Check\_Ident\_New Column

Tobias Opialla

June 21, 2019

Table of Contents

[load packages 1](#_Toc13248994)

[read raw data 2](#_Toc13248995)

[format raw data 2](#_Toc13248996)

[get pairings for comparison 3](#_Toc13248997)

[Plot RIs to check mis-annotations 3](#_Toc13248998)

[plot Deltas vs Library to check movements 6](#_Toc13248999)

[compare vs library entry 8](#_Toc13249000)

[write new table 16](#_Toc13249001)

# load packages

library(tidyverse)

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.0 v purrr 0.3.2  
## v tibble 2.1.3 v dplyr 0.8.2  
## v tidyr 0.8.3 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(cowplot)

##   
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggplot2':  
##   
## ggsave

library(ggplot2)

# read raw data

First round, the raw data had a fluke in the Sucrose annotation. In this run of the script the fluke was ***not*** corrected.

if(params$use\_uncorrected\_input){  
 rm(df)  
 for(file in dir(pattern = "IdentRef.tsv")){  
   
 cdf=read.delim(file)  
 cdf$file=file  
 if(!is.data.frame(df)){  
 df=cdf  
 }else{  
 df=rbind(df,cdf)  
 rm(cdf)  
 }  
 }  
}else{  
 rm(df)  
 for(file in dir(pattern = "cdf-Ident.\*\\.tsv")){  
 if(file=="CTPLI-e19154to\_005\_IdentC.cdf-IdentRef.tsv"){  
 next()  
 }else{  
 cdf=read.delim(file)  
 cdf$file=file  
 if(!is.data.frame(df)){  
 df=cdf  
 }else{  
 df=rbind(df,cdf)  
 rm(cdf)  
 }  
 }  
 }  
}

# format raw data

df=df %>% filter(!grepl("Unknown",Name))  
df$ident=gsub(".\*(Ident[ABCD]).\*","\\1",df$file)  
df=arrange(df,ident)  
df$n\_MEOX=gsub(".\*(\\d+)MEOX.\*","\\1",df$Name)  
df$n\_MEOX=ifelse(nchar(df$n\_MEOX)>2,0,df$n\_MEOX) %>% as.numeric()  
df$n\_TMS=gsub(".\*(\\d+)TMS.\*","\\1",df$Name)  
df$n\_TMS=ifelse(nchar(df$n\_TMS)>2,0,df$n\_TMS) %>% as.numeric()  
df$RI\_library=gsub(".\*RI:(\\d+).\*","\\1",df$Name) %>% as.numeric() # alkanes give NAs

## Warning in function\_list[[k]](value): NAs introduced by coercion

df$RI\_library=ifelse(is.na(df$RI\_library),  
 gsub("c","",df$Name) %>% as.numeric()\*100,  
 df$RI\_library) #for alkanes

## Warning in function\_list[[k]](value): NAs introduced by coercion

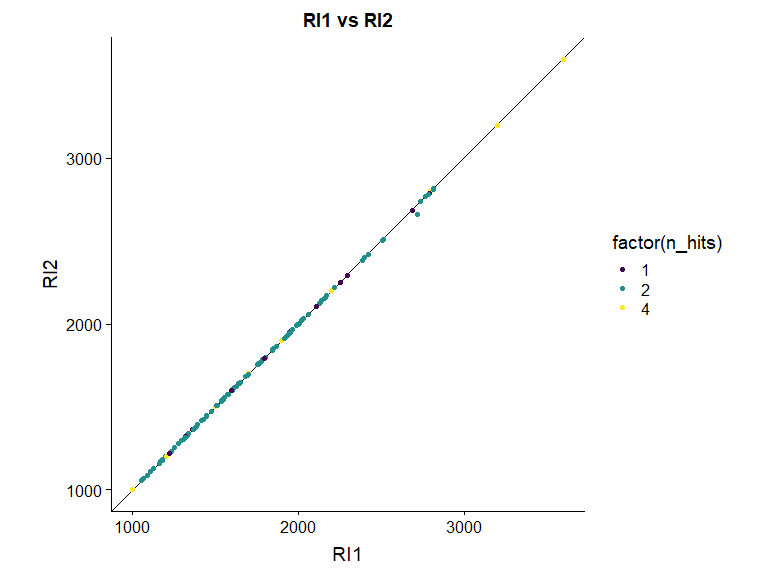
df$fc\_RI\_vs\_lib=df$RetentionIndex/df$RI\_library

# get pairings for comparison

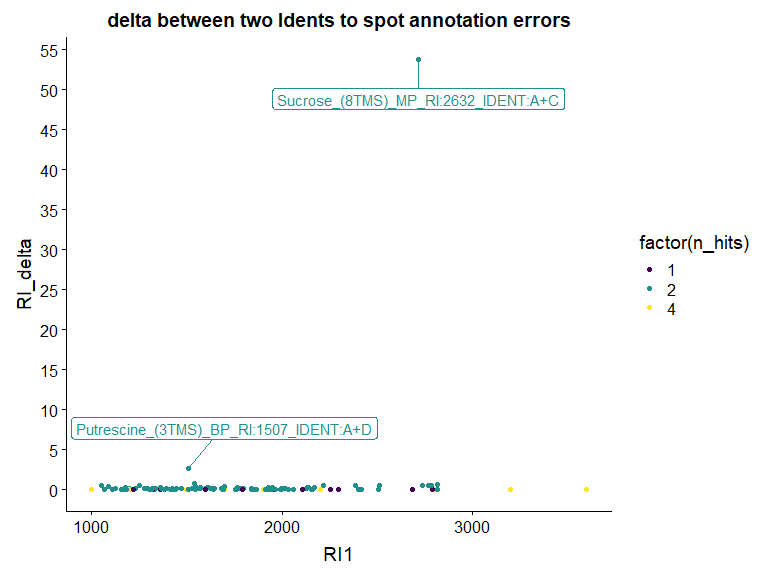
df\_sum=df %>% group\_by(Name,n\_TMS,n\_MEOX) %>% summarise(  
 RI1=first(RetentionIndex),  
 Rt1=first(RetentionTime),  
 fc\_RI\_vs\_lib1=first(fc\_RI\_vs\_lib),  
 Ident1=first(ident),  
   
 RI2=last(RetentionIndex),  
 Rt2=last(RetentionTime),  
 fc\_RI\_vs\_lib2=last(fc\_RI\_vs\_lib),  
 Ident2=last(ident),  
   
 n\_hits=length(RetentionIndex)  
   
) %>% mutate(fc\_RI\_vs\_lib\_mean=mean(c(fc\_RI\_vs\_lib1,fc\_RI\_vs\_lib2),na.rm = T),  
 RI\_new\_mean=mean(c(RI1,RI2),na.rm = T))  
  
df\_sum$RI\_fold\_change=df\_sum$RI1/df\_sum$RI2  
df\_sum$Rt\_fold\_change=df\_sum$Rt1/df\_sum$Rt2  
  
df\_sum$RI\_delta=abs(df\_sum$RI1 - df\_sum$RI2)  
df\_sum$Rt\_delta=abs(df\_sum$Rt1 - df\_sum$Rt2)

# Plot RIs to check mis-annotations

ggplot(df\_sum, aes(x=RI1,y=RI2,color=factor(n\_hits)))+  
 geom\_abline(slope = 1,intercept = 0)+  
 geom\_point()+  
 coord\_equal()+  
 scale\_colour\_viridis\_d()+  
 ggtitle("RI1 vs RI2")



ggplot(df\_sum, aes(x=RI1,y=RI\_delta,color=factor(n\_hits)))+  
 geom\_abline(slope = 1,intercept = 0)+  
 geom\_point()+  
 #coord\_equal()+  
 scale\_colour\_viridis\_d()+  
 scale\_y\_continuous(breaks=seq(0,60,5))+  
 ggtitle("delta between two Idents to spot annotation errors")+  
 ggrepel::geom\_label\_repel(data=df\_sum %>% filter(RI\_delta>2),aes(label=Name),nudge\_y = c(5,-5),show.legend = F)

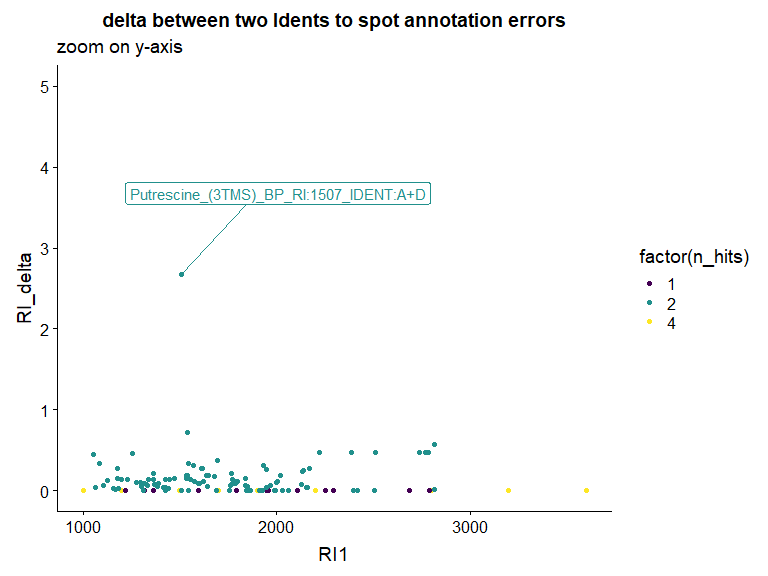


there are is one **obvious** error, zoom in more on y

ggplot(df\_sum, aes(x=RI1,y=RI\_delta,color=factor(n\_hits)))+  
 geom\_abline(slope = 1,intercept = 0)+  
 geom\_point()+  
 #coord\_equal()+  
 scale\_colour\_viridis\_d()+  
 ggtitle("delta between two Idents to spot annotation errors", subtitle = "zoom on y-axis")+  
 ylim(NA,5)+  
 ggrepel::geom\_label\_repel(data=df\_sum %>% filter(RI\_delta>2),aes(label=Name),nudge\_x = 500,nudge\_y = 1,show.legend = F)

## Warning: Removed 1 rows containing missing values (geom\_point).

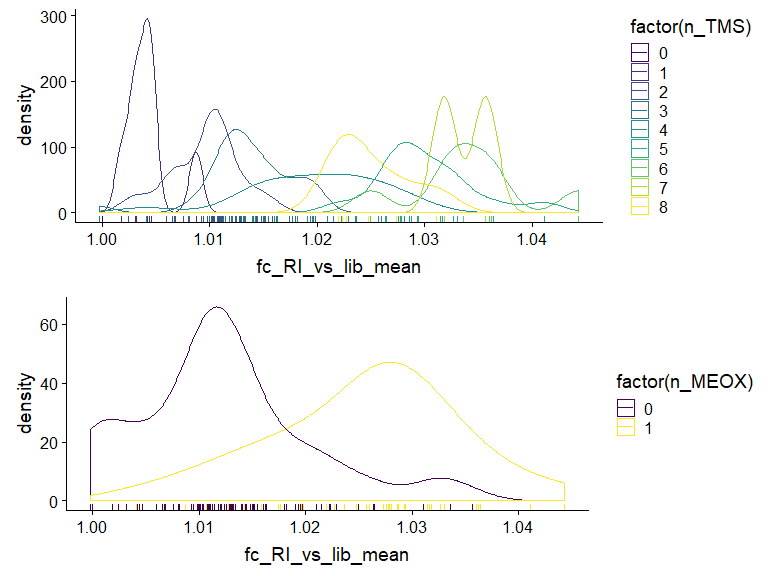
## Warning: Removed 1 rows containing missing values (geom\_label\_repel).



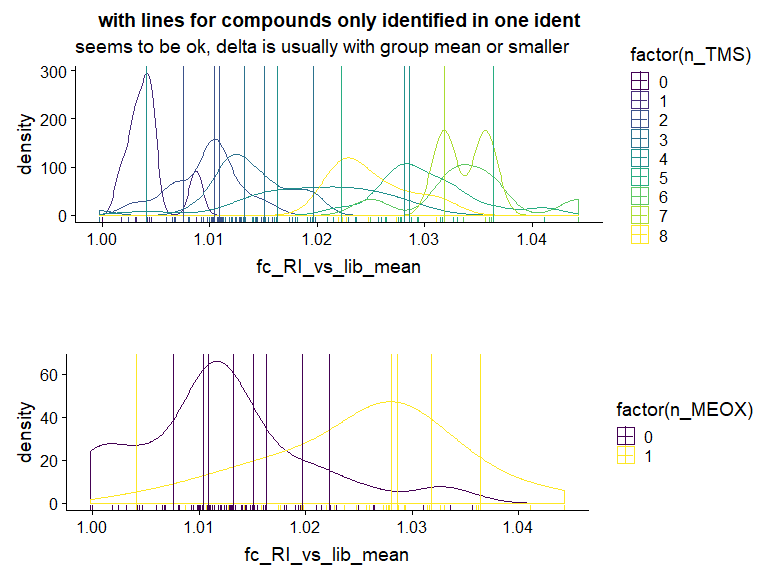
A delta in RI of ca 2.5 is very acceptable

# plot Deltas vs Library to check movements

p.TMS=ggplot(df\_sum,aes(x=fc\_RI\_vs\_lib\_mean,color=factor(n\_TMS)))+  
 geom\_rug()+  
 geom\_density()+  
 scale\_colour\_viridis\_d()  
  
  
p.MEOX=ggplot(df\_sum,aes(x=fc\_RI\_vs\_lib\_mean,color=factor(n\_MEOX)))+  
 geom\_rug()+  
 geom\_density()+  
 scale\_colour\_viridis\_d()  
  
  
plot\_grid(p.TMS,p.MEOX,ncol = 1, align = "h",axis="l")

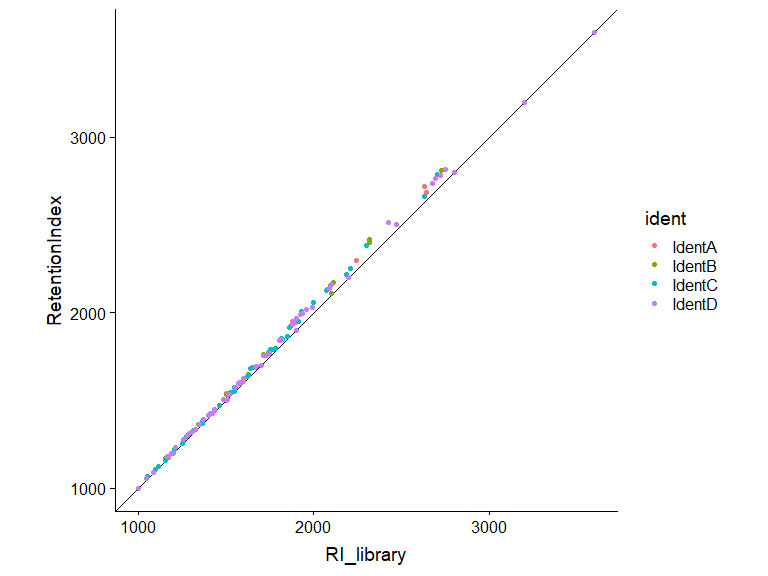


p.TMS=ggplot(df\_sum,aes(x=fc\_RI\_vs\_lib\_mean,color=factor(n\_TMS)))+  
 geom\_rug()+  
 geom\_density()+  
 geom\_vline(data=df\_sum %>% filter(n\_hits<2),aes(xintercept=fc\_RI\_vs\_lib\_mean,color=factor(n\_TMS)))+  
 scale\_colour\_viridis\_d()+  
 ggtitle("with lines for compounds only identified in one ident",  
 subtitle = "seems to be ok, delta is usually with group mean or smaller")  
  
  
p.MEOX=ggplot(df\_sum,aes(x=fc\_RI\_vs\_lib\_mean,color=factor(n\_MEOX)))+  
 geom\_rug()+  
 geom\_density()+  
 geom\_vline(data=df\_sum %>% filter(n\_hits<2),aes(xintercept=fc\_RI\_vs\_lib\_mean,color=factor(n\_MEOX)))+  
 scale\_colour\_viridis\_d()  
  
  
plot\_grid(p.TMS,p.MEOX,ncol = 1, align = "h",axis="l")

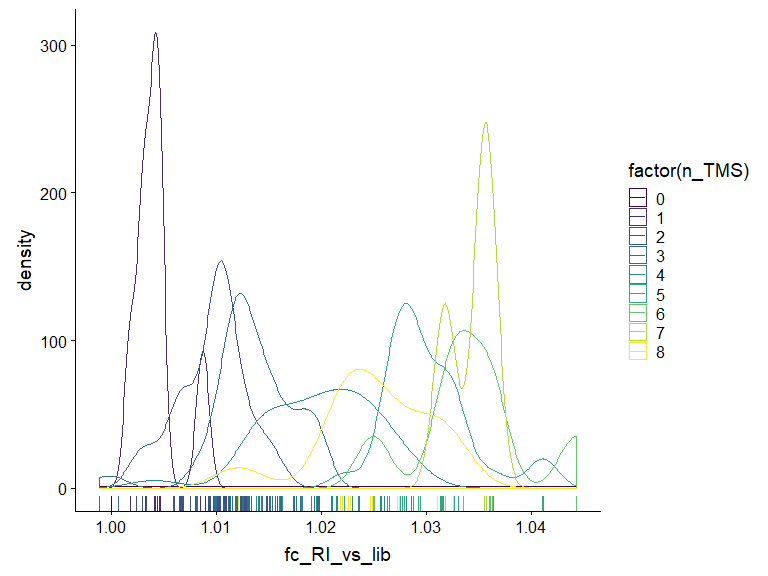


# compare vs library entry

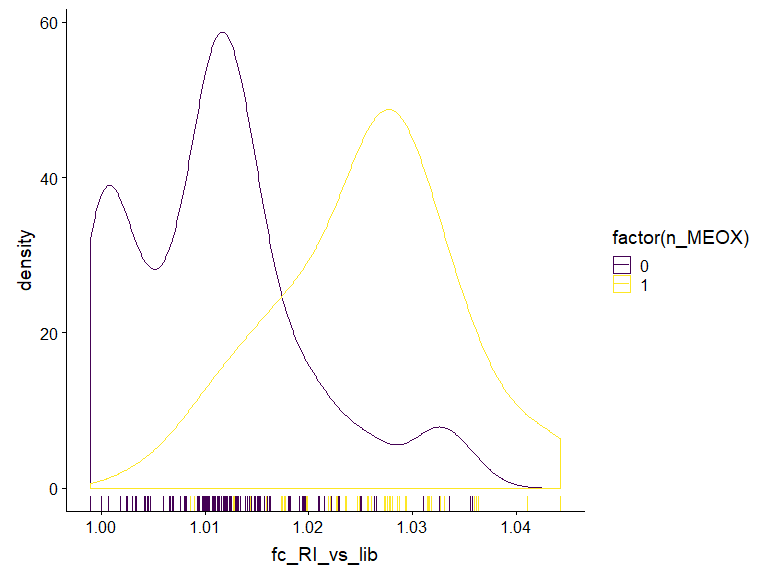
ggplot(df,aes(x=RI\_library,y=RetentionIndex,color=ident))+  
 geom\_abline(slope = 1,intercept = 0)+  
 geom\_point()+  
 coord\_equal()



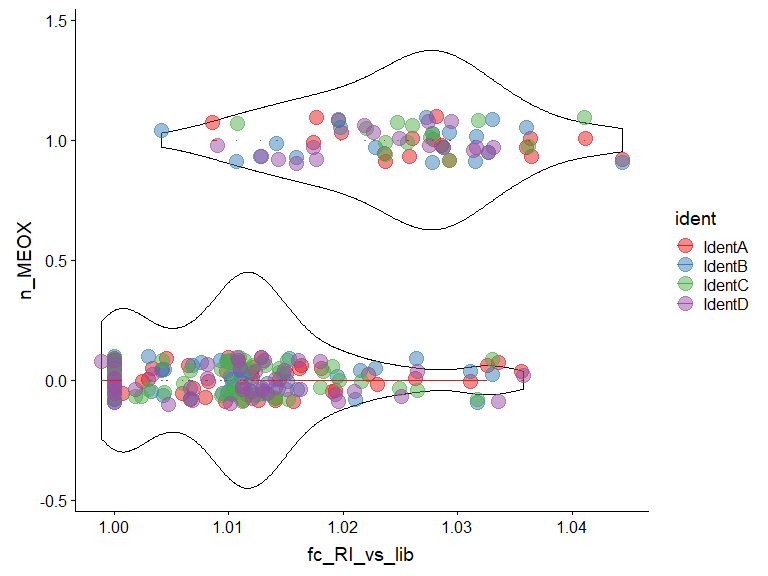
ggplot(df,aes(x=fc\_RI\_vs\_lib,color=factor(n\_TMS)))+  
 geom\_rug()+  
 geom\_density()+  
 scale\_colour\_viridis\_d()



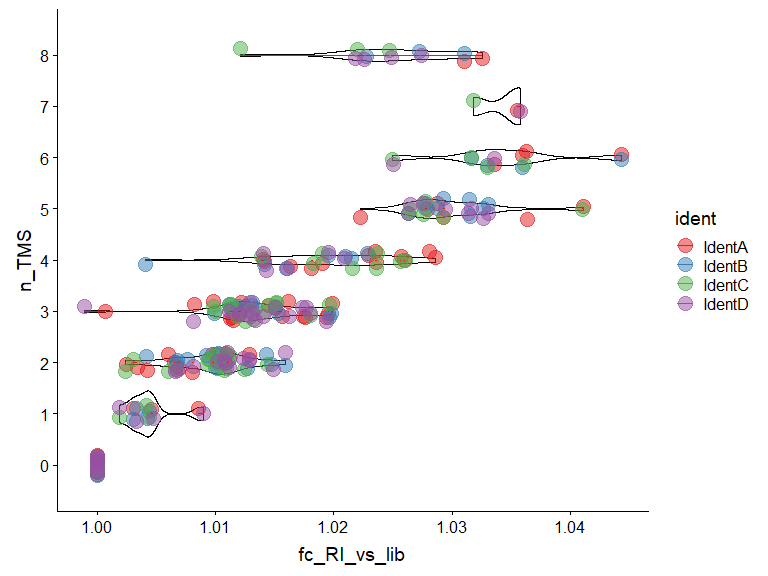
ggplot(df,aes(x=fc\_RI\_vs\_lib,color=factor(n\_MEOX)))+  
 geom\_rug()+  
 geom\_density()+  
 scale\_colour\_viridis\_d()



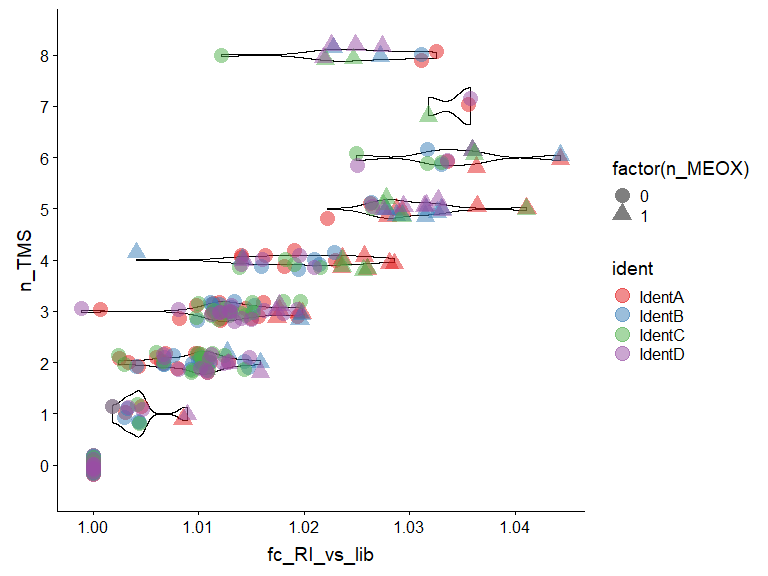
ggplot(df,aes(y=fc\_RI\_vs\_lib,x=n\_MEOX,color=ident))+  
 geom\_violin(color="black",aes(group=n\_MEOX))+  
 geom\_jitter(size=5,alpha=0.5,width = 0.1,height = 0)+  
 scale\_color\_brewer(palette = "Set1")+  
 geom\_line(aes(group=Name))+  
 coord\_flip()



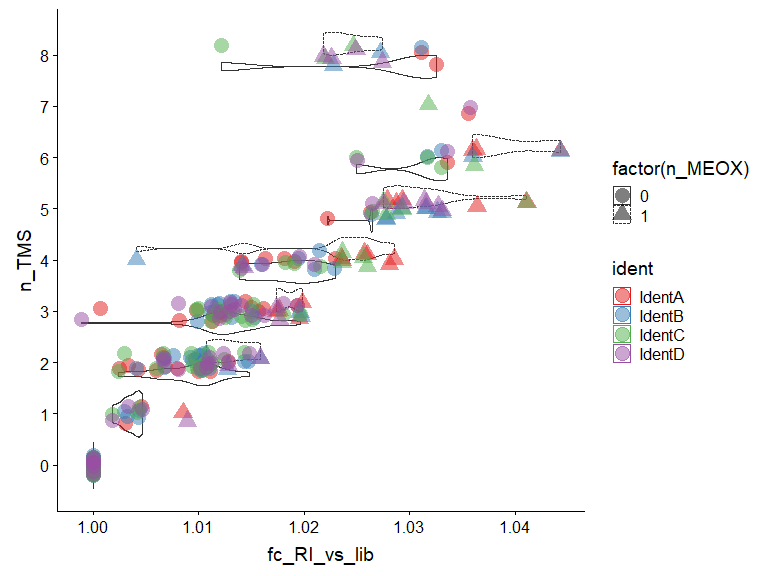
ggplot(df,aes(y=fc\_RI\_vs\_lib,x=n\_TMS,color=ident))+  
 geom\_violin(color="black",aes(group=n\_TMS))+  
 geom\_jitter(size=5,alpha=0.5,width = 0.2,height = 0)+  
 scale\_color\_brewer(palette = "Set1")+  
 scale\_x\_continuous(breaks = 0:8)+  
 geom\_line(aes(group=Name))+  
 coord\_flip()



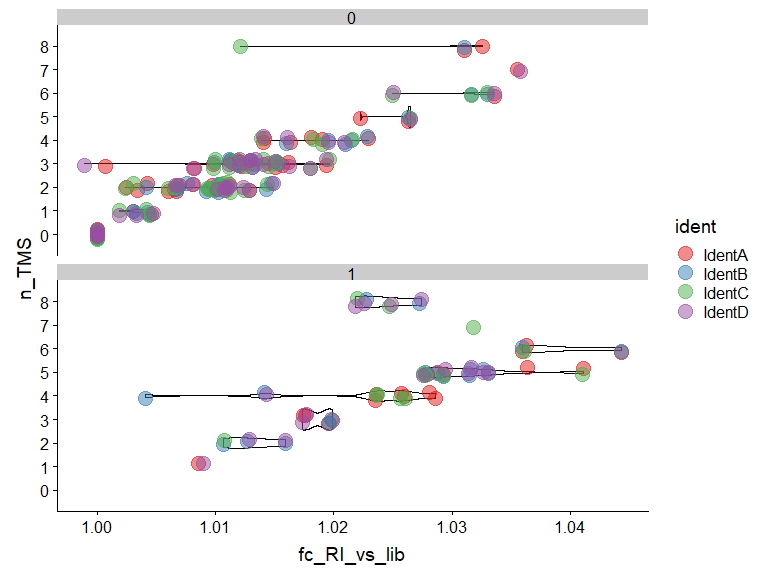
ggplot(df,aes(y=fc\_RI\_vs\_lib,x=n\_TMS,color=ident))+  
 geom\_violin(color="black",aes(group=n\_TMS))+  
 geom\_jitter(aes(shape=factor(n\_MEOX)),size=5,alpha=0.5,width = 0.2,height = 0)+  
 scale\_color\_brewer(palette = "Set1")+  
 scale\_x\_continuous(breaks = 0:8)+  
 coord\_flip()



ggplot(df,aes(y=fc\_RI\_vs\_lib,x=n\_TMS,color=ident))+  
 geom\_violin(aes(group=paste0(n\_TMS,n\_MEOX),linetype=factor(n\_MEOX)),fill=NA,scale = "width")+  
 geom\_jitter(aes(shape=factor(n\_MEOX)),size=5,alpha=0.5,width = 0.2,height = 0)+  
 scale\_color\_brewer(palette = "Set1")+  
 scale\_x\_continuous(breaks = 0:8)+  
 coord\_flip()



ggplot(df,aes(y=fc\_RI\_vs\_lib,x=n\_TMS,color=ident))+  
 geom\_violin(color="black",aes(group=n\_TMS))+  
 geom\_jitter(size=5,alpha=0.5,width = 0.2,height = 0)+  
 scale\_color\_brewer(palette = "Set1")+  
 coord\_flip()+  
 scale\_x\_continuous(breaks = 0:8)+  
 facet\_wrap(~n\_MEOX,ncol = 1)



The only compound in question is Sucrose. It’s likely that the measurement in IdentA is correct, the one in IdentC is probably the one found by the search and wronly locked.

# write new table

Write output for new ident

#make sure, that it's only written from the corrected input  
if(!params$use\_uncorrected\_input){  
 df\_output=df\_sum %>% ungroup %>% select(Name,RI\_new\_mean) %>% filter(!str\_detect(Name,"^c\\d\\d$"))  
 df\_output$RI\_new\_mean=round(df\_output$RI\_new\_mean)  
 write.table(df\_output,paste0("New\_RIs\_",Sys.Date(),".tsv"),row.names = F,col.names = T,sep = "\t")  
}