Variance All the Way Down: Exploring the Impact of RNA-Seq Pipeline Choices on Differential Expression Variance

Hunter Schuler and Art Tay

Analysis

Assume there are n samples of G gene counts. Let B_{gi} denote the count for gene g in sample i reported to the NIH database, and let C_{giX} denote the count obtained from pipeline with choices X. Similar let D_g and E_{gX} denote the p-values obtained from `edgeR`. Now,

$$Y_{1X}^2 = \frac{1}{nG}\sum_{i=1}^{n}\sum_{g=1}^{G}(C_{giX} - B_{gi})^2 \tag{1}$$

and

$$Y_{2X}^2 = \frac{1}{G}\sum_{g=1}^{G}(E_{gX} - D_g)^2 \tag{2}$$

Our primary analysis will focus on the two following regression models:

$$Y_{1X} = \beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{1\leq i<j\leq p}\beta_{ij}(X_i \times X_j) + \epsilon \tag{3}$$

and

$$Y_{2X} = \beta_0 + \sum_{i=1}^{p}\beta_i X_i + \sum_{1\leq i<j\leq p}\beta_{ij}(X_i \times X_j) + \epsilon \tag{4}$$

where p is the number of pipeline choices from tbl-1. The first model studies the effect of each pipeline choice, include all pairwise interactions, on the average square deviation from the official NIH count matrix. The second model does the same, but for the p-values from a differential expression analysis.

Frequentist

```
# Meta Data
sample_names <- c(
    "gene",
    "SRR31476642",
    "SRR31476643",
    "SRR31476644",
    "SRR31476645",
    "SRR31476646",
    "SRR31476647",
    "SRR31476648",
    "SRR31476649",
    "SRR31476650"
)
```

```r
treatments <- c(
    "DMSO",
    "DMSO",
    "DMSO",
    "DMSO",
    "EPZ015666",
    "EPZ015666",
    "EPZ015666",
    "DMSO",
    "DMSO"
)

factors <- c("aligner", "trim_poly_g", "trim_poly_x", "norm_method")
```

Counts

```r
# Read in data.
count_sd_df <- read.csv("./data/gen_samples/count_sd_df.csv")
count_sd_df <- count_sd_df |>
    mutate(across(any_of(factors), ~ as.factor(.)))

# Classic LM
lm_fit <- count_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent)) |>
    (\(x) glm(count_sd ~ (.)^2, family = gaussian(), data = x))()
summary(lm_fit)
```

```
Call:
glm(formula = count_sd ~ (.)^2, family = gaussian(), data = x)

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2284.93325   23.02627  99.232   <2e-16 ***
alignersalmon               128.83274   78.18917   1.648   0.1081
min_phred                    -1.86181    0.85911  -2.167   0.0369 *
min_length                   -0.59808    0.50857  -1.176   0.2473
trim_poly_g1                 -1.72228    8.16305  -0.211   0.8341
trim_poly_x1                  1.69684    8.24527   0.206   0.8381
alignersalmon:min_phred      -1.43600    3.52913  -0.407   0.6865
alignersalmon:min_length     -0.10910    0.59610  -0.183   0.8558
alignersalmon:trim_poly_g1  -12.20600   24.28720  -0.503   0.6183
alignersalmon:trim_poly_x1         NA         NA      NA       NA
min_phred:min_length          0.01663    0.01904   0.873   0.3884
min_phred:trim_poly_g1        0.09148    0.24875   0.368   0.7152
min_phred:trim_poly_x1        0.17504    0.25240   0.694   0.4924
min_length:trim_poly_g1      -0.02757    0.11239  -0.245   0.8076
min_length:trim_poly_x1      -0.19574    0.10719  -1.826   0.0761 .
trim_poly_g1:trim_poly_x1    -0.96535    1.49856  -0.644   0.5235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.34726)
```

```
    Null deviance: 26856.5  on 50  degrees of freedom
Residual deviance:   192.5  on 36  degrees of freedom
AIC: 244.47

Number of Fisher Scoring iterations: 2
```

```
# Log-normal GLM
glm_log_fit <- count_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent)) |>
    (\(x) glm(count_sd ~ (.)^2, family = gaussian(link = "log"), data = x))()
summary(glm_log_fit)
```

```
Call:
glm(formula = count_sd ~ (.)^2, family = gaussian(link = "log"),
    data = x)

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 7.734e+00  1.033e-02 748.646   <2e-16 ***
alignersalmon               5.550e-02  3.411e-02   1.627   0.1124
min_phred                  -8.322e-04  3.855e-04  -2.159   0.0376 *
min_length                 -2.669e-04  2.282e-04  -1.169   0.2500
trim_poly_g1               -7.718e-04  3.663e-03  -0.211   0.8343
trim_poly_x1                7.862e-04  3.701e-03   0.212   0.8330
alignersalmon:min_phred    -5.872e-04  1.538e-03  -0.382   0.7049
alignersalmon:min_length   -4.838e-05  2.591e-04  -0.187   0.8529
alignersalmon:trim_poly_g1 -5.149e-03  1.058e-02  -0.486   0.6296
alignersalmon:trim_poly_x1         NA         NA      NA       NA
min_phred:min_length        7.405e-06  8.547e-06   0.866   0.3920
min_phred:trim_poly_g1      4.110e-05  1.117e-04   0.368   0.7150
min_phred:trim_poly_x1      7.768e-05  1.133e-04   0.686   0.4974
min_length:trim_poly_g1    -1.241e-05  5.043e-05  -0.246   0.8070
min_length:trim_poly_x1    -8.790e-05  4.810e-05  -1.828   0.0759 .
trim_poly_g1:trim_poly_x1  -4.331e-04  6.724e-04  -0.644   0.5235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.355166)

    Null deviance: 26856.46  on 50  degrees of freedom
Residual deviance:   192.79  on 36  degrees of freedom
AIC: 244.55

Number of Fisher Scoring iterations: 3
```

```
# Quasi GLM
quasi_fit <- count_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent)) |>
    (\(x) glm(count_sd ~ (.)^2, family = quasi(), data = x))()
summary(quasi_fit)
```

```
Call:
```

```
glm(formula = count_sd ~ (.)^2, family = quasi(), data = x)

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2284.93325   23.02627  99.232   <2e-16 ***
alignersalmon               128.83274   78.18917   1.648   0.1081
min_phred                    -1.86181    0.85911  -2.167   0.0369 *
min_length                   -0.59808    0.50857  -1.176   0.2473
trim_poly_g1                 -1.72228    8.16305  -0.211   0.8341
trim_poly_x1                  1.69684    8.24527   0.206   0.8381
alignersalmon:min_phred      -1.43600    3.52913  -0.407   0.6865
alignersalmon:min_length     -0.10910    0.59610  -0.183   0.8558
alignersalmon:trim_poly_g1  -12.20600   24.28720  -0.503   0.6183
alignersalmon:trim_poly_x1         NA         NA      NA       NA
min_phred:min_length          0.01663    0.01904   0.873   0.3884
min_phred:trim_poly_g1        0.09148    0.24875   0.368   0.7152
min_phred:trim_poly_x1        0.17504    0.25240   0.694   0.4924
min_length:trim_poly_g1      -0.02757    0.11239  -0.245   0.8076
min_length:trim_poly_x1      -0.19574    0.10719  -1.826   0.0761 .
trim_poly_g1:trim_poly_x1    -0.96535    1.49856  -0.644   0.5235
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 5.34726)

    Null deviance: 26856.5  on 50  degrees of freedom
Residual deviance:   192.5  on 36  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 2
```

P-Values

```r
# Read in data.
DE_sd_df <- read.csv("./data/gen_samples/DE_sd_df.csv")
DE_sd_df <- DE_sd_df |>
    mutate(across(any_of(factors), ~ as.factor(.)))

# Classic LM
lm_fit <- DE_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent, effect_size_sd)) |>
    (\(x) glm(p_value_sd ~ (.)^2, family = gaussian(), data = x))()
summary(lm_fit)
```

```
Call:
glm(formula = p_value_sd ~ (.)^2, family = gaussian(), data = x)

Coefficients: (5 not defined because of singularities)
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        3.725e-01  1.499e-03 248.488   <2e-16
alignersalmon                      5.976e-03  5.215e-03   1.146   0.2686
min_phred                          6.631e-05  6.504e-05   1.019   0.3231
min_length                        -9.363e-06  2.585e-05  -0.362   0.7219
```

```
trim_poly_g1                            -3.163e-04  3.668e-04  -0.862   0.4013
trim_poly_x1                            -5.951e-04  4.046e-04  -1.471   0.1607
norm_methoddefault                       1.805e-01  6.182e-04 292.046   <2e-16
norm_methodRLE                           1.803e-01  7.686e-04 234.522   <2e-16
norm_methodTMM                           1.806e-01  1.337e-03 135.071   <2e-16
norm_methodupperquartile                 1.806e-01  7.139e-04 252.963   <2e-16
alignersalmon:min_phred                 -3.086e-04  2.108e-04  -1.464   0.1625
alignersalmon:min_length                 9.020e-05  2.924e-05   3.085   0.0071
alignersalmon:trim_poly_g1              -1.848e-03  1.567e-03  -1.180   0.2553
alignersalmon:trim_poly_x1                      NA         NA      NA       NA
alignersalmon:norm_methoddefault                NA         NA      NA       NA
alignersalmon:norm_methodRLE                    NA         NA      NA       NA
alignersalmon:norm_methodTMM                    NA         NA      NA       NA
alignersalmon:norm_methodupperquartile          NA         NA      NA       NA
min_phred:min_length                    -1.541e-07  1.087e-06  -0.142   0.8891
min_phred:trim_poly_g1                   1.935e-06  1.365e-05   0.142   0.8891
min_phred:trim_poly_x1                   8.002e-06  1.547e-05   0.517   0.6121
min_phred:norm_methoddefault            -7.192e-05  2.742e-05  -2.623   0.0185
min_phred:norm_methodRLE                -6.687e-05  3.345e-05  -1.999   0.0629
min_phred:norm_methodTMM                -7.941e-05  5.367e-05  -1.480   0.1584
min_phred:norm_methodupperquartile      -8.571e-05  3.476e-05  -2.466   0.0253
min_length:trim_poly_g1                  5.124e-06  5.202e-06   0.985   0.3393
min_length:trim_poly_x1                  1.372e-05  5.124e-06   2.677   0.0165
min_length:norm_methoddefault            1.189e-05  8.708e-06   1.365   0.1911
min_length:norm_methodRLE                1.357e-05  7.284e-06   1.863   0.0810
min_length:norm_methodTMM                1.517e-05  1.114e-05   1.362   0.1920
min_length:norm_methodupperquartile      7.880e-06  8.766e-06   0.899   0.3820
trim_poly_g1:trim_poly_x1                3.023e-05  6.288e-05   0.481   0.6372
trim_poly_g1:norm_methoddefault         -2.603e-05  1.023e-04  -0.254   0.8024
trim_poly_g1:norm_methodRLE              5.479e-05  8.612e-05   0.636   0.5337
trim_poly_g1:norm_methodTMM             -9.192e-05  3.695e-04  -0.249   0.8067
trim_poly_g1:norm_methodupperquartile    6.595e-05  8.865e-05   0.744   0.4677
trim_poly_x1:norm_methoddefault          2.769e-04  1.433e-04   1.933   0.0712
trim_poly_x1:norm_methodRLE              2.484e-04  1.370e-04   1.813   0.0886
trim_poly_x1:norm_methodTMM              4.243e-04  3.597e-04   1.179   0.2554
trim_poly_x1:norm_methodupperquartile    3.021e-04  1.279e-04   2.361   0.0312

(Intercept)                             ***
alignersalmon
min_phred
min_length
trim_poly_g1
trim_poly_x1
norm_methoddefault                      ***
norm_methodRLE                          ***
norm_methodTMM                          ***
norm_methodupperquartile                ***
alignersalmon:min_phred
alignersalmon:min_length                **
alignersalmon:trim_poly_g1
alignersalmon:trim_poly_x1
alignersalmon:norm_methoddefault
alignersalmon:norm_methodRLE
alignersalmon:norm_methodTMM
```

```
alignersalmon:norm_methodupperquartile
min_phred:min_length
min_phred:trim_poly_g1
min_phred:trim_poly_x1
min_phred:norm_methoddefault           *
min_phred:norm_methodRLE               .
min_phred:norm_methodTMM
min_phred:norm_methodupperquartile     *
min_length:trim_poly_g1
min_length:trim_poly_x1                *
min_length:norm_methoddefault
min_length:norm_methodRLE              .
min_length:norm_methodTMM
min_length:norm_methodupperquartile
trim_poly_g1:trim_poly_x1
trim_poly_g1:norm_methoddefault
trim_poly_g1:norm_methodRLE
trim_poly_g1:norm_methodTMM
trim_poly_g1:norm_methodupperquartile
trim_poly_x1:norm_methoddefault        .
trim_poly_x1:norm_methodRLE            .
trim_poly_x1:norm_methodTMM
trim_poly_x1:norm_methodupperquartile  *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 5.911428e-09)

    Null deviance: 2.1702e-01  on 50  degrees of freedom
Residual deviance: 9.4583e-08  on 16  degrees of freedom
AIC: -808.65

Number of Fisher Scoring iterations: 2
```

```r
# Log-normal GLM
glm_log_fit <- DE_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent, effect_size_sd)) |>
    (\(x) glm(p_value_sd ~ (.)^2, family = gaussian(link = "log"), data = x))()
summary(glm_log_fit)
```

```
Call:
glm(formula = p_value_sd ~ (.)^2, family = gaussian(link = "log"),
    data = x)

Coefficients: (5 not defined because of singularities)
                                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)                        -9.870e-01  3.049e-03 -323.752   <2e-16
alignersalmon                       1.032e-02  1.001e-02    1.031   0.3177
min_phred                           1.437e-04  1.331e-04    1.080   0.2963
min_length                         -1.338e-05  5.085e-05   -0.263   0.7958
trim_poly_g1                       -6.221e-04  7.339e-04   -0.848   0.4091
trim_poly_x1                       -9.970e-04  8.029e-04   -1.242   0.2323
norm_methoddefault                  3.947e-01  1.547e-03  255.158   <2e-16
norm_methodRLE                      3.942e-01  1.776e-03  222.019   <2e-16
```

```
norm_methodTMM                            3.948e-01  2.745e-03  143.827   <2e-16
norm_methodupperquartile                  3.947e-01  1.693e-03  233.193   <2e-16
aligner:min_phred                        -5.372e-04  4.047e-04   -1.327   0.2030
aligner:min_length                        1.597e-04  5.609e-05    2.847   0.0117
aligner:trim_poly_g1                     -3.186e-03  3.007e-03   -1.060   0.3050
aligner:trim_poly_x1                             NA         NA       NA       NA
aligner:norm_methoddefault                       NA         NA       NA       NA
aligner:norm_methodRLE                           NA         NA       NA       NA
aligner:norm_methodTMM                           NA         NA       NA       NA
aligner:norm_methodupperquartile                 NA         NA       NA       NA
min_phred:min_length                     -3.110e-07  2.091e-06   -0.149   0.8836
min_phred:trim_poly_g1                    6.166e-07  2.624e-05    0.023   0.9815
min_phred:trim_poly_x1                    1.527e-05  2.958e-05    0.516   0.6129
min_phred:norm_methoddefault             -1.528e-04  7.010e-05   -2.179   0.0446
min_phred:norm_methodRLE                 -1.444e-04  7.907e-05   -1.826   0.0866
min_phred:norm_methodTMM                 -1.634e-04  1.127e-04   -1.450   0.1665
min_phred:norm_methodupperquartile       -1.759e-04  8.119e-05   -2.167   0.0457
min_length:trim_poly_g1                   9.045e-06  1.038e-05    0.871   0.3964
min_length:trim_poly_x1                   2.235e-05  1.008e-05    2.217   0.0414
min_length:norm_methoddefault             1.946e-05  1.914e-05    1.017   0.3244
min_length:norm_methodRLE                 2.243e-05  1.683e-05    1.333   0.2014
min_length:norm_methodTMM                 2.623e-05  2.326e-05    1.128   0.2760
min_length:norm_methodupperquartile       1.299e-05  1.921e-05    0.676   0.5087
trim_poly_g1:trim_poly_x1                 4.868e-05  1.242e-04    0.392   0.7002
trim_poly_g1:norm_methoddefault           9.489e-05  2.362e-04    0.402   0.6932
trim_poly_g1:norm_methodRLE               2.261e-04  2.103e-04    1.075   0.2983
trim_poly_g1:norm_methodTMM              -4.858e-05  7.186e-04   -0.068   0.9469
trim_poly_g1:norm_methodupperquartile     2.541e-04  2.146e-04    1.184   0.2537
trim_poly_x1:norm_methoddefault           4.959e-04  3.430e-04    1.446   0.1675
trim_poly_x1:norm_methodRLE               4.520e-04  3.337e-04    1.354   0.1944
trim_poly_x1:norm_methodTMM               7.801e-04  7.187e-04    1.085   0.2938
trim_poly_x1:norm_methodupperquartile     5.391e-04  3.203e-04    1.683   0.1118

(Intercept)                              ***
aligner
min_phred
min_length
trim_poly_g1
trim_poly_x1
norm_methoddefault                       ***
norm_methodRLE                           ***
norm_methodTMM                           ***
norm_methodupperquartile                 ***
aligner:min_phred
aligner:min_length                       *
aligner:trim_poly_g1
aligner:trim_poly_x1
aligner:norm_methoddefault
aligner:norm_methodRLE
aligner:norm_methodTMM
aligner:norm_methodupperquartile
min_phred:min_length
min_phred:trim_poly_g1
min_phred:trim_poly_x1
```

```
min_phred:norm_methoddefault           *
min_phred:norm_methodRLE               .
min_phred:norm_methodTMM
min_phred:norm_methodupperquartile     *
min_length:trim_poly_g1
min_length:trim_poly_x1                *
min_length:norm_methoddefault
min_length:norm_methodRLE
min_length:norm_methodTMM
min_length:norm_methodupperquartile
trim_poly_g1:trim_poly_x1
trim_poly_g1:norm_methoddefault
trim_poly_g1:norm_methodRLE
trim_poly_g1:norm_methodTMM
trim_poly_g1:norm_methodupperquartile
trim_poly_x1:norm_methoddefault
trim_poly_x1:norm_methodRLE
trim_poly_x1:norm_methodTMM
trim_poly_x1:norm_methodupperquartile
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 6.591249e-09)

    Null deviance: 2.1702e-01  on 50  degrees of freedom
Residual deviance: 1.0546e-07  on 16  degrees of freedom
AIC: -803.1

Number of Fisher Scoring iterations: 2
```

```r
# Quasi GLM
quasi_fit <- DE_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent, effect_size_sd)) |>
    (\(x) glm(p_value_sd ~ (.)^2, family = quasi(), data = x))()
summary(quasi_fit)
```

```

Call:
glm(formula = p_value_sd ~ (.)^2, family = quasi(), data = x)

Coefficients: (5 not defined because of singularities)
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          3.725e-01  1.499e-03 248.488   <2e-16
alignersalmon                        5.976e-03  5.215e-03   1.146   0.2686
min_phred                            6.631e-05  6.504e-05   1.019   0.3231
min_length                          -9.363e-06  2.585e-05  -0.362   0.7219
trim_poly_g1                        -3.163e-04  3.668e-04  -0.862   0.4013
trim_poly_x1                        -5.951e-04  4.046e-04  -1.471   0.1607
norm_methoddefault                   1.805e-01  6.182e-04 292.046   <2e-16
norm_methodRLE                       1.803e-01  7.686e-04 234.522   <2e-16
norm_methodTMM                       1.806e-01  1.337e-03 135.071   <2e-16
norm_methodupperquartile             1.806e-01  7.139e-04 252.963   <2e-16
alignersalmon:min_phred             -3.086e-04  2.108e-04  -1.464   0.1625
alignersalmon:min_length             9.020e-05  2.924e-05   3.085   0.0071
alignersalmon:trim_poly_g1          -1.848e-03  1.567e-03  -1.180   0.2553
```

```
alignersalmon:trim_poly_x1                    NA         NA      NA       NA
alignersalmon:norm_methoddefault              NA         NA      NA       NA
alignersalmon:norm_methodRLE                  NA         NA      NA       NA
alignersalmon:norm_methodTMM                  NA         NA      NA       NA
alignersalmon:norm_methodupperquartile        NA         NA      NA       NA
min_phred:min_length                  -1.541e-07  1.087e-06  -0.142   0.8891
min_phred:trim_poly_g1                 1.935e-06  1.365e-05   0.142   0.8891
min_phred:trim_poly_x1                 8.002e-06  1.547e-05   0.517   0.6121
min_phred:norm_methoddefault          -7.192e-05  2.742e-05  -2.623   0.0185
min_phred:norm_methodRLE              -6.687e-05  3.345e-05  -1.999   0.0629
min_phred:norm_methodTMM              -7.941e-05  5.367e-05  -1.480   0.1584
min_phred:norm_methodupperquartile    -8.571e-05  3.476e-05  -2.466   0.0253
min_length:trim_poly_g1                5.124e-06  5.202e-06   0.985   0.3393
min_length:trim_poly_x1                1.372e-05  5.124e-06   2.677   0.0165
min_length:norm_methoddefault          1.189e-05  8.708e-06   1.365   0.1911
min_length:norm_methodRLE              1.357e-05  7.284e-06   1.863   0.0810
min_length:norm_methodTMM              1.517e-05  1.114e-05   1.362   0.1920
min_length:norm_methodupperquartile    7.880e-06  8.766e-06   0.899   0.3820
trim_poly_g1:trim_poly_x1              3.023e-05  6.288e-05   0.481   0.6372
trim_poly_g1:norm_methoddefault       -2.603e-05  1.023e-04  -0.254   0.8024
trim_poly_g1:norm_methodRLE            5.479e-05  8.612e-05   0.636   0.5337
trim_poly_g1:norm_methodTMM           -9.192e-05  3.695e-04  -0.249   0.8067
trim_poly_g1:norm_methodupperquartile  6.595e-05  8.865e-05   0.744   0.4677
trim_poly_x1:norm_methoddefault        2.769e-04  1.433e-04   1.933   0.0712
trim_poly_x1:norm_methodRLE            2.484e-04  1.370e-04   1.813   0.0886
trim_poly_x1:norm_methodTMM            4.243e-04  3.597e-04   1.179   0.2554
trim_poly_x1:norm_methodupperquartile  3.021e-04  1.279e-04   2.361   0.0312

(Intercept)                           ***
alignersalmon
min_phred
min_length
trim_poly_g1
trim_poly_x1
norm_methoddefault                    ***
norm_methodRLE                        ***
norm_methodTMM                        ***
norm_methodupperquartile              ***
alignersalmon:min_phred
alignersalmon:min_length              **
alignersalmon:trim_poly_g1
alignersalmon:trim_poly_x1
alignersalmon:norm_methoddefault
alignersalmon:norm_methodRLE
alignersalmon:norm_methodTMM
alignersalmon:norm_methodupperquartile
min_phred:min_length
min_phred:trim_poly_g1
min_phred:trim_poly_x1
min_phred:norm_methoddefault          *
min_phred:norm_methodRLE              .
min_phred:norm_methodTMM
min_phred:norm_methodupperquartile    *
min_length:trim_poly_g1
```

```
min_length:trim_poly_x1                *
min_length:norm_methoddefault
min_length:norm_methodRLE              .
min_length:norm_methodTMM
min_length:norm_methodupperquartile
trim_poly_g1:trim_poly_x1
trim_poly_g1:norm_methoddefault
trim_poly_g1:norm_methodRLE
trim_poly_g1:norm_methodTMM
trim_poly_g1:norm_methodupperquartile
trim_poly_x1:norm_methoddefault        .
trim_poly_x1:norm_methodRLE            .
trim_poly_x1:norm_methodTMM
trim_poly_x1:norm_methodupperquartile  *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 5.911428e-09)

    Null deviance: 2.1702e-01  on 50  degrees of freedom
Residual deviance: 9.4583e-08  on 16  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 2
```

Effect Size

```r
# Read in data.
DE_sd_df <- read.csv("./data/gen_samples/DE_sd_df.csv")
DE_sd_df <- DE_sd_df |>
    mutate(across(any_of(factors), ~ as.factor(.)))

# Classic LM
lm_fit <- DE_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent, p_value_sd)) |>
    (\(x) glm(effect_size_sd ~ (.)^2, family = gaussian(), data = x))()
summary(lm_fit)
```

```
Call:
glm(formula = effect_size_sd ~ (.)^2, family = gaussian(), data = x)

Coefficients: (5 not defined because of singularities)
                                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)                          1.586e+00  1.236e-02  128.263   <2e-16
alignersalmon                        2.029e-02  4.300e-02    0.472   0.6434
min_phred                            5.672e-04  5.364e-04    1.057   0.3060
min_length                           3.590e-04  2.132e-04    1.684   0.1115
trim_poly_g1                         1.834e-03  3.025e-03    0.606   0.5529
trim_poly_x1                         7.713e-04  3.336e-03    0.231   0.8201
norm_methoddefault                  -6.262e-01  5.098e-03 -122.843   <2e-16
norm_methodRLE                      -6.286e-01  6.339e-03  -99.170   <2e-16
norm_methodTMM                      -6.226e-01  1.103e-02  -56.456   <2e-16
norm_methodupperquartile            -6.285e-01  5.887e-03 -106.753   <2e-16
```

```
aligner:salmon:min_phred                -4.455e-04  1.738e-03   -0.256   0.8010
aligner:salmon:min_length                1.373e-04  2.411e-04    0.569   0.5771
aligner:salmon:trim_poly_g1             -2.198e-03  1.292e-02   -0.170   0.8670
aligner:salmon:trim_poly_x1                     NA         NA       NA       NA
aligner:salmon:norm_methoddefault               NA         NA       NA       NA
aligner:salmon:norm_methodRLE                   NA         NA       NA       NA
aligner:salmon:norm_methodTMM                   NA         NA       NA       NA
aligner:salmon:norm_methodupperquartile         NA         NA       NA       NA
min_phred:min_length                    -1.061e-05  8.968e-06   -1.183   0.2541
min_phred:trim_poly_g1                  -1.052e-04  1.126e-04   -0.935   0.3639
min_phred:trim_poly_x1                   2.339e-06  1.276e-04    0.018   0.9856
min_phred:norm_methoddefault            -7.699e-05  2.261e-04   -0.341   0.7379
min_phred:norm_methodRLE                -1.194e-04  2.758e-04   -0.433   0.6710
min_phred:norm_methodTMM                -2.276e-04  4.426e-04   -0.514   0.6141
min_phred:norm_methodupperquartile      -4.199e-05  2.866e-04   -0.146   0.8854
min_length:trim_poly_g1                 -1.820e-05  4.290e-05   -0.424   0.6771
min_length:trim_poly_x1                 -1.183e-07  4.226e-05   -0.003   0.9978
min_length:norm_methoddefault            7.390e-06  7.181e-05    0.103   0.9193
min_length:norm_methodRLE                6.560e-05  6.007e-05    1.092   0.2910
min_length:norm_methodTMM                1.131e-05  9.183e-05    0.123   0.9035
min_length:norm_methodupperquartile     -5.206e-05  7.229e-05   -0.720   0.4818
trim_poly_g1:trim_poly_x1               -1.131e-05  5.185e-04   -0.022   0.9829
trim_poly_g1:norm_methoddefault          2.237e-03  8.438e-04    2.651   0.0174
trim_poly_g1:norm_methodRLE              1.739e-03  7.102e-04    2.449   0.0262
trim_poly_g1:norm_methodTMM             -3.160e-04  3.047e-03   -0.104   0.9187
trim_poly_g1:norm_methodupperquartile    1.538e-03  7.310e-04    2.104   0.0516
trim_poly_x1:norm_methoddefault         -1.034e-03  1.182e-03   -0.875   0.3947
trim_poly_x1:norm_methodRLE             -1.177e-03  1.130e-03   -1.042   0.3131
trim_poly_x1:norm_methodTMM              7.284e-04  2.967e-03    0.246   0.8092
trim_poly_x1:norm_methodupperquartile   -8.725e-04  1.055e-03   -0.827   0.4204

(Intercept)                             ***
aligner:salmon
min_phred
min_length
trim_poly_g1
trim_poly_x1
norm_methoddefault                      ***
norm_methodRLE                          ***
norm_methodTMM                          ***
norm_methodupperquartile                ***
aligner:salmon:min_phred
aligner:salmon:min_length
aligner:salmon:trim_poly_g1
aligner:salmon:trim_poly_x1
aligner:salmon:norm_methoddefault
aligner:salmon:norm_methodRLE
aligner:salmon:norm_methodTMM
aligner:salmon:norm_methodupperquartile
min_phred:min_length
min_phred:trim_poly_g1
min_phred:trim_poly_x1
min_phred:norm_methoddefault
min_phred:norm_methodRLE
```

```
min_phred:norm_methodTMM
min_phred:norm_methodupperquartile
min_length:trim_poly_g1
min_length:trim_poly_x1
min_length:norm_methoddefault
min_length:norm_methodRLE
min_length:norm_methodTMM
min_length:norm_methodupperquartile
trim_poly_g1:trim_poly_x1
trim_poly_g1:norm_methoddefault        *
trim_poly_g1:norm_methodRLE            *
trim_poly_g1:norm_methodTMM
trim_poly_g1:norm_methodupperquartile  .
trim_poly_x1:norm_methoddefault
trim_poly_x1:norm_methodRLE
trim_poly_x1:norm_methodTMM
trim_poly_x1:norm_methodupperquartile
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.020262e-07)

    Null deviance: 2.6529e+00  on 50  degrees of freedom
Residual deviance: 6.4324e-06  on 16  degrees of freedom
AIC: -593.45

Number of Fisher Scoring iterations: 2
```

```
# Log-normal GLM
glm_log_fit <- DE_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent, p_value_sd)) |>
    (\(x) glm(effect_size_sd ~ (.)^2, family = gaussian(link = "log"), data = x))()
summary(glm_log_fit)
```

```

Call:
glm(formula = effect_size_sd ~ (.)^2, family = gaussian(link = "log"),
    data = x)

Coefficients: (5 not defined because of singularities)
                                    Estimate Std. Error  t value Pr(>|t|)
(Intercept)                        4.550e-01  1.222e-02   37.228   <2e-16
alignersalmon                      1.694e-02  4.387e-02    0.386   0.7045
min_phred                          6.045e-04  5.298e-04    1.141   0.2707
min_length                         3.283e-04  2.133e-04    1.539   0.1433
trim_poly_g1                       2.528e-03  2.917e-03    0.867   0.3989
trim_poly_x1                       6.724e-04  3.347e-03    0.201   0.8433
norm_methoddefault                -4.961e-01  4.135e-03 -119.973   <2e-16
norm_methodRLE                    -4.982e-01  5.681e-03  -87.692   <2e-16
norm_methodTMM                    -4.929e-01  1.088e-02  -45.309   <2e-16
norm_methodupperquartile          -4.986e-01  5.099e-03  -97.783   <2e-16
alignersalmon:min_phred           -2.898e-04  1.770e-03   -0.164   0.8720
alignersalmon:min_length           1.148e-04  2.454e-04    0.468   0.6462
alignersalmon:trim_poly_g1        -1.053e-03  1.318e-02   -0.080   0.9373
alignersalmon:trim_poly_x1                NA         NA       NA       NA
```

```
alignersalmon:norm_methoddefault               NA         NA       NA       NA
alignersalmon:norm_methodRLE                   NA         NA       NA       NA
alignersalmon:norm_methodTMM                   NA         NA       NA       NA
alignersalmon:norm_methodupperquartile         NA         NA       NA       NA
min_phred:min_length                    -1.079e-05  9.162e-06   -1.178   0.2560
min_phred:trim_poly_g1                  -1.255e-04  1.146e-04   -1.096   0.2894
min_phred:trim_poly_x1                   1.861e-05  1.316e-04    0.141   0.8893
min_phred:norm_methoddefault            -1.066e-04  1.772e-04   -0.602   0.5559
min_phred:norm_methodRLE                -1.591e-04  2.411e-04   -0.660   0.5185
min_phred:norm_methodTMM                -2.417e-04  4.297e-04   -0.563   0.5816
min_phred:norm_methodupperquartile      -6.738e-05  2.530e-04   -0.266   0.7934
min_length:trim_poly_g1                 -1.078e-05  4.089e-05   -0.264   0.7955
min_length:trim_poly_x1                 -1.418e-05  4.031e-05   -0.352   0.7296
min_length:norm_methoddefault            4.799e-05  6.699e-05    0.716   0.4841
min_length:norm_methodRLE                1.055e-04  5.331e-05    1.979   0.0653
min_length:norm_methodTMM                5.002e-05  8.962e-05    0.558   0.5845
min_length:norm_methodupperquartile     -1.085e-05  6.810e-05   -0.159   0.8754
trim_poly_g1:trim_poly_x1                2.503e-05  5.043e-04    0.050   0.9610
trim_poly_g1:norm_methoddefault          1.774e-03  7.428e-04    2.389   0.0296
trim_poly_g1:norm_methodRLE              1.206e-03  5.947e-04    2.028   0.0595
trim_poly_g1:norm_methodTMM             -8.748e-04  3.110e-03   -0.281   0.7821
trim_poly_g1:norm_methodupperquartile    1.055e-03  6.145e-04    1.716   0.1054
trim_poly_x1:norm_methoddefault         -8.201e-04  1.017e-03   -0.807   0.4317
trim_poly_x1:norm_methodRLE             -9.160e-04  9.469e-04   -0.967   0.3478
trim_poly_x1:norm_methodTMM              1.007e-03  2.975e-03    0.339   0.7393
trim_poly_x1:norm_methodupperquartile   -6.844e-04  8.551e-04   -0.800   0.4352

(Intercept)                             ***
alignersalmon
min_phred
min_length
trim_poly_g1
trim_poly_x1
norm_methoddefault                      ***
norm_methodRLE                          ***
norm_methodTMM                          ***
norm_methodupperquartile                ***
alignersalmon:min_phred
alignersalmon:min_length
alignersalmon:trim_poly_g1
alignersalmon:trim_poly_x1
alignersalmon:norm_methoddefault
alignersalmon:norm_methodRLE
alignersalmon:norm_methodTMM
alignersalmon:norm_methodupperquartile
min_phred:min_length
min_phred:trim_poly_g1
min_phred:trim_poly_x1
min_phred:norm_methoddefault
min_phred:norm_methodRLE
min_phred:norm_methodTMM
min_phred:norm_methodupperquartile
min_length:trim_poly_g1
min_length:trim_poly_x1
```

```
min_length:norm_methoddefault
min_length:norm_methodRLE               .
min_length:norm_methodTMM
min_length:norm_methodupperquartile
trim_poly_g1:trim_poly_x1
trim_poly_g1:norm_methoddefault         *
trim_poly_g1:norm_methodRLE             .
trim_poly_g1:norm_methodTMM
trim_poly_g1:norm_methodupperquartile
trim_poly_x1:norm_methoddefault
trim_poly_x1:norm_methodRLE
trim_poly_x1:norm_methodTMM
trim_poly_x1:norm_methodupperquartile
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 4.074639e-07)

    Null deviance: 2.6529e+00  on 50  degrees of freedom
Residual deviance: 6.5194e-06  on 16  degrees of freedom
AIC: -592.77

Number of Fisher Scoring iterations: 2
```

```r
# Quasi GLM
quasi_fit <- DE_sd_df |>
    select(-c(runtime_sec, gene_overlap_percent, p_value_sd)) |>
    (\(x) glm(effect_size_sd ~ (.)^2, family = quasi(), data = x))()
summary(quasi_fit)
```

```
Call:
glm(formula = effect_size_sd ~ (.)^2, family = quasi(), data = x)

Coefficients: (5 not defined because of singularities)
                                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)                          1.586e+00  1.236e-02  128.263   <2e-16
alignersalmon                        2.029e-02  4.300e-02    0.472   0.6434
min_phred                            5.672e-04  5.364e-04    1.057   0.3060
min_length                           3.590e-04  2.132e-04    1.684   0.1115
trim_poly_g1                         1.834e-03  3.025e-03    0.606   0.5529
trim_poly_x1                         7.713e-04  3.336e-03    0.231   0.8201
norm_methoddefault                  -6.262e-01  5.098e-03 -122.843   <2e-16
norm_methodRLE                      -6.286e-01  6.339e-03  -99.170   <2e-16
norm_methodTMM                      -6.226e-01  1.103e-02  -56.456   <2e-16
norm_methodupperquartile            -6.285e-01  5.887e-03 -106.753   <2e-16
alignersalmon:min_phred             -4.455e-04  1.738e-03   -0.256   0.8010
alignersalmon:min_length             1.373e-04  2.411e-04    0.569   0.5771
alignersalmon:trim_poly_g1          -2.198e-03  1.292e-02   -0.170   0.8670
alignersalmon:trim_poly_x1                  NA         NA       NA       NA
alignersalmon:norm_methoddefault            NA         NA       NA       NA
alignersalmon:norm_methodRLE                NA         NA       NA       NA
alignersalmon:norm_methodTMM                NA         NA       NA       NA
alignersalmon:norm_methodupperquartile      NA         NA       NA       NA
min_phred:min_length                -1.061e-05  8.968e-06   -1.183   0.2541
```

```
min_phred:trim_poly_g1                -1.052e-04  1.126e-04   -0.935   0.3639
min_phred:trim_poly_x1                 2.339e-06  1.276e-04    0.018   0.9856
min_phred:norm_methoddefault          -7.699e-05  2.261e-04   -0.341   0.7379
min_phred:norm_methodRLE              -1.194e-04  2.758e-04   -0.433   0.6710
min_phred:norm_methodTMM              -2.276e-04  4.426e-04   -0.514   0.6141
min_phred:norm_methodupperquartile    -4.199e-05  2.866e-04   -0.146   0.8854
min_length:trim_poly_g1               -1.820e-05  4.290e-05   -0.424   0.6771
min_length:trim_poly_x1               -1.183e-07  4.226e-05   -0.003   0.9978
min_length:norm_methoddefault          7.390e-06  7.181e-05    0.103   0.9193
min_length:norm_methodRLE              6.560e-05  6.007e-05    1.092   0.2910
min_length:norm_methodTMM              1.131e-05  9.183e-05    0.123   0.9035
min_length:norm_methodupperquartile   -5.206e-05  7.229e-05   -0.720   0.4818
trim_poly_g1:trim_poly_x1             -1.131e-05  5.185e-04   -0.022   0.9829
trim_poly_g1:norm_methoddefault        2.237e-03  8.438e-04    2.651   0.0174
trim_poly_g1:norm_methodRLE            1.739e-03  7.102e-04    2.449   0.0262
trim_poly_g1:norm_methodTMM           -3.160e-04  3.047e-03   -0.104   0.9187
trim_poly_g1:norm_methodupperquartile  1.538e-03  7.310e-04    2.104   0.0516
trim_poly_x1:norm_methoddefault       -1.034e-03  1.182e-03   -0.875   0.3947
trim_poly_x1:norm_methodRLE           -1.177e-03  1.130e-03   -1.042   0.3131
trim_poly_x1:norm_methodTMM            7.284e-04  2.967e-03    0.246   0.8092
trim_poly_x1:norm_methodupperquartile -8.725e-04  1.055e-03   -0.827   0.4204

(Intercept)                           ***
alignersalmon
min_phred
min_length
trim_poly_g1
trim_poly_x1
norm_methoddefault                    ***
norm_methodRLE                        ***
norm_methodTMM                        ***
norm_methodupperquartile              ***
alignersalmon:min_phred
alignersalmon:min_length
alignersalmon:trim_poly_g1
alignersalmon:trim_poly_x1
alignersalmon:norm_methoddefault
alignersalmon:norm_methodRLE
alignersalmon:norm_methodTMM
alignersalmon:norm_methodupperquartile
min_phred:min_length
min_phred:trim_poly_g1
min_phred:trim_poly_x1
min_phred:norm_methoddefault
min_phred:norm_methodRLE
min_phred:norm_methodTMM
min_phred:norm_methodupperquartile
min_length:trim_poly_g1
min_length:trim_poly_x1
min_length:norm_methoddefault
min_length:norm_methodRLE
min_length:norm_methodTMM
min_length:norm_methodupperquartile
trim_poly_g1:trim_poly_x1
```

```
trim_poly_g1:norm_methoddefault        *
trim_poly_g1:norm_methodRLE            *
trim_poly_g1:norm_methodTMM
trim_poly_g1:norm_methodupperquartile  .
trim_poly_x1:norm_methoddefault
trim_poly_x1:norm_methodRLE
trim_poly_x1:norm_methodTMM
trim_poly_x1:norm_methodupperquartile
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasi family taken to be 4.020262e-07)

    Null deviance: 2.6529e+00  on 50  degrees of freedom
Residual deviance: 6.4324e-06  on 16  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 2
```

Bayesian

We know that

$$Y^2 \xrightarrow{d} \mathcal{N}(\mu, \sigma^2) \tag{5}$$

by the central limit theorem since Y^2 is an average. This is not completely accurate because $Y^2 > 0$, but if $\mu >> 0$, then the truncation is inconsequential. Using the 1-1 transformation formula we can derive that the distribution of Y must be:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(y^2-\mu)^2}{2\sigma^2}} \cdot 2y \tag{6}$$

Unfortunately this doesn't have a close form expectation, which makes it difficult to model $\mathbb{E}\ Y = X\beta$. Since a mean and variance function can be derived, it is possible to fit a model with something like general estimating equation, but there a two key problems. First, the mean function is an integral which most likely needs to be approximated. Second, the necessary link function results in a non-linear relationship between the $\hat{\beta}$s and Y making interpretation difficult.

Instead, we will build from the fact that $Y \geq 0$. There are several common likelihoods that have support $[0, \infty)$ such as the log-normal, gamma, weibull, etc. Since we are looking to model $\mathbb{E}\ Y = X\beta$, the log-normal is the simplest choice since the default parameterization is a location-scale family.

Consider the following Bayesian Hierarchical Model:

$$\begin{aligned} Y_i &\sim \log -\mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i &= X_i\beta \\ \sigma_i &= a \cdot \mu_i^b \\ \beta &\sim \mathcal{N}(0, 100) \\ a &\sim \text{Gamma}(c, d) \\ b &\sim \mathcal{N}(0, 10) \end{aligned} \tag{7}$$

This set up has a couple of key advantages.

1. The interpretation is still linear on the Y scale since we are modeling $\mathbb{E}\ Y_i = \mu_i = X_i\beta$.

2. Natural parameter shrinkage via the prior on β. Handles multicollinearity and high dimensionality of X.

3. Does not assume constant variance. Specifically, we are applying the variance-power law from the Tweedie family of distributions, which the log-Normal is a member.

$$\text{Var } Y \propto (\mathbb{E}\ Y)^p \tag{8}$$

   $a > 0$ and represents a common variance scale ie if $b = 0$ we recover the classical log-Normal regression model. $b \in \mathbb{R}$ where $b > 0$ indicates over-dispersion and $b < 0$ indicates under-dispersion.

4. We can use the posterior predictive distribution to check whether the model is consistent with the fact that $Y^2 \sim \mathcal{N}$.