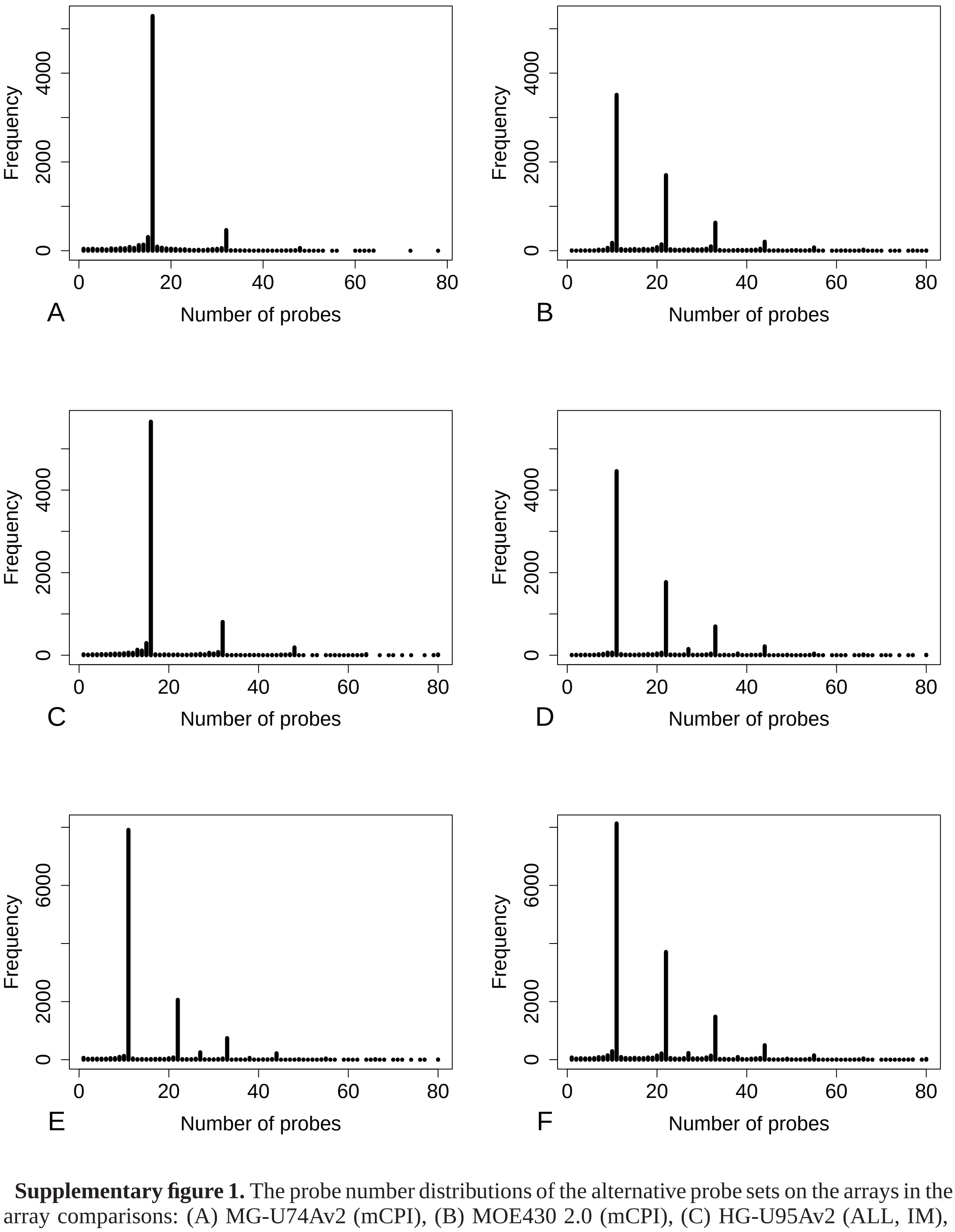
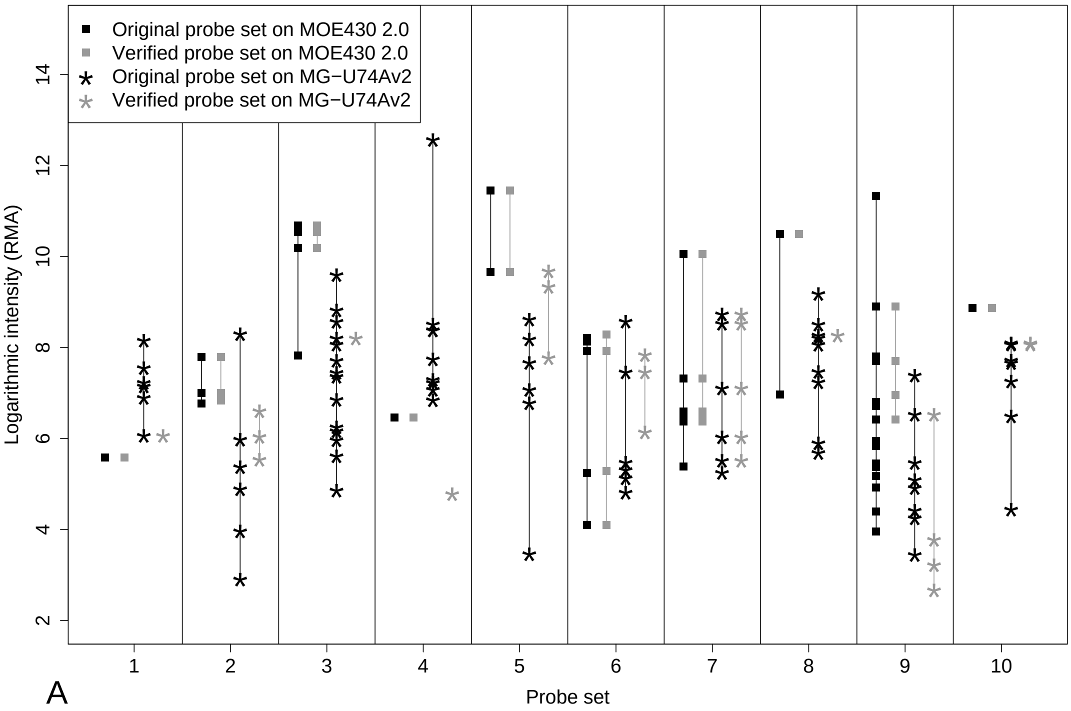
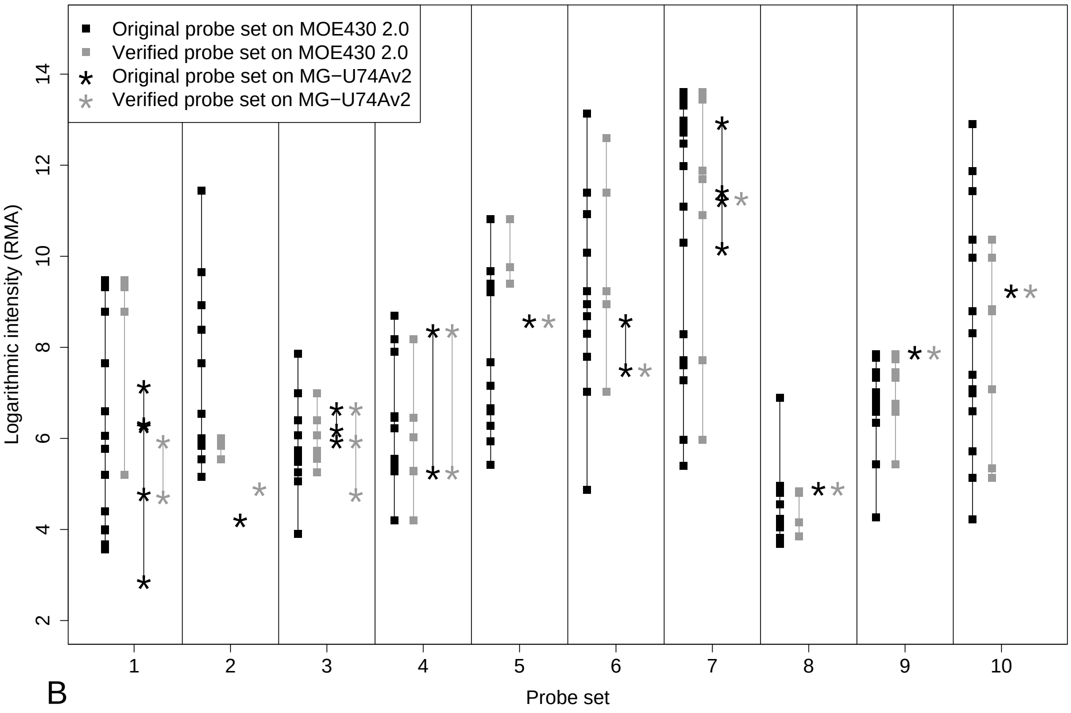
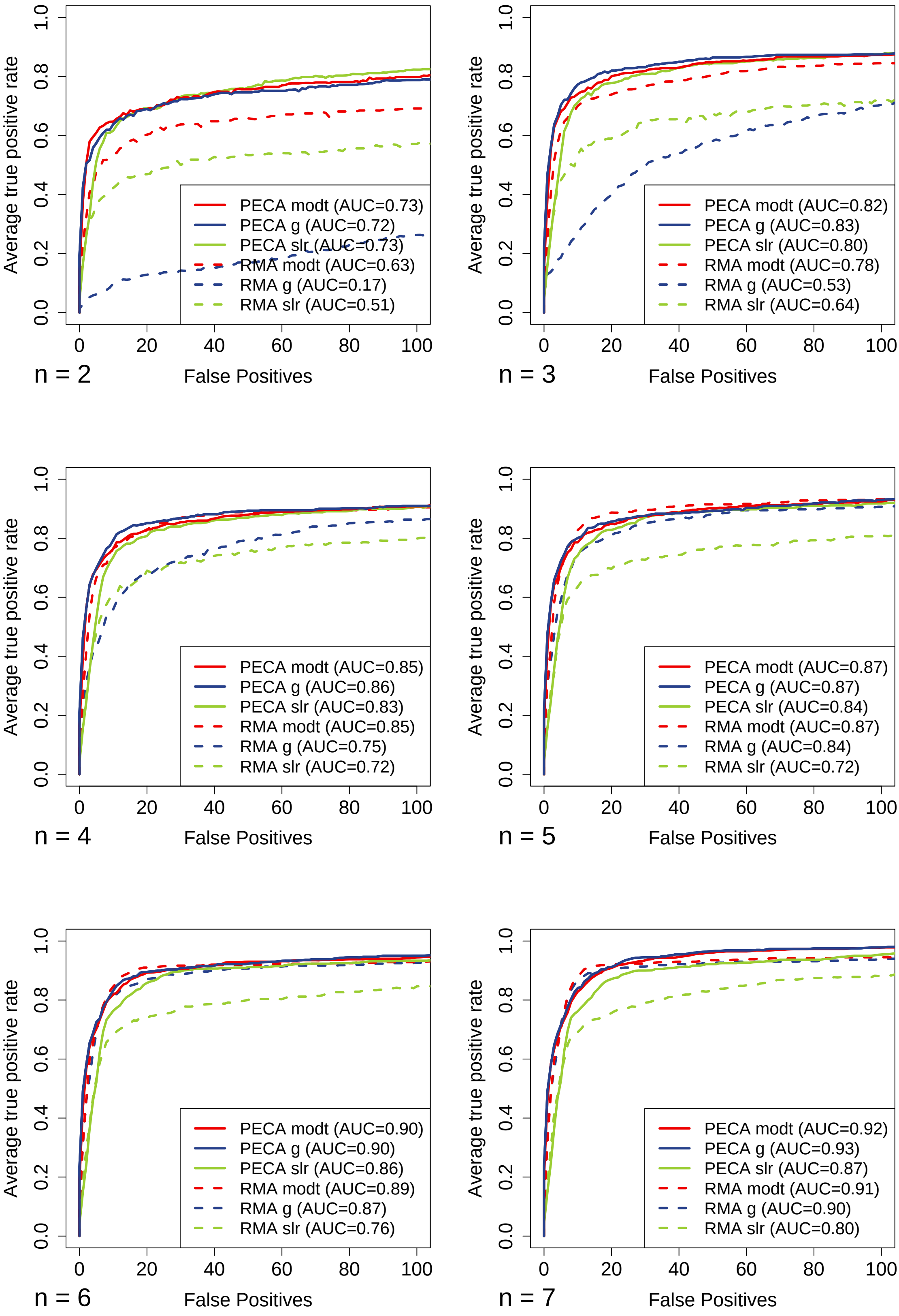


Supplementary figure 1. The probe number distributions of the alternative probe sets on the arrays in the different array comparisons: (A) MG-U74Av2 (mCPI), (B) MOE430 2.0 (mCPI), (C) HG-U95Av2 (ALL, IM), (D) HG-U133A (ALL, IM), (E) HG-U133A (hESC), and (F) HG-U133Plus2.0 (hESC). The manufacturer-defined probe sets on the MOE430 2.0, HG-U133A and HG-U133Plus2.0 arrays contained 11 probes, whereas the MG-U74Av2 and HG-U95Av2 arrays were originally designed to have probe sets of size 16. An alternative probe set contains all the verified probes on an array that match to a unique GeneID. A typical size of an alternative probe set was the size of the original Affymetrix probe set or its multiplier. The proportion of alternative sets with less than 5 probes varied between 0.4 % (MOE430 2.0) and 2.1 % (MG-U74Av2). Thus, small probe sets were not overrepresented on any array. The total numbers of alternative sets included in the array comparisons were 7735 in the mCPI array comparison, 8240 in the ALL and IM array comparisons, and 12661 in the hESC array comparison.

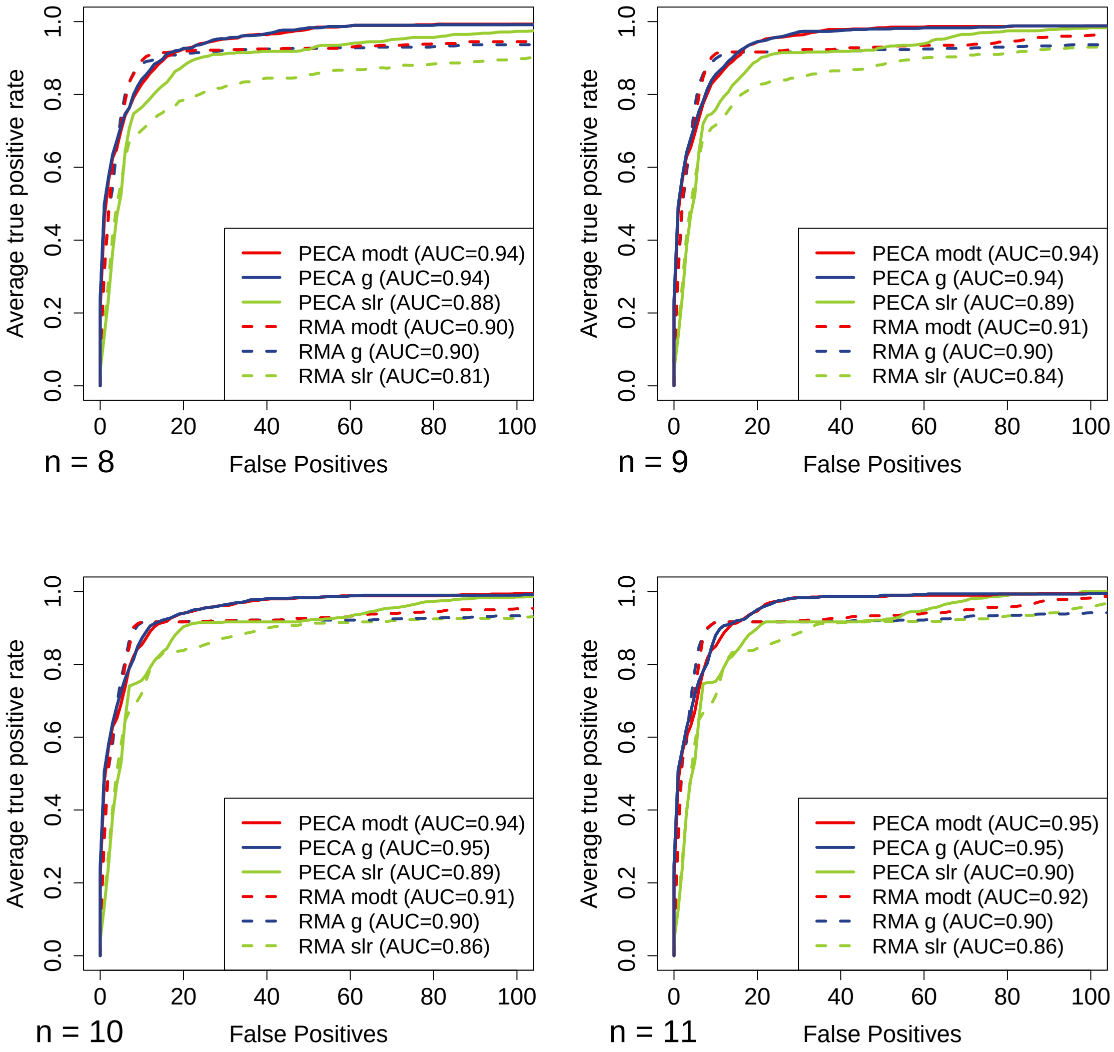




Supplementary figure 2. Variability of the RMA-based intensity values of the probe sets corresponding to the same GeneID for the 10 GeneIDs with the largest numbers of probe sets on the (A) MG-U74Av2 and (B) MOE430 2.0 arrays. The intensity values are shown for the original (black) and verified (grey) probe sets corresponding to the same GeneID on both arrays. The closer the points are along a vertical line, the better is the agreement between the intensity values for the same GeneID on an array. It can be noticed that the probe verification can improve the consistency of the measurements within an array.



Supplementary figure 3.



Supplementary figure 3. (cont.) Averaged Receiver Operator Characteristic (ROC) curves for the PECA signal log-ratio, PECA Hedges' g -statistic, PECA modified t-statistic, RMA signal log-ratio, RMA Hedges' g -statistic, and RMA modified t-statistic (29). The different analysis methods were applied to the Affymetrix HG-U95Av2 spike-in data containing two groups of 12 samples (www.affymetrix.com). In this carefully controlled experiment, it is known that 12 spiked genes are differentially expressed between the two groups, whereas all the other genes are not. Since the truth is known, it is easy to determine true positives (TP) and false positives (FP). We randomly sampled 100 subsets of each possible size from 2 to 11 and determined the average ROC curve over them. The average TP was calculated over the sampled subsets for each FP value, and then plotted against FP. Similarly as Cope et al. (*Bioinformatics* **19**, 185-193, 2003), we restricted the analysis up to 100 FPs, since the lists of genes with more than 100 errors are typically not useful. As a summary statistic, we report the average area under the curve (AUC) up to 100 FPs . The AUCs were standardized so that the largest possible value is 1. The PECA signal log-ratios and Hedges' g -values outperformed the signal log-ratios and Hedges' g -values calculated from the RMA-normalized intensity values, especially with the smallest sample sizes. Moreover, the PECA-estimated Hedges' g performed at least equally well as the modified t-statistic calculated from the RMA-normalized intensity values, being clearly better with sample sizes 2 and 3. The PECA-estimated modified t-statistic could not improve the performance further.