

```

1  setwd("~/04_Champion_Lab/02_N-terminal_Acetylation/OnePot")
2
3  #import libraries
4  library(readxl)
5  library(stringr)
6  library(tidyr)
7  library(dplyr)
8  library(seqinr)
9  library(Peptides)
10 library(cleaver)
11 library(ggplot2)
12 library(scales)
13
14 #set colors
15 red <- hue_pal()(3)[1]
16 green <- hue_pal()(3)[2]
17 blue <- hue_pal()(3)[3]
18
19
20 #read in data. one pot, enriched data (trypsin), genome
21 op <- read.csv("Data/OnePotSummarisedData.csv")
22 enr <- read.csv("Data/EnrichedSummarisedData.csv")
23 gen <- read.csv("Data/GenomeSummarisedData_GASTVCP.csv")
24
25
26
27 #filters
28 minStart <- 2
29 minLength <- 5
30 maxLength <- 50
31
32 #filter datasets
33 opN <- op[op$Start <= minStart &
34           op$Length >= minLength &
35           op$Length <= maxLength,]
36 enr <- enr[enr$Start <= minStart &
37            enr$Length >= minLength &
38            enr$Length <= maxLength,]
39 gen <- gen[gen$Length >= minLength &
40            gen$Length <= maxLength,]
41
42 #only keep columns of interest
43 keeps <- c("Peptide", "IEP")
44 opN <- opN[,names(opN) %in% keeps]
45 enr <- enr[,names(enr) %in% keeps]
46 gen <- gen[,names(gen) %in% keeps]
47
48 #set labels
49 opN$category <- "One Pot (This Work)"
50 enr$category <- "Enriched (Thompson et al.)"
51 gen$category <- "Genome"
52
53 #create combined dataset with category labels
54 combined <- bind_rows(opN, enr, gen)
55 combined$category <- factor(combined$category, levels = c("Genome",
56                                                            "One Pot (This Work)",
57                                                            "Enriched (Thompson et al.)"))
58
59 #create plots and save
60 ggplot(combined) +
61   geom_density(aes(x = IEP, color = category),
62                linewidth = 1.4) +
63   theme_bw(base_size = 15) +
64   theme(panel.grid = element_blank(),
65         axis.text.y = element_blank(),
66         axis.ticks.y = element_blank()) +
67   labs(x = "Isoelectric Point",
68        y = "N-term Peptide Frequency",
69        color = element_blank())

```

```

70 gggsave("plots/GASCTPV/DensityPlotNterminalPeptidesALL.png",
71         width = 12, height = 8)
72
73
74 ggplot(combined[combined$category %in% c("Genome", "One Pot (This Work)", "Enriched
(Thompson et al.)"),]) +
75   geom_density(aes(x = IEP, color = category),
76               linewidth = 1.4) +
77   theme_bw(base_size = 15) +
78   theme(panel.grid = element_blank(),
79         axis.text.y = element_blank(),
80         axis.ticks.y = element_blank()) +
81   labs(x = "Isoelectric Point",
82        y = "N-term Peptide Frequency",
83        color = element_blank())
84
85 gggsave("plots/GASCTPV/DensityPlotNterminalPeptides_OP-ENR-GenR.png",
86         width = 12, height = 8)
87
88
89
90 ggplot(combined) +
91   geom_histogram(aes(x = IEP, fill = category),
92                 linewidth = 1.4, color = "black", binwidth = 0.2) +
93   theme_bw(base_size = 15) +
94   theme(panel.grid = element_blank(),
95         legend.position = "none") +
96   labs(x = "Isoelectric Point",
97        y = "N-term Peptide Frequency",
98        color = element_blank()) +
99   facet_grid(category~.)
100
101 gggsave("plots/GASCTPV/HistogramPlotNterminalPeptides_All.png",
102         width = 12, height = 16)
103
104
105
106 ggplot(combined[combined$category %in% c("Genome", "One Pot (This Work)", "Enriched
(Thompson et al.)"),]) +
107   geom_histogram(aes(x = IEP, fill = category),
108                 linewidth = 1.4, color = "black", binwidth = 0.2) +
109   theme_bw(base_size = 30) +
110   theme(panel.grid = element_blank(),
111         legend.position = "none") +
112   labs(x = "Isoelectric Point",
113        y = "N-term Peptide Frequency",
114        color = element_blank()) +
115   facet_grid(category~.)
116
117 gggsave("plots/GASCTPV/HistogramPlotNterminalPeptides_OP-ENR-GenR_sizedUp.png",
118         width = 12, height = 16)
119
120
121
122
123
124 ###Neo vs N-term Genome and OP
125 op <- read.csv("Data/OnePotSummarisedData.csv")
126 gen <- read.csv("Data/GenomeSummarisedData_allPeptides_GASTVCP.csv")
127
128 #filters
129 minLength <- 5
130 maxLength <- 50
131
132
133 #filter datasets
134 op <- op[op$Length >= minLength &
135         op$Length <= maxLength,]
136

```

```

137 #determine whether peptide is Nterminal
138 op$NtermPeptide <- F
139 op$NtermPeptide[op$Start <=2] <- T
140
141 #name columns to match PEAKS format
142 op$mass <- op$Mass
143 op$Accession <- op$Accessions
144 op$Peptide <- op$str_seq
145
146 #pull out only columns that are in the genome dataset
147 op <- op[,which(names(op) %in% names(gen))]
148
149
150 #set categories
151 op$origin <- "One Pot"
152 gen$origin <- "Genome"
153 op$cat <- "All"
154 gen$cat <- "All"
155 op$facetVar <- "One Pot All"
156 gen$facetVar <- "Genome All"
157 gen_Nterm <- gen[gen$NtermPeptide,]
158 gen_Nterm$facetVar <- "Genome N-termini"
159 gen_Nterm$cat <- "N-termini"
160 op_Nterm <- op[op$NtermPeptide,]
161 op_Nterm$facetVar <- "One Pot N-termini"
162 op_Nterm$cat <- "N-termini"
163
164
165 gen_NonNterm <- gen[!gen$NtermPeptide,]
166 gen_NonNterm$facetVar <- "Genome Neo Peptides"
167 gen_NonNterm$cat <- "Neo Peptides"
168 op_NonNterm <- op[!op$NtermPeptide,]
169 op_NonNterm$facetVar <- "One Pot Neo Peptides"
170 op_NonNterm$cat <- "Neo Peptides"
171
172 #bind rows together
173 master <- bind_rows(gen, op, gen_Nterm, gen_NonNterm, op_Nterm, op_NonNterm)
174
175 #create and save plots
176 ggplot(master) +
177   geom_histogram(aes(x = IEP, fill = facetVar),
178                 linewidth = 1.4, color = "black", binwidth = 0.2)+
179   theme_bw(base_size = 20) +
180   theme(panel.grid = element_blank(),
181         legend.position = "none") +
182   labs(x = "Isoelectric Point",
183        y = "Peptide Frequency",
184        color = element_blank()) +
185   facet_grid(facetVar~., scales = "free")
186
187
188 ggsave("plots/GASCTPV/Histogram_IEP_neo_Ntermini_All_OPvsGen_6x1.png", width = 12,
189        height = 20)
190
191
192
193
194 ggplot(master) +
195   geom_histogram(aes(x = IEP, fill = cat),
196                 linewidth = 1.4, color = "black", binwidth = 0.2)+
197   theme_bw(base_size = 20) +
198   theme(panel.grid = element_blank(),
199         legend.position = "none") +
200   labs(x = "Isoelectric Point",
201        y = "Peptide Frequency",
202        color = element_blank()) +
203   facet_grid(cat~origin, scales = "free")
204
205 ggsave("plots/GASCTPV/Histogram_IEP_neo_Ntermini_All_OPvsGen_3x2.png", width = 12,

```

```

206         height = 12)
207
208
209
210
211
212
213
214
215
216
217
218
219
220 ggplot(master) +
221     geom_histogram(aes(x = mass, fill = facetVar),
222                    linewidth = 1.4, color = "black", binwidth = 100)+
223     theme_bw(base_size = 20) +
224     theme(panel.grid = element_blank(),
225            legend.position = "none") +
226     labs(x = "Mass (Da)",
227           y = "Peptide Frequency",
228           color = element_blank()) +
229     facet_grid(facetVar~., scales = "free")
230
231
232 ggsave("plots/GASCTPV/Histogram_mass_neo_Ntermini_All_OPvsGen_6x1.png", width = 12,
233         height = 20)
234
235
236
237
238 ggplot(master) +
239     geom_histogram(aes(x = mass, fill = cat),
240                    linewidth = 1.4, color = "black", binwidth = 100)+
241     theme_bw(base_size = 20) +
242     theme(panel.grid = element_blank(),
243            legend.position = "none") +
244     labs(x = "Mass (Da)",
245           y = "Peptide Frequency",
246           color = element_blank()) +
247     facet_grid(cat~origin, scales = "free")
248
249 ggsave("plots/GASCTPV/Histogram_mass_neo_Ntermini_All_OPvsGen_3x2.png", width = 12,
250         height = 12)
251
252
253
254 ggplot(master) +
255     geom_histogram(aes(x = Length, fill = facetVar),
256                    linewidth = 1.4, color = "black", binwidth = 1)+
257     theme_bw(base_size = 20) +
258     theme(panel.grid = element_blank(),
259            legend.position = "none") +
260     labs(x = "Length (AAs)",
261           y = "Peptide Frequency",
262           color = element_blank()) +
263     facet_grid(facetVar~., scales = "free")
264
265
266 ggsave("plots/GASCTPV/Histogram_length_neo_Ntermini_All_OPvsGen_6x1.png", width = 12,
267         height = 20)
268
269
270
271
272 ggplot(master) +
273     geom_histogram(aes(x = Length, fill = cat),
274                    linewidth = 1.4, color = "black", binwidth = 1)+

```

```

275     theme_bw(base_size = 20) +
276     theme(panel.grid = element_blank(),
277           legend.position = "none") +
278     labs(x = "Length (AAs)",
279          y = "Peptide Frequency",
280          color = element_blank()) +
281     facet_grid(cat~origin, scales = "free")
282
283 ggsave("plots/GASCTPV/Histogram_length_neo_Ntermini_All_OPvsGen_3x2.png", width = 12,
284        height = 12)
285
286
287 #same as above but only one pot data for retention time analysis
288 op <- read.csv("Data/OnePotSummarisedData.csv")
289
290 #filters
291 minLength <- 5
292 maxLength <- 50
293
294
295
296 op <- op[op$Length >= minLength &
297         op$Length <= maxLength,]
298
299 op$NtermPeptide <- F
300 op$NtermPeptide[op$Start <=2] <- T
301
302 op$mass <- op$Mass
303 op$Accession <- op$Accessions
304
305 op$Peptide <- op$str_seq
306
307
308 op$RTfacet <- "Neo Peptides"
309 op$RTfacet[op$NtermPeptide] <- "N-termini"
310
311 ggplot(op) +
312   geom_histogram(aes(x = RT, fill = RTfacet),
313                 linewidth = 1.4, color = "black", binwidth = 1)+
314   theme_bw(base_size = 40) +
315   theme(panel.grid = element_blank(),
316         legend.position = "none") +
317   labs(x = "RT (min)",
318        y = "Peptide Frequency",
319        title = "One Pot Peptides",
320        color = element_blank()) +
321   scale_fill_manual(values = c(green,blue)) +
322   facet_grid(RTfacet~., scales = "free")
323
324 ggsave("plots/GASCTPV/RT_histogram_onePot_neovsNterm.png", width = 12, height = 12)
325
326 ggplot(op) +
327   geom_density(aes(x = RT, color = RTfacet,
328                   fill = RTfacet),
329               lwd = 3, alpha = 0.3)+
330   theme_bw(base_size = 40) +
331   theme(panel.grid = element_blank(),
332         legend.position = "top",
333         axis.ticks.y = element_blank(),
334         axis.text.y = element_blank()) +
335   labs(x = "RT (min)",
336        y = "Peptide Frequency (Density)",
337        title = element_blank(),
338        color = element_blank()) +
339   scale_color_manual(values = c(green,blue)) +
340   scale_fill_manual(guide = "none",
341                     values = c(green,blue))
342
343

```

```

344 ggsave("plots/GASCTPV/RT_Density_onePot_neovsNterm.png", width = 12, height = 10)
345
346
347
348 #create dataframe for amino acid analysis
349 master2 <- master[master$cat != "All",]
350
351 #all amino acids
352 AAs <- c("A", "G", "I", "L", "P", "V", "F", "W", "Y", "D", "E", "R", "H", "K", "S", "T",
353 "C", "M", "N", "Q")
354
355 #create variables for counting and percentage
356 for (AA in AAs) {
357   master2[[paste0(AA, "_count")]] <- str_count(master2$Peptide, AA)
358   master2[[paste0(AA, "_pct")]] <- str_count(master2$Peptide, AA) / master2$Length
359 }
360
361 #create long strings of all peptides in each category to count AAs
362 OP_NtermString <- paste(master2$Peptide[master2$cat == "N-termini" &
363   master2$origin == "One Pot"], collapse = "")
364
365 OP_NeoString <- paste(master2$Peptide[master2$cat == "Neo Peptides" &
366   master2$origin == "One Pot"], collapse = "")
367
368 GEN_NtermString <- paste(master2$Peptide[master2$cat == "N-termini" &
369   master2$origin == "Genome"], collapse = "")
370
371 GEN_NeoString <- paste(master2$Peptide[master2$cat == "Neo Peptides" &
372   master2$origin == "Genome"], collapse = "")
373
374
375 #create AA dataframe
376 AA_df <- data.frame(expand.grid(AA = AAs, cat = unique(master2$cat), origin = unique(
377   master2$origin)))
378 AA_df$AA <- as.character(AA_df$AA)
379 AA_df$count = NA
380 AA_df$pct = NA
381
382 #loop through each category and count each AA and calculate percentage.
383 for (i in 1:nrow(AA_df)) {
384   if (AA_df$cat[i] == "N-termini" &
385     AA_df$origin[i] == "One Pot") {
386     AA_df$count[i] <- str_count(OP_NtermString, AA_df$AA[i])
387     AA_df$pct[i] <- AA_df$count[i] / nchar(OP_NtermString)
388   } else if (AA_df$cat[i] == "N-termini" &
389     AA_df$origin[i] == "Genome") {
390     AA_df$count[i] <- str_count(GEN_NtermString, AA_df$AA[i])
391     AA_df$pct[i] <- AA_df$count[i] / nchar(GEN_NtermString)
392   } else if (AA_df$cat[i] == "Neo Peptides" &
393     AA_df$origin[i] == "One Pot") {
394     AA_df$count[i] <- str_count(OP_NeoString, AA_df$AA[i])
395     AA_df$pct[i] <- AA_df$count[i] / nchar(OP_NeoString)
396   } else if (AA_df$cat[i] == "Neo Peptides" &
397     AA_df$origin[i] == "Genome") {
398     AA_df$count[i] <- str_count(GEN_NeoString, AA_df$AA[i])
399     AA_df$pct[i] <- AA_df$count[i] / nchar(GEN_NeoString)
400   }
401 }
402
403 #categorize AAs
404 AA_df$AAType <- NA
405 AA_df$AAType[AA_df$AA %in% c("A", "G", "I", "L", "M", "V")] <- "Aliphatic"
406 AA_df$AAType[AA_df$AA %in% c("F", "W", "Y")] <- "Aromatic"
407 AA_df$AAType[AA_df$AA %in% c("P", "S", "T", "C", "N", "Q")] <- "Polar"
408 AA_df$AAType[AA_df$AA %in% c("D", "E")] <- "Negative"
409 AA_df$AAType[AA_df$AA %in% c("R", "H", "K")] <- "Positive"
410
411 AA_df$AAType <- factor(AA_df$AAType, levels = c("Aliphatic", "Polar", "Aromatic",

```

```

411 "Negative",
412                                     "Positive"))
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

```