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/

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.

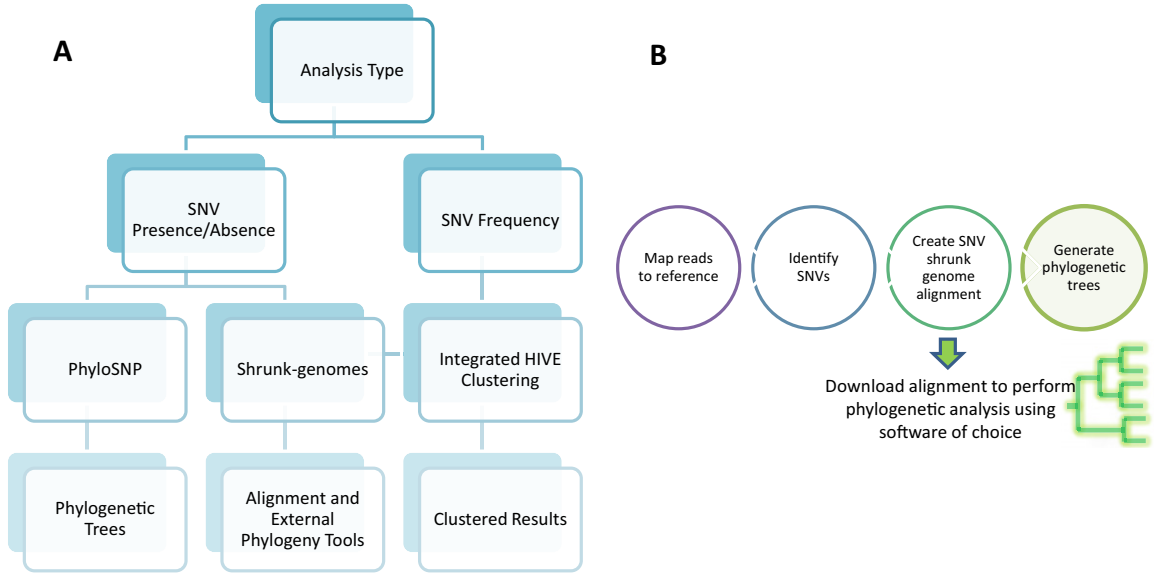

Fig. 1. Overview of PhyloSNP analysis. A) Demonstrates the ways that a user can analyze their dataset once samples have been aligned and profiled for SNPs in any typical NGS analysis workflow. Three options are available, PhyloSNP, shrunk-genome alignments, and integrated HIVE Clustering. The first two 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.

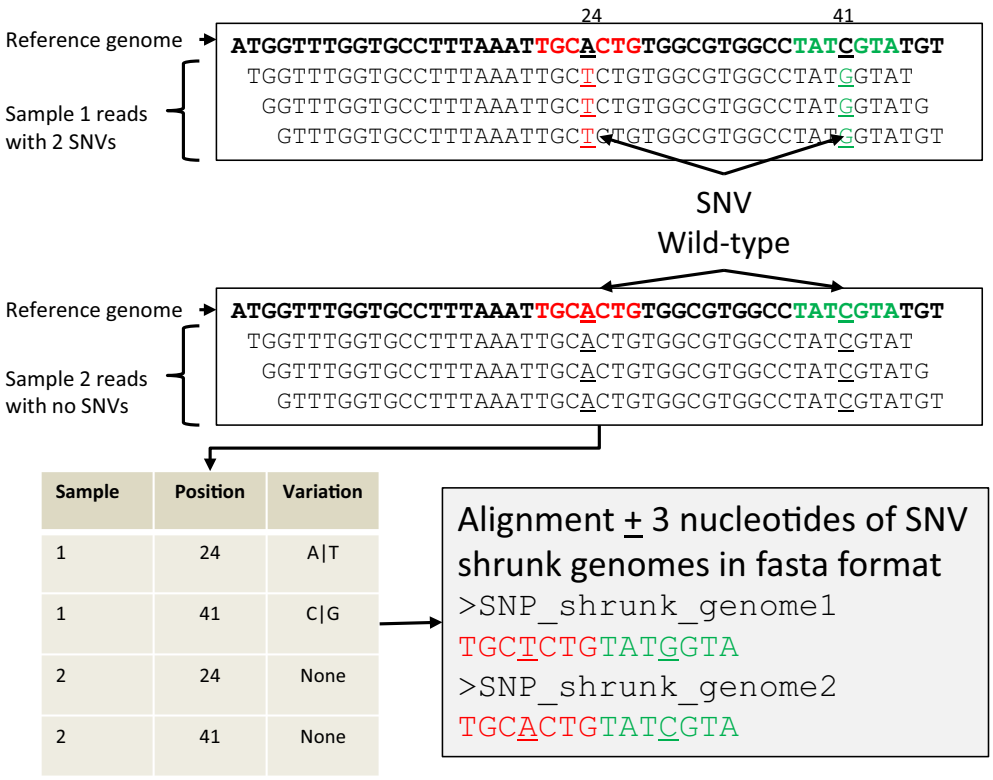

Fig. 2. Process used by PhyloSNP to obtain shrunk-genomes. Reads from individual samples are first mapped to the reference and variations are identified to create a table containing variation position. This table is then used to extract a user defined number of nucleotides from the reference upstream and downstream of the SNP. The concatenated nucleotides obtained from all the samples can be used as the input alignment into phylogenetic programs to build trees. Nucleotides at that variation site are underlined; the two stretches of nucleotides that include the variation are colored red and green.

Nucleotide polymorphisms at each position are represented by a set of four values ranging from 0 to 1 describing the content of each nucleotide. A variety of parameters is provided for choosing which positions will be examined: for example, positions may be discarded if a sample has insufficient read coverage, or if SNP frequencies are below a given threshold; positions within a given delta from an SNP may be included; frequencies outside a specified range may be declared zero. The resulting shrunk vectors of frequencies are compared pairwise using a distance function selected by the user, producing a distance matrix. Using cosine similarity as the distance effectively normalizes the frequencies in each sample. Euclidean distance, another p-norm, or the Canberra metric can be used if SNP frequencies in different samples are comparable without normalization. Euclidean distance was chosen for this pipeline as it traditionally is the most commonly used distance function for hierarchal clustering based on continuous distance values. The neighbor-joining algorithm [13,14] is then used to produce phylograms from the distance matrix.

2.3.1. High-performance Integrated Virtual Environment, HIVE

The High-performance Integrated Virtual Environment, HIVE, was used to find the SNPs from the data shown in this paper. HIVE is a cloud-based web environment optimized for management and analysis of next-gen data. From the main HIVE interface, the SRA read files and reference genomes were uploaded via the internal file-loading utility. The read files were then aligned to the reference genomes using the natively developed HIVE-hexagon aligner which employs a number of novel algorithmic enhancements to quickly and accurately compute alignments in the massively parallel HIVE execution environment [15]. HIVE-hexagon was run using the following parameters: minimum overlap of 40, 5% mismatch. After alignment, HIVE-heptagon (base-caller and SNP profiler) was used to retain desired SNPs based on specification of parameters such that minimum coverage is 35, minimum quality is 26, and the frequency threshold is 40%. The HIVE-heptagon SNP profiler calculates and compares the frequency of individual nucleotide bases at each position with the number of bases in the reference genome or consensus sequence at that position, outputs quality and sequencing noise profiles and reports statistical significance of all findings. The abovementioned parameters have been empirically determined to provide good algorithm performance [16].

3. Results and discussion

3.1. Usage

3.1.1. Presence/absence data matrix

PhyloSNP takes as input variation data across several genomes or contigs and outputs tree files. Intermediate output includes PHYLIP formatted SNP data which can be easily converted to NEXUS format for further analysis in phylogenetic analysis programs other than PHYLIP.

Input to the program can be multiple files with only one column (SNP position), one merged file with two columns (1st column having SNP positions; 2nd column with sample number/name) or files in standard VCF format (http://vcftools.sourceforge.net/specs.html). If additional columns are present in the file, they are ignored by the program. All SNP positions should refer to a single reference sequence. An example genome file is provided in the PhyloSNP zip file download. A three-step process needs to be followed for uploading and analysis of SNP data. First, the folder where the SNP .csv or .vcf file(s) are located is selected. Next, the user specifies the number of replicate data sets they would like to generate for bootstrap tree analysis. Once the folder and replicates have been selected, the user clicks the submit button to generate the tree. Troubleshooting and detailed instructions are provided in the readme file found in the program's folder.

When operating PhyloSNP in a UNIX environment shell, the following perl command prompt generates tree files: perl phylosnp.pl dir (or files). This command will run PhyloSNP and search for a folder of files or individual files and perform the standard tree generation. The default parameters will generate a directory named output that stores the binary sequences of datasets for downstream phylogenetic analysis. For documentation and help, user can type perl phylosnp.pl --help.

3.1.2. Shrunk-genomes alignment

The shrunk-genomes function takes variation data across several genomes and a reference genome as input and outputs concatenated molecular sequences of each genome (Fig. 2). These concatenated sequences are Fasta formatted alignments for use in downstream phylogenetic analysis programs. These alignments can be easily transformed into other alignment formats such as Clustal, PHYLIP, or MEGA format using http://www.phylogeny.fr/version2_cgi/data_converter.cgi or http://www.ibi.vu.nl/programs/convertalignwww/.

Input to the program can be multiple files with two columns (1st column having SNP positions; 2nd column with SNP changes as N|N), or files in standard VCF format. Similar to the previously described method a three-step process needs to be followed for uploading and analysis of SNP data, wherein the folder where the SNP .csv or .vcf file(s) are located is selected for the first step. Next, the user specifies the location of the reference genome in Fasta format and selects the number of flanking base pairs around each SNP (delta default is 0). After the folder, reference file and position delta have been selected, the submit button is clicked to generate the concatenated sequences file. Troubleshooting and detailed instructions are provided in the readme file found in the program's folder.

When operating Shrunk-genomes in a UNIX environment, the following command prompt generates tree files: perl shrunk-genomes.pl dir (or files) --reference-file = referenceFasta > outputFasta. This command will run shrunk-genomes and search for a folder of files or individual files and perform the standard sequence concatenation. The default parameters will generate an outputFasta file located in the parent directory of the sequence folder or files for downstream phylogenetic analysis. For documentation and help, user can type perl shrunk-genomes.pl --help.

3.1.3. HIVE-integrated clustering and shrunk-genomes

To launch HIVE's hierarchical clustering analysis tool, the user chooses a set of alignment profiler calculations already performed on HIVE and a set of reference genes from the reference genome used by the alignments. There are selectable parameters for the distance function, clustering algorithm, noise filtering, minimum and maximum SNP frequency thresholds, minimum read coverage, and position deltas; by default, Euclidean distance and neighbor-joining clustering are used. PhyloSNP's shrunk-genome algorithm has been integrated into the HIVE clustering tool, and the user can choose to save the shrunk genome in Fasta format or as a table with SNP frequencies.

The clustering tool's main output is an interactive, downloadable SVG phylogram. A subset of samples can be selected from the phylogram to compare their SNP frequencies side by side as described in sections below. The shrunk genome sequence, distance matrix, and SNP frequency tables produced by the clustering tool can be downloaded for exhaustive phylogenetic analysis by external software.

3.1.4. Downstream analysis

To ensure compatibility in several downstream programs, datasets were tested in MEGA [4], Clustal [12] and Splitstree [5]. After testing, it was shown that the shrunk-genomes data can be easily used as inputs to each of these programs for specific analysis.

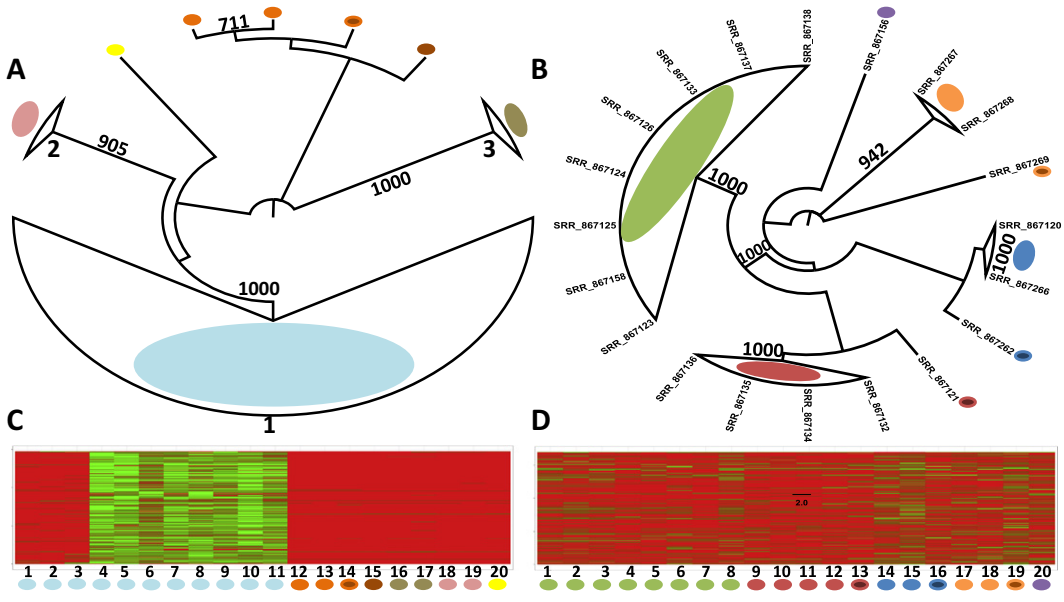

Fig. 3. Phylogenetic trees and heatmaps of virus and bacteria data generated through the various PhyloSNP functions. A) Phylogenetic tree generated from concatenated Poliovirus OPV-3 sequences using the shrunk-genomes method of PhyloSNP with a position delta of five. Concatenated sequences were then arranged into phylogenetic tree using ClustalW2 with bootstrap support shown on the branches (out of 1000). Three main groups were discovered using this analysis, as noted on the diagram. SNPs were obtained from the HIVE pipeline using match length 40 and 5% mismatch with the remaining parameters at default. B) Phylogenetic tree generated from concatenated *B. pseudomallei* sequences using the shrunk-genomes method of PhyloSNP with a position delta of five. The generated shrunk-genome alignment was used to create the phylogenetic tree using Neighbor Joining method as implemented in ClustalW2. Bootstrap support is shown on the branches (out of 1000). Alignment and SNP profiling parameters for obtained data were the same as shown in 3A. C) Heatmap of identified SNPs across Poliovirus OPV-3 samples. Nearly 3000 SNPs were identified, the majority of which came from samples 3–10, which group into the largest cluster as demonstrated in 3A. D) Heatmap of identified SNPs across *B pseudomallei* samples. Through SNP alignment and profiling with set parameters, HIVE Heptagon was able to identify over 67,000 unique SNPs among all 20 data samples. Green represents the presence of a SNP at a specific position for a particular sample, while red notes the absence of a SNP. NCBI-SRA accession numbers of heatmap from left to right are as follows: 1:SRR_867123, 2:SRR_867124, 3:SRR_867125, 4:SRR_867126, 5:SRR_867133, 6:SRR_867137, 7:SRR_867138, 8:SRR_867158, 9:SRR_867132, 10:SRR_867134, 11: SRR_867135, 12:SRR_867136, 13:SRR_867121, 14:SRR_867120, 15:SRR_867266, 16:SRR_867262, 17:SRR_867267, 18:SRR_867268, 19:SRR_867269, and 20:SRR_867156.

3.2. Example analysis results

3.2.1. Virus

Poliovirus is a causative agent of poliomyelitis, an infectious paralytic disease affecting mostly children. Inactivated and live attenuated vaccines made from viruses of all three serotypes of poliovirus are used for its prevention [17–19]. While almost eradicated in most countries, poliovirus is still endemic in some developing countries, where live oral poliovirus vaccine (OPV) is widely used. This study includes deep-sequence data for 20 samples of type 3 OPV which was used to create a phylogenetic tree using the presence/absence approaches and the qualitative approach. The sample analyzed included vaccine batches made by different manufacturers from different seed viruses. Each stock of the vaccine virus has a characteristic pattern of SNPs (SNP profile) that can be used both for identification of the source of the stock, and for vaccine quality control by monitoring molecular consistency of vaccine manufacture. The results of comparison of SNP profiles presented here demonstrated that batches of OPV-3 produced from the same seed virus clustered together on the trees (Fig. 3a). For this dataset the tree topology was similar regardless of which of the three approaches was used. This suggests that such analysis could be used for classification of OPV-3 batches and identification of their origin.

3.2.2. Bacteria

B. pseudomallei data demonstrates the power of PhyloSNP's shrunk-genomes script, being able to generate a concatenated genomic dataset sharing over 67,000 unique SNPs. Physiologically, *B. pseudomallei* causes melioidosis, which is a deadly infectious disease. Several *B. pseudomallei* strains were sequenced to investigate the diversity of the isolates. Previously Ooi et al., has shown that the chromosomes in these organisms are highly mosaic and contain strand-specific genes to account for varying conditions [20]. The resulting analyses show the relationships between the genome samples as well as the contribution of SNPs to the diversity between samples (Fig. 3b). The heatmap generated in Fig. 3d shows that these SNPs are varied in position and are relatively evenly distributed across all samples. This provides a broader genomic representation with less genetic linkage, and thus, lessens the impact of recombination. There does not appear to be any strong association between the number of SNPs and their relatedness (Figs. 3b and d), as the only grouping of samples with moderate similarities in the amount of SNPs appears in the grouping between samples 14 and 15, with a loose association of that pair with sample 16. Interestingly samples 15 and 16 were both obtained from patients that originated in Darwin, Australia, while the third sample, sample 14, was isolated in 1949 [SRP023117]. What is surprising, however, is that these two samples do not appear to be

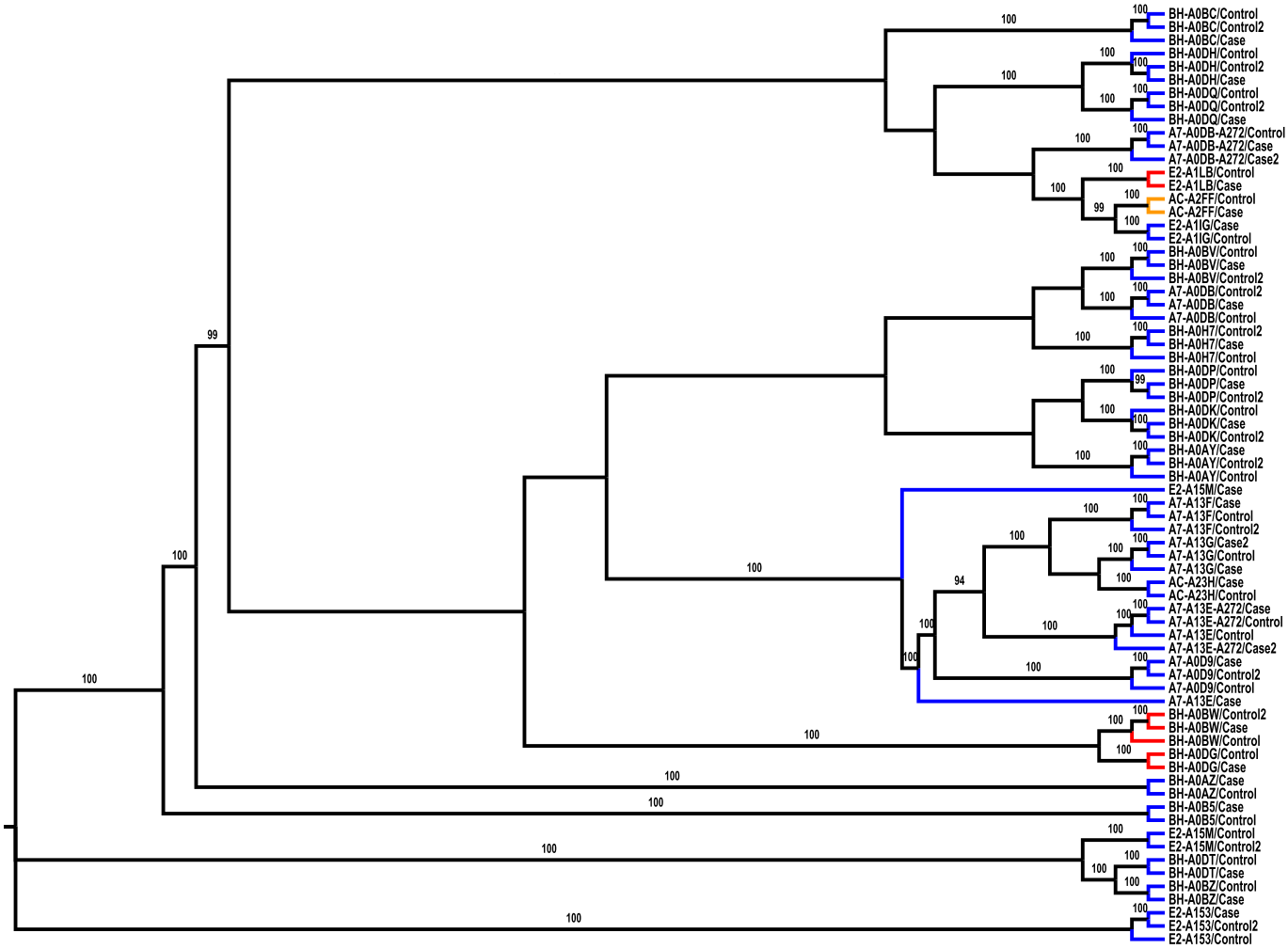

Fig. 4. Phylogenetic tree generated from concatenated *Homo sapiens* sequences using the shrunk-genomes method of PhyloSNP with a position delta of zero. Concatenated sequences were then arranged into phylogenetic tree using Neighbor Joining method as implemented in ClustalW2. Bootstrap support is shown at the nodes (out of 100). Blue corresponds to Caucasians, red to African-Americans, and orange to Asians.

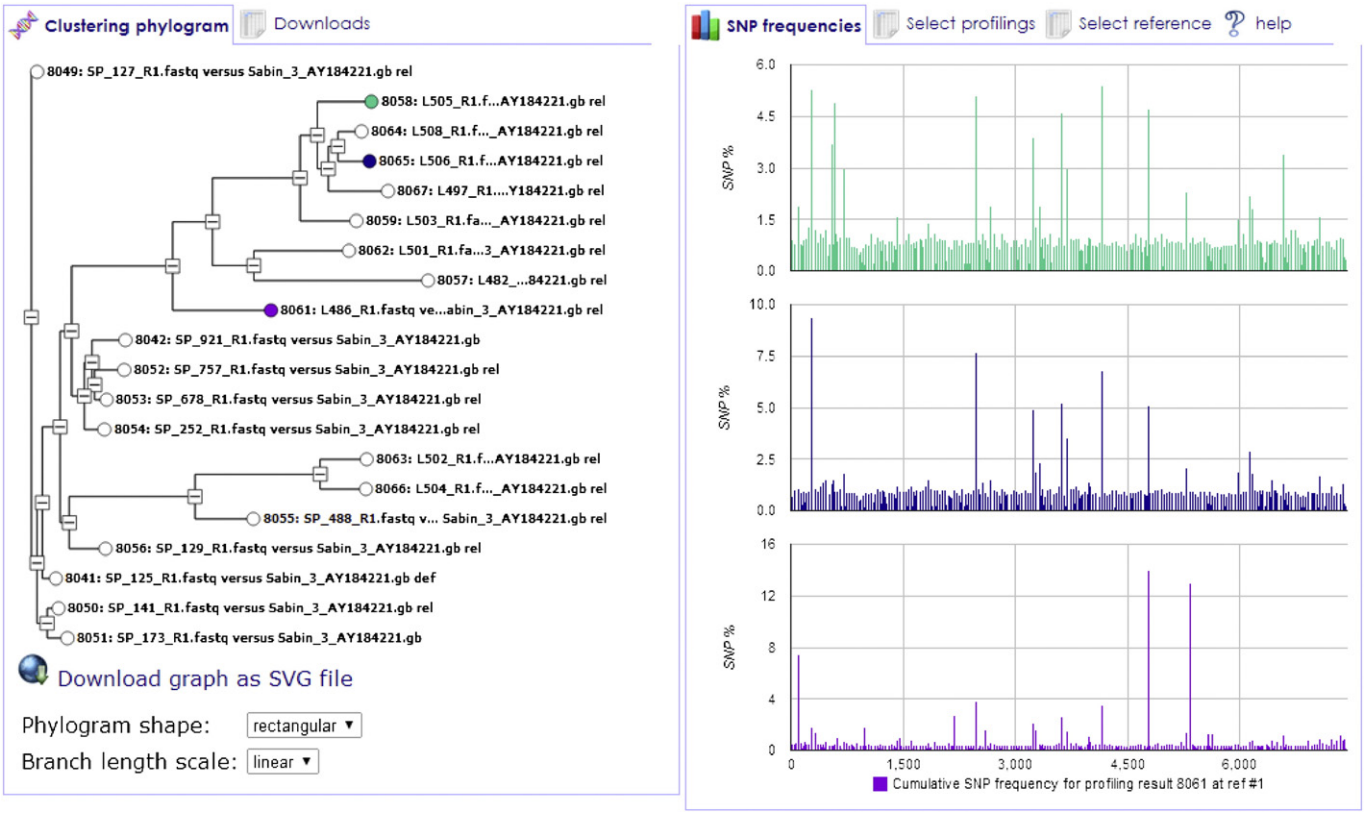

Fig. 5. Screenshot of HIVE clustering analysis run on 20 Poliovirus OPV-3 samples with a minimum mutation frequency threshold of 0.005, position delta of 5, using Euclidean distance and the neighbor-joining algorithm. On the left is the interactive phylogram. On the right are SNP frequency graphs for the three selected samples (L505, L506, and L501).

closely related with the three other Darwin strains (samples 18–20). More analyses need to be performed with additional samples and better epidemiological and clinical annotations to better understand their relationships.

3.2.3. Human

With the advent of NGS techniques, the next step in patient care will come in the form of personalized medicine. Ever since the human genome project was completed, researchers have been looking for ways to cater care on an individual basis [21]. With access to tools allowing the discovery of SNPs, we are able to map the genetic mutations that make up the differences between each person. We believe that classification of patients based on their SNP profiles can provide clues in terms of prognostics, diagnostics and therapeutics. With PhyloSNP, we are able to generate a SNP shrunk genome for a large dataset of 71 genomic samples (tumor and normal) from 25 breast cancer patients and produce phylogenetic trees from these results with high confidence (Fig. 4). SNP data obtained from this study is available from Curated Short Read (CSR) archive [16]. In the figure it can be seen that all samples from the same patient cluster together (except for one patient who has many case and control samples; all other patients have less numbers of samples) which is a good indication of the validity of the algorithm (Fig. 5). Additionally, it can be seen that ethnicity does not appear to be a determinant on how individuals group. The implications of studies such as these are enormous. With the ability to generate phylogenetic relationships, an individual can be grouped into a phylogenetic cluster of known patients and it is feasible to create HMMER models [22] using branch specific alignments which then can be used to place new patients into existing groups. Within this cluster, each patient can receive treatment that is unique to their genetic sequence allowing for more accurate and beneficial drug treatments with fewer adverse side effects. The possibilities of personalized medicine are still in the infant stages; however, PhyloSNP can provide a pivotal role in preliminary classification of patients. (See Fig. 5.)

4. Conclusion

The PhyloSNP phylogenetic analysis tool serves as a way to interpret deep-sequencing data about subtle differences between samples, which enables generation of hypotheses to help understand the reasons for the variability among them. The dual function of the tool, through preliminary phylogenetic tree construction and the generation of concatenated genomes, allows for a large variety of uses and increased versatility in genomic analysis including the ability to analyze virus, bacteria, and human genomes through either a presence/absence matrix, through concatenated genomes or NGS analysis platform integrated clustering analysis.

4.1. Access

PhyloSNP is available from http://hive.biochemistry.gwu.edu/dna.cgi?cmd=phylosnp. To use HIVE's computationally intensive tools users need to register at http://hive.biochemistry.gwu.edu/dna.cgi?cmd=userReg. Users can also install HIVE on their own hardware or use HIVE-in-a-box which is a low cost alternative to analyze NGS data using pre-determined workflows. For additional details users are encouraged to contact the HIVE team (http://hive.biochemistry.gwu.edu/dna.cgi?cmd=contact).

Authors' contributions

RM conceived, designed and coordinated the study. RM, VS, KeC, and KoC were involved in developing/refining the different methods and providing useful comments. KoC provided the poliovirus dataset and

KeC and ECN helped with acquiring the bacterial data. WJF and AR developed the specific algorithm and were responsible for software design and implementation, and writing of the manuscript. All authors read and approved the final manuscript.

Disclaimer

The contributions of WJF are an informal communication and represent his own best judgment. These comments do not bind or obligate FDA.

Acknowledgments

PhyloSNP development is supported by High-performance Integrated Virtual Environment (HIVE). The HIVE project led by Dr. Mazumder and Dr. Simonyan supports Big Data analysis which includes next-generation sequence analysis and storage (http://hive.biochemistry.gwu.edu/). We would like to thank Hayley Dingerdissen and Pan Yang for providing useful comments. This project is supported in part by the Research Participation Program at the CBER administered by the ORISE through an interagency agreement between the USDOE, the USFDA and CTSI-CN/GW/VT (IXXS90459N) collaborative research grant to RM.

References

[1] A. Vignal, D. Milan, M. SanCristobal, A. Eggen, A review on SNP and other types of molecular markers and their use in animal genetics, Genet. Sel. Evol. 34 (2002) 275–305.
[2] P. Leekitcharoenphon, R.S. Kaas, M.C. Thomsen, C. Friis, S. Rasmussen, F.M. Aarestrup, snpTree — a web-server to identify and construct SNP trees from whole genome sequence data, BMC Genomics 13 (Suppl. 7) (2012) S6.
[3] A. Van Geystelen, R. Decorte, M.H. Larmuseau, AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications, BMC Genomics 14 (2013) 101.
[4] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: Molecular Evolutionary Genetics Analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729.
[5] D.H. Huson, SplitsTree: analyzing and visualizing evolutionary data, Bioinformatics 14 (1998) 68–73.
[6] S.N. Gardner, T. Slezak, Scalable SNP analyses of 100+ bacterial or viral genomes, J. Forensic Res. (2010). http://dx.doi.org/10.4172/2157-7145.1000107.
[7] J. Felsenstein, PHYLIP — Phylogeny Inference Package (version 3.2), Cladistics 5 (1989) 164–166.
[8] V. Simonyan, R. Mazumder, High-performance Integrated Virtual Environment clouds (HIVE) for extra-large (XL) data analysis, in: g.a. Comparative sequence, genome assembly, genome scale computational methods session (Eds.), The 2011 International Conference on Bioinformatics and Computational Biology., Las Vegas, Nevada, 2011.
[9] W.M. Fitch, Toward defining the course of evolution: minimum change for a specific tree topology, Syst. Biol. 20 (1971) 406–416.
[10] G. Cardona, F. Rossello, G. Valiente, Extended Newick: it is time for a standard representation of phylogenetic networks, BMC Bioinformatics 9 (2008) 532.
[11] R_core_team, R: a language and Environment for Statistical Computing, R Foundation for Statistical Computing, 2013. http://www.R-project.org .
[12] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, Bioinformatics 23 (2007) 2947–2948.
[13] N. Saitou, M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Mol. Biol. Evol. 4 (1987) 406–425.
[14] J.A. Studier, K.J. Keppler, A note on the neighbor-joining algorithm of Saitou and Nei, Mol. Biol. Evol. 5 (1988) 729–731.
[15] H. Dingerdissen, A. Voskanian, L. Santana-Quintero, R. Mazumder, V. Simonyan, HIVE: Highly Optimized Efficient Approaches of Next-gen, Best poster award. Bio-IT Conference, Bio-IT, Boston, 2013, (http://hive.biochemistry.gwu.edu/HIVE_AlgorithmicsPoster.pdf).
[16] C. Cole, K. Krampis, K. Karagiannis, J. Almeida, W.J. Faison, M. Motwani, Q. Wan, A. Golikov, Y. Pan, V. Simonyan, R. Mazumder, Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data, BMC Bioinformatics 15 (2014) 28.
[17] P. Ragonese, B. Fierro, G. Salemi, G. Randisi, D. Buffa, M. D'Amelio, A. Aloisio, G. Savettieri, Prevalence and risk factors of post-polio syndrome in a cohort of polio survivors, J. Neurol. Sci. 236 (2005) 31–35.
[18] B.T. Mayer, J.N. Eisenberg, C.J. Henry, M.G. Gomes, E.L. Ionides, J.S. Koopman, Successes and shortcomings of polio eradication: a transmission modeling analysis, Am. J. Epidemiol. 177 (2013) 1236–1245.
[19] K. Chumakov, E. Ehrenfeld, E. Wimmer, V.I. Agol, Vaccination against polio should not be stopped, Nat. Rev. Microbiol. 5 (2007) 952–958.
[20] W.F. Ooi, C. Ong, T. Nandi, J.F. Kreisberg, H.H. Chua, G. Sun, Y. Chen, C. Mueller, L. Conejero, M. Eshaghi, R.M. Ang, J. Liu, B.W. Sobral, S. Korbsrisate, Y.H. Gan, R.W. Titball, G.J. Bancroft, E. Valade, P. Tan, The condition-dependent transcriptional landscape of *Burkholderia pseudomallei*, PLoS Genet. 9 (2013) e1003795.
[21] F.S. Collins, V.A. McKusick, Implications of the Human Genome Project for medical science, JAMA 285 (2001) 540–544.
[22] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37.