RDM analysis for Human populations

Jing Peng

Here we use the Root Distance Method (RDM) to infer the phylogeny for 49 Human populations. This is a Human Genome Diversity Cell Line Panel (HGDP) data set from ALlele FREquency Database (ALFRED). More details about the dataset can be found at http://alfred.med.yale.edu or in the manuscript. Since this is a large dataset, I saved the workspace in R, which can reduce loading time. I also put the necessary RDM functions in the saved workspace for convenience. For general use, the RDM source code can be found in Supplemental Materials.

We have 6 datasets in the worksapce 'HumanData.RData', 'mat_allele_freq', 'log_mat_allele_freq', 'sqrt_mat_allele_freq', 'ang_mat_allele_freq', 'mat_allele_freq_noExtreme', and 'log_mat_allele_freq_noExtreme'. The first one is the original dataset, but I replace the cell with value 0/1 with 0.01/0.99 such that the transformation can be done. The second to fourth datasets are transformed dataset using logit, square root, and angular transformation, respectively. The last two datasets are similar with the first two, but instead of replacing 0/1 with close number, we delete those extreme values. Again, we pre-processing the data and save them because the file is too large to be loaded efficiently.

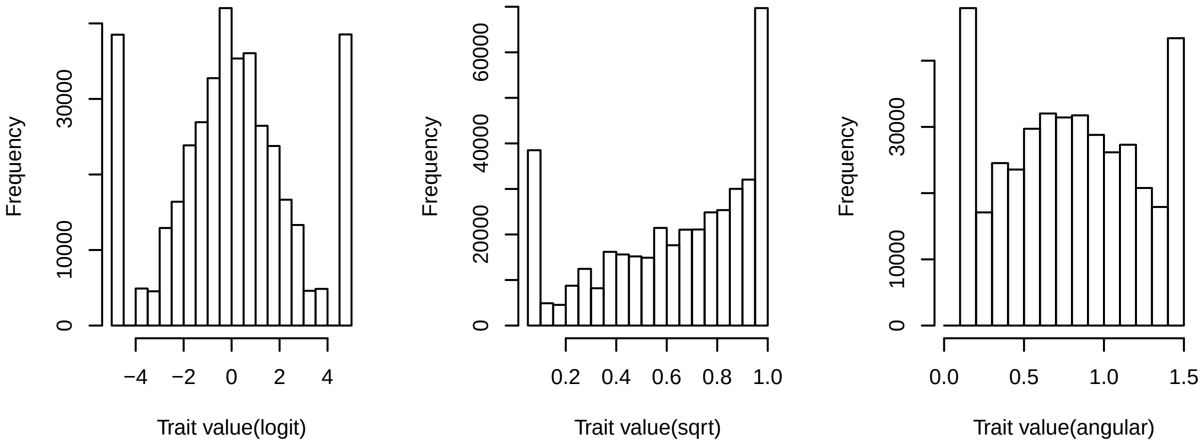
```
load("HumanData.RData")
dim(mat_allele_freq)
```

```
## [1]     49 402430
```

```
mat_allele_freq[1:7,1:7]
```

```
##                         V1        V2    V3        V4        V5         V6
## Balochi         0.01959184 0.9000000 0.180 0.3893878 0.5604082 0.04040816
## Han             0.06000000 0.2600000 0.010 0.6000000 0.6500000 0.25000000
## Bantu_speakers  0.47473684 0.4457895 0.952 0.7131579 0.0100000 0.01000000
## Italians        0.01000000 0.5600000 0.010 0.2233333 0.7047059 0.01000000
## Biaka                   NA        NA    NA 0.8500000 0.0100000 0.01000000
## Mandenka        0.67000000 0.4800000 0.010 0.6400000 0.0500000 0.01000000
## Mbuti           0.67000000 0.1700000 0.990 0.9900000 0.0100000 0.01000000
##                        V7
## Balochi         0.7257143
## Han             0.7744444
## Bantu_speakers  0.6052632
## Italians        0.9200000
## Biaka           0.7000000
## Mandenka        0.5700000
## Mbuti           0.6300000
```

```
par(mfrow=c(1,3))
hist(as.numeric(log_mat_allele_freq['Han',]),xlab='Trait value(logit)',nclass=15,main='')
hist(as.numeric(sqrt_mat_allele_freq['Han',]),xlab='Trait value(sqrt)',nclass=15,main='')
hist(as.numeric(ang_mat_allele_freq['Han',]),xlab='Trait value(angular)',nclass=15,main='')
```



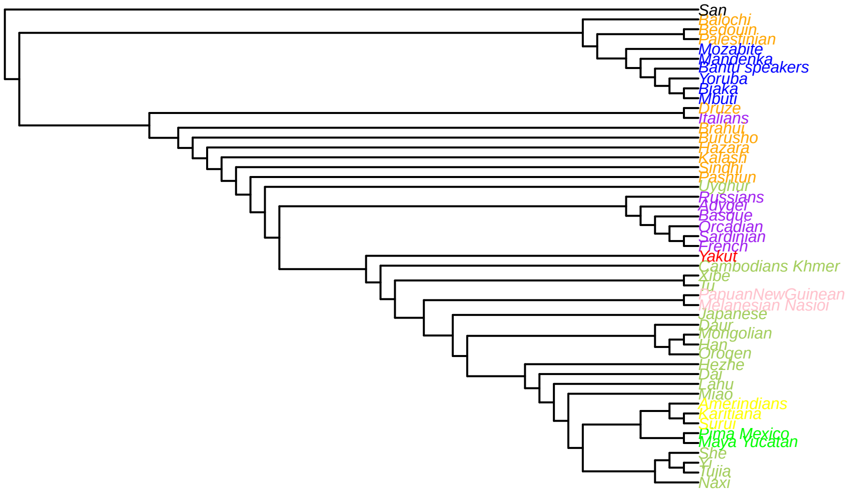
We can see that the dataset has 402,430 sites (columns) and 49 populations (rows). Angular transformation and logit transformation can make data approximately normally distributed, which may be preferred for the following analysis. We also noticed that there are some extreme values for this population, but we may keep that for constructing the tree because they are informative.

```
group_Africa<-list('Mbuti','Biaka','Yoruba','Bantu_speakers','Mandenka','Mozabite')
group_Europe<-list('Italians','Russians','Adygei','Basque','Orcadian','Sardinian','French')
group_Asia<-list('Bedouin','Palestinian','Balochi','Druze','Brahui',
                 'Burusho','Hazara','Kalash','Pashtun','Sindhi')
group_EastAsia<-list('Tujia','Yi','She','Naxi','Miao','Lahu','Hezhe','Dai','Oroqen','Daur',
                     'Han','Mongolian','Cambodians_Khmer','Japanese','Xibe','Tu','Uyghur')
group_Siberia<-list('Yakut')
group_Oceania<-list('PapuanNewGuinean','Melanesian_Nasioi')
group_NorthAmerica<-list('Maya_Yucatan','Pima_Mexico')
group_SouthAmerica<-list('Surui','Karitiana','Amerindians')
```

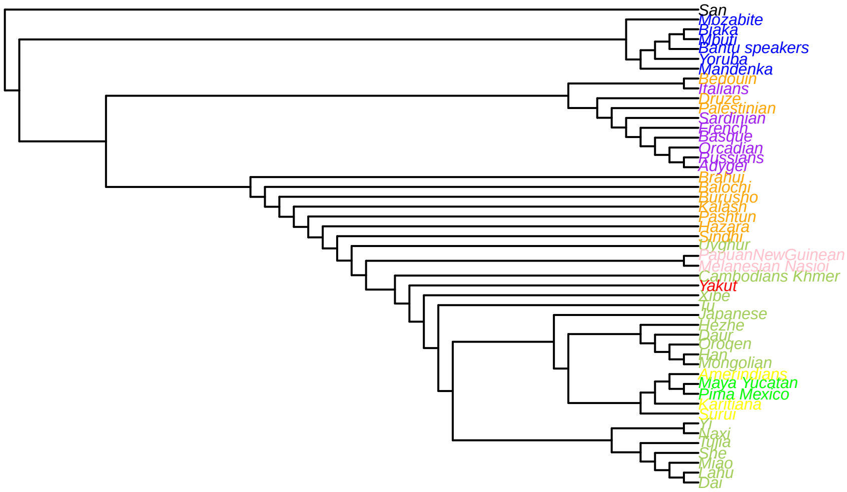
We can then assign group labels according to the geographical location where samples come from, and see if the tree can explain that difference. We can call “RDM” function directly. The first argument is (transformed) allele frequency matrix. Ideally, the data is approximately normal, but multivariate normality is hard to check. Usually we use logit transformed or angular tranformed data from closely related populations, which is justified in evolutionary biology that genetic drift assumption can be satisfied. More details about the assumptions can be checked in our manuscript. RDM needs the specification of outgroup for carrying out the analsyis, and this can be either the population name or a numerical row number of the outgroup data. The “use=” option let users specify which part of data is used to compute the covariance matrix, and details can be checked in ‘cov’ function in R.

```
rd_tre1<-RDM(log_mat_allele_freq,outgroup='San',use="pairwise.complete.obs")
tipcol<-rep("black",length(rd_tre1$tip.label))



tipcol[vapply(group_Africa, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'blue'
tipcol[vapply(group_Europe, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'purple'
tipcol[vapply(group_Asia, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'orange'
tipcol[vapply(group_EastAsia, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'darkolivegreen3'
tipcol[vapply(group_Siberia, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'red'
tipcol[vapply(group_Oceania, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'pink'
tipcol[vapply(group_NorthAmerica, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'green'
tipcol[vapply(group_SouthAmerica, function(x) which(rd_tre1$tip.label==x), numeric(1))]<-'yellow'
```

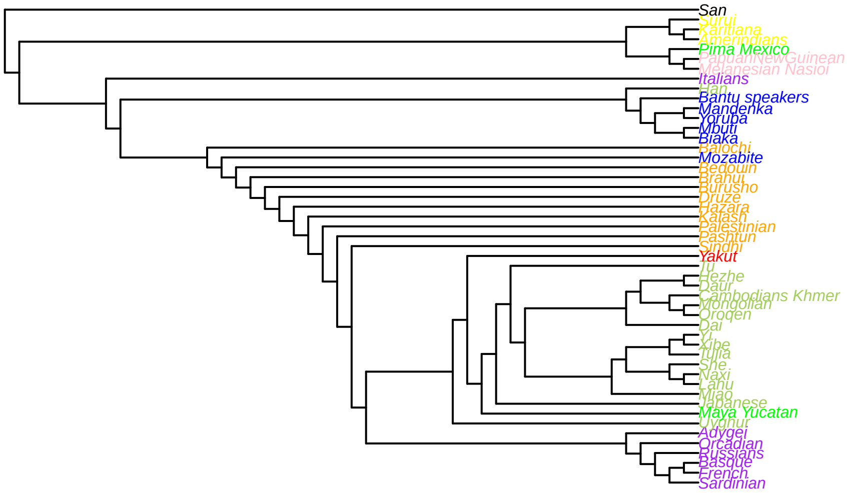
```
plot(rd_tre1,
     use.edge.length = FALSE,
     tip.color=tipcol,
     cex = 0.5)
```



We can try another tranformation (angular):



To accommendate one of our reviewers comments, we also build the tree after deleting extreme values:



The tree actually changed in this case: the two Americans groups (South American, one population from North America) and populations from Oceania are grouped as monophyletic , and they are more closely related to each other than to Europeans, Africans or Asians. Other weird things include Italians and Hans become ancestral population of Africans. My guess is that we still need those extreme gene frequencies 0 or 1 when we estimate covariance matrix. Actually those sites are very informative on how far two populations are related, and the covariance will be influenced by them. Especially if we check the number of sites with extreme values for every pair of populations, 24 paris of them have over 50% of sites with 0/1 gene frequencies.