

Figure 1: Comparison of family-wise error rate (FWER) obtained by 1000 resampling runs and estimated by the heuristic method presented here. The test groups of size $T = 20$, $T = 100$ and $T = 250$ were chosen randomly from all GO-annotated probe-sets on Affymetrix’s HG-133A (solid lines, 15652 annotated probe-sets). The dotted line indicates the diagonal

Family-wise error rate (FWER)

If there is no a priori expectation of any term to be enriched in the test group, controlling the family-wise error rate (FWER) is regarded as the appropriate multiple testing correction [1]. The FWER provides the probability to get *any* false discovery:

$$FWER(\alpha) = Pr(NFD(\alpha) > 0) . \quad (1)$$

Here $NFD(\alpha)$ denotes the number of false discoveries given a threshold of α for the single-test p -values. If all tests were independent, we could calculate the family-wise error rate for the list of terms from term $t = 1$ to term $t = i$

iteratively by

$$\begin{aligned} FWER_i(\alpha) &= FWER_{i-1}(\alpha) \\ &+ Pr(p_i \leq \alpha)(1 - FWER_{i-1}(\alpha)) , \end{aligned} \tag{2}$$

starting the iteration with $FWER_0(\alpha) = 0$, and $Pr(p_i \leq \alpha)$ being defined as above. The iteration ends at the end of the list of annotated terms and yields the family-wise error for the entire list. However, there are strong correlations between the annotation of terms due to the graph-structure and also correlations between annotations of single genes (e.g. certain functions correspond to certain cellular locations). Therefore the joint probability of term t having a p-value below the threshold *and* no p-value of any other term in the list before passing the threshold is overestimated by simple multiplication of the probabilities $Pr(p_t \leq \alpha)(1 - FWER_{t-1}(\alpha))$. We find empirically, that the joint probabilities can be very well estimated by $\beta Pr(p_i \leq \alpha)(1 - FWER_{i-1}(\alpha))$ with the parameter $\beta = \frac{1}{4}$. This estimate leads to the following heuristic adaptation of Equation 2:

$$\begin{aligned} FWER_i(\alpha) &= FWER_{i-1}(\alpha) \\ &+ \beta Pr(p_i \leq \alpha)(1 - FWER_{i-1}(\alpha)) . \end{aligned} \tag{3}$$

We tested this heuristic formula with a variety of data sets including different Affymetrix chips, gene groups of small customized microarrays and all annotated UniGene genes and found a good agreement of the estimated $FWER(\alpha)$ by this method with the $FWER(\alpha)$ obtained by resampling. Figure S1 shows examples of this study. Since the $FWER(\alpha)$ is already strictly monotonic with increasing α we can utilize the $FWER(\alpha)$ as an adjusted p-value p_{FWER} controlling the probability of having any false discoveries in the list. For instance, if one chooses a threshold of $p_{FWER} \leq 0.05$, one obtains a list where the probability of having one or more falsely discovered terms is 5%.

Comparison of our method to estimate FWER with other methods

Figure S2 shows that our approach approximates the FWER obtained by resampling simulations very well, especially in the important region of adjusted p-values below 0.05. The other methods (Holm, Sidak, Bonferoni) are

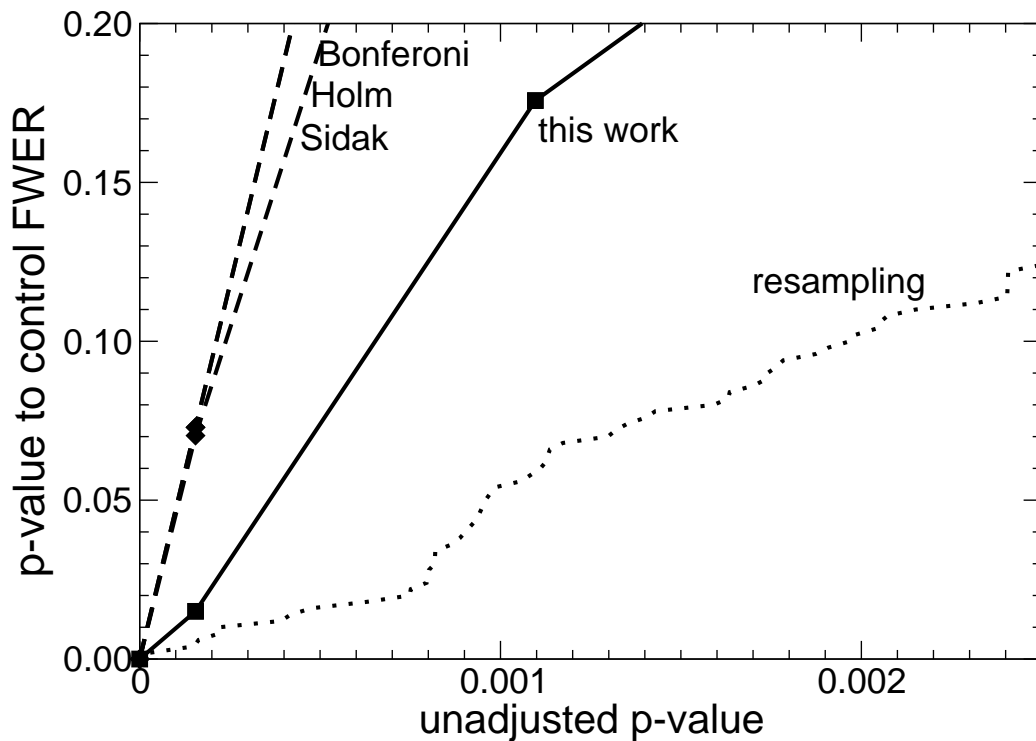


Figure 2: Adjusted p-value to control the FWER for the group of down-regulated genes in the cerebellum in the Huntington’s disease mouse model. The FWER estimated by our approach (solid) is in good agreement with the results from resampling simulations (dotted), especially in the important region of low p-values. Squares indicate the values of the adjusted p-values for the GO-terms. The adjusted p-values by Bonferoni, Holm and Sidak (dashed) are similar to each other and about 2.5-6 fold too conservative when compared to resampling simulations.

too conservative resulting in adjusted p-values that are around 2.5-6 times larger, and they thus miss significant results.

References

- [1] Dudoit S., Shaffer J.P., and Boldrick J.C., 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**: 71–103.