Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 7, July 2021: 6230- 6234

Research Article

Computational Analysis on Single Nucleotide Polymorphisms (SNPs) in Chromosome 6 of Human Reference Genome Using R Programming

S. Balamurugan
Ph. D. Research Scholar, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai
Email: sivabala76@gmail.com

Dr. S. Prasanna
Professor and Head, Department of Computer Applications, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Pallavaram, Chennai
Email: prasanna.scs@velsuniv.ac.in

ABSTRACT
Single Nucleotide Polymorphisms (SNPs) are point mutations in the DNA sequence represent a common form of genetic variation present all over the genome sequences in human, which makes everyone as unique. There are a remarkable number of SNPs on the human genome, estimated at over 10 million average. These are frequently associated with the development of genetic diversity and some may lead to genetic diseases. So computational descriptive analysis are required for focus on SNPs according to their potentially deleterious impacts to human health. With these requirements in mind, the present study is developed a web application on descriptive analysis of SNPs in chromosome 6 of human reference genome. The results display that how SNPs are occurred in chromosome 6 and this will be useful to the biologist for their further research on SNPs. The web application is developed using R programming along with Bioconductor packages. Later the App will be hosted for the user's further analysis.

Keywords: Human reference genome; Single Nucleotide Polymorphism (SNPs); R programming, Bioconductor

INTRODUCTION
In the fields, human medical genetics and population genetics, single-nucleotide polymorphisms (SNPs) play a major role and analysis on SNPs through genome-wide association studies (GWAS) reveals the evolutionary history and heritable risk for common diseases [1,2]. SNPs occur throughout the DNA sequence, particularly once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. Consequently, an individual genome differs from the reference genome at ~ 4.1 million to 5.0 million sites [3]. Databases like dbSNP, 1000 genome project, HapMap Project provide human genetic variants data which help us to predict disease risk and population spread. The study shows that the common variants are distributed mostly in the entire human population and only the rarer variants are restricted to a specific population [4]. Generally, the periodic DNA has 1.8 times higher SNP density than rest of the genome [5]. In this regard, the present study concentrates on the computational descriptive analysis of SNPs in chromosome 6 of human reference genome and develop a web application to describe the variants. The primary goal of the project was generally agreed to be the

development of a public resource of genetic variation to support the next generation of association studies relating genetic variation to diversity and disease.

The development of web application is done with Shiny server. Shiny is an open source R package of RStudio that is used to build interactive web applications with R. R is a language and environment for statistical computing and graphics. RStudio is a free and open-source integrated development environment (IDE) for R. The user-interface definition (UI) file called ui.R and a server script file called server.R files work together to create R shiny web application [6]. The source code ui.R is used to set-up what the user will actually see in the web app and also used to accept input from the user. Server.R does the computational R work and contains the instructions that your computer needs to build the app. The pursued SNP injection in chromosome 6 of human reference genome is done through the App helps the users to visualize and explore the presence of SNPs at ease.

MATERIALS AND METHODS

R programming language and RStudio is utilized as an interface (https://www.r-project.org). Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data [7,8]. Bioconductor packages along with Shiny framework (https://CRAN.R-project.org/package=shiny) are utilized for the development of App. Shiny apps have two components: a user-interface (ui) script and a server script. The ui script controls the layout and appearance of an application. In ui.R, functions like mainPanel (), sidebarPanel (), navbarPage (), tabsetPanel (), tabPanel (), selectInput (), submitButton (),actionButton (), verbatimTextOutput () and downloadButton () are used for laying out the user interface. The server.R script contains the instructions that the computer needs to build an application. In server.R, the functions like renderPlot (), renderPrint (), renderDataTable (), downloadHandler () and reactive () are used for the application development [6].

Retrieval of 6th chromosome sequence of Human reference genome and SNP data

The latest genome feature annotation and variation files available for the human GRCh38 assemblies in NCBI are used to retrieve chromosome 6 sequence of human reference genome using the bioconductor package. As like chromosome 6 of human reference genome, its corresponding SNPs location (SNPlocs) data are retrieved from NCBI through another bioconductor package [9]. Both data are stored as vector variable, which are useful for the retrieval and manipulation to infer the results. RSQLite embeds the SQLite database engine, providing a DBI-compliant interface. The DBI package specifies a common interface between R and database management systems (DBMS). sqldf () transparently sets up a database, imports the data frames into that database. These packages support in dealing data in R environment.

RESULTS AND DISCUSSION

Much of genetic variation in the human genome is in the form of SNPs which is the result of point mutations that produce single base-pair differences (substitutions or deletions) among chromosome sequences. There are many laboratories and computational approaches to finding single nucleotide polymorphisms (SNPs) within a genome, but all involve some form of comparative analysis of the same DNA segment from different individuals or from different haplotypes. For these studies, Next-generation sequencing projects (NGS) and Genome wide association (GWAS) studies have become an important means for understanding and discovering susceptibility genes for complex diseases and other genetic variation [5]. In this regard, the present study developed a web application on computational descriptive analysis of SNPs in chromosome 6 of human reference genome for easy access by the biologist. For the visualization, the screenshot of the developed application is shown in Figure 1. Generally, lots of human

genome sequences are available in the databases of NCBI, but it is difficult to retrieve exact sequences by all the biologist. In this regard, this application will help biologists to get genes and SNPs of chromosome 6 of human reference genome. Therefore, to show the results of application, chromosome 6 and its 3494 genes entries are displayed in Figure 2.

Initially, the SNPs of human reference genome GRCh38 is retrieved from NCBI dbSNP Build 151 and injected in the human reference genome in order to show how do the changes due to SNPs occur in an individual. While injecting SNPs, the possible nucleotide including their corresponding IUPAC ambiguity code are replaced [10]. The obtained SNP injected genome can be considered as an individual sequenced genome model. In future, if any genome of individual genome sequences is available, this application will be helpful to compare everyone with the reference genome sequences to analyze genetic variation and genetic disease risk factors. The comparative results of nucleotide frequencies are made between the 6^{th} chromosome of reference genome and its corresponding SNP injected reference genome and the results are displayed in the web application. To show the results, here the screenshot of nucleotide frequency of chromosome 6 of reference genome and its corresponding SNP injected genome in dot plots are shown in Figure 3. It helps biologist to visualize the replacement of nucleotides clearly and one can use this app to get SNPs information in chromosome 6 of human reference genome for their further research.

CONCLUSIONS

The present study is the development of a web application to do computational descriptive genome wide association studies (GWAS) on Single Nucleotide polymorphism (SNPs) in Human reference genome using R programming. Generally, SNPs are used to track the inheritance of disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease, but most SNPs have no effect on disease. At present the advancement in sequencing technologies like next generation sequencing (NGS) and genome-wide association studies (GWAS) of human genetic variation can inform understanding of common human diseases. These advancements made the greatest challenges of storing huge data and analyzing and interpreting these data to get sensible solutions for many challenges like identification of hereditary diseases. In the present preliminary study, SNPs in the 6^{th} chromosome of human reference genome to show how do the changes due to SNPs occur in an individual. While injecting SNPs, the probable nucleotide, their subsequent IUPAC ambiguity code are substituted. The obtained SNP injected chromosome 6 of reference genome can be considered as an individual sequenced genome model. The work will be continued to do the SNPs descriptive analysis for all chromosomes and will be published elsewhere. In future, if any genome of personal chromosome sequence is accessible, this application will be helpful to compare with the reference genome sequences to analyze genetic variation, if possible.

REFERENCES

1. Fareed M and Afzal M (2013), Single nucleotide polymorphism in genome-wide association of human population: A tool for broad spectrum service. The Egyptian Journal of Medical Human Genetics, 14: 123–134.
2. Gonzaga-Jauregui C, Lupski J R and Gibbs R A (2012), Human genome sequencing in health and disease. Annual Review of Medicine, 63: 35–61.

3. Madsen B E et.al., (2007), A periodic pattern of SNPs in the human genome. Genome Research, 17(10): 1414–1419.

4. The 1000 Genomes Project Consortium (2015), A global reference for human genetic variation. Nature, 526: 68–74.

5. Tuzun E et al., (2005), Fine-scale structural variation of the human genome. Nature Genetics, 37: 727–732.

6. Sivaprakasam B and Sadagopan P (2019), Development of an Interactive Web Application "Shiny App for Frequency Analysis on Homo sapiens Genome (SAFA-HsG)". Interdisciplinary Science: Computational Life Sciences 11, 723–729.

7. Pagès H, Aboyoun P, Gentleman R, DebRoy S (2017), Biostrings: string objects representing biological sequences, and matching algorithms. R package version, 2(44): 1.

8. Huber W, Carey VJ, Gentleman R et al (2015), Orchestrating high throughput genomic analysis with Bioconductor. Nature Methods, 12: 115–121.

9. Pagès H (2018), SNPlocs.Hsapiens.dbSNP151.GRCh38: SNP locations for Homo sapiens (dbSNP Build 151). R package version 0.99.20.

10. Johnson A D (2010), An extended IUPAC nomenclature code for polymorphic nucleic acids. Bioinformatics, 26: 1386–1389.

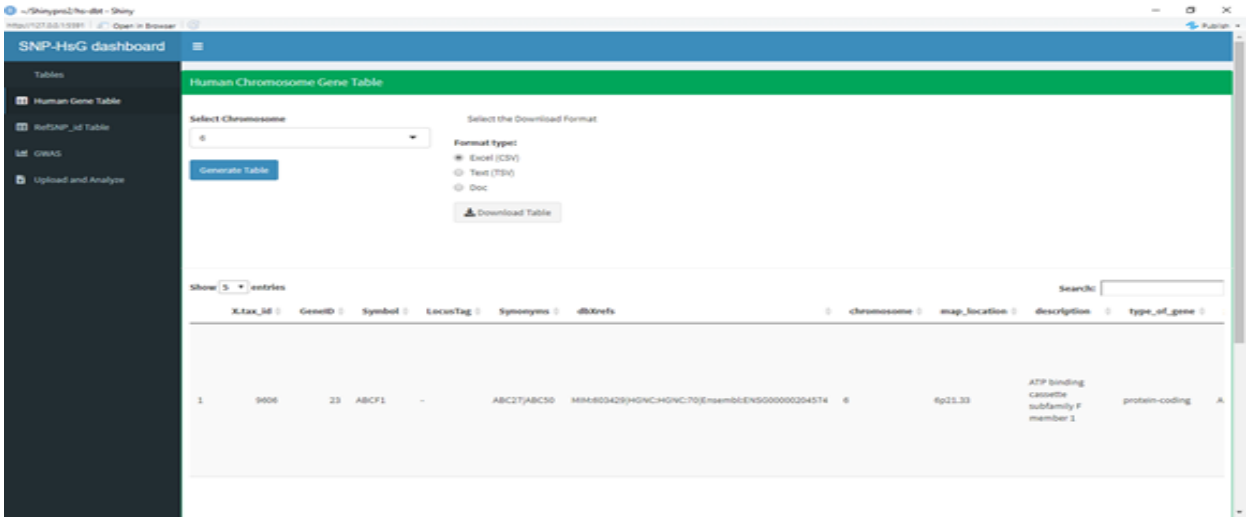

Figure 1: The screenshot of homepage of the application with selected chromosome 6.

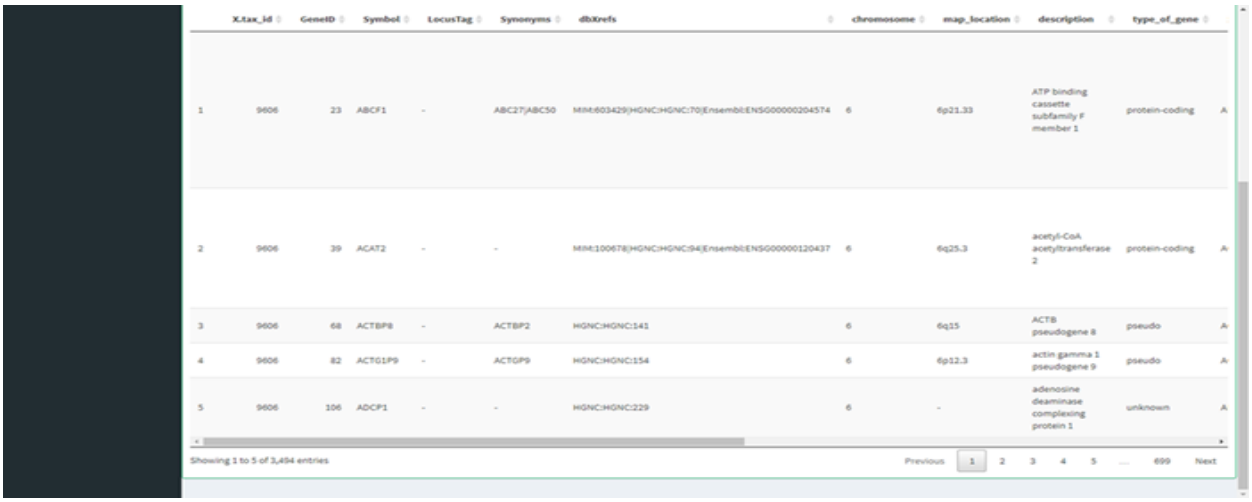

Figure 2: The screenshot displays genes of chromosome 6.

Figure 3: The screenshot of descriptive analysis of nucleotide frequencies of reference genome and SNPs injected genome in dot plot.