```r
setwd("~/04_Champion_Lab/02_N-terminal_Acetylation/OnePot")

#import libraries
library(readxl)
library(stringr)
library(tidyr)
library(dplyr)
library(seqinr)
library(Peptides)
library(cleaver)
library(ggplot2)
library(scales)

#set colors
red <- hue_pal()(3)[1]
green <- hue_pal()(3)[2]
blue <- hue_pal()(3)[3]


#read in data. one pot, enriched data (trypsin), genome
op <- read.csv("Data/OnePotSummarisedData.csv")
enr <- read.csv("Data/EnrichedSummarisedData.csv")
gen <- read.csv("Data/GenomeSummarisedData_GASTVCP.csv")



#filters
minStart <- 2
minLength <- 5
maxLength <- 50

#filter datasets
opN <- op[op$Start <= minStart &
            op$Length >= minLength &
            op$Length <= maxLength,]
enr <- enr[enr$Start <= minStart &
             enr$Length >= minLength &
             enr$Length <= maxLength,]
gen <- gen[gen$Length >= minLength &
             gen$Length <= maxLength,]

#only keep columns of interest
keeps <- c("Peptide", "IEP")
opN <- opN[,names(opN) %in% keeps]
enr <- enr[,names(enr) %in% keeps]
gen <- gen[,names(gen) %in% keeps]

#set labels
opN$category <- "One Pot (This Work)"
enr$category <- "Enriched (Thompson et al.)"
gen$category <- "Genome"

#create combined dataset with category labels
combined <- bind_rows(opN, enr, gen)
combined$category <- factor(combined$category, levels = c("Genome",
                                                          "One Pot (This Work)",
                                                          "Enriched (Thompson et al.)"))
#create plots and save
ggplot(combined) +
  geom_density(aes(x = IEP, color = category),
               linewidth = 1.4)+
  theme_bw(base_size = 15) +
  theme(panel.grid = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  labs(x = "Isoelectric Point",
       y = "N-term Peptide Frequency",
       color = element_blank())

```

```r
ggsave("plots/GASCTPV/DensityPlotNterminalPeptidesALL.png",
       width = 12, height = 8)


ggplot(combined[combined$category %in% c("Genome", "One Pot (This Work)", "Enriched (Thompson et al.)"),]) +
  geom_density(aes(x = IEP, color = category),
               linewidth = 1.4)+
  theme_bw(base_size = 15) +
  theme(panel.grid = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank()) +
  labs(x = "Isoelectric Point",
       y = "N-term Peptide Frequency",
       color = element_blank())

ggsave("plots/GASCTPV/DensityPlotNterminalPeptides_OP-ENR-GenR.png",
       width = 12, height = 8)



ggplot(combined) +
  geom_histogram(aes(x = IEP, fill = category),
               linewidth = 1.4, color = "black", binwidth = 0.2)+
  theme_bw(base_size = 15) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Isoelectric Point",
       y = "N-term Peptide Frequency",
       color = element_blank()) +
  facet_grid(category~.)

ggsave("plots/GASCTPV/HistogramPlotNterminalPeptides_All.png",
       width = 12, height = 16)



ggplot(combined[combined$category %in% c("Genome", "One Pot (This Work)", "Enriched (Thompson et al.)"),]) +
  geom_histogram(aes(x = IEP, fill = category),
                 linewidth = 1.4, color = "black", binwidth = 0.2)+
  theme_bw(base_size = 30) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Isoelectric Point",
       y = "N-term Peptide Frequency",
       color = element_blank()) +
  facet_grid(category~.)

ggsave("plots/GASCTPV/HistogramPlotNterminalPeptides_OP-ENR-GenR_sizedUp.png",
       width = 12, height = 16)




###Neo vs N-term Genome and OP
op <- read.csv("Data/OnePotSummarisedData.csv")
gen <- read.csv("Data/GenomeSummarisedData_allPeptides_GASTVCP.csv")

#filters
minLength <- 5
maxLength <- 50


#filter datasets
op <- op[op$Length >= minLength &
           op$Length <= maxLength,]

```

```r
#determine whether peptide is Nterminal
op$NtermPeptide <- F
op$NtermPeptide[op$Start <=2] <- T

#name columns to match PEAKS format
op$mass <- op$Mass
op$Accession <- op$Accessions
op$Peptide <- op$str_seq

#pull out only columns that are in the genome dataset
op <- op[,which(names(op) %in% names(gen))]


#set categories
op$origin <- "One Pot"
gen$origin <- "Genome"
op$cat <- "All"
gen$cat <- "All"
op$facetVar <- "One Pot All"
gen$facetVar <- "Genome All"
gen_Nterm <- gen[gen$NtermPeptide,]
gen_Nterm$facetVar <- "Genome N-termini"
gen_Nterm$cat <- "N-termini"
op_Nterm <- op[op$NtermPeptide,]
op_Nterm$facetVar <- "One Pot N-termini"
op_Nterm$cat <- "N-termini"


gen_NonNterm <- gen[!gen$NtermPeptide,]
gen_NonNterm$facetVar <- "Genome Neo Peptides"
gen_NonNterm$cat <- "Neo Peptides"
op_NonNterm <- op[!op$NtermPeptide,]
op_NonNterm$facetVar <- "One Pot Neo Peptides"
op_NonNterm$cat <- "Neo Peptides"

#bind rows together
master <- bind_rows(gen, op, gen_Nterm, gen_NonNterm, op_Nterm, op_NonNterm)

#create and save plots
ggplot(master) +
  geom_histogram(aes(x = IEP, fill = facetVar),
                 linewidth = 1.4, color = "black", binwidth = 0.2)+
  theme_bw(base_size = 20) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Isoelectric Point",
       y = "Peptide Frequency",
       color = element_blank()) +
  facet_grid(facetVar~., scales = "free")


ggsave("plots/GASCTPV/Histogram_IEP_neo_Ntermini_All_OPvsGen_6x1.png", width = 12,
       height = 20)




ggplot(master) +
  geom_histogram(aes(x = IEP, fill = cat),
                 linewidth = 1.4, color = "black", binwidth = 0.2)+
  theme_bw(base_size = 20) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Isoelectric Point",
       y = "Peptide Frequency",
       color = element_blank()) +
  facet_grid(cat~origin, scales = "free")

ggsave("plots/GASCTPV/Histogram_IEP_neo_Ntermini_All_OPvsGen_3x2.png", width = 12,
```

```
       height = 12)
```

```
ggplot(master) +
  geom_histogram(aes(x = mass, fill = facetVar),
                 linewidth = 1.4, color = "black", binwidth = 100)+
  theme_bw(base_size = 20) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Mass (Da)",
       y = "Peptide Frequency",
       color = element_blank()) +
  facet_grid(facetVar~., scales = "free")

ggsave("plots/GASCTPV/Histogram_mass_neo_Ntermini_All_OPvsGen_6x1.png", width = 12,
       height = 20)
```

```
ggplot(master) +
  geom_histogram(aes(x = mass, fill = cat),
                 linewidth = 1.4, color = "black", binwidth = 100)+
  theme_bw(base_size = 20) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Mass (Da)",
       y = "Peptide Frequency",
       color = element_blank()) +
  facet_grid(cat~origin, scales = "free")

ggsave("plots/GASCTPV/Histogram_mass_neo_Ntermini_All_OPvsGen_3x2.png", width = 12,
       height = 12)
```

```
ggplot(master) +
  geom_histogram(aes(x = Length, fill = facetVar),
                 linewidth = 1.4, color = "black", binwidth = 1)+
  theme_bw(base_size = 20) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Length (AAs)",
       y = "Peptide Frequency",
       color = element_blank()) +
  facet_grid(facetVar~., scales = "free")

ggsave("plots/GASCTPV/Histogram_length_neo_Ntermini_All_OPvsGen_6x1.png", width = 12,
       height = 20)
```

```
ggplot(master) +
  geom_histogram(aes(x = Length, fill = cat),
                 linewidth = 1.4, color = "black", binwidth = 1)+
```

```
  theme_bw(base_size = 20) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "Length (AAs)",
       y = "Peptide Frequency",
       color = element_blank()) +
  facet_grid(cat~origin, scales = "free")

ggsave("plots/GASCTPV/Histogram_length_neo_Ntermini_All_OPvsGen_3x2.png", width = 12,
       height = 12)


#same as above but only one pot data for retention time analysis
op <- read.csv("Data/OnePotSummarisedData.csv")

#filters
minLength <- 5
maxLength <- 50



op <- op[op$Length >= minLength &
           op$Length <= maxLength,]

op$NtermPeptide <- F
op$NtermPeptide[op$Start <=2] <- T

op$mass <- op$Mass
op$Accession <- op$Accessions

op$Peptide <- op$str_seq


op$RTfacet <- "Neo Peptides"
op$RTfacet[op$NtermPeptide] <- "N-termini"

ggplot(op) +
  geom_histogram(aes(x = RT, fill = RTfacet),
                 linewidth = 1.4, color = "black", binwidth = 1)+
  theme_bw(base_size = 40) +
  theme(panel.grid = element_blank(),
        legend.position = "none") +
  labs(x = "RT (min)",
       y = "Peptide Frequency",
       title = "One Pot Peptides",
       color = element_blank()) +
  scale_fill_manual(values = c(green,blue)) +
  facet_grid(RTfacet~., scales = "free")

ggsave("plots/GASCTPV/RT_histogram_onePot_neovsNterm.png", width = 12, height = 12)

ggplot(op) +
  geom_density(aes(x = RT, color = RTfacet,
                   fill = RTfacet),
               lwd = 3, alpha = 0.3)+
  theme_bw(base_size = 40) +
  theme(panel.grid = element_blank(),
        legend.position = "top",
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank()) +
  labs(x = "RT (min)",
       y = "Peptide Frequency (Density)",
       title = element_blank(),
       color = element_blank()) +
  scale_color_manual(values = c(green,blue)) +
  scale_fill_manual(guide = "none",
                    values = c(green,blue))
```

```r
ggsave("plots/GASCTPV/RT_Density_onePot_neovsNterm.png", width = 12, height = 10)



#create dataframe for amino acid analysis
master2 <- master[master$cat != "All",]

#all amino acids
AAs <- c("A", "G", "I", "L", "P", "V", "F", "W", "Y", "D", "E", "R", "H", "K", "S", "T",
"C", "M", "N", "Q")

#create variables for counting and percentage
for (AA in AAs) {
  master2[[paste0(AA, "_count")]] <- str_count(master2$Peptide, AA)
  master2[[paste0(AA, "_pct")]] <- str_count(master2$Peptide, AA) / master2$Length
}


#create long strings of all peptides in each category to count AAS
OP_NtermString <- paste(master2$Peptide[master2$cat == "N-termini" &
                                        master2$origin == "One Pot"], collapse = "")

OP_NeoString <- paste(master2$Peptide[master2$cat == "Neo Peptides" &
                                          master2$origin == "One Pot"], collapse = "")

GEN_NtermString <- paste(master2$Peptide[master2$cat == "N-termini" &
                                          master2$origin == "Genome"], collapse = "")

GEN_NeoString <- paste(master2$Peptide[master2$cat == "Neo Peptides" &
                                         master2$origin == "Genome"], collapse = "")


#create AA dataframe
AA_df <- data.frame(expand.grid(AA = AAs, cat = unique(master2$cat), origin = unique(
master2$origin)))
AA_df$AA <- as.character(AA_df$AA)
AA_df$count = NA
AA_df$pct = NA

#loop through each category and count each AA and calcualte percentage.
for (i in 1:nrow(AA_df)) {
  if (AA_df$cat[i] == "N-termini" &
      AA_df$origin[i] == "One Pot") {
    AA_df$count[i] <- str_count(OP_NtermString, AA_df$AA[i])
    AA_df$pct[i] <- AA_df$count[i] / nchar(OP_NtermString)
  } else if (AA_df$cat[i] == "N-termini" &
             AA_df$origin[i] == "Genome") {
    AA_df$count[i] <- str_count(GEN_NtermString, AA_df$AA[i])
    AA_df$pct[i] <- AA_df$count[i] / nchar(GEN_NtermString)
  } else if (AA_df$cat[i] == "Neo Peptides" &
             AA_df$origin[i] == "One Pot") {
    AA_df$count[i] <- str_count(OP_NeoString, AA_df$AA[i])
    AA_df$pct[i] <- AA_df$count[i] / nchar(OP_NeoString)
  } else if (AA_df$cat[i] == "Neo Peptides" &
             AA_df$origin[i] == "Genome") {
    AA_df$count[i] <- str_count(GEN_NeoString, AA_df$AA[i])
    AA_df$pct[i] <- AA_df$count[i] / nchar(GEN_NeoString)
  }
}

#categorize AAs
AA_df$AAType <- NA
AA_df$AAType[AA_df$AA %in% c("A", "G", "I", "L", "M", "V")] <- "Aliphatic"
AA_df$AAType[AA_df$AA %in% c("F", "W", "Y")] <- "Aromatic"
AA_df$AAType[AA_df$AA %in% c("P", "S", "T", "C", "N", "Q")] <- "Polar"
AA_df$AAType[AA_df$AA %in% c("D", "E")] <- "Negative"
AA_df$AAType[AA_df$AA %in% c("R", "H", "K")] <- "Positive"

AA_df$AAType <- factor(AA_df$AAType, levels = c("Aliphatic", "Polar", "Aromatic",
```

```
    "Negative",
411                                                 "Positive"))
412
413   #plot and save
414   ggplot(AA_df) +
415     geom_bar(aes(x = AA, y = pct, fill = cat),
416              stat = "identity", position = "dodge", color = "black") +
417     facet_grid(origin~AAType, scales = "free", space = "free") +
418     theme_bw(base_size = 25) +
419     theme(panel.grid = element_blank(),
420           legend.position = "top") +
421     scale_fill_manual(values = c(green, blue)) +
422     labs(x = "Amino Acid", y = "Proportion", fill = element_blank())
423
424   ggsave("plots/GASCTPV/AAfrequency_facetByOrigin.png", width = 15, height = 8)
425
426
427
```