HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

APPLICATION NOTE

AncestryPainter: A Graphic Program for Displaying Ancestry Composition of Populations and Individuals

Qidi Feng [1,2,#,a], Dongsheng Lu [1,#,b], Shuhua Xu [1,2,3,4,5,6,*,c]

[1] CAS Key Laboratory of Computational Biology, Max Planck Independent Research Group on Population Genomics, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[4] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China
[5] Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China
[6] Human Phenome Institute, Fudan University, Shanghai 201203, China

Received 25 February 2018; revised 13 April 2018; accepted 9 May 2018
Available online 22 November 2018

Handled by Yu Xue

KEYWORDS

Graphic program;
Admixed populations;
Ancestry composition;
AncestryPainter;
Admixture proportions

Abstract **Ancestry composition** of populations and individuals has been extensively investigated in recent years due to advances in the genotyping and sequencing technologies. As the number of populations and individuals used for ancestry inference increases remarkably, say more than 100 populations or 1000 individuals, it is usually challenging to present the ancestry composition in a traditional way using a rectangular graph. To address this issue, we developed a program, ***AncestryPainter***, which can illustrate the ancestry composition of populations and individuals with a rounded and nice-looking graph to save space. Individuals are depicted as length-fixed bars partitioned into colored segments representing different ancestries, and the population of interest can be highlighted as a pie chart in the center of the circle plot. In addition, *AncestryPainter* can also be applied to display personal ancestry in a way similar to that for displaying population ancestry. *AncestryPainter* is publicly available at http://www.picb.ac.cn/PGG/resource.php.

Introduction

Ancestry composition of individuals indicates the average proportion of contributing ancestries to their entire genomes [1]. Global ancestry for individuals is frequently inferred in population genetics to understand the genetic relationship

[*] Corresponding author.
E-mail: xushua@picb.ac.cn (Xu S).
[#] Equal contribution.
[a] ORCID: 0000-0002-8012-7698.
[b] ORCID: 0000-0002-8348-5159.
[c] ORCID: 0000-0002-1975-1002.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2018.05.002

and history of gene flow of populations [2,3], figure out ancestry composition of admixed populations [4–6], or detect population stratification [7,8]. Ancestry of a population or an individual can be estimated in different ways, through simple calculation based on allele frequency of the ancestral source populations with K-mean clustering, or using model-based methods such as *ADMIXTURE* [1], *structure* [9], and *frappe* [2,10]. To use these methods, it is assumed that the observed individuals are drawn from a population with contributions from K ancestries. Then the programs estimate the ancestry proportions contributed by K ancestral source populations, where K is a potential number of ancestral source populations. In theory, the sum of the ancestry proportions of each individual attributed to K ancestral source populations is expected to be 100%. However, it does not necessarily mean that all of the ancestral source populations truly contribute ancestry to the target individual or population and statistical tests can be applied to estimate the significance of ancestry contributions.

Several computational tools are available to graphically display the estimated ancestry proportions of individuals or populations [9,11]. These programs display the ancestry of an individual as a fixed-length line segment partitioned into K colored components, with the length of each colored component corresponding to the proportion of each ancestry. However, all these programs align individuals in a rectangular graph, which is not able to accommodate a large number of individuals or populations in a single print page, thus making it difficult to publish these results. To solve this problem, we developed *AncestryPainter*, a computational program for displaying the ancestry composition of individuals and populations in a circular graph, and highlighting the ancestry composition of individuals that are of particular interest to the users in the center of the circle as a pie chart (**Figure 1**). Moreover, *AncestryPainter* automatically aligns populations based on the proportions of their representative ancestries, *i.e.*, the ancestry accounting for the largest proportion in the population. Therefore, the graph generated by *AncestryPainter* is not only nice-looking, but also efficient in ancestry visualization.

Method

The graphic program of *AncestryPainter* is written in Perl. By running the Perl program, an R script is produced based on the input ancestry composition of each population and then automatically run to plot the figure. The final figure is a circle with fixed perimeter and radius by automatically adjusting the relative space of each input population. It is of note that in *AncestryPainter*, for the purpose of presentation, same space is allocated to each population, regardless of the number of individuals included in each population.

AncestryPainter automatically sorts the populations/individuals according to ancestry proportions when displaying the figure. Firstly, the ancestry that accounts for the largest proportion in the population is determined (defined as the representative ancestry of the population) and the input populations are categorized into K groups accordingly. Then, populations in each of the K groups are plotted in a descending order of the representative ancestry proportions. Individuals within each population are also sorted based on a descending order of its representative ancestry proportion.

Implementation

AncestryPainter requires users to provide a matrix of ancestry proportions of each individual as input data, which can be obtained from any analysis of personal or population ancestry. The program also requires a predefined group identifier of each individual to classify and label the individuals in the graph. The predefined group identifier could be population label, sampling location, or phenotypic classification. For instance, *ADMIXTURE* output file [1] that includes a matrix of ancestry proportions (.Q file) can be used directly as the input data of *AncestryPainter*. In theory, ancestry proportions obtained using any approach could be provided as the input data for *AncestryPainter* as long as the proportions across ancestral populations sum to one, and the input format is modified to meet software requirements. Users can run *AncestryPainter* with a command line under Linux, Mac, or Windows operating systems. It produces and automatically runs an R script to plot a figure file for displaying personal or population ancestry composition. The output of *AncestryPainter* includes a figure (pdf or png), the R script, the color scheme used for plotting the figure, and a file containing the ancestry proportions as well as population label of each individual.

Display options and features

Figure 1 exemplifies the display for the ancestry composition of global populations using *AncestryPainter* with the Uyghur population highlighted in the center. Dataset used to generate Figure 1 included 2345 individuals representing 203 populations from the Human Origins dataset [12]. The ancestry proportions of each population were obtained using *ADMIXTURE* program, assuming 8 pseudo ancestries (K = 8). Default color scheme was used in the figure and populations were automatically sorted according to admixture proportions. As shown in Figure 1, Uyghurs are composed of four major ancestries, with representative ancestral source populations from West Eurasia, South Asia, East Asia, and Siberia, respectively. *AncestryPainter* allows users to freely modify the default settings such as the color scheme, the order of populations/individuals, and what populations/individuals to be included with command options. Users can also specify a certain population or individual to be highlighted in the center of the graph, or plot without any population or individual highlighted in the center. If there are too many populations to present, users can remove the black lines between populations for a better display.

In addition to displaying ancestry composition of individuals, *AncestryPainter* could be applied more extensively to illustrate any data that can be presented in a bar plot. As long as the assigned proportions for all items sum to 100%, the data can be displayed in a circular graph in the same way as shown for the personal or population ancestry composition.

Future developments

In this study, we developed *AncestryPainter*, a computational program that can be used to illustrate the ancestry compositions of populations and individuals with a rounded and nice-looking graph in a spatially-efficient way. In the future, we will

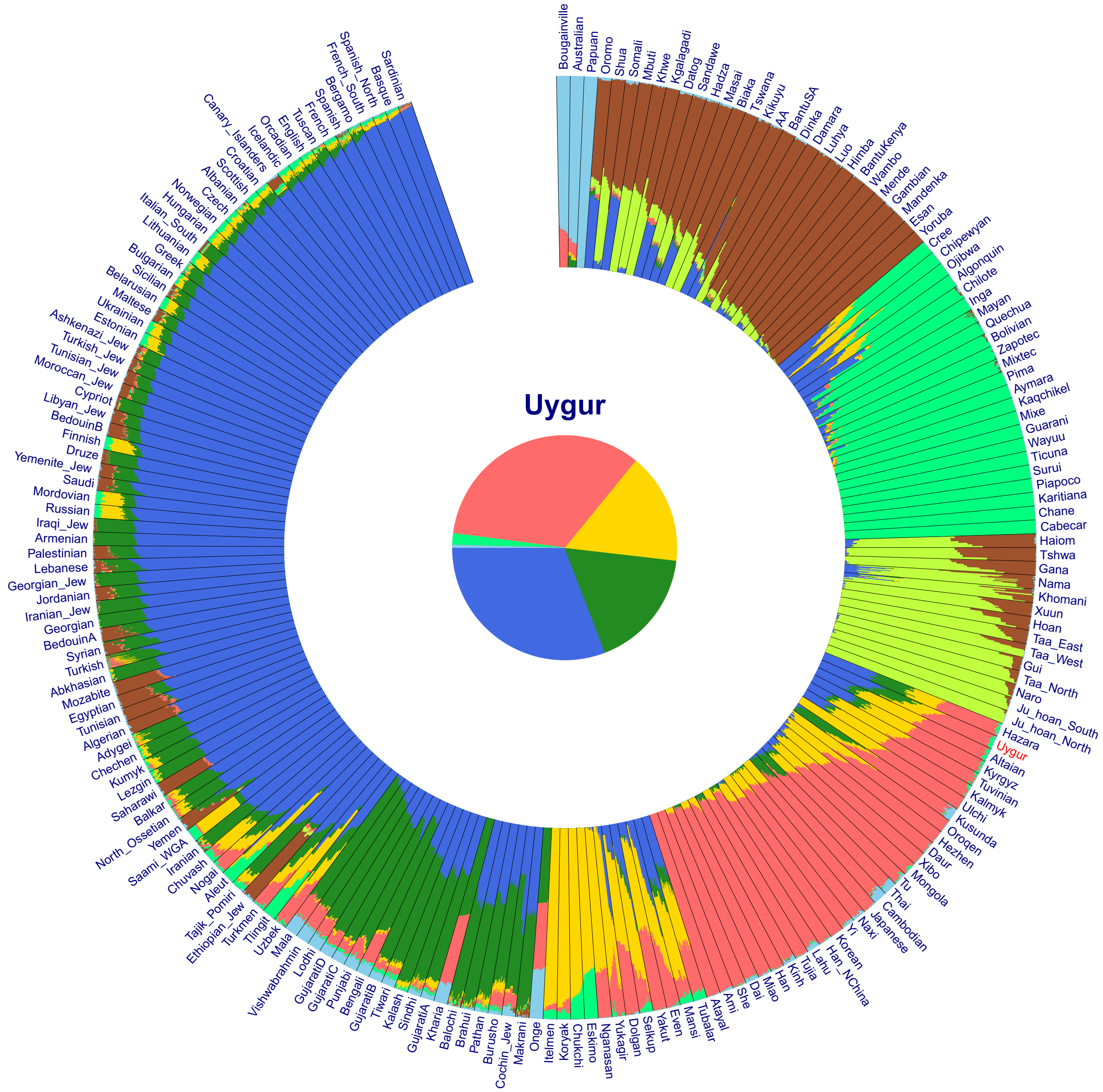

Figure 1 An example graph taken from the output of *AncestryPainter*

The input of this figure is produced by running *ADMIXTURE* assuming eight ancestral source populations ($K = 8$) based on the Human Origins datasets [12] including Africans, Americans, Oceanians, West Eurasians, South Asians, Central Asians/Siberians, and East Asians. The Uyghur population is highlighted in the center of the graph. The command line used for this result is "perl AncestryPainter.pl -i data.ind -q data.Q -t Uygur". This example is also included in the *AncestryPainter* program package, which can be downloaded at http://www.picb.ac.cn/PGG/resource.php.

implement more functions, for instance, to allow displaying ancestry compositions of populations assuming different number of ancestral populations (multiple Ks) in one image using multiple concentric circles, so that ancestry compositions at different levels could be presented in an explicit and efficient way.

Authors' contributions

SX conceived and designed the study. DL and QF developed the program and wrote the computer code. QF drafted the manuscript and SX revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This study was supported by the Strategic Priority Research Program (Grant No. XDB13040100) and Key Research Program of Frontier Sciences (Grant No. QYZDJ-SSW-SYS009) of the Chinese Academy of Sciences, the National Natural Science Foundation (Grant Nos. 91731303, 31771388, and 31711530221), the National Science Fund for Distinguished Young Scholars (Grant No. 31525014), the Program of Shanghai Academic Research Leader (Grant No. 16XD1404700), the National Key R&D Program (Grant No. 2016YFC0906403), and the Shanghai Municipal Science and Technology Major Project (Grant No. 2017SHZDZX01), China. SX is Max-Planck Independent Research Group Leader and member of CAS Youth Innovation Promotion Association. SX also gratefully acknowledges the support of the "Ten-Thousand Talents" Program for Young Top-notch Talents, China. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–64.

[2] Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319:1100–4.

[3] Consortium HP-AS, Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, et al. Mapping human genetic diversity in Asia. Science 2009;326:1541–5.

[4] Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, et al. Genetic history of Xinjiang's Uyghurs suggests bronze age multiple-way contacts in Eurasia. Mol Biol Evol 2017;34:2572–82.

[5] Lu D, Lou H, Yuan K, Wang X, Wang Y, Zhang C, et al. Ancestral origins and genetic history of Tibetan highlanders. Am J Hum Genet 2016;99:580–94.

[6] Zhang C, Lu Y, Feng Q, Wang X, Lou H, Liu J, et al. Differentiated demographic histories and local adaptations between Sherpas and Tibetans. Genome Biol 2017;18:115.

[7] Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet 2009;85:775–85.

[8] Xu S, Yin X, Li S, Jin W, Lou H, Yang L, et al. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet 2009;85:762–74.

[9] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–59.

[10] Alkan C, Kavak P, Somel M, Gokcumen O, Ugurlu S, Saygi C, et al. Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. BMC Genomics 2014;15:963.

[11] Rosenberg NA. distruct: a program for the graphical display of population structure. Mol Ecol Resour 2003;4:137–8.

[12] Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 2014;513:409–13.