TimeZone_v1.0

Package contents:

TimeZone_v1.0.exe (the main program)

makeblastdb.exe (program to format genome sequences for BLAST)

blastn.exe (program for standalone Blast operations)

clustalw.exe (program for multiple sequence alignments)

convert.exe (program to calculate Tajima D and Fu & Li D* statistics)

maxchi.exe (program to detect presence of recombination using MaxChi statistic)

phylpro.exe (program to detect presence of recombination using PhylPro statistic)

Documentation.pdf (this documentation)

Contents:

Introduction:

The TimeZone is a software package comprising of a set of analytical tools that can be applied for a genome-wide prediction of recent positive selection in genes with detection of specific adaptive mutations therein. This comprehensive tool outputs genome-wide information of mutations in coding genes that emerged under positive selection and their association with specific strain groups.

Features:

(1) TimeZone includes modified version of Zonal Phylogeny Software (ZPS) (Chattopadhyay et al. 2007) that, along with a set of approaches, was shown to be successful in the identification of potential genes undergoing short-term adaptive evolution.

(2) TimeZone is designed for performing large-scale analysis that can be used for microbial genome-wide association studies.

(3) TimeZone acts as a bridge to allow systematic attempts to detect and functionally compare natural allelic variants. This approach, therefore, will enable to understand the physiologically relevant properties of gene products, and also provide insights on genome-wide networks of co-evolving genes.

(4) TimeZone analysis determines a complete diversity and selection profile of genes across the entire genome, thereby allowing possibility of identifying clonal markers with improved discriminatory power.

Applications:

(1) Based on the user defined reference genome (fully annotated), orthologous sequences from any number of fully annotated or draft genomes are extracted according to the user-specified gene sequence identity and length coverage threshold. The output can be used to detect the sets of omnipresent (core) and patchy (mosaic) genes of the genomes under study.

(2) The sequences carrying non-ACGT characters are excluded for analysis and an output file is generated with the list of genes and strains carrying non-ACGT characters to allow special attention for analysis. Also the annotated pseudogenes in the reference genome are excluded from the analysis. Though, the orthologous copies of all genes with premature stop codons in other genomes are defined as such and analyzed. An output file with the list of genes and strains with premature stop codons is generated to allow further analysis on gene inactivation.

(3) TimeZone incorporates Zonal Phylogeny (ZP) analysis that can separate structural variants of the coded proteins in two types (tree zones): evolutionarily long-term (fixed) and evolutionarily short-term (recent) (Chattopadhyay et al. 2007). Using this ZP algorithm, allelic diversity of each zone can be computed based on diversity indices (number of representatives and their frequency). A significantly higher diversity in the short-term zone recommends the gene to be under positive selection for recent emergence of structural variants.

(4) TimeZone detects the presence of convergent (hotspot) amino acid mutations – a powerful indication of positive selection (Chattopadhyay et al. 2009; Christin et al.

2010). Information of hotspot positions and of the strains accumulating such mutations is generated as a separate file to predict adaptive mutations along specific clonal lineages.

(5) Presence of gene recombination is detected by either MaxChi (Smith 1992) or PhylPro (Weiller 1998) at the 95% for each gene set. A list of genes showing recombination will be generated and can be used for further analysis.

(6) TimeZone calculates the proportion of structural to silent mutations in the terminal branches (tips) relative to the internal branches (twigs). This analysis provides an additional test for genes under recent positive selection (Bush et al. 2000).

(7) TimeZone computes rates of structural (nonsynonymous) mutations (dN) and silent (synonymous) mutations (dS) as well as their ratio (dN/dS) (Nei and Gojobori 1986). The calculated dN/dS value for each gene is compared with their simulated values of 100 nonparametric bootstrap runs to assess the predominance of dN over dS at the 95% significance level.

(8) TimeZone also determines Tajima D (Tajima 1989) and Fu & Li D* (Fu and Li 1993) values for each gene set. The presence of either selective sweep by an advantageous allelic variant, or any excess of deleterious or recently emerged allelic variants would result a negative Tajima D value. In the Fu and Li D* statistic, an excess of mutations in external branches would give rise to a negative value suggesting possible recent selective pressure for accumulating mutations. The calculation of Fu and Li D* here assumes the infinite site model.

(9) TimeZone creates a list of potential positively-selected genes (after being adjusted for the effect of recombination), which is developed based on meeting at least one of the following criteria:

(a) significantly higher allelic diversity in the evolutionarily recent zone;

(b) occurrence of evolutionarily recent structural hotspot mutations;

(c) significant high ratio of nonsynonymous to synonymous mutations in the terminal branches than in internal branches; and

(d) dN/dS values significantly higher than 1 (detected through nonparametric bootstrapping).

Apart from these things, the average pairwise divergence (π), number of haplotypes for any gene set etc. can also be calculated by TimeZone (a detail list of outputs is provided in Table 1).

Installation:

Version and operating system - The present version of TimeZone (TimeZone_v1.0) is available for 64 bit Windows systems (OS XP or higher).

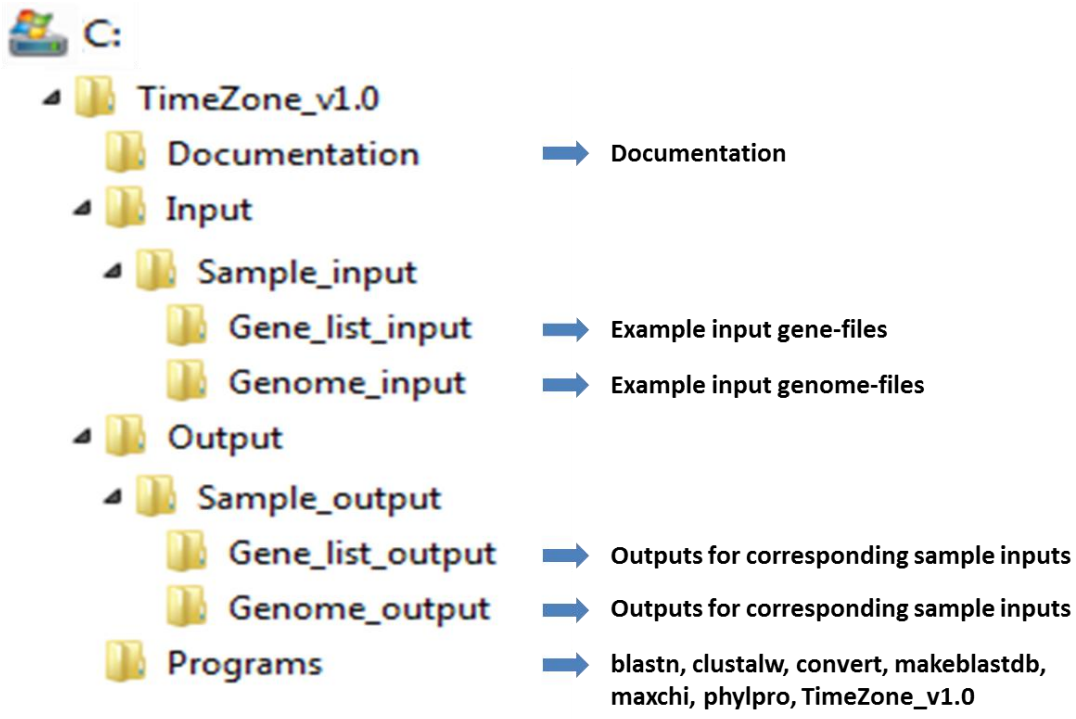
Minimum system requirements - 2 GB RAM, Dual-Core or equivalent processor, 56 MB of hard-disk space for program installation.

Source for installer download - https://sourceforge.net/projects/timezone1/

The stepwise installation procedure is given below:

1. Download “TimeZone_v1.0_installer” from https://sourceforge.net/projects/timezone1/. Clicking on the installer will automatically install “TimeZone_v1.0” on computer, preferably on C-drive within newly created “TimeZone_v1.0” folder. Some Windows operating systems might require administrative privilege to allow installation through User Account Control.

2. The installed folder “TimeZone_v1.0” includes 4 subfolders and they are: “Input”, “Output”, “Programs”, and “Quick_Start_Guide”. The subfolder “Programs” includes the applications ‘TimeZone_v1.0.exe’, ‘makeblastdb.exe’, ‘blastn.exe’, ‘clustalw.exe’, ‘convert.exe’, ‘maxchi.exe’ and ‘phylpro.exe’. Example as follows:



3. Another application required for functionality of TimeZone is the Windows version of PAUP* 4.0, a commercial software (Swofford 2000). Obtain 'paupWin32' from http://paup.csit.fsu.edu/downl.html, or more specifically only the application 'win-paup4b10-console.exe' (without any installation of paupWin32). Copy it from original PAUP directory to "TimeZone v1.0\Programs" directory, and rename it as 'paup.exe'.

4. Install Windows version of TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) or TreeViewX (http://darwin.zoology.gla.ac.uk/~rpage/treeviewx/download.html) as the default phylogenetic tree-viewing software to view the phylogenetic trees. To make it as a default tree-viewing software right-click any tree file (with 'dnd' extension) and select 'open with' mode. Then choose the 'TreeView' or 'TreeViewX' icon and select 'Always use the selected program to open this kind of file' option and finally click 'ok'. The Zonal phylogeny trees hyperlinked to the output spreadsheet file summarizing TimeZone analysis results now will be open on in tree viewing window by a single click . Alternatively, users can choose any other tree-viewing software that can read Newick-formatted tree-topologies.

Using TimeZone:

TimeZone enables users to perform analysis according to their need of dataset. Two types of datasets can be used as inputs:

A. Whole genome analysis: This feature enables users to perform analysis on all genes of any fully annotated genome (as reference) with the orthologs of each gene from any number of genomes. The users can specify the threshold value for the sequence identity

and length coverage to detect orthologous gene set. Only for the reference genome users need to provide the fully annotated genome but for other genomes it may be fully annotated genome or any draft genome of users' choice.

B. Gene list analysis: This feature is perfect for users looking for TimeZone analysis on genes set of their interest. Here users can provide any number of gene sets according to their choice from any number of organisms.

Input files:

Store the **input files in "TimeZone_v1.0\Input"** folder:

The input files needed for TimeZone analysis are as follows:

*** For analysis of genomes –**

(a) A file of annotated complete genome sequence in GenBank format (for the use as reference). *See sample file "genome1_ref.txt" in the "Genome_input" folder under "Sample_input".*

(b) A text file listing the filenames (with full extension) of other genomes to be analyzed. *See sample file "genome_set.txt" in the "Genome_input" folder under "Sample_input".*

(c) Files of the genomes to be analyzed. The genomes can be complete or draft or in contigs, but must be either in GenBank format or in Fasta format (as single sequence or multiple contigs). *See sample files "genome2.txt"through "genome7.txt" in the "Genome_input" folder under "Sample_input".*

*** For analysis of a list of genes –**

(a) Fasta-formatted nucleotide sequences of the genes to be analyzed with extension ".fasta". *See sample file "gene1.fasta" through "gene5.fasta" in the "Gene_list_input" folder under "Sample_input". (see Alerts)*

(b) A text file listing the genes to be analyzed. *See sample file "genelist.txt" in the "Gene_list_input" folder under "Sample_input".*

Running TimeZone:

Start TimeZone by –

(a) clicking "TZ", a shortcut icon for TimeZone in Desktop or in Start Menu;

(b) clicking "TimeZone_v1.0" application in installed "TimeZone_v1.0\Programs" folder;

(c) typing "TimeZone_v1.0" or "TimeZone_v1.0.exe" inside the "C:\TimeZone_v1.0\Programs" directory from command prompt window.

Input commands as prompted in TimeZone run window:

(a) Specify name of a folder to store the results-output (will be created in "TimeZone_v1.0\Output" folder).

(b) Choose analysis type (option 1 for analysis of genomes, option 2 for analysis of a list of genes).

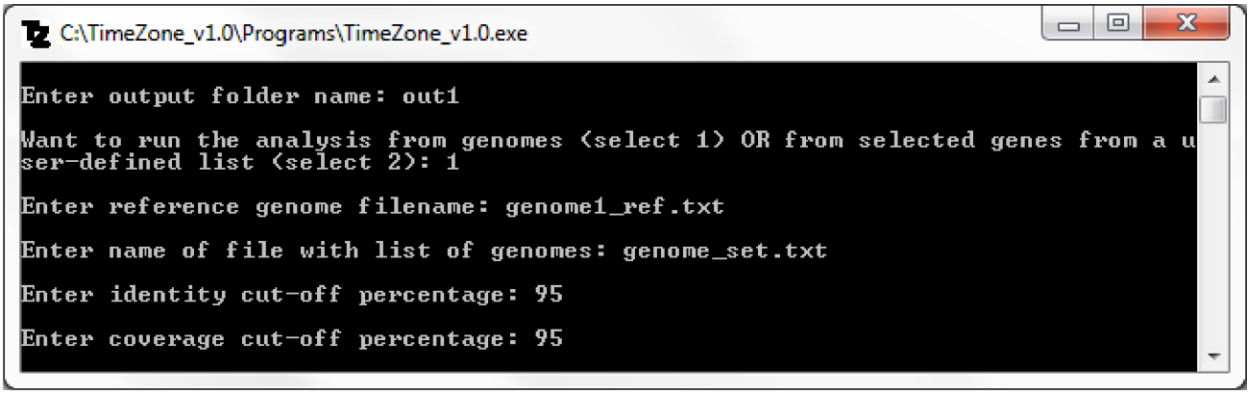
***For analysis of genomes (option 1) –**

(c) Specify reference genome filename with extension (Example: genome1_ref.txt).

(d) Specify name with extension of file containing list of genomes (Example: genome_set.txt).

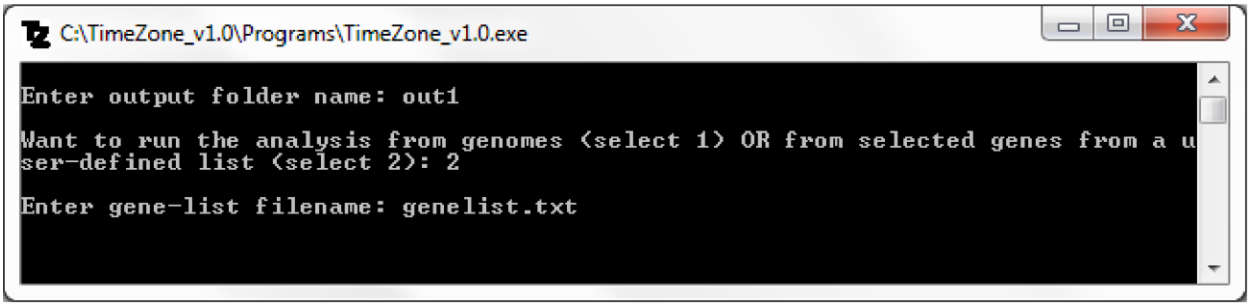
(e) Specify cut-off value for % sequence-identity (numbers only, from 0 to 100) (Example: 95).

(f) Specify cut-off value for % coverage of sequence-length (numbers only, from 0 to 100) (Example: 95). (See following screenshot of TimeZone running in the genome-analysis mode)



***For analysis of a list of genes (option 2) –**

(c) Specify name with extension of file containing list of genes (Example: genelist.txt). (See following screenshot of TimeZone running in the gene-list mode)



Output files:

Output-results will be created in the user-specified folder within **"TimeZone_v1.0\Output"**.

(a) For each gene, the output files include:

(i) Aligned (both *.aln and *.fasta) DNA and protein sequence datasets and phylogenetic trees (*.dnd) (ClustalW-derived);

(ii) ZP trees including all sequences (*-zp_tree_uni.all) as well as the unique (alleles) one (*-zp_tree_uni.dnd);

(iii) List of all synonymous and nonsynonymous (with corresponding amino acid) mutations along each branch of ZP tree (*-mutations.txt); and,

(iv) ZP analysis summary with details of structural (hotspot) mutations (*-ZP_hotspot-analysis.txt).

(b) Two spreadsheet outputs of genes and strains ('Genes&Strains-with_nonACGT.xls' & 'Genes&Strains-with_stop-codons.xls') that include sequences with non-ACGT characters and premature stop-codons respectively.

(c) A complete summary spreadsheet file ('TimeZone-summary.xls') of the analyzed genes from "CDS-details.xls" (for option 1, i.e. analysis of genomes) or from user-specified list (for option 2, i.e. analysis of a set of genes) is created to incorporate 5 groups of results:

(i) Protein-coding gene annotation information (with respect to the reference genome);

(ii) Zonal phylogeny and haplotype diversity analysis;

(iii) Analysis of structural hotspot mutations;

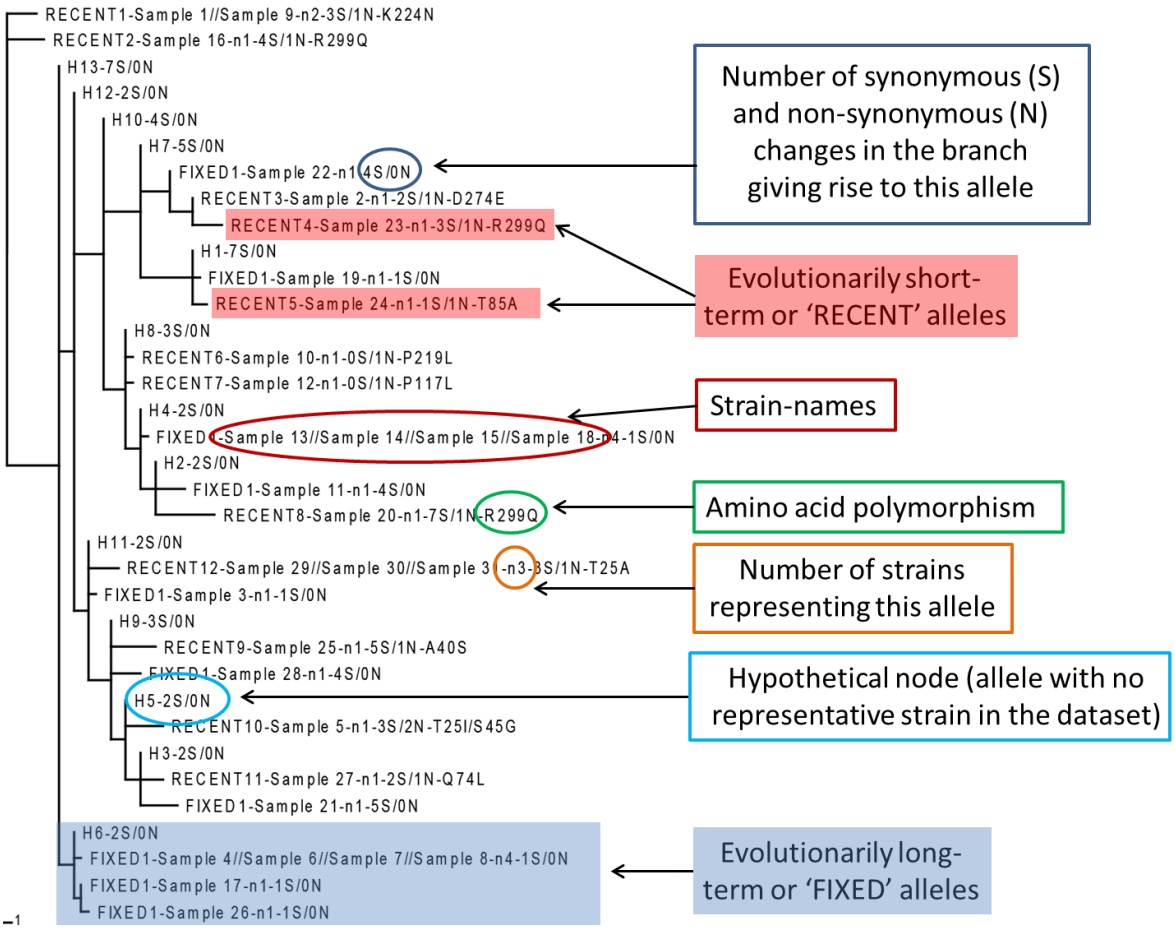
(iv) Synonymous/nonsynonymous mutations in internal vs. terminal branches; and,

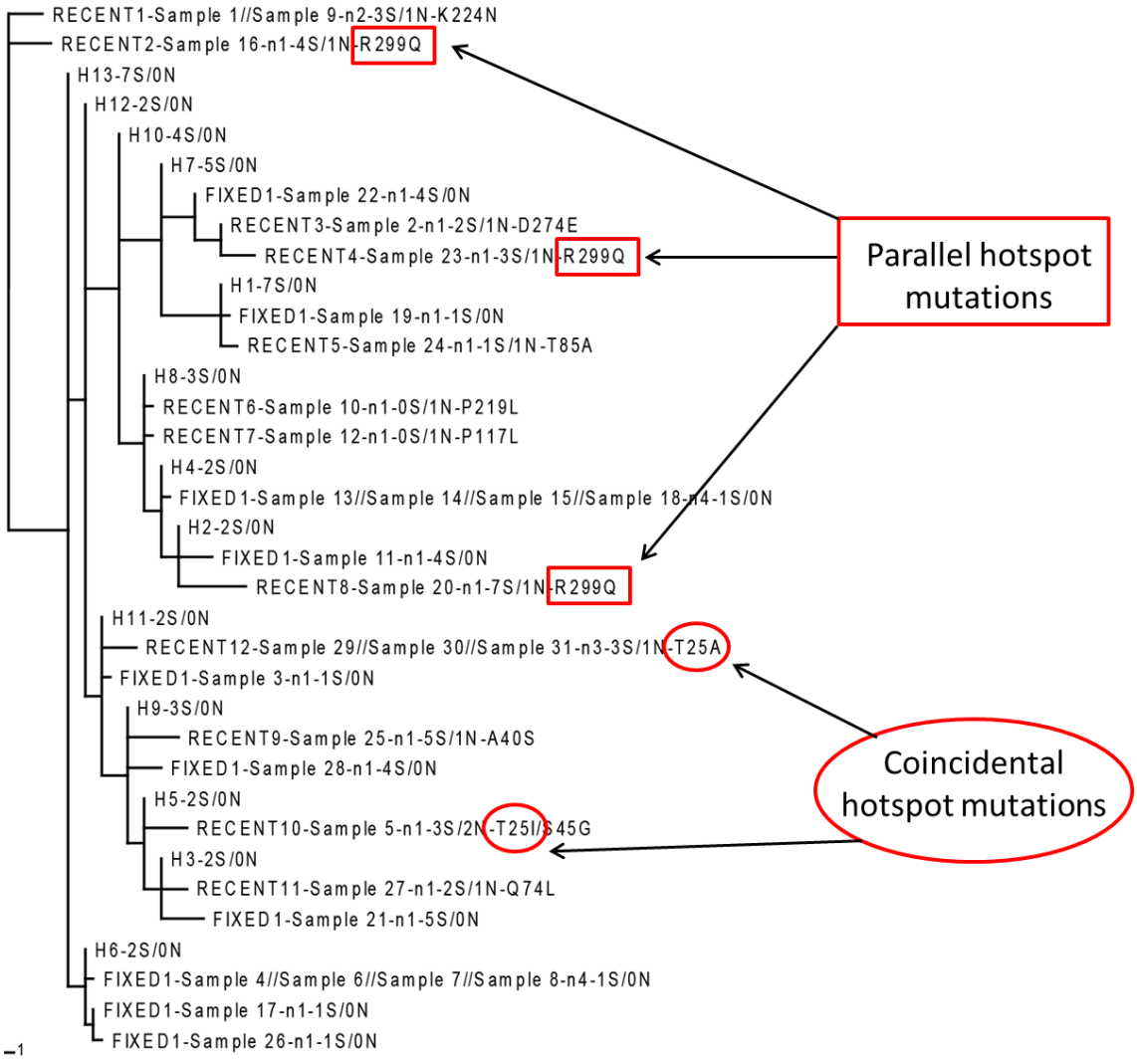
(v) Overall sequence diversity.

(d) A spreadsheet file, 'Recombinants-list.xls', includes probable recombinants (from a short-listed set of positively-selected genes) as detected by any of the two statistics, MaxChi and Phylpro.

(e) A list of selection-candidate genes, 'Selection-Candidates.xls', after being adjusted for the effect of recombination.

Example of a zonal phylogeny for gene list analysis from 'Selection-Candidates.xls' for gene3 (by clicking the "gene3-tree" cell in spreadsheet under ZP-Tree column heading): Information along each branch includes zonal specification ('FIXED' or 'RECENT' or 'H' denoting hypothetical alleles with no representatives in the sample), strain-names, number of strains, number and nature of substitutions ('S' for synonymous and 'N' for non-synonymous), and finally the amino acid change(s), if any. The examples of 'FIXED' and 'RECENT' alleles are marked by blue and red boxes respectively in the tree shown below.





In the above tree, 2 convergent (hotspot) amino acid mutation positions can be observed in short-term or recent zone: one between Sample 16, Sample 20 and Sample 23 at position 299 and another between Sample 5 and Sample 29/Sample 30/Sample 31 at position 25. The first mutation is of parallel type i.e. changes are same (R→Q) whereas the second one is an example of coincidental type i.e. changes are different (T→A and T→I).

Table 1. Description of each column-header of the Output-files mentioned in (c), (d), and (e).

Section	Column	Header	Description
	1	Gene#	Serial number of the order in which the genes were extracted
PROTEIN -CODING GENE ANNOTATION INFO	2	Gene Name	Name of gene (or locus-tag in absence of name) in reference genome annotation
	3	GI	Unique CDS identifier of gene in reference genome
	4	Strand	Gene present in plus (+) or minus (-) strand of reference genome
	5	CDS Region	Nucleotide-based position of gene in reference genome
	6	Product	Function of encoded protein as annotated in reference genome
	7	Protein Length (AA)	Amino acid length of encoded protein
	8	Representatives	Strains having the gene (i.e. qualifying the user-set identity/coverage thresholds)
	9	# of Seqs	Total number of orthologous sequences for the gene in sample set
ZONAL PHYLOGENY AND HAPLOTYPE DIVERSITY	10	ZP-Tree	Zonal Phylogeny tree (hyperlinked to the topology file stored in the same folder)
	11	#Strains-PRI	Number of strains in Primary (evolutionarily long-term/fixed) phylogenetic zone
	12	#Strains-EXT	Number of strains in External (evolutionarily short-term/recent) phylogenetic zone
	13	#Haplo-PRI	Number of haplotypes (or alleles) in Primary zone
	14	#Haplo-EXT	Number of haplotypes (or alleles) in External zone
	15	Haplo-Ratio	Ratio of “#Haplo-EXT” to total number of haplotypes
	16	Simp-PRI (SE)	Primary zone value Simpson’s diversity index (with standard error in parentheses)
	17	Simp-EXT (SE)	External zone value Simpson’s diversity index (with standard error in parentheses)
	18	Z-PRI_vs_EXT	Z-test value of “Simp-PRI (SE)” vs. “Simp-EXT (SE)”
	19	EXT>PRI diversity at P<0.05	If External diversity is significantly higher (“sig”) or not (“non-sig”)
STRUCTURAL HOTSPOT MUTATIONS	20	#HS-pos	Number of amino acid positions with structural hotspot changes
	21	#HS-mut	Number of structural hotspot changes
	22	#NonHS-mut	Number of non-hotspot structural changes
	23	HSfreq	Ratio of hotspot changes to total number of amino acid changes
	24	#Para-PRI	Number of parallel hotspot changes in Primary zone
	25	#Coinc-PRI	Number of coincidental hotspot changes in Primary zone
	26	#HS-PRI	Total number of hotspot changes in Primary zone

	27	#NonHS-PRI	Total number of non-hotspot changes in Primary zone
	28	HSfreq-PRI	Ratio of hotspot changes to total number of amino acid changes in Primary zone
	29	#Para-EXT	Number of parallel hotspot changes in External zone
	30	#Coinc-EXT	Number of coincidental hotspot changes in External zone
	31	#HS-EXT	Total number of hotspot changes in External zone
	32	#NonHS-EXT	Total number of non-hotspot changes in External zone
	33	HSfreq-EXT	Ratio of hotspot changes to total number of amino acid changes in External zone
SYN/NONSYN MUTATIONS IN INTERNAL VS. TERMINAL BRANCHES	34	#Tips	Number of tips or terminal branches
	35	#Twigs	Number of twigs or internal branches
	36	Tips-Syn	# of synonymous mutations in tips
	37	Tips-Nonsyn	# of non-synonymous mutations in tips
	38	Twigs-Syn	# of synonymous mutations in twigs
	39	Twigs-Nonsyn	# of non-synonymous mutations in twigs
	40	dN/dS-Tips (SE)	dN/dS value of tips (with standard error in parentheses)
	41	dN/dS-Twigs (SE)	dN/dS value of twigs (with standard error in parentheses)
	42	Z-Tips_vs_Twigs	Z-test value of “dN/dS-Tips (SE)” vs. “dN/dS-Twigs (SE)”
	43	Tips>Twigs dN/dS at P<0.05	If dN/dS of tips is significantly higher (“sig”) or not (“non-sig”)
OVERALL SEQUENCE DIVERSITY	44	#Syn	Number of synonymous changes
	45	#Nonsyn	Number of non-synonymous changes
	46	Pi (SE)	Average pairwise diversity value π/nucleotide (with standard error in parentheses)
	47	dS (SE)	Rate of synonymous changes (with standard error in parentheses)
	48	dN (SE)	Rate of non-synonymous changes (with standard error in parentheses)
	49	dN/dS	Ratio of dN to dS
	50	Bootstrap P	P-value of non-parametric bootstrapping (100 runs)
	51	dN/dS-based Selection	“Positive” (if Bootstrap P<0.05), “Purifying” (if Bootstrap P>0.95) or “Neutral”
	52	Tajima D	Value of Tajima D statistic
	53	Fu & Li D*	Value of Fu & Li D* statistic
RECOMBINATION TESTS	54	Rec-MaxChi	“YES” (if MaxChi statistic-based P<0.05 for any breakpoint detected) or “NO”
	55	Rec-Phylpro	“YES” (if Phylopro statistic-based P<0.05 for any breakpoint detected) or “NO”

Alerts:

For input gene files in case of gene-list analysis,

1. The first 10 characters of any two sequence-identifiers (i.e. Fasta header) in a gene-dataset should not be identical to run ClustalW.
2. Spaces or '-' in the sequence-identifier (i.e. Fasta header) within a gene dataset must be either removed or replaced by underscores ('_') to run PAUP*.
3. The Fasta-formatted gene sequence files must be without any stop-codons at the end (since the program excludes sequences with stop-codons from analysis).
4. PAUP* requires a minimum of 4 alleles to reconstruct an unrooted ML tree, hence should provide at least 4 alleles in gene list analysis. For genome wide analysis there should be at least 4 genomes in total.
5. Before any next run, users are advised to remove all the input files of the previous run from "TimeZone_v1.0\Input" folder to avoid any possible redundancy in filenames.

References:

Bush, R.M., Smith, C.B., Cox, N.J. & Fitch, W.M. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. *Proc Natl Acad Sci U S A.* **97**, 6974-6980 (2000).

Chattopadhyay, S., Dykhuizen, D.E. & Sokurenko, E.V. ZPS: visualization of recent adaptive evolution of proteins. *BMC Bioinformatics.* **8**, 187 (2007).

Chattopadhyay, S. *et al.* High frequency of hotspot mutations in core genes of Escherichia coli due to short-term positive selection. *Proc Natl Acad Sci U S A.* **106**, 12412-12417 (2009).

Christin, P-A, Weinreich, D.M. & Besnard, G. Causes and evolutionary significance of genetic convergence. *Trends Genet.* **26**, 400-405 (2010).

Fu, Y.X. & Li, W.H. Statistical tests of neutrality of mutations. *Genetics.* **133**, 693-709 (1993).

Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* **3**, 418-426 (1986).

Smith, J.M. Analyzing the mosaic structure of genes. *J Mol Evol.* **34**, 126-129 (1992).

Swofford, D.L. PAUP*: Phylogenetic Analysis Using Parsimony and Other Methods (software). *Sinauer Associates, Sunderland, MA.* (2000).

Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* **123**, 585-595 (1989).

Weiller, G.F. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Mol Biol Evol.* **15**, 326-335 (1998).