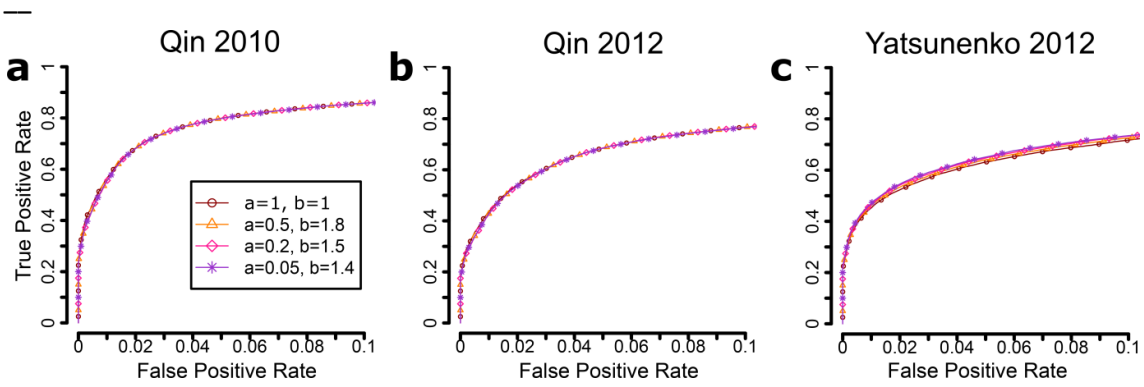
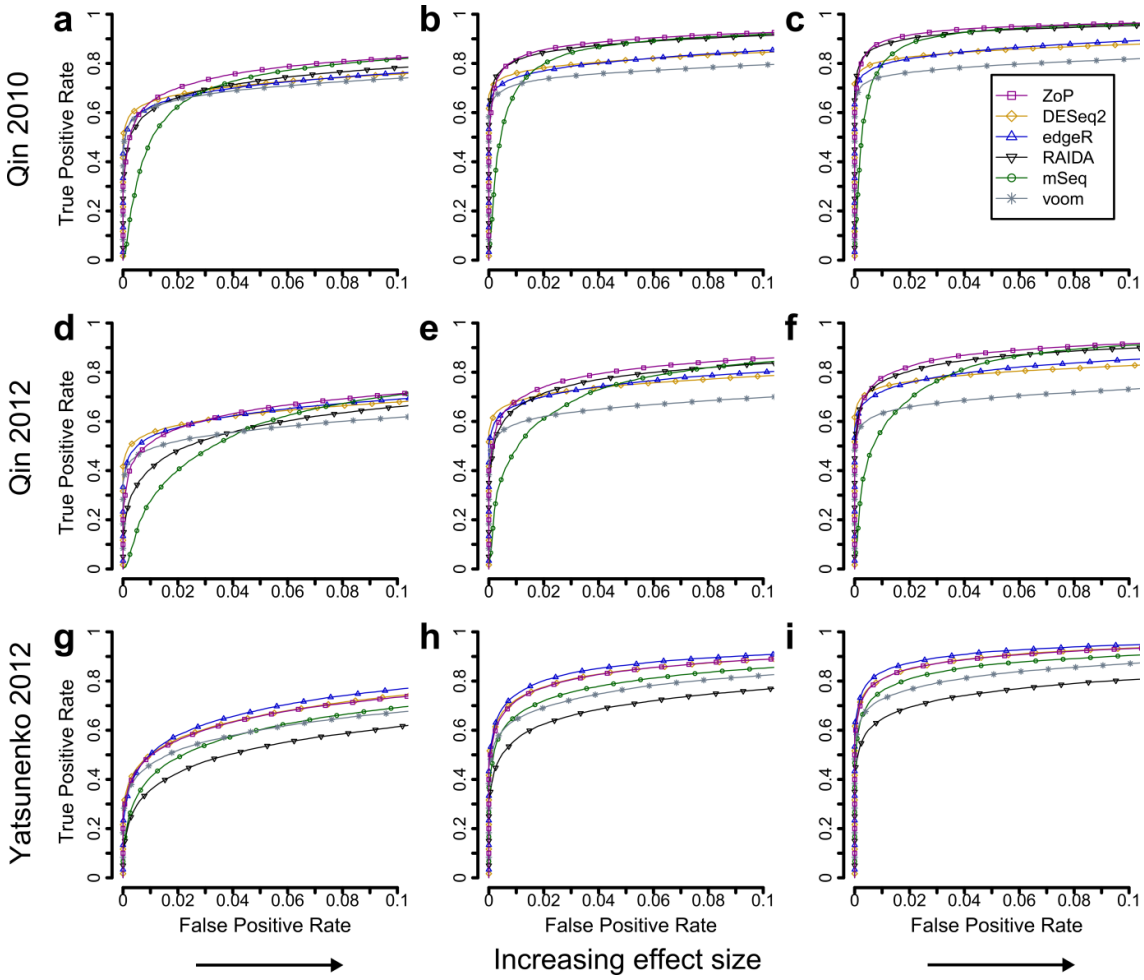


Supplementary Figure 1:

Impact of the prior distribution on p_i. Each ROC curve was generated from the ZoP model with different priors on the zero-inflation parameter p_i. The priors were Beta(1,1) (red circles), Beta(0.5,1.8) (orange triangles), Beta(0.2,1.5) (pink diamonds) and Beta(0.05,1.4) (purple stars). The results were generated from 100 realizations on resampled data, with a group size of 5 and an effect size of 3, from each of the three datasets, Qin 2010 (a), Qin 2012 (b) and Yatsunenko 2012 (c).

Supplementary Table 1:

Results from the enrichment analysis of zero-inflated genes in GO-terms in excel format. Sheets 1-3 correspond to the results for all GO-terms in each of the three datasets, Qin 2010, Qin 2012 and Yatsunenko.



Supplementary Figure 2:

The difference in performance between models is smaller at lower effect sizes. ROC curves showing the performance of the six included methods for increasing effect size (fold-change) across the three datasets. The effect size was set to 2, 3 and 4 and the group size fixed to 10. The ROC curves represent an average over 100 repetitions on resampled data. The included methods were, ZoP, DESeq2, edgeR, RAIDA, metagenomeSeq (mSeq) and voom.

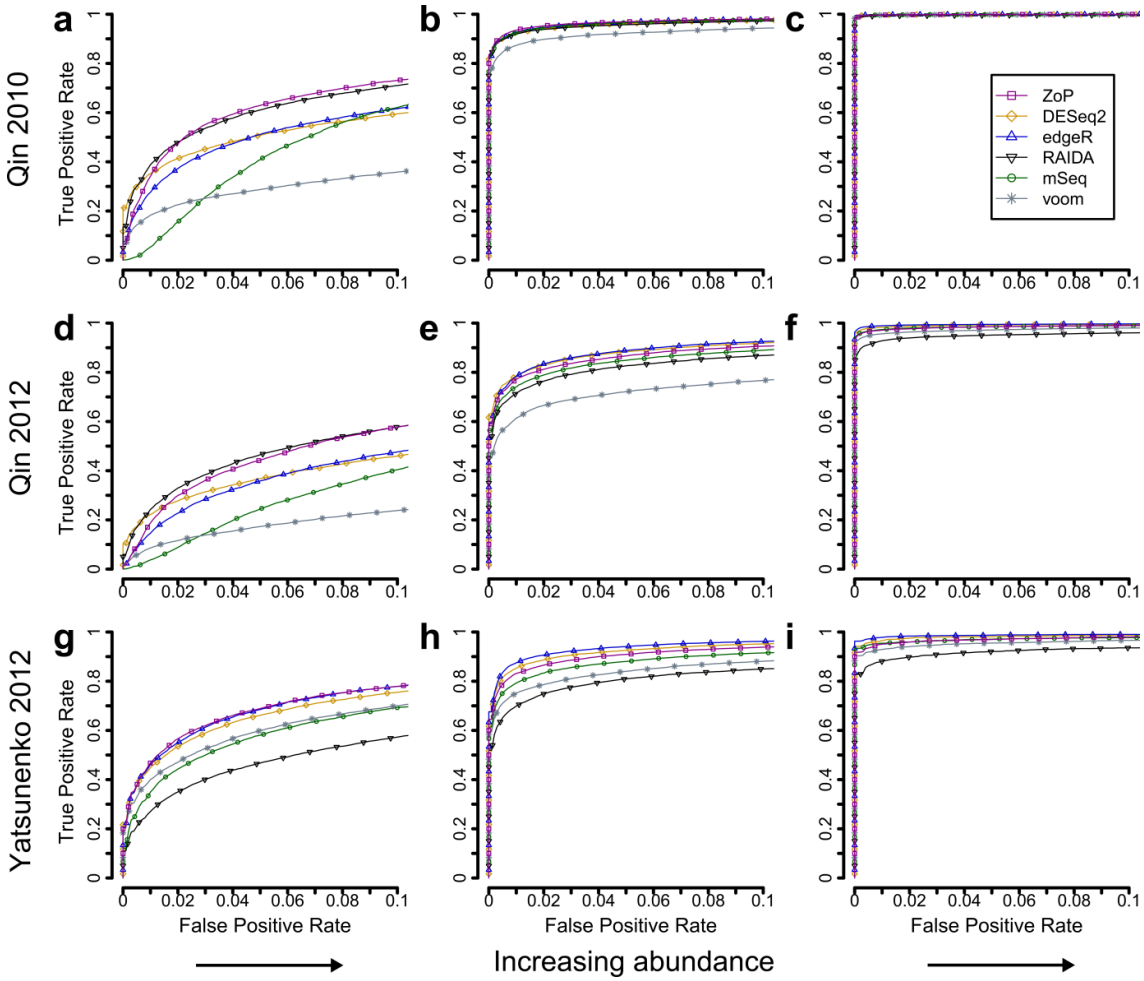
Supplementary Table 2

Supplementary table 2: $AUC_{0.1}$ results for increasing effect sizes on resampled data. The values correspond to the ROC curves in Supplementary Figure 2. Group size was fixed to 10. Values represent averages over 50 repetitions.

Dataset:	Qin 2010			Qin 2012			Yatsunenko		
Effect:	2	3	4	2	3	4	2	3	4
ZoP	0.75	0.88	0.93	0.63	0.78	0.86	0.63	0.82	0.88
DESeq2	0.71	0.81	0.85	0.62	0.74	0.79	0.63	0.82	0.88
edgeR	0.70	0.80	0.85	0.62	0.74	0.80	0.65	0.84	0.90
RAIDA	0.71	0.87	0.92	0.55	0.76	0.84	0.50	0.68	0.74
mSeq	0.69	0.83	0.89	0.53	0.70	0.78	0.57	0.77	0.84
voom	0.69	0.76	0.78	0.56	0.65	0.69	0.58	0.75	0.81

Supplementary Table 2:

$AUC_{0.1}$ results for increasing effect sizes on resampled data. The values correspond to the ROC curves in Supplementary Figure 2. Group size was fixed to 10. Values represent averages over 100 repetitions.



Supplementary Figure 3:

Zero-inflated models have the highest performance at low gene abundances. ROC curves showing the performance of the six included methods at different gene abundances across the three datasets. The gene abundance cut-offs in average observed counts were i) $\bar{y}<1000$, ii) $1000<\bar{y}<5000$ and iii) $\bar{y}>5000$ for Qin 2012 and Qin 2010 and i) y 10, ii) $10<\bar{y}<50$ and iii) $\bar{y}>50$ for Yatsunenko. The group and effect sizes were fixed to 10 and 3 respectively. The ROC curves represent an average over 100 repetitions on resampled data. The included methods were, ZoP, DESeq2, edgeR, RAIDA, metagenomeSeq (mSeq) and voom.

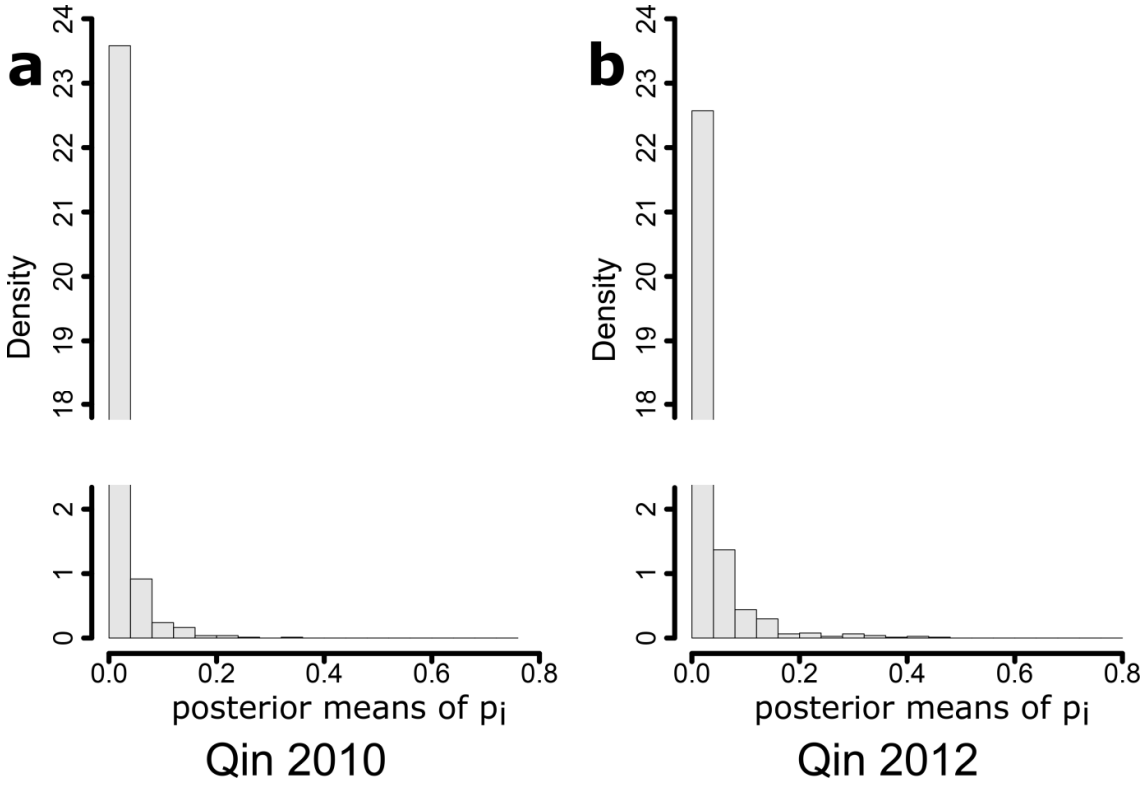
Supplementary Table 3

Supplementary table 3: $AUC_{0.1}$ results for increasing gene abundance on resampled data. The values correspond to the ROC curves in Supplementary Figure 3. Abundance cut-offs in terms of average count correspond to i) $\bar{y} < 1000$, ii) $1000 < \bar{y} < 5000$ and iii) $\bar{y} > 5000$ for Qin 2012 and Qin 2010 and i) $\bar{y} < 10$, ii) $10 < \bar{y} < 50$ and iii) $\bar{y} > 50$ for Yatsunenko. Group and effect sizes were fixed to 10 and 3 respectively. Values represent averages over 50 repetitions

Dataset:	Qin 2010			Qin 2012			Yatsunenko		
Abundance:	low	med	high	low	med	high	low	med	high
ZoP	0.58	0.96	1.00	0.41	0.85	0.98	0.65	0.89	0.97
DESeq2	0.49	0.95	1.00	0.35	0.86	0.99	0.63	0.91	0.98
edgeR	0.47	0.96	1.00	0.33	0.87	0.99	0.65	0.93	0.99
RAIDA	0.57	0.95	1.00	0.43	0.80	0.95	0.44	0.79	0.91
mSeq	0.37	0.96	1.00	0.22	0.83	0.98	0.54	0.86	0.97
voom	0.27	0.91	1.00	0.16	0.70	0.97	0.57	0.82	0.95

Supplementary Table 3

$AUC_{0.1}$ results for increasing gene abundance on resampled data. The values correspond to the ROC curves in Supplementary Figure 3. Abundance cut-offs in terms of average count correspond to i) $\bar{y}<1000$, ii) $1000<\bar{y}<5000$ and iii) $\bar{y}>5000$ for Qin 2012 and Qin 2010 and i) $\bar{y}<10$, ii) $10<\bar{y}<50$ and iii) $\bar{y}>50$ for Yatsunenko. Group and effect sizes were fixed to 10 and 3 respectively. Values represent averages over 100 repetitions.



Supplementary Figure 4

Posterior mean of the zero-inflation parameter p_i for each gene in downsampled versions of Qin 2010 dataset (panel a) and Qin 2012 (panel b) dataset. Counts were removed in each sample until the total counts corresponded to that of the median depth in the Yatsunenko dataset $6*10^4$. The y-axes have been cut as there is a large proportion of genes with almost no excess zeros in each dataset.

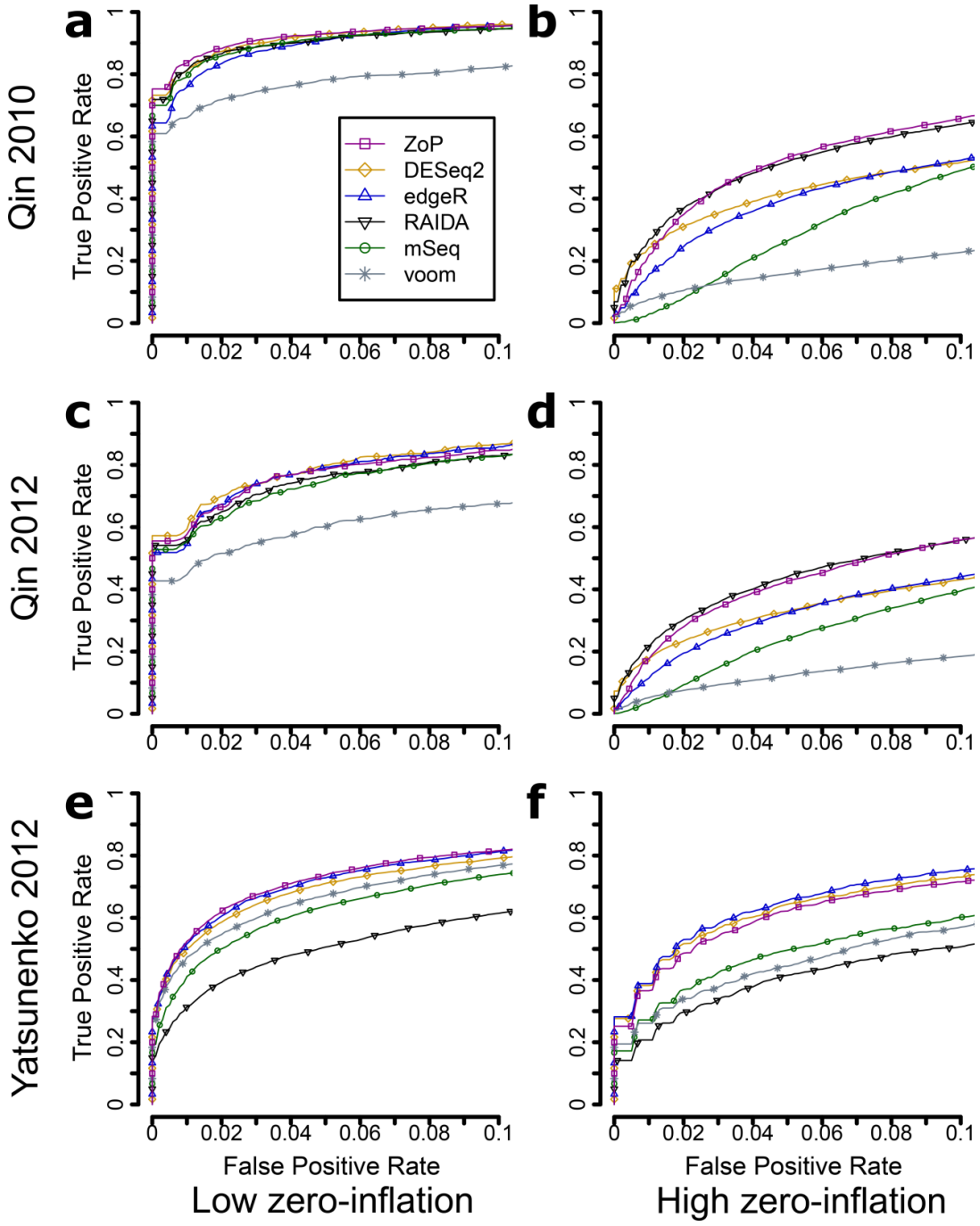
Supplementary Table 4

Supplementary table 4: $AUC_{0.1}$ results on low abundant genes with low and high zero-inflation. The values correspond to the ROC curves in Supplementary Figure 5. The low abundant genes from Figure 6 were split into low zero-inflation ($p_i <= 0.05$) and high zero-inflation ($p_i > 0.05$) based on estimates from the full data set. Group and effect sizes were fixed to 10 and 3 respectively. Values represent averages over 100 repetitions.

Dataset:	Qin 2010		Qin 2012		Yatsunenko 2010	
Zero-inflation:	low	high	low	high	low	high
ZoP	0.91	0.48	0.75	0.39	0.70	0.58
DESeq2	0.91	0.39	0.77	0.31	0.67	0.60
edgeR	0.88	0.36	0.76	0.30	0.69	0.62
RAIDA	0.89	0.48	0.73	0.41	0.48	0.38
mSeq	0.89	0.25	0.72	0.22	0.60	0.46
voom	0.76	0.15	0.58	0.12	0.64	0.43

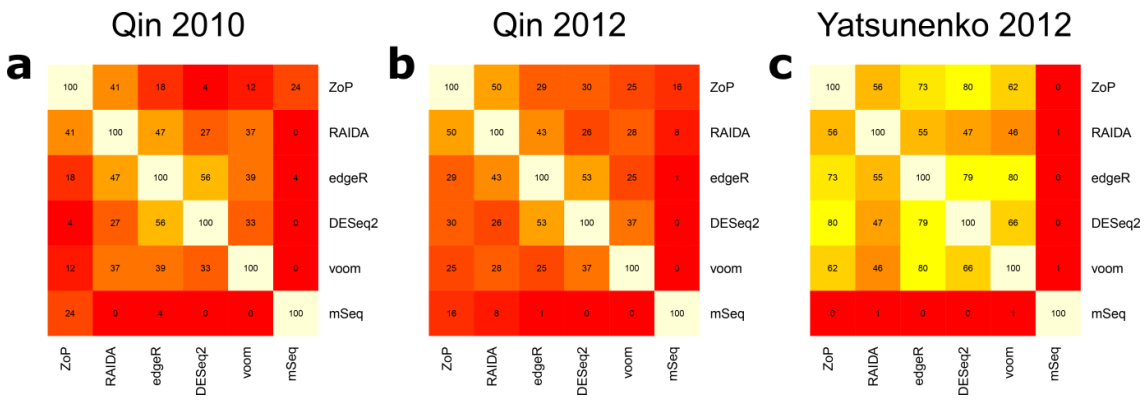
Supplementary Table 4

$AUC_{0.1}$ results on low abundant genes with low and high zero-inflation. The values correspond to the ROC curves in Supplementary Figure 5. The low abundant genes from Figure 6 were split into low zero-inflation ($p_i <= 0.05$) and high zero-inflation ($p_i > 0.05$) based on estimates from the full data set. Group and effect sizes were fixed to 10 and 3 respectively. Values represent averages over 100 repetitions.



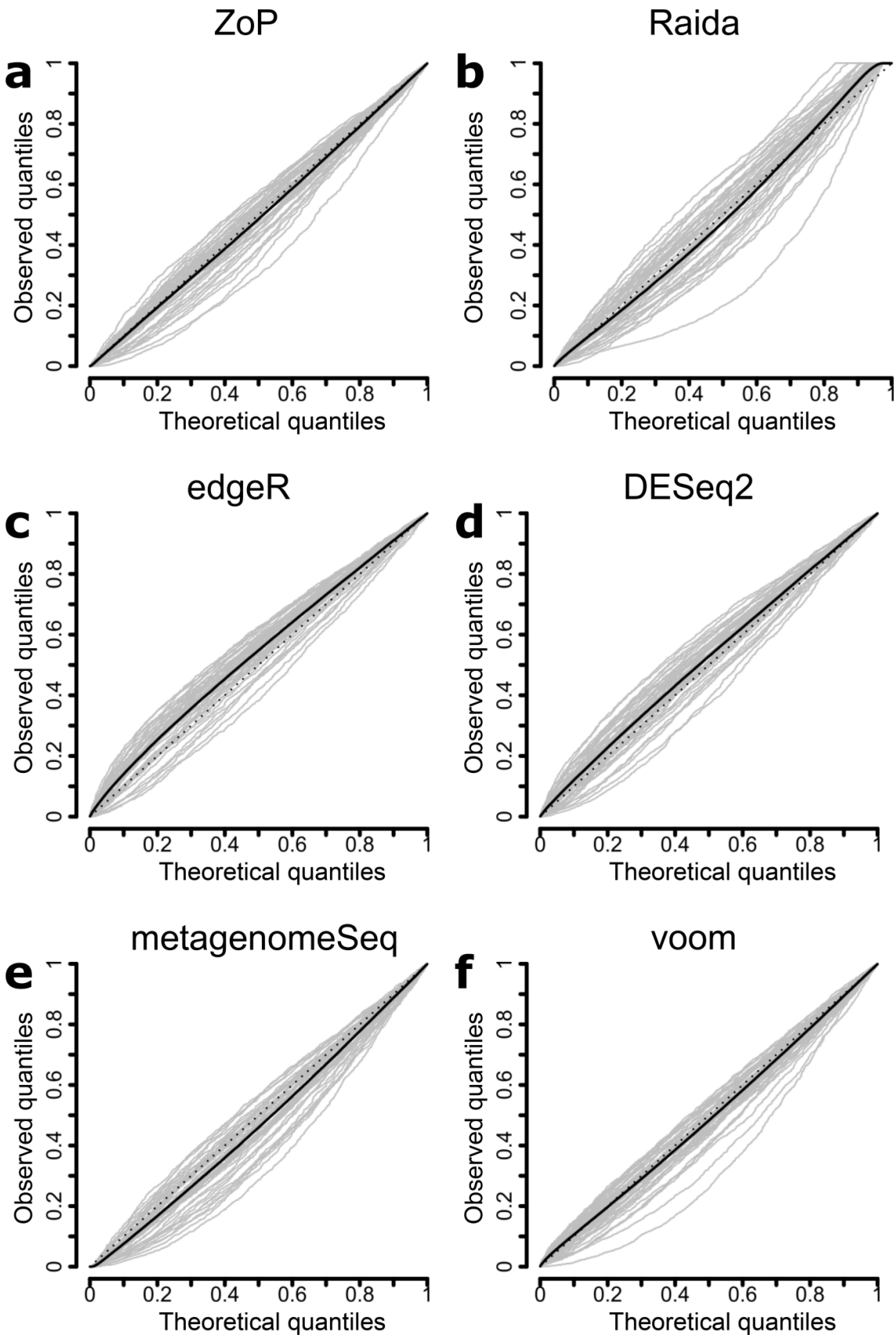
Supplementary Figure 5

Performance of the methods on low abundant genes with low and high zero-inflation. The low abundant genes from Figure 6 were split into low zero-inflation (p_i <= 0.05) and high zero-inflation (p_i > 0.05) based on estimates from the full data set. The number of genes with low abundance in the low and high zero-inflation categories were 194 and 702 for Qin 2010, 339 and 851 for Qin 2012 and, 622 and 203 for Yatsuneko 2012. The effect size was set to 3 and the group size fixed to 10. The ROC curves represent an average over 100 repetitions on resampled data. The included methods were, ZoP, DESeq2, edgeR, RAIDA, metagenomeSeq (mSeq) and voom.



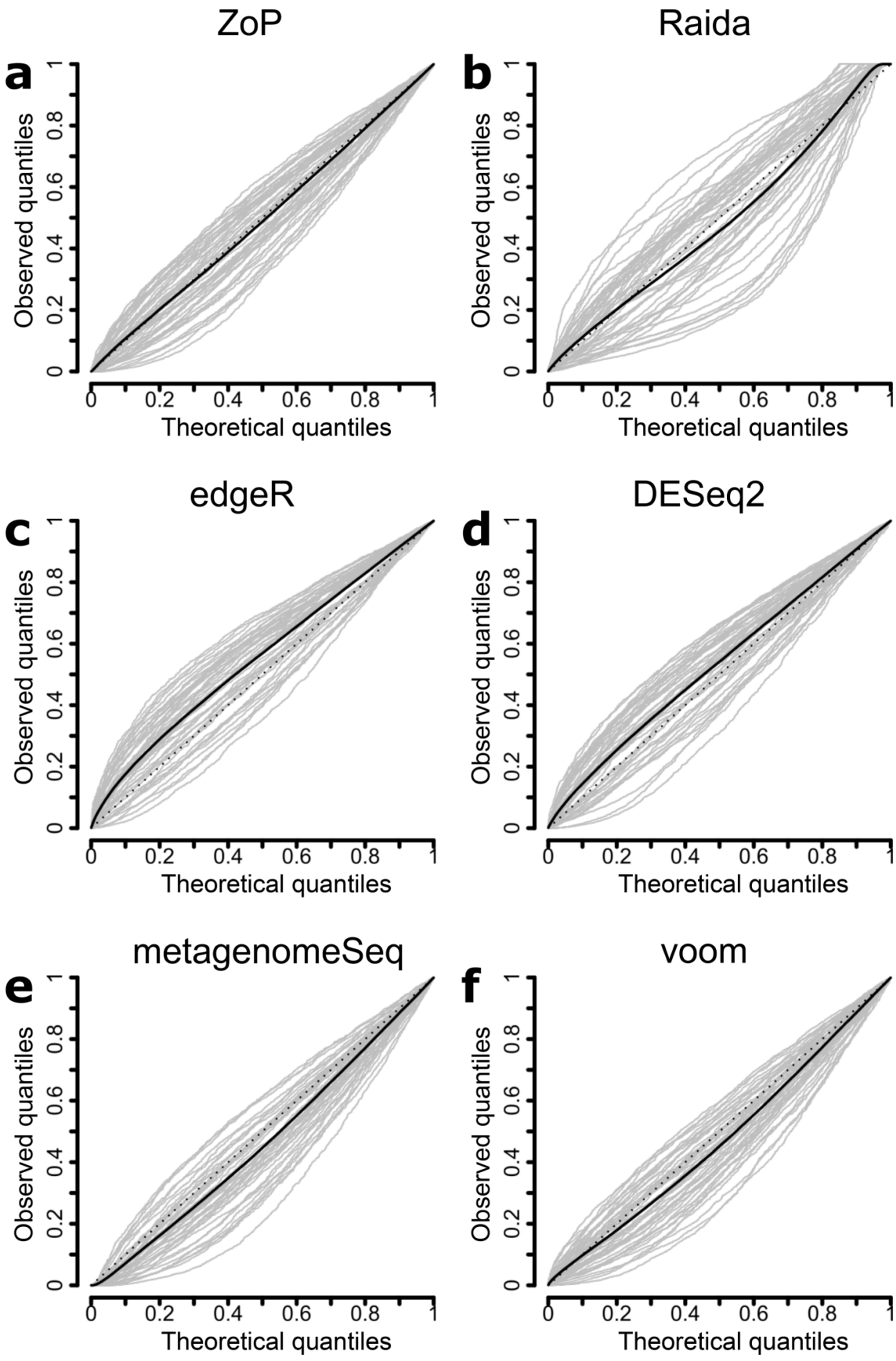
Supplementary Figure 6

Comparison of top ranked genes for each method and dataset. The numbers correspond to the number of overlapping genes among the top 100 most significant genes for each pair of methods. For each dataset 10 samples were randomly selected per condition to be compared. For Qin 2010 (panel a) the data consisted of 10 patients with inflammatory bowel disease (samples V1.UC-10, O2.UC-12, O2.UC-19, V1.CD-12, O2.UC-22, V1.CD-15, V1.UC-14, O2.UC-16, O2.UC-20, and V1.CD-6) vs 10 control patients (samples V1.CD-3, V1.CD-4, V1.CD-11, V1.UC-7, V1.CD-9, V1.UC-8, V1.UC-9, V1.CD-2, V1.UC-19, and V1.CD-8). For Qin 2012 (panel b) the data consisted of 10 randomly selected healthy lean women (samples NLF001, NLF005, NLF007, NLF008, NLF009, NLF010, NLF011, NLF012, NLF013 and, NLF014) vs 10 randomly selected diabetic lean women (DLF001, DLF003, DLF004, DLF005, DLF006, DLF009, DLF010, DLF012, DLF013 and, DLF014). For Yatsunenko 2012 (panel c) the data consisted of 10 breast-fed infants (samples USinfTw4.1, USinfTw4.2, USinfTw6.1, USinfTw6.2, USinfTw19.1, USinfTw19.2, USinfTw20.1, USinfTw20.2, USinfTw21.1, and USinfTw21.2) vs 10 formula fed infants (samples USinfTw2.1, USinfTw3.2, USinfTw5.1, USinfTw8.2, USinfTw9.2, USinfTw10.1, USinfTw12.2, USinfTw15.2, USinfTw16.1, and USinfTw17.1).



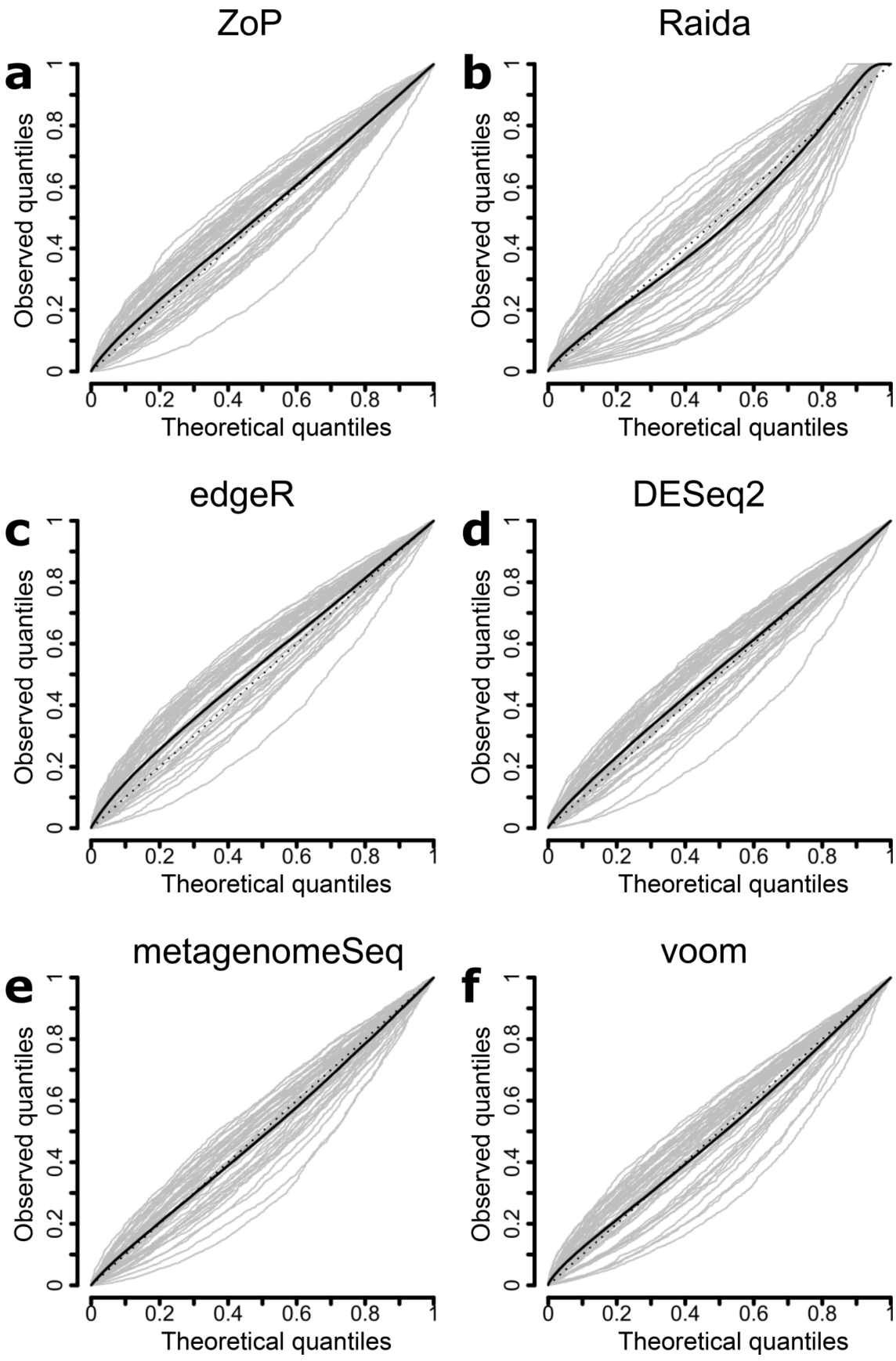
Supplementary Figure 7

Quantile-quantile (qq) plots of p-values under the null-hypothesis for each method for the Qin 2010 dataset, 10+10 samples. The p-values were calculated from resampled data generated as previously described but without adding any effects to the genes. The average qq-line, solid black, is calculated by averaging over 100 resampled datasets. The dashed line represents the theoretical average under a uniform p-value distribution. Each grey line represents a single realisation of the data.


a
ZoP
b
Raida
c
edgeR
d
DESeq2
e
metagenomeSeq
f
voom
Observed quantiles
Theoretical quantiles
0
0.2
0.4
0.6
0.8
1

Supplementary Figure 8

Quantile-quantile (qq) plots of p-values under the null-hypothesis for each method for the Qin 2012 dataset, 10+10 samples. The p-values were calculated from resampled data generated as previously described but without adding any effects to the genes. The average qq-line, solid black, is calculated by averaging over 100 resampled datasets. The dashed line represents the theoretical average under a uniform p-value distribution. Each grey line represents a single realisation of the data.


ZoP
Raida
edgeR
DESeq2
metagenomeSeq
voom
a
b
c
d
e
f
Observed quantiles
Theoretical quantiles
0
0.2
0.4
0.6
0.8
1

Supplementary Figure 9

Quantile-quantile (qq) plots of p-values under the null-hypothesis for each method for the Yatsunenko 2012 dataset, 10+10 samples. The p-values were calculated from resampled data generated as previously described but without adding any effects to the genes. The average qq-line, solid black, is calculated by averaging over 100 resampled datasets. The dashed line represents the theoretical average under a uniform p-value distribution. Each grey line represents a single realisation of the data.