Sensitivity, specificity and reproducibility of RNA-Seq differential expression calls

Paweł P. Łabaj and David P. Kreil
Chair of Bioinformatics Research Group, Boku University Vienna, Austria
pawel.labaj@boku.ac.at

Introduction The MAQC[1] and SEQC[2, 3] projects have introduced a key resource for testing future developments of microarray and RNA-seq analysis tools, as required in clinical and regulatory settings. In this study, based on SEQC data set, we investigate the sensitivity, specificity and reproducibility of RNA-Seq differential expression calls. Going beyond general results of the original SEQC studies[2, 3] I will extend and complement the comparative analysis by considering differential expression tests that are closer to typical 'real world' experiments. In particular I will concentrate on comparisons of samples A and C, where C consists of 3 parts of sample A and 1 part of sample B.[1, 2] This pair of samples has the smallest average effect strength ('signal') amongst the different possible pair-wise comparisons of the MAQC/SEQC ABCD samples. Exploring the effect of RNA-Seq pipeline choices, we now also consider all 55,674 known AceView genes,[17] rather than the 23,437 genes of the originally published comparison with Affymetrix HGU133Plus2.0 microarrays. A key result of our study is thus as comprehensive benchmark of alternative methods for gene expression estimation and differential expression calling, representative of the wide range of tools now available and reflecting the rapid development of the field. The presented metrics assess sensitivity, specificity, and reproducibility for both genome wide analysis and the identification of top candidates for further follow-up.

Results Comparing the SEQC samples A and C we are expecting more genes with a stronger expression in sample C because it contains, in addition to RNA from sample A, also RNA of genes expressed in sample B. Benchmark results compare a wide range of tools for expression estimation (**EE**), including r-make[4], Subread[5], TopHat2[6]/Cufflinks2[7], SHRiMP2[8]/BitSeq,[9] and kallisto[10], in combination with a range of established tools for differential expression calls (**DEC**), including limma[11], edgeR,[12] and DESeq2.[13] Tools were selected to provide a good overview of the current state of the art in RNA-Seq data analysis.

Depending on the methods for expression estimation and DE calling, the number of detected differentially express genes vary roughly between 7,000 – 10,000 (Fig. 1). Sensitivity in general depends less on the method for differential expression calling, while more variation is observed for the different approaches in estimation of expression levels. Remarkably, there nevertheless is only limited agreement of the lists of genes identified by different methods for differential expression calls, with a typical pairwise agreement of 56–67%. To investigate these discrepancies we examined *M(A)* plots, where genes are represented by dots coloured according to which methods identified them as differentially expressed (Fig. 2). In the left panel (for AvsC) we can identify areas where different DEC methods are particularly sensitive. Variation in the sensitivity of DEC methods for different effect strengths (*M*) and gene abundances (*A*) reflects the range of approaches to data normalization and statistics used for DE calling. Among the examined DEC methods, DESeq2 appears to be the most conservative in calling DE genes of low abundance (low average expression). This may be appropriate considering the relatively high variance of low count data that is characteristic of weakly expressed genes in RNA-Seq.[14] Also weakly expressed genes might be relatively more affected by site-specific variation arising during library preparation[3], as seen in the right panel of the Fig. 2, which shows a same–same sample comparison – genes identified there as 'differentially expressed' are false positives (FPs) in a search for biologically relevant differences.

The SEQC study design[1 – 3] provides us the unique possibility to further examine the site-specific effects. In particular, we can calculate an eFDR (empirical False Discovery Rate) by comparing the cross-site sensitivities for AvsC, CvsC and AvsA (Fig. 3, and Fig. 4 left panel). The number of false positives (FPs) can be reduced when appropriate methods[15, 16] are applied to remove the unwanted variation by analysing the experiment in context of similar experiments obtained from the public repositories. In our study we can use different sequencing site to mimic such 'context'. We have applied the PEER tool[16], which has performed the best in the SEQC study[3], to remove the unwanted variation. The eFDR has been

reduced noticeably from typical eFDR reaching up to 50% to not crossing in general the 20% (Fig. 4 left vs middle panel). As the eFDR is strongly dependent on the combination of EE method and DEC method, even after PEER some sequencing site pairs obtain more than 60% eFDR (outlier sites for kallisto). As the eFDR level is still not satisfactory the further filtering is needed as was shown in MAQC[1] and SEQC[2, 3] studies. In terms of RNA-Seq, unlike for microarrays, in addition to filter for small changes also the filter for small expression levels is required (see Methods for threshold details; reasoning, approach and consequences will be extended in the full version of the manuscript). This is the direct consequence of the sampling nature of NGS[14]. Application of the dedicated filters which fix the EE+DEC pipelines sensitivity for intra-site AvsC comparison to about 3000 differentially expressed genes reduced the eFDR for a typical site pair below 2.5% for almost all cases. Just for SHRiMP2/BitSeq and kallisto used together with edgeR the typical eFDR is higher but still below 5 %. Adding filtration by removing the FPs not only lower the eFDR but also increase the agreement between DEC methods as now the method specific FPs has been removed. The agreement has increased from 60-67% (after PEER correction) to the level of 86-94% depending on site, EE and DEC method.

In medical and life sciences the goal is to produce the accurate gene signatures – lists of differentially expressed genes which can be reproduced in the other laboratory. This challenge can be seen in different ways depending on the study design and the next steps which will be taken with the provided gene signatures. In terms of the whole genome studies the interest is in the accuracy of the list of all differentially expressed genes. Based on our study we can conclude that agreement between sequencing sites depends strongly on the selected DEC method when no addition filtering is used: typically 50-68% for limma, 66-72% for edgeR and 72-78% for DESeq2. Application of the additional filtering, although reduce the sensitivity, increase the agreement and makes that all DEC method have more similar ranges: typically 77-80% for limma, 81-83% for edgeR and 82-84% for DESeq. There are studies, however, where the full list of differentially express genes is not of interest. More important is list of ‘top’ candidates which can be further tested in follow-up studies. As here not the sensitivity but rather specificity is of the main concern, the filtering is more than welcome. In Figure 5 the summarized agreement (on y-axis in %) between topN (where N is on x-axis) differentially expressed genes (sorted by the effect strength) is shown. The different panels represent different DEC methods while different colours in violin plots represents different expression estimation methods (as specified in the legend). For the short top lists the agreement is strongly dependent on combination of EE and DEC methods. These differences are getting smaller when lists are getting longer with almost all combinations reaching 80% agreement for top200 and crossing 90% agreement for top1000. For a good performance for the short lists the solution might be to use of even stronger filters on average expression, but then the ‘true’ candidates with weaker average expression can be filtered out, what for many studies can be a big loss. That is why a better approach is to consider a different combination of EE and DEC method, eg. SHRiMP2/BitSeq + edgeR).

Conclusions Going beyond the general comparison presented in SEQC study[2, 3], we present here the benchmarking for scenarios which better represent the effect sizes of typical experiments. We have in details examined the sensitivity, specificity, and reproducibility of the RNA-Seq differential expression calls for a comparison of the SEQC samples A and C. We have shown that application of appropriate procedures and filters improves the reproducibility of both the genome wide analysis as well as the identification of top candidates. We also have shown that it is important to benchmark different analysis tools and pick the one which fits the best for our scenario.

In particular, it is crucial to analyse results in the context of similar experiments, as such an approach allows to apply tools like PEER which can identify and remove hidden confounders, having a great influence on the eFDR without changing the overall landscape of sensitivity. We have shown, however, that further filtering of FPs is required to obtain acceptable level of eFDR. A cost of an improved specificity is the decreased sensitivity. The good news is, however, that both for the genome wide studies as well as for ones when the top candidates are identified the results have been improved. When we consider the full list of genes called as differentially expressed, both the agreement between sites for the same DEC method as well as the agreement between different DEC methods improves

noticeably with filtering, making analysis results more robust and easier to reproduce across laboratories. The improvement from filtering can also be seen for the top ranked candidates with the strongest expression change. Here we can recommend in general the use of DESeq2 tool for DE calling especially in combination with BitSeq. This combination performed particularly well for the shorter lists of the most highly-ranked 50–200 differentially expressed genes. Different aspects of performance, however, vary across tool combinations. In general, pipelines relying on Tophat2/Cufflinks2 for estimation of expression levels performed the worst, while newer tools such as BitSeq (or kallisto) performed better.

Future work for the full manuscript: For the conference presentation and the full proceedings manuscript, the analysis will be extended to explore DE calling by dedicated methods for BitSeq and Cufflinks, and examine BitSeq DE calling for kallisto bootstrapping results.

Methods In this study the SEQC data set has been used (which is described in details and summarized elsewhere)[2]. Here the sequencing data of samples A and C of six Illumina HiSeq 2000 sites have been used.

The expression profiles of AceView[17] genes has been assessed by selected tools representing the state of the art approaches for expression profiling. The gene expression profiles were assessed in the form of read counts. R-make (based on STAR) and Subread perform the alignment to the genome what is followed by counting the reads which are falling into the gene regions. The TopHat2 with the G option represents the hybrid approach, where reads are first aligned to the virtual transcriptome and then mapped back to the genome. The gene and transcript expressions are then estimated with Cufflinks2 based on the genome based alignments. BitSeq uses directly the transcriptome alignments (here provided by SHRiMP2) to assess the transcripts abundances. These where then sum up per gene to obtain the read count estimates for genes. Kallisto represents the alignment free approach, where transcript abundances are assessed directly from reads based on graph pre-built with use of the transcript sequences. Also here the transcript expression estimates where sum up per gene to obtain the read count estimates for genes.

Gene expression estimates for all samples were used to detect latent variables using PEER package[16]. The covariates associated with sample type were included for inference and the inferred hidden confounders were removed from the signal.

Differential expression analysis has been performed with use of three dedicated R packages: limma, edgeR and DEseq2. In all three cases the suggested way of analysis has been performed (in terms of limma it includes TMM+voom pre-processing). The Benjamini-Hochberg adjustment for multiple testing has been performed. The genes were called differentially expressed when q-val <0.05. When filtering has been applied in addition: gene effect strength has to be higher than 2 (abs(log2FC) > 1) and the Average Expression has to be higher than dedicated threshold. Average expression threshold was selected for each combination of expression estimation and DE calling method separately in order to fix the average intra-site AvsC sensitivity at level of 3000 genes. On average 45th percentile with SD of 2.3 has been used (lowest for limma than DESeq2 and edgeR; lowest for Subread, then kallisto, TH2G, BitSeq and r-make). The same thresholds have been applied to inter-site DE calling. The DE analysis has been focused on down-regulated genes in A versus C comparison, as the strength of the up-regulated signal is limited by design of sample C as 3 parts of A and one part of B.

Overall agreement between lists of differentially expressed genes has been calculated as ratio of lists intersection and lists union. Agreement of topN candidates has been calculated as ratio of intersection of compared topN lists and the N, where differentially expressed candidates have been order by the change strength.

Bibliography

1. Shi, L. *et al*. Nat. Biotechnology 24, 1151–1161 (2006)
2. Su, Z. *et al.* Nature Biotechnology 32, 903–914 (2014)
3. Li, S. *et al.* Nature Biotechnology 32, 888–895 (2014)
4. Dobin, A. *et al*. Bioinformatics 29, 15–21 (2013)
5. Liao, Y. *et al*. Nucleic Acids Res. 41, e108 (2013)
6. Trapnell, C. *et al*. Nat. Biotechnol. 31, 46–53 (2013)
7. Kim, D. *et al*. Genome Biol. 14, R36 (2013)
8. David, M. *et al.* Bioinformatics 27, 1011–1012 (2011)
9. Glaus, P. *et al.* Bioinformatics 28, 1721–1728 (2012).
10. Bray, N. *et al.* arXiv:1505.02710 (2015)
11. Smyth, G.K. *et al.* 397–420 (Springer, New York, 2005).
12. Robinson, M.D. *et al.* Bioinformatics 26, 139–140 (2010).
13. Love, M.I. *et al.* Genome Biology, 15, pp. 550
14. Łabaj, P.P. *et al*. Bioinformatics 27, i383–i391 (2011)
15. Leek, J.T. *et al*. Bioinformatics 28, 882–883 (2012).
16. Stegle, O. *et al*. PLoS Comput. Biol. 6, e1000770 (2010).
17. Thierry-Mieg, D. *et al.* Genome Biol. 7, S12 (2006).

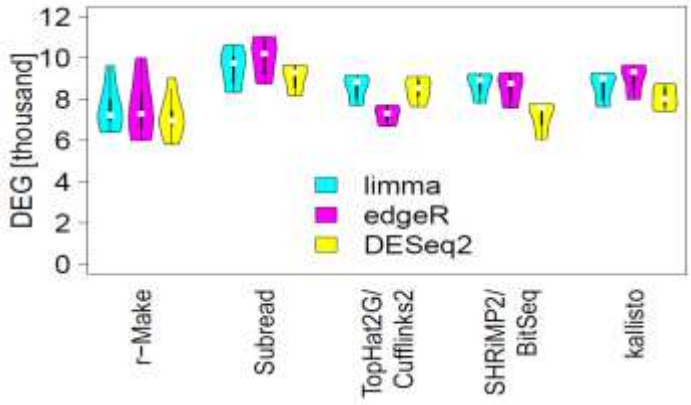

Figure 1. Intra-site differential expression call sensitivity. For each expression estimate method (x-axis) and each DE calling method (colour) all intra-site A versus C comparisons are presented in a form of the violin plot. Y-axis represents the sensitivity as a number of differentially expressed genes (with q.val<0.05).

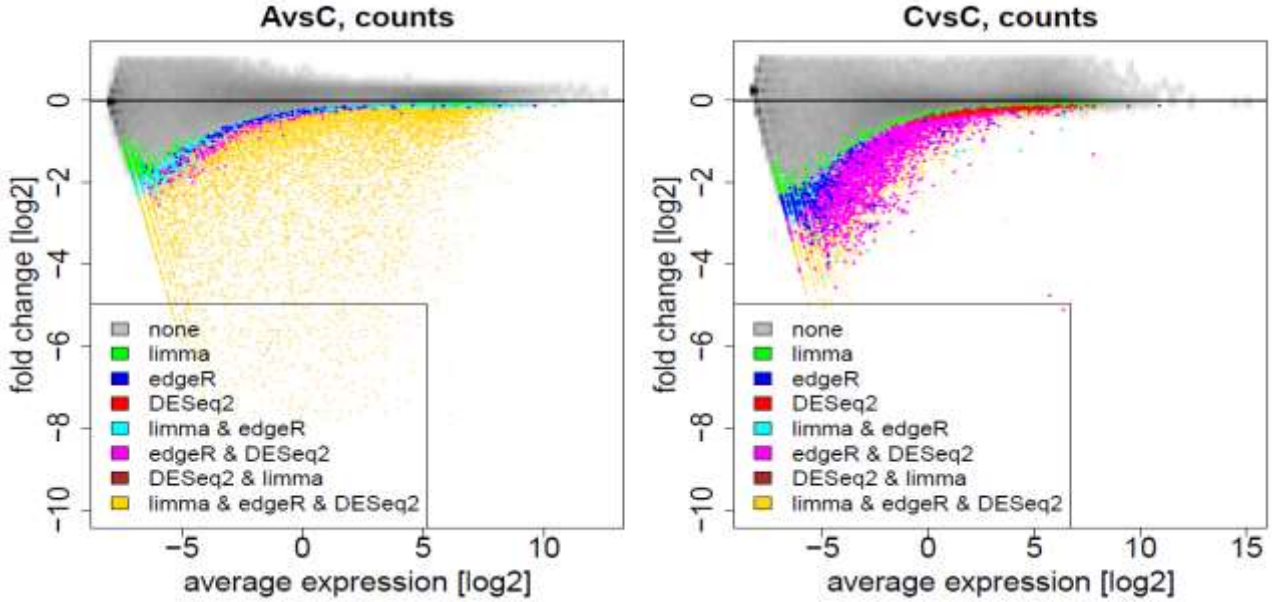

Figure 2. Left panel represents the overlap of the DE calling by different DEC methods for AvsC intra-site comparison, while right panel shows the results for the inter-site AvsA comparison. The overlap between calling as DE by different DEC methods is encoded by different colours. Grey clouds represent not down-regulated genes.

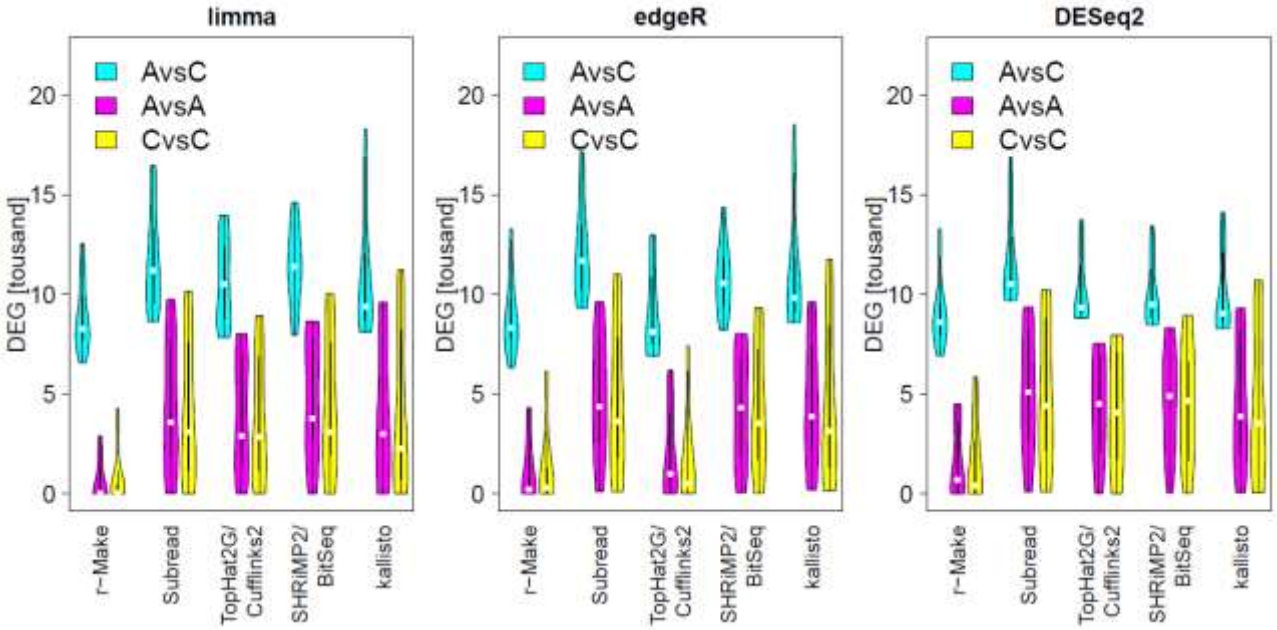

Figure 3. Inter-site differential expression call sensitivity, including false-positives from same–same comparisons. For each expression estimate method (*x*-axis) and each DE calling method (panel) all inter-site A versus C comparisons (cyan) as well as A versus A (magenta) and C versus C (yellow) are presented in a form of violin plots. The same–same comparisons show the sensitivity of methods to picking up false positives.

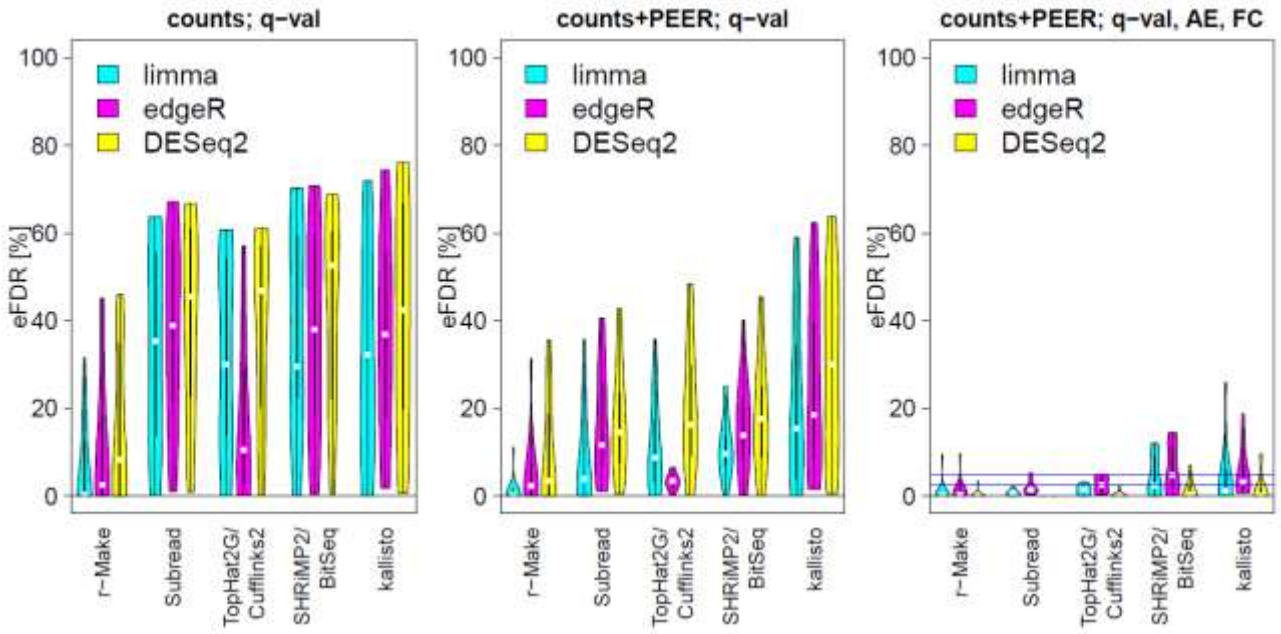

Figure 4. eFDR. For each expression estimate method (x-axis) and each DE calling method (colour) eFDR has been estimated as ratio between inter-site A versus A plus C versus C and A versus C. The left panel represents results based on not corrected counts with DE calling by q-val threshold. In the middle panel hidden confounders have been removed by PEER from count expression estimates. In the right panel additional DE calling filters has been applied (as described in methods).

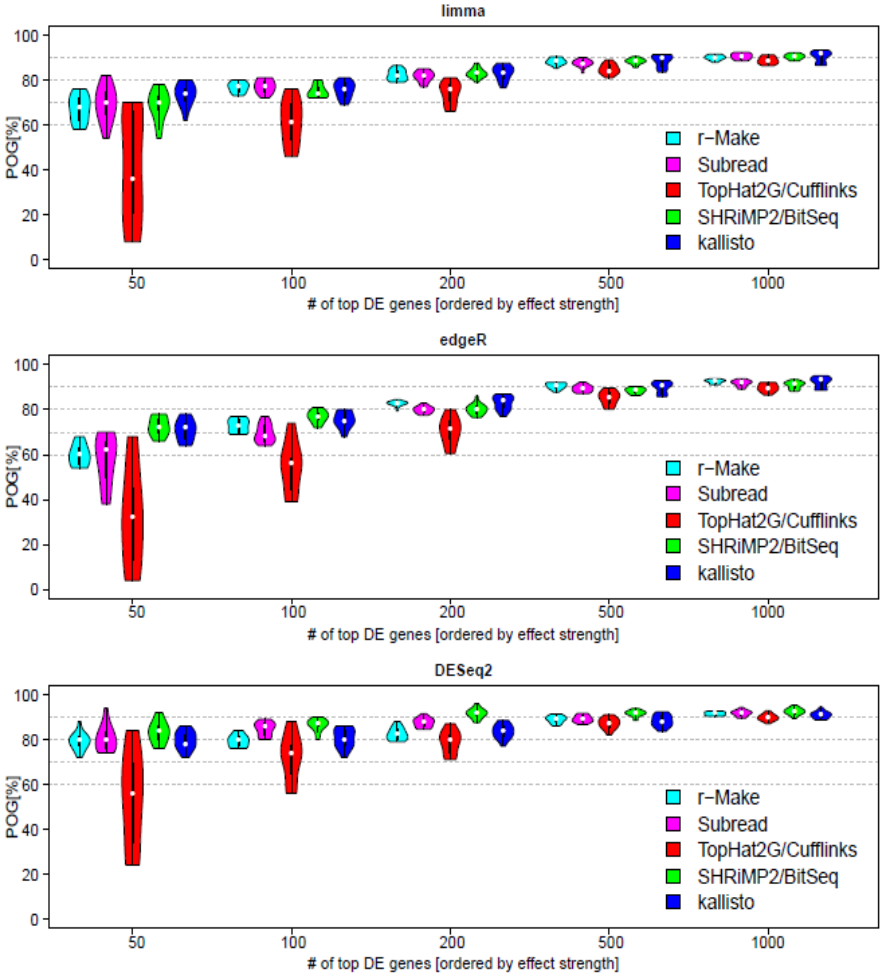

Figure 5. Inter-site reproducibility of differential expression calls. Comparing the identities and the directions of change for DEGs across sites, agreement is plotted for lists including the top-ranked genes as sorted by effect size (*x*-axis). The observed response violin plots depend on expression estimate pipeline, DE calling pipeline and filter choice, showing more variation and lower agreement levels for shorter lists. Results for BitSeq and DESeq2 seems to be the most robust. Agreement for top1000 genes cross 90% irrespectively from the pipeline choices.
Presented results were obtained based on expression estimates after removing the hidden confounders by PEER. For DE calling additional filters for average expression and effect strength have been applied.