Manual for AncSid

Version 1.0

1. Introduction

The main function of this AncSid is to identify species of ancient animals based on next generation sequencing data. There are several key parameters that will influence the identification accuracy including the sequencing similarity, query coverage, and the screening scope of deamination induced mutations. We strongly recommended to use a similarity with ≥98% but <100% and query coverage >96% for running this script. The deamination based screening can be turned off, but this function will significantly improve the ability of species identification, especially for highly damaged DNAs.

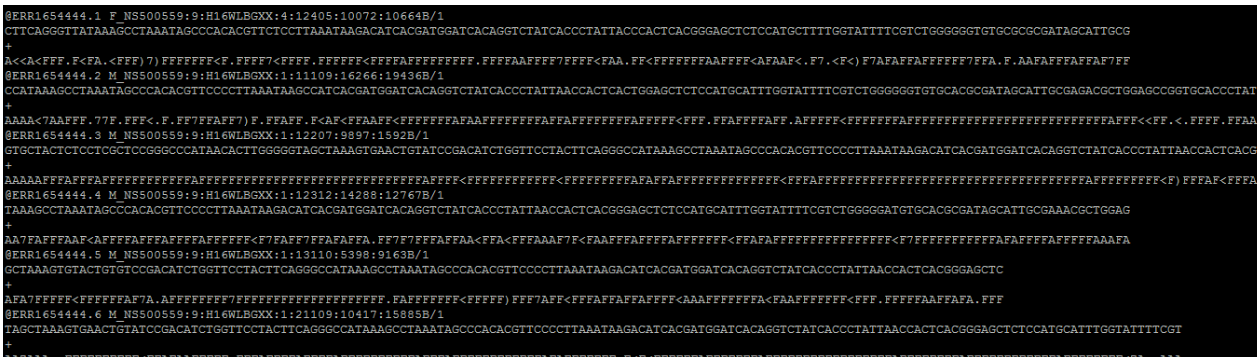
2. Parameters

Parameters	Parameter Type	Description
-i	string	Fastq file after adapter filtration
-name	string	Sample name that used for identify different samples
-formatdb	string	Absolute path of Formatdb. If you have formatdbed database,then ignore this
-blastall	string	Blastall Absolute path
-Rpath	string	R Absolute path
-database	string	MtDNA database
-coveragemin	num	The minimal query coverage
-coveragemax	num	The maximal query coverage
-similaritymin	num	The minimal sequence similarity
-similaritymax	num	The maximal sequence similarity
-mutationsite	num	The first and last N bases that used for screening reads with the C->T and G->A changes
-ct_mutationnum	num	More than N C-to-T changes within the -mutationsite <num>. This should be used with the parameter -mutationsite
-ga_mutationnum	num	More than N G-to-A changes within the -mutationsite <num>. This should be used with the parameter -mutationsite
-delete	string	Deleting the temp files, yes or no
-o	string	The output directory
-h	string	Show the help message

3. Example:

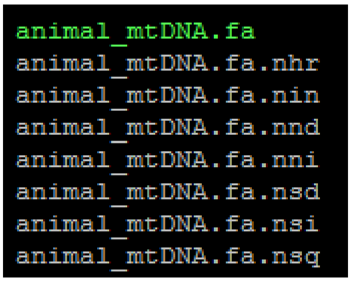
```
perl pipeline.pl -i path/test.fq.gz -name sample1 -formatdb path/formatdb -blastall path/blastall -Rpath path/R -database path/animal_mtDNA.fa -coveragemin 0.1 -coveragemax 1.00 -similaritymin 10 -similaritymax 100 -mutationsite 10 -ct_mutationnum 1 -ga_mutationnum 1 -delete yes -o path/
```

(1) The format of test.fq.gz: This is a normal Fastq file with a format as following:



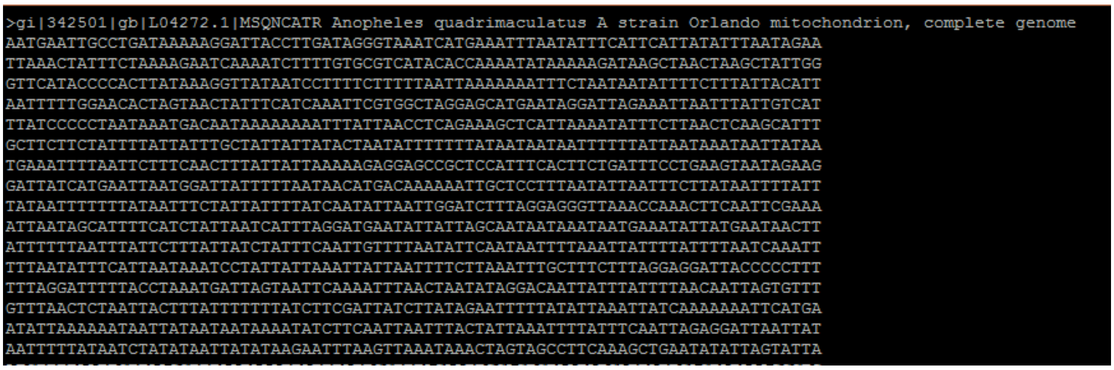
```
@ERR1654444.1 F_NS500559:9:H16WLBGXX:4:12405:10072:10664B/1
CTTCAGGGTTATAAAGCCTAAATAGCCCACACGTTCTCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTACCCACTCACGGGAGCTCTCCATGCTTTTGGTATTTTCGTCTGGGGGGTGTGCGCGCGATAGCATTGCG
+
A<<A<FFF.F<FA.<FFF)7)FFFFFFF<F.FFFF7<FFFF.FFFFFF<FFFFAFFFFFFFFFF.FFFFAAFFFF7FFFF<FAA.FF<FFFFFFFFAAFFFF<AFAAF<.F7.<F<)F7AFAFFAFFFFFF7FFA.F.AAFAFFFAFFAF7FF
@ERR1654444.2 M_NS500559:9:H16WLBGXX:1:11109:16266:19436B/1
CCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGCCATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACTGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGTGCACCCTAT
+
AAAA<7AAFFF.77F.FFF<.F.FF7FFAFF7)F.FFAFF.F<AF<FFAAFF<FFFFFFFAFAAFFFFFFFFAFFAFFFFFFFFFAFFFFF<FFF.FFAFFFFAFF.AFFFFF<FFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFF<<FF.<.FFFF.FFAA
@ERR1654444.3 M_NS500559:9:H16WLBGXX:1:12207:9897:1592B/1
GTGCTACTCTCCTCGCTCCGGGCCCATAACACTTGGGGGTAGCTAAAGTGAACTGTATCCGACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACG
+
AAAAAFFFAFFFAFFFFFFFFFFFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFFF<FFFFFFFFFFF<FFFFFFFFFAFAFFAFFFFFFFFFFFFFFF<FFFAFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFAFFFFFFFFF<F)FFFAF<FFFA
@ERR1654444.4 M_NS500559:9:H16WLBGXX:1:12312:14288:12767B/1
TAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGATGTGCACGCGATAGCATTGCGAAACGCTGGAG
+
AA7FAFFFAAF<AFFFFAFFFAFFFFAFFFFFF<F7FAFF7FFAFAFFA.FF7F7FFFAFFAA<FFA<FFFAAAF7F<FAAFFFAFFFFAFFFFFFF<FFAFAFFFFFFFFFFFFFFFFF<F7FFFFFFFFFFFAFAFFFFAFFFFFFAAAFA
@ERR1654444.5 M_NS500559:9:H16WLBGXX:1:13110:5398:9163B/1
GCTAAAGTGTACTGTGTCCGACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTC
+
AFA7FFFFF<FFFFFFAF7A.AFFFFFFFF7FFFFFFFFFFFFFFFFFFF.FAFFFFFFF<FFFFF)FFF7AFF<FFFAFFAFFAFFFF<AAAFFFFFFFFA<FAAFFFFFFF<FFF.FFFFFAAFFAFA.FFF
@ERR1654444.6 M_NS500559:9:H16WLBGXX:1:21109:10417:15885B/1
TAGCTAAAGTGAACTGTATCCGACATCTGGTTCCTACTTCAGGGCCATAAAGCCTAAATAGCCCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGT
+
```

(2) formatdb: This includes several files that are generated by formatdb, which are used for blast search:



```
animal_mtDNA.fa
animal_mtDNA.fa.nhr
animal_mtDNA.fa.nin
animal_mtDNA.fa.nnd
animal_mtDNA.fa.nni
animal_mtDNA.fa.nsd
animal_mtDNA.fa.nsi
animal_mtDNA.fa.nsq
```

(3) animal_mtDNA.fa: This is the mtDNA database for ancient animals with the normal Fasta format:

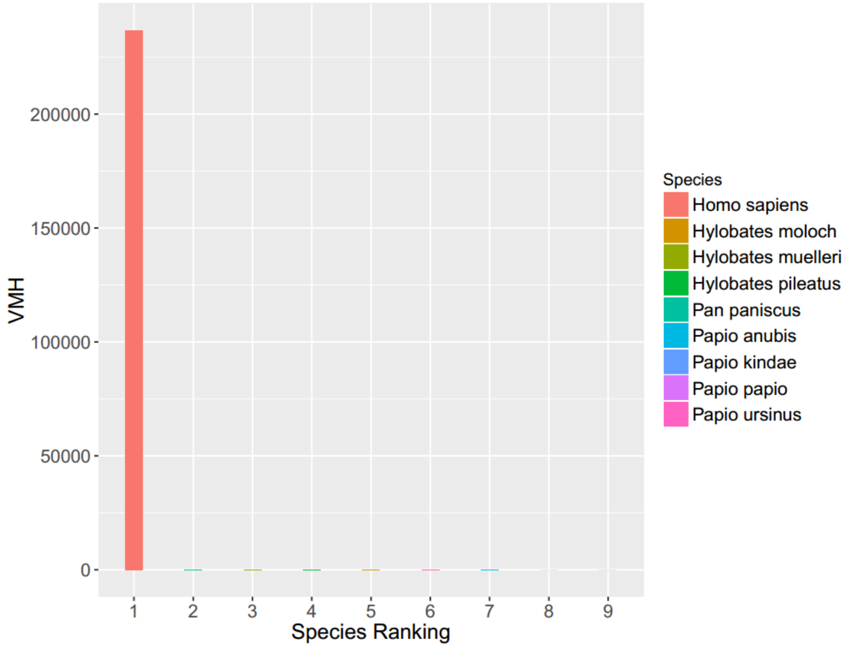


```
>gi|342501|gb|L04272.1|MSQNCATR Anopheles quadrimaculatus A strain Orlando mitochondrion, complete genome
AATGAATTGCCTGATAAAAAGGATTACCTTGATAGGGTAAATCATGAAATTTAATATTTCATTCATTATATTTAATAGAA
TTAAACTATTTCTAAAAGAATCAAAATCTTTTGTGCGTCATACACCAAAATATAAAAAGATAAGCTAACTAAGCTATTGG
GTTCATACCCCACTTATAAAGGTTATAATCCTTTTCTTTTTAATTAAAAAAATTTCTAATAATATTTTCTTTATTACATT
AATTTTTGGAACACTAGTAACTATTTCATCAAATTCGTGGCTAGGAGCATGAATAGGATTAGAAATTAATTTATTGTCAT
TTATCCCCCTAATAAATGACAATAAAAAAAATTTATTAACCTCAGAAAGCTCATTAAAATATTTCTTAACTCAAGCATTT
GCTTCTTCTATTTTATTATTTGCTATTATTATACTAATATTTTTTTATAATAATAATTTTTTATTAATAAATAATTATAA
TGAAATTTTAATTCTTTCAACTTTATTATTAAAAAGAGGAGCCGCTCCATTTCACTTCTGATTTCCTGAAGTAATAGAAG
GATTATCATGAATTAATGGATTATTTTTAATAACATGACAAAAAATTGCTCCTTTAATATTAATTTCTTATAATTTTATT
TATAATTTTTTTATAATTTCTATTATTTTATCAATATTAATTGGATCTTTAGGAGGGTTAAACCAAACTTCAATTCGAAA
ATTAATAGCATTTTCATCTATTAATCATTTAGGATGAATATTATTAGCAATAATAAATAATGAAATATTATGAATAACTT
ATTTTTTAATTTATTCTTTATTATCTATTTCAATTGTTTTAATATTCAATAATTTTAAATTATTTTATTTTAATCAAATT
TTTAATATTTCATTAATAAATCCTATTATTAAATTATTAATTTTCTTAAATTTGCTTTCTTTAGGAGGATTACCCCCTTT
TTTAGGATTTTTACCTAAATGATTAGTAATTCAAAATTTAACTAATATAGGACAATTATTTATTTTAACAATTAGTGTTT
GTTTAACTCTAATTACTTTATTTTTTTATCTTCGATTATCTTATAGAATTTTTATATTAAATTATCAAAAAAATTCATGA
ATATTAAAAAATAATTATAATAATAAAATATCTTCAATTAATTTACTATTAAATTTTATTTCAATTAGAGGATTAATTAT
AATTTTTATAATCTATATAATTATATAAGAATTTAAGTTAAATAAACTAGTAGCCTTCAAAGCTGAATATATTAGTATTA
```

4. Output files:

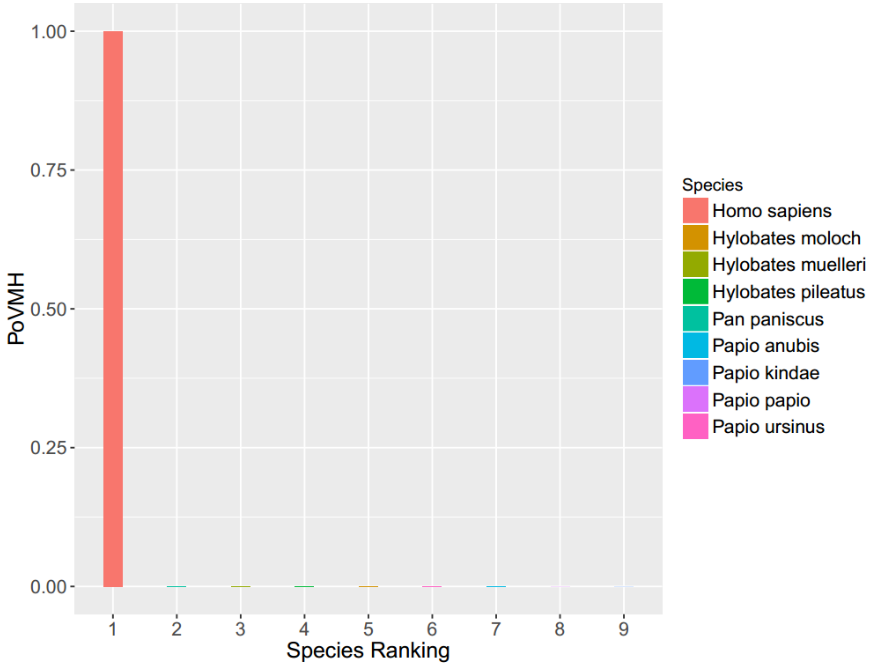
The AncSid will finally generate 3 files:

1. Result.sample1.Mappedcounts_Rank.pdf



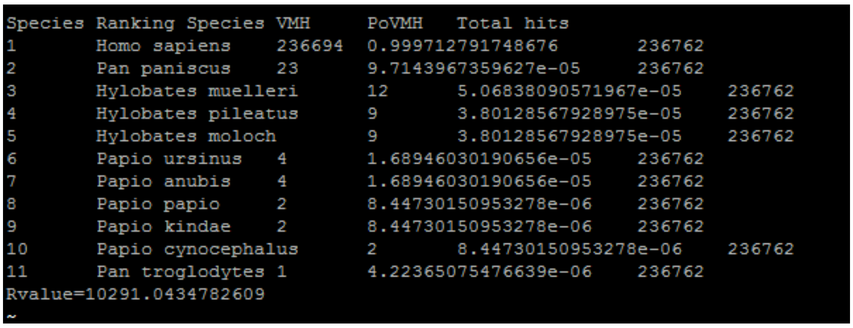
This pdf file presents the result of top10 species ranking information. The Y axis showed the VMH.

2. Result.sample1.Mappedratio_Rank.pdf



This pdf file also presents the result of top10 species ranking information. But the Y axis showed the PoVMH.

3. Result.sample1.txt



```
Species Ranking Species VMH     PoVMH   Total hits
1       Homo sapiens    236694  0.999712791748676       236762
2       Pan paniscus    23      9.7143967359627e-05     236762
3       Hylobates muelleri      12      5.06838090571967e-05    236762
4       Hylobates pileatus      9       3.80128567928975e-05    236762
5       Hylobates moloch        9       3.80128567928975e-05    236762
6       Papio ursinus   4       1.68946030190656e-05    236762
7       Papio anubis    4       1.68946030190656e-05    236762
8       Papio papio     2       8.44730150953278e-06    236762
9       Papio kindae    2       8.44730150953278e-06    236762
10      Papio cynocephalus      2       8.44730150953278e-06    236762
11      Pan troglodytes 1       4.22365075476639e-06    236762
Rvalue=10291.0434782609
~
```

This text showed the detailed information of the top10 species ranking. Including the valid mapping hit number, proportion of the valid mapping hits, the total hits number and the R value.