```r
setwd("~/04_Champion_Lab/02_N-terminal_Acetylation/OnePot/NAT5")

#import libraries
library(ggplot2)
library(stringr)
library(dplyr)
library(tidyr)
library(ggrepel)
library(scales)

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

#import peptides
peptides <- read.csv("db.protein-peptides.csv")

#filter out contaminants and clean sequences
peptides <- peptides[!grepl("CONTAM", peptides$Accession),]
peptides$Peptide <- str_remove(peptides$Peptide, "^.\\.")
peptides$Peptide <- str_remove(peptides$Peptide, "\\..$")

#identify peptides as Light, heavy or unlabelled
peptides$NtermAcet_L <- grepl("^[A-Z]\\(\\+42\\.01\\)", peptides$Peptide)
peptides$NtermAcet_H <- grepl("^[A-Z]\\(\\+45\\.03\\)", peptides$Peptide)
peptides$UnLabelled <- !peptides$NtermAcet_H & !peptides$NtermAcet_L

#convert to long data format (one row per inj/pep)
peptidesLong <- peptides  %>% pivot_longer(which(grepl("^Area.L", names(peptides))),
                                           values_to = "Area",
                                           names_to = "sample")

#remove extra columns
drops <- c("USED", "Quality", "Significance","Avg..ppm","X1.K0","Avg..Area",
           "Sample.Profile..Ratio.","Area.WT","Area.Del",
           "Area.Comp","Group.Profile..Ratio.","Max.Ratio", "X.Vector",
           names(peptides)[grepl("Intensity.", names(peptides))],
           names(peptides)[grepl("X.Spec", names(peptides))],
           names(peptides)[grepl("Area\\.", names(peptides))])
peptidesLong <- peptidesLong[,!names(peptidesLong) %in% drops]

#parse meta data
peptidesLong$Inj <- str_remove(peptidesLong$sample, "Area.L_x[0-9]+_")
peptidesLong$strain <- str_remove(str_extract(peptidesLong$Inj, "^[^_]+_"), "_")
peptidesLong$Biorep <- str_extract(peptidesLong$Inj, "br[0-9]")
peptidesLong$Techrep <- str_extract(peptidesLong$Inj, "tr[0-9]")

#pull N-terminal peptides
NacetpeptidesLong <- peptidesLong[peptidesLong$Start <= 2,]

#parse accession and gene info
NacetpeptidesLong$Accession2 <- str_remove(str_extract(NacetpeptidesLong$Accession,
"^[^\\|]*\\|"), "\\|")
NacetpeptidesLong$gene <- str_remove(str_remove(str_extract(NacetpeptidesLong$Accession,
"^[^\\|]*\\|[^\\|]*\\|"),
                                                "^[^\\|]*\\|"), "\\|")
#pull out list of accessions with light N-terminal acetylation
accessionsWithNtermAcet <- unique(NacetpeptidesLong$Accession2[NacetpeptidesLong$
NtermAcet_L])
NacetpeptidesLong <- NacetpeptidesLong[NacetpeptidesLong$Accession2 %in%
accessionsWithNtermAcet,]

#create sequence without heavy or light for peptide pair matching
NacetpeptidesLong$acetSeq[NacetpeptidesLong$NtermAcet_L] <- str_replace(NacetpeptidesLong
$Peptide[NacetpeptidesLong$NtermAcet_L],

                                                                        "\\(\\+42\\.01\\)
                                                                        ",
```

```
                                                                    "(ac)")
NacetpeptidesLong$acetSeq[NacetpeptidesLong$NtermAcet_H] <- str_replace(NacetpeptidesLong
$Peptide[NacetpeptidesLong$NtermAcet_H],
                                                                    "\\(\\+45\\.03\\)
                                                                    ",
                                                                    "(ac)")

NacetpeptidesLong$acetSeq[NacetpeptidesLong$UnLabelled] <- NacetpeptidesLong$Peptide[
NacetpeptidesLong$UnLabelled]
NacetpeptidesLong$noAcetSeq <- str_remove(NacetpeptidesLong$acetSeq, "\\(ac\\)")

#strain variables
STRAIN <- "1839"
DELSTRAIN <- paste0("d", STRAIN)
COMPSTRAIN <- paste0("c", STRAIN)


#pull out peptides with these strains of interest (search was with other injections as
well)
peps <- NacetpeptidesLong[NacetpeptidesLong$strain == "WT" |
                            grepl(STRAIN, NacetpeptidesLong$strain),]
#categorize peptides
peps$category <- ifelse(peps$UnLabelled, "Unlabelled",
                        ifelse(peps$NtermAcet_H, "Heavy", "Light"))

#group peptides and calculate the mean area for each one in each category
peps2 <- peps %>% group_by(Peptide, category) %>%
  mutate(rows = n(),
         IDs = sum(!is.na(Area)),
         meanArea = mean(Area, na.rm = T))

#only light peptides
lights <- peps2[peps2$NtermAcet_L,]

#decide if the peptide is the best flier for each accession based on largest mean area
peps3 <- lights %>% group_by(Accession2, category) %>%
  mutate(bestFlierArea = max(meanArea, na.rm = T),
         bestFlier = bestFlierArea == meanArea)
#pull out best fliers
bestFliers <- peps3[peps3$bestFlier &
                      !is.na(peps3$bestFlier),]



#find the heavy version of each best flier and pull the area.
bestFliers$HeavyArea <- NA
for (i in 1:nrow(bestFliers)) {
  temp <- peps[peps$Inj == bestFliers$Inj[i] &
                 peps$acetSeq == bestFliers$acetSeq[i] &
                 peps$NtermAcet_H &
                 peps$Accession2 == bestFliers$Accession2[i],]
  if (nrow(temp) == 1) {
    bestFliers$HeavyArea[i] <- temp$Area[1]
  } else if (nrow(temp) < 1) {
    bestFliers$HeavyArea[i] <- NA
  } else {
    print(nrow(temp))
    print(i)
    print(bestFliers$acetSeq[i])
  }
}

#calculat pct acet (light / (light + heavy )) * 100
bestFliers$pctAcet <- NA
bestFliers$pctAcet[!is.na(bestFliers$Area) &
                     !is.na(bestFliers$HeavyArea)] <- 100 *
  (bestFliers$Area[!is.na(bestFliers$Area) &
                     !is.na(bestFliers$HeavyArea)] / (bestFliers$Area[!is.na(bestFliers$
```

```
                      Area) &
                                                                    !is.na(bestFliers
                                                                    $HeavyArea)] +
                                                        bestFliers$HeavyArea[!is.na(
                                                        bestFliers$Area) &
                                                                             !is.na(
    bestFliers$HeavyArea)]))

    #for NAs, set 0 and 100 % depending on whether heavy or light is NA
    bestFliers$pctAcet[!is.na(bestFliers$Area) &
                         is.na(bestFliers$HeavyArea)] <- 100


    bestFliers$pctAcet[is.na(bestFliers$Area) &
                         !is.na(bestFliers$HeavyArea)] <- 0




    #combine all bioreps by protein and peptide sequence
    #calculate mean area of each strain and each heavy/light peptide
    #count the number of Non NA values for each
    #calculate sd of area
    #calculate mean and sd of pct acet for each strain
    #calculate fold chain betwee del/wt and del/comp
    #t-tests for each as well, based on pct acet between del/wt and del/comp (only for valid
    ones)
    CombinedTandBreps <- bestFliers %>% group_by(Accession2, gene, acetSeq) %>%
      summarise(mean_HeavyArea_WT = mean(HeavyArea[strain == "WT"], na.rm = T),
                mean_Area_WT = mean(Area[strain == "WT"], na.rm = T),
                mean_HeavyArea_DEL = mean(HeavyArea[strain == DELSTRAIN], na.rm = T),
                mean_Area_DEL = mean(Area[strain == DELSTRAIN], na.rm = T),
                mean_HeavyArea_COMP = mean(HeavyArea[strain == COMPSTRAIN], na.rm = T),
                mean_Area_COMP = mean(Area[strain == COMPSTRAIN], na.rm = T),
                count_H_WT = sum(!is.na(HeavyArea[strain == "WT"])),
                count_L_WT = sum(!is.na(Area[strain == "WT"])),
                count_H_DEL = sum(!is.na(HeavyArea[strain == DELSTRAIN])),
                count_L_DEL = sum(!is.na(Area[strain == DELSTRAIN])),
                count_H_COMP = sum(!is.na(HeavyArea[strain == COMPSTRAIN])),
                count_L_COMP = sum(!is.na(Area[strain == COMPSTRAIN])),
                sd_HeavyArea_WT = sd(HeavyArea[strain == "WT"], na.rm = T),
                sd_Area_WT = sd(Area[strain == "WT"], na.rm = T),
                sd_HeavyArea_DEL = sd(HeavyArea[strain == DELSTRAIN], na.rm = T),
                sd_Area_DEL = sd(Area[strain == DELSTRAIN], na.rm = T),
                sd_HeavyArea_COMP = sd(HeavyArea[strain == COMPSTRAIN], na.rm = T),
                sd_Area_COMP = sd(Area[strain == COMPSTRAIN], na.rm = T),
                mean_pctAcet_WT = mean(pctAcet[strain == "WT"], na.rm = T),
                mean_pctAcet_DEL = mean(pctAcet[strain == DELSTRAIN], na.rm = T),
                mean_pctAcet_COMP = mean(pctAcet[strain == COMPSTRAIN], na.rm = T),
                sd_pctAcet_WT = sd(pctAcet[strain == "WT"], na.rm = T),
                sd_pctAcet_DEL = sd(pctAcet[strain == DELSTRAIN], na.rm = T),
                sd_pctAcet_COMP = sd(pctAcet[strain == COMPSTRAIN], na.rm = T),
                count_pctAcet_WT = sum(!is.na(pctAcet[strain == "WT"])),
                count_pctAcet_DEL = sum(!is.na(pctAcet[strain == DELSTRAIN])),
                count_pctAcet_COMP = sum(!is.na(pctAcet[strain == COMPSTRAIN])),
                FC_pctAcet_WTDEL = mean_pctAcet_WT / mean_pctAcet_DEL,
                FC_pctAcet_COMPDEL = mean_pctAcet_COMP / mean_pctAcet_DEL,
                pval_WTDEL = ifelse(is.character(try(t.test(pctAcet[strain == "WT"],
                                                            pctAcet[strain == DELSTRAIN])[[3
                                                            ]],
                                                    silent = T)),
                                    NA, t.test(pctAcet[strain == "WT"],
                                               pctAcet[strain == DELSTRAIN])[[3]]),
                pval_COMPDEL = ifelse(is.character(try(t.test(pctAcet[strain == COMPSTRAIN],
                                                              pctAcet[strain == DELSTRAIN])[[
                                                              3]],
                                                      silent = T)),
                                      NA, t.test(pctAcet[strain == COMPSTRAIN],
                                                 pctAcet[strain == DELSTRAIN])[[3]])
```

```
  )



#require at least 2 obs of WT and Comp (out of 6)
minWTandCompObs <- 2
CombinedTandBreps2 <- CombinedTandBreps[CombinedTandBreps$count_pctAcet_WT >=
minWTandCompObs &
                                          CombinedTandBreps$count_pctAcet_COMP >=
                                          minWTandCompObs,]

#manually set a new veriable for pvalue filtering, for values of 100%WT and 0 % Del set
to arbitrary significant
CombinedTandBreps2$pvalCategoryWTDEL <- CombinedTandBreps2$pval_WTDEL
CombinedTandBreps2$pvalCategoryWTDEL[CombinedTandBreps2$mean_pctAcet_WT == 100 &
                                       CombinedTandBreps2$mean_pctAcet_DEL == 0] <- 0.001

#calculate %RSD of pct acet
#if there are missing values, use appropriate area to calculate instead of pct acet
CombinedTandBreps2$RSD_pctAcet_WT <- CombinedTandBreps2$sd_pctAcet_WT /
CombinedTandBreps2$mean_pctAcet_WT
for (i in 1:nrow(CombinedTandBreps2)) {
  if ((CombinedTandBreps2$RSD_pctAcet_WT[i] == 0 &
       !is.na(CombinedTandBreps2$RSD_pctAcet_WT[i]))|
      (CombinedTandBreps2$mean_pctAcet_WT[i] == 0 &
       !is.na(CombinedTandBreps2$mean_pctAcet_WT[i]))) {
    if (CombinedTandBreps2$mean_pctAcet_WT[i] == 100) {
      CombinedTandBreps2$RSD_pctAcet_WT[i] <- CombinedTandBreps2$sd_Area_WT[i] /
        CombinedTandBreps2$mean_Area_WT[i]
    } else if (CombinedTandBreps2$mean_pctAcet_WT[i] == 0) {
      CombinedTandBreps2$RSD_pctAcet_WT[i] <- CombinedTandBreps2$sd_HeavyArea_WT[i] /
        CombinedTandBreps2$mean_HeavyArea_WT[i]
    }
  }
}

CombinedTandBreps2$RSD_pctAcet_DEL <- CombinedTandBreps2$sd_pctAcet_DEL /
CombinedTandBreps2$mean_pctAcet_DEL
for (i in 1:nrow(CombinedTandBreps2)) {
  if ((CombinedTandBreps2$RSD_pctAcet_DEL[i] == 0 &
       !is.na(CombinedTandBreps2$RSD_pctAcet_DEL[i])) |
      (CombinedTandBreps2$mean_pctAcet_DEL[i] == 0 &
       !is.na(CombinedTandBreps2$mean_pctAcet_DEL[i]))) {
    if (CombinedTandBreps2$mean_pctAcet_DEL[i] == 100) {
      CombinedTandBreps2$RSD_pctAcet_DEL[i] <- CombinedTandBreps2$sd_Area_DEL[i] /
        CombinedTandBreps2$mean_Area_DEL[i]
    } else if (CombinedTandBreps2$mean_pctAcet_DEL[i] == 0) {
      CombinedTandBreps2$RSD_pctAcet_DEL[i] <- CombinedTandBreps2$sd_HeavyArea_DEL[i] /
        CombinedTandBreps2$mean_HeavyArea_DEL[i]
    }
  }
}

CombinedTandBreps2$RSD_pctAcet_COMP <- CombinedTandBreps2$sd_pctAcet_COMP /
CombinedTandBreps2$mean_pctAcet_COMP
for (i in 1:nrow(CombinedTandBreps2)) {
  if ((CombinedTandBreps2$RSD_pctAcet_COMP[i] == 0 &
       !is.na(CombinedTandBreps2$RSD_pctAcet_COMP[i])) |
      (CombinedTandBreps2$mean_pctAcet_COMP[i] == 0 &
       !is.na(CombinedTandBreps2$mean_pctAcet_COMP[i]))) {
    if (CombinedTandBreps2$mean_pctAcet_COMP[i] == 100) {
      CombinedTandBreps2$RSD_pctAcet_COMP[i] <- CombinedTandBreps2$sd_Area_COMP[i] /
        CombinedTandBreps2$mean_Area_COMP[i]
    } else if (CombinedTandBreps2$mean_pctAcet_COMP[i] == 0) {
      CombinedTandBreps2$RSD_pctAcet_COMP[i] <- CombinedTandBreps2$sd_HeavyArea_COMP[i] /
        CombinedTandBreps2$mean_HeavyArea_COMP[i]
    }
  }
```

```
}

#calculate the absolute percentage change between WT and del
CombinedTandBreps2$absPctChange_WTvDel <- abs(CombinedTandBreps2$mean_pctAcet_WT -
CombinedTandBreps2$mean_pctAcet_DEL)

#export
write.table(CombinedTandBreps2, "AllMeasuredAcetylation1839.tsv", sep = "\t", row.names =
 F)

#create plotting dataframe requiring more acetlation in WT and comp than del
plot <- CombinedTandBreps2[CombinedTandBreps2$FC_pctAcet_WTDEL > 1 &
                           !is.na(CombinedTandBreps2$FC_pctAcet_WTDEL) &
                           CombinedTandBreps2$FC_pctAcet_COMPDEL > 1 &
                           !is.na(CombinedTandBreps2$FC_pctAcet_COMPDEL),]

#filters. At least 20 absolute pct diff between WT and del
#max p val of 0.05
abs_pct_change_min <- 20
pval_cutoff <- 0.05

#filtering
plot <- plot[plot$absPctChange_WTvDel >= abs_pct_change_min,]
plot <- plot[plot$pvalCategoryWTDEL <= pval_cutoff,]

#pivot to longer format for plotting
plotLong <- plot %>% pivot_longer(cols = (starts_with("mean_pctAcet_")|
                                            starts_with("count_pctAcet_") |
                                            starts_with("sd_pctAcet_") |
                                            starts_with("RSD_pctAcet_")),
                                  names_to = c(".value", "strain"),
                                  names_pattern = "(.*)_pctAcet_(.*)")

#set order
plotLong$strain <- factor(plotLong$strain, levels = c("WT", "DEL", "COMP"))

#set NA RSDs to high (1)
plotLong$RSD[is.na(plotLong$RSD)] <- 1

#plot
ggplot(plotLong) +
  geom_line(aes(x = strain, y = mean,
                group = acetSeq, color = gene),
            lwd = 1.2) +
  geom_point(aes(x = strain, y = mean,
                 group = acetSeq, fill = gene,
                 size = ifelse(RSD < 0.2,
                               "<20%", ifelse(RSD < 0.5,
                                              "<50%", ">50%")),
                 shape = ifelse(pvalCategoryWTDEL <= 0.001,
                                "<= 0.001", ifelse(pvalCategoryWTDEL <= 0.01,
                                                   "<= 0.01", ifelse(pvalCategoryWTDEL <=
                                                    0.05,
                                                                     "<= 0.05", "> 0.05"
                                                    )))),
             stroke = 1.4) +
  theme_bw(base_size = 20) +
  theme(legend.position = "right",
        panel.grid = element_blank()) +
  labs(y = "Percent N-Acetylation",
       x = element_blank(),
       shape = "p-value\n(WT vs DEL)",
       size = "%RSD",
       title = paste0(STRAIN, " (2 Bio Rep x 3 Tech Rep)")) +
  scale_colour_discrete(guide = "none") +
  scale_fill_discrete(guide = "none") +
  scale_size_manual(values = c(5,3.5,2)) +
  scale_shape_manual(values = c(21, 22, 23, 24)) +
  geom_label_repel(data = plotLong[plotLong$strain == "WT",],
```

```
                    aes(x = strain, y = mean,
                        label = gene, color = gene),
                    nudge_x = -1.3, alpha = 0.9,
                    size = 3.5,
                    max.overlaps = 50) +
  coord_cartesian(ylim = c(0,120))




#plot
ggplot(plotLong) +
  geom_line(aes(x = strain, y = mean,
                group = acetSeq, color = ifelse(pvalCategoryWTDEL <= 0.001,
                                                "<= 0.001", ifelse(pvalCategoryWTDEL <=
                                                0.01,
                                                                   "<= 0.01", ifelse(
                                                                   pvalCategoryWTDEL <=
                                                                   0.05,
                                                                                     "<=
0.05", "> 0.05")))),
            lwd = 1.2) +
  geom_point(aes(x = strain, y = mean,
                 group = acetSeq,
                 size = ifelse(RSD < 0.2,
                               "<20%", ifelse(RSD < 0.5,
                                              "<50%", ">50%")),
                 fill = ifelse(pvalCategoryWTDEL <= 0.001,
                               "<= 0.001", ifelse(pvalCategoryWTDEL <= 0.01,
                                                  "<= 0.01", ifelse(pvalCategoryWTDEL <=
                                                  0.05,
                                                                    "<= 0.05", "> 0.05"
                                                  )))),
             stroke = 1.4, shape = 21) +
  theme_bw(base_size = 30) +
  theme(legend.position = "right",
        panel.grid = element_blank()) +
  labs(y = "Percent N-Acetylation",
       x = element_blank(),
       color = "p-value\n(WT vs DEL)",
       fill = "p-value\n(WT vs DEL)",
       size = "%RSD",
       title = element_blank())+
  scale_color_discrete(guide = "none") +
  scale_size_manual(values = c(5,3.5,2)) +
  geom_label_repel(data = plotLong[plotLong$strain == "WT",],
                   aes(x = strain, y = mean,
                       label = gene, color = ifelse(pvalCategoryWTDEL <= 0.001,
                                                    "<= 0.001", ifelse(pvalCategoryWTDEL
                                                    <= 0.01,
                                                                       "<= 0.01", ifelse(
                                                                       pvalCategoryWTDEL
                                                                       <= 0.05,

"<= 0.05", "> 0.05")))),
                   nudge_x = -1.3,
                   size = 3.5,
                   max.overlaps = 50) +
  coord_cartesian(ylim = c(0,120)) +
  scale_y_continuous(breaks = c(0, 25, 50, 75, 100))

ggsave("1839_acetylation_change_20pctAbs_pval0_05.png", width = 15, height = 12)
#export plotted table
write.table(plot, "1839_acetylation_reported.tsv", sep = "\t", row.names = F)

#categorize rsd
plotLong$RSDCat <- ifelse(plotLong$RSD < 0.2,
                          "<20%", ifelse(plotLong$RSD < 0.5,
```

```
                                                "<50%", ">50%"))
#plot one color
ggplot(plotLong) +
  geom_hline(yintercept = c(0,100), linetype = 2, lwd = 1, color = "grey") +
  geom_line(aes(x = strain, y = mean,
                group = acetSeq),
            lwd = 1.4, color = "black") +
  geom_line(aes(x = strain, y = mean,
                group = acetSeq),
            lwd = 0.8, color = green) +
  geom_point(aes(x = strain, y = mean,
                 group = acetSeq,
                 size = RSDCat,
                 shape = RSDCat),
             stroke = 1.4,
             fill = green) +
  theme_bw(base_size = 30) +
  theme(legend.position = "right",
        panel.grid = element_blank()) +
  labs(y = "Percent N-Acetylation",
       x = element_blank(),
       color = "p-value\n(WT vs DEL)",
       fill = "p-value\n(WT vs DEL)",
       size = "%RSD",
       shape = "%RSD",
       title = element_blank())+
  scale_color_discrete(guide = "none") +
  scale_size_manual(values = c(6,4,2)) +
  geom_label_repel(data = plotLong[plotLong$strain == "WT",],
                   aes(x = strain, y = mean,
                       label = gene),
                   nudge_x = -1.3,
                   size = 3.5,
                   max.overlaps = 50) +
  coord_cartesian(ylim = c(0,120)) +
  scale_y_continuous(breaks = c(0, 25, 50, 75, 100)) +
  scale_shape_manual(values = c(21,22,23)) +
  scale_x_discrete(labels = c("WT", expression(Delta * italic(emp1)),
                              expression(Delta * italic(emp1)* "/comp")))

ggsave("1839_acetylation_change_20pctAbs_pval0_05_ONECOLOR.png", width = 15, height = 12)


#nonCanonicalN-terms
#pull accessions and genes
peptidesLong$Accession2 <- str_remove(str_extract(peptidesLong$Accession, "^[^\\|]*\\|")
, "\\|")
peptidesLong$gene <- str_remove(str_remove(str_extract(peptidesLong$Accession,
"^[^\\|]*\\|[^\\|]*\\|"),
                                                   "^[^\\|]*\\|"), "\\|")

#pull acetylated peptides starting after position 2
nonCanonicals <- peptidesLong[!peptidesLong$UnLabelled,]
nonCanonicals <- nonCanonicals[nonCanonicals$Start > 2,]

#filter out extra strains
nonCanonicals <- nonCanonicals[nonCanonicals$strain != "dCb",]

#light peptides
LightNonCans <- nonCanonicals[nonCanonicals$NtermAcet_L,]
#same acetylation string as above
LightNonCans$acetSeq <- str_replace(LightNonCans$Peptide,
                                    "\\(\\+42\\.01\\)",
                                    "(ac)")
#heavy peptides
HeavyNonCans <- nonCanonicals[nonCanonicals$NtermAcet_H,]
HeavyNonCans$acetSeq <- str_replace(HeavyNonCans$Peptide,
                                    "\\(\\+45\\.03\\)",
                                    "(ac)")
```

```r

#combine heavy and lights
NonCans2 <- LightNonCans %>% left_join(HeavyNonCans %>% select(c("Accession2", "acetSeq", "strain", "Inj", "Area")),
                                       by = c("Accession2", "acetSeq", "strain", "Inj"),
                                       suffix = c("_L", "_H"))

#set missing data to 0 for calc
NonCans2$Area_H[is.na(NonCans2$Area_H)] <- 0
NonCans2$Area_L[is.na(NonCans2$Area_L)] <- 0

#percent acet
NonCans2$pctAcet <- NonCans2$Area_L / (NonCans2$Area_H + NonCans2$Area_L)

#count number of valid values and mean + sd of pct acet
NonCans3 <- NonCans2 %>% group_by(Accession2, acetSeq, strain, Start) %>%
  summarise(countValid = sum(!is.nan(pctAcet)),
            meanPctAcet = mean(pctAcet, na.rm = T),
            sdPctAcet = sd(pctAcet, na.rm = T))

#set order
NonCans3$strain <- factor(NonCans3$strain, levels = c("WT", "d1839", "c1839"))

#pull out WT to pick examples
lightNonCansWT <- LightNonCans[LightNonCans$strain == 'WT',] %>%
  group_by(Accession2, acetSeq) %>%
  summarise(valid = sum(!is.na(Area)))

#require 2 valid WT measurements
lightNonCansWT <- lightNonCansWT[lightNonCansWT$valid > 1,]
lightNonCansAccessions <- unique(lightNonCansWT$Accession2)

#pull out the peptide with the maximum nubmer of valid WT values for each accession
lightNonCansWTKeeps <- lightNonCansWT %>% group_by(Accession2) %>%
  slice_max(valid, with_ties = F) %>%
  ungroup()
NonCans4 <- NonCans3[NonCans3$acetSeq %in% lightNonCansWTKeeps$acetSeq,]

#create string for plotting
NonCans4$startString <- paste0("Start: ", NonCans4$Start)
ggplot(NonCans4) +
  geom_point(aes(x = strain, y = meanPctAcet, color = Accession2,
                 size = countValid)) +
  geom_errorbar(aes(x = strain, ymin = meanPctAcet - sdPctAcet,
                    ymax = meanPctAcet + sdPctAcet, color = Accession2), width = 0.2) +
  geom_line(aes(x = strain, y = meanPctAcet, color = Accession2, group = acetSeq)) +
  theme_bw(base_size = 8) +
  theme(legend.position = "none") +
  coord_cartesian(ylim = c(0,1.1)) +
  facet_wrap(c("Accession2", "acetSeq", "startString"))

ggsave("AllNonCanonicalExamples.png", width = 15, height = 15)

#pull out examples
Examples <- c("MMAR_0359", "MMAR_1369", "MMAR_1730", "MMAR_3090", "MMAR_3818", "MMAR_5064", "MMAR_2201",
              "MMAR_1025", "MMAR_4849")

ExamplePlot <- NonCans4[NonCans4$Accession2 %in% Examples,]

ExamplePlot$countValid <- factor(ExamplePlot$countValid)

ExamplePlot$Accession2 <- factor(ExamplePlot$Accession2, levels = Examples)

#plot
ggplot(ExamplePlot) +
  geom_hline(yintercept = c(0,1), linetype = 2, lwd = 1, color = "grey") +
  geom_errorbar(aes(x = strain, ymin = meanPctAcet - sdPctAcet,
                    ymax = meanPctAcet + sdPctAcet,), width = 0.25, lwd = 1.5) +
```

```
    geom_errorbar(aes(x = strain, ymin = meanPctAcet - sdPctAcet,
                      ymax = meanPctAcet + sdPctAcet, color = Accession2), width = 0.2) +
    geom_line(aes(x = strain, y = meanPctAcet, group = acetSeq), lwd = 3) +
    geom_line(aes(x = strain, y = meanPctAcet, color = Accession2, group = acetSeq), lwd =
    1) +
    geom_point(aes(x = strain, y = meanPctAcet, fill = Accession2,
                   size = countValid), shape = 21, stroke = 2) +
    theme_bw(base_size = 20) +
    theme(legend.position = "none",
          panel.grid = element_blank()) +
    coord_cartesian(ylim = c(0,1.2)) +
    facet_wrap(c("Accession2", "acetSeq", "startString")) +
    scale_size_manual(values = c(5, 6, 7,8)) +
    scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1)) +
    labs(x = element_blank(), y = "Proportion Acetylated") +
    scale_x_discrete(labels = c("WT", expression(Delta * italic(emp1)),
                                expression(Delta * italic(emp1)* "/comp")))


ggsave("NonCanonicalN-terminaiExamples.png", width = 15, height = 15)
```