Expected and Observed Count Values - Biosample, Pipeline

Nate Olson

2017-04-06

Objective

Evaluate the overall fit for biological replicates and pipelines. Want to see which has greater variability. Is there a pipeline or biological sample effect. The pipeline effect would be due to differences in feature definition. The biological sample effect would indicate a potential issue titration generation.

Loading Data

Feature level count data

```
## Subsetting features to only include pre, post, and full features
count_exp_df <- readRDS("../data/expected_count_values_feature_df.rds")
feature_cat <- readRDS("../data/feature_categories_df.rds")
count_exp_df <- left_join(count_exp_df, feature_cat) %>%
      filter(cat %in% c("cat_full", "cat_pre", "cat_post"))
```

Genus level count data

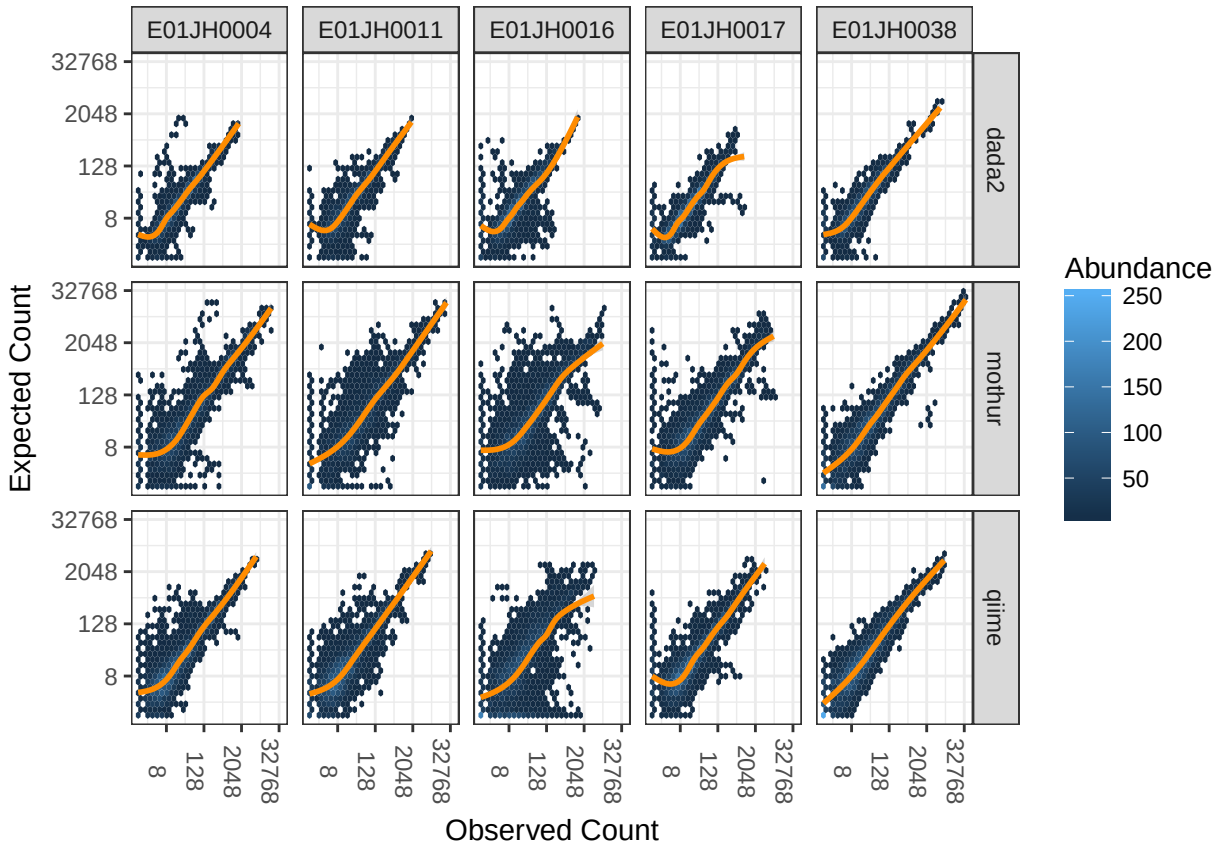
```
## Subsetting features to only include pre, post, and full features
genus_exp_df <- readRDS("../data/expected_count_values_genus_df.rds")
genus_cat <- readRDS("../data/genus_categories_df.rds")
genus_exp_df <- left_join(genus_exp_df, genus_cat) %>%
      filter(cat %in% c("cat_full", "cat_pre", "cat_post"))
```

Observed - Expected By Pipeline and Biological Replicate

Feature level analysis

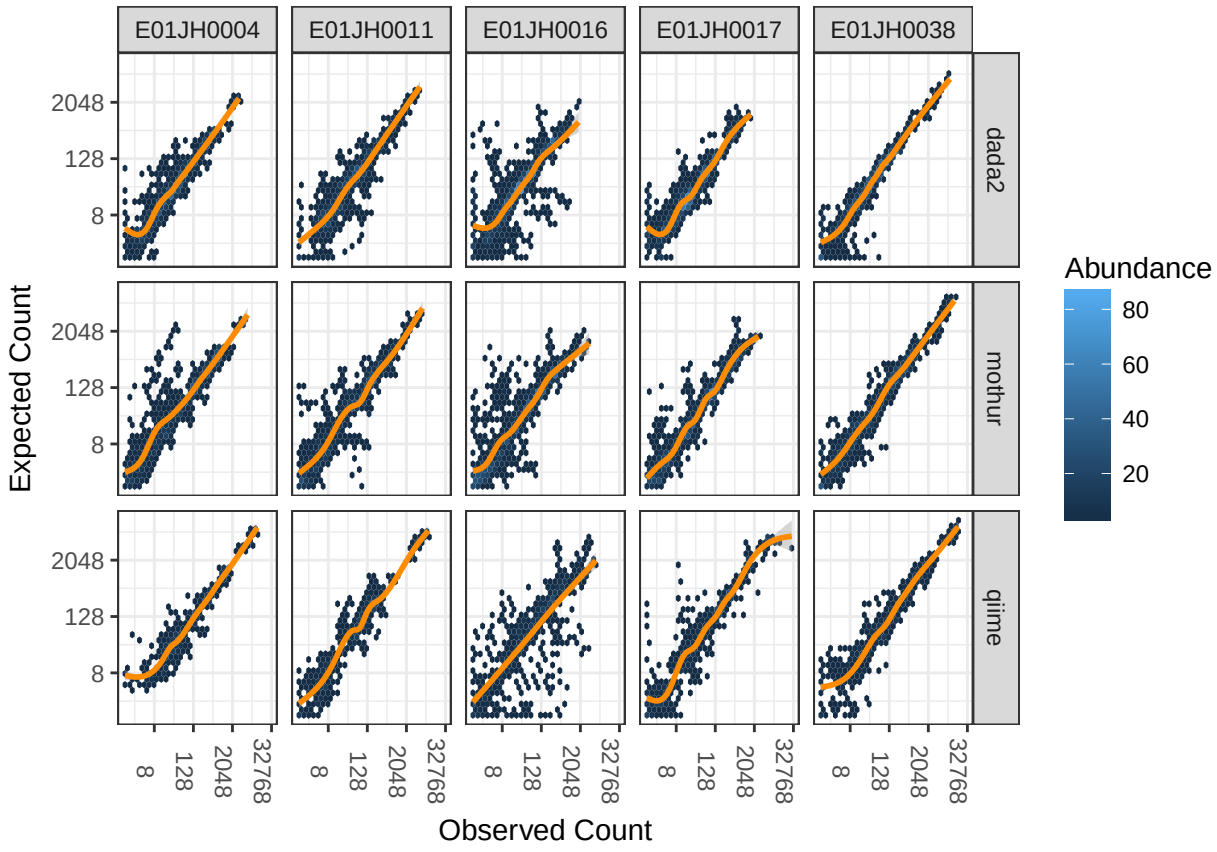
Overall relationship between the observed and expected values by pipeline and biological replicate. Orange line is a fitted smoothing function (loess, local polynomial regression) to highlight the relationship between the observed and expected counts.

```
count_exp_df %>% ggplot() +
      #geom_point(aes(x = obs_count + 1, y = exp_count + 1),
      #           fill = "darkblue", color = "white", alpha = 0.5, shape = 21) +
      geom_hex(aes(x = obs_count + 1, y = exp_count + 1)) +
      geom_smooth(aes(x = obs_count + 1, y = exp_count + 1), color = "darkorange") +
      #geom_abline(aes(intercept = 0, slope =1), color = "darkorange") +
     facet_grid(pipe~biosample_id)+ theme_bw() +
      labs(y = "Expected Count", x = "Observed Count", fill = "Abundance") +
      scale_y_continuous(trans = "log2") +
      scale_x_continuous(trans = "log2") +
      theme(axis.text.x = element_text(angle = 270))
```



Genus level analysis

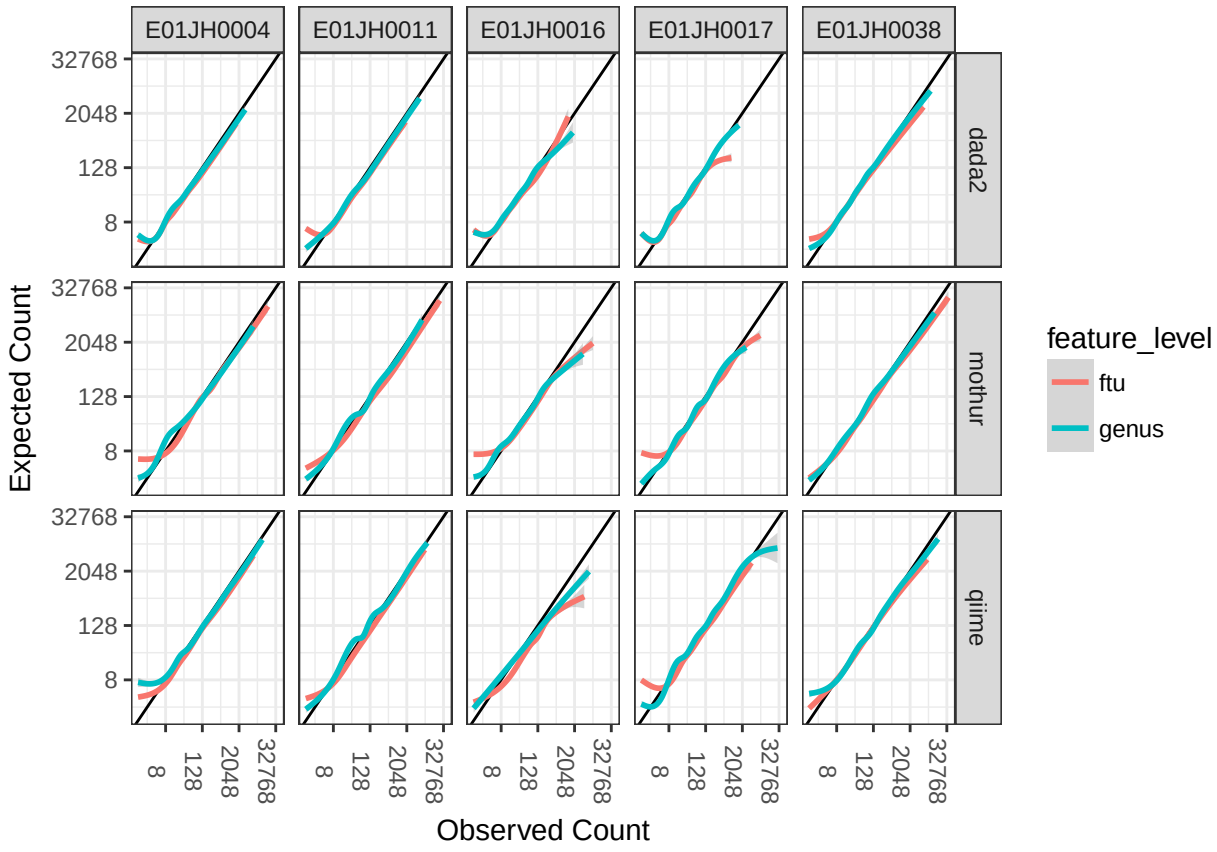
```
genus_exp_df %>% ggplot() +
      #geom_point(aes(x = obs_count + 1, y = exp_count + 1),
      #           fill = "darkblue", color = "white", alpha = 0.5, shape = 21) +
      geom_hex(aes(x = obs_count + 1, y = exp_count + 1)) +
      geom_smooth(aes(x = obs_count + 1, y = exp_count + 1), color = "darkorange") +
      #geom_abline(aes(intercept = 0, slope =1), color = "darkorange") +
     facet_grid(pipe~biosample_id)+ theme_bw() +
      labs(y = "Expected Count", x = "Observed Count", fill = "Abundance") +
      scale_y_continuous(trans = "log2") +
      scale_x_continuous(trans = "log2") +
      theme(axis.text.x = element_text(angle = 270))
```



Comparison of Raw and Genus Feature Levels

Comparison of raw feature level and features aggregated to the genus level. Black line indicates the expected 1-to-1 relationship between the expected and observed counts.

```
exp_df <-bind_rows(ftu = count_exp_df, genus = genus_exp_df,.id = "feature_level")
exp_df %>% ggplot() +
      geom_abline(aes(intercept = 0, slope = 1)) +
      geom_smooth(aes(x = obs_count + 1, y = exp_count + 1, color = feature_level)) +
     facet_grid(pipe~biosample_id)+ theme_bw() +
      labs(y = "Expected Count", x = "Observed Count", fill = "Abundance") +
      scale_y_continuous(trans = "log2") +
      scale_x_continuous(trans = "log2") +
      theme(axis.text.x = element_text(angle = 270))
```



Attributing Variability

Use a mixed effects model to characterize the impact of pipeline and biological replicate.

```
count_fit <- lme(residual ~ exp_count, random =  ~ biosample_id | pipe, data = count_exp_df)
summary(count_fit)
```

```
## Linear mixed-effects model fit by REML
##  Data: count_exp_df
##        AIC     BIC    logLik
##   1157525 1157692 -578744.5
##
## Random effects:
##  Formula: ~biosample_id | pipe
##  Structure: General positive-definite, Log-Cholesky parametrization
##                         StdDev     Corr
## (Intercept)              6.026856 (Intr) b_E01JH0011 b_E01JH0016
## biosample_idE01JH0011   19.689711  0.866
## biosample_idE01JH0016   16.481109 -0.984 -0.925
## biosample_idE01JH0017   10.169504 -0.986 -0.918       0.994
## biosample_idE01JH0038   14.215994  0.874  0.996      -0.932
## Residual               332.118363
##
## (Intercept)            b_E01JH0017
## biosample_idE01JH0011
## biosample_idE01JH0016
```

```
## biosample_idE01JH0017
## biosample_idE01JH0038 -0.924
## Residual
##
## Fixed effects: residual ~ exp_count
##                  Value Std.Error    DF   t-value p-value
## (Intercept)  2.2135314 1.4044643 80104   1.57607   0.115
## exp_count   -0.1611089 0.0019481 80104 -82.69975   0.000
##  Correlation:
##           (Intr)
## exp_count -0.187
##
## Standardized Within-Group Residuals:
##           Min            Q1           Med            Q3           Max
## -42.741038745  -0.024047778   0.003725774   0.042497452  58.101900470
##
## Number of Observations: 80108
## Number of Groups: 3
```

```
genus_fit <- lme(residual ~ exp_count, random =  ~ biosample_id | pipe, data = genus_exp_df)
summary(genus_fit)
```

```
## Linear mixed-effects model fit by REML
##  Data: genus_exp_df
##      AIC      BIC  logLik
##   291794 291936.2 -145879
##
## Random effects:
##  Formula: ~biosample_id | pipe
##  Structure: General positive-definite, Log-Cholesky parametrization
##                       StdDev    Corr
## (Intercept)            15.18247 (Intr) b_E01JH0011 b_E01JH0016 b_E01JH0017
## biosample_idE01JH0011  62.63127 -0.972
## biosample_idE01JH0016  17.22937 -0.922  0.877
## biosample_idE01JH0017  56.35939 -0.979  0.982       0.850
## biosample_idE01JH0038  74.97435  0.915 -0.967      -0.893      -0.909
## Residual              360.53285
##
## Fixed effects: residual ~ exp_count
##                 Value Std.Error    DF   t-value p-value
## (Intercept)  6.146195 2.9849425 19960   2.05907  0.0395
## exp_count   -0.121639 0.0031498 19960 -38.61805  0.0000
##  Correlation:
##           (Intr)
## exp_count -0.26
##
## Standardized Within-Group Residuals:
##           Min            Q1           Med            Q3           Max
## -71.355314383  -0.060620251  -0.009148277   0.037792641  12.831559609
##
## Number of Observations: 19964
## Number of Groups: 3
```

Comparison of variance components

```
var_count <- varcomp(count_fit, scale = TRUE)
var_genus <- varcomp(genus_fit, scale = TRUE)
data_frame(comp = names(var_count),
           raw_feature = round(var_count * 100,2),
           genus = round(var_genus * 100,2))
```

```
## # A tibble: 2 × 3
##     comp raw_feature genus
##    <chr>       <dbl> <dbl>
## 1   pipe        0.03  0.18
## 2 Within       99.97 99.82
```

Session information

```
s_info <- devtools::session_info()
print(s_info$platform)
```

```
##  setting  value
##  version  R version 3.3.3 (2017-03-06)
##  system   x86_64, darwin15.6.0
##  ui       unknown
##  language (EN)
##  collate  en_US.UTF-8
##  tz       America/New_York
##  date     2017-04-06
```

```
s_info$packages %>% filter(`*` == "*") %>% select(-`*`) %>%
      knitr::kable()
```

package	version	date	source
ape	4.1	2017-02-14	CRAN (R 3.3.2)
bbmle	1.0.18	2016-02-11	CRAN (R 3.3.2)
Biobase	2.34.0	2016-11-07	Bioconductor
BiocGenerics	0.20.0	2016-11-07	Bioconductor
BiocParallel	1.8.1	2016-11-07	Bioconductor
Biostrings	2.42.1	2016-12-19	Bioconductor
DESeq	1.26.0	2016-11-28	Bioconductor
DESeq2	1.15.28	2017-02-02	bioc (readonly/DESeq2@125913)
dplyr	0.5.0	2016-06-24	CRAN (R 3.3.2)
edgeR	3.16.5	2017-02-02	Bioconductor
forcats	0.2.0	2017-01-23	CRAN (R 3.3.2)
foreach	1.4.3	2015-10-13	CRAN (R 3.3.1)
GenomeInfoDb	1.10.3	2017-03-28	Bioconductor
GenomicAlignments	1.10.1	2017-03-28	Bioconductor
GenomicRanges	1.26.4	2017-03-28	Bioconductor
ggplot2	2.2.1	2016-12-30	CRAN (R 3.3.2)
glmnet	2.0-5	2016-03-17	CRAN (R 3.3.1)
hexbin	1.27.1	2015-08-19	CRAN (R 3.3.1)
IRanges	2.8.2	2017-03-28	Bioconductor
knitr	1.15.1	2016-11-22	CRAN (R 3.3.2)
lattice	0.20-34	2016-09-06	CRAN (R 3.3.3)
limma	3.30.13	2017-03-28	Bioconductor
locfit	1.5-9.1	2013-04-20	CRAN (R 3.3.1)

package	version	date	source
Matrix	1.2-8	2017-01-20	CRAN (R 3.3.3)
metagenomeSeq	1.16.0	2016-11-07	Bioconductor
modelr	0.1.0	2016-08-31	cran (@0.1.0)
nlme	3.1-131	2017-02-06	CRAN (R 3.3.3)
permute	0.9-4	2016-09-09	CRAN (R 3.3.1)
phyloseq	1.19.1	2017-01-04	Bioconductor
ProjectTemplate	0.7	2016-08-11	CRAN (R 3.3.1)
purrr	0.2.2	2016-06-18	CRAN (R 3.3.1)
RColorBrewer	1.1-2	2014-12-07	CRAN (R 3.3.1)
readr	1.1.0	2017-03-22	CRAN (R 3.3.2)
readxl	0.1.1	2016-03-28	cran (@0.1.1)
Rqc	1.8.0	2016-11-07	Bioconductor
Rsamtools	1.26.1	2016-11-07	Bioconductor
S4Vectors	0.12.2	2017-03-28	Bioconductor
sads	0.3.1	2016-05-13	CRAN (R 3.3.2)
savR	1.12.0	2016-11-07	Bioconductor
ShortRead	1.32.1	2017-03-28	Bioconductor
stringr	1.2.0	2017-02-18	CRAN (R 3.3.2)
SummarizedExperiment	1.4.0	2016-11-07	Bioconductor
tibble	1.2	2016-08-26	CRAN (R 3.3.1)
tidyr	0.6.1	2017-01-10	CRAN (R 3.3.2)
tidyverse	1.1.1	2017-01-27	CRAN (R 3.3.2)
vegan	2.4-2	2017-01-17	CRAN (R 3.3.2)
XVector	0.14.1	2017-03-28	Bioconductor