Genomics 104 (2014) 1–7

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Methods

# Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes

William J. Faison [a,1], Alexandre Rostovtsev [b,1], Eduardo Castro-Nallar [c], Keith A. Crandall [c], Konstantin Chumakov [b], Vahan Simonyan [b], Raja Mazumder [a,d,*]

[a] The Department of Biochemistry & Molecular Medicine, George Washington University Medical Center, Washington, DC 20037, USA
[b] Center for Biologics Evaluation and Research, US Food and Drug Administration, 1451 Rockville Pike, Rockville, MD 20852, USA
[c] Computational Biology Institute, George Washington University, Ashburn, VA 20147, USA
[d] McCormick Genomic and Proteomic Center, George Washington University, Washington, DC 20037, USA

ARTICLE INFO

*Article history:*
Received 11 February 2014
Accepted 4 June 2014
Available online 12 June 2014

*Keywords:*
Phylogenetic
Next-generation sequencing
SNP
SNV
Cancer genomics
Vaccine quality control

ABSTRACT

Next-generation sequencing data can be mapped to a reference genome to identify single-nucleotide polymorphisms/variations (SNPs/SNVs; called SNPs hereafter). In theory, SNPs can be compared across several samples and the differences can be used to create phylogenetic trees depicting relatedness among the samples. However, in practice this is difficult because currently there is no stand-alone tool that takes SNP data directly as input and produces phylogenetic trees. In response to this need, PhyloSNP application was created with two analysis methods 1) a quantitative method that creates the presence/absence matrix which can be directly used to generate phylogenetic trees or creates a tree from a shrunk genome alignment (includes additional bases surrounding the SNP position) and 2) a qualitative method that clusters samples based on the frequency of different bases found at a particular position. The algorithms were used to generate trees from *Poliovirus*, *Burkholderia* and human cancer genomics NGS datasets.
Availability: PhyloSNP is freely available for download at http://hive.biochemistry.gwu.edu/dna.cgi?cmd=phylosnp.

## 1. Introduction

With the advent of next-generation sequencing (NGS), large-scale data analysis has become the preferred way to study genomic data. NGS has also allowed researchers to identify and study single-nucleotide polymorphisms/variations (SNPs/SNVs; called SNPs hereafter), the individual base pair mutations in a genome [1]. These individual mutations are the basis for making every individual organism unique among a set of nearly identical sequences. While these mutations only occur rarely (at the mutation rate of a given organism that can vary by orders of magnitude depending on the organism), they are numerous enough to provide sufficient data to be compared among several samples and the differences can then be used to create phylogenetic trees of the data set samples. SNP analysis not only provides an emerging way to study and understand gene mutations, but quantitative profiles of mutations could facilitate a better understanding of diseases that affect certain populations, allowing scientists to develop personalized treatment plans for a group of individuals.

Groups such as Leekitcharoenphon et al. and Van Geystelen et al. have developed applications to automate the generation of SNP trees; however, the scope of each is somewhat limited [2,3]. While it does provide a plethora of information, snpTree does not allow for user uploaded data and in its current implementation allows analysis of bacterial genomes only. AMY-tree, a different tree-building software, does analyze full human genomes, but the program is limited as it was developed to specifically determine the relative position of the Y chromosome for lineage purposes. There are many popular tools (for example, MEGA, SplitsTree and others [4,5]) which generate phylogenetic trees similar to PhyloSNP, however, all of the aforementioned programs either work on smaller datasets or require a multiple sequence alignment to generate the phylogenetic trees. Other extant programs can generate phylogenetic trees based upon clustering algorithms [6]. However, these algorithms do not cluster based on similarity to a reference genome but rather by finding shared SNPs across all data samples, using conserved k-mers as the comparison. While a fast method for clustering data, the SNPs found between the data samples are only likely SNPs and are not based on actual mapped reads from mapping and profiling steps in NGS analysis.

* Corresponding author at: The Department of Biochemistry & Molecular Medicine, George Washington University Medical Center, Washington, DC 20037, USA.
*E-mail addresses:* Jamie_Faison@gwmail.gwu.edu (W.J. Faison), Alexandre.Rostovtsev@fda.hhs.gov (A. Rostovtsev), Ecastron@gwmail.gwu.edu (E. Castro-Nallar), Kcrandall@gwu.edu (K.A. Crandall), Konstantin.Chumakov@fda.hhs.gov (K. Chumakov), Vahan.Simonyan@fda.hhs.gov (V. Simonyan), Mazumder@gwu.edu (R. Mazumder).

[1] Equal contributors.

http://dx.doi.org/10.1016/j.ygeno.2014.06.001
0888-7543/© 2014 Elsevier Inc. All rights reserved.

PhyloSNP creates a SNP presence and absence data matrix which can be used as an input in PHYLIP pars program [7] to generate a phylogenetic tree. Another option is to create a SNP shrunk-genome alignment utilizing the presence/absence matrix where the user defines how many flanking bases around the SNP to use to create the alignment and once the alignment is done, it can be used by any phylogenetic tree building program that requires a multiple sequence alignment. An additional algorithm, presented here provides a more qualitative approach to this problem by taking into consideration the frequencies of the different bases in a particular position to cluster samples, thereby providing a novel comparative genomics approach. This is important for datasets where a cutoff for the presence or absence of a variation may be initially unclear. The tool builds interactive phylograms based on the frequency of SNPs in sets of reads, allowing the user to quickly see the effect that changes in algorithmic parameters for selecting polymorphisms would have on the tree. The integrated clustering tool additionally outputs shrunk genomes in the same format as the standalone PhyloSNP for integration with the user's phylogenetic workflow. A flowchart describing all three approaches is shown in Fig. 1a. PhyloSNP is designed to be easily integrated into NGS analysis platforms such as the High-performance Integrated Virtual Environment (HIVE) [8] where the output of aligner and SNP profiler tools become the input to the program (Fig. 1b). The program allows for diverse analysis over several genome types of unlimited size. This, therefore, allows a user not only the simplicity of having one tool for all genome types, but assuming access to a capable machine, also allows the analysis of extremely large scale data sets that were previously impossible to parse as one experiment.

## 2. Methods

### 2.1. Datasets

A simple graphical user interface was created to help the user quickly generate phylogenetic trees or shrunk-genomes from SNP data with a few button clicks. The program has two dependencies which are both freely available. They are Perl (http://www.perl.org/) and PHYLIP [7]. *Burkholderia pseudomallei* data used in this study were obtained from NCBI SRA (accession: SRP023117). A *Poliovirus OPV-3* data set was obtained with permission from the Chumakov lab and the Human data set was obtained from TCGA (http://cancergenome.nih.gov/) breast cancer study.

### 2.2. Quantitative approach: presence/absence data matrix generation

When running the PhyloSNP portion to directly generate phylogenetic trees, the PHYLIP package utilizes the pars algorithm which conducts discrete character parsimony on the inputted dataset. The algorithm executes Wagner parsimony upon the data with multistate characters between data [7]. In the case of PhyloSNP, the characters of each data sample are the SNPs discovered across all data samples, and the character states are the presence (1) or absence (0) of a SNP in a particular position. The Wagner method generates phylogenetic trees by adding samples to a tree based on the smallest distance between data samples and a hypothetical outgroup. More specifically, the presence/absence matrix is converted into PHYLIP format for final analysis which includes bootstrap replicates of the provided data file, estimation of the maximum parsimony [9] tree for the dataset, merging of all generated trees and producing a best fitting tree along with bootstrap values followed by estimation of a phylogenetic tree with branch lengths which is provided in Newick format [10]. To visualize the SNP data, R [11] is used to generate a heatmap using packages ggplot and reshape.

#### 2.2.1. Shrunk-genome alignment generation

Following SNP identification, the resultant data files were downloaded to the local user environment and run through both PhyloSNP utilities, Phylogenetic Trees and shrunk-genomes, with delta positions of 0, 5, and 10 bases surrounding each SNP to generate concatenated genomes for further analysis. The shrunk-genomes are generated in PhyloSNP by the method as described in Fig. 2. These concatenated genomes and the resultant alignment were then used create phylogenetic trees using the neighbor-joining method as implemented in ClustalW [12] and trees were visualized in FigTree (http://tree.bio.ed.ac.uk/software/figtree 2012).

### 2.3. Qualitative approach: HIVE-integrated analysis

HIVE's integrated fast hierarchical clustering is based on the frequency of polymorphisms found in a sample's reads at given positions.

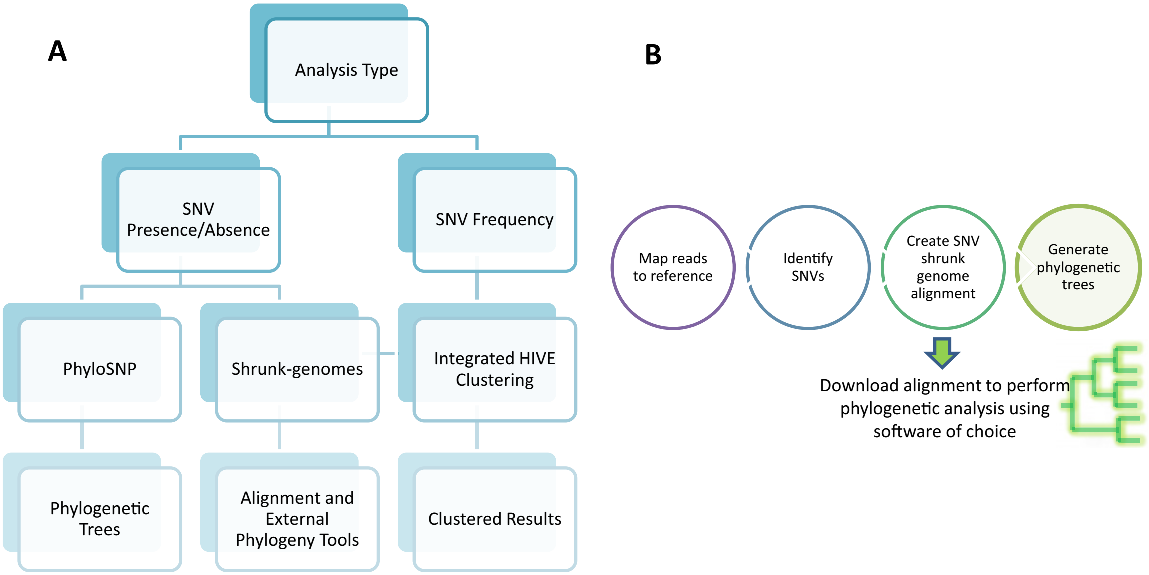
A
Analysis Type
SNV Presence/Absence
SNV Frequency
PhyloSNP
Shrunk-genomes
Integrated HIVE Clustering
Phylogenetic Trees
Alignment and External Phylogeny Tools
Clustered Results
B
Map reads to reference
Identify SNVs
Create SNV shrunk genome alignment
Generate phylogenetic trees
Download alignment to perform phylogenetic analysis using software of choice

**Fig. 1.** Overview of PhyloSNP analysis. A) Demonstrates the ways that a user can analyze their dataset once samples have been aligned and profiled for SNPs in any typical NGS analysis workflow. Three options are available, PhyloSNP, shrunk-genome alignments, and integrated HIVE Clustering. The first two options pertain to the presence and absence of SNPs and are part of a downloadable client side pipeline, while the latter is part of an integrated HIVE pipeline that clusters based on frequencies of SNPs. B) Overview of PhyloSNP pipeline.