Let $X_{i1}, \ldots X_{in} \sim N(\mu_i, \sigma^2)$ and $Y_{i1}, \ldots Y_{in} \sim N(\theta_i, \sigma^2)$ independently, for $i = 1, \ldots m$. Think of $X_{i1}, \ldots X_{in}$ as the expression levels of gene i for a set of n healthy patients, and $Y_{i1}, \ldots Y_{in}$ the expression levels of gene i for a set of n sick patients. We will test the m hypotheses

$$H_i : \ \mu_i = \theta_i, \quad i = 1, \ldots m.$$

Specifically for each gene i, we will compute a t statistic

$$T_i = \frac{\bar{X}_i - \bar{Y}_i}{\sqrt{\frac{1}{n}(S^2_{X_i} + S^2_{Y_i})}},$$

where $\bar{X}_i = \frac{1}{n}\sum_{j=1}^n X_{ij}$, and $S^2_{X_i} = \frac{1}{n-1}\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2$, and similarly for $Y_{i1}, \ldots Y_{in}$.

We claim that screening out genes from this set in pre-processing, based on total variances, is OK—in that, if we were to compute say FDR estimates using Benjamini-Hochberg on the reduced set, then our estimates would still be valid. For this to be true, we would need to make sure that the distribution of the null t statistics (null p-values) is unchanged after screening. Note that what happens to the distribution of the non-null t statistics (non-null p values) is important for power purposes, but does not concern us when it comes to FDR control.

Hence, it suffices to check that for the null genes i, in which $\mu_i = \theta_i$, the total variance

$$V_i = \frac{1}{2n}\left(\sum_{j=1}^n (X_{ij} - M_i)^2 + \sum_{j=1}^n (Y_{ij} - M_i)^2\right),$$

where M_i denotes the mean of the joint sample $X_{i1}, \ldots X_{in}, Y_{i1}, \ldots Y_{in}$, is independent of T_i, the t statistic for the ith gene. This can be seen in more than one way. From a classical stats perspective: under the null in which the means are the same and therefore $X_{i1}, \ldots X_{in}, Y_{i1}, \ldots Y_{in}$ are drawn independently from the same normal model, M_i is a complete, sufficient statistic for σ^2. But T_i is ancillary to σ^2; this can be seen by dividing numerator and denominator by σ, and noticing that it is then entirely expressed in terms of normal variates within unit variance. Hence by Basu's theorem, the two are independent.