# Cancer outlier differential gene expression detection Supplementary Information

Baolin Wu*
Division of Biostatistics
School of Public Health
University of Minnesota
A460 Mayo Building, MMC 303
Minneapolis, MN, 55455, USA

June 14, 2006

## 1 Outlier robust t-statistic

The outlier robust t-statistic (ORT) for detecting cancer genes with outlier disease samples is defined as

$$t_j^* = \frac{\sum_{i\in R_j}(x_{ij} - \text{med}_{1j})}{\text{median}\{|x_{ij} - \text{med}_{1j}|_{i\le n_1},\ |x_{ij} - \text{med}_{2j}|_{i>n_1}\}}, \quad j = 1, \cdots, p \tag{1}$$

where $R_j$ is the set of "outlier disease samples" for gene $j$. For detecting down-expressed cancer genes, we define

$$R_j = \{i > n_1 : x_{ij} < q_{25}(x_{kj} : k = 1, \cdots, n_1) - \text{IQR}(x_{kj} : k = 1, \cdots, n_1)\} \tag{2}$$

For detecting over-expressed cancer genes, we define

$$R_j = \{i > n_1 : x_{ij} > q_{75}(x_{kj} : k = 1, \cdots, n_1) + \text{IQR}(x_{kj} : k = 1, \cdots, n_1)\} \tag{3}$$

## 2 Simulation study

### 2.1 False and true positive rates comparison

Figures 1 through 4 compare the false/true positive rates estimated from 1000 simulations for the four cancer gene outlier detection methods.

*Email: baolin@biostat.umn.edu, phone: (612)624-0647, fax: (612)626-0660

Figure 1: False/true positive rates comparisons: $n = 25, \mu = 1$

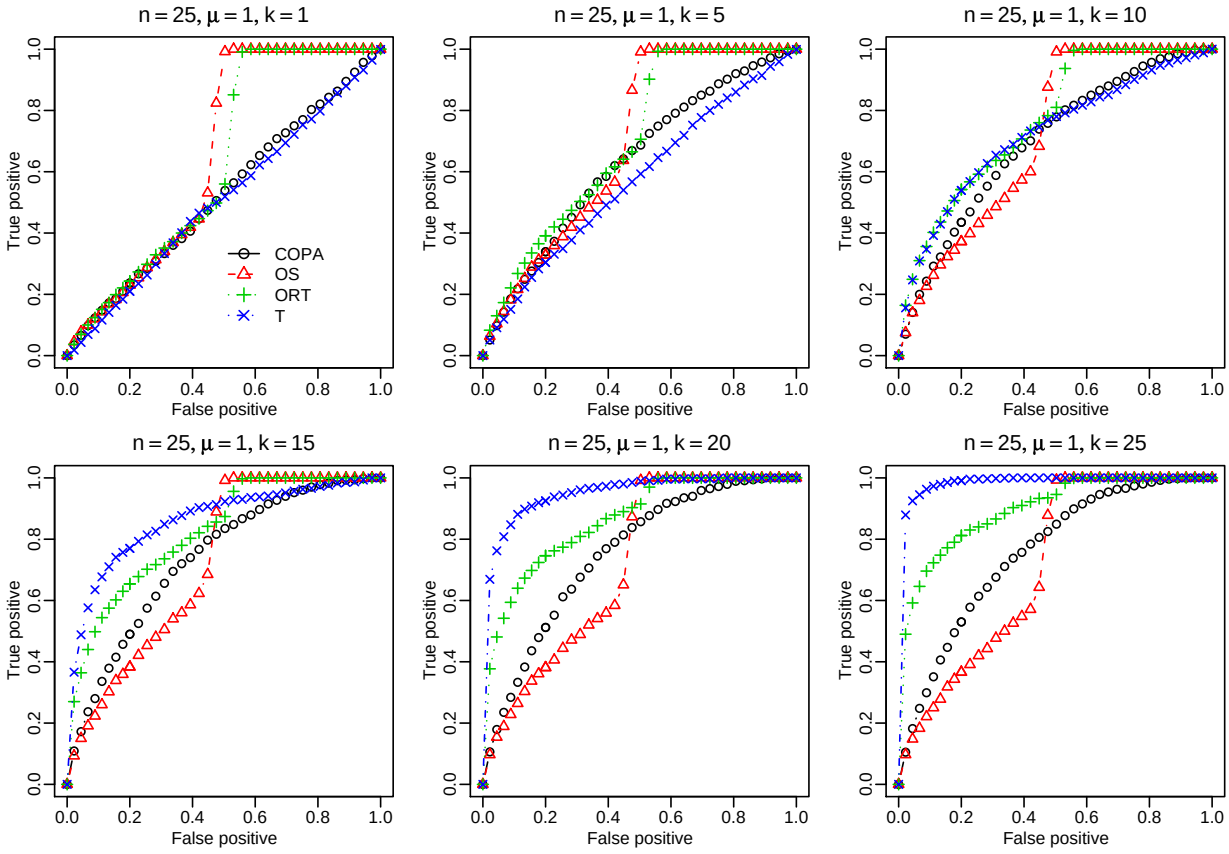


Figure 2: False/true positive rates comparisons: $n = 25, \mu = 2$

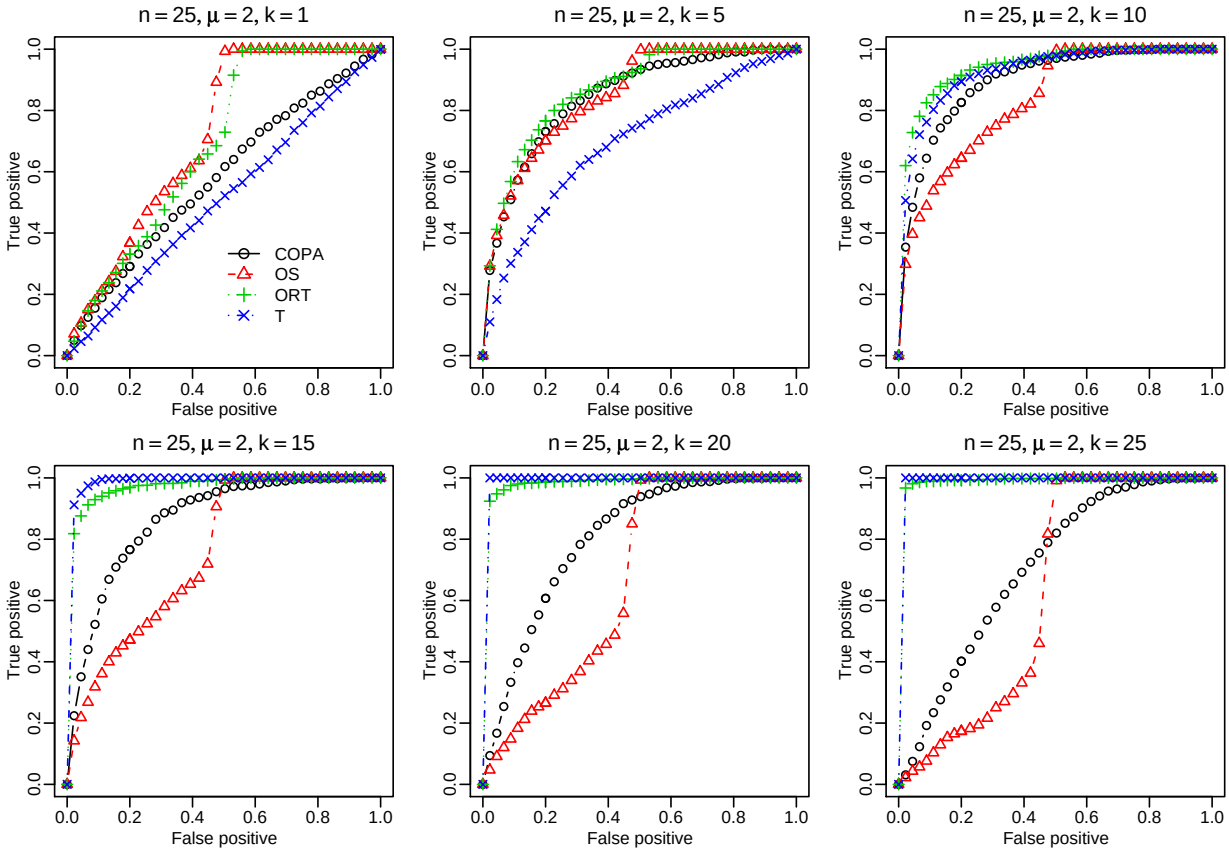


Figure 3: False/true positive rates comparisons: $n = 15, \mu = 1$

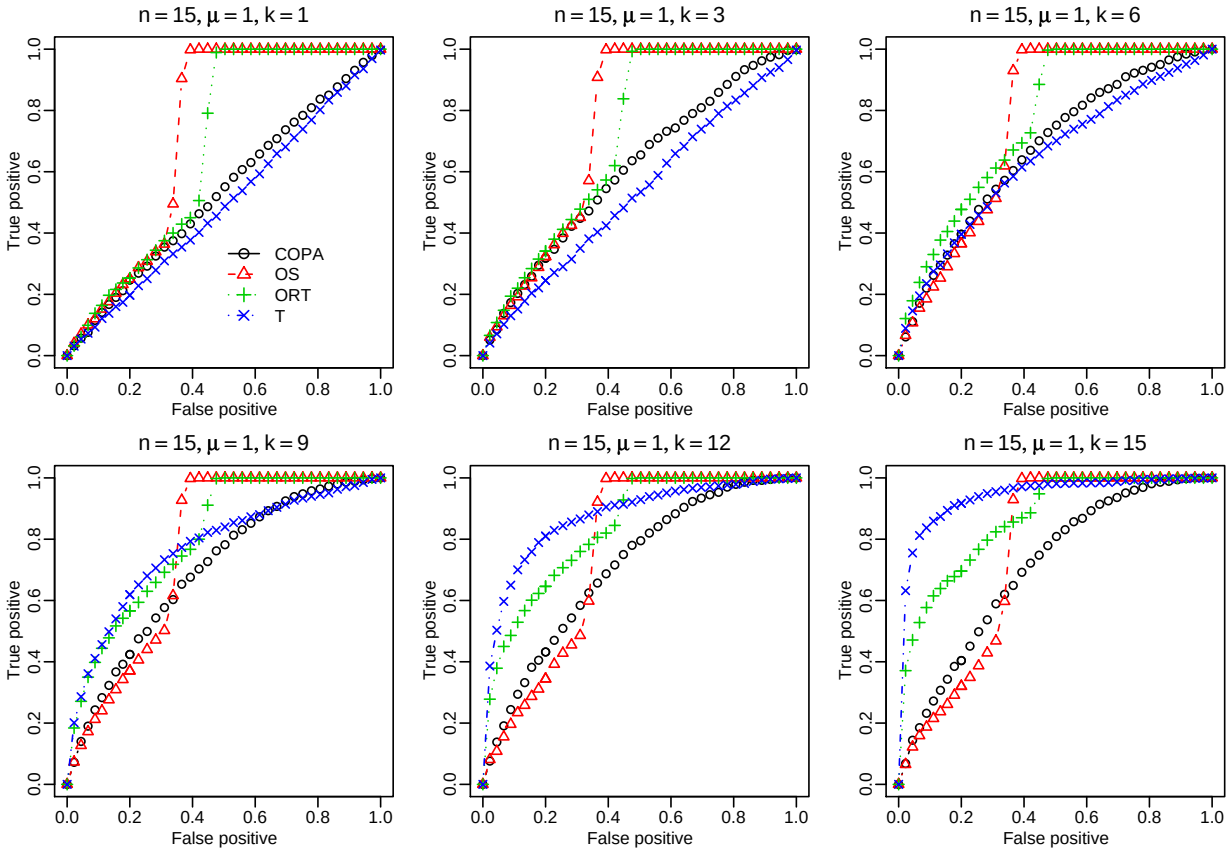


Figure 4: False/true positive rates comparisons: $n = 15, \mu = 2$

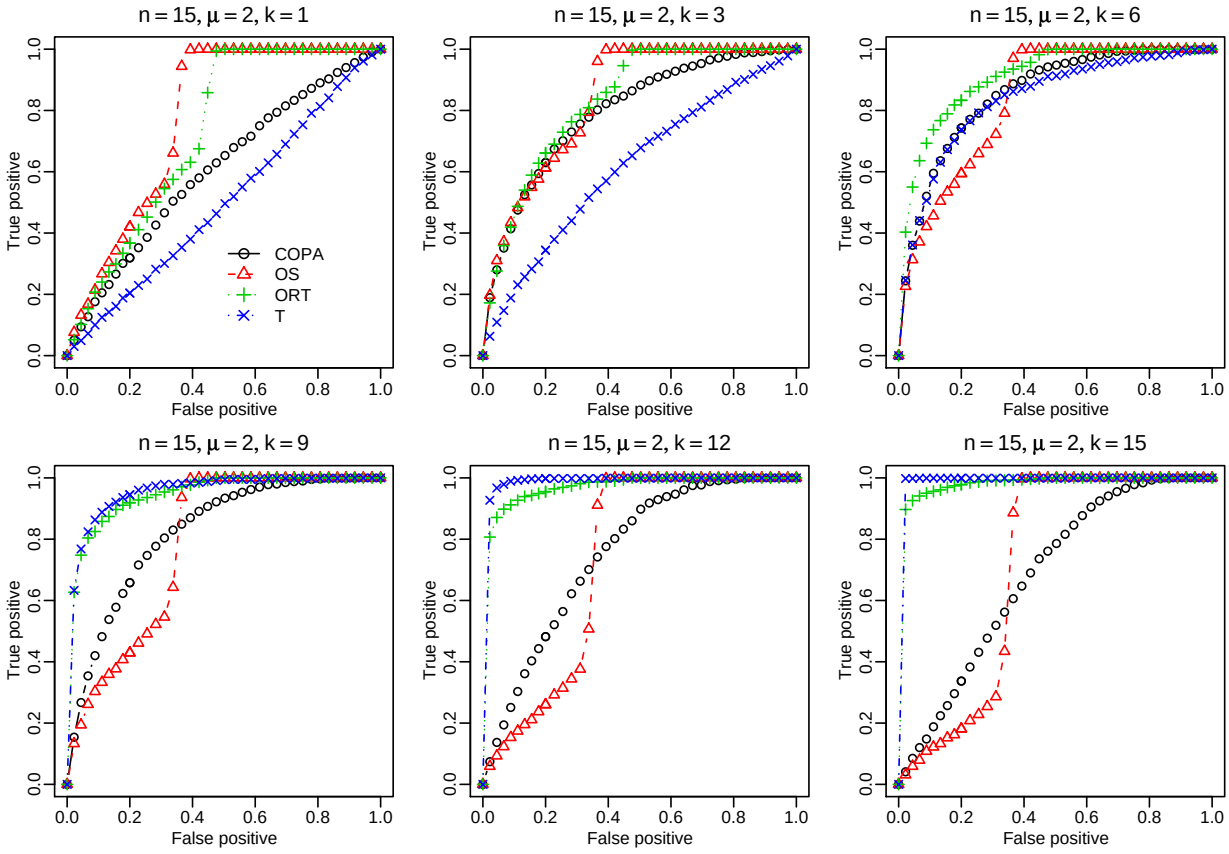
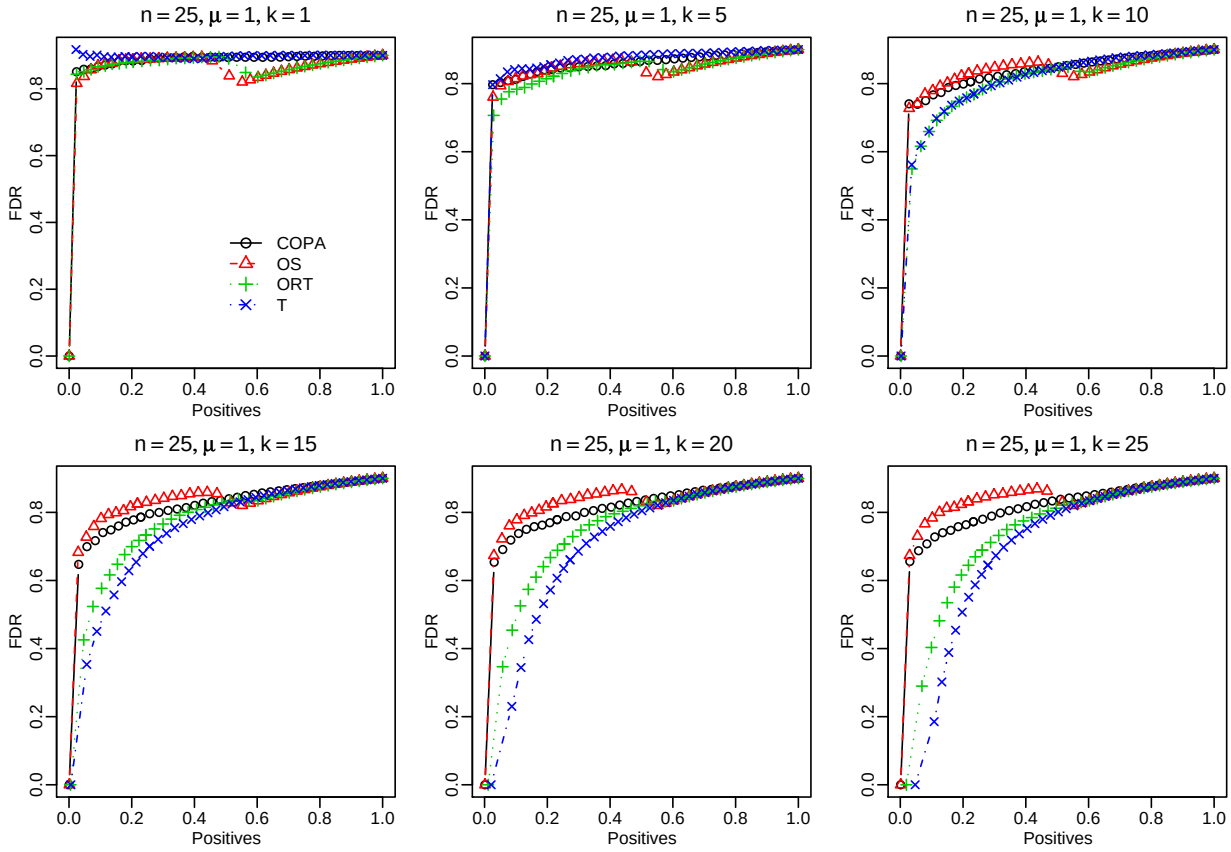


Figure 5: False discovery rates comparisons: $n = 25, \mu = 1, p = 1000, \pi_0 = 0.9$




## 2.2 False discovery rates comparison

Figures 5 through 16 compare the false discovery rates estimated from 1000 simulations for the four cancer gene outlier detection methods.

# 3 Breast cancer microarray data analysis

The breast cancer microarray data reported by West *et al.* (2001) contained the expression levels of 7129 genes from 49 breast tumor samples. Each sample had a binary outcome describing the status of lymph node involvement in breast cancer. Among them, 25 tumor samples had no positive lymph nodes discovered and 24 tumor samples had identifiably positive nodes. The gene expressions, obtained from the Affymetrix human HuGeneFL GeneChip, can be downloaded from `http://data.cgt.duke.edu/west.php`. We normalize the data using quantile normalization (Bolstad *et al.*, 2003), and then log transform the intensities for followup statistical analysis. In the cancer gene outlier detection, we treat the negative group as the normal class. We applied the t-statistic, COPA, OS, and the proposed

Figure 6: False discovery rates comparisons: $n = 25, \mu = 1, p = 1000, \pi_0 = 0.8$

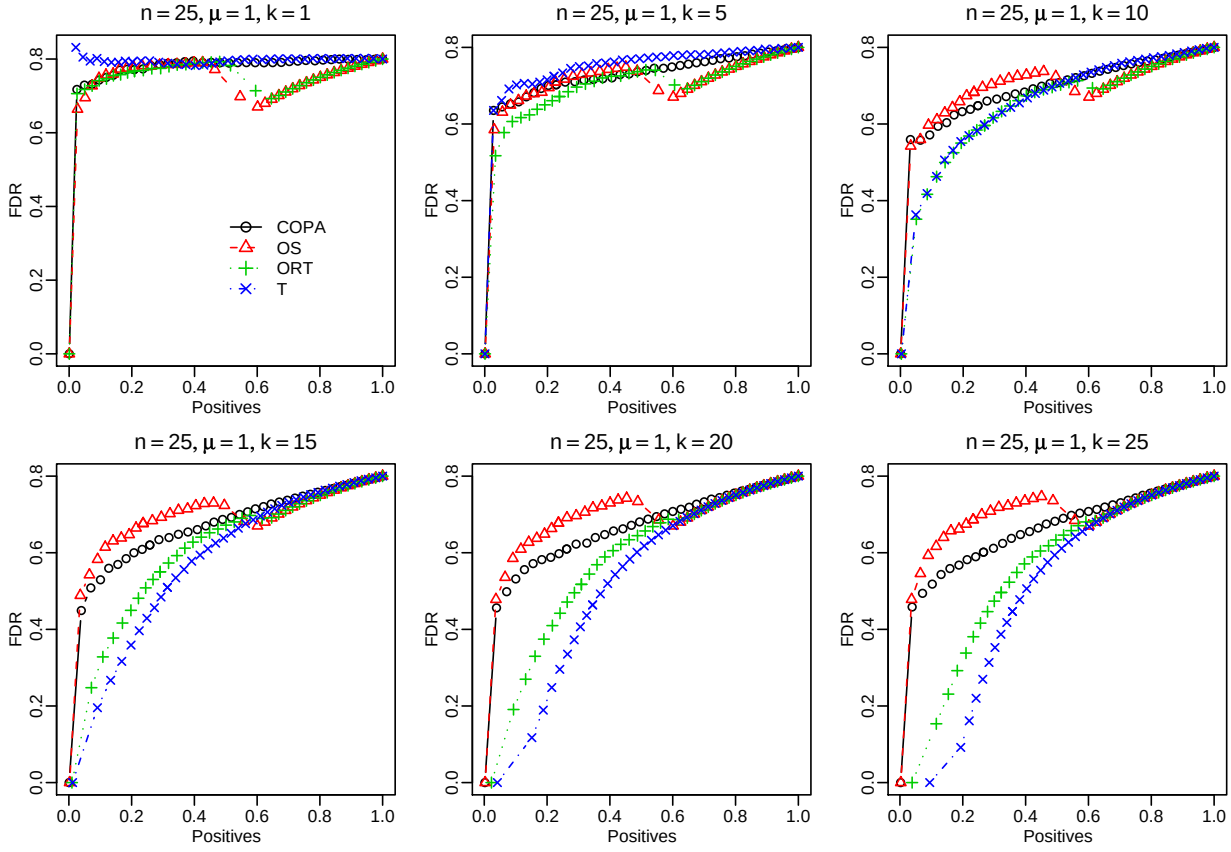


Figure 7: False discovery rates comparisons: $n = 25, \mu = 1, p = 1000, \pi_0 = 0.7$

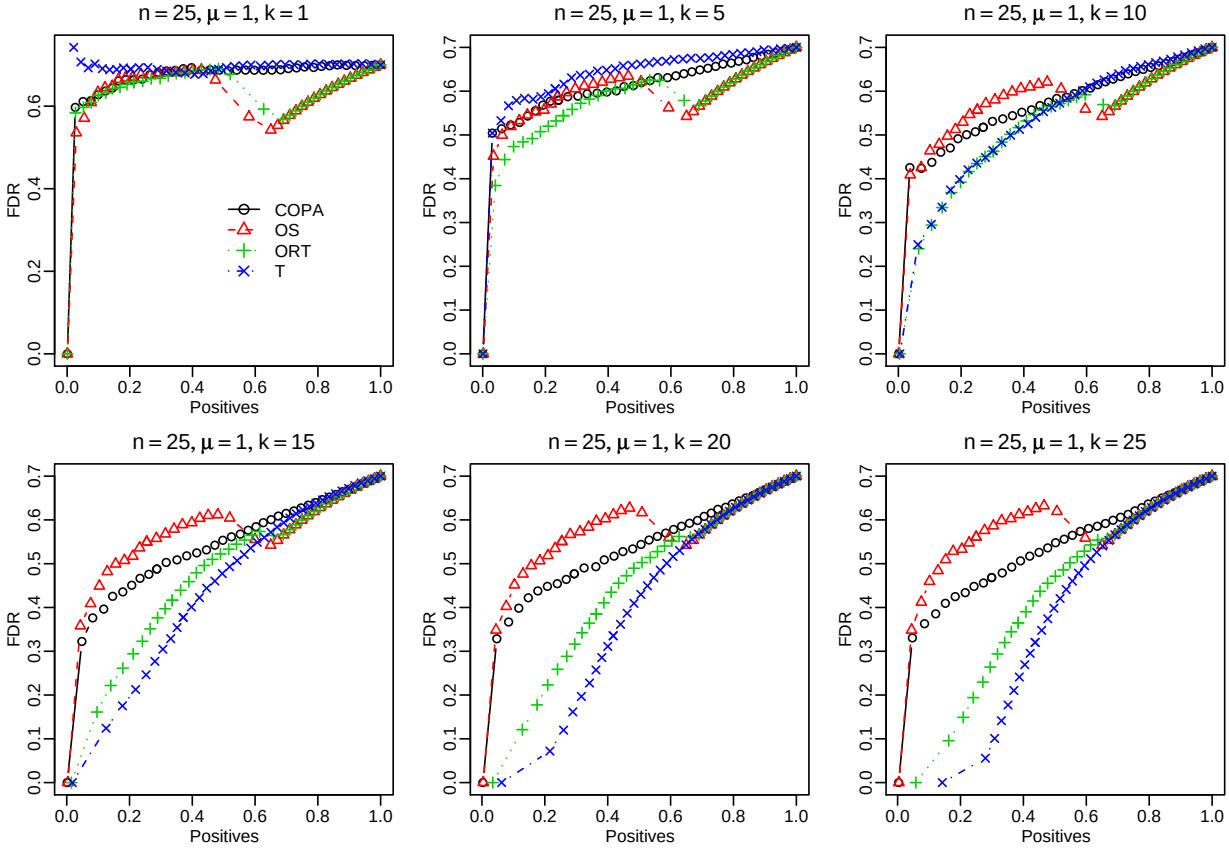


Figure 8: False discovery rates comparisons: $n = 25, \mu = 2, p = 1000, \pi_0 = 0.9$

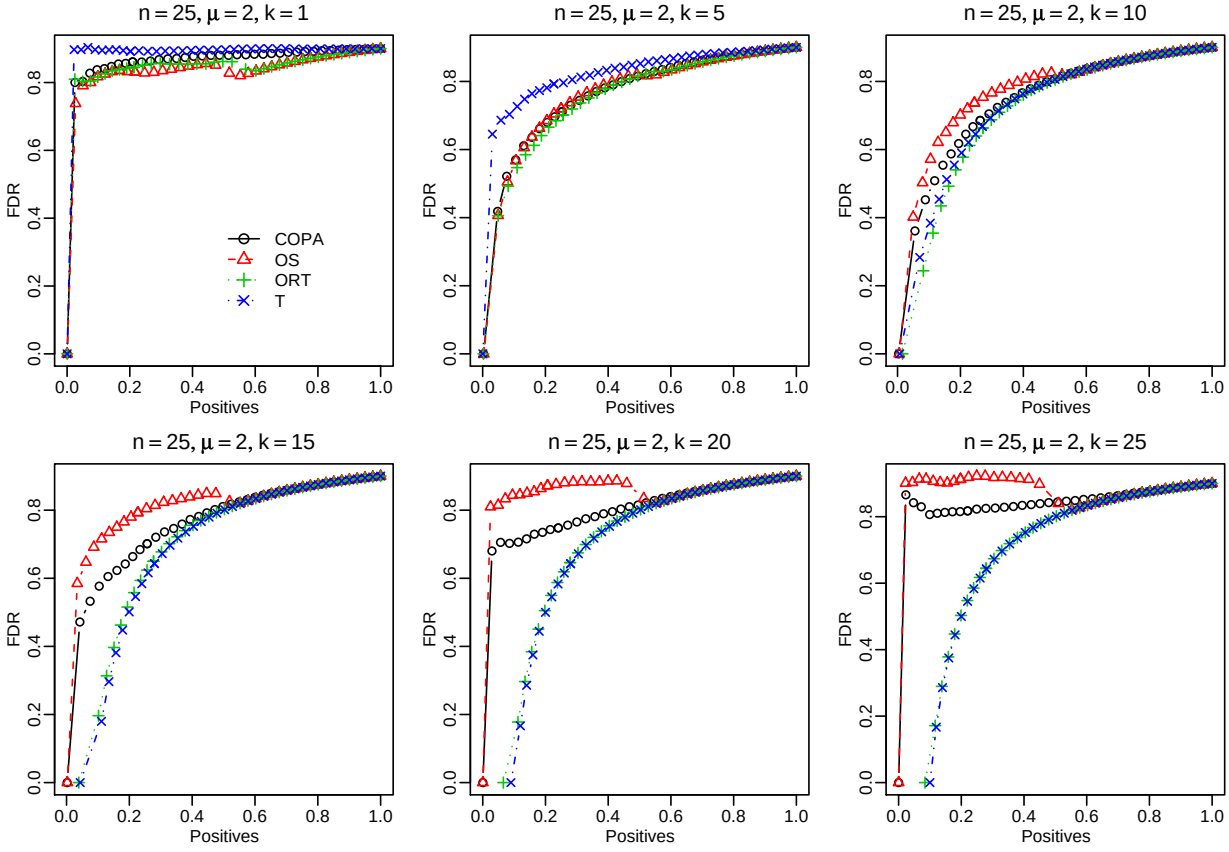


Figure 9: False discovery rates comparisons: $n = 25, \mu = 2, p = 1000, \pi_0 = 0.8$

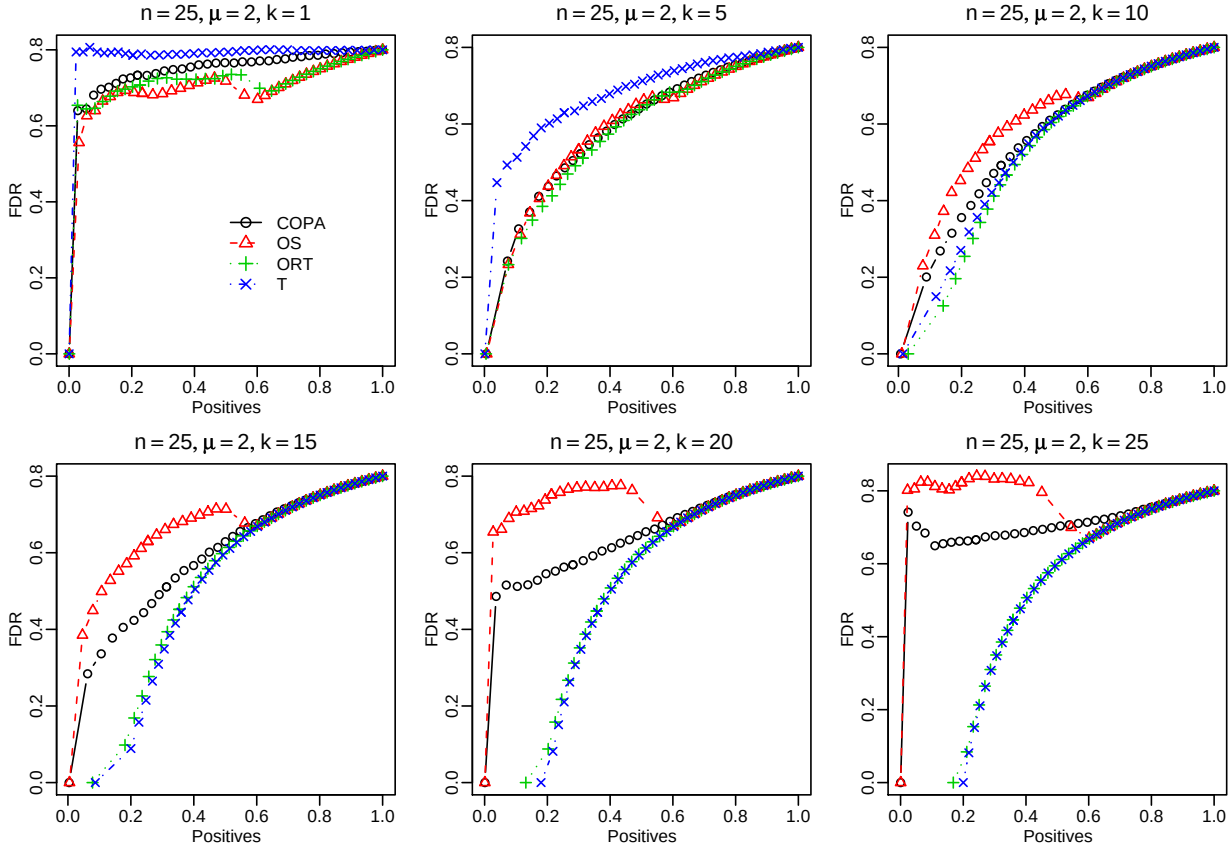


Figure 10: False discovery rates comparisons: $n = 25, \mu = 2, p = 1000, \pi_0 = 0.7$

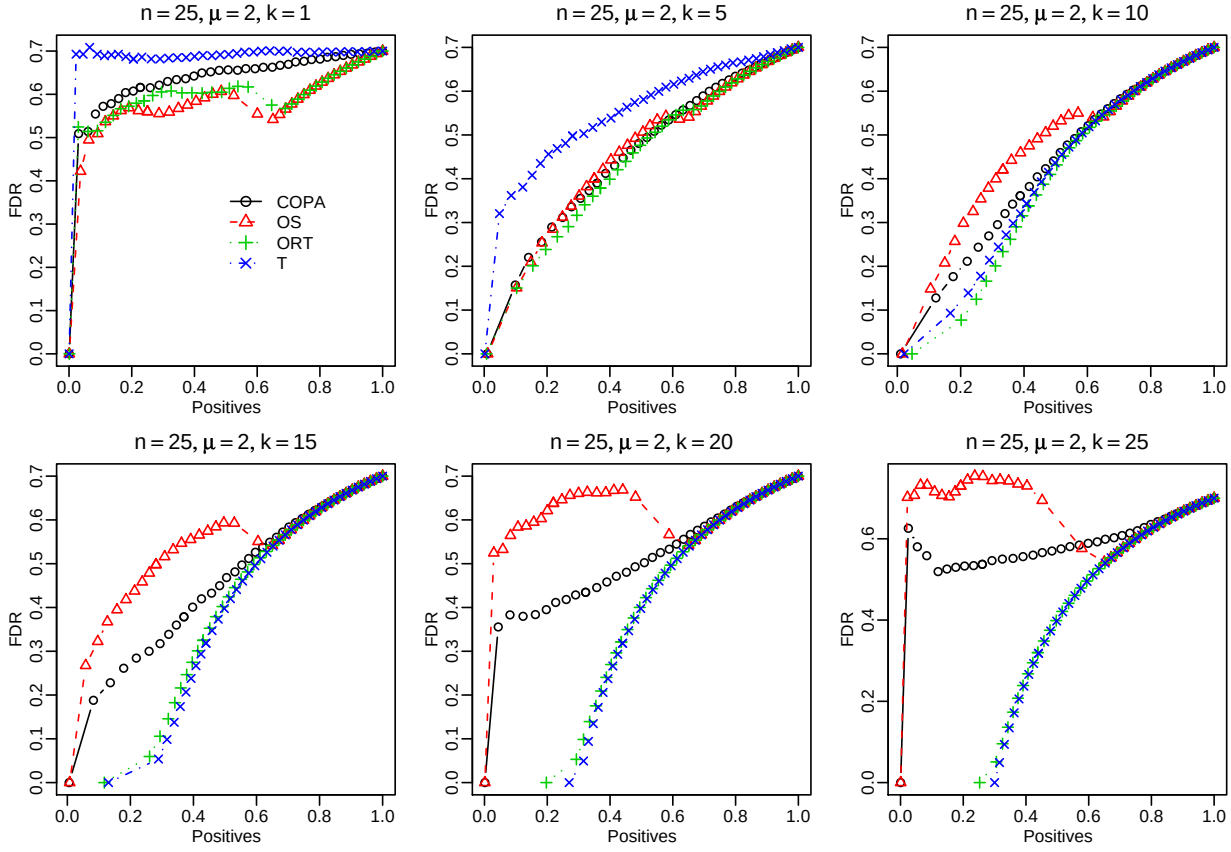


Figure 11: False discovery rates comparisons: $n = 15, \mu = 1, p = 1000, \pi_0 = 0.9$

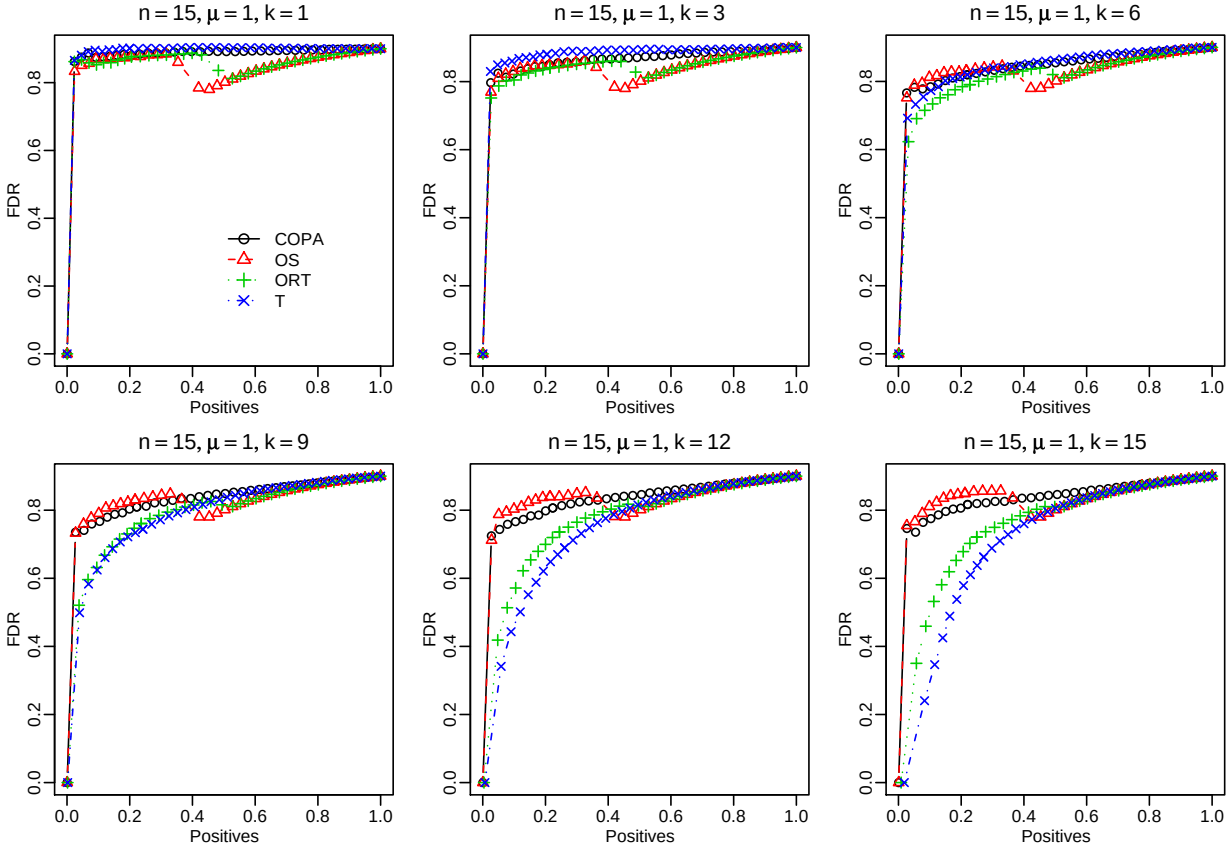


Figure 12: False discovery rates comparisons: $n = 15, \mu = 1, p = 1000, \pi_0 = 0.8$

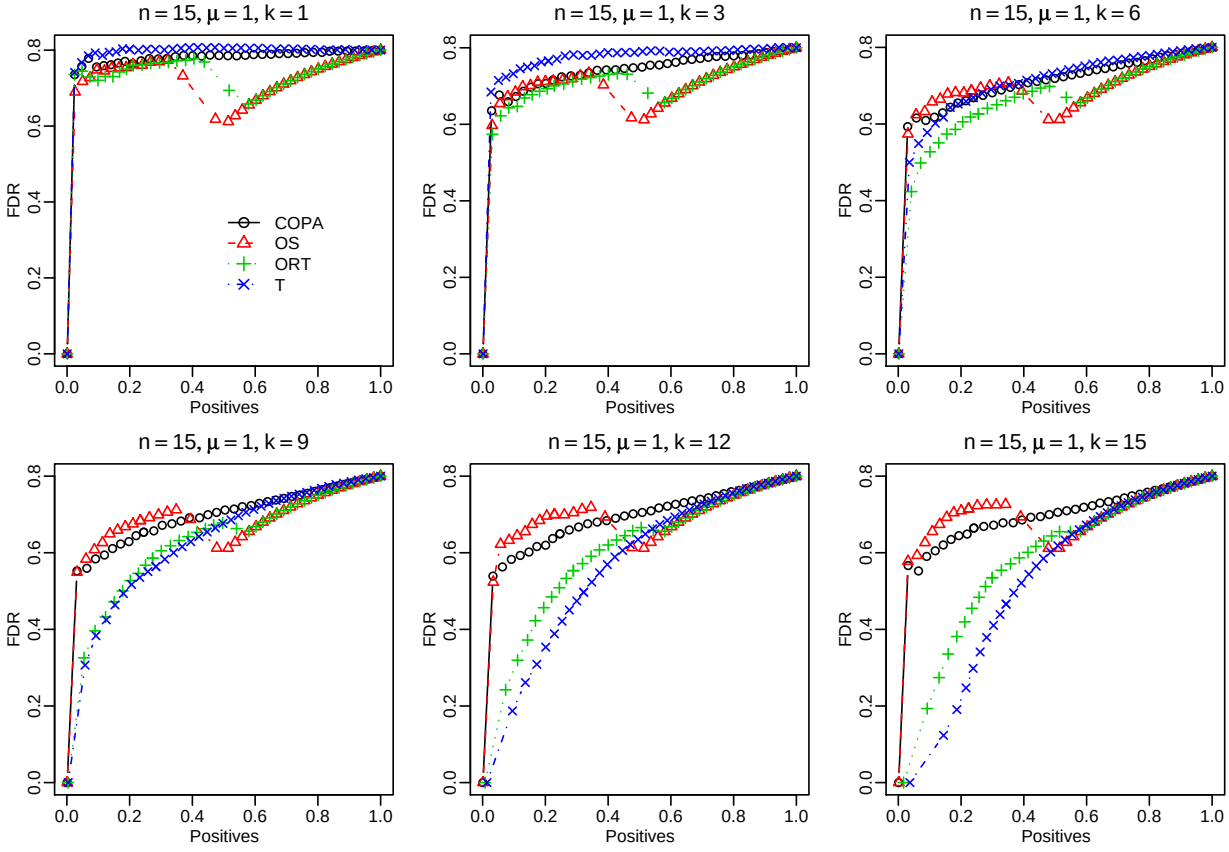


Figure 13: False discovery rates comparisons: $n = 15, \mu = 1, p = 1000, \pi_0 = 0.7$

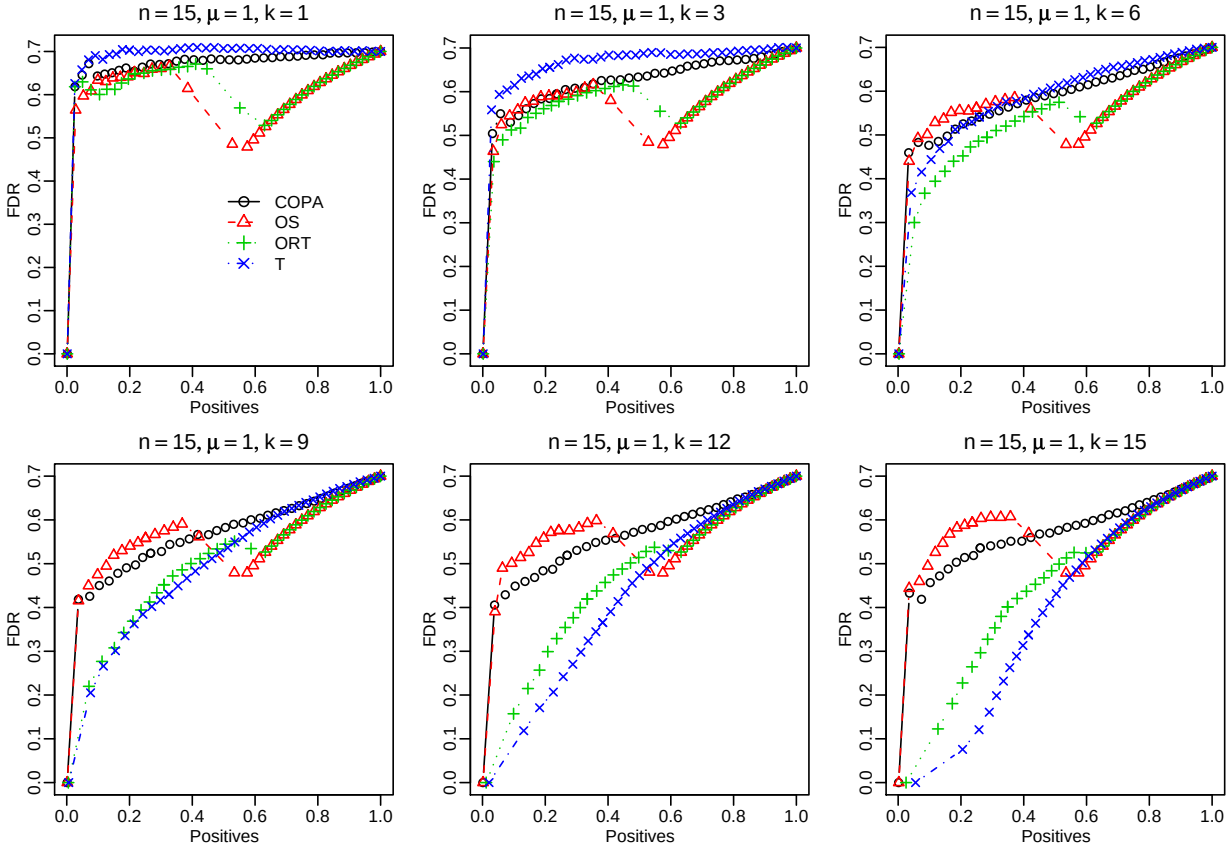


Figure 14: False discovery rates comparisons: $n = 15, \mu = 2, p = 1000, \pi_0 = 0.9$

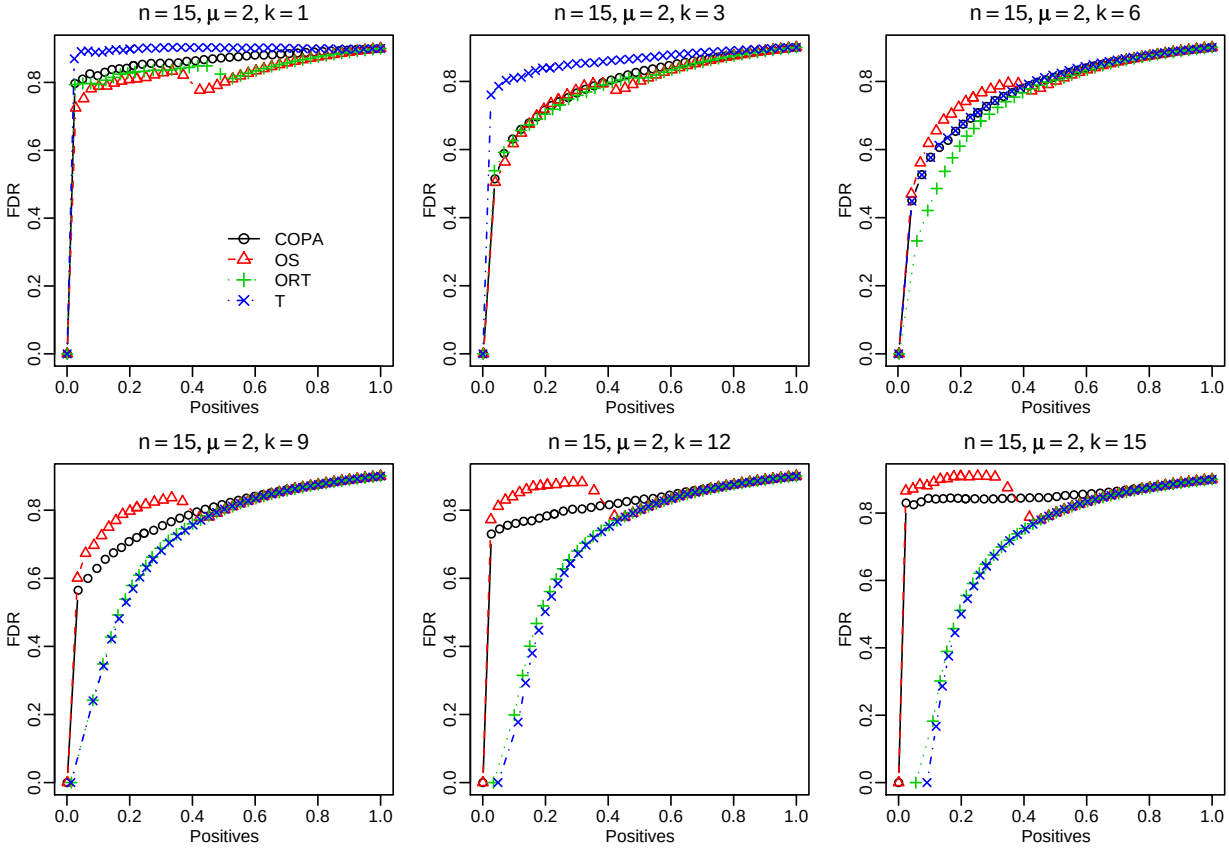


Figure 15: False discovery rates comparisons: $n = 15, \mu = 2, p = 1000, \pi_0 = 0.8$

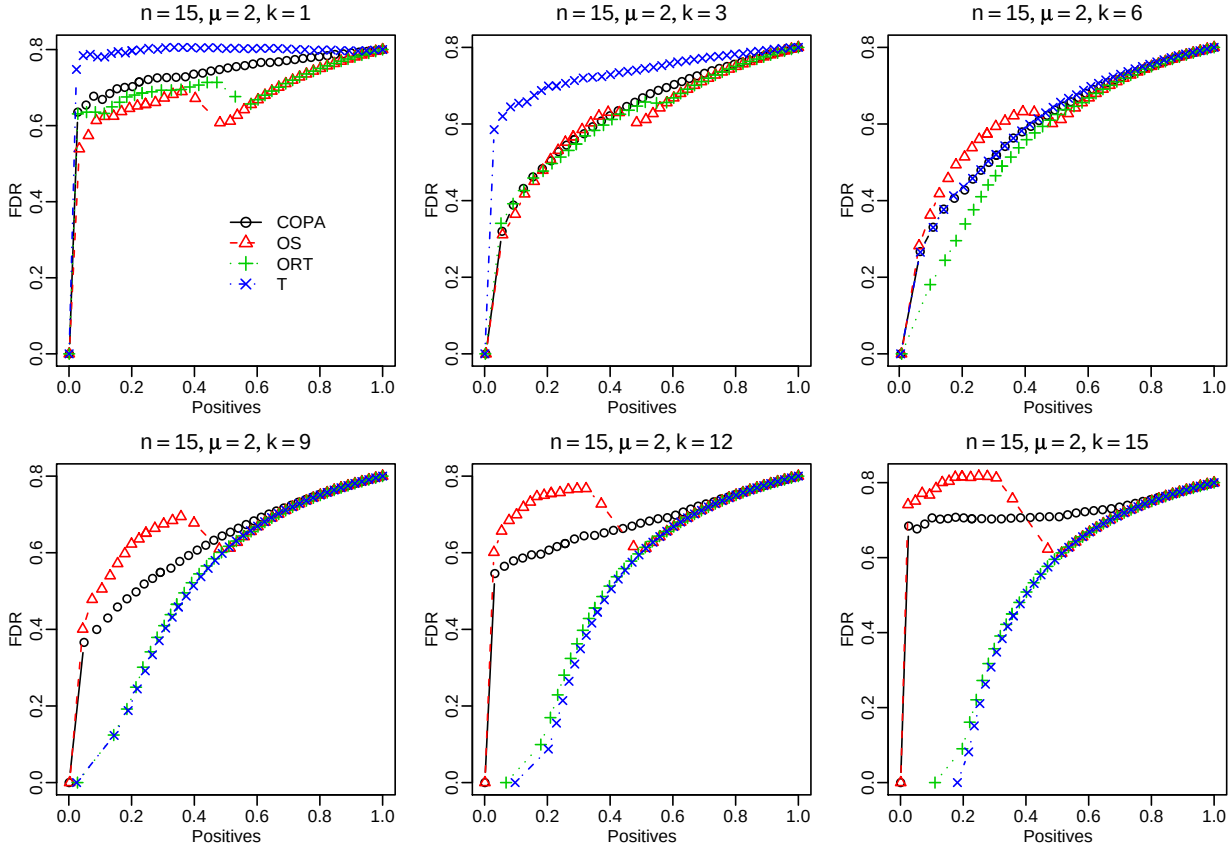
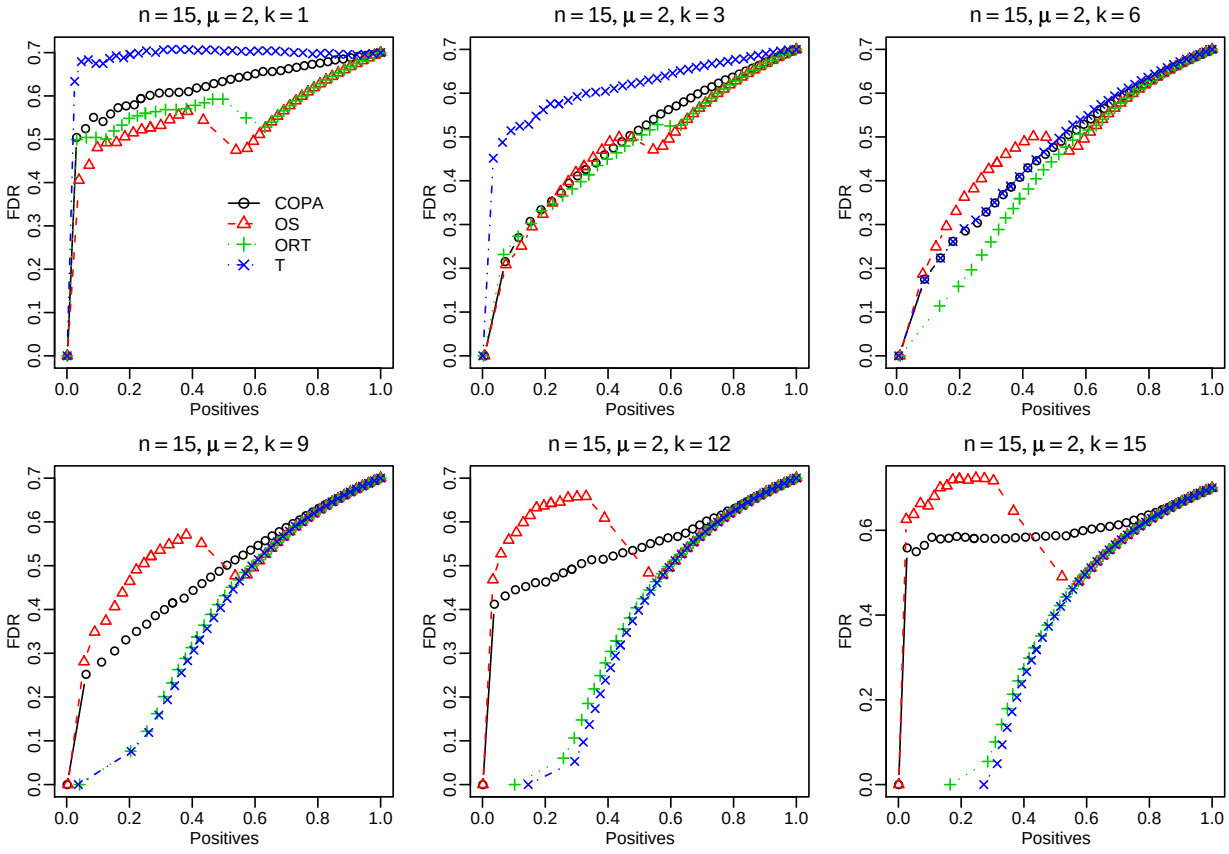


Figure 16: False discovery rates comparisons: $n = 15, \mu = 2, p = 1000, \pi_0 = 0.7$



ORT to detect cancer genes with outlier disease samples. We rank the genes based on each test statistic.

## 3.1 Cancer genes with over-expressed outlier disease samples

Table 1 to 4 list the top 25 genes identified by each outlier detection statistic.

Table 1: Top 25 (over-expressed) genes identified by the outlier robust t-statistic (ORT). Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 1 | Hs.7195 | GABRG2 |
| 2 | Hs.477891 | CPB1 |
| 3 | Hs.2012 | TCN1 |
| 4 | Hs.446291 | MSR1 |
| 5 | Hs.98428 | HOXB6 |
| 6 | Hs.1290 | C9 |
| ***7*** | Hs.435561 | ATM |
| 8 | Hs.196983 | SSFA2 |
| ***9*** | Hs.390729 | ERBB4 |
| 10 | Hs.129944 | ESM1 |
| 11 | Hs.437040 | PTPN21 |
| 12 | Hs.79387 | PSMC5 |
| 13 | Mm.29182 | Taldo1 |
| 14 | Hs.75294 | CRH |
| 15 | Hs.144795 | KCNMA1 |
| 16 | Hs.487325 | PRKACB |
| ***17*** | Hs.724 | THRA |
| ***18*** | Hs.327527 | SMARCA4 |
| ***19*** | Hs.460996 | TRADD |
| ***20*** | Hs.534310 | CTAG1B |
| ***21*** | Hs.477887 | AGTR1 |
| ***22*** | Hs.350229 | CASC3 |
| 23 | Hs.271003 | LOC440118 |
| 24 | Hs.352243 | CLCNKB |
| 25 | Hs.2210 | ZNHIT3 |

## 3.2 Cancer genes with down-expressed outlier disease samples

Table 5 shows the expression profiles of the identified genes with down-expressed disease samples that have been confirmed associated with the breast cancer in previous studies.

Table 2: Top 25 genes identified by the t-statistic. The first gene has no annotation. Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 2 | Hs.438918 | ACVR1B |
| 3 | Hs.434973 | GYPB |
| 4 | Rn.103750 | Fos |
| 5 | Hs.352243 | CLCNKB |
| 6 | Hs.423935 | RDBP |
| 7 | Hs.152213 | WNT5A |
| 8 | Hs.197320 | TLE1 |
| 9 | Hs.388034 | RXRB |
| 10 | Hs.534311 | CYP2D6 |
| 11 | Hs.198072 | PDE4B |
| 12 | Hs.491582 | PLAT |
| 13 | Hs.208597 | CTBP1 |
| 14 | Hs.20131 | NR6A1 |
| 15 | Hs.193725 | PSMD5 |
| 16 | Hs.131269 | RARRES1 |
| 17 | Hs.555883 | NF1 |
| ***18*** | Hs.435561 | ATM |
| 19 | Hs.376046 | BTN3A2 |
| 20 | Hs.123034 | GPR12 |
| 21 | Hs.512676 | RPS25 |
| 22 | Hs.74565 | APLP1 |
| ***23*** | Hs.338207 | FRAP1 |
| ***24*** | Hs.487046 | SOD2 |
| 25 | Hs.307905 | RELB |

Figure 17 shows the expression profiles of these genes.

Table 6 to 8 list the top 25 genes identified by each outlier detection statistic.

# References


Bolstad,B., Irizarry,R., Astrand,M. and Speed,T. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics,* **19** (2), 185–193.

West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,John A.,J., Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *PNAS,* **98** (20), 11462–11467.

Figure 17: Oncogene outlier detection for breast cancer microarray data: 8 top ranking breast genes that are identified by ORT and confirmed associated with breast cancer in the literature are plotted. The lymph node negative samples (LN-) serve as the normal group, and the lymph node positive samples (LN+) are treated as the disease group in the outlier detection analysis. We have added some jittering to the horizontal positions to distinguish among close points. The title lists the gene names. Within the parentheses are those outlier statistics that ranked the gene in top 25.

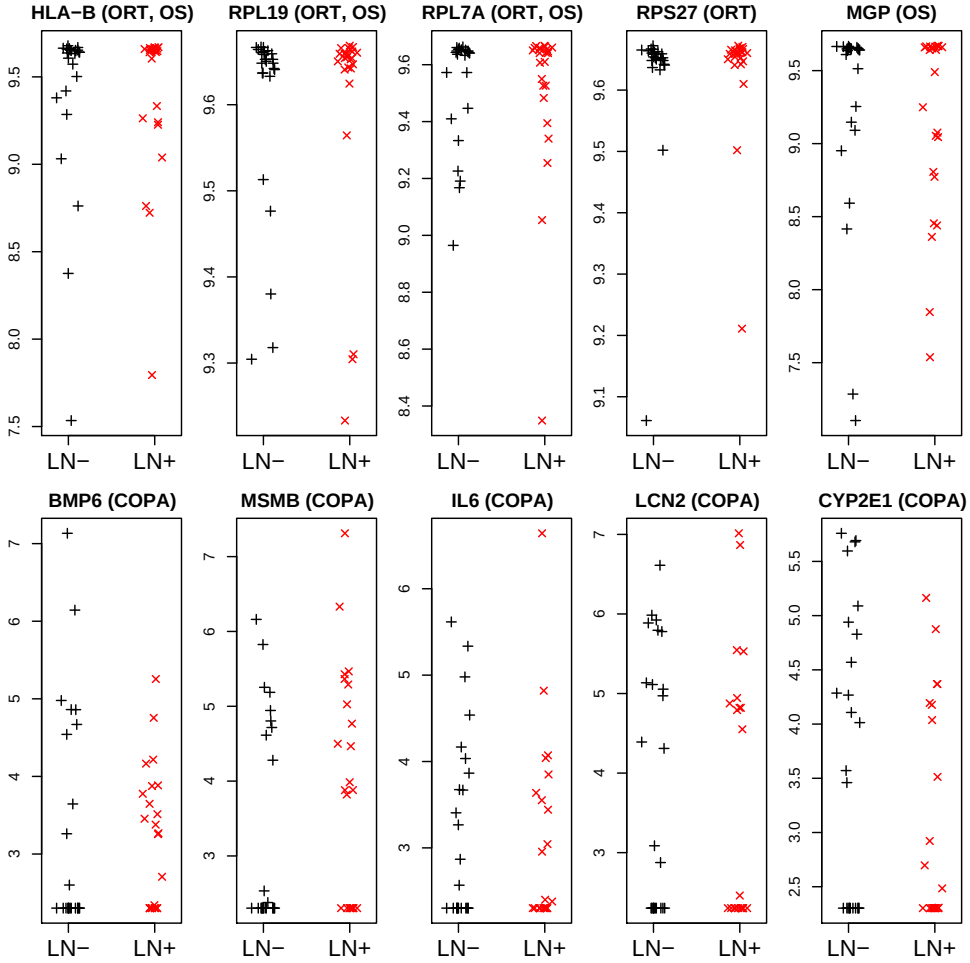


Table 3: Top 25 (over-expressed) genes identified by the outlier sum (OS) statistic. Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 1 | Hs.497200 | PLA2G4A |
| 2 | Hs.449076 | PWP2H |
| 3 | Hs.2012 | TCN1 |
| 4 | Hs.255462 | MSMB |
| ***5*** | Hs.512234 | IL6 |
| 6 | Hs.477891 | CPB1 |
| 7 | Hs.437040 | PTPN21 |
| 8 | Hs.196983 | SSFA2 |
| 9 | Hs.165258 | NR4A2 |
| 10 | Hs.9914 | FST |
| 11 | Hs.98428 | HOXB6 |
| 12 | Hs.369009 | SLC18A2 |
| 13 | Hs.464985 | RIT2 |
| ***14*** | Hs.477887 | AGTR1 |
| ***15*** | Hs.435714 | PAK1 |
| ***16*** | Hs.350229 | CASC3 |
| 17 | Hs.408458 | WWP2 |
| 18 | Hs.54415 | CSN3 |
| 19 | Hs.154658 | PSD |
| 20 | Hs.272499 | DHRS2 |
| 21 | Hs.3109 | ARHGAP4 |
| 22 | Hs.555888 | PSG5 |
| 23 | Hs.487325 | PRKACB |
| 24 | Hs.372360 | PTHB1 |
| 25 | Hs.54505 | AQP6 |

Table 4: Top 25 (over-expressed) genes identified by the cancer outlier profile analysis (COPA). Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 1 | Hs.20131 | NR6A1 |
| 2 | Hs.381285 | ZNF45 |
| 3 | Hs.435044 | TBC1D22A |
| 4 | Hs.446291 | MSR1 |
| 5 | Hs.497200 | PLA2G4A |
| 6 | Hs.160411 | TSHR |
| 7 | Hs.285671 | BMP6 |
| 8 | Mm.29182 | Taldo1 |
| 9 | Hs.46 | PTAFR |
| 10 | Hs.255462 | MSMB |
| 11 | Hs.183109 | MAOA |
| 12 | Hs.449076 | PWP2H |
| 13 | Hs.256067 | PRKAA2 |
| 14 | Hs.2012 | TCN1 |
| 15 | Hs.272011 | B4GALT1 |
| 16 | Hs.514477 | LLGL2 |
| ***17*** | Hs.512234 | IL6 |
| 18 | Hs.73078 | DAZL |
| 19 | Hs.65734 | ARNTL |
| 20 | Hs.75294 | CRH |
| ***21*** | Hs.204238 | LCN2 |
| 22 | Hs.288867 | XPA |
| 23 | Hs.432458 | PRG4 |
| 24 | Hs.165258 | NR4A2 |
| 25 | Hs.442182 | ABCC6 |

Table 5: Genes (down-expressed) ranked in top 25 by the outlier detection statistics and confirmed associated with breast cancer in previous studies. The last four columns also list the ranking of each gene by the four methods.

| Methods | Rank | UniGene ID | Gene Name | T | COPA | OS | ORT |
|---|---|---|---|---|---|---|---|
| T | 18 | Hs.435561 | ATM | | 3668 | 5048 | 4995 |
| | 23 | Hs.338207 | FRAP1 | | 762 | 336 | 71 |
| | 24 | Hs.487046 | SOD2 | | 5275 | 5048 | 4995 |
| COPA | 6 | Hs.285671 | BMP6 | 4756 | | 5048 | 4995 |
| | 8 | Hs.255462 | MSMB | 1350 | | 5048 | 4995 |
| | 12 | Hs.512234 | IL6 | 3447 | | 5048 | 4995 |
| | 18 | Hs.204238 | LCN2 | 3367 | | 5048 | 4995 |
| | 23 | Hs.12907 | CYP2E1 | 542 | | 5048 | 4995 |
| OS | 9 | Hs.365706 | MGP | 3483 | 5394 | | 31 |
| | 14 | Hs.77961 | HLA-B | 5418 | 5226 | | 6 |
| | 19 | Hs.499839 | RPL7A | 4194 | 5346 | | 21 |
| | 20 | Hs.381061 | RPL19 | 5146 | 4187 | | 19 |
| ORT | 6 | Hs.77961 | HLA-B | 5418 | 5226 | 14 | |
| | 19 | Hs.381061 | RPL19 | 5146 | 4187 | 20 | |
| | 21 | Hs.499839 | RPL7A | 4194 | 5346 | 19 | |
| | 25 | Hs.504517 | RPS27 | 4471 | 5157 | 30 | |

Table 6: Top 25 (down-expressed) genes identified by the outlier robust t-statistic (ORT). Several genes have no annotation. Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 4 | Hs.520640 | ACTB |
| 5 | Hs.178551 | RPL8 |
| ***6*** | Hs.77961 | HLA-B |
| 7 | Hs.520640 | ACTB |
| 8 | Hs.356502 | RPLP1 |
| 9 | Hs.546356 | RPL13A |
| 10 | Hs.80545 | RPL37 |
| 11 | Hs.408054 | RPL12 |
| 12 | Hs.497353 | MED6 |
| 13 | Hs.156367 | RPS29 |
| 15 | Hs.505705 | MYL6 |
| 16 | Hs.247828 | RPL23AP7 |
| 17 | Hs.300141 | RPL39 |
| 18 | Hs.400295 | RPL30 |
| ***19*** | Hs.381061 | RPL19 |
| 20 | Hs.418241 | MT2A |
| ***21*** | Hs.499839 | RPL7A |
| 22 | Hs.242947 | RPL41 |
| 23 | Hs.356366 | RPS2 |
| 24 | Hs.397609 | RPS16 |
| ***25*** | Hs.504517 | RPS27 |

Table 7: Top 25 (down-expressed) genes identified by the outlier sum (OS) statistic. Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 4 | Hs.520640 | ACTB |
| 5 | Hs.449621 | IGKC |
| 6 | Hs.520640 | ACTB |
| 7 | Hs.178551 | RPL8 |
| 8 | Hs.348935 | IGLL1 |
| ***9*** | Hs.365706 | MGP |
| 10 | Hs.356502 | RPLP1 |
| 11 | Hs.418241 | MT2A |
| 12 | Hs.80545 | RPL37 |
| 13 | Hs.546356 | RPL13A |
| ***14*** | Hs.77961 | HLA-B |
| 15 | Hs.497353 | MED6 |
| 16 | Hs.408054 | RPL12 |
| 17 | Hs.247828 | RPL23AP7 |
| 18 | Hs.505705 | MYL6 |
| ***19*** | Hs.499839 | RPL7A |
| ***20*** | Hs.381061 | RPL19 |
| 21 | Hs.300141 | RPL39 |
| 22 | Hs.400295 | RPL30 |
| 23 | Hs.374596 | TPT1 |
| 24 | Hs.242947 | RPL41 |
| 25 | Hs.156367 | RPS29 |

Table 8: Top 25 (down-expressed) genes identified by the cancer outlier profile analysis (COPA). Those in bold face font have been studied previously and confirmed associated with breast cancer in the literature.

| Ranking | UniGene ID | Gene Name |
|---|---|---|
| 1 | Hs.20131 | NR6A1 |
| 2 | Hs.381285 | ZNF45 |
| 3 | Hs.435044 | TBC1D22A |
| 4 | Hs.446291 | MSR1 |
| 5 | Hs.497200 | PLA2G4A |
| ***6*** | Hs.285671 | BMP6 |
| 7 | Hs.160411 | TSHR |
| ***8*** | Hs.255462 | MSMB |
| 9 | Hs.46 | PTAFR |
| 10 | Hs.183109 | MAOA |
| 11 | Hs.449076 | PWP2H |
| ***12*** | Hs.512234 | IL6 |
| 13 | Mm.29182 | Taldo1 |
| 14 | Hs.65734 | ARNTL |
| 15 | Hs.73078 | DAZL |
| 16 | Hs.256067 | PRKAA2 |
| 17 | Hs.272011 | B4GALT1 |
| ***18*** | Hs.204238 | LCN2 |
| 19 | Hs.514477 | LLGL2 |
| 20 | Hs.432458 | PRG4 |
| 21 | Hs.2012 | TCN1 |
| 22 | Hs.464985 | RIT2 |
| ***23*** | Hs.12907 | CYP2E1 |
| 24 | Hs.165258 | NR4A2 |
| 25 | Hs.288867 | XPA |