1 Additional results on t-tests

	Simulation model									
	Poisson		NB1		NB2		NBP		NBQ	
Test	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR	FDR	TDR
t	0.034	0.934	0.075	0.242	0.155	0.001	0.070	0.009	0.096	0.003
t (log)	0.046	0.918	0.122	0.186	0.188	0.001	0.179	0.008	0.256	0.003
t (sqrt)	0.070	0.956	0.100	0.228	0.173	0.001	0.158	0.008	0.232	0.002
t (asin)	0.070	0.956	0.100	0.228	0.173	0.001	0.158	0.008	0.232	0.002
t_W	0.012	0.602	0.029	0.031	0.143	0.000	0.046	0.001	0.121	0.000
t_W (log)	0.009	0.562	0.010	0.013	0.150	0.000	0.034	0.000	0.078	0.000
t_W (sqrt)	0.011	0.743	0.012	0.028	0.200	0.000	0.062	0.001	0.054	0.000
t_W (asin)	0.011	0.743	0.012	0.028	0.200	0.000	0.062	0.001	0.055	0.000
t_l	0.073	0.966	0.127	0.385	0.210	0.002	0.129	0.048	0.161	0.014
t_l (log)	0.098	0.963	0.167	0.339	0.266	0.005	0.237	0.073	0.302	0.026
t_l (sqrt)	0.164	0.986	0.158	0.404	0.247	0.003	0.230	0.090	0.336	0.027
t_l (asin)	0.164	0.986	0.158	0.404	0.247	0.003	0.230	0.090	0.336	0.027

Table 1: Comparing different t tests for differential gene expressions. t, t_W and t_l refer to equal-variance, Welch and limma t-tests respectively. The tests were applied to untransformed, log transformed, square root transformed and arcsin-root transformed RNA-Seq counts.

Table 1 summarizes the false discovery rates (FDR) and true discovery rates (TDR) of different t-tests applied to simulated data (see Section 7 for details on the simulation models and the interpretation of FDR and TDR). Since the RNA-Seq counts y for individual genes are very small compared to the total count m, the arcsin-root transformation $\sqrt{m}\arcsin(\sqrt{y}/\sqrt{m})$ and the square root transformation $\sqrt{y}$ are almost identical. For data simulated under non-Poisson models, all the t-tests yielded inflated false discovery rates except Welch t-test for the NB1 data. For data simulated under the Poisson model, the Welch t-test is less powerful than the t-test assuming equal variance. Assuming unequal variance, the Welch t-test consumes one more degree of freedom to estimate an additional variance parameter. With only three samples in each group, this additional degree of freedom induced noticeable power loss. The limma moderated t-test yielded higher TDR as well as inflated FDR. When estimating the count variance, limma shrinks the variance estimates for each gene towards a common value, but does not address the mean-variance dependence. For the data sets we simulated (which mimic the distribution of the Arabidopsis data set), there are many genes with small mean counts and thus small count variances. By shrinking the variance estimates towards a common value, limma underestimates the variances of the counts with large means. All these

t-tests do not address the dependence of the variance on the mean and they are not adequate for explaining the count variability in RNA-Seq data, especially when the range of the mean counts is large.

2 Additional MA plot for the Arabidopsis example

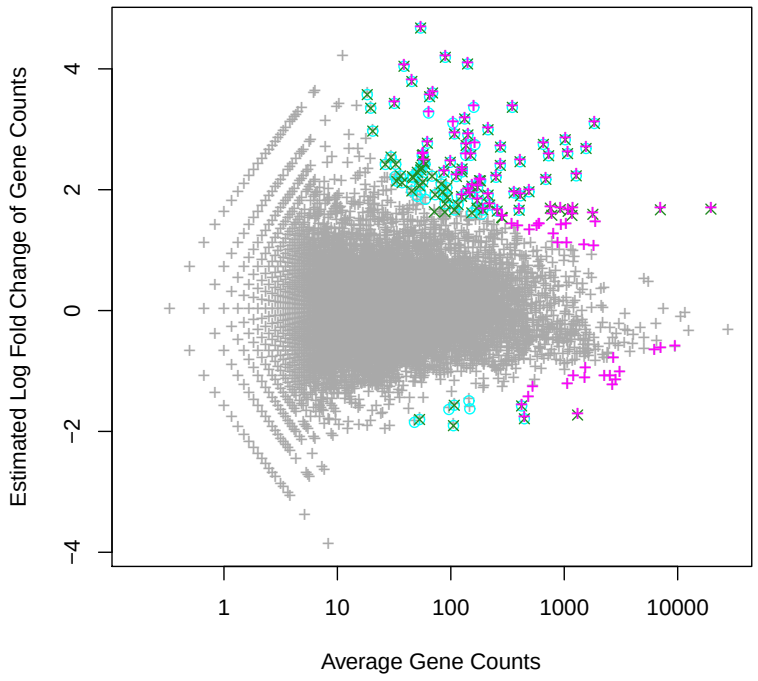

Figure 1: Scatter plot of estimated log fold change versus estimated mean of the gene counts. Green crosses, cyan circles, and magenta pluses highlight the top 100 differentially expressing genes identified by edgeR (trend), DESeq and NBP respectively.

3 Mean-variance plots of typical simulated data sets

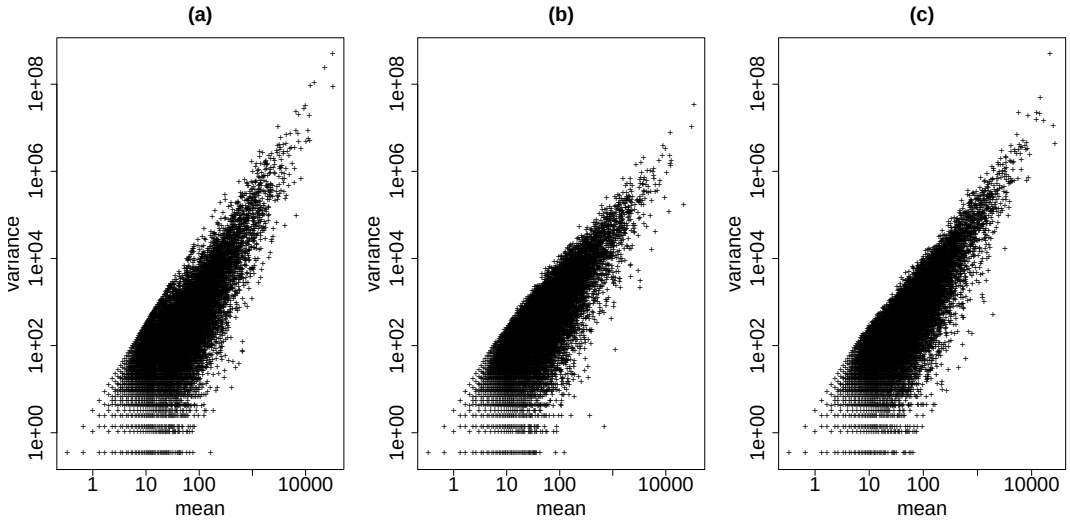

Figure 2: Scatter plot of sample variance vs. sample mean (log-log scale) of gene counts in (a) the Arabidopsis data (b) a data set simulated according to the NBP model and (c) a data set simulated according to the NBQ model. Gene counts in (b) and (c) were simulated to match the Arabidopsis data as much as possible under respective model specifications (see Section 7 of the paper for further details). The mean-variance plots of the two typical simulated data sets resemble the mean-variance plot of the Arabidopsis data.