frontiers | Frontiers in Plant Science

TECHNOLOGY AND CODE
published: 01 July 2022
doi: 10.3389/fpls.2022.927407

HAPPE: A Tool for Population Haplotype Analysis and Visualization in Editable Excel Tables

Cong Feng[1], Xingwei Wang[1,2,3], Shishi Wu[1,2,3], Weidong Ning[1,4], Bo Song[1], Jianbin Yan[1]* and Shifeng Cheng[1]**

[1] Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences (CAAS), Shenzhen, China, [2] State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, Henan University, Kaifeng, China, [3] Shenzhen Research Institute of Henan University, Shenzhen, China, [4] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China

Haplotype identification, characterization and visualization are important for large-scale analysis and use in population genomics. Many tools have been developed to visualize haplotypes, but it is challenging to display both the pattern of haplotypes and the genotypes for each single SNP in the context of a large amount of genomic data. Here, we describe the tool HAPPE, which uses the agglomerative hierarchical clustering algorithm to characterize and visualize the genotypes and haplotypes in a phylogenetic context. The tool displays the plots by coloring the cells and/or their borders in Excel tables for any given gene and genomic region of interest. HAPPE facilitates informative displays wherein data in plots are easy to read and access. It allows parallel display of several lines of values, such as phylogenetic trees, *P* values of GWAS, the entry of genes or SNPs, and the sequencing depth at each position. These features are informative for the detection of insertion/deletions or copy number variations. Overall, HAPPE provides editable plots consisting of cells in Excel tables, which are user-friendly to non-programmers. This pipeline is coded in Python and is available at https://github.com/fengcong3/HAPPE.

Keywords: haplotype, SNPs, phylogenetic clustering, visualization, Excel

OPEN ACCESS

Edited by:
Nunzio D'Agostino, University of Naples Federico II, Italy

Reviewed by:
Zhenyu Jia, University of California, Riverside, United States
Vivek Shrestha, The University of Tennessee, Knoxville, United States

****Correspondence:***
Bo Song
songbo01@caas.cn
Jianbin Yan
jianbinlab@caas.cn
Shifeng Cheng
chengshifeng@caas.cn

Specialty section:
This article was submitted to Plant Bioinformatics, a section of the journal Frontiers in Plant Science

Received: *24 April 2022*
Accepted: *13 June 2022*
Published: *01 July 2022*

Citation:
Feng C, Wang X, Wu S, Ning W, Song B, Yan J and Cheng S (2022) HAPPE: A Tool for Population Haplotype Analysis and Visualization in Editable Excel Tables. Front. Plant Sci. 13:927407. doi: 10.3389/fpls.2022.927407

INTRODUCTION

The decrease in DNA sequencing costs and routine bioinformatics platforms have led to the explosion of genomic datasets in recent years. To date, more than 700 plant reference genomes have been sequenced and assembled (Marks et al., 2021; Sun et al., 2022). Furthermore, the size of the plant population used for resequencing studies increases dramatically, resulting in the rapid accumulation of plant genomic datasets. Recently, the genomes of as many as 3,365 accessions of chickpea have been sequenced (Varshney et al., 2021). The reductions in sequencing costs have also promoted populational genomic studies of plants with a large genome, such as gingko (∼ 9.8 Gb) (Zhao et al., 2019) and common wheat (∼16 Gb) (Cheng et al., 2019; Hao et al., 2020; Zhou et al., 2020).

The nearby single nucleotide polymorphism (SNPs) sites in the genome are usually inherited together, and the combination of SNPs is termed the haplotype (The International HapMap Consortium, 2003). The identification of haplotypes is important in the analyses of genomes

because it is beneficial for reducing the cost of genotyping, reducing the complexity of association studies, and providing a foundation for haplotype-designed breeding. Haplotype-based association studies have been demonstrated to be efficient in the identification of genes related to complex traits (Todesco et al., 2020). Many tools have been developed to characterize and visualize the haplotypes in genomes. However, the increase in either population size or genome size has caused some problems in displaying genome haplotypes. Although some of the existing tools, such as VIVA (Tollefson et al., 2019), inPHAP (Jäger et al., 2014), and Haploscope (San Lucas et al., 2012), can efficiently produce heatmaps or plots showing the haplotypes in the genome, they rarely display the genotypes with details for each single SNP or any other values or information that can be essential for the researchers to better understand the results. Several interactive tools have been developed to allow zooming in to obtain more details of the haplotypes, but outputting these figures is also not user-friendly.

Here, we present a tool, HAPPE (Haplotype plot in Excel table), which generates visuals of the haplotypes and genotypes in plots consisting of colored cells and/or their borders in Excel tables in a phylogenetic context clustered by a clustering algorithm (aggregative hierarchical clustering). HAPPE automatically detects the haplotypes and displays the genotypes of each single SNP using a distinguished color system in the cells of Excel for any given gene or genomic region of interest. HAPPE also allows the parallel display of other customized information, including the *P* values of genome-wide association studies, the entries of genes or SNPs, and the sequencing depths at each position, thereby enabling the visualization and detection of copy number variations. The users can easily select and copy the values out of the tables to recreate new plots or for the purpose of secondary analyses. Overall, the clustering, characterization, and display of genotypes and haplotypes in editable cells in Excel tables provides an easier method for users to read and access datasets.

METHODS

Data Input

Several dependences should be preinstalled before the installation of HAPPE: bcftools (Danecek et al., 2021), bgzip and tabix in htslib (Bonfield et al., 2021). HAPPE can take input files of genotypes in a format of VCF compressed with bgzip along with its corresponding index file. A list of samples together with their colors should be provided. HAPPE can also take an input of sequencing depth at the loci to be displayed, but these values need to be calculated in advance using a third-party program, such as a mosdepth (Pedersen and Quinlan, 2018).

Analysis

HAPPE filters and processes the samples and variants according to the list of samples and genomic regions if they were provided by the users; otherwise, it keeps all samples and variants by default. The annotation information is then extracted from the INFO field for each variant followed by the conversion of genotypes to a numeric matrix with −1, 0, 0.9 and 1 representing reference allele, data not available, heterozygous allele and alternative allele, respectively. The distance matrix between samples is measured by Euclidean distance, and the linkage matrix is then calculated using the ward linkage method. Using this method, a tree is generated, and the samples are grouped using dynamicTreeCut (Langfelder et al., 2008). HAPPE can also calculate the normalized depth of each window (50 bp by default) following the formula:

$$\text{Normalized depth of this window} = \frac{\textit{total depth of this window} \,/\, \textit{window size}}{\textit{average depth of this sample}}$$

Visualization

HAPPE uses the openpyxl module[1] to edit Excel tables. HAPPE first generates a phylogenetic tree in Excel, the width of which can be customized (1,000 columns of Excel cells with a consistent width by default). Next, the label and annotation of samples are displayed, and the cells are correspondingly colored according to the color code provided. Then, the genotypes and the annotation of each variant are displayed and colored blue or red, representing the presence and absence of the variant, respectively. Finally, a heatmap is produced on the right to show the variation in sequencing depth at each window.

Datasets

We used a published dataset of *Setaria viridis* (Mamidi et al., 2020) to test this pipeline. We first downloaded the reads from the National Center of Biotechnology Information under the accession numbers BioProject PRJNA560514 and PRJNA265547, mapped the reads to the reference genome and called SNPs following the methods described in the manuscript (Mamidi et al., 2020). Finally, two genes (Sevir.5G085400 and Sevir.5G394700) were selected for visualization using HAPPE.

RESULTS

Framework

HAPPE consists of three major steps: data input, analysis and visualization (**Figure 1**). The tool takes one compulsory input file, the genotyping dataset in compressed Variant Call Format (VCF), and several other optional input files to display additional information of the individuals or SNPs. For example, a file of read alignment in "bam" or "cram" format is required to display the sequencing depth (**Figure 2**). The genotyping datasets can be either annotated or not. If the VCF file has been annotated by SnpEff (Cingolani et al., 2012), the annotations can be displayed in the plot and tables. Each individual should be given a unique identity, and additional information, including ancestral group, geographical location and read depth, can also be added by setting the corresponding options (see the usage of HAPPE).

The samples and variants are then filtered according to the input list of samples and the range of coordinates followed by the

[1]https://foss.heptapod.net/openpyxl/openpyxl

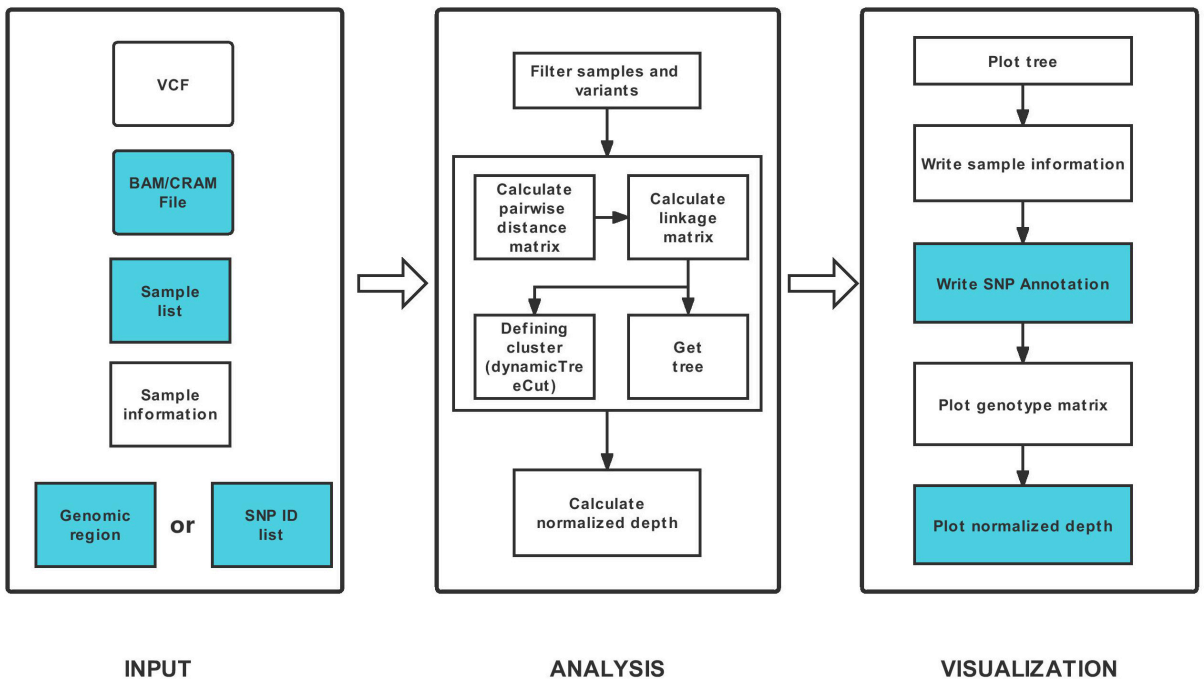


FIGURE 1 | Workflow of HAPPE. The inputs and outputs colored in white are required, whereas the others are optional.

computing of pairwise distance between samples. The distance matrix is then used to calculate the linkage between samples. A linkage matrix is built, and a tree is generated to show the relationship between the samples.

Three compulsory regions, a tree, a list of samples and a main panel of genotype matrix, are displayed in the output table of HAPPE. Additional information of samples, including ancestral group, geographical location and read depth, and annotation of SNPs, including putative effects, gene names, amino acid changes and gene functions, can also be displayed if the corresponding options are properly assigned.

Examples

To better illustrate the features of HAPPE, we applied it to a subset of the *S. viridis* genomic dataset (Mamidi et al., 2020). Sevir.5G085400, a gene associated with seed shattering, was selected as an example. The genotyping dataset in VCF format was input into HAPPE, which selects the regions according to the coordinates assigned in the options "-k" and "-r." The samples were then automatically clustered according to the genotypes in the selected regions, and a tree was drawn on the left of the main panel by coloring the cell border on the path of the tree into black. The number of cells used to show the tree is automatically adjusted proportionally to the branch length of the tree.

Other than the main panel showing the haplotypes, the plot can be extended to display additional information by setting a customized option "-i" to add features to the samples and an option of "-I" to add features to SNPs, such as the name and functions of the genes in which the SNPs are located. In the example shown in this study (**Figure 2A**), we added the labels of individuals in the population and the locations where they were collected shown in latitude and longitude (**Figure 2B**), allowing a quick projection to the world map in Excel to generate a more direct view of the geographical distribution (**Figure 2C**) of the individuals with different haplotypes. For the SNPs, we added the substitution type, annotation as well as the gene names and functions (**Figure 2D**). These factors are critical and allow the readers to have a quick view of the potential effects of the genotypes and haplotypes. A heatmap showing the sequence depth of each SNP in different individuals is shown on the right of the main panel, which can be helpful for the detection of copy number variants or the identification of problematic samples. In the example shown in **Figure 2**, a deletion event of this gene was detected in the accession AGTWO, as evidenced by the low depth of reads (**Figures 2A,E**). This deletion was also confirmed in the genome browser Integrative Genomics Viewer[2].

In another example, ten haplotypes were identified in Sevir.5G394700, a gene conferring the angle of *S. viridis* leaves (**Figure 3**). Several of the haplotypes were group specific. For example, haplotype 1 was predominantly found in the "Central_East" subgroup (**Figures 3B,C**), whereas haplotypes 6 and 8 were specific to the "West_Coast" subgroup. In addition, haplotypes 7 and 9 were specific to the "Central" subgroup.

DISCUSSION

Genome sequencing and analysis were predominantly performed by bioinformaticians or programmers, but this situation has

[2]https://github.com/igvteam/igv-webapp

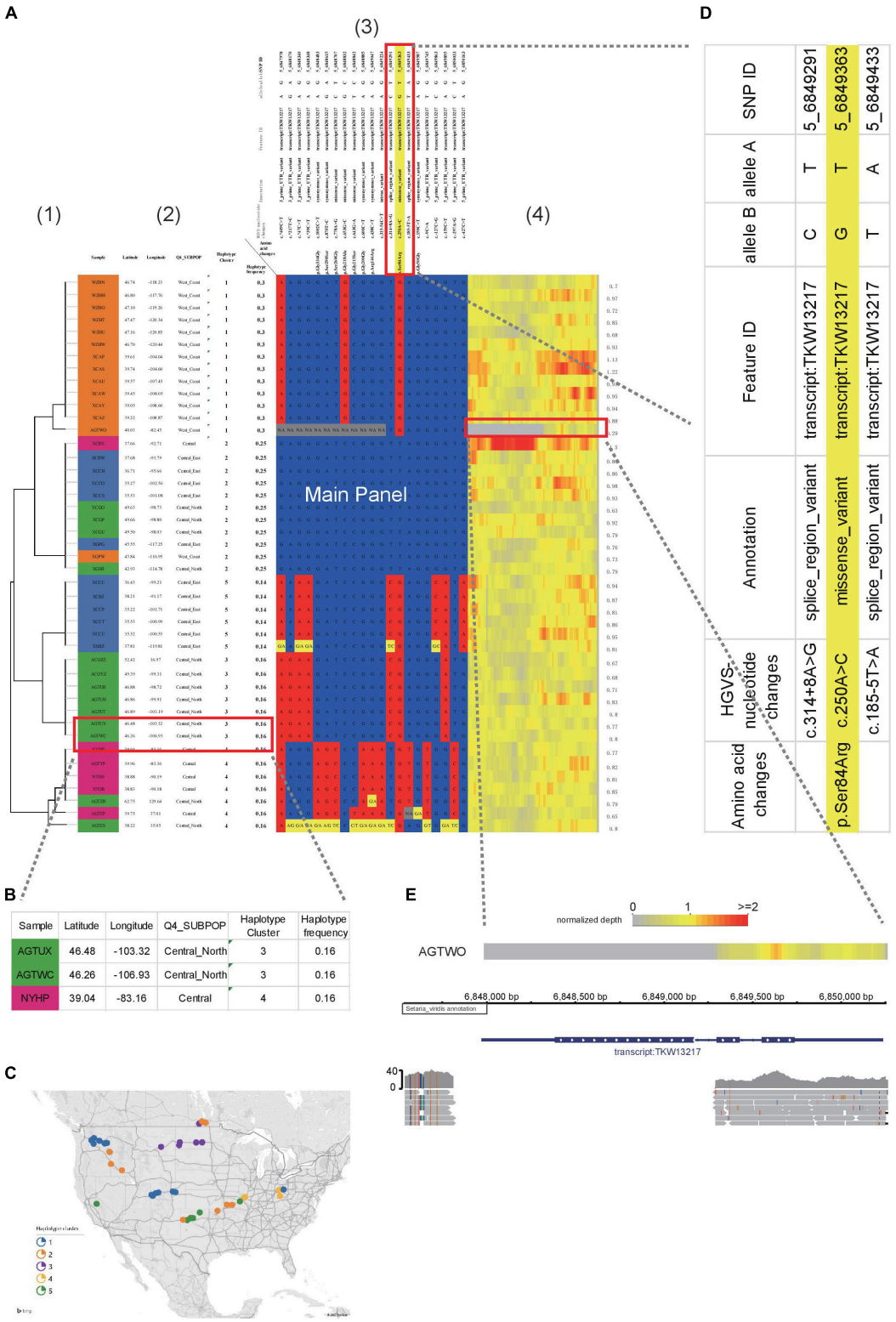
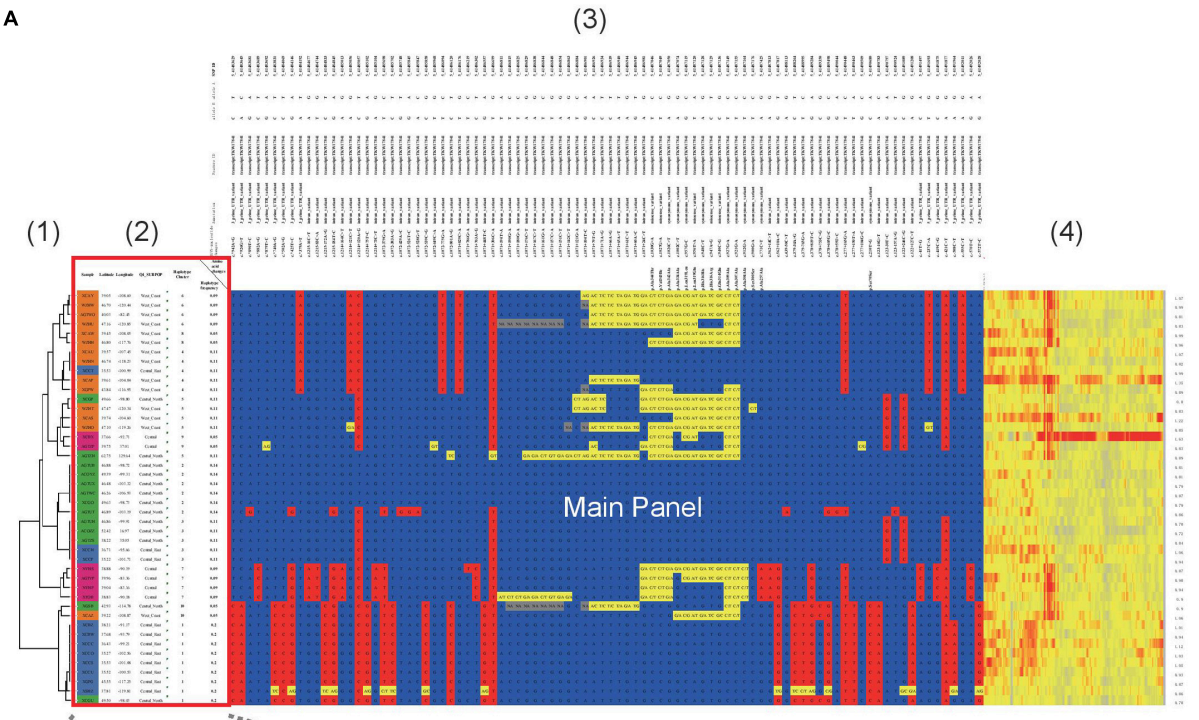


FIGURE 2 | Example of HAPPE on the display of *S. viridis* gene Sevir.5G085400. **(A)** Haplotypes of the *S. viridis* gene Sevir.5G085400, which shows additional information, including (1) a tree, (2) descriptions of samples, (3) annotations of SNPs and (4) a heatmap of read depth. **(B)** Magnified view of Panel 2 showing the descriptions of the samples. **(C)** A map generated using the geographical coordinates shown in the table. **(D)** Zoomed-in view of Panel 3 showing the annotations of SNPs. **(E)** A deletion event shown by the heatmap of read depth in Panel 3, which was further confirmed in a genome browser. More details can be found in **Supplementary Tables 1–3**.



Sample	Latitude	Longitude	Q4_SUBPOP	Haplotype Cluster	Haplotype frequency
XCAY	39.05	-108.60	West_Coast	6	0.09
WZHW	46.70	-120.44	West_Coast	6	0.09
AGTWO	40.03	-82.45	West_Coast	6	0.09
WZHU	47.16	-120.85	West_Coast	6	0.09
XCAW	39.45	-108.05	West_Coast	8	0.05
WZHH	46.80	-117.76	West_Coast	8	0.05
XCAU	39.57	-107.45	West_Coast	4	0.11
WZHN	46.74	-118.23	West_Coast	4	0.11
XCCT	35.53	-100.99	Central_East	4	0.11
XCAP	39.61	-104.04	West_Coast	4	0.11
XGPW	43.84	-116.95	West_Coast	4	0.11
XCGP	49.66	-98.80	Central_North	5	0.11
WZHT	47.47	-120.34	West_Coast	5	0.11
XCAS	39.74	-104.60	West_Coast	5	0.11
WZHO	47.10	-119.26	West_Coast	5	0.11
XCBX	37.66	-92.71	Central	9	0.05
AGTZP	39.75	37.01	Central	9	0.05
AGTZH	62.75	129.64	Central_North	5	0.11
AGTUB	46.88	-98.72	Central_North	2	0.14
ACOYZ	49.39	-99.31	Central_North	2	0.14
AGTUX	46.48	-103.32	Central_North	2	0.14
AGTWC	46.26	-106.93	Central_North	2	0.14
XCGO	49.63	-98.73	Central_North	2	0.14
AGTUT	46.89	-103.19	Central_North	2	0.14
AGTUH	46.86	-99.91	Central_North	3	0.11
ACOZZ	52.42	16.97	Central_North	3	0.11
AGTZS	38.22	35.05	Central_North	3	0.11
XCCH	36.71	-95.66	Central_East	3	0.11
XCCP	35.22	-101.71	Central_East	3	0.11
NYHS	38.88	-90.19	Central	7	0.09
AGTYP	39.96	-83.36	Central	7	0.09
NYHP	39.04	-83.16	Central	7	0.09
XTOB	38.83	-90.18	Central	7	0.09
XGSB	42.93	-114.78	Central_North	10	0.05
XCAZ	39.22	-108.87	West_Coast	10	0.05
XCBZ	38.21	-91.17	Central_East	1	0.2
XCBW	37.68	-93.79	Central_East	1	0.2
XCCC	36.43	-99.21	Central_East	1	0.2
XCCO	35.27	-102.56	Central_East	1	0.2
XCCS	35.53	-101.08	Central_East	1	0.2
XCCU	35.52	-100.53	Central_East	1	0.2
XGPG	45.55	-117.25	Central_East	1	0.2
XSHZ	37.81	-119.81	Central_East	1	0.2
XCGU	49.50	-98.03	Central_North	1	0.2

FIGURE 3 | Example of HAPPE on the display of the *S. viridis* gene Sevir.5G394700. **(A)** Haplotypes of the *S. viridis* gene Sevir.5G394700 with additional information, including (1) a tree, (2) descriptions of samples, (3) annotations of SNPs and (4) a heatmap of read depth. **(B)** Zoomed-in view of Panel 2 showing the details of the samples. **(C)** Distribution of haplotypes of the gene in different subgroups. More details can be found in **Supplementary Tables 4–7**.

changed considerably. With the continuous reduction in DNA sequencing prices, whole-genome sequencing is now routinely applied in the works of many clinicians and breeders who are not well-versed in the coding and visualization of datasets. Instead, these researchers favor the use of Excel more than R, Python or any other coding languages for data analysis and visualization. In recent years, more emphasis has been placed on deep collaborations between wet and dry biologists. In these collaborations, the bioinformaticians can help visualize the datasets. However, these visuals are usually generated in an image format, such as png and pdf. Thus, it is difficult for the readers to identify more details, particularly when datasets are large. Although the plots can also be generated into svg format allowing infinite zooming in, the real-time data cannot be shown in the plot either, which potentially hampers the communication and collaboration between different teams. It is occasionally necessary to select a subset of the data for further exploration, but the users have to return to the raw tables to access the data of interest, which could be impossible for large datasets. The tool described in this work reads VCF datasets and outputs the plots consisting of editable cells in Excel tables, enabling the readers to access the data directly from the plot. The additional information can also be displayed in the extension regions (**Figures 2**, **3**) either by setting corresponding options in the commands or editing the table directly in Excel tables, so all the information of samples or the annotation of SNPs can be shown in the plot and table. This feature is very friendly to the non-programmer readers.

Code Availability and Usage of HAPPE

The application of HAPPE depends on the preinstallation of several tools, including bcftools and bgzip, the path of which can be given in a config file. HAPPE reads genotype datasets in compressed VCF format and selects the samples according to a list provided to the parameter of "-k" and the sites according to the list provided to "-r" or "-s." The description of samples and the annotation of SNPs can be provided optionally through the parameters "-i" and "-I," respectively. The users also have the option to exclusively keep the SNPs in the coding or non-coding regions or those resulting in changes in amino acids by selecting a parameter among "-f," "-n" and "-x." More details of usage can be found on the page of HAPPE[3]. The code of HAPPE is available on github (see text footnote 3) and Python Package Index (PyPI)[4] and can be installed using the pip command. It should be noted that the number of samples and sites that can be displayed by HAPPE are constrained by the limitation of the Excel table, which allows only 18,278 columns and 1,048,576 rows at most. Thus, the number of SNPs and samples to be shown should be less than 18,278 and 1,048,576, respectively.

[3] https://github.com/fengcong3/HAPPE

[4] https://pypi.org/project/HAPPE/

CONCLUSION

HAPPE produces haplotype plots consisting of editable cells of Excel tables with customized extension information to help readers understand more details of the data shown in the plot. The output, an Excel table, is user-friendly to non-programmer readers and users and facilitates efficient communication and collaboration between different teams. Given the application of WGS in routine research due to the decreasing sequencing price, we believe that HAPPE will be widely used in various studies and collaborations.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SC and JY conceptualized the work and revised the manuscript. CF, XW, and SW coded HAPPE. BS, WN, XW, and SW tested the tool. BS and CF wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by National Key Research and Development Program of China (2019YFA0707003), Guangdong Innovation Research Team Fund (Grant No. 2014ZT05S078), Guangdong "ZhuJiang Talent Innovation" project (2019ZT08N628), and NSFC "Excellent Young Talent" (32022006). Projects subsidized by special funds for Science Technology Innovation and Industrial Development of Shenzhen Dapeng New District (Grant No. PT202101-01).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.927407/full#supplementary-material

REFERENCES

Bonfield, J. K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., et al. (2021). HTSlib: c library for reading/writing high-throughput sequencing data. *GigaScience* 10:giab007. doi: 10.1093/gigascience/giab007

Cheng, H., Liu, J., Wen, J., Nie, X., Xu, L., Chen, N., et al. (2019). Frequent intra- and inter-species introgression shapes the landscape of genetic variation in bread wheat. *Genome Biol.* 20:136. doi: 10.1186/s13059-019-1744-x

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of Drosophila

melanogaster strain w ^{1118}; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008. doi: 10.1093/gigascience/giab008

Hao, C., Jiao, C., Hou, J., Li, T., Liu, H., Wang, Y., et al. (2020). Resequencing of 145 Landmark Cultivars Reveals Asymmetric Sub-genome Selection and Strong Founder Genotype Effects on Wheat Breeding in China. *Mol. Plant* 13, 1733–1751. doi: 10.1016/j.molp.2020.09.001

Jäger, G., Peltzer, A., and Nieselt, K. (2014). inPHAP: interactive visualization of genotype and phased haplotype data. *BMC Bioinform.* 15:200. doi: 10.1186/1471-2105-15-200

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., et al. (2020). A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci. *Nat. Biotechnol.* 38, 1203–1210. doi: 10.1038/s41587-020-0681-2

Marks, R. A., Hotaling, S., Frandsen, P. B., and VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nat. Plants* 7, 1571–1578. doi: 10.1038/s41477-021-01031-8

Pedersen, B. S., and Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* 34, 867–868. doi: 10.1093/bioinformatics/btx699

San Lucas, F. A., Rosenberg, N. A., and Scheet, P. (2012). Haploscope: a tool for the graphical display of haplotype structure in populations. *Genet. Epidemiol.* 36, 17–21. doi: 10.1002/gepi.20640

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., and Guo, L. (2022). Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.* 27, 391–401. doi: 10.1016/j.tplants.2021.10.006

The International HapMap Consortium (2003). The International HapMap Project. *Nature* 426, 789–796. doi: 10.1038/nature02168

Todesco, M., Owens, G. L., Bercovich, N., Légaré, J.-S., Soudi, S., Burge, D. O., et al. (2020). Massive haplotypes underlie ecotypic differentiation in sunflowers. *Nature* 584, 602–607. doi: 10.1038/s41586-020-2467-6

Tollefson, G. A., Schuster, J., Gelin, F., Agudelo, A., Ragavendran, A., Restrepo, I., et al. (2019). VIVA (VIsualization of VAriants): a VCF File Visualization Tool. *Sci. Rep.* 9:12648. doi: 10.1038/s41598-019-49114-z

Varshney, R. K., Roorkiwal, M., Sun, S., Bajaj, P., Chitikineni, A., Thudi, M., et al. (2021). A chickpea genetic variation map based on the sequencing of 3,366 genomes. *Nature* 599, 622–627. doi: 10.1038/s41586-021-04066-1

Zhao, Y.-P., Fan, G., Yin, P.-P., Sun, S., Li, N., Hong, X., et al. (2019). Resequencing 545 ginkgo genomes across the world reveals the evolutionary history of the living fossil. *Nat. Commun.* 10:4201. doi: 10.1038/s41467-019-12133-5

Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., et al. (2020). Triticum population sequencing provides insights into wheat adaptation. *Nat. Genet.* 52, 1412–1422. doi: 10.1038/s41588-020-00722-w

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.