1 Marginal targeted power from simulation results under nominal FDR 0.2

Table S1 compares the power-related quantities when DE genes are declared with nominal FDR 0.2. Compared with Table 2 in the manuscript, the marginal powers are higher because a more liberal cutoff is used. However there are significantly more false positives, and as a result the FDC is higher.

Table S1: Marginal targeted power analysis results from simulations when DE genes are declared with nominal FDR 0.2.

(A) *Cheung data*

N	FDRn	FDRa	power	$\bar{n}_{TD}$	$\bar{n}_{FD}$	FDC
3	0.20	0.51	0.33	128.71	138.95	1.08
5	0.20	0.36	0.45	185.51	106.26	0.57
7	0.20	0.28	0.54	223.07	87.90	0.39
10	0.20	0.21	0.62	261.30	72.00	0.28

(B) *Bottomly data*

N	FDRn	FDRa	power	$\bar{n}_{TD}$	$\bar{n}_{FD}$	FDC
3	0.20	0.28	0.66	434.55	175.62	0.40
5	0.20	0.20	0.75	501.53	127.68	0.25
7	0.20	0.16	0.78	532.01	107.97	0.20
10	0.20	0.13	0.82	558.18	92.63	0.17

2 Number of all genes and DE genes stratified by mean counts

Figure S1 shows the Distribution of all genes and DE genes in all strata defined by mean counts. White bars represent all genes, and blue bars represent the DE genes. The number of genes in the first stratum (with average counts between 0 and 10) contain much more genes than other strata.

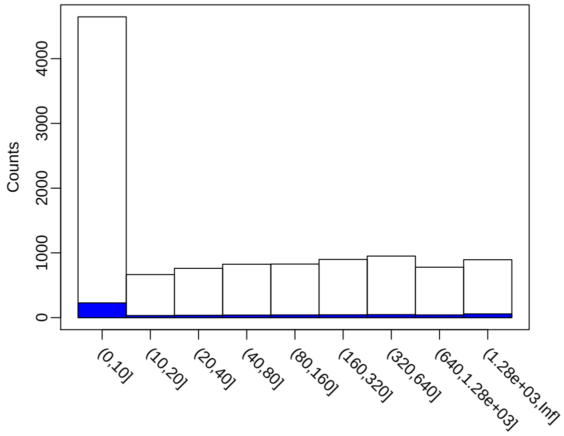

Figure S1: Distribution of all genes and DE genes in all strata

3 Results from *Bottomly data*

Figure S2 shows the power analysis results for *Bottomly data*. The *Bottomly data* have smaller biological variations compared with *Cheung data*. So under the same sample size, effect size, and nominal FDR, the DE detection is easier for *Bottomly data*: higher power, more true discoveries, and lower false discovery cost.

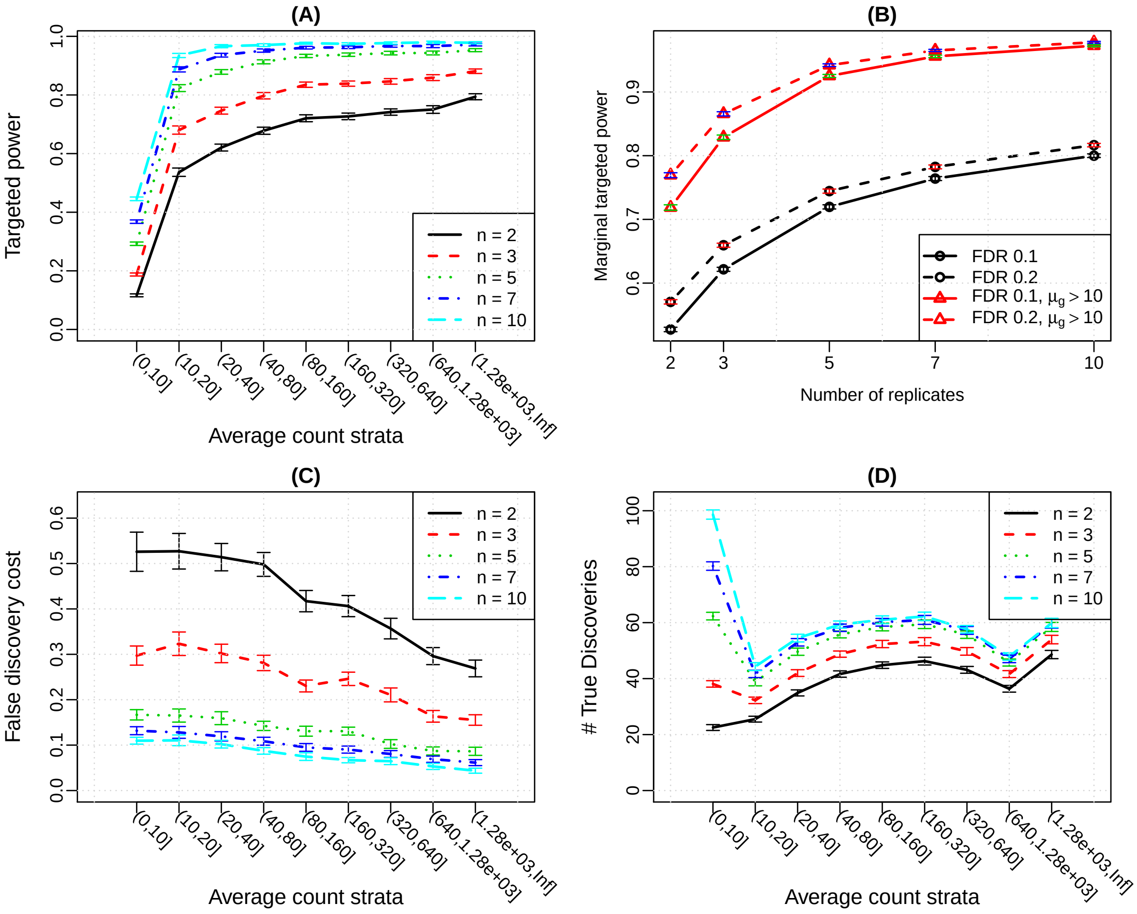

Figure S2: Power analyses results for *Bottomly* data. (A) Targeted power stratified by mean counts; (B) Marginal targeted power versus sample size, with and without filtering out genes with small counts. (C) False discovery cost stratified by mean counts; (D) Number of true discoveries stratified by mean counts.

4 Results with higher Δ

Δ is a threshold for log fold changes for defining biologically meaningful DE genes. These genes are deemed true positives and used to compute targeted powers. All results presented in the manuscript are based on $\Delta = 0.5$. This section provides a set of results using $\Delta = 1$.

Table S2 shows the comparison of power-related quantities when DE genes are declared with $\Delta = 1$ and nominal FDR 0.1. Compared with Table 2 in the manuscript, the marginal targeted powers are higher because the DE genes have greater effect sizes. However, the number of true discoveries decreases and the FDC is higher under this more stringent criteria for defining true DE.

Table S2: Marginal targeted power analysis results from simulations when DE genes are declared with $\Delta = 1$, and with nominal FDR 0.1.

(A) *Cheung data*

N	FDRn	FDRa	power	$\bar{n}_{TD}$	$\bar{n}_{FD}$	FDC
3	0.10	0.48	0.38	104.27	100.75	0.97
5	0.10	0.31	0.56	158.91	73.73	0.46
7	0.10	0.22	0.66	188.93	58.19	0.31
10	0.10	0.15	0.72	209.41	45.06	0.22

(B) *Bottomly data*

N	FDRn	FDRa	power	$\bar{n}_{TD}$	$\bar{n}_{FD}$	FDC
3	0.10	0.24	0.75	345.67	130.35	0.38
5	0.10	0.15	0.80	373.41	85.16	0.23
7	0.10	0.11	0.83	388.49	67.61	0.17
10	0.10	0.08	0.85	402.21	53.71	0.13

Figure S3 and S4 show the power analysis results for *Cheung data* and *Bottomly data*, respectively. Compared with the results from using $\Delta = 0.5$, the powers are considerably higher, whereas the numbers of true discoveries are lower.

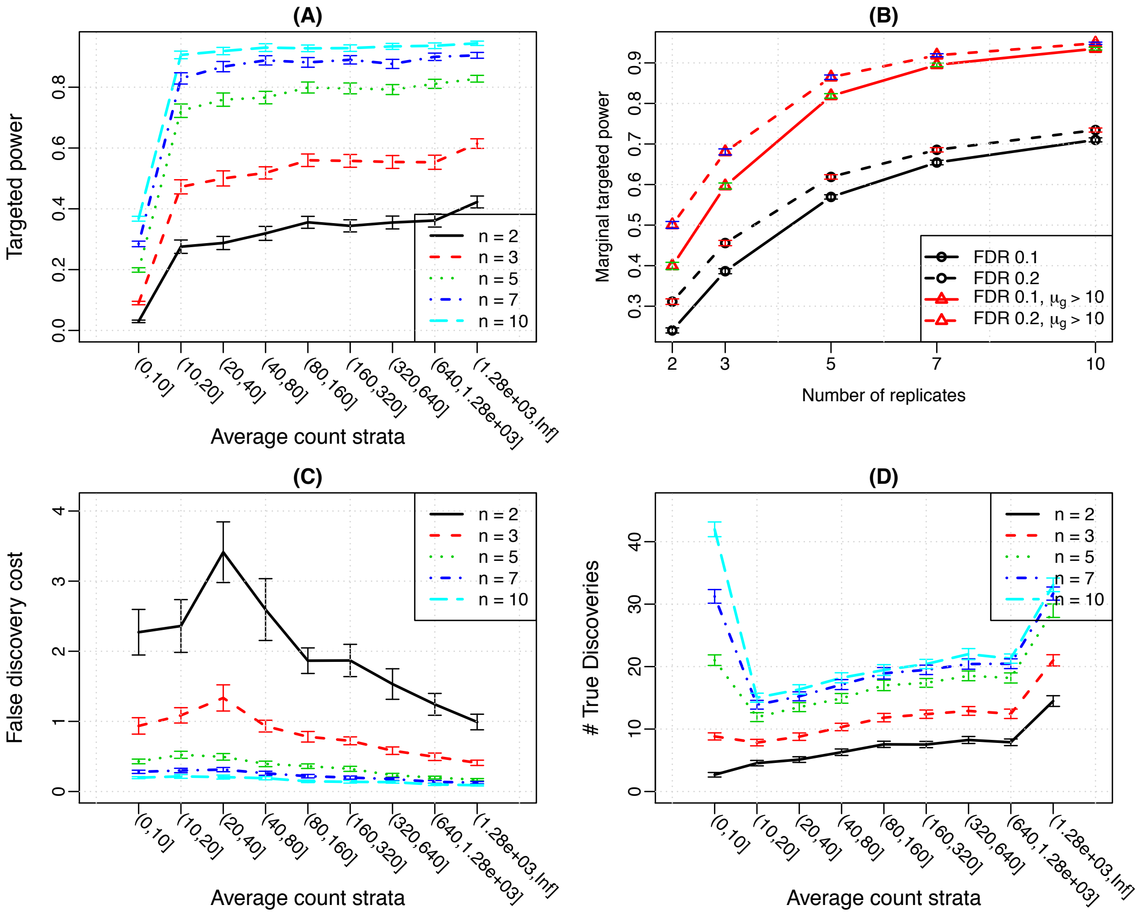

Figure S3: Power analyses results for *Cheung data*, using $\Delta = 1$ to define DE genes.

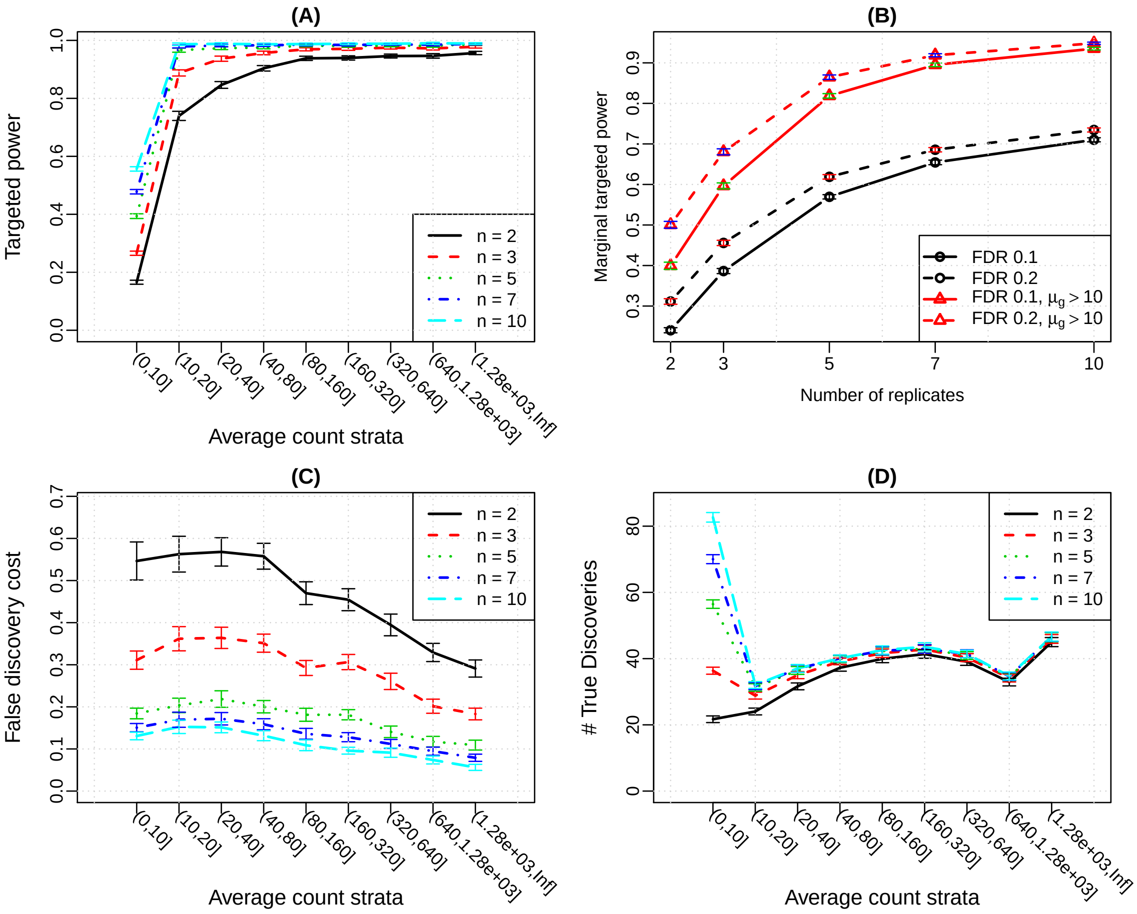

Figure S4: Power analyses results for *Bottomly* data, using $\Delta = 1$ to define DE genes.

5 Stratified by dispersion

Figure S5 and S6 show the power analysis results for *Cheung data* and *Bottomly data*, respectively, stratified by the over-dispersion of the genes. The results show that in general, the targeted powers are lower and FDCs are higher for genes with greater dispersion. These imply that genes with greater dispersion (or larger biological variations) are more difficult to detect. They also show that larger sample sizes result in higher powers and lower FDCs, which are consistent with the results when stratifying by mean counts.

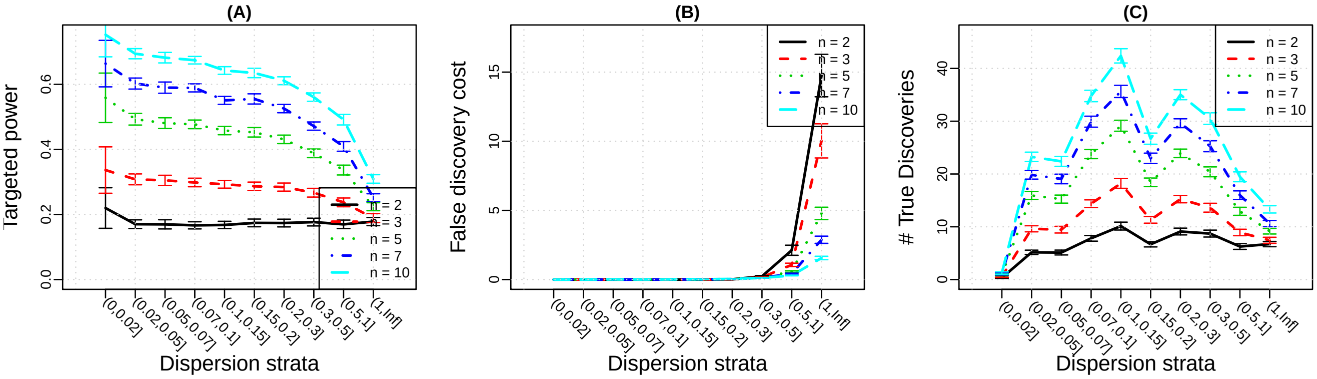

Figure S5: Power analyses results for *Cheung data*, stratified by dispersion.

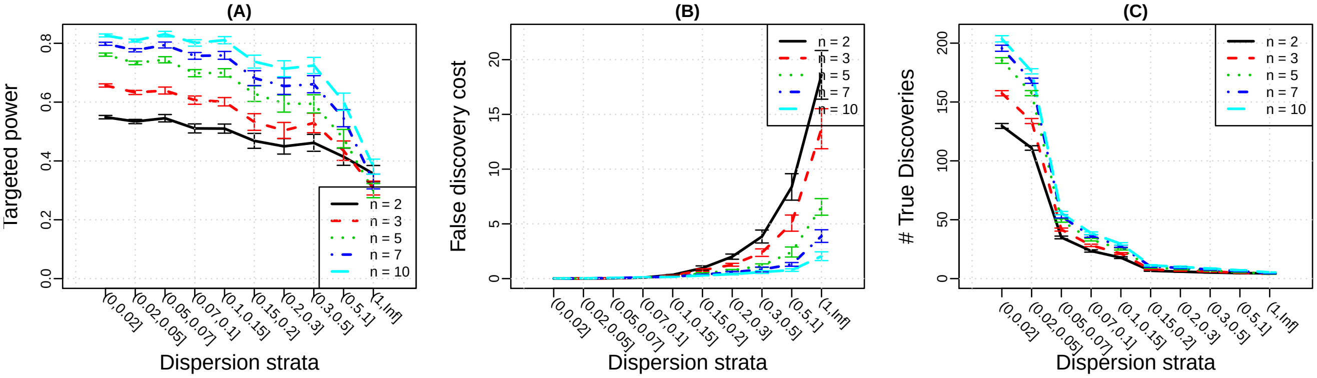

Figure S6: Power analyses results for *Bottomly* data, stratified by dispersion.

6 Results using different stratification

We performed additional simulations using a different stratification for mean counts. We further divide the (0,10] stratum into three strata: (0,2], (2,5], and (5,10]. The results from *Cheung data* (stratified power and number of true discoveries) are shown in Figure S7.

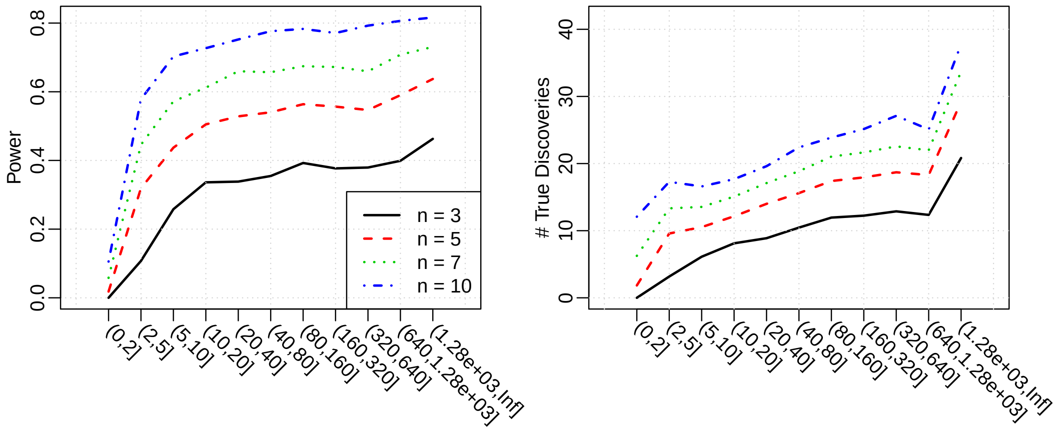

Figure S7: Targeted power and number of true discoveries, stratified by average counts, under different sample sizes. These are similar to Figures 1 and 3 in the main text, but with different stratification.

Compared with the results in the main text (Figures 1 and 4), we observe that (1) the stratified power curves become smoother, and do not show the sharp increase after the first stratum; (2) The high number of true discoveries in the (0,10] stratum disappeared. These are the results of dividing one stratum into three. However these results still suggest that using a cutoff of 10 to filter is reasonable, especially when sample sizes are small. When there are 10 samples in each group, filtering by 5 might also be reasonable choice.

7 Comparison of `PROPER` and `ssize.fdr`

The `ssize.fdr` R package (Liu and Hwang, 2007) is frequently used to perform sample size calculations for microarray experiments. People may wonder whether this package can also be applied to RNA-seq experiments. In this subsection, we explore the consequence of using this package as compared to our PROPER package.

We used the same simulation settings as in the manuscript, based on the *Cheung data* and the *Bottomly data*. For `ssize.fdr`, we used the `ssize.twoSampVary` function with effect size `deltaMean` equal to 0 and `deltaSE` equal to 1.5. The function also required specification of distributional parameters for the variance of gene expression levels, which was assumed to follow a inverse gamma distribution. For microarray data, the variance of gene expression is approximately the squared coefficient of variation in the untransformed data. This is analogous to the dispersion parameter ϕ for RNA-seq data (Wu *et al.*, 2013). We therefore used the ϕ values from the *Cheung data* and the *Bottomly data* to estimate the distributional parameters in the inverse gamma distribution. For PROPER, we considered the marginal power (by setting $\Delta = 0$) instead of the marginal targeted power since `ssize.fdr` only computes marginal power.

The comparison results are shown in Figure S8 below. They indicate that compared to PROPER, ssize.fdr mostly overestimates the marginal power from both the *Cheung data* and the *Bottomly data*. This is because that ssize.fdr does not take into account the sequencing depth information, and assumes that the power of detecting DE genes only depends on the effect sizes.

To summarize, the power calculation method developed by Liu and Hwang (2007) for microarray data is not applicable for RNA-seq data and may lead to erroneous results.

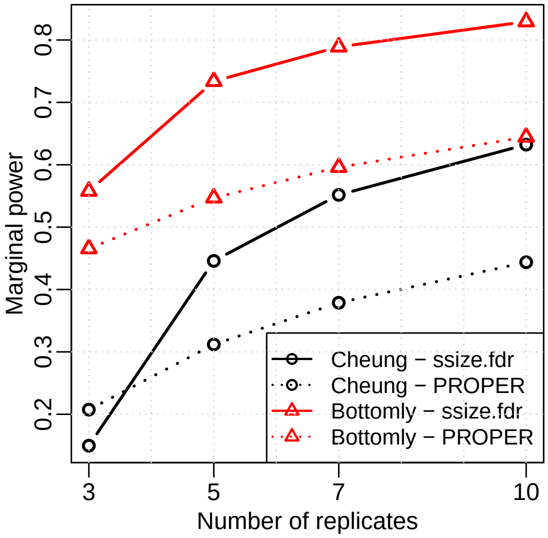

Figure S8: Marginal power versus sample size for PROPER and ssize.fdr.

References

Liu, P. and Hwang, J. G. (2007). Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics*, **23**(6), 739–746.

Wu, H., Wang, C., and Wu, Z. (2013). A new shrinkage estimator for dispersion improves differential expression detection in rna-seq data. *Biostatistics*, **14**(2), 232–243.