**PhyloSNP**

The PhyloSNP application offers two methods for analysis. The first is a quantitative method that generates a presence/absence matrix for each position of a reference from the SNP profile which can then be used to build phylogenetic trees. The quantitative approach can also directly create a tree from an alignment of shrunk genomes, short concatenated genomes of regions flanking reported SNPs. The second qualitative method clusters samples based on the frequencies of different bases reported with respect to reference position.

Like all HIVE tools, PhyloSNP can analyze datasets without imposing size limitations, and therefore can accommodate a diverse set of inputs. This benefits the user not only by providing a single tool to handle the variety of all possible input genomes, but also by facilitating analysis of big data which was previously an impossible task.

Other applications can automate the generation of SNP trees with a limited scope. SNPTree provides useful information but does not allow users to upload custom data and only accommodates bacterial analysis. AMY-tree can analyze human genomes but was specifically developed to determine lineages based on Y chromosome comparisons. Other popular tools like MEGA and SplitsTree generate phylogenetic trees like PhyloSNP but are either restricted by size of dataset or require the user to supply a multiple sequence alignment as input. Additional programs capable of building phylogenies from clustering algorithms do not cluster based on comparison to a reference. PhyloSNP adds to the pool of existing tools by allowing users a single tool which can accomplish a variety of complex computations despite data size. PhlyoSNP provides a number of parameters to facilitate easy upload of information and customization of desired analysis.

For example, a user can perform a shrunk-genomes data analysis from local PhyloSNP directory by specifying the position of interest, input dataset and reference genome, the name of the file in which to dump the results and the size of the region around the SNPs to capture. This can take up to several hours to execute depending on the size and amount of the files, so it is recommended to run this overnight.

**Questions**

**(1) What specifically does the tool do?**

The PhyloSNP application offers two methods for analysis. The first is a quantitative method that generates a presence/absence matrix for each position of a reference from the SNP profile which can then be used to build phylogenetic trees. The quantitative approach can also directly create a tree from an alignment of shrunk genomes, short concatenated genomes of regions flanking reported SNPs. The second qualitative method clusters samples based on the frequencies of different bases reported with respect to reference position.

**(2) Why tool is useful? I.e., why would a user be interested in using this tool from a biological perspective, not just from what it does?**

Like all HIVE tools, PhyloSNP can analyze datasets without imposing size limitations, and therefore can accommodate a diverse set of inputs. This benefits the user not only by providing a single tool to handle the variety of all possible input genomes, but also by facilitating analysis of big data which was previously an impossible task.

**(3) How does the tool differ from similar extant tools?**

Other applications can automate the generation of SNP trees with a limited scope. SNPTree provides useful information but does not allow users to upload custom data and only accommodates bacterial analysis. AMY-tree can analyze human genomes but was specifically developed to determine lineages based on Y chromosome comparisons. Other popular tools like MEGA and SplitsTree generate phylogenetic trees like PhyloSNP but are either restricted by size of dataset or require the user to supply a multiple sequence alignment as input. Additional programs capable of building phylogenies from clustering algorithms do not cluster based on comparison to a reference.

**(4) Brief description of ALL parameters (list).**

**PhyloSNP**

**-output-dir=DIRNAME:** Directory where output will be written ("output" by default)

**-accession-col=NAME:** Name of accession column ("Accession" by default)

**-position-col=NAME:** Name of position column ("Position" by default)

**-genome-col=NAME:** Name of the genome name column ("Name" by default)

**-replicates=NAME:** Number of replicates for bootstrapping (100 by default)

**Shrunk-genomes:**

**-accession-col=NAME:** Name of accession column ("Accession" by default; not used for FASTA format references)

**-position-col=NAME:** Name of position column ("Position" by default)

**-genome-col=NAME:** Name of the sample genome name column ("Name" by default)

**-change-col=NAME:** Name of the SNP change column ("Change" by default)

**-letter-col=NAME:** Name of the reference genome letter column ("Letter" by default)

**-reference-file=NAME:** Name of reference genome file; needs to be in FASTA or CSV format

**-reference-name=NAME:** Name of a specific reference inside a FASTA file with multiple references

**-reference-col=NAME:** Column containing reference names used in the FASTA reference genome

**-reference-map=NAME:** CSV file with two columns: the values used in the --reference-col column in the input files, and the names used in the FASTA reference genome NOTE: if you use --reference-map option, you must also specify --reference-col.

**-position-delta:** Position range around each SNP to output (0 by default, so only positions that have an SNP are output).

**(5) Walkthrough of single use case. Any example you would like to use.**

Shrunk-Genomes Data Analysis

To start an analysis, use ssh hive3 to move to Hive3 to perform analysis and navigate to your PhyloSNP directory. Enter perl shrunk-genomes.pl dir\_of\_your\_files --position-col=POS(change) --change-col=Change(change) --reference-col=CHROM(chrom) --reference-file=hg19.fa --position-delta=0(change) > output\_file.fasta &. Note where it says (change) change to the appropriate column name you found above, and the number of flanking positions desired in the position-delta. Also note that the & command allows the user to run shrunk-genomes.pl in the background after logging off. If this is your first time using the data set it is recommended to add nohup at the beginning of the above command to output any errors to the output file, however, this will require downloading the output file, removing the shrinking genomes... line and re-uploading the fasta file. This can take up to several hours to execute depending on the size and amount of the files, so it is recommended to run this overnight.