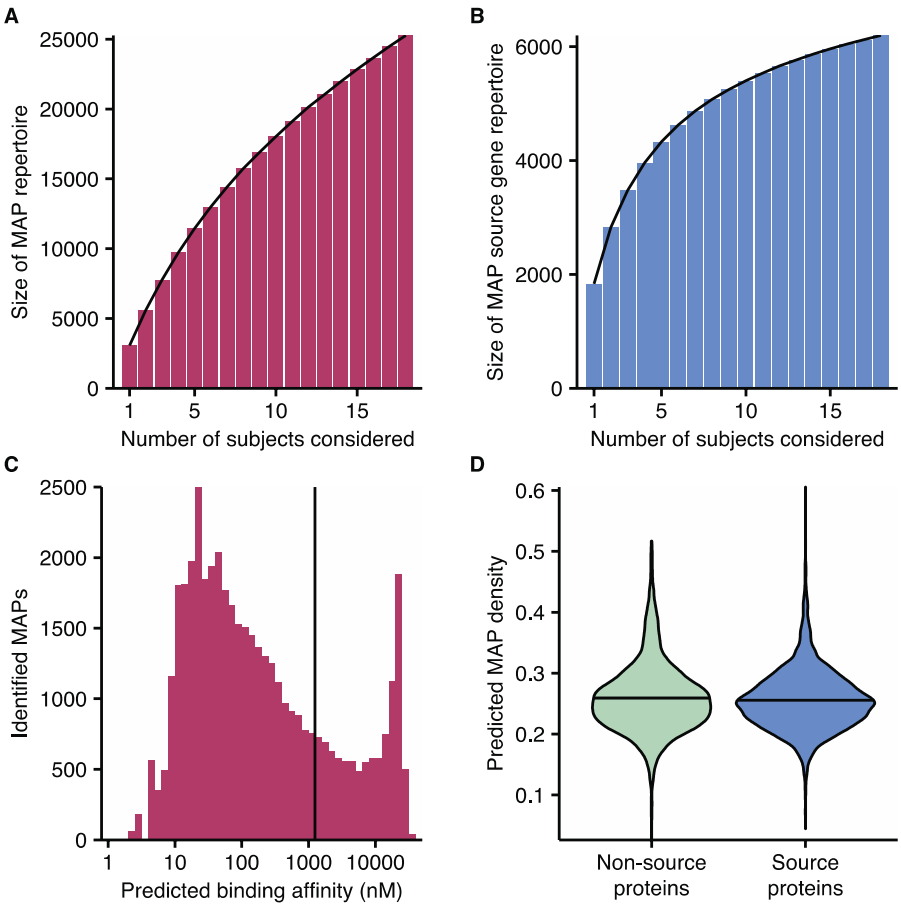


Supplemental Figure 1. Supplemental characterization of the immunopeptidome. The number of unique identifications of MAPs (A) and MAP source genes (B) was counted for each number of randomly selected HLA allotypes. Results show the average of 1000 simulations. Note: shared alleles between subjects increased the redundancy of peptides identified between subjects. (C) The predicted binding affinities of identified MAPs prior to application of the ≤ 1250nM filter. (D) Density of nonamer peptides predicted to bind any of the 27 HLA allotypes studied with an affinity ≤ 1250 nM in source and non-source proteins. No greater density of MAPs was found in MAP source genes (P = 0.99, one-tailed Student's t-Test); mean shown as horizontal line.

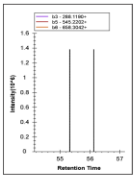
A

B-LCL # 1
HLA-A*03:01; A*29:02
HLA-B*08:01; B*44:03

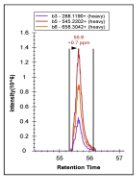
Non source proteins

Peptide 1: AESQEL**L**TF
HLA-B*44:03
Predicted affinity score: 22

Endogenous



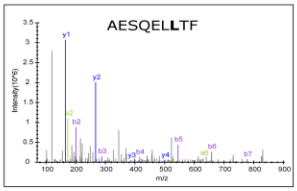
Heavy synthetic peptide



MS/MS Endogenous

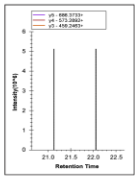
Not detected

MS/MS Heavy synthetic peptide

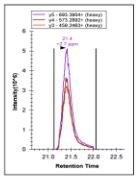


Peptide 2: EESH**L**NRHF
HLA-B*44:03
Predicted affinity score: 33

Endogenous



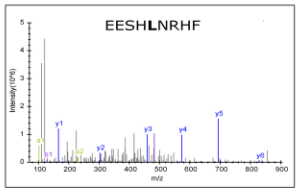
Heavy synthetic peptide



MS/MS Endogenous

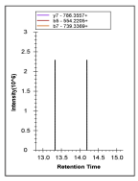
Not detected

MS/MS Heavy synthetic peptide

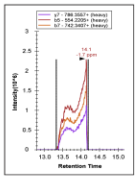


Peptide 3: HESAE**G**KEY
HLA-B*44:03
Predicted affinity score: 27

Endogenous



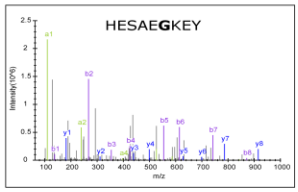
Heavy synthetic peptide



MS/MS Endogenous

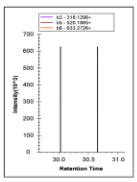
Not detected

MS/MS Heavy synthetic peptide

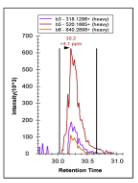


Peptide 4: TESSD**I**TEY
HLA-B*44:03
Predicted affinity score: 38

Endogenous



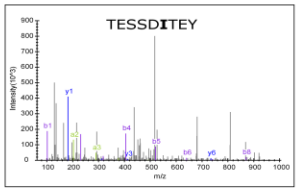
Heavy synthetic peptide



MS/MS Endogenous

Not detected

MS/MS Heavy synthetic peptide



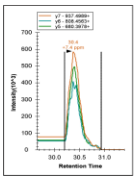
B

B-LCL # 1
HLA-A*03:01; A*29:02
HLA-B*08:01; B*44:03

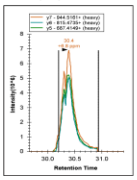
Source proteins

Peptide 1: AEIEQK**I**KEY
HLA-B*44:03
Predicted affinity score: 14

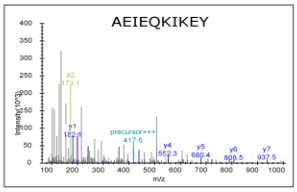
Endogenous



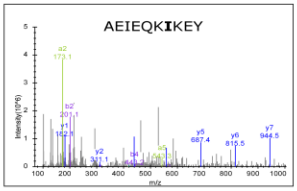
Heavy synthetic peptide



MS/MS Endogenous

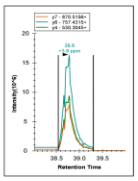


MS/MS Heavy synthetic peptide

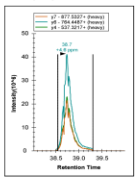


Peptide 2: EEINLQRN**I**
HLA-B*44:03
Predicted affinity score: 96

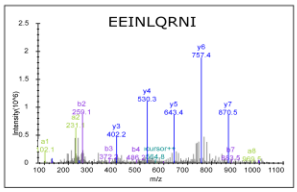
Endogenous



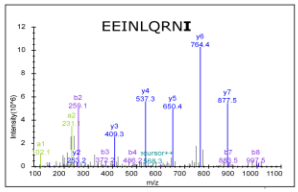
Heavy synthetic peptide



MS/MS Endogenous

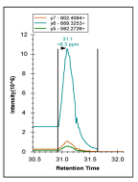


MS/MS Heavy synthetic peptide

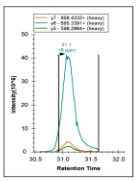


Peptide 3: EEIP**V**SSHY
HLA-B*44:03
Predicted affinity score: 15

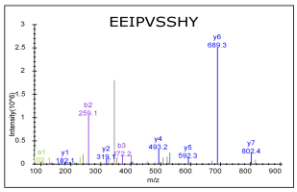
Endogenous



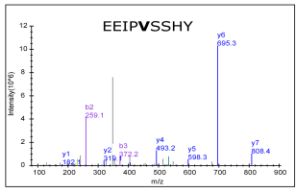
Heavy synthetic peptide



MS/MS Endogenous

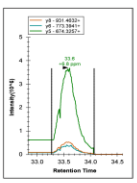


MS/MS Heavy synthetic peptide

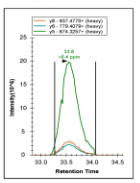


Peptide 4: EESA**V**PERSW
HLA-B*44:03
Predicted affinity score: 45

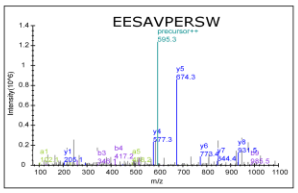
Endogenous



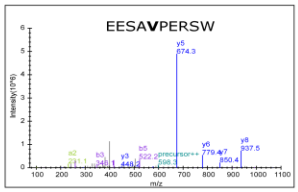
Heavy synthetic peptide



MS/MS Endogenous



MS/MS Heavy synthetic peptide



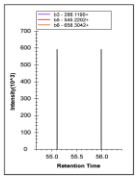
C

B-LCL # 2
HLA-A*03:01; A*29:02
HLA-B*08:01; B*44:03

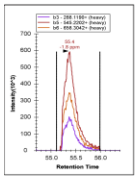
Non source proteins

Peptide 1: AESQELL**TF**
HLA-B*44:03
Predicted affinity score: 22

Endogenous



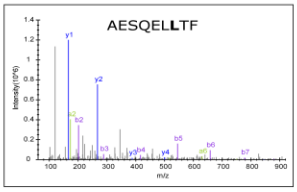
Heavy synthetic peptide



MS/MS Endogenous

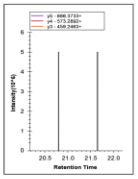
Not detected

MS/MS Heavy synthetic peptide

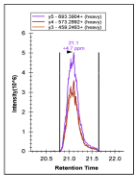


Peptide 2: EESHL**NRHF**
HLA-B*44:03
Predicted affinity score: 33

Endogenous



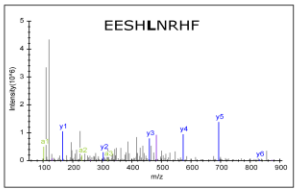
Heavy synthetic peptide



MS/MS Endogenous

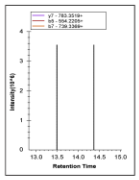
Not detected

MS/MS Heavy synthetic peptide

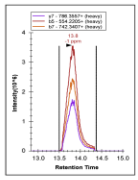


Peptide 3: HESAEG**KEY**
HLA-B*44:03
Predicted affinity score: 27

Endogenous



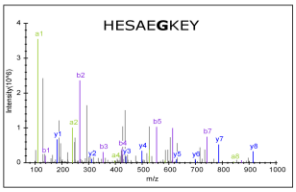
Heavy synthetic peptide



MS/MS Endogenous

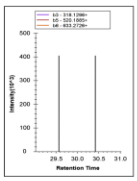
Not detected

MS/MS Heavy synthetic peptide

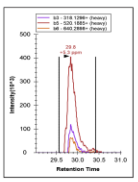


Peptide 4: TESSDI**TEY**
HLA-B*44:03
Predicted affinity score: 38

Endogenous



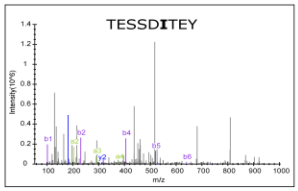
Heavy synthetic peptide



MS/MS Endogenous

Not detected

MS/MS Heavy synthetic peptide



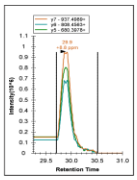
D

B-LCL # 2
HLA-A*03:01; A*29:02
HLA-B*08:01; B*44:03

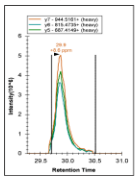
Source proteins

Peptide 1: AEIEQKI**KEY**
HLA-B*44:03
Predicted affinity score: 14

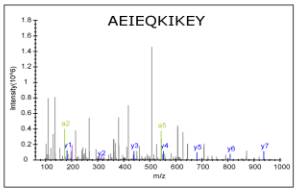
Endogenous



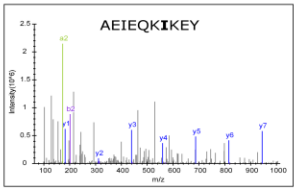
Heavy synthetic peptide



MS/MS Endogenous

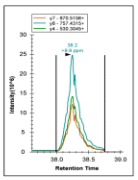


MS/MS Heavy synthetic peptide

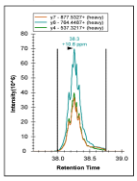


Peptide 2: EEINLQRNI
HLA-B*44:03
Predicted affinity score: 96

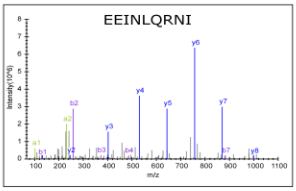
Endogenous



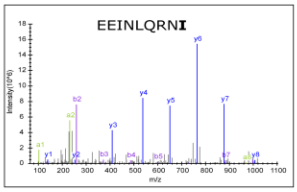
Heavy synthetic peptide



MS/MS Endogenous

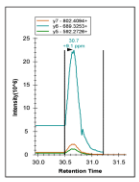


MS/MS Heavy synthetic peptide

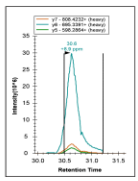


Peptide 3: EEIPV**SSHY**
HLA-B*44:03
Predicted affinity score: 15

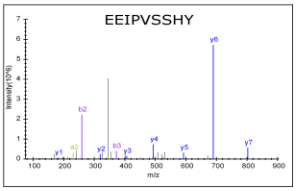
Endogenous



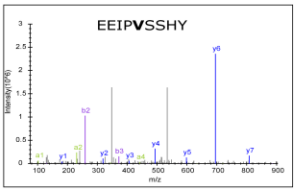
Heavy synthetic peptide



MS/MS Endogenous

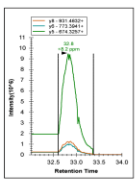


MS/MS Heavy synthetic peptide

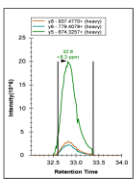


Peptide 4: EESAV**PERSW**
HLA-B*44:03
Predicted affinity score: 45

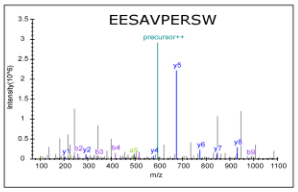
Endogenous



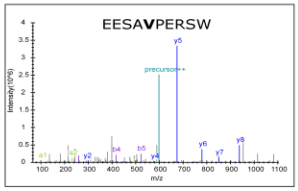
Heavy synthetic peptide



MS/MS Endogenous



MS/MS Heavy synthetic peptide



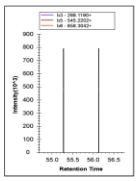
E

B-LCL # 6
HLA-A*02:01; A*11:01
HLA-B*40:01; B*44:03

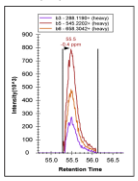
Non source proteins

Peptide 1: AESQEL**L**TF
HLA-B*44:03
Predicted affinity score: 22

Endogenous



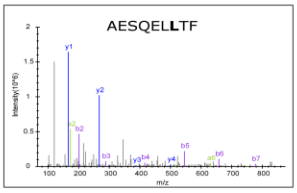
Heavy synthetic peptide



MS/MS Endogenous

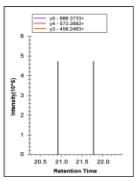
Not detected

MS/MS Heavy synthetic peptide

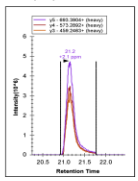


Peptide 2: EESH**L**NRHF
HLA-B*44:03
Predicted affinity score: 33

Endogenous



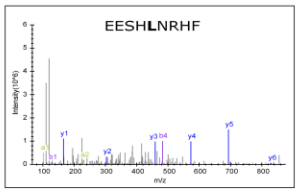
Heavy synthetic peptide



MS/MS Endogenous

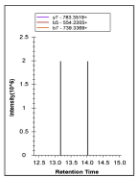
Not detected

MS/MS Heavy synthetic peptide

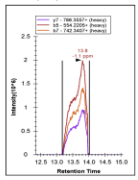


Peptide 3: HESAE**G**KEY
HLA-B*44:03
Predicted affinity score: 27

Endogenous



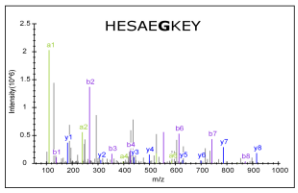
Heavy synthetic peptide



MS/MS Endogenous

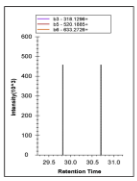
Not detected

MS/MS Heavy synthetic peptide

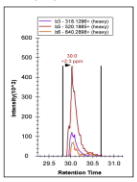


Peptide 4: TESSD**I**TEY
HLA-B*44:03
Predicted affinity score: 38

Endogenous



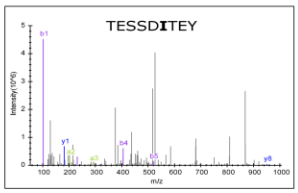
Heavy synthetic peptide



MS/MS Endogenous

Not detected

MS/MS Heavy synthetic peptide



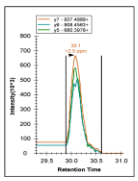
F

B-LCL # 6
HLA-A*02:01; A*11:01
HLA-B*40:01; B*44:03

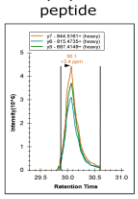
Source proteins

Peptide 1: AEIEQK**I**KEY
HLA-B*44:03
Predicted affinity score: 14

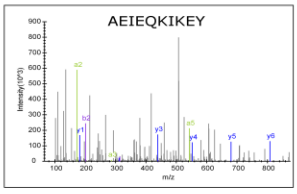
Endogenous



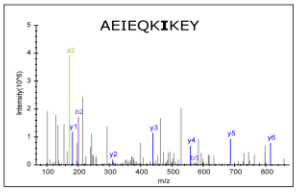
Heavy synthetic peptide



MS/MS Endogenous

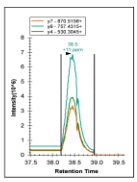


MS/MS Heavy synthetic peptide

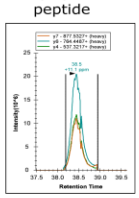


Peptide 2: EEINLQRN**I**
HLA-B*44:03
Predicted affinity score: 96

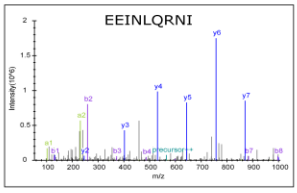
Endogenous



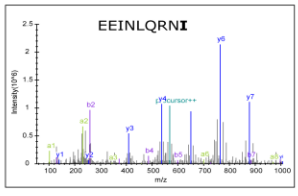
Heavy synthetic peptide



MS/MS Endogenous

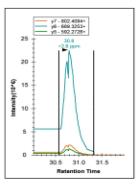


MS/MS Heavy synthetic peptide

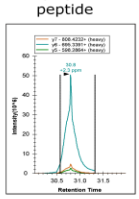


Peptide 3: EEIP**V**SSHY
HLA-B*44:03
Predicted affinity score: 15

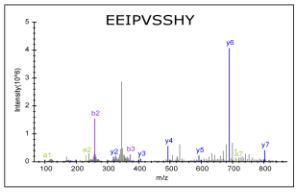
Endogenous



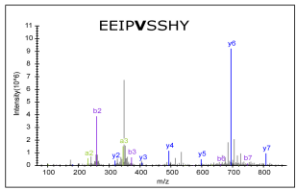
Heavy synthetic peptide



MS/MS Endogenous

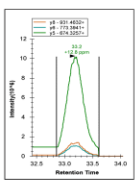


MS/MS Heavy synthetic peptide

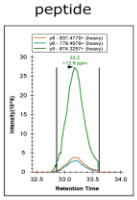


Peptide 4: EESA**V**PERSW
HLA-B*44:03
Predicted affinity score: 45

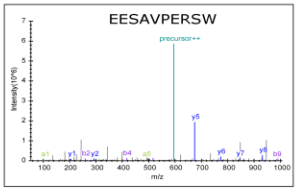
Endogenous



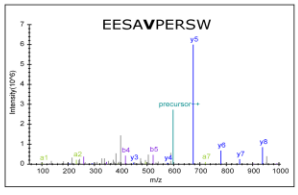
Heavy synthetic peptide



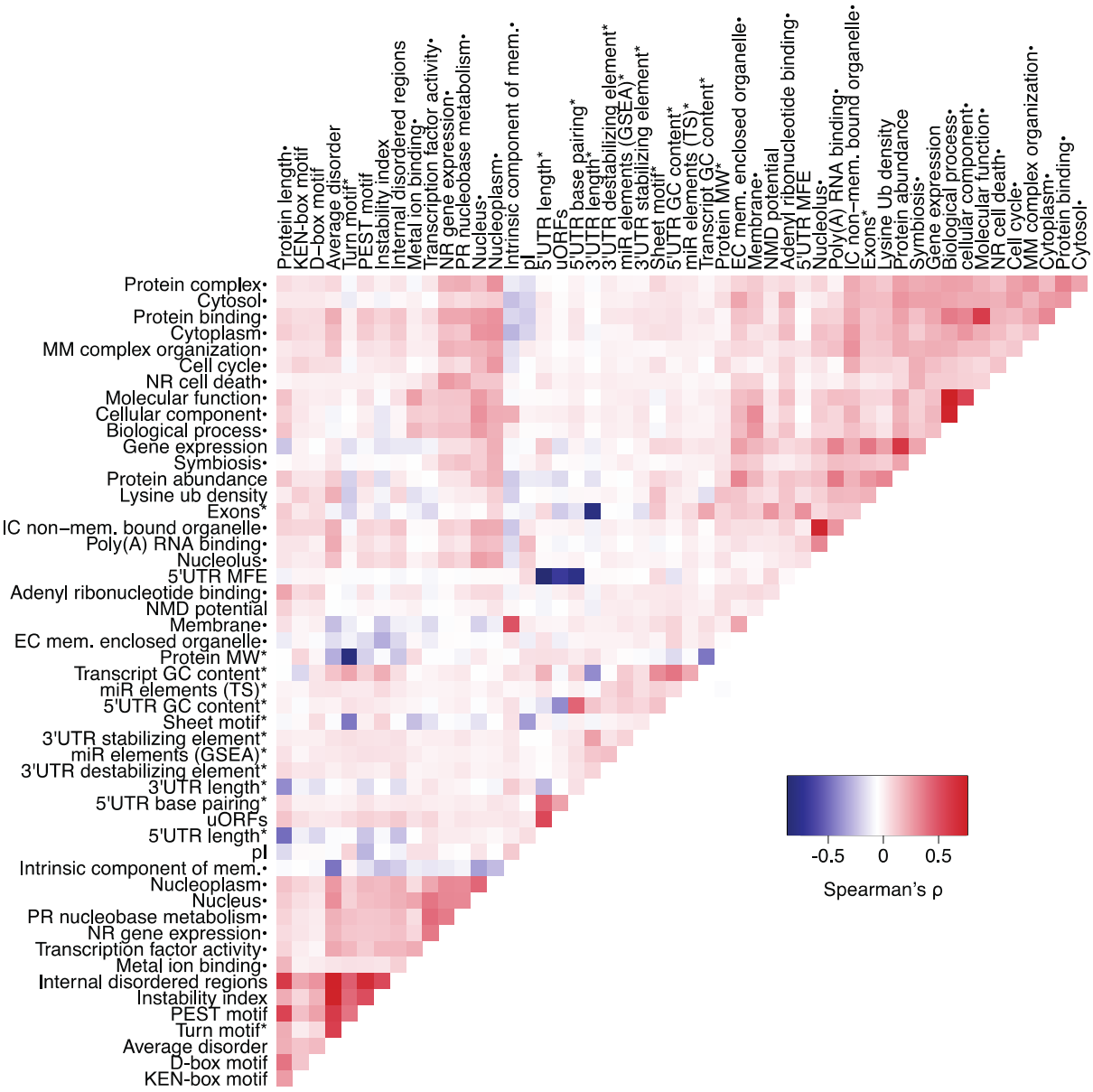
MS/MS Endogenous



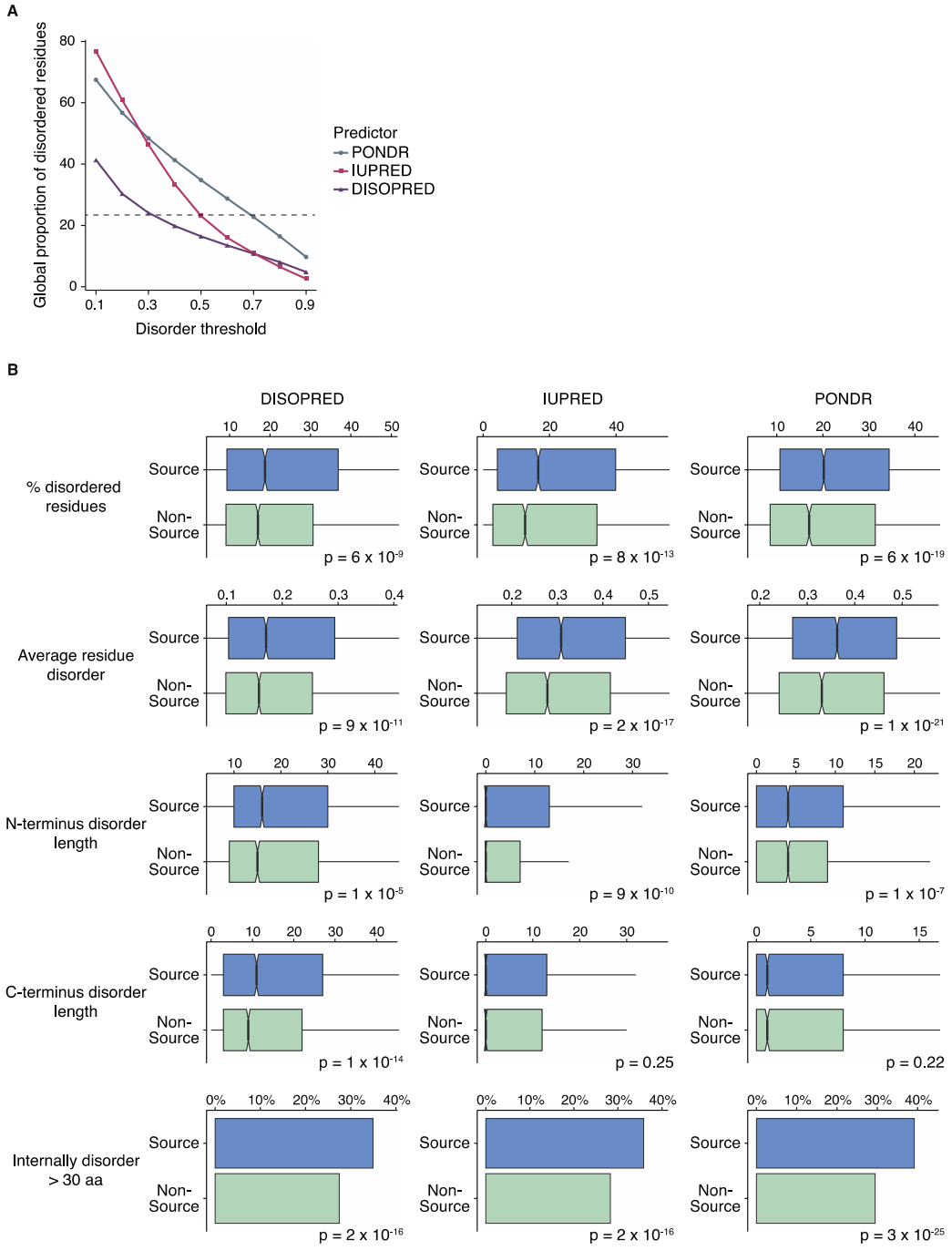
MS/MS Heavy synthetic peptide



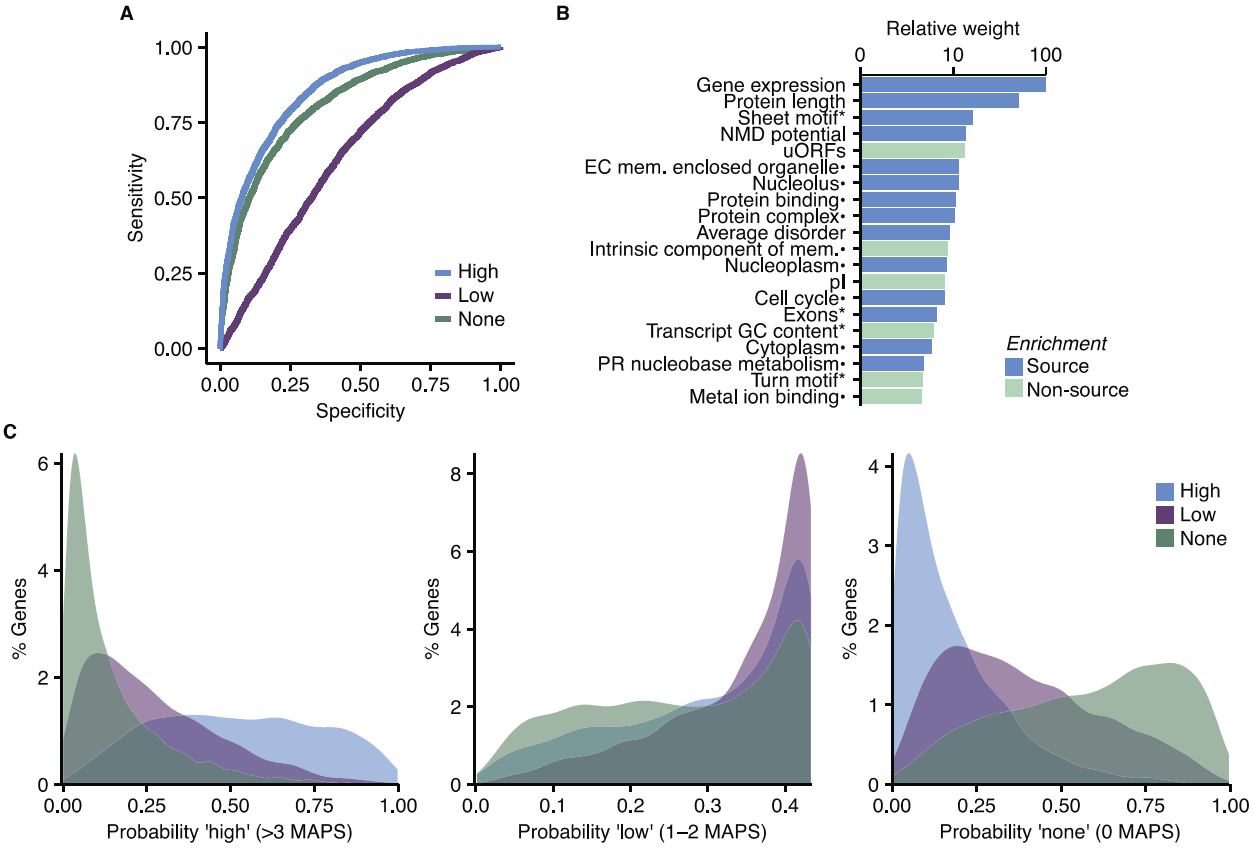
Supplemental Figure 2. IS-PRM analyses of peptides coded by MAP source and non-source genes. Isotopically labeled peptides were spiked in mild acid extracts from three different B-LCLs (BCL1, BCL2 and BCL6) expressing the B*44:03 allotype. Extracted ion chromatograms for fragment ions corresponding to light and heavy peptide ions are shown on the top panels along with their MS/MS spectra in middle and bottom panels, respectively. Data for peptides coded by non-source genes (AESQEL**L**TF, EESH**L**NRHF, HESAE**G**KEY, and TESSD**I**TEY; bold type indicate ^{13}C, ^{15}N-labeled amino acids) are depicted in panels A (BCL-1), C (BCL-2) and E (BCL-6). Data for peptides with the amino acid sequence of MAPs presented by B*44:03MAPs (AEIEQK**I**KEY, EEINLQRN**I**, EEIP**V**SSHY and EESA**V**PERSW) are depicted in panels B (BCL-1), D (BCL-2) and F (BCL-6).



Supplemental Figure 3. Correlation matrix of all model input variables using Spearman's coefficient ρ. Variables normalized by the length of the corresponding UTR, transcript or protein are denoted by *, GO terms denoted by •. EC : extracellular, GSEA: Gene Set Enrichment Analysis database, IC : intracellular, Mem. : membrane, MFE : minimum free energy, MM : macromolecular, NMD : nonsense mediated decay, NR : negative regulation of, PR : positive regulation of, TS: TargetScan.



Supplemental Figure 4. Protein disorder predicted by three complementary methods: PONDR VL-XT, DISOPRED and IUPRED. (A) Global proportion of disordered residues as a function of the threshold for each predictor. Cutoff values for each predictor were chosen to roughly equate proteome wide disorder : 0.3, 0.5 and 0.7 for DISOPRED, IUPRED and PONDR VL-XT, respectively. (B) Prediction of 5 metrics of disorder: the proportion of disordered residues, average residue disorder, the length of N- and C-terminus disorder and occurrence of internally disordered regions longer than 30 amino acids (1-3). Wilcoxon rank sum or Fisher tests were used for continuous and count data respectively; notched boxplots show the median values of each population extending to the interquartile range.



Supplemental Figure 5. An ordered logistic regression model to predict whether MAP output for a gene will be high, low or nonexistent. (A) Model performance measured by a ROC plot; the AUCs are 0.86 ± 0.02 for the high category, 0.64 ± 0.01 for the none category and 0.81 ± 0.02 for the low category. (B) The relative weight of the top 20 features contributing to prediction scores. Variables normalized by the length of the corresponding UTR, transcript or protein are denoted by *, GO terms denoted by •. EC : extracellular, Mem. : membrane, MFE : minimum free energy, MM : macromolecular, NMD : nonsense mediated decay, PR : positive regulation of. (C) Prediction scores for each category grouped by experimentally defined source classification.

Supplemental Methods. R source code to build the a two-class logistic regression model from data of 18 B-LCL donors (Supplemental Tables 2 & 6 required) and generate the panels of Figure 5.

```
options(stringsAsFactors = FALSE)
options(echo=TRUE)

# Prepare package library
libs = c('caret','gdata','pROC','plotROC','cowplot','gridExtra')
lapply(libs, require, character.only = TRUE)

library('colorout')

model_df = read.xls('Supplemental_Tables.xlsx', sheet =
'Supplemental Table 6', skip = 1)
maps = read.xls('Supplemental_Tables.xlsx', sheet = 'Supplemental
Table 2', skip = 1)

# Data preparation
# Log Transformations
logvar = c('Gene.expression','No..exons.','mRNA.length',
      'X5.UTR.length.','X5.UTR.base.pairing.','X3.UTR.length.',
      'X3.UTR.destabilizing.elements.',
      'X3.UTR.stabilizing.elements.',
      'X3.UTR.AU.rich.elements.','miR.elements..GSEA..',
      'miR.elements..TargetScan..','Protein.abundance',
      'Protein.length','N.terminus.disorder','C.terminus.disorder',
      'Lysine.Ub.density')

model_df[,logvar] = log10(model_df[,logvar] + 0.001)

# Remove near zero variance variables
model_df = model_df[, !nearZeroVar(model_df, saveMetrics = T)$nzv]

# Identify correlated predictors > 0.7
spear = cor(model_df[,sapply(model_df, is.numeric)], use =
'complete.obs', method = 'spearman')

cor_rem = c('mRNA.length','X5.UTR.minimum.free.energy',
      'X3.UTR.GC.content.','X3.UTR.AU.rich.elements.','Hydropathy.',
      'Helix.motif.','Disordered.residues.','GO.0003723',
      'GO.0010557','GO.0010628','GO.0031328','GO.0032991',
      'GO.2000113')

model_df = model_df[,!c(colnames(model_df) %in% cor_rem)]

findCorrelation(spear, cutoff = 0.7, names = T, exact = T)

# Check for linear combinations
findLinearCombos(model_df[,sapply(model_df, is.numeric)])

# Assign Source and Non-source genes
model_df$Source = factor(ifelse(model_df$Ensembl.gene.ID %in%
maps$Gene.ID, 'Source', 'NonSource'), levels =
c('Source','NonSource'))
```

```
# Number of replicates for model building
times = 1000

### Build Model

# Control function for logistic regression including variable
importance
ctrl = trainControl('cv', number = 10,
      summaryFunction = twoClassSummary,
      classProbs = TRUE)

# Control function for logistic regression with recursive feature
elimination
ctrlFS = trainControl('cv', number = 10,
      summaryFunction = twoClassSummary,
      classProbs= TRUE, allowParallel = T)

glmRFE = lrFuncs

glmRFE$fit = function (x, y, first, last, ...)
{
      tmp <- if (is.data.frame(x))
      x
      else as.data.frame(x)
      tmp$Class <- y
      glm(Class ~ ., data = tmp, family = binomial(link = 'logit'))
}

glmRFE$summary = twoClassSummary

ctrlRFE = rfeControl(functions = glmRFE,
      verbose = F, allowParallel = T)

cNames = c('Rep', 'Accuracy', 'Kappa', 'AccuracyLower',
'AccuracyUpper',
'AccuracyNull','AccuracyPValue','McnemarPValue','Sensitivity','Speci
ficity','Pos_Pred_Value','Neg_Pred_Value','Prevalence','Detection_Ra
te', 'Detection_Prevalence','Balanced_Accuracy','ROC_AUC')
metrics = data.frame(setNames(replicate(length(cNames),NA, simplify
= F), cNames))

gene_pred = model_df[,c('Ensembl.gene.ID','Source')]

pep_count =
unique(maps[,c('Peptide.Sequence','Gene.ID')])['Gene.ID']
gene_pred$num_maps = sapply(gene_pred$Ensembl.gene.ID, function(x)
sum(pep_count == x))

var_imp = data.frame(Variable = colnames(model_df)[sapply(model_df,
is.numeric)])

rfe_freq = data.frame(Variable = colnames(model_df)[sapply(model_df,
is.numeric)])

coef = list()

for (iter in 1:times){
```

```
      # Data splitting
      inTraining = createDataPartition(model_df$Source, p = 0.8,
list = FALSE)
      model_df_mod = model_df[,c(colnames(model_df)[sapply(model_df,
is.numeric)],'Source')]
      cTrain = model_df_mod[inTraining, ]
      cTest = model_df_mod[-inTraining, ]
      genes_cTest = model_df[-inTraining, 'Ensembl.gene.ID']

      # Build Model
      lrFitVI = train(Source ~ ., data = cTrain, method = 'glm',
family = binomial(link = 'logit'), trControl = ctrl, metric = 'ROC',
preProc = c('center','scale'))

      coef[[iter]] = summary(lrFitVI)$coefficients

      # Save results
      lrPred = data.frame(Ensembl.gene.ID = genes_cTest,
predict(lrFitVI, cTest, type = 'prob'), pred = predict(lrFitVI,
cTest) )

      gene_pred[,paste('Rep',iter,sep ='')] =
lrPred$Source[match(gene_pred$Ensembl.gene.ID,
lrPred$Ensembl.gene.ID)]

      metrics[iter, ] = c(iter,
            confusionMatrix(lrPred$pred, cTest$Source)$overall,
            confusionMatrix(lrPred$pred, cTest$Source)$byClass,
            roc(response = cTest$Source, predictor = lrPred$Source,
levels = c('NonSource', 'Source'))$auc[1] )

      var_imp[,paste('Rep',iter,sep ='')] =
varImp(lrFitVI)$importance

      # Build Recursive feature elimination model
      lrRFE = rfe(Source ~ ., data = cTrain, sizes = c(10),
            rfeControl = ctrlRFE, metric = 'ROC', trControl =
ctrlFS,
            maximize = F)

      rfe_freq[,paste('Rep',iter,sep ='')] = rfe_freq$Variable %in%
predictors(lrRFE)
}

gene_pred$Mean_pred = apply(gene_pred[,grep('Rep',
colnames(gene_pred))], 1, mean, na.rm = TRUE)

gene_pred$Source_pred = factor(ifelse(gene_pred$Mean_pred > 0.5,
'Source','Non-Source'))

rfe_freq$Freq = apply(rfe_freq[,grep('Rep', colnames(rfe_freq))], 1,
sum) / times

var_imp$Mean_imp = apply(var_imp[,grep('Rep', colnames(var_imp))],
1, mean)

# Calulate average model metrics
```

```
apply(apply(metrics, 2, as.numeric), 2, mean, na.rm = TRUE)

levels(gene_pred$Source) = levels(gene_pred$Source_pred)

## Figure 5
theme_uni = theme_classic() +
theme(text = element_text(size=8),
      axis.text = element_text(size=8))

# Separation of classes with model predictions
A = ggplot() +
      geom_density(aes(x = Mean_pred, fill = Source), data =
gene_pred, alpha = 0.5) +
      theme_uni + guides(fill = FALSE) +
      scale_fill_manual(values = c('#B4D0C0','#678EC1')) +
      scale_y_continuous(expand= c(0,0)) +
      ylab('% Genes') + xlab('Probability of MAP generation')

# Probability of MAP generation by number of MAPs generated
gene_pred = data.frame(gene_pred[order(gene_pred$Mean_pred,
decreasing = T),], Rank = c(1:nrow(gene_pred)))

B = ggplot(aes(x = Mean_pred, y = num_maps, ymax = num_maps, ymin =
num_maps), data = gene_pred) +
      theme_uni +
      scale_y_continuous(expand= c(0,0)) +
      geom_crossbar(size = 1.5) +
      ylab('MAPs generated') + xlab('Probability of MAP generation')

gene_pred$D = ifelse(gene_pred$num_maps == 0, 0, 1)

ci = paste('AUC =',round(mean(as.numeric(metrics$ROC_AUC)), digits =
2),'±',
      round(qnorm(0.975)*(mean(as.numeric(metrics$ROC_AUC)) /
sqrt(nrow(model_df)) ), digits = 2 ) )

C = ggplot(gene_pred, aes(m = Mean_pred, d = D)) +
      geom_text(aes(x = 0.5, y = 0.1, label = ci), stat =
"identity", size = 3) +
      geom_roc(labels = FALSE, size.point = 0) +
      theme_uni +
      xlab('Specificity') + ylab('Sensitivity')

## D Recursive feature elimination logistic regression limited
## to 10 variables, reported frequency of each variable in top 10.
## frequencies above 0.05 are shown

fv = rfe_freq[rfe_freq$Freq > 0.05, ]
fv$Full_Name = gsub('.',' ',gsub('GO.','GO:',
gsub('X','',fv$Variable)), fixed = TRUE)

fv_plot = ggplot(aes(x = reorder(Full_Name, Freq), y = Freq), data =
fv) + geom_bar(stat = 'identity') +
      theme_uni + ylab('Frequncy of selection\nin 10-variable
model') + xlab('') +
      scale_y_continuous(expand = c(0,0)) +
      theme(axis.text.x = element_text(hjust = 0.5)) +
```

```
      coord_flip()
D =  ggdraw(switch_axis_position(fv_plot, 'x') )

## E All variables
var_imp$Mean_imp_log10 = log10(var_imp$Mean_imp)
var_imp$Full_Name = gsub('.',' ',gsub('GO.','GO:'

vi_plot = ggplot() +
      geom_bar(data = var_imp ,
            aes(x = reorder(var_imp$Full_Name, Mean_imp_log10), y =
Mean_imp_log10),
            stat = 'identity') +
      theme_uni + ylab('Relative weight')  + xlab('') + theme_uni +
      scale_y_continuous(expand= c(0,0), limits = c(0,2),
            breaks =c(log10(1),log10(10),log10(100))) +
      theme(axis.text.x = element_text(hjust = 0.5),
            axis.text.y = element_text(size = 7)) +
      coord_flip()

E = ggdraw(switch_axis_position(vi_plot, 'x') )

# P values for each variable
apply(data.frame(t(do.call(cbind, lapply(coef, function(x)
x[,'Pr(>|z|)']))))),2,mean)

# Make Figure 5
plotz = list(A,B,C,D,E)
ggsave("Figure5.pdf", marrangeGrob(grobs = plotz, nrow=2, ncol=3,
top = NULL, layout_matrix = cbind(c(1,2),c(3,4),c(5,5))),
      height = 17.8 * .67, width = 18, units = 'cm')
```

Supplemental references

(1) Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. *Proteins* 2001;42:38-48.

(2) Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 2015;31:857-863.

(3) Dosztanyi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 2005;21:3433-3434.