Bioinformatics, 31(18), 2015, 3048–3050
doi: 10.1093/bioinformatics/btv324
Advance Access Publication Date: 21 May 2015
Applications Note

OXFORD

Genome analysis

PopGeV: a web-based large-scale population genome browser

Xinyi Shi[1,2,3,†], Jing Peng[1,4,†]Xiaohan Yu[5], Xiaohong Zhang[5], Dongye Li[4], Baohui Liu[1], Fanjiang Kong[1] and Xiaohui Yuan[1,*]

[1]The Key Lab of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, China, [2]University of Chinese Academy of Sciences, Beijing, China, [3]School of Computer Science and Technology, Heilongjiang University, Harbin, China, [4]College of Electronic and Information, Northeast Agricultural University, Harbin, China and [5]School of Computer Science and Technology, Wuhan University of Technology, Wuhan, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on March 16, 2015; revised on May 14, 2015; accepted on May 15, 2015

Abstract

Motivation: The development of high-throughput sequencing technology has made it possible for more and more researchers to use population sequencing data to mine genes associated with specific traits. However, the massive amounts of sequencing data have also brought new challenges to the researchers. The question of how to browse population genomic data in an easy and intuitive manner must be addressed. Web-based genome browsers allow user to conveniently view the results of genomic analyses, but heavy usage can reduce the response speed of the webpage, which limits its usefulness in the display of large-scale genome data. IndexedDB technology is a good solution to this problem; it supports web browsers and so creates local databases. In this way, data can be read from the local storage, achieving a smooth display of population genomic data.
Results: PopGeV has the following characteristics. First, it uses a new encoding method for compression of population SNP and INDEL data. IndexedDB technology is used to download the results to local storage so that users can browse the results smoothly even when the network traffic is heavy. Second, PopGeV identify similar genomic regions between two individuals based on SNP data. Population diversity indexes are calculated when comparing two populations. Third, user defined annotation information can be integrated for user-friendly mining of gene functions. Simulation shows that PopGeV can smoothly display analysis results of population genome containing over 500 individuals with 2 millions SNP data.
Availability and implementation: PopGeV is available at www.soyomics.com/popgev/
Contact: yuanxh@iga.ac.cn

1 Introduction

The development of high-throughput sequencing technology has decreased the cost of sequencing sharply, and more and more researchers are using population resequencing technology to study gene function and the evolution of species. Increases in the quantity of data collected in genomic studies have also posed great challenges to biomedical researchers. First, powerful computing resources and efficient algorithms are needed to extract meaningful results from population sequencing data. Second, user-friendly visualization tools for the results of genomic analysis would be of great use to biology researchers. Currently, several different genome visualization tools have been developed to address different specific needs, such as stand-alone program (Fiume *et al.*, 2012; Robinson *et al.*,

2011) and web-based browsers (Kent *et al.*, 2002; Kumagai *et al.*, 2013; Skinner *et al.*, 2009). As the quantity of genomic data collected in studies increases, population genome analysis tools are becoming more and more important. Considering the computation capacity, more and more users have chosen to perform data analysis on the cloud server, and the results of analysis can be returned in HTML format. This suggests web-based visualization tools may become even more popular in the future.

When using web-based browsers to display the results of population genome analyses, two issues must be resolved. First, the large amount of data transmitted will produce corresponding delays; second, web browsers are not as flexible as stand-alone programs in terms of user interaction. Here, HTML5 technology was used to develop a visualization tools for population genome analysis. Canvas and SVG (Scalable Vector Graphics) are used as the primary display technology, SVG is a way to achieve vector image display, making the visualization clearer than ever and improving the interactive performance. The IndexedDB local database technology reduces the number of interactions between the webpage and the server, so that the users can smoothly browse the analysis results. Simulation shows that this software can display genomic variants of more than 500 soybean genomes smoothly.

In terms of software features, in addition to basic functions such as displaying the SNP and INDEL variation in each individual, PopGeV also identify recombination sites within the population and analyze the genetic relationships between individuals. Population diversity indexes such as Fst, Tajima's D are calculated, and these can be used to show difference between two populations.

2 Methods

PopGeV places all data and programs on the server side; the user accesses the results of analysis through web browsers. On the server side, an APACHE2 webserver with MySQL database support must be installed. On the client end, only a web browser that supports HTML5 needs to be installed. In order to improve the performance, the server uses Apache Zip for compression of webpage; the use of Zip compression can greatly reduce the number of bytes in network transmission, significantly increasing speed of loading webpages. When user visits the website first time, the genomic data on web server will be downloaded to local disk, browser uses IndexedDB technology for data storage. When the user opens the page again, IndexedDB storage data are used to quickly restore the last operation status. In this way, PopGeV uses storage and display methods different from those of currently available genome browsers. This renders browsing of population genome data very smooth.

3 Results

3.1 SNP and IDNEL display

By reading the population variation files in VCF file format and corresponding annotation files, PopGeV can extract SNP and INDEL details and display them on a floating window. In order to reduce the amount of data transmitted, PopGeV uses hexadecimal encoding to compress the genotype of each SNP locus in the population. PopGeV can also read the annotation results generated by ANNOVAR (Wang *et al.*, 2010), and use different colors to mark the SNP and INDEL variations. When the user places the mouse over an SNP, a floating window appears showing the details of that variation.

3.2 Population comparisons

For population comparisons, the users usually want to know whether different individuals and populations are the same or similar at which genomic regions on the whole-genome scale. Here, the variation block (VB) algorithm proposed by Kim *et al.* (2014) was used to determine recombination blocks (RCB). Variation block can be used to analyze the genomic differences between individuals and between sub-populations, and to evaluate transitive relationships between genetic regions in genealogical analysis. PopGeV can also measure the differences between two populations using indicators such as Fst, Pi and Tajima's D as shown in Figure 1. Combining these indicators and information of recombination regions, the differences between population genomes can be clearly identified.

3.3 Annotations

Researchers often use population SNP data and traits to perform genome-wide association studies in order to identify SNP loci associated with traits. PopGeV uses Manhattan plots to show GWAS results. Gene annotation files are in the format used by public databases such as Phytozome (Goodstein *et al.*, 2012). PopGeV can also display customized information, such as QTL information and series of genes involved in a particular pathway. This customized information will help to accurately locate genes associated with traits.

3.4 Performance

To assess the software's processing capabilities, data on 500 soybean genomes, each containing 2 millions SNP and INDEL data, were simulated. When users visit the website first time, it will cost several minutes to retrieve data from web server. However, when users visit the website again, the webpage loading time will be less than one second even there are over 500 genomes.

4 Conclusion

The rapid growth of the amount of sequencing data produced by genomic studies has raised challenges for data analysis and visualization techniques. PopGeV uses the HTML5 IndexDB technology to store data locally. In this way, it exploits the convenience of a web browser and prevents frequent network communications, producing a smooth display of massive quantities of genetic data. This software

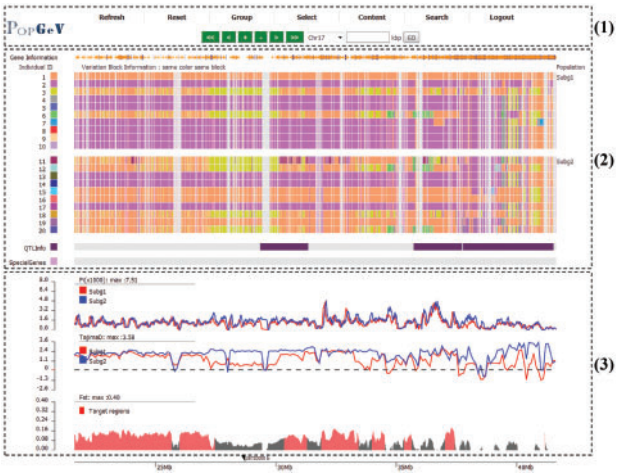

Fig. 1. Comparison between two groups. Block (1) is the menu for operating, block (2) shows the RCB information for each group; block (3) shows the group difference indexes

provides a platform that bioinformatics researchers and experimental biology investigators can use to share the results of their research.

Funding

"Hundred Talents Program" of Chinese Academy of Sciences.

Conflict of Interest: none declared.

References

Fiume,M. *et al*. (2012) Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res.*, **40**, W615–W621.

Goodstein,D.M. *et al*. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Kim,Y.H. *et al*. (2014) Variation block-based genomics method for crop plants. *BMC Genomics*, **15**, 477.

Kumagai,M. *et al*. (2013) TASUKE: a web-based visualization program for large-scale resequencing data. *Bioinformatics*, **29**, 1806–1808.

Robinson,J.T. *et al*. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Skinner,M.E. *et al*. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164–e164.