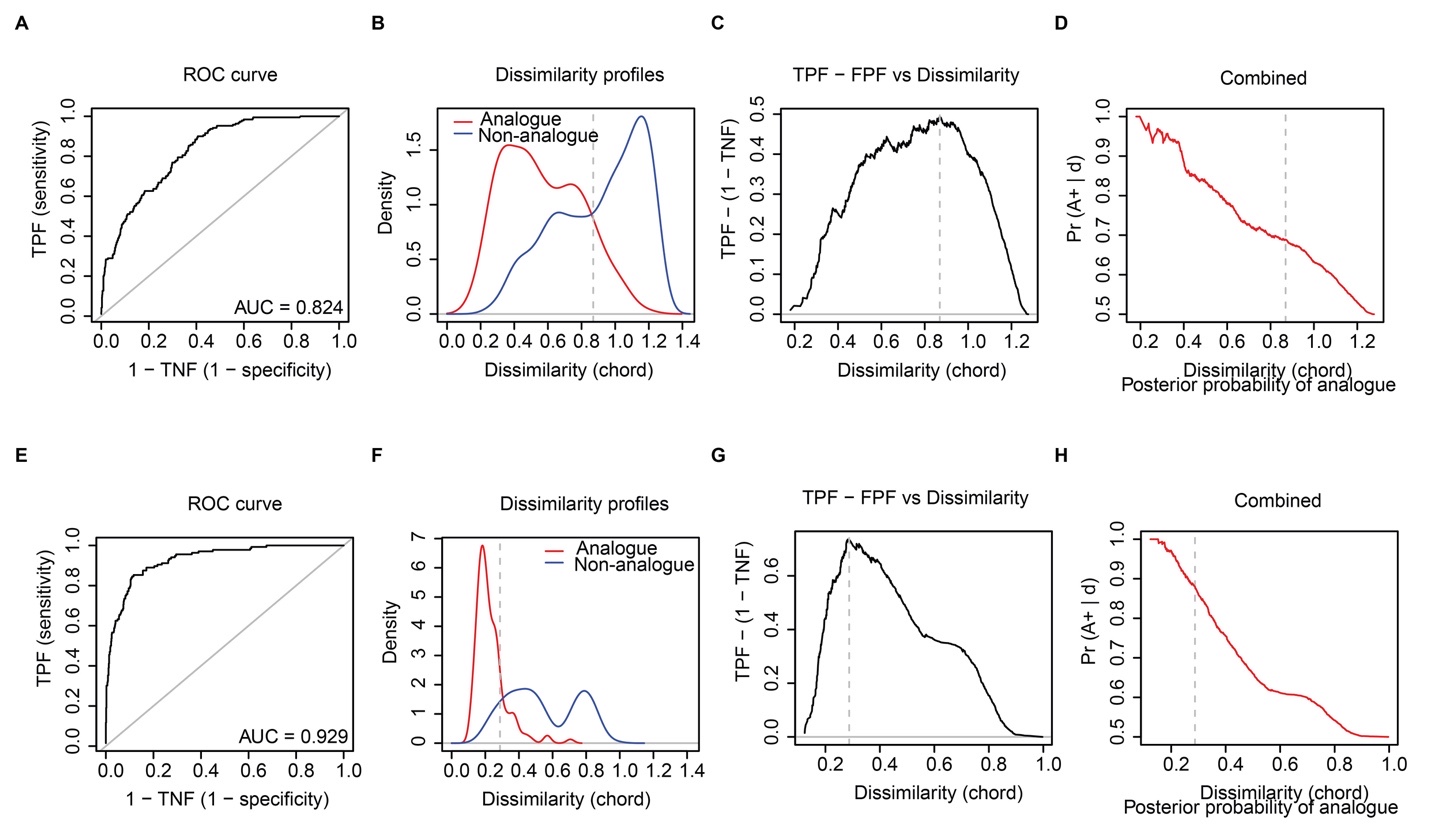
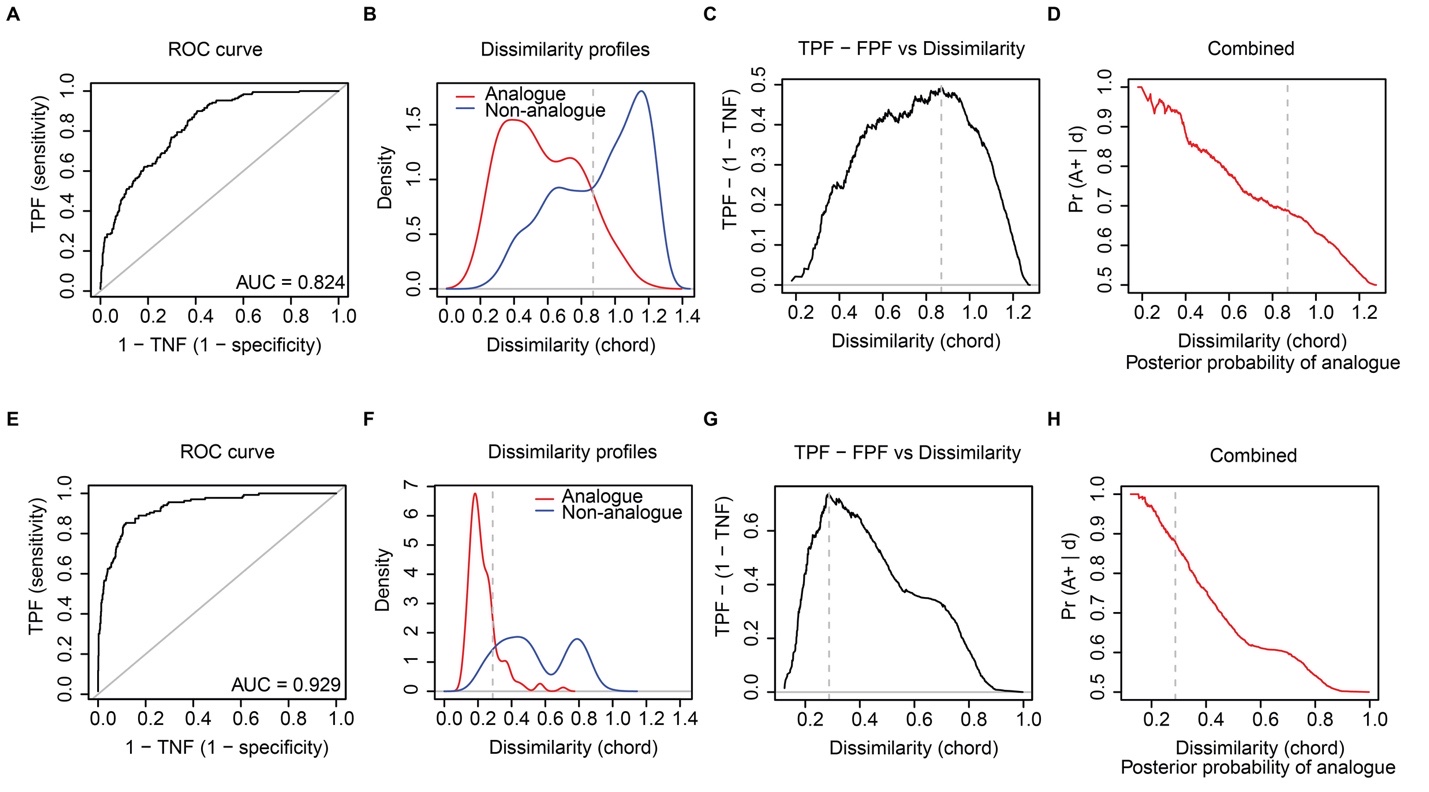
Supplementary Material

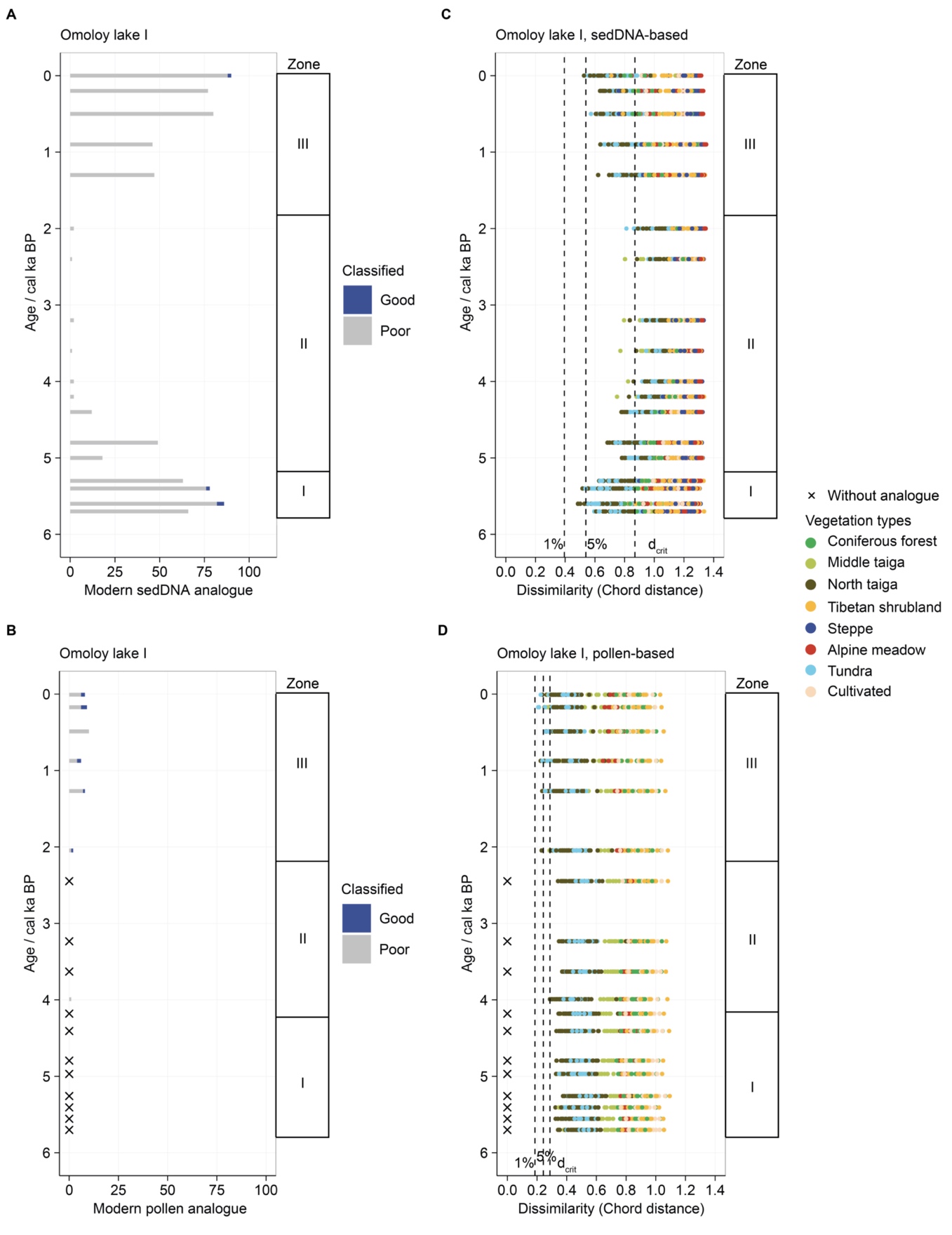
1. **Supplementary Figures**

****

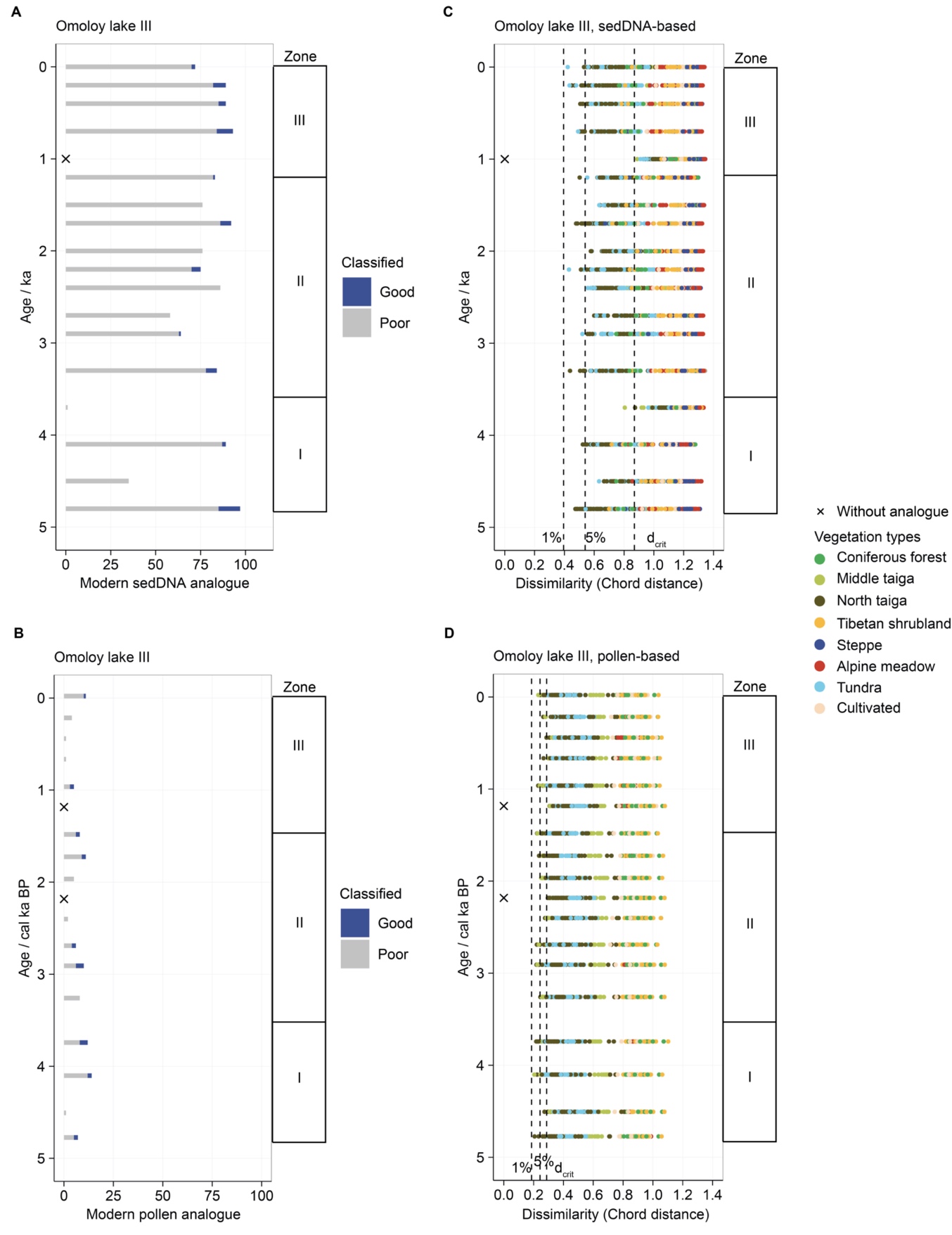
**Supplementary Figure 1** | Receiver operating characteristic (ROC) curve analysis with chord distance applied to the log(x+1) transformed modern sedDNA and modern pollen data for Omoloy lake I.The results indicate that vegetation types can be well discriminated based on modern sedDNA (A-D) and modern pollen assemblages (E-H). The ROC curve and the area under curve (AUC) value assess the optimal dissimilarity threshold (*dcrit*) (A and E). Dissimilarity curves show the kernel density estimates of the distribution of pair-wise dissimilarities for analogue and non-analogue samples (B and F). The sensitivity (true positive fraction: TPF) and the specificity (true negative fraction: TNF) against dissimilarity suggest a good performance of ROC curve analysis (C and G). The posterior probability that any two samples are analogues is calculated based on TPF and false positive fraction (FPF) (D and H). Vertical dashed lines mark *dcrit*. All curves are based on the results of combined vegetation types.



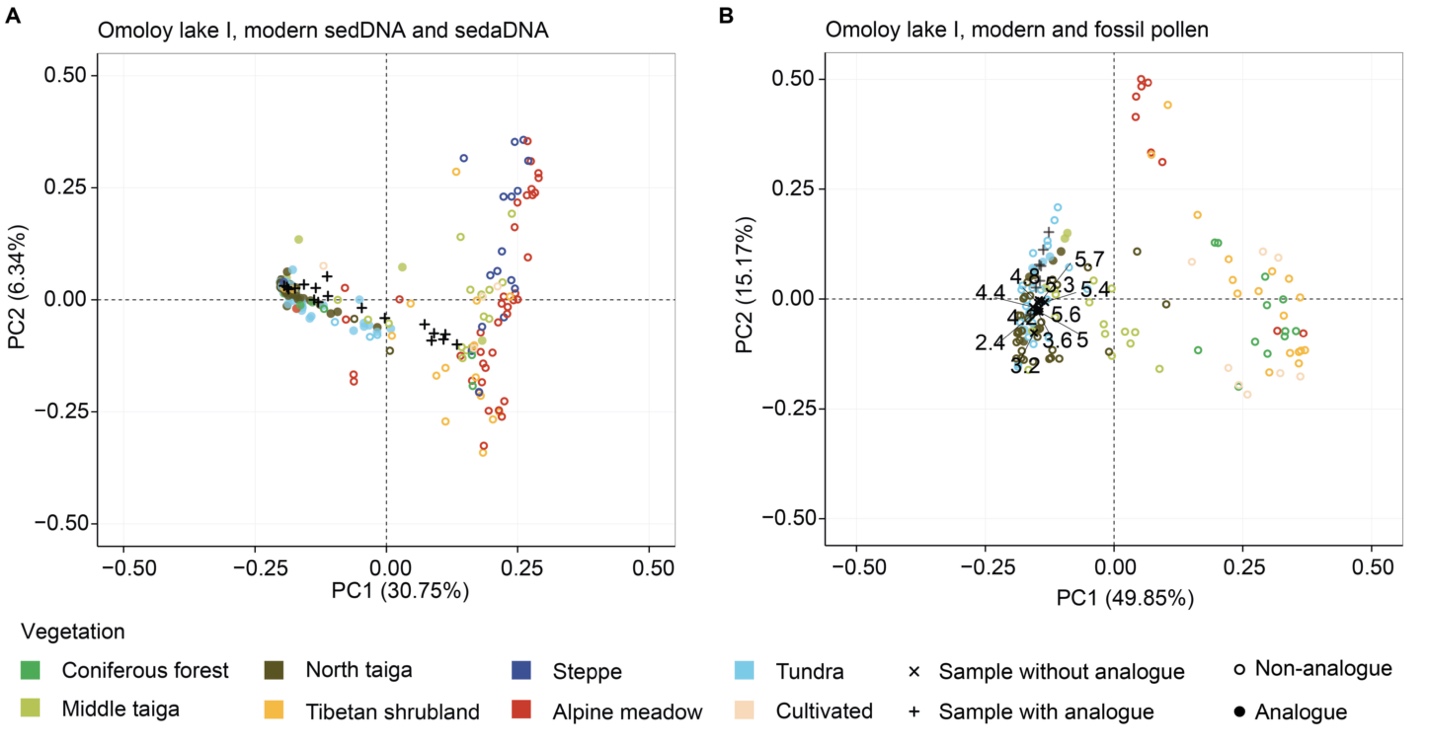
**Supplementary Figure 2** | Receiver operating characteristic (ROC) curve analysis with chord distance applied to the log(x+1) transformed modern sedDNA and modern pollen data for Omoloy lake III. The results indicate that vegetation types can be well discriminated based on modern sedDNA (A-D) and modern pollen assemblages (E-H). The ROC curve and the area under curve (AUC) value assess the optimal dissimilarity threshold (*dcrit*) (A and E). Dissimilarity curves show the kernel density estimates of the distribution of pair-wise dissimilarities for analogue and non-analogue samples (B and F). The sensitivity (true positive fraction: TPF) and the specificity (true negative fraction: TNF) against dissimilarity suggest a good performance of ROC curve analysis (C and G). The posterior probability that any two samples are analogues is calculated based on TPF and false positive fraction (FPF) (D and H). Vertical dashed lines mark *dcrit*. All curves are based on the results of combined vegetation types.



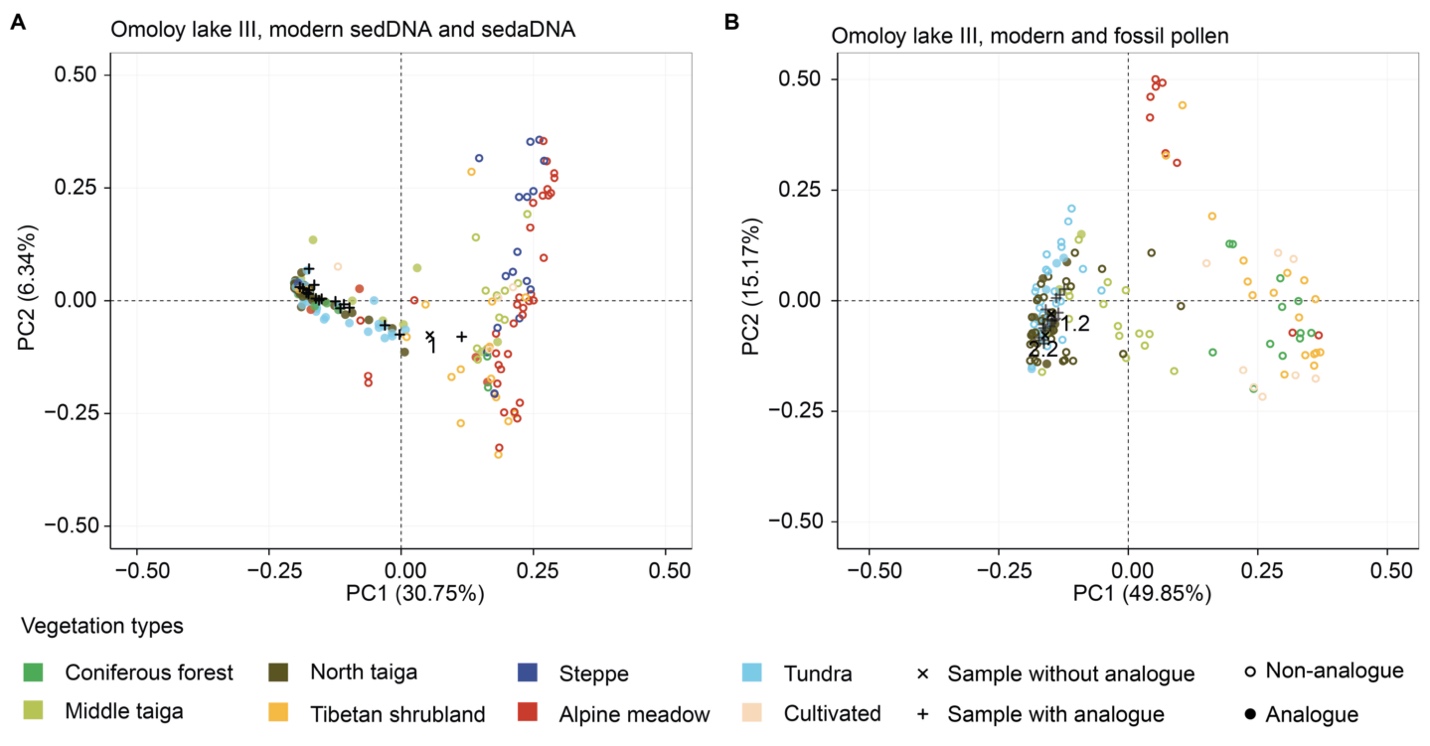
**Supplementary Figure 3 |** Quality and number of the modern analogues (left) and reconstructed vegetation types (right) for Omoloy lake I based on sedDNA (top) and pollen (bottom). The number of modern analogues is estimated via receiver operating characteristic (ROC) curve analysis. The modern analogues are classified as close (<1% percentile), good (1–5%), or poor (>5%). Fossil sediment samples without modern analogues are marked with an x. The point plots illustrate the reconstructed vegetation types based on sedDNA (C) and pollen data (D). Vertical dashed lines mark dissimilarity of 1%, 5%, and the optimal dissimilarity threshold (*dcrit*).



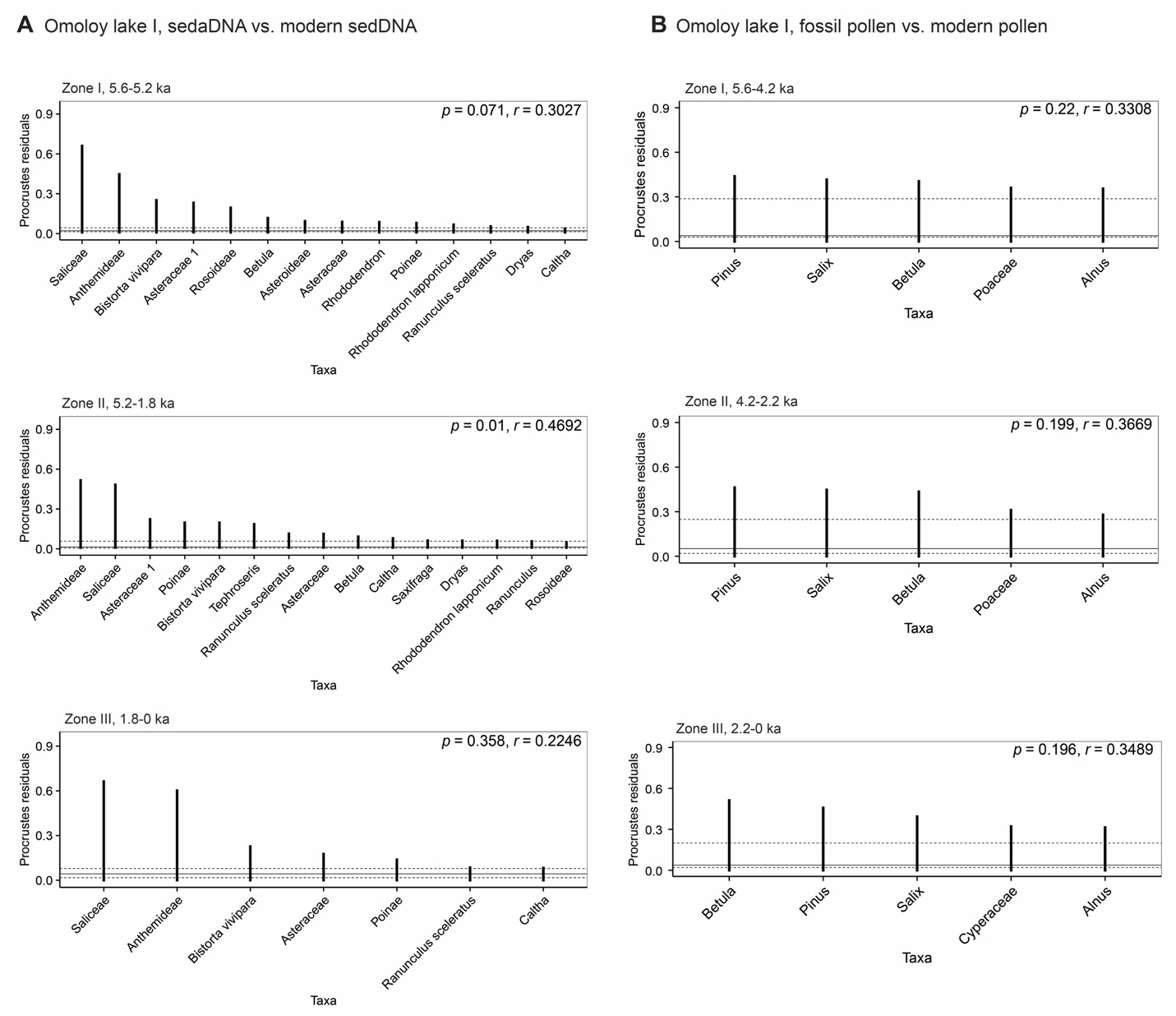
**Supplementary Figure 4 |** Quality and number of the modern analogues (left) and reconstructed vegetation types (right) for Omoloy lake III based on sedDNA (top) and pollen (bottom). The number of modern analogues is estimated via receiver operating characteristic (ROC) curve analysis. The modern analogues are classified as close (<1% percentile), good (1–5%), or poor (>5%). Fossil sediment samples without modern analogues are marked with an x. The point plots illustrate the reconstructed vegetation types based on sedDNA (C) and pollen data (D). Vertical dashed lines mark dissimilarity of 1%, 5%, and the optimal dissimilarity threshold (*dcrit*).



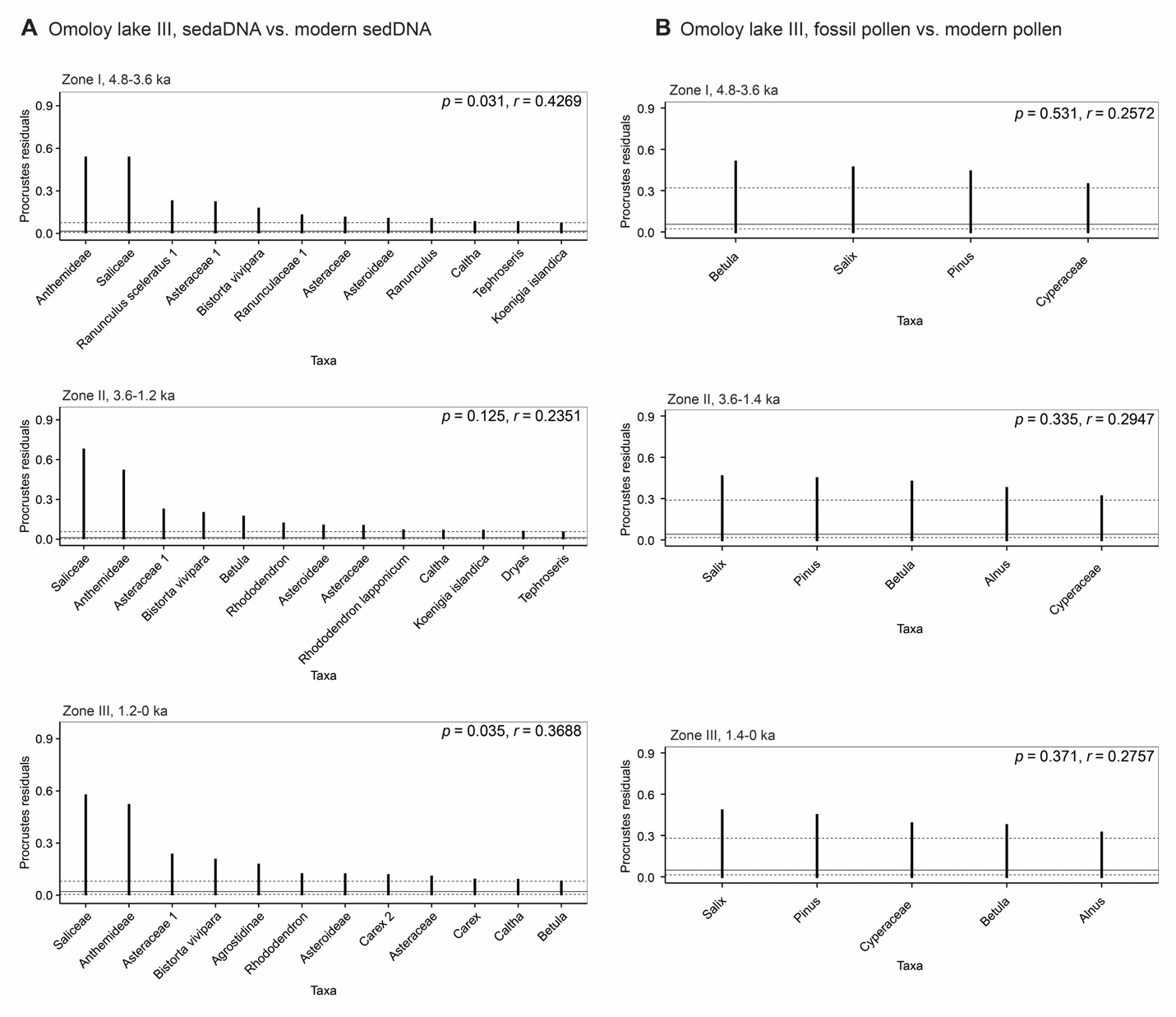
**Supplementary Figure 5 |** Plot of the principal component analysis (PCA) showing which modern samples (coloured according to vegetation type) have been matched to sediment assemblages (black) of Omoloy lake I. (A) PCA site scores explain 37.09% of the total variance of the log(1+x) chord transformed modern sedDNA training-set. (B) PCA site scores explain 65.02% of the total variance of the log-chord transformed modern pollen training-set.



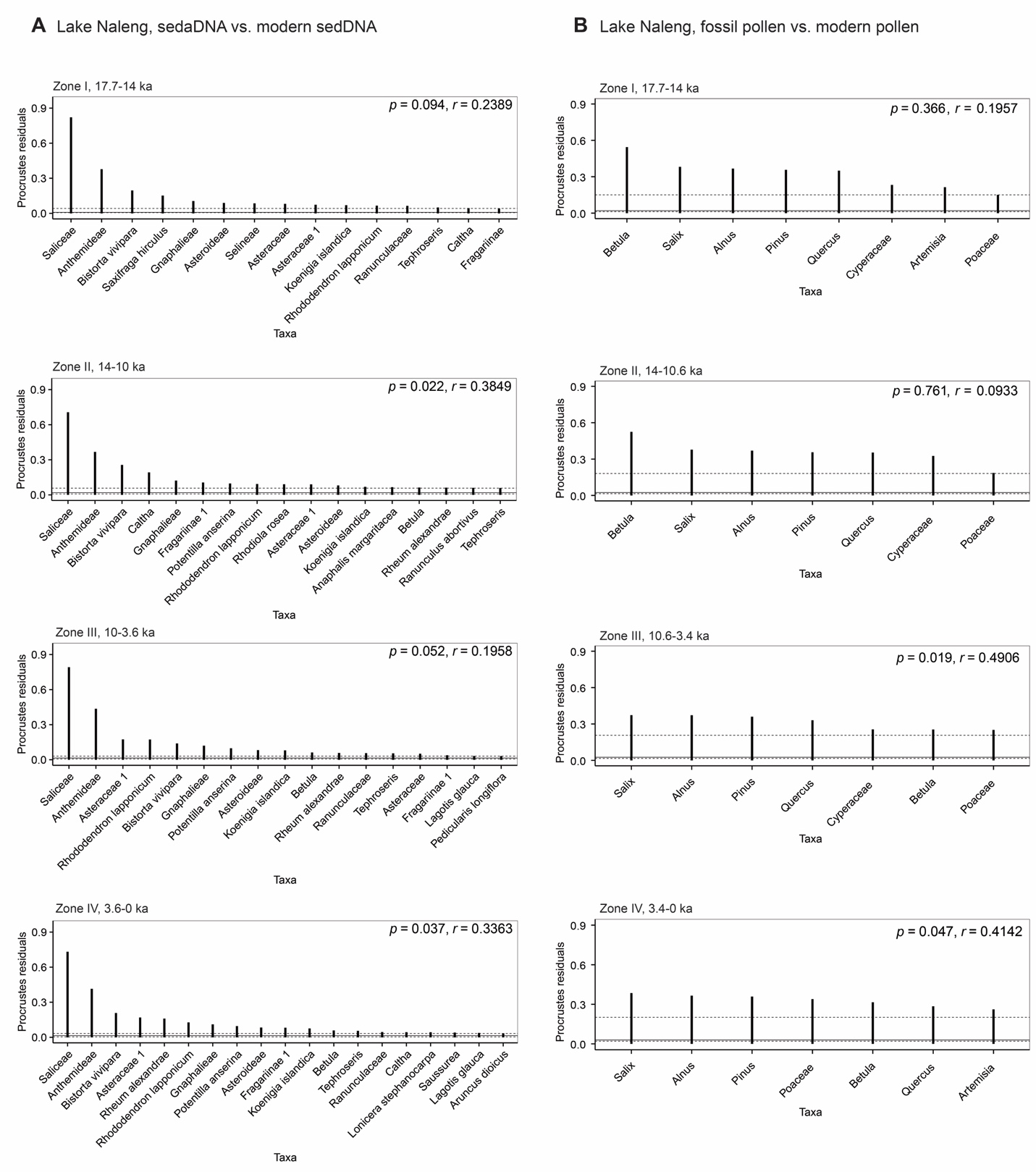
**Supplementary Figure 6 |** Plot of the principal component analysis (PCA) showing which modern assemblages (coloured according to vegetation type) have been matched to sediment assemblages (black) of Omoloy lake III. (A) PCA site scores explain 37.09% of the total variance of the log(1+x)-chord transformed modern sedDNA training-set. (B) PCA site scores explain 65.02% of the total variance of the log-chord transformed modern pollen training-set.



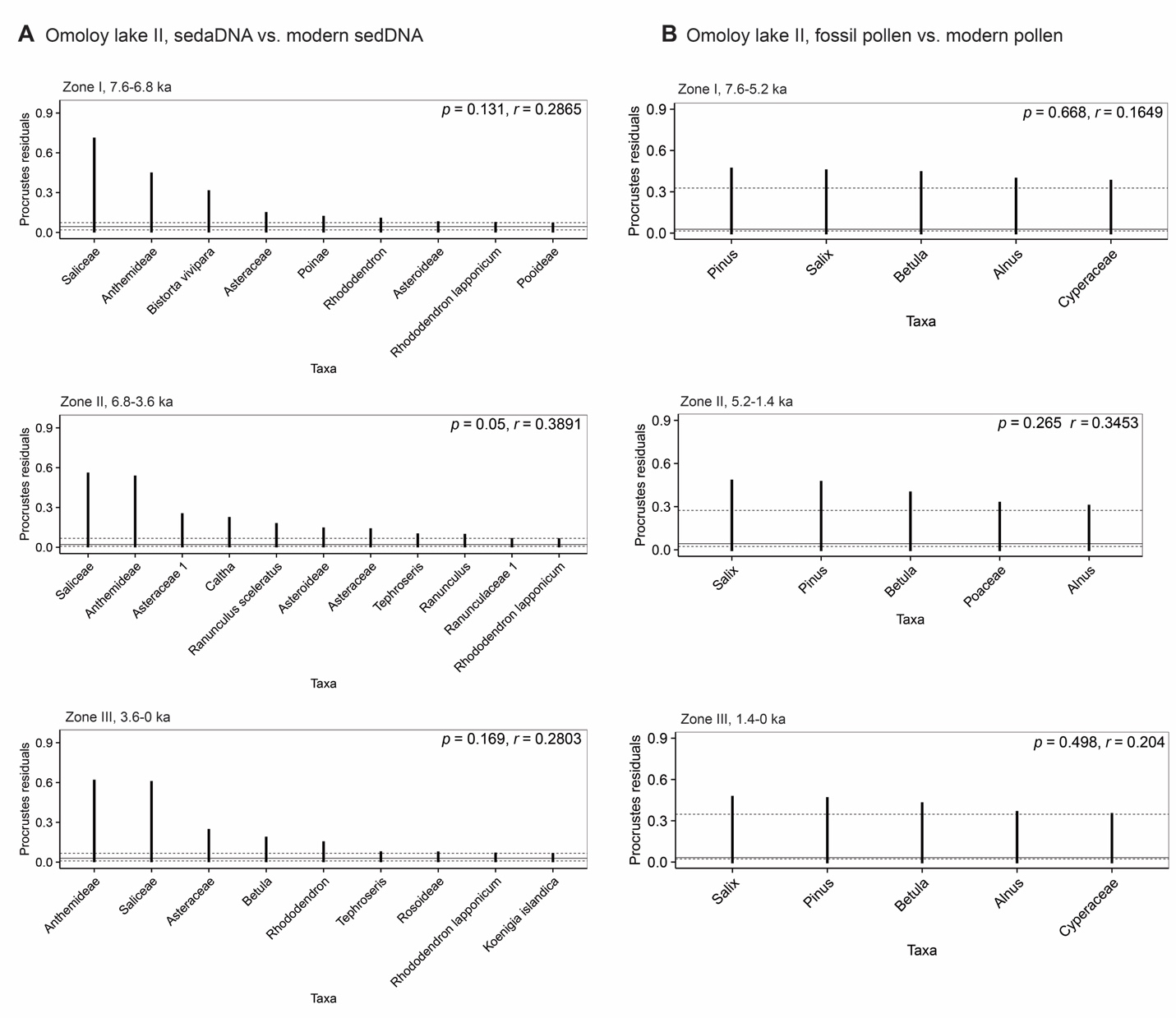
**Supplementary Figure 7 |** Procrustes analysis showing the residuals of principal component analysis (PCA) species scores between fossil data and their modern training-set for sedDNA (left) and pollen (right) for Omoloy lake I. Dashed and solid lines are the first, second, and third quartiles. The *p*-value indicates the likelihood of the relationship occurring by chance and *r* is the correlation between the two ordination results by superimposition. Taxa whose residuals are greater than the third quartile are shown.

****

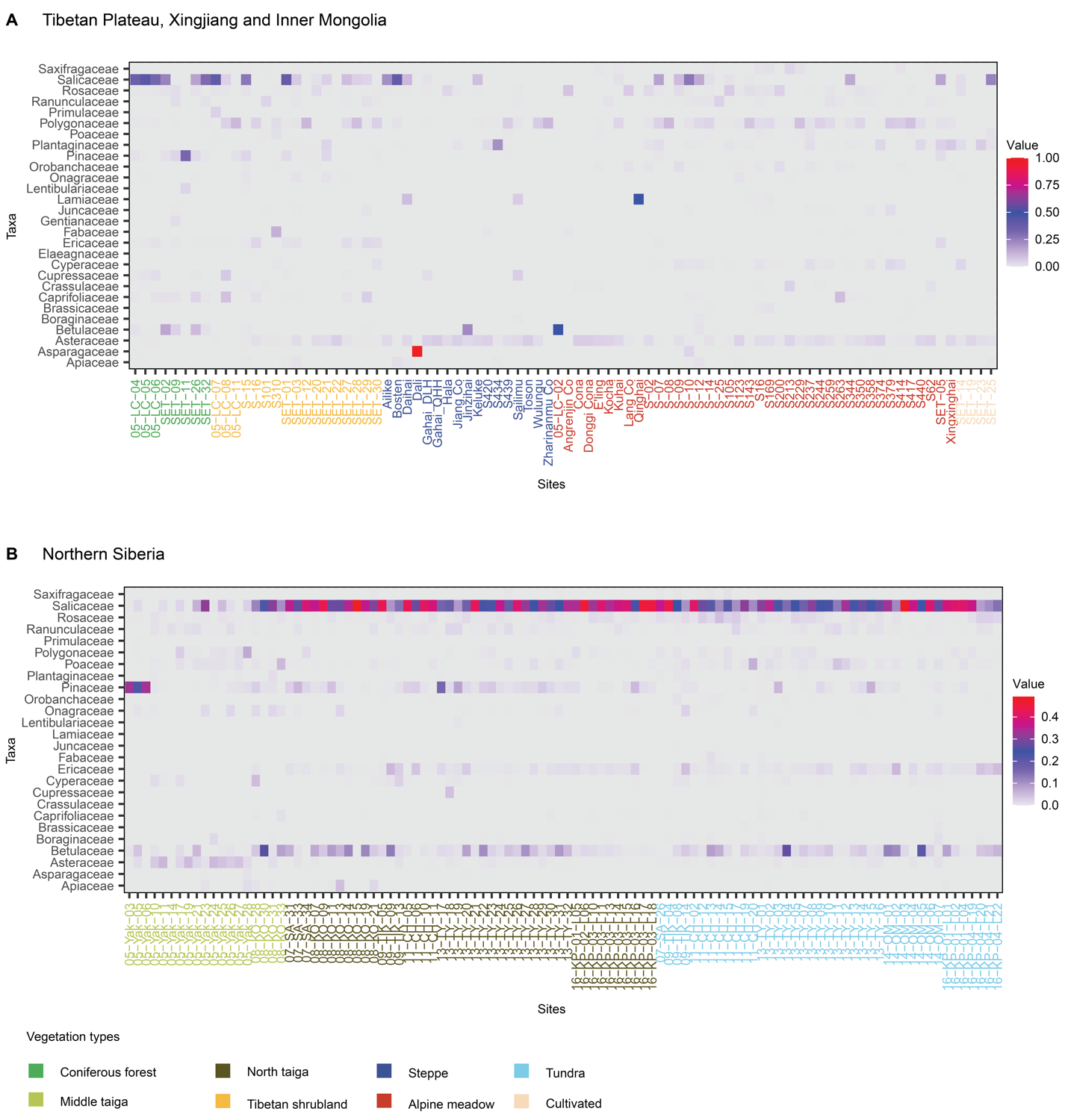
**Supplementary Figure 8 |** Procrustes analysis showing the residuals of principal component analysis (PCA) species scores between fossil data and their modern training-set for sedDNA (left) and pollen (right) for Omoloy lake III. Dashed and solid lines are the first, second, and third quartiles. The *p*-value indicates the likelihood of the relationship occurring by chance and *r* is the correlation between the two ordination results by superimposition. Taxa whose residuals are greater than the third quartile are shown.



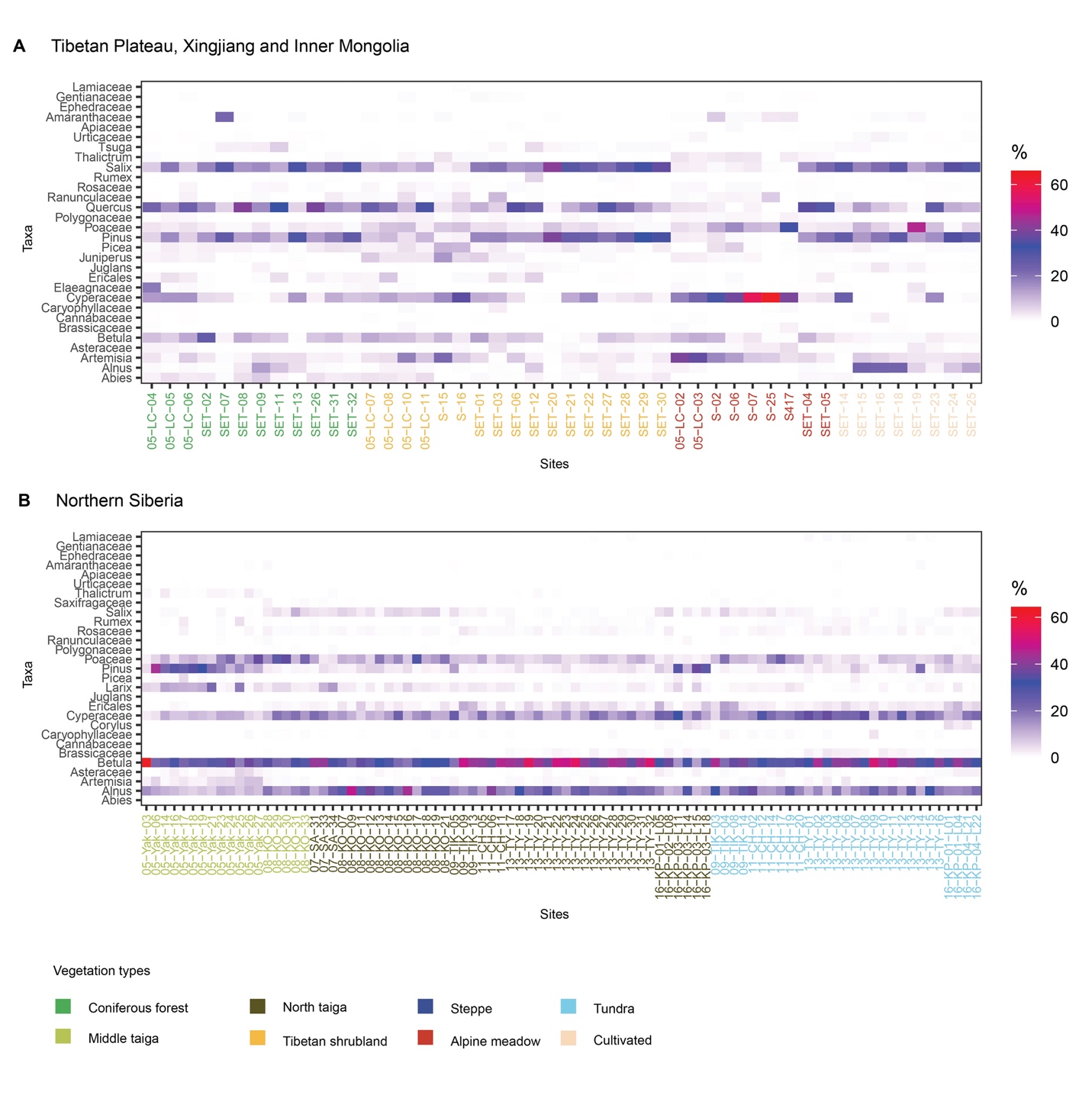
**Supplementary Figure 9 |** Procrustes analysis showing the residuals of principal component analysis (PCA) species scores between fossil data and their modern training-set for sedDNA (left) and pollen (right) for Lake Naleng. Dashed and solid lines are the first, second, and third quartiles. The *p*-value indicates the likelihood of the relationship occurring by chance and *r* is the correlation between the two ordination results by superimposition. Taxa whose residuals are greater than the third quartile are shown.



**Supplementary Figure 10 |** Procrustes analysis showing the residuals of principal component analysis (PCA) species scores between fossil data and their modern training-set for sedDNA (left) and pollen (right) for Omoloy lake II. Dashed and solid lines are the first, second, and third quartiles. The *p*-value indicates the likelihood of the relationship occurring by chance and *r* is the correlation between the two ordination results by superimposition. Taxa whose residuals are greater than the third quartile are shown.



**Supplementary Figure 11 |** Relative read abundance of the modern sedDNA data. The percentages of taxa on the family-level are present. The taxonomic resolution used in statistical analyses is described in Supplementary Data 2.



**Supplementary Figure 12 |** Relative abundance of the modern pollen data. The percentages of taxa on the family-level are present. The taxonomic resolution used in statistical analyses is described in Supplementary Data 2.

1. **Supplementary Tables**

**Supplementary Table 1 | The overview of modern sedDNA and sedaDNA datasets**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Rarefied modern sedDNA data | | | | Rarefied sedaDNA data | | | |
|  | Original | | Selection / max. 2% at least + 5 occurrences | | Original | | Selection / max. 2% at least +  5 occurrences | |
| Core | Sample size | Number of taxa | Sample size | Number of taxa | Sample size | Number of taxa | Sample size | Number of taxa |
| Lake Naleng | 190 | 385 | 190 | 125 | 71 | 153 | 71 | 87 |
| Lake Omoloy I | 190 | 385 | 190 | 117 | 18 | 156 | 18 | 64 |
| Lake Omoloy II | 190 | 385 | 190 | 115 | 18 | 85 | 18 | 51 |
| Lake Omoloy III | 190 | 385 | 190 | 119 | 18 | 110 | 18 | 59 |

**Supplementary Table 2 | The overview of modern and fossil pollen datasets**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Modern pollen data | | | | Fossil pollen data | | | |
|  | Original | | Selection / max. 2% at least + 5 occurrences | | Original | | Selection / max. 2% at least +  5 occurrences | |
| Core | Sample size | Number of taxa | Sample size | Number of taxa | Sample size | Number of taxa | Sample size | Number of taxa |
| Lake Naleng | 136 | 54 | 136 | 33 | 191 | 40 | 191 | 29 |
| Lake Omoloy I | 136 | 54 | 136 | 29 | 18 | 33 | 18 | 20 |
| Lake Omoloy II | 136 | 54 | 136 | 29 | 18 | 34 | 18 | 20 |
| Lake Omoloy III | 136 | 54 | 136 | 29 | 18 | 37 | 18 | 21 |

**Supplementary Table 3 |** Overview of the receiver operating characteristic (ROC) curve analysis of the modern sedDNA and modern pollen training-set corresponding to Omoloy lakes I and III.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Data** | **Vegetation** | **In** | **Out** | **Opt. Dis.** | **AUC** | **SE** | **p-value** |
| Omoloy lake I | | | | | | | |
| sedDNA | Coniferous forest | 8 | 182 | 1.008 | 0.880 | 0.079 | 2.81E-04 |
| sedDNA | Cultivated | 4 | 186 | 0.817 | 0.991 | 0.034 | 8.09E-04 |
| sedDNA | Alpine meadow | 40 | 150 | 0.625 | 0.672 | 0.051 | 8.74E-04 |
| sedDNA | Middle taiga | 19 | 171 | 0.602 | 0.796 | 0.063 | 2.38E-05 |
| sedDNA | North taiga | 44 | 146 | 0.793 | 0.746 | 0.046 | 7.46E-07 |
| sedDNA | Tibetan shrubland | 17 | 173 | 1.033 | 0.598 | 0.076 | 1.82E-01 |
| sedDNA | Steppe | 17 | 173 | 0.813 | 0.724 | 0.072 | 2.39E-03 |
| sedDNA | Tundra | 41 | 149 | 0.504 | 0.914 | 0.031 | 5.05E-16 |
| sedDNA | Combined | 190 | 1330 | 0.870 | 0.824 | 0.019 | 2.07E-47 |
|  |  |  |  |  |  |  |  |
| pollen | Coniferous forest | 12 | 124 | 0.384 | 0.948 | 0.045 | 3.29E-07 |
| pollen | Cultivated | 8 | 128 | 0.270 | 0.996 | 0.016 | 2.68E-06 |
| pollen | Alpine meadow | 9 | 127 | 0.314 | 0.892 | 0.072 | 9.11E-05 |
| pollen | Middle taiga | 18 | 118 | 0.257 | 0.934 | 0.041 | 3.27E-09 |
| pollen | North taiga | 43 | 93 | 0.204 | 0.974 | 0.017 | 7.71E-19 |
| pollen | Tibetan shrubland | 17 | 119 | 0.440 | 0.943 | 0.040 | 3.88E-09 |
| pollen | Tundra | 29 | 107 | 0.282 | 0.934 | 0.033 | 8.59E-13 |
| pollen | Combined | 136 | 816 | 0.287 | 0.929 | 0.015 | 7.77E-58 |
| Omoloy lake III | | | | | | | |
| sedDNA | Coniferous forest | 8 | 182 | 1.008 | 0.880 | 0.079 | 2.81E-04 |
| sedDNA | Cultivated | 4 | 186 | 0.817 | 0.991 | 0.034 | 8.09E-04 |
| sedDNA | Alpine meadow | 40 | 150 | 0.625 | 0.672 | 0.051 | 8.64E-04 |
| sedDNA | Middle taiga | 19 | 171 | 0.602 | 0.796 | 0.063 | 2.38E-05 |
| sedDNA | North taiga | 44 | 146 | 0.793 | 0.749 | 0.046 | 5.95E-07 |
| sedDNA | Tibetan shrubland | 17 | 173 | 1.033 | 0.600 | 0.076 | 1.74E-01 |
| sedDNA | Steppe | 17 | 173 | 0.813 | 0.724 | 0.072 | 2.36E-03 |
| sedDNA | Tundra | 41 | 149 | 0.523 | 0.909 | 0.032 | 1.08E-15 |
| sedDNA | Combined | 190 | 1330 | 0.870 | 0.824 | 0.019 | 2.49E-47 |
|  |  |  |  |  |  |  |  |
| pollen | Coniferous forest | 12 | 124 | 0.384 | 0.948 | 0.045 | 3.29E-07 |
| pollen | Cultivated | 8 | 128 | 0.270 | 0.996 | 0.016 | 2.68E-06 |
| pollen | Alpine meadow | 9 | 127 | 0.314 | 0.892 | 0.072 | 9.11E-05 |
| pollen | Middle taiga | 18 | 118 | 0.257 | 0.934 | 0.041 | 3.27E-09 |
| pollen | North taiga | 43 | 93 | 0.204 | 0.974 | 0.017 | 7.71E-19 |
| pollen | Tibetan shrubland | 17 | 119 | 0.440 | 0.943 | 0.040 | 3.88E-09 |
| pollen | Tundra | 29 | 107 | 0.282 | 0.934 | 0.033 | 8.59E-13 |
| pollen | Combined | 136 | 816 | 0.287 | 0.929 | 0.015 | 7.77E-58 |

Combined: all vegetation types

In: the number of analogues;

Out: the number of non-analogues;

Opt. Dis.: the optimal dissimilarity;

AUC: the area under the ROC curve;

SE: the standard error of AUC;

*p*-value: a Wilcoxon rank

**Supplementary Table 4 |** Results of Procrustes analysis and associated PROTEST indicating the significant sequences/taxa fit between fossil assemblages and modern training-sets for Omoloy lakes I and III

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *p*-value | *r* | m12 | RMSE | Age zone |
| Omoloy lake I | | | | | |
| Modern sedDNA vs. sedaDNA | 0.071 | 0.3027 | 0.9083 | 0.1285 | 5.6-5.2 ka |
| 0.01 | 0.4692 | 0.7799 | 0.117 | 5.2-1.8 ka |
| 0.358 | 0.2246 | 0.9495 | 0.1875 | 1.8-0 ka |
|  |  |  |  |  |  |
| Modern pollen vs. Fossil pollen | 0.22 | 0.3308 | 0.8905 | 0.211 | 5.6-4.2 ka |
| 0.199 | 0.3669 | 0.8654 | 0.2134 | 4.2-2.2 ka |
| 0.196 | 0.3489 | 0.8783 | 0.2096 | 2.2-0 ka |
|  |  |  |  |  |  |
| Omoloy lake III | | | | | |
| Modern sedDNA vs. sedaDNA | 0.031 | 0.4269 | 0.8178 | 0.1348 | 4.8-3.6 ka |
| 0.125 | 0.2351 | 0.9447 | 0.1375 | 3.6-1.2 ka |
| 0.035 | 0.3688 | 0.864 | 0.1342 | 1.2-0 ka |
|  |  |  |  |  |  |
| Modern pollen vs. Fossil pollen | 0.531 | 0.2572 | 0.9338 | 0.2416 | 4.8-3.6 ka |
| 0.335 | 0.2947 | 0.9131 | 0.2252 | 3.6-1.4 ka |
| 0.371 | 0.2757 | 0.924 | 0.2205 | 1.4-0 ka |

*p*-value: Classical t-test statistic

*r*: Correlation between the two ordination results

m12: Procrustes rotation sum of squares

RMSE: Root mean square error

1. **Supplementary Codes**

**Supplementary Code 1: Analogue matching**

# MAT for sedDNA-sedaDNA / modern-fossil pollen

# Here is an example for sedDNA-sedaDNA.

#

library(readxl)

library(readr)

library(reshape2)

library(data.table)

library(analogue)

library(dplyr)

library(tidyr)

library(forcats) # "fct\_inorder"

# set path

setwd("~/")

# modern vegetation information

# modern sites: 190

# a subset of SI1: Environmental data of lake localities from https://doi.org/10.3897/BDJ.8.e57089

#

veg=as.data.frame(read\_excel("Supplementary-Data1.xlsx", sheet = "sedDNA-sites"))

# sedDNA data, the rarefied dataset

# modern sites: 190

# the rarefied dataset, the raw data was downloaded from https://doi.org/10.3897/BDJ.8.e57089, SI3: Taxa list of # identified plant sequences and occurrences

finaleper=as.data.frame(read\_excel("Supplementary-Data2.xlsx", sheet = "sedDNA-rarefied counts-m"))

## set the rownames

rownames(finaleper)=finaleper$sequence.x

## set the frame

mdna=finaleper[-c(1:3)]

mdna=as.data.frame(t(mdna)) # sites x species

# check if total count per site = base count (1000)

rowSums(mdna)

## core name

core=c("NC", "OM12A", "OM02B", "OM20B")

# data type

dataT="DNA\_roc"

# option: 2% + 5 times

opt="p002-s5"

# transformation

trans="log1p"

# MAT and ROC

# sedaDNA data

# Omoloy lakes: https://doi.org/10.3389/fevo.2020.560243

# Naleng (NC): in review

for (corei in core) {

# Prepaer aDNA

print(corei)

pdata=read.csv(paste0(corei, "\_c10\_rarefied\_mean\_specieslevel\_C10\_05\_18.csv"), stringsAsFactors = FALSE, check.names = FALSE)

sdata=pdata[-c(2:3)]

rownames(sdata)=sdata$sequence.x

sdata$sequence.x=NULL

## remove sequence = 0

sdata=sdata[apply(sdata, 1, function(x) !all(x==0)),]

sdata=as.data.frame(t(sdata))

sdata$age=as.numeric(rownames(sdata))

sdata=sdata[order(sdata$age, decreasing = TRUE), ]

rownames(sdata)=sdata$age

sdata$age=NULL

# data%, remove NaN

# adna=sdata[complete.cases(sdata), ] # ages x seq

## store ages

adna=sdata

age=as.numeric(rownames(adna))

# Calculate %

## sample/age x species

mdna.per=prop.table(as.matrix(mdna), 1)

adna.per=prop.table(as.matrix(adna), 1)

## check if

rowSums(mdna.per) # should be 1

rowSums(adna.per) # should be 1

# Filter

mdna.per=as.data.frame(mdna.per)

adna.per=as.data.frame(adna.per)

## combine modern and fossil data

m.f.data=analogue::join(mdna.per, adna.per, verbose = TRUE)

mrd=m.f.data[[1]]

frd=m.f.data[[2]]

m.f.rd=rbind(mrd, frd)

## max >= 2% and times >= 5

max.abb <- apply(m.f.rd, 2, max)

n.occ <- colSums(m.f.rd > 0)

spp.want <- which(max.abb >= 0.02 & n.occ >= 5)

## select spp.want from modern and fossil data

mrd.want=mrd[, spp.want]

frd.want=frd[, spp.want]

## save (SI Table)

write.csv(mrd.want, paste0(corei, "\_", dataT, "\_", opt, "mDNA\_relabun.csv"))

write.csv(frd.want, paste0(corei, "\_", dataT, "\_", opt, "aDNA\_relabun.csv"))

# Transformation, log(1+x)

mrd.ln <- log1p(mrd.want)

frd.ln <- log1p(frd.want)

# Analog matching

## combine data

m.f.ln=analogue::join(mrd.ln, frd.ln, verbose = TRUE)

mrd.ln=m.f.ln[[1]]

frd.ln=m.f.ln[[2]]

## chord distance

core.analog <- analog(mrd.ln, frd.ln, method = "chord",

keep.train = T, na.rm=T)

## store analog results

outfile=summary(core.analog)

## store the analog

core.analogues=as.data.frame(core.analog[["analogs"]])

core.train=as.data.frame(core.analog[["train"]])

## output the data

write.csv(core.analogues, paste0("analogs\_", corei,"\_", dataT, "\_",opt,"\_",trans,".csv"))

write.csv(core.train, paste0("train\_", corei,"\_", dataT, "\_",opt,"\_",trans,".csv"))

# Assign the vegetation types to analog

vegT=veg[c("ana.Site", "Vegetation\_type")]

names(vegT)=c("ana.Site", "VEG\_TYPE")

vegT=vegT[vegT$ana.Site %in% rownames(mrd.ln), ]

## format vegT

vegType=as.vector(vegT$VEG\_TYPE)

names(vegType)=as.vector(vegT$ana.Site)

# Calcualte the dcrit values for each vegetation type and combined vegetation types

core.analog.roc <- roc(core.analog, groups = vegType, method = "chord")

## make table

core.analog.roc.sta=core.analog.roc$statistics

core.analog.roc.sta$Dtype=dataT

## output

write.csv(core.analog.roc.sta, paste0(corei,"\_", dataT, "\_ROC\_statistic\_", opt, "\_",trans,".csv"))

# Plot ROC curve

pdf(file = paste0(corei,"\_", dataT, "\_ROC\_curves\_", opt, "\_",trans, ".pdf"), width = 8, height = 8)

opar <- par(mfrow = c(2,2))

plot(core.analog.roc, abline.col="gray", inGroup.col = "red", outGroup.col = "blue")

par(opar)

dev.off()

# Find the cutoff value

dcrit=core.analog.roc$roc[["Combined"]]$optimal

cutoff.value=dcrit

# check the analog quality

## plot

core.mdc=minDC(core.analog, probs = c(0.01, 0.02, 0.05, 0.1))

pdf(file = paste0(corei, "\_mdc\_r1000\_", opt,"\_",trans, ".pdf"), width=6,height=4)

plot(core.mdc, use.labels = TRUE, xlab = "Age / cal yr BP", main = " ", cex.axis=.5, cex.lab=.5)

title(main = list(paste0("Percentiles of the dissimilarities: ", corei, " sedaDNA vs. sedDNA training-set"), cex = .5,

col = "black", font = .5))

dev.off()

## make a table

core.mdc.sta=as.data.frame(core.mdc$minDC)

names(core.mdc.sta)="Dissimilarity"

write.csv(core.mdc.sta, paste0(corei,"\_", dataT,"\_mdc\_", opt, "\_",trans,".csv"))

core.mdc.quantiles=as.data.frame(core.mdc$quantiles)

names(core.mdc.quantiles)="quantiles"

core.mdc.quantiles$core=corei

write.csv(core.mdc.quantiles, paste0(corei,"\_", dataT,"\_mdc\_quantiles\_", opt, "\_",trans,".csv"))

# Output: modern analogues

core.cma <- cma(core.analog, cutoff = cutoff.value, prob = c(0.01, 0.025, 0.05, 0.1))

closeSites.core.ana=core.cma[["close"]] # list

# output the modern analog

core.age=as.character(age)

buff.closeSites=NULL

for (agei in core.age) {

print(agei)

if (length(closeSites.core.ana[[agei]])>0) {

buff.closeSites=rbind(buff.closeSites, data.frame(T=agei,ana.Site=names(closeSites.core.ana[[agei]]),

ana.value=closeSites.core.ana[[agei]],

num.ana=length(closeSites.core.ana[[agei]])))

} else {

next

}

}

if (is.null(buff.closeSites) != TRUE) {

rownames(buff.closeSites)=rep(1:dim(buff.closeSites)[1])

# add veg. type

core.veg=merge(buff.closeSites, veg, by="ana.Site")

core.veg=core.veg[order(core.veg$T, decreasing = TRUE), ]

# for mapping and plot veg.

write.csv(buff.closeSites, paste0(corei, "\_", dataT, "\_modernAnalog\_Veg\_", opt, "\_",trans,".csv"))

}

}

**Supplementary Code 2: Ordination**

# Ordination

# sedDNA-sedaDNA / modern-fossil pollen

# Here is an example for sedDNA-sedaDNA.

library(readr)

library(data.table)

library(analogue)

library(adespatial)

library(dplyr)

library(reshape2)

library(sf)

library(maps)

library(rnaturalearth)

library(rnaturalearthdata) # world

library(ggspatial) # annotation\_

library(ggrepel) # needed for geom\_text\_repel()

library(gridExtra)

library(grDevices)

library(grid)

library(maptools)

library(rgdal)

library(forcats) # "fct\_inorder"

library(RColorBrewer) # colorbar

library(easyGgplot2)

library(ggtext)

# set path

setwd("~/")

# modern vegetation information

# modern sites: 190

# a subset of SI1: Environmental data of lake localities from https://doi.org/10.3897/BDJ.8.e57089

#

veg=as.data.frame(read\_excel("Supplementary-Data1.xlsx", sheet = "sedDNA-sites"))

# sedDNA data, the rarefied dataset

# modern sites: 190

# the rarefied dataset, the raw data was downloaded from https://doi.org/10.3897/BDJ.8.e57089, SI3: Taxa list of identified plant sequences and occurrences

finaleper=as.data.frame(read\_excel("Supplementary-Data2.xlsx", sheet = " Supplementary-Data2.xlsx "))

## set the rownames

rownames(finaleper)=finaleper$sequence.x

## set the frame

mdna=finaleper[-c(1:3)]

mdna=as.data.frame(t(mdna)) # sites x species

# check if total count per site = base count (1000)

rowSums(mdna)

## core name

core=c("NC", "OM12A", "OM02B", "OM20B")

# data type

dataT="DNA\_roc"

# option: 2% + 5 times

opt="p002-s5"

# transformation

trans="log1p"

# Forward selection

PCApre="PCAforward"

# sedaDNA data

# Omoloy lakes: https://doi.org/10.3389/fevo.2020.560243

# Naleng (NC): in review

##################

# ggplot colour, shapes, theme, codes in final position

##################

# prepare the colour

mycols=c(

# blue

rgb(61, 80, 147, maxColorValue = 255),

# red

rgb(199, 58, 43, maxColorValue = 255),

# orange

rgb(243, 186, 68, maxColorValue = 255),

# dark green

rgb(88, 83, 34, maxColorValue = 255),

# grass green

rgb(184, 196, 83, maxColorValue = 255),

# light green

rgb(79, 171, 88, maxColorValue = 255),

# light blue

rgb(126, 201, 229, maxColorValue = 255),

# yellow

rgb(245, 218, 186, maxColorValue = 255),

# purple

rgb(145, 140, 203, maxColorValue = 255),

# analogue quality

rgb(19, 49, 88, maxColorValue = 255),

rgb(54, 100, 168, maxColorValue = 255),

# < 1%

rgb(119, 186, 199, maxColorValue = 255),

# 1-5%

rgb(218, 214, 108, maxColorValue = 255),

# >5%

rgb(194, 194, 194, maxColorValue = 255)

)

cols.ana = c("Close" = mycols[2], "Good" = mycols[1], "Poor" = mycols[14])

# prepare the theme

my.theme <- theme(# Title

plot.title = element\_text(#margin=margin(b=-20),

hjust = 0,

size = 8, family = "Helvetica", colour = "black"),

#aspect.ratio=4/2,

plot.margin = unit(c(0.4,0.4,0.4,0.4), "cm"),

panel.border = element\_rect(colour = "black", fill=NA, size = 0.11),

# y axis right

axis.line.y.right = element\_line(size = 0.233, colour = "black"),

axis.ticks.y.right = element\_blank(),

axis.ticks.length.y.right = element\_blank(),

axis.text.y.right = element\_blank(),

axis.title.y.right = element\_blank(),

# y axis left

axis.line.y.left = element\_line(size = 0.233),

axis.ticks.y.left = element\_line(size = 0.233, colour = "black"),

axis.ticks.length.y.left = unit(.10, "cm"),

axis.text.y.left = element\_text(size = 8, family = "Helvetica", colour = "black"),

axis.title.y.left = element\_text(size = 8, family = "Helvetica", colour = "black"),

# x axis top

axis.line.x.top = element\_line(size = 0.233),

axis.ticks.x.top = element\_blank(),

axis.text.x.top = element\_blank(),

axis.title.x.top = element\_blank(),

# x axis bottom

axis.line.x.bottom = element\_line(size = 0.233),

axis.ticks.x.bottom = element\_line(size = 0.233, colour = "black"),

axis.ticks.length.x.bottom = unit(.10, "cm"),

axis.text.x.bottom = element\_text(size = 8, family = "Helvetica", colour = "black"),

axis.title.x.bottom = element\_text(size = 8, family = "Helvetica", colour = "black"),

# legend

legend.position = "right",

#legend.background = element\_rect(color = "black", fill = "grey90", size = 1, linetype = "solid"),

legend.text = element\_text(size = 8, family = "Helvetica"),

legend.title = element\_text(size = 8, family = "Helvetica"),

# grid

#panel.grid = element\_blank()

panel.grid.major = element\_line(size = 0.1),

panel.grid.minor.y = element\_blank(),

panel.grid.minor.x = element\_blank()

)

cols=c("Coniferous forest" = mycols[6],

"Middle taiga" = mycols[5],

"North taiga" = mycols[4],

######### shrub

"Tibetan shrubland" = mycols[3],

######## steppe

"Steppe" = mycols[1],

######## Alpine meadow

"Alpine meadow" = mycols[2],

# tundra

"Tundra" = mycols[7],

# others

"Cultivated" = mycols[8]

)

cols.fill=c("Coniferous forest" = mycols[6],

"Middle taiga" = mycols[5],

"North taiga" = mycols[4],

######### shrub

"Tibetan shrubland" = mycols[3],

######## steppe

"Steppe" = mycols[1],

######## Alpine meadow

"Alpine meadow" = mycols[2],

# tundra

"Tundra" = mycols[7],

# others

"Cultivated" = mycols[8]

)

#

# Ordination

for (corei in core) {

print(corei)

# fossil composition

cdata=read.csv(paste0(corei, "\_c10\_rarefied\_mean\_specieslevel\_C10\_05\_18.csv"), stringsAsFactors = FALSE, check.names = FALSE)

tdata=cdata[-c(2:3)]

rownames(tdata)=tdata$sequence.x

tdata$sequence.x=NULL

tdata=as.data.frame(t(tdata))

tdata$age=as.numeric(rownames(tdata))

tdata=tdata[order(tdata$age, decreasing = TRUE), ]

rownames(tdata)=tdata$age

tdata$age=NULL

# buff, fossil data

fdata=tdata

# Calculate %

## sample/age x species

mrdata=prop.table(as.matrix(mdna), 1)

frdata=prop.table(as.matrix(fdata), 1)

## check if

rowSums(mrdata) # should be 1

rowSums(frdata) # should be 1

# data frame

mrdata=as.data.frame(mrdata)

frdata=as.data.frame(frdata)

# joint data

m.f.data=analogue::join(mrdata, frdata, verbose = TRUE)

#

mrd=m.f.data[[1]]

frd=m.f.data[[2]]

m.f.rd=rbind(mrd, frd)

# max > 2% and present in 5 sites

max.abb <- apply(m.f.rd, 2, max)

n.occ <- colSums(m.f.rd > 0)

spp.want <- which(max.abb > 0.02 & n.occ >= 5)

#

mrd.p002.s5=mrd[, spp.want]

frd.p002.s5=frd[, spp.want]

#

mrd.st=mrd.p002.s5

frd.st=frd.p002.s5

# matrix

mrd.st=as.matrix(mrd.st)

frd.st=as.matrix(frd.st)

#

mrd.st.ln <- log1p(mrd.st)

frd.st.ln <- log1p(frd.st)

#

mrd.st.ln.norm <- decostand(mrd.st.ln, "norm")

frd.st.ln.norm <- decostand(frd.st.ln, "norm")

mrd.st.ln.norm=as.data.frame(mrd.st.ln.norm)

frd.st.ln.norm=as.data.frame(frd.st.ln.norm)

# joint data

m.f.bcbox.data=analogue::join(mrd.st.ln.norm, frd.st.ln.norm, verbose = TRUE)

#

mrd.bcbox=m.f.bcbox.data[[1]]

frd.bcbox=m.f.bcbox.data[[2]]

# PCA for log-chord transformed training-set

mrd.box.cox.chord.pca=rda(mrd.bcbox)

#

mrd.box.cox.chord.pca.sum=summary(mrd.box.cox.chord.pca)

pro.ex=as.data.frame(mrd.box.cox.chord.pca.sum[["cont"]][["importance"]])

pro.ex.pc1=round(pro.ex$PC1[2]\*100, digits = 2)

pro.ex.pc2=round(pro.ex$PC2[2]\*100, digits = 2)

# process the fossil data with the predict.PCA

f.m.pred=as.data.frame(predict(mrd.box.cox.chord.pca, newdata=frd.bcbox, type = "wa", scaling = 3, model = "CA"))

# data1

f.m.pred.pc12=f.m.pred[c("PC1", "PC2")]

f.m.pred.pc12$age=rownames(f.m.pred.pc12)

f.m.pred.pc12$age=as.numeric(f.m.pred.pc12$age)

write.csv(f.m.pred.pc12, paste0("fPCA\_", corei,"\_", dataT, "\_",opt,"\_",trans,".csv"))

# summary

perta.pca=mrd.box.cox.chord.pca

perta.pca.site <- scores(perta.pca, display = "sites", scaling = 3)

perta.pca.spp <- scores(perta.pca, display = "species", scaling = 3)

# output the PCA data

m.site.pc=as.data.frame(perta.pca.site)

m.sp.pc=as.data.frame(perta.pca.spp)

m.sp.pc$sequence.x=rownames(m.sp.pc)

m.sp.pc.m=merge(m.sp.pc, sp.rarefied.full[1:3], by = "sequence.x")

m.sp.pc.f=merge(m.sp.pc, cdata[1:3], by = "sequence.x")

names(m.sp.pc.f)=c("sequence.x", "PC1", "PC2", "Species\_R", "Family\_R")

m.sp.pc.name=rbind(m.sp.pc.m, m.sp.pc.f)

# remove duplicated

deduped.m.sp.pc <- unique(m.sp.pc.name[, 1:5] )

rownames(deduped.m.sp.pc)=make.names(deduped.m.sp.pc$Species\_R, unique = T)

deduped.m.sp.pc$taxa=rownames(deduped.m.sp.pc)

# add veg

m.site.pc$ana.Site=rownames(m.site.pc)

veg.msite=subset(veg, veg$ana.Site %in% m.site.pc$ana.Site)

m.site.pc=merge(m.site.pc, veg.msite, by="ana.Site")

# modern PCA + Veg infor

m.pdata=m.site.pc[c("ana.Site", "PC1", "PC2", "Vegetation\_type", "Geographic region")]

names(m.pdata)=c("ana.Site", "PC1", "PC2", "Veg\_Plot", "Region")

#

write.csv(m.pdata, paste0("mPCA\_", corei,"\_", dataT, "\_",opt,"\_",trans,".csv"))

# Full ana and non-ana for AM

fam=read.csv(paste0("analogs\_", corei,"\_", dataT, "\_",opt,"\_",trans,".csv"), check.names = F)

names(fam)[1]="ana.Site"

#

fam.mdata <- melt(setDT(fam), id=c("ana.Site"), variable.factor = FALSE)

names(fam.mdata)=c("ana.Site", "age", "ana.value")

fam.mdata=type\_convert(fam.mdata)

#

core.quant=read.csv(paste0(corei,"\_", dataT,"\_mdc\_quantiles\_", opt, "\_",trans,".csv"))

names(core.quant)[1]="qua"

fam.5=fam.mdata

# dcrit

dcrit=read.csv(paste0(corei,"\_", dataT, "\_ROC\_statistic\_", opt, "\_",trans,".csv"))

names(dcrit)[1]="vegG"

dcrit.value=as.data.frame(dcrit[dcrit$vegG == "Combined", "Opt..Dis."])

names(dcrit.value)="dcrit.value"

#

fam.5$analogues=ifelse(fam.mdata$ana.value < core.quant$quantiles[1], "Close",

ifelse(fam.mdata$ana.value >= core.quant$quantiles[1] & fam.mdata$ana.value < core.quant$quantiles[3], "Good",

ifelse(fam.mdata$ana.value >= core.quant$quantiles[3] & fam.mdata$ana.value < dcrit.value$dcrit.value, "Poor", "Non-analogue")))

fam.an=subset(fam.5, !(fam.5$analogues %in% c("Non-analogue")))

fam.ana=subset(fam.5, fam.5$analogues %in% c("Non-analogue"))

# as analogue one time at least

anS=as.data.frame(unique(fam.an$ana.Site))

names(anS)=c("ana.Site")

anS.veg=subset(m.pdata, m.pdata$ana.Site %in% anS$ana.Site)

# never as analogue

non.anS=subset(m.pdata, !(m.pdata$ana.Site %in% anS$ana.Site))

# good ana

fam.good=subset(fam.5, fam.5$analogues %in% c("Close", "Good"))

ana.good=as.data.frame(unique(fam.good$ana.Site))

names(ana.good)=c("ana.Site")

ana.good=subset(m.pdata, m.pdata$ana.Site %in% ana.good$ana.Site)

# read ana

fd.an=read.csv(paste0(corei, "\_", dataT, "\_modernAnalog\_Veg\_", opt, "\_",trans,".csv"))

names(fd.an)[2]="age"

fd.an.T=as.data.frame(unique(fd.an$age))

names(fd.an.T)="age"

# fossil samples having analogue

fd.an.pca=subset(f.m.pred.pc12, f.m.pred.pc12$age %in% fd.an.T$age)

fd.non.an.pca=subset(f.m.pred.pc12, !(f.m.pred.pc12$age %in% fd.an.T$age))

# Veg level

anS.veg$fVEG <- factor(anS.veg$Veg\_Plot, levels=c("Coniferous forest", "Tibetan shrubland", "Steppe", "Alpine meadow", "Middle taiga",

"North taiga", "Tundra", "Cultivated"))

non.anS$fVEG <- factor(non.anS$Veg\_Plot, levels=c("Coniferous forest", "Tibetan shrubland", "Steppe", "Alpine meadow", "Middle taiga",

"North taiga", "Tundra", "Cultivated"))

ana.good$fVEG <- factor(ana.good$Veg\_Plot, levels=c("Coniferous forest", "Tibetan shrubland", "Steppe", "Alpine meadow", "Middle taiga",

"North taiga", "Tundra", "Cultivated"))

# plot

if (corei %in% c("OM12A", "OM02B")) {

cor.geo.p=corei

p1=ggplot() +

# x and y

scale\_x\_continuous(name = paste0("PC1 (", pro.ex.pc1, "%)"), limits = c(-0.5, 0.5)) +

scale\_y\_continuous(name = paste0("PC2 (", pro.ex.pc2, "%)"), limits = c(-0.5, 0.5)) +

# pc1 = 0, pc2 = 0

geom\_hline(yintercept = 0, lty = 2, size = 0.15) +

geom\_vline(xintercept = 0, lty = 2, size = 0.15) +

# analgoues

geom\_point(data =anS.veg, aes(x = PC1, y = PC2, color = fVEG, fill = fVEG), shape = 16, show.legend = FALSE, size = 1.5, alpha = 0.8) +

# label good

#geom\_text\_repel(data = ana.good, aes(x = PC1, y = PC2, label = ana.Site, color = fVEG), nudge\_y = -0.01, size = 2.5, segment.size = 0.1, show.legend = FALSE) +

# non-ana

geom\_point(data = non.anS, aes(x = PC1, y = PC2, color = fVEG, fill = fVEG), shape = 1, size = 1) +

#geom\_text\_repel(data = non.anS, aes(x = PC1, y = PC2, label = ana.Site, color = fVEG),

# nudge\_y = -0.01, size = 2, segment.size = 0.1, show.legend = FALSE) +

#fossil samples prediction, am

geom\_point(data = fd.an.pca, aes(x = PC1, y = PC2), color = "black", shape = 3, size = 1) +

scale\_color\_manual(name = "Vegetation types",

values = cols) +

scale\_fill\_manual(name = "Vegetation types",

values = cols.fill) +

labs(title = cor.geo.p) +

# theme

theme\_bw() +

my.theme

} else if (corei %in% c("NC")){

cor.geo.p=corei

p1=ggplot() +

# x and y

scale\_x\_continuous(name = paste0("PC1 (", pro.ex.pc1, "%)"), limits = c(-0.5, 0.5)) +

scale\_y\_continuous(name = paste0("PC2 (", pro.ex.pc2, "%)"), limits = c(-0.5, 0.5)) +

# pc1 = 0, pc2 = 0

geom\_hline(yintercept = 0, lty = 2, size = 0.15) +

geom\_vline(xintercept = 0, lty = 2, size = 0.15) +

# analgoues

geom\_point(data =anS.veg, aes(x = PC1, y = PC2, color = fVEG, fill = fVEG), shape = 16, show.legend = FALSE, size = 1.5, alpha = 0.8) +

# label good

#geom\_text\_repel(data = ana.good, aes(x = PC1, y = PC2, label = ana.Site, color = fVEG), nudge\_y = -0.01, size = 2.5, segment.size = 0.1, show.legend = FALSE) +

# non-ana

geom\_point(data = non.anS, aes(x = PC1, y = PC2, color = fVEG, fill = fVEG), shape = 1, size = 1) +

#geom\_text\_repel(data = non.anS, aes(x = PC1, y = PC2, label = ana.Site, color = fVEG),

# nudge\_y = -0.01, size = 2, segment.size = 0.1, show.legend = FALSE) +

#fossil samples prediction, am

geom\_point(data = fd.an.pca, aes(x = PC1, y = PC2), color = "black", shape = 3, size = 1) +

geom\_text\_repel(data = fd.an.pca[1:21,], aes(x = PC1, y = PC2, label = age),

color = "black", nudge\_y = -0.01, size = 2, segment.size = 0.1) +

# fossil non

geom\_point(data = fd.non.an.pca, aes(x = PC1, y = PC2), color = "black", shape = 4, size = 1) +

geom\_text\_repel(data = fd.non.an.pca, aes(x = PC1, y = PC2, label = age),

color = "black", nudge\_y = -0.01, size = 2, segment.size = 0.1) +

# legend

#scale\_shape\_manual(name = "Vegetation types",

# values = shapes) +

scale\_color\_manual(name = "Vegetation types",

values = cols) +

scale\_fill\_manual(name = "Vegetation types",

values = cols.fill) +

labs(title = cor.geo.p) +

# theme

theme\_bw() +

my.theme

} else {

cor.geo.p=corei

p1=ggplot() +

# x and y

scale\_x\_continuous(name = paste0("PC1 (", pro.ex.pc1, "%)"), limits = c(-0.5, 0.5)) +

scale\_y\_continuous(name = paste0("PC2 (", pro.ex.pc2, "%)"), limits = c(-0.5, 0.5)) +

# pc1 = 0, pc2 = 0

geom\_hline(yintercept = 0, lty = 2, size = 0.15) +

geom\_vline(xintercept = 0, lty = 2, size = 0.15) +

# analgoues

geom\_point(data =anS.veg, aes(x = PC1, y = PC2, color = fVEG, fill = fVEG), shape = 16, show.legend = FALSE, size = 1.5, alpha = 0.8) +

# label good

#geom\_text\_repel(data = ana.good, aes(x = PC1, y = PC2, label = ana.Site, color = fVEG), nudge\_y = -0.01, size = 2.5, segment.size = 0.1, show.legend = FALSE) +

# non-ana

geom\_point(data = non.anS, aes(x = PC1, y = PC2, color = fVEG, fill = fVEG), shape = 1, size = 1) +

#geom\_text\_repel(data = non.anS, aes(x = PC1, y = PC2, label = ana.Site, color = fVEG),

# nudge\_y = -0.01, size = 2, segment.size = 0.1, show.legend = FALSE) +

#fossil samples prediction, am

geom\_point(data = fd.an.pca, aes(x = PC1, y = PC2), color = "black", shape = 3, size = 1) +

# fossil non

geom\_point(data = fd.non.an.pca, aes(x = PC1, y = PC2), color = "black", shape = 4, size = 1) +

geom\_text\_repel(data = fd.non.an.pca, aes(x = PC1, y = PC2, label = age),

color = "black", nudge\_y = -0.01, size = 2, segment.size = 0.1) +

scale\_color\_manual(name = "Vegetation types",

values = cols) +

scale\_fill\_manual(name = "Vegetation types",

values = cols.fill) +

labs(title = cor.geo.p) +

# theme

theme\_bw() +

my.theme

}

#ggplot2

ggsave(paste0(corei, "\_",dataT, "\_",PCApre, "\_", opt, "\_",trans,".pdf"), plot = p1, width = 15, height = 10, units = "cm")

}