1 APPENDIX

1.1 Nelder-Mead optimization

a. Initialize candidate $\{\pi_i^*\}_{i=1}^S$.

b. Estimate sample mean and variance, i.e., μ_{Tg}^* and σ_{Tg}^{*2} by

$$\mu_{Tg}^* = \frac{1}{S}\sum_{i=1}^{S} \log_2[\{y_{ig} - (1-\pi_i^*)n_{ig}\}/\pi_i^*],$$

and,

$$\sigma_{Tg}^{*2} = \frac{1}{(S-1)}\sum_{i=1}^{S} \left(\log_2[\{y_{ig} - (1-\pi_i^*)n_{ig}\}/\pi_i^*] - \mu_{Tg}^*\right)^2.$$

c. With $y_{ig}, \hat{\mu}_{Ng}, \hat{\sigma}_{Ng}^2, \mu_{Tg}^*, \sigma_{Tg}^{*2}, \pi_i^*$, we evaluate the likelihood value of observing the mixed sample expression Y, i.e., $\prod_i^S \prod_g^G f_{Y_{ig}}(y_{ig})$ as

$$\begin{aligned} f_{Y_{ig}}(y) &\propto \int_0^y \frac{1}{(y-t')\sqrt{\hat{\sigma}_{Ng}^2}} \\ &\times \exp\left[-\frac{\{\log_2(y-t') - (\log_2(1-\pi_i^*) + \hat{\mu}_{Ng})\}^2}{2\hat{\sigma}_{Ng}^2}\right] \\ &\times \frac{1}{t'\sqrt{\sigma_{Tg}^{*2}}} \exp\left[-\frac{\{\log_2 t' - (\log_2 \pi_i^* + \mu_{Tg}^*)\}^2}{2\sigma_{Tg}^{*2}}\right] dt'. \end{aligned}$$

Here, we note that mixed sample expression Y does not follow $\log_2 Normal$ distribution, as opposed to the $Normal$ distribution with log transformed data. As we do not have a closed form for the density of Y_{ig}, we approximate it by a numerical integration.

d. Obtain the next candidate $\{\pi_i^*\}_{i=1}^S$ following the Nelder-Mead rule.

e. If changes in the current π_i^* compared to the previous π_i^* are less than 1%, stop and return to $\pi_i^*, i = 1, \ldots, S$ otherwise, go back to step b.

1.2 Linear model (LM) for log-transformed data

We use Y', N', and T' for the $\log_2$-transformed data; these values correspond to Y, N, and T for the raw-measured data. We follow a similar procedure as in DeMix to estimate the corresponding parameters, which maximize the likelihood of observing the mixed sample expression Y', i.e., $\prod_i^S \prod_g^G f_{Y'_{ig}}(y'_{ig})$ as:

$$\begin{aligned} f_{Y'_{ig}}(y') &= \frac{1}{\sqrt{2\pi(\pi_i^2\hat{\sigma}_{T'g}^2 + (1-\pi_i)^2\hat{\sigma}_{N'g}^2)}} \\ &\times \exp\left[-\frac{\{\pi_i\hat{\mu}_{T'g} + (1-\pi_i)\hat{\mu}_{N'g} - y'_{ig}\}^2}{2\{\pi_i^2\hat{\sigma}_{T'g}^2 + (1-\pi_i)^2\hat{\sigma}_{N'g}^2\}}\right]. \end{aligned}$$

We further deconvolve y'_{ig} into n'_{ig} and t'_{ig} by searching for values of t'_{ig} that maximizes the following function:

$$\text{argmax}_{t'_{ig}}\, \phi'(t'_{ig}|\hat{\mu}_{T'g}, \hat{\sigma}_{T'g}^2)\phi'\Big(\frac{y'_{ig} - \hat{\pi}_i t'_{ig}}{1-\hat{\pi}_i}\Big|\hat{\mu}_{N'g}, \hat{\sigma}_{N'g}^2\Big),$$

where $\phi'(\cdot|\mu, \sigma^2)$ is a normal density with mean μ and variance σ^2.

2 SUPPLEMENTAL TABLES

Table 1. The estimated proportions along with their 95% confidence intervals (CI) according to the four datasets in the main text.

			DeMix		LM	
GSE19830	Sample	Brain	Est. (%)	95% CI	Est. (%)	95% CI
	1	70	70	(62, 79)	99	(98, 100)
	2	70	72	(65, 78)	99	(98, 100)
	3	70	71	(65, 78)	99	(98, 100)
	4	25	22	(11, 32)	41	(38, 44)
	5	25	21	(13,29)	44	(41, 47)
	6	25	21	(11, 30)	42	(40, 44)
	7	35	32	(27, 37)	52	(49, 55)
	8	35	31	(26, 34)	53	(50, 56)
	9	35	33	(29, 36)	52	(50, 54)
	10	34	31	(26, 35)	52	(50, 54)
	11	34	31	(25, 36)	51	(49, 53)
	12	34	31	(27, 34)	52	(49, 55)
MAQC	Sample	Brain	Est. (%)	95% CI	Est. (%)	95% CI
MAQC 1	1	25	28	(22, 33)	39	(38, 40)
	2	25	30	(24, 36)	43	(41, 44)
	3	25	32	(26, 37)	45	(44, 47)
	4	25	31	(22, 41)	45	(44, 46)
	5	25	28	(24, 33)	40	(39, 42)
	6	75	73	(68, 78)	99	(98, 100)
	7	75	71	(66, 77)	99	(97, 100)
	8	75	72	(66, 78)	98	(97, 99)
	9	75	72	(66, 78)	99	(98, 100)
	10	75	75	(69, 81)	99	(98, 100)
MAQC 3	1	25	34	(32, 37)	54	(51, 57)
	2	25	26	(25, 28)	35	(34, 36)
	3	25	33	(31, 36)	50	(48, 52)
	4	25	26	(25, 27)	37	(35, 38)
	5	25	25	(23, 26)	32	(31, 34)
	6	75	70	(68, 73)	98	(97, 98)
	7	75	76	(74, 79)	99	(98, 100)
	8	75	71	(67, 74)	100	(99, 100)
	9	75	70	(66, 73)	100	(98, 100)
	10	75	77	(74, 80)	98	(97, 99)
AFFY	Sample	Brain	Est. (%)	95% CI	Est. (%)	95% CI
AFFY	1	25	36	(31, 42)	52	(48, 57)
	2	25	36	(30, 41)	54	(48, 61)
	3	25	36	(30, 42)	53	(48, 57)
	4	50	53	(46, 59)	82	(74, 91)
	5	50	51	(45, 58)	75	(68, 83)
	6	50	47	(42, 51)	70	(62, 78)
	7	50	48	(42, 53)	75	(65, 84)
	8	50	54	(50, 57)	76	(70, 82)
	9	50	50	(45, 54)	74	(65, 83)
	10	50	53	(47, 59)	81	(74, 87)
	11	50	59	(54, 64)	87	(78, 97)
	12	50	54	(48, 6)	81	(73, 89)
	13	75	72	(68, 76)	95	(87, 100)
	14	75	72	(66, 78)	95	(89, 100)
	15	75	74	(70, 78)	97	(93, 100)

DeMix : Deconvolution model using raw measured data
LM : Deconvolution model using log-transformed data

Table 2. We consider one gene with the raw measured intensity values of $T = 1,024$ and $N = 32$, with varying π. The $\log 2$-transformed measures are denoted by T' and N', respectively. The true proportion is denoted by π, and the LM-based estimate is denoted by π'.

T	N	T'	N'	Y'	$\pi' = \frac{Y'-N'}{T'-N'}$	π	$\pi' - \pi$
1,024	32	10	5	7.84	0.64	0.2	0.44
1,024	32	10	5	8.36	0.72	0.3	0.42
1,024	32	10	5	8.74	0.79	0.4	0.39
1,024	32	10	5	9.04	0.84	0.5	0.34
1,024	32	10	5	9.29	0.88	0.6	0.28
1,024	32	10	5	9.50	0.91	0.7	0.21
1,024	32	10	5	9.68	0.94	0.8	0.14

Table 3. We consider one gene with the raw measured intensity values of $T = 32$ and $N = 1,024$ with varying π. The $\log 2$-transformed measures are denoted by T' and N', respectively. The true proportion is denoted by π, and the LM-based estimate is denoted by π'.

N	T	N'	T'	Y'	$\pi' = \frac{Y'-N'}{T'-N'}$	π	$\pi' - \pi$
1,024	32	10	5	9.68	0.05	0.2	-0.15
1,024	32	10	5	9.50	0.08	0.3	-0.22
1,024	32	10	5	9.29	0.12	0.4	-0.28
1,024	32	10	5	9.04	0.16	0.5	-0.34
1,024	32	10	5	8.74	0.21	0.6	-0.39
1,024	32	10	5	8.36	0.27	0.7	-0.43
1,024	32	10	5	7.84	0.36	0.8	-0.44

3 SUPPLEMENTAL FIGURES

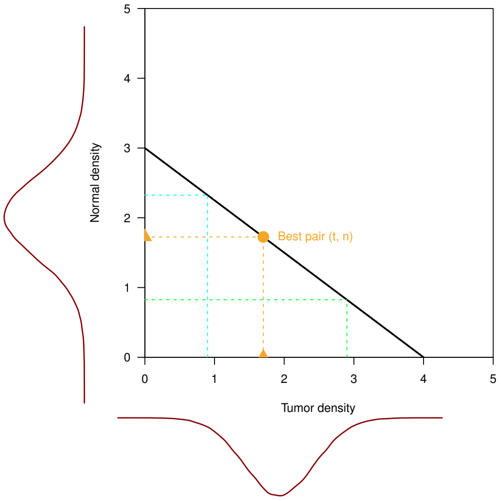


Fig. 1. Geometric interpretation of formula (2). The black solid line represents a linear equation: $y_{ig} = \hat{\pi}_i t_{ig} + (1 - \hat{\pi}_i) n_{ig}$. The $\log_2$ Normal distributions of T and N are next to the x-axis and y-axis. The intersection of the yellow and black lines corresponds to the best pair (t, n) that maximizes the product of the probabilities of observing each value. The blue and green lines correspond to less likely pairs of (t, n).

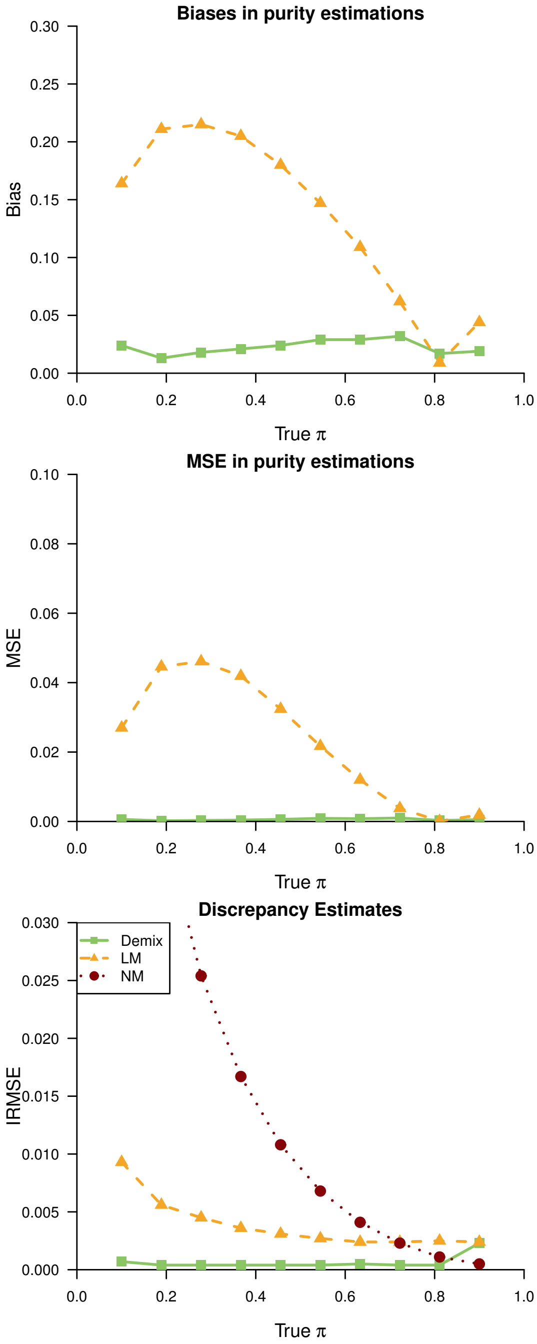


Fig. 2. Simulation results for data scenario 2. We assumed there is no reference gene among 2,000 genes, and generated data from 10 type A and type B matched samples.

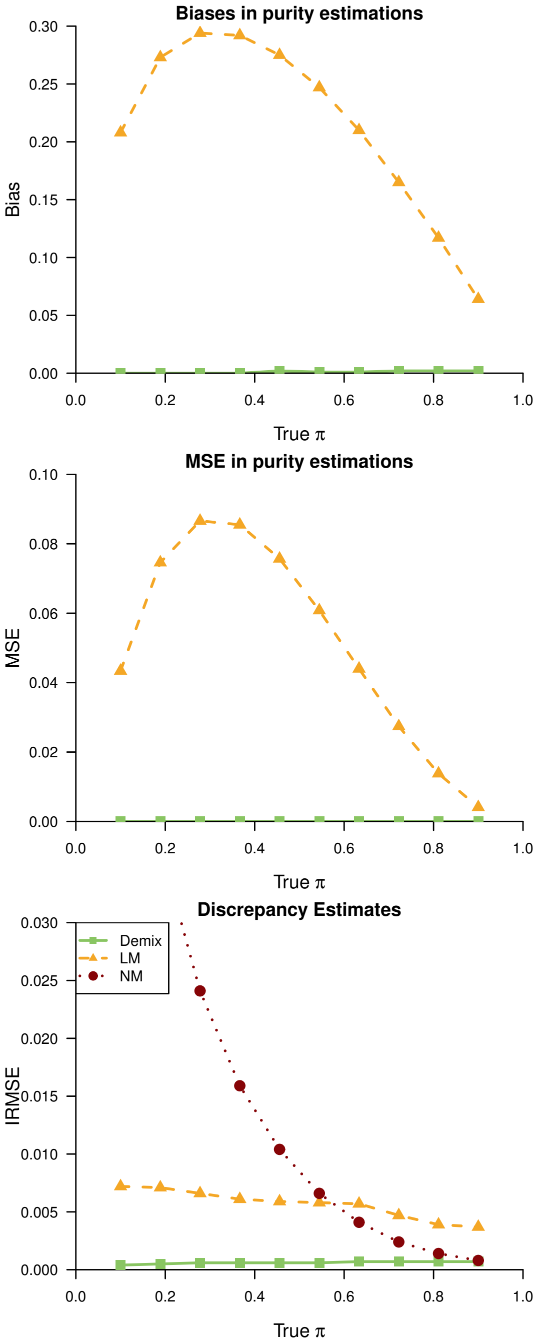


Fig. 3. Simulation results for data scenario 3. We assumed there are 50 reference genes among 2,000 genes, and generated data from 10 type A and type B unmatched samples.

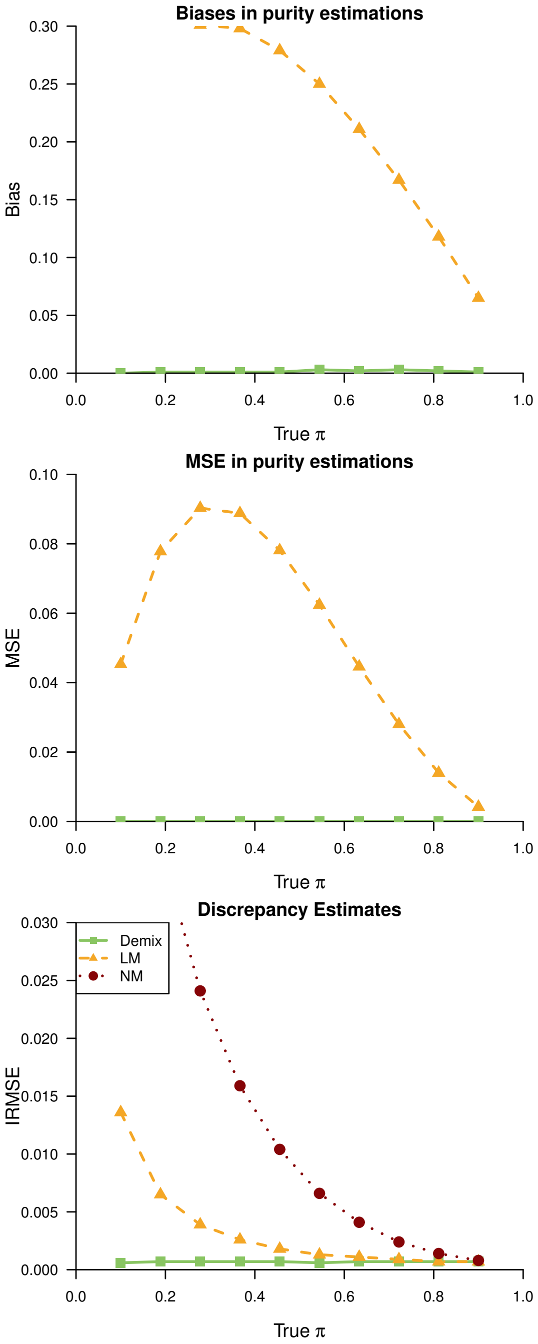


Fig. 4. Simulation results for data scenario 4. We assumed there are 50 reference genes among 2,000 genes, and generated data from 10 type A and type B matched samples.

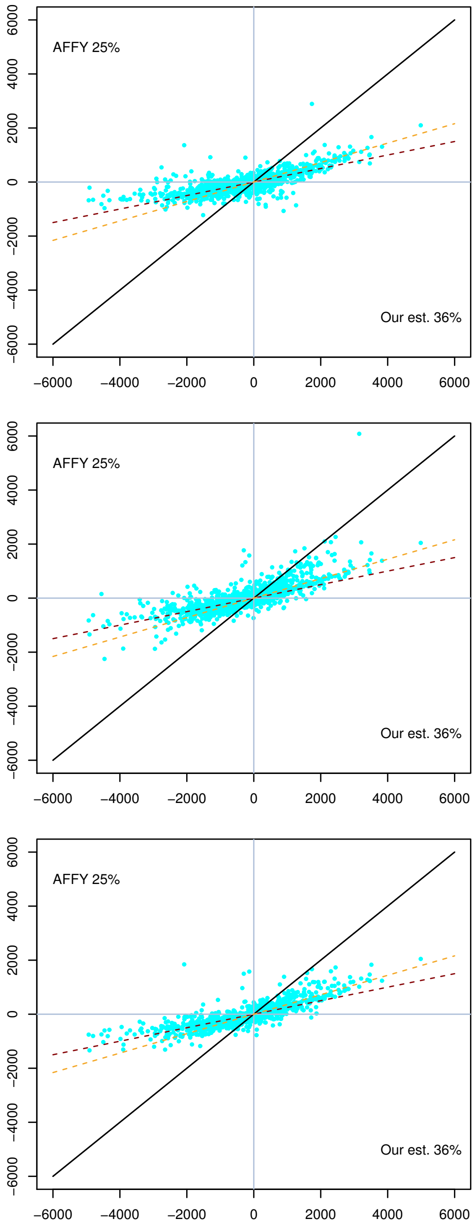
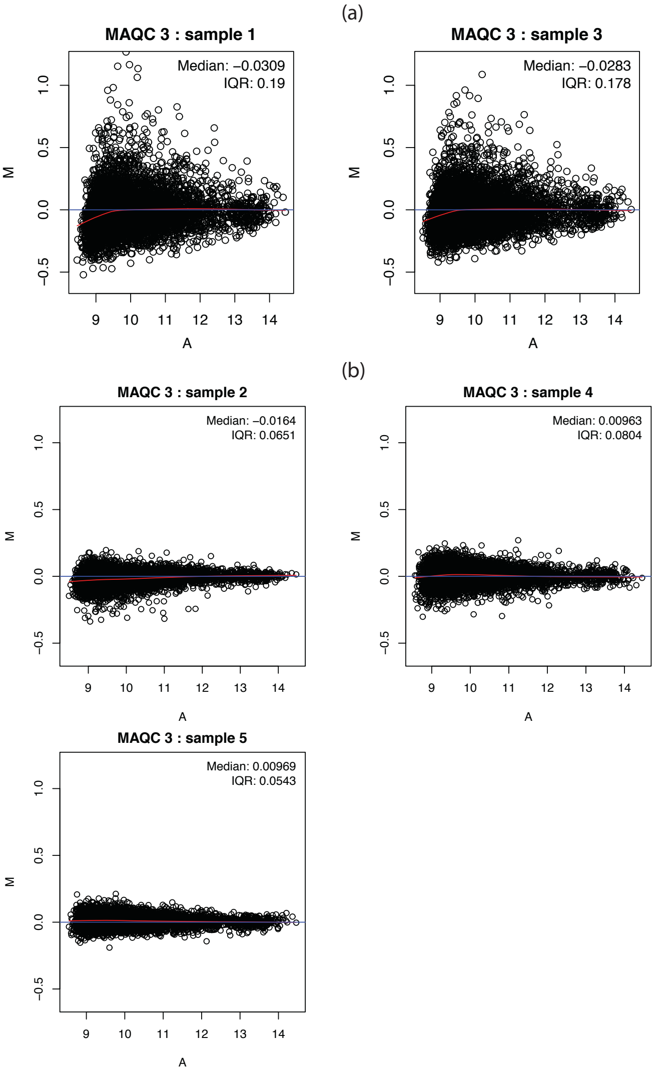


Fig. 5. Scatter plots of transcript abundance in raw measured cell type A expressions - cell type B expressions vs mixed sample expressions - cell type B expressions in the Affymetrix dataset for samples with π equal to 25% (as is shown in the red dashed lines). Our estimates of π's are shown in the yellow solid lines.

Fig. 6. MA plots of five samples in MAQC 3 at the 25% ratio using an average of sample 2, 4, 5 as a reference.



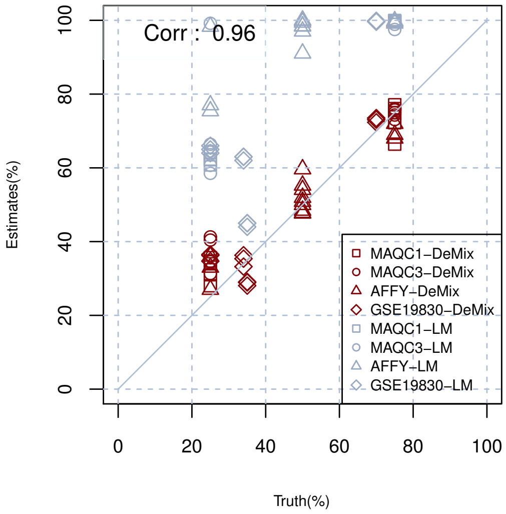
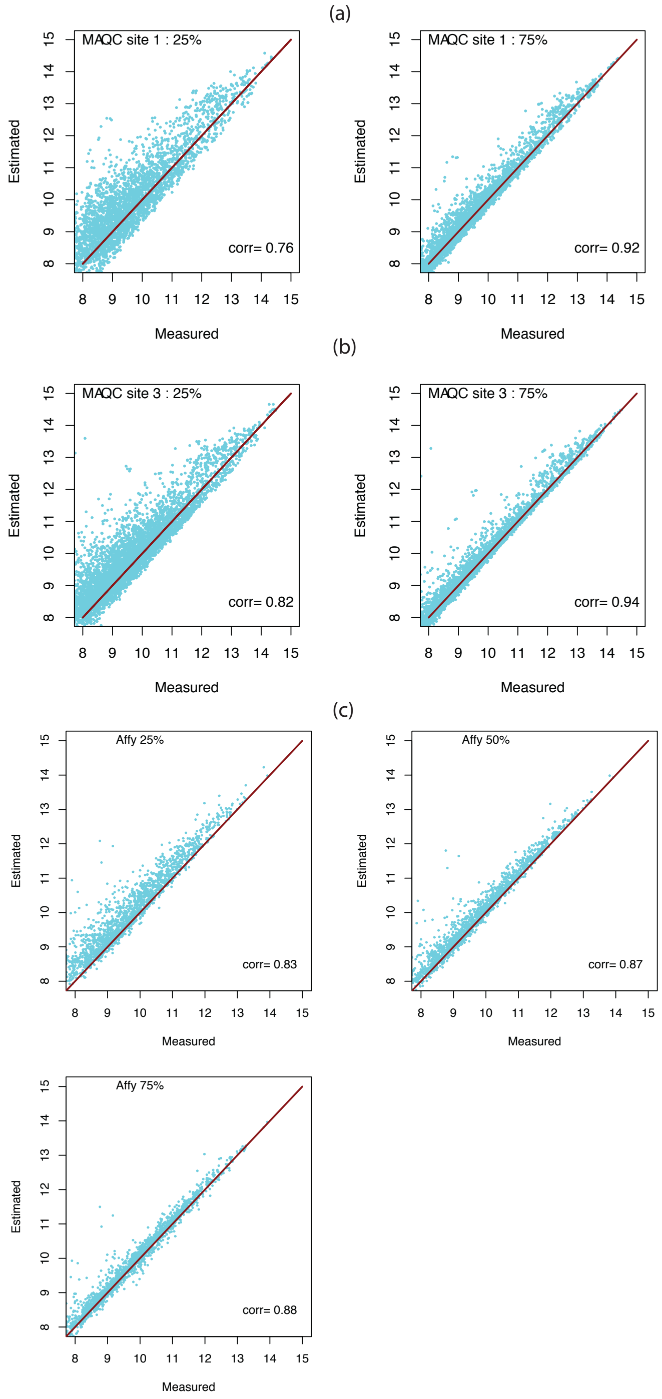


Fig. 7. We analyzed the four sets of validation data using all genes with expression levels above 2^6 in the normal samples. We used 19,268 (35%) probesets for MAQC1, 26,770 (49%) probesets for MAQC3, 13892 (43%) probesets for Affymetrix, and 14,369 (46%) probesets for GSE19830.

Fig. 8. Scatter plots of mean pure expressions vs deconvolved expressions corresponding to (a) cell type B from MAQC1 (b) cell type B from MAQC3 (c) heart tissue from Affymetrix.


(a)

(b)

(c)

Fig. 9. Performance of DeMix and PERT on the four validation data sets. The list of reference genes consists of 10 randomly selected DE genes among genes that yield the mean difference between two tissue types greater than 2 ($\log_2$ transformed). We iterated the random sampling 100 times.

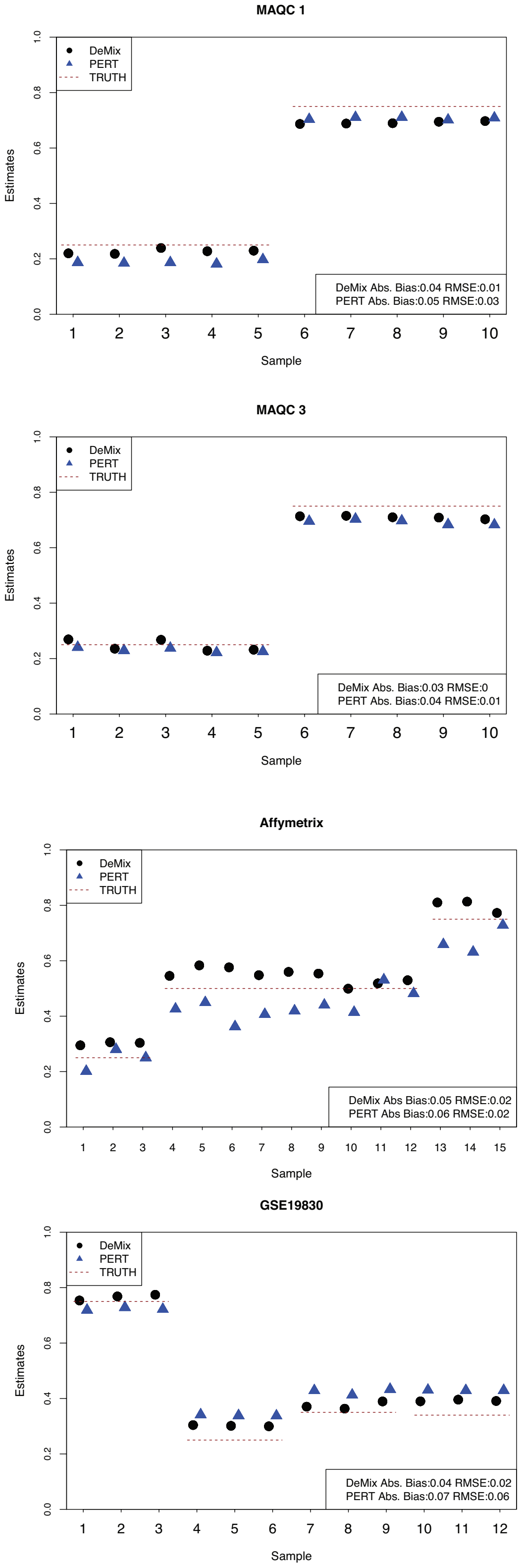
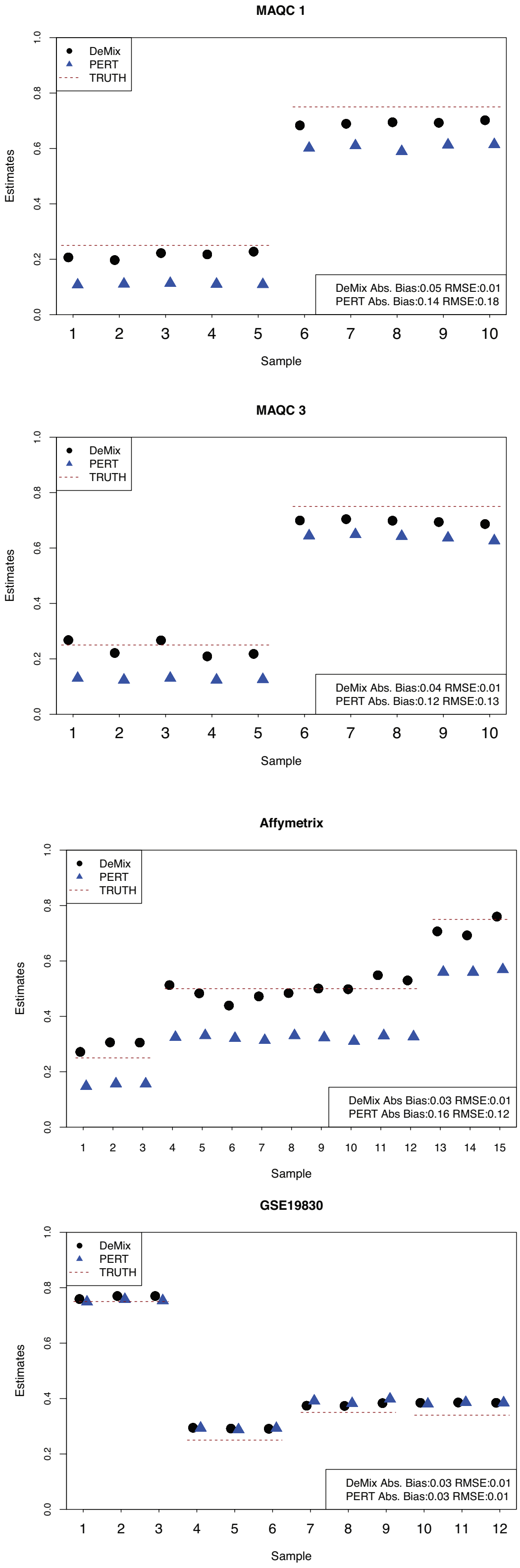


Fig. 10. Performance of DeMix and PERT on the four validation data sets. The list of reference genes consists of 1000 randomly selected genes among genes that yield the mean difference between two tissue types greater than 2 ($\log_2$ transformed). We iterated the random sampling 100 times.



REFERENCES

Erkkilä, T., Lehmusvaara, S., Ruusuvuori, P. Visakorpi, Tapio., Shmulevich, I., and Lähdesmäki, H. (2010) Probabilistic analysis of gene expression measurements from heterogeneous tissues, *Bioinformatics*, **26**, 2571-2577.

Qiao W, Quon G, Csaszar E, Yu M, Morris Q, et al. (2012). PERT: A method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput Biol*, **8**, e1002838. doi:10.1371/journal.pcbi.1002838