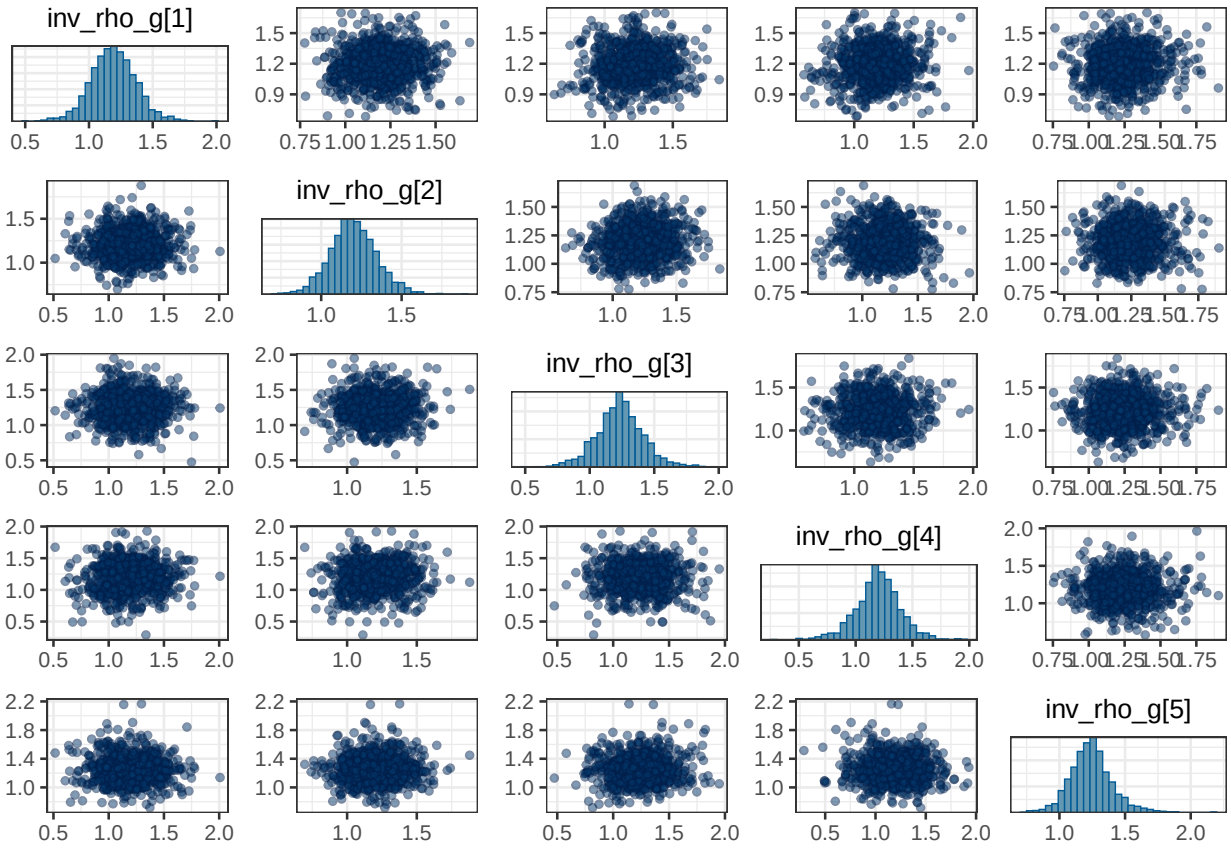
Demo analysis for TraDIS infection screens with cmdstanr

Laura Jenniches

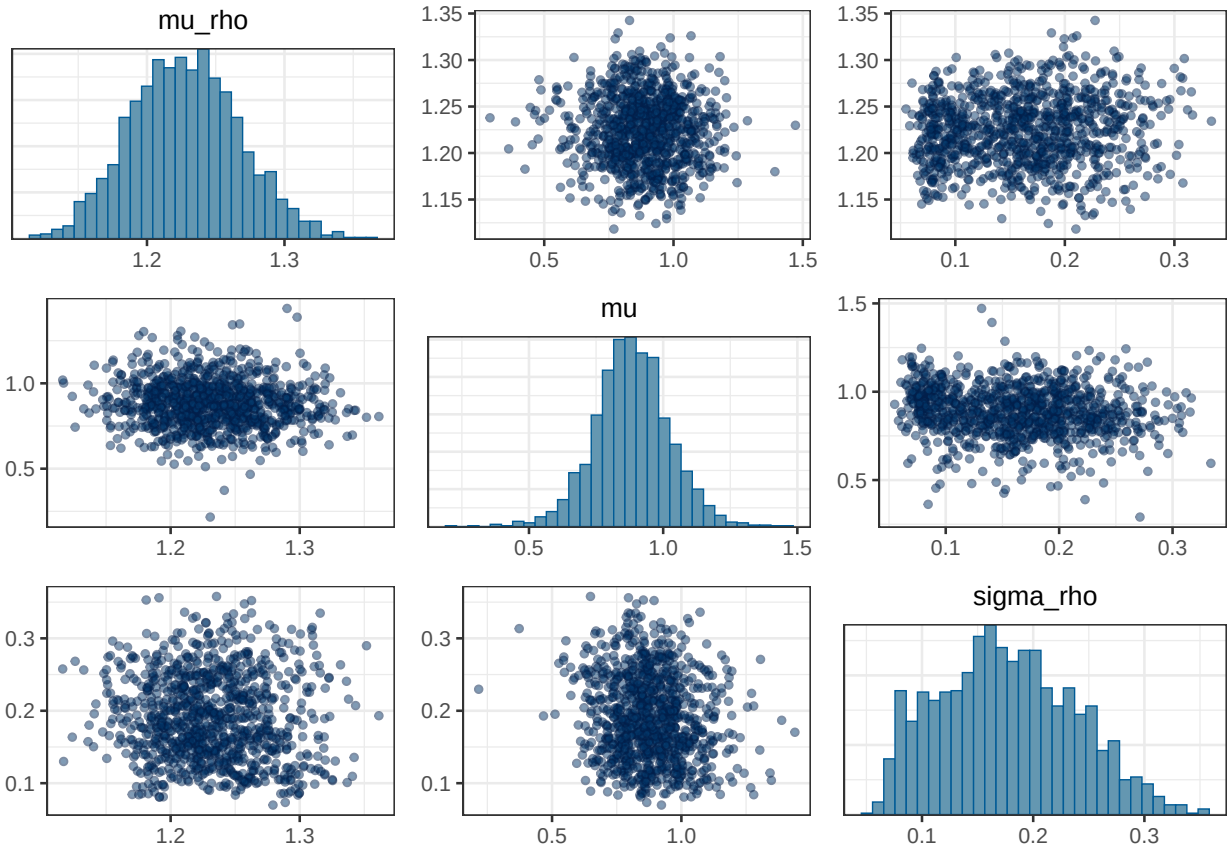
08/1/2025

1) Load demo count data. counts_pseudo_demo.csv contains 200 selected (neutral) pseudo genes of ten input-output pairs for normalization. counts_all_demo.csv contains 200 selected genes (100 with a fitness effect and 100 neutral) for which the fitness scores will be determined.
2) Create dfit files which contain the information required for the stan model.
3) Create data list for cmdstanr (normalization).
4) Run cmdstan and save fit results. Alternatively, data can be exported with stan_rdump and the sampling can be performed with cmdstan. Make sure cmdstan is installed: https://mc-stan.org/cmdstanr/articles/cmdstanr.html
5) Bayesian diagnostics: pair and trace plots, NUTS energy plots. There may be some divergent transitions (pair and trace plots) because fitting a hierarchical Bayesian model to a small number of genes is suboptimal. This results in an inefficient exploration of the parameter space (NUTS energy plot). For a full dataset, there should be no divergent transitions. This step can be accelerated when initializing the parameters adequately.



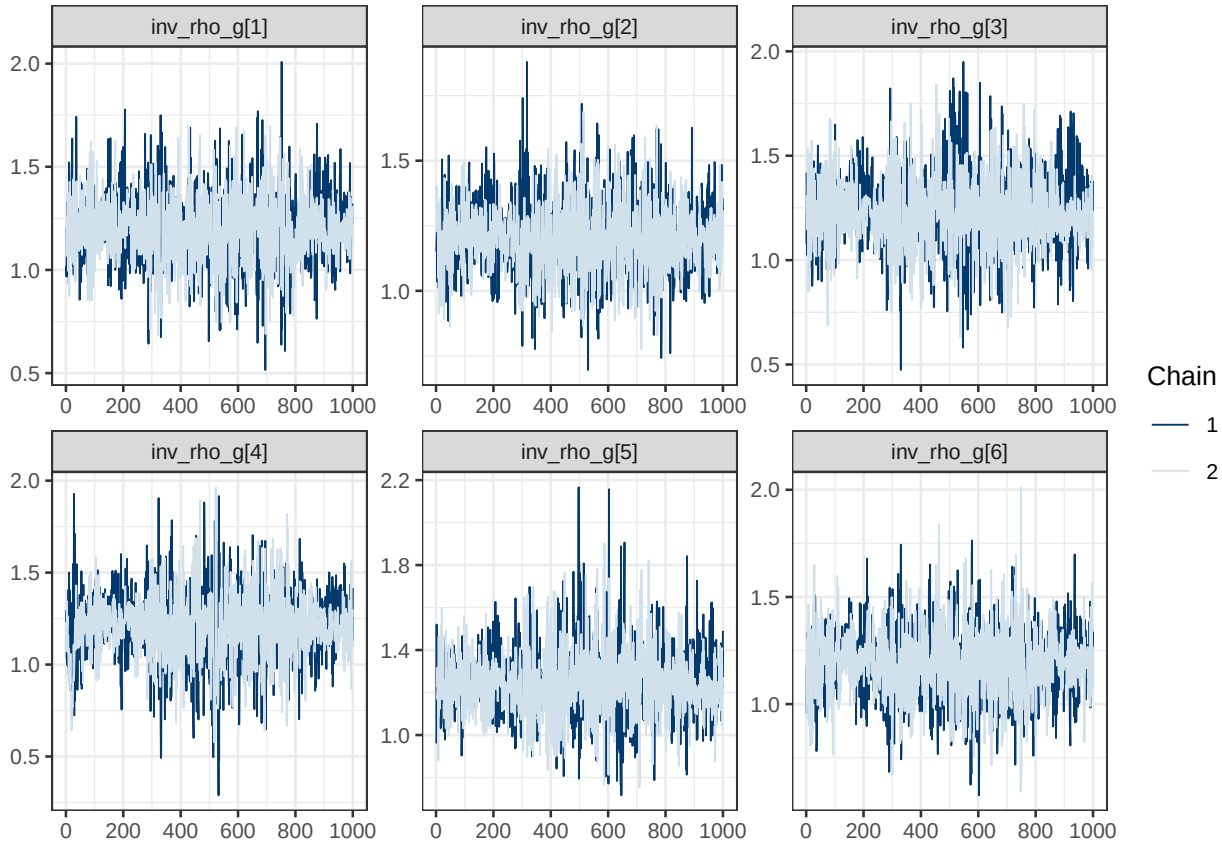
```
## pdf
##   2
```

```
## pdf
##   2
```



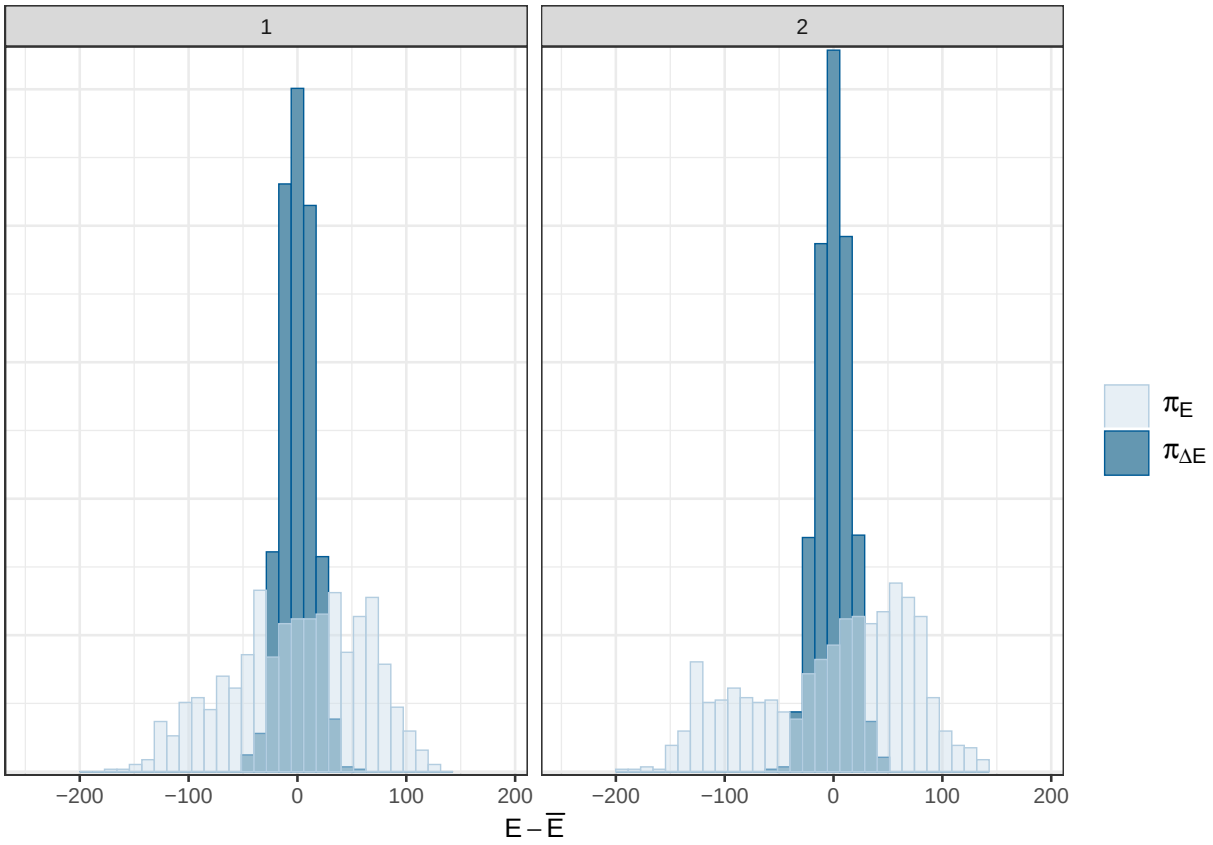

```
## pdf
##   2
```

```
## pdf
##   2
```


inv_rho_g[1]
inv_rho_g[2]
inv_rho_g[3]
inv_rho_g[4]
inv_rho_g[5]
inv_rho_g[6]
Chain
1
2

```
## pdf
##   2
```

```
## pdf
##   2
```




```
## pdf
##   2
```

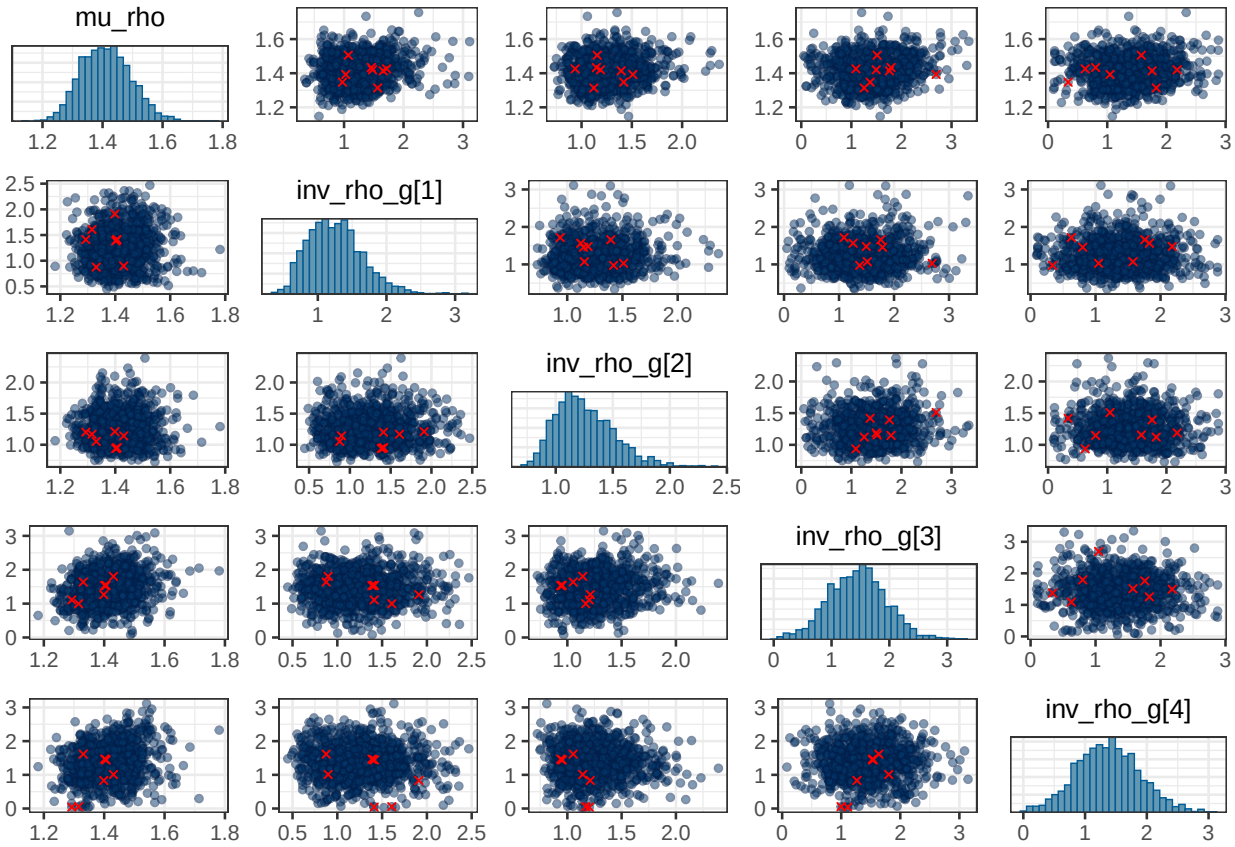
```
## pdf
##   2
```

6) Extract technical (normalization) parameters from fitting ZINB_pair2_norm. With the demo dataset, the following results should be optained: a_in ≈ (0.05, -0.68), b ≈ -0.35.

7) Create data list for cmdstanr to calculate fitness scores (logFCs) for the 200 genes demo dataset.

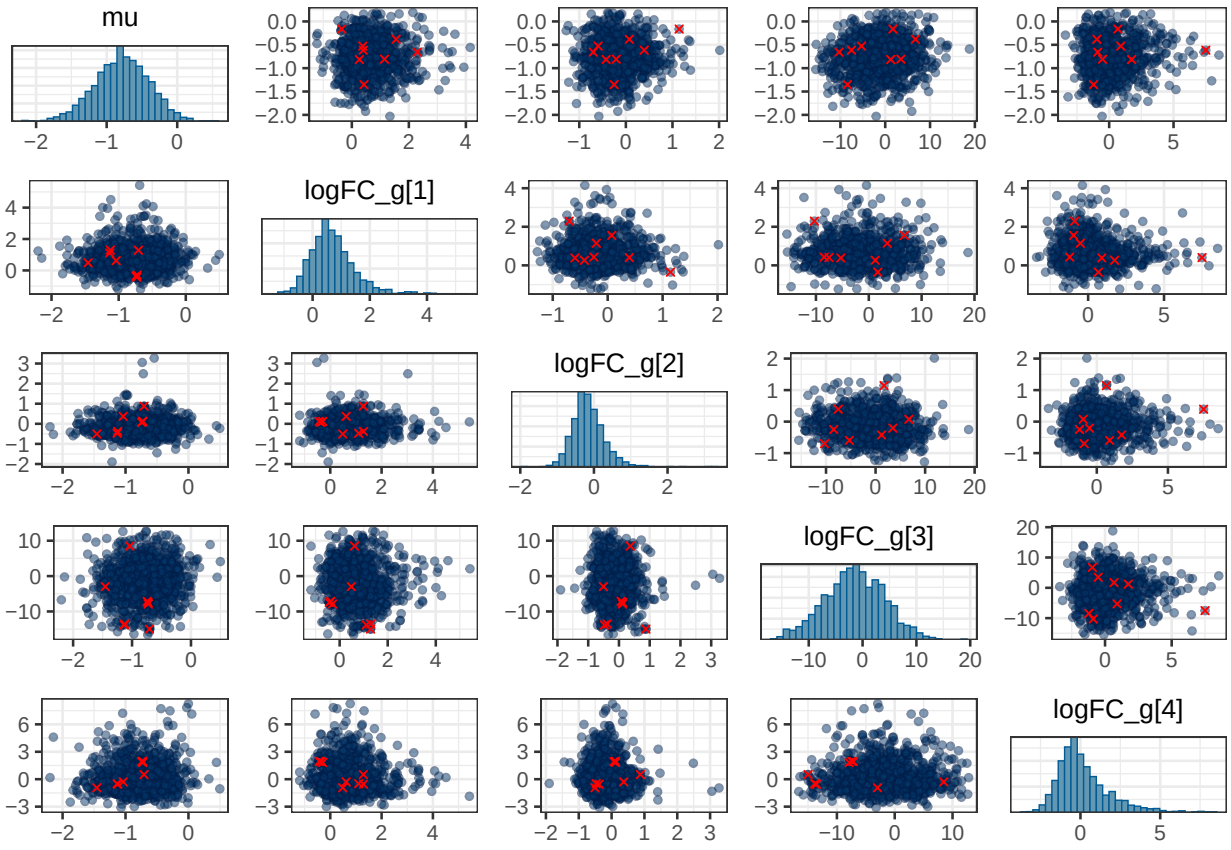
8) Run cmdstan and save fit results. Alternatively, data can be exported with stan_rdump and the sampling can be performed with cmdstan. Make sure cmdstan is installed: https://mc-stan.org/cmdstanr/articles/cmdstanr.html

9) Bayesian diagnostics: pair and trace plots, NUTS energy plots. There may be some divergent transitions (pair and trace plots) because fitting the hierarchical model to a small number of genes is suboptimal. The divergences are not due to misspecifications because they are distributed all over the parameter space. The small number of genes also results in an inefficient exploration of the parameter space (NUTS energy plot). For a full dataset, there should be no divergent transitions. This step can be accelerated when initializing the parameters adequately.



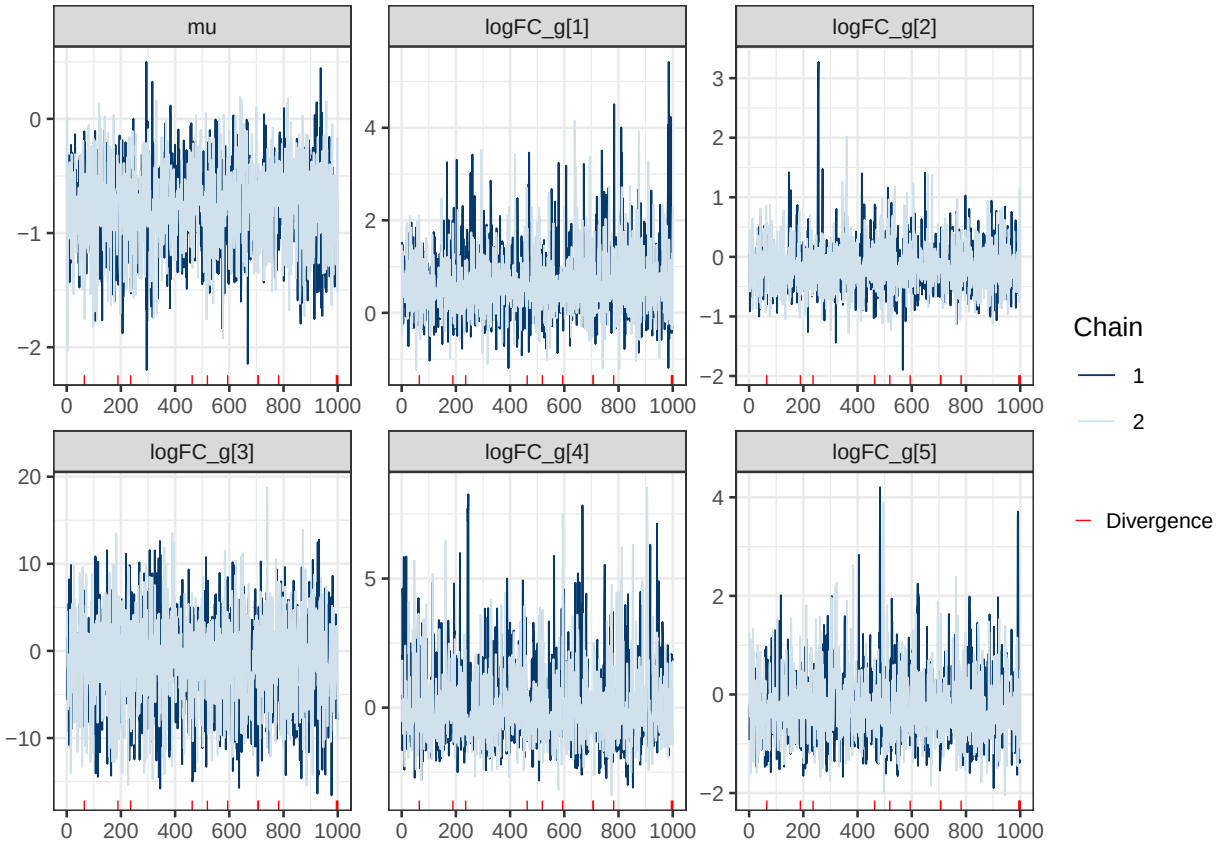

```
## pdf
##   2
```

```
## pdf
##   2
```



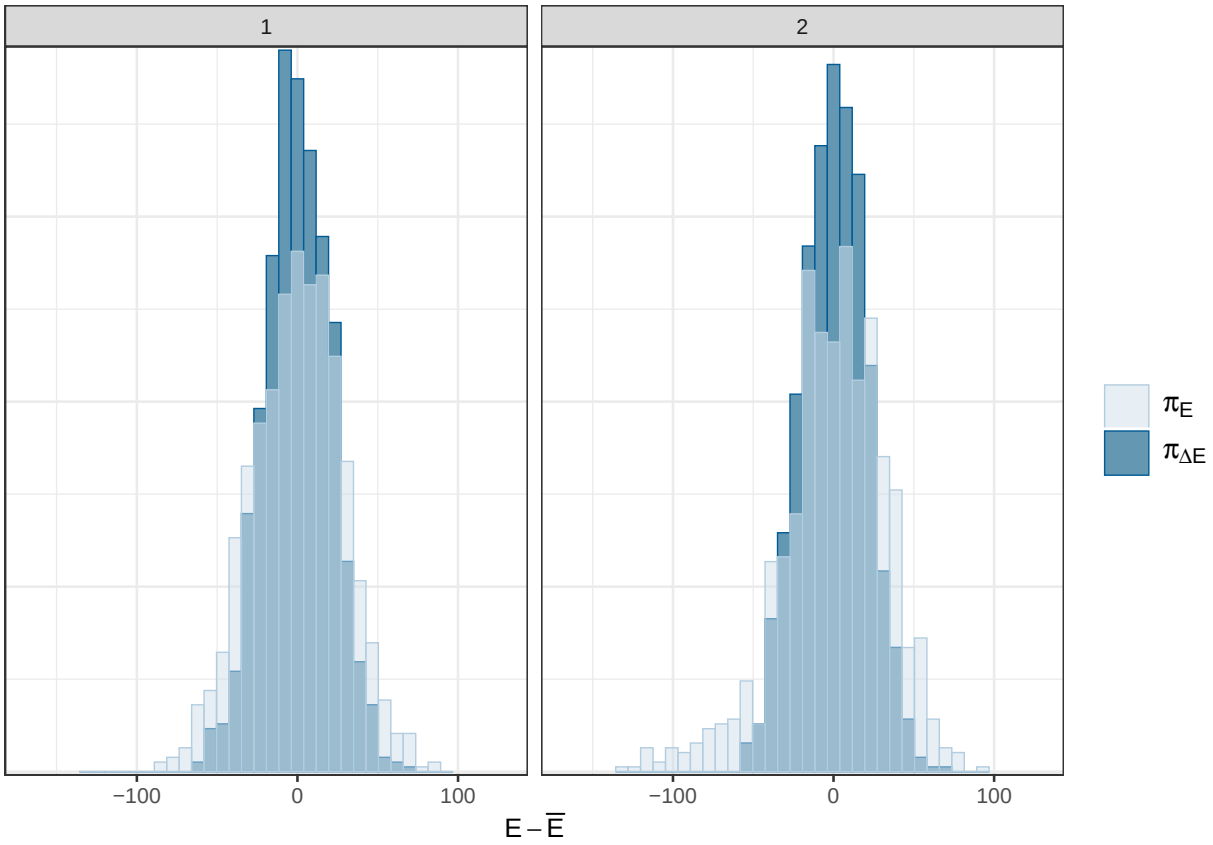

```
## pdf
##   2
```

```
## pdf
##   2
```


mu
logFC_g[1]
logFC_g[2]
logFC_g[3]
logFC_g[4]
logFC_g[5]
0 200 400 600 800 1000
Chain
1
2
Divergence

```
## pdf
##   2
```

```
## pdf
##   2
```

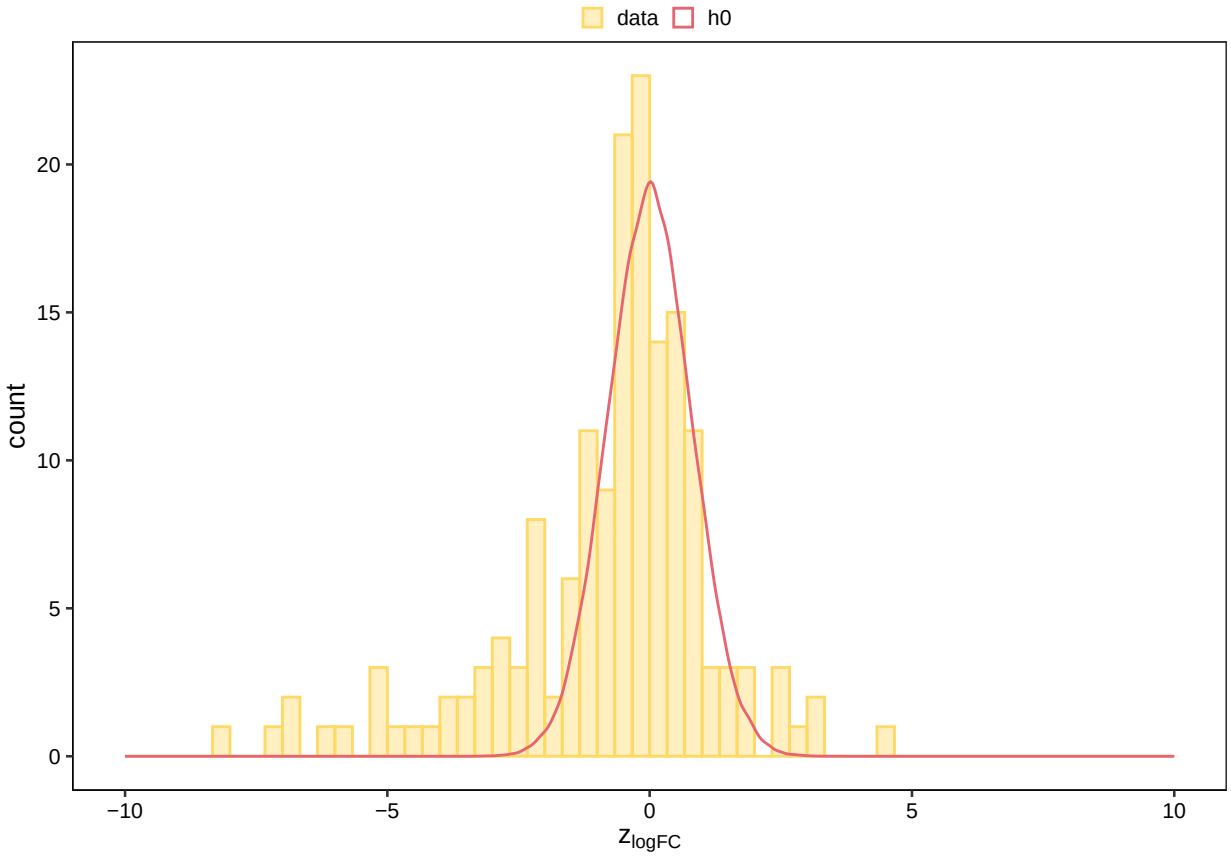



```
## pdf
##   2
```

```
## pdf
##   2
```

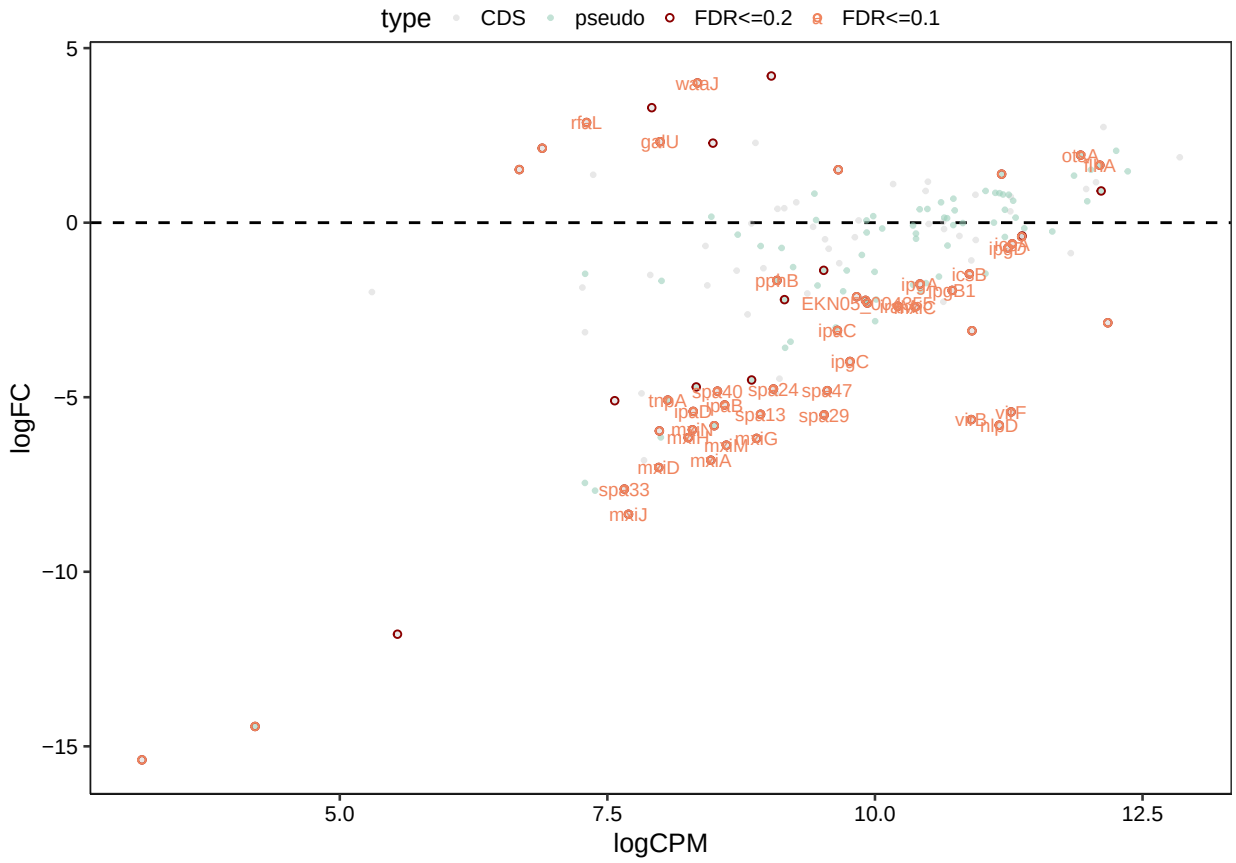
10) Extract fitness scores (logFCs) from fit and calculate mean input expression (logCPM). The natural logarithm is used here.

11) Compare the distribution of Z scores to the expected distribution under the null hypothesis (red line). Here, the fraction of log-fold changes in agreement with no fitness effect (null hypothesis) has been set to 0.7 because the demo dataset contains a very high fraction of genes with fitness effects. For a real dataset, this value should be around 0.9.




```
## pdf
##   2
```

12) The MA plot shows that the data is centered around zero, i.e. the normalization with the small demo dataset was successfull. The natural logarithm was converted to log2.


type
CDS
pseudo
FDR<=0.2
FDR<=0.1
logFC
logCPM
5
0
−5
−10
−15
5.0
7.5
10.0
12.5

```
## pdf
##   2
```