Gene Reports 23 (2021) 101033

Contents lists available at ScienceDirect

Gene Reports

journal homepage: www.elsevier.com/locate/genrep

Development of shiny dashboard application for "genome-wide association study on analysis of SNPs injected in *Homo sapiens* genome (snips-HsG)"

Balamurugan Sivaprakasam *, Prasanna Sadagopan

Department of Computer Applications, Vel's Institute of Science, Technology and Advanced Studies, Pallavaram, Chennai 600 117, Tamil Nadu, India

ARTICLE INFO

Keywords:
Human reference genome
SNPs injected genome
GC-content
R and bioconductor
Isochore family
Shiny dashboard

ABSTRACT

Human genome information in databases is growing exponentially, and in future, the collection of these data will exceed 5v's (variety, volume, veracity, value, and velocity). Much of genetic variation exist among genomes of human individuals is mostly in the form of SNPs. Therefore, studying SNPs among individuals is important to analyze genetic risk and diversity. In this regard, the present preliminary study has developed an application on human reference genome before and after SNPs injection using R and hosted in https://snip-hsg.shinyapps.io/home/. In it, initially the comparison of nucleotide frequency between two copies of genomes are made and showed the nucleotide bias due to the SNPs sites. Secondly, genes of two copies of genomes are grouped into five different isochore families, L1 (<37% GC), L2 (37–42%), H1 (42–47%), H2 (47–52%), H3 (>52%) based on GC %. The results through histograms showed that there are differences in the GC% distribution among chromosomes and this study may help the biologist to examine the nucleotide frequency including GC% and isochore family differences comparatively between genomes. Thirdly, in the statistical analysis, the Shapiro–Wilk test provided the evidence that the data have non-normal distributions. Therefore, the values are considered for non-parametric test to validate the significance relationship between these two variables. From the Wilcoxon rank sum test and Chi-square test, the significance difference between gene sets of two copies of genome are projected that the two samples of measurements come from different distributions. The present descriptive and inferential analysis on SNPs in human genome may be useful for the biologists to access SNPs comfortably. One can use this application for comparison, easy access, downloading and visualizing.

1. Introduction

After completion of human genome project (International Human Genome Sequencing Consortium, 2001; International Human Genome Sequencing Consortium, 2004), it is understood that the prediction of possible genetic risk factors for inherited diseases is one of the major challenges in human medical genetics and the study on diversity of individuals (human genetic variants) is another major task in population genetics (Makalowski, 2001; Gonzaga-Jauregui et al., 2012). In these fields of research, single-nucleotide polymorphisms (SNPs) play a major role and analysis of SNPs through genome-wide association studies (GWAS) reveals the evolutionary history and heritable risk for common diseases (Fareed and Afzal, 2013). Similarly, the studies (Tuzun et al., 2005; The 1000 Genomes Project Consortium, 2015) revealed that the epigenomic approaches that encompass functional annotation of regulatory elements to focus on the disease risk-associated SNPs also occur in non-coding regions of the genome. Therefore, 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 (Madsen et al., 2007). Databases like dbSNP, 1000 genome project, HapMap Project provide human genetic variants data which help us to predict disease risk and population spread (The 1000 Genomes Project Consortium, 2015; Sherry et al., 2001). In this regard, the present study focuses on injection of SNPs from the build dbSNP151 in human reference genome sequences. Both the reference and SNPs injected genomes are used for descriptive and inferential analysis comparatively on their nucleotide

Abbreviations: HGP, Human Genome Project; PGP, Personal Genome Project; SNPs, Single-Nucleotide Polymorphisms; GWAS, Genome-Wide Association Studies; GC, Guanine and Cytosine; GC%, Percent GC content; IDE, Integrated Development Environment; UI, User-Interface; SNPlocs, SNP locations.

* Corresponding author at: Department Computer Applications, Vel's Institute of Science, Technology and Advanced Studies (VISTAS), Pallavaram, Chennai 600 117, India.

E-mail addresses: sivabala76@gmail.com (B. Sivaprakasam), prasanna.scs@velsuniv.ac.in (P. Sadagopan).

https://doi.org/10.1016/j.genrep.2021.101033
Received 9 October 2020; Received in revised form 15 January 2021; Accepted 21 January 2021
Available online 3 February 2021

frequency.

Guanine and Cytosine (GC) content in human genome is one of the important genomic features that is possibly related from a functional point of view, like gene density, methylation rate etc. (Romiguier et al., 2010; Piovesan et al., 2019). The long genomic segments that are homogeneous in their GC composition are called isochores. The human genome was described as a mosaic of isochores of alternating low and high GC contents. Human isochores have been classified into five families, L1 (<37% GC), L2 (37–42%), H1 (42–47%), H2 (47–52%), H3 (>52%) (Bernardi, 2000). The isochore family classification shows that normally in a human genome, the GC content is in the range of 35% to 60% (Romiguier et al., 2010). This GC content in a coding gene of human genome, if found less than 35% or more than 60%, can be indication of some genetic defects. This also signifies that important coding regions (gene-rich) are elevated in GC-content, which are more stable and resistant to mutation compared to gene-poor regions (Vinogradov, 2003). Nevertheless, it is still unsure whether this came about through random mutation or through a pattern of selection. There is also discussion on methods used to find out whether the relationship between GC-content and coding region are accurate and unbiased (Sémon et al., 2005).

Generally, the segmentation algorithms are used for identifying isochore families in a genomic sequence (Zhang et al., 2005; Elhaik et al., 2010a; Elhaik et al., 2010b; Arhondakis et al., 2020). There are many web-based tools available to identify isochore families like, Emboss isochore (Madeira et al., 2019), IsoFinder (Oliver et al., 2004) and IsoXpressor (Ayad et al., 2020). However, only the person who has knowledge on biology and computer science can access those resources. In this context, to our best knowledge, there has been no web application available for the classification of genes based on GC% into isochore families for the chromosomes of human reference and SNPs injected genome. With this lacuna in mind, we have developed a web application snips-HsG with the descriptive and inferential statistical analysis of GC% and isochore families of all the genes of two copies of genomes through a shiny dashboard to provide results to the end users.

The development of application is done with shiny server. Shiny is an open-source R package of RStudio that is used to build interactive applications with R programming language (Sivaprakasam and Sadagopan, 2019). The pursued SNPs injection in human reference genome and analysis of two copies of genome are done in the application. It may help the biologists with a little computer knowledge to visualize and explore the presence of SNPs in human genome comfortably.

2. Materials and methods

2.1. Retrieval of human reference genome, genes from GTF file and SNP data

The latest reference genome sequence (GRCh38) and annotation file (GTF) of *Homo sapiens* are available in NCBI database (Team T.B.D, 2014). Using the Biostrings-based genome data packages provided by Bioconductor project, those sequences are accessed and are stored in the basic containers like DNAString and DNAStringSet (Huber et al., 2015; Pagès et al., 2017). These are memory efficient string containers, and its string-matching algorithms are used for fast manipulation of large sets of biological sequences. These require the BSgenome package to provide the infrastructure needed to support them and work properly. In the present study, reference genome (1–22, X and Y chromosome) sequences of *Homo sapiens* is obtained from BSgenome.Hsapiens.NCBI.GRCh38 and are stored in Biostrings objects. The genome annotation file in GTF format is retrieved from NCBI using rtracklayer package of Bioconductor. The rtracklayer package is an interface between R and genome browsers and is used to create, manipulate genomic views, and import/export sequences to and from a browser like NCBI. In the present study, only the required genome features like seq_name, seq_range_start, seq_range_end (calculated by start position + width - 1), type and geneID, are retrieved. Based on the start and end positions of each genes, which are provided in GTF file, the gene sequences of all 24 chromosomes of reference genome are retrieved using "Granges" package of Bioconductor. The predefined algorithms in R and Bioconductor are used for analysis of the real time world data like genome sequences of human (Alekseyenko and Lee, 2007). In this regard, the present study used NCList construction algorithm, which is in Granges package do fast performance for genomic intervals retrieval.

SNP locations (SNPlocs) and alleles data for *Homo sapiens* are retrieved from NCBI dbSNP Build 151 (created by NCBI, 2018) through another bioconductor package "SNPlocs.Hsapiens.dbSNP151.GRCh38" (Pagès, 2018). The SNPlocs class is a container for storing known SNP locations for *Homo sapiens*. SNPlocs objects are generally made in advance by a volunteer and made available to the Bioconductor community as "SNPlocs data packages". In dbSNP, every submitted variation has stable and unique SNP ID number ("ss#"). More than one record of a variation will likely be submitted to dbSNP and are grouped into a single reference SNP ("rs#"), which is also a distinct and stable identifier. Both data are stored as vector variable, which are used for the retrieval and manipulation to infer the results.

2.2. SNPs injection in human reference genome and generation of "SNPs injected genome"

The stored SNP data for 1–22, X and Y chromosomes from NCBI dbSNP Build 151 (SNPlocs.Hsapiens.dbSNP151.GRCh38) are injected in the reference genome sequences (BSgenome.Hsapiens.NCBI.GRCh38) using the Bioconductor function injectSNPs(). Those SNPs are perhaps landed at their corresponding correct position in reference genome sequences and the changes of base pairs are made. The SNPs in the altered genome are represented by an IUPAC ambiguity code at each SNP location (Johnson, 2010). Ultimately, the altered genome with mapped SNPs is considered as "SNPs injected genome". In the present study, these original reference genome and SNPs injected genome are considered for nucleotide frequency including GC% for descriptive and inferential analysis. As like in reference genome, the gene sequences of all 24 chromosomes of SNPs injected genome are also retrieved using "Granges" package with NCList construction algorithm of Bioconductor (Alekseyenko and Lee, 2007) and both sets of genes are considered for nucleotide frequency analysis.

2.3. Data storage

The obtained data are assigned in tables and are saved as dataframe in RSQLite database for further analysis. RSQLite embeds the SQLite database engine in R, providing a DBI-compliant interface. SQLite is a public-domain, single-user, very light-weight database engine that implements a decent subset of the SQL 92 standard, including the core table creation, updating, insertion, and selection operations, plus transaction management. dplyr is a new package which provides a set of tools for efficiently manipulating datasets in R. dplyr is the next iteration of plyr, focusing on only data frames. sqldf() transparently sets up a database, imports the data frames into that database, performs the SQL select or other statement and returns the result using a heuristic to determine which class to assign to each column of the returned data frame.

2.4. Nucleotide frequency analysis and GC% in two copies of genome

The DNA data of 1–22, X and Y chromosome sequences of human reference and SNPs injected genomes are considered for nucleotide frequency analysis. In this regard, the alphabetFrequency() and vcountpattern() present in Biostring of Bioconductor computes the frequency of each A, T, G, C base pair in both the genomes. The predefined naïve exact algorithm applied in the packages vcountpattern and alphabetfrequency, which do fast performance for nucleotide frequency

calculation are used. Similarly, GC% for sequences of two copies of genome are calculated by using the formula "G+C/width of the sequence *100" (Gao and Zhang, 2006; Cohen et al., 2005). To do the descriptive analysis on nucleotide changes, plots are created using ggplot2 and the analysis can be made for each chromosome of two copies of genome in the same window.

2.5. Descriptive and inferential statistics on frequency of GC% and isochore families between genomes

To check the normality of the variables, the study used shapiro.test for the data of two variables such as GC % of reference genome and SNPs injected genome. The Shapiro-Wilk test is based on the degree of linearity in a Q-Q plot (Mishra et al., 2019). This is for the reason that, which inferential statistics is suitable to test the significance based on whether the variables are parametric or non-parametric. For the above paired - two independent samples, to test the significance, the study used Wilcoxon rank sum test and finally the Chi-square test is used to check the two samples of measurements come from different distributions (Fagerland and Sandvik, 2009; Al-jouie et al., 2015).

2.6. snips-HsG shinydashboard application development

With all the above-mentioned data and methodology as shown in Fig. 1, the study developed an application snips-HsG using R programming and Shiny with RStudio as an interface (https://www.r-project.org). Bioconductor packages along with shiny framework are also utilized for the same (Sivaprakasam and Sadagopan, 2019). Shiny is an R package that makes it easy to build interactive web apps straight from R (https://CRAN.R-project.org/package=shiny). Every shiny application has two main components: a user-interface script and a server script. The ui.R script controls the layout and appearance of an application. Server. R script contains the instructions that the computer needs to build an application and codes for back-end like data retrieval, manipulation, and wrangling.

3. Results and discussion

SNPs are of unique importance for studying human genomic variations (Madsen et al., 2007; Sherry et al., 2001) and in this regard, the present study developed a convenient data interface and shinydashboard application on descriptive and inferential analysis of SNPs in human reference genome and SNPs injected genome for easy access, comparison, download and visualization for the biologists. The application is available in https://snip-hsg.shinyapps.io/home/. For the visualization, the screenshot of homepage of the application is shown in Fig. 2. The rest of results in the form of plots and statistical images of nucleotide frequency, GC% and isochore family analysis are described and shown as follows.

3.1. Comparison in nucleotide frequency data between reference and SNPs injected genomes

In the present study on web application development for genomic variation statistics and visualization, the chromosome wise (1–22, X and Y) comparison on nucleotide frequencies is made between the reference genome and SNPs injected reference genome and the results are displayed in the web application based upon user's selection. Initially, the nucleotide count is generated for each 24 chromosomes of reference genome and SNPs injected genome, which are stored separately as tables in RSQLite database and therefore the users can download the resultant data in csv, tsv and doc formats. The same data are displayed in plots for user's easy visualization. To show the results, the screenshot of nucleotide frequency of chromosome 6 of the two copies of genome in dot plots are shown in Fig. 3.

The study showed the comparison between nucleotide frequency of two copies of genome and observed certain nucleotide bias due to the SNPs sites (Fig. 3). The above aspects of comparison are scattered as previously suggested (Tian et al., 2011; Koonin and Galperin, 2003; Asthana et al., 2007). As per user's selection, the nucleotide frequency is displayed, and it helps biologist to visualize the chromosomes comparatively. This is the preliminary study and this kind of comparison on SNPs are generally used for mapping of the probes and used for any downstream analysis (Ciobanu et al., 2010; Manconi et al., 2014). In future, if any genome of an individual genome sequences is available, this application may extend to compare anyone with the reference genome sequence to analyze genetic variation based on nucleotide frequency.

3.2. Grouping and comparison on isochore family data of the two copies of genome

The GC% in human genome is one of the important genomic features that are possibly related from a functional point of view (Gao and Zhang, 2006; Cohen et al., 2005). In this regard, GC% for all the chromosomes between two copies of genome are compared and the results show that there is not much deviation in GC% (see the results in the application). Therefore, the study is extended to group and compare isochore families based on GC% on each gene sequences of both reference and SNPs injected genomes. To the best of our knowledge, no work was done on isochore families in the reference genome before and after injected SNPs comparatively. Regarding GC% study at genomic level, the present results are in agreement with a recent work (Piovesan et al., 2019) and revealed that both isochore size and average GC% of isochore families are conserved in between genomes, supporting the concept that isochores represent a fundamental level of genome organization (Cozzi et al., 2015). As previously described (Cohen et al., 2005; Rouchka and States, 2002; Costantini et al., 2006), GC% for each gene of two copies of genome are generated and based on it, each gene is grouped in to five different isochore families, L1 (<37% GC), L2 (37–42%), H1 (42–47%), H2 (47–52%), H3 (>52%). From the results, the present study verified, and the distribution confirms that they belong in the five families (Fig. 4).

As explained earlier (Rouchka and States, 2002), the distribution of GC% are compared by constructing histograms and are visualized in the application. As the histograms show, there are differences in the GC% distribution among chromosomes. Chromosomes, such as 6, 8, 9, 19 and Y appear with bimodal distribution and rest of the other chromosomes appear with unimodal distributions in the GC%. It is also observed that in none of the cases were there more than two peaks in the distribution of GC fragments. Our results show the difficulty of defining isochore boundaries based on GC fragments alone. But this comparison may help the biologist to examine the GC% and isochore family's differences comparatively between genomes. Generally, in a genome the functional region's GC content is 35% to 60% and if found lesser or more, then it may be the indication for some genetic defect (Vinogradov, 2003; Sémon et al., 2005). So, based on the report of GC%, the present study may help biologists to study on disease risk.

3.3. Statistical analysis of GC% based isochore families between sets of genes of two copies of genome

As described in Section 2.5, GC% and the grouped isochore families (L1, L2, H1, H2 and H3) of reference and SNPs injected genomes are considered as paired, independent, categorical, and ordinal variables (Vetter and Mascha, 2018). To relate the two copies of genome, present study has done the preliminary statistical analysis on isochore families of genes of each chromosome comparatively as follows.

The study obtained the summary of histogram with mean, median, upper, and lower quartiles for each chromosome of genomes (see the results from application). These results indicated that values of SNPs injected genome were lower than those of reference genome, which

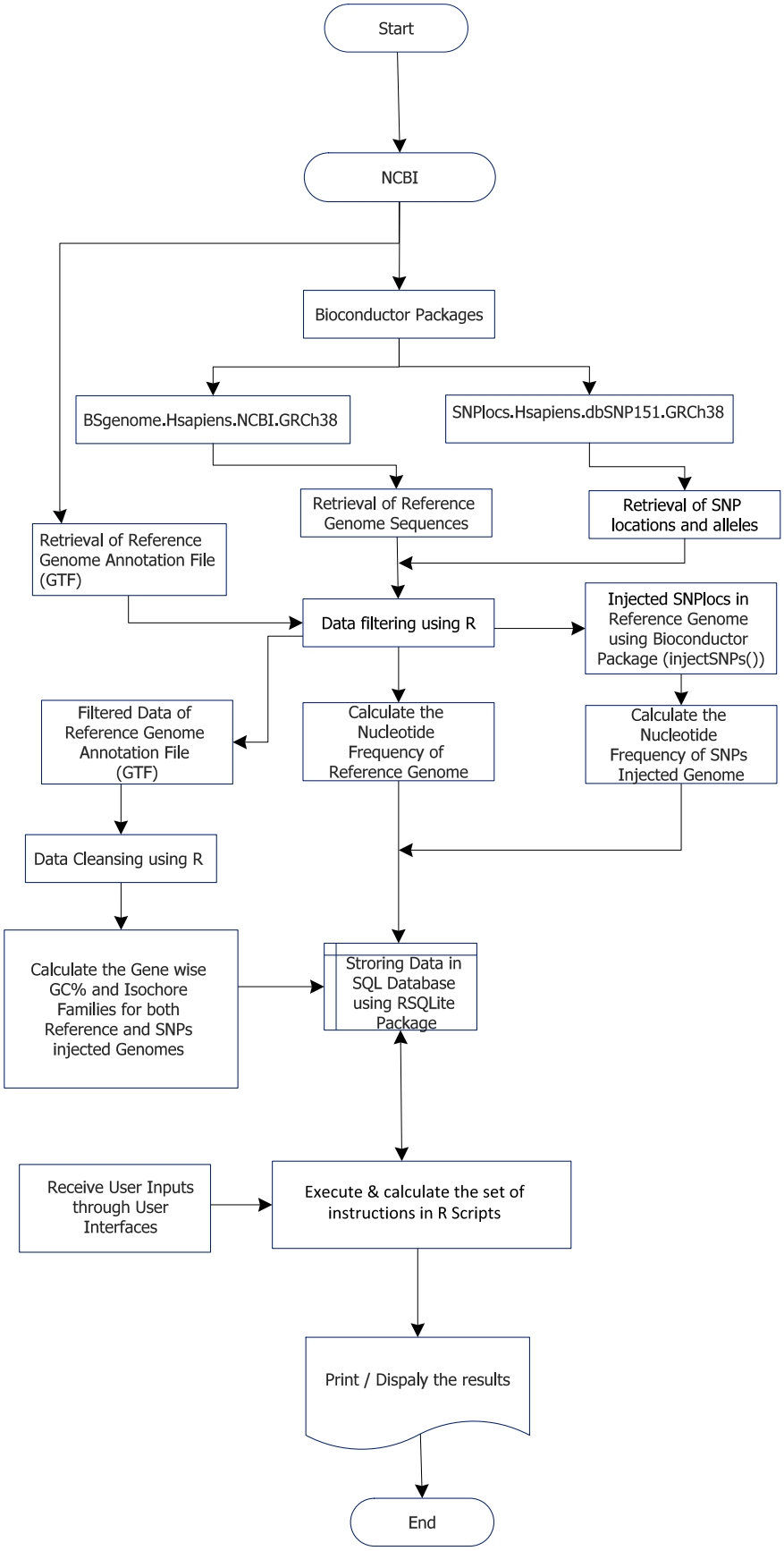

Fig. 1. Flowchart illustrating the different strategies used in this study.

Fig. 2. For the visualization, the screenshot of homepage of the developed application is shown.

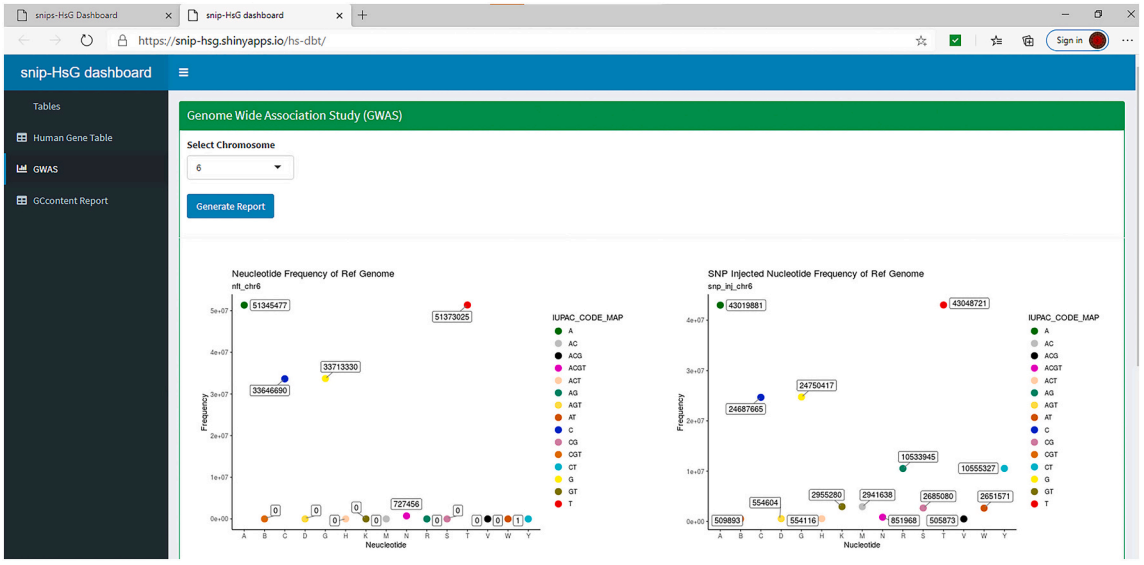

Fig. 3. The screenshot of nucleotide frequency of chromosome 6 of reference genome and its corresponding SNPs injected genome in dot plots are shown.

shows that the gene sets of reference genome are more conserved than SNPs injected genome. As SNPs in the dbSNP database were mapped (Neininger et al., 2019) and the SNP density was visualized through histograms to compare the differences (Fig. 4), which may help the biologists in genome wide SNPs study.

The Shapiro–Wilk test has more statistical power to detect a non-normal (asymmetric) distribution (Yap and Sim, 2011; Barton and Peat, 2014) and R used the AS R94 algorithm for the test (Royston, 1992). The Shapiro–Wilk test is based on the correlation between the data and the corresponding normal scores (Ghasemi and Zahediasl, 2012; Hanusz et al., 2016). In the present study, the obtained Shapiro-Wilk test results for the reference and SNPs injected genomes are W = 0.88386 & *p*-value = 0.008309 and W = 0.88322 & p-value = 0.008058, respectively. In these results, the *P* value less than 0.05 provides evidence that the distribution is significantly different from normal and they have potentially non-normal distributions, which supports the findings by Oztuna et al. (Oztuna et al., 2006). Therefore, the values are considered for non-parametric test to validate the significance relationship between these two variables.

In this regard, the present study used Wilcoxon rank sum test to test the significance difference between two gene sets of human reference and SNPs injected genomes. The Wilcoxon rank-sum test is a nonparametric test for assessing whether two samples of measurements come from the same distribution (Fang et al., 2012). In this test, the study considered two sets of random variables (GC%), which are independent with each other (Pozzoli et al., 2008). The null hypothesis is two samples of measurements come from the same distribution ($\mu 1 = \mu 2$), and the alternative hypothesis is two samples of measurements come from different distribution ($\mu 1 \neq \mu 2$) (Li and Johnson, 2014). The obtained result is W = 519 and *p*-value is 6.413e-05 (see in the web application). As the p-value turns out to be 6.413e-05, and is less than the 0.05 significance level, we reject the null hypothesis. It is predicted that the two samples of measurements come from the different distribution. This was done with reference to the results of Wenhua LV, 2015 (Wenhua, 2015),

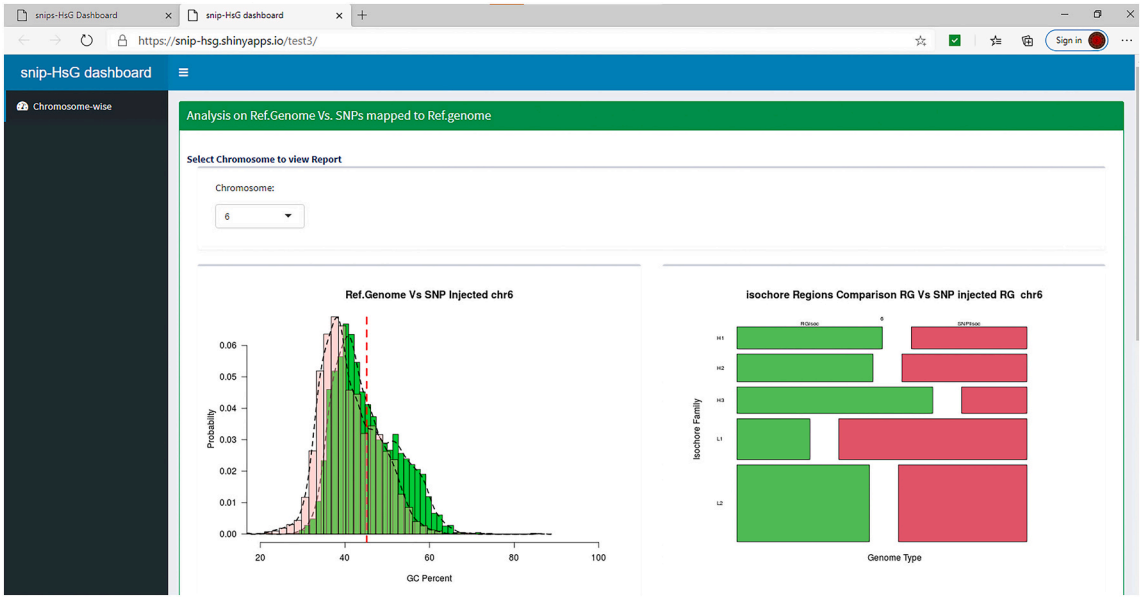

Fig. 4. To relate reference genome and SNPs injected genome, the isochore families of genes of chromosome 6 are comparatively shown in histogram and mosaic plots.

which compared the evolutionary features of human essential gene sets with human housekeeping genes. For visualization, the comparison of GC% based isochore families between chromosome wise genes of two copies of genome and their statistical results of chromosome 6 are shown in the Fig. 5. As the results from Wenhua LV, 2015 (Wenhua, 2015) and Liao BY et al. 2006 (Liao and Zhang, 2006), the present study compared only the gene sets of two copies of genome. The study can further be extended to all the non-coding and transcripts which are excluded in this study and can trace evolutionary rates comparatively in the reference and SNPs injected genomes.

The present study next wished to verify whether the SNPs in genes between two copies of genomes were differently distributed or same depending on isochore type. After grouped isochore families (L1, L2, H1, H2 and H3) in two copies of genome, the Chi-square test is applied to validate the same (Al-jouie et al., 2015). Null hypothesis: frequencies of isochore families in reference genome and SNP injected reference genome, the observations are distributed similarly; Alternative hypothesis: Frequencies of isochore families in reference genome and SNP injected reference genome, the observations are differently distributed. Since the p-value here is less than 0.05 (Fig. 5), the null hypothesis (distribution of any two given chromosomes is similar) is rejected and the result again shows that observations are differently distributed.

The present preliminary study is a holistic approach on all chromosomes and the study will narrow down into a particular chromosome and particular genes like HLA. Since the HLA genes are residing in 6th chromosome, all the represented pictures are highlighted from 6th chromosome.

4. Conclusions

The present preliminary study developed a shinydashboard application to do descriptive and inferential analysis on SNPs in human

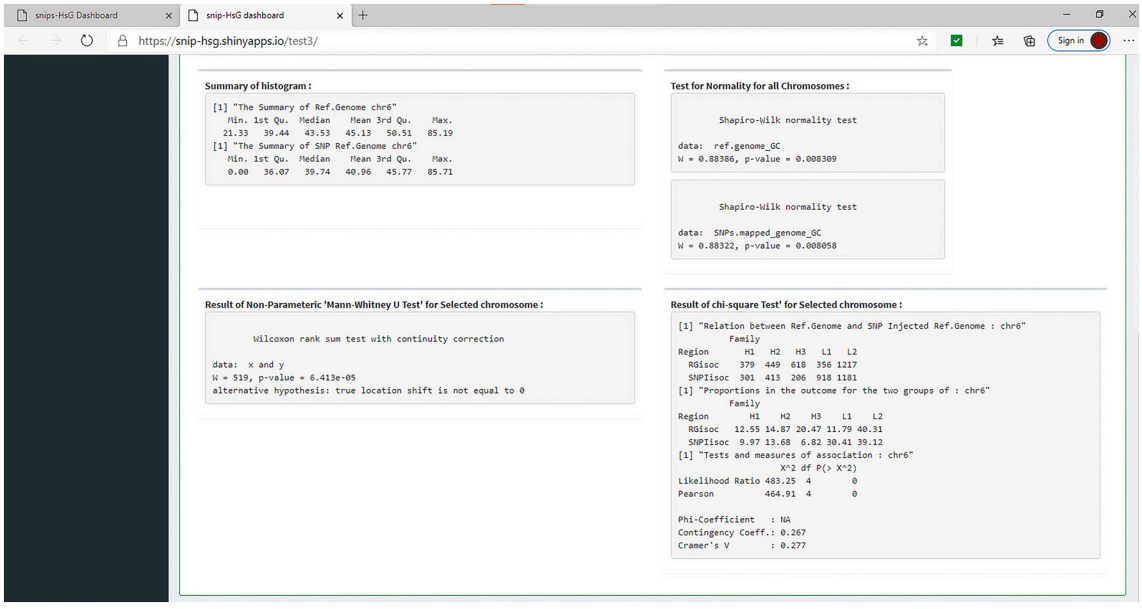

Fig. 5. For visualization, the statistical results of chromosome 6 is shown.

reference genome before and after SNPs injection using R programming with Bioconductor packages, which is hosted in https://snip-hsg.shinyapps.io/home/. Generally, SNPs are used to track the inheritance of diseases and genetic diversity, which can be done commonly through nucleotide frequency analysis (Sherry et al., 2001). In this regard, the comparison of nucleotide frequency between reference and SNPs injected genomes showed that the nucleotide bias due to the SNPs sites.

When SNPs occur within a gene or in a regulatory region near a gene, they may play a role in tracking diseases and diversity (Madsen et al., 2007; Sherry et al., 2001). Hence, all the genes of two copies of genomes are categorized in to five separate isochore families, L1 (<37% GC), L2 (37–42%), H1 (42–47%), H2 (47–52%), H3 (>52%) based on GC%. The comparative results through histograms showed that there are discrepancies in the GC% distribution among chromosomes and this assessment may help the biologist to examine the nucleotide frequency including GC% and isochore family differences comparatively between genomes.

Statistical analysis is used for genome data distribution comparatively (Yap and Sim, 2011; Hanusz et al., 2016; Li and Johnson, 2014). In this regard, the Shapiro–Wilk test results provided the evidence that the data of two copies of genomes have non-normal distributions. Hence, the values are tested with non-parametric Wilcoxon rank sum test and Chi-square test to validate the significance relationship between these two variables. Finally using these tests, the significance difference between gene sets of two copies of genome are projected that the two samples of measurements come from different distributions.

The application may help biologists with little computer knowledge to visualize, download and descriptively analyze the reference and SNPs injected genome comfortably. There are other parameters apart from GC % responsible for disease risk and diversity studies (Vinogradov, 2003; Sémon et al., 2005; Arhondakis et al., 2020). Using only GC% as a parameter in this work is a major limitation and the study will be extended with other parameters and the work will be published elsewhere.

CRediT authorship contribution statement

Balamurugan Sivaprakasam: Methodology, Data curation, Writing and Reviewing. **Prasanna Sadagopan**: Conceptualization and Supervision

Declaration of competing interest

The authors declare no conflict of interest.

References

Alekseyenko, A.V., Lee, C.J., 2007. Nested containment list (NCList): a new algorithm for accelerating interval query of genome alignment and interval databases. Bioinformatics. 23 (11), 1386–1393. https://doi.org/10.1093/bioinformatics/btl647.

Al-jouie, A., Esfandiari, M., Ramakrishnan, S., Roshan, U., 2015. Chi8: a GPU program for detecting significant interacting SNPs with the Chi-square 8-df test. BMC Res. Notes. 8, 436. https://doi.org/10.1186/s13104-015-1392-5.

Arhondakis, S., Milanesi, M., Castrignanò, T., Gioiosa, S., Valentini, A., Chillemi, G., 2020. Evidence of distinct gene functional patterns in GC-poor and GC-rich isochores in *Bos taurus*. Anim. Genet. 51 (3), 358–368. https://doi.org/10.1111/age.12917.

Asthana, S., Roytberg, M., Stamatoyannopoulos, J., Sunyaev, S., 2007. Analysis of sequence conservation at nucleotide resolution. PLoS Comput. Biol. 3 (12), e254. https://doi.org/10.1371/journal.pcbi.0030254.

Ayad, L.A.K., Dourou, A.M., Arhondakis, S., Pissis, S.P., 2020. IsoXpressor: a tool to assess transcriptional activity within isochores. Genome Biol. Evol. 12 (9), 1573–1578. https://doi.org/10.1093/gbe/evaa171.

Barton, B., Peat, J., 2014. Medical Statistics. John Wiley & Sons.

Bernardi, G., 2000. Isochores and the evolutionary genomics of vertebrates. Gene. 241 (1), 3–17. https://doi.org/10.1016/s0378-1119(99)00485-0.

Ciobanu, D.C., Lu, L., Mozhui, K., Wang, X., Jagalur, M., Morris, J.A., Taylor, W.L., Dietz, K., Simon, P., Williams, R.W., 2010. Detection, validation, and downstream analysis of allelic variation in gene expression. Genetics. 184 (1), 119–128. https://doi.org/10.1534/genetics.109.107474.

Cohen, N., Dagan, T., Stone, L., Graur, D., 2005. GC composition of the human genome: in search of isochores. Mol. Biol. Evol. 22 (5), 1260–1272. https://doi.org/10.1093/molbev/msi115.

Costantini, M., Clay, O., Auletta, F., Bernardi, G., 2006. An isochore map of human chromosomes. Genome Res. 16 (4), 536–541. https://doi.org/10.1101/gr.4910606.

Cozzi, P., Milanesi, L., Bernardi, G., 2015. Segmenting the human genome into isochores. Evol. Bioinformatics Online 11, 253–261. https://doi.org/10.4137/EBO.S27693.

Elhaik, E., Graur, D., Josic, K., 2010a. Comparative testing of DNA segmentation algorithms using benchmark simulations. Mol. Biol. Evol. 27, 1015–1024.

Elhaik, E., Graur, D., Josic, K., Giddy, L., 2010b. Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acid Res. 38 (15), e158.

Fagerland, M.W., Sandvik, L., 2009. The Wilcoxon-Mann-Whitney test under scrutiny. Stat. Med. 28, 1487–1497. https://doi.org/10.1002/sim.3561.

Fang, Z., Du, R., Cui, X., 2012. Uniform approximation is more appropriate for Wilcoxon rank-sum test in gene set analysis. PLoS One 7 (2), e31505. https://doi.org/10.1371/journal.pone.0031505.

Fareed, M., Afzal, M., 2013. Single nucleotide polymorphism in genome-wide association of human population: a tool for broad spectrum service. Egypt. J. Med. Hum. Genet. 14, 123–134. https://doi.org/10.1016/j.ejmhg.2012.08.001.

Gao, F., Zhang, C.T., 2006. GC-profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic Acids Res. 34, W686–W691. https://doi.org/10.1093/nar/gkl040.

Ghasemi, A., Zahediasl, S., 2012. Normality tests for statistical analysis: a guide for non-statisticians. Int. J. Endocrinol. Metab. 10 (2), 486–489. https://doi.org/10.5812/ijem.3505.

Gonzaga-Jauregui, C., Lupski, J.R., Gibbs, R.A., 2012. Human genome sequencing in health and disease. Annu. Rev. Med. 63, 35–61. https://doi.org/10.1146/annurev-med-051010-162644.

Hanusz, Z., Tarasinska, J., Zielinski, W., 2016. Shapiro–Wilk test with known mean. REVSTAT – Statistical Journal. 14 (1), 89–100.

Huber, W., Carey, V.J., Gentleman, R., Anders, S., Carlson, M., Benilton, S., et al., 2015. Orchestrating high throughput genomic analysis with bioconductor. Nat. Methods 12, 115–121. https://doi.org/10.1038/nmeth.3252.

International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature. 409 (6822), 860–921. https://doi.org/10.1038/35057062.

International Human Genome Sequencing Consortium, 2004. Finishing the euchromatic sequence of the human genome. Nature. 431, 931–945. https://doi.org/10.1038/nature03001.

Johnson, A.D., 2010. An extended IUPAC nomenclature code for polymorphic nucleic acids. Bioinformatics. 26, 1386–1389. https://doi.org/10.1093/bioinformatics/btq098.

Koonin, E.V., Galperin, M.Y., 2003. Sequence — evolution — function. Computational approaches in comparative genomics. Springer US. https://doi.org/10.1007/978-1-4757-3783-7.

Li, H., Johnson, T., 2014. Wilcoxon's signed-rank statistic: what null hypothesis and why it matters. Pharm. Stat. 13 (5), 281–285. https://doi.org/10.1002/pst.1628.

Liao, B.Y., Zhang, J., 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol. Biol. Evol. 23 (3), 530–540.

Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al., 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47 (W1), W636–W641. https://doi.org/10.1093/nar/gkz268.

Madsen, B.E., Villesen, P., Wiuf, C., 2007. A periodic pattern of SNPs in the human genome. Genome Res. 17 (10), 1414–1419. https://doi.org/10.1101/gr.6223207.

Makalowski, W., 2001. The human genome structure and organization. Acta Biochim. Pol. 48 (3), 587–598.

Manconi, A., Orro, A., Manca, E., Armano, G., Milanesi, L., 2014. A tool for mapping single nucleotide polymorphisms using graphics processing units. BMC Bioinformatics. 15 (S10) https://doi.org/10.1186/1471-2105-15-S1-S10.

Mishra, P., Pandey, C.M., Singh, U., Gupta, A., Sahu, C., Keshri, A., 2019. Descriptive statistics and normality tests for statistical data. Ann. Card. Anaesth. 22, 67–72. https://doi.org/10.4103/aca.ACA_157_18.

Neininger, K., Marschall, T., Helms, V., 2019. SNP and indel frequencies at transcription start sites and at canonical and alternative translation initiation sites in the human genome. PLoS One 14 (4), e0214816. https://doi.org/10.1371/journal.pone.0214816.

Oliver, J.L., Carpena, P., Hackenberg, M., Bernaola-Galván, P., 2004. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res. 32 (2), W287–W292. https://doi.org/10.1093/nar/gkh399.

Oztuna, D., Elhan, A.H., Tuccar, E., 2006. Investigation of four different normality tests in terms of type I error rate and power under different distributions. Turk. J. Med. Sci. 36 (3), 171–176.

Pagès, H., 2018. SNPlocs.Hsapiens.dbSNP151.GRCh38: SNP Locations for *Homo sapiens* (dbSNP Build 151). R Package Version 0.99.20.

Pagès, H., Aboyoun, P., Gentleman, R., DebRoy, S., 2017. Biostrings: string objects representing biological sequences and matching algorithms. R package version. 2.42.1.

Piovesan, A., Pelleri, M.C., Antonaros, F., Strippoli, P., Caracausi, M., Vitale, L., 2019. On the length, weight and GC content of the human genome. BMC. Res. Notes. 12 (106), 1–7. https://doi.org/10.1186/s13104-019-4137-z.

Pozzoli, U., Menozzi, G., Fumagalli, M., Cereda, M., Comi, G.P., Cagliani, R., Bresolin, N., Sironi, M., 2008. Both selective and neutral processes drive GC content evolution in the human genome. BMC Evol. Biol. 8 (99), 1–12. https://doi.org/10.1186/1471-2148-8-99.

Romiguier, J., Ranwez, V., Douzery, E.J.P., Galtier, N., 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20, 1001–1009. https://doi.org/10.1101/gr.104372.109.

Rouchka, E.C., States, D.J., 2002. Compositional Analysis of Homogeneous Regions in Human Genomic DNA, Report Number: WUCSE-2002-2. All Computer Science and Engineering Research. https://openscholarship.wustl.edu/cse_research/1137.
Royston, J.P., 1992. Approximating the Shapiro-Wilk W-test for non-normality. Stat. Comput. 2, 117–119.
Sémon, M., Mouchiroud, D., Duret, L., 2005. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Hum. Mol. Genet. 14 (3), 421–427. https://doi.org/10.1093/hmg/ddi038.
Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., Sirotkin, K., 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 29 (1), 308–311. https://doi.org/10.1093/nar/29.1.308.
Sivaprakasam, B., Sadagopan, P., 2019. Development of an interactive web application "shiny app for frequency analysis on *Homo sapiens* genome (SAFA-HsG)". Interdiscip. Sci. 11, 723–729. https://doi.org/10.1007/s12539-019-00340-z.
Team T.B.D, 2014. BSgenome.Hsapiens.NCBI.GRCh38: Full Genome Sequences for *Homo sapiens* (GRCh38). R Package Version. 1.3.1000.
The 1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. Nature. 526, 68–74. https://doi.org/10.1038/nature15393.
Tian, X., Strassmann, J.E., Queller, D.C., 2011. Genome nucleotide composition shapes variation in simple sequence repeats. Mol. Biol. Evol. 28 (2), 899–909. https://doi.org/10.1093/molbev/msq266.
Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., et al., 2005. Fine-scale structural variation of the human genome. Nat. Genet. 37 (7), 727–732. https://doi.org/10.1038/ng1562.
Vetter, T.R., Mascha, E.J., 2018. Unadjusted bivariate two-group comparisons: when simpler is better. Anesth. Analg. 126 (1), 338–342. https://doi.org/10.1213/ANE.0000000000002636.
Vinogradov, A.E., 2003. DNA helix: the importance of being GC-rich. Nucleic Acids Res. 31 (7), 1838–1844. https://doi.org/10.1093/nar/gkg296.
Wenhua, L.V., 2015. Comparing the evolutionary conservation between human essential genes, human orthologs of mouse essential genes and human housekeeping genes. Brief. Bioinform. 16 (6), 922–931. https://doi.org/10.1093/bib/bbv025.
Yap, B.W., Sim, C.H., 2011. Comparisons of various types of normality tests. J. Stat. Comput. Simul. 81 (12), 2141–2155.
Zhang, C.T., Gao, F., Zhang, R., 2005. Segmentation algorithm for DNA sequences. Phys. Rev. E. Stat. Nonlin. Soft Matter Phys. 72, 041917.