Supplementary File 1 – An alternative approach of RPE expression estimation

An alternative approach of RPE expression estimation is based on the weighted-average assumption (equation S1). Whereas the 1-fold cut-off threshold becomes approximately 3-fold in the alternative approach, the resulting number of significantly expressed genes in RPE at 52hpf, as well as the ranking of the fold changes between the RPE and retina remain the same. The detailed derivation based on the weighted-average assumption is shown below.

1. Derivation of RPE expression based on the weighted-average assumption

\[
Expr_{WRR52} = \frac{Yield_{WR52} \times Expr_{WR52_i} + (Yield_{WRR52} - Yield_{WR52}) \times Expr_{RPE_i}}{Yield_{WRR52}}, i = 1, \cdots, n
\]  

(S1)

This leads to the estimated RPE expression as

\[
\hat{\theta}_{oi}^* = \frac{Yield_{WRR52} \times Expr_{WR52_i} - Yield_{WR52} \times Expr_{WR52_i}}{Yield_{WRR52} - Yield_{WR52}}, i = 1, \cdots, n
\]  

(S2)

Comparing equation (S2) with equation (2) in the manuscript, it can be shown that

\[
\hat{\theta}_{oi}^* = c \times \hat{\theta}_{oi}, i = 1, \cdots, n
\]  

(S3)

where
\[ c = \frac{\text{YieldWR52}}{\text{YieldWR52} - \text{YieldWR52}} \approx 3.1038 \]  

(S4)

i.e., the estimated RPE expression based on (S1) equals to the estimated RPE expression (2) multiplied by a constant c.

For the same reason, the corresponding standard deviations of RPE expression based on (S1) are also the original standard deviations multiplied by c. Therefore, the confidence interval of RPE based on (S1) is (c*lower bound, c*upper bound) if the original confidence interval of RPE based on (3) is (lower bound, upper bound). Thus, a gene will still be inferred as significantly expressed in RPE52 under the assumption (S1) when this gene is inferred as significantly expressed in RPE52 using the lower 90% confidence limit of the estimate greater than 0 as the criterion in the original formulation (3), and vice versa. In other words, the same set of significant genes expressed in RPE52 will be obtained, regardless of which one of the two approaches is used. The above property is called invariance in statistical estimation theory.

2. Fold change calculation and cut-off threshold selection

The fold change between expression of RPE52 and that of WR52 based on (S1) is also the fold change of equation (4) multiplied by a constant,

\[ \frac{\theta_{2i}^*}{\theta_{2i}} = c \times \frac{\theta_{2i}}{\theta_{2i}} = c \times \left( \frac{\theta_{1i}}{\theta_{2i}} - 1 \right) = c \times r_i - c \]  

(S5)
Moreover, $r_i$ can still be estimated through the pivotal quantity $Q_i$. Therefore, the ranking or the order of the genes that is sorted according to the fold change is the same regardless of which one of the two approaches is used.

Now the only difference between these two approaches is the threshold selection. Since a lower 90% confidence limit of $\frac{\theta_{0i}}{\theta_{2i}} > 1$ is used to infer significant over-expression in the original approach, this gene will be inferred to be significantly over-expressed if lower 90% confidence limit of $\frac{\theta^*_{0i}}{\theta_{2i}} > c$. Since $c > 1$, this suggests that the original criterion is more conservative compared to the calculation based on (S1), i.e, fewer genes will be inferred as over-expressed by the original approach if the same cut-off is used in both approaches. While a higher 90% confidence limit of $\frac{\theta_{0i}}{\theta_{2i}} < 1$ is used to infer significant under-expression of the genes in the original approach, this gene will be inferred to be significantly under-expressed if upper 90% confidence limit of $\frac{\theta^*_{0i}}{\theta_{2i}} < c$. In this case, more genes will be inferred as under-expressed if the same cut-off is used in both approaches.

Actually, the threshold of original method approximately equals to a three-fold confidence limit cutoff between RPE52 and WR52 based on the alternative approach. The three-fold cut-off threshold has been commonly used in many biological studies\textsuperscript{2-9}. If
one uses one-fold confidence limit cutoff in the alternative approach (equals to \( 1/c \approx 0.3221 \) in the original approach), 1878 genes will be inferred as significantly over-expressed, and 1700 genes with a fold change > 0 will be inferred as significantly under-expressed in RPE52 compared with WR52.

In short, the original approach is a conservative implementation of the alternative approach and the respective cut-off thresholds are still set at widely used levels. Furthermore, the number of genes significantly expressed in RPE52 and the ranking of the genes based on the fold change comparison remain the same. The complete gene list with their RPE52 estimations, fold changes and respective 90% confidence limits based on the original approach is shown in Supplementary Table S4. In theory, one can use a less stringent threshold to obtain more genes for further experimental characterization.

References