In order to extract the effect of interest and filter noise resulting from differences in dye labeling efficiency and variation between/within each array, we apply the two-stage strategy: (1) we use the standard normalization procedure to normalize log ratios of the raw data, and (2) we employe an ANOVA model for the normalized log ratios to estimate the significant levels of the effect of interest.

For the first stage, the composite normalization strategies between/within each array in Yang et al. (2002) can be summarized by the following statistics:

$$\log_2 Y_{gr} = \left[\left[\frac{\log_2 Ratio_{igr} - \log_2 \alpha_i(A)}{\beta_i}\right]\right]$$

The term $\log_2 Y_{gr}$ is the normalized log ratio, which is normalized by the location normalization within array and scale normalization between arrays, $Ratio_{igr} = R_{igr} / G_{igr}$, where R_{igr} and G_{igr} represent the raw data of red and green fluorescent intensity in array i for gene g and replicate r, $A = \log_2 \sqrt{R_{igr} G_{igr}}$, and β_i is the scale normalization, the normalization between arrays, which can be estimated by the median absolute deviation *MAD* (See Yang et al. 2002; Yang and Speed 2002). The $\log_2(\alpha_i(A))$ term means intensity-dependent normalization, the normalization within array, which can be estimated by the LOWESS fit method.

For the second stage, we first write down the model and then explain the implication. The ANOVA model for the normalized log ratios is, according to Eq. (1) of Landgrebe et al. (2006):

$$\log_2 Y_{gr} = \gamma_{Rg} - \gamma_{Gg} + \tau_{kg} - \tau_{k'g} + \varepsilon_{gr},$$

where γ_{Rg} and γ_{Gg} represent the dye-gene interaction term for the dye effect on Cy5 and

Cy3 dyes, respectively, $\tau_{kg} - \tau_{k'g}$ indicates the term of interest that reflects gene-specific differences in expression for two varieties (k, k') (like two varieties of BY and RM strains in this paper), and ε_{gr} denotes the noise. For modeling aspects and applications, see Landgrebe et al. (2006).

The above model can be represented in the matrix form and the corresponding test statistics can be derived for detecting the significant difference in gene expression between BY and RM:

$$\underset{\sim}{Z}_g = X \underset{\sim}{\beta}_g + \underset{\sim}{\varepsilon}_g ,$$

$$t_g = \left(\tau_{1g} - \tau_{2g}\right) \Big/ \sqrt{s_g^2 l^T \left(X' X\right)^{-1} l} ,$$

where $Z_{gr} = \log_2 Y_{gr}$, $\underset{\sim}{\beta}_g^T = \left[\gamma_{Rg} \;\; \gamma_{Gg} \;\; \tau_{1g} \;\; \tau_{2g}\right]$, X is the corresponding design matrix of the model; $l^T = (0\ 0\ 1\ \text{-}1)$, and s_g^2 is the sample variance of g. Similar to Landgrebe et al. (2006), we assume that the error is normally distributed with mean 0 and a common variance, so that the test statistics follows a t-distribution with the degree of freedom of $v_g = n_g$-rank(X), where n_g is the number of observations on gene g and rank(X) is the rank of the design matrix. The corresponding significant level, the p-value, can be calculated from the above test statistic for each of the genes under study.