DEUS: an R package for accurate small RNA profiling based on differential expression of unique sequences

Tim Jeske [1,2,], Peter Huypens [3,4], Laura Stirm [4,5], Selina Höckele [3,4], Christine M. Wurmser [6], Anja Böhm [4,5], Cora Weigert [4,5,7], Harald Staiger [3,4,5,8], Christoph Klein [2], Johannes Beckers [3,4,9], and Maximilian Hastreiter [10,11,*]*

[1]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München GmbH, 85764 Neuherberg, Germany

[2] Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, 80337 München, Germany

[3] Institute of Experimental Genetics, Helmholtz Zentrum München GmbH, 85764 Neuherberg, Germany

[4]German Center for Diabetes Research (DZD), 85764 Neuherberg, Germany

[5] Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Zentrum München at the University of Tübingen, 72076 Tübingen, Germany

[6]Chair of Animal Breeding, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85354 Freising, Germany

[7]Division of Pathobiochemistry and Clinical Chemistry, Department of Internal Medicine IV, University Hospital Tübingen, 72076 Tübingen, Germany

[8] Institute of Pharmaceutical Sciences, Department of Pharmacy and Biochemistry, University Tübingen, 72076 Tübingen, Germany

[9]Chair of Experimental Genetics, Technische Universität München, Wissenschaftszentrum Weihenstephan, 85354 Freising, Germany

[10]Institute of Computational Biology, Helmholtz Zentrum München GmbH, 85764 Neuherberg, Germany

[11] Chair of Genome-oriented Bioinformatics, Technische Universität München, Wissenschaftzentrum Weihenstephan, 85354 Freising, Germany

* To whom correspondence should be addressed. Tel: +49 89 3187 3581; Fax: +49 89 3187 3585; Email: tim.jeske@helmholtz-muenchen.de (TJ), m.hastreiter@tum.de (MH)

"The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors".

ABSTRACT

The accurate analysis of small RNA populations remains a challenging task. Major obstacles arise when these short RNA sequences map to multiple locations in the genome, align to regions that are not annotated or underwent post-transcriptional changes which hamper accurate mapping. In order to tackle these issues, we present a novel profiling strategy that circumvents the need for read mapping to a reference genome and utilizes the actual read sequence as a primary identifier. After differential expression analysis of sequence counts, the sequences are clustered by similarity and annotated against user defined feature databases. This strategy enables a more comprehensive and concise representation of the small RNA population without any data loss or data convolution. We validated our pipeline on 150 samples from various mouse and human biomaterials and show that data convolution and data loss are substantial when employing mapping-based methods. Interestingly, we observed substantially higher percentages of multi-mapping reads in biomaterials with a potential carrier function, such as plasma, sperm and exosomes when compared to somatic cell types and colorectal cancer cell lines. Our pipeline is implemented as an open-source R package and freely available at http://ibis.helmholtz-muenchen.de/deus/.

INTRODUCTION

The growing impact of next generation sequencing (NGS) technologies has been particularly noteworthy in the discovery and characterization of a plethora of small non-coding RNAs (sncRNA). These sncRNAs, including microRNAs (miRNA) (1), piwi-associated RNAs (piRNA) (2,3), small nucleolar RNAs (snoRNA) (4,5) and tRNA fragments (tRFs) (6,7), have been implicated in the regulation of cellular processes, such as gene expression, alternative splicing, silencing of transposable elements, chromatin remodeling and nucleic acid methylation. Besides their prominent role in the regulation of genome integrity and functionality, these sncRNAs can be released as exosomes able to affect gene expression in distal tissue cell types (8,9). In addition, these sncRNAs have also been implicated as a carrier of paternally acquired information able to affect the offspring's metabolic health outcome (10-12). The emerging role of sncRNAs as a carrier of epigenetic information between neighboring and more distal cell types and from one generation to the next holds great promise for the development of potential diagnostic biomarkers and epigenetic treatment strategies for complex disorders such as obesity, type 2 diabetes and cancer (13,14). Evidently, in order to gain more insight in the biological relevance of this expanding universe of sncRNAs, we also continuously need to improve our analysis strategies in ways that allow a comprehensive but concise representation of the expressed sncRNA sequences which often can undergo a myriad of posttranscriptional modifications, such as cleavage (15,16), trimming (17,18), elongation (19,20) and editing (21).

Here, we present our method that analyzes differential expression of unique sequences (DEUS) for profiling sncRNA sequence data. DEUS differs from most sncRNA analysis tools as it does not rely on read mapping and conducts annotation after differential expression analysis on the read counts. Using 150 samples from various mouse and human biomaterials, we provide evidence that data convolution and data loss are considerable when using a mapping-based strategy and provide several arguments how this can easily be avoided by using DEUS.

MATERIAL AND METHODS

Implementation of Differential Expression of Unique Sequences (DEUS)

The input reads for small RNA profiling by DEUS should resemble the actual sncRNA sequences like they were isolated from the biological samples. For this reason, we recommend to remove all reads with a length that equals or exceeds the sequencing length after adapter trimming. Next, we use the ShortRead (22) package to identify unique reads in each of the input FASTQ files. This step generates a typical RNA-seq count matrix, but utilizes the actual read sequence instead of the gene feature as identifier. The count table is then used as input for the DESeq2 (23) analysis to calculate statistically significant read count differences between samples from different experimental conditions. Adjusted p-values for the differentially expressed (DE) unique sequences are calculated using the Independent Hypothesis Weighting (IHW) method (24), which increases statistical power compared to the Benjamini-Hochberg method by taking data-driven weights into account. DE unique sequences are subsequently annotated by BLASTn (25) searches against user defined BLAST databases. Subsequently, the CD-Hit clustering algorithm (26,27) is applied to classify significant DE reads into subgroups of highly similar sequences. During the clustering process, reads are sequentially processed and become either classified as a member of an existing cluster or as a new cluster representative. Finally, a comprehensive summary table is generated by combining results from differential expression analysis, BLASTn annotation and cluster assignment. To easily explore the content of the table the user can define an individual set of terms that represent feature classes of interest. The given terms will be integrated as columns each containing the number of BLAST hits that match the corresponding term. DEUS also automatically generates a PCA plot, a sample distance map and a MA plot. The latter shows the log2 fold changes (M) of each DE sequence versus the mean of its normalized counts (A), allowing easy and fast visual inspection of the fold change and expression value distributions. We implemented each of the described steps as a customizable function in the R package DEUS. This modular design allows the user to customize our pipeline, tailored to the specific needs of the project.

Library preparation

Total RNA served as input material for small RNA library preparation using NEBNext Small RNA Library Prep Set for Illumina (New England BioLabs) according to the manufacturer’s guidelines. Following a brief denaturation step, 3′ adapters were ligated to the input RNA for 1 hour, followed by hybridization of the reverse transcription primers and subsequent ligation of the 5′ adapters. Next, reverse transcription was performed using ProtoScript II for 1 hr at 50°C and the sequence and index primers were added by PCR amplification for 11-15 cycles using LongAmp Taq 2× master according to the manufacturer’s recommendations.

Illumina 2500 HiSeq sequencing

The libraries were sequenced on a HiSeq2500 (Illumina inc., San Diego, CA, USA) using HiSeq Rapid v2 chemistry (Illumina). Raw data was collected by the Illumina HiSeq Control Software (version 2.2.58). Illumina Real-Time Analysis tool (version 1.18.64) was used for image analysis and base calling. The single-end sequence reads with 50bp read length were demultiplexed and FASTQ files were generated with CASAVA BCL2FASTQ Conversion Software (version 1.8.3).

Publically available data sources

sncRNA sequencing data derived from human colorectal cancer cell lines and their released exosomes (28) was obtained from NCBI GEO (29,30), dataset accession number: GSE67004, ID:200067004.

Read trimming

Trim Galore (v0.0.4, http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to trim reads of all analyzed data sets in two steps. In the first step, all adapters were removed and only reads shorter than the maximum sequencing length were kept. In the second step, reads were trimmed for a minimum quality threshold of 20.

Read mapping

To compute the proportion of reads mapping to multiple genomic locations (MMAP fraction), we used STAR v2.5.2a (40) to map the trimmed reads to the primary assembly of the human reference genome GRCh38 for human data sets and to the mouse reference genome GRCm38 for mouse data sets. STAR was configured to match at least 16 nucleotides with a maximum of 5% mismatches over the mapped length having the splicing function switched off.

Read annotation

We applied our DEUS pipeline to all data sets and computed the number of reads that had no BLAST hit (NA fraction). We used the NCBI BLASTn 2.6.0+ and manually compiled available sncRNA databases to generate comprehensive BLAST databases. For human data sets we included the Ensembl CDS and ncRNA databases (31), DASHR, the database of small human noncoding RNAs (32), and the RetrogeneDB (33). Overall, the human database comprises a total of 193,634 sequences. For mouse data sets, Ensembl CDS and ncRNA databases (31), miRBase (34), RetrogeneDB (33), piRNABank (35) and GtRNAdb (36,37) were used, resulting in a total of 126,153 sequences.

Statistical analysis

The statistical analysis in Figure 2 was performed using GraphPad Prism (GraphPad Software, La Jolla California USA). Data represent mean ± SD. The box and whiskers graphs show the 25th to 75th percentiles, with whiskers extending from the smallest up to the largest value. The line plotted in the middle of the box represents the median. One-way ANOVA, post-hoc Bonferroni multiple comparisons test, with a confidence interval of 99.9% was used to evaluate statistical significance between groups, showing the least significant p-value in the comparison between samples belonging to somatic cell types and colorectal cell lines on one hand and carrier sncRNA samples such as plasma, sperm and exosomes on the other.

RESULTS

Major differences between mapping-based and DEUS small RNA profiling strategies

DEUS deviates from mapping-based small RNA profiling methods in several aspects (Figure 1A). Most notably, DEUS circumvents the need for read mapping and, therefore, facilitates sncRNA profiling even when a reference genome is not available. Instead, DEUS starts with the identification of unique sequences and employs the actual nucleotide sequence as an identifier in the downstream analysis. As a consequence, DEUS does not discard reads that map to regions of the genome that are not annotated (NA reads) (Figure 1B). DEUS annotation does not merge read counts as opposed to mapping-based profiling methods where actual read counts become abstracted to feature counts (Figure 1A). Such praxis could theoretically convolute the biological interpretation of the data, especially when reads would map to multiple genomic locations or when reads map to different spatial coordinates of the same genomic feature as previously reported for tRfs (6) (Figure 1B). In contrast, DEUS does can provide multiple annotations per unique sequence without causing any data convolution.

Similarly, reads that display discrete sequence or length variations may also pose problems when analyzed by mapping-based tools as, depending on the stringency criteria used for read mapping, these reads either become a gateway to data convolution or are no longer retained in the analysis (Figure 1B). DEUS, on the other hand, groups highly similar sequences using the CD-Hit clustering algorithm (Figure 1A), allowing feature-based data interpretation without losing any input reads. Additionally, aligned sequence clusters visualize discrete sequence variations in an unrivalled fashion. Depending on the underlying data set, the clustering of multiple sequences into similarity clusters leads to a significant data compression. During our analysis we observed compression rates ranging from about 40% up to 80%. Although the initial order of the sequences will have an effect on the number of clusters, we did only observe minor deviations (< 3%) in compression rate when randomizing the order of sequences before clustering. Table 1 shows a typical DEUS output file for two clusters. The ten sequences belonging to cluster 807, display several sequence variations of miR-340, some of which are regulated in the opposite direction as compared to the original miR sequence. The lower panel shows a smaller cluster of four sequences that are a typical example of multi-mapping reads as they are annotated both as piRNAs and tRNAs with 100% sequence identity.

Estimation of data loss and data convolution using mapping-based profiling strategies

In order to estimate data convolution and data loss using mapping-based profiling strategies, we collected small RNA sequence data from human plasma (n=19), human blood mononuclear cells (PBMCs, n=16), human myocytes (n=20), human sperm (n=14), human colorectal cancer cell lines (n=9), human exosomes (n=9), mouse sperm (n=17) and mouse B-cells (n=46). After adapter trimming we mapped the reads of these 150 samples to determine the MMAP fraction allowing 5% mismatches and computed the fraction of NA reads using BLAST based on our compiled sncRNA databases requiring 100% sequence identity. This analysis revealed a weighted average of 61.5 ± 20.1% MMAP reads for the various biomaterials (Supplementary Table 1), supporting the notion that data convolution is substantial when applying mapping-based profiling methods. Interestingly, we found that the percentage of MMAP reads is significantly higher in biomaterials with potential carrier function, such as plasma, sperm and exosomes when compared to somatic cell types and colorectal cancer cell lines (Figure 2, 78.1 ± 7.4% vs 44.9 ± 12.4%, respectively; One-way ANOVA, post-hoc Bonferroni, $p<0.001$). Next, we applied DEUS to these 150 samples in order to estimate the percentage of NA reads. We computed the percentage of NA reads using stringent BLAST criteria against our compiled database as detailed in the material and method section. We observed a weighted average of 44.7 ± 17.2% NA reads for these 150 samples (Supplementary Table 1). The percentage of NA reads was notably higher in mouse as compared to human samples (65.8 ± 7.6% for mouse versus 37.6 ± 12.9% for human samples) reflecting the difference in the number of sncRNA sequences in our mouse and human annotation databases. When we restricted the comparison of NA read fractions to human samples, we observed on average 47.1 ± 11.1% NA reads in human plasma, sperm and exosomes compared to 28.2 ± 4.5% in somatic cell types and human colorectal cancer cell lines (Supplementary Table 1). Taken together, these findings indicate that mapping-based inflicted read count convolution and data loss may be more considerable when dealing with such biological materials.

DISCUSSION

A general approach to analyze sncRNA data encompasses the evaluation of differential expression between conditions of interest. For this purpose, several software packages, such as miRDeep (38), tDRmapper (39), sRNAnalyzer (40) and sRNAtoolbox (41), have been developed. A common step shared by these sncRNA profiling tools is the alignment of reads to a reference genome, followed by their annotation, feature count quantification and the subsequent statistical evaluation between experimental conditions (42). However, the analysis of the expressed sncRNA populations poses several hurdles because these short reads are more likely to map to multiple locations in the genome, or map to genomic coordinates that are not annotated and may deviate from the originating feature sequence due to editing and post-transcriptional processing steps. As such, the use of feature counts for differential expression analysis provides a gateway to data convolution and data loss as read counts specific to feature variants such as 5' and 3' coverage of tRfs (6), isomiRs (20,43,44) and shortened piRNAs (45,46) are usually pooled together when using mapping-based strategies. The general consensus regarding MMAP reads is that these should not be discarded from the analysis due to the inherent loss of considerable amounts of information (47). Besides *at random* assignment of a feature to a MMAP read, other strategies include the assignment of either absolute or probabilistic read counts to each of the MMAP features. Evidently, either of these feature count strategies has substantial impact on the downstream analysis and its biological interpretation. For this reason, sncRNA analysis generally applies stringent mapping criteria in order to limit the number of MMAP reads. The inevitable trade-off is the inherent loss of reads that discretely differ from the genomic feature sequence. Data loss inflicted by mapping-based strategies also occurs when reads are mapped to genomic regions that are not annotated.

Therefore, we developed a simple alternative method which provides an excellent compromise to address many of these problems which are inherent to mapping-based strategies. DEUS conducts differential expression analysis based on the unique sequences and their respective read counts. Consequently, DEUS does not cause any data loss or data convolution due to NA and MMAP reads. Another major advantage of DEUS is the use of a final clustering step which provides unique insight into the potential sequence variations among members of the same cluster sequence, allowing their representation in an unprecedented and more concise manner. In accord with previous studies, we show that sncRNA data sets from various mouse and human biomaterials are plagued by substantial amounts of MMAP reads (>50%) and noticeable amounts of NA reads (~40%) (48). We also provide evidence that the percentages of MMAP and NA reads are typically higher in biomaterials with potential carrier function, such as plasma, sperm and exosomes in comparison to somatic cell types and cancer cell lines. This is relevant as these biomaterials with potential carrier function have been implicated in the onset of systemic diseases such as obesity and type 2 diabetes and, therefore, may serve as an interesting source for the discovery of novel biomarkers as well as the development of prevention and treatment strategies (8-10,12).

In summary, DEUS provides an unprecedented way to profile and visualize sncRNA data. DEUS clearly diverges from mapping-based analysis strategies, hampered by substantial data loss and convolution of feature counts. DEUS circumvents the need for a reference genome and, therefore, facilitates sncRNA profiling in virtually any organism. We believe that our DEUS pipeline considerably improves the analysis of sncRNA-seq data, being applicable in various existing pipelines and returning intuitively interpretable results.

DATA AVAILABILITY

The DEUS R package and accompanying documentation is available at http://ibis.helmholtz-muenchen.de/deus/. Sequencing data on human colorectal cell lines and their released exosomes is available at GEO accession number: GSE67004. The remaining sncRNA sequencing data is available on request as these data sets are part of active research projects. Please direct your data requests to HS (harald.staiger@med.uni-tuebingen.de) for human plasma and human blood mononuclear cells, to CW (cora.weigert@med.uni-tuebingen.de) for human myocytes, to JB (beckers@helmholtz-muenchen.de) for human and mouse sperm and to Marcin Lyszkiewicz for mouse B-cells (marcin.lyszkiewicz@med.uni-muenchen.de).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

AUTHOR CONTRIBUTION

The conceptualization of the DEUS pipeline was done by TJ, PH and MH. TJ and MH performed bioinformatics analysis on all samples, and compiled the software into an R package. CMW conducted high-throughput sequencing. LS, SH, AB, CW, HS, CK and JB provided datasets for the validation of the pipeline. PH wrote the introduction. The remaining text in the manuscript was written by TJ, PH and MH with the input from all authors.

ACKNOWLEDGEMENT

We thank Natalia Zietara, Marcin Lyszkiewicz in the group of Christoph Klein at the Department of Pediatrics at the Dr. von Hauner Children's Hospital for providing RNA-seq data of B-cells.

FUNDING

The work was supported by funding of the Care-for-Rare Foundation to TJ and by grants of the DZD - German Center for Diabetes Research and the Helmholtz Alliance AMPro to JB.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

FIGURES

FIGURE 1.

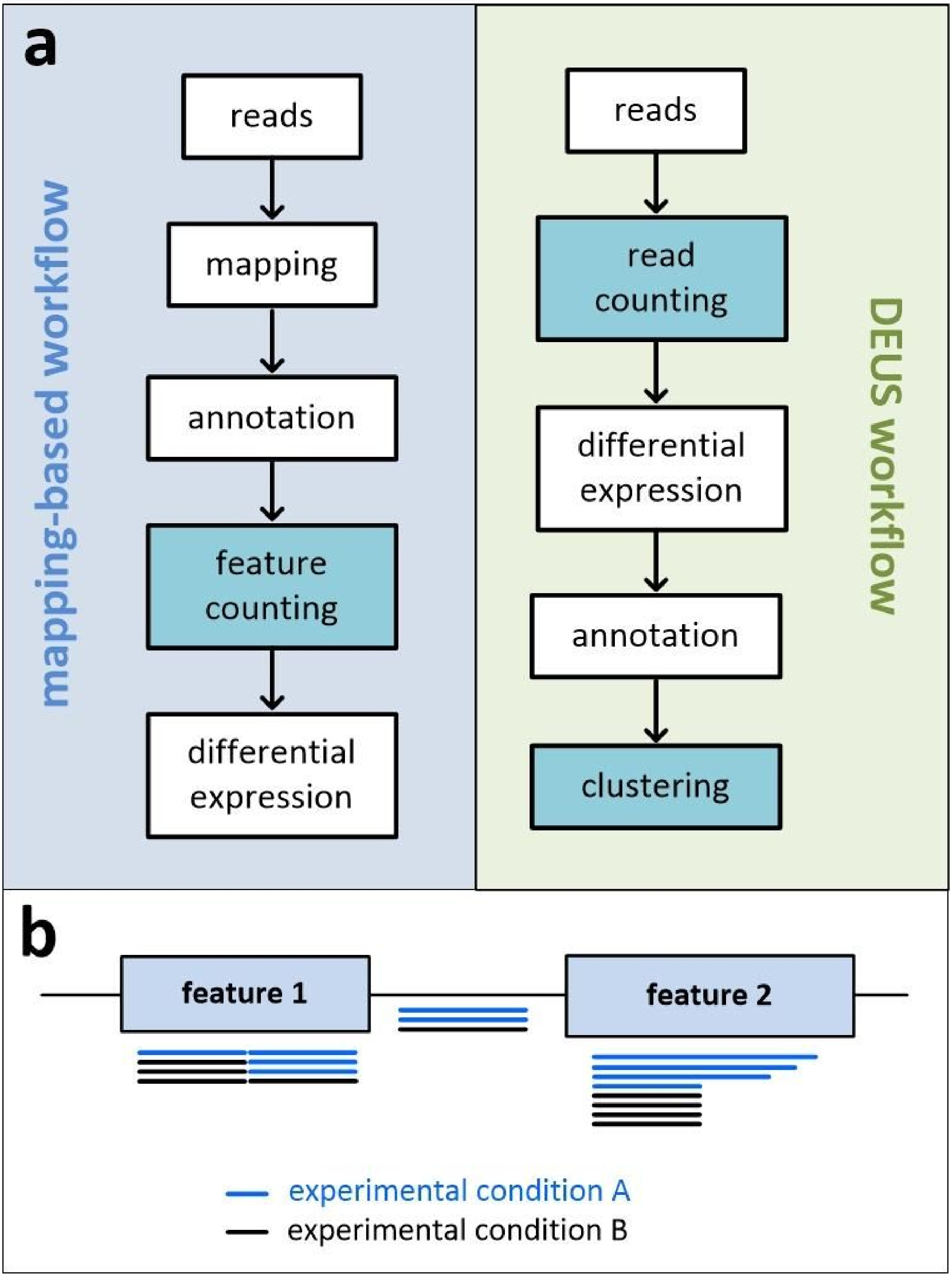
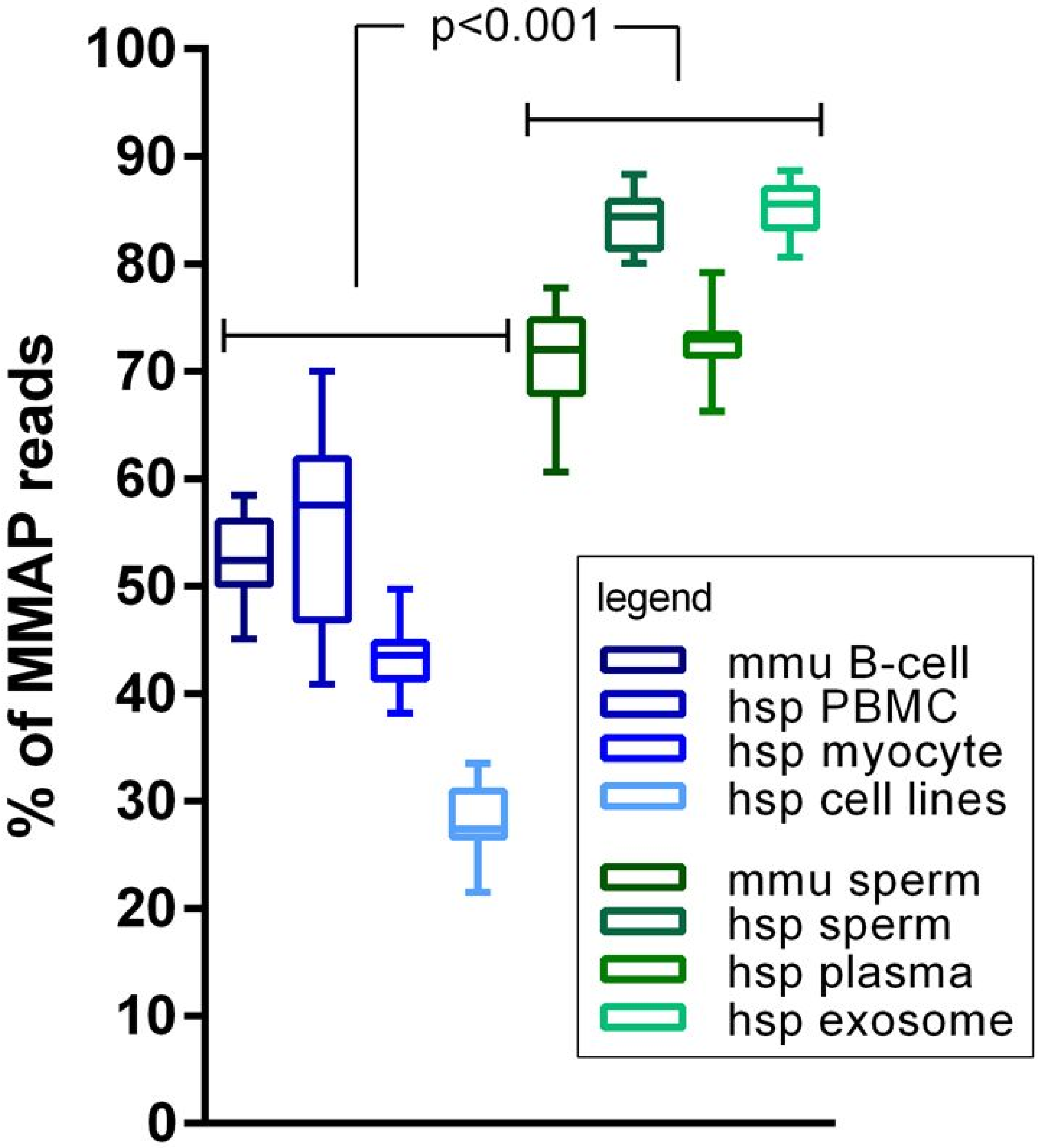


FIGURE 2.



REFERENCES

1. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853-858.
2. Girard, A., Sachidanandam, R., Hannon, G.J. and Carmell, M.A. (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, **442**, 199-202.
3. Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K. and Hannon, G.J. (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science*, **316**, 744-747.

4. Ender, C., Krek, A., Friedlander, M.R., Beitzinger, M., Weinmann, L., Chen, W., Pfeffer, S., Rajewsky, N. and Meister, G. (2008) A human snoRNA with microRNA-like functions. *Mol Cell*, **32**, 519-528.

5. Taft, R.J., Glazov, E.A., Lassmann, T., Hayashizaki, Y., Carninci, P. and Mattick, J.S. (2009) Small RNAs derived from snoRNAs. *RNA*, **15**, 1233-1240.

6. Lee, Y.S., Shibata, Y., Malhotra, A. and Dutta, A. (2009) A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev*, **23**, 2639-2649.

7. Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A.Z. and Kay, M.A. (2010) Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, **16**, 673-695.

8. Ying, W., Riopel, M., Bandyopadhyay, G., Dong, Y., Birmingham, A., Seo, J.B., Ofrecio, J.M., Wollam, J., Hernandez-Carretero, A., Fu, W. *et al.* (2017) Adipose Tissue Macrophage-Derived Exosomal miRNAs Can Modulate In Vivo and In Vitro Insulin Sensitivity. *Cell*, **171**, 372-384 e312.

9. Thomou, T., Mori, M.A., Dreyfuss, J.M., Konishi, M., Sakaguchi, M., Wolfrum, C., Rao, T.N., Winnay, J.N., Garcia-Martin, R., Grinspoon, S.K. *et al.* (2017) Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*, **542**, 450-455.

10. Chen, Q., Yan, W. and Duan, E. (2016) Epigenetic inheritance of acquired traits through sperm RNAs and sperm RNA modifications. *Nat Rev Genet*, **17**, 733-743.

11. Grandjean, V., Fourre, S., De Abreu, D.A., Derieppe, M.A., Remy, J.J. and Rassoulzadegan, M. (2015) RNA-mediated paternal heredity of diet-induced obesity and metabolic disorders. *Sci Rep*, **5**, 18193.

12. Zhang, Y., Zhang, X., Shi, J., Tuorto, F., Li, X., Liu, Y., Liebers, R., Zhang, L., Qu, Y., Qian, J. *et al.* (2018) Dnmt2 mediates intergenerational transmission of paternally acquired metabolic disorders through sperm small non-coding RNAs. *Nat Cell Biol*, **20**, 535-540.

13. Czech, M.P., Aouadi, M. and Tesz, G.J. (2011) RNAi-based therapeutic strategies for metabolic disease. *Nat Rev Endocrinol*, **7**, 473-484.

14. Anastasiadou, E., Jacob, L.S. and Slack, F.J. (2018) Non-coding RNA networks in cancer. *Nat Rev Cancer*, **18**, 5-18.

15. Kawaji, H., Nakamura, M., Takahashi, Y., Sandelin, A., Katayama, S., Fukuda, S., Daub, C.O., Kai, C., Kawai, J., Yasuda, J. *et al.* (2008) Hidden layers of human small RNAs. *BMC Genomics*, **9**, 157.

16. Thompson, D.M., Lu, C., Green, P.J. and Parker, R. (2008) tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA*, **14**, 2095-2103.

17. Zhang, Y., Guo, R., Cui, Y., Zhu, Z., Zhang, Y., Wu, H., Zheng, B., Yue, Q., Bai, S., Zeng, W. *et al.* (2017) An essential role for PNLDC1 in piRNA 3' end trimming and male fertility in mice. *Cell Res*, **27**, 1392-1396.

18. Ding, D., Liu, J., Dong, K., Midic, U., Hess, R.A., Xie, H., Demireva, E.Y. and Chen, C. (2017) PNLDC1 is essential for piRNA 3' end trimming and transposon silencing during spermatogenesis in mice. *Nat Commun*, **8**, 819.

19. Morin, R.D., O'Connor, M.D., Griffith, M., Kuchenbauer, F., Delaney, A., Prabhu, A.L., Zhao, Y., McDonald, H., Zeng, T., Hirst, M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, **18**, 610-621.

20. Zhou, H., Arcila, M.L., Li, Z., Lee, E.J., Henzler, C., Liu, J., Rana, T.M. and Kosik, K.S. (2012) Deep annotation of mouse iso-miR and iso-moR variation. *Nucleic Acids Res*, **40**, 5864-5875.

21. Kawahara, Y., Zinshteyn, B., Sethupathy, P., Iizasa, H., Hatzigeorgiou, A.G. and Nishikura, K. (2007) Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science*, **315**, 1137-1140.

22. Morgan, M., Anders, S., Lawrence, M., Aboyoun, P., Pages, H. and Gentleman, R. (2009) ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics*, **25**, 2607-2608.

23. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

24. Ignatiadis, N., Klaus, B., Zaugg, J.B. and Huber, W. (2016) Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat Methods*, **13**, 577-580.

25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

26. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658-1659.

27. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150-3152.

28. Cha, D.J., Franklin, J.L., Dou, Y., Liu, Q., Higginbotham, J.N., Demory Beckler, M., Weaver, A.M., Vickers, K., Prasad, N., Levy, S. *et al.* (2015) KRAS-dependent sorting of miRNA to exosomes. *Elife*, **4**, e07197.

29. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res*, **39**, D1005-1010.

30. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res*, **41**, D991-995.

31. Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Giron, C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res*, **46**, D754-d761.

32. Leung, Y.Y., Kuksa, P.P., Amlie-Wolf, A., Valladares, O., Ungar, L.H., Kannan, S., Gregory, B.D. and Wang, L.S. (2016) DASHR: database of small human noncoding RNAs. *Nucleic Acids Res*, **44**, D216-222.

33. Kabza, M., Ciomborowska, J. and Makalowska, I. (2014) RetrogeneDB--a database of animal retrogenes. *Mol Biol Evol*, **31**, 1646-1648.

34. Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, **42**, D68-73.

35. Sai Lakshmi, S. and Agrawal, S. (2008) piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res*, **36**, D173-177.

36. Chan, P.P. and Lowe, T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res*, **44**, D184-189.

37. Chan, P.P. and Lowe, T.M. (2009) GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res*, **37**, D93-97.

38. Friedlander, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S. and Rajewsky, N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, **26**, 407-415.

39. Selitsky, S.R. and Sethupathy, P. (2015) tDRmapper: challenges and solutions to mapping, naming, and quantifying tRNA-derived RNAs from human small RNA-sequencing data. *BMC Bioinformatics*, **16**, 354.

40. Wu, X., Kim, T.K., Baxter, D., Scherler, K., Gordon, A., Fong, O., Etheridge, A., Galas, D.J. and Wang, K. (2017) sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. *Nucleic Acids Res*, **45**, 12140-12151.

41. Rueda, A., Barturen, G., Lebron, R., Gomez-Martin, C., Alganza, A., Oliver, J.L. and Hackenberg, M. (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res*, **43**, W467-473.

42. Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W. and Robinson, M.D. (2013) Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc*, **8**, 1765-1786.

43. Lee, L.W., Zhang, S., Etheridge, A., Ma, L., Martin, D., Galas, D. and Wang, K. (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. *RNA*, **16**, 2170-2180.

44. Tan, G.C., Chan, E., Molnar, A., Sarkar, R., Alexieva, D., Isa, I.M., Robinson, S., Zhang, S., Ellis, P., Langford, C.F. *et al.* (2014) 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res*, **42**, 9424-9435.

45. Berninger, P., Jaskiewicz, L., Khorshid, M. and Zavolan, M. (2011) Conserved generation of short products at piRNA loci. *BMC Genomics*, **12**, 46.

46. Oey, H.M., Youngson, N.A. and Whitelaw, E. (2011) The characterisation of piRNA-related 19mers in the mouse. *BMC Genomics*, **12**, 315.

47. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol*, **17**, 13.

48. Johnson, N.R., Yeoh, J.M., Coruh, C. and Axtell, M.J. (2016) Improved Placement of Multi-mapping Small RNAs. *G3 (Bethesda)*, **6**, 2103-2111.

TABLE AND FIGURES LEGENDS

Table 1. DEUS output files.

Example of two sequence clusters in a typical DEUS output file. The output file lists the actual sequence, sequence ID, log2 fold change, p-value, IHW p-value, average read counts for each experimental group, standard deviation (SD), cluster ID, sequence length, blast e-value and the number of annotation hits for several user specified small RNA species classes in consecutive columns as well as a comma separated feature list.

Supplementary Table 1. Overview of the percentage of MMAP and NA reads.

The table lists the number of reads, their average length after adapter trimming as well as the percentage of MMAP and NA reads for each of the analyzed RNA-seq samples.

Figure 1. Major differences between mapping-based and DEUS small RNA profiling strategies.

(a) Schematic representation of the workflow of mapping-based pipelines compared to DEUS. Left panel: Mapping-based workflows rely on read mapping, followed by feature annotation and statistical evaluation of these feature counts to identify DE features. Right panel: The DEUS pipeline first counts the occurrences of unique sequences and then conducts statistical evaluation on read counts using the actual nucleotide sequence as an identifier. The subsequent sequence annotation and clustering step enables an accurate and comprehensive representation of the data.

(b) Schematic representation of scenarios that result in data convolution or data loss when applying mapping-based sncRNA profiling strategies. Mapping-based workflows ignore reads that map to non-annotated genome regions and foster data convolution as variant-specific read counts are usually summed up during subsequent feature counting even if these reads align at different spatial coordinates of the same genomic feature or exhibit discrete variations in nucleotide sequence or sequence length.

Figure 2. Significant higher abundance of MMAP reads in sperm, plasma and exosomes as compared to somatic cell types and cell lines.

Box and whiskers plot showing the percentage of MMAP reads in mouse (mmu) B-cells, human (hsp) PBMCs, human myocytes, human colorectal cancer cell lines, mouse and human sperm, human plasma and exosomes released from human colorectal cancer cell lines.