Optimizing the noise versus bias trade-off for Illumina Whole Genome Expression BeadChips

Wei Shi[1], Alicia Oshlack[1] and Gordon K Smyth[1,2]

[1] *The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052,*
[2] *The Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC 3010, Australia*

1 Supplementary Figures and Tables

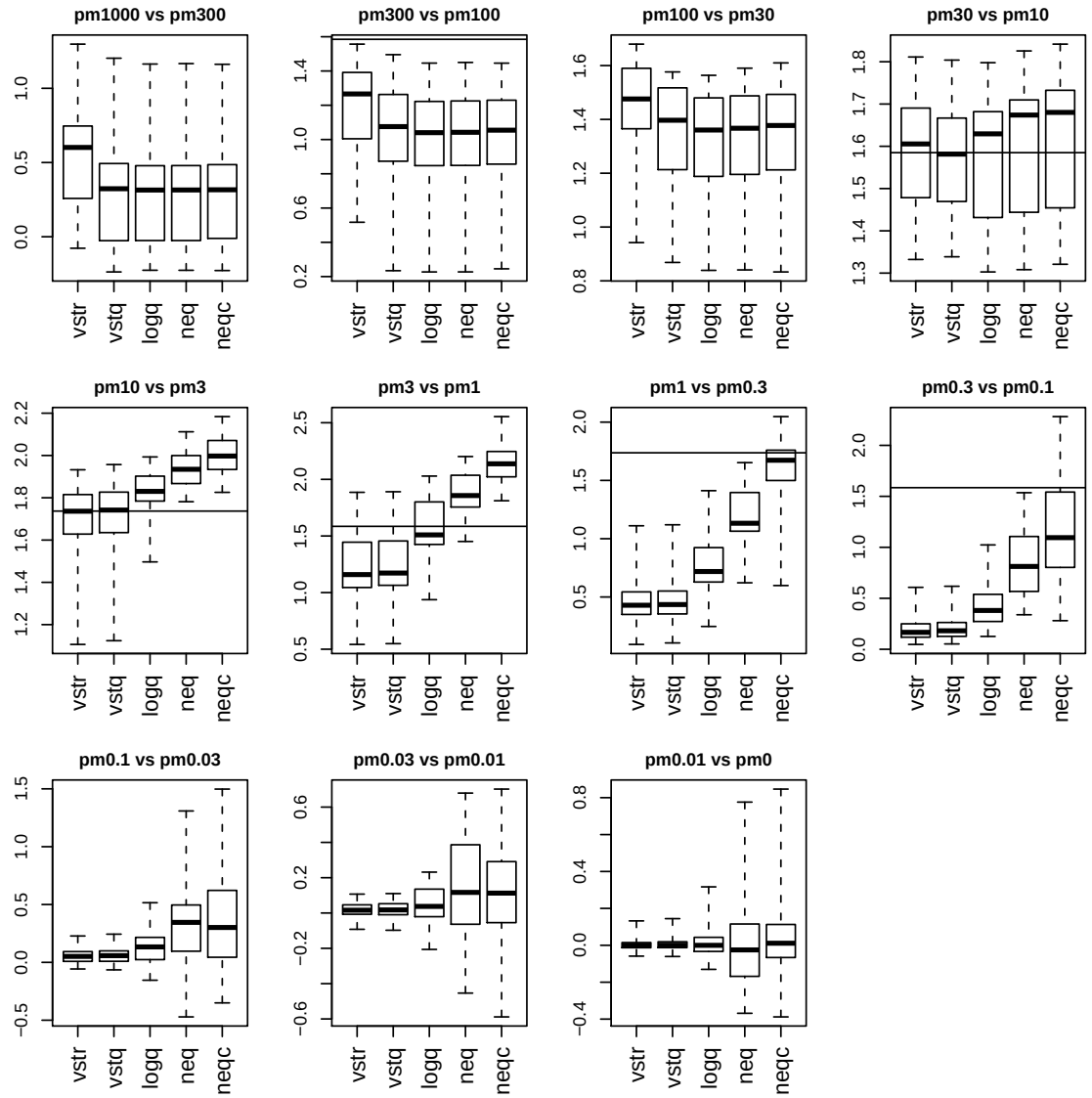

Figure S1: Observed $\log_2$ fold-changes between closest nominal spike-in concentrations for each pre-processing strategy. The horizontal lines represent the ideal $\log_2$ fold-changes. Results shown are for dataset S1 (Results are similar for dataset S2).

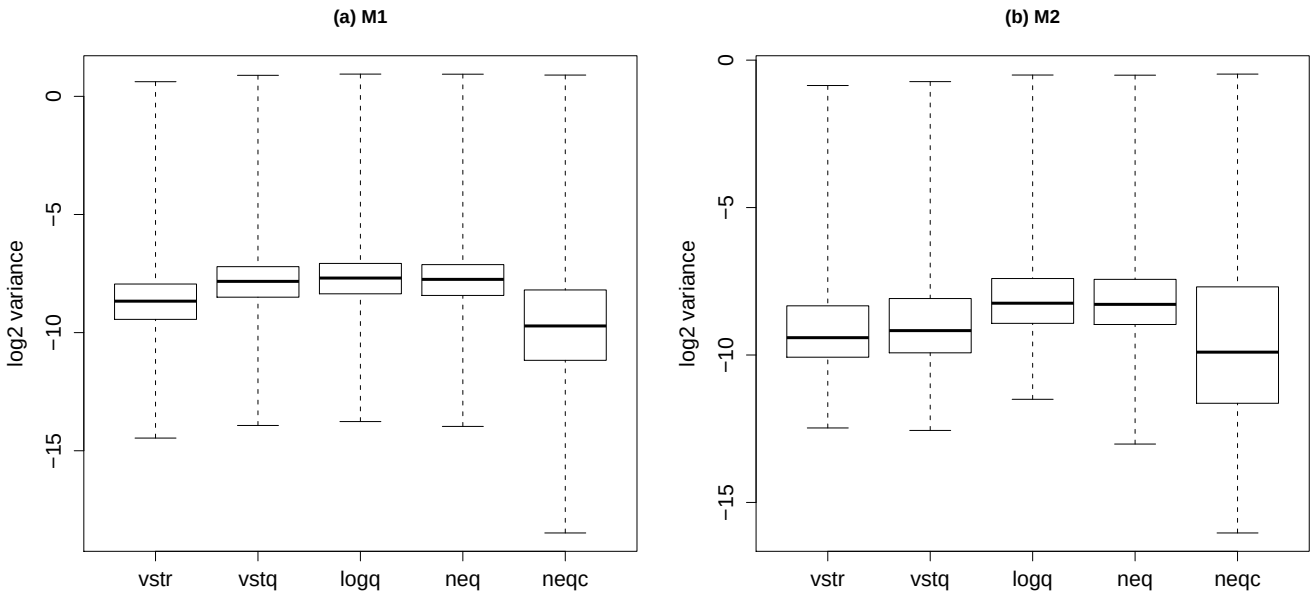

Figure S2: Precision comparison for alternative pre-processing strategies when their total offsets are forced equal. (a) Dataset M1. (b) Dataset M2.

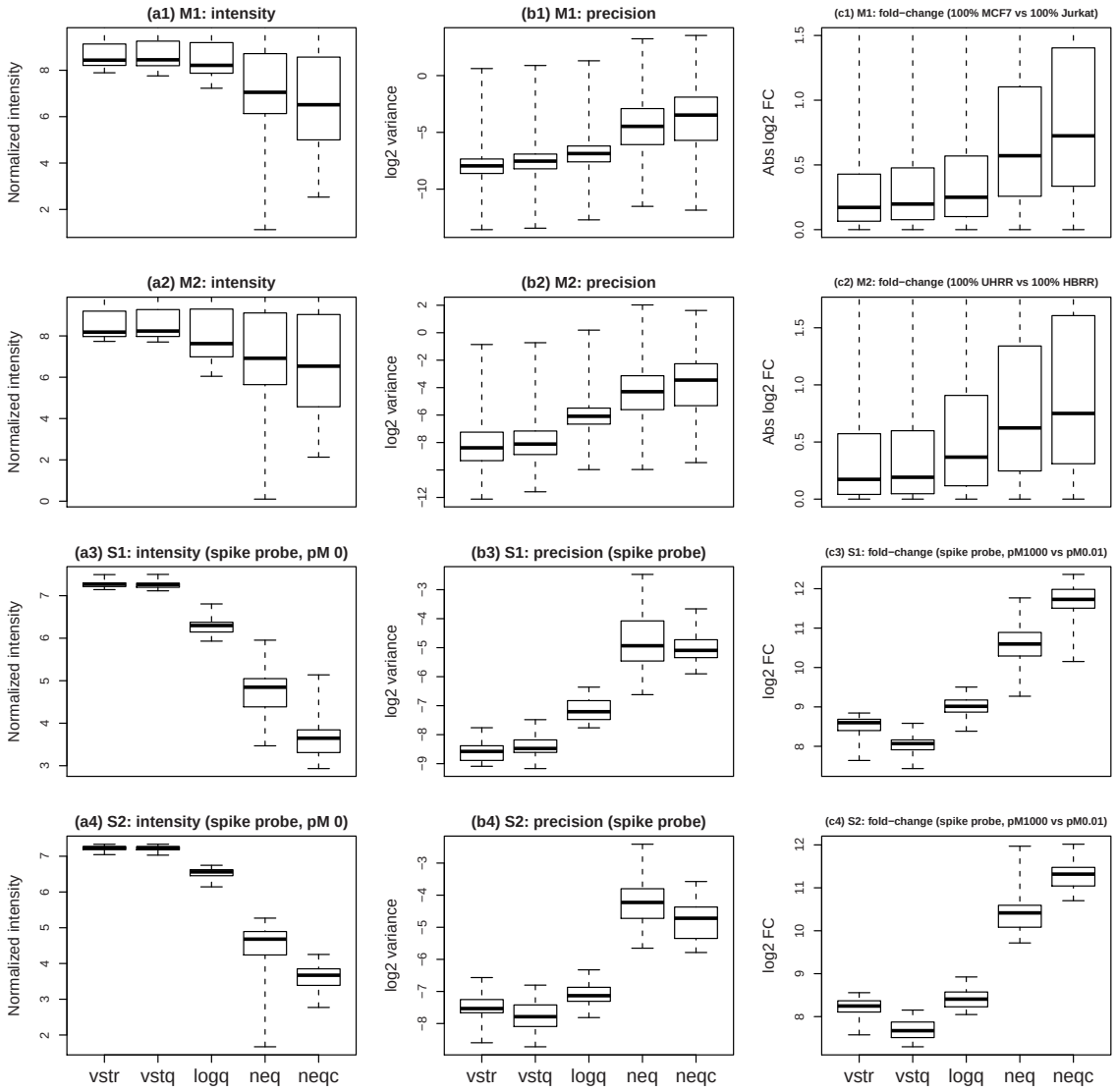

Figure S3: Intensity range, precision and fold-change range measured for each pre-processing strategy using filtered data.

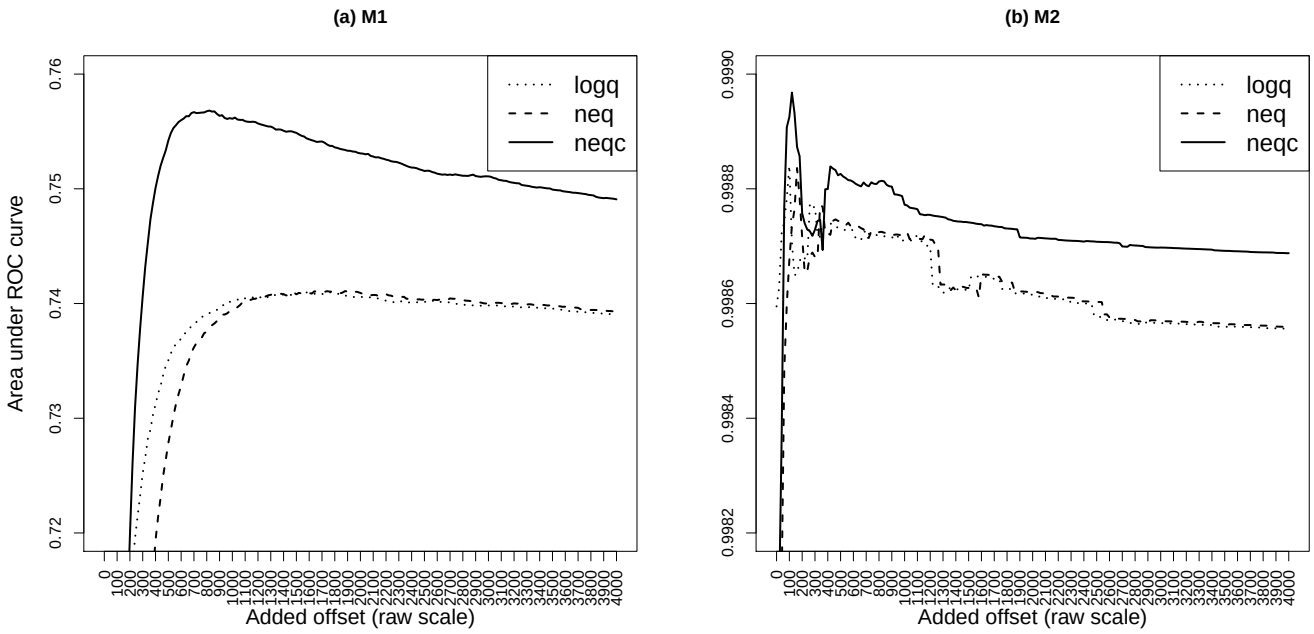

Figure S4: Change of AUC values with the increase of added offset for logq, neq and neqc strategies using filtered data. (a) Dataset M1. (b) Dataset M2.

Table S1: Innate offset, precision and bias measured for each pre-processing strategy using filtered data on mixture datasets M1 and M2

Strategy	M1			M2		
	Innate offset	$\log_2$ var	90% logFC	Innate offset	$\log_2$ var	90% logFC
vstr	264	**-7.9**	0.84	227	**-8.4**	1.22
vstq	253	-7.5	0.92	226	-8.1	1.27
logq	190	-6.9	1.07	90	-6.1	1.76
neq	25	-4.5	1.98	14	-4.3	2.47
neqc	**11**	-3.5	**2.54**	**7**	-3.5	**2.99**

Note that the innate offset is measured as the 1% quantile of normalized intensities for the filtered data.

Table S2: Comparing different strategies using filtered data when their false discovery rates are forced equal

Dataset	Strategy	Added offset	Total offset	AUC	90% logFC
M1	vstr	-	264	0.712	0.84
	vstq	-	253	0.693	0.92
	logq	182	375	0.712	0.80
	neq	345	370	0.712	0.80
	neqc	**177**	**188**	0.712	**1.10**
M2	vstr	-	227	0.9987	1.22
	vstq	-	226	0.9988	1.27
	logq	70	160	0.9988	1.50
	neq	147	161	0.9988	1.50
	neqc	**59**	**66**	0.9988	**1.98**

Table S3: Comparing different strategies using filtered data when their total offsets are forced equal

Dataset	Strategy	Added offset	Total offset	AUC	90% logFC
M1	vstr	-	264	0.712	0.84
	vstq	-	253	0.693	0.92
	logq	74	264	0.690	0.93
	neq	239	264	0.690	0.93
	neqc	**253**	264	**0.734**	**0.96**
M2	vstr	-	227	0.9987	1.22
	vstq	-	226	**0.9988**	1.27
	logq	136	226	0.9986	1.36
	neq	212	226	0.9986	1.36
	neqc	**219**	226	0.9987	**1.39**

2 A case study

In this section, we use a case study to illustrate how to perform the normalization using the neqc pre-processing strategy. A publicly available dataset from Gene Expression Omnibus database (accession number GSE16997) was used in this case study. This database can be accessed via the URL: http://www.ncbi.nlm.nih.gov/geo/.

This dataset used two Illumina HumanWG-6 version 3 BeadChips which each includes six arrays. There are four different biological samples in this dataset: mammary stem cells (MS), progenitor luminal cells (pL), mature luminal cells (mL) and fibroblast-enriched stromal cells (stroma). Each sample has three replicates.

To perform the neqc normalization, the programming software "R" (www.r-project.org) and Bioconductor R package "limma" (http://www.bioconductor.org/packages/release/bioc/html/limma.html) have to be downloaded and installed onto the computer.

Now we show how to read in data, perform the neqc normalization, filter out non-expressed probes and discover differentially expressed probes.

Read in data:

```
> library(limma)
> x <- read.ilmn(files="probe profile.txt",ctrlfiles="control probe profile.txt",
+ other.columns="Detection")
```

Read in sample information:

```
> targets <- readTargets()
> targets
```

```
   Ptnumber Age Digest Subpopulation SampleNo SentrixBarcode SampleSection SecP   Type
1  08RMH263  39    9hr  P5(Myo/stem)        1     4380071023             A   P5     MS
2  08RMH263  39    9hr   P6(Stromal)        2     4380071023             B   P6 stroma
3  08RMH263  39    9hr P7(MatureLum)        3     4380071023             C   P7     mL
4  08RMH263  39    9hr P8(ProgenLum)        4     4380071023             D   P8     pL
5  08RMH313  57    9hr  P5(Myo/stem)        5     4380071023             E   P5     MS
6  08RMH313  57    9hr   P6(Stromal)        6     4380071023             F   P6 stroma
7  08RMH313  57    9hr P7(MatureLum)        7     4380071027             A   P7     mL
8  08RMH313  57    9hr P8(ProgenLum)        8     4380071027             B   P8     pL
9  08RMH434  21    5hr  P5(Myo/stem)        9     4380071027             C   P5     MS
10 08RMH434  21    5hr   P6(Stromal)       10     4380071027             D   P6 stroma
11 08RMH434  21    5hr P7(MatureLum)       11     4380071027             E   p7     mL
12 08RMH434  21    5hr P8(ProgenLum)       12     4380071027             F   p8     pL
```

Perform neqc normalization (the neqc() function implemented the neqc pre-processing strategy) :

```
> y <- neqc(x)
```

Filter out probes which were not expressed in all samples:

```
> expressed <- apply(y$other$Detection < 0.05,1,any)
> y <- y[expressed,]
```

Carry out differential expression analysis and get the number of differentially expressed genes under false discovery rate of 5%:

```
> ct <- factor(targets$Type)
> design <- model.matrix(~0+ct)
> colnames(design) <- levels(ct)
> fit <- lmFit(y,design)
> contrasts <- makeContrasts(MS-mL, MS-pL, mL-pL, levels=design)
> contrasts.fit <- eBayes(contrasts.fit(fit, contrasts))
> summary(decideTests(contrasts.fit, method="global"))
   MS - mL MS - pL mL - pL
-1    2724    2377    1235
0    23273   23924   25941
1     2461    2157    1282
```

Display top 10 differentially expressed genes between samples "MS" and "mL".

```
> topTable(contrasts.fit, coef=1)
          PROBE_ID    SYMBOL logFC AveExpr     t  P.Value adj.P.Val    B
13343 ILMN_1766707     IL17B  3.16    6.78  52.5 6.71e-13  1.91e-08 17.5
3907  ILMN_1783149     CDH23  3.26    7.60  41.4 6.20e-12  8.82e-08 16.4
25182 ILMN_1728496      SYT9 -1.98    6.50 -33.8 4.06e-11  2.99e-07 15.2
5191  ILMN_1708303   CYP4F22 -3.04    6.77 -33.7 4.20e-11  2.99e-07 15.2
15261 ILMN_1669819 LOC402569 -1.67    6.40 -32.4 6.03e-11  3.43e-07 14.9
1091  ILMN_1777998  ARHGAP25  3.77    7.03  31.4 8.14e-11  3.86e-07 14.7
18639 ILMN_1676088     MSRB3  5.45    8.51  30.8 9.63e-11  3.92e-07 14.6
8474  ILMN_2413323       GRP  5.58    7.53  28.8 1.82e-10  6.47e-07 14.1
26055 ILMN_1811426     TMTC1  4.50    7.72  28.3 2.13e-10  6.74e-07 14.0
24370 ILMN_1701933      SNCA  4.18    7.07  26.8 3.58e-10  9.62e-07 13.6
```