Table of contents:

Software Declaration ........................................................................................................ 4
Preface ................................................................................................................................. 4
Acknowledgements ............................................................................................................ 4
eGPS software development team .................................................................................. 5
Copyright ............................................................................................................................ 7
How to cite eGPS ............................................................................................................... 7
How to contribute to this manual ..................................................................................... 8
List of people who make this manual available ................................................................ 8
Part 1: Quick Start ............................................................................................................... 9
First time user .................................................................................................................... 9
Installing and launching eGPS ......................................................................................... 9
Uninstall eGPS ................................................................................................................... 9
Features and support ........................................................................................................ 10
Functional modules of eGPS ......................................................................................... 10
Technical support and updates ....................................................................................... 11
Reporting bugs .................................................................................................................. 11
Operation guidance course ............................................................................................... 12
Operation introduction .................................................................................................... 12
Preparation before analysis .............................................................................................. 13
Viewing alignment .......................................................................................................... 16
Viewing circRNA results ................................................................................................. 19
Opening the gene-to-gene-tree module ......................................................................... 24
Analyzing protein omics data ....................................................................................... 26
Analyzing genomic data ................................................................................................. 28
Testing neutrality ............................................................................................................. 31
Estimating evolutionary distance .................................................................................. 31
Building tree from alignment ......................................................................................... 34
Building trees with distance data ................................................................................... 36
Demography builder and coalescent-based simulation .................................................. 37
Running eGPS command-line version .......................................................................... 38
Plug-in ............................................................................................................................... 39
Part 2: Multi-omic Analysis

Genomics........................................................................................................... 42
VCF snapshot ........................................................................................................... 44
VCF tools parameters ............................................................................................ 47
VCF analysis ........................................................................................................... 50
Transcriptomic ....................................................................................................... 53
circRNA viewer ....................................................................................................... 53
RNA-seq data analysis .......................................................................................... 53
Protein omics ......................................................................................................... 54

Part 3: Visualizing and Exploring Data

eGPS tree viewer .................................................................................................... 55
Toolbar ...................................................................................................................... 55
Left-side operations navigating bar ........................................................................ 56
Popup menu ............................................................................................................ 58
Viewing alignment .................................................................................................. 62
Viewing evolutionary distance ................................................................................ 63
Viewing genomic location in genome browser .................................................... 63

Part 4: Evolutionary Analysis

Pipeline to reconstruct phylogenetic tree ............................................................. 64
Candidate gene to evolutionary analysis ............................................................... 65
Align multiple sequences ...................................................................................... 67
Evolutionary analysis for genomic regions of interest .......................................... 67
Genomic evolutionary analysis: construct species tree for whole genome sequence alignment .................................................................................................................. 69

Part 5: Population History Modelling

Build Model .......................................................................................................... 71
Simulate ................................................................................................................... 72

Part 6: Plugin development..................................................................................... 75
Part 7: update recordes.......................................................................................... 76
V1.01 ....................................................................................................................... 76
V1.0.2 ..................................................................................................................... 76

Appendix: Frequently Asked Questions ................................................................ 77
Glossary ................................................................................................................... 77
Reference ................................................................................................................ 79
**Manual usage guidance:** The first part is the software declaration including general instructions. The second part “Quick Start” is to get started quickly. If you use the software for the first time, you should start from the second part which includes how to use the software such as installing and some illustrated tutorials. These tutorials cover the features of eGPS. The second part “Quick Start” offers tutorials for the users to get started quickly. The rest parts are detailed descriptions of each function including the illustration of each module’s parameters and the format statements of the input and output files.

**Software Declaration**

**Preface**

Multi-omic and evolutionary analyses have been booming recently, and interactive figures and software are believed to be important for these studies. Moreover, cloud computing became popular recently both in scientific research and industry, which provided a convenient way to manage big data and acquire computation resources. Therefore, we combine the advantages of cloud computing and desktop application, and design highly interactive user interfaces and pipelines to bridge the gap between multi-omic and evolutionary analyses. A number of sophisticated methods are integrated and implemented in the eGPS 1.0 platform, which is different from the widely used biological software.

Third-party plugins are supported and developers take all credits for contributing the plugins. This makes it easier for the community to implement new modules and encourages the sharing of them. A command-line eGPS 1.0 is also provided for high-throughput and streamline analysis. The eGPS 1.0 cloud and desktop application is freely available at [http://www.egps-software.org/](http://www.egps-software.org/).

eGPS 1.0 desktop application was developed in Java, so it can run across different computer platforms. The multithread analyses are also supported, by switching modules. The grid-computing will be implemented in the next versions.

**Acknowledgements**

We thank the following individuals for their contributions to this project: Dr. Ya-Ping Zhang, Dr. Yonggang Yao, Dr. Peng Shi, Dr. Xuemei Lv and the administration office members of the Strategic Priority Research Program (the evolutionary Genotype-Phenotype Systems biology) at Kunming Institute of Zoology, Chinese Academy of Sciences for their constructive suggestions and
coordination; Dr. Shuhua Xu, Dr. Sijia Wang, Dr. Yong Zhang and their students for testing the software. This work was fully supported by a grant from the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13000000, evolutionary Genotype-Phenotype Systems biology, abbreviated as eGPS). The software was named after the grant.

Thank following users employ eGPS and provide effective feedback:

<table>
<thead>
<tr>
<th>eGPS version</th>
<th>Source</th>
<th>Suggestions/bugs</th>
<th>Resolve status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0.0</td>
<td>QQ Group: Yun zhong yue</td>
<td>At the time of installation, 360 (antivirus software) notified of a security risk.</td>
<td>Please ignore or turn off the anti-virus software, eGPS is safe and non-toxic.</td>
</tr>
<tr>
<td>1.0.0</td>
<td>E-mail: -</td>
<td>Alignment viewer cost a lot of time to load large file.</td>
<td>Refactored the module so that it can load large files quickly.</td>
</tr>
<tr>
<td>1.0.1</td>
<td>QQ Group: Hong</td>
<td>When used on MAC OS, the startup script does not have certain permissions.</td>
<td>We offer a green and free-installation version for MAC OS.</td>
</tr>
<tr>
<td>1.0.1</td>
<td>QQ Group: Hong</td>
<td>When importing the PRO file for analysis, clicking the Run button does not respond.</td>
<td>Some data is not suitable for analysis when using the DEG-mass spectrum module. If this happens, a notification dialog will pop up.</td>
</tr>
</tbody>
</table>

**eGPS software development team**

**List of Developers:**
Dalang Yu, Lili Dong, Fangqi Yan, Hailong Mu, Bixia Tang, Xiao Yang, Tao Zeng, Qing Zhou, Feng Gao, Zhonghuang Wang, Ziqian Hao, Hongen Kang, Yi Zheng, Hongwei Huang, Yuzhang Wei, Wei Pan, YAOchen Xu, Junwei Zhu, Shilei Zhao, Ciran Wang, Pengyu Wang, Long Dai, Mushan Li, Li Lan, Yiwei Wang, Hua Chen, Yi-Xue Li, Yun-Xin Fu, Zhen Shao, Yiming Bao, Fang-Qing Zhao, Luo-Nan Chen, Guo-Qing Zhang, Wenming Zhao, Haipeng Li

**List of development units:**
1) CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Yueyang Road 320, Shanghai 200031, China.
2) BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.
3) Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China.
4) CAS Laboratory of Genomic and Precision Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China.
5) Shanghai Center for Bioinformation Technology, Shanghai 200235, China.
6) Key Laboratory of Systems Biology, Center for Excellence in Molecular Cell Science, Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China.
7) Department of Biostatistics and Data Science, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA.
8) Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China.
9) School of Life Science and Technology, ShanghaiTech University, Shanghai 200031, China.
10) Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.
11) University of Chinese Academy of Sciences, Beijing 100049, Beijing, China.
Copyright

This software is protected under the copyright law. No part of this manual or program design may be reproduced without written permission from copyright holders. Please e-mail all inquiries to lihaipeng@picb.ac.cn.

EGPS SOFTWARE LICENSE TERMS
EGPS DESKTOP APPLICATION
These license terms are an agreement between CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health (SINH), Chinese Academy of Sciences and you. Please read them. They apply to the software named above. By copying, downloading, or using the software, you accept these terms. If you do not accept them, do not use the software. Instead, please delete and uninstall the software from your computer.

SCOPE OF LICENSE. This agreement gives you rights to download, copy, and use the eGPS software for free. SINH reserves all other rights. You may not 1) work around any technical limitations in the software; 2) reverse engineer, decompile or disassemble the software; 3) use the software in any way that is against the law; 4) rent or lease the software; or 5) use the software for commercial software hosting services.

BACKUP COPY. You may make backup copies as many as you want.

REDISTRIBUTION. You can copy the software, together with this license, to anyone who you know or may not know.

LIMITATION ON AND EXCLUSION OF DAMAGES. SINH has no responsibilities for any kinds of damages on your properties because of copying, downloading, and using the eGPS software.

UPGRADE. You can visit www.egps-software.org for the latest version.

How to cite eGPS

How to contribute to this manual

You are very welcomed to edit this manual and share your experience on the eGPS software with other users. The word version of manual can be downloaded from [www.egps-software.org](http://www.egps-software.org). Please turn on Track Changes when you edit the manual. Or you can highlight your changes/edits. Please email us the changed parts, together with your full name.

List of people who make this manual available

You are very welcomed to edit this manual and share your experience on the eGPS software with other users. The word version of manual can be downloaded from [www.egps-software.org](http://www.egps-software.org). Please turn on Track Changes when you edit the manual. Or you can highlight your changes/edits. Please email us the changed parts, together with your full name.

<table>
<thead>
<tr>
<th>Name</th>
<th>Email address</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAN, Li</td>
<td><a href="mailto:lanl@big.ac.cn">lanl@big.ac.cn</a></td>
</tr>
<tr>
<td>LI, Haipeng</td>
<td><a href="mailto:lihaipeng@picb.ac.cn">lihaipeng@picb.ac.cn</a></td>
</tr>
<tr>
<td>MU, Hailong</td>
<td><a href="mailto:mhl_dyx@126.com">mhl_dyx@126.com</a></td>
</tr>
<tr>
<td>YAN, Fangqi</td>
<td><a href="mailto:yfq0818@163.com">yfq0818@163.com</a></td>
</tr>
<tr>
<td>YU, Dalang</td>
<td><a href="mailto:yudalang2017@sibs.ac.cn">yudalang2017@sibs.ac.cn</a></td>
</tr>
</tbody>
</table>

*Listed in alphabet.
Part 1: Quick Start

First time user

First of all, thank you very much for choosing eGPS in your research. This user manual will provide you with comprehensive documentation for the eGPS desktop application. New users of eGPS may wish to read and follow the walkthrough tutorial, which demonstrates the major features and functions of eGPS that you may find useful.

At the same time, you can quickly locate the function modules you interest by viewing the main functions of eGPS.

Installing and launching eGPS

On Windows operating system, we provide exe format installers for the 32-bit and 64-bit operating systems. Users only need to follow the prompts.

On the Mac OS operating system, we provide a green, free-install version of the 64-bit operating system. For the first time, please unzip the file and right click to open it. EGPS is green and non-toxic software, please feel reassured to use.

On the Ubuntu operating systems, Currently, the eGPS provided at this stage is a green free installation software that comes with all the operating environments. Users only need to download the Zip file from the official website and then extract contents and execute running file.

The latest version of Installers and Zip files can be downloaded from www.egps-software.org.

Installation Precautions: When selecting an installation path, please do not include any non-English symbol in the path.

Uninstall eGPS

For MAC OS and Ubuntu, you need to do is to delete the folder where the software is located. For Windows, you can uninstall eGPS as natural way. For instance, in Windows10, go to Programs and Features and click uninstall.
Features and support

Functional modules of eGPS

The complete functional modules in the eGPS desktop application are shown in Table 1.

<table>
<thead>
<tr>
<th>classification</th>
<th>emphasis</th>
<th>Module Name</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolution</td>
<td>Sequence aligner</td>
<td>Multiple sequence Aligner</td>
<td>Multiple sequences</td>
</tr>
<tr>
<td>Evolution</td>
<td>Sequence alignment</td>
<td>Alignment viewer</td>
<td>Multiple sequence alignment(MSA)</td>
</tr>
<tr>
<td>Evolution</td>
<td>Evolutionary distance</td>
<td>Genetic distance viewer</td>
<td>Genetic distance(DIST)</td>
</tr>
<tr>
<td></td>
<td>Evolution phylogeny</td>
<td>Tree viewer</td>
<td>Phylogenetic tree(TREE)</td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>VCF snapshot</td>
<td>Variant Call Format (VCF)</td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>VCF tools</td>
<td></td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>Genetic diversity</td>
<td></td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>Neutrality test</td>
<td>Variant Call Format (VCF)</td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>PBS statistic</td>
<td></td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>HKA score</td>
<td></td>
</tr>
<tr>
<td>Omic study</td>
<td>Genomics</td>
<td>FST statistic</td>
<td></td>
</tr>
<tr>
<td>Omic study</td>
<td>Transcriptomics</td>
<td>DEG-RNAsEq</td>
<td>RNA expression profile (rnaExp)</td>
</tr>
<tr>
<td>Omic study</td>
<td>Proteomics</td>
<td>DEG-massSpectrum</td>
<td>Proteomics expression matrix(PRO)</td>
</tr>
<tr>
<td>Evolution</td>
<td>Population genetics</td>
<td>Simulator</td>
<td>Population history model (SIMU)</td>
</tr>
<tr>
<td>Omic study</td>
<td>Transcriptomics</td>
<td>circRNA viewer</td>
<td>Circular RNA list (LIST)</td>
</tr>
<tr>
<td>Evolution</td>
<td>Evolution phylogeny</td>
<td>Gene to gene tree</td>
<td>/</td>
</tr>
<tr>
<td>Evolution</td>
<td>Evolutionary distance sequences alignment</td>
<td>Gene to genetic distance</td>
<td>/</td>
</tr>
<tr>
<td>Evolution</td>
<td>Evolutionary distance sequences alignment</td>
<td>Gene to alignment</td>
<td>/</td>
</tr>
</tbody>
</table>
Technical support and updates

The latest version of eGPS will be available at the website http://www.egps-software.org/

Reporting bugs

Welcome to report bugs when any problems such as unexplained dialogs, documentation inconsistencies, or program crashes happened. Please report them to us by clicking the ‘Report Bugs’ link in ‘Help’ menu.
Operation guidance course

Operation introduction

This walkthrough provides a variety of concise tutorials that explain how to perform common tasks in the eGPS. Sample data files are required for each tutorial and can be found in the examples folder.

In the tutorial, we use the following conventions:

- Keystrokes are represented by bold letters (e.g., F1).
- *Italicized* words indicate the name of a menu, toolbar or window (e.g., *File* menu).
- *Italicized bold* words indicate individual commands that are found in menus, submenus, and toolbars (e.g., *Import Data* command submenu).
- For the sake of brevity, a series of menu/button clicks are separated by a pipe symbol (|) to display a series of command indications (for example, ‘*File | Import Data*’ means you should click ‘file’ main menu item, then click ‘Import Data’ submenu item.

Main interface of eGPS

The eGPS analysis interface consists *menus, tool bars, the status bar*, a main window with several *analysis panels* and a *history panel* to store records of important analysis procedures.

When a researcher executes an analysis process, progress animation will demonstrate at status bar. After the procedure finish properly, the link of the output datasets will be added to the history panel to form a record which allowing annotation and flagging. The history log files will be stored locally and can be copied to and viewed in another computer where eGPS desktop application is installed.

Tool tips will be displayed when the mouse is hovering on records, which provides a wealth of information including the tool name and running parameters. Records are initially grouped by date and can be searched by entering partial keyword. When user searches, the history dynamically updates to show the matching records and the analysis panel will be showed as a separate tab after double clicking the record of interest.

Among analysis panels, there is a special and unique locked *data panel* which can be divided into three regions: a *data area* on the left where to drag files into and shows the
files already dragged; an *information area* in middle to display information of input files; a *method area* which holds several buttons on the right for the corresponding suitable approaches (eGPS will set the suitable analysis methods as buttons). Every analysis panel except data panel holds a functional module and these panels can be dragged out of mainframe becoming a floating window, which facilitates users to easily compare the results and parameters.

Figure 1. The main user interface of the eGPS desktop application.

**Preparation before analysis**

**Importing data**

Users have three ways to import data into eGPS:

- Directly drag the files to be analyzed and drop files to *Data area*.
- Import data using the shortcut button ![shortcut button](image) in the *toolbar*.
- Click *File | Import Data* to import data.

After importing data, the data information and suitable methods will be displayed in the *information*
area and method area. The figure above shows the eGPS response after dragging into a series of VCF files.

Note: Users can only import the same type of data format at a time. If you drag a file in another format, eGPS will give a warning message and will not add other files to the data area. If you want to change the data file, you need to click the Select all button and then click the Delete Selected button in the bottom left corner.

Table 2 shows how eGPS classifies the data: firstly, we divided data into data type then subdivided data type into data format. Only the same type of data format could be imported at one time.
Table 2. Data formats supported in the eGPS desktop application.

<table>
<thead>
<tr>
<th>Module Name</th>
<th>Data type</th>
<th>Data format</th>
<th>Data example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple sequence Aligner</td>
<td>Multiple sequences</td>
<td>fasta/msf/clustalw/paml/mega/phylip</td>
<td>mtCDNA.fas</td>
</tr>
<tr>
<td>Alignment viewer</td>
<td>Multiple sequence alignment(MSA)</td>
<td>fasta/msf/clustalw/paml/mega/phylip</td>
<td>mtCDNA.fas</td>
</tr>
<tr>
<td>Genetic distance viewer</td>
<td>Genetic distance(DIST)</td>
<td>dist</td>
<td>testGeneticDistance.dist</td>
</tr>
<tr>
<td></td>
<td>Multiple sequence alignment(MSA)</td>
<td>aligned fasta</td>
<td>mtCDNA.fas</td>
</tr>
<tr>
<td>Phylogenetic tree(TREE)</td>
<td>NEXUS</td>
<td>nwk/nhx/tre/etree</td>
<td>Influenza_A_virus.nex</td>
</tr>
<tr>
<td>Tree viewer</td>
<td>NEXUS</td>
<td>nex/nexus</td>
<td>testGeneticDistance.dist</td>
</tr>
<tr>
<td>VCF snapshot</td>
<td>Variant Call Format (VCF)</td>
<td>Uncompressed vcf / Gzip compressed vcf / Bgzip compressed vcf</td>
<td>vcf.sample.test.chr20.vcf</td>
</tr>
<tr>
<td>VCF tools</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genetic diversity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutrality test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PBS statistic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HKA score</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FST statistic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DEG-RNAseq</td>
<td>RNA expression profile (rnaExp)</td>
<td>rnaExp</td>
<td>GDS2676.rmaExp</td>
</tr>
<tr>
<td>DEG-massSpectrum</td>
<td>Proteomics expression matrix(PRO)</td>
<td>pro</td>
<td>MAP_testing.pro</td>
</tr>
<tr>
<td>Simulator</td>
<td>Population history model (SIMU)</td>
<td>simu</td>
<td>sample.simu</td>
</tr>
<tr>
<td>circRNA viewer</td>
<td>Circular RNA list (LIST)</td>
<td>list</td>
<td>1519652983250_out_jav.list</td>
</tr>
</tbody>
</table>

Setting parameters

There are two ways to set eGPS parameters, including global settings and specific settings for each module. Global Settings can be called up by clicking on Options | Preferences, or by entering the shortcut key Control + R. The construction parameters of the phylogenetic tree are in the global
settings, while the parameters of the omics analysis are mostly in their respective modules. For example, if you want to build a distance based phylogenetic tree from multiple alignment sequences data, you can set genetic distance calculation method and tree reconstruction method by clicking Options | Preferences | Genetic distance and Tree build method respectively.

**Viewing alignment**

Firstly, drag aligned fasta file to eGPS, as mentioned before, we have three ways to import data.

- Drag and drop data into data area, as shown in the following figure.
- Click first shortcut button in toolbar, use choose file dialog.
- Click File | Import Data as before.

When the fasta file is dragged in, there are two ways we can get into Alignment viewer.

1. Click directly on the right View alignment button, as shown in the following illustration.
2. **Tools | Alignment viewer**

Finally, we enter the *Alignment Viewer* module to display the data as shown in the following figure.
1. The data will be displayed block by block. The number of sequences and columns displayed in each block depends on the current font size and window size. The font size can be adjusted via the buttons on the toolbar.

2. When the mouse hovers over the nucleotide, the number of columns corresponding to the current nucleotide is displayed on the screen. When each base is clicked, the border of the base turns red.

3. The base color scheme is provided on the left, and the color matching effect is as described as the string.

4. Below sequences is a histogram showing Consensus for each column. Tooltips of the percentage will be displayed when the mouse is over.

Export data:

We can save our data by clicking File | Save As or toolbar shortcut. Temporarily, we support to export fasta/msf/clustalW/paml/mega/phylip/nexus format file.
Viewing circRNA results

eGPS can be employed to interactively visualize the results of de novo circular RNA identification from the eGPS cloud.

1. Import data
In this module, we have two ways to import data into eGPS.

(1) Drag circRNA file (.list) into program, then click avoidable method turn to circRNA viewer tab to visualize circRNA, figures are showed below.

(2) Without dragging in any files, simply click Tools | circRNA viewer, circRNA viewer tab will not display any figure. We can enter a Job ID which is obtained from eGPS cloud, then click Download circRNA button to download data file.

Two files (one ending with _out_jav.list and another ending with _out_library_length.list) will be downloaded, and two files will be added to our program. The circRNA genesis procedure will be
Filter circRNAs to display target ones:

Users can selectively display circular RNA through adjust various filter parameters in *Filter parameters* Panel such as min expression.

Users can control the number of circRNA genesis panel and current presented panel through operations in *Layout* panel. The eGPS allows exhibiting maximum number of 4 circRNA genesis
3. circRNA viewer display panel (right side):
The circRNA viewer supports locking a single circRNA genesis panel. If the circRNA in the view is locked, the locked circRNA will not be switched when the current page go to next, as shown in the figure below.
Among the genomic interval, the menu bar can be displayed by clicking the left mouse button.

Users can view this interval in a genome browser (such as Ensembl, UCSC), construct a phylogenetic tree of homologous sequences between closely related species, and view the alignment of multiple sequence alignments.

By moving the mouse to genomic interval in the genesis panel, users can see the relationship between the circRNA ring region and linear genomic region:
4. circRNA viewer data format:
Two files are needed to visualize both circular RNA and the producing genesis. When only dragging the file ending with _out_jav.list, clicking the genesis panel will not display the circular circRNA.

When dragging only files ending with _out_library_length.list, the view cannot jump.

5. circRNA viewer data export and print:
When the data is displayed on the interface, we can click the button to save the data to our local disk or print the view as shown below:
Users can save the graphics with jpg, png, pdf, and svg file formats.

**Opening the gene-to-gene-tree module**

Open Gene to gene tree function module: **Tools | Gene to gene tree**

Users can enter gene symbol or genomic location in **Gene & species set panel**.
You can set species set group by clicking **Choose species set** button. The alignment will be obtained from eGPS Cloud or Ensembl REST API.
If you enter the gene symbol eGPS allows users to choose specific genomic regions.

Finally click **Build gene tree** button to start the task, the real-time construction progress will display at **status bar**.
When the task accomplishes, we can analyze the phylogenetic tree. The detailed manual to operate tree please refer to Part 3 “eGPS tree viewer”.

Analyzing protein omics data

The MAP method implemented in eGPS enable discovering differential expression genes from processed pair-ware mass spectrometric data.
Firstly, import files with correct suffix (.pro), clicking the suitable method *DEG-mass spectrum* to get in analysis module.

We leave parameters as default, clicking *analysis* button to start calculation procedure. After the progress bar progress running from 0% to 100%, an interactive MA Plot appears in the right area. We can place the mouse over an red point, a wealth of tooltip will display information about each point.
eGPS allows users to click on genes of interest to build a phylogenetic tree. We can click on the point with the left mouse button and select **Build gene tree** and click to build a phylogenetic tree.

At the same time, users can also click on the **Open genome browser** to view this gene on the genome browser.

### Analyzing genomic data

The Variant Call Format (VCF) file is widely used in bioinformatics to store sequence variations. But it is not easy to take a glance of it, manipulate (*i.e.* filtering) it and process it in a user-friendly way partially due to the largeness of the file. So, at the most cases, researchers require additional programming skills to accomplish these operations by the run mode of command line in Linux.

Taking the advantage of eGPS, researchers can use “VCF snapshot” to quickly snapshot the file and this could be even ultra-faster when the file compressed by bgzip; use “VCF tools” to manipulate the file with various filtering conditions, and other suitable methods of functional modules to analyze the file to get summary statistics (such as statistics of genetic diversity and positive selection) of each slide window.

EGPS supports selecting individuals in an intuitive way and restructuring specific genomic regions by incorporating a bed file.

After executing the process task in functional module, eGPS displays the results with interactive point plot powered by Echart3.0 (http://echarts.baidu.com/).

### Viewing VCF file

Import VCF (Variant Call Format) file (e.g. chr22.100000.lines.vcf) to **data area**.

If the files you dragged are the standard VCF file specification, the analysis method for the VCF file will be listed in the **method area**.

**VCF snapshot** is in the first one, and click this button will go to this module.

After the successfully loading data, eGPS will read the file according to the default parameters and display it to the interface with the form of a table.
Filtering VCF file

Click VCF tools button to go to VCF tools tab, eGPS provides a serious filtering options to help user to filter variant records. For example, if users want to only keep SNP records, then you can go to Variant Type/ Site ID Filter panel select Keep only SNPs check box. Finally click Save Filtered Data As to run the task.
Analyzing VCF file

Same as previous operations, click on the corresponding analysis module button, for example, select the *Genetic diversity* module, and then choose the statistics of interest for calculation. The default setting is to calculate Theta and Pi. We leave parameters as default and click on the *analysis* button. The result is shown below:
Here the abscissa of each point represents a region on the genome, and the ordinate represents the calculated statistic for that region. Users can also click on points of interest to build a phylogenetic tree.

In addition to the Genetic diversity module, you can also click on the FST Statistic module to analyze population genetic structures.

Please keep this state to practice the tutorial for neutral testing.

**Testing neutrality**

User can employ the eGPS to get the summary statistics from VCF data by sliding window analysis. Similar to the previous tutorial, you can choose the Neutrality Test, HKA score, PBS value and other modules for analysis.

Also you can conveniently compare these genomic statistics by drag the tab out of main frame.

**Estimating evolutionary distance**

1 Estimating Evolutionary Distances Using Pairwise Distance

In eGPS, you can estimate evolutionary distances between sequences by computing the proportion of nucleotide differences between each pair of sequences.
Import aligned fasta file (multiple sequence alignment file), configure calculation parameter (Option | Preferences | Genetic distance), select p-distance, we leave other parameters as default. Finally, we click Calculate distance button as shown in the figure.
After a short display of the progress bar, the distance results will be displayed in a new tab with a matrix table. Keep this window open so that we can compare the results of the following steps.

2 Compute and Compare Distances Using Other Models/Methods

eGPS also supports a wide collection of models for estimating evolutionary distances. Here we compare evolutionary distances calculated by using different models/methods.

Repeat tutorial above, but select the *Jukes/Cantor model* under the Model/Method pull-down instead of the *p-distance model*, leaving all the other options the same. Again, leave the results window open for comparison.

Repeat the analysis, this time selecting the *Tamura-Nei* model under the Model/Method pull-down, leaving all the other options the same. Again, leave the results window open for comparison.

Users are now able to compare the three open result windows which contain the distances estimated by the different methods. eGPS supports to drag all tab out of main frame so they could conveniently data.
Building tree from alignment

In this example, we will use eGPS to construct NJ phylogenetic tree and get familiar with Tree Viewer functional module.

Firstly, import aligned fasta file (multiple sequence alignment), and then configure genetic distance calculate parameters (Option | Preferences | Genetic distance). Choose p-distance, then click Tree build method

Select Neighbor-Joining. We leave other parameters as default (see the following picture).
Finally click the **Build tree** button.

After the progress bar indicating the task is finished, the phylogenetic tree will appear in a new tab. To select a branch, use the left mouse button to click on it. After selecting it, you can perform a series of operations, such as increasing the label of the leaf node, changing the color of the line, and so on. Users can also change the layout of the tree. For example, to become a circular layout, click on **Circular layout**. For more details on how to use it, see the eGPS tree viewer in Part 3.
Building trees with distance data

Import distance matrix file into eGPS, then configure the tree reconstruction method. *Option | Preferences | Tree build method* Select *Neighbor-Joining*. Finally, click *Build tree* button.

For the genetic distance file format, we use the format compatible with the phylip program. The eGPS also allows user to annotated data at header lien start with ‘#’ symbol. See part 4 for details.
Demography builder and coalescent-based simulation

We provide a very user-friendly tool called “simulator” to build demographic models by clicking mouse and dragging model icons and simulate the sequences under specific models with animation of monitor parameters.

Click **Tools | Simulator**, drag corresponding icons from **Build Model | Tools Menu** to construct population history. We leave parameters in **Parameters** as default.

Click sub panel **Simulator** (left top in this module), because only one group available at the current time while group historical modules were being set up. So one group appears in the **Sample information** column, and we set the sample size of this group to 15. Other parameters are set to default, click **Run only** to generate simulation data, and the histogram on the right will show the distribution of some statistics. The user can also click **Run & Save data as** to simulate and save the data.
Running eGPS command-line version

The current command line version of eGPS can calculate some statistics for VCF files. For example, to calculate the FST between two groups, users can run on various operating systems as long as Oracle JRE 1.8 has been installed.

For example, if we run the command line version in the window 10 operating system.
You can see the help information by using \texttt{--h} option.

The command line above is:

\begin{verbatim}
java -cp \eGPS_v1.0.jar egps.run.FST -i1 \sampleFiles_oneChromPerFile\pop1_SampleIDs.txt -i2 \sampleFiles_oneChromPerFile\pop2_SampleIDs.txt -mq 20 -sz 500 -st 100 -v ",\sampleFiles_oneChromPerFile\vcf.sample.test.chr20.vcf;\sampleFiles_oneChromPerFile\vcf.sample.test.chr21.vcf;\sampleFiles_oneChromPerFile\vcf.sample.test.chr22.vcf" -inb \sampleFiles_oneChromPerFile\sample.bed
\end{verbatim}

Also other statistics can be calculated! For example:

\begin{verbatim}
java -cp \eGPS_v1.0.jar egps.run.HKA -mq 2
0 -sz 5000 -st 1000 -v \HKA_test_data\test_vcf.recode.filter.vcf -i1 \HKA_test_data\BlackMuntjac_SampleIDs.txt -i2 \HKA_test_data\IndianMuntjac_SampleIDs.txt
\end{verbatim}

\section*{Plug-in}

\subsection*{File format convertor}

This module supports to convert two file formats. We took a lot effort to design a friendly user interface, and more file formats will be supported very soon.

- .emf file to widely used .maf file
- .etree file to widely used .nwk file

1. There are two ways to import data:
   Click \textit{Add files} button to open choose file dialog. You can import one or several files at once, but make sure to keep import file format consistent.
Or you can click add directory to import all files in the directory.

2. Choose the destined file format. When we import emf file we can only select maf format and nwk for etree format.

3. when data file is imported, an convert item will be added in the middle panel.

Click button to open directory choose dialog to set output directory.

5. Click Convert or Convert All to start convert task. Convert All means convert all files; Convert means only convert selected file.
We can terminate the conversion process by clicking the delete button on right top side when the files have been converted.

**Population structure index $F_{ST}$ calculator**

Launch the eGPS, make sure the installation path “config/plugin” directory including swingFSTDemo.jar file. Click **Plugins** | **swingFSTDemo** to get in the module.

Once users enter the number of each genotypes, you can demonstrate how to calculate the $F_{ST}$. Now eGPS supports to calculate results by incorporate up to three populations. To see the genetic structure of the two groups, set the input value of the third group to null. You can click the calculate button to display the results.
Part 2: Multi-omic Analysis

Genomics

The functional modules of the genomics are designed to processing multiple various processing of VCF files, including:

- Rapidly snapshot the contents of the VCF file
- Filter VCF variant records with a variety of conditions
- Calculate the statistics for each sliding window

After importing the VCF file, we can select some common parameters in the information panel, as shown below.

1. Select the individuals for analysis
   After importing the VCF file, the data area will display the individual information contained in
the file. Users can click on the checkbox in the table (middle top) to select the individuals you are interested in. The button can perform an inverse selection on the currently checked status.

If user wants to select multiple individuals at once, you can input the text file containing the individual names by clicking the Open button under the Setting individuals panel. Note that the suffix name of the file needs to be txt, here is an example file “pop1_SampleIDs.txt”.

| HG00096  |
| HG00097  |
| HG00099  |
| HG00100  |
| HG00101  |
| HG00102  |

2. Set target area for analysis

Users can control the target area where they are interested to analyze by setting options in included and excluded chromosomal regions. eGPS supports the BED file format. Please see the details at https://genome.ucsc.edu/FAQ/FAQformat.html#format1.
VCF snapshot

Extensible panels and parameters on the left side.

File progress bar:
1) To indicate approximate location of the entire file for the current displayed data, they are expressed as a percentage value. 0% means locating in file first place, 100% means going to end.
2) Drag the progress bar directly, the data displaying on the right table will automatically change. For example, dragging to the 20% table will display the data text corresponding to approximately 20% position of the file.
3) You can also change the value by clicking the arrow to the left of % for precise adjustment. However, clicking the arrow does not immediately refresh the data in the table on the right. You need to manually click Show Data again to update the data.

Begin pos and End pos:
1) They are used to set the specific location of the first record of the currently displayed data and the specific location of the last record of the currently displays the data. The input format is chromosome number, ‘:’ and position. The chromosome name should be consistent with VCF.
2) When importing the bgzip compressed VCF file, Begin pos and End pos become editable. Other type of VCF (uncompressed VCF and VCF) could not be used at this option. Users can modify the chromosome number and position. After clicking Show Data button, the table on the right will display the corresponding data (The begin position cannot be greater than or equal to the end position).

Num of lines shown:
Set the total number of rows of displayed data. If the dragged file is a bgzip compressed VCF file, and Begin pos and End pos not null. This item will be invalid.

Columns shown:

1) VCF field check box
   The check box name corresponding to VCF field. If this box is checked, the column in the right table, else the column will be hide.
2) All selected samples:
   If this option is checked, all individuals’ information will be displayed and MAX Samples Shown option will be invalid.
3) MAX samples shown:
   It is used to set the maximum number of individuals in the display data. If the number of individual in the file is less than the total number, the number of individuals in the file shall prevail.
4) Position format:
   It is used to set whether the display data uses “,” ” to divide the number. This item only takes effect for the POS column.

Supplemental:

When there are several VCF files in the data panel, this module will only read the first file. And when you set Begin pos and End pos manually, file progress bar will not be updated.

The Begin pos and End pos option is an interval, variants records fitting in this interval will be displayed.

Place the mouse on the head of each column, users can adjust the width of each column.
In addition to the first nine columns, the other Samples columns are bound to each other. For example, when the width of the first Samples column is adjusted, and the widths of all subsequent Samples columns will also change.

When the imported file is a bgzip compressed file, eGPS will automatically create its corresponding tbi index file. During this process, the snapshot button is disabled. You need to wait for the tbi index file to be accomplished. Process shown in the following figure.
Users can select or unselect individual samples in the Sample Information panel of the my data module. For example, if you uncheck “HG0096 HG0097 HG0099” three sample IDs, snapshot will filter the column on the right when display the data.

When you successfully open the snapshot, then you delete the file referenced by the snapshot in the eGPS data panel, the Show Data button and File progress bar will be disable until new VCF file be imported.

When you successfully open the snapshot, then you delete the file referenced by the snapshot in the eGPS data panel, and re-add another VCF file. After you switch back to the snapshot, the snapshot will automatically refresh the table and the data displayed at this time is the latter file.
VCF tools parameters

Position filter

Chromosome <string> Identifiers for CHROM. (e.g. 1;2;3)
Include sites with identifiers matching CHROM. More than one value separated by semicolon can be input to include multiple chromosomes.

From-bp <integer> Position. (e.g. 12345)
Specify a lower bound for a range of sites to be processed. Sites with positions less than this value will be excluded. This option can only be used in conjunction with one chromosome. This option can be used with or without to-bp.

To-bp <integer> Position. (e.g. 12345)
Specify an upper bound for a range of sites to be processed. Sites with positions greater than this value will be excluded. This option can only be used in conjunction with one chromosome. This option can be used with or without from-bp.

Not-chromosome < string > Identifiers for CHROM. (e.g. 1;2;3)
Exclude sites with identifiers matching CHROM. More than one value separated by semicolon can be input to include multiple chromosomes.

Positions <filename>
Include a set of sites on the basis of a list of positions in a file. Each line of the input file should contain a (tab-separated) chromosome and position. Lines that start with a "#" are command lines and will be ignored.

Exclude positions <filename>
Exclude a set of sites on the basis of a list of positions in a file. Each line of the input file should contain a (tab-separated) chromosome and position. Lines that start with a "#" are command lines and will be ignored.

Interval <integer> Minimum distance of two sites. (e.g. 50) bp, kb, mb
Make sure that no two sites are within the specified distance from one another.

Quality filter

Min quality score <float> Quality score. (e.g. 20.0)
Include only sites with Quality value above this threshold.
Variant type filter

Keep only SNPS
Include sites that contain a SNP.

Remove SNPs
Exclude sites that contain a SNP.

Keep only indels
Include sites that contain an indel.

Remove indels
Exclude sites that contain an indel.

Keep structural variants <string> Key words for ALT. (e.g. DEL;DUP;DUP:TANDEM;CNV)
Include sites that contain a special structural variant. More than one value separated by semicolon can be input to include several kinds of structural variants.

Remove structural variants <string> Key words for ALT. (e.g. DEL;DUP;DUP:TANDEM;CNV)
Exclude sites that contain special structural variants. More than one value separated by semicolon can be input to include several kinds of structural variants.

Site ID filter

Site ID<string> Site IDs. (e.g. rs6054257; rs6040355)
Include site(s) with matching ID. More than one value separated by semicolon can be input to include multiple SNPs.

Sites <filename>
Include a list of sites given in a file. The file should contain a list of site IDs, with one ID per line. Lines that start with a "#" are command lines and will be ignored.

Exclude sites <filename>
Exclude a list of SNPs given in a file. The file should contain a list of SNP IDs, with one ID per line. Lines that start with a "#" are command lines and will be ignored.

Filter flag filter

Remove filtered all
Remove all sites with a FILTER flag other than PASS.

Keep filters <string> Filter flag. (e.g. q10;q20)
Include all sites marked with a specific FILTER flag. More than one value separated by semicolon can be input to specify multiple FILTER flags.

**Remove filters** <string> Filter flag. (e.g. q10;q20)
Exclude all sites marked with a specific FILTER flag. More than one value separated by semicolon can be input to specify multiple FILTER flags.

**INFO field filter**

**Keep INFO** <string> INFO flag. (e.g. AA;AC)
Include all sites with a specific INFO flag. This option only filters on the presence of the flag and not its value. More than one value separated by semicolon can be input to specify multiple INFO flags.

**Remove INFO** <string> INFO flag. (e.g. AA;AC)
Exclude all sites with a specific INFO flag. This option only filters on the presence of the flag and not its value. More than one value separated by semicolon can be input to specify multiple INFO flags.

**Allele filter**

**MAF** <float> Lower bound for MAF. (e.g. 0.2)
Include only sites with a Minor Allele Frequency greater than or equal this value. This option can be used with or without **Max MAF**.

**Max MAF** <float> Upper bound for MAF. (e.g. 0.8)
Exclude only sites with a Minor Allele Frequency less than or equal this value. This option can be used with or without MAF.

**Min alleles** <integer> Lower bound for alleles. (e.g. 2)
Include only sites with a number of alleles greater than or equal to this value. This option can be used with or without **Max alleles**.

**Max alleles** <integer> Upper bound for alleles. (e.g. 3)
Include only sites with a number of alleles less than or equal to this value. This option can be used with or without **Min alleles**.

**Genotype Value Filter**
Min meanDP <float> Lower bound for mean DP. (e.g. 50.0)
Include only sites with mean depth values (over all included individuals) greater than or equal to this value. This option requires that the "DP" FORMAT tag is included for each site. This option can be used with or without Min meanDP.

Max meanDP <float> Upper bound for mean DP. (e.g. 50.0)
Include only sites with mean depth values (over all included individuals) less than or equal to this value. This option requires that the "DP" FORMAT tag is included for each site. This option can be used with or without Max meanDP.

HWE <float>
Assess sites for Hardy-Weinberg Equilibrium using a Chi-square test or Fisher’s exact test (choose one method when programming). Sites with a p-value below the threshold defined by this option are taken to be out of HWE, and therefore excluded.

Max missing <float> Upper bound for missing data. (e.g. 0.8)
Exclude sites on the basis of the proportion of missing data (defined to be between 0 and 1, where 0 allows sites that are completely missing and 1 indicates no missing data allowed).

Phased
Exclude all sites that contain unphased genotypes.

VCF analysis

When data existed in the data panel, we can turn to each analysis module and the analysis button is activated. There are four common parameters for each data analysis module.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Window size</td>
<td>The length of chromosomal regions (i.e. windows) analyzed at one time. The window size usually ranges from a few hundreds to thousands base pairs.</td>
</tr>
<tr>
<td>Window step</td>
<td>The number of base pairs that window slides along chromosome. If window step is less than window size, the windows will be overlapping and more detailed information will be given.</td>
</tr>
<tr>
<td>Minimum quality score</td>
<td>Include only sites with Quality score above this</td>
</tr>
</tbody>
</table>
threshold. Quality score
= \text{-10} \times \log(1-p), \text{ where } p \text{ is the probability of variant being present given the read data. For Example: } 99\% \text{ confidence (1\% error rate)} = \text{Quality score 20}; 99.9\% \text{ confidence (0.1\% error rate)} = \text{Quality score 30.}

| Output with P-value | If you set this parameter, the output file will contain a column whose value is p value. The p value calculate equation is its value’s rank divide total size. |

In addition, when calculating statistics, there is an option need to consider. Click \texttt{Options | Preferences | Insdel & missing data.}

The user can set the processing mode of indel: 
\texttt{ins/dels ignored} means that only point mutations are considered; \texttt{ins/dels treated as fifth nucleotide} means that indel is regarded as the fifth nucleotide. If you choose later, the validate length of slide window will minus REF allele type’s length.
Meanwhile, user can also set the treatment of genotype deletions, users can set tolerance for missing data.

\textbf{Genetic diversity}

There are two checkboxes for the user to select the diversity statistics that need to be calculated (Watterson’s theta and Pi).

Watterson’s theta (1) is an estimator for describing the genetic diversity in a population. It is a measure of the population mutation rate (i.e. \(4Ne\mu\)) by counting the number of polymorphic sites, where \(Ne\) is the effective population size and \(\mu\) the mutation rate per site per generation.
Pi (2) is the average number of SNPs found in two pairwise comparisons of sequences in the sample, and divided by the length of considered region.

**Testing neutrality**

There are four checkboxes for the users to select the statistics for the neutrality test that needs to be calculated: Fu & Li’s D (3), Fu’s Fs (4), Fay & Wu’s H (5) and Tajima’s D (6).

In *Ancestral state* panel, we can set the way to infer allele ancestral state. By default, the eGPS will assign REF field in the VCF file as an ancestral state i.e. *Ref alleles as ancestral alleles*. *No ancestral/derived alleles inferred* means use folded site frequency spectrum, and Fu’s Fs statistic will be disable. *Ancestral alleles inferred from outgroup* means infer ancestral allele state from outgroup. In this case, users need to provide a MAF file with the same name as the VCF but a different suffix name. The MAF file should contain species in the VCF.

If *Ancestral alleles inferred from outgroup* be selected. *Outgroup information* panel would be usable, which enable researchers to set name of the reference genome and the outer genome (genome assembly).

**HKA score**

This method calculates the statistics of neutrality by incorporating another population from different species (7, 8). This module needs to provide an individual name file for different population. The user needs to put the individual name and end the file with a .txt suffix.

Information sites means that substitutions between populations or within-populations polymorphisms. If substitutions between populations (divergence) is 0, we also remove the window. If substitutions within-populations is 0, we set it as 0.1 to take advantage of its full information.

**PBS value**

This method calculates the statistics of neutrality by incorporating two different species groups of the same species (9). This module needs three files for three different populations. The user needs to put the individual names and end the file with a .txt suffix. The first population is the target one to infer positive selection.

SNP Density threshold means that minimum SNP frequencies where the sites are considered if they satisfy greater than $(1.0 - P_{00})$. $P_{00}$ is the frequency of monotypic sites in a genomic region.
FST Static

Calculate fixation index (10) between two populations. This module needs to provide an individual name file for different population. The user needs to put the individual name and end the file with a .txt suffix.

Transcriptomic

circRNA viewer

This module visualizes de novo identified circRNA genesis procedures (11-13). The walkthrough tutorial has a nearly complete explanation for this module. Here is a summary for the Filter parameters column on left side. The main operation here is to filter out the circRNA that does not meet the conditions required by the user. Each parameter can be found in the software tooltip.

<table>
<thead>
<tr>
<th>Filter parameter</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>min expression</td>
<td>The minimum expression (BSJ reads number) of circRNA. Default 10.</td>
</tr>
<tr>
<td>max expression</td>
<td>The maximum expression (BSJ reads number) of circRNA. Default 999999999.</td>
</tr>
<tr>
<td>top total expression</td>
<td>Only the top expression circRNA that contain X% of BSJ reads.</td>
</tr>
<tr>
<td>top expression</td>
<td>Only the expression top X% of circRNA.</td>
</tr>
<tr>
<td>max isoform</td>
<td>The maxium number of considering isoform, default 10. High value will make the quantification slower.</td>
</tr>
<tr>
<td>chr-region</td>
<td>Circular RNA that locate this region will be included. Default &quot;&quot;, no limitation.</td>
</tr>
</tbody>
</table>

RNA-seq data analysis

After importing data, researchers need to provide classification (condition statement file) file.
This file contains column names to classify data into two groups. Here is an example:

| GSM146300  |
| GSM146308  |
| GSM146310  |
| GSM146312  |
| GSM146314  |
| GSM146334  |

Then you can set the differential significant test method and fold change statistic calculation method. After you run the analysis procedure, in interactive volcano plot will be display.

**Protein omics**

All parameter statements of proteomic data analysis can be found in the module’s tooltip when you mouse hovering over “?” label. This analysis method mainly includes the following parameters. Please refer for [http://bioinfo.sibs.ac.cn/shaolab/MAP/index.php](http://bioinfo.sibs.ac.cn/shaolab/MAP/index.php) details.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Norm flag</td>
<td>/</td>
</tr>
<tr>
<td>Win size</td>
<td>Size of the sliding window to scan the ratio-intensity plot (similar to the M-A plot widely used for microarray data analysis). By default: 400 proteins.</td>
</tr>
<tr>
<td>Step size</td>
<td>Step size of the sliding window. By default: 100 proteins.</td>
</tr>
<tr>
<td>central</td>
<td>The fraction of proteins with the weakest intensity changes of each window used to model technical variations.</td>
</tr>
</tbody>
</table>
Part 3: Visualizing and Exploring Data

eGPS tree viewer

This is a highly interactive phylogenetic tree viewer. A rich tree decorating operations, such as tree layouts, zooming, rotating, clade annotating, bootstrap values displaying, branches and leaf labels decorating, tree shape modification, un-do/re-do and high quality outputs are implemented by the eGPS and most operations can be completed by simple mouse clicking and dragging actions. In addition, the tree edits can be saved for future session. eGPS has a user-friendly interface and it can easily handle/color/label a tree with a very large number of species/taxa.

Toolbar

The toolbar of Tree viewer is composed of five parts. The first part includes opening input file
(Because the input files eGPS are unified managed in Data Panel, this button will be disable when this tab is activated), saving and printing current tree. The second part contains cutting, copying, deleting and pasting internal node or leaf. The third part has two button for zoom in and zoom out the tree. The forth part includes the functions of un-do and re-do, which makes users easily to see the effect of decoration or revision. Additionally, the functionalities of first four parts are the same as corresponding menu items in menu bar. The role of fourth part is formatting branch and leaf labels. The font face and font size of selected leaf labels can be changed by selecting specific item. The color of selected branch and leaf labels can be changed by selecting specific color.

**Left-side operations navigating bar**

The left-side operations navigating bar contains the operations implemented by tree viewer, such as tree layouts, zooming and rotating, selecting and formatting of branches and leaf labels, clade annotating and bootstrap values displaying, etc.

**Tree layouts**

The eGPS incorporates a number of tree layouts, *i.e.*, rectangular, slanted, circular (internal and external) or radial cladograms and rectangular, circular (external) or radial phylograms. Each toggle button represents one layout and current selected button is highlighted. The buttons for making phylograms will be automatically enabled or disabled according to whether the information of branch lengths is available or not.

Different zooming in/out and rotating strategies are available for different layouts. Horizontal and Vertical Zooming In/Out are only available for rectangular or slanted layouts. Whole Zooming In/Out is available for all layouts. Additionally, the user can redo the edits using redo button.

For rectangular or slanted layouts, 90-degree rotating operations are available. For circular or radial layouts, arbitrary degree between 0 and 360 rotating operations are available. The user can input or drag the cursor to specific degree.

Furthermore, the user can drag the tree to any position with holding down the mouse, and this is a convenient adjustment after zooming or rotating.
Operations

After clicking to select the internal node (a light blue rectangle appears), the whole clade, only branch or only leaf labels can be selected by clicking ‘Select Clade’ button ( ), ‘Select Branches’ button ( ) or ‘Select Leaf Labels’ button ( ), respectively. The ‘Swap Leaf Labels’ button ( ) is used to display leaf labels in a reverse order and this operation is defined as one of tree shape modification operations.

The ‘Ladderize Up’ button ( ) is used to ladderize up the whole tree or specific clade. The ‘Ladderize Down’ button ( ) is used to ladderize down the whole tree or specific clade. To ladderize up/down the whole tree, the tree root should be selected first. And the internal node of specific clade should be selected to ladderize up/down the specific clade. Ladderizing up and down operations are defined as two of tree shape modification operations.

The ‘Search’ button ( ) opens the following dialog to search specific leaf labels. It is very useful when the number of species/taxa in the tree is large. The user can select leaf labels in pull-down menu (The ‘All leaf labels’ item in pull-down menu is used to select all leaf labels, which will result in selecting the whole labels), or they can type in the text box to search leaf labels. Multiple leaf labels can be selected automatically based on regular expression.

Display

After clicking to select the branch (two light blue small rectangles appear), the branch width and style of selected branch can be changed easily by clicking buttons showed in left picture. By default, clicking these buttons will change the width or style of all tree branches. Tree view provides three different branch styles, and the user can click the branch style button one or more times to change branch style successively.

After clicking to select (a light blue border appears) the leaf label, the font face and font size of selected leaf label can be changed by selecting specific item in toolbar. Additionally, the buttons showed in left picture can be used to increase/decrease the font size, make the selected leaf label bold or italicize the selected leaf label.
To unselect the specific leaf label, the user can click the blank area near the label. By default, these buttons will change the style of all leaf labels.

Multiple branches or leaf labels can be selected by holding down the Ctrl key while clicking them. The most convenient way to select the whole tree is using the searching dialog. Then the colors and other properties of selected branches or leaf labels can be modified simultaneously. By default, the color of branches or leaf labels of whole tree will be changed.

If the information of bootstrap values is available, a dialog will appear after clicking the ‘Show or Hide Bootstrap Values’ button ().

In numeric view, bootstrap values are showed as numbers. Thermometric view graphically displays the bootstrap value in a two colored thermometer, and the colors of thermometer can be changed easily. In three-interval view, bootstrap values will be divided into three groups according to their values. Each group is then represented as a colored circle. As expected, the colors of thermometer and the circles representing three groups can be specified by user. The ‘Only for right panel’ option is used to confirm whether to show bootstrap values only in right panel or not.

The ‘Turn on or off Leaf Labels’ button () is used to turn on or off the leaf labels of whole tree. If the total number of leaves is larger than one hundred, the leaf labels are turned off automatically.

The ‘Turn on or off Branch Length’ button () is used to turn on or off the branch length if the information of branch length is available.

The ‘Turn on or off Root’ button () is used to turn on or off the root edge for rectangular layout.

**Popup menu**

eGPS provides three popup menus, which are used to extend the functionalities or implement the functionalities in different ways.
Leaf popup menu

Clicking the leaf label to select it (a light blue border appears), and the popup menu for the current leaf will show after right clicking.

The leaf popup menu includes the following items.

- The ‘Rename…’ item provides a dialog to input new name of the current leaf.
- The ‘Recolor…’ item opens a dialog to select new color for the current leaf label. The utility of this item is equivalent to the color button in toolbar.
- The ‘Addicon…’ item provides a dialog to add a leaf icon. The leaf icon can be added to make the tree more illustrative. The ‘Only for right panel’ option is used to confirm whether to show leaf icon only in right panel or not (see 4.2). See 5.3d for an example.
- The ‘Swap with sibling’ item is used to swap the current leaf with its sibling leaf. This operation is defined as one of tree shape modification operations.
- The ‘Hide label’ item is used to turn off the current leaf label.
- The ‘Hide other labels’ item is used to turn off the labels of other leaves. ‘Show or hide leaf
labels’ button (_button) in left-side operations navigating bar is used to turn on or off all leaf labels.

✓ The ‘Cut’, ‘Copy’, ‘Paste’ and ‘Delete’ items have the same functionalities as the corresponding menu items in the menu bar or toolbar.

✓ The ‘New sibling as new leaf...’ item opens the following dialog to add sibling leaves. ‘Add sibling’ button is used to add a new sibling leaf. ‘Delete sibling’ button is used to delete a sibling leaf. At least one sibling leaf should be added. If the information of branch length is available, the ‘Branch length’ text field is not allowed to be null. This operation is defined as one of tree shape modification operations.

✓ The ‘New children as new leaf...’ item opens the following dialog to add children for current leaf. ‘Add child’ button is used to add a new child. ‘Delete child’ button is used to delete a child. At least two children should be added. If the information of branch length is available, the ‘Branch length’ text fields are not allowed to be null. This operation is defined as one of tree shape modification operations.

✓ The ‘Reroot’ item is to do a mid-point rooting using current selected leaf.

**Internal node popup menu**

Clicking the internal node to select it (a light blue rectangle appears), the popup menu for the current internal node will show after right clicking. The internal node is in the place of the intersections of branches. When the mouse cursor is hovering over, the additional information about the internal node is showing.

The internal node popup menu includes the following items.
The ‘Select clade’ item selects the branches and leaf labels of current clade.
The ‘Select all branches’ item selects the branches of current clade.
The ‘Select all leaf labels’ item selects the leaf labels of current clade. The above items make multiple selections much convenient and they have the same functionalities as corresponding buttons in the left-side operations navigating bar.
The ‘Show clade in another tab’ item creates a new tab in right-side displaying panel. And the new tab displays the clade as a tree.
The ‘Cut’, ‘Copy’, ‘Paste’ and ‘Delete’ items have the same functionalities as the corresponding menu items in the menu bar or toolbar.
The ‘Turn on/off leaf labels’ item is used to turn on or off the leaf labels of current clade.
The ‘Swap the outset leaves’ item is used to swap the first child with the last child of current internal node. This operation is defined as one of tree shape modification operations.
The ‘New sibling…’ item opens the following dialog to add sibling internal node for current internal node. The children number of sibling internal node can be modified easily by clicking ‘Add child’ or ‘Delete child' buttons. And the minimum number of children is two. If the information of branch length is available, the branch lengths of new added internal nodes and their children must be specified. This operation is defined as one of tree shape modification operations.
The ‘Reroot’ item is to do a mid-point rooting using current selected internal node.
The ‘Ladderize up’ item is used to ladderize up the current clade.
The ‘Ladderize down’ item is used to ladderize down the current clade.
The ‘Annotation’ item provides the dialog (a) to annotate internal node. For all layouts, ‘Internal node’ annotation (b) is used to annotate the internal node. Parameters, such as shape style, the color and font of annotation text can be specified to achieve desired plot. For rectangular and slanted layouts, it is easy to draw bars and annotation text on the side of a clade using ‘Rightside’ annotation (c). Additionally, ‘Both’ of the two annotations can be specified (d) or deleted (a). The ‘Only for right panel’ option is used to confirm whether to show annotations only in right panel (see above) or not.

Leaf icon popup menu

If the leaf icon has been added, the leaf icon popup menu will appear after right clicking the icon.

The leaf icon popup menu includes two items.

The ‘Reset icon’ item opens a dialog to reset icon. The source, width and height of leaf icon can be reset easily.

The ‘Remove icon’ item removes the leaf icon.

Viewing alignment

The guidance course has given full information on how to use, please see the guidance course: View alignment.
Viewing evolutionary distance

When the software shows the pairwise evolutionary genetic distance, *Decimal places* can specify the number of display digits for the decimal point. *Upper right* and *Lower left* can display the position of the distance matrix, above or below the left.

**Viewing genomic location in genome browser**

Genomic region and candidate to genome browser

The omics analysis method can get a lot of significant genomic regions and candidate genes. Take the advantage of eGPS, users can jump from these places to the genome browser. Clicking *Option | Preference | genome browser*, we can set parameters in the preference panel.

The *Favorite genome browser* option allows you to choose the genome browser. Now we support the UCSC and Ensembl browsers and their mirrors. At the same time, the user needs to select a query species which needs to be supported by the corresponding genome browser.
### Part 4: Evolutionary Analysis

#### Pipeline to reconstruct phylogenetic tree

The pipeline of constructing a phylogenetic tree by distance method is to perform multiple alignment on multiple sequences first, and then calculate the paired genetic distance matrix from the obtained multiple sequence alignment. Finally construct a phylogenetic tree through this genetic distance.

Now eGPS supports the construction of phylogenetic trees from multiple sequence alignments and genetic distances. All settings for building a phylogenetic tree are in Preferences, which you can set by clicking *Options | Preferences*.

If you input an aligned fasta file, you can click on the *Genetic distance* node to select the parameters of intersect. The following table is a description of each of the currently available parameters.

<table>
<thead>
<tr>
<th>Estimate variance</th>
<th>Use this to specify whether to compute Distances only or Distances and Standard Errors. Currently, we support to estimate variance with bootstrap. If you select the latter, then you also need to set Bootstrap replications (e.g. 200).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substitution model</td>
<td>eGPS can handle nucleotide sequence evolutionary distance, and also consider both transition and transversion. We offer several model/methods for user (14, 15).</td>
</tr>
<tr>
<td>Gap/missing data treatment</td>
<td>These are options for handling gaps or missing data. You may choose to remove all sites containing alignment gaps and missing information before the calculation begins (Complete-deletion option). Otherwise, you may choose to retain all such sites initially, excluding them as necessary in the pairwise distance estimation (Pairwise-deletion option), alternatively you may use Partial Deletion (Site coverage) as a percentage.</td>
</tr>
</tbody>
</table>
eGPS provides following models/methods to estimate genetic distances:
Number of differences, p distance (the ratio of difference to validate length), The ratio of transversion to difference, distance of Kimura 2-parameter model (16), distance of jukes-cantor model (17), distance of Tajima Nei model (18), distance of Tamura Nei (19), distance of Tamura 3-parameter model (20).

Then click **Tree build method** node to construct NJ tree (21), Swift NJ tree (22) or UPGMA tree.

If researchers input an evolutionary distance file (.dist file), then you can directly enter **Tree build method** node, setting suitable parameters and click **build tree** button in the **method area**.

Besides, researchers are also allowed to reconstruct the tree from whole genome multiple sequence alignment by importing an MAF format file with selected species.

File format specification
1. Alignment: a simple aligned fasta format.
2. Distance data: we design the file format compatible with the Phylip format.

Input distance data example: every element is separate with tab or space separator.

```plaintext
# some annotation
5
A 0.131646 0.095159 0.126667 0.118657
B 0.131646 0 0.077834 0.102978 0.086919
C 0.095159 0.077834 0 0.023185 0.021227
D 0.126667 0.102978 0.023185 0 0.006619
E 0.118657 0.086919 0.021227 0.006619 0
```

First validate line only contains an integer indicate the number of OUTs.
Following line start with an OUT name and elements of a square matrix. The square matrix is symmetric matrix with diagonal element is 0.
Note: OUT name should not contain any separator.

While the output data is all separate with tab.

**Candidate gene to evolutionary analysis**

Through clicking **Option | Preference**

We can set **Genetic distance** and **Tree build method** parameters so users can conveniently jump omics analysis to evolutionary analysis. Current eGPS has following approaches to jump:
<table>
<thead>
<tr>
<th>Approach</th>
<th>Annotation</th>
<th>Picture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein omic to phylogenetic</td>
<td>Click point of interest to build tree.</td>
<td></td>
</tr>
<tr>
<td>RNA-seq to phylogenetic</td>
<td>Click point of interest to build tree.</td>
<td></td>
</tr>
<tr>
<td>circRNA viewer to phylogenetic</td>
<td>Click rectangular of interest to build tree.</td>
<td></td>
</tr>
<tr>
<td>Genomic analysis results to phylogenetic</td>
<td>Click point of interest to build tree.</td>
<td></td>
</tr>
<tr>
<td>Gene to gene tree</td>
<td>Directly enter genomic location or gene name.</td>
<td></td>
</tr>
</tbody>
</table>

In order to build phylogenetic tree, the eGPS desktop obtain data from the eGPS cloud or Ensembl alignment REST (23).

Click **Connection** node to set the maximum time to waiting when connecting and transferring data from web server.

Click on **Species set** to set up the species set group and the query genome for the gene tree. Currently, we support data from eGPS cloud and Ensembl. If the user obtains the gene tree from gene symbol, **the regions to reconstruct** option can be set and user can choose to use the entire region to create trees or only use exons to create trees. eGPS uses Ensembl’s LOOKUP REST and specifies the query genome to query the gene information. We extract the first isoform's exon information for tree building.

All in all, eGPS offers graphic workflow to reconstruct, visualize and modify phylogenetic trees,
which makes it convenient to connect multi-step analyses.

**Align multiple sequences**

Almost the first step of all evolutionary data analysis (like phylogenetic tree reconstruction, computing evolutionary distances or inferring population history) is to construct the multiple sequence alignment. Here we packaged two well-known multiple sequence aligners in eGPS with GUI to help researchers accomplish the initial step.

For detailed information about the aligner, please refer to:
- MAAF (https://mafft.cbrc.jp/alignment/software/)
- ClustalW (http://www.clustal.org/clustal2/)

**Evolutionary analysis for genomic regions of interest**

eGPS provides options to get specific genomic region when users enter gene symbol in gene to alignment/genetic distance/gene tree modules. It obtains the gene annotation information from the Ensembl REST service. The options including *Whole region of gene, Exons, Introns, Coding sequence and 4-fold degenerate sites*. For the later four options, it extracts sequence according to first transcript information.

Here we demonstrate this function by walkthrough an example:

4DTv rates (the rate of transversion at fourfold degenerate sites, a site is four-fold degenerate if all possible changes at the site are synonymous) stands for fourfold synonymous third-codon transversion can be calculated to assess the genetic distances between paralogous pairs. D4DTv ranges from 0 for recently duplicated peptides, to ~0.5 for paralogs with an ancient evolutionary past.

Following is the operating steps:

1. Setting parameters
   
   Click *Option | Preferences | Genetic distance* and *Option | Preferences | Species set*
   
   Set *considered gene regions* as *4-fold degenerate sites* in *Species set* and *Model/Method* as *The rate of transversion* in *Genetic distance*. 

---

67
2. Go to gene to genetic distance module

*Tools | Gene to Genetic distance*

Enter the gene symbol (example “INS”) in *Gene symbol* input text field. Click *Get genetic distances* button.

Note: The extracted genomic region is based on the sequence of the reference species, so other sequences may have the following problems. Please pay attention to whether the non-reference gene sequence meets the conditions of data analysis.

1. When the user selects the acquisition region as the coding sequence, the sequence start characters obtained from Ensembl are ATG, and the termination triplet sequence also corresponds to the stop codon. However, the non-reference sequence needs to be noted whether the starting string is ATG and whether the ending string is a stop codon. For the sequence obtained from the eGPS
Cloud, if the coding gene is in the negative chain, it will be converted into a positive chain; therefore, the starting character of the reference gene acquired by eGPS is not ATG.

2. When the user selects 4-fold degenerate sites, it is also necessary to consider whether the important position of the reference sequence or the non-reference sequence is broken by the gap.

**Genomic evolutionary analysis: construct species tree for whole genome sequence alignment**

Multiple sequence alignment (MSA) is generally the initial material for constructing the phylogenetic tree. It is generated by three or more sequences and will be supplemented by a number of Gaps (generally denoted by the symbol '-') so that all sequences reach the same length. The file format used to store it is generally *fasta, ggc (msf), clustalW, paml, mega, phylip*, and so on.

The MSA usually stores information as a series of combinations of lengths close to genes (typically 1k ~ 1000k). So is there a genome-wide alignment for multiple genomes? What is its format? Multiple alignment between entire genomes refers to the alignment generated by multiple genome (DNA level) sequences. The most widely used format is MAF format (multiple alignment format, see [MAF file format in UCSC, https://genome.ucsc.edu/FAQ/FAQformat.html#format5](https://genome.ucsc.edu/FAQ/FAQformat.html#format5)). In addition to MAF, there is an EMF format, the former is mainly produced by UCSC, the latter is Ensembl. You can use the converter plugin module converts EMF to MAF.

The MAF consists of a header and a block, each of which is equivalent to an MSA. The difference between them is that MSA is usually a combination of a proteins or some short-segment DNA, and there is generally no complicated situation such as genomic rearrangements and opposite positive and negative strands.

eGPS accept MAF([https://genome.ucsc.edu/FAQ/FAQformat.html#format5](https://genome.ucsc.edu/FAQ/FAQformat.html#format5)) file to reconstruct the phylogenetic tree to infer species tree. The tree build method is distance-based method. So researcheres just need to import maf file(s) into eGPS, choose the species of interest, set *Genetic distance* and *tree build method*. And then click the *Build tree* button.
1. For first time reconstruction, eGPS will run through the whole VCF to generate a configuration file which records the maximum number of species assembly names. So it may take a long time if VCF file is very large. Users are allowed to produce a configuration file like following text and make sure the suffix of file is “.config”, the file name is same as VCF file name.

```plaintext
##eGPS multiple alignment file configure file format(MAFconfig)
version=1.0 totalSize=81
# list of genomes
pteAle1
ochPri3
echTel2
cavPor3
cheMyd1
anaPla1
chiLan1
```

Lines start with “#” will be treat as annotation.

2. EGPS will collect needed information block by block, if a block does not contain sequence of interest, the block will be discard. Therefore, if no block contains all the sequences selected by the user, the tree will not succeed because information is not enough to build a tree.

**Part 5: Population History Modelling**

The eGPS provides a very user-friendly tool called “simulator” to build demographic models by clicking mouse and dragging model icons and simulate the evolutionary events with animation of monitor parameters. These operations overcome the weakness of the command line run mode that
researchers need to carefully supervise the parameters before simulation. Now, the eGPS has implemented the coalescent-based backward simulation for which researchers can set the specific mutation rate.

**Build Model**

The eGPS implemented a highly interactive, full of tooltips population history modeling module. In the modeling stage, researchers can choose suitable icons to build complex population history models. In this modeling procedure, researchers will be guided with wealthy and friendly tooltips. We categorize various parameters into different groups by their functions.

The horizontal axis represents time and 0 represents now. Users can drag and drop the appropriate population history icons to build a model of interest. Through different combinations of modules, users can obtain any single population model and multiple population models. By constructing leaf populations that start at different times, the user can simulate the DNA sequence of the ancestral sample. The population history models from top to bottom in the figure are:

<table>
<thead>
<tr>
<th>Icon</th>
<th>Model</th>
</tr>
</thead>
</table>

71
<table>
<thead>
<tr>
<th>Model</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential growth model.</td>
<td></td>
</tr>
<tr>
<td>Linear growth model.</td>
<td></td>
</tr>
<tr>
<td>Logistic growth model.</td>
<td></td>
</tr>
<tr>
<td>Instantaneous growth model.</td>
<td></td>
</tr>
<tr>
<td>Exponential decline model.</td>
<td></td>
</tr>
<tr>
<td>Linear decline model.</td>
<td></td>
</tr>
<tr>
<td>Instantaneous decline model.</td>
<td></td>
</tr>
<tr>
<td>Constant size model.</td>
<td></td>
</tr>
</tbody>
</table>

Note: The differences between the three models have only a visual effect.

**Simulate**

After building the group history model, you can turn to this simulation section which has four parameter setting fields. In this stage, researchers can conveniently and intuitively set the parameters. When researcher clicks button to simulate, the distribution of monitor parameters will dynamically change.
The **Sample information** column shows the number of population and the population name at the current time. This name is built-in and cannot be changed. The description of other setting parameters is as follows:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference current N</td>
<td>Current population size for reference, you can choose either Set ‘Ref. Pop.’ or ‘Ref. cur. N’.</td>
</tr>
<tr>
<td>Set Ref. Pop.</td>
<td>Choose the reference population to determine ref current N</td>
</tr>
<tr>
<td>Ref. cur. N</td>
<td>To set current effective population size as reference, default 100000.</td>
</tr>
<tr>
<td>Mutation information</td>
<td>Mutation information for simulation, you can choose either ‘Fixed theta’ or ‘Fixed K’.</td>
</tr>
<tr>
<td>Fixed theta</td>
<td>$4 \times \text{refN0} \times \text{mutRate}$. Usually range between 0.1 and 50.</td>
</tr>
<tr>
<td>Fixed K</td>
<td>Fixed number of segregating sites (K)</td>
</tr>
<tr>
<td>Others</td>
<td></td>
</tr>
<tr>
<td>Simulated trees in output file</td>
<td>Whether the trees representing the history of the samples are also output</td>
</tr>
<tr>
<td>Num of iterations</td>
<td>The number of simulation iterations</td>
</tr>
</tbody>
</table>

Run & Save data as button means run the simulation procedure and save results into a file.  
Run only button means only run the simulation procedure.

Note: these buttons would be disabled if user does not establish the population history models.

After finishing the simulation procedure, the eGPS will give summary statistics and offer the option for saving the results with a text file. Furthermore, the eGPS provides a beneficial way to compare different results by organizing the draggable sub tabs.

The output file format is the same as Hudson’s *ms* program here comes with an example file.

The example of output file:

```plaintext
TotalSampleSize 4 (Pop2 4) ; theta 0.5; ref current N 50,000
//
(Pop2_3:1.0560117,(Pop2_4:0.1998489,(Pop2_1:0.0788691,Pop2_2:0.0788691):0.1209798):0.8561628):0;
segsites: 1
positions: 0.8245964
```
The first line of the output is the parameter information. Each iteration is preceded by a line with just “//” on it. If users checked the *Simulated trees in output file* option. The tree of the simulated sample will be output.

The numbers of segregating (or polymorphic) sites are given in lines starting with “segsites”. The positions of SNPs are given in a scale of (0,1). The positions are randomly and independently assigned from a uniform distribution. Simulated haplotypes are given in the following lines. The ancestral state is coded with a “0”, and the mutant, or derived state, indicated with a “1”.

<table>
<thead>
<tr>
<th>0</th>
<th>0</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
</table>

//
(Pop2_1:1.473843,(Pop2_4:0.3295514,Pop2_2:0.3295514):0.3569735,Pop2_3:0.6865249):0.7873181):0;

segsites: 5

positions: 0.0929961 0.335626 0.6107916 0.7173615 0.7985285

10100

00011

01011

00011
Part 6: Plugin development

Third-party developers can develop third-party applications by means of the eGPS published application programming interface, and the application will be displayed under the Plugin menu in eGPS main frame. The information of the third-party developer and its website will be dynamically displayed in the status bar at bottom left corner when the Plugin module is activated.

We fixed the path and class name of the plugin entrance class, we named it egps.plugin.unified.Main.java

The entrance class needs to inherit the published abstract class whose name is AbstractSwingPlugin. And developers need to rewrite the following three methods.

<table>
<thead>
<tr>
<th>Method signatures</th>
<th>Annotation</th>
<th>Example</th>
</tr>
</thead>
</table>
| String getTabName() | Set the name of the functional module (also is the tab name) | @Override<br>public String getTabName() {<br>    return "eGPS plugin demo";
}
| JComponent getViewPanel() | Specific content on the graphical interface | --- |
| String[] getTeamAndAuthors() | Set developer information, format is “organization”, “developers” and “website” | @Override<br>public String[]<br>getTeamAndAuthors() {
    String[] newString = {"University|Institute","Jack,Tom","www.abcd.com"};
}

If you prefer to write a graphical interface in Javafx, you can use the getViewPanel() method to return a JFXPanel instead of JComponent.

If you want to set the icon for the module button under the Plugin menu bar, you can put any size PNG image file in the same directory and the name of Jar file and image file requiring the same name.
Part 7: update records

V1.01

New features:

1. Increase Alignment View export format, now we support ClustW/ GCG(MSF)/ MEGA/PAML/ PHYLIP.
2. Add new module geneToAlignment, reconstruct the codes of geneToAlignment/ geneToDistance/ geneToGenetree.
4. Add Log.
5. Add new module multiple sequence aligner which package two programs : ClustalW and Mafft.
6. Add new options when obtaining alignment, calculate Genetic distance and infer phylogenetic tree from gene symbol: 1. whole gene body length 2. exons 3. introns 4. coding sequence 4. 4-fold degenerate sites
7. Add new options for Genetic distance: The transversion rate. When obtaining alignment of 4-fold degenerate sites, 4DTv rates (the rate of transversion at fourfold degenerate sites) could be computed.

Fixed bug:
1. considered gene regions not wok for exons.
2. For swingx jar library: JXTaskPanel in MAC OS is grey.
3. Alignemnt View display problem.
4. These are two dialog when files in Recent file have been deleted.
5. When save GeneToGeneticDist file in distance viewer, the number of OUT is not correct.

V1.0.2

New features:
1. Genetic distance viewer module adds new display mode of heat map for enhanced visualization, and support for exporting vector graphics.
2. In alignment viewer module, exporting NEXUS format file is supported.
3. For alignment viewer module, loading large file in short time is permitted.
4. In the evolution analysis, add new features to jump between modules, for example: from the Alignment viewer to the Distance viewer and the Tree viewer.
5. Provides green, free installation version for MAC OS.

Fixed bug:
1. Fix the Neuxs file format parse problem, now supports tree and data block.
2. In the history root node, when you click the right mouse button to mark, a null pointer error occurs.
3. When dragging an error file, two dialogs will pop up twice.
4. When the MSA is exported to, for example, the Phylip file format, since the officially defined format is that the sequence noun cannot exceed 12 characters, it is now changed to a loose format.
5. Some data is not suitable for DEG-MassSpecturm module to analyze, add a dialog to tell users.

Appendix: Frequently Asked Questions

1. Is eGPS open source? Is it charged?
The eGPS provides Plugin development support. Plugin developers own full credits for their plugins. The eGPS is free of use.
2. When I installed eGPS Desktop, there was a prompting from an anti-virus software: "The software you downloaded contains a virus". What is going on?
We recommend shutting down all anti-virus software before installing eGPS Desktop. eGPS Desktop is a security software that does not contain any malicious programs that infringe your computer.

Glossary

Multiple sequence alignment (MSA)
https://en.wikipedia.org/wiki/Multiple_sequence_alignment
A multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences, generally protein, DNA, or RNA.
Aligned fasta file

Reference

