Phage Differential Abundance Using GPD count data

Ilhan Cem Duru

2024-08-02

Load necessary libraries

```
library(DESeq2)
library(ggplot2)
library(reshape2)
library(apeglm)
```

Import count and metadata and create DESeqDataSet object

```
countdata <- read.table("features_reads_raw_count.tsv", header=TRUE, row.names=1)
head(countdata)
```

```
##        C102 C103 C104 C105 C1 C107 C111 C114 C116 C118 C119 C123 C124 C134 C135
## ivig_1    0    0    0    0  0    0    0    0    0    0    0    0    0    0    0
## ivig_2    0    0    0    0  0    0    0    0    0    0    0    0    0    0    0
## ivig_3    0    0    0    0  0    0    0    0    0    0    0    0    0    0    0
## ivig_6    0    0    0    0  0    0    0    0    0    0    0    0    0    0    0
## ivig_7    0    0    0    0  0    0    0    0    0    0    0    0    0    0    0
## ivig_8    0    0    0    0  0    0    0    0    0    0    0    0    0    0    0
##        C136 C137 C140 C142 C146 C147 C148 C15 C152IIP C18 C19 C20 C21 C23 C24
## ivig_1    0    0    0    0    0    0    0   0       0   0   0   0   0   0   0
## ivig_2    0    0    0    0    0    0    0   0       0   0   0   0   0   0   0
## ivig_3    0    0    0    0    0    0    0   0       0   0   0   0   0   0   0
## ivig_6    0    0    0    0    0    0    0   0       0   0   0   0   0   0   0
## ivig_7    0    0    0    0    0    0    0   0       0   0   0   0   0   0   0
## ivig_8    0    0    0    0    0    0    0   0       0   0   0   0   0   0   0
##        C26 C28 C30 C32 C33 C34old C35 C40 C44 C46 C47 C48 C49 C5 C51 C54 C59
## ivig_1   0   0   0   0   0      0   0   0   0   0   0   0   0  0   0   0   0
## ivig_2   0   0   0   0   0      0   0   0   0   0   0   0   0  0   0   0   0
## ivig_3   0   0   0   0   0      0   0   0   0   0   0   0   0  0   0   0   0
## ivig_6   0   0   0   0   0      0   0   0   0   0   0   0   0  0   0   0   0
## ivig_7   0   0   0   0   0      0   0   0   0   0   0   0   0  0   0   0   0
## ivig_8   0   0   0   0   0      0   0   0   0   0   0   0   0  0   0   0   0
##        C65 C68 C69 C70 C7 C72 C74 C75 C76II C80 C82 C85 C86 C87 C88 C89 C90 C9
## ivig_1   0   0   0   0  0   0   0   0     0   0   0   0   0   0   0   0   0  0
## ivig_2   0   0   0   0  0   0   0   0     0   0   0   0   0   0   0   0   0  0
## ivig_3   0   0   0   0  0   0   0   0     0   0   0   0   0   0   0   0   0  0
## ivig_6   0   0   0   0  0   0   0   0     0   0   0   0   1   0   0   0   0  0
## ivig_7   0   0   0   0  0   0   0   0     0   0   0   0   0   0   0   0   0  0
## ivig_8   0   0   0   0  0   0   0   0     0   0   0   0   0   0   0   0   0  0
##        C95 C96 C98 P100 P103 P104IIP P105 P107 P10old P114 P115 P116 P118 P119
## ivig_1   0   0   0    0    0       0    0    0      0    0    0    0    0    0
## ivig_2   0   0   0    0    0       0    0    0      0    0    0    0    0    0
## ivig_3   0   0   0    0    0       0    0    0      0    0    0    0    0    0
## ivig_6   0   0   0    0    0       0    0    0      0    0    0    0    0    0
## ivig_7   0   0   0    0    0       0    0    0      0    0    0    0    0    0
## ivig_8   0   0   0    0    0       0    0    0      0    0    0    0    0    0
##        P11old P120 P12 P14 P15 P16 P17 P18 P19 P20 P24 P26 P28 P31 P34 P37 P38
## ivig_1      0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## ivig_2      0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## ivig_3      0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## ivig_6      0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## ivig_7      0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## ivig_8      0    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##        P4 P42 P43 P45 P46 P47 P48 P50 P5 P51 P52 P53 P56 P57 P58 P59 P60 P61
## ivig_1  0   0   0   0   0   0   0   0  0   0   0   0   0   0   0   0   0   0
## ivig_2  0   0   0   0   0   0   0   0  0   0   0   0   0   0   0   0   0   0
## ivig_3  0   0   0   0   0   0   0   0  0   0   0   0   0   0   0   0   0   0
## ivig_6  0   0   0   0   0   0   0   0  0   0   0   0   0   0   0   0   0   0
## ivig_7  0   0   0   0   0   0   0   0  0   0   0   0   0   0   0   0   0   0
## ivig_8  0   0   0   0   0   0   0   0  0   0   0   0   0   0   0   0   0   0
##        P62P P63 P66 P67 P68 P69 P70 P71 P72 P73 P74 P77 P79 P8 P83 P85 P87 P88
## ivig_1    0   0   0   0   0   0   0   0   0   0   0   0   0  0   0   0   0   0
## ivig_2    0   0   0   0   0   0   0   0   0   0   0   0   0  0   0   0   0   0
## ivig_3    0   0   0   0   0   0   0   0   0   0   0   0   0  0   0   0   0   0
## ivig_6    0   0   0   0   0   0   0   0   0   0   0   0   0  0   0   0   0   0
## ivig_7    0   0   0   0   0   0   0   0   0   0   0   0   0  0   0   0   0   0
## ivig_8   24   0   0   0   0   0   0   0   0   0   0   0   0  0   0   0   0   0
##        P9 P94 P95II P99
## ivig_1  0   0     0   0
## ivig_2  0   0     0   0
## ivig_3  0   0     0   0
## ivig_6  0   0     0   0
## ivig_7  0   0     0   0
## ivig_8  0   0     0   0
```

```
metadata <- read.table("pd_meta_with_ffq_and_scfa_only_oursamples_3variables.csv",
                       header=TRUE, row.names=1,sep=",")
head(metadata)
```

```
##      Group gender age_at_stool_collection      BMI
## C102     C      M                       69 30.27371
## C103     C      F                       66 23.05175
## C104     C      M                       71 29.56590
## C105     C      M                       58 26.17134
## C1       C      M                       73 26.25072
## C107     C      M                       64 24.09908
```

```
# include relevant covariates in the design
#formula to account for potential confounding factors
dds <- DESeqDataSetFromMatrix(countData = countdata,
                              colData = metadata,
                              design = ~ gender + age_at_stool_collection + Group)

dds
```

```
## class: DESeqDataSet
## dim: 142809 136
## metadata(1): version
## assays(1): counts
## rownames(142809): ivig_1 ivig_2 ... uvig_598943 uvig_598946
## rowData names(0):
## colnames(136): C102 C103 ... P95II P99
## colData names(4): Group gender age_at_stool_collection BMI
```

```
#filter out low count phages (total count < 30, 136 samples in total
#so 30/136=0.22 per sample) and minimum number of samples with at least 1 count.
#At least 15 samples should have at least 1 count. By this way we can be
#efficient in terms of computational time, because if only couple of samples
#have counts, it is not possible to make any statistical inference.
keep <- rowSums(counts(dds)) >= 30 & rowSums(counts(dds) >= 1) >= 15
dds <- dds[keep,]
dds
```

```
## class: DESeqDataSet
## dim: 41533 136
## metadata(1): version
## assays(1): counts
## rownames(41533): ivig_14 ivig_16 ... uvig_598842 uvig_598850
## rowData names(0):
## colnames(136): C102 C103 ... P95II P99
## colData names(4): Group gender age_at_stool_collection BMI
```

Run DESeq2

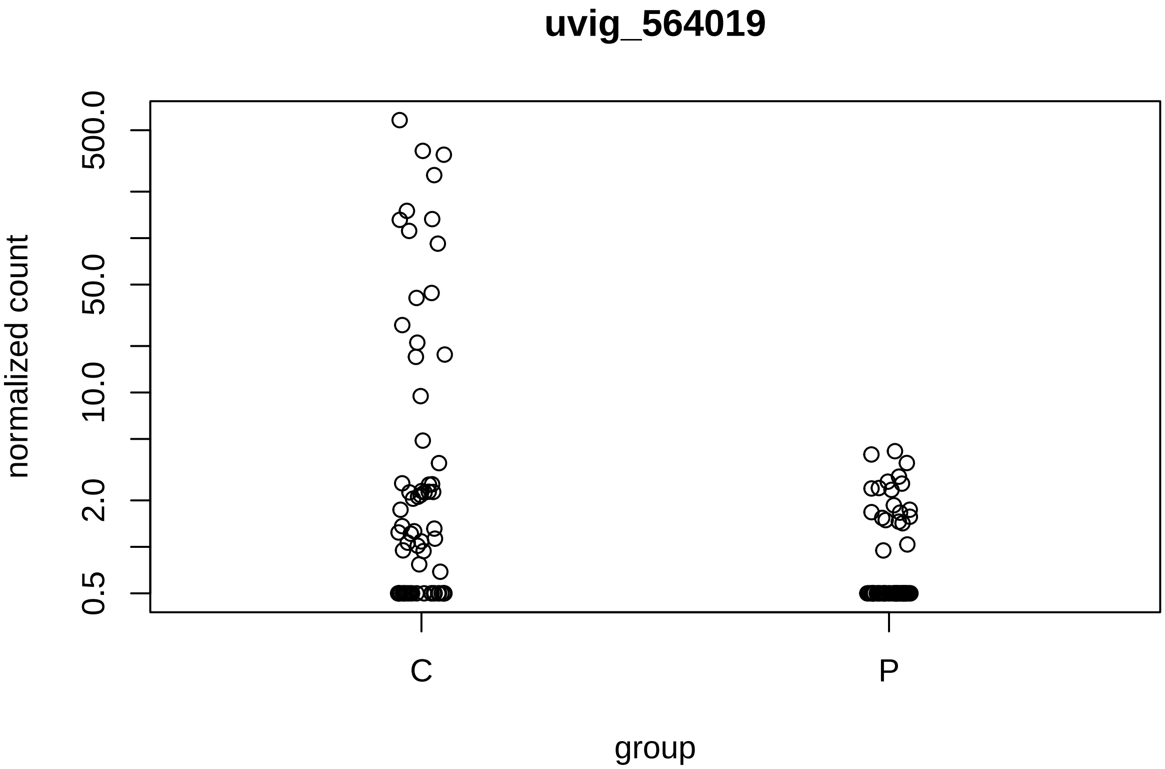
```
dds <- DESeq(dds)
#resultsNames(dds)
# Apply shrinkage to log fold changes using apeglm method
resLFC <- lfcShrink(dds, coef="Group_P_vs_C", type="apeglm")

# Filter out low fold2 changes and high p-values
res_clean <- resLFC[!is.na(resLFC$log2FoldChange) & !is.na(resLFC$padj), ]
res_filtered <- res_clean[abs(res_clean$log2FoldChange) > 0.25 & res_clean$padj < 0.05, ]


# Order results by adjusted p-value
resOrdered <- res_filtered[order(res_filtered$padj),]
head(resOrdered)
```

```
## log2 fold change (MAP): Group P vs C
## Wald test p-value: Group P vs C
## DataFrame with 6 rows and 5 columns
##              baseMean log2FoldChange     lfcSE      pvalue        padj
##             <numeric>      <numeric> <numeric>   <numeric>   <numeric>
## uvig_564019   17.7048       -5.87094  0.679614 4.20874e-19 9.09257e-15
## uvig_127743  194.7976        3.29836  0.493847 1.14606e-14 1.23798e-10
## uvig_355255  344.7418        3.51615  0.640567 1.83620e-13 1.32231e-09
## uvig_285529   76.6986        4.86810  1.044309 9.67346e-12 5.22463e-08
## uvig_196      20.8174        3.91775  0.657503 2.28185e-11 9.85944e-08
## uvig_240452   62.6316        2.83556  0.460793 4.18946e-11 1.50849e-07
```

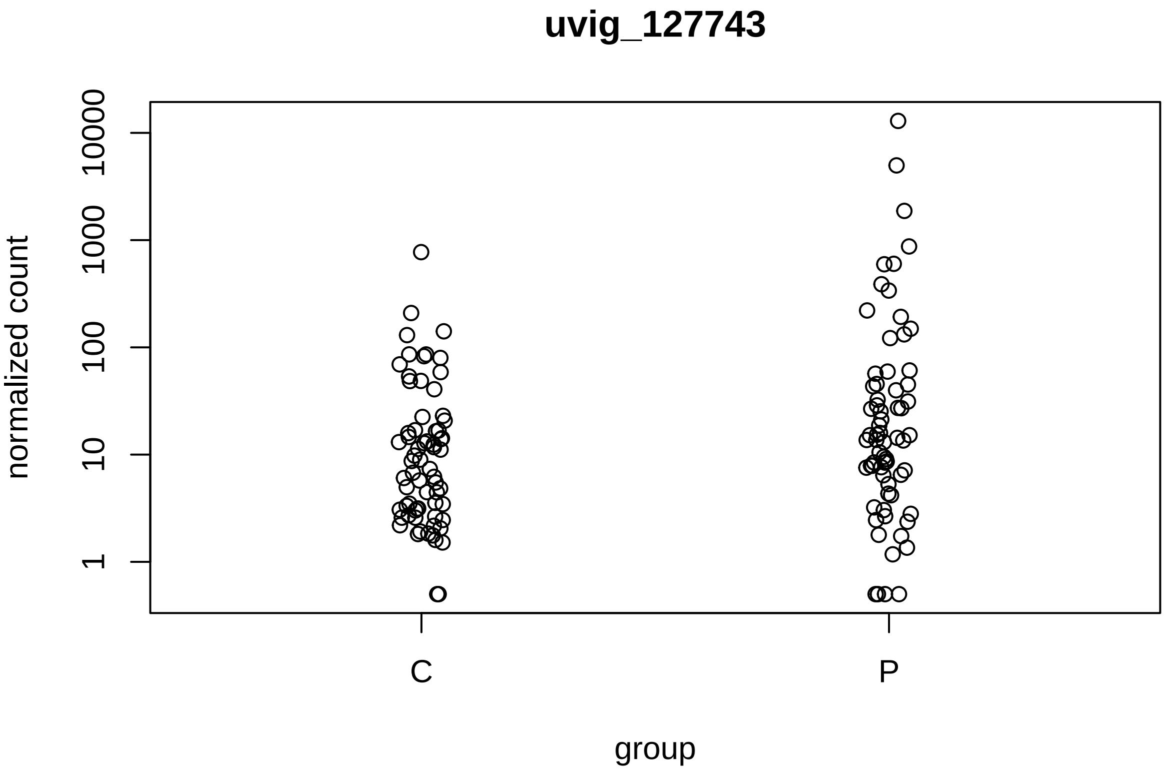
```
# Summarize results
summary(res_filtered, alpha=0.05)
```

```
##
## out of 1859 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 820, 44%
## LFC < 0 (down)     : 1039, 56%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 1)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
#plot one of the phages with higher abundance in Control group (minus log2fold)
plotCounts(dds, gene="uvig_564019", intgroup="Group")
```



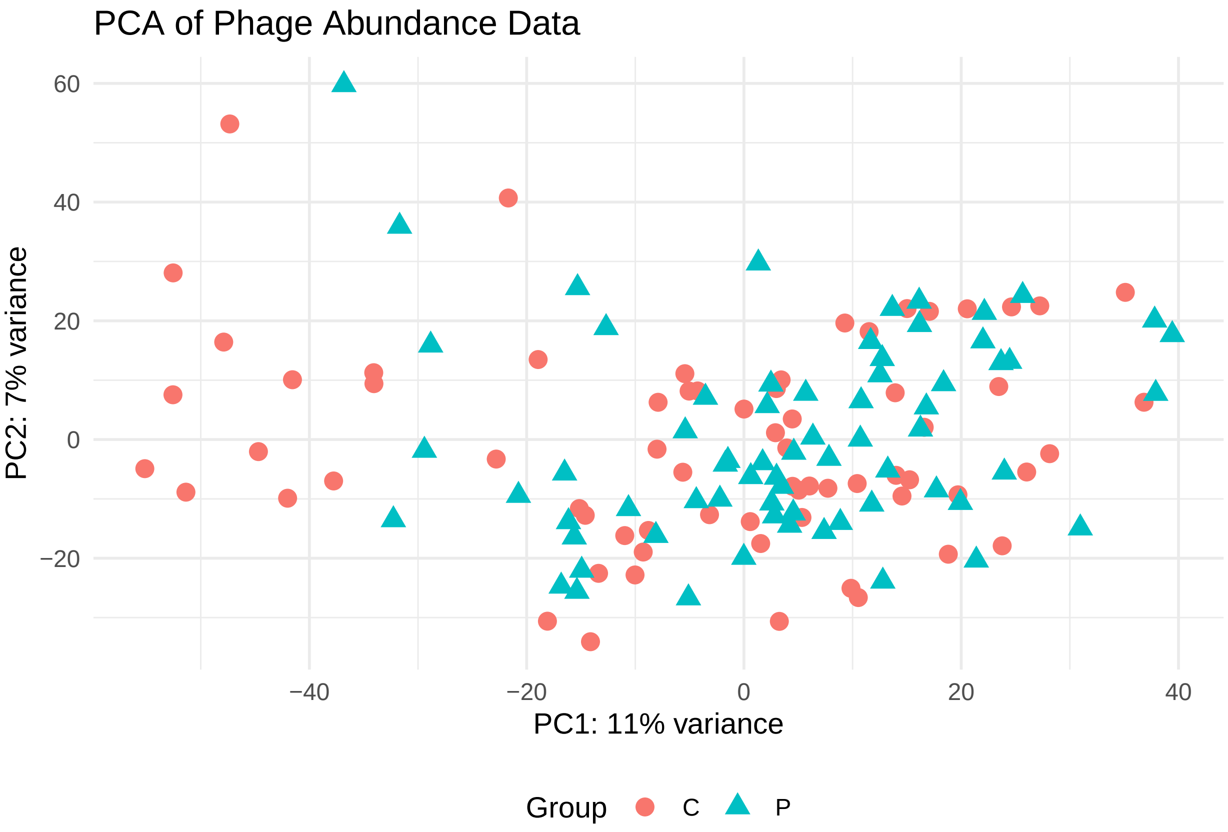
```
#plot one of the phages with higher abundance in Patient group (plus log2fold)
plotCounts(dds, gene="uvig_127743", intgroup="Group")
```



# plot

PCA

```
# 1. Perform variance stabilizing transformation
vsd <- vst(dds, blind=FALSE)

# 2. Calculate PCA
pcaData <- plotPCA(vsd, intgroup=c("Group"), returnData=TRUE)

# 3. Calculate the percentage of variance explained by each principal component
percentVar <- round(100 * attr(pcaData, "percentVar"))

# 4. Create the PCA plot
ggplot(pcaData, aes(x = PC1, y = PC2, color = Group, shape = Group)) +
  geom_point(size = 3) +
  xlab(paste0("PC1: ", percentVar[1], "% variance")) +
  ylab(paste0("PC2: ", percentVar[2], "% variance")) +
  ggtitle("PCA of Phage Abundance Data") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



```r
# 5. Save the plot
ggsave("PCA_plot.png", width = 8, height = 6)
```

Export significant results

```r
sigResults <- subset(resLFC, padj < 0.05)
write.csv(as.data.frame(sigResults), file = "significant_phages_shrinkage.csv")
# also export normalized counts
normalized_counts <- counts(dds, normalized=TRUE)
write.csv(as.data.frame(normalized_counts), file = "normalized_counts.csv")
# export all results
write.csv(as.data.frame(resLFC), file = "all_phages_shrinkage.csv")
```

Session Info for reproducibility

```r
sessionInfo()
```

```
## R version 4.4.1 (2024-06-14)
## Platform: x86_64-pc-linux-gnu
## Running under: Ubuntu 22.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.20.so;  LAPACK version 3.10.0
##
```

```
## locale:
##  [1] LC_CTYPE=en_GB.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_GB.UTF-8        LC_COLLATE=en_GB.UTF-8
##  [5] LC_MONETARY=en_GB.UTF-8    LC_MESSAGES=en_GB.UTF-8
##  [7] LC_PAPER=en_GB.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_GB.UTF-8 LC_IDENTIFICATION=C
##
## time zone: Europe/Helsinki
## tzcode source: system (glibc)
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] apeglm_1.26.1               reshape2_1.4.4
##  [3] ggplot2_3.5.1               DESeq2_1.44.0
##  [5] SummarizedExperiment_1.34.0 Biobase_2.64.0
##  [7] MatrixGenerics_1.16.0       matrixStats_1.3.0
##  [9] GenomicRanges_1.56.1        GenomeInfoDb_1.40.1
## [11] IRanges_2.38.1              S4Vectors_0.42.1
## [13] BiocGenerics_0.50.0
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.5            xfun_0.46               lattice_0.22-5
##  [4] numDeriv_2016.8-1.1     vctrs_0.6.5             tools_4.4.1
##  [7] generics_0.1.3          parallel_4.4.1          tibble_3.2.1
## [10] fansi_1.0.6             highr_0.11              pkgconfig_2.0.3
## [13] Matrix_1.6-5            lifecycle_1.0.4         GenomeInfoDbData_1.2.12
## [16] farver_2.1.2            compiler_4.4.1          stringr_1.5.1
## [19] tinytex_0.52            munsell_0.5.1           codetools_0.2-19
## [22] htmltools_0.5.8.1       yaml_2.3.10             pillar_1.9.0
## [25] crayon_1.5.3            MASS_7.3-61             BiocParallel_1.38.0
## [28] DelayedArray_0.30.1     emdbook_1.3.13          abind_1.4-5
## [31] bdsmatrix_1.3-7         tidyselect_1.2.1        locfit_1.5-9.10
## [34] digest_0.6.36           mvtnorm_1.2-5           stringi_1.8.4
## [37] dplyr_1.1.4             labeling_0.4.3          fastmap_1.2.0
## [40] grid_4.4.1              colorspace_2.1-1        cli_3.6.3
## [43] SparseArray_1.4.8       magrittr_2.0.3          S4Arrays_1.4.1
## [46] utf8_1.2.4              withr_3.0.1             scales_1.3.0
## [49] UCSC.utils_1.0.0        rmarkdown_2.27          XVector_0.44.0
## [52] httr_1.4.7              coda_0.19-4.1           evaluate_0.24.0
## [55] knitr_1.48              bbmle_1.0.25.1          rlang_1.1.4
## [58] Rcpp_1.0.13             glue_1.7.0              rstudioapi_0.16.0
## [61] jsonlite_1.8.8          R6_2.5.1                plyr_1.8.9
## [64] zlibbioc_1.50.0
```