OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–14
https://doi.org/10.1093/bib/bbac582
Problem Solving Protocol

CODA: a combo-Seq data analysis workflow

Marta Nazzari, Duncan Hauser, Marcel van Herwijnen, Mírian Romitti, Daniel J. Carvalho, Anna M. Kip and Florian Caiment
Corresponding author. Florian Caiment, Tel.: +31433881218; E-mail: florian.caiment@maastrichtuniversity.nl

Abstract

The analysis of the combined mRNA and miRNA content of a biological sample can be of interest for answering several research questions, like biomarkers discovery, or mRNA–miRNA interactions. However, the process is costly and time-consuming, separate libraries need to be prepared and sequenced on different flowcells. Combo-Seq is a library prep kit that allows us to prepare combined mRNA–miRNA libraries starting from very low total RNA. To date, no dedicated bioinformatics method exists for the processing of Combo-Seq data. In this paper, we describe CODA (Combo-seq Data Analysis), a workflow specifically developed for the processing of Combo-Seq data that employs existing free-to-use tools. We compare CODA with exceRpt, the pipeline suggested by the kit manufacturer for this purpose. We also evaluate how Combo-Seq libraries analysed with CODA perform compared with conventional poly(A) and small RNA libraries prepared from the same samples. We show that using CODA more successfully trimmed reads are recovered compared with exceRpt, and the difference is more dramatic with short sequencing reads. We demonstrate how Combo-Seq identifies as many genes and fewer miRNAs compared to the standard libraries, and how miRNA validation favours conventional small RNA libraries over Combo-Seq. The CODA code is available at https://github.com/marta-nazzari/CODA.

Keywords: CODA, Combo-Seq, exceRpt, RNA-Seq, mRNA, miRNA

Introduction

The analysis of the RNA content of a biological sample, referred to as transcriptomics, has become a routine practice for many fields of biology. The transcriptome comprises several types of RNA, called biotypes [1], whose composition varies depending on the type of sample or cell model [2–5]. Some of the most frequently studied biotypes are messenger RNAs (mRNA) and micro RNAs (miRNAs) due to their link with protein expression levels or for their biomarker potential [6–11]. mRNAs generally possess a 3′ poly(A) tail and are usually 1 kilobase or longer [12–14], while miRNAs are short non-coding RNAs that are 20–22 nucleotides long [15].

Currently, simultaneous mRNA and miRNA analysis from the same sample is often performed by preparing separate sequencing libraries for the two RNA species. These libraries follow two very different protocols for selecting the desired RNA: mRNA libraries protocols either perform positive poly(A) selection, capturing all RNA species that possess a 3′ poly(A) tail (so called 'poly(A) libraries'), or perform a negative rRNA selection, by using baits targeting the ribosomal RNA (rRNA) to deplete these species from the total RNA (termed 'ribodepleted libraries'). To sequence the miRNA content of a sample, a small RNA library needs to be prepared, which selects small RNAs by performing a size selection. The separate preparation of two libraries can pose a problem for samples that have very low starting material or RNA content. Moreover, mRNA and miRNA sequencing libraries need to be sequenced on separate flowcells, due to the different number of cycles required (since the insert sizes differ greatly), and by the type of reads generated. In fact, mRNA reads are sequenced paired-end, while short RNA libraries single-end. If longer and shorter fragments were mixed in the same flowcell, the short fragments would tend to outcompete the longer fragments, which would result in the former being overrepresented and latter being undersequenced. When sequencing continues through the full fragment, there can be a sharp decline in base quality and the sequencing run could be potentially aborted [16]. Lastly, when different library prep kits are used, barcodes and barcode collision must be considered to confirm compatibility for multiplexing.

One commercially available library prep kit that aims to overcome these limitations is the NEXTFLEX® Combo-Seq™ library prep kit that allows to prepare combined mRNA/miRNA libraries starting from very little input total RNA (between 5 ng and 100 ng) [17]. In this method, poly(A) RNAs are first selectively retrotranscribed; RNA–DNA hybrids are then digested by RNase H into small fragments; the sample then contains mRNA-derived fragments and short RNAs of comparable length that are further processed in the same way. As such, miRNAs but also other similarly short RNA species, like small nucleolar RNAs (snoRNAs), can be captured. The final library contains sequences of homogeneous size that can be then sequenced in a single flowcell.

Marta Nazzari is a PhD candidate in the Department of Toxicogenomics at the University of Maastricht, the Netherlands.
Duncan Hauser and Marcel van Herwijnen are laboratory technicians in the Department of Toxicogenomics at the University of Maastricht, the Netherlands.
Mírian Romitti is a postdoctoral fellow at the Institute of Interdisciplinary Research in Molecular Human Biology (IRIBHM), Université Libre de Bruxelles.
Daniel J. Carvalho is a PhD candidate in the Department of Instructive Biomaterials Engineering, MERLN Institute for Technology-Inspired Regenerative Medicine at the University of Maastricht, the Netherlands.
Anna M. Kip is a postdoctoral fellow in the Department of Toxicogenomics at the University of Maastricht, the Netherlands.
Florian Caiment is a senior researcher in the Department of Toxicogenomics at the University of Maastricht.
Received: September 2, 2022. **Revised:** November 23, 2022. **Accepted:** November 28, 2022

Recently, Illumina commercialized the v1.5 S4 35 cycle kit [18] for the NovaSeq6000 sequencer, which generates short reads. As the average insert length of a Combo-Seq library is 21–22 nt, it possesses a very suitable length for being sequenced on a 35-cycle flowcell, further reducing the price per sample to study the full transcriptome at very high throughput.

To date, no dedicated pipeline exists to process Combo-Seq generated datasets, and the manufacturer of Combo-Seq recommends using exceRpt [19]. This toolkit was developed for the analysis of extracellular RNAs but can be adapted to the analysis of WGS/exome and long RNA-seq data according to the authors [20]. The published works we could find using Combo-Seq libraries employ exceRpt for their bioinformatics analysis [17,21,22]. As it was not originally developed for Combo-Seq, it presents some problems when used on this type of data: the references use custom gene annotations that group together gene counts based on the biotype, making it a hybrid between a gene and transcript level analysis. They are custom made by the developers and available only for human (hg19 and hg38) and mouse (mm10). The user is thus bound to using them and they cannot be changed or updated, as it is not possible to prepare a reference using a genome version downloaded from common repositories such as Ensembl and Gencode. As a consequence, the applicability of Combo-Seq libraries is reduced to only two species with an outdated reference.

To our knowledge, no independent evaluation of Combo-Seq has been performed, and users could wonder if it provides results comparable to using a combination of poly(A) and miRNA libraries. For example, mRNA is fragmented in a different way (enzymatically in the Combo-Seq protocol or chemically in most poly(A) libraries), and mRNA fragments undergo a different size selection, since Combo-Seq retains short poly(A)-derived fragments, while standard poly(A) libraries usually include 300 nt long inserts and longer (and thus exclude the shorter mRNA fragments from the pool).

In this work, we developed a custom-tailored workflow for the processing of Combo-Seq data which uses existing tools commonly used in RNA-Seq data analysis and compared it to exceRpt. We generated Combo-Seq libraries from two different *in vitro* cell models and sequenced them in 100- or 35-cycle flowcells. We processed them with CODA or exceRpt and noticed how exceRpt discards part of the reads during the trimming step. We show that this is more dramatic as the average read length decreases and it is more biased toward some RNA species. We also provide an evaluation of Combo-Seq performance compared to conventional poly(A) and small RNA libraries prepared from the same RNA samples. We performed differential expression (DE) analysis to compare the dysregulated genes and miRNA after benzo[a]pyrene treatment that can be identified with the different libraries. We show that the DE genes partially overlap between the two types of libraries, while there is no overlap of DE miRNAs. In addition, we performed miRNA RT-qPCR validation to solve discrepancies between conventional small RNA libraries and Combo-Seq quantification.

Methods

Thyroid follicles differentiation and enrichment

Mouse embryonic stem cell-derived thyroid follicles were differentiated and enriched as described previously [23, 24] (see also Supplementary Methods).

Datasets

The datasets used in this paper were obtained from two sources: (1) the human epithelial follicular cell line Nthy-ori 3-1 was seeded at a density of 40 000 cells/cm^2 in a 6-well plate and exposed for 24 h in triplicate to 1 or 10 μM benzo[a]pyrene (BAP) (B-1760, Sigma) dissolved in DMSO (dimethyl sulfoxide,1029521000, Merck) (final concentration of DMSO 0.5%). Six DMSO controls were included. Cells were cultured in RPMI 1640 Medium, GlutaMAX™ Supplement (61870036, Gibco) with 10% FBS (fetal bovine serum) and 100 U/mL penicillin–streptomycin (15140122, Gibco). (2) Enriched thyroid follicles were exposed to DMSO 0.5% for 24 h (five biological replicates). For culture, the differentiation medium (see Supplementary Methods) was supplemented with 8-Br-cAMP (0.3 nM) and TGF-βRI inhibitor SB431542 (10 μM) (1614, Tocris). At the end of the exposure time, cells were lysed in QIAzol Lysis Reagent (79306, Qiagen). Total RNA was extracted using the Direct-zol RNA miniprep (R2051, Zymo Research) for Nthy-ori 3-1 cells and with the miRNAeasy Micro Kit (217084, Qiagen) for the follicles.

Libraries preparation and sequencing

An amount of 50 ng (Nthy-ori 3-1) and 20 ng (thyroid follicles) of total RNA were used as input for the NEXTFLEX® Combo-Seq™ mRNA/miRNA Kit (NOVA-5139-53, PerkinElmer). All RNA integrity number (RIN) values, as calculated by the Agilent software [25], were 8 or higher. tRNA fragments and Y RNA fragments were depleted with NEXTFLEX® tRNA/YRNA blocker. A total of 13 and 16 PCR cycles were performed for Nthy-ori 3-1 and follicles, respectively. Nthy-ori 3-1 samples were sequenced on an S2 Illumina flowcell 100 cycles (v1.5) (Illumina) in single-end mode; follicles samples were sequenced on an S4 Illumina flowcell 35 cycles (v1.5) (Illumina) in single-end mode. Throughout this paper, we will sometimes refer to the RNA-Seq data derived from the Nthy-ori 3-1 and the follicles as '1×100 dataset' and '1×35 dataset', respectively ('1×' means that both libraries were sequenced in single-end mode).

Poly(A) libraries were prepared on an automated system (Zephyr G3® NGS) with the NEXTFLEX® Rapid Directional RNA-Seq Kit 2.0 (NOVA-5198-02, PerkinElmer), NEXTFLEX® Poly(A) Beads 2.0 (NOVA-512992, PerkinElmer) and NEXTFLEX® Unique Dual Index Barcodes (NOVA-512923, PerkinElmer) using 1 μg of total RNA extracted from Nthy-ori 3-1 samples and performing 10 PCR cycles. The libraries were sequenced on an S1 Illumina flowcell 200 cycles (v1.5) (Illumina) in paired-end mode.

miRNA libraries were prepared manually with the NEXTFLEX® Small RNA-Seq Kit v3 (NOVA-5132, PerkinElmer) from 100 ng of total Nthy-ori 3-1 RNA and performing 18 PCR cycles. They were sequenced on an S4 Illumina flowcell 35 cycles (v1.5) (Illumina) in single-end mode.

All prepared libraries were quantified on a Qubit 2.0 Fluorometer (ThermoFisher), and quality control performed on the 2200 TapeStation System (Agilent) or BioAnalyzer 2100 expert (Agilent). The sequencing was done with the NovaSeq 6000 Sequencing System (Illumina).

Data analysis

RNA-Seq data processing

Data from Combo-Seq libraries were processed with exceRpt (v4.6.3, 2018-03-18) or CODA. When using exceRpt, we followed the parameters suggested by PerkinElmer (Supplementary Table S1) [19].

CODA is composed of three steps. Reads are trimmed using Cutadapt (v3.4) [26] and used as input in two different steps: miRNAs are quantified with miRge3.0 (v3.0) [27], while genes using RSEM (v1.3.3) with the —STAR option (v2.7.9a). The STAR parameters that are hard-coded in RSEM follow the ENCODE3's STAR-RSEM pipeline [28], and we opted for these options over the default STAR because they allow read alignment to more loci and permit fewer base mismatches than the default (Supplementary Table S2) [29, 30]. The primary assemblies of the human (GRCh38) and mouse (GRCm39) genomes were downloaded from Gencode (https://www.gencodegenes.org/) [31]. For miRNA detection, the human (v22) and mouse (v22) annotations were obtained from miRBase (https://www.mirbase.org/) [32]. BBMap (v38.94) [33], FastQC (v0.11.5) [34] and multiQC (v1.11) [35] are used for quality control. The CODA code is available at https://github.com/marta-nazzari/CODA.

Demultiplexed data from small RNA libraries were processed as suggested by PerkinElmer [36] and used as input for miRge3.0. Demultiplexed data from poly(A) libraries were processed using a modified version of the Omics Data Analysis Framework for regulatory application (R-ODAF) [37, 38]: reads were trimmed with fastp [39], and aligned and mapped using RSEM with the —STAR option.

Pipelines comparison

exceRpt performs adapter trimming and reads size selection in several steps. First, the 3′ adapter sequence is removed, followed by a combined 5' 4 N adapter trimming and exclusion of the 5′ trimmed inserts shorter than 15 nt. As only reads that are 3′ adapter trimmed are reported in the summary statistics by exceRpt, we retrieved the number of reads passing both trimming and size-selection filters from the output .log files. The count of reads trimmed by CODA is the 'Reads passing filters' statistic output by Cutadapt.

The length distribution of trimmed reads was retrieved from the summary file 'exceRpt_ReadLengths.txt' for the exceRpt pipeline, or after running FastQC for CODA.

To compare CODA and exceRpt, we analysed the genes and miRNA counts separately. Since exceRpt uses custom-made annotations, a one-to-one feature comparison with the Gencode and miRge3.0 ones is not possible. For this reason, we summed all biotypes of the same gene in a single count for exceRpt (see also Supplementary Methods). For CODA, we used the gene counts output by RSEM. In addition, as exceRpt filters out reads mapping to all primary endogenous rRNA genes, we mapped all genes to the corresponding biotype using the biomaRt R package [40, 41] and removed all rRNA counts from both datasets. To compare miRNA counts, we kept only the annotations that overlap between the miRge3.0 and exceRpt outputs, as trying to manipulate the annotations to match the discordant annotations could introduce a bias. The differences between the two workflows are reported in detail in Supplementary Table S3.

Principal component analysis and correlation analysis

Principal component analyses were performed on variance-stabilized expression levels of normalized gene and miRNA read counts using the R package PCAtools (v2.4.0) [42]. Pearson correlation was used to calculate the correlation between genes read count. To calculate the miRNA expression correlation among samples prepared with Combo-Seq and small RNA library prep kit, we ranked the miRNAs based on level of expression (miRNA with the highest read count = highest rank). When multiple miRNAs had the same read count, they were assigned the same rank with the highest value. miRNAs for which the read count was 0 in all samples were removed. To evaluate the correlation between miRNAs, we used the non-parametric Spearman correlation.

DE and gene ontology analysis

DE analysis was performed with R using the DESeq2 [43] and edgeR [44]. To select relevant DE genes and miRNA, stringent filtering was applied using a modified version of the R-ODAF. Briefly, a gene was considered expressed if its count per million (CPM) value is ≥1 in at least 75% of the replicates of either group (i.e. BAP or DMSO). In addition, DE genes and miRNAs identified by DESeq2 were filtered to remove spurious spikes (for details, see the paper by Verheijen et al. [37]). To increase statistical power, all BAP samples were grouped together and compared to the DMSO controls. Gene ontology (GO) (2021) [45, 46] and Reactome (2022) [47] enrichment analyses were performed using the web-based tool Enrichr [48], and the FDR was set to 0.01.

miRNAs reverse transcription-qPCR

Total RNA from the six Nthy-ori 3-1 DMSO control samples was used for cDNA synthesis with the TaqMan® Advanced miRNA cDNA Synthesis Kit (A28007, Applied Biosystems) according to manufacturer's protocol. The synthesized cDNA was used for qPCR using the TaqMan™ Fast Advanced Master Mix (4444556, Applied Biosystems) and TaqMan™ Advanced miRNA Assay (A25576, Applied Biosystems) following the manufacturer's protocol for hsa-miR-122-5p (477855_mir), hsa-miR-361-3p (478055_mir) and hsa-miR-622 (479106_mir). The program used for the qPCR reaction was 20 s at 95 °C (1 cycle), 3 s at 95 °C–30 s at 60 °C (40 cycles) on a CFX Connect™ Real-Time System (Bio-Rad). Each sample was analysed in four technical replicates. For each sample a technical replicate was retained if its Ct value difference from its closest other replicates was lower than 0.6.

Results

Description of CODA

For the processing of Combo-Seq data, we developed a pipeline named 'CODA' (Combo-Seq Data Analysis), composed of three different steps (Figure 1). The first step uses Cutadapt to trim the 5′ and 3′ adapters and discard reads that are shorter than 15 nt. Since Combo-Seq generates libraries from both polyadenylated RNA species and miRNA, the workflow splits in two and the trimmed files are used as input for gene and miRNA mapping and quantification. To analyse the reads that derive from poly(A)-tailed species, the trimmed reads are aligned to the reference genome with STAR and quantified using RSEM. To identify miRNAs, the trimmed reads are used as input for miRge3.0. The genes and miRNAs count files output by both tools for each sample are then merged into a single table for genes and for miRNAs. The pipeline outputs a report with useful QC metrics that can be inspected by the user.

Comparison of CODA with exceRpt

Trimming and read length distribution

Since the first step of both pipelines is adapter trimming, we evaluated the number of retained trimmed reads and their length distribution. Figure 2A shows the read length distribution of the 1 × 100 samples trimmed with CODA or exceRpt. The maximum read length of samples processed with CODA (97 nt) is longer than the maximum length identified by exceRpt (90 nt). The median number of reads that pass adapter trimming is 89.44% and 90.90% of total input reads when using exceRpt or CODA,

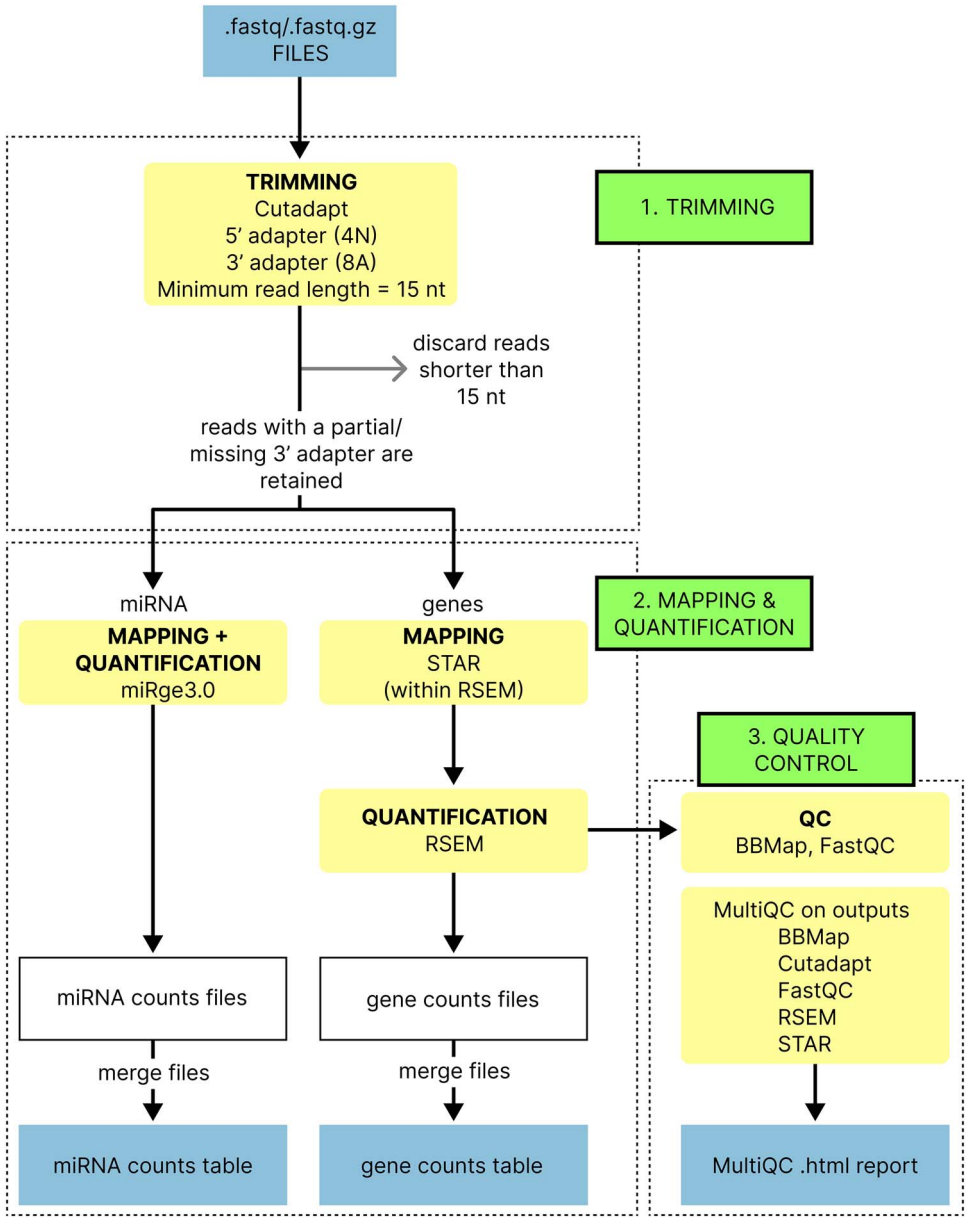

Figure 1. Schematic representation of CODA: fastq or fastq.gz sequencing files are used as input and 5′ and 3′ sequencing adapters are removed using Cutadapt. Reads shorter than 15 nt are also discarded. This trimming step retains also reads with a partial or missing 3′ adapter (a point further discussed in the section 'Trimming and read length distribution'). Mapping and quantification of miRNA is then performed using miRge3.0. Genes mapping and quantification is then performed with RSEM using STAR as aligner (which follows the criteria of the ENCODE3's STAR-RSEM pipeline). As each tool outputs a single file per sample, the count files are then merged into a single table for genes or miRNAs. The last step uses the BBMap suite and FastQC to gather some summary statistics on the trimmed/mapped reads and MultiQC is used to compile all information into a .html report.

respectively (Figure 2B). On average, an additional 1.46% of total raw reads are retained by CODA and their length is between 91 and 97 nt. This increase is even more evident with datasets that have shorter reads, like the one generated from thyroid follicles sequenced on a 1 × 35-cycle flowcell (Figure 2C). The maximum read length when using CODA is 44 nt, while it is 37 nt when using exceRpt. The median percentage of reads that successfully pass 5′ and 3′ adapter trimming when using exceRpt is 73.39% of total sequenced reads. This number increases to 89.12% when the same samples are trimmed with CODA, retaining an additional 15.73% of the raw reads. The difference in maximum read length identified by the two methods likely lies in the choice of trimmer. When sequencing single-end libraries (like Combo-Seq), if a fragment is longer than the total number of cycles of the chosen flowcell, only part of the fragment will be sequenced, thus partly or completely excluding the 3′ adapter (since single-end libraries are always sequenced from the 5′ end). Cutadapt has the option to retain such reads, while the one used by exceRpt with the options set by the exceRpt developers (fastx -M 7) allows only 1 nt mismatch (Supplementary Figure S1) and if the reads are not clipped, they are discarded (determined by the -c flag in the fastx command) [49]. Effectively, exceRpt discards reads where the 3′ adapter is either 6A or shorter or missing altogether. This factor is the most impactful on the trimmed read length distribution, as 1 × 100 dataset reads trimmed with two other trimmers retaining non-clipped reads show a similar distribution as Cutadapt, with a peak at the maximum read length, 97 nt (Supplementary Figure S2) (see also Supplementary Methods).

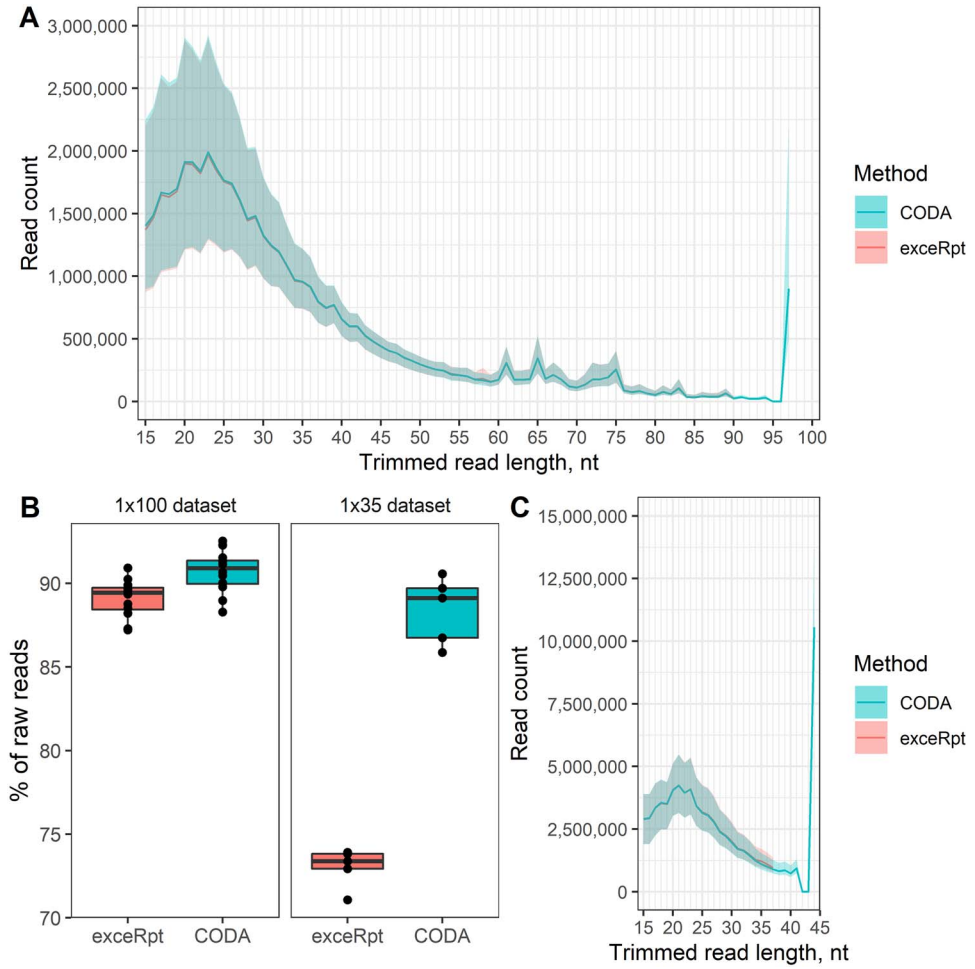

Figure 2. (**A**) Read length distribution of 1 × 100 dataset processed with exceRpt (red) or trimmed with CODA (blue). (**B**) Distribution of trimmed reads expressed as percentage of total raw read count. (**C**) Read length distribution of 1 × 35 dataset processed with exceRpt (red) or trimmed with CODA (blue). For plots A and C, the line represents the average count, while the edges of the shaded area correspond to the highest and lowest count among the replicates.

Comparison of mapping and quantification

Since Combo-Seq libraries capture both poly(A) species and miRNAs, our pipeline performs separately gene and miRNA mapping and quantification. We compared the reads mapping to genes or miRNA using CODA and exceRpt.

The mapping efficiency is comparable, evidenced by similar percentages of trimmed reads mapping to genes or miRNA in both datasets (Table 1). However, since the number of reads successfully passing trimming is higher when using CODA, samples processed with it have more total mapping reads compared to exceRpt, especially in the 1 × 35 dataset, going from a median 52.01 M (exceRpt) to 60.69 M (CODA) reads. The median count of miRNA-mapping reads is instead comparable between the two methods, supporting the observation that, for shorter inserts like miRNAs, the 3′ adapters are fully sequenced, and the reads properly trimmed.

Comparison of genes and miRNAs

To analyse how many and which genes are identified by either pipeline, we first analysed the overlap of the genes which have a raw read count greater than 0 in the 1 × 100 and 1 × 35 datasets processed either with CODA or exceRpt. While the total read count is higher when the samples are processed with CODA, exceRpt maps the reads to more genes (Supplementary Figure 3A–C, G). This is not the case for miRNAs, where a comparable number of mapped features are detected by both pipelines (Supplementary Figure 3D–F, H). We then analysed the raw read count distribution per RNA biotype in the 1 × 100 dataset (Supplementary Figure S4A) and observed that exceRpt assigns reads to low-expressed genes belonging to several biotypes, whose count is instead 0 in CODA. The read count distribution for protein coding genes appears to be bimodal in both pipelines, with two peaks identifiable for 'low' and 'high' expressed genes (Supplementary Figure S4B).

Looking at the PCA plots, samples cluster along PC1 according to the pipelined use both for genes (Figure 3A and B) and miRNAs (Figure 3C and D), for both the 1 × 100 and 1 × 35 datasets, showing how the processing method is the biggest source of variation (Supplementary Figure S5). Additionally, clustering along PC2 reflects the BAP-treated versus control condition for the Nthy-ori 3-1 dataset.

Gene counts correlation is stronger among samples analysed with the same pipeline, which remains relatively high across the two methods for the 1 × 100 dataset (Figure 4A). In the 1 × 35 dataset, the correlation among biological replicates analysed with the same pipeline is also high but shows a lower value across methods (Figure 4B). The relatively higher correlation between

Table 1. Median counts and proportions of reads mapped to the reference transcriptome (or to exceRpt Gencode annotations) and miRNA and quantified with CODA or with exceRpt in the 1 × 100 and 1 × 35 datasets

	1 × 100 dataset				1 × 35 dataset			
	Read count		Percentage of trimmed reads		Read count		Percentage of trimmed reads	
	CODA	exceRpt	CODA	exceRpt	CODA	exceRpt	CODA	exceRpt
Total trimmed reads	46.32 M	45.50 M	100%	100%	73.37 M	61.69 M	100%	100%
Genes mapping reads	40.07 M	39.36 M	86.50%	86.42%	60.69 M	52.01 M	84.14%	85.60%
miRNA mapping reads	0.13 M	0.12 M	0.27%	0.25%	1.40 M	1.39 M	1.80%	5.95%

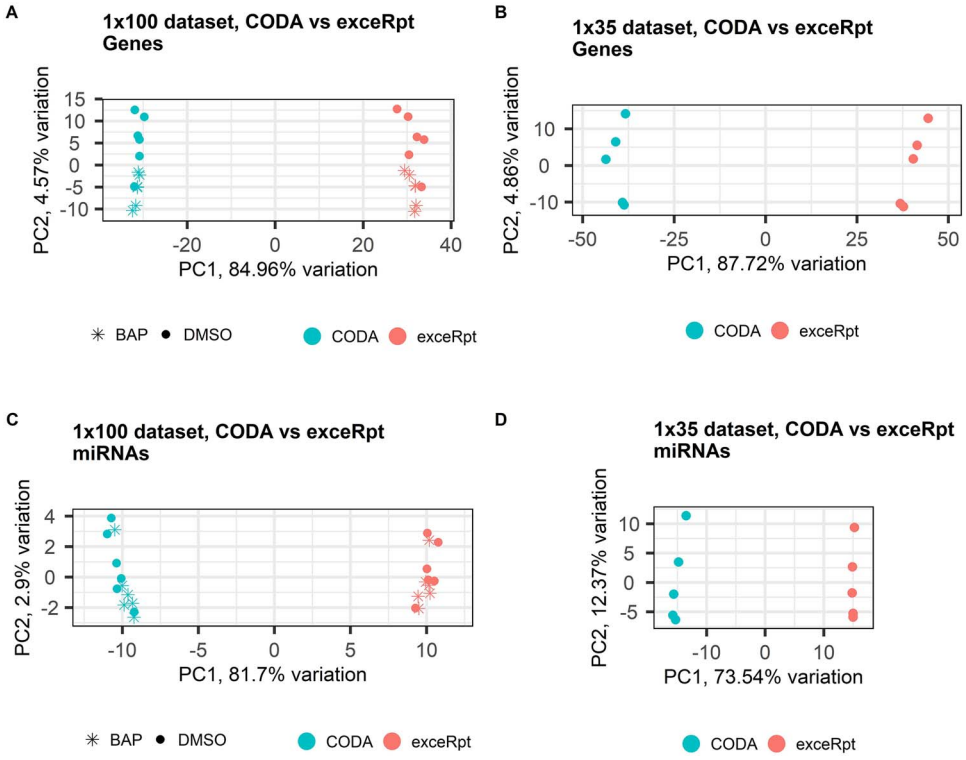

Figure 3. PCA plots showing PC1 and PC2 of PCA analysis carried out on variance-stabilized normalized gene counts for (**A**) 1 × 100 and (**B**) 1 × 35 samples processed with either pipeline. Plots showing PC1 and PC2 of PCA analysis carried out on variance-stabilized transformed miRNA counts for (**C**) 1 × 100 and (**D**) 1 × 35 samples processed with either pipeline (cyan = CODA, red = exceRpt).

CODA and exceRpt in the 1 × 100 dataset compared to the 1 × 35 one could be explained by the fact that the median total gene read count is very similar (from 39.36 M with exceRpt to 40.07 M reads with CODA, with an increase of around 0.71 M reads). For the 1 × 35 dataset, instead, there is a gain of ∼8.68 M reads per sample (around +17%, from a median 52.01 M with exceRpt to 60.69 M with CODA).

Correlation of normalized miRNA counts for both datasets is higher among samples analysed with the same pipeline and between the same sample analysed with CODA or exceRpt (Figure 4C and D).

To investigate the correlation difference among samples in the 1 × 100 dataset compared to the one in the 1 × 35 dataset, we analysed the RNA biotype composition of the Nthy-ori 3-1 and follicles DMSO control samples. To reduce the background noise and highlight the most consistent differences, we focused on biotypes representing at least 1% of total mapping reads on average (Figure 5A and B, Supplementary Table S4). Reads mapping to lncRNA and snoRNA are mainly between 60 and 100 nt long and are thus identified in the 1 × 100 dataset samples by both pipelines. Mitochondrial rRNA (Mt rRNA) reads show two peaks, at 89 and 91 nt, but the 91 nt peak is not identified by exceRpt (Figure 5C). snoRNA are almost completely missed by exceRpt. Possibly, these reads and other biotypes on average longer than 44 nt do not have a complete 3′ adapter, and are thus discarded by exceRpt. By recovering longer reads with an incomplete adapter, CODA also recovers protein coding reads, which constitute between 40% and 80% of the reads with an incomplete or partial adapter (Figure 5D).

DE analysis was performed to evaluate the number of DE genes and miRNA after BAP exposure of Nthy-ori 3-1 cells (1 × 100 dataset). MA-plots of the dataset processed with either workflow showed comparable distributions for the genes (Supplementary Figure S6A and B), while highly expressed miRNAs are characterized by a higher log2 fold change when analysed by CODA (Supplementary Figure S6C and D). In addition, the two pipelines identify a comparable number of DE genes and miRNA (CODA: 1201 DE genes, 1 DE miRNA; exceRpt: 1251 DE genes, 0 DE miRNA) (FDR = 0.01) (Supplementary Figure S6E and F) but only a partial overlap between genes (Supplementary Figure S6G).

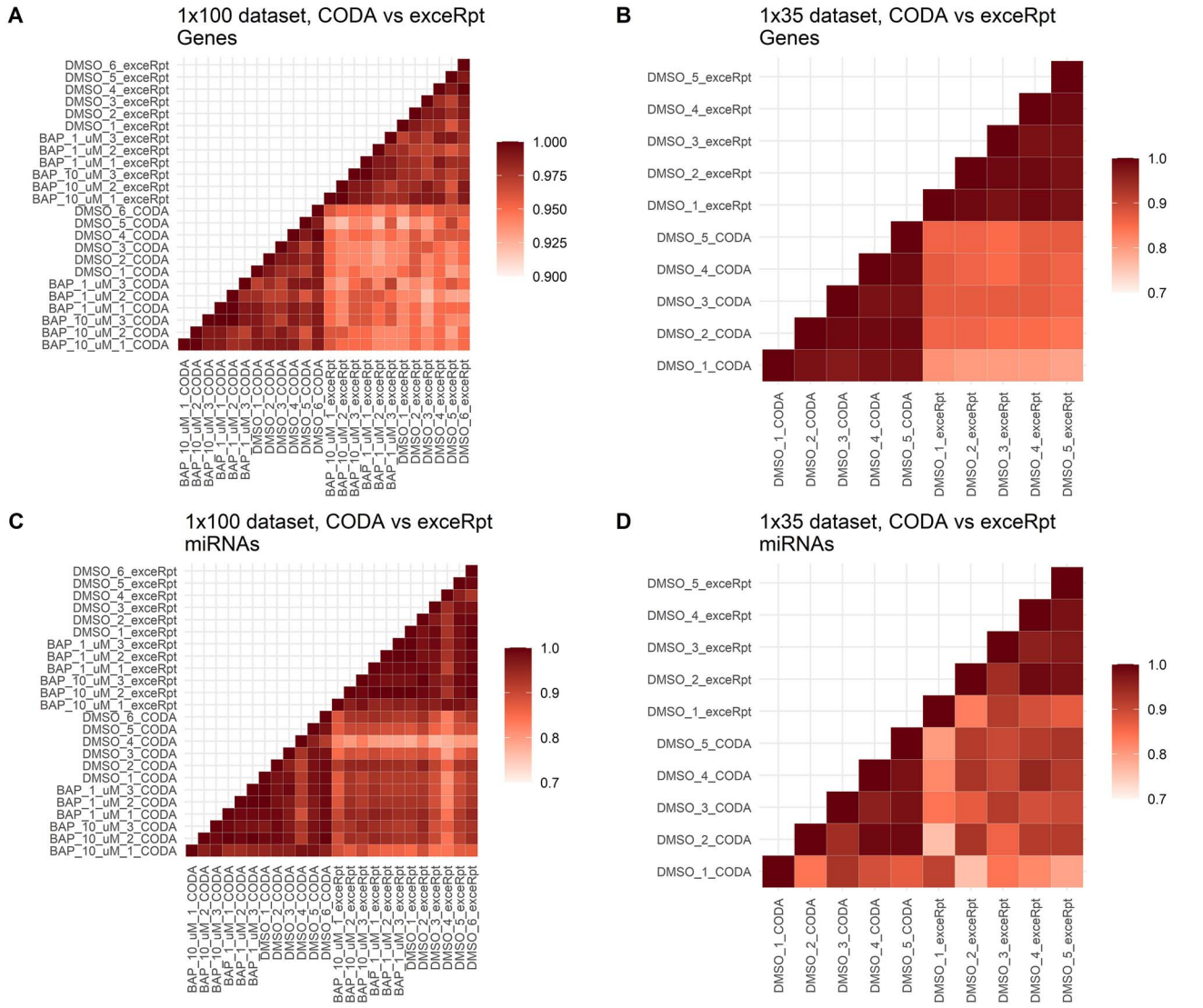

Figure 4. Pearson correlation of normalized gene counts for (**A**) 1 × 100 and (**B**) 1 × 35 samples. Pearson correlation of normalized miRNA counts for (**C**) 1 × 100 and (**D**) 1 × 35 samples.

GO and Reactome analysis show how the *P*-adjusted values for the same pathways are mostly lower in CODA (Supplementary Figure S7A and C) and showing enrichment for more terms (Supplementary Figure S7B and D). As BAP is known to be a strong inducer of the cytochrome P450 enzymes CYP1A1 and CYP1B1 via the activation of the aryl hydrocarbon receptor (AHR) [50], we analysed the expression of these two genes: both result DE at a comparable level (CYP1A1: p-adj = 5.2e-11 in CODA, p-adj = 4.2e-10 in exceRpt; CYP1B1: p-adj = 1.5e-05 in CODA, p-adj = 2.3e-10 in exceRpt) (Supplementary Figure S7E and F). Taken together, these results show how DE results are comparable between pipelines, with CODA showing higher sensitivity in gene functional enrichment analysis.

Evaluation of Combo-Seq compared to poly(A) and small RNA libraries

To evaluate the genes and miRNA identified by Combo-Seq libraries, we compared them to conventional poly(A) libraries (for genes) and small RNA libraries (for miRNAs) prepared with the same Nthy-ori 3-1 input RNA. We evaluated the number of expressed genes and miRNA using both libraries and performed DE analysis to identify genes and miRNA dysregulated upon BAP treatment.

The median total gene read count is 42.4 M and 35.0 M reads for Combo-Seq and poly(A) libraries, respectively (Supplementary Figure S8A), while for miRNAs it is 0.13 M (Combo-Seq) and 2.78 M (small RNA libraries) reads (Supplementary Figure S8B). Samples cluster along PC1 based on library preparation method and along PC2 based on treatment both for genes (Figure 6A) and miRNAs (Figure 6B), showing how the type of library is the greatest source of variation (Supplementary Figure S9). In addition, the greater spread along PC2 for Combo-Seq samples could be attributed to the libraries preparation over different batches, as opposed to the poly(A) and small RNA libraries, which were prepared in single batches, or to the low number of miRNA-mapping reads in Combo-Seq samples, as background noise tends to increase with small sample sizes. Correlation of genes normalized counts is not very high between the two different libraries compared to the correlation within the same methods (Figure 6C). Due to the very different sequencing depths of the two datasets, normalization for library size would tend to overestimate miRNA count in the samples prepared with Combo-Seq (Supplementary

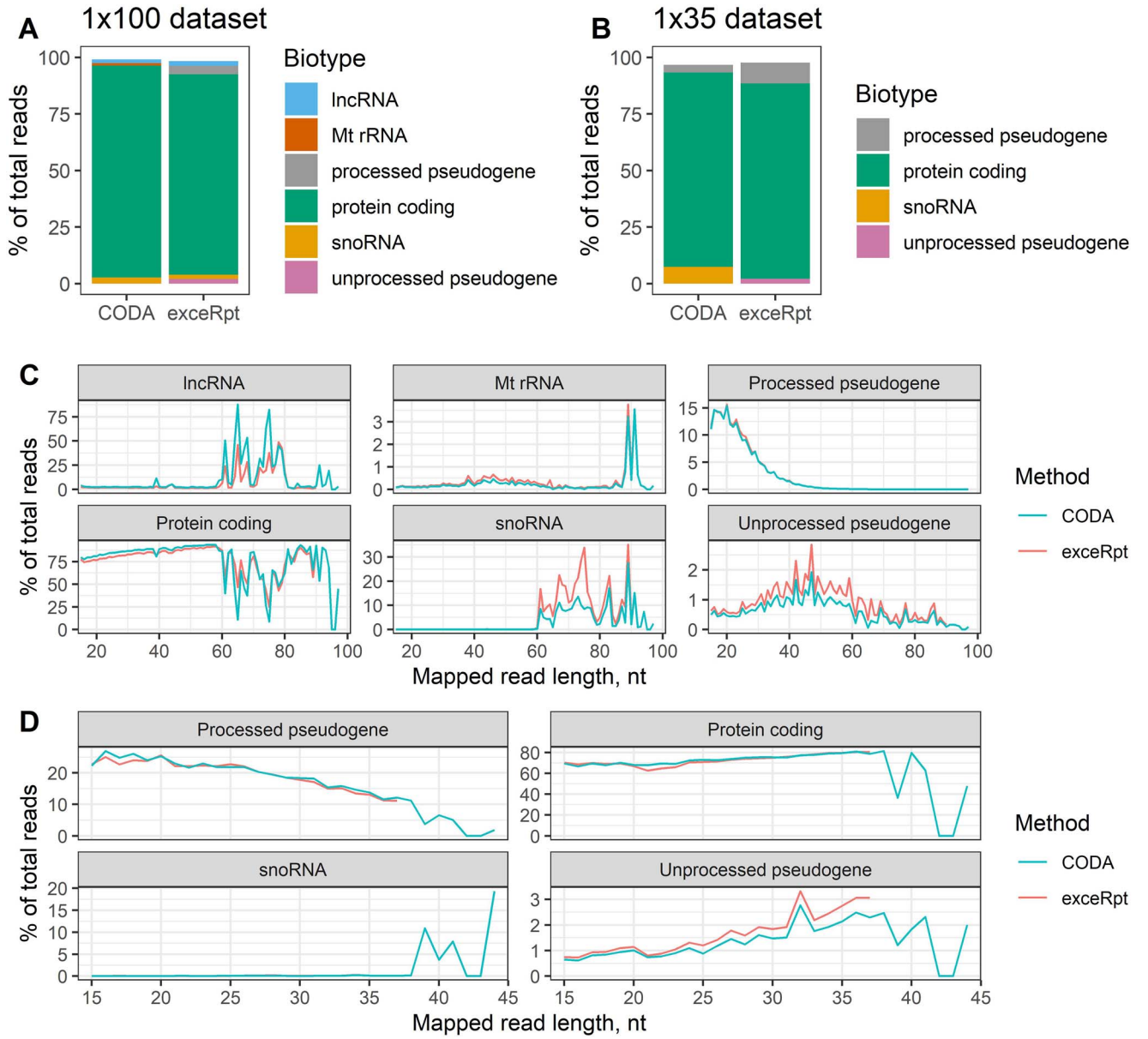

Figure 5. Average biotype composition of (**A**) 1 × 100 (six replicates) and (**B**) 1 × 35 (five replicates) DMSO control samples processed with either CODA or exceRpt. The values are expressed as percentage of total gene read counts. Read length distribution for (**C**) 1 × 100 and (**D**) 1 × 35 datasets expressed as percentage of total mapped reads grouped per biotype. Only the biotypes representing at least 1% of total reads on average are reported.

Figure S10). We then ranked the miRNAs in each sample based on their level of expression (highest read count = highest rank) and calculated Spearman correlation of the ranks across samples. The correlation plot confirms the PCA results, showing a high correlation among samples prepared with the same type of library (Figure 6D).

To evaluate how the type of library affects genes and miRNA detection, we compared the transcripts that are thus considered expressed. On average, as many genes are identified in samples prepared with poly(A) as in Combo-Seq libraries (3% more on average) (Supplementary Figure S11A–C). On the other hand, small RNA libraries identify 1.8 times more miRNAs than Combo-Seq on average, and similarly to genes, most of the ones detected by Combo-Seq overlap with the other library (Supplementary Figure S11D and F). Regression analysis of the average normalized read count of expressed genes and miRNAs for each group is not very strong (average $R^2 = 0.60$ for genes and $R^2 = 0.69$ for miRNAs) (Supplementary Figure S12).

Next, to identify DE genes and miRNA in the BAP-treated samples compared to the DMSO control, we performed DE analysis. When the samples are prepared with poly(A) libraries or Combo-Seq libraries, 4462 or 1186 genes result DE, respectively, 967 of which overlap between the two methods (Figure 7A). GO analysis shows that most top 10 terms are shared by Combo-Seq and poly(A) and are related to the processes of protein localization to telomeres (GO:0070203, GO:1904851, GO:1904816), regulation of apoptosis (GO:0042981, GO:0043065, GO:0043069), extracellular matrix organization (GO:0030198, GO:0097435, GO:0030334, GO:0030335) and regulation of protein localization to Cajal body (GO:1904871, GO:1904869) (Figure 7B). Reactome enrichment analysis shows similar results, where on average the top hits tend to have a lower *P*-adjusted value in the poly(A)

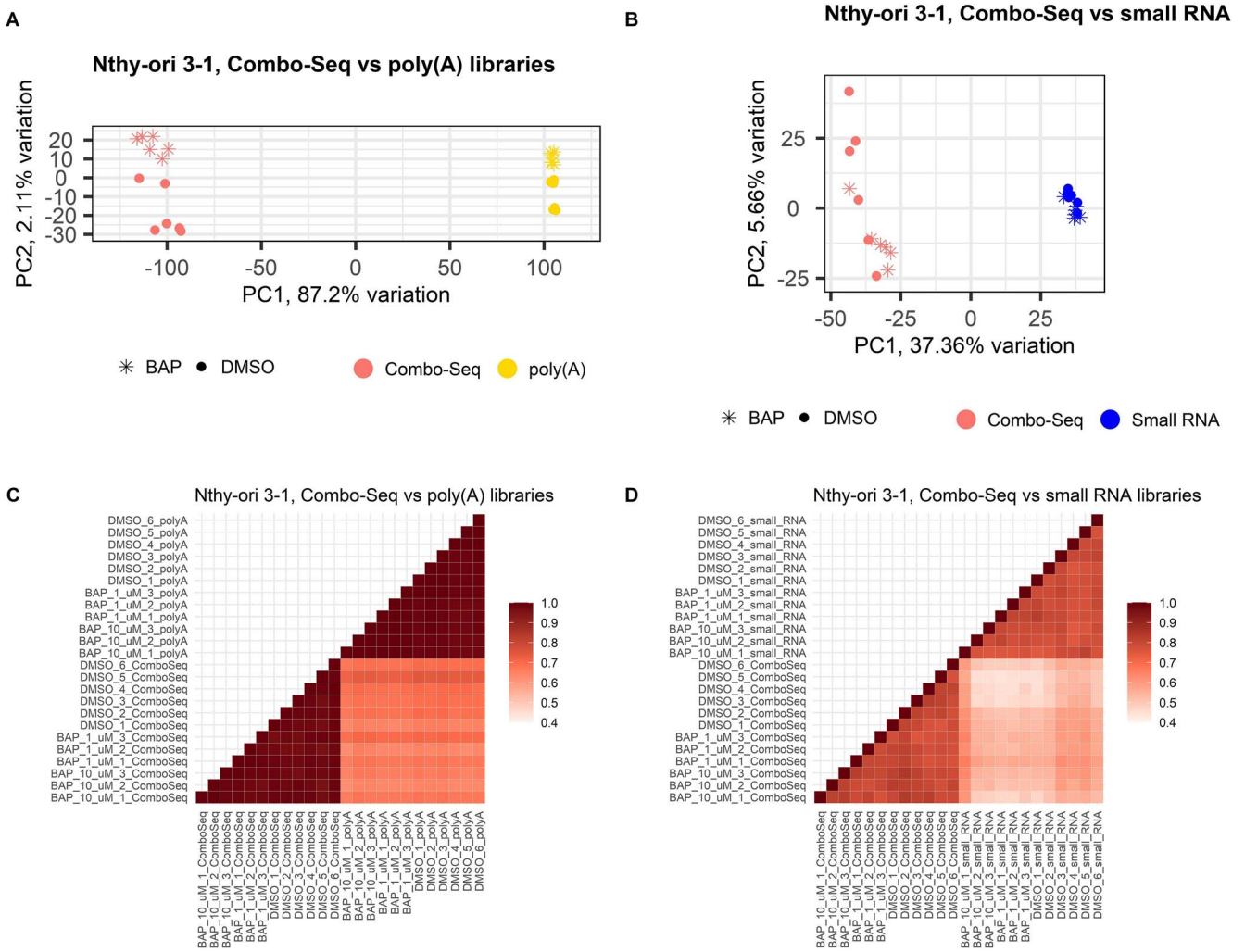

Figure 6. (**A**) PCA plot of variance-stabilized transformed gene counts of Nthy-ori 3-1 samples prepared using Combo-Seq or poly(A) libraries. (**B**) PCA plot of variance-stabilized transformed miRNA counts of Nthyori 3-1 samples prepared using Combo-Seq or small RNA libraries. (**C**) Person correlation of normalized gene counts for Nthy-ori 3-1 samples prepared with Combo-Seq or poly(A) libraries. (**D**) Spearman correlation of ranked miRNA counts of Nthy-ori 3-1 samples prepared using the Combo-Seq kit a small RNA kit. Combo-Seq libraries were sequenced on a 1 × 100 single-end flowcell, poly(A) libraries on a 2 × 200 paired-end flowcell and small RNA libraries on a 1 × 35 single-end flowcell.

dataset compared to Combo-Seq (Figure 7C). Globally, a greater enrichment in both GO (biological pathway) and Reactome terms results from poly(A) samples (Supplementary Figure 13). BAP is a genotoxic compound able to induce apoptosis. For this reason, we analysed the expression of genes known to be induced by TP53, by BAP treatment, or labelled as pro-apoptotic (Table 2). It must be pointed out that neither of the BAP concentrations tested resulted cytotoxic on Nthy-ori 3-1 cells after treatment for up to 72 h (data not shown). A partial overlap of the dysregulated genes can be observed (up: *BMF*, *CDKN1A*, *CYP1A1*, *CYP1B1*, *FAS*, *MDM2*; down: *BNIP3*, *BOK*, *GADD45A*), while some genes result dysregulated in either dataset (*BAK1*, *BCL2L11*, *BID*, *DDB2*, *RRMB2*).

Using Combo-Seq samples processed with CODA, we identify only one DE miRNA, while we identify nine DE miRNA from small RNA libraries samples. Interestingly, there is no overlap between the DE miRNA in the Combo-Seq and small RNA groups (Figure 7D). In addition, the only DE miRNA in the Combo-Seq group (hsa-miR-3654) is not expressed in the small RNA one. Vice versa, only six out of nine DE miRNA (hsa-miR-1268a/1268b, hsa-miR-186-5p, hsa-miR-222-3p, hsa-miR-30a-3p, hsa-miR-30c-2-3p, hsa-miR-92a-1-5p) are expressed in the Combo-Seq group. To discern which of the two libraries most truthfully detects the miRNAs in our samples, we validated the expression of three miRNAs using RT-qPCR in the DMSO control samples. We selected miRNAs for which the two datasets disagree in either direction (i.e. the rank is high in one dataset a low in the other) or are in concordance (Figure 7E): we selected hsa-miR-622, hsa-miR-122-5p and hsa-miR-361-3p (Supplementary Figure S14). RT-qPCR shows how hsa-miR-361-3p is the only miRNA detected in the DMSO control samples, at consistent levels among biological replicates. On the other hand, hsa-miR-122-5p and hsa-miR-622 are not detected (Figure 7F).

Intriguingly, hsa-miR-622 is coded within the keratin 18 pseudogene 27 (KRT18P27) (Supplementary Figure S15A). It is possible that a fragment of KRT18P27 mRNA generated early in the protocol of Combo-Seq library preparation, when poly(A) species are retrotranscribed and the RNA-cDNA hybrid is fragmented by RNaseH, is then erroneously recognized as an miRNA. Analysis of the isomiRs of hsa-miR622 reveals how there is a wide distribution of isomiRs across the six DMSO replicates detected at a low level (Supplementary Figure S15B). Interestingly, hsa-miR-622 read count calculated by exceRpt is comparable to the

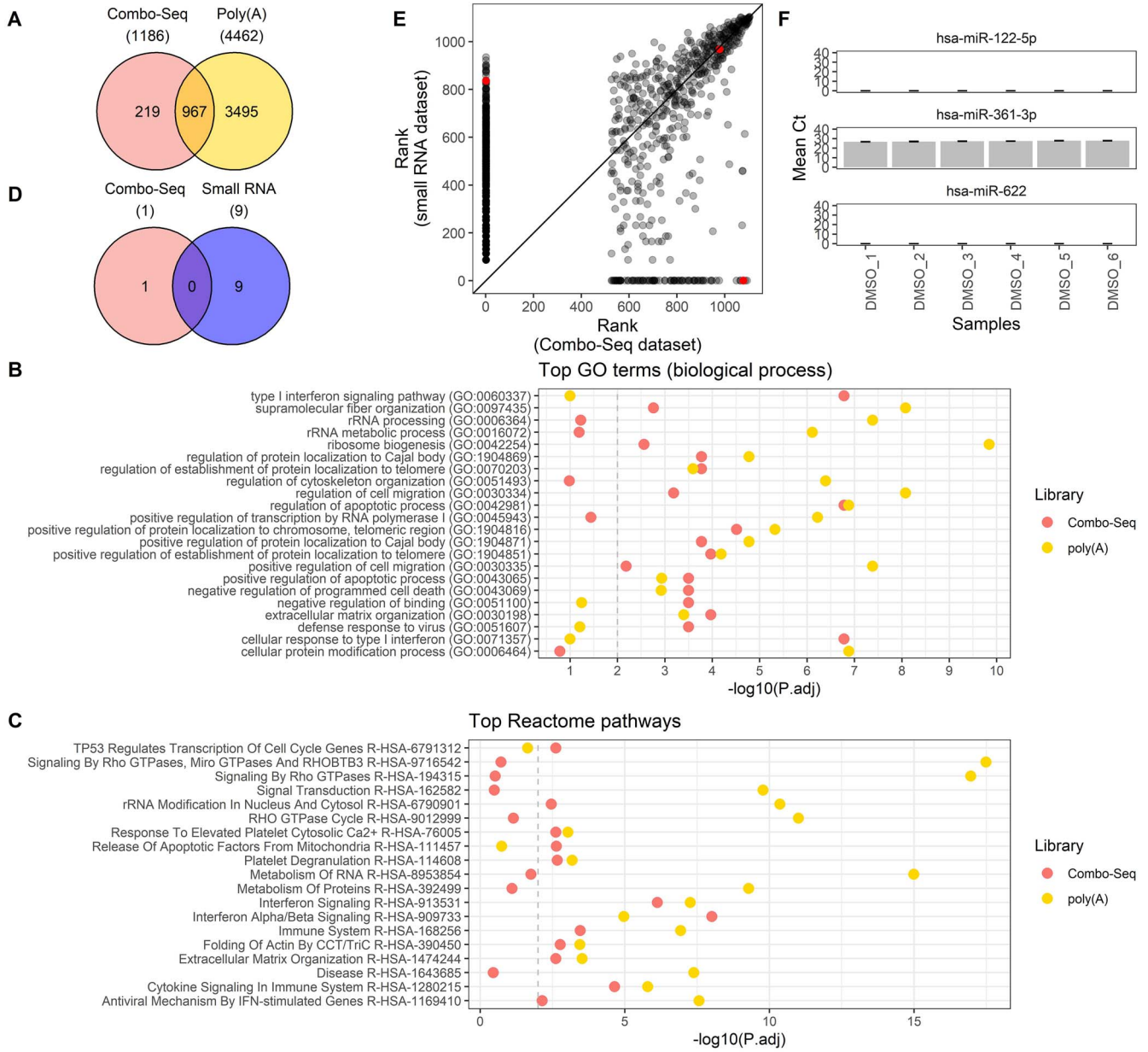

Figure 7. (**A**) Overlap of DE genes after BAP treatment compared to the DMSO control in datasets prepared with Combo-Seq libraries (red) or poly(A) (yellow) libraries. (**B**) Results of GO (biological process) and (**C**) Reactome pathway analyses performed on the DE genes in BAP versus DMSO samples prepared with either Combo-Seq (red) or poly(A) (yellow) libraries. The top 10 GO terms with lowest *P*-adjusted value in each group were selected and then plotted together. If two or more terms had the same *P*-adjusted value, all terms were reported. The dotted grey line corresponds to the set FDR value of 0.01. (**D**) Overlap of DE miRNA after BAP treatment compared to the DMSO control in datasets prepared with Combo-Seq libraries (red) or small RNA (blue) libraries. (**E**) Scatterplot representing the rankings of miRNA in mean read count of Nthy-ori 3-1 DMSO control samples. The mean read count was calculated as the average of the replicate samples prepared with either a Combo-Seq or small RNA library prep kit. miRNAs were then ranked based on their level of expression in each condition (most highly expressed miRNA = highest rank). Each dot in the plot represents a miRNA, and miRNAs for which the mean count was 0 in both conditions were removed. A total of 1104 miRNAs were ranked and miRNAs with the same level of expression were assigned the same rank. The miRNA selected for qPCR validation are highlighted in red. (**F**) RT-qPCR analysis of the selected miRNAs. The bar represents the average Ct value for each sample, and the error bars represent the mean ± sd. Each sample was measured in four technical replicates.

one calculated by CODA, supporting the hypothesis that this is not a consequence of a mature miRNA processing, but that at least part of the fragments derives from KRT18P27 fragmentation (Supplementary Figure S15C).

Discussion

Combo-Seq is a library prep kit for RNA-Seq that allows us to prepare combined mRNA-miRNA libraries starting from the same sample with very little minimum input (down to 5 ng of total RNA). For these reasons, it represents a convenient solution for simultaneously analysing both RNA species from a single sample, even for samples that contain little RNA, such as biopsies, extracellular fluids or organoids. In addition, it provides useful information about the relative mRNA and miRNA content of a cell, which to our knowledge cannot be provided by any library preparation kit currently available in the market. However, no specific bioinformatic pipeline was developed for the processing

Table 2. List of genes that are induced by BAP, involved in apoptosis, or induced by TP53. The data refer to the differential expression analysis carried out in the Nthy-ori 3-1 samples prepared either with Combo-Seq or poly(A) library kit. If the gene results upregulated, it is reported in red, if downregulated, in blue. The name of the library is reported, as well as the Ensembl gene ID and a brief description of the protein coded by the gene

Combo-Seq	Poly(A)	Ensembl gene ID	Protein function
BAD	BAD	ENSG00000002330	Proapoptotic member of the BCL-2 family
BAK1	BAK1	ENSG00000030110	Proapoptotic member of the BCL-2 family
BAX	BAX	ENSG00000087088	Proapoptotic member of the BCL-2 family
BBC3	BBC3	ENSG00000105327	Proapoptotic member of the BCL-2 family
BCL2L11	BCL2L11	ENSG00000153094	Proapoptotic member of the BCL-2 family
BID	BID	ENSG00000015475	Proapoptotic member of the BCL-2 family
BIK	BIK	ENSG00000100290	Proapoptotic member of the BCL-2 family
BMF	BMF	ENSG00000104081	Proapoptotic member of the BCL-2 family
BNIP3	BNIP3	ENSG00000176171	Proapoptotic member of the BCL-2 family
BOK	BOK	ENSG00000176720	Proapoptotic member of the BCL-2 family
CDKN1A	CDKN1A	ENSG00000124762	Inhibitor of cyclin-dependent kinase 2 and 4 complexes. Regulated by p53.
CYP1A1	CYP1A1	ENSG00000140465	Member of cytochrome P450 family. Induced by the AHR after binding by BAP.
CYP1B1	CYP1B1	ENSG00000138061	Member of cytochrome P450 family. Induced by the AHR after binding by BAP.
DDB2	DDB2	ENSG00000134574	Involved in DNA repair. Regulated by p53.
FAS	FAS	ENSG00000026103	Member of TNF-receptor superfamily. Necessary for the formation of the death-inducing signalling complex (DISC), involved in apoptosis. Regulated by p53.
GADD45A	GADD45A	ENSG00000116717	Involved in DNA repair mechanism. Regulated by p53.
MDM2	MDM2	ENSG00000135679	Oncogene. Codes for a nuclear-localized E3 ubiquitin ligase. It targets tumour suppressor proteins (like p53) for proteasomal degradation. Regulated by p53.
RRMB2	RRMB2	ENSG00000048392	Necessary for DNA synthesis. Regulated by p53.

of this data. To this purpose, the manufacturer recommends using the exceRpt pipeline with some modifications [19]. Nonetheless, it presents some limitations when adopted for the processing of Combo-Seq data.

In this paper, we illustrated CODA, a pipeline we developed for the processing of Combo-Seq data. It is modular and implements free-to-use tools often employed in RNA-Seq processing analysis. The first step is adapter trimming with Cutadapt, which we chose because its manual clearly states that it can handle partial adapters, a key point especially critical for shorter sequencing reads. In addition, Cutadapt is regularly supported and updated, and offers a clear and extensive documentation. After trimming, our pipeline performs separate alignment and quantification of miRNAs and genes: miRNA detection is carried out with miRge3.0. Gene mapping and quantification is done with RSEM based on the ENCODE3's STAR-RSEM pipeline. It is important to note that, although we selected certain tools, the strength point of CODA is the control that the user has over each step that is carried out and it is possible to change each tool according to the user's preferences. CODA can then be considered a guideline on how to analyse sequencing data deriving from Combo-Seq libraries, and the end user is free to use it or set up their own.

To compare CODA to exceRpt, we generated Combo-Seq libraries from two different cell models and compared the processing of the two pipelines. We showed that, because of the chosen trimmer, the maximum read length of trimmed reads when using CODA is higher than the one with exceRpt, and it results in more reads successfully passing. This is more dramatic the shorter the sequenced reads are. This tends to affect gene-mapping reads, rather than miRNA mapping ones: in fact, when the same samples are processed with CODA, the absolute number of reads mapping to genes increases, especially for shorter sequencing reads, where the proportion of reads with an incomplete/missing adapter increases. On the other hand, the number of reads mapped to miRNAs is almost the same. The two pipelines are comparable at the mapping stages, showing similar percentages of reads passing trimming that are aligned to genes or miRNA, possibly because they both use STAR as mapper for genes, while miRNA mapping, being more stringent and less ambiguous, can be performed by different aligners with similar results. In addition, more genes are assigned reads when samples are processed with exceRpt rather than CODA. This difference may be due to the quantification step: while exceRpt adopts its own quantification algorithm [20], CODA uses RSEM, which employs an Expectation–Maximization (EM) algorithm in its statistical model, whereby assignment of multimapping reads is determined by estimating the level of expression of deriving from unambiguously mapping reads [29]. We hypothesize that while exceRpt attributes every read to its highest scoring location, RSEM instead allocates reads mapping to very low expressed genes to higher expressed paralogues.

We also observed that the read length distribution in Combo-Seq libraries is not homogeneous for all RNAs: some biotypes tend to generate fragments longer than the Combo-Seq average (which is 21–22 nt [51]). As such, most reads coming from these species will have an incomplete/partial adapter when sequenced on a low number of cycles (e.g. using a 1 × 35 flowcell like we did). If the pipeline used to process the data cannot retain these reads, they will be lost. This can lead to an incorrect estimation of the RNA biotype composition of a sample, and loss of potentially interesting data [52, 53]. Although two different RNA extraction kits were used during the Nthy-ori 3-1 and thyroid follicles processing, the generated RNA-Seq data is comparable [54, 55]. We believe then the results obtained by the analysis of the 1 × 100 and 1 × 35 datasets have been minimally influenced by this factor.

We also compared how Combo-Seq libraries perform in comparison to standard poly(A) and small RNA libraries for the analysis of mRNA or miRNA, respectively. We showed how the type of library is the main source of variation when comparing the two datasets. We noticed that conventional poly(A) libraries identify

around 4% more genes, while small RNA libraries identify almost twice as many miRNAs. The difference in genes identified is most likely due to the chemistry underlying the library preparation kit, as very different RNA inputs from the same sample prepared with the same kit show very similar percentages of exonic-, intronic- and intergenic-mapping reads [56]. In addition, for this work, the poly(A) libraries were prepared in a single batch in an automated system. The Combo-Seq libraries were prepared manually over different batches and the protocol includes several steps. It is then possible that the variation introduced during the Combo-Seq libraries preparation reflects in a variability in gene expression among biological replicates, which would affect the DE analysis. Considering the miRNAs, most of the variability in detection probably arises from the difference in total miRNA read counts. Indeed, the number of recovered miRNAs from Combo-seq is dependent on the miRNA content of the cells, which seems to be low in the selected Nthy-ori 3-1 cell line. The sequencing depth would therefore need to be much higher to reach a number of reads comparable to the small RNA libraries. Admittedly, the input RNA used for library preparation is a possible confounder, as we did not test how the number of miRNA mapping reads changes with different amounts of input, and using a lower amount, which increases the miRNA/mRNA ratio in Combo-Seq libraries [57], could result in a greater number of miRNA mapping reads and thus in more miRNAs identified. Small RNA libraries prepared with a high RNA input perform similarly when detecting highly expressed miRNAs compared to using a low input, and detect more low-expressed ones [58].

DE analysis shows how four times more DE genes are identified in poly(A) libraries, but GO top hits are almost the same. Nine versus one DE miRNAs result also from the analysis in small RNA and Combo-Seq libraries, respectively. While we observed an overlap in the DE genes between the two methods, we did not get a similar result for the DE miRNAs. In addition, miRNA validation by RT-qPCR is concordant with small RNA libraries: hsa-miR-622 was detected by Combo-Seq only at high levels, but its expression was absent in RT-qPCR. We hypothesize that at least part of the reads assigned to hsa-miR-622 in Combo-Seq samples may instead derive from the fragmentation of the KRT18P27 transcript. hsa-miR-122-5p was detected at low levels and only by small RNA libraries, which as already discussed have a greater read coverage.

In conclusion, Combo-Seq is a convenient solution to capture both poly(A)-tailed and small RNAs starting from very little material and from a single RNA aliquot. In addition, it requires less time and money per sample than the combination of conventional separated poly(A) and small RNA libraries. However, it presents some inconsistencies when compared to standard poly(A) and small RNA libraries, which researchers should be aware of and evaluate when choosing how to prepare their samples.

Key Points

- We developed CODA, a pipeline for the processing of RNA-Seq libraries prepared with the Combo-Seq kit. It makes use of established RNA-Seq data analysis software and allows for control and customization over the different processing steps. The pipeline can be used as such, using the necessary files available in the CODA GitHub repository, or modified according to the user's need.
- Compared to exceRpt, CODA increases the number of reads passing the initial trimming phase. The gain is more dramatic as the number of sequencing cycles decrease. Longer reads are enriched for some RNA biotypes like snoRNA, lncRNA and mtRNA.
- DE analysis following BAP treatment using Combo-Seq libraries identifies fewer DE genes and miRNAs than matched poly(A) and small RNA libraries, respectively. There is a partial overlap of DE genes, but no overlap of DE miRNAs.
- RT-qPCR validation of three miRNAs shows concordance with small RNA libraries over Combo-Seq.

Data availability

The data underlying this article is available in BioStudies with accession E-MTAB-12078.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825745 (SCREENED).

References

1. Boivin V, Faucher-Giguere L, Scott M, *et al.* The cellular landscape of mid-size noncoding RNA. *Wiley Interdiscip Rev RNA* 2019;**10**(4):e1530.
2. Godoy PM, Bhakta NR, Barczak AJ, *et al.* Large differences in small RNA composition between human biofluids. *Cell Rep* 2018;**25**(5):1346–58.
3. Potemkin N, Cawood SMF, Treece J, *et al.* A method for simultaneous detection of small and long RNA biotypes by ribodepleted RNA-Seq. *Sci Rep* 2022;**12**(1):621.
4. Nolte-'t Hoen EN, Buermans HP, Waasdorp M, *et al.* Deep sequencing of RNA from immune cell-derived vesicles uncovers the selective incorporation of small non-coding RNA biotypes with potential regulatory functions. *Nucleic Acids Res* 2012;**40**(18):9272–85.
5. Boivin V, Deschamps-Francoeur G, Couture S, *et al.* Simultaneous sequencing of coding and noncoding RNA reveals a human transcriptome dominated by a small number of highly expressed noncoding genes. *RNA* 2018;**24**(7):950–65.
6. Stoicea N, Du A, Lakis DC, *et al.* The MiRNA journey from theory to practice as a CNS biomarker. *Front Genet* 2016;**7**:11.
7. Wittmann J, Jack HM. Serum microRNAs as powerful cancer biomarkers. *Biochim Biophys Acta* 2010;**1806**(2):200–7.
8. Backes C, Meese E, Keller A. Specific miRNA disease biomarkers in blood, serum and plasma: challenges and prospects. *Mol Diagn Ther* 2016;**20**(6):509–18.
9. Scholer N, Langer C, Dohner H, *et al.* Serum microRNAs as a novel class of biomarkers: a comprehensive review of the literature. *Exp Hematol* 2010;**38**(12):1126–30.
10. da Silva JL, Cardoso Nunes NC, Izetti P, *et al.* Triple negative breast cancer: a thorough review of biomarkers. *Crit Rev Oncol Hematol* 2020;**145**:102855.
11. Arantes L, De Carvalho AC, Melendez ME, *et al.* Serum, plasma and saliva biomarkers for head and neck cancer. *Expert Rev Mol Diagn* 2018;**18**(1):85–112.

12. Anvar SY, Allard G, Tseng E, *et al.* Full-length mRNA sequencing uncovers a widespread coupling between transcription initiation and mRNA processing. *Genome Biol* 2018;**19**(1):46.
13. Ramberg S, Hoyheim B, Ostbye TK, *et al.* A de novo full-length mRNA transcriptome generated from hybrid-corrected PacBio long-reads improves the transcript annotation and identifies thousands of novel splice variants in Atlantic Salmon. *Front Genet* 2021;**12**:656334.
14. Liu D, Graber JH. Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation. *BMC Bioinformatics* 2006;**7**:77.
15. Eddy SR. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* 2001;**2**(12):919–29.
16. Illumina. *How Short Inserts Affect Sequencing Performance*. 2020. https://support.illumina.com/bulletins/2020/12/how-short-inserts-affect-sequencing-performance.html (13 May 2022, date last accessed).
17. Verheijen MCT, Krauskopf J, Caiment F, *et al.* iPSC-derived cortical neurons to study sporadic Alzheimer disease: a transcriptome comparison with post-mortem brain samples. *Toxicol Lett* 2022;**356**:89–99.
18. Illumina. *Introducing the NovaSeq™ 6000 v1.5 reagents*. 2020. https://support.illumina.com/bulletins/2020/11/introducing-the-novaseq--6000-v1-5-reagents.html (28 April 2022, date last accessed).
19. PerkinElmer Inc. *NEXTFLEX® Combo-Seq Analysis Guidelines*. 2020. https://perkinelmer-appliedgenomics.com/wp-content/uploads/2020/06/NOVA-5139-AG_v01_NEXTFLEX-Combo-seq-Analysis-Guideline.pdf (9 December 2021, date last accessed).
20. Rozowsky J, Kitchen RR, Park JJ, *et al.* exceRpt: a comprehensive analytic platform for extracellular RNA profiling. *Cell Syst* 2019;**8**(4):352–357.e3.
21. Abdelhamid RF, Ogawa K, Beck G, *et al.* piRNA/PIWI protein complex as a potential biomarker in sporadic amyotrophic lateral sclerosis. *Mol Neurobiol* 2022;**59**(3):1693–705.
22. Zheng T, Ellinghaus D, Juzenas S, *et al.* Genome-wide analysis of 944 133 individuals provides insights into the etiology of haemorrhoidal disease. *Gut* 2021;**70**:1538–49.
23. Antonica F, Kasprzyk DF, Opitz R, *et al.* Generation of functional thyroid from embryonic stem cells. *Nature* 2012;**491**(7422):66–71.
24. Romitti M, Eski SE, Fonseca BF, *et al.* Single-cell trajectory inference guided enhancement of thyroid maturation in vitro using TGF-beta inhibition. *Front Endocrinol* 2021;**12**:657195.
25. Mueller O, Lightfoot S, Schroeder A. *RNA Integrity Number (RIN) – Standardization of RNA Quality Control*. January 21, 2016; https://www.agilent.com/cs/library/applications/5989-1165EN.pdf.
26. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 2011;**17**(1):10–12.
27. Patil AH, Halushka MK. miRge3.0: a comprehensive microRNA and tRF sequencing analysis pipeline. *NAR Genom Bioinform* 2021;**3**(3):lqab068.
28. Li B. *rsem-Prepare-Reference Documentation Page*. May 10, 2022. https://deweylab.github.io/RSEM/rsem-prepare-reference.html.
29. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011;**12**:323.
30. Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21.
31. Frankish A, Diekhans M, Ferreira AM, *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res* 2019;**47**(D1):D766–73.
32. Griffiths-Jones S, Saini HK, van Dongen S, *et al.* miRBase: tools for microRNA genomics. *Nucleic Acids Res* 2008;**36**(Database issue):D154–8.
33. Bushnell B. *BBMap*, December 12, 2021. sourceforge.net/projects/bbmap/.
34. Andrews S. *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]*, 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
35. Ewels P, Magnusson M, Lundin S, *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**(19):3047–8.
36. PerkinElmer Inc. *NEXTflex™ Small RNA Trimming Instructions*, March 30, 2022. https://perkinelmer-appliedgenomics.com/wp-content/uploads/marketing/NEXTFLEX/miRNA/NEXTflex_Small_RNA_v3_Trimming_Instructions.pdf.
37. Verheijen MC, Meier MJ, Asensio JO, *et al.* R-ODAF: omics data analysis framework for regulatory application. *Regul Toxicol Pharmacol* 2022;**131**:105143.
38. CEFIC C4 team. *Omics Data Analysis Framework for Regulatory Application (R-ODAF)*. 2021. https://github.com/R-ODAF/Main (7 December 2021, date last accessed).
39. Chen S, Zhou Y, Chen Y, *et al.* Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**(17):i884–90.
40. Durinck S, Spellman PT, Birney E, *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 2009;**4**(8):1184–91.
41. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.
42. Blighe K, Lun A. *PCAtools: PCAtools: Everything Principal Components Analysis*, 2021. https://github.com/kevinblighe/PCAtools.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**(12):550.
44. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**(1):139–40.
45. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;**25**(1):25–9.
46. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;**49**(D1):D325–d334.
47. Gillespie M, Jassal B, Stephan R, *et al.* The reactome pathway knowledgebase 2022. *Nucleic Acids Res* 2022;**50**(D1):D687–92.
48. Chen EY, Tan CM, Kou Y, *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128. (17 December 2021, date last accessed).
49. Hannon GJ. *FASTX-Toolkit*. 2010. http://hannonlab.cshl.edu/fastx_toolkit.
50. Atsdr. *Toxicological profile for polycyclic aromatic hydrocarbons*. Atlanta, GA: U.S. Department of Health and Human Services, Public Health Service, 1995.
51. Allen K, Morris A, Piehl S, *et al. Combined mRNA & microRNA NGS Library Prep Enables a more Complete Characterization of Cell-free RNA*, 2018.
52. Liang J, Wen J, Huang Z, *et al.* Small nucleolar RNAs: insight into their function in cancer. *Front Oncol* 2019;**9**:587.
53. Calvo Sánchez J, Köhn M. Small but mighty—the emerging role of snoRNAs in Hematological malignancies. *Noncoding RNA* 2021;**7**(4):68.
54. Marczyk M, Fu C, Lau R, *et al.* The impact of RNA extraction method on accurate RNA sequencing from

formalin-fixed paraffin-embedded tissues. *BMC Cancer* 2019;**19**(1):1189.
55. Scholes AN, Lewis JA. Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses. *BMC Genomics* 2020;**21**(1):249.
56. Wang L, Felts SJ, Van Keulen VP, *et al.* Exploring the effect of library preparation on RNA sequencing experiments. *Genomics* 2019;**111**(6):1752–9.
57. PerkinElmer. A.G.-. *NEXTFLEX® Combo-Seq™ mRNA/miRNA Kit (v20.04)*. https://perkinelmer-appliedgenomics.com/wp-content/uploads/2020/07/NOVA-5139-01-02-NEXTFLEX-Combo-Seq-Kit-AG052004_28.pdf.
58. Yeri A, Courtright A, Danielson K, *et al.* Evaluation of commercially available small RNASeq library preparation kits using low input RNA. *BMC Genomics* 2018;**19**(1):331.
59. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* 2016;**9**:88.
60. Martin M. *Algorithms and Tools for the Analysis of High Throughput DNA Sequencing Data*. Technischen Universität Dortmund, Dortmund, Germany, 2014.
61. Stephens M. False discovery rates: a new deal. *Biostatistics* 2017;**18**(2):275–94.