G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2014;12(4):289-292
http://dx.doi.org/10.5808/GI.2014.12.4.289

APPLICATION NOTE

MAP: Mutation Arranger for Defining Phenotype-Related Single-Nucleotide Variant

In-Pyo Baek[1], Yong-Bok Jeong[2], Seung-Hyun Jung[1], Yeun-Jun Chung[1]*

[1]Department of Microbiology, Integrated Research Center for Genome Polymorphism (IRCGP), The Catholic University of Korea College of Medicine, Seoul 137-701, Korea,
[2]Quantum Technology, Seongnam 462-807, Korea

Next-generation sequencing (NGS) is widely used to identify the causative mutations underlying diverse human diseases, including cancers, which can be useful for discovering the diagnostic and therapeutic targets. Currently, a number of single-nucleotide variant (SNV)-calling algorithms are available; however, there is no tool for visualizing the recurrent and phenotype-specific mutations for general researchers. In this study, in order to support defining the recurrent mutations or phenotype-specific mutations from NGS data of a group of cancers with diverse phenotypes, we aimed to develop a user-friendly tool, named mutation arranger for defining phenotype-related SNV (MAP). MAP is a user-friendly program with multiple functions that supports the determination of recurrent or phenotype-specific mutations and provides graphic illustration images to the users. Its operation environment, the Microsoft Windows environment, enables more researchers who cannot operate Linux to define clinically meaningful mutations with NGS data from cancer cohorts.

Keywords: cancer, mutation, next-generation sequencing (NGS), sequence snalysis, single-nucleotide variant (SNV), software

Availability: The MAP installation package with an online manual is available at the website http://www.ircgp.com/MAP/index.html.

Introduction

It is widely accepted that mutation is the main causal factor for diverse tumorigenesis [1, 2]. The recent development of next-generation sequencing (NGS) technologies has revolutionized the speed and throughput of DNA sequencing, which facilitates the discovery of new driver mutations [3]. For example, the *ARID1A* gene is mutated in 83% of gastric cancers with microsatellite instability and the *SF3B1* gene is somatically mutated in 9.7% of chronic lymphocytic leukemia patients [4, 5]. In colorectal cancers, although the majority of recurrent mutations have been previously known, some novel mutations, such as the mutations in the *SOX9* and *FAM123B* genes, were also identified by NGS technology, which may have implications for colorectal tumorigenesis [6]. In melanoma, approximately 40% was found to have causal mutations affecting B-Raf function [7]. Along with the well-known driver mutations, the novel mutations identified by NGS will be useful for predicting the prognosis and molecular characterization of cancers [8, 9].

Recently, a number of mutation calling tools, such as VarScan [10], Genome Analysis ToolKit [11], and MuTect [12], have been developed to detect single-nucleotide variants (SNVs) or indels from NGS data. However, SNV calling from a single cancer case is not the final step for defining clinically meaningful mutations. To define the pathogenic SNVs in human cancers, a commonly used approach is to identify the phenotype-associated recurrent mutations in the study group. Due to the data size burden of NGS output, profiling the mutations associated with a certain phenotype or prognosis in a study group by a researcher who does not have a bioinformatics background or facilities is very difficult. Currently, there is no user-friendly tool supporting methods for the detection of recurrent or phenotype-specific mutations.

In this study, in order to support the definition of recurrent mutations or phenotype-specific mutations from

Received October 16, 2014; Revised November 16, 2014; Accepted November 20, 2014

***Corresponding author:** Tel: +82-2-2258-7343, Fax: +82-2-537-0572, E-mail: yejun@catholic.ac.kr

This is 2014 KOBIC best paper awarded.

NGS data of a group of cancers with diverse phenotypes, we aimed to develop a user-friendly tool, named mutation arranger for defining phenotype-related SNV (MAP).

Description

MAP software was designed as a standalone program, compatible in Microsoft Windows environments with a user-friendly graphic interface, and compiled codes of MAP can be easily installed. Both Variant Call Format [13] and its annotation files, such as ANNOVAR output [14], can be used as input files for MAP. When a user loads the input file, called Sample Descriptor file (GSF), MAP software displays the sample set information and filter options (Fig. 1). Once data are loaded, the user can filter the data using the 'Analysis Options' tab in MAP software based on annotation information. By using the sample set and analysis options, MAP software generates the summary metrics for multiple samples. Summary metrics include the mutation status for each sample, total mutation frequency, p-values for differences between phenotype groups based on clinical information, genomic position, reference/observed sequence, and additional information, such as read depth by default. MAP software also displays more detailed information in summary metrics based on the user-selected 'Analysis Options.' In the 'Analysis Result' tab, MAP software provides the frequency- or p-value-based filter function. This functionality supports detection of the recurrent mutations in study subjects and identification of the phenotype-specific mutation that occurs significantly in a certain phenotypic subclass. Then, MAP software represents the user-selected mutations graphically in the 'Visualization' tab. To sort the variants by genomic position, p-value, or frequency, users can use the sort function in the 'Analysis Result' tab. Details of the program download, installation, and analysis procedures are available in the user manual.

The major steps of MAP are as follows.

- Generating mutation summary metrics in study subjects
- Determination of recurrent mutations by frequency
- Determination of phenotype-specific mutations by association test
- Graphical illustration of user-selected mutations

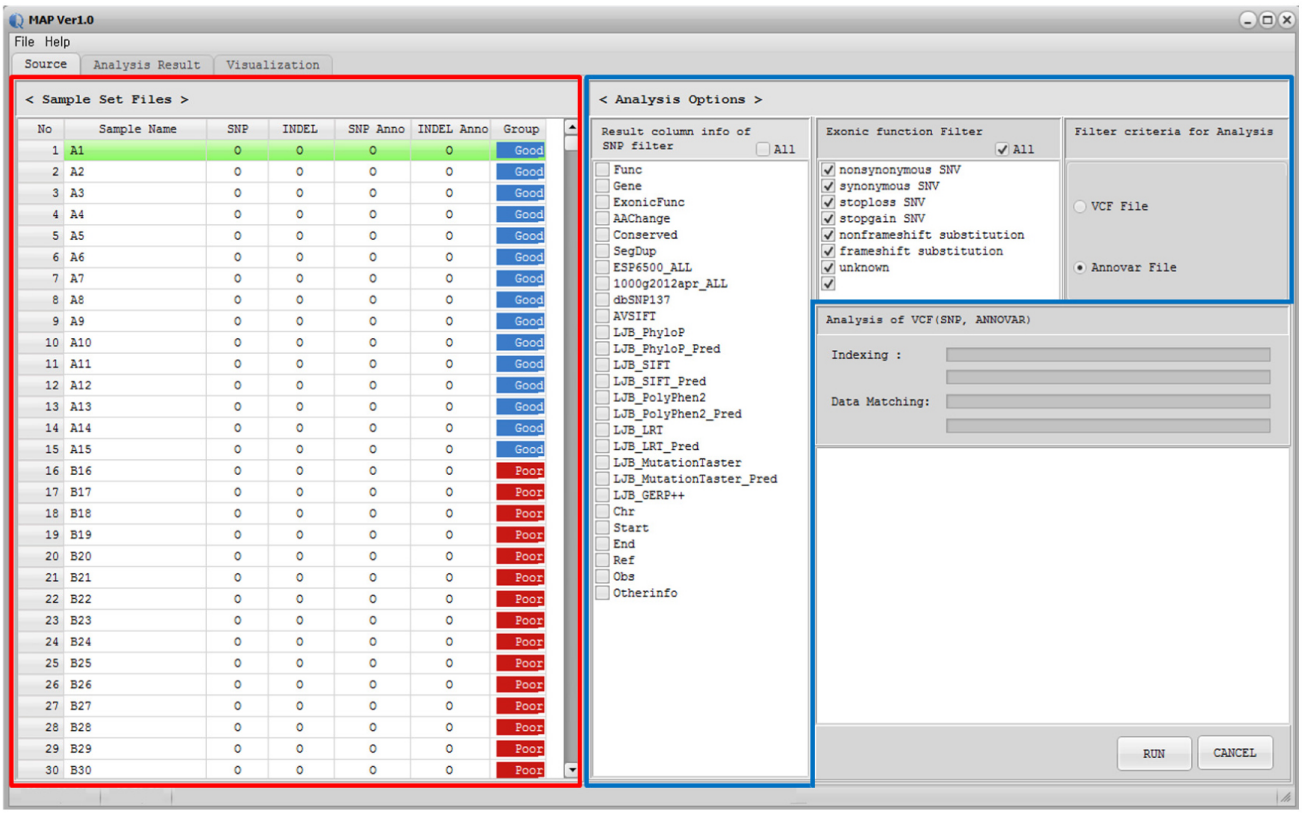

Fig. 1. Screenshot of source tab in MAP software. Once data are loaded, the Sample Set Files panel and Analysis Options panel are displayed in the Source tab. In the Sample Set Files tab, MAP software displays the input file status and phenotype information. MAP software supports the filter function in the Analysis Options tab based on annotation information. Examples of the Sample Set Files tab and Analysis Options tab are shown in red and blue boxes, respectively. MAP, mutation arranger for defining phenotype-related single-nucleotide variant.

Prerequisites

Mutation calling files from Genome Analysis ToolKit [11] can be used directly as input files for 'Sample Descriptor.' Alternatively, manually prepared standard Variant Call Format [13] is also available. Annotation information files, such as ANNOVAR outputs, can be used as input files for Sample Descriptor. GSF is a comma-separated values text file that describes the input files, such as GATK output files and ANNOVAR output files, and phenotype information. MAP imports this Sample Descriptor file as the input file. Details are available in the user manual.

Generation of summary metrics

When users load the input file, MAP software displays the sample set information and filter options in the 'Source' tab. If any file is not prepared properly, MAP software informs the user in the 'Sample Set Files' tab. After data are loaded, the user can filter the data using the 'Analysis Option' tab based on annotation information. MAP software supports eight filter options for exonic function. By using the mutation calls of each sample, MAP generates the summary metrics (Fig. 2A). Based on user-selected filters in the 'Analysis Options' tab, summary metrics represent additional annotation information, such as gene name, function, known variant, and amino acid change. The additional annotations can be hidden using the SNP filter option in the 'Analysis Option' tab. These mutation metrics are exported into a comma-separated values text file for further investigation.

Determination of recurrent or phenotype-specific mutations

In the 'Analysis Result' tab, MAP software defines the recurrent mutations in study subjects by frequency. The frequency threshold can be determined arbitrarily by the user in the 'Frequency Filter' function. MAP also defines the phenotype-specific mutations using the mutation metrics and phenotype information. Phenotype-specific mutations

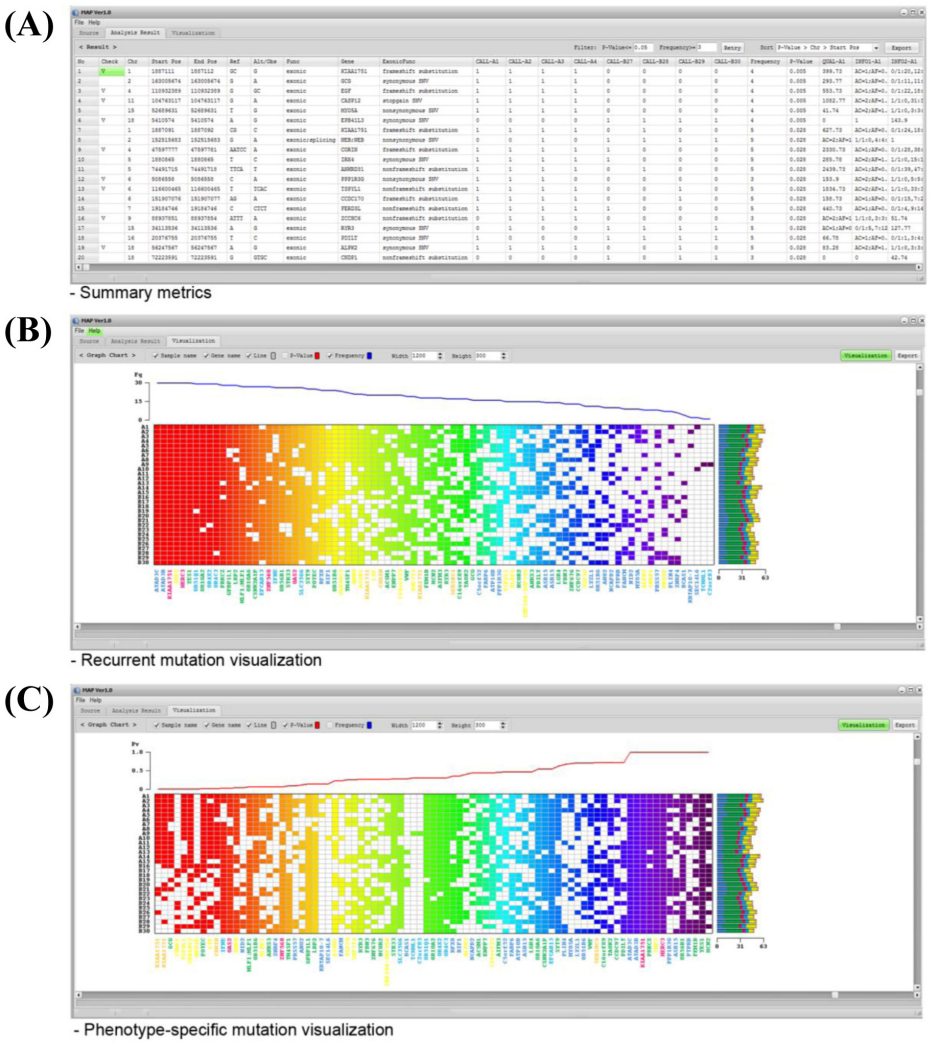

Fig. 2. Examples of MAP software output. MAP software generates the summary metrics and graphical illustration of user-selected mutations. (A) Example of summary metrics. Summary metrics include mutation status for each sample, total mutation frequency, p-values for differences between phenotype groups based on clinical information, genomic position, reference/observed sequence, and additional information, such as read depth by default. (B, C) Examples of frequency (B)- and p-value (C)-based visualization, respectively. The left and right of the heat-map represent the sample name and mutational spectrum, respectively. The bottom of the heat-map represents the gene name by color for the mutational spectrum. Frequency or p-value plots are represented on top of the heat-map. MAP, mutation arranger for defining phenotype-related single-nucleotide variant.

that occur more frequently in one of two clinical conditions (e.g., drug-responder vs. non-responder) are determined by chi-square test under the user-defined significance level.

Visualization

The graphic user interface is implemented in the software for users to easily handle large-scale data or access the analysis result (Fig. 2B and 2C). The recurrent or phenotype-specific mutations can also be demonstrated visually in heat-map style. Furthermore, MAP software can support the visualization for additional information at a time. For example, the left and right of the heat-map represent the sample name and mutational spectrum, respectively. The bottom of the heat-map represents the gene name by color for mutational spectrum. Frequency or p-value plots are represented on top of the heat-map (Fig. 2B and 2C). This graphical result can be exported into an image file for further investigation.

Conclusion

MAP is a user-friendly program with multiple functions that supports the determination of recurrent or phenotype-specific mutations and provides graphic illustration images to the users. Its operation environment, Microsoft Windows, enables more researchers who cannot operate Linux to define clinically meaningful mutations with NGS data from cancer cohorts.

Acknowledgments

This research was supported by a grant from the KRIBB Research Initiative Program. This study was supported by a grant from the Ministry for Health, Welfare and Family Affairs (A120175) and from the Cancer Evolution Research Center project (2012 R1A5A2047939).

References

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;144:646-674.
2. Stadler ZK, Schrader KA, Vijai J, Robson ME, Offit K. Cancer genomics and inherited risk. *J Clin Oncol* 2014;32:687-698.
3. Kahvejian A, Quackenbush J, Thompson JF. What would you do if you could sequence everything? *Nat Biotechnol* 2008;26:1125-1133.
4. Wang K, Kan J, Yuen ST, Shi ST, Chu KM, Law S, *et al*. Exome sequencing identifies frequent mutation of *ARID1A* in molecular subtypes of gastric cancer. *Nat Genet* 2011;43:1219-1223.
5. Quesada V, Conde L, Villamor N, Ordonez GR, Jares P, Bassaganyas L, *et al*. Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nat Genet* 2012;44:47-52.
6. Kim TM, Lee SH, Chung YJ. Clinical applications of next-generation sequencing in colorectal cancers. *World J Gastroenterol* 2013;19:6784-6793.
7. Davies MA, Samuels Y. Analysis of the genome to personalize therapy for melanoma. *Oncogene* 2010;29:5545-5555.
8. Kovtun IV, Cheville JC, Murphy SJ, Johnson SH, Zarei S, Kosari F, *et al*. Lineage relationship of Gleason patterns in Gleason score 7 prostate cancer. *Cancer Res* 2013;73:3275-3284.
9. Javle M, Rashid A, Churi C, Kar S, Zuo M, Eterovic AK, *et al*. Molecular characterization of gallbladder cancer using somatic mutation profiling. *Hum Pathol* 2014;45:701-708.
10. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, *et al*. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 2009;25:2283-2285.
11. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, *et al*. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 2011;43:491-498.
12. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, *et al*. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 2013;31:213-219.
13. 1000 Genomes. VCF (variant call format) version 4.0. 1000 Genomes, 2011. Accessed 2014 Nov 1. Available from: http://www.1000genomes.org/wiki/Analysis/vcf4.0.
14. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.