Voom variance modelling

The above linear model is fitted, by ordinary least squares, to the log-cpm values y_{gi} for each gene. This yields regression coefficient estimates $\hat{\beta}^*_{gj}$, fitted values $\hat{\mu}_{gi} = x_i^T \hat{\beta}_g$ and residual standard deviations s_g.

Also computed is the average log-cpm $\bar{y}_g$ for each gene. The average log-cpm is converted to an average log-count value by

$$\tilde{r} = \bar{y}_g + \log_2(\tilde{R}) - \log_2(10^6)$$

where $\tilde{R}$ is the geometric mean of the library sizes plus one.

To obtain a smooth mean-variance trend, a loess curve is fitted to square-root standard deviations $s_g^{1/2}$ as a function of mean log-counts $\tilde{r}$ (Figure 2ab). Square-root standard deviations are used because they are roughly symmetrically distributed. The lowess curve [44] is statistically robust [45] and provides a trend line through the majority of the standard deviations. The lowess curve is used to define a piecewise linear function lo() by interpolating the curve between ordered values of $\tilde{r}$.

Next the fitted log-cpm values $\hat{\mu}_{gi}$ are converted to fitted counts by

$$\hat{\lambda}_{gi} = \hat{\mu}_{gi} + \log_2(R_i + 1) - \log_2(10^6).$$

The function value $\mathrm{lo}(\hat{\lambda}_{ga})$ is then the predicted square-root standard deviation of y_{gi}.

Finally, the voom precision weights are the inverse variances $w_{gi} = \mathrm{lo}(\hat{\lambda}_{gi})^{-4}$ (Figure 2c). The log-cpm values y_{gi} and associated weights w_{gj} are then input into the standard limma linear modeling and empirical Bayes differential expression analysis pipeline.