Supplementary Methods and Results

Sequencing bias due to transcript length

Several investigators (notably Oshlack and Wakefield, 2009) have pointed out that a long transcript will tend to have higher aggregate read counts than a short transcript, even if the two have equal expression, as the long transcript has more opportunities for sequences from fragmented reads to appear. Oshlack and Wakefield (2009) demonstrated empirically that the proportion of differentially expressed transcripts in a number of experiments was positively associated with transcript length. Clearly when using read counts to compare expression *across* transcripts, some normalization by transcript length is desirable. However, within a transcript/gene, such normalization may not be highly consequential.

As a simple example, consider a genes with read count Y_i, transcript length λ_i, and assume all samples have equal library sizes. For simplicity, also suppose that the counts do not exhibit overdispersion, i.e. are distributed as Poisson. If the mean count value is λ, the transcript follows a standard mean-variance relationship $\log(\mathrm{var}(Y_i)) = \log(\mathrm{mean}(Y_i)) = \log(\lambda)$. Now consider the mean-variance relationship for the length-normalized read count $Y_i'=Y_i/l_i$, where l_i is the library size. We have $\log(\mathrm{E}(Y_i'))= -\log(l_i)+\log(\lambda)$, whereas

$\log(\mathrm{var}(Y_i'))=-2\log(l_i)+\log(\lambda)=-\log(l_i)+\log(\mathrm{E}(Y_i'))$. Thus the normalization has taken a perfect unit mean-variance relationship and introduced an extraneous offset by $\log(l_i)$. This phenomenon appears to lie behind the observation in Oshlack and Wakefield (2009) that the mean variance relationship is much more strongly affected by transcript length after normalization (see the authors' Figure 2b vs. 2a). However, in doing so, the researcher has not really learned more about the transcript variability than was already apparent from the transcript mean. In other words, it is not clear that the normalization provides extra information that is useful for differential expression analysis. Using the CEU data, Supplementary Figure 2 also illustrates that transcript length is only a modest determinant of the mean-variance (and therefore the mean-overdispersion) relationship.

We recognize that this discussion is elementary, but perhaps necessary, as a number of researchers may, for example, immediately compute RPKM values (reads per kilobases mapped per million reads, Mortazavi *et al.*, 2008) before performing any downstream analysis. For differential expression analysis, especially when attempting to use the mean-variance relationship in a manner similar to our constrained approach, pre-such normalization may be counterproductive.

Note that the BBSeq model, as well as the negative binomial approaches described by others, appropriately considers the library size, which typically varies considerably across samples. However, it is often useful to have a simple quantity to represent expression level across samples, and for this purpose we recommend using the read proportions y_{ij}/s_j (example in Supplementary Figure 8).

Handling outliers

When fitting the free model, a small percentage of transcripts (typically 5% or fewer) exhibit outlying ψ estimates, in terms of their residuals from the mean-overdispersion model. This outcome is relatively insensitive to starting values, and tends to occur in genes with a majority of zero counts. For such low expression genes, the power to detect differential expression is low, and the outlying values are of little consequence, except for the danger of declaring false positives by underestimating the overdispersion. We detect outlying ψ values by applying the mean absolute deviation method of Davies and Gather (1993), with a cutoff of 5.2 median absolute deviation units. An extremely simple approach is to simply impute the outlying values at the mean of ψ among the non-outliers, which appears to be conservative, and was used for the results in the paper.

Another type of outlier occurs for individual samples. Outlying high read counts compared to other samples can produce spurious results, as well as an excessive proportion of zero counts (so that otherwise unimpressive non-zero values gain undue weight). These count values may be correct, but can produce apparently highly significant results for the BBSeq constrained model when fitting the model to data with large sample sizes, even if a single large value occurs in only one experimental condition. The free model does not appear very susceptible to this phenomenon, as the outlying value results in a larger estimate of the overdispersion parameter ϕ (or equivalently ψ), reducing the apparent significance. An extremely simple approach is to compare the ratio of highest read counts (or read proportions) compared to the second-highest values, and "flag" the constrained gene significance as potentially suspect if the ratio exceeds a specified threshold. By default we use a threshold of 5.0. If more than 95% of the read count values are zero, we also flag the gene. In our full example CEU dataset, for example, approximately 4% of the genes were flagged in this manner, and can be subject to further scrutiny if declared significant by the constrained model.

Finally, for the free model we found that spurious Wald statistics can arise if all zero counts appear in one of the experimental conditions. For simple two-group testing, for this small number of genes we perform simple pooled-variance t-testing to obtain approximate *p*-values.

Software settings and maximization

For edgeR (v.2.2.5), we followed recommendations from the user manual for choosing `prior.n` so that the total degrees of freedom (`prior.n*df`) associated with the prior is about 50, subject to `prior.n` not going below 1. For baySeq (v.1.4.0), we analyzed the data assuming a Negative Binomial distribution (not Poisson). As in the user manual, we obtained priors using `getPriors.NB`, and then acquired posterior likelihoods using `getLikelihoods.NB`. For DESeq (v1.5.1), we used the standard mean-variance estimation and `nBinomtest`. BBSeq performs R `optim()` optimization and Hessian matrix estimation. We found that for total sample sizes larger than 4, outliers were reduced (although at the expense of more computation) by using the conjugate gradient "CG" option.

Fitting of the overdispersed GLM proceeded using the functions `glm` and `glm.binomial.disp` from the `dispmod` package, as described in http://cran.r-project.org/web/packages/dispmod/dispmod.pdf. The approach fits an overdispersed binomial logistic model as described by Willams (1982). For a subset of 1000 genes in each of several simulations from Dataset 2, we also fit an alternative overdispersed Poisson quasi-likelihood using `glm` and `glm.poisson.disp`, with $\log(l_i)$ as a covariate. For these datasets, we confirmed that the two GLM approaches provided very similar results.

Null distributional approximations

Empirical investigation of BBSeq using small/moderate subsamples of the CEU dataset (in the range of 2 to 5 samples per experimental group) revealed that random groupings of the data resulted in Wald statistics with variance exceeding 1.0. By analogy with linear regression, we reasoned that estimation of ψ per gene may result in variance estimates relatively uncorrelated with the coefficients *B*, and thus result in a *t* distribution with approximate degrees of freedom $n-p$. The constrained model uses all genes to estimate ψ, and so we use a standard normal approximation for the Wald statistic. For genes with low expression and many zero counts, we noted remaining extra variability in the Wald statistics. We reasoned that the effective degrees of freedom for the free model was limited to the samples that showed non-zero counts, as only these values are very informative about the parameters, except in extreme differential expression scenarios. Thus we further reduced the degrees of freedom by the total number of observed zero counts per gene, with a minimum assumed degree of freedom of 1. We note that this effort pertains to the *p*-values only and that for all but the low-expressing genes the ordering of genes based on the absolute values of Wald statistics is unaffected by the degrees of freedom used. Moreover, the ROC curves or similar constructions (such as the sex-chromosome detection shown in Figure 6) are also invariant to these assumptions. Results from the null simulations described in section 3.1.2 are given in Supplementary Table 1 below.

		Constrained	Free	trend	tag	common	DESeq	GLM
$n_1=n_2=5$	α=0.05	0.046	0.047	0.015	0.054	0.055	0.016	0.074
	α =0.001	0.001	0.001	0.000	0.007	0.008	0.000	0.009
$n_1=n_2=2$	α =0.05	0.048	0.034	0.003	0.023	0.024	0.005	0.086
	α =0.001	0.001	0.001	0.000	0.001	0.001	0.000	0.013

Supplementary Table 1. Empirical type I errors for the simulations in Dataset 2, based on 20 simulated datasets, each with 32,027 genes.

Specifying multiple experimental factors

Here we describe how multiple experimental factors are specified, following standard linear modeling. As a simple example, consider a hypothetical RNA-Seq experiment with 4 cell lines, and age as a continuous covariate. The `X-builder` routine in BBSeq will automatically produce “dummy” indicator columns for each cell line, and four design matrices: $X_{Indicator} = \vec{1}$, $X_{Discrete} = \{\vec{1}, cell2, cell3, cell4\}$ (treating the first level as a reference); $X_{Quant} = \left\{\vec{1}, age\right\}$, $X_{full} = \{\vec{1}, cell2, cell3, cell4, age\}$. Standard log-likelihood ratio comparisons are then used to test for the significance of each of the factors cell line and age, either in isolation or in the presence of the other factor, using chi-square approximations and the appropriate degrees of freedom.

An example with two factors: exposure to etoposide

Few purpose-built packages for RNA-Seq analysis enable the simultaneous analysis of multiple experimental factors, and we illustrate here the use of BBSeq for such an example. In the CEU dataset, 42 of the samples had accompanying inhibitory concentration (IC_{50}) scores on a cell-death assay (Huang et al., 2007) resulting from experimental exposure to the cytotoxic cancer drug etoposide (downloaded from the Pharmacogenomics Knowledge Base `www.pharmgkb.org`). Using these values as a quantitative covariate, in addition to the qualitative covariate sex and the sexXIC_{50} interaction, we are able to fit richer models using both the free and constrained models. Supplementary Table 1 shows the *p*-values from the free model for the most significant genes using each of these predictors in the combined model. In order to illustrate the two-factor analyses resulting from the "full" vs. "reduced" models, we also ran the *X*-builder function and computed the likelihood ratio tests. For example (Supplementary Table 1), the likelihood-ratio based *p*-value for the sex covariate is computed by fitting the larger main effects model with sex and IC_{50} compared to the smaller model with IC_{50} alone. In this example the likelihood ratio approach is not easier or more parsimonious than examining the Wald statistics. However, for factors with multiple levels (such as an ANOVA analysis), it is often of interest to obtain a *p*-value for the entire factor.

The vignette for the R BBSeq package similarly contains an example quantitative score for a smaller subset of CEU samples. Supplementary Table 1 is used entirely for illustration here, and we do not comment on the biological plausibility of the genes described. However, we do note that, as expected, many of the "sex" genes are genes on the X chromosome identified in Table 1. The gene *GSTM5* (the top gene for the sexXIC_{50} interaction) is extremely interesting, as the glutathione S-transferases are known metabolizing enzymes for a variety of xenobiotics (see OMIM entry `www.ncbi.nlm.nih.gov/omim/138385`). Analyses such as these illustrate the untapped potential, even in existing RNA-Seq databases.

LRT_{sex}		$Wald_{sex}$		LRT_{IC50}		$Wald_{IC50}$		$Wald_{sexXIC50}$	
Gene	*P*value	Gene	*P*value	Gene	*P*value	Gene	*P*value	Gene	*P*value
XIST	7.00E-13	*PNPLA4*	7.62E-10	*KRT17*	1.65E-07	*PDGFA*	9.42E-17	*GSTM5*	3.07E-09
NLGN4X	1.80E-10	*XIST*	4.06E-09	*PEX26*	2.97E-07	*JUP*	1.99E-15	*ERBB4*	9.00E-08
PNPLA4	6.07E-10	*NLGN4X*	2.94E-07	*RASAL2*	3.47E-07	*KRT17*	3.38E-15	*JUP*	1.40E-07
EIF1AX	2.81E-07	*KDM5C*	1.39E-06	*BCL11B*	1.33E-06	*BCL11B*	2.62E-14	*IL1R2*	2.72E-07
COLQ	2.11E-06	*LOXHD1*	1.58E-06	*ARHGEF6*	1.41E-06	*RASAL2*	3.04E-14	*NLRP11*	2.82E-07
DDX43	2.25E-06	*EIF1AX*	5.19E-06	*PDGFA*	2.15E-06	*SERPINA1*	3.46E-14	*AKR1C1*	3.59E-07
KDM5C	2.29E-06	*PTTG2*	6.46E-06	*CYP4F3*	2.32E-06	*CYP4F3*	1.54E-13	*SYPL2*	4.66E-07
KDM6A	3.02E-06	*NCRNA00183*	1.84E-05	*SERPINA1*	2.71E-06	*HOMER3*	3.94E-13	*AKR1C2*	1.02E-06
NCRNA00183	3.52E-06	*PRKX*	3.15E-05	*PGF*	4.34E-06	*RICH2*	2.97E-12	*IL18R1*	1.16E-06
RPS4X	8.72E-06	*KDM6A*	4.73E-05	*ASTN2*	5.64E-06	*PRF1*	3.90E-12	*FYB*	1.27E-06

Supplementary Table 2. *P*-values based on the likelihood ratio test (LRT) statistics and Wald statistics for sex, etoposide IC_{50}, and the sexXIC_{50} interaction, using the 42 CEU samples with etoposide IC_{50} scores. Wald statistics are based on a combined model using sex, etopside IC_{50}, and the sexXIC_{50} interaction as predictors. The LRT performs similar analyses (although without the interaction) using the likelihood ratio approach.

References

Davies, P.L. and Gather, U. (1993). "The identification of multiple outliers" *J. Amer. Statist. Assoc.*,**88**, 782-801.

Huang, R.S. *et al*., (2007) “A genome-wide approach to identify genetic variants that contribute to etoposide-induced cytotoxicity”. *PNAS*, **104**(23), 9758-9763.

Mortazavi, A. *et al.*, (2008) “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. *Nature Methods* , 5, 621 - 628 (2008)

Oshlack, A. and Wakefield, M.J. (2009) “Transcript length bias in RNA-seq data confounds systems biology”. *Biology Direct*, **4**:14 .

Williams, D. A. (1982), Extra-binomial variation in logistic linear models, *Applied Statistics*, **31**, 144–148.