SCI-VCF: A cross-platform application to summarise, compare and design interactive visualisations of the variant call format

Venkatesh Kamaraj[1,2,*], Himanshu Sinha[1,2,3,*]

[1] Centre for Integrative Biology and Systems Medicine (IBSE), IIT Madras, Chennai, India

[2] Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras, Chennai, India

[3] Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, IIT Madras, Chennai, India

[*] Corresponding authors

ABSTRACT

Summary: As genomics advances swiftly and its applications extend to diverse fields, bioinformatics tools must enable researchers and clinicians to work with genomic data irrespective of their programming expertise. We introduce SCI-VCF, a comprehensive toolkit for summarising, comparing, and visualising genetic variants from the variant call format. With an intuitive GUI, SCI-VCF aims to bridge the approachability gap in genomics that arises from the predominantly existing command-line utilities.

Availability and Implementation: SCI-VCF is written with R and is freely available at https://github.com/HimanshuLab/SCI-VCF. For installation-free access, users can access the online version of SCI-VCF at https://ibse.shinyapps.io/sci-vcf-online/. A detailed user manual is also available online at https://himanshulab.github.io/SCI-VCF-docs/.

Contact: sinha@iitm.ac.in, vengatesh.vengatesh89@gmail.com

Supplementary information: Data used to create the visualisations in the article are presented as sample VCF files within SCI-VCF.

INTRODUCTION

Genomics is a rapidly evolving field that has profoundly impacted biomedical and public health research. With the advent of next generation sequencing (NGS), an increasing number of genomes from various species are getting sequenced, enabling the applications of genomics to proliferate into diverse disciplines such as medicine (Hudson, 2011), agriculture (Wang et al., 2017), microbiology (Padmanabhan et al., 2013), ecology (Klaper and Thomas, 2004), evolutionary biology (Rokas and Abbot, 2009), psychology (Hoehe and Morris-Rosendahl, 2018), anthropology (Benn Torres, 2020), etc. This is accelerated further by combining genomics with the advancements in data science and artificial intelligence (Navarro et al., 2019). Such interoperability is facilitated by well-defined data formats that represent diverse biological entities. One such data structure is the Variant Call Format (VCF).

VCF is a standard file format to store the sequence variations in a genome like single nucleotide polymorphisms (SNP), insertions, deletions (INDEL), structural variants (SV), and other assorted variants (Danecek et al., 2011). It is widely adopted and is used in many bioinformatics tools that analyse genomic data. While the VCF has many advantages for storing and sharing genomic variants, it can be complex and arduous to understand, particularly for those unfamiliar with bioinformatics and genomic data analysis. Although several tools allow users to inspect and interpret VCF files, they are predominantly command-line utilities and require significant programming expertise from the user's end.

Data visualisation is a ubiquitous method for quickly understanding complex information, but tools that effectively visualise a VCF file are scarce. VIVA (Tollefson et al., 2019), vcfR (Knaus and Grünwald, 2017), and the vcflib (Garrison et al., 2022) are some existing frameworks that allow the visualisation of VCF. Effective use of these frameworks requires proficiency in programming by the user end. Feature comparisons of these open-source variant visualisation software are presented in Table 1.

As the applications of genomics burgeon into diverse fields, it calls for tools and software that empower clinicians and researchers to work with genomic data formats irrespective of their programming expertise. Here, we introduce SCI-VCF, a comprehensive toolkit with an intuitive Graphical User Interface (GUI) that lets users summarise, interpret, and compare genomic variants from VCF files. It also equips users to design interactive visualisations of the VCF in numerous ways. SCI-VCF is platform-agnostic and works seamlessly across any operating system. A web version of the tool is also made available for enhanced accessibility. Although it is infeasible for a single GUI-based tool to support the wide-spanning analytic capabilities of the VCF, SCI-VCF provides a well-founded framework that simplifies the core components of VCF analyses, thus increasing the approachability of genomics to novices.

METHODS

SCI-VCF is written in R, a free and open-source programming language for statistical computing, data analysis, and visualisation. The user interface and the routines are designed and developed in Shiny, an R framework to build interactive applications. VCF files are

parsed in SCI-VCF using the *vcfR* library, a widely used bioinformatics tool. The static visualisations of the VCF files are developed with *ggplot2* (Wickham, 2011), which operates on a grammatical theory of graphics. The static plots are transformed into interactive visualisations using *Plotly* for R. The area-proportional Venn diagrams are made with the *eulerr* package. The entire codebase is containerised in Docker to improve portability and enable deployment of SCI-VCF in different platforms, irrespective of the operating system. An online version of the application is hosted publicly for installation-free use.

Although the structure of a VCF file is standardised, with the progress of genomics, many discrepancies arise with the variant calls, posing challenges in data interpretation and comparison studies. One key riff is the representation of multiallelic variants. While some VCFs depict alternate alleles of a multiallelic site as a comma-separated entry in a single row, other formats may have them as individual rows. Duplicate entries transpire into a file when variants called in different samples are combined. The sample level genotype information may or may not be present. To overcome these hurdles, SCI-VCF, by default, breaks down comma-separated multiallelic sites into individual entries and removes any duplicate variants present in the VCF file. All variant metrics are computed sample-agnostic to support files that do not contain genotype information. The tool also allows these default settings to be changed if the user wills.

SCI-VCF lets the user summarise and compare the contents of variant call format in a guided GUI setting. Both compressed and uncompressed VCF files are accepted as valid inputs. The summarise and the compare workflows are separated into respective modules in the navigation bar. Submodules capturing the result of individual analyses are assembled under each module's sidebars (Figure 1A). Every plot generated can be a customised such that it equips the user to change various aspects of the visualisation. All the data and visualisations generated by the tool are downloadable for further examination. Sample VCF files are provided to get started with SCI-VCF from the ground up.

RESULTS AND DISCUSSION

The variant call format is fundamentally a tab-delimited text file, with each row comprising information about a genomic variant site and metadata of these variants presented at the

beginning of the file (Figure 1B). The first eight columns of the VCF are fixed and are named CHROM, POS, ID, REF, ALT, QUAL, FILTER, and INFO. After breaking multiallelic sites into individual variants and removing duplicates, the first four columns can be used to identify unique variants present in a VCF file. The nature of REF and ALT entries enables the classification of variants into SNPs, INDELs, and other types.

The summary of a VCF file is generated by classifying variants and summing up unique entries in each category. Statistics derived from this summary, like the ratio of transitions to transversions, are generally used as quality control metrics to assess the confidence in the variants captured by the variant calling algorithm. The occurrence of variants throughout the genome is presented as an interactive histogram to recognise regions that are depleted or enriched with genetic variations (Figure 1C). This plot aids in identifying conserved sections of the genome, possibilities of selection pressures, and genomic loci linked to fundamental biological functions and salient clinical features. With the distribution of INDEL sizes captured by SCI-VCF, it is possible to gauge the prevalence of frameshift mutations and structural variants (Figure 1D). The spread of each variant type across different chromosomes is represented in diverse plots. By illustrating the synopsis and distribution of specific variant types across the genome, SCI-VCF helps researchers gain valuable insights into the genetic diversity, evolutionary history, and disease susceptibility of an individual, a population, or a species.

Two VCFs can be compared using SCI-VCF by interpreting the first eight columns of the files as two-dimensional heterogeneous tabular datasets. The area-proportional Venn diagram created assists in a quick overview of the commonalities, dissimilarities, and relationships between the two variant calls (Figure 1E). The shared and unique variant sets between the two VCFs are summarised to advance further investigation by the user (Figure 1F). Comparing VCFs and recapitulating the resulting variant sets enables researchers and clinicians to comprehend the genetic diversity across individuals and populations and gain crucial insights about them.

Novel variants with potential clinical and biological significance are uncovered by comparing the VCFs of an individual or population with a reference database like gnomAD (Karczewski

et al., 2020). Variant calling pipelines can be validated for consistency and accuracy by comparing their results with available high-confidence variant sets for thoroughly studied samples like the Genome In A Bottle benchmark sets (Zook et al., 2016). The genetic basis of various diseases is understood by juxtaposing the variants in individuals with and without a particular disease. This procedure helps in identifying the mutations associated with the condition. A similar approach is also used to study genotype-phenotype association and pin down genetic markers for phenotypic traits. SCI-VCF equips users with the framework to effortlessly compare VCFs and summarise the overlapping and dissimilar variants between individuals and populations. These comparisons and summaries are pivotal for understanding genetics, unravelling disease mechanisms, developing targeted interventions, and advancing genomic research.

All data visualisations in SCI-VCF are presented as interactive plots to enhance exploratory data analysis with the variant call format. With the help of features like input selection, zooming, panning, and tooltips, researchers can explore different aspects and dive deeper into specific areas of interest to uncover patterns, outliers, and relationships that may not be apparent in static plots. All graphics made with the tool are supplemented with plot customisation features to improve the effectiveness, clarity, and visual impact of the data visualisations. As points of interest from the displayed interactive plots can be customised and saved locally, users can extract publication-ready visualisations from SCI-VCF.

The versatility of SCI-VCF makes it suitable for local installation allowing personal use and server deployment permitting communal use. SCI-VCF can be used without any installation as an online tool. If required, it can also be installed in different ways as per the user's requirements. As the essential R libraries are downloaded automatically upon first use, local installation is practicable with R put in place. For better package management, a Conda virtual environment with the dependencies of SCI-VCF is created, which can be reproduced easily for improved environment handling. A containerised form of the application is available as a Docker image which helps run SCI-VCF irrespective of platforms, making it straightforward to port the tool to remote HPC clusters. Advanced users can also customise parts of the SCI-VCF algorithm to switch preprocessing steps according to their analysis necessity.

CONCLUSION

We have developed SCI-VCF as a cross-platform application to summarise and compare the genetic variants from the widely accepted variant call format. Users can also create customisable interactive visualisations based on the contents of VCF files. The guided GUI setting of the tool empowers researchers and clinicians to perform exploratory genomic data analysis on VCF files irrespective of their programming expertise.

The user-friendly and intuitive design of SCI-VCF increases the approachability of genomics to newcomers and introduces genomic data analysis expeditiously. SCI-VCF is deployed at https://ibse.shinyapps.io/sci-vcf-online/, enabling its usage without installation. Users can download and install SCI-VCF locally or on remote servers using the code available on GitHub (https://github.com/HimanshuLab/SCI-VCF). A well-documented user guide is available online at (https://himanshulab.github.io/SCI-VCF-docs/) for improved user support and system understanding.

ACKNOWLEDGEMENTS

We acknowledge Ayam Gupta, Veerendra Gadekar, Harshita Agarwal and Centre for Integrative Biology and Systems Medicine (IBSE) members for their valuable discussions and comments.

FUNDING INFORMATION

The work was supported by the Department of Biotechnology, Govt. of India [BT/GenomeIndia/2018] and the Centre for Integrative Biology and Systems Medicine, IIT Madras [BIO/18-19/304/ALUM/KARH] to H.S.

REFERENCES

Benn Torres, J., 2020. Anthropological perspectives on genomic data, genetic ancestry, and race. Am. J. Phys. Anthropol. 171, 74–86. https://doi.org/10.1002/ajpa.23979

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R., 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. Bioinformatics 27, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Garrison, E., Kronenberg, Z.N., Dawson, E.T., Pedersen, B.S., Prins, P., 2022. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. PLOS Comput. Biol. 18, e1009123. https://doi.org/10.1371/journal.pcbi.1009123

Hoehe, M.R., Morris-Rosendahl, D.J., 2018. The role of genetics and genomics in clinical psychiatry. Dialogues Clin. Neurosci. 20, 169–177.

Hudson, K.L., 2011. Genomics, Health Care, and Society. N. Engl. J. Med. 365, 1033–1041. https://doi.org/10.1056/NEJMra1010517

Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., Gauthier, L.D., Brand, H., Solomonson, M., Watts, N.A., Rhodes, D., Singer-Berk, M., England, E.M., Seaby, E.G., Kosmicki, J.A., Walters, R.K., Tashman, K., Farjoun, Y., Banks, E., Poterba, T., Wang, A., Seed, C., Whiffin, N., Chong, J.X., Samocha, K.E., Pierce-Hoffman, E., Zappala, Z., O'Donnell-Luria, A.H., Minikel, E.V., Weisburd, B., Lek, M., Ware, J.S., Vittal, C., Armean, I.M., Bergelson, L., Cibulskis, K., Connolly, K.M., Covarrubias, M., Donnelly, S., Ferriera, S., Gabriel, S., Gentry, J., Gupta, N., Jeandet, T., Kaplan, D., Llanwarne, C., Munshi, R., Novod, S., Petrillo, N., Roazen, D., Ruano-Rubio, V., Saltzman, A., Schleicher, M., Soto, J., Tibbetts, K., Tolonen, C., Wade, G., Talkowski, M.E., Neale, B.M., Daly, M.J., MacArthur, D.G., 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443. https://doi.org/10.1038/s41586-020-2308-7

Klaper, R., Thomas, M.A., 2004. At the Crossroads of Genomics and Ecology: The Promise of a Canary on a Chip. BioScience 54, 403–412. https://doi.org/10.1641/0006-3568(2004)054[0403:ATCOGA]2.0.CO;2

Knaus, B.J., Grünwald, N.J., 2017. vcfr: a package to manipulate and visualize variant call format data in R. Mol. Ecol. Resour. 17, 44–53. https://doi.org/10.1111/1755-0998.12549

Navarro, F.C.P., Mohsen, H., Yan, C., Li, S., Gu, M., Meyerson, W., Gerstein, M., 2019. Genomics and data science: an application within an umbrella. Genome Biol. 20, 109. https://doi.org/10.1186/s13059-019-1724-1

Padmanabhan, R., Mishra, A.K., Raoult, D., Fournier, P.-E., 2013. Genomics and metagenomics in medical microbiology. J. Microbiol. Methods 95, 415–424. https://doi.org/10.1016/j.mimet.2013.10.006

Rokas, A., Abbot, P., 2009. Harnessing genomics for evolutionary insights. Trends Ecol. Evol. 24, 192–200. https://doi.org/10.1016/j.tree.2008.11.004

Tollefson, G.A., Schuster, J., Gelin, F., Agudelo, A., Ragavendran, A., Restrepo, I., Stey, P., Padbury, J., Uzun, A., 2019. VIVA (VIsualization of VAriants): A VCF File Visualization Tool. Sci. Rep. 9, 12648. https://doi.org/10.1038/s41598-019-49114-z

Wang, W., Cao, X.H., Miclăuș, M., Xu, J., Xiong, W., 2017. The Promise of Agriculture Genomics. Int. J. Genomics 2017, e9743749. https://doi.org/10.1155/2017/9743749

Wickham, H., 2011. ggplot2. WIREs Comput. Stat. 3, 180–185.

https://doi.org/10.1002/wics.147

Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., Henaff, E., McIntyre, A.B.R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., Saghbini, M., Dzakula, Z., Hastie, A., Cao, H., Deikus, G., Schadt, E., Sebra, R., Bashir, A., Truty, R.M., Chang, C.C., Gulbahce, N., Zhao, K., Ghosh, S., Hyland, F., Fu, Y., Chaisson, M., Xiao, C., Trow, J., Sherry, S.T., Zaranek, A.W., Ball, M., Bobe, J., Estep, P., Church, G.M., Marks, P., Kyriazopoulou-Panagiotopoulou, S., Zheng, G.X.Y., Schnall-Levin, M., Ordonez, H.S., Mudivarti, P.A., Giorda, K., Sheng, Y., Rypdal, K.B., Salit, M., 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci. Data 3, 160025. https://doi.org/10.1038/sdata.2016.25

FIGURES

Figure 1: (A) Workflow methodology of SCI-VCF enumerating the sub-modules of the summarise and compare workflows. (B) Composition of the standard VCF. Metadata and the information of variants in the first eight columns are mandatory. Genotype information is optional and can be present across multiple columns. (C) The occurrence of variants in chromosome 20 in the GIAB sample HG002 is presented as an interactive histogram. The tooltip functionality captures the bin position and the number of variants in that genomic range. (D) The size distribution of INDELs in chromosomes 20, 21 and 22 of HG002. Insertions and deletions are represented in different colours, with the tooltip furnishing more details. (E) An area-proportional Venn diagram depicting the number of unique and common variants in chromosomes 20, 21, and 22 of the GIAB samples HG002 and HG003. (F) An interactive doughnut plot depicting the summary of variants unique to HG003 contrasted against HG002.

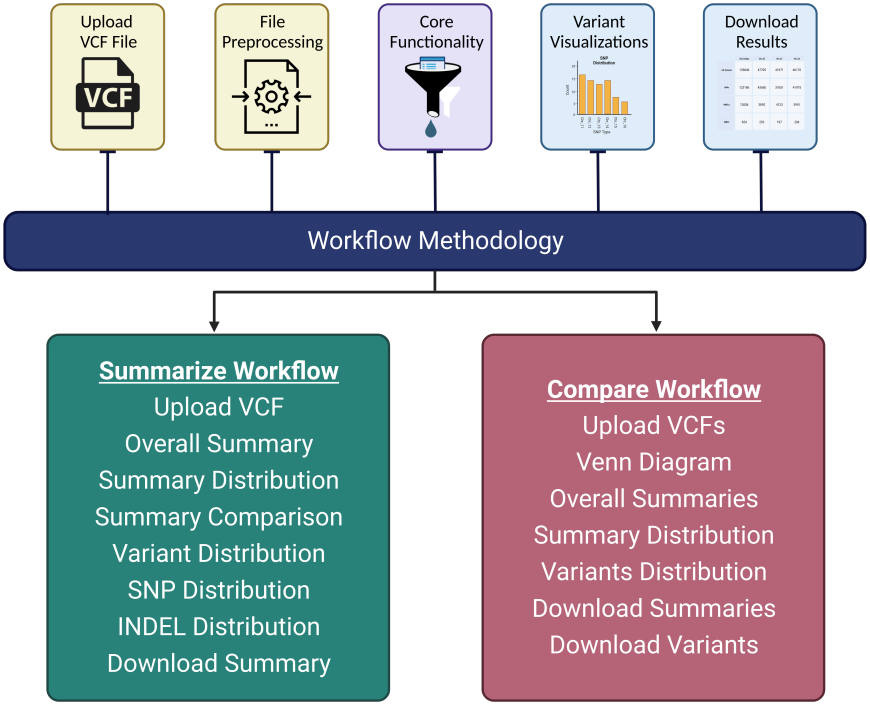
TABLES

Table 1: Comparison of features of VCF visualisation tools.

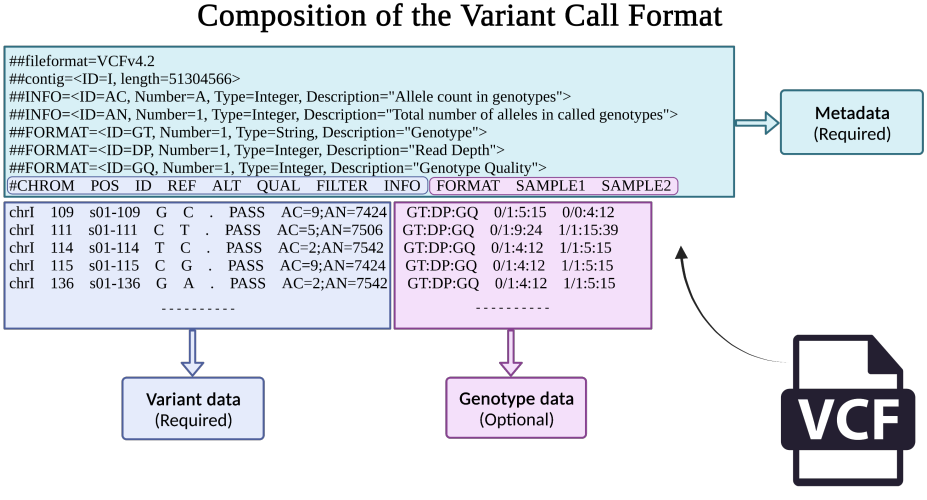
Features	**SCI-VCF**	**VIVA**	**vcflib**	**vcfR**
Technical				
Language	R	Julia	C++	R
Environment (OS)	Windows, Mac, Linux	Windows, Mac, Linux	Windows, Mac, Linux	Windows, Mac, Linux
Graphical User Interface	Yes	No	No	No
Online version	Yes	No	No	No
Summarise				
Variant level summary	Yes	No	Yes	Yes*
Sample level summary	No	Yes	Yes	Yes*
Compare				
Variant level comparison	Yes	No	Yes	Yes*
Unique/intersection set summary	Yes	No	Yes*	Yes*
Visualise				
Interactive plots	Yes	Yes	No	Yes*
Visualise Genomic regions	Yes	Yes	No	Yes*
Output				
Publication quality graphics	Yes	Yes	Yes	Yes*
Export results as tabular data	Yes	Yes	Yes*	Yes*
* Multi-step process involving other tools/libraries				

Figure 1.

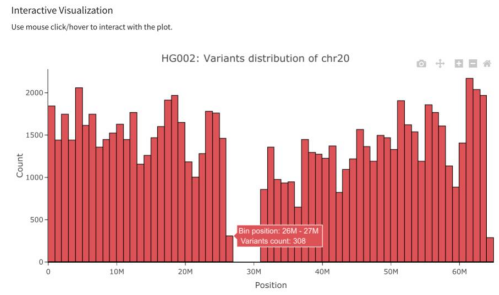
A)



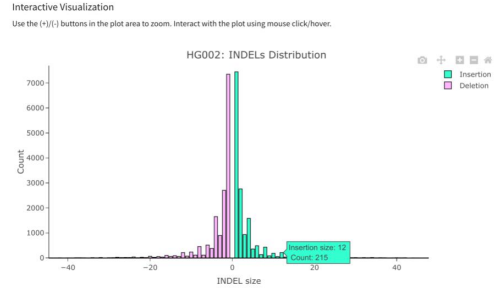
B)



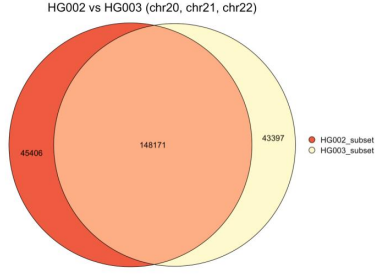
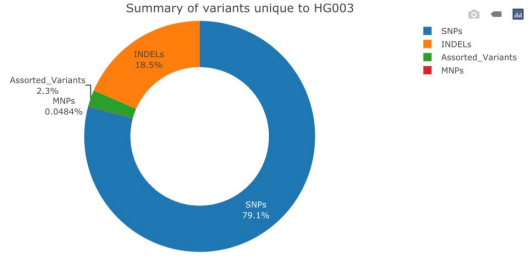
C)



D)



E)



F)