VCF2PopTree: a one-click client-side software to construct population phylogeny from genome-wide SNPs

Sankar Subramanian[1*], Umayal Ramasamy[1,2] and David Chen[2]

[1]*GeneCology Research Centre, The University of the Sunshine Coast, 90 Sippy Downs Drive, Sippy Downs Qld 4556, Australia*

[2]*School of Information and Communication Technology, Griffith University, 170 Kessels Road, Nathan, Qld 4111, Australia*

[*] Corresponding author

Address for correspondence:

GeneCology Research Centre
University of the Sunshine Coast
90 Sippy Downs Drive
Sippy Downs QLD 4556
Australia
Phone: + 61-7-5430 2873
Fax: +61-7- 5430 2881
E-mail: ssankara@usc.edu.au

Running head: VCF2PopTree

Keywords: VCF, Phylogeny, UPGMA, Neighbour-Joining, MEGA and PHYLIP

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27682v1 | CC BY 4.0 Open Access | rec: 25 Apr 2019, publ: 25 Apr 2019

Abstract

In the past decades a number of software programs have been developed to deduce the phylogenetic relationship between populations. However, these programs are not suited for large-scale whole genome data. Recently, a few standalone or web applications have been developed to handle genome-wide data, but they were either computationally intensive, dependent on third party software or required significant time and resource of a web server. In the post-genomic era, researchers are able to obtain bioinformatically processed high-quality publication-ready whole genome data for many individuals in a population from next generation sequencing companies due to the reduction in the cost of sequencing and analysis. Such genotype data is typically presented in the Variant Call Format (VCF) and there is no simple software available that uses this data to construct the phylogeny of populations in a short time. To address this limitation, we have developed a one-click user-friendly software, *VCF2PopTree* that uses gnome-wide SNPs to construct and display phylogenetic trees in seconds to minutes. For example, it reads a 1 GB VCF file and draws a tree in less than 5 minutes. *VCF2PopTree* accepts genotype data from a local machine, constructs a tree using UPGMA and Neighbour-Joining algorithms and displays it on a web-browser. It also produces pairwise-diversity matrix in MEGA and PHYLIP file formats as well as trees in the *Newick* format which could be directly used by other popular phylogenetic software programs. The software including the source code, a test VCF input file and short documentation are available at: http://sankarsubramanian.net/dat/index.html.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27682v1 | CC BY 4.0 Open Access | rec: 25 Apr 2019, publ: 25 Apr 2019

1. Introduction

One of the major tasks in genetics and evolutionary biology is to deduce the ancestral relationship between populations and species. For this purpose, a number of mathematical and statistical algorithms have been developed. To implement these algorithms, computationally efficient software programs were developed. However, these software such as *MEGA* (Kumar, Stecher and Tamura, 2016), *PHYLIP* (Felsenstein, 2005), *PAUP* (Wilgenbusch and Swofford, 2003), *RaxML* (Stamatakis, 2006) and *BEAST* (Drummond, *et al.*, 2012) are suited only for sequence-based alignments. With the advent of the next generation sequence techniques, large-scale whole genome data containing millions of Single Nucleotide Variations (SNVs) are generated for populations. The whole genome data is typically presented in the Variant Call Format (VCF) and there was a need for genetic software to construct population phylogeny using these large-scale data. To address this limitation a number of software programs have been developed in the recent past. However, these programs were either computationally intensive and time consuming, heavily dependent on third party software or required significant time and resource of a web server.

Computer programs that handle genome-wide VCF data are either standalone or web server-based applications. Typically, standalone applications such as *SNPhylo* (Lee, et al., 2014), *VCF-Kit* (Cook and Andersen, 2017), and *VCFtoTree* (Xu, *et al.*, 2017) are software pipelines that need to be installed in a local computer. Furthermore, these programs are dependent on a series of other software such as *bwa, samtools* and/or *MUSCLE*. Therefore, an adequate level of computer expertise required to implement and run the standalone programs. On the other hand, web server-based programs such as *SNiPlay* (Dereeper*, et al.*, 2015), and *CSI Phylogeny* (Kaas*, et al.*, 2014) take significant amount of time to produce a tree. This is partly due to the time taken to upload the large-data set to server from the user's local machine, which depends on the web traffic and internet speed. Furthermore, both standalone and server-based applications perform a series of data processing steps through software pipelines, which also cause significant time delay.

Due to the reduction in the cost of sequencing and bioinformatic analysis, it is now possible to obtain the processed whole genome data for many individuals. Using standard bioinformatic data processing pipelines most of the sequencing service providers deliver high quality publication-ready genotype data for whole genomes in the form of VCF files. Hence population geneticists now need a simple program that reads this data in VCF files and construct a phylogenetic tree in a short time as there is no need of any data processing routines. Therefore, the current study is aimed to the address this important limitation in genomic research. Hence, we developed a JavaScript based client-side software to infer phylogenetic relationship using genome-wide SNP data.

Methods

Implementation

The software, *VCF2PopTree* was written in JavaScript, which runs purely within the user's computer/browser. The input VCF file typically contains the genotype information which are coded by '0's and '1's, is read by *VCF2PopTree* from the user's local computer. The genotype data should be at least from four individuals in order to build a tree. *VCF2PopTree* is designed to read and process the input data line-by-line so it is able to handle large data files without running out of memory. Using the genotype data, the normalised counts of nucleotide differences for all possible pairwise combinations are computed. The pairwise divergence matrix is then used to infer the phylogenetic relationship using the *UPGMA* (Sokal and Michener, 1958) and *Neighbour-Joining* (Saitou and Nei, 1987) algorithms and the resulting tree is presented in the popular *Newick* or parenthetical format. The *Newick* formatted phylogeny is used to draw the tree on the browser using the JavaScript package, d3.phylogram.js.

Features

VCF2PopTree performs three different tasks namely: construct and draws a phylogenetic tree, produce a tree file and generates pairwise diversity matrix. There are two radio buttons to infer phylogenetic relationship between populations using UPGMA and Neighbour-Joining algorithms and the latter method produces an unrooted tree (Figure 1). Two more radio buttons are provided to draw the phylogenetic tree in a rectangular or circular style. Apart from drawing trees *VCF2PopTree* also produces the tree file in the popular *newick* format by checking the radio button "Newick format". Finally, this program produces pairwise diversity matrix in the popular MEGA and PHYLIP formats and the last two radio buttons should be used for this purpose respectively.

Results and Discussion

Performance

VCF2PopTree is a simple and straight forward program to use, which requires only a single click to view the phylogeny of a population using the default settings. *VCF2PopTree* is designed to run on personal computers with moderate specification. To display a phylogenetic tree, it takes a few seconds to minutes depending on the size of the data as well as the number of samples/individuals. For example, it takes only 26 seconds and 92 seconds to display the phylogeny of 10 and 22 individuals respectively by reading the data from VCF files of 0.25 GB and 0.5 GB respectively using a *Windows* computer with 4GB RAM and *Intel Core i5* processor. The display time was 4.52 minutes for a 1 GB file with 56 genomes using a 16GB RAM computer. Clearly this suggests that for reading large files the limitation is only the RAM. *VCF2PopTree* is compatible with all population browsers including *Chrome*, *Opera*, *Edge and Firefox* and works equally efficient in *Mac*, *Windows* and *Linux* (*Ubuntu*). Furthermore, it displays the tree in a mobile phone (*iPhone* and *Android*) if the input file size is small.

Usage Example

Since this is a client-side software the http://sankarsubramanian.net/dat/VCF2PopTree.html has to be downloaded to the local computer (by right clicking and selecting "Save link as") for faster performance. However, the program also works by directly clicking the link, which will be slow as the VCF file has to be uploaded to the server. To examine the functionality of the software we obtained a VCF file (test.vcf) from the Simons Genome Project containing about half a million SNPs from ten human populations (Mallick*, et al.*, 2016). To test the software both http://sankarsubramanian.net/dat/VCF2PopTree.html and http://sankarsubramanian.net/dat/test.vcf.gz have to be downloaded and the test.vcf.gz has to be unzipped. To infer the phylogenetic relationship between the 10 populations the pairwise diversity matrix needs to be computed. Since the diversities are expressed on a *per site* basis, the total genome size has to be provided in the "Enter the genome size" text box. By default, the human genome size has been entered in this box. Once the size of the genome in base pairs in entered one of the radio buttons for "UPGMA" or "Neighbour-Joining tree" and that of "Rectangular tree" or "circular tree" should be selected. After selecting the options to construct a phylogenetic tree, the "choose file" button should be clicked to open a VCF file, which in turn reads and displays the tree on the browser as shown in Figure 1A. Note that an alert window will pop-up if any file other than a VCF was selected.

To obtain the tree file in the *newick* format, radio buttons for "UPGMA" or "Neighbour-Joining tree" and that of "Newick tree format" have to be selected. This will present the parenthetical formatted tree in a text area (Figure 1B). In order to obtain the pairwise diversity matrix, radio buttons "Pair-wise diversity - MEGA" or "Pair-wise diversity - PHYLIP" need to be selected. This will produce the pairwise diversity matrix in MEGA (Kumar*, et al.*, 2016) or PHYLIP (Felsenstein, 2005) formats in a text area (Figure 1C). The pairwise diversity matrix could be copied and pasted on to a text file, which could then be used as an input for programs such as MEGA, PHYLIP or any other software that accepts these formats. Hence users are now able to use MEGA and other popular gene specific software to edit or manipulate trees based on whole genome data. Similarly, the whole genome based *newick* tree generated by *VCF2PopTree* could be further manipulated by the tree editing software such as *TreeGraph* (Stover and Muller, 2010) or *FigTree* (http://tree.bio.ed.ac.uk/software/figtree/).

Conclusions

VCF2PopTree is unique with respect to handling whole genome data from populations and it reads data directly from the local machine and is independent of operating systems and browsers. Importantly, this program does not require high performance computational resources, third party software tools, a web server or internet connectivity. It is the ultrafast software available at present to deduce and draw population phylogeny in seconds to minutes. Apart from building and displaying trees this program also produces distant matrix and *Newick* trees and thus facilitates whole genome based phylogenetic analysis through other popular software. Therefore, *VCF2PopTree* could be a valuable phylogenetic tree building software for researchers and students in the fields of Genetics, Ecology, Evolutionary Biology and Medicine.

Funding

This project was funded by a Linkage grant awarded to SS by the Australian Research Council (LP160100594).

Acknowledgements

The authors are thankful to all who tested the various versions of this software.

Conflict of Interest

None declared

References

COOK DE, ANDERSEN EC. 2017 VCF-kit: assorted utilities for the variant call format. Bioinformatics. 33(10):1581-1582.

DEREEPER A, HOMA F, ANDRES G, SEMPERE G, SARAH G, HUEBER Y, DUFAYARD JF, RUIZ M. 2015 SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. Nucleic Acids Res. 43(W1):W295-300.

DRUMMOND AJ, SUCHARD MA, XIE D, RAMBAUT A. 2012 Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 29(8):1969-1973.

FELSENSTEIN J. 2005 Phylogeny Inference Package (Version 3.2). Cladistics. 5:164-166.

KAAS RS, LEEKITCHAROENPHON P, AARESTRUP FM, LUND O. 2014 Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One. 9(8):e104984.

KUMAR S, STECHER G, TAMURA K. 2016 MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 33(7):1870-1874.

LEE TH, GUO H, WANG X, KIM C, PATERSON AH. 2014 SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics. 15:162.

MALLICK S, LI H, LIPSON M, MATHIESON I, GYMREK M, RACIMO F, ZHAO M, CHENNAGIRI N, NORDENFELT S, TANDON A, SKOGLUND P, LAZARIDIS I, SANKARARAMAN S, FU Q, ROHLAND N, RENAUD G, ERLICH Y, WILLEMS T, GALLO C, SPENCE JP, SONG YS, POLETTI G, BALLOUX F, VAN DRIEM G, DE KNIJFF P, ROMERO IG, JHA AR, BEHAR DM, BRAVI CM, CAPELLI C, HERVIG T, MORENO-ESTRADA A, POSUKH OL, BALANOVSKA E, BALANOVSKY O, KARACHANAK-YANKOVA S, SAHAKYAN H, TONCHEVA D, YEPISKOPOSYAN L, TYLER-SMITH C, XUE Y, ABDULLAH MS, RUIZ-LINARES A, BEALL CM, DI RIENZO A, JEONG C, STARIKOVSKAYA EB, METSPALU E, PARIK J, VILLEMS R, HENN BM, HODOGLUGIL U, MAHLEY R, SAJANTILA A, STAMATOYANNOPOULOS G, WEE JT, KHUSAINOVA R, KHUSNUTDINOVA E, LITVINOV S, AYODO G, COMAS D, HAMMER MF, KIVISILD T, KLITZ W, WINKLER CA, LABUDA D, BAMSHAD M, JORDE LB, TISHKOFF SA, WATKINS WS, METSPALU M, DRYOMOV S, SUKERNIK R, SINGH L, THANGARAJ K, PAABO S, KELSO J, PATTERSON N, REICH D. 2016 The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 538(7624):201-206.

SAITOU N, NEI M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406-425.

SOKAL R, MICHENER C. 1958 A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin. 38:1409–1438.

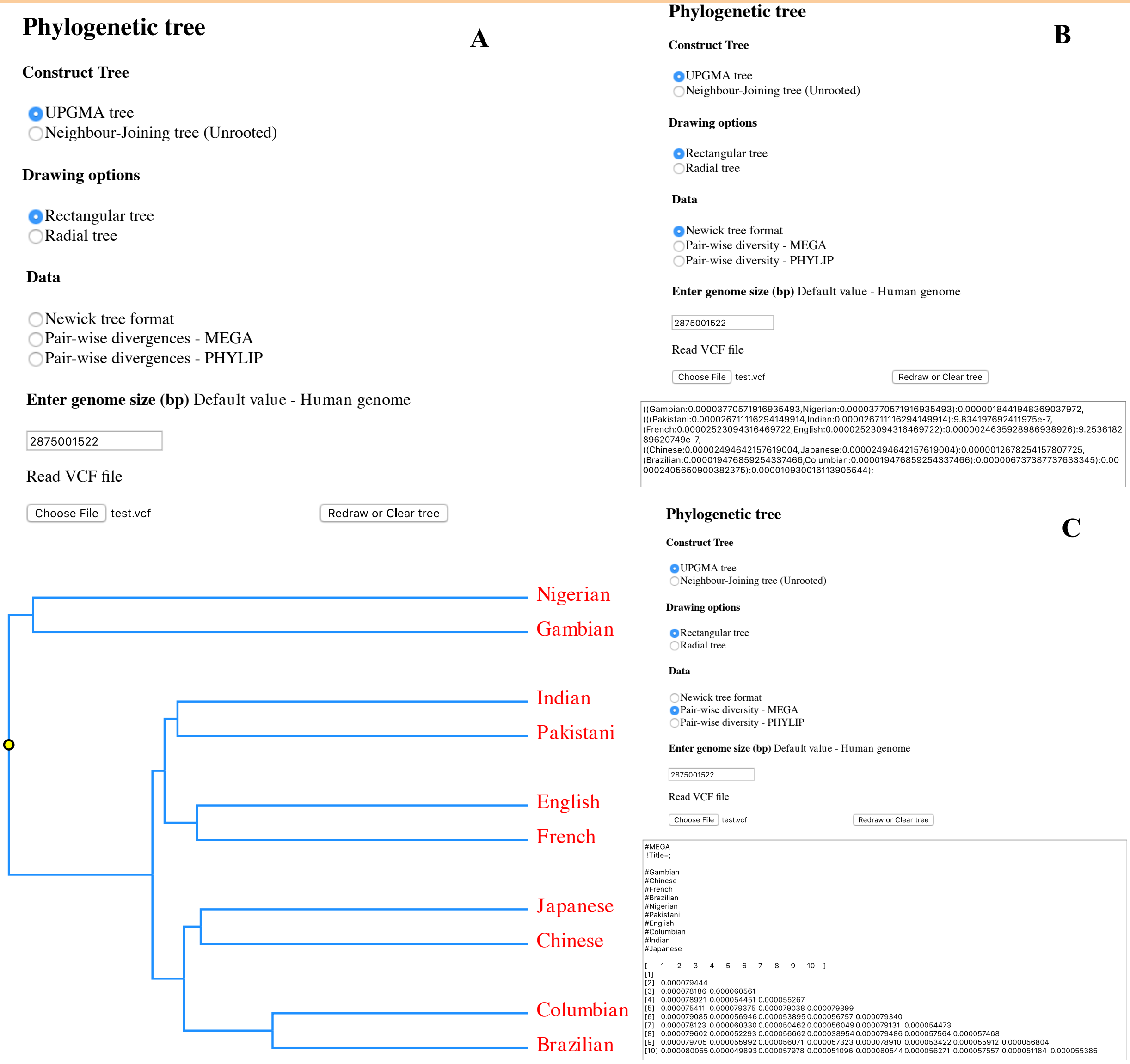
STAMATAKIS A. 2006 RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 22(21):2688-2690.

STOVER BC, MULLER KF. 2010 TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 11:7.

WILGENBUSCH JC, SWOFFORD D. 2003 Inferring evolutionary trees with PAUP*. Curr Protoc Bioinformatics. Chapter 6:Unit 6 4.

XU D, JABER Y, PAVLIDIS P, GOKCUMEN O. 2017 VCFtoTree: a user-friendly tool to construct locus-specific alignments and phylogenies from thousands of anthropologically relevant genome sequences. BMC Bioinformatics. 18(1):426.

Figure 1. Screen shot of *VCF2PopTree* on the Google Chrome browser. Three displays are shown. **(A)** UPGMA tree **(B)** *Newick* tree **(C)** Pairwise divergences in MEGA format



A

Phylogenetic tree

Construct Tree

● UPGMA tree
○ Neighbour-Joining tree (Unrooted)

Drawing options

● Rectangular tree
○ Radial tree

Data

○ Newick tree format
○ Pair-wise divergences - MEGA
○ Pair-wise divergences - PHYLIP

Enter genome size (bp) Default value - Human genome

2875001522

Read VCF file

B

Phylogenetic tree

Construct Tree

● UPGMA tree
○ Neighbour-Joining tree (Unrooted)

Drawing options

● Rectangular tree
○ Radial tree

Data

● Newick tree format
○ Pair-wise diversity - MEGA
○ Pair-wise diversity - PHYLIP

Enter genome size (bp) Default value - Human genome

2875001522

Read VCF file

```
((Gambian:0.00003770571916935493,Nigerian:0.00003770571916935493):0.0000018441948369037972,
(((Pakistani:0.000026711116294149914,Indian:0.000026711116294149914):9.834197692411975e-7,
(French:0.000025230943164697222,English:0.000025230943164697222):0.0000024635928986938926):9.2536182
89620749e-7,
((Chinese:0.000024946421576190004,Japanese:0.000024946421576190004):0.0000012678254157807725,
(Brazilian:0.000019476859254337466,Columbian:0.000019476859254337466):0.0000067373877376333345):0.00
0002405650900382375):0.0000010930016113905544);
```

C

Phylogenetic tree

Construct Tree

● UPGMA tree
○ Neighbour-Joining tree (Unrooted)

Drawing options

● Rectangular tree
○ Radial tree

Data

○ Newick tree format
● Pair-wise diversity - MEGA
○ Pair-wise diversity - PHYLIP

Enter genome size (bp) Default value - Human genome

2875001522

Read VCF file

```
#MEGA
 !Title=;

#Gambian
#Chinese
#French
#Brazilian
#Nigerian
#Pakistani
#English
#Columbian
#Indian
#Japanese

[    1    2    3    4    5    6    7    8    9    10  ]
[1]
[2]  0.000079444
[3]  0.000078186 0.000060561
[4]  0.000078921 0.000054451 0.000055267
[5]  0.000075411 0.000079375 0.000079038 0.000079399
[6]  0.000079085 0.000056946 0.000053895 0.000056757 0.000079340
[7]  0.000078123 0.000060330 0.000050462 0.000056049 0.000079131  0.000054473
[8]  0.000079602 0.000052293 0.000056662 0.000038954 0.000079486 0.000057564 0.000057468
[9]  0.000079705 0.000055992 0.000056071 0.000057323 0.000078910 0.000053422 0.000055912 0.000056804
[10] 0.000080055 0.000049893 0.000057978 0.000051096 0.000080544 0.000056271 0.000057557 0.000051184 0.000055385
```

Figure 1

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27682v1 | CC BY 4.0 Open Access | rec: 25 Apr 2019, publ: 25 Apr 2019