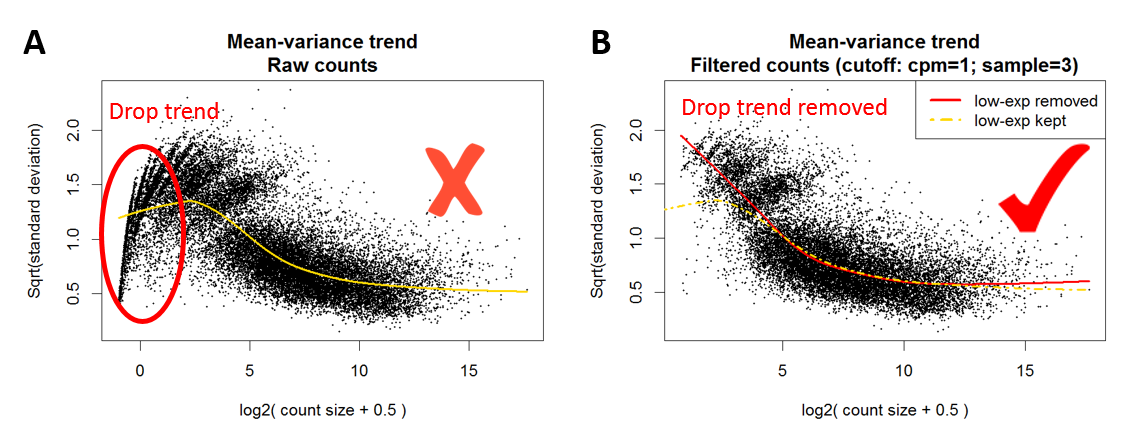
The low expressed transcripts are filtered based on count per million (CPM) reads:

where is read count of transcript and is the library size of the sample.

* An expressed transcript must have samples with expression CPM.
* An expressed gene must have at least one expressed transcript.

The sample number cut-off and CPM cut-off are determined by the mean-variance trend plot. Read counts are assumed to be negative binomial distributed of which the mean and variance has a relation:

where is read count, is the mean and is the overdispersion. The expression variance decreasing monotonically with the increasing of the mean. In real RNA-seq data, the low expressed transcripts confound with noise and the expression distribution is different to the expressed transcripts. In the mean-variance trend plot, the low expressed transcripts cause a drop trend in low expression region. Therefore, the cut-offs and to filter the low expression can be optimised until the drop trend in low expression region disappeared in the mean-variance trend plot.



**Figure**: Mean-variance trend plot before (A) and after (B) low expression filters. Each black point in the figure represents a gene/transcript. The red and yellow lines are the fitted trends of the genes/transcripts.