

mtDNA analysis

Albert Ko

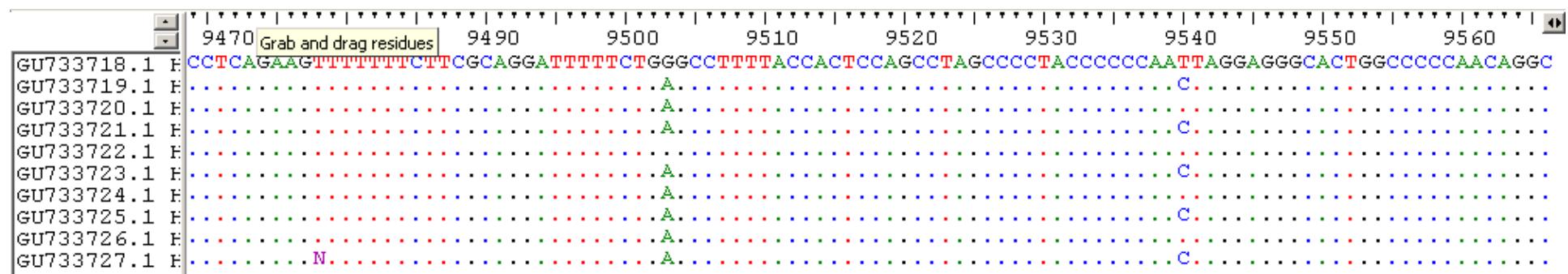
August 4, 2017

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

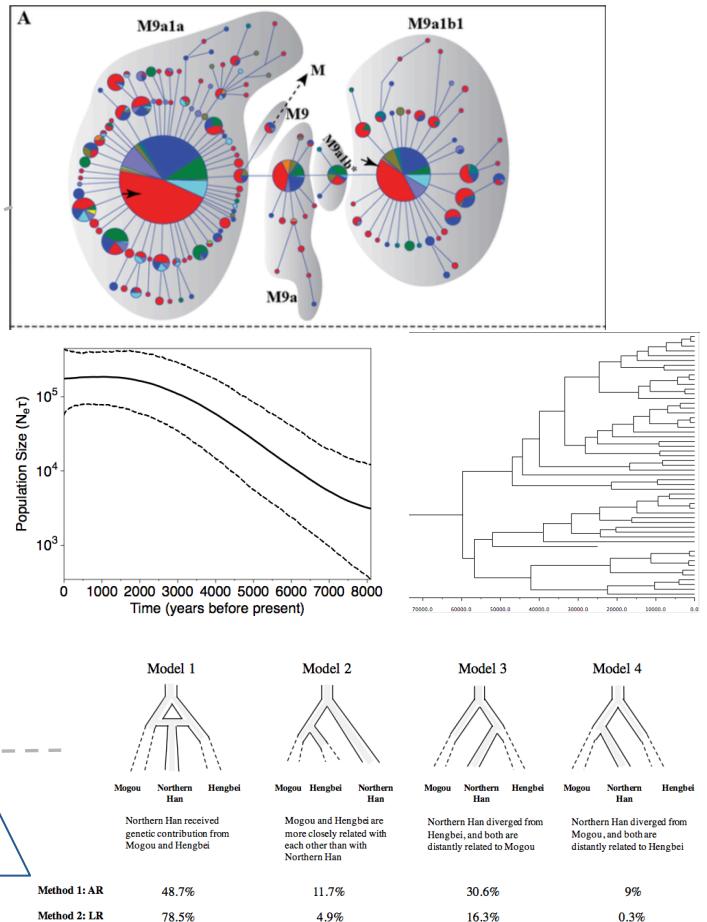
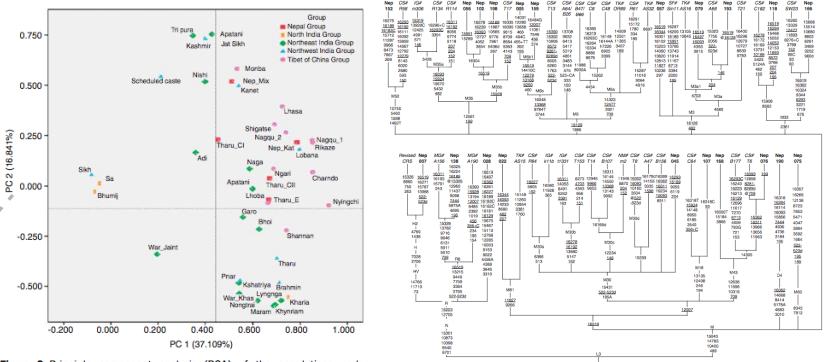
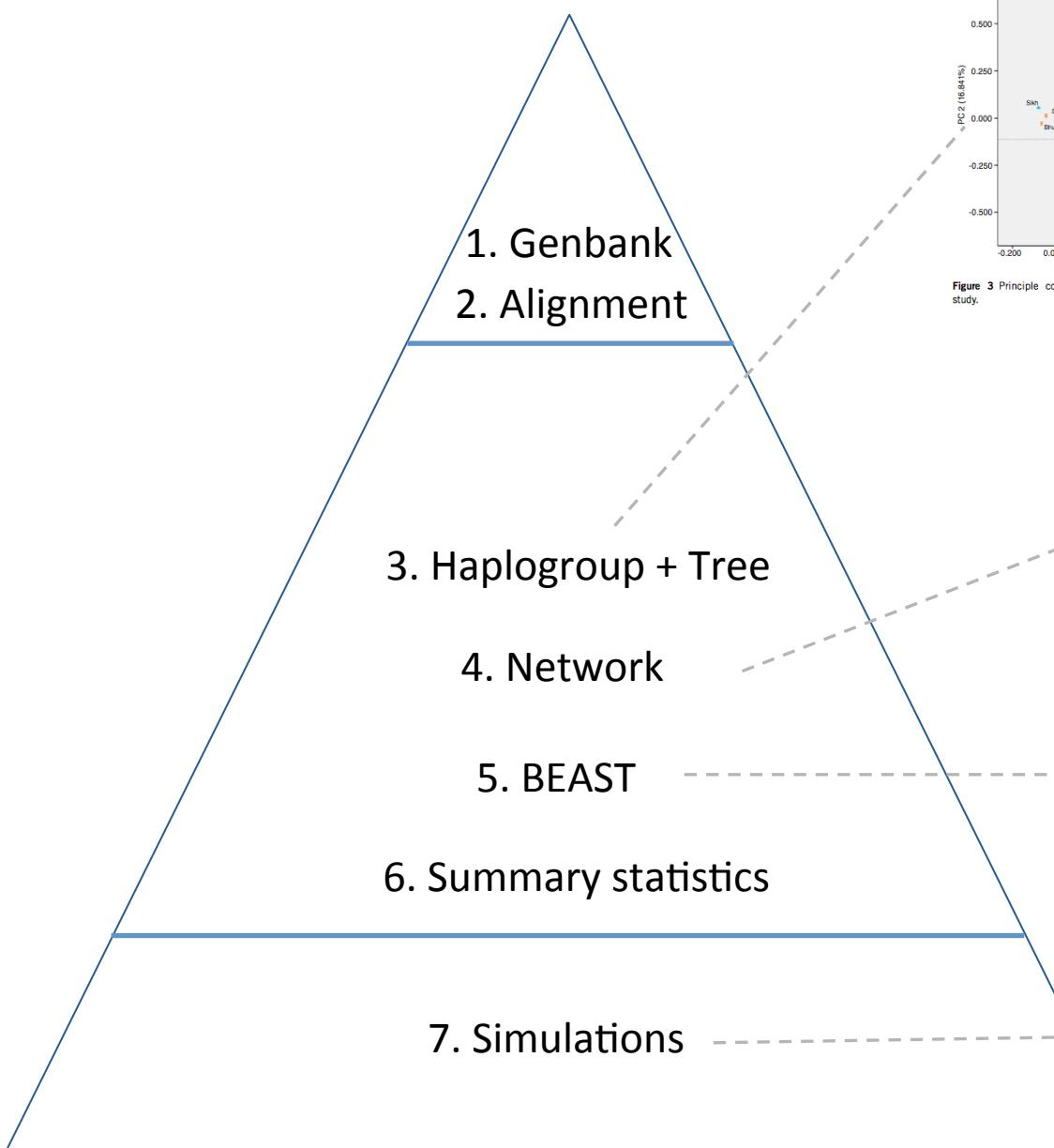
The screenshot shows a sequence alignment interface with 10 sequence entries on the left and a position scale from 9470 to 9560 at the top. Each entry consists of a sequence ID, a status indicator (F), and a sequence of DNA bases. The sequences are highly similar, with minor variations highlighted in different colors (red, green, blue). The alignment shows a repeating motif of TTTTCT followed by a variable sequence.

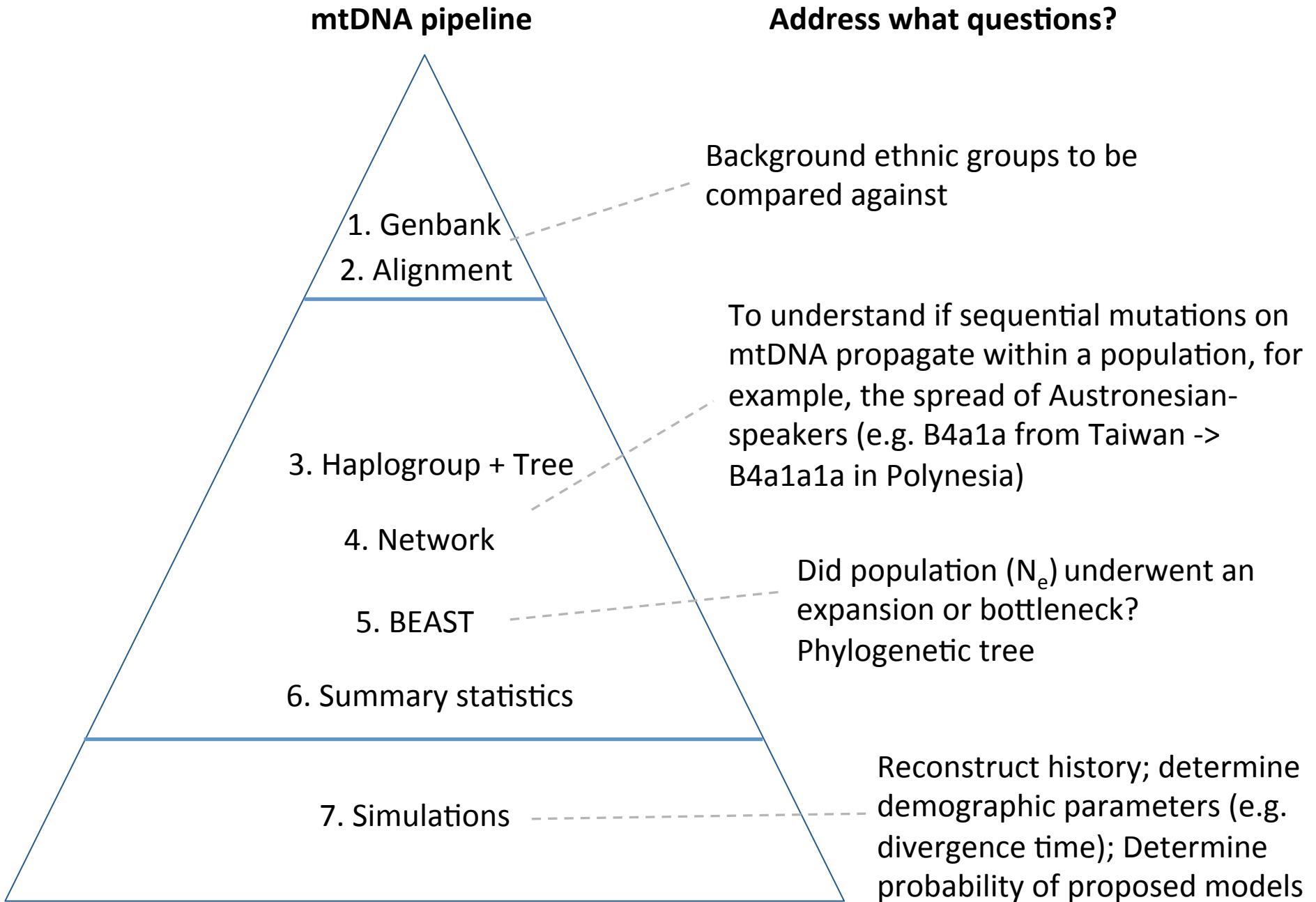
Sequence ID	Status	Sequence
GU733718.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGGGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAATTAGGAGGGCAGTGGCCCCAACAGGC
GU733719.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAACTAGGAGGGCAGTGGCCCCAACAGGC
GU733720.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAATTAGGAGGGCAGTGGCCCCAACAGGC
GU733721.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAACTAGGAGGGCAGTGGCCCCAACAGGC
GU733722.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGGGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAATTAGGAGGGCAGTGGCCCCAACAGGC
GU733723.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAACTAGGAGGGCAGTGGCCCCAACAGGC
GU733724.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAATTAGGAGGGCAGTGGCCCCAACAGGC
GU733725.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAACTAGGAGGGCAGTGGCCCCAACAGGC
GU733726.1	F	CCTCAGAAGTTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAATTAGGAGGGCAGTGGCCCCAACAGGC
GU733727.1	F	CCTCAGAAGNTTTCTCGAGGATTTCTGAGCCTTTACCACTCCAGCCTAGCCCCTACCCCCAAACTAGGAGGGCAGTGGCCCCAACAGGC

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



mtDNA pipeline





Retrieve Accession nos. from Genbank

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

The screenshot shows a web browser window with the title "GenBank Home". The URL is https://www.ncbi.nlm.nih.gov/genbank/. The main content area displays the GenBank homepage, featuring the NCBI logo, a search bar with the query "GU733718", and navigation links for GenBank, Submit, Genomes, WGS, Metagenomes, TPA, TSA, INSDC, and Other. Below this is a section titled "GenBank Overview" with a sub-section "What is GenBank?". It describes GenBank as a collection of publicly available DNA sequences from various sources like Nucleic Acids Research, International Nucleotide Sequence Database Collaboration, DDBJ, ENA, and GenBank at NCBI. It also mentions the release cycle, growth statistics, and the availability of previous releases.

GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2013 Jan 1;41(D1):D42–62). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [file site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample [GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

The screenshot shows the NCBI GenBank search results for the "Homo sapiens isolate Mam1 mitochondrial genome". The page title is "Homo sapiens isolate Mam1 mitochondrial genome, complete genome". The accession number is GU733718.1. The genome is 16566 bp long and is circular. The taxon is Homo sapiens isolate Mam1 Mitochondrion. The source is mitochondria from human. The reference is Bulykaryan et al., 2006. The authors are G. Bulykaryan, D.O. Li, N.V. Bauchet, M., Finstermeier, K., and Stonking, K. The title is "High-throughput sequencing of complete human mtDNA genomes from the". The journal is Genome Res. 21 (1), e11 (2011).

The screenshot shows a web browser window with the following details:

- Title Bar:** Homo sapiens isolate Mam1 | NCBI
- URL:** https://www.ncbi.nlm.nih.gov/nucleotide/GU337181?report=fasta
- Header:** NCBI Resources How To
- Search Bar:** Nucleotide
- Advanced Search:** Advanced
- Format Selection:** FASTA (selected)
- Sequence View:** The main area displays the mitochondrial genome sequence for Homo sapiens isolate Mam1. The sequence starts with: ACGCCACCCGGCCATCTTCAAGCACACACAGCGCTTCATACCCCATACCGGAAACCAAAACCCAA.

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

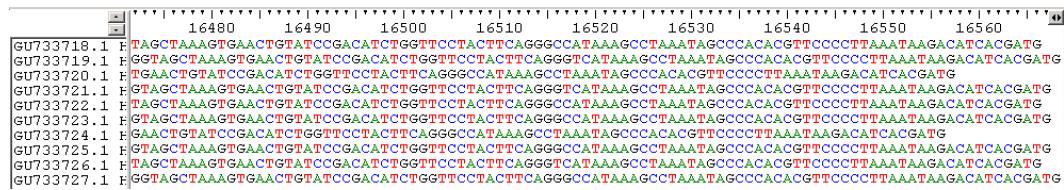


```
1. bash  
Albert:- Albert$ bash retrieve.sh > temp.fas
```

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



```
GU733718.1 F TAGCTAAAGTGAACGTGATTCGGACATCGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733719.1 F GGTAGCTAAAGTGAACGTGATTCGGACATTCGGACATCTGGTTCCACTCTCAGGGTCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733720.1 F TGAACTGTATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733721.1 F GTAGCTAAAGTGAACGTGATTCGGACATCTGGTTCCACTCTCAGGGTCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733722.1 F TAGCTAAAGTGAACGTGATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733723.1 F STAGCTAAAGTGAACGTGATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733724.1 F GAACTGTATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733725.1 F TAGCTAAAGTGAACGTGATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733726.1 F TAGCTAAAGTGAACGTGATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG  
GU733727.1 F GGTAGCTAAAGTGAACGTGATTCGGACATCTGGTTCCACTCTCAGGGCCAATAAGCTAAATAGCCCCACAGGTTCCCTTAAAATAGACATCACGATG
```

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.

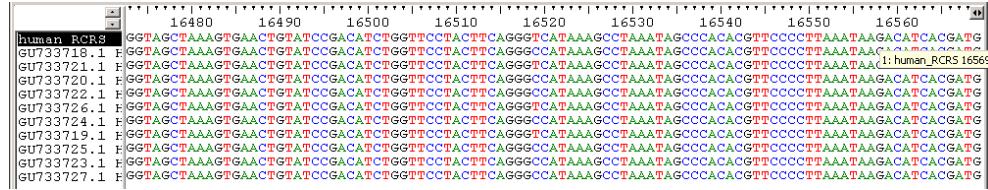
combined 11 seqs, max length 16569, avg length 16565
00:00:00    4 MB(0%) Iter 1 100.00% K-mer dist pass 1
00:00:00    4 MB(0%) Iter 1 100.00% K-mer dist pass 2
00:00:48   407 MB(2%) Iter 1 100.00% Align node
00:00:48   407 MB(2%) Iter 1 100.00% Root alignment
00:01:20   414 MB(2%) Iter 2 100.00% Refine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

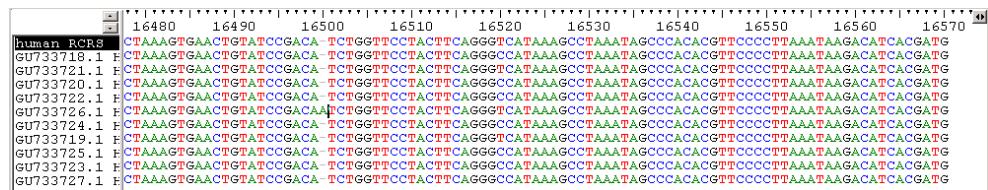
Control + F to find gaps (indicated by “-”)

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

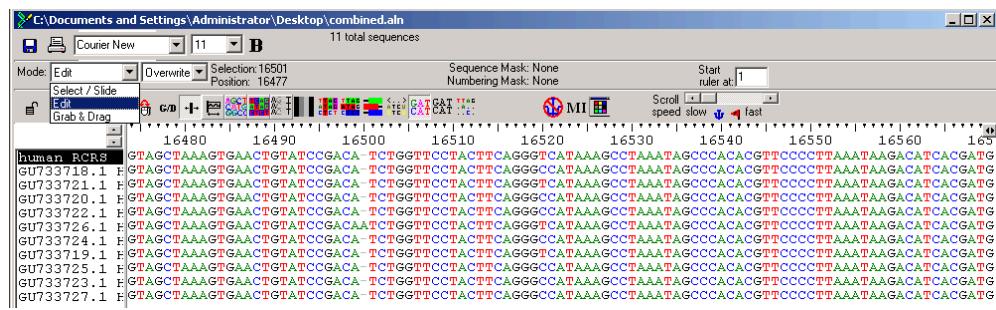


Then this is perfectly aligned to the human reference.

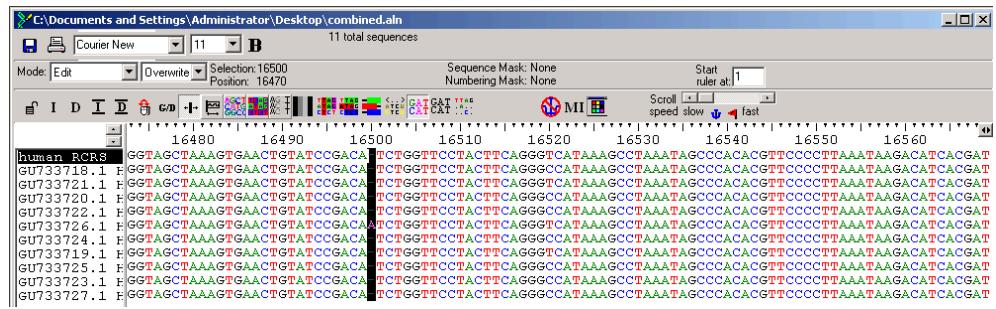
What happens when there is a gap? For example, there is a gap at position 14500. GU733726.1 shows additional A that is missing in reference, you can delete this.



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)

The screenshot shows the PhyloTree.org homepage. At the top, there's a search bar and a link to 'PhyloTree 17'. Below the header, a note says 'Please cite the mtDNA tree as follows:' followed by a citation: 'van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum Mutat* 30(2):E386-E394. http://www.phylotree.org. doi:10.1002/humu.20921'.

The main content area features a detailed phylogenetic tree of human mtDNA. A note below the tree states: 'This website provides a comprehensive phylogenetic tree of worldwide human mitochondrial DNA variation, currently comprising over 5,400 nodes (haplogroups) with their defining mutations. As such, it gives a detailed view of the genetic evolution of humankind from a matrilineal perspective. This mtDNA tree serves as a framework for evolutionary, anthropological, medical, forensic, and genealogical researchers. Since its launch in 2008, the tree has been updated periodically to incorporate information from newly sequenced mitogenome sequences.' Below the tree, a link reads 'mtDNA tree Build 17 (18 Feb 2016)'.

On the left side, there's a sidebar with links to 'What is new?', 'rCRS-oriented version of Build 17', 'Previous Builds of the mtDNA tree', and 'Database of 24,275 entire mtDNA sequences considered for tree construction'. Another section titled 'Additional resources' includes links to 'The revised Cambridge Reference Sequence (rCRS), annotated', 'The Reconstructed Sapiens Reference Sequence (RSRS), annotated', 'Differences between the rCRS and the RSRS', 'Genomic organization of human mtDNA, linearized view', and 'MtDNA sequences of human's closest relatives'.

At the bottom, a small note says 'Maintained by Mannis van Oven (mannis@gmail.com)'.

- To call or determine a haplogroup, you can use the following methods:
- Haplogrep2 @ <http://haplogrep.uibk.ac.at/>
- Mitotoools @ <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.

The screenshot shows the Haplogrep 2.0 software interface. At the top, there's a menu bar with 'File', 'Edit', 'Testdata', 'Export', 'Help', and 'About'. Below the menu, a toolbar includes icons for 'Open', 'Load Testdata', 'Export', 'Apply Best Hit', 'Check for Recombination', 'Check for Phantom-Mutations', and 'Check for Haplogroup Discordance'. The title bar says 'Haplogrep 2.0 - Division of Genetic Epidemiology - Medical University Innsbruck'.

The main window contains a table of samples with columns: ID, Range, Haplogroup, Quality, W, E, and Polymorphisms. Some rows are highlighted in yellow. Below the table is a 'Change Haplogroup of Sample' dialog with a legend. The legend shows nine colored circles corresponding to different haplogroups: E1a1a1a (blue), E1a1a1 (orange), E1a1a (red), E1a1 (green), E1a1a1b (purple), E1a1a1c (pink), E1a (yellow), E1a1b (brown), and E1a1c (grey).

On the right side, there are two tabs: 'Lineage' and 'Errors and Warnings'. The 'Lineage' tab shows a phylogenetic tree with nodes labeled with mutation sites: 14766T, Elala, 16291T, Elalol, 8843C, and Elalala. The 'Errors and Warnings' tab shows a table with columns: Expected, Found, Remaining, Reason, and AAC.

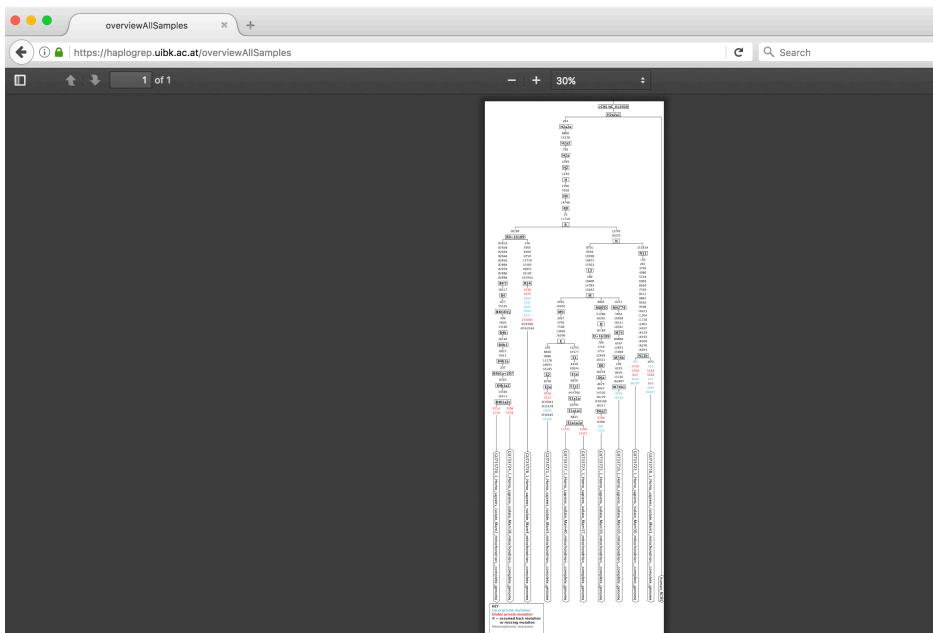
To view tree -> select Export -> "Graphical Phylogenetic Tree"

The screenshot shows the HaploGrep 2.0 software interface. In the top navigation bar, there are tabs for 'Open', 'Load Testdata', 'Export', 'Apply Best Hit', 'Check for Recombination', 'Check for Phantom-Mutations', and 'Check for Haplotype Discordance'. The 'Export' tab is currently selected. A dropdown menu is open under 'Export' containing the following options:

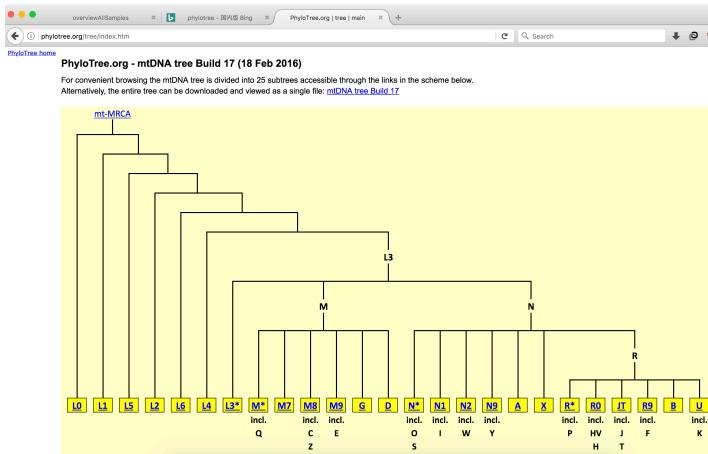
- Haplotype (hdt)
- Haplotype extended (btx)
- Graphical Phylogenetic Tree (New!)
- View Multiple Alignment Format (New!)
- View VCF (no dels)
- FASTA (Download)
- Network.exe (Download)

Below the menu, there is a table titled 'Polymorphisms' with columns for Y, W, E, and Polymorphisms. The table contains several rows of data, with the last row highlighted in green.

This will create a haplogroup tree based on your samples

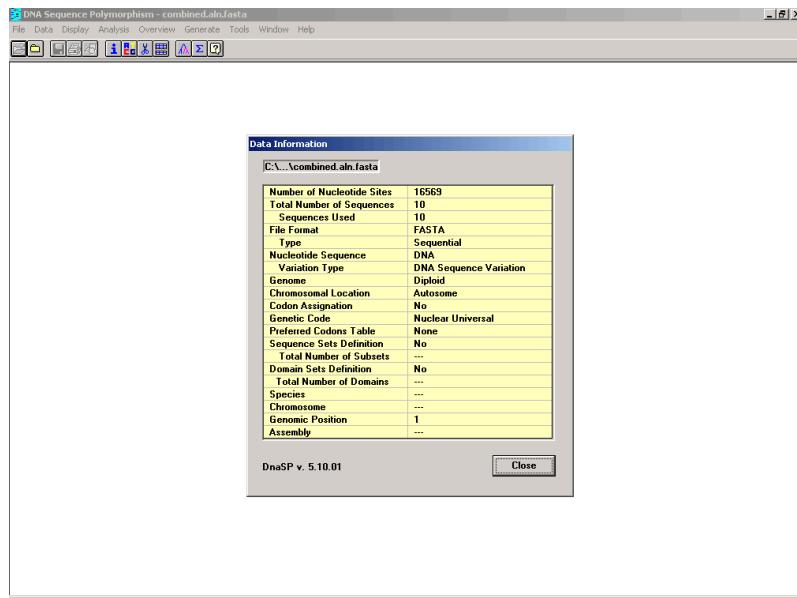


According to the format of the Phylotree,

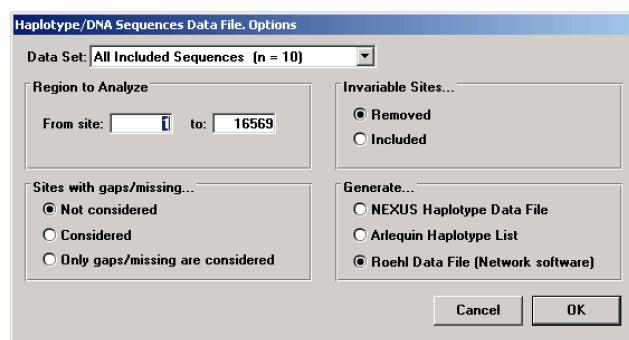
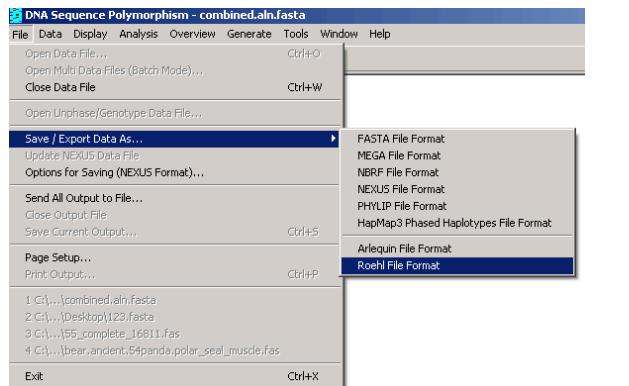


Network

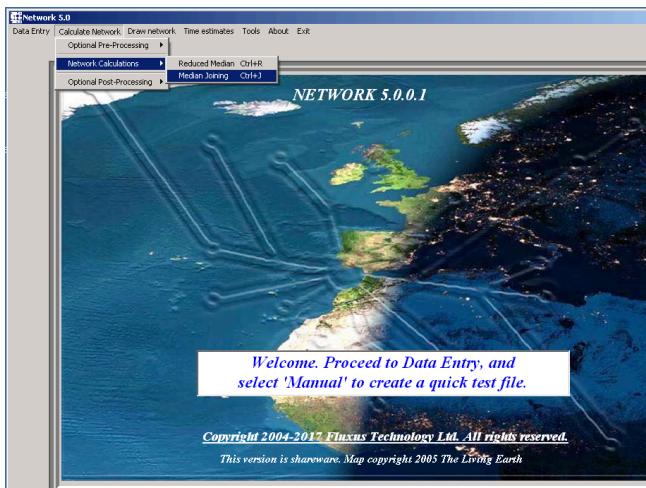
- Download DNAsP @ <http://www.ub.edu/dnasp/>



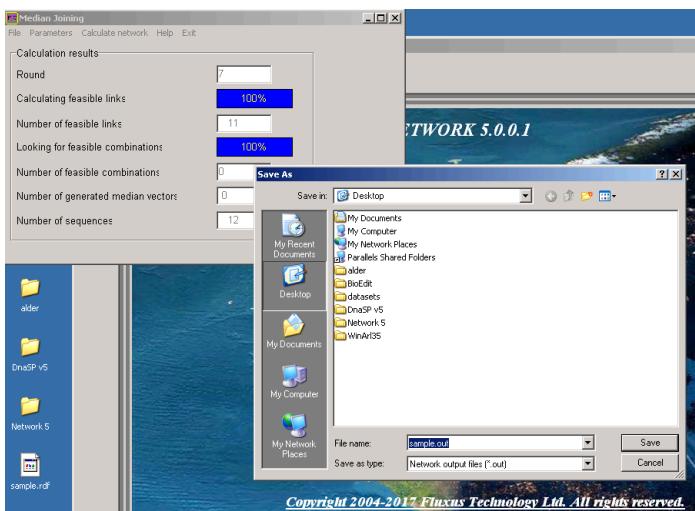
Then select save as Roehl file format, click ok -> save as "sample.rdf"



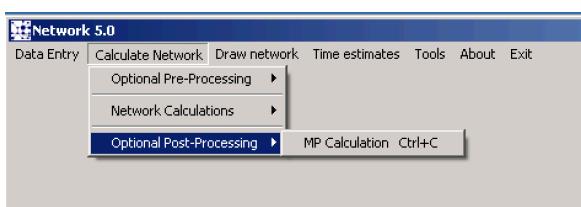
Open Network -> select the median-joining method,



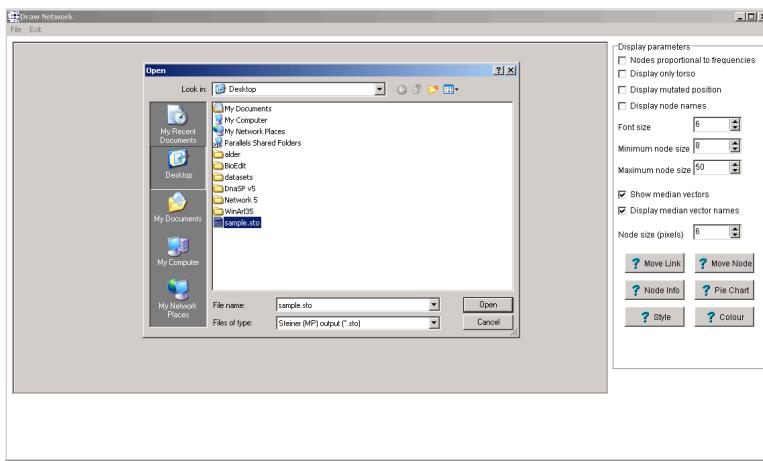
Select calculate network -> save as "sample.out"



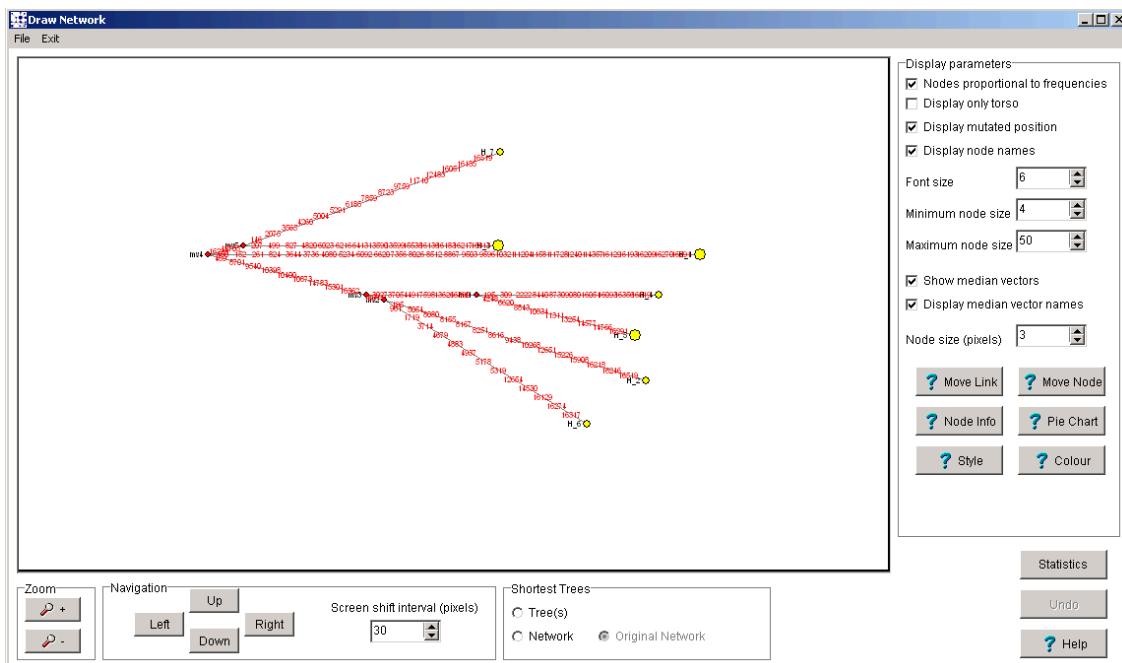
Select optional post-processing -> MP calculation -> save as "sample.sto"



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

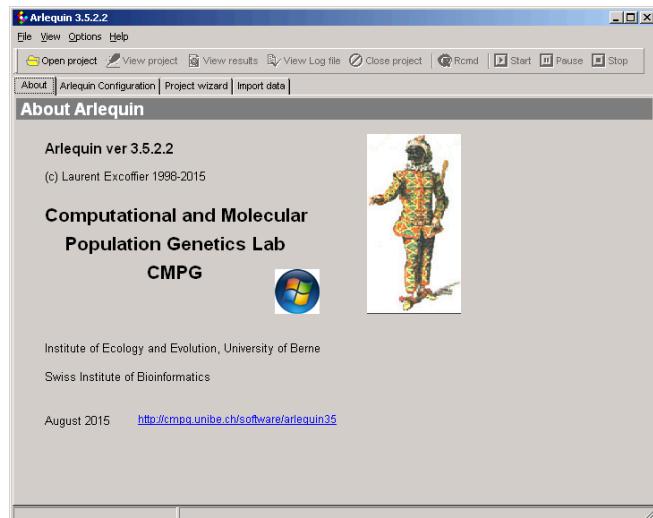
- Arlequin v3.5.2.2 @ <http://cmpg.unibe.ch/software/arlequin35/>
- Goto download section, get winarl35.zip (windows graphical interface)

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.

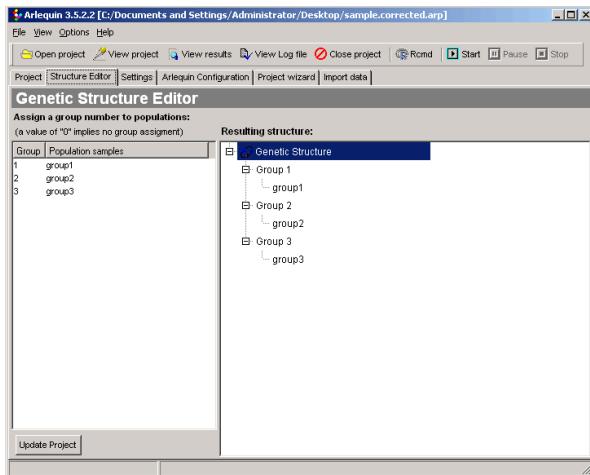
```
< > sample.corrected.arp *
```

```
1 [Profile]
2   Title = "Population"
3   NbSamples = 10
4   DataType = DNA
5   GenotypicData = 0
6   LocusSeparator = NONE
7   MissingData = "?"
8   CompDistMatrix = 1
9
10
11 [Data]
12
13 [[Samples]]
14   SampleName = "group1"
15   SampleSize = 4
16   SampleData= {
17     GU733718 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
18     GU733719 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
19     GU733720 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
20     GU733721 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
21   }
22   SampleName = "group2"
23   SampleSize = 3
24   SampleData= {
25     GU733722 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
26     GU733723 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
27     GU733724 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
28   }
29   SampleName = "group3"
30   SampleSize = 3
31   SampleData= {
32     GU733725 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
33     GU733726 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
34     GU733727 1 GATCACAGGTCTATCACCCTATTAAACCACTCACGGGAGCTCTCATGCATTGGTATTTCTGGTCTGGGGGGTGTGCA
35   }
36 }
```

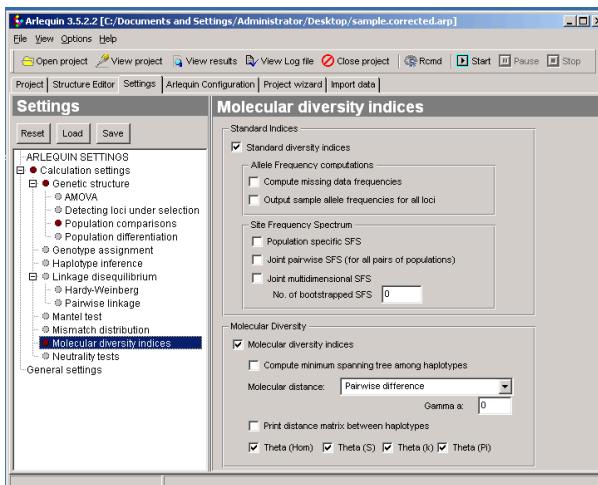
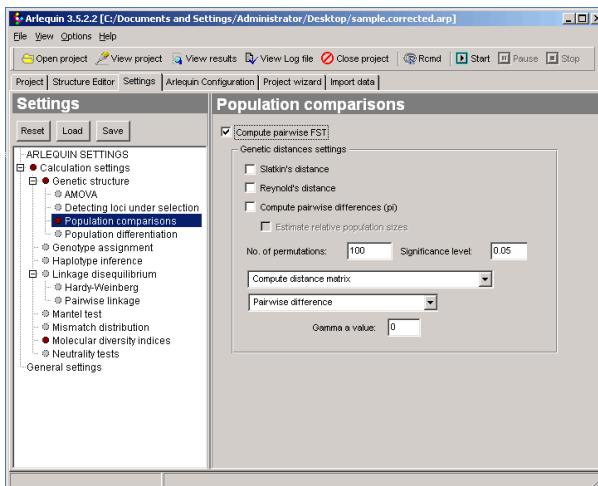
To start Arlequin, run WinArl35.exe



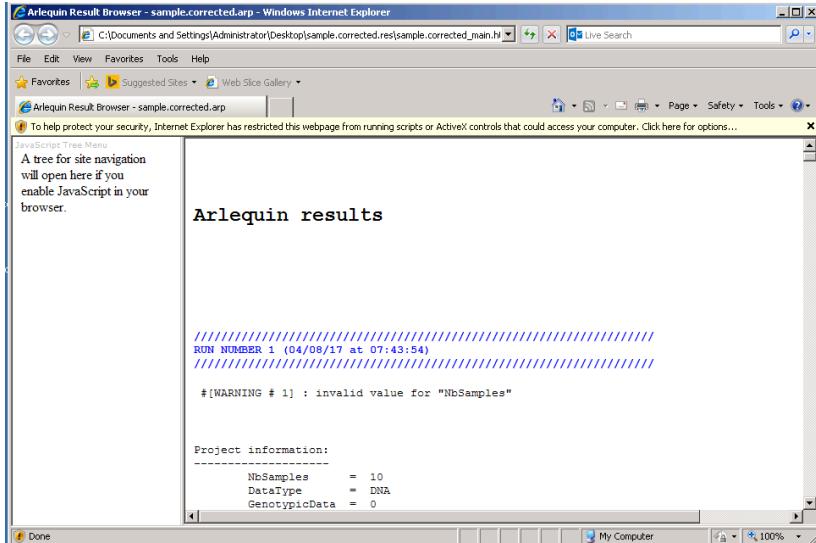
Open Project *.arp, then goto “Structure Editor” tab and assign group1 as 1, group2 as 2, group3 as 3, then press “Update Project” below.



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



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



We see that Pairwise difference (Δ) 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.

