Asia Pacific allergy

pISSN 2233-8276 · eISSN 2233-8268

Educational & Teaching Material

doi: 10.5415/apallergy.2011.1.2.93
Asia Pac Allergy 2011;1:93-97

Introduction to bioinformatics: sequencing technology

Soyeon Ahn*

Medical Research Collaborating Center, Seoul National University Bundang Hospital, Seongnam 463-707, Korea

Bioinformatics, the study of integrating high throughput biological data and statistical model through intensive computation, has been attracting great interest in recent times and Sequencing is at the very center of it. The large amount of information obtained from sequencing has deepened our understanding and fundamental knowledge of organisms. This review will aim to provide a brief summary of new sequencing technology, current issues, and projects focused on medical applications. The article is organized in three parts. Part I explains common terminologies and background of sequencing technology, and Part II compares distinct features of currently available platforms. Part III contains applications in various medical fields.

Key words: Bioinformatics, Next-generation sequencing; Massively parallel sequencing

Part I. Introduction to sequencing technology

On April 15, 1953, Francis Crick and James Watson proposed the double helical structure of the DNA molecular structure [1]. Since then, methods have been devised to determine the sequence of DNA residues, which serves as a blueprint of organism. A conventional sequencing means the Sanger-type sequencing, which is capillary-based, laboratory-intensive work. The human genome project (HGP) accelerated progress in sequencing, but sequencing remained a cumbersome procedure despite of its importance.

The first commercial next-generation sequencing (NGS) was launched in 2005 [2], about 50 years after the discovery of the DNA structure by Crick and Watson. Although NGS is the current generation sequencing method, it does not have a prototype model/platform yet. Generally it refers to a type of sequencing that does not need bacterial artificial chromosome cloning but runs automatically based on enzymological amplification [3]. Alternatively, high-throughput sequencing or massively parallel sequencing is often used to describe extremely vast amounts of outputs. The Roche/454 FLX Titiniaum and the Illumina/HiSeq2000 are the most commonly available platforms.

Recently a new sequencing method termed next-NGS (NNGS) or no-amplication NGS had been developed. Compared to NGS, NNGS requires no amplification step; therefore it is a single-molecular-based technology, hence it is often defined as the third generation sequencing replacing the first generation automated Sanger method and the second generation NGS [4]. Helicos is the first commercially available NNGS platform but Pacific Bioscience is emerging as a new pathfinder. An even newer fourth generation sequencer has been developed, which

Correspondence: Soyeon Ahn
Medical Research Collaborating Center, Seoul National University Bundang Hospital, 166 Gumi-ro, Bundang-gu, Seongnam 463-707, Korea
Tel: +82-31-787-4894
Fax: +82-31-787-4044
E-mail: ahnsoyeon@gmail.com

Received: July 1, 2011
Accepted: July 6, 2011

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government [NRF-2009-351- C00107].

This is an Open Access article distributed under the terms of the Creative Commons Attribution. Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2011. Asia Pacific Association of Allergy, Asthma and Clinical Immunology.

is a platform that enables sequencing without imaging. Up to third generation sequencing, sequencers depend upon enzymatic cascade and optical (fluorescence) detection, which is high-cost-and-low-efficiency. Ion Torrent is the first fourth generation sequencer, making a breakthrough by utilizing digital detection on chips.

The main driving force behind the development of new generations of sequencers has been the reduction of cost. The-thousand-dollar-human-genome has been a catchword for the new sequencing technology. DNA sequencing costs has been decreasing at a very high rate [5]. Note that the HGP cost $3 billion and now it is estimated to cost less than $1,000 within a few years. Nonetheless, sequencing the whole genome is expensive and cost-ineffective. An alternative approach is to sequence specific regions of DNA rather than whole genomes, a strategy called targeted sequencing. For example, exome (-targeted-) sequencing, which focuses on the exome (exonic regions, which is ~5% of human genomes) has been popular because of its effectiveness to identify potential mutations. More than 20 rare Mendelian disorders have been identified so far utilizing this method, including Miller syndrome (family-based design) [6], Kabuki syndrome (unrelated individuals) [7], and even a set of ion channel mutations believed to cause Mendelian disorders. See [8] for a review on recent works.

From the technological point-of-view, the *in vitro* cloning step clonally amplified polymerase chain reaction (PCR) enabled NGS due to its time-saving processing. The most common methods are emulsion PCR (emPCR) and bridge PCR. emPCR is a method for DNA amplification using waters and oil emulsion to amplify without loss of DNA molecules. Bridge PCR is performed on a slide where affixed primers provide series of DNA amplification. Clonally amplified PCR has opened a new era in NGS but ironically it is destined to discontinue due to development of single-molecular-based sequencing technology. For a comprehensive review of the sequencing techniques, please refer to [9-11].

Regardless of different platforms that use diverse tactics to shorten time and cost, the only product and by-product obtained from sequencing is read. Read, or sequence tag is a very short DNA sequence that is assumed to be a copy of the true genome sequence. In practice, however, it also contains footprints of individual variants and systematic errors of current platforms. The short length of the read was recognized as one of the obstacles of early NGS technology. The first read sequence of Solexa (Illumina, USA) was 20-30 bases on average, which could be likened to like assembling 10^8-piece puzzle sets to reconstruct the human template. Even though the Sanger method still produces relatively long reads (300-1,000 bases), a short read length is not a barrier anymore for some platforms: the 454 FLX Titanium (Roche, USA) produces 400 bases, and the PacBio RS (Pacific Biosciences, USA) produces 1,000 bases on average. Also, Illumina (USA) has been successfully proven that rather short sequences (e.g., 50 bases) are sufficient for re-sequencing purpose and plausible even for *de novo* sequencing [12]. Currently most reads from NGS are single-end but there are some modifications for longer read length and accurate alignment, such as mate-paired/paired-end, strobed reads. Those reads (paired-end or strobed read) acquire local positional information and tell reads located nearby, whereas single-end reads are only aligned to one position on the genome like an island. There are subtle differences between mate-paired (library) and paired-end (sequencing) in terms of how the library is made [13].

Beside the length of read, another aspect to consider is the amount of reads. The (average) coverage is defined as the average number of times a position in the genome is actually sequenced. In rare cases, the percent coverage is used to represent the percentage that a position is sequenced at least once. Reads are chopped randomly as shotgun style and there are repeated regions in the genome like retrotransposon. It is, therefore, generally agreed that at least 20-30 coverage is required to resequence the human genome acceptably. It should be noted that coverage is not uniformly distributed, possibly for several reasons: not randomly sheared fragment, not uniformly amplified DNA molecules due to genomic sequences composition differences [14] or chromatin status [15].

Having sufficient reads is not the only important issue in sequencing and bioinformatics. After obtaining reads, proper modeling plays an essential role in dissecting data to abstract important results from a vast amount of sequencing output. For example, assembly and alignment is the key procedure to match a read into its real location in the genome. See [16] for computational resources like clouding computing and [17, 18] for sequence-specific analysis and integrative approach.

Part II. Platform comparisons

As of July 2011, 6 sequencing platforms are commercially available. Table 1 summarizes distinctive attributes of some sequencing platforms. See [13] for the most updated comparisons in terms of technical aspects.

doi: 10.5415/apallergy.2011.1.2.93

Each sequencing platform has its advantages and deficiencies. It is sometimes recommended to combine data from different platforms to overcome limitation and maximize efficiency [19].

Part III. Applications

Box 1. Sequencing Application

Depending on input materials: Genome sequencing, RNA-seq (transcriptome, exome), ChIP-seq (methylome, transcript-DNA binding), small-RNA-seq (miRNA, piRNA, siRNAs), and ClIP-seq (transcript-RNA binding) [20].

Depending on purpose: *de novo* sequencing, (targeted) re-sequencing for variants and structural variants detection, transcriptome analysis, epigenetic changes and methylation pattern, Metagenomics including Microbial diversity and Paleogenomics, and so on.

Medical interests mostly focus on re-sequencing to find variants linked with diseases or specific phenotype. Some large-scale projects are listed below.

- The Cancer Genome Atlas project was launched by the National Cancer Institute and the National Human Genome Research Institute to provide genetic underpinning of cancer by extensive sequencing. Data is open to the research community. Recently it released detailed ovarian cancer data confirming that the mutation of TP53, BRCA1 and BRCA2 are highly associated with ovarian cancer [21].

- The 1000 Genomes Project started to provide understanding of the human genome variants (SNPs, structural variants, and their haplotypes) from population-scale sequencing by international collaboration. In 2010, the project reported its pilot phase and the updated data are released monthly. For the details of the 1000 Genomes Project, refer to [22] presenting the overview of human genome sequence variation studies and the pilot phase of the project.

- Metagenomics of the Human Intestinal Tract project aims to link human health and intestinal microbiota. The human gut has the potential to diagnose the health of individuals and a large study was done defining the minimal gut metagenome of 124 European individuals [23].

- The International Rare Disease Research Consortium, launched in early 2011 aiming to diagnose rare diseases by 2020, announced €100-million (US $140-million) call for research proposals [24].

There are several next-generation studies based on new techniques. For example, a web-based genome-wide association study was done by one direct-to-consumer (DTC) company. In many cases, medical science is slow. Multicenter genetic study on Parkinson's diseases spanned 6 years [25]. Now it can be reproduced within about 1 year once the cohort was constructed from the customer database of the personal genetics company [26]. Of course, this study was approved by an external IRB. Another good example is integration of genome-sequencing analysis and social-network analysis [27]. This study is a convergence of the classical and the modern. Combining

Table 1. Comparison of sequencing platforms

	454 FLX +	HiSeq 2000	PacBio *RS*	Ion Torrent 316
Company	Roche (USA)	Illumina (USA)	Pacific Biosciences (USA)	Life Technologies (USA)
Sequencing method	Synthesis (pyrosequencing)	Synthesis (cyclic reversible terminator)	Realtime sequencing	Synthesis (H^+ detection) on the chip
Amplification	emPCR	BridgePCR	None	emPCR
Run time	23 h	11 days (dual flow cell)	0.5 - 2 h	2 h
Reads Mb/run	1,000	540,000-600,000	5 - 10	>100
Reagent cost/run	$6,200	~$20,000	$110 - 900	$750
Reagent cost/Mb	$7	>$0.04	$11 - 180	<$7.5
Read length	500-1,000 (mode 700)	2*100 (paired-end reads)	860 - 1,100	>200
Primary errors	Indel	Substitution	CG deletion	Indel
Pros	Long read length	Highest throughput and lowest cost per Mb	Longest read length, no amplification error	Low cost per sample
Cons	High capital cost and high cost per Mb	High capital cost and high computation needs	Error rates, comparatively small outputs, high cost per Mb	High cost per Mb

 doi: 10.5415/apallergy.2011.1.2.93

sequencing and epidemiology uncovered a tuberculosis outbreak. The analysis comprised relatively very small bacterial sequencing, social-network questionnaires from 41 patients, and a linked-network model.

Apart from the academic challenges stated in previous paragraphs, business models such as the personal genome industry or DTC have been using sequencing. If you want to know yourself, DTC companies such as 23andMe, deCODEme and Navigenics will calculate a set of disease risks under $500 once you provide DNA samples, like saliva or a cheek swab. However, you might want to read J. Craig Venter's opinion for current limitations [28].

The human leukocyte antigen (HLA) polymorphism detection is a specific application of sequencing for diagnosis. Many genes located in the major histocom patibility complex (MHC) on chromosome 6 are related to immunological functions such as HLA expression. Clinically, matching HLA haplotypes is essential for further therapy of bone marrow transplanting. The comparison between the first generation Sanger sequencing and the second generation NGS was done from the study on HLA polymorphism [29]. Comprehensive comparison results from selected genes (Class I and II genes) showed that NGS outperforms the conventional Sanger method in terms of timing and cost but there are still considerable issues regarding notable sequencing alignment error and the need for intensive computational support. Another effort on sequencing MHC region showed that direct sequencing of the whole gene regions can reveal several variants in Tubulin beta of patients with acute myeloid leukaemia undergoing HLA-matched allogenic hematopoietic stem cell transplantation [30].

Summary

Massively parallel sequencing guides us to new phase of medical research and application. For example, detection of rare single nucleotide variants is a direct clinical usage of new sequencing technology. It also enables to diagnosing structural variants detection such as chromosome rearrangement or genome-wide variants from somatic diseases like cancer [31, 32]. There is no standard platform and researchers should consider the capacities of different platforms depending upon the aim of research. As the cost of sequencing is gradually decreasing and more integrative work is being developed, we will gain deeper understanding of human biology. A new paradigm of clinical study is about to begin.

ACKNOWLEDGEMENTS

The author gratefully acknowledges comments from Professor Yoon-Seok Chang, Seoul National University Bundang Hospital.

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government [NRF-2009-351-C00107].

REFERENCES

1. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1953;171:737-8.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005;437:376-80.
3. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008;9:387-402.
4. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet 2010;11:31-46.
5. National Human Genome Research Institute. DNA sequencing costs. Available from: www.genome.gov/sequencingcosts.
6. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 2010;42:30-5.
7. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 2010;42:790-3.
8. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. Hum Genet 2011;129:351-70.
9. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet 2008;24:133-41.
10. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol 2008;26:1135-45.
11. Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol 2009;25:195-203.
12. Illumina. *De novo* assembly with the genome analyzer. Available from: http://www.illumina.com/Documents/products/technotes/

doi: 10.5415/apallergy.2011.1.2.93

technote_denovo_assembly.pdf.
13. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour 2011. doi: 10.1111/j.1755-0998.2011.03024.x. [Epub ahead of print].
14. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res 2008;36:e105.
15. Teytelman L, Özaydin B, Zill O, Lefrançois P, Snyder M, Rine J, Eisen MB. Impact of chromatin structures on DNA processing for genomic analyses. PLoS One 2009;4:e6700.
16. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP. Computational solutions to large-scale data management and analysis. Nat Rev Genet 2010;11:647-57.
17. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat Methods 2009;6:S22-32.
18. Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet 2010;11:476-86.
19. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM. Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458:97-101.
20. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH; modENCODE Consortium. Unlocking the secrets of the genome. Nature 2009;459:927-30.
21. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011;474:609-15.
22. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.
23. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J; MetaHIT Consortium, Bork P, Ehrlich SD, Wang J. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59-65.
24. Abbott A. Rare-diseases project has global ambitions. Available from: http://www.nature.com/news/2011/110404/full/472017a.html.
25. Sidransky E, Nalls MA, Aasly JO, Aharon-Peretz J, Annesi G, Barbosa ER, Bar-Shira A, Berg D, Bras J, Brice A, Chen CM, Clark LN, Condroyer C, De Marco EV, Dürr A, Eblan MJ, Fahn S, Farrer MJ, Fung HC, Gan-Or Z, Gasser T, Gershoni-Baruch R, Giladi N, Griffith A, Gurevich T, Januario C, Kropp P, Lang AE, Lee-Chen GJ, Lesage S, Marder K, Mata IF, Mirelman A, Mitsui J, Mizuta I, Nicoletti G, Oliveira C, Ottman R, Orr-Urtreger A, Pereira LV, Quattrone A, Rogaeva E, Rolfs A, Rosenbaum H, Rozenberg R, Samii A, Samaddar T, Schulte C, Sharma M, Singleton A, Spitz M, Tan EK, Tayebi N, Toda T, Troiano AR, Tsuji S, Wittstock M, Wolfsberg TG, Wu YR, Zabetian CP, Zhao Y, Ziegler SG. Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. N Engl J Med 2009;361:1651-61.
26. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, Wojcicki A, Eriksson N. Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. PLoS Genet 2011;7:e1002141.
27. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, Rempel S, Moore R, Zhao Y, Holt R, Varhol R, Birol I, Lem M, Sharma MK, Elwood K, Jones SJ, Brinkman FS, Brunham RC, Tang P. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 2011;364:730-9.
28. Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. Nature 2009;461:724-6.
29. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, Walker R, Hsiao T, McLaughlin L, D'Arcy M, Gai X, Goodridge D, Sayer D, Monos D. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 2010;71:1033-42.
30. Proll J, Danzer M, Stabentheiner S, Niklas N, Hackl C, Hofer K, Atzmuller S, Hufnagl P, Gully C, Hauser H, Krieger O, Gabriel C. Sequence capture and next generation resequencing of the MHC region highlights potential transplantation determinants in HLA identical haematopoietic stem cell transplantation. DNA Res 2011. doi: 10.1093/dnares/dsr008. [Epub ahead of print].
31. Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C, Lindgren A, Kirby A, Liu S, Muddukrishna B, Ohsumi TK, Shen Y, Borowsky M, Daly MJ, Morton CC, Gusella JF. Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am J Hum Genet 2011;88:469-81.
32. Mardis ER. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med 2009;1:40.

 doi: 10.5415/apallergy.2011.1.2.93