Top 10

Caleb Easterly

April 26, 2018

The goal here is to estimate the top 10 'things' that are changing across WS and NS, by fold change. The fold change is calculated by the sum of the spectral counts attributed to each 'thing' - in eggNOG mapper and BLAST, the counts are aggregated by protein; in MEGAN, they are aggregated by eggNOG orthologous group; and in metaGOmics and Unipept, they are assigned to GO terms. In all cases, the Laplace correction is made, which is just adding 1 to all observed counts. This prevents division by zero, and allows for fold change estimation when one 'thing' was seen in one sample but not in another.

Reading in data

Necessary packages:

```
library(dplyr)
library(kableExtra)
```

Peptides

These are the counts used in eggNOG-mapper and BLAST results interpretation.

```
peptidesNS <- read.delim("peptide_shaker_outputs/737NS_Peptide_Shaker_Peptide_Report.tabular",
                         stringsAsFactors = FALSE) %>%
    select(peptide = Sequence, countNS = "X.Validated.PSMs")

peptidesWS <- read.delim("peptide_shaker_outputs/737WS_Peptide_Shaker_Peptide_Report.tabular",
                         stringsAsFactors = FALSE) %>%
    select(peptide = Sequence, countWS = "X.Validated.PSMs")

peptides_all <- full_join(peptidesNS, peptidesWS, by = "peptide")
peptides_all[is.na(peptides_all)] <- 0
```

eggNOG mapper results

Join peptides to counts, and calculate ratios.

```
eggnog <- read.delim("eggnogmap_results/diamond_annotations.tabular",
                     stringsAsFactors = FALSE,
                     header = FALSE) %>%
    select(peptide = V1, protein = V2, gene = V5, go = V6, ko = V7, desc = V13)

eggnog_w_counts <- left_join(eggnog,
                             peptides_all,
                             by = "peptide") %>%
    filter(!is.na(countWS) & !is.na(countNS)) %>%
    group_by(protein, desc, gene) %>%
    summarize(sumCountWS = sum(countWS) + 1, sumCountNS = sum(countNS) + 1) %>%
    mutate(log2ratio = log2(sumCountWS/sumCountNS)) %>%
    arrange(-log2ratio) %>%
    select(protein, gene, sumCountWS, sumCountNS, log2ratio, desc)
```

Top 10:

Table 1: eggNOG-

protein	gene	sumCountWS	sumCountNS	log2ratio	desc
866776.HMPREF9321_0304	OCAR_4246	92	1	6.523562	Peptidase propeptide and YPEB
866776.HMPREF9321_1872	MT2607	68	1	6.087463	Orn lys arg decarboxylase
655813.HMPREF8579_0938	IRAE	60	1	5.906891	Glycosyl hydrolase family 70
888048.HMPREF8577_1396	DPS	60	1	5.906891	DNA protection during starvatio
866776.HMPREF9321_0978	NIFJ	55	1	5.781360	Oxidoreductase required for the
384765.SIAM614_14165	ENO	52	1	5.700440	Catalyzes the reversible conversio
866776.HMPREF9321_1481	YGFH	44	1	5.459432	succinate CoA transferase
889201.HMPREF9422_0254	RPLL	43	1	5.426265	Seems to be the binding site for
866776.HMPREF9321_1437		41	1	5.357552	Pfam:YadA
655813.HMPREF8579_0100	CLPL	40	1	5.321928	ATP-dependent Clp protease AT

The Pfam:YadA refers to the YadA head domain in the trimeric autotransporter adhesin protein family.

Blast

```
blast <- read.delim('blast_results/blastp_vs_nr_current.tabular',
                    stringsAsFactors = FALSE,
                    header = FALSE) %>%
    select(peptide = V1, protein = V2, desc = V25)

blast_counts <- left_join(blast, peptides_all, by = "peptide") %>%
    group_by(protein, desc) %>%
    filter(!is.na(countWS) & !is.na(countNS)) %>%
    summarize(sumCountWS = sum(countWS) + 1, sumCountNS = sum(countNS) + 1) %>%
    mutate(log2ratio = log2(sumCountWS/sumCountNS)) %>%
    arrange(-log2ratio)
```

Print results:

```
## Warning in kable_styling(., latex_options = "scale_down"): Please specify
## format in kable. kableExtra can customize either HTML or LaTeX outputs. See
## https://haozhu233.github.io/kableExtra/ for details.
```

protein	desc
WP_027333991	phosphopyruvate hydratase [Mycoplasma elephantis]
WP_074657392	50S ribosomal protein L7/L12 [Streptococcus gallolyticus]<>LSU ribosomal protein L12P [Streptococcus
BAV80289	DNA protection during starvation protein [Streptococcus sp. NPS 308]
WP_080980855	DNA starvation/stationary phase protection protein [Streptococcus pseudopneumoniae]
WP_080977080	DNA starvation/stationary phase protection protein [Streptococcus pseudopneumoniae]
WP_057894291	elongation factor G [Lactobacillus brantae]<>elongation factor G [Lactobacillus brantae DSM 23927]
WP_080569238	DNA starvation/stationary phase protection protein [Streptococcus oralis]

protein	desc
KYF38032	Peptide deformylase [Streptococcus mitis]
CTN69507	L-lactate dehydrogenase [Streptococcus pneumoniae]<>L-lactate dehydrogenase [Streptococcus pneumon
WP_003009034	MULTISPECIES: DNA starvation/stationary phase protection protein [Streptococcus]<>ferritin-like pro

metaGOmics

We take the top 10 results that have a FDR-corrected q value less than 0.05.

```r
metagomics <- read.delim("metaGOmics_results/go_compare_149_150.txt",
                         comment.char = "#") %>%
    select(go = GO.acc,
           name = GO.name,
           log2ratio =  Laplace.corr..Log.2..fold.change,
           p = Laplace.corr..q.value)

metagomics_filt <- metagomics %>% filter(p < 0.05) %>%
    arrange(-log2ratio)
```

Top 10 results:

Table 3: MetaGOmics: Top 10 fold changes

go	name	log2ratio	p
GO:0047112	pyruvate oxidase activity	7.231802	0
GO:0016623	oxidoreductase activity, acting on the aldehyde or oxo group of donors, oxygen as acceptor	7.231802	0
GO:0004867	serine-type endopeptidase inhibitor activity	6.756717	0
GO:0009611	response to wounding	6.756342	0
GO:1902011	poly(ribitol phosphate) teichoic acid metabolic process	6.538750	0
GO:1902012	poly(ribitol phosphate) teichoic acid biosynthetic process	6.538750	0
GO:0008730	L(+)-tartrate dehydratase activity	6.416641	0
GO:0050256	ribitol-5-phosphate 2-dehydrogenase activity	6.360788	0
GO:0008886	glyceraldehyde-3-phosphate dehydrogenase (NADP+) (non-phosphorylating) activity	6.178950	0
GO:0005518	collagen binding	6.019751	0

MEGAN

```r
megan <- read.delim("MEGAN_outputs/737NSvsWS_EGGNOGcount.csv") %>%
    select(og = X.Datasets, countNS = X737_NS_BLASTOutput_2StepCombined,
           countWS = X737_WS_BLASTOutput_2StepCombined) %>%
    mutate(corrWS = countWS + 1,
           corrNS = countNS + 1,
           log2ratio = log2(corrWS/corrNS)) %>%
    arrange(-log2ratio)
```

Top 10 results:

og
ENOG410Z98C Streptococcal surface antigen repeat

og
COG0028 acetolactate synthase
COG0119 Catalyzes the condensation of the acetyl group of acetyl-CoA with 3-methyl-2-oxobutanoate (2-oxoisovalerate) to
COG3525 ec 3.2.1.52
ENOG411248X Cell Wall
COG3579 aminopeptidase c
COG0242 Removes the formyl group from the N-terminal Met of newly synthesized proteins. Requires at least a dipeptide
COG1621 Hydrolase
ENOG410YESU
COG1026 peptidase

Note that EC 3.2.1.52 is a beta-hexosaminidase and ENOG410YESU is involved in cell wall/membrane/envelope biogenesis.

Unipept

```
unipept_results_NS <- paste('unipept_results/',
                               list.files("unipept_results/", pattern = "^737NS.*\\.csv"),
                               sep = "")
unipept_results_WS <- paste('unipept_results/',
                               list.files("unipept_results/", pattern = "^737WS.*\\.csv"),
                               sep = "")
unipeptNS <- lapply(unipept_results_NS, function(i) {
        read.delim(i, sep = ',', as.is = TRUE)}) %>%
    bind_rows() %>%
    select(-X) %>%
    rename(peptides = X.peptides)
unipeptWS <- lapply(unipept_results_WS, function(i) {
    read.delim(i, sep = ',', as.is = TRUE)}) %>%
    bind_rows() %>%
    select(-X) %>%
    rename(peptides = X.peptides)

unipept_all <- inner_join(unipeptNS, unipeptWS, by = c("GO.term", "Name")) %>%
    mutate(lapCountNS = peptides.x + 1, lapCountWS = peptides.y + 1,
           log2ratio = log(lapCountWS/lapCountNS)) %>%
    select(GO.term, Name, lapCountWS, lapCountNS, log2ratio) %>%
    arrange(-log2ratio)
```

Top 10:

Table 5: Unipept: top 10 Fold Changes

GO.term	Name	lapCountWS	lapCountNS	log2ratio
GO:0042586	peptide deformylase activity	59	2	3.384390
GO:0008662	1-phosphofructokinase activity	42	2	3.044522
GO:2001059	D-tagatose 6-phosphate catabolic process	88	5	2.867899
GO:0009024	tagatose-6-phosphate kinase activity	88	5	2.867899
GO:0009611	response to wounding	35	2	2.862201
GO:0004084	branched-chain-amino-acid transaminase activity	50	3	2.813411
GO:0006091	generation of precursor metabolites and energy	33	2	2.803360

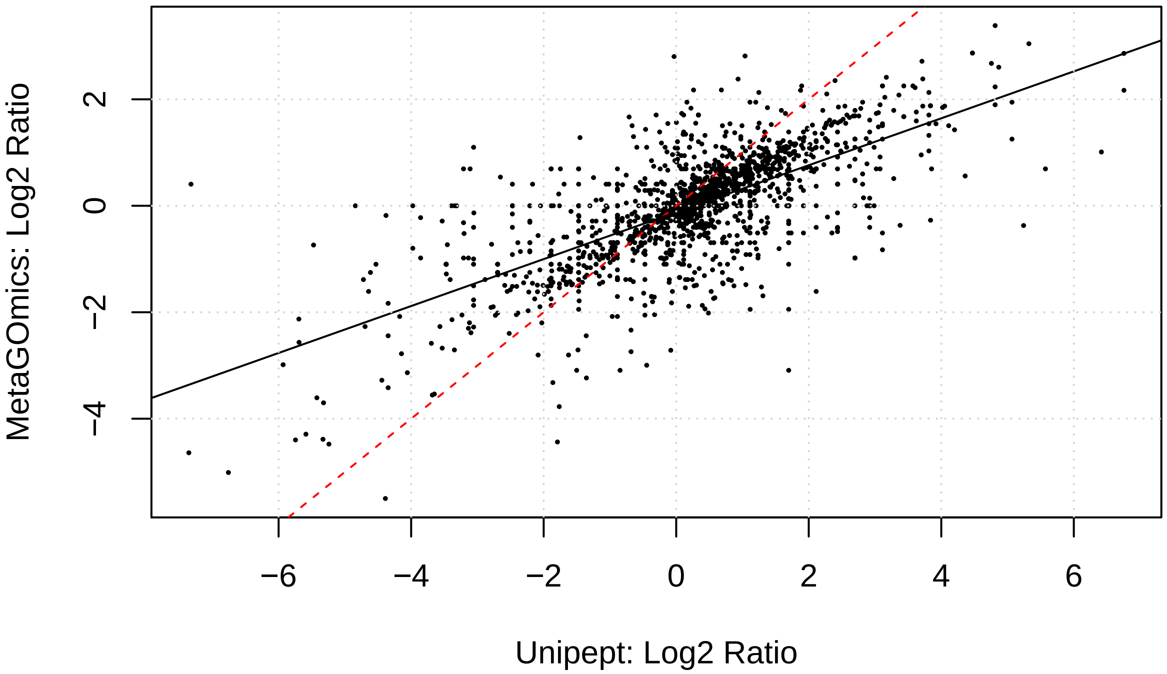
GO.term	Name	lapCountWS	lapCountNS	log2ratio
GO:0009374	biotin binding	166	11	2.714093
GO:0008888	glycerol dehydrogenase [NAD+] activity	29	2	2.674149
GO:0004564	beta-fructofuranosidase activity	27	2	2.602690

Unipept and MetaGOmics

```
um <- inner_join(unipept_all, metagomics, by = c("GO.term" = "go"))
```

```
## Warning: Column `GO.term`/`go` joining character vector and factor,
## coercing into character vector
```

```
plot(log2ratio.x ~ log2ratio.y, data = um, pch = 20, cex = 0.3,
     xlab = "Unipept: Log2 Ratio",
     ylab = "MetaGOmics: Log2 Ratio")
mod <- lm(log2ratio.x ~ log2ratio.y, data = um)
abline(0, 1, col = "red", lty = 2)
abline(coef(mod))
grid()
```



```
cor(um$log2ratio.x, um$log2ratio.y)
```

```
## [1] 0.6930295
```