Appendix C: R script to create an identification table

#DATA IMPORT (obistat output)

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

data=read.table("minicircles\_assignation\_stat.txt",header=T,sep="\t",colClasses=c("character","character","numeric","numeric"))

#CREATION OF THE IDENTIFICATION TABLE

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

#Create an empty dataframe

#==============================

#row names are the names of the sample

#column names depend on the list of species identified among the samples

sp\_names=sort(unique(data[,2]))[sort(unique(data[,2]))!="None"]

minicircles\_identification=as.data.frame(matrix(0,nrow=length(unique(data[,1])),ncol=length(sp\_names)+3))

rownames(minicircles\_identification)=unique(data[,1])

colnames(minicircles\_identification)=c("Identification","Score",sp\_names,"None")

#Computing of the proportion of sequences assigned to each species in each sample

#==============================

#takes into account the abundance of each sequence

for (i in unique(data[,1])){

for (j in unique(data[,2])){

if (length(data$total[data[,1]==i&data[,2]==j])>0)

minicircles\_identification[i,j]=data$total[data[,1]==i&data[,2]==j]/sum(data$total[data[,1]==i])

}

}

#Samples identification

#==============================

#the sample is identified as the species to wich the highest proportion of sequences has been assigned

minicircles\_identification$Identification=colnames(minicircles\_identification)[max.col(minicircles\_identification)]

#Computing of an identification score

#==============================

#the identification score is defined as the proportion of the identified species among all specific

#assignations (i.e. excluding sequences that did not find a specific match in the references)

minicircles\_identification$Score=apply(minicircles\_identification[,sp\_names],1,FUN=max)/rowSums(minicircles\_identification[sp\_names])

#Rename columns and ouptut the table in a tab-delimited file

#==============================

colnames(minicircles\_identification)=c("Identification","Score",paste("prop.",sp\_names),"None")

write.table(minicircles\_identification,file = "minicircles\_identifications.txt",col.names = T,row.names = T,sep='\t',quote = F)