G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2017;15(4):178-182
https://doi.org/10.5808/GI.2017.15.4.178

APPLICATION NOTE

IVAG: An Integrative Visualization Application for Various Types of Genomic Data Based on R-Shiny and the Docker Platform

Tae-Rim Lee[§], Jin Mo Ahn[§], Gyuhee Kim[§], Sangsoo Kim*

Department of Bioinformatics and Life Science, Soongsil University, Seoul 06978, Korea

Next-generation sequencing (NGS) technology has become a trend in the genomics research area. There are many software programs and automated pipelines to analyze NGS data, which can ease the pain for traditional scientists who are not familiar with computer programming. However, downstream analyses, such as finding differentially expressed genes or visualizing linkage disequilibrium maps and genome-wide association study (GWAS) data, still remain a challenge. Here, we introduce a dockerized web application written in R using the Shiny platform to visualize pre-analyzed RNA sequencing and GWAS data. In addition, we have integrated a genome browser based on the JBrowse platform and an automated intermediate parsing process required for custom track construction, so that users can easily build and navigate their personal genome tracks with in-house datasets. This application will help scientists perform series of downstream analyses and obtain a more integrative understanding about various types of genomic data by interactively visualizing them with customizable options.

Keywords: docker, genome browser, genome-wide association study, RNA sequencing, Shiny, visualization

Availability: A docker image of IVAG can be downloaded at https://hub.docker.com/r/leetaerim/ivag/. Pre-processed example input data and the manual file are available at https://github.com/jmoa/IVAG.

Introduction

Since its advent, high throughput next-generation sequencing (NGS) technology has revolutionized the genomics research area, including transcriptome analysis and genome-wide association studies (GWASs) taking advantage of accelerated sequencing speed with reduced cost [1, 2]. Even though many bioinformatics software programs have been developed to handle and analyze the massive data generated from NGS, most of them are based on a command-line interface and require quite a high level of computational power [3], which creates a high barrier for wet lab biologists to enter into this field. Thanks to web-based analysis platforms, including Galaxy [4] and BIOEXPRESS [5], this barrier has been lowered. However, there still are problems. Downstream analyses, such as finding differentially expressed genes (DEGs), conducting Gene Ontology (GO) enrichment analysis, calculating linkage disequilibrium (LD), annotating gene information into GWAS results, and finally visualizing the resulting data, still require significant computer programming skills.

In this study, we present a dockerized application, IVAG. It provides a user-friendly web interface in which all downstream analyses mentioned above can be carried out without any programming knowledge. Detailed parameters for each analysis step can be adjusted with simple click-and-drag operation. IVAG interactively outputs publication-quality plots in response to the given parameters, and all of these plots can be downloaded. Also, a variety of data types, ranging from RNA sequencing (RNA-seq) and GWAS results to sequence read alignments, gene annotation, variant call information, and peak information, can be uploaded into the embedded genome browser and then

Received October 30, 2017; Accepted November 17, 2017

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

[§]These authors contributed equally to this work.

visualized together to help users gain greater integrative insights into their data. Furthermore, IVAG is lightweight, allowing it to be deployed on a desktop computer, as well as a server application.

Methods

IVAG is mostly written in the R programming language [6] and dockerized [7] with all required dependencies to avoid compatibility issues. It uses the Shiny package [8] to build a user-friendly web interface and several other packages to analyze and visualize RNA-seq and GWAS data (Supplementary Table 1). VCFtools [9] and PLINK (v1.90b4.6) [10] were used for the LD analysis. The JBrowse platform [11] was integrated into IVAG, and all intermediate steps required for custom track construction were automated using a custom BASH script. Three publicly available plugins (Supplementary Table 2), with slight modification, were incorporated into the genome browser to build GWAS, GC content, and browser extensible data (BED) tracks. Gene transfer format (GTF)-to-general feature format 3 (GFF3) format conversion was carried out with Cufflinks (2.2.1) [12]. Binary sequence alignment map (BAM) and variant call format (VCF) files were sorted with SAMtools [13]. Example data were prepared using publicly available RNA-seq and GWAS data (Supplementary Table 3).

Results

Workflow

Fig. 1 shows a graphical overview of the pre-analysis steps and the IVAG workflow. The blue line (Fig. 1A) denotes a schematic representation of the external pipelines required for RNA-seq and GWAS data. These parts are prerequisites for downstream analyses prior to IVAG analysis. The red box in the right panel (Fig. 1B) illustrates the IVAG workflow. The orange items are input files for IVAG, and their detailed formats are described in Supplementary Figs. 1–29.

RNA-seq

IVAG DEG analysis requires raw count RNA sequence data that can be generated using open source software, such as Htseq [14]. DEG analysis outputs a DEG results table generated with user-specified parameters based on the R Bioconductor package edgeR [15]. The output table consists of multiple columns, such as log2 fold-change, log2 count per million, associated p-value, and associated false discovery rate (Fig. 2A), and it can be visualized as a heatmap, a volcano plot, and a principal component analysis plot (Fig. 2B). The heatmap is generated using raw count data, which are converted to counts per million and normalized to a have row-based percentage value. The volcano plot is generated using log2 fold-change and the associated false discovery rate. Principal component analysis is generated using the raw count of each sample normalized to the log2 count per million. The heatmap and volcano plot can be interactively updated based on user-specified filtering criteria, such as false discovery rate or absolute fold-change. GO enrichment analysis uses the DEG analysis results table, GO annotation file, and GTF file. The DEG analysis results table can be generated using IVAG DEG analysis, or a pre-analyzed DEG analysis result can also be used. The GO annotation file consists of two columns: gene ID and GO category. A GTF file is needed to generate the gene length of each gene in the DEG analysis table. However, it can be omitted if a user wishes not to take gene length bias into account. IVAG GO enrichment analysis outputs over-represented and under-represented GO terms among DEGs (Fig. 2C) using the R Bioconductor package goseq [16]. It also shows a histogram of DEGs in each GO category based on its ontology: biological process, cellular component, and molecular function.

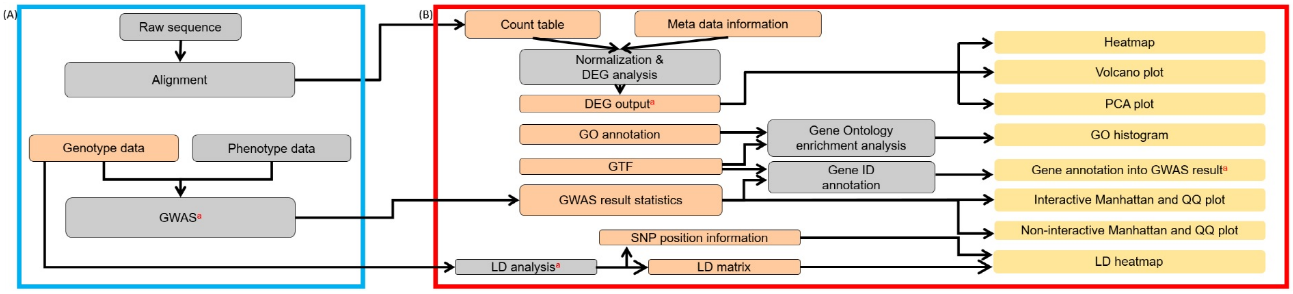

Fig. 1. Graphical overview of IVAG workflow. (A) External pre-calculation and automated pipelines for RNA sequencing and genome-wide association study (GWAS) analysis. (B) Schematic representation of the App pipeline. DEG, differentially expressed gene; GO, Gene Ontology; GTF, gene transfer format; SNP, single nucleotide polymorphism; LD, linkage disequilibrium; PCA, principal component analysis; QQ, quantile-quantile. [a]These data can be uploaded directly to the genome browser. The orange items are input files for IVAG, while the yellow ones are output files.

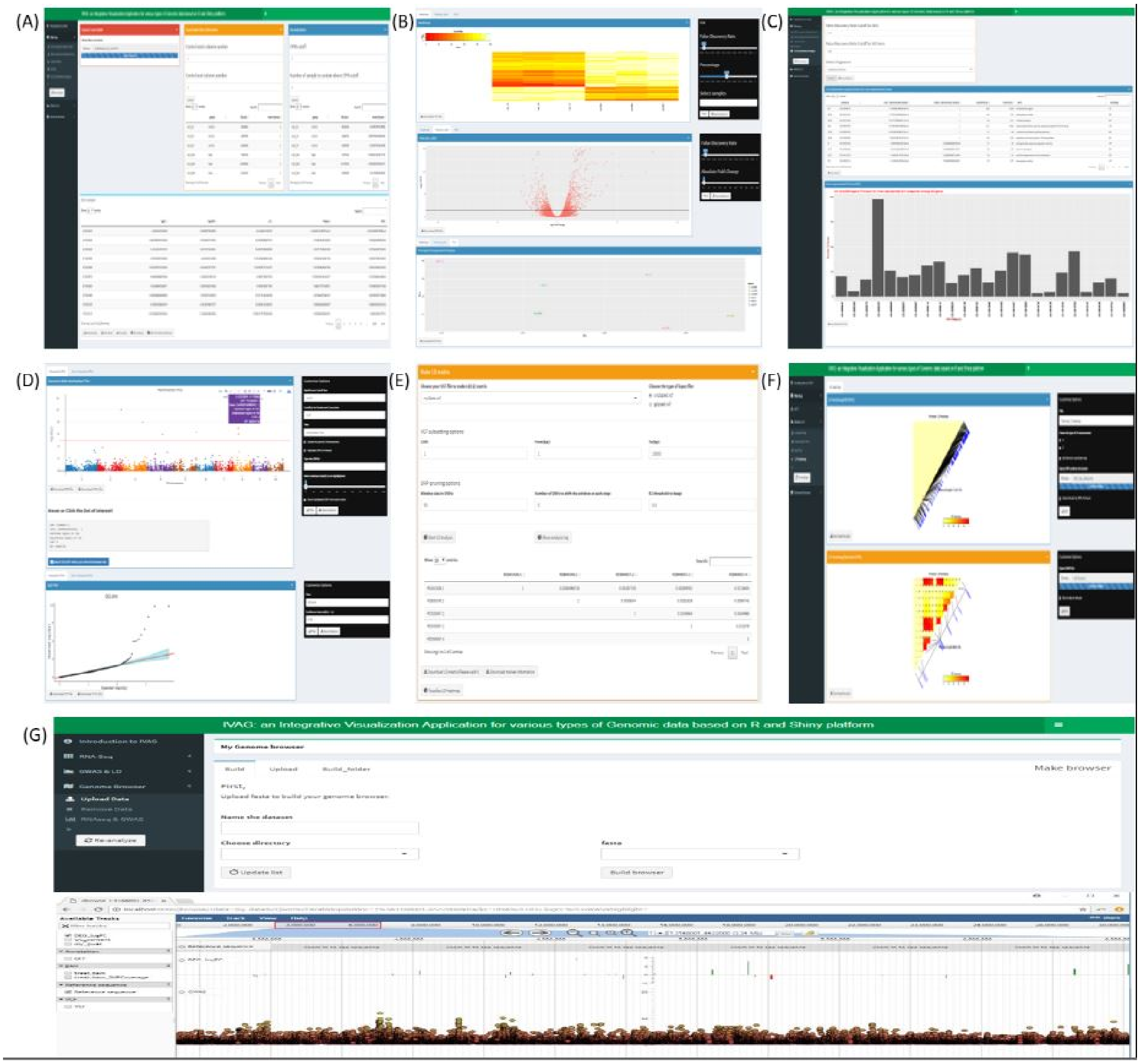

Fig. 2. Functions and results of IVAG. (A) Single-factor differential expression analysis. (B) Heatmap, volcano, and principal component analysis plot drawn with specified parameters. (C) Result of gene ontology enrichment analysis. Histogram shows how many differentially expressed genes are allocated to specific Gene Ontology categories. (D) Manhattan and quantile-quantile plots drawn with customizable options. (E) Linkage disequilibrium (LD) analysis generating LD matrix. (F) Pairwise LD heatmap. A group of single nucleotide polymorphisms of interest can be the subset. (G) Genome browser track with integrated view of differentially expressed gene and genome-wide association study results.

GWAS

Gene ID annotation requires a tab-separated GWAS result file comprising marker ID, chromosome ID, base position, and p-value columns in order, and a GTF file that contains strand and position information of genes. It returns a new GWAS result file in which gene, upstream, and downstream columns are added. Both GWAS result files, before and after this annotation, can be visualized in Manhattan and quantile-quantile plot with customizable options (Fig. 2D). One can see all information for a specific dot of one's interest if he clicks on the interactive plots. The LD analysis part is read in a VCF file with several detailed options to generate an LD matrix and a marker information file, which can be visualized in the LD heatmap (Fig. 2E and 2F).

Genome browser

Constructing a custom genome browser with a reference genome sequence is the first step. After selecting one of the genome browsers configured in IVAG, various types of genomic data, including GTF, GFF3, BAM, BED, BigWig, and VCF, can be uploaded and visualized all together (Fig. 2G). Also, this genome browser receives RNA-seq and GWAS results generated from IVAG as inputs.

Discussion

IVAG is an easy-to-use, web-based application with three modules, including RNA-seq, GWAS, and a genome browser. This application enables scientists with little computational proficiency to analyze and visualize their data easily. Some web applications provide similar functions for RNA-seq and GWAS, but they have some limitations. For example, DEApp [17] provides differential expression analysis using three different methods—edgeR, limma-voom, and DESeq2—while a heatmap or a principal component analysis plot is not provided. START [18] can output several plots, but it does not offer a GO enrichment analysis function. LocusTrack [19] can visualize GWAS data and annotate multiple tracks on them, but it is limited to only one species, human. Zbrowse [20] can be used over every species. However, because it focuses on plotting multiple GWAS results in one panel to enable users to detect genotype-environment interactions, the number of markers that can be plotted for one trait is limited to 5,000. IVAG is not limited to a specific organism or the number of markers [14]. Most importantly, IVAG combines a genome browser with analysis and visualization modules so that users can analyze, visualize, and finally navigate their entire data interactively in one application. We offer only two analysis and visualization modules now, but several more modules are in development and will be included in the near future.

ORCID: Tae-Rim Lee: http://orcid.org/0000-0003-0684-6552; Jin Mo Ahn: http://orcid.org/0000-0002-9073-9911; Gyuhee Kim: http://orcid.org/0000-0002-4054-979X; San gsoo Kim: http://orcid.org/0000-0001-9836-9823

Authors' contribution

Conceptualization: SK
Formal analysis: TRL, JMA, GK
Funding acquisition: SK
Writing – original draft: TRL, JMA, GK, SK
Writing – review & editing: TRL, SK

Acknowledgments

This work was funded by a program (PJ01167402) from the RDA (Rural Development Administration) and a program (NRF-2012M3A9D1054705) from the NRF (National Research Foundation of Korea) and computationally supported by a program (NRF-2010-0018156, NTIS1711048528) from KISTI GSDC.

Supplementary materials

Supplementary data including three tables and 29 figures can be found with this article online at http://www.genominfo.org/src/sm/gni-15-178-s001.pdf.

References

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* 2014; 30:418-426.
2. D'Agostino N, Tripodi P. NGS-based genotyping, high-throughput phenotyping and genome-wide association studies laid the foundations for next-generation breeding in horticultural crops. *Diversity* 2017;9:38.
3. Seemann T. Ten recommendations for creating usable bioinformatics command line software. *Gigascience* 2013;2:15.
4. Goecks J, Nekrutenko A, Taylor J; Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 2010;11:R86.
5. Korean Bioinformation Center. Bioexpress. Daejeon: Korean Bioinformation Center, 2017. Accessed 2017 Nov 1. Available from: https:// bioexpress.kobic.re.kr.
6. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017.
7. Merkel D. Docker: lightweight Linux containers for consistent development and deployment. *Linux J* 2014;2014:2.
8. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R. R package version 1.0.5. The Comprehensive R Archive Network, 2017. Accessed 2017 Nov 1. Available from: https://CRAN.R-project.org/package=shiny.
9. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al*. The variant call format and VCFtools. *Bioinformatics* 2011;27:2156-2158.
10. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, *et al*. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;81:559-575.
11. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, *et al*. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 2016;17:66.
12. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, *et al*. Transcript assembly and quantification by

RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; 28:511-515.

13. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078-2079.
14. Anders S, Pyl PT, Huber W. HTSeq: a Python framework to work with high-throughput sequencing data. *Bioinformatics* 2015;31:166-169.
15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139-140.
16. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* 2010;11:R14.
17. Li Y, Andrade J. DEApp: an interactive web interface for differential expression analysis of next generation sequence data. *Source Code Biol Med* 2017;12:2.
18. Nelson JW, Sklenar J, Barnes AP, Minnier J. The START App: a web-based RNAseq analysis and visualization resource. *Bioinformatics* 2017;33:447-449.
19. Cuellar-Partida G, Renteria ME, MacGregor S. LocusTrack: Integrated visualization of GWAS results and genomic annotation. *Source Code Biol Med* 2015;10:1.
20. Ziegler GR, Hartsock RH, Baxter I. Zbrowse: an interactive GWAS results browser. *PeerJ Comput Sci* 2015;1:e3.