Development of a high-performance bioinformatics pipeline for rice exome sequencing

H. Ichida,[*1] Y. Shirakawa,[*1] R. Morita,[*1] Y. Hayashi,[*1] and T. Abe[*1]

Heavy-ion beams cause DNA double strand breaks in a cell along with the beam direction and induce mutations, including insertions, deletions, inversions, and base substitutions in genome. Recent advances in massively parallel (aka "next generation") DNA sequencing technology enabled us to perform a comprehensive analysis of genomic mutations in base-pair resolution. We have been establishing heavy ion beam-irradiated rice progenies as a bioresource that are subjected to screening experiments to meet scientific and agricultural needs. Although among crops, rice has a relatively small genome size (approximately 400 million bases per haploid), most parts are filled with repetitive and junk elements and only 10% is predicted to encode proteins. We have developed a system to enrich rice exons, which are genomic regions transcribed and translated into proteins, and to determine their nucleotide sequences. By this strategy, called exome sequencing, it is possible to reduce sequencing costs significantly by eliminating meaningless sequences and focusing on the regions encoding proteins.

In the present study, we have developed a high-performance bioinformatics pipeline for analyzing rice exome sequencing results, and we identify the most significant mutations without the need for prior knowledge. The pipeline was implemented on Hokusai GreatWave, which is a parallel computing platform operated by Advanced Center for Computing and Communication, RIKEN. In the pipeline, raw sequencing reads (100–150 bp in length) are mapped to reference Nipponbare sequences using Burrows-Wheeler Aligner (BWA) software[1)], sorted, and realigned, and the data is stored in the standard BAM format. Programs with 3 distinct algorithms (GATK[2)], Pindel[3)], and Bedtools[4)]) are used to identify the mutations. We implemented a filter to remove duplicated and unreliable mutation candidates. The list of mutations were stored in tab-delimited text file and accessible from generic spreadsheet software. In parallel, the quality of raw reads were checked by FastQC[5)] program and the results were stored as an HTML-formatted report for review. The entire process was executed automatically and in parallel through a batch job controlling system. The deployment of an in-house bioinformatics pipeline enabled reliable comparison of results from different experiments (batches), since the pipeline is controlled by a versioning system; therefore, it is possible to reproduce exactly the same program and database versions any time.

To test the validity of variant calling and the following filtration, we analyzed the exome sequencing results from 8 individual rice mutants obtained via carbon- and neon-ion irradiations. As the result, there were 62,024 mutation candidates in GATK and Pindel outputs, but most of those mutations are common among the mutants and are likely to be intra-cultivar polymorphisms between our parental Nipponbare line and the reference sequence. Our newly implemented filtering program effectively removed such variations and identified 117 line-specific mutations, which are likely to be caused by heavy-ion beam irradiations. PCR and sequencing analysis showed that among the randomly-chosen 87 loci, 85 had the defined mutations. These results indicates that the process for narrowing down the candidates, which was implemented in the pipeline is reasonable and greatly improves the efficiency, as confirmed in subsequent experiments. This pipeline is a fundamental resource for rapid, comprehensive, and cost-effective genome-wide mutation screening and analysis of rice.

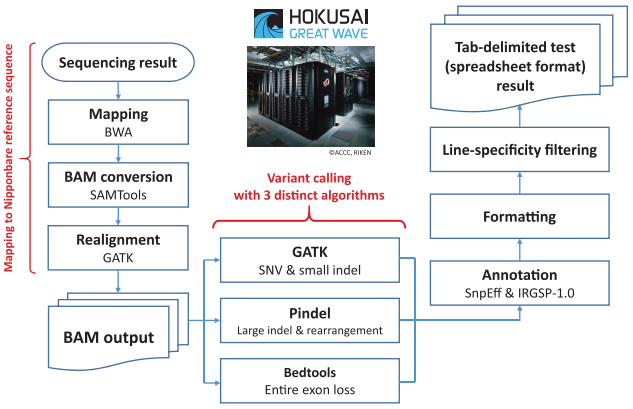

Fig. 1. A flow chart of the developed bioinformatics pipeline for exome sequencing datasets

References

1) H. Li and R. Durbin: Bioinformatics **25**, 1754–60 (2009).
2) A. McKenna et al.: Genome Res. **20**, 1297–1303 (2010).
3) K. Ye et al.: Bioinformatics **25**, 2865-28-71 (2009).
4) A. R. Quinlan and I. M. Hall: Bioinformatics **26**, 841-842 (2010).
5) S. Andrews: available from *http://www.bioinformatics.babraham.ac.uk/projects/fastqc.*

*1 RIKEN Nishina Center