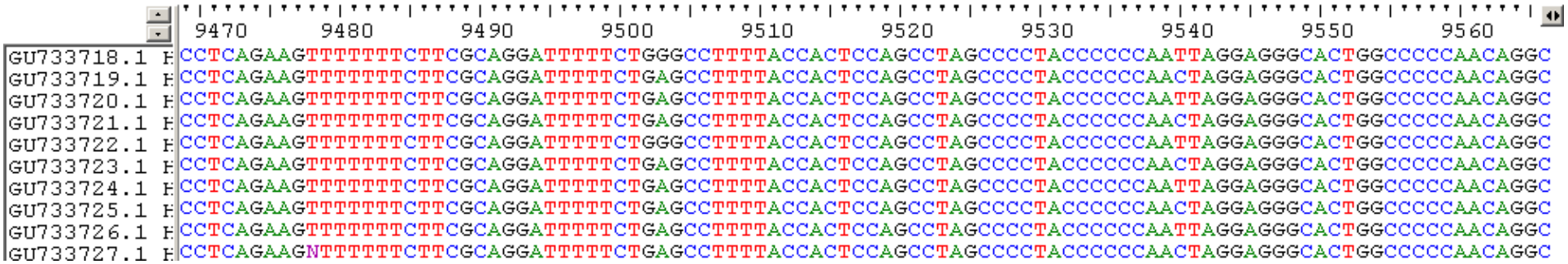
mtDNA analysis

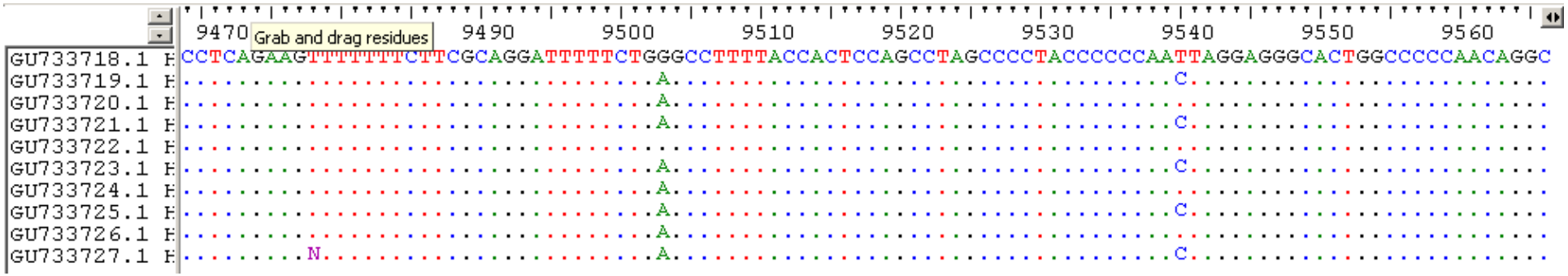
Albert Ko

August 4, 2017

Think about what analysis can you do from just looking at differences among sequences?



Think about what analysis can you do from just looking at differences among sequences?



mtDNA pipeline

1. Genbank
2. Alignment
3. Haplogroup + Tree
4. Network
5. BEAST
6. Summary statistics
7. Simulations

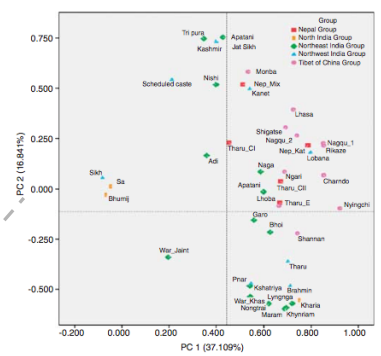
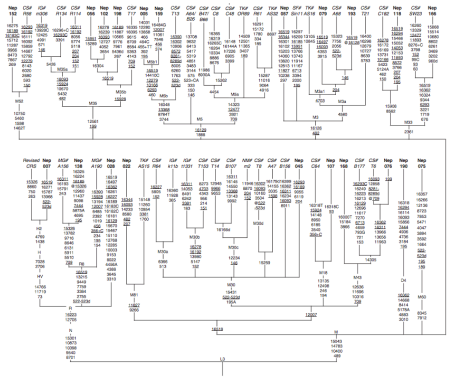
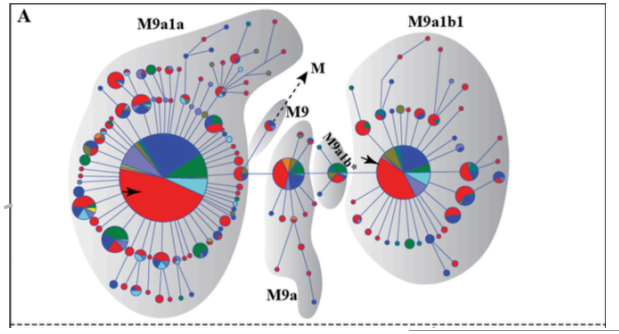
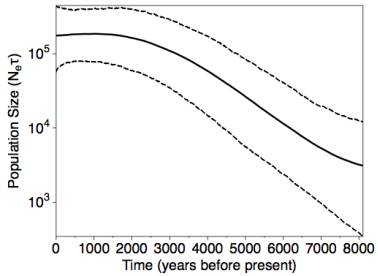
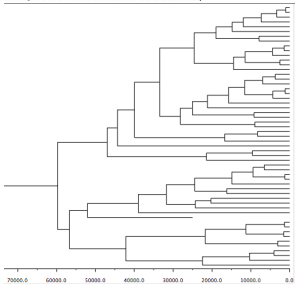


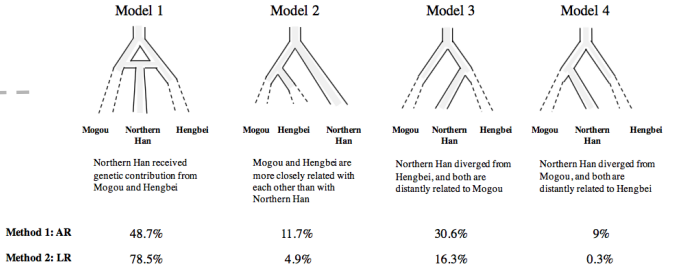
Figure 3 Principle component analysis (PCA) of the populations under study.



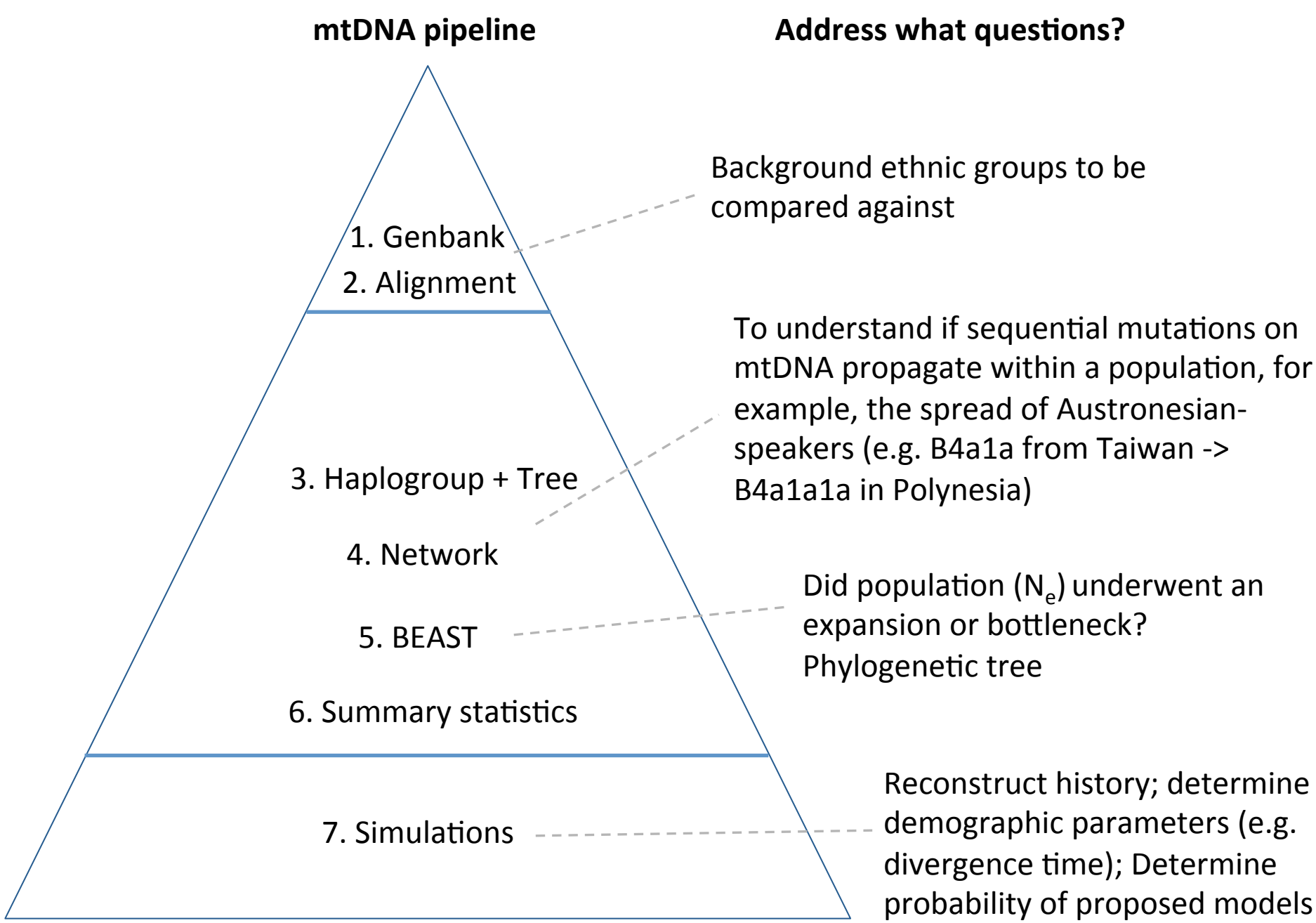






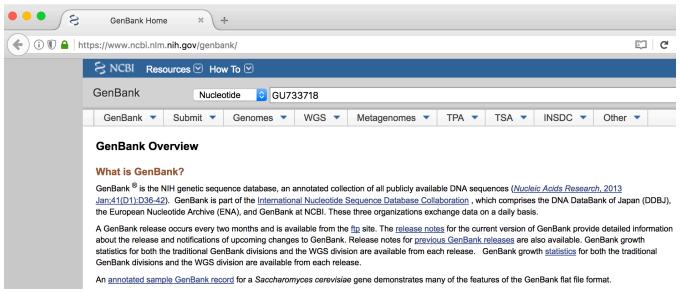


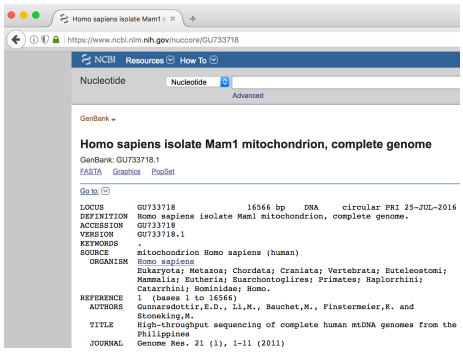
Method 1: AR	48.7%	11.7%	30.6%	9%
Method 2: LR	78.5%	4.9%	16.3%	0.3%

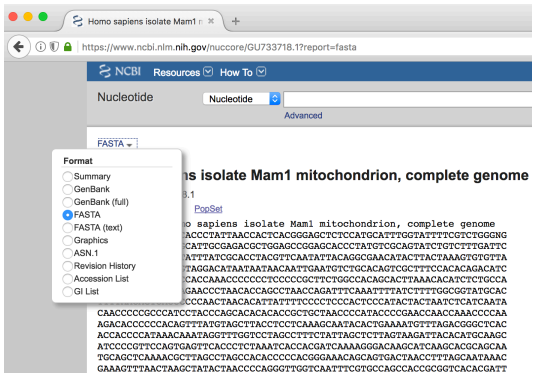

mtDNA pipeline
Address what questions?
1. Genbank
2. Alignment
3. Haplogroup + Tree
4. Network
5. BEAST
6. Summary statistics
7. Simulations
Background ethnic groups to be compared against
To understand if sequential mutations on mtDNA propagate within a population, for example, the spread of Austronesian-speakers (e.g. B4a1a from Taiwan -> B4a1a1a in Polynesia)
Did population (N_e) underwent an expansion or bottleneck? Phylogenetic tree
Reconstruct history; determine demographic parameters (e.g. divergence time); Determine probability of proposed models

Retrieve Accession nos. from Genbank

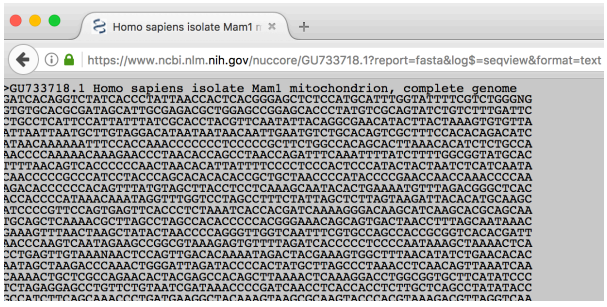
- https://www.ncbi.nlm.nih.gov/genbank/







Choose save as… Fasta (text) will give you a text file page



Batch retrieve Genbank sequences

Place the following 10 accession no., by rows into a file, save as “list.txt”

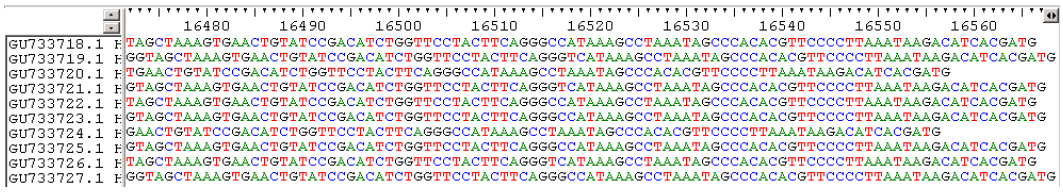
GU733718
GU733719
GU733720
GU733721
GU733722
GU733723
GU733724
GU733725
GU733726
GU733727

Run retrieve bash script in the terminal

This downloads each sequence by order, in fasta format, into file “temp.fas”

```
Albert:- Albert$ bash retrieve.sh > temp.fas
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   313  100   313    0     0     83      0  0:00:03  0:00:03 --:--:--    83
100 16873    0 16873    0     0   2475      0 --:--:--  0:00:06 --:--:--  9719
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   313  100   313    0     0    177      0  0:00:01  0:00:01 --:--:--   177
100 16876    0 16876    0     0   2836      0 --:--:--  0:00:05 --:--:--  4246
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   313  100   313    0     0    421      0 --:--:-- --:--:-- --:--:--   422
100 16864    0 16864    0     0   4259      0 --:--:--  0:00:03 --:--:--  5604
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   313  100   313    0     0    566      0 --:--:-- --:--:-- --:--:--   567
100 16874    0 16874    0     0   8212      0 --:--:--  0:00:02 --:--:-- 18441
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100   313  100   313    0     0    485      0 --:--:-- --:--:-- --:--:--   486
100 16874    0 16874    0     0   3013      0 --:--:--  0:00:05 --:--:--  4509
```

Now, if you look at “temp.fas” in BioEdit, the sequences are not uniformly aligned



mtDNA Sequence alignment

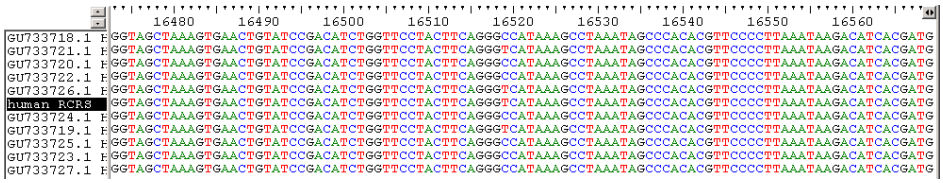
So we concatenate “temp.fas” with rCRS (human mtDNA reference)

```
Albert:- Albert$ cat rcrs.fas temp.fas > combined.fas
```

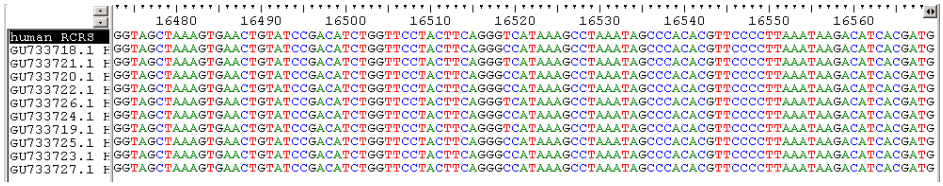
Then align all 11 sequences using muscle program

```
Albert:- Albert$ ./muscle3.8.31_i86darwin64 -in combined.fas -out combined.aln

MUSCLE v3.8.31 by Robert C. Edgar

http://www.drive5.com/muscle
This software is donated to the public domain.
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

combined 11 seqs, max length 16569, avg  length 16565
00:00:00      4 MB(0%)  Iter   1  100.00%  K-mer dist pass 1
00:00:00      4 MB(0%)  Iter   1  100.00%  K-mer dist pass 2
00:00:48    407 MB(2%)  Iter   1  100.00%  Align node
00:00:48    407 MB(2%)  Iter   1  100.00%  Root alignment
00:01:20    414 MB(2%)  Iter   2  100.00%  Refine tree
00:01:20    414 MB(2%)  Iter   2  100.00%  Root alignment
00:01:20    414 MB(2%)  Iter   2  100.00%  Root alignment
00:02:47    414 MB(2%)  Iter   3  100.00%  Refine biparts
Albert:- Albert$
```

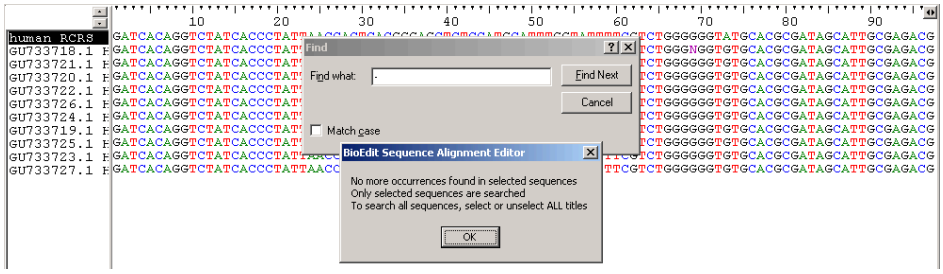
Now, when we look at output in BioEdit, the 11 sequences are aligned to 16569 bp.



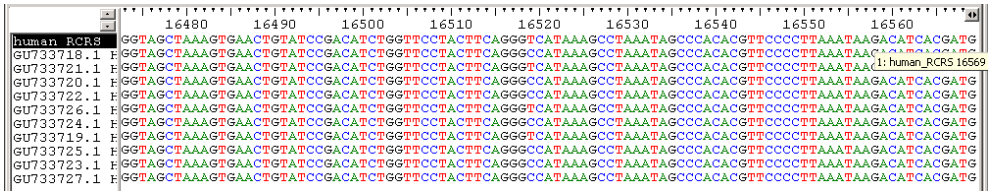
Move the human reference to the first row



Control + F to find gaps (indicated by “-“)

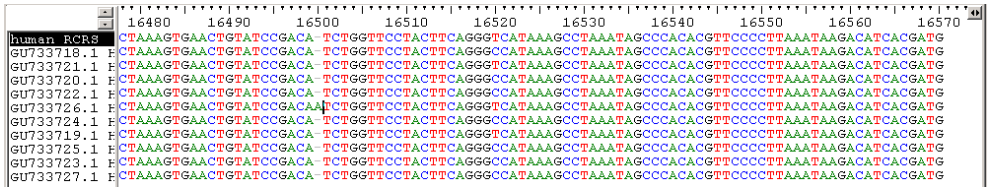


Make sure there are no gap in the human rCRS and total length is 16569 bp.

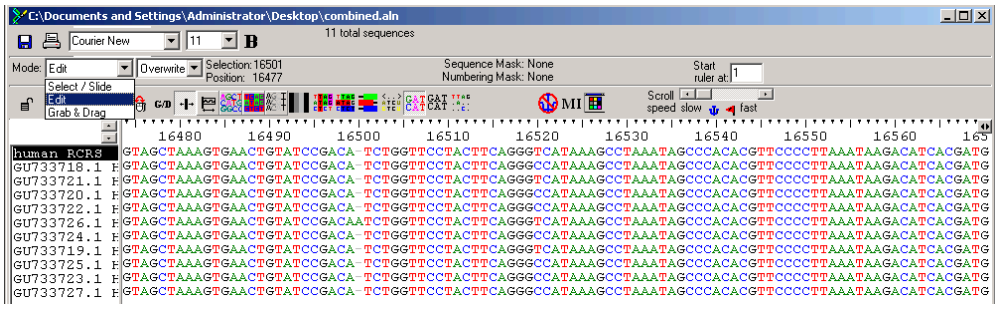


Then this is perfectly aligned to the human reference.

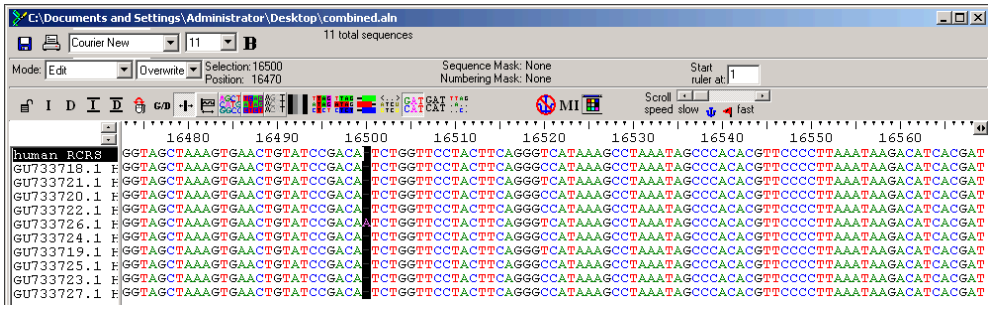
What happens when there is a gap? For example, there is a gap at position 14500. GU733726.1 shows additional A that is missing in reference, you can delete this.



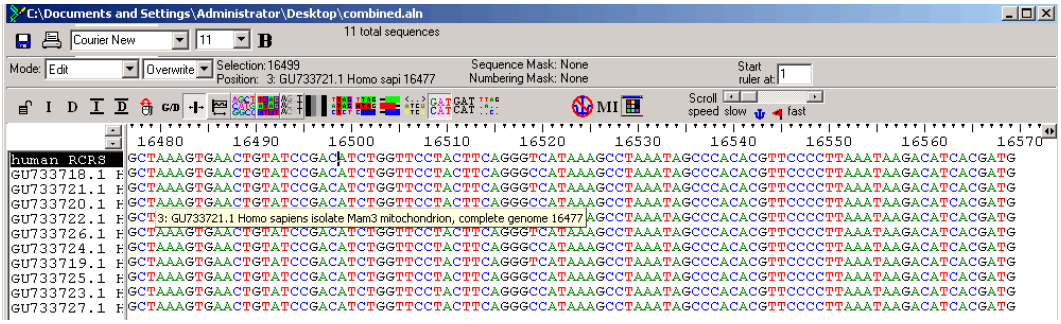
Select Edit,



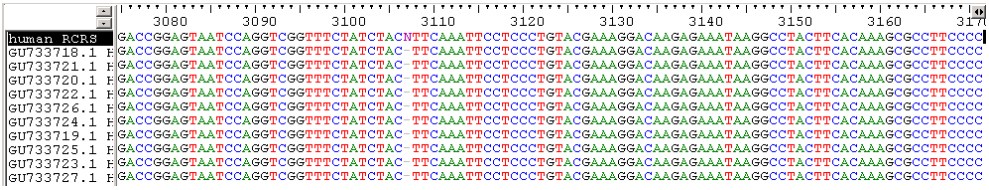
Then highlight position 14500, and press delete.



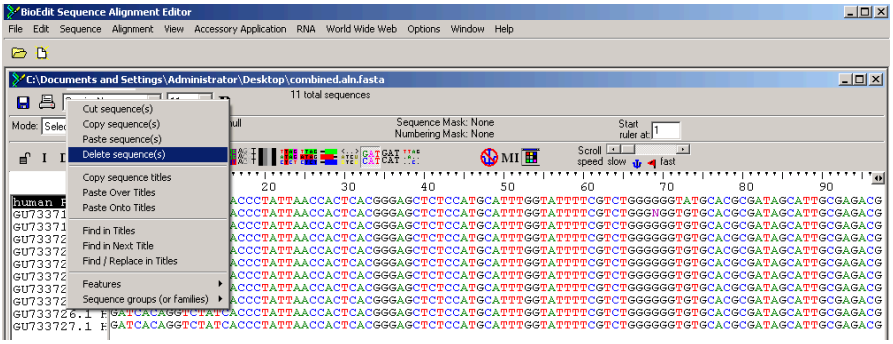
This deletes the entire column (including the extra A nucleotide) and returns 11 sequences to 16569 bp, same as the human reference.

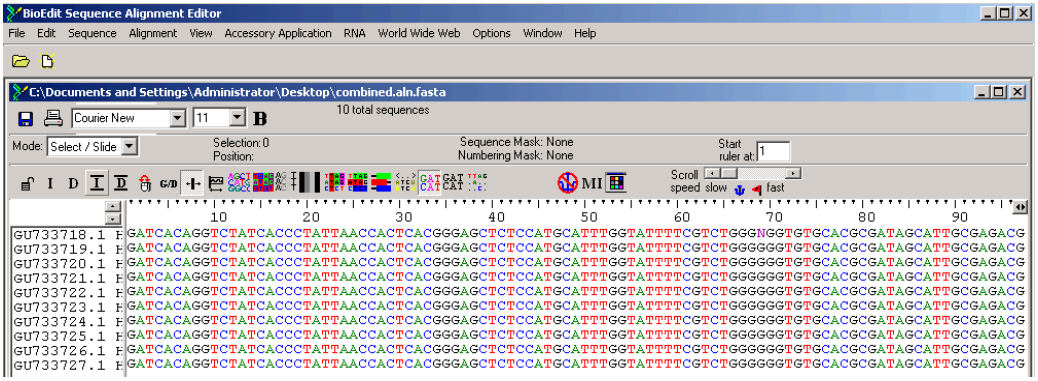


Generally, besides ACGT, the gaps ("-" or "?") are usually created during alignment, and missing data ("N") means the position could not be properly sequenced, and could be due to poor sample quality. Here, the human reference has an N at position 3107, this is normal and is kept there for historical reasons.



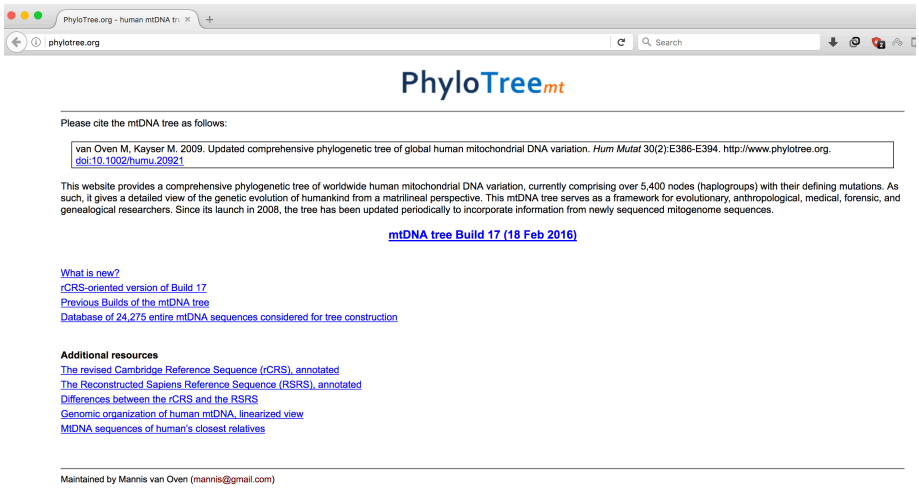
Once alignment to human reference is complete, you can delete the human reference by highlighting the human rCRS -> right click, select delete sequence(s) -> then save.





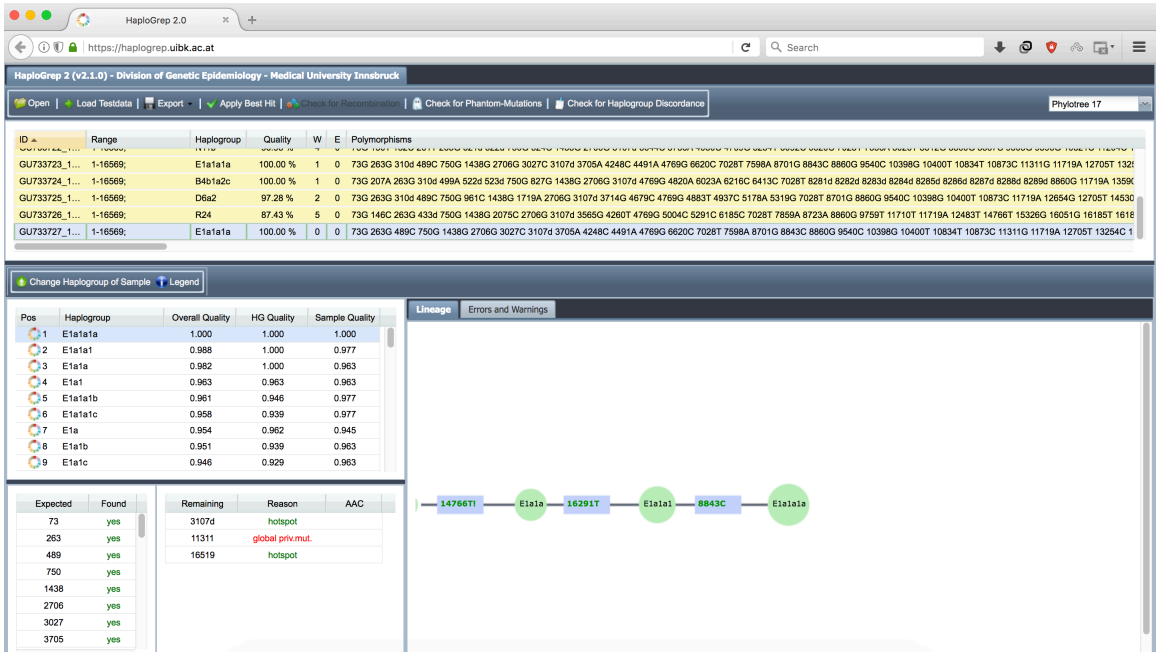
Haplogroup

- mtDNA haplogroups are defined in Phylotree @ http://phylotree.org/
- Latest version is build 17 (Feb 2016)

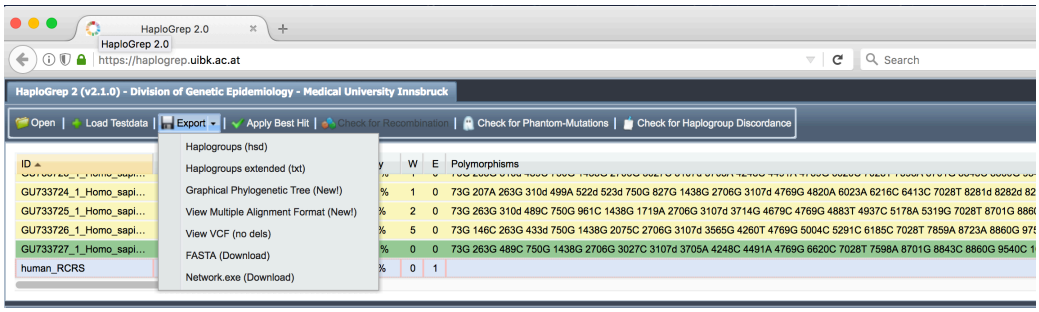


- To call or determine a haplogroup, you can use the following methods:
- Haplogrep2 @ http://haplogrep.uibk.ac.at/
- Mitotools @ http://www.mitotool.org/
- Website @ https://dna.jameslick.com/mthap/

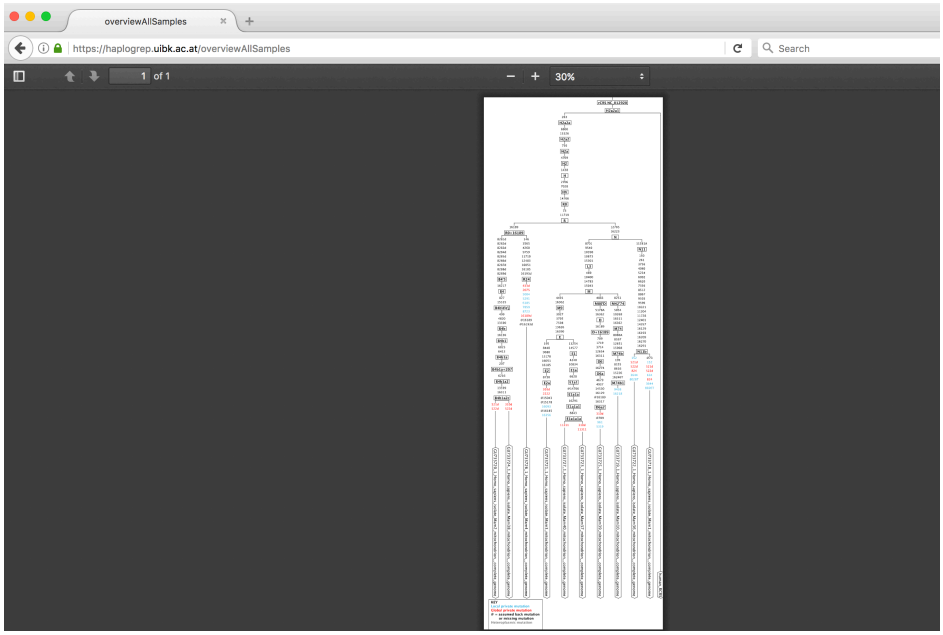
For example, in Haplogrep2 -> rename combined.aln to combined.aln.fasta, then open file. This will automatically determine haplogroups for all samples in the file. Here, you can examine each sequence in detail.



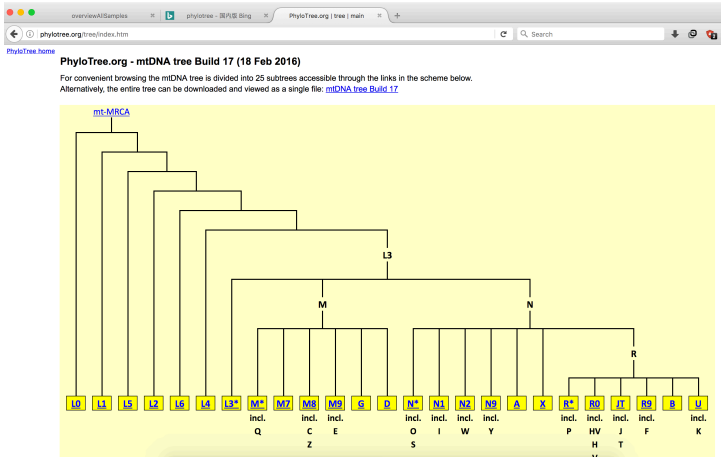
To view tree -> select Export -> “Graphical Phylogenetic Tree”



This will create a haplogroup tree based on your samples

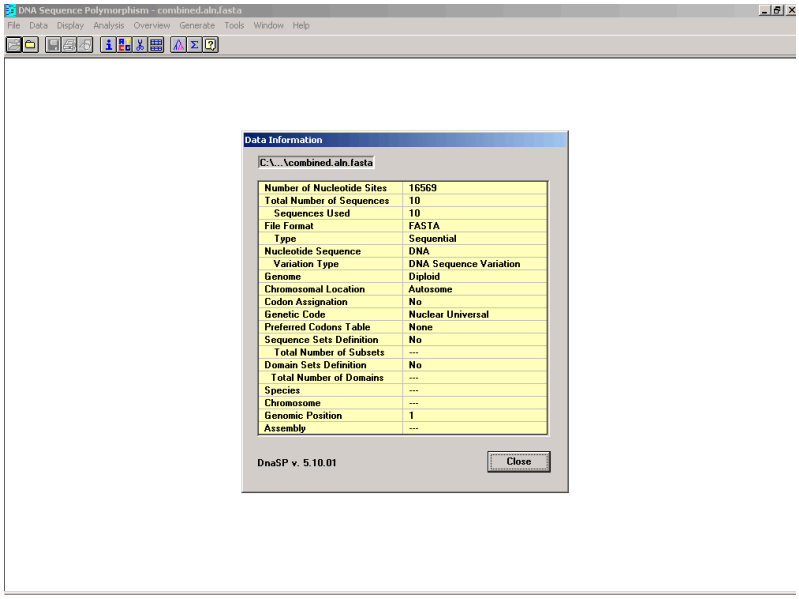


According to the format of the Phylotree,

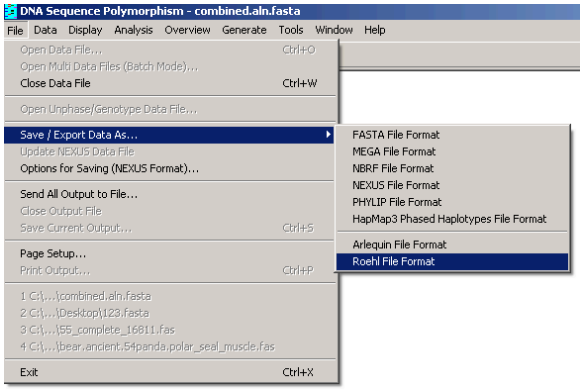


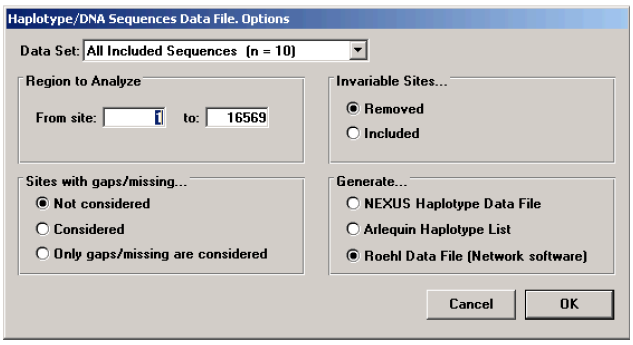
Network

- Download DNAsp @ http://www.ub.edu/dnasp/

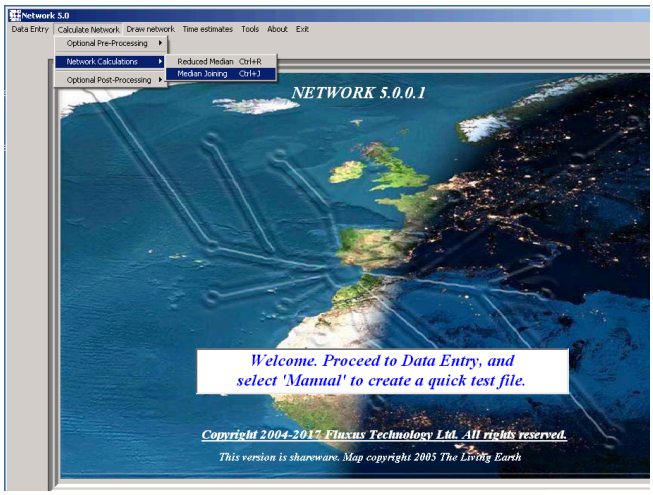


Then select save as Roehl file format, click ok -> save as "sample.rdf"

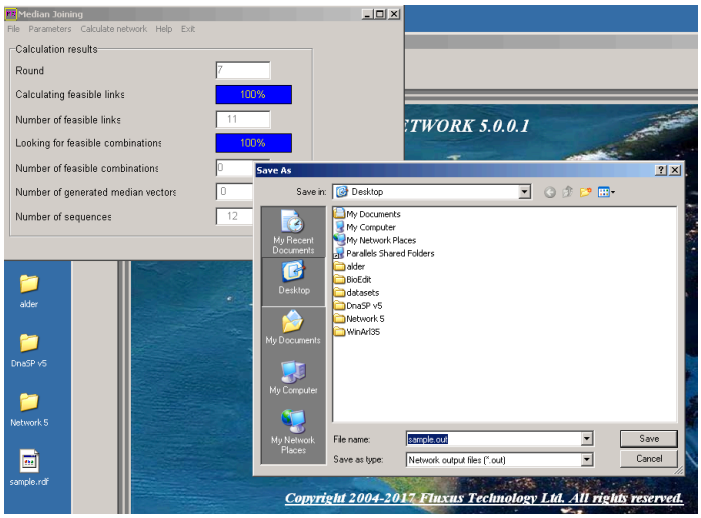




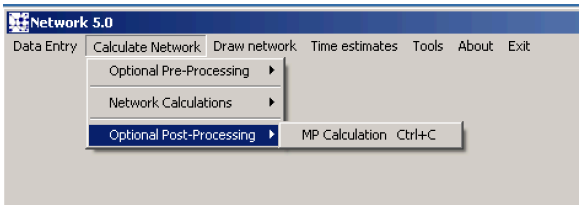
Open Network -> select the median-joining method,



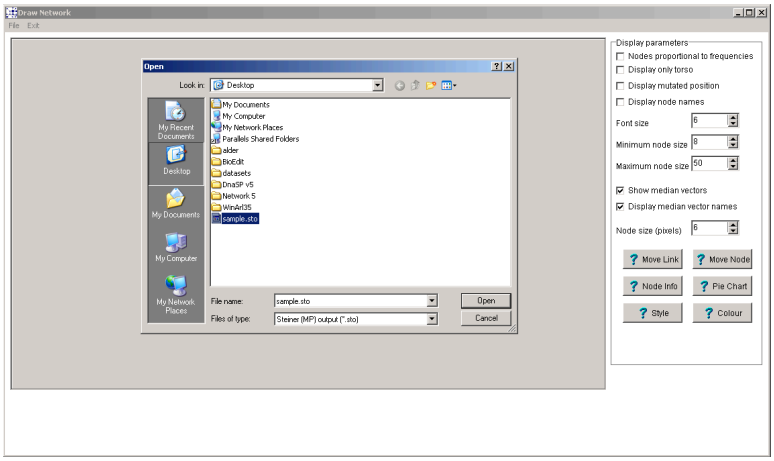
Select calculate network -> save as “sample.out”



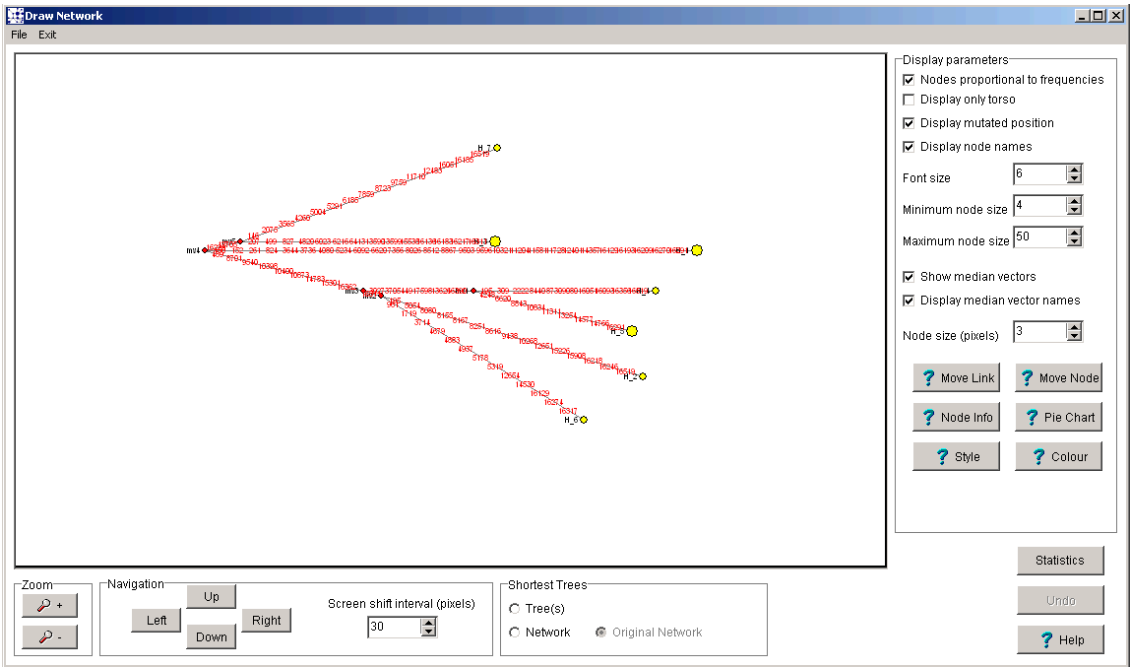
Select optional post-processing -> MP calculation -> save as “sample.sto”



Select draw network -> select sample.sto



You can now view the samples in terms of a network, where the distance between samples are now defined by mutations; and circle is the proportion of samples with the same haplotype (there are 10 samples, but drawn here 7 haplotypes (branches): H_1 to H_7, if you compare this result to haplogroup tree assumedly the program automatically removes gaps with missing data and made some sequences identical to each other)



- Try to look for a characteristic “Star Pattern”, this occurs when the branches are coming out of a central point, and suggests population expansion.
- Be careful when you rotate branches, make sure to double-click + select branch first, because the distances are relative to each other!

Calculate Summary Statistics in Arlequin 3.5

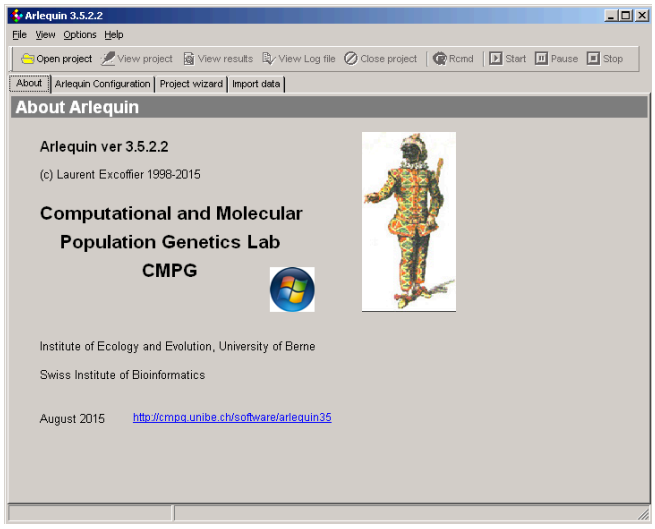
- Arlequin v3.5.2.2 @ http://cmpg.unibe.ch/software/arlequin35/
- Goto download section, get winarl35.zip (windows graphical interface)

You need to prepare input file in the following Arlequin format, and save as (*.arp). This file indicates that we have 3 groups: group1, group2, group3. It says the sample size in the group are: 4, 3, 3 and for every group, includes the sequence label, then 1 means there is 1 sequence of that type, followed by the sequence itself.

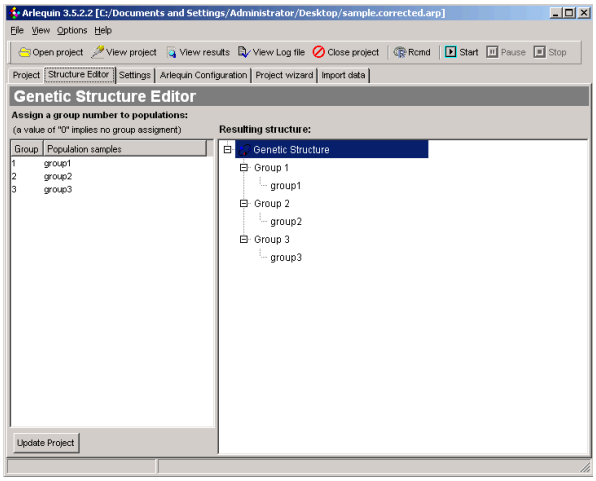
sample.corrected.arp

```
[Profile]
   Title = "Population"
   NbSamples = 10
   DataType = DNA
   GenotypicData = 0
   LocusSeparator = NONE
   MissingData = "?"
   CompDistMatrix = 1


[Data]

[[Samples]]
   SampleName = "group1"
   SampleSize = 4
   SampleData= {
       GU733718 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGNGGTGTGCA
       GU733719 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
       GU733720 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
       GU733721 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
}
   SampleName = "group2"
   SampleSize = 3
   SampleData= {
       GU733722 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
       GU733723 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
       GU733724 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
}
   SampleName = "group3"
   SampleSize = 3
   SampleData= {
       GU733725 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
       GU733726 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
       GU733727 1 GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCTGGGGGGTGTGCA
}

```

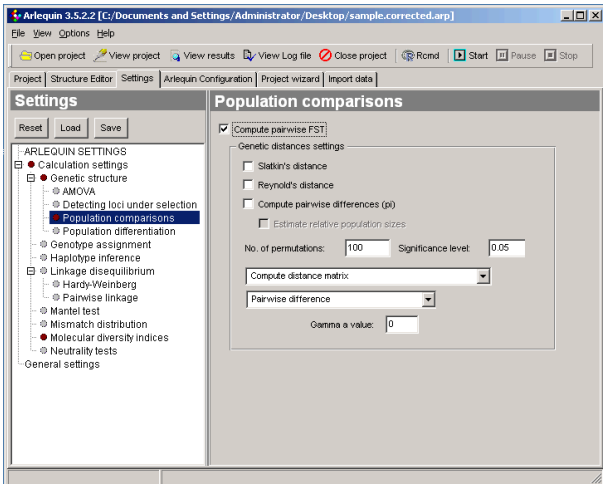
To start Arlequin, run WinArl35.exe

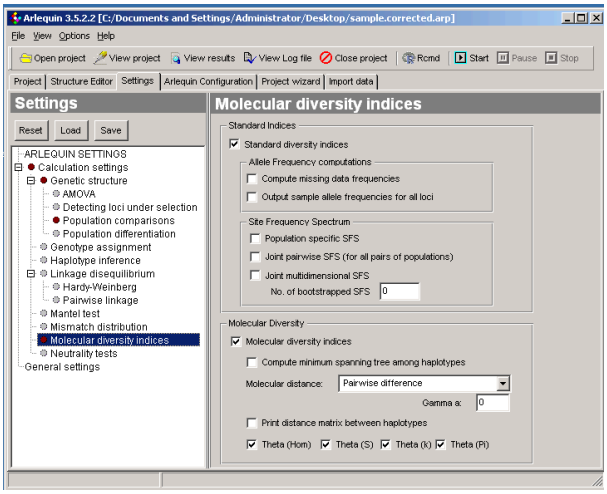


Open Project *.arp, then goto “Structure Editor” tab and assign group1 as 1, group2 as 2, group3 as 3, then press “Update Project” below.

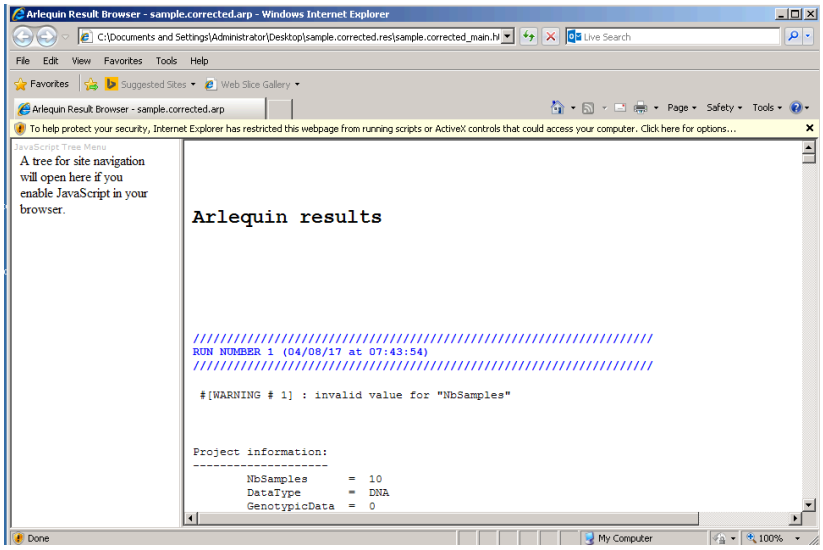


Under Settings tab, tick “Population comparisons” and “ Molecular diversity indicies”





To run the analysis, press “Start” in upper right corner. This will generate a webpage containing the results.



We see that Pairwise difference (Pi) for group1 (50) > group3 (39) > group2 (34), or nucleotide diversity 0.003018 (50/16568) > 0.002374 > 0.002052.

Molecular diversity indexes

Statistics	group1	group2	group3	Mean	s.d.
No. of transitions	79	46	55	60.000	17.059
No. of transversions	5	2	1	2.667	2.082
No. of substitutions	84	48	56	62.667	18.903
No. of indels	12	3	3	6.000	5.196
No. of ts. sites	79	46	55	60.000	17.059
No. of tv. sites	5	2	1	2.667	2.082
No. of subst. sites	84	48	56	62.667	18.903
No. private subst. sites	36	0	25	20.333	18.448
No. of indel sites	12	3	3	6.000	5.196
Pi	50.000	34.000	39.333	41.11111	8.14680

The genetic distance (Fst) result is negative (equivalent to zero) with P-value > 0.05, overall means that the three groups are not significantly different from each other.

Population pairwise FSTs

Distance method: Pairwise difference

	1	2	3
1	0.00000		
2	-0.07522	0.00000	
3	-0.09892	-0.10000	0.00000

FST P values

Number of permutations : 110

	1	2	3
1	*		
2	0.56757+-0.0360	*	
3	0.99099+-0.0030	0.69369+-0.0305	*

For example, if the Fst were significantly positive, such as the following table

	group1	group2	group3
group1	0.00000	0.08256	0.01909
group2	0.08256	0.00000	0.05123
group3	0.01909	0.05123	0.00000

We can proceed to generate a Multi-Dimensional Scaling (MDS) plot, which explains the position of groups relative to each other.