspxAutoDetectCookieSupport=1)
- Cancer Informatics (http://www.la-press.com/journal.php?pa=description&journal_id=10)
- Molecular Systems Biology (http://www.nature.com/msb/index.html
- PLoS Computational Biology (http://www.ploscompbiol.org/home.action)
- International Journal of Bioinformatics Research and Applications (http://www.inderscience.com/browse/index.php?journalcode=ijbra)

Sequencing DNA, Metagenomics, and Ecogenomics

Next generation sequencing methods will not only map genomes. They will be used to map the environment. This is called ecogenomics. To humans the environment can mean various microbial, plant, and animal interactions around us. Microbial interaction is especially critical to our health. Gut bacteria are a natural environment within us. Metagenomics is a methodology that sequences the whole set of microbes in our food tract. Researchers are realizing that the human genome is complemented by such environmental genomes. A new term, 'ecogenomics' is now used to describe these concepts. Metagenomics and ecogenomics are for mapping the variations of environmental genetic factors.

Mapping Expression using DNA Sequencing

DNA sequencing technologies were mostly used for mapping genotypes. However, they are now used to map RNA expression levels in cells. Cells produce various types of RNA. mRNA is the most abundant and important. In the past, microarray and DNA chips were used to measure expression levels. They are not accurate and take many bioinformatic adjustments before producing reliable expression data. New sequencing technologies can measure expression levels much more accurately. By sequencing the RNAs, we can now quantify the expression levels by precisely knowing the RNA sequences. Sequencing technologies will restructure the expression analyses in the future.

Linking Genome Information On-line

Sequencing a genome is basically the production of data, whereas analyzing the whole genome takes human minds networking their hypotheses, proofs, and discoveries, i.e. genomics is a scientific endeavor beyond mechanical sequencing. Therefore, a worldwide effort is required to link all the genome information for proper management and utilization. The internet is the best infrastructure for genome information exchange. Bioinformatics resources should be available as freely as possible for all nations, including those underdeveloped and developing. Genome sequencing and associated analyses should be done freely in certain instances by the support of local governments and international organizations. For maximum efficiency, an adequate data and information license should also be required. Some researchers propose an openfree sharing of bioinformatics analysis tools, as well as the genome sequences (under proper permission). One such movement is Free Genomics (http://freegenomics.org).

The following are on-line genomics sites:
- Genomics portal: http://genomics.org
- Personal Genome Project: http://personalgenomes.org
- openfree Genomics Project: http://personalgenome.net
- Personal Genome sequencing company: http://www.knome.com
- Personal Genome SNP typing: http://decodeme.com
- Google's Personal Genome Typing: http://23andme.com
- The Sanger Centre: http://sanger.ac.uk
- General Omics site: http://omics.org
- Korean Genome Data Site: http://koreagenome.org
- Korean Bioinformation Center: http://kobic.kr

Conclusion

We have examined the current trends in genomics and variomics. In 2009 and onwards, personal genome projects will produce an unprecedented amount of biological data. New bioinformatics technologies will be required to handle them. New sequencing technologies will drive the next decades of biology and transform medical practices. Fast sequencing brought us interesting and unexpected applications such as metagenomics and ecogenomics.

Acknowledgements

SK was supported by Soongsil University Research Fund. JB, GH, and RR were supported by KRIBB/KOBIC fund from the MEST of Korea. The authors thank Maryana Bhak for editing the manuscript.

References

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P.,

Roe, B.A., Sanger, F., Schreier, P.H., Smith, A.J., Staden, R., and Young, I.G. (1981). Sequence and organization of the human mitochondrial genome. *Nature* 290, 457-465.

Chan, E.Y. (2005). Advances in sequencing technology. *Mutat. Res.* 573, 13-40.

Church, G.M. (2005). The personal genome project. *Mol. Syst. Biol.* 1, 2005.0030.

Gupta, P.K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26, 602-611.

IHGSC. (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945.

Kim, T.M., Yim, S.H., and Chung. Y. (2008). Copy number variations in the human genome: potential source for individual diversity and disease association studies. *Genomics & Informatics* 6, 1-7.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L., and Venter, J.C. (2007). The diploid genome sequence of an individual human. *PLoS Biol.* 5, e254.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133-141.

Metzker, M.L. (2005). Emerging technologies in DNA sequencing. *Genome Res.* 15, 1767-1776.

Park, H.S., Kim, J.I., Cho, S.I., Sung, J.H., Kim, H.L., Ju, Y.S., Bayasgalan, G., Lee, M.K., and Seo, J.S. (2008). Genome-wide Linkage Study for Plasma HDL Cholesterol Level in an Isolated Population of Mongolia. *Genomics & Informatics* 6, 8-13.

Porreca, G.J., Shendure, J., and Church, G.M. (2006). Polony DNA sequencing. *Curr. Protoc. Mol. Biol.* Chapter 7 Unit 7 8.

Ring, H.Z., Kwok, P.Y., and Cotton, R.G. (2006). Human Variome Project: an international collaboration to catalogue human genetic variation. *Pharmacogenomics* 7, 969-972.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* 265, 687-695.

Shendure, J., Mitra, R.D., Varma, C., and Church, G.M. (2004). Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.* 5, 335-344.

Sung, J.H., Lee, M.K., and Seo, J.S. (2008). Inbreeding coefficients in two isolated Mongolian populations - GENDISCAN Study. *Genomics & Informatics* 6, 14-17.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A., and Rothberg, J.M. (2008). The complete genome of an individual by massively parallel DNA sequencing. *Nature* 452, 872-876.