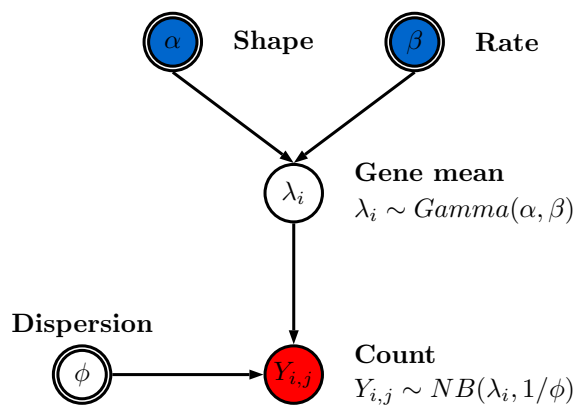# Additional figures

2017-05-02

Figure 1: Diagram of the Simple simulation model. Input parameters are indicated with double borders, blue shading shows those that can be estimated from real data. Red shading indicates the final output.
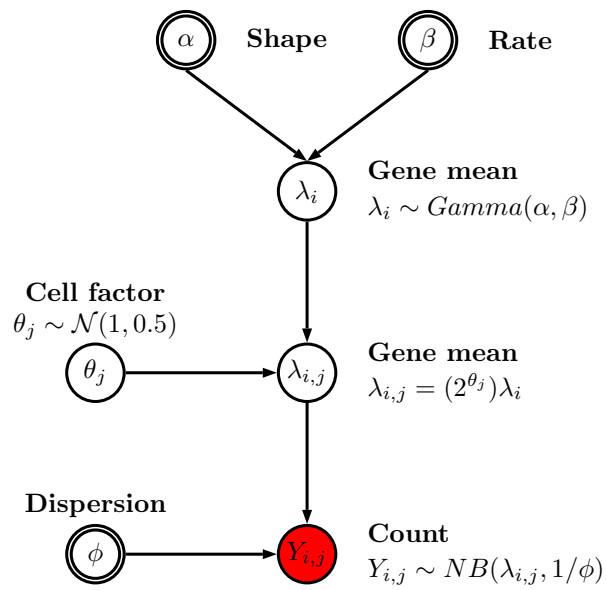
Figure 2: Diagram of the core Lun simulation model. Input parameters are indicated with double borders, blue shading shows those that can be estimated from real data. Red shading indicates the final output.

Gene mean

λ_i

Plate variance

σ_k^2

Plate factor
$\omega_k \sim \mathcal{N}(-m^p \sigma_k^2/2, \sqrt{m^p \sigma_k^2})$

ω_k

m^p

Plate modifier

$\lambda_{i,k}$

Plate mean
$\lambda_{i,j} = (e^{\omega_k})\lambda_i$

Library size

l_j

Lib size factor
$\kappa_j = m^l(l_j/mean(l_j))$

κ_j

m^l

Lib size modifier

$\lambda_{i,j}$

Cell mean
$\lambda_{i,j} = \kappa_j \lambda_{i,k}$

Gene dispersion

ϕ_i

$Y'_{i,j}$

True count
$Y'_{i,j} \sim NB(\lambda_{i,j}, 1/\phi_i)$

Zero indicator
$1^Z_{i,j} \sim Ber(z_i)$

Zero proportion

z_i

$1^Z_{i,j}$

$Y'_{i,j}$

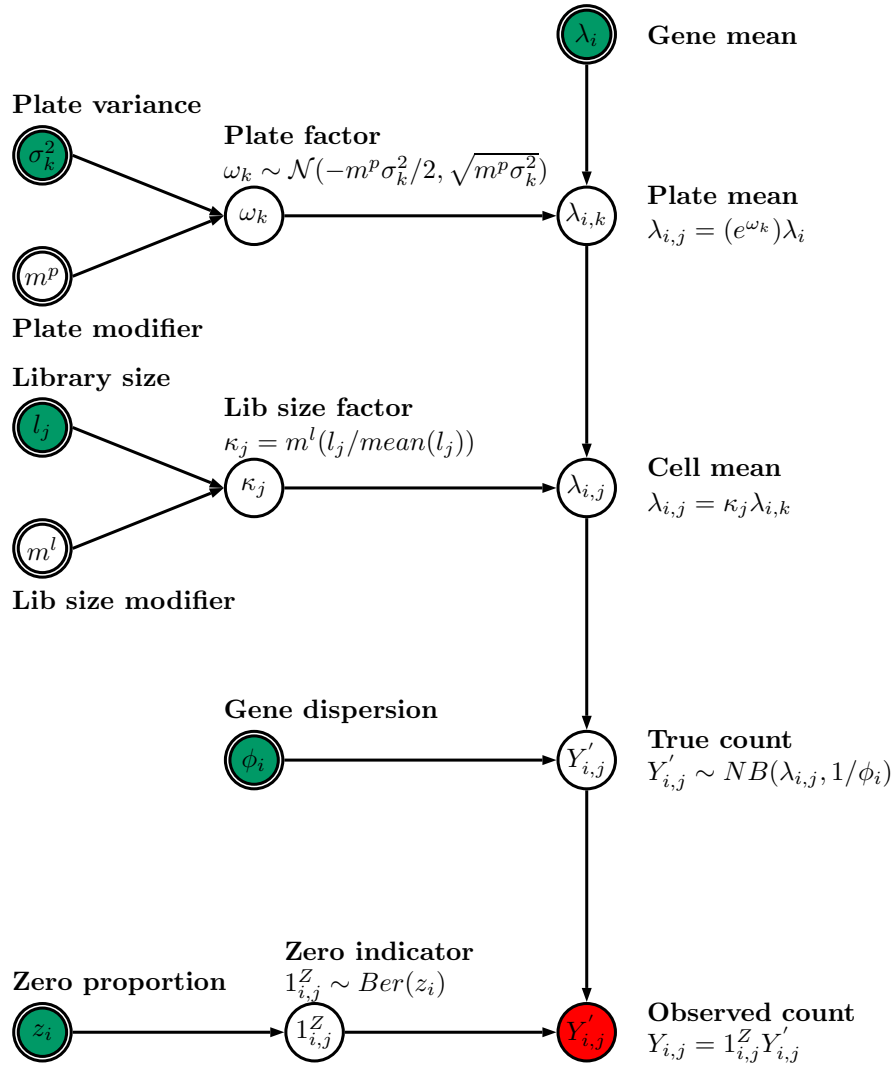**Observed count**
$Y_{i,j} = 1^Z_{i,j} Y'_{i,j}$

Figure 3: Diagram of the Lun 2 simulation model. Input parameters are indicated with double borders, green shading shows those that can be estimated and sampled from real data. Red shading indicates the final output.
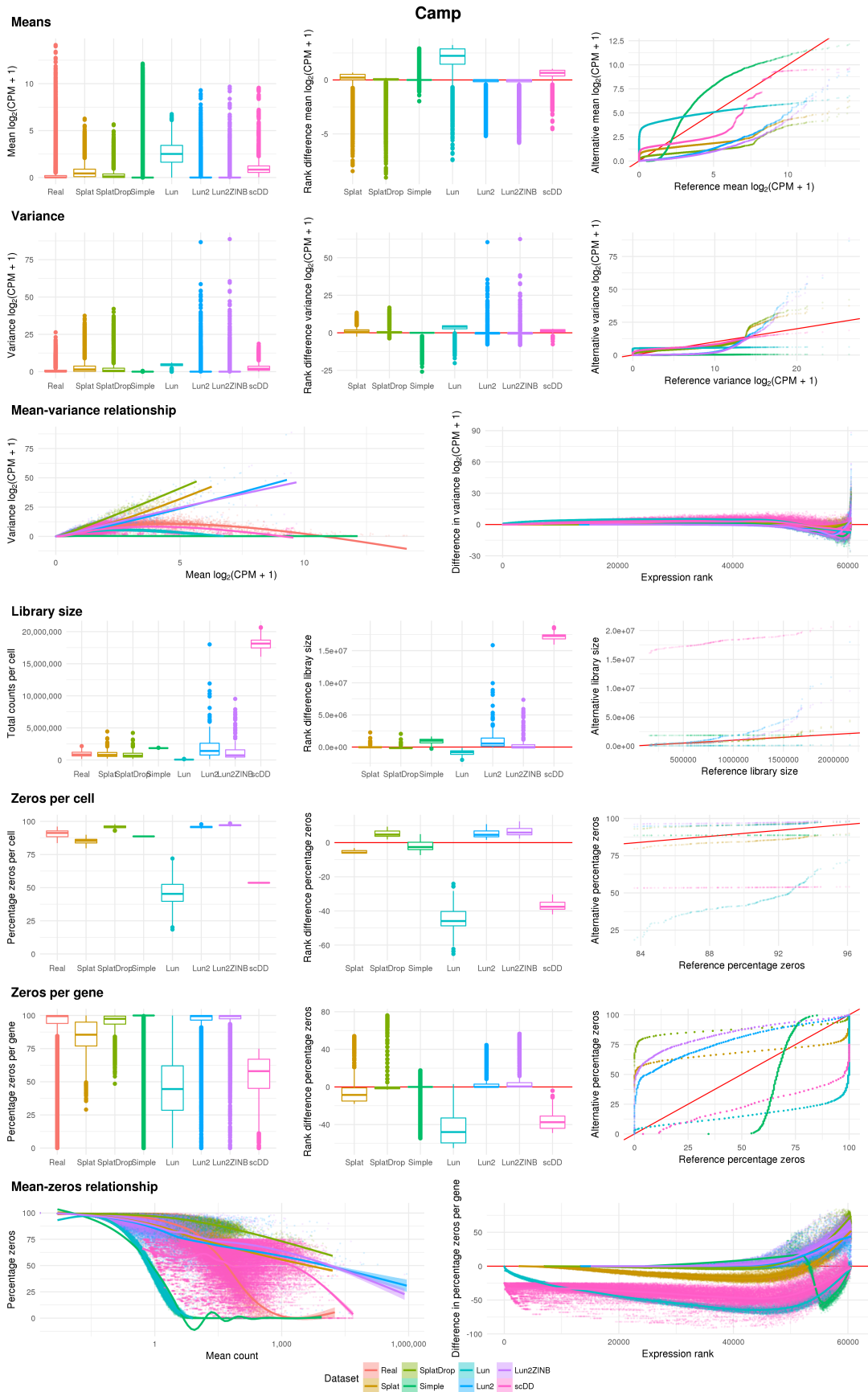
Figure 4: Output of Splatter's comparison functions for simulations based on the Camp dataset.
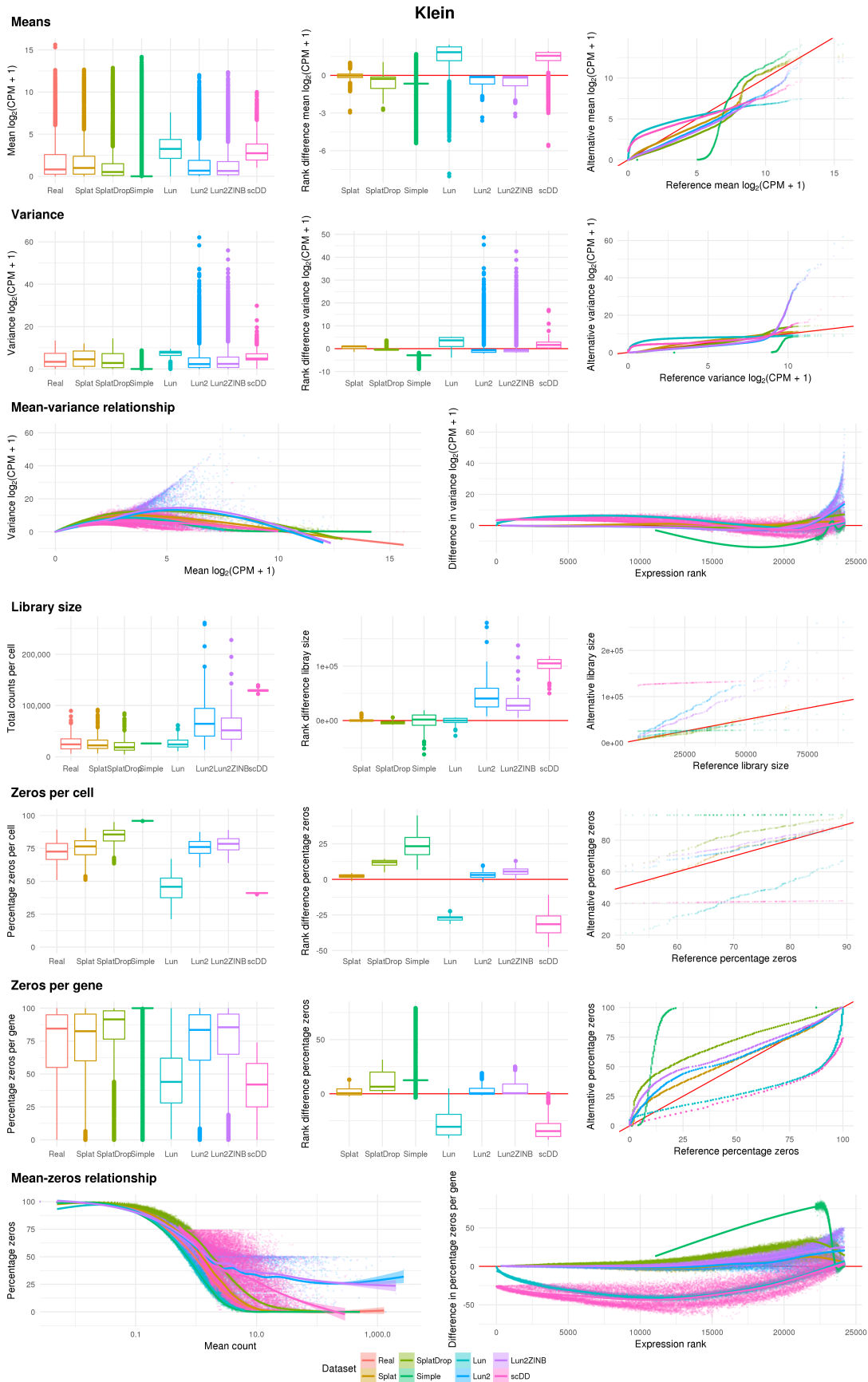
Figure 5: Output of Splatter's comparison functions for simulations based on the Klein dataset.
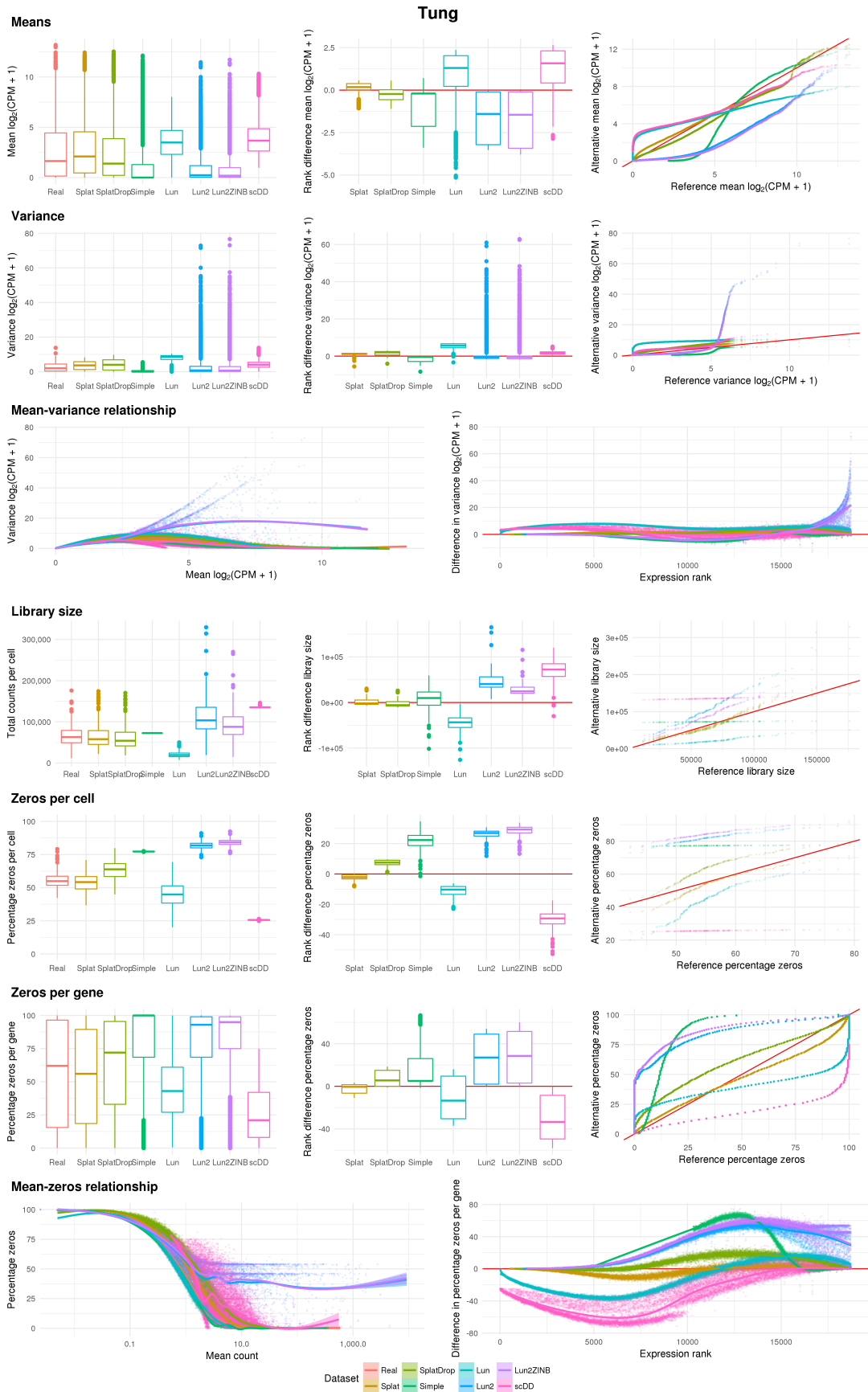
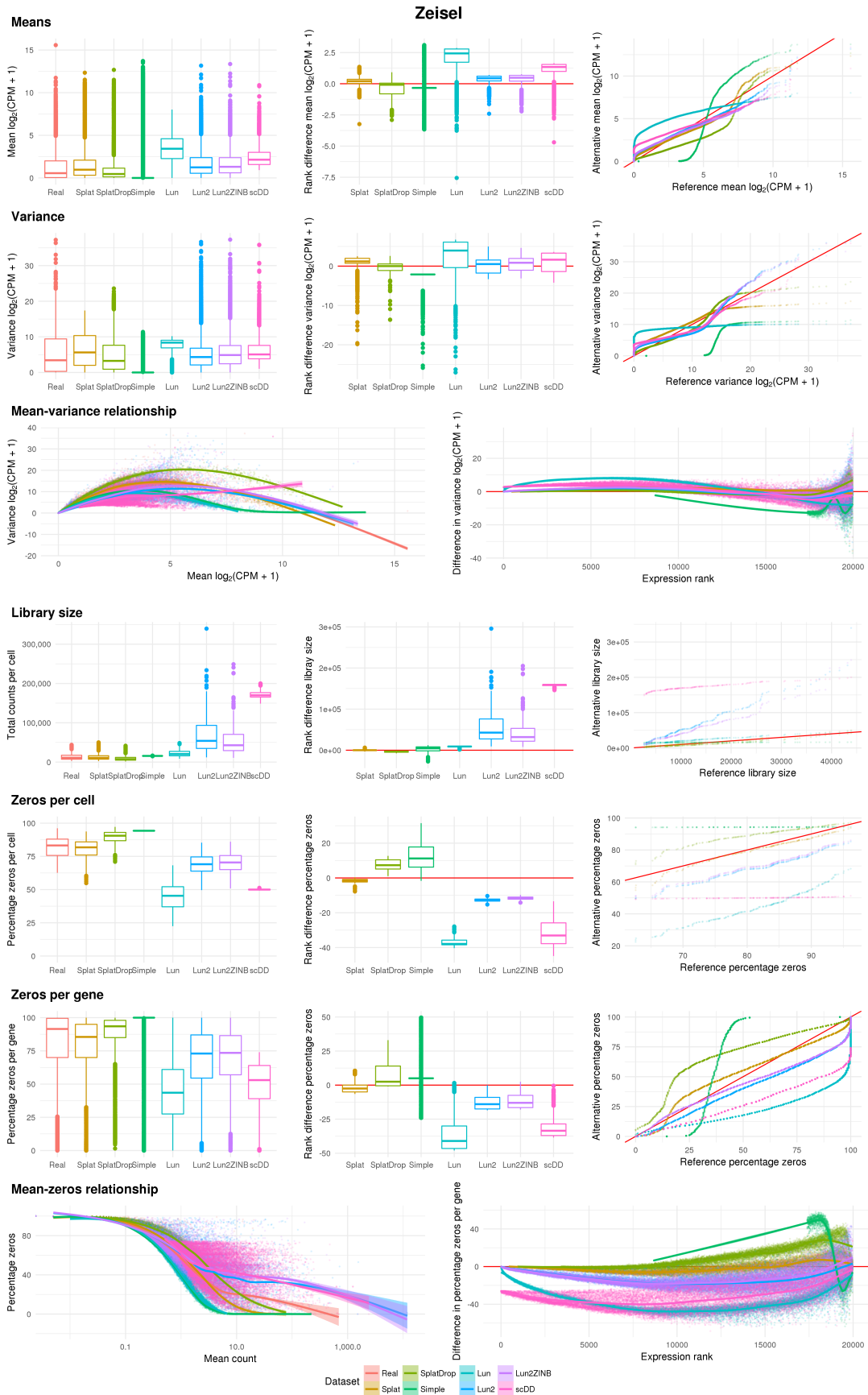Figure 6: Output of Splatter's comparison functions for simulations based on the Tung dataset.

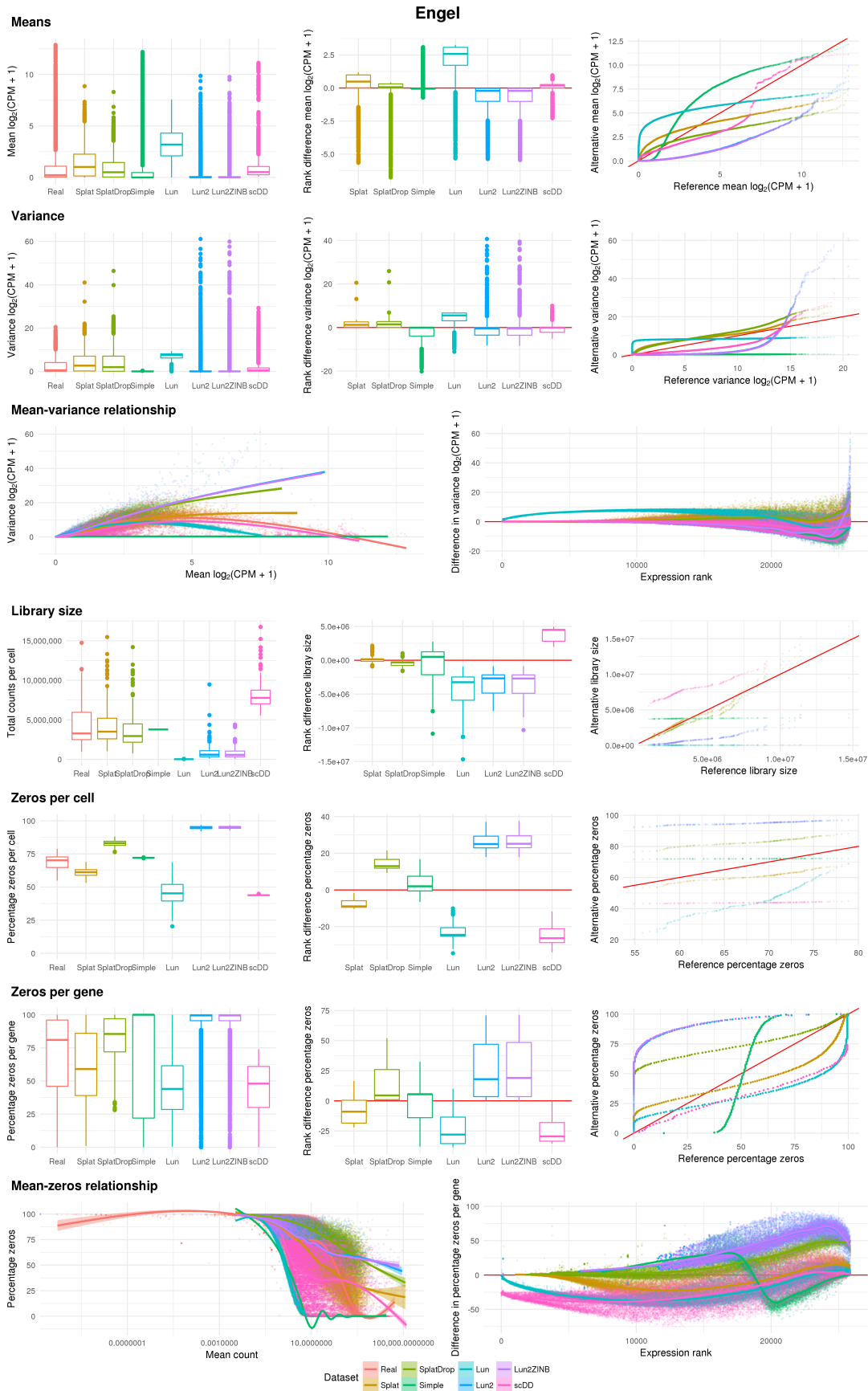Figure 7: Output of Splatter's comparison functions for simulations based on the Zeisel dataset.

Figure 8: Output of Splatter's comparison functions for simulations based on the Engel dataset.
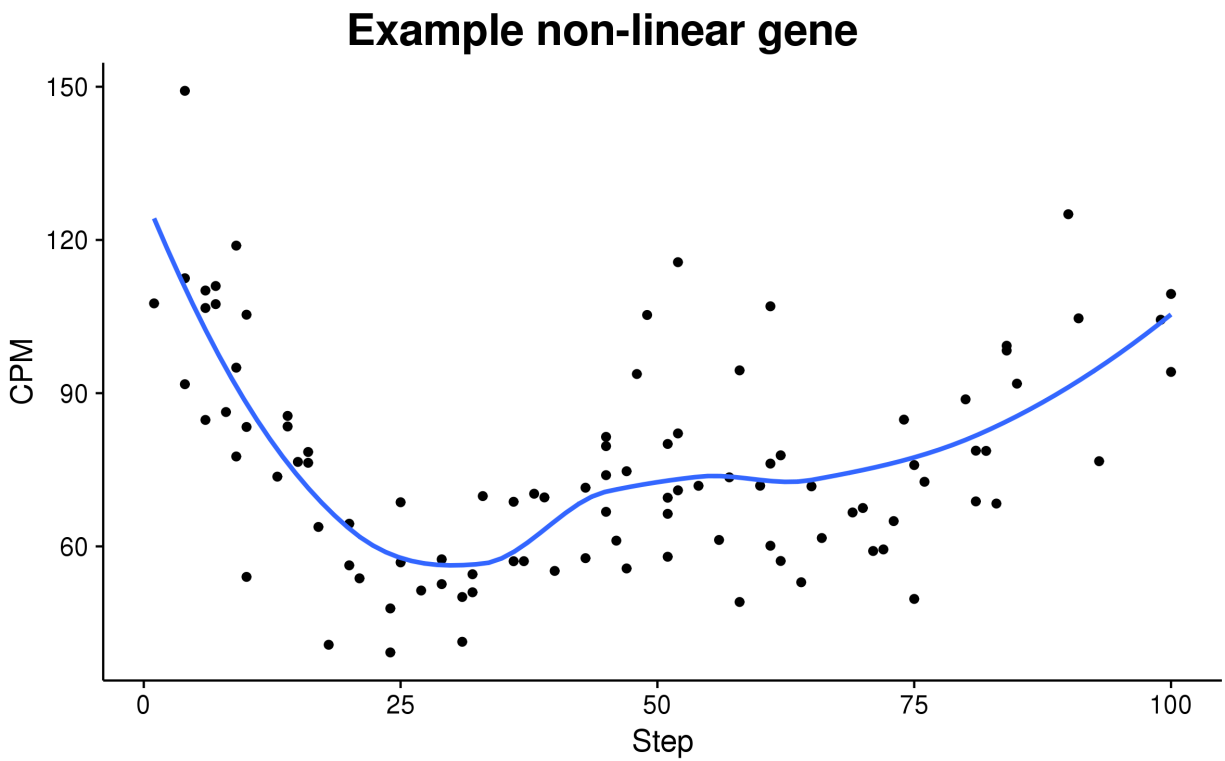
Figure 9: Example of a simulated gene that changes expression in a non-linear way over the course of a path. Each point is a cell with the x-axis showing progression along the path and the y-axis the simulated expression level. The blue line shows a LOESS fit.
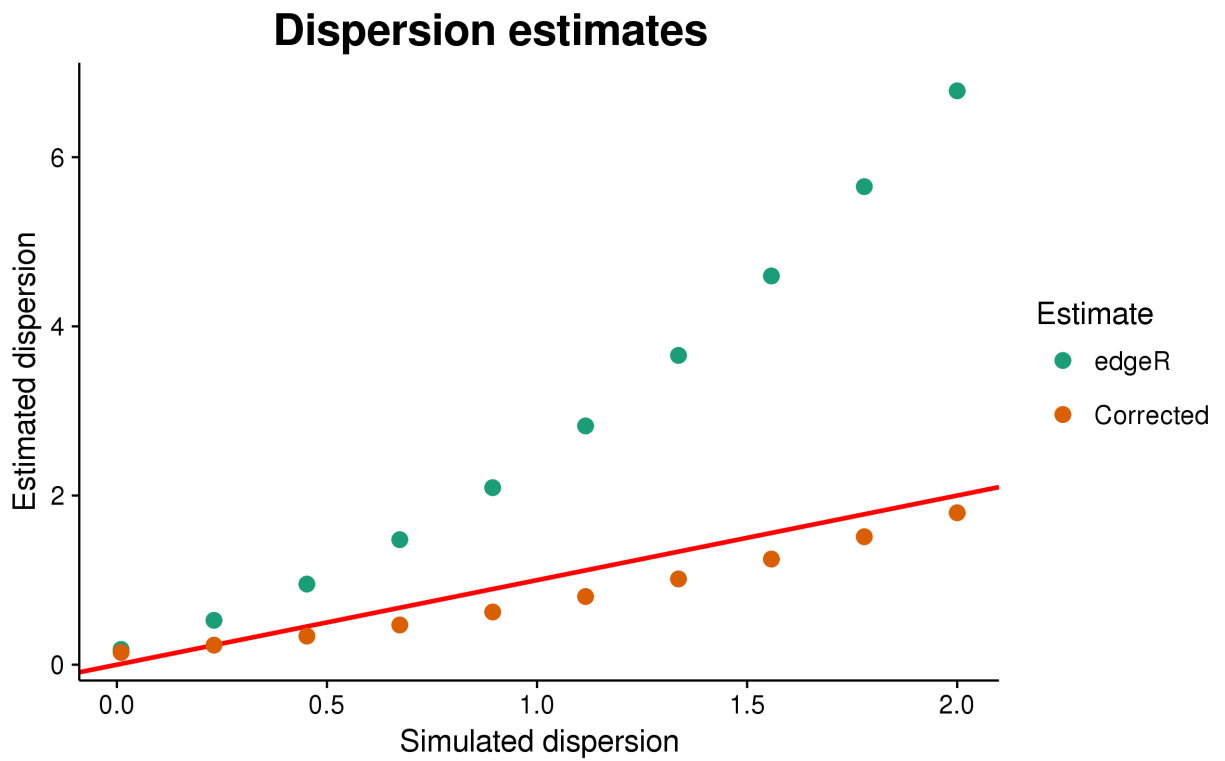
Dispersion estimates

Figure 10: Correction of edgeR dispersion estimates. Scatter plot of estimated dispersions against the true simulated values. Estimates of common dispersion obtained from edgeR (green) can be inflated for single-cell data. The Splat simulation uses a linearly corrected value (orange) in its estimation procedure. The red line shows the true values.
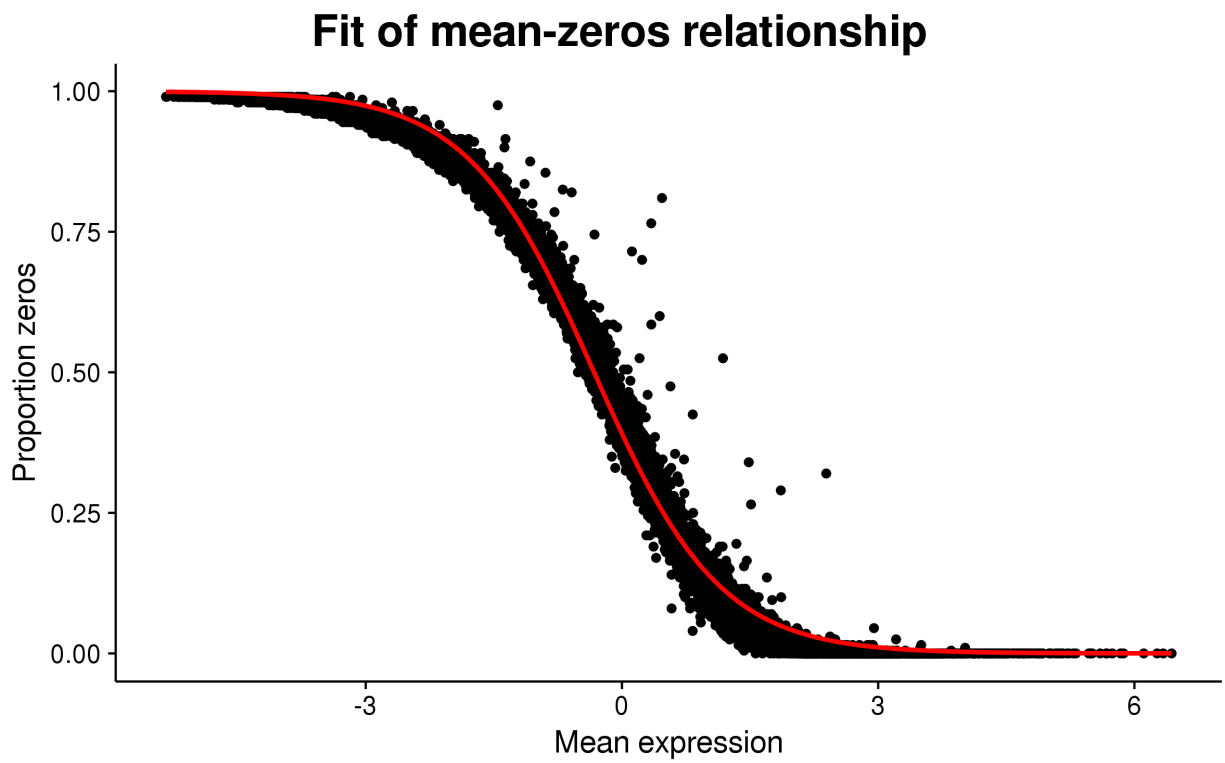
Figure 11: Fitting the mean-zeros relationship. Points show genes in the Tung dataset and the red line is a logistic function fitted to the data using Splat's estimation procedure.