Journal of Computer Science & Systems Biology - Open Access

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jcsb.1000053

JCSB/Vol.3 Issue 2

Challenges and Strategies for Next Generation Sequencing (NGS) Data Analysis

Vivek Thakur[1] and Rajeev Varshney[1,2]*

[1]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru-502324, India*
[2]*Generation Challenge Program (GCP), c/o CIMMYT, 06600 Mexico DF, Mexico*

Abstract

Next Generation Sequencing (NGS) technologies enable the resequencing of entire - genomes or the sampling of entire transcriptomes more efficiently and economically and with greater depth than ever before. Rather than sequencing individual genomes, now it is possible to sequence hundreds or even thousands of related genomes to sample genetic diversity within and between germplasm pools. Identifying and tracking genetic variation is now so efficient and precise that thousands of variants can be tracked within large populations. While NGS technologies have significant implications on crop breeding, optimization and availability of appropriate methodologies and tools to analyze, visualize and intrepret NGS data are still in their infancy. With an objective to help crop genetics and breeding community to utilize NGS data for crop improvement, Generation Challenge Programme, Mexico and ICRISAT, India organized an international workshop entitled "Next Generation Sequencing (NGS) Data Analysis" during Jul 21-23, 2009. More than 30 scientists from nine countries (USA, UK, France, Korea, Japan, Australia, Mexico, Philippines and India) participated in the workshop. The workshop had four technical sessions and two brainstorming sessions in which participants shared their experience/views on data generation and analysis by using NGS technologies for accelerating plant genome research to understand genome dynamics as well as to facilitate crop breeding. This article presents a report on this international workshop. All presentations made in this workshop have been made available online (http://www.icrisat.org/bt-publicdomain-ngs.htm).

Keywords: Next generation sequencing; Assembly; SNP discovery; RNA-seq

Abbreviations: ACPFG: Australian Centre for Plant Functional Genomics; BGI: Beijing Genome Institute; CIRAD: French International Agricultural Research Centre for Developing Countries; EST: Expressed Sequence Tag; Gbp: Giga base pairs; IT: Information Technology; NCGR: National Centre for Genome Resources; NBRI: National Botanical Research Institute; NIAS: National Institute of Agricultural Sciences; NRCPB: National Research Centre for Plant Biotechnology; PCA: Principal Component Analysis; PCR: Polymerase Chain Reaction; SCRI: Scottish Crops Research Institute; SNP: Single Nucleotide Polymorphism; SSR: Simple Sequence Repeat; TSL: The Sainsbury Laboratory; UMN: University of Minnesota

Introduction

This workshop was organized to discuss about strategies to analyze NGS data, especially for applications in crop genetics and breeding and to plan a pipeline for NGS data analysis with open-source softwares. The meeting had presentations and brainstorming sessions on key themes of NGS: i) Data generation and its assembly, ii) SNP discovery, iii) RNA-sequencing, and iv) NGS data management. The workshop opened with an introductory talk by Rajeev Varshney (ICRISAT, India) who presented technological overview and the current challenges (Varshney et al., 2009). A critical report on different topics (technical sessions) has been provided in this article.

Genomic resources, tools for assembly and visualization

David Studholme (TSL, UK) provided details on sequencing and analysis of some of the plant pathogens causing diseases of agronomic importance (eg. Banana Xanthomonas Wilt) (MacLean et al., 2009). He discussed identification of pathogenic sequence elements by comparing genome in question with those of its non-pathogenic close relatives. Research groups at TSL have successfully used Illumina (aka Solexa) for generating draft genomes of microbes with typical coverage of >=95%.

Dave Edwards (ACPFG, Australia) discussed emerging applications of paired-end libraries for identification of genes and promoter elements, orienting the genome fragments, BAC finishing, etc. Xavier Argout (CIRAD, France) reported use of deep sequencing data of short-RNA fraction to identify candidate miRNAs and to predict their targets.

Several other speakers briefed about recently generated NGS resources for plants/crops. TSL researchers have developed draft genomes of several microbial plant-pathogens, whereas at ACPFG a database of tags (TagDB) of several crops has been developed (http://flora.acpfg.com.au/tagdb/cgi-bin/index). Dave Marshall (SCRI, UK) informed that for potato sequencing project (http://www.potatogenome.net/index.php), BGI has completed >65X paired-end Solexa sequencing of heterozygous diploid line and its assembly is being optimized, whereas for barley, substantial progress towards physical map construction and genome sequencing has been made (Schulte et al., 2009). Xavier Argout (CIRAD, France) informed about generation of 454 reads of Oil-Palm short-RNAs for miRNAs discovery, and generation of RNA-seq data for Mediterranean/Tropical crops. Tilak Sharma (NRCPB, India) shared the availability of 454/FLX reads and their assembly for two varieties of pigeonpea.

Jimmy Woodward (NCGR, USA) discussed about Alpheus

***Corresponding author:** Rajeev Varshney, International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru-502324, India, E-mail: r.k.varshney@cgiar.org

Received February 13, 2010; **Accepted** April 08, 2010; **Published** April 08, 2010

Citation: Thakur V, Varshney R (2010) Challenges and Strategies for Next Generation Sequencing (NGS) Data Analysis. J Comput Sci Syst Biol 3: 040-042. doi:10.4172/jcsb.1000053

(Miller et al., 2008), a pipeline that can map/assemble data from multiple NGS platforms to genomic/transcriptomic reference and can call SNPs with high specificity. On the other hand, David Studholme informed use of open-source softwares, such as MAQ (Li et al., 2008), by TSL researchers for successful assembly of Illumina tags from microbes at very low error rate.

Regarding the visualization tools, Dave Edwards demonstrated a feature-rich visualizer for TagDB search results. Alpheus too comes along with visualizer where the alignment as well as variants are displayed. Among the open source visualizers, EagleView (http://bioinformatics.bc.edu/marthlab/EagleView) and AMOS/Hwakeye ([http://sourceforge.net/apps/mediawiki/amos/index.php?title=Hawkeye) were suggested to be quite useful by David Studholme. Dave Marshall demonstrated a new Solexa alignment viewer, Tablet (Milne et al., 2010), whose features are very promising. For viewing assembly of circular genomes, Cgview (http://wishart.biology.ualberta.ca/cgview/) was recommended.

SNP discovery

Single nucleotide polymorphisms (SNPs) or variants are important genomic resources which can be used in variety of analysis including molecular plant breeding. Dave Edwards described features of autoSNPdb (http://autosnpdb.qfab.org/), a SNP database developed specifically for major cereals. He described characterization of SNP abundance rate and SNP distribution in cereal genomes, the latter showed interesting patterns in flanking regions of coding sequences and SSRs. He also talked about a software, autoSNP (Barker et al., 2003), for SNP identification in 454/FLX or EST assembly.

Dan MacLean (TSL, UK) and Vivek Thakur (ICRISAT, India) both discussed strategies for accurate detection of polymorphisms from Illumina tags, however, in different species, namely microbial plant-pathogens and chickpea, respectively. Dan McLean presented the work on optimization of parameters of MAQ (Li et al., 2009) to identify highly accurate SNPs. Vivek Thankur presented comparison of multiple open-source tools for SNP identification from NGS data wherein he reported differences in tools' performance.

Jimmy Woodward discussed in details the *Medicago* HapMap project, being run in collaboration with Nevin Young (UMN, US). It aims to define the haplotype structure of *Medicago truncatula*, gather associated phenotypic data and examine in detail genomic role of legume-rhizobium symbiosis. In this project the objectives are to sequence 32-lines deeply to approximately 30X coverage for the purpose of variant detection.

RNA-sequencing

NGS is now increasingly being used for quantitative transcriptomics and identification of novel transcripts, and is considered to be an alternate of microarrays. Tsuyoshi Tanaka (NIAS, Japan) discussed use of Illumina sequencing in rice for detection of transcripts variants and validation of expression at those loci, where the transcript evidence from traditional approaches have failed. Both Jimmy Woodward and Mehar Asif (NBRI, India) discussed digital gene expression (DGE) studies however on different crops. Jimmy Woodward talked on homoeolog-specific digital gene expression in soybean through Illumina RNA-Sequencing. He reported the comparison of expression profile of high and low protein lines to identify a set of differentially expressed genes. In addition, an investigation of role of photoprotection in promoting polypoid diversification was discussed. Mehar Asif discussed the use of 454/FLX sequencing of *Gossypium* transcriptome to identify differentially expressed genes in germplasms characterized for traits like drought and fiber thickness. Her analysis involved comparison of count of aligned tag using R statistics, followed by PCA and Biplot analysis. Another interesting report was high (but imperfect) correlation between microarray and digital gene expression.

Costs, outsourcing, hardware and data management

As the NGS equipments requires major investment/maintenance, several speakers shared their experiences and opinions in individual presentations as well as in the brainstorming session. On the idea of setting-up own platform, they cautioned for following: 1) the rate of advancement of NGS technologies is very fast, 2) whether the sequencing needs are for long term and/or full usage of equipment is anticipated, 3) spectrum of sequencing needs is often wide and a single platform may fail to meet that.

In order to maximize the utility with NGS technologies, effective experiment design was advocated. The basic design issues, such as choice of tissues, should also be carefully selected. Another important design issue is to reduce the sequencing space, which can be achieved by transcriptome sequencing, sequencing PCR amplicons, and hybridization capture technique. One other possibility is using multiplexing of samples for sequencing.

It was further suggested to plan informatics before sequencing. As an alternative to purchasing extensive in-house IT infrastructure, it was advised to explore collaborations with national/international computing centers. Typically the image files and analyzed results are too large in size and their storage is a challenge, hence appropriate strategies were recommended such as removing unusable intermediate files.

Another aspect of data management, which was under focus, was quality control (QC), as unclean data is a frequent problem. Few recommendations in this regard were: Basic QC (eg. redundancy, distribution of quality scores and read lengths) on routine basis, trimming 15+ nt off the 3'-end for better mapping. Other recommendations for data management include: having own copy of data apart from that of service provider, consulting online user-groups/forums in addition to literature for experiment planning and analysis, versioning of analyzed results, use of Laboratory Information Management System (LIMS) or electronic notebooks (such as GIT, CVS, Subversion, etc.) and project management softwares like BaseCamp.

Few speakers shared their expertise on the hardware requirements for NGS. They were of the opinion that the specifications depend on the specific project, size of datasets, type of analysis and how often analysis will be performed. The hardware bottlenecks include memory (RAM), processing power and data storage. They suggested that the RAM bottlenecks could be overcome by splitting the task into smaller pieces, *e.g.* assembly of different chromosome arms separately.

Journal of Computer Science & Systems Biology - Open Access

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jcsb.1000053

JCSB/Vol.3 Issue 2

Summary

After an intensive discussions for three days, participants reached on consensus on several aspects: (i) the read- depth and base quality should be considered key for the detection of reliable SNPs, (ii) the number of SNPs one can expect is variable, dependent on species and strain, (iii) one can make SNP sets *ad infinitum*, (iv) after aligning the NGS data, it is imperative to visualize and assess the identified SNPs, (v) do not believe the aligners blindly, rather consult biologists on data analysis, (vi) instead of investing money in purchasing the machines, resources should be invested in the planning of experiments, data analysis and interpreting the results.

The workshop ended with a wrap-up session by Rajeev Varshney, who advocated for sharing NGS data with public and cautious use of NGS tools for analysis like SNP detection, where accuracy is very important. He anticipated increased use of NGS for miRNA discovery and digital gene expression in the years to come.

Acknowledgments

Authors would like to thank all workshop participants and resource persons and especially David Studholme (TSL, UK), David Edwards (UQ/ACPFG, Australia), Dan MacLean (TSL, UK), and Jimmy Woodward (NCGR, USA) for their valuable contributions as well as useful discussions. The workshop was funded by Generation Challenge Program (GCP, www.generationcp.org).

Competing Interests

None

References

1. Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19: 421-422. » CrossRef » PubMed » Google Scholar
2. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851-1858. » CrossRef » PubMed » Google Scholar
3. MacLean D, Jones JDG, Studholme DJ (2009) Application of 'next-generation' sequencing technologies to microbial genetics. Nat Rev Micro 7: 287-296. » CrossRef » PubMed » Google Scholar
4. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, et al. (2010) Tablet--next generation sequence assembly visualization. Bioinformatics 26: 401-402. » CrossRef » PubMed » Google Scholar
5. Miller NA, Kingsmore SF, Farmer A, Langley RJ, Mudge J, et al. (2008) Management of High-Throughput DNA Sequencing Projects: Alpheus. J Comput Sci Syst Biol 1: 132-148. » CrossRef » PubMed » Google Scholar
6. Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, et al., (2009) The International Barley Sequencing Consortium-at the threshold of efficient access to the barley genome. Plant Physiol 149: 142-147. » CrossRef » PubMed » Google Scholar
7. Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 27: 522-530. » CrossRef » PubMed » Google Scholar