

# ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences

Wentao Yang

August 10, 2017

## 1 Introduction

This vignette is intended to give a brief introduction of the **ABSSeq** R package by analyzing the simulated data from Sonesson et al. [4]. For details about the approach, consult Yang [3]. Currently, **ABSSeq** integrates two modules: basic model for pairwise comparison and linear model for complex design.

RNA-Seq quantifies gene expression with reads count, which usually consists of conditions (or treatments) and several replicates for each condition. The expected expression of each gene is estimated from number of read count, proportional to the expectation value of the true concentration of count. As a result, a normalization method need to be apply on the original counts. The normalized counts usually have enormous variation across genes and compared conditions. The reliable identification of differential expression (DE) genes from such data requires a probabilistic model to account for ambiguity caused by sample size, biological and technical variations, levels of expression and outliers.

**ABSSeq** infers differential expression directly by the counts difference between conditions. It assumes that the sum counts difference between conditions follow a Negative binomial distribution with mean  $\mu$  (proportional to expression level) and dispersion factor  $r$  (size). The  $\mu$  and  $r$  is determined by variation in the experiment, i.e., biological variation, sequencing and mapping biases. Typically, the number of replicates in a study is small and not enough to reveal all variation. To overcome this problem, a common solution is to borrow information across genes. Here, we use local regression to smooth dispersion across genes. The smoothed dispersions are then used to produce pseudocounts in the  $\mu$  estimation to accounts for dynamic dispersions across expression levels, which in turn moderates the fold-change.

**ABSSeq** tests counts difference directly against a baseline estimated from the data set ( $\mu$ ), and therefore reports p-values related to magnitude of difference (fold-change). In addition, **ABSSeq** moderates the fold-changes by two steps: the expression level and gene-specific dispersion, that might facilitate the gene ranking by fold-change and visualization (Heatmap). New alternative approach (named **aFold**) was introduced, which calls DE genes via log fold-change (see second Section for example). **aFold** uses a polynomial function to model the uncertainty (or variance) of read count, and thus takes into consideration the variance of expression levels across treatments and genes. **aFold** produces accurate estimation of fold changes. In combination with the linear model from

limma [2], aFold is capable to analyze data set with complex experimental design (see last Section for example).

## 2 Pairwise study

We firstly import the ABSSeq package.

```
> library(ABSSeq)
```

Then, we load a simulated data set. It is a list object and contains three elements: the counts matrix, denoted by 'counts', the groups, denoted by 'groups' and differential expression genes, denoted by 'DEs'.

```
> data(simuN5)
> names(simuN5)
```

```
[1] "counts" "groups" "DEs"
```

simuN5 is simulated from Negative binomial distribution with means and variances from Pickrell's data [5] and added outliers randomly [4]. simuN5 includes group information.

```
> simuN5$groups
```

```
[1] 0 0 0 0 0 1 1 1 1 1
```

An ABSDataSet object is required for ABSSeq and could be constructed with the ABSDataSet function by providing counts matrix and defined groups. Here, we can also initiate a paired comparison for specific samples, such as data for cancer and normal tissue from same individuals, by setting the **paired** parameter in ABSDataSet object.

```
> obj <- ABSDataSet(simuN5$counts, factor(simuN5$groups))
> pairedobj <- ABSDataSet(simuN5$counts, factor(simuN5$groups), paired=TRUE)
```

The ABSDataSet function also includes the parameter for the normalization method, which has a default as **qtotal**. **qtotal** assesses the influence of DE on data structure to normalize the data. Additional choices of normalization methods are also provided, that are, **total** by total reads count, **geometric** from DESeq [6], **quantile** by reads count in the first three quartiles from baySeq [?], **TMM** from edgeR [1] and **user** through size factors provided by users. The normalization method can be showed and revised by **normMethod**.

```
> obj1 <- ABSDataSet(simuN5$counts, factor(simuN5$groups),
+                   normMethod="user", sizeFactor=runif(10,1,2))
> normMethod(obj1)
```

```
[1] "user"
```

```
> normMethod(obj1) <- "geometric"
> normMethod(obj1)
```

```
[1] "geometric"
```

The size factors could be estimated from an `ABSDataset` object via the function `normalFactors` and retrieved by the function `sFactors`.

```
> obj <- normalFactors(obj)
> sFactors(obj)
```

```
[1] 1.2734782 1.1399902 0.7160235 1.0910921 1.1519218 0.9613578 0.9136937
[8] 1.0693247 0.8215518 0.8615662
```

Reads count after normalization could be retrieved by the function `counts`.

```
> head(counts(obj,norm=TRUE))
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
1	58.10857	14.035208	47.48448	0.9165129	1.736229	50.96958
2	1448.00279	950.008145	1219.23377	2216.1281496	2915.128492	6301.50409
3	2655.71877	2203.527664	2402.15589	1610.3131343	1930.686649	3455.52932
4	25.12803	9.649206	12.56942	19.2467705	18.230405	17.68332
5	1487.26533	3606.171269	3523.62751	2450.7554475	5106.249491	12778.80122
6	844.93004	816.673669	530.70886	140.2264710	1336.896331	1432.34915

	[,7]	[,8]	[,9]	[,10]
1	6.566752	0.000000	7.303252	32.49895
2	58186.893621	4996.611543	12128.267654	11214.46042
3	3185.969102	3298.343423	29429.672355	39291.23517
4	191.530262	4.675848	24.344174	55.71249
5	8367.136306	19423.474029	15542.538105	15958.14700
6	956.556852	1356.931190	1415.613738	1449.68548

With the size factors, we can calculate the absolute counts difference between conditions, mean ( $\mu$ ), size factor ( $r$ ) and moderate log2 of fold-change for each gene. It can be done by function `callParameter` as

```
> obj=callParameter(obj)
```

If we want to see correlation between the absolute log2 fold-change (with or without moderation) and expression level in same conditions, we can use function `plotDiffToBase`.

```
> obj <- callDEs(obj)
> plotDiffToBase(obj)
```

In the end, we model the counts differences with Negative binomial distribution and calculate the pvalue for each gene. DE calling is performed by the function `callDEs`, which reports pvalues as well as adjusted pvalues, that can be accessed by `results` with names of `pvalue` and `adj.pvalue`. Noticely, this function also provides fold-change moderation according to gene-specific dispersion by utilizing `aFold`, which will report fold-changes closer to gene's dispersion. In the end, `ABSseq` produces three kinds fold-changes: the original (denoted by 'rawFC'), corrected by expression level (denoted by 'lowFC') and moderated by expression level and gene-specific dispersion (denoted by 'foldChange'), which are stored in the `ABSDataset` object and could be also retrieved by `results`.

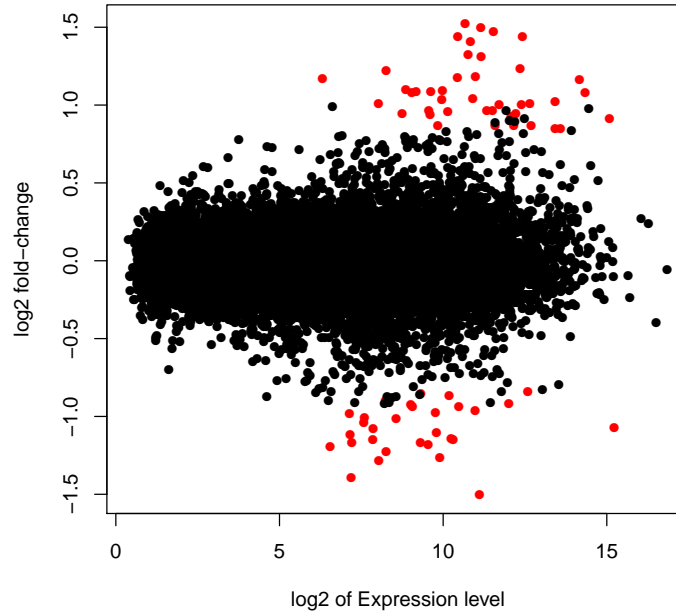


Figure 1: 'Absolute log2 fold-change against expression level'-plot for count data. We show the fitted and raw data with different colors.

```
> obj <- callDEs(obj)
> head(results(obj, c("rawFC", "lowFC", "foldChange", "pvalue", "adj.pvalue")))
```

	rawFC	lowFC	foldChange	pvalue	adj.pvalue
1	-0.2093404	-0.1146444	-0.04566872	7.449780e-01	1.00000000
2	2.8970309	2.6628526	0.35171829	4.419554e-02	1.00000000
3	1.9811932	0.9029868	0.52921978	2.124437e-02	0.81191274
4	0.8638418	0.5941570	0.20714118	1.528670e-01	1.00000000
5	2.2173452	2.1036125	1.01251221	5.787243e-05	0.01602295
6	1.1568894	0.9863872	0.43788761	6.973058e-02	1.00000000

The `results` function can be used to access all information in an `ABSDataset`.

```
> head(results(obj))
```

	Amean	Bmean	baseMean	absD	Variance	rawFC	lowFC
1	3.557138	3.347798	61.52524	25	1.174319e+03	-0.2093404	-0.1146444
2	10.654283	13.551314	23458.65032	84079	5.006097e+08	2.8970309	2.6628526
3	11.056413	13.037606	5524.26166	10567	1.904550e+06	1.9811932	0.9029868
4	4.097485	4.961327	114.26121	209	5.895087e+03	0.8638418	0.5941570
5	11.543451	13.760796	18263.75752	55896	1.884588e+07	2.2173452	2.1036125
6	9.196377	10.353267	1771.36352	2942	2.371328e+05	1.1568894	0.9863872

	foldChange	pvalue	adj.pvalue	trimmed
1	-0.04566872	7.449780e-01	1.00000000	0
2	0.35171829	4.419554e-02	1.00000000	0
3	0.52921978	2.124437e-02	0.81191274	2
4	0.20714118	1.528670e-01	1.00000000	0
5	1.01251221	5.787243e-05	0.01602295	0
6	0.43788761	6.973058e-02	1.00000000	0

In addition to a step-by-step analysis in above, DE calling could be simply performed by the function `ABSSeq`, which runs a default analysis by calling above functions in order and returns a `ABSDataset` object with all information.

```
> data(simuN5)
> obj <- ABSDataset(simuN5$counts, factor(simuN5$groups))
> obj <- ABSSeq(obj)
> res <- results(obj, c("Amean", "Bmean", "foldChange", "pvalue", "adj.pvalue"))
> head(res)
```

	Amean	Bmean	foldChange	pvalue	adj.pvalue
1	3.557138	3.347798	-0.04566872	7.449780e-01	1.00000000
2	10.654283	13.551314	0.35171829	4.419554e-02	1.00000000
3	11.056413	13.037606	0.52921978	2.124437e-02	0.81191274
4	4.097485	4.961327	0.20714118	1.528670e-01	1.00000000
5	11.543451	13.760796	1.01251221	5.787243e-05	0.01602295
6	9.196377	10.353267	0.43788761	6.973058e-02	1.00000000

Moreover, `ABSSeq` also allows testing on user-defined baseline for counts difference by giving a same value to `minRates` and `maxRates` as

```
> data(simuN5)
> obj <- ABSDataset(simuN5$counts, factor(simuN5$groups), minRates=0.2, maxRates=0.2)
> #or by slot functions
> #minRates(obj) <- 0.2
> #maxRates(obj) <- 0.2
> obj <- ABSSeq(obj)
> res <- results(obj, c("Amean", "Bmean", "foldChange", "pvalue", "adj.pvalue"))
> head(res)
```

	Amean	Bmean	foldChange	pvalue	adj.pvalue
1	3.557138	3.347798	-0.04566872	6.752807e-01	1.0000000000
2	10.654283	13.551314	0.35171829	2.332833e-02	0.4206189759
3	11.056413	13.037606	0.52921978	2.920906e-03	0.1225305179
4	4.097485	4.961327	0.20714118	9.451169e-02	0.6318826949
5	11.543451	13.760796	1.01251221	9.567908e-07	0.0002432787
6	9.196377	10.353267	0.43788761	1.716688e-02	0.3609419096

`ABSSeq` penalizes the dispersion estimation by adding a common dispersion value to the observed dispersion for each gene, which is obtained by quantile estimation on observed dispersions. This penalized dispersion could be provided by user as

```

> data(simuN5)
> obj <- ABSDataSet(simuN5$counts, factor(simuN5$groups), minDispersion=0.1)
> #or by slot functions
> #minimalDispersion(obj) <- 0.2
> obj <- ABSSeq(obj)
> res <- results(obj, c("Amean", "Bmean", "foldChange", "pvalue", "adj.pvalue"))
> head(res)

```

	Amean	Bmean	foldChange	pvalue	adj.pvalue
1	3.557138	3.347798	-0.04566872	0.7296174095	1.0000000
2	10.654283	13.551314	0.35171829	0.0458284520	1.0000000
3	11.056413	13.037606	0.52921978	0.0381019750	1.0000000
4	4.097485	4.961327	0.20714118	0.1558080086	1.0000000
5	11.543451	13.760796	1.01251221	0.0004845806	0.1829512
6	9.196377	10.353267	0.43788761	0.0925008540	1.0000000

In addition, ABSSeq provides special parameter estimation for data set without replicates. It firstly treats the two groups as replicates and separates genes into two sets according to fold-change cutoff (depends on expression level). The set with fold-change under cutoff is used to estimate the dispersion for each gene by local regression as well as fold-change moderation. Here is the example, which replaces the `callParameter` by `callParameterwithoutReplicates`.

```

> data(simuN5)
> obj <- ABSDataSet(simuN5$counts[, c(1,2)], factor(c(1,2)))
> obj <- ABSSeq(obj)
> res <- results(obj, c("Amean", "Bmean", "foldChange", "pvalue", "adj.pvalue"))
> head(res)

```

	Amean	Bmean	foldChange	pvalue	adj.pvalue
1	6.152216	4.164848	-1.31253092	0.003349715	0.018035481
2	10.771782	10.164162	-0.58497715	0.130788605	0.311984536
3	11.646445	11.377250	-0.26251482	0.792564421	1.000000000
4	4.969144	3.660384	-0.75156419	0.070505912	0.196517484
5	10.810358	12.087692	1.23750156	0.001260609	0.008184431
6	9.995210	9.946187	-0.04703356	0.999960287	1.000000000

### 3 Detecting DE via aFold

Recently, ABSSeq integrates a new method for DE detection: aFold. aFold utilizes a polynomial function to model the uncertainty of observed reads count and moderate the fold change calculation. aFold takes into account variations among samples and genes and reports DE and fold changes in a reliable way. The fold changes produced by aFold may help the experimentalist to avoid arbitrary choice of cut-off thresholds and may enhance subsequent downstream functional analyses. Here is the example for how to use aFold in ABSSeq.

```

> data(simuN5)
> obj <- ABSDataSet(counts=simuN5$counts, groups=factor(simuN5$groups))
> obj <- ABSSeq(obj, useaFold=TRUE)

```

```
> res <- results(obj,c("Amean","Bmean","foldChange","pvalue","adj.pvalue"))
> head(res)
```

	Amean	Bmean	foldChange	pvalue	adj.pvalue
1	3.557138	3.347798	-0.04566872	7.533376e-01	9.339060e-01
2	10.654283	13.551314	0.35171829	1.551471e-02	1.952503e-01
3	11.056413	13.037606	0.52921978	2.710494e-04	9.874281e-03
4	4.097485	4.961327	0.20714118	1.540658e-01	5.890517e-01
5	11.543451	13.760796	1.01251221	3.237986e-12	9.839527e-10
6	9.196377	10.353267	0.43788761	2.586298e-03	5.693054e-02

## 4 PCA analysis via aFold

aFold model also stabilizes variances across expression levels, which could be used for principal component analysis (PCA). Here is an example. Noticeably, the group information is not necessary for the `ABSDataset` object under PCA analysis.

```
> data(simuN5)
> obj <- ABSDataset(counts=simuN5$counts)
> ##as one group
> cond <- as.factor(rep("hex",ncol(simuN5$counts)))
> ##normalization
> cda <- counts(obj,T)
> ##variance stabilization
> sds <- genAFold(cda,cond)
> ##sds is list vector, which contains variance stabilized read counts in 3rd element
> ##or expression level adjusted counts in 4th element. 3rd element is more sensitive
> ##to difference between samples than the 4th one. Here we use the 4th element for a
> ##PCA analysis.
> ## log transformation
> ldat <- log2(sds[[4]])
> ## PCA analysis
> PCA <- prcomp(t(ldat), scale = F)
> ## Percentage of components
> percentVar <- round(100*PCA$sdev^2/sum(PCA$sdev^2),1)
> ## plotting
> pc1=PCA$x[,1]
> pc2=PCA$x[,2]
> #plot(pc1,pc2,main="",pch=16,col="black",xlab="PC1",ylab="PC2",cex=1.2)
```

## 5 DE analysis with complex design

In combination with linear model from `limma` [2], aFold is capable to analyze data set with complex experimental design, which is performed by the function `ABSSeqlm`. Here is an example.

```
> data(simuN5)
> groups<-factor(simuN5$groups)
```

```
> obj <- ABSDataSet(counts=simuN5$counts)
> design <- model.matrix(~0+groups)
> res <- ABSSeqlm(obj,design,condA=c("groups0"),condB=c("groups1"))
> head(res)
```

	basemean	logFC	pvalue	p.adj
1	4.669945	-0.04281206	7.854513e-01	9.545837e-01
2	14.180418	0.35069190	2.575496e-02	4.982625e-01
3	13.941520	0.36647548	1.979309e-02	4.349244e-01
4	5.901813	0.19783008	2.084245e-01	8.358346e-01
5	13.815285	0.99967063	2.064820e-10	7.145999e-08
6	10.369845	0.42200472	7.289263e-03	2.346690e-01

Noticely, the parameters `condA` and `condB` could contain multiple conditions (factors) to run a comparison between multiple conditions. The function `ABSSeqlm` could be also used for analysis of variance (ANOVA) across conditions. To run ANOVA, all conditions (factors) are imported to the parameter `condA` in the function `ABSSeqlm` (without `condB`).

```
> res <- ABSSeqlm(obj,design,condA=c("groups0","groups1"))
> head(res)
```

	basemean	logFC	pvalue	p.adj
1	4.669945	0.0302727	8.489800e-01	0.999768961
2	14.180418	0.2479766	1.188042e-01	0.999768961
3	13.941520	0.2591373	1.030984e-01	0.999768961
4	5.901813	0.1398870	3.789073e-01	0.999768961
5	13.815285	0.7068739	8.734188e-06	0.003022757
6	10.369845	0.2984024	6.051828e-02	0.999768961

The linear model is performed by `lmFit` from `limma` [2], which could be suppressed via the parameter `lmodel` as

```
> res <- ABSSeqlm(obj,design,condA=c("groups0"),condB=c("groups1"),lmodel=FALSE)
> head(res)
```

	basemean	logFC	pvalue	p.adj
1	4.669945	-0.04281206	7.854513e-01	9.545837e-01
2	14.180418	0.35069190	2.575496e-02	4.982625e-01
3	13.941520	0.36647548	1.979309e-02	4.349244e-01
4	5.901813	0.19783008	2.084245e-01	8.358346e-01
5	13.815285	0.99967063	2.064820e-10	7.145999e-08
6	10.369845	0.42200472	7.289263e-03	2.346690e-01

## References

- [1] Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. Bioinformatics 26.1 (2010): 139-140.



- [2] Gordon K. Smyth.. *Linear models and empirical Bayes methods for assessing differential expression in microarray experiments*. Statistical Applications in Genetics and Molecular Biology 3.1 (2004): 1-25.
- [3] Wentao Yang, Philip Rosenstielb and Hinrich Schulenburg. *ABSSeq: a new RNA-Seq analysis method based on modelling absolute expression differences*. BMC Genomics 2016 17:541.
- [4] Sonesson C, Delorenzi M *A comparison of methods for differential expression analysis of RNA-seq data*. BMC Bioinformatics 2013, 14(1):91.
- [5] Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK *Understanding mechanisms underlying human gene expression variation with RNA sequencing* Nature 2010, 464(7289):768-772.
- [6] Anders S, Huber W *Differential expression analysis for sequence count data*. Genome Biol 2010, 11(10):R106.