KungFuPanda
User Guide

1. Overview

The KungFuPanda is a program for examination of accelerated evolution. The software was written in Java and suitable for different computer platform. It was designed for both of the high performance computing environment, such as computer cluster, and a personal computer. The software can analyze the accelerated evolution of an internal branch in a user-defined tree, based on a user-defined configure file.

2. Setup KungFuPanda Programs

The software can be freely downloaded from http://www.picb.ac.cn/evolgen/softwares/. Then please extract all the files to a folder.

3. Running KungFuPanda Programs

The running of KungFuPanda consists mainly of following steps.

1. Preparation of the multi-alignment files
2. Establish chainIndex files
3. Set up configure file
4. Running of KungFuPanda
5. Running of finalTest

3.1 The multi-alignment files

The multi-alignment files (in MAF format) are required for the program. Below is an example of the multi-alignment file.

> chr1.hg19.panTro3.ponAbe2.rheMac2.mm9.rn4.cavPor3.oryCun2.bosTau6.equCab2.canFam2.loxAfr3.monDom5.galGal3.taeGut1.anoCar2.synNet.axt.maf

chr1 is the file name of the multi-alignment file. It specifies the certain chromosome. hg19.panTro3.ponAbe2.rheMac2.mm9.rn4.cavPor3.oryCun2.bosTau6.equCab2.canFam2.loxAfr3.monDom5.galGal3.taeGut1.anoCar2.synNet.axt.maf is the suffix name of the file. It shows the species in the multi-alignment file. Quality scores can be added to the multi-alignment files when necessary. Please consult the related documents.

Unfortunately, KungFuPanda does not support the compressed MAF files in the current version. The compressed MAF files could be supported in future.

3.2 The chain index files

It is required to index the multi-alignment files. Users can use command establishChainIndex (which is included in the KungFuPanda software package) to generate chain index files.

Below is an example of the process.

```
establishChainIndex /data/multiAlign/
chr1.hg19.panTro3.ponAbe2.rheMac2.mm9.rn4.cavPor3.oryCun2.bosTau6.equCab2.canF
am2.loxAfr3.monDom5.galGal3.taeGut1.anoCar2.synNet.axt.maf
```

In the command line, the directories of the multi-alignment file and its name should be specified. And the output files (with the prefix of chainIndex_) are generated in the same directory. Below is an example of the index file.

```
chainIndex_chr1.hg19.panTro3.ponAbe2.rheMac2.mm9.rn4.cavPor3.oryCun2.bosTau6.eq
uCab2.canFam2.loxAfr3.monDom5.galGal3.taeGut1.anoCar2.synNet.axt.maf
```

Users can also get the help information of establishChainIndex by entering the command line with no parameters.

3.3 The configure file

The default control file for KungFuPanda is a text file. Below is a copy of the control file. Lines beginning with "#" are treated as comments.

```
#This configure file is set up for a genome-wide screening for placental-accelerated
sequences (PAS)
listOfAllChrs    chr1   249250621
listOfAllChrs    chr2   243199373
listOfAllChrs    chr3   198022430
listOfAllChrs    chr4   191154276
listOfAllChrs    chr5   180915260
listOfAllChrs    chr6   171115067
listOfAllChrs    chr7   159138663
listOfAllChrs    chr8   146364022
listOfAllChrs    chr9   141213431
listOfAllChrs    chr10  135534747
listOfAllChrs    chr11  135006516
```

```
listOfAllChrs     chr12 133851895
listOfAllChrs     chr13 115169878
listOfAllChrs     chr14 107349540
listOfAllChrs     chr15 102531392
listOfAllChrs     chr16 90354753
listOfAllChrs     chr17 81195210
listOfAllChrs     chr18 78077248
listOfAllChrs     chr19 59128983
listOfAllChrs     chr20 63025520
listOfAllChrs     chr21 48129895
listOfAllChrs     chr22 51304566

winSize              100
slidingSpace       20

minPercOfSpeciesWithNormalSites 0.5
minPercOfSeqLengthForKeySpecies 0.3
maxNumOfMissingSpecies 4
maxPercOfInsDelsAmongNonMissingSpecies 0.3

alignDir       /data/multiAlign/
alignSuffix
hg19.panTro3.ponAbe2.rheMac2.mm9.rn4.cavPor3.oryCun2.bosTau6.equCab2.canFam2.l
oxAfr3.monDom5.galGal3.taeGut1.anoCar2.synNet.axt.maf

workingDir     /result/

define human          hg19
define chimpanzee panTro3
define orangutan    ponAbe2
define rhesus        rheMac2
define mouse          mm9
define rat             rn4
define guineaPig    cavPor3
define rabbit        oryCun2
define cow             bosTau6
define horse         equCab2
define dog            canFam2
define elephant     loxAfr3
define opossum      monDom5
define chicken      galGal3
define zebraFinch taeGut1
define lizard        anoCar2
```

```
tree
(((((((human:6.4,chimpanzee:6.4):9.3,orangutan:15.7):13.9,rhesus:29.6):61.4,(((mouse:25.2,
rat:25.2):46.9,guineaPig:72.1):14.3,rabbit:86.4):4.6):6.4,(cow:84.6,(horse:82.5,dog:82.5):2.1):
12.8):7.3,elephant:104.7):71.4,opossum:176.1):148.4,((chicken:106.4,zebraFinch:106.4):16
8.5,lizard:274.9):49.6):0

node root
(human,chimpanzee,orangutan,rhesus,mouse,rat,guineaPig,rabbit,cow,horse,dog,elephant
,opossum,chicken,zebraFinch,lizard)
node lizardBirds                    (chicken,zebraFinch,lizard)
node birds                           (chicken,zebraFinch)
node opossumPlacentalMammals
(human,chimpanzee,orangutan,rhesus,mouse,rat,guineaPig,rabbit,cow,horse,dog,elephant
,opossum)
node placentalMammals
(human,chimpanzee,orangutan,rhesus,mouse,rat,guineaPig,rabbit,cow,horse,dog,elephant
)
node euaLau
(human,chimpanzee,orangutan,rhesus,mouse,rat,guineaPig,rabbit,cow,horse,dog)
node euarchontoglires
(human,chimpanzee,orangutan,rhesus,mouse,rat,guineaPig,rabbit)
node catarrhini                     (human,chimpanzee,orangutan,rhesus)
node hominidae                       (human,chimpanzee,orangutan)
node humChimp                         (human,chimpanzee)
node rodent                          (mouse,rat,guineaPig,rabbit)
node sciurognathi                   (mouse,rat,guineaPig)
node murinae                         (mouse,rat)
node laurasiatheria                 (cow,horse,dog)
node perissodactyla                  (horse,dog)

fastEvolBranches opossumPlacentalMammals-placentalMammals

forNeutralEvolRate branches root-lizardBirds root-opossumPlacentalMammals
forNeutralEvolRate branches lizardBirds-lizard
forNeutralEvolRate branches lizardBirds-birds
forNeutralEvolRate branches birds-zebraFinch
forNeutralEvolRate branches birds-chicken
forNeutralEvolRate branches opossumPlacentalMammals-opossum
forNeutralEvolRate branches placentalMammals-elephant placentalMammals-euaLau
forNeutralEvolRate branches euaLau-euarchontoglires euaLau-laurasiatheria
forNeutralEvolRate branches euarchontoglires-catarrhini euarchontoglires-rodent
forNeutralEvolRate branches catarrhini-rhesus catarrhini-hominidae
forNeutralEvolRate clusters hominidae
forNeutralEvolRate branches rodent-rabbit rodent-sciurognathi
```

```
forNeutralEvolRate branches sciurognathi-guineaPig
forNeutralEvolRate branches sciurognathi-murinae
forNeutralEvolRate branches murinae-mouse murinae-rat
forNeutralEvolRate branches laurasiatheria-cow laurasiatheria-perissodactyla
forNeutralEvolRate branches perissodactyla-horse
forNeutralEvolRate branches perissodactyla-dog
```

The control variables are described below.

listOfAllChrs specifies the size of the reference genome. If you use hg19 as the reference genome, you don't need to set the listOfAllChrs. It's the default list.

winSize and **slidingSpace** specifies the size of sliding window and its sliding step.

minPercOfSpeciesWithNormalSites: As in this example, KungFuPanda only analyzes sites, at least half of which species are not insdel and lowQS. If the percentage of normal DNA characters at a site is smaller than the setting value (exclusive), this site will be discarded.

minPercOfSeqLengthForKeySpecies: The internal branch is determined or confined by three key clusters. Within each cluster, at least one species has to have a minimum length (exclusive) of minPercOfSeqLengthForKeySpecies (compared to winSize). If one of clusters does not meet the requirement (seqLength in all the species within the cluster is smaller than the setting value), this window will be discarded.

maxNumOfMissingSpecies: If the number of missing species (*i.e.* sequence is completely missing for the species) is larger than the setting value (exclusive), the window will be discarded. It is noted that different analysis uses different genomes, and different genomes have different quality. So this number may be highly variable among different analysis.

maxPercOfInsDelsAmongNonMissingSpecies: After excluding missing species, the window will be ignored if the percentage of remaining insdels among all the remaining characters is larger than the setting value (exclusive).

alignDir and **alignSuffix** specify the location of the alignment files (including chainIndex files) and their suffix name.

workingDir specifies the location of the output files.

define: The common name of species and their genome assemble name can be associated, and thus the common name could be used below.

tree: Phylogenetic trees (rooted) is used here in the Newick format.

node defines each node in the tree. It has to be defined by leaves.

fastEvolBranches specifies the fast evolved branch that will be tested. It has to be an internal branch.

forNeutralEvolRate specifies the branch sets (see the method section of the paper). One branch could be used as a branch set.

3.4 Running of KungFuPanda

Below is an example how to run KungFuPanda.

```
./kungFuPanda -config /data/config.txt -sigLevel 0.01 -useMaxNeutralRate yes
-recordMutRate true -chrs chr1 > /result/chr1.txt
```

The parameters are described below.

-config: after this parameter, users should specify the exact location of the configure file. Detailed explanation of the configure file is shown in 3.3.

-sigLevel: Here it should be specified the significant level of the result. If the significant level is higher than the setting value, the result will not be outputted.

-useMaxNeutralRate [yes|no]: This parameter shows whether to use the maximun neutral rate. If not, the average neutral rate will be used.

-recordMutRate [true|false]: This parameter shows whether to record the mutation rate.

-chrs: This parameter specifies which chromosome will be analyzed.

The result of KungFuPanda program will be printed out directly on the screen. It is recommended that users output the result to a file on hard-disk.

2 sets of output file will be generated after the running process. They are the results of KungFuPanda (file name is specified by users) and the mutation rate files (e.g. mutRate_chr1_0.txt), which are required for finalTest.

Users can get more help information of kungFuPanda by entering the command line with no parameters.

3.5 Running of finalTest

Given weighted substitution rate, finalTest program can calculate the alpha value for each branch for normalization of KungFuPanda results.

```
./finalTest -dir /data/result/ -config config.txt -sigLevel 0.01 -useMaxNeutralRate yes >
/data/result/results.txt
```

The parameters are described below.

-dir : Users should put the directory of mutation rate files (which are generated through KungFuPanda program) under this parameter.

-config : after this parameter, it should be specified the exact location of the configure file. Detailed explanation of the configure file is shown in 3.3.

-sigLevel : Here is the significant level of the result. If the significant level is higher than the setting value, the result will not be outputted.

-useMaxNeutralRate [yes|no]: This parameter shows whether to use the maximun neutral rate. If not, the average neutral rate will be used.

The result of finalTest program will be output directly to screen. It is recommended that users re-direct the result to a file on hard-disk.

More information of finalTest can be attained by entering the command line with no parameters.

3.5 How to view alignments for interested regions

Use this command ./alignmentUtil for the further help information. It could output the alignment for the user defined region.

4. FAQ

Q1: Class cannot be found!
A1: The program should be run at the directory where the program is. It can run in other directory when the absolute classpath is provided. It should be easy to revise the command batch files, for example, kungFuPanda. Please consult your IT department or Java documents or your friends on how to do it.

Q2: OpenJava *vs.* Java.
A2: Please make sure you are using Java, instead of OpenJava. You can find the information of Java version when you run the program. Search for something like open

java or OpenJava. Please avoid using OpenJava. Visit www.java.com for downloading the correct java runtime environment.

Q3: I only have compressed MAF alignment files.
A3: You need to un-compress these files first. The compressed MAF alignment files may be supported in the future.

Q4: How to run the program on Windows?
A4: Please make the window batch files for yourself by following the enclosed Linux batch files.

5. How to cite

Yuting Wang, Guangyi Dai, Zhili Gu, Guopeng Liu, Ke Tang, Yi-Hsuan Pan, Yujie Chen, Haoshan Chen, Su Feng, Shou Qiu, Hongduo Sun, Xin Lin, Lili Zhang, Qian Li, Chuan Xu, Yanan Mao, Yong E. Zhang, Philipp Khaitovich, Yan-Ling Wang, Qunxiu Liu, Jing-Dong J. Han, Zhen Shao, Gang Wei, Chun Xu, Naihe Jing, Haipeng Li (2019) **Modulation of social hierarchy by a placental-accelerated enhancer of *Lhx2*** (to be submitted).