PLOS ONE

RESEARCH ARTICLE

AlignStatPlot: An R package and online tool for robust sequence alignment statistics and innovative visualization of big data

Alsamman M. Alsamman[1], Achraf El Allali[2]*, Morad M. Mokhtar[1,2], Khaled Al-Sham'aa[3], Ahmed E. Nassar[1,3], Khaled H. Mousa[1,3], Zakaria Kehel[3]*

1 Agricultural Genetic Engineering Research Institute, Giza, Egypt, **2** African Genome Center, Mohammed VI Polytechnic University, Ben Guerir, Morocco, **3** International Center for Agriculture Research in the Dry Areas, Giza, Egypt

* Achraf.Elallali@um6p.ma (AEA); Z.Kehel@cgiar.org (ZK)

Citation: Alsamman AM, El Allali A, Mokhtar MM, Al-Sham'aa K, Nassar AE, Mousa KH, et al. (2023) AlignStatPlot: An R package and online tool for robust sequence alignment statistics and innovative visualization of big data. PLoS ONE 18(9): e0291204. https://doi.org/10.1371/journal.pone.0291204

Editor: Nikolas Pontikos, University College London Institute of Ophthalmology, UNITED KINGDOM

Received: January 18, 2023

Accepted: August 23, 2023

Published: September 20, 2023

Data Availability Statement: The online server is available from: https://bioinformatics.um6p.ma/AlignStatPlot, and the R package is open for download and use from Github: https://github.com/AlsammanAlsamman/alignstatplot and Zenodo: https://doi.org/10.5281/zenodo.8133002.

Funding: The author(s) received no specific funding for this work.

Abstract

Multiple sequence alignment (MSA) is essential for understanding genetic variations controlling phenotypic traits in all living organisms. The post-analysis of MSA results is a difficult step for researchers who do not have programming skills. Especially those working with large scale data and looking for potential variations or variable sample groups. Generating bi-allelic data and the comparison of wild and alternative gene forms are important steps in population genetics. Customising MSA visualisation for a single page view is difficult, making viewing potential indels and variations challenging. There are currently no bioinformatics tools that permit post-MSA analysis, in which data on gene and single nucleotide scales could be combined with gene annotations and used for cluster analysis. We introduce "AlignStatPlot," a new R package and online tool that is well-documented and easy-to use for MSA and post-MSA analysis. This tool performs both traditional and cutting-edge analyses on sequencing data and generates new visualisation methods for MSA results. When compared to currently available tools, AlignStatPlot provides a robust ability to handle and visualise diversity data, while the online version will save time and encourage researchers to focus on explaining their findings. It is a simple tool that can be used in conjunction with population genetics software.

Background

Multiple sequence alignment (MSA) is fundamental to the study of genetic variations leading to phenotypic variations in all living organisms. It can be used to identify sequence regions that lead to differences in gene structure and thus gene functionality. MSA analysis is used to study inter- and intra-diversity in order to understand the population structure of the collected DNA samples, which may indicate the origins of evolution and emergence of species [1]. Several dynamic programming algorithms [2] have been used in a variety of programming tools to improve efficiency in various molecular genetic studies. Despite the simple structure of the MSA output, it contains a wealth of information about sequence structure and uniqueness and can be used to extract incomparable information for a wide range of genetic applications.

Competing interests: The authors have declared that no competing interests exist.

When gene structure and annotation data are combined with MSA results, a more detailed picture of the location of sequence variations in genes emerges, which can be used to assess mutational effects and identify gene functionality. The use of gene annotation with sequence alignment to study susceptibility genes and identify pathogenic mutations is useful in cancer genetics [3]. In crop science, MSA is also applied to identify the molecular basis of biotic or abiotic resistance in cultivated crops, which enables varietal improvement through marker-assisted selection [4]. MSA analysis is useful for evaluating gene classes and gene structures when studying gene families, and the addition of gene annotations helps identify structural domains and functional regions [5]. The MSA results could be converted to advanced genomic data formats such as variant calling format (VCF) or haplotype map (HAPMAP), so that diversity studies and genome-wide association studies can be performed. Few bioinformatics tools such as SNP-sites [6] are used to convert MSA format to VCF format.

Despite the abundance of tools for processing MSA results, there are some challenges that researchers face on a daily basis. MSA visualization tools are extremely useful when dealing with small sets of sequences with short lengths, such as short exons or partial genes. For sequences with tens of thousands of characters or huge datasets, it is difficult to visualize the data on a single page, making it difficult to search for potential indels and variations. Most articles show only a few sequences or parts of the sequences studied to keep plot sizes small. This task has become so overwhelming that tools and pipelines are required to obtain conclusive results and understandable, publication-ready figures. Despite the abundance of tools for processing MSA results, there are some challenges that researchers face on a daily basis. To date, there are no bioinformatics tools that allow post-MSA analysis, where information on sequence variations on genes and SNP scales could be used for cluster analysis and combined with gene annotations. The generation of bi-allelic data and the comparison of wild and alternative gene forms are crucial steps in population genetics. In this paper, we introduce a new R package and web-based tool called "AlignStatPlot", which is a well-documented MSA and post-MSA analysis tool. This tool generates new visualization techniques for MSA results and performs both traditional and novel analysis techniques on sequencing data. AlignStatPlot is a simple analysis tool that can be combined with population genetics software to help genetic researchers search for genetic variation that controls the manifestation of disease or stress tolerance. In addition, the tool is also freely available online for those who do not want to install the package through R programming language.

Methods

Analytical procedure

The proposed R package includes several comprehensive data analysis tools that allow users to perform sequence alignment, regular and innovative analyses, and data visualization (Fig 1). The tool was also made available to the public as an online tool to make it more accessible and user-friendly for researchers (Fig 2). AlignStatPlot was written primarily in the R programming language (98%), with 2% of the code written in C to increase the tool's versatility with complex and large data sets. Using the online devtools R package, our R package can be easily installed from the github repository. Fig 3 depicts the total network of data analysis offered by the R package. AlignStatPlot performs sequence alignment analysis for DNA sequences in FASTA format. It can perform MSA analysis using Clustalw, ClustalOmega, [7] and Muscle [8]. The extracted sequence alignments are then formatted and used to provide a statistical overview of alignment performance and sequence similarity (Table 1). The R package circlize is used to provide an overview of sequence alignment [9] (Fig 1A and 1B). Classic visualizations, including phylogenetic and similarity matrices, are then automatically generated. When

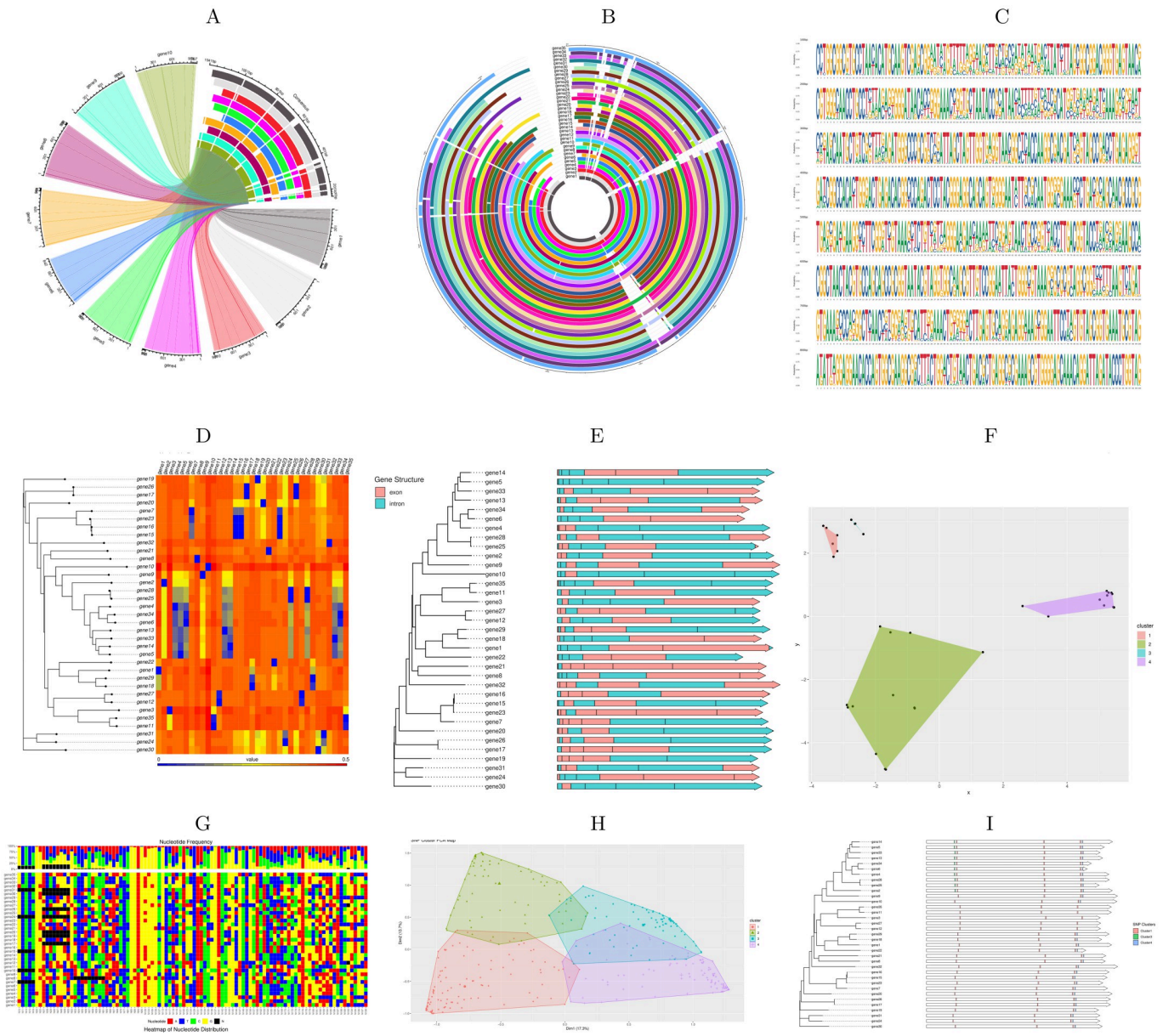

Fig 1. Some sequence alignment statistics visualization were generated using AlignStatPlot, both with online and local tools. The figures include (A) MSA analysis results for a low number of sequences (15 sequences) and (B) for a large number of sequences (15–300 sequences), showing shared regions between aligned sequences. Additionally, (C) displays nucleotide frequency across the MSA, (D) represents the heatmap of the sequence dissimilarity matrix, (E) integrates the phylogenetic tree with sequence annotation, (F) showcases the PCA analysis performed on the studied samples using their sequence variation, and (G) presents nucleotide frequency across the MSA. Furthermore, there is a clustering analysis of MSA-generated SNPs visualized as PCA (H), and their location on gene sequences combined with the phylogenetic tree (I).

https://doi.org/10.1371/journal.pone.0291204.g001

gene annotations are provided, combined plots are created to visualize possible shared aspects of sequence similarity, phylogenetic clusters, and gene structures (Fig 1C–1E). Matrices are generated to describe these variations, which are used for cluster analysis, including principal component analysis (PCA) across genes (Fig 1F). Nucleotide variations are then identified across sequences and large amounts of missing and non-biallelic nucleotides are removed to improve the next analysis procedures and focus on nucleotide variations that may contribute

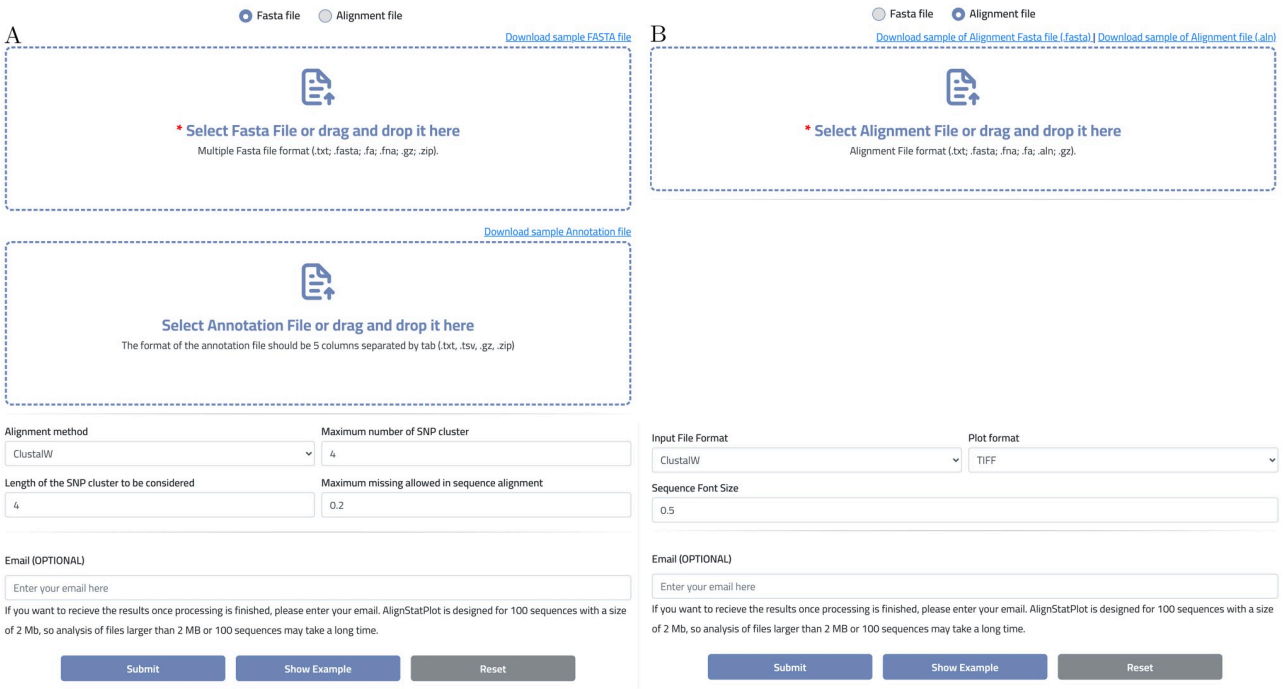

Fig 2. The AlignStatPlot R package offers an online implementation accessible at https://bioinformatics.um6p.ma/AlignStatPlot. This user-friendly platform employs interactive forms to facilitate the entire analysis pipeline. Users can input DNA sequences in Fasta format, along with an optional annotation file, and select their preferred sequence alignment tool (A). Moreover, for datasets consisting of fewer than 300 sequences (DNA or protein), users have the option to directly provide the sequence alignment, enabling the generation of circular format plots, which are particularly valuable (B). This feature enhances the visualization of sequence alignments, facilitating the exploration and analysis of the data.

https://doi.org/10.1371/journal.pone.0291204.g002

to gene evolution and diversity. SNP clustering is one of the new analyses introduced in the R package AlignStatPlot. The analysis uses filtered biallelic nucleotide variation and data clustering to detect possible nucleotide groups with correlated genetic variation across the sequences under consideration (Fig 1C–1E). This type of analysis may reveal a new way to study linkage disequilibrium phenomena at the gene level.

For simplicity, users can do all the above-discussed steps with just one function called "AlignStatPlot". In addition, users have the ability to specify their own analysis with a variety of built-in functions, all of which are well documented. Users can also visualize the sequence

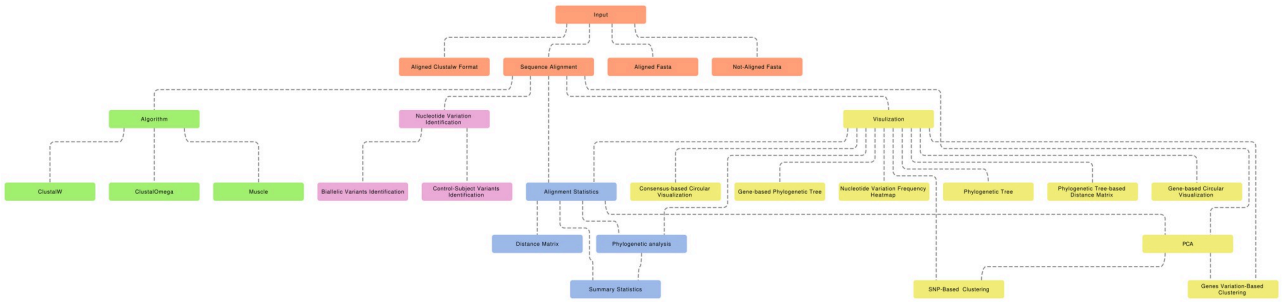

Fig 3. The AlignStatPlot flowchart illustrates the analysis workflow, showcasing the network of steps involved, as well as the possible input options and expected results and visualizations.

https://doi.org/10.1371/journal.pone.0291204.g003

Table 1. Information about the sequences used to validate the AlignStatPlot package.

Data set	Genes	Study	Seq. count	Min. Len.	Max. Len.	Min. GC. Per.	Max. GC. Per.
Rice	COL4	[14]	130	993	999	73.67%	73.97%
	DPL1	[15]	117	853	854	38.57%	38.80%
	DTH7	[14]	125	2221	2268	47.29%	47.69%
	DTH8	[14]	136	875	897	70.06%	71.46%
	Ghd7	[14]	131	774	777	65.25%	66.02%
	Hd1	[14]	116	1181	1506	55.64%	59.11%
	Hd6	[14]	132	973	1002	42.42%	42.75%
	PhyB	[14]	125	3516	3516	52.19%	52.45%
	Se5	[14]	140	870	870	60.46%	60.57%
	LABA1	[16]	247	5029	5050	42.03%	42.13%
	TPP7	PopSet: 2106161045	475	2165	2320	58.06%	59.91%
	gammaTMT	PopSet: 2169220100	475	3300	3390	44.22%	44.75%
Maize	KRN2	PopSet: 2214448971	176	4709	5186	49.35%	50.60%
BRCA	BRCA2_1	[17]	28	5002	5002	34.23%	34.37%
	BRCA2_2	[18]	57	668	1184	23.96%	34.80%
	BRCA2_3	[19]	57	758	894	35.60%	39.51%
	BRCA2_4	[20]	72	696	864	31.72%	33.80%
	BRCA2_5	[21]	134	943	4359	31.47%	47.87%

https://doi.org/10.1371/journal.pone.0291204.t001

alignment by selecting the second option, which requires only the FASTA or clustalw sequence alignment format of the data. This step is accessible via the plotAlignCircle function or the online implementation (Fig 2). The online implementation of AlignStatPlot is hosted on a LAMP server running Linux 5.4.0–89-generic x86 64 (Ubuntu 20.04.3 LTS), Apache (version 2.4.41), and the R (v4.1.2) compiler. The LAMP server is powered by a machine with 32 GB of memory, 16-core CPUs and a 10 TB hard drive. HTCondor (v9.5.0) is used to manage and schedule tasks and processes. When the server's job queue is full, jobs are routed to Africa's fastest high-performance computer (TOUBKAL-POWEREDGE C6420, CRC-STACKHPC, XEON PLATNIUM 8276L 28C 2.2GHZ, MELLANOX INFINIBAND HDR100).

Using PCA analysis for exploring the genetic data

Principal Component Analysis (PCA) is a widely used approach for investigating correlations among samples within a dataset. AlignStatPlot provides two distinct PCA plots to facilitate this analysis, utilizing the genetic variation data obtained from the PCA analysis. The first PCA plot delves into the relationship between samples, offering valuable insights into genetic diversity. By examining the interrelatedness of samples, we gain a deeper understanding of their genetic makeup and evolutionary history (Fig 1F). The second PCA plot specifically focuses on the detected Single Nucleotide Polymorphisms (SNPs) across genes. Within this plot, it is possible to observe the clustering of different SNPs into groups across the studied genes. These SNP clusters may exhibit similarities in terms of their location, variation, or inheritance patterns, shedding light on potential functional connections (Fig 1H). AlignStatPlot further augments the traditional PCA plots by providing an additional SNP clustering analysis plot. This plot not only pinpoints the location of shared SNPs but also reveals the specific groups to which they belong. By integrating this plot with the phylogenetic tree generated through MSA analysis, we gain deeper insights into the gene structure and annotation, enriching our understanding of the dataset (Fig 1I). This comprehensive approach enables us to capture both the

genetic relationships among samples and the significance of SNP clusters within the gene context. It allows for meaningful comparisons with gene structure, annotation, and phylogenetic information, fostering a more comprehensive analysis of the dataset.

Case study

We processed multiple gene sets to validate and demonstrate the potential use of AlignStatPlot in the fields of medicine, microbiology, and plant science (Table 1). We analyzed sequence data from several studies focused on the gene *BRCA1*, which regulates breast cancer progression. We included genes important to plant sciences, such as COL4 [10], DPL1 [11], and DTH7 [12] in rice and KRN2 [13] in maize. In microbiology, 16S and 18S rRNA sequences have been used to demonstrate the utility of AlignStatPlot in the study of prokaryotic diversity, especially when large numbers of genes are studied.

Results

Sequence alignment to study the genetic diversity of different genome samples is a common task for biology researchers. This task has become so overwhelming that tools and pipelines are required to obtain conclusive results and understandable, publication-ready figures. We introduce "AlignStatPlot," a new R package and online tool that is well-documented and easy-to use for MSA and post-MSA analysis. This tool performs both traditional and cutting-edge analyses on sequencing data and generates new visualisation methods for MSA results. We tested our tool on a variety of gene sets (Table 1). More than 3273 sequences were analyzed using AlignStatPlot (Table 1, and S1 Table). The length of the gene sequences ranged from 696 to 5050 bp. Both the online and local versions were useful in analyzing these sequences and provided the expected results (Fig 1 and Table 1). For large gene sets, AlignStatPlot generated plots that indicated indels and shared sequences between genes (Fig 4). In the *BRCA* gene,

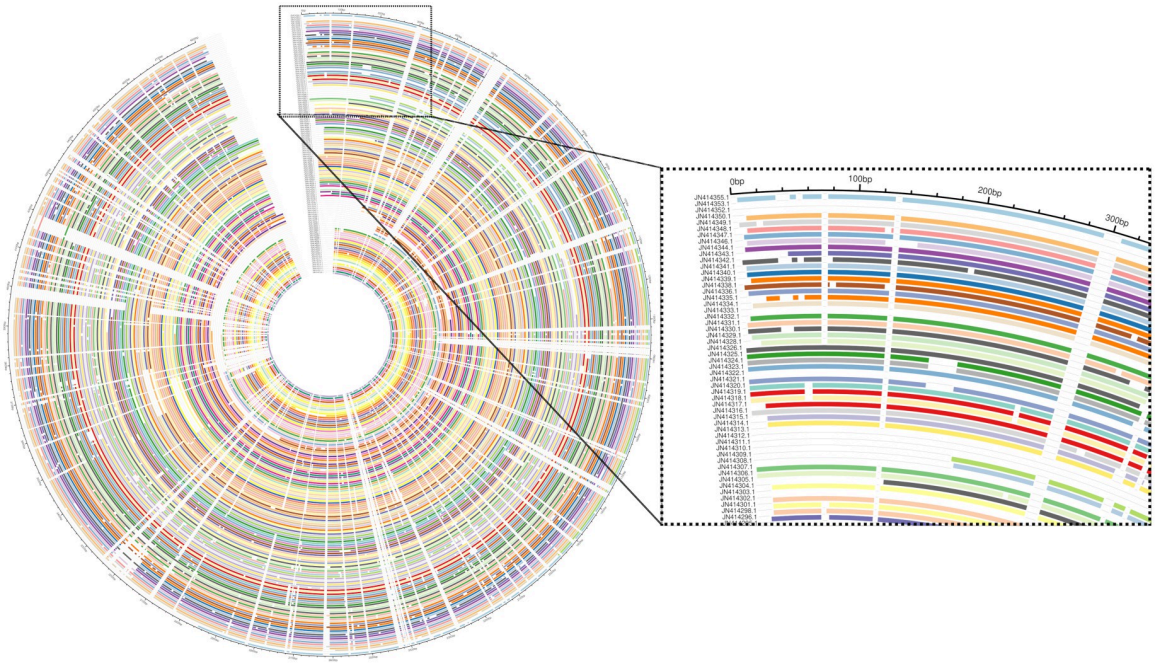

Fig 4. The sequence alignment visualization of large gene sets analyzed with the AlignStatPlot package.

https://doi.org/10.1371/journal.pone.0291204.g004

numerous indels were found and were produced by some groups of sequences sharing a particular region. This type of pattern is uncommon in barcoding genes like 16S and 18S rRNA, which may be why it is useful in studies of microbial diversity, where mostly SNPs are the key factor for isolate identification [22]. Our tool provides two types of similarity matrices, one clustering genes based on their correlation of genetic variation and the other based on their order in the phylogenetic tree (S1 and S2 Figs). Both plots provide two distinct views. While clustering genes based on genetic variation provides a population structure-like view, phylogenetic ordering allows researchers to estimate the rate of similarity of clustered genes to other genes and determine how much similarity exists within and between genes (S3 and S4 Figs). Similar methods have been used to study the genetic diversity and evolution of viruses by comparing viral sequences of different historical strains using MSA and clustering analysis [23]. This method is widely used in gene-based sequencing to study gene diversity in plants [24].

PCA analysis is a common strategy for studying correlated genetic elements in samples. Gene clustering using PCA is another method to show correlated genes in the same data set. This method is very useful for studying functional genes as well as for bacterial gene diversity

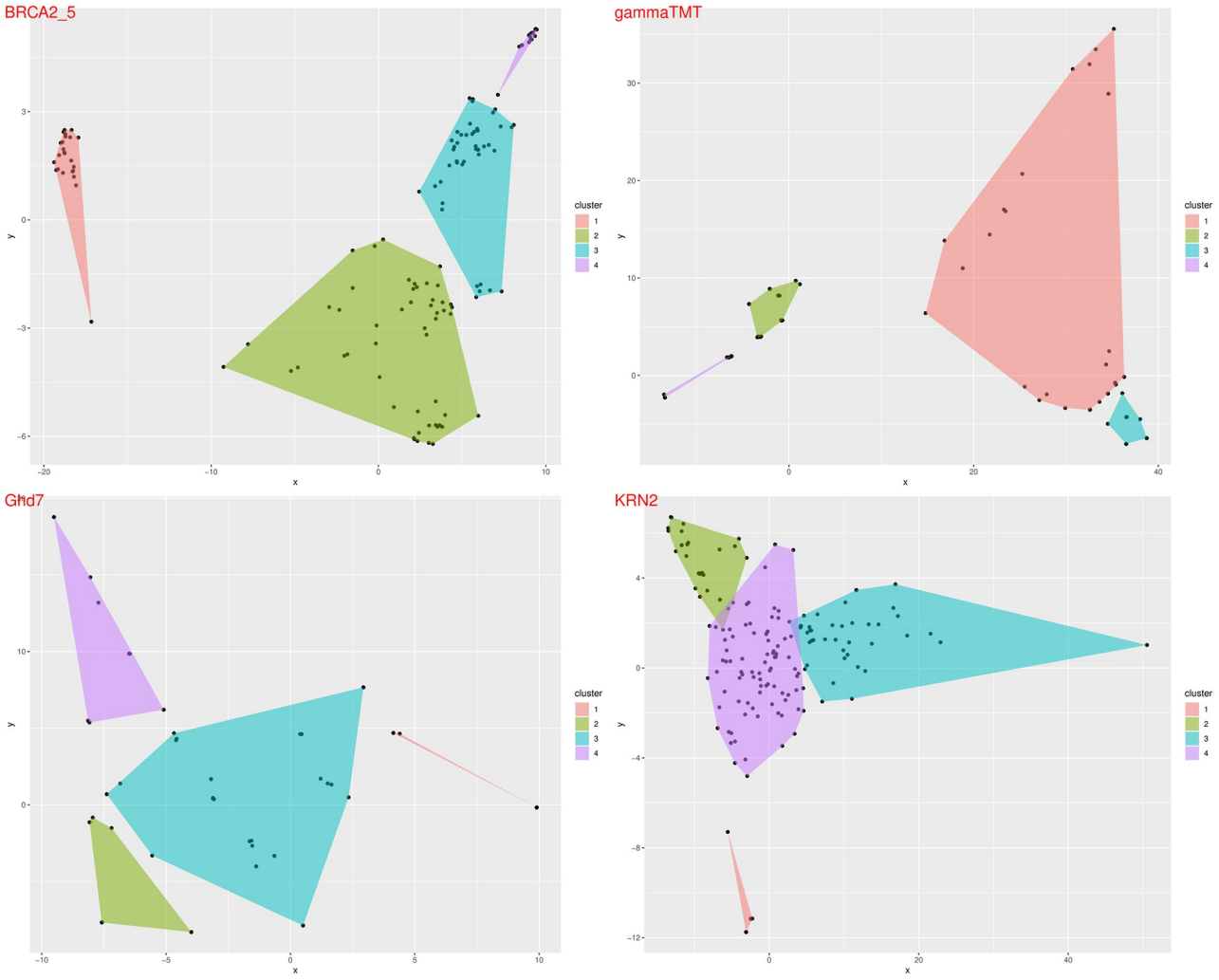

Fig 5. The PCA plot constructed for some genes with the AlignStatPlot package based on the findings of the MSA analysis of genetic variation.

https://doi.org/10.1371/journal.pone.0291204.g005

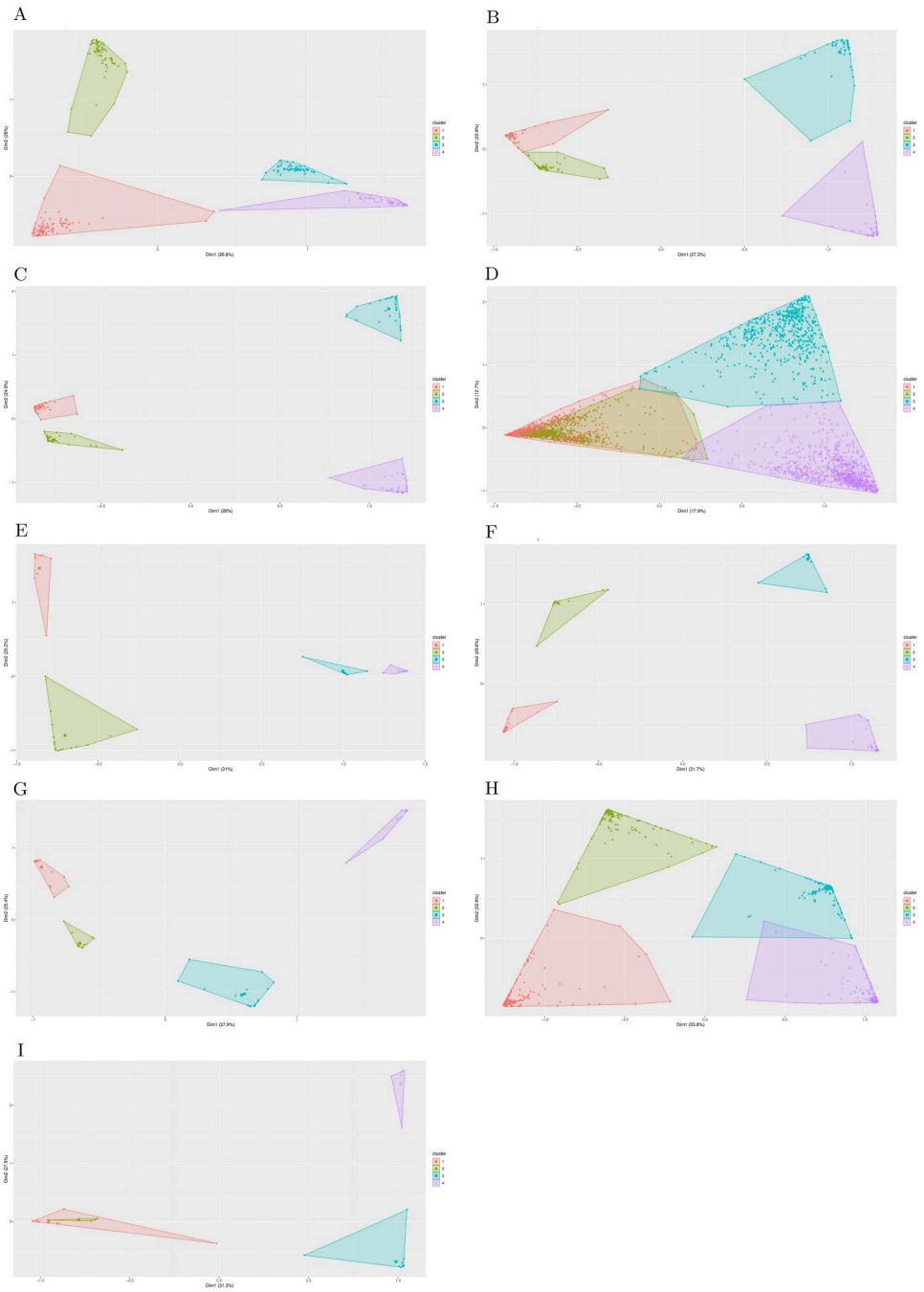

Fig 6. Based on the results of the MSA study, the PCA plot was created for certain SNPs using the AlignStatPlot tool.

https://doi.org/10.1371/journal.pone.0291204.g006

analysis using 16S rRNA [25]. Two different PCA methods are available in AlignStatPlot, one for studying clustering of genes based on SNP variation and the other for studying clustering of SNPs based on variation in genes (Fig 1). Gene sets such as BRAC2 and KRN2 showed distinct groups in PCA clustering of genes based on SNP variation (Fig 5). Clustering of SNPs based on their variation in different studies is a new analysis provided by AlignStatPlot. It

shows that SNPs are clustered across genes regardless of their location. Such an analysis could provide a linkage disequilibrium-like view of nucleotide variation, but at the gene scale rather than the genome scale (Fig 1).

Different SNP cluster groups were detected in the majority of the genes studied. These SNPs could be linked by their location, variation, or inheritance. Alignstaplot provides an additional plot for SNP clustering analysis that shows the location of these shared SNPs as well as the group to which they belong (Figs 1 and 6). This plot is combined with the phylogenetic tree generated by the MSA analysis to provide additional information about the analysis. SNP clustering analysis has been used previously to examine large groups and detect possible genome-level correlations [26, 27]. To our knowledge, such an analysis has never been used at the gene level, and it was also generated on the fly with minimal effort. AlignStatPlot will produce both a phylogenetic tree plot and a statistical summary file. If the annotation for the genes is provided, a phylogenetic tree with exons, introns, and other gene components will be generated (Fig 4). Which in useful in several studies such as gene family analyses [28]. The similar gene structure was reflected by the phylogenetic clustering in some of the investigated genes (Fig 4). The supplementary data contains a detailed discussion of the validation data.

Conclusion

AlignStatPlot has the potential to be successfully integrated in a variety of genomics fields, including medical, crop, and microbial genetics. The tool generated MSA analysis methods that were both traditional and advanced. It includes several analysis procedures that make use of the MSA analysis output in an easy-to-use manner. The tool can be easily combined with several population genetics tools that process bi-allelic data. The online tool will make it easier for researchers who do not have a programming background to produce publishable results.

Supporting information

S1 Table. An overview of the data analysis. An overview of the data analysis findings that were used to verify the alignstatplot tool.
(DOCX)

S1 Fig. Sequence similarity matrix of the studied case study data. Correlation of sequence similarity generated using MSA analysis.
(JPG)

S2 Fig. Sequence similarity matrix of the studied case study data with tree of group A. Correlation of sequence similarity generated using MSA analysis combined with phylogenetic tree group A.
(JPG)

S3 Fig. Sequence similarity matrix of the studied case study data with tree of group B. Correlation of sequence similarity generated using MSA analysis combined with phylogenetic tree group B.
(JPG)

S4 Fig. Phylogenetic tree combined with with gene structure. Phylogenetic tree combined with with the gene structure of some of the case studied data.
(JPG)

S5 Fig.
(JPG)

Author Contributions

Conceptualization: Alsamman M. Alsamman, Achraf El Allali, Khaled Al-Sham'aa, Zakaria Kehel.

Data curation: Alsamman M. Alsamman.

Formal analysis: Morad M. Mokhtar.

Funding acquisition: Zakaria Kehel.

Investigation: Zakaria Kehel.

Methodology: Alsamman M. Alsamman, Achraf El Allali, Morad M. Mokhtar, Khaled Al-Sham'aa.

Project administration: Zakaria Kehel.

Resources: Achraf El Allali, Zakaria Kehel.

Software: Alsamman M. Alsamman, Achraf El Allali, Morad M. Mokhtar, Khaled Al-Sham'aa, Ahmed E. Nassar, Khaled H. Mousa, Zakaria Kehel.

Validation: Khaled H. Mousa.

Visualization: Morad M. Mokhtar, Ahmed E. Nassar, Khaled H. Mousa, Zakaria Kehel.

Writing – original draft: Alsamman M. Alsamman, Ahmed E. Nassar, Zakaria Kehel.

Writing – review & editing: Alsamman M. Alsamman, Achraf El Allali, Morad M. Mokhtar, Khaled Al-Sham'aa, Ahmed E. Nassar, Khaled H. Mousa, Zakaria Kehel.

References

1. Santoferrara LF, Rubin E, Mcmanus GB. Global and local DNA (meta) barcoding reveal new biogeography patterns in tintinnid ciliates. Journal of Plankton Research. 2018; 40(3):209–221. https://doi.org/10.1093/plankt/fby011
2. Carrillo H, Lipman D. The multiple sequence alignment problem in biology. SIAM journal on applied mathematics. 1988; 48(5):1073–1082. https://doi.org/10.1137/0148063
3. Xiao L, Wei F, Liang F, Li Q, Deng H, Tan S, et al. TSC22D2 identified as a candidate susceptibility gene of multi-cancer pedigree using genome-wide linkage analysis and whole-exome sequencing. Carcinogenesis. 2019; 40(7):819–827. https://doi.org/10.1093/carcin/bgz095 PMID: 31125406
4. Dracatos PM, Bartoˈ J, Elmansour H, Singh D, Karafiátová M, Zhang P, et al. The coiled-coil NLR Rph1, confers leaf rust resistance in barley cultivar Sudan. Plant physiology. 2019; 179(4):1362–1372. https://doi.org/10.1104/pp.18.01052 PMID: 30593453
5. Chu Y, Xiao S, Su H, Liao B, Zhang J, Xu J, et al. Genome-wide characterization and analysis of bHLH transcription factors in Panax ginseng. Acta Pharmaceutica Sinica B. 2018; 8(4):666–677. https://doi.org/10.1016/j.apsb.2018.04.004 PMID: 30109190
6. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial genomics. 2016; 2(4). https://doi.org/10.1099/mgen.0.000056 PMID: 28348851
7. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics. 2003;(1):2–3.
8. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research. 2004; 32(5):1792–1797. https://doi.org/10.1093/nar/gkh340 PMID: 15034147
9. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. Bioinformatics. 2014; 30(19):2811–2812. https://doi.org/10.1093/bioinformatics/btu393 PMID: 24930139
10. Zhang L, Li Q, Dong H, He Q, Liang L, Tan C, et al. Three CCT domain-containing genes were identified to regulate heading date by candidate gene-based association mapping and transformation in rice. Scientific reports. 2015; 5(1):1–11. https://doi.org/10.1038/srep07663 PMID: 25563494

11. Mizuta Y, Harushima Y, Kurata N. Rice pollen hybrid incompatibility caused by reciprocal gene loss of duplicated genes. Proceedings of the National Academy of Sciences. 2010; 107(47):20417–20422. https://doi.org/10.1073/pnas.1003124107 PMID: 21048083

12. Ye J, Niu X, Yang Y, Wang S, Xu Q, Yuan X, et al. Divergent Hd1, Ghd7, and DTH7 alleles control heading date and yield potential of japonica rice in Northeast China. Frontiers in plant science. 2018; 9:35. https://doi.org/10.3389/fpls.2018.00035 PMID: 29434613

13. Mural RV, Schnable JC. Can the grains offer each other helping hands? Convergent molecular mechanisms associated with domestication and crop improvement in rice and maize. Molecular Plant. 2022; 15(5):793–795. https://doi.org/10.1016/j.molp.2022.04.003 PMID: 35421584

14. Cui Y, Wang J, Feng L, Liu S, Li J, Qiao W, et al. A combination of long-day suppressor genes contributes to the northward expansion of rice. Frontiers in Plant Science. 2020; 11:864. https://doi.org/10.3389/fpls.2020.00864 PMID: 32612630

15. Xun X, Song G, Fumin Z. Genetic and Geographic Patterns of Duplicate DPL Genes Causing Genetic Incompatibility Within Rice: Implications for Multiple Domestication Events in Rice. Rice Science. 2021; 28(1):58–68. https://doi.org/10.1016/j.rsci.2020.11.007

16. Hua L, Wang DR, Tan L, Fu Y, Liu F, Xiao L, et al. LABA1, a domestication gene associated with long, barbed awns in wild rice. The Plant Cell. 2015; 27(7):1875–1888. https://doi.org/10.1105/tpc.15.00260 PMID: 26082172

17. Ozmen O, Kul S, Risvanli A, Ozalp G, Sabuncu A, Kul O. Single nucleotide variations of the canine RAD51 domains, which directly binds PALB2 and BRCA2. Japanese Journal of Veterinary Research. 2017; 65(2):75–82.

18. Sadlier RA, Debar L, Chavis M, Bauer AM, Jourdan H, Jackman TR. Epibator insularis, a new species of scincid lizard from l'Ile Walpole, New Caledonia. Pacific Science. 2019; 73(1):143–161. https://doi.org/10.2984/73.1.7

19. Lebedev VS, Shenbrot GI, Krystufek B, Mahmoudi A, Melnikova MN, Solovyeva EN, et al. Phylogenetic relations and range history of jerboas of the Allactaginae subfamily (Dipodidae, Rodentia). Scientific reports. 2022; 12(1):1–15. https://doi.org/10.1038/s41598-022-04779-x PMID: 35039544

20. Bannikova AA, Chernetskaya D, Raspopova A, Alexandrov D, Fang Y, Dokuchaev N, et al. Evolutionary history of the genus Sorex (Soricidae, Eulipotyphla) as inferred from multigene data. Zoologica Scripta. 2018; 47(5):518–538. https://doi.org/10.1111/zsc.12302

21. Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, et al. Impacts of the Cretaceous Terrestrial Revolution and KPg extinction on mammal diversification. science. 2011; 334 (6055):521–524. https://doi.org/10.1126/science.1211028 PMID: 21940861

22. Martinez-Porchas M, Villalpando-Canchola E, Suarez LEO, Vargas-Albores F. How conserved are the conserved 16S-rRNA regions? PeerJ. 2017; 5:e3036. https://doi.org/10.7717/peerj.3036 PMID: 28265511

23. Joshi LR, Mohr KA, Gava D, Kutish G, Buysse AS, Vannucci FA, et al. Genetic diversity and evolution of the emerging picornavirus Senecavirus A. Journal of General Virology. 2020; 101(2):175–187. https://doi.org/10.1099/jgv.0.001360 PMID: 31859611

24. Haerinasab M, Eslami-Farouji A. Contribution to the knowledge of the genetic diversity and taxonomy of some Iranian Trifolium species. Genetic Resources and Crop Evolution. 2022; 69(2):699–717. https://doi.org/10.1007/s10722-021-01254-w

25. Klemetsen T, Willassen NP, Karlsen CR. Full-length 16S rRNA gene classification of Atlantic salmon bacteria and effects of using different 16S variable regions on community structure analysis. MicrobiologyOpen. 2019; 8(10):e898. https://doi.org/10.1002/mbo3.898 PMID: 31271529

26. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, et al. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS genetics. 2007; 3(9): e160. https://doi.org/10.1371/journal.pgen.0030160 PMID: 17892327

27. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C. dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics. 2004; 20(8):1233–1240. https://doi.org/10.1093/bioinformatics/bth069 PMID: 14871870

28. Zhao P, Wang D, Wang R, Kong N, Zhang C, Yang C, et al. Genome-wide analysis of the potato Hsp20 gene family: identification, genomic organization and expression profiles in response to heat stress. BMC genomics. 2018; 19(1):1–13. https://doi.org/10.1186/s12864-018-4443-1 PMID: 29347912