exRNAQC004: RNA purification kit performance (mRNA level), part of phase 1

Annelien Morlion - on behalf of exRNAQC Consortium

15/12/2020

Contents

1 Introduction **2**

1.1 Experimental setup . . . 2

1.2 Metric selection . . . 3

1.3 RMarkdown set-up . . . 3

2 Annotation **3**

3 Sequencing and preprocessing **3**

3.1 Filtering . . . 3

3.2 Downsampling . . . 3

3.3 Strandedness . . . 5

3.4 Duplicate rate . . . 5

3.4.1 Comparison of technical replicates before and after duplicate removal . . . 5

3.4.2 Duplicate removal . . . 8

3.4.3 Link between duplication and plasma input volume? . . . 9

3.5 Total number of reads . . . 9

3.6 Gene count conversion . . . 9

4 Performance metrics **13**

4.1 Duplication level . . . 13

4.2 RNA concentration . . . 13

4.3 RNA yield . . . 13

4.4 Efficiency of kits . . . 13

4.5 Filter threshold . . . 16

4.5.1 Cutoff examples . . . 17

4.5.2 95% SP elimination cutoffs . . . 19

4.5.3 Impact of filtering . . . 19
4.5.4 Robustness of cutoff . . . 20
4.6 Number of genes . . . 20
4.7 Coverage . . . 20
4.8 ALC . . . 25
4.8.1 Individual ALC plots . . . 25
4.8.2 Overview ALC . . . 28
4.9 Overview . . . 28
4.9.1 Correlation between metrics . . . 28
4.9.2 Comparison of kits . . . 28

5 Selection for phase 2 **33**

6 Spikes **34**
6.1 ERCC . . . 34
6.1.1 Linear models . . . 34
6.1.2 Recovery of spikes . . . 35

1 Introduction

1.1 Experimental setup

For the evaluation of the different RNA isolation kits in the first phase of exRNAQC, blood was drawn from 1 healthy volunteer. We tested 8 different kits:

- miRNeasy Serum/Plasma Kit (abbreviated to MIR; Qiagen, 217184)
- miRNeasy Serum/Plasma Advanced Kit (abbreviated to MIRA; Qiagen, 217204)
- mirVana PARIS Kit (abbreviated to MIRV (and MIRVE); Life Technologies, AM1556)
- NucleoSpin miRNA Plasma Kit (abbreviated to NUC; Macherey-Nagel, 740981.50)
- QIAamp ccfDNA/RNA Kit (abbreviated to CCF; Qiagen, 55184)
- Plasma/Serum Circulating and Exosomal RNA Purification Kit/Slurry Format (abbreviated to CIRC; Norgen Biotek Corp., 42800)
- Maxwell RSC miRNA Plasma and Exosome Kit (Promega, AX5740) in combination with the Maxwell RSC Instrument (abbreviated to MAX; Promega, AS4500)
- MagNA Pure 24 Total NA Isolation Kit (Roche, 07 658 036 001) in combination with the MagNA Pure instrument (abbreviated to MAP; Roche, 07 290 519 001)

Most kits allow a range of plasma input volumes. Therefore, we tested both the minimum and maximum input volume recommended by the supplier. The input volume in ml directly follows the abbreviated name in the plots in this report. This yields 15 unique combinations of kit and input volumes, with 3 technical replicates processed for every combination.

1.2 Metric selection

Nine performance metrics were evaluated. Kits for phase 2 were eventually selected based on transforming metrics for sensitivity and reproducibility to robust z-scores (see Selection for phase 2).

- Sensitivity: Absolute number of genes detected (after setting a count cutoff that removes 95% of single positives between technical replicates).
- Reproducibility: pairwise ALC (area-left-of-curve) calculation between technical replicates.

1.3 RMarkdown set-up

First, basic parameters are set up in this RMarkdown, such as loading dependencies, setting paths and setting up a uniform plot structure.

2 Annotation

Sample annotation with info about kit, used input volume, eluate volume etc.

Sequin spike-in controls are added to plasma prior to RNA isolation, and External RNA Control Consortium (ERCC) spike-in controls to the RNA eluate prior to library prep. Original spike concentrations in mix are taken from providers' annotation files (Garvan Institute of Medical Research for Sequins and ThermoFisher Scientific for ERCCs)

3 Sequencing and preprocessing

- Three runs:
  - NSQ_Run479-93008916
  - NSQ_Run481-93380289
  - NSQ_Run482-93738645
- Sequenced on 2018/08/24 - 2018/09/03 (NextSeq)
- Original amount of paired reads (min= 24,910,761, mean= 31,939,163, max= 54,316,378)

3.1 Filtering

Quality filtering of sequenced reads (keep PE reads were at least 80% of nt in both reads have phred score 20)

3.2 Downsampling

Randomly downsample everything to the lowest number of paired end reads (at FASTQ level) to make sure the comparison of metrics is fair. E.g. if one sample is sequenced deeper it is likely to yield more genes compared to a sample that was sequenced less deep.

As the lowest number of reads is 21,370,152 (in MIR0.2 sample), we downsampled all samples to **21M paired end reads**.

Table 1: Sample annotation. PlInputV: plasma input volume; Repl: technical replicate; RNAinput: RNA input volume used for library prep; Sequin: dilution of Sequin spike stock; ERCC: dilution of ERCC spike stock

UniqueID	Biotype	RNAisolation	PlInputV	EluateV	RNAinput	Repl	Tube	LibraryPrep	Sequin	ERCC	Abbrevation
RNA003584L1	artiPPP	miRNeasySPkit	200	14	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MIR0.2
RNA003587L1	artiPPP	miRNeasySPAkit	200	20	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MIRA0.2
RNA003590L1	artiPPP	miRNeasySPAkit	600	20	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MIRA0.6
RNA003593L1	artiPPP	QIAamp	1000	14	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	CCF1
RNA003596L1	artiPPP	QIAamp	4000	14	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	CCF4
RNA003599L1	artiPPP	mirVana	100	100	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MIRV0.1
RNA003602L1	artiPPP	mirVana	625	100	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MIRV0.625
RNA003605L1	artiPPP	NucleoSpin	300	30	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	NUC0.3
RNA003608L1	artiPPP	NucleoSpin	900	30	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	NUC0.9
RNA003611L1	artiPPP	Norgen	250	100	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	CIRC0.25
RNA003614L1	artiPPP	Norgen	5000	100	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	CIRC5
RNA003617L1	artiPPP	MagnaPure	2000	50	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MAP2
RNA003620L1	artiPPP	MagnaPure	4000	50	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MAP4
RNA003623L1	artiPPP	Maxwell	100	50	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MAX0.1
RNA003626L1	artiPPP	Maxwell	500	50	8.5	RNA1	EDTA	RNAAccess	1/1000000	1/500000	MAX0.5
RNA003585L1	artiPPP	miRNeasySPkit	200	14	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MIR0.2
RNA003588L1	artiPPP	miRNeasySPAkit	200	20	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MIRA0.2
RNA003591L1	artiPPP	miRNeasySPAkit	600	20	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MIRA0.6
RNA003594L1	artiPPP	QIAamp	1000	14	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	CCF1
RNA003597L1	artiPPP	QIAamp	4000	14	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	CCF4
RNA003600L1	artiPPP	mirVana	100	100	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MIRV0.1
RNA003603L1	artiPPP	mirVana	625	100	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MIRV0.625
RNA003606L1	artiPPP	NucleoSpin	300	30	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	NUC0.3
RNA003609L1	artiPPP	NucleoSpin	900	30	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	NUC0.9
RNA003612L1	artiPPP	Norgen	250	100	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	CIRC0.25
RNA003615L1	artiPPP	Norgen	5000	100	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	CIRC5
RNA003618L1	artiPPP	MagnaPure	2000	50	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MAP2
RNA003621L1	artiPPP	MagnaPure	4000	50	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MAP4
RNA003624L1	artiPPP	Maxwell	100	50	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MAX0.1
RNA003627L1	artiPPP	Maxwell	500	50	8.5	RNA2	EDTA	RNAAccess	1/1000000	1/500000	MAX0.5
RNA003586L1	artiPPP	miRNeasySPkit	200	14	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MIR0.2
RNA003589L1	artiPPP	miRNeasySPAkit	200	20	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MIRA0.2
RNA003592L1	artiPPP	miRNeasySPAkit	600	20	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MIRA0.6
RNA003595L1	artiPPP	QIAamp	1000	14	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	CCF1
RNA003598L1	artiPPP	QIAamp	4000	14	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	CCF4
RNA003601L1	artiPPP	mirVana	100	100	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MIRV0.1
RNA003604L1	artiPPP	mirVana	625	100	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MIRV0.625
RNA003607L1	artiPPP	NucleoSpin	300	30	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	NUC0.3
RNA003610L1	artiPPP	NucleoSpin	900	30	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	NUC0.9
RNA003613L1	artiPPP	Norgen	250	100	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	CIRC0.25
RNA003616L1	artiPPP	Norgen	5000	100	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	CIRC5
RNA003619L1	artiPPP	MagnaPure	2000	50	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MAP2
RNA003622L1	artiPPP	MagnaPure	4000	50	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MAP4
RNA003625L1	artiPPP	Maxwell	100	50	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MAX0.1
RNA003628L1	artiPPP	Maxwell	500	50	8.5	RNA3	EDTA	RNAAccess	1/1000000	1/500000	MAX0.5

3.3 Strandedness

A strand specific protocol was used. To test if this worked as expected, we used RSeQC on BAM output files after STAR alignment to infer strandedness:

- infer_experiment.py from RSeQC/2.6.4-intel-2018a-Python-2.7.14
- look at fraction of reads explained by fr-firststrand (i.e. reverse in htseq)
  - category 1+-,1-+,2++,2– (e.g. 1+- read 1 '+' mapped to + strand while gene is on '-' strand: is what we expect in our case)

MagnaPure RNA isolation kit has worst performance

- 70-90% on correct strand (while in other kits: 97-99%)
- The low % strandedness is an issue in all chromosomes (not only mitochondrial)
- **DNA-contamination?**
  - Protocol co-isolates DNA and RNA -> Possible that there is still DNA left after DNase treatment?
    - We definitely applied the DNase treatment to all kits.
    - DNase treatment on MAP samples was done together with Maxwell (which does have a good performance)
    - Not enough enzyme? Incompatible DNase treament?
  - We tried to remove DNA again using a different method, but there still seems to be DNA contamination
- **Not everything is DNA**, otherwise strandedness would be closer to 50%
  - e.g. if ratio RNA to DNA is 50/50, you expect strandedness to be close to 75%
- Strange that MAP4 has better strandedness than MAP2
- **However, DNA contamination is problematic as we cannot be sure that what we pick up is indeed coming from exRNA**
  - We could remove everything that maps to the antistrand, but it is still not really fair as DNA will also contribute reads from other strand
  - => **we will leave MagnaPure kits out of analyses**
- Remark: although not relevant for exRNA quantification, in some cases it is an advantage to have a kit that isolates both DNA and RNA

3.4 Duplicate rate

Low amount of input RNA, such as in plasma, results in many PCR duplicates. After duplicate removal (allowing up to 2 substitutions to account for sequencing errors), technical replicate counts are closer together and the cutoff for eliminating 95% of single positives (see Filter threshold) is considerably lower.

3.4.1 Comparison of technical replicates before and after duplicate removal

Gene counts of 2 technical replicates are plotted against each other R-squared determined based on linear regression of the log counts of these genes (higher R2 = better)

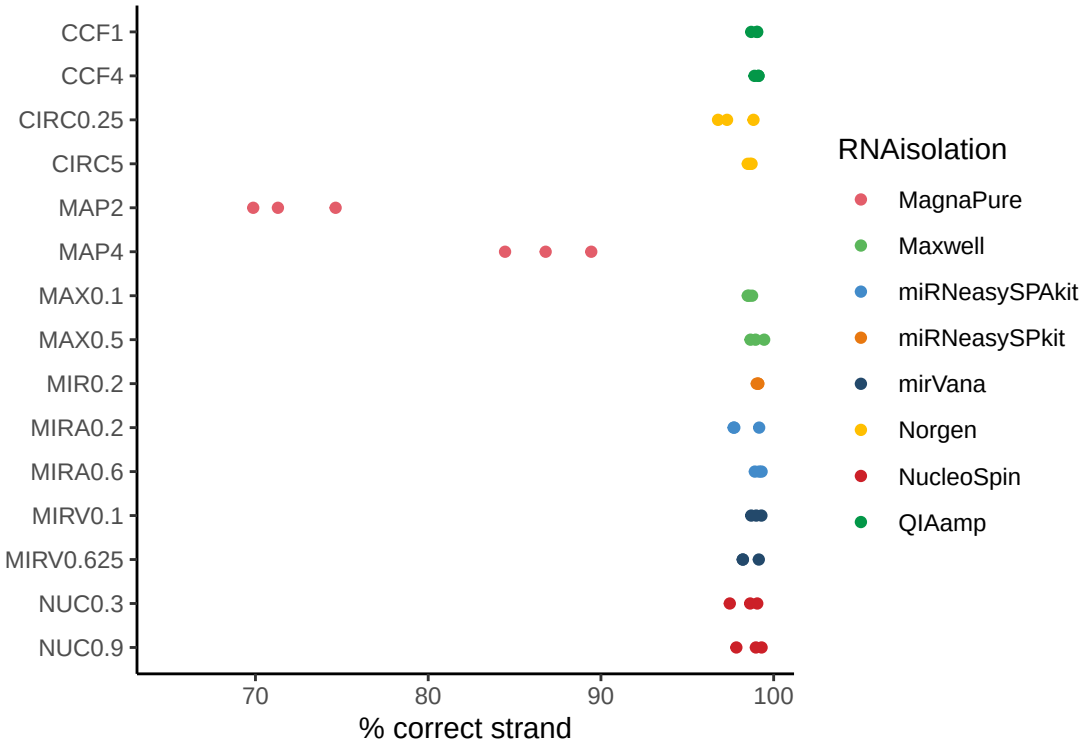


Figure 1: MagnaPure kits excluded based on strandedness (reads coming from other strand may indicate DNA contamination). (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

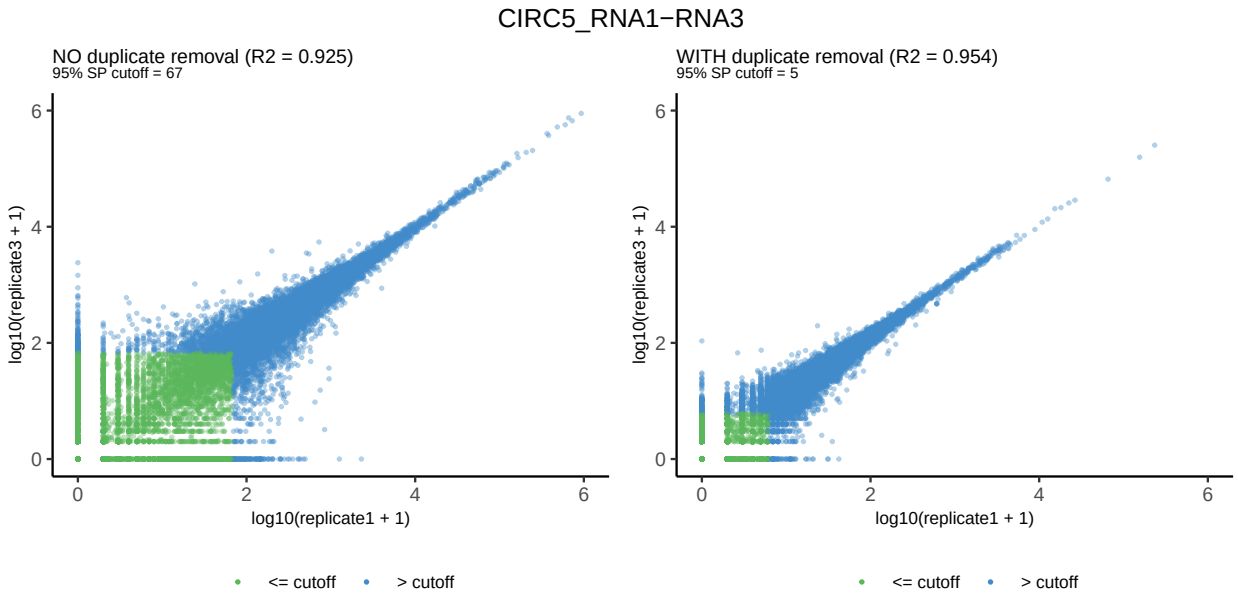


Figure 2: Pairwise RNA count comparison of first and third replicate of the CIRC5 kit without (left) and with (right) duplicate removal. R2 is the coefficient of determination (linear model that fits log10 values). The 95% SP cutoff removes at least 95% of single positives. Single positives are 0 in one replicate and > 0 in other. Green dots show data points that are filtered out with this cutoff. (CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input)

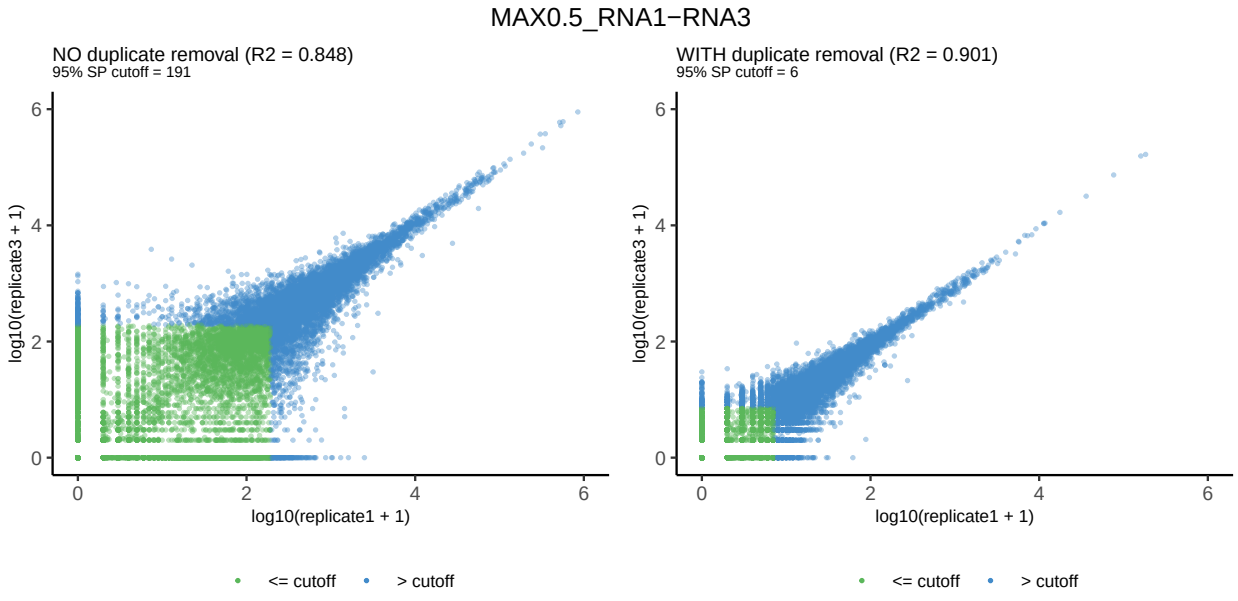


Figure 3: Pairwise RNA count comparison of first and third replicate of the MAX0.5 kit without (left) and with (right) duplicate removal. R2 is the coefficient of determination (linear model that fits log10 values). The 95% SP cutoff removes at least 95% of single positives. Single positives are 0 in one replicate and > 0 in other. Green dots show data points that are filtered out with this cutoff. (MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input)

3.4.2 Duplicate removal

- How to remove (PCR and optical) duplicates?
  - command line: clumpify dedupe=true flag (BBMap/38.26-foss-2018b)
  - parameters: p=20, k=31, s=2 (multiple passes, k=31 default, allowing up to 2 substitutions)
  - each time on first 60 nt of both reads for a pair (to account for lower quality towards end, full 75nt length of unique reads are recovered afterwards)
- Duplicate % is estimated by dividing the number of reads after clumpify by the number of reads after subsampling (no other filtering was applied between these steps)
  - **min=78.8%, mean=92.8%, max=97.9%**
- **Remark: differences in % duplication have a high impact when translating it to number of usable non-duplicated reads!**. In the most extreme case: only 2.1% of mapped reads is usable while for others > 21% is usable!
- CCF4 and MAP4 seem to be the best (but for MAP4 this could be related to DNA contamination see Strandedness)
- Note that Clumpify duplicate removal acts on read sequences (not on mapped reads) so this is not equal to the number of mappable reads

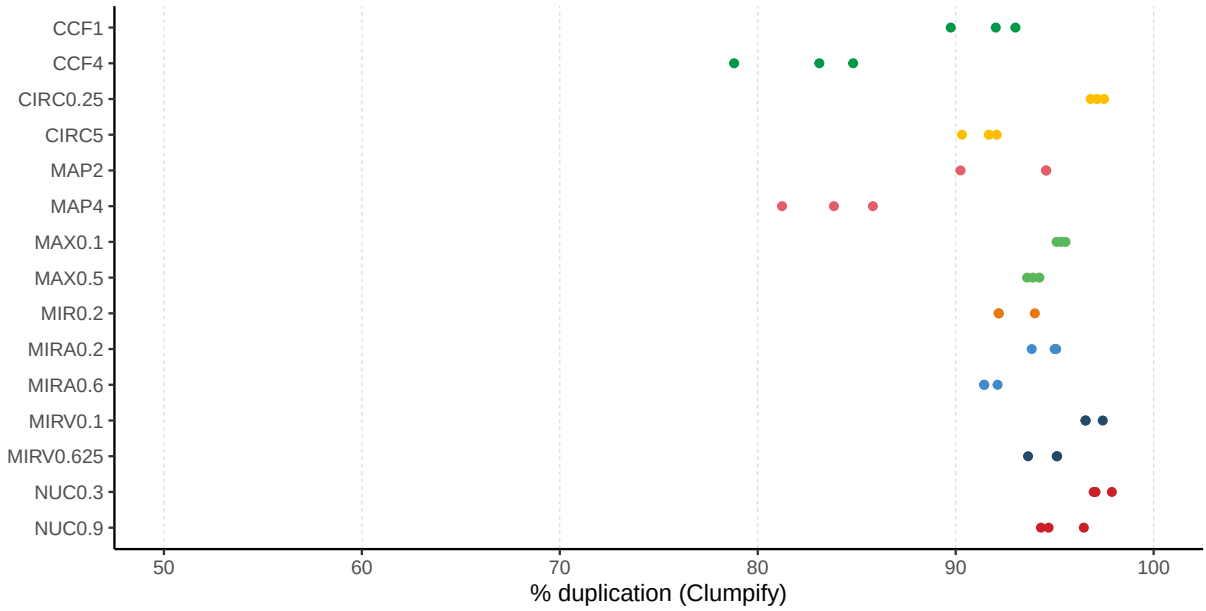


Figure 4: Duplicate percentage. Based on number of reads remaining after Clumpify duplicate removal. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

3.4.3 Link between duplication and plasma input volume?

In plot below, lines connect samples in which RNA was isolated with same kit using lowest and highest input volume. Each time 3 technical replicates: the low and high input samples that were sequenced in same run are connected (but other connections within same kit would also be ok).

Within a kit: higher input V, lower % of duplication. However, this does not explain differences between kits: e.g. MIR0.2 % duplication is almost on same level as CIRC5 (while the plasma input volume is 25x lower). It depends on the RNA concentration and diversity after RNA isolation (see RNA concentration).

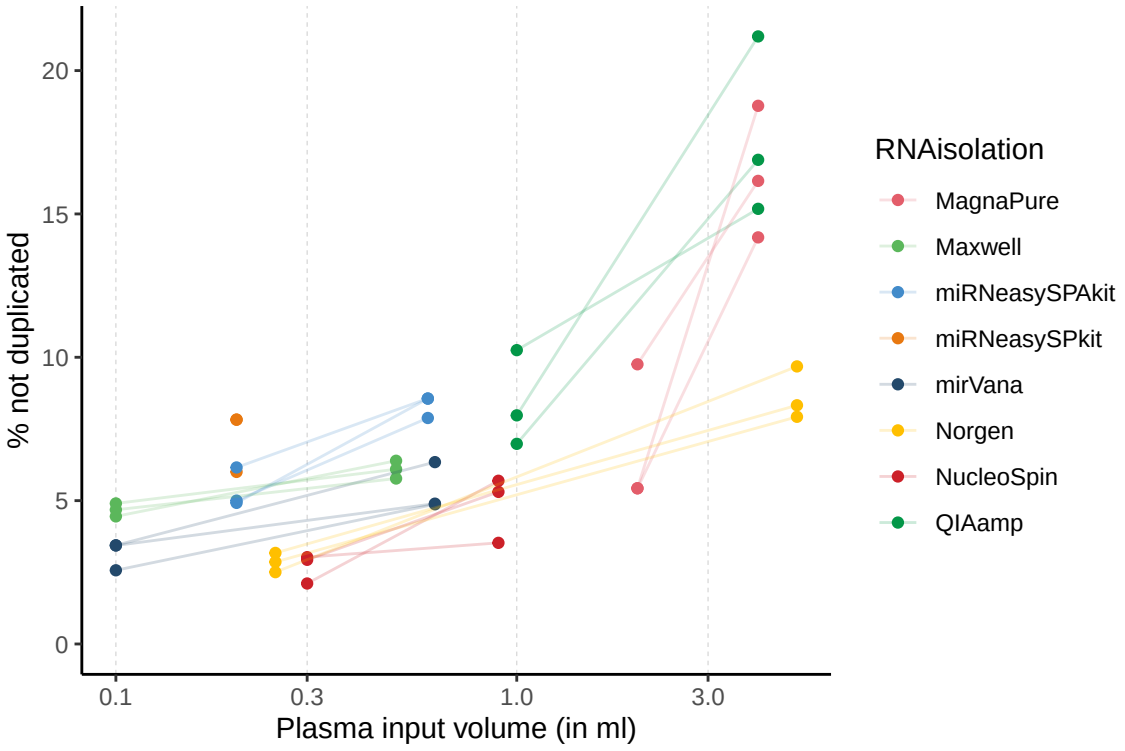


Figure 5: Unique reads vs plasma input volume

3.5 Total number of reads

After 21M subsampling & duplicate removal: min= 443,090; max= 4,450,430; mean= 1,520,642 read pairs

3.6 Gene count conversion

Our pipeline converts reads to spike and transcript counts using Kallisto, based on Ensemblv91. For further processing, we gathered these count and TPM dataframes for all samples, and calculated counts per million (CPM). To aggregate counts at gene level, transcripts counts (or TPM values) are grouped per gene and summed. We also summed spike counts per sample (separate summation for Sequin and ERCC spikes)

MAP samples filtered out (see Strandedness)

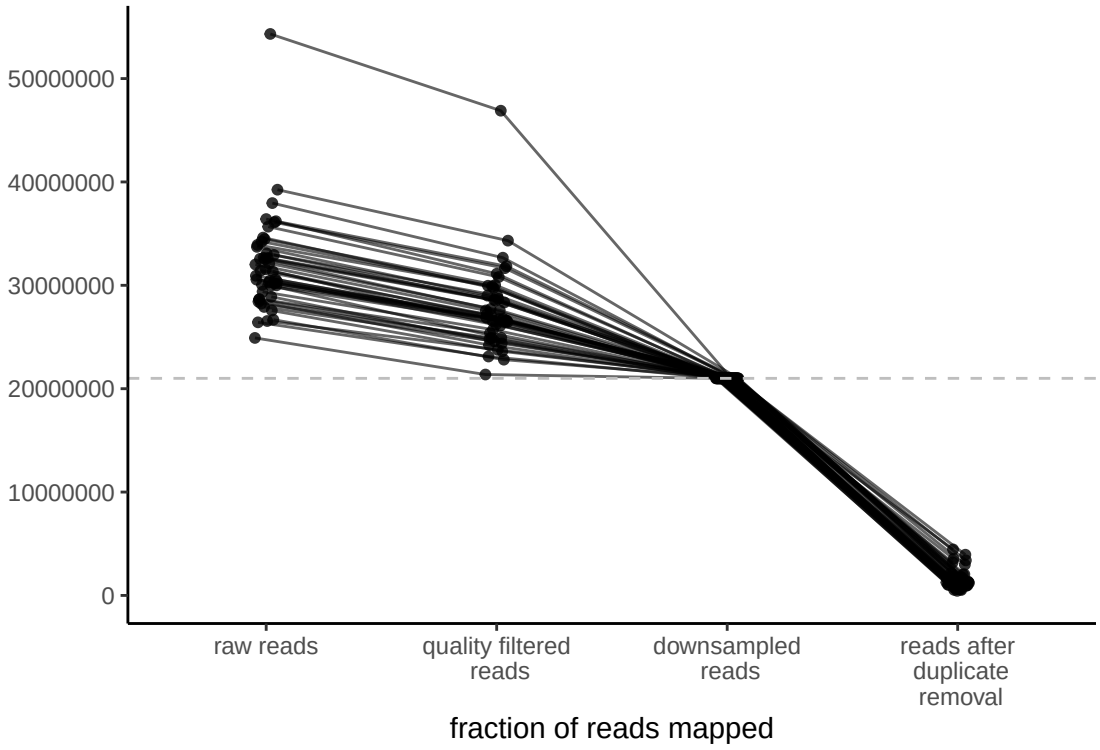


Figure 6: Number of reads at different stages in the data processing workflow. Raw reads: total number of read pairs in raw FASTQ files at start; quality filtered reads: number of read pairs where at least 80% of bases in both reads have a phred score of 20 or higher; reads after downsampling: all samples were downsampled to 21M paired end reads; reads after duplicate removal: number of downsampled read pairs remaining after Clumpify duplicate removal.

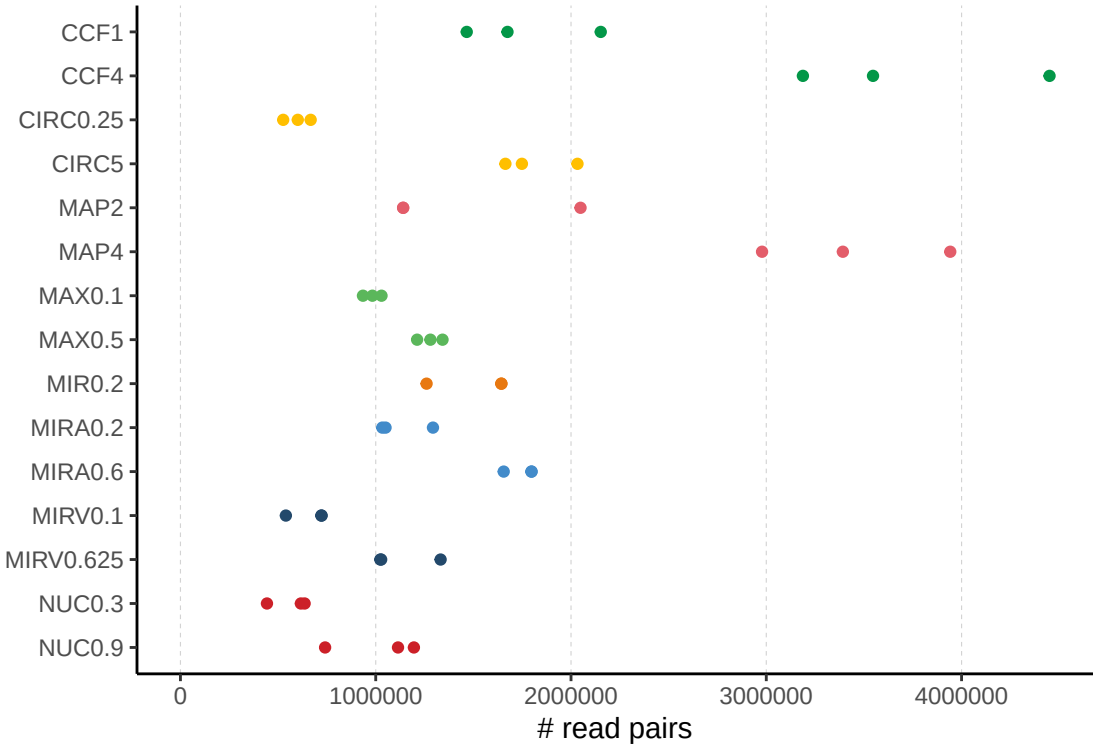


Figure 7: Number of paired end reads remaining after duplicate removal. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

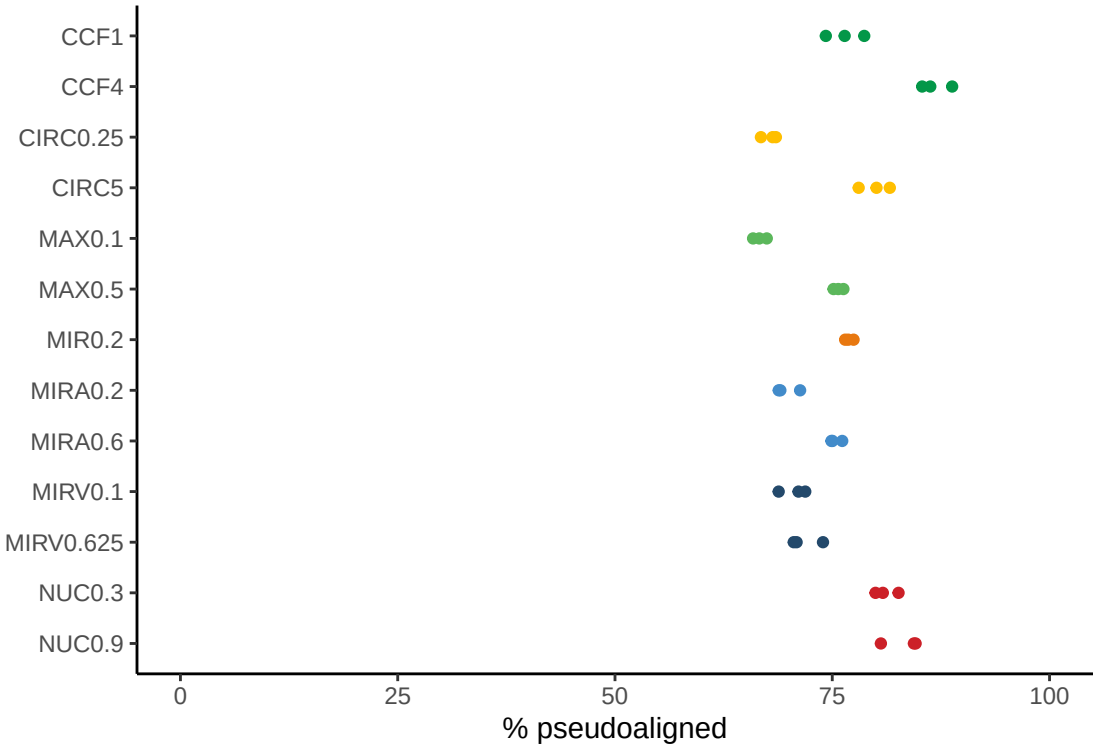


Figure 8: % of reads that are pseudoaligned by kallisto - after duplicate removal with Clumpify. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

4 Performance metrics

Nine performance metrics are calculated.

In order to compare different kits in a uniform way using these metrics, we calculated **z-scores**. Before z-score transformation, we made sure that a **higher value always corresponds to better performance**. To account for the low sample size, we calculated **robust z-scores**.

4.1 Duplication level

If the used fraction of the eluate contains more RNA (in absolute numbers and in terms of diversity), there will be less duplicates (see Duplicate rate) & more reads remaining **Scoring: % unique (100% - % duplication) to make sure higher is better**

4.2 RNA concentration

ERCC spikes were each time added in the same amount (2 microL) to 12 microL of eluate after RNA isolation. The ratio of endogenous RNA to ERCC reflects the relative concentration of endogenous RNA in the eluate. The higher the endogenous RNA concentration in used fraction of eluate, the less ERCCs, the higher the ratio endo/ERCC. Remember that some kits have a much larger eluate volume after RNA isolation. A larger total eluate volume results in more diluted endogenous RNA (lower concentration) and therefore less endogenous RNA in library prep (given constant input volume for all library preparations).

Scoring: the higher the concentration, the better

4.3 RNA yield

For RNA sequencing purposes, we are most interested in the concentration of the eluate as we can only use a limited amount of volume during library prep. However, by multiplying the relative RNA concentrations above with the total eluate volume, we get an idea of the relative RNA yield in the eluate after RNA isolation. In case the total eluate volume is larger, the RNA will be more diluted (this is for example the case for MIRV: 100 microL eluate compared to only 12 microL for CCF).

If the RNA yield is high, but the eluate volume is large, further concentrating the total eluate before library prep might give better results for your experiment. However, we did not evaluate this in our study. **Scoring: the higher the yield, the better**

4.4 Efficiency of kits

Based on the previous plot with RNA yield in eluate, we observe differences in efficiencies among kits (kit with low input volume might isolate input RNAs more efficiently). By **correcting the yield for the plasma input volume**, we obtain a better picture:

- **with more input, you expect to have more yield in eluate. To correct for this: divide yield by input volume**
- (endogenous/ERCC)*EluateV / PlasmaInputV
- e.g. CCF1 has 10x more plasma input (=> more RNA) than MAX0.1, but this does not result in 10x more yield. MAX0.1 seems to extract the lower volume better than CCF1
- Although the yield is higher within a given kit when using the maximum input volume, this sometimes seems to be associated with a lower efficiency than the minimal input volume

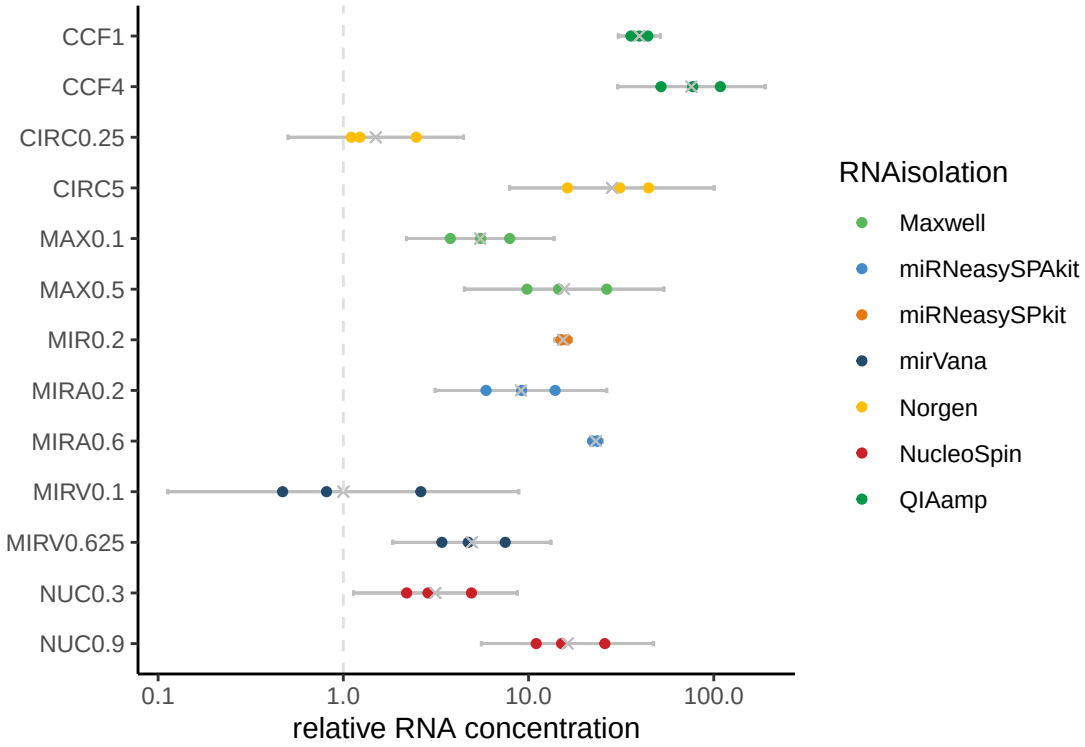


Figure 9: Relative RNA concentration in eluate after RNA purification. Concentration: ratio of endogenous RNA to ERCC spikes. Values were first log transformed and rescaled to average of MIRV0.1, then transformed back to linear scale. Mean per kit (cross) and 95% confidence intervals shown (grey lines). (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

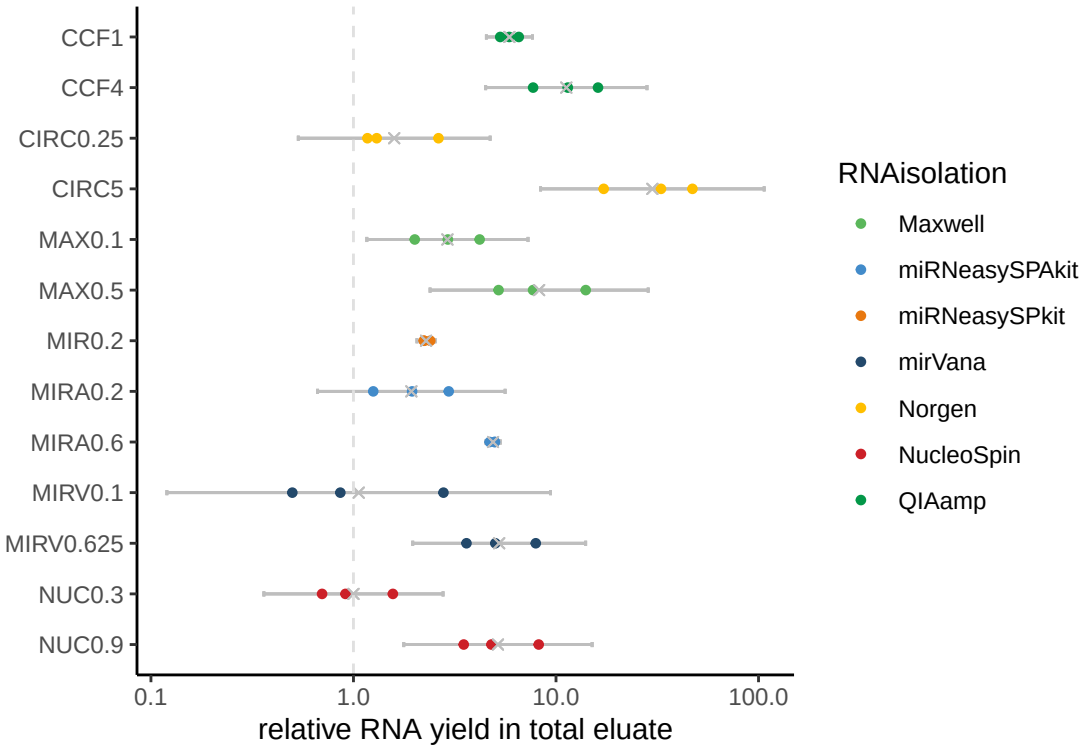


Figure 10: Relative RNA yield of kits. Yield: eluate volume corrected RNA concentration. Values were first log transformed and rescaled to the average of MIRV0.1, then transformed back to linear scale. Mean per kit (cross) and 95% confidence intervals shown (grey lines). (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

Remark: while we could also look at Sequin/ERCC ratio corrected for input and eluate volume (should give similar results), we decided to look at endogenous RNA as a more representative metric as this is the biomaterial of interest.

There is a clear difference in kit efficiency, with a difference of factor 10 or more.

Note the variability between technical replicates: for some kits the performance on the three replicates is very similar (e.g. MIR0.2 and MIRA0.6), for others it is quite variable (e.g. MIRV0.1)

If some adjustments would be made to kits with low input volume but high efficiency (i.e. increase allowed plasma input V and keep eluate V as small as possible), **the overall performance may further improve**. Of note, we did not evaluate this in our study.

Scoring: the higher the efficiency, the better

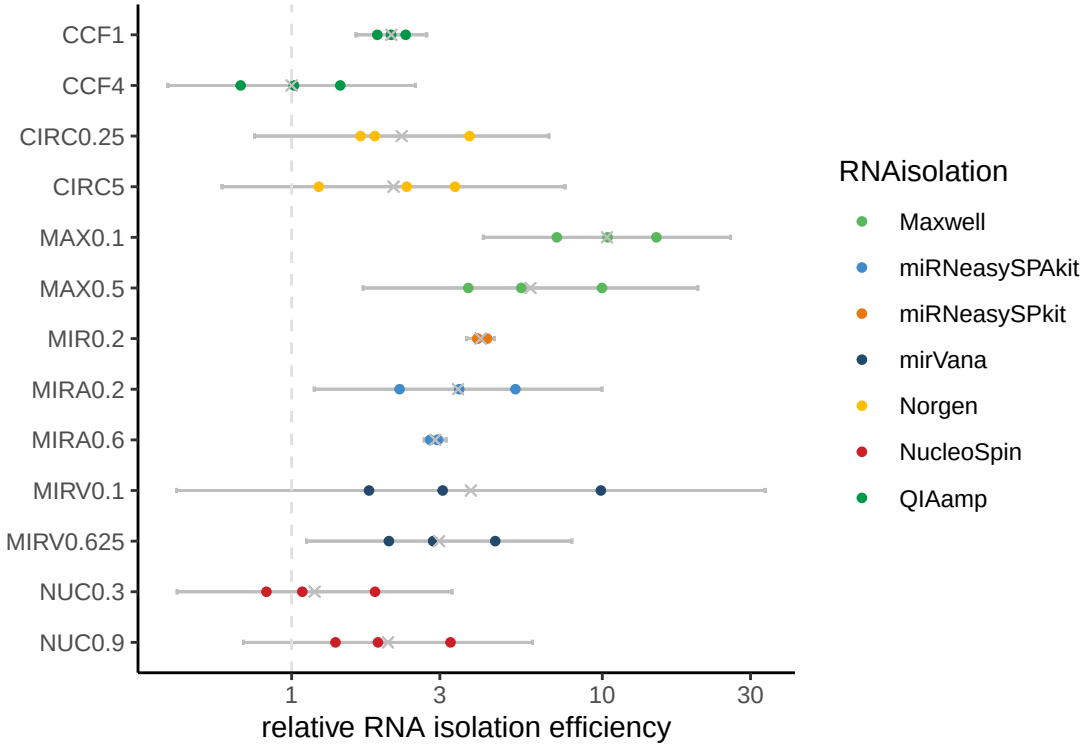


Figure 11: Relative efficiency of kits. Efficiency: plasma input volume corrected RNA yield. Values were first log transformed and rescaled to the average of CCF4, then transformed back to linear scale. Mean per kit (cross) and 95% confidence intervals shown. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

4.5 Filter threshold

- We want to have a filter threshold that eliminates 95% of Single Positives between technical replicates, i.e. genes detected in only one technical replicate (cf. miRQC study of Mestdagh et al., 2014, Nature

Methods):

  - All pseudocounts in Kallisto < 1 are first rounded down to 0
  - Pairwise comparison of technical replicates (3 pairs per kit-volume combination)
  - Determine threshold at which at least 95% of the single positives are removed (this threshold can be a decimal number as a result of kallisto quantification)
  - Take median cutoff per combination of kit and volume

- **95% SP elimination cutoff (which is specific for each kit-volume combination) will be used for filtering throughout ALL analyses**

  - This is our proposed strategy to make data comparable, we do not claim that this is the only way to do this

4.5.1 Cutoff examples

This tab shows two examples of pairwise kit-volume comparisons together with their cutoff and R-squared value (based on linear model (y=x) of log counts). Histograms show the relative amount of RNAs with counts in that bin. For an overview of the cutoffs for each kit-volume combination, see next tab 95% SP elimination cutoffs.

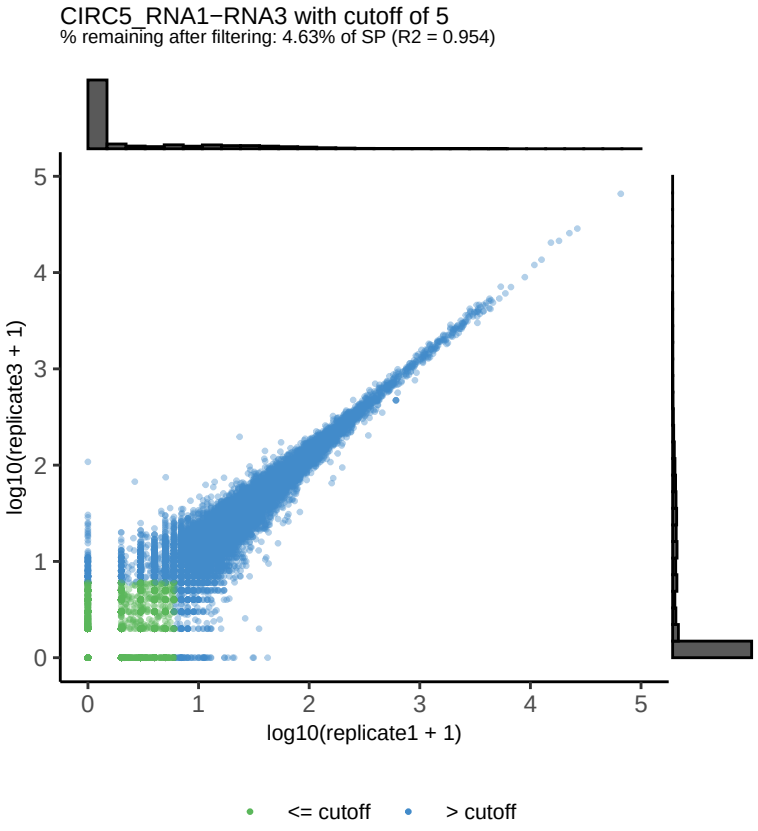


Figure 12: Pairwise RNA count comparison of first and third replicate of the CIRC5 kit. The coefficient of determination is 0.954 (linear model that fits log10 values). Single positives are 0 in one replicate and > 0 in other. The cutoff of 5 removes 95.37% of single positives. Green dots show data points that are filtered out with this cutoff. (CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input)

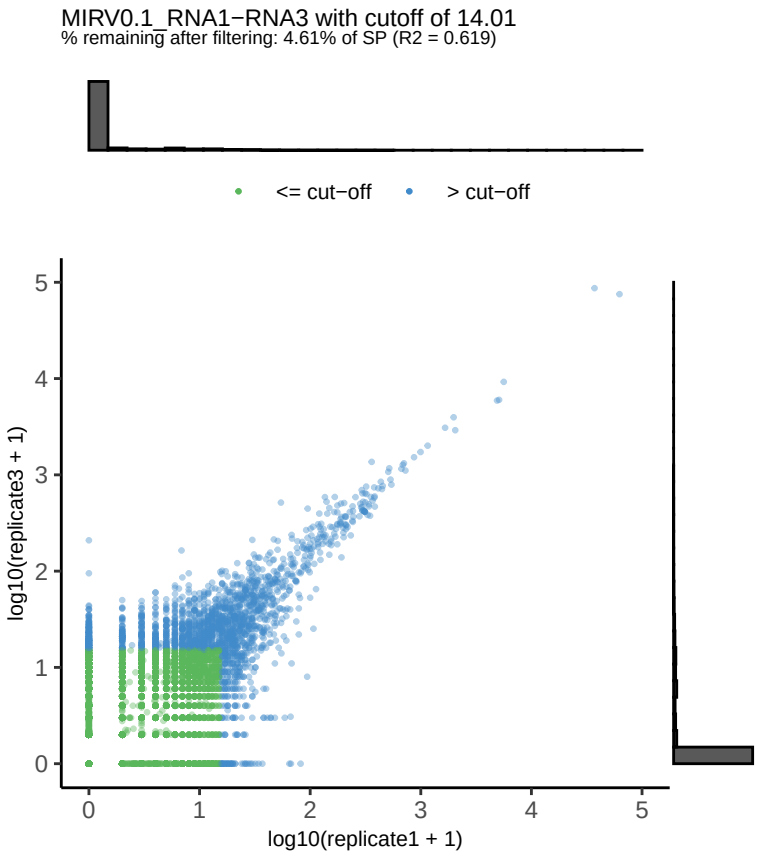


Figure 13: Pairwise RNA count comparison of first and third replicate with MIRV0.1 kit. Coefficient of determination is 0.619 (based on linear model that fits log10 values). Single positives are 0 in one replicate and > 0 in other. The cutoff of 14.01 removes 95.6% of single positives. Green dots show data points that are filtered out with this cutoff. (MIRV0.1: mirVana PARIS kit, 0.1 ml input)

4.5.2 95% SP elimination cutoffs

- If all counts smaller than or equal to cutoff are eliminated, at least 95% of single positives are removed, resulting in data that is highly reproducible
- **Cutoff is always higher for lower input volume within same kit** (with lower input volume, there is more variation in which genes are detected in each replicate)
- **Cutoffs are close to each other** BUT 1 count difference can already have a major impact on the number of genes filtered out
- We use the median cutoff per kit-volume combination for filtering in further analyses (see table below)
- **Scoring: take the negative of the cutoff values (so that higher = better precision)**

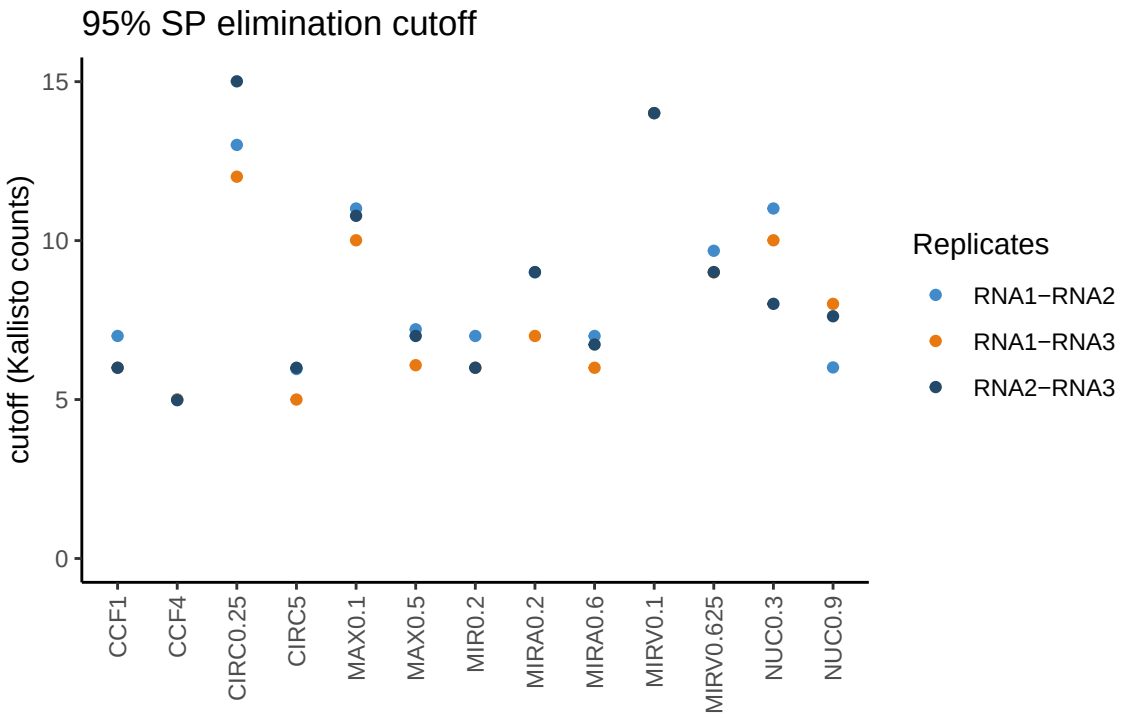


Figure 14: Count threshold that removes 95% of single positives for each pairwise comparison of replicates. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

4.5.3 Impact of filtering

- After applying the repeatability cutoff, remaining counts per sample: min=178,242 mean=896,101 max=3,600,928

Table 2: Median cutoff per kit

Kit	cutoff
CCF1	6.00
CCF4	5.00
CIRC0.25	13.01
CIRC5	5.96
MAX0.1	10.78
MAX0.5	7.00
MIR0.2	6.00
MIRA0.2	9.00
MIRA0.6	6.73
MIRV0.1	14.01
MIRV0.625	9.01
NUC0.3	10.01
NUC0.9	7.62

- **Scoring: data retention: more % of counts remaining = better precision**
  - Is related to the cutoff & initial amount of reads (after duplicate removal)

4.5.4 Robustness of cutoff

We tested how robust these 95% single positive elimination cutoffs are at different downsampling levels - For some kits, this cutoff is very stable, while for others it keeps on increasing with a higher subsampling level. - Also differences within same purification kit: stable cutoff for CIRC5, but cutoff increases in CIRC0.25 with higher subsampling levels. - Most variability in MIRV0.1, CIRC0.25, NUC0.3 & MAX0.1.

Within a kit, cut-off more stable for high than for low input volume, but more related to RNA concentration in eluate than plasma input volume. For example, MIR0.2 has a slightly more stable cut-off than MIRA0.2 and a much more stable cut-off than CIRC0.25. Possible explanation: less RNA in eluate -> more stochastic variation in RNA between replicates -> sequencing deeper does not remove stochastic variation, it just increases counts and therefore cut-off.

4.6 Number of genes

- Filter: only keep **protein coding genes** that reach the median 95% SP cutoff per kit in terms of counts (Kallisto)
- Observations before and after filtering:
  - Quite some variability between kits
  - Overall trend is that a higher input volume within a kit results in a higher amount of detected genes
- **Scoring: more genes that reach reproducibility threshold = better**

4.7 Coverage

Look at how many % of the transcriptome is covered at least once (based on genomeCoverageBed (BEDtools 2.27))

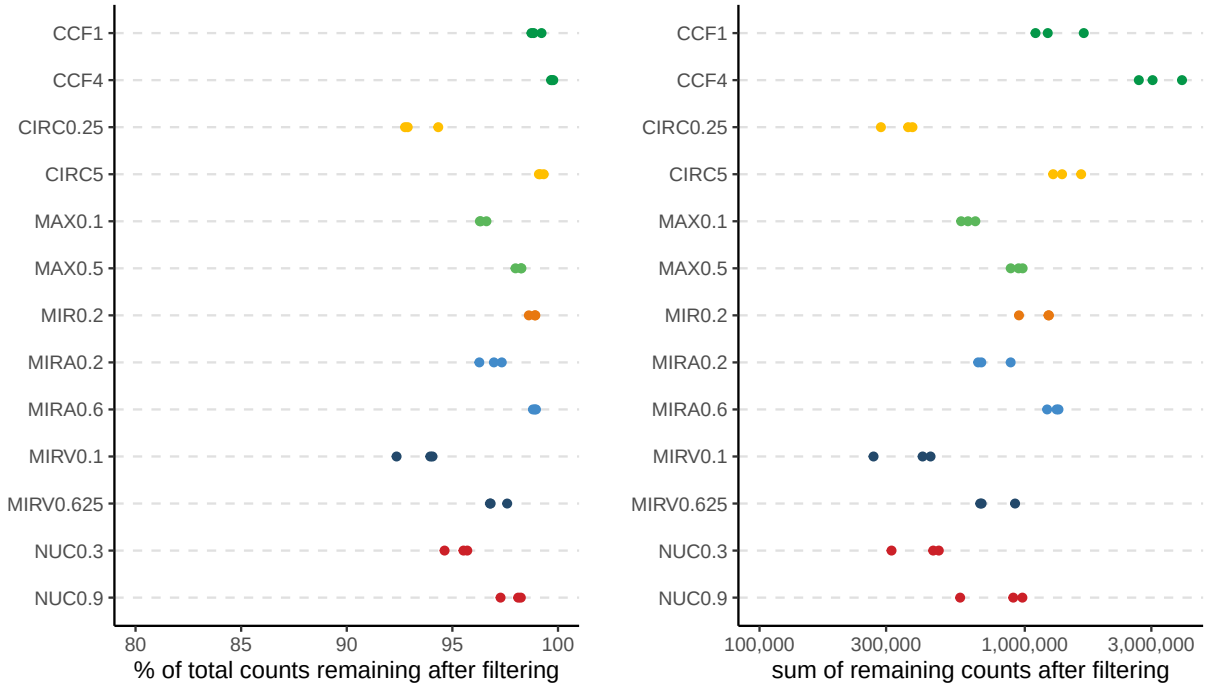


Figure 15: Impact of filtering (filter removes 95% of single positives). Left: % of total counts that are kept after applying filter; right: sum of counts that are not filtered out. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

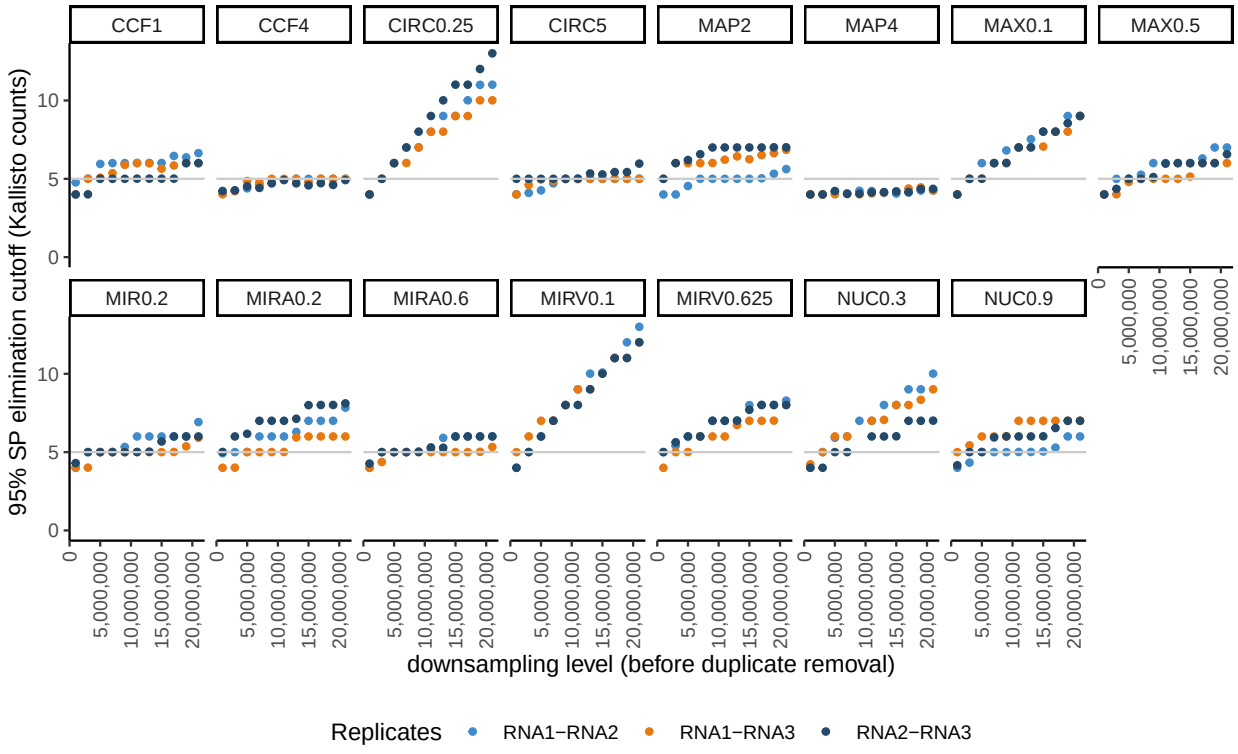


Figure 16: Robustness of filter threshold at different downsampling levels. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

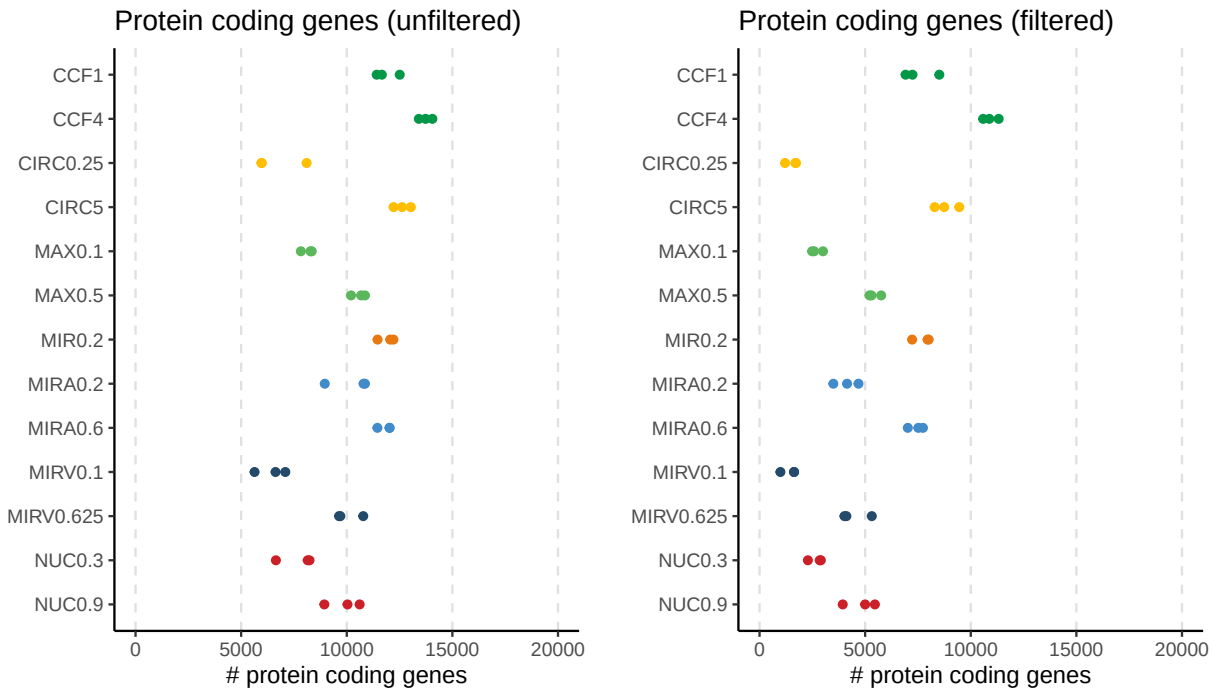


Figure 17: Number of protein coding genes that are detected. Left: all protein coding genes that are detected with at least one count; right: protein coding genes that are reproducibly detected (precision threshold that eliminates 95% of single positives). (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

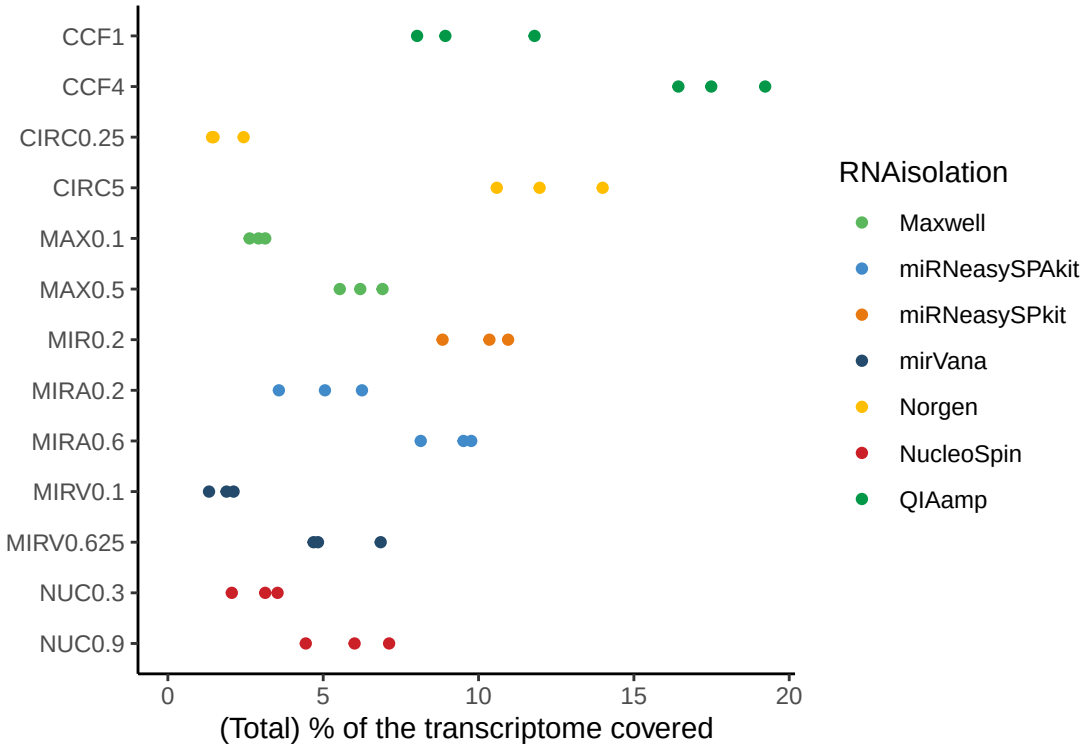


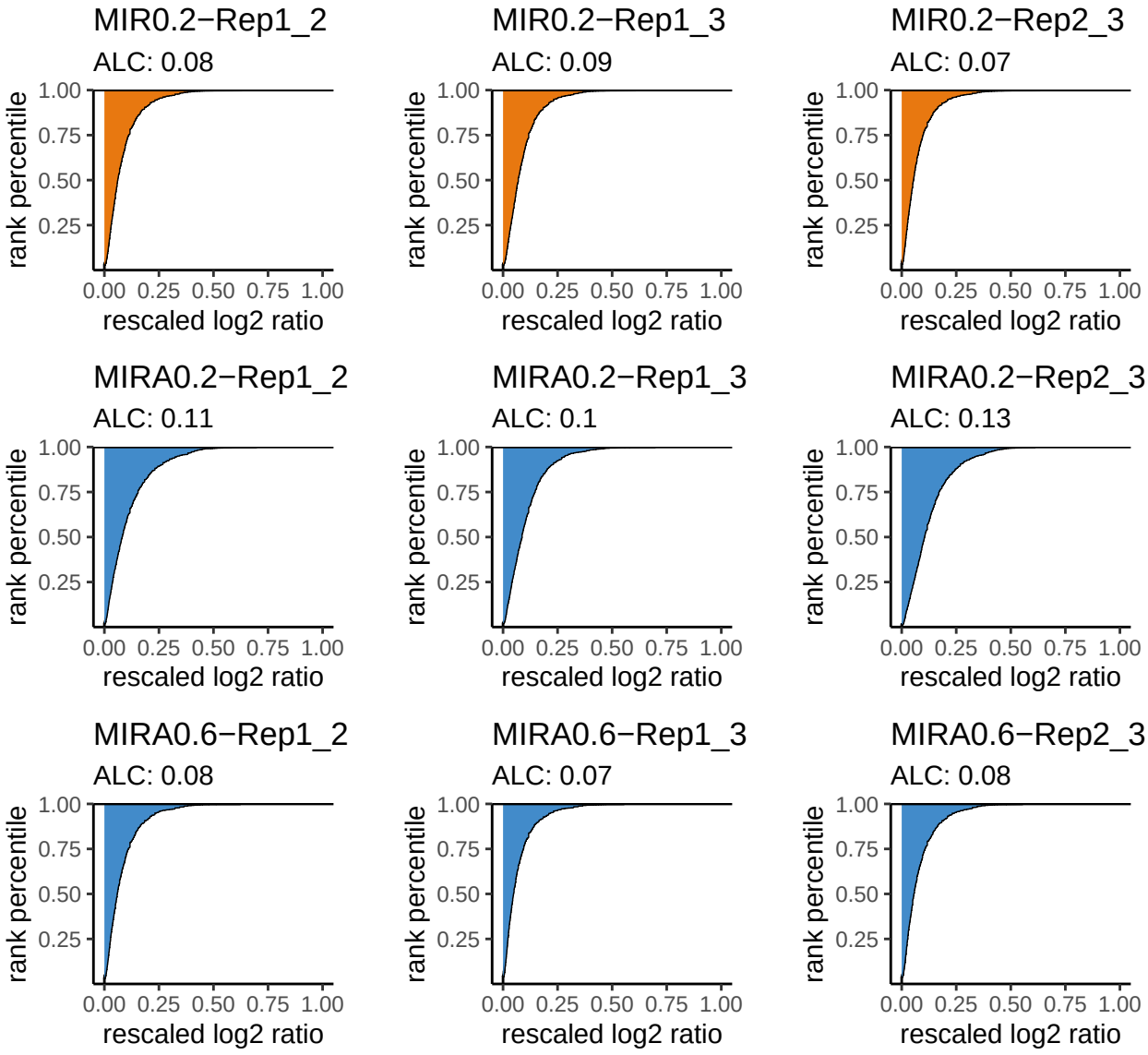
Figure 18: Percentage of the transcriptome that is covered at least once. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

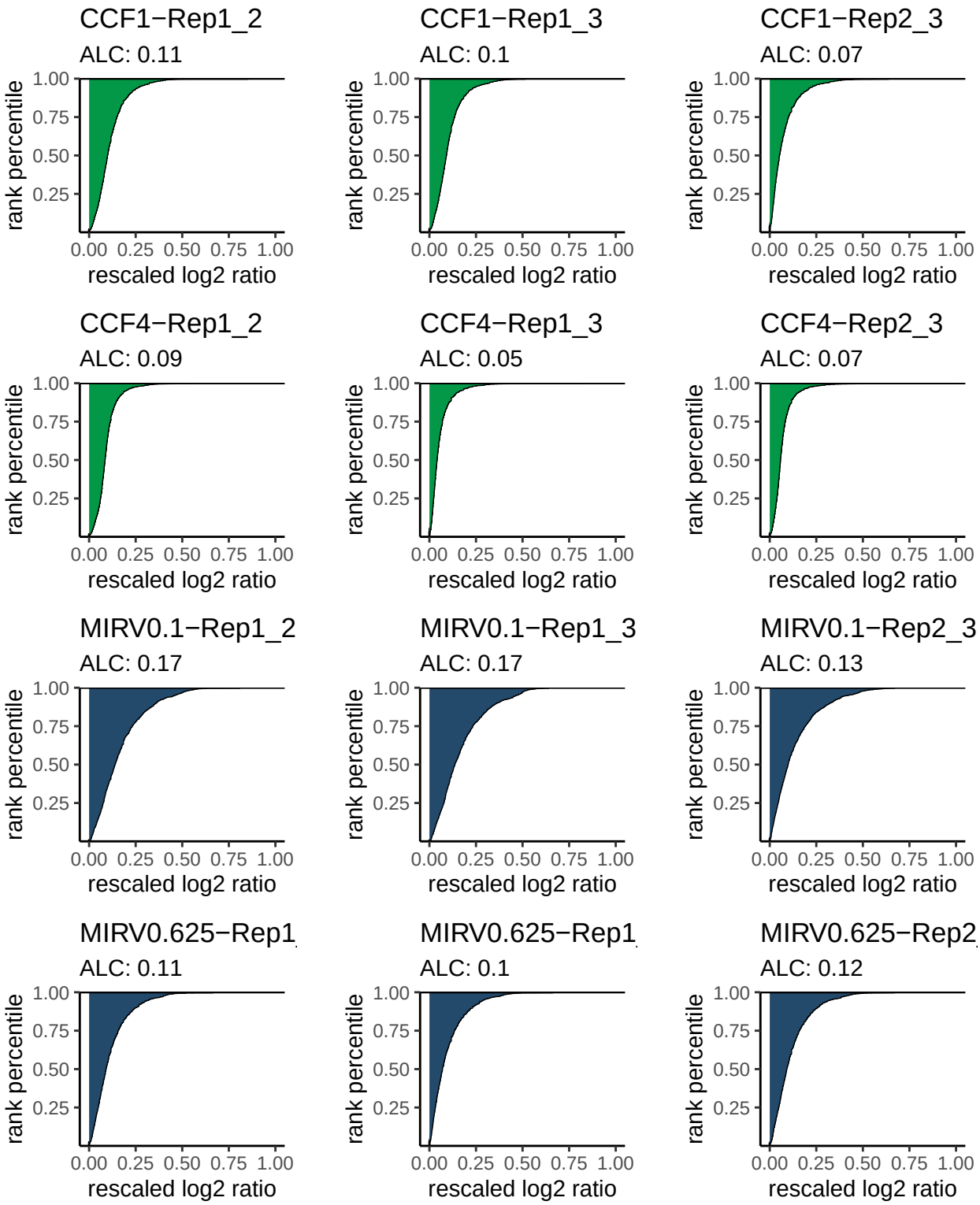
4.8 ALC

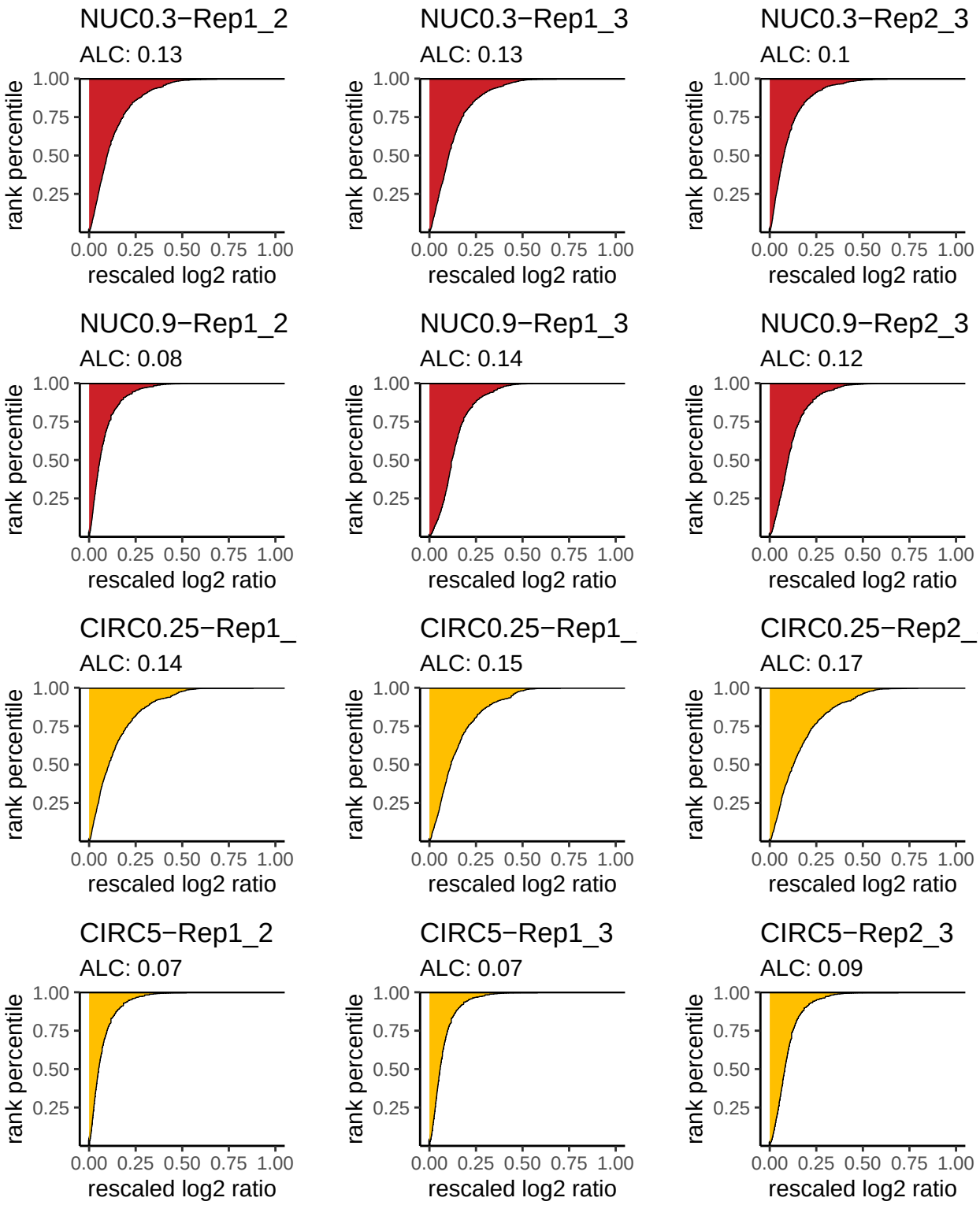
ALC (area-left-of-curve) **as repeatability metric** (see Mestdagh et al., 2014, Nature Methods). Compare two technical replicates at a time, only considering genes that reach the precision threshold (which eliminates 95% of single positives) in at least one of the samples and 1 count in the other sample. Replace all zero counts by NA. Determine log2 ratios of counts for each gene. Plotted are the (percentage) ranks of these absolute log2 ratios. The faster the curve reaches 100% (the smaller the ALC), the better. A small ALC indicates that the replicates give similar counts for each gene.

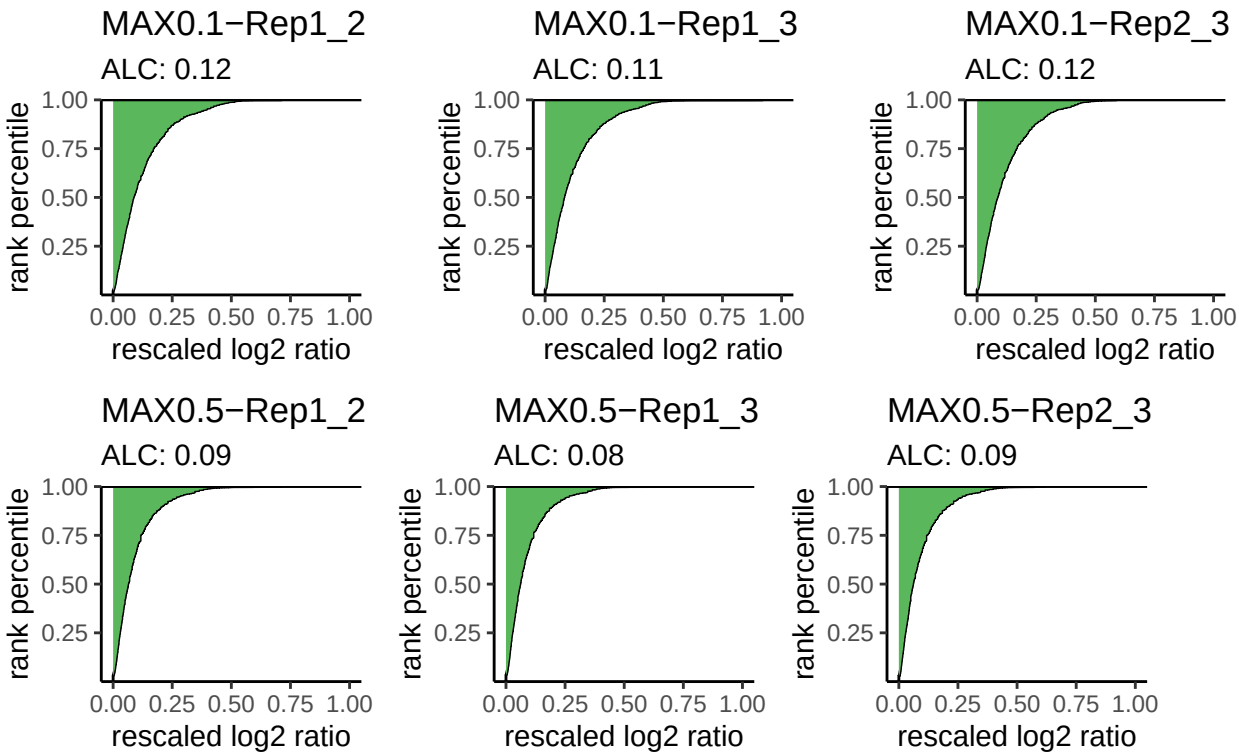
Scoring: lower ALC = better precision

4.8.1 Individual ALC plots




CCF1−Rep1_2
ALC: 0.11
CCF1−Rep1_3
ALC: 0.1
CCF1−Rep2_3
ALC: 0.07
CCF4−Rep1_2
ALC: 0.09
CCF4−Rep1_3
ALC: 0.05
CCF4−Rep2_3
ALC: 0.07
MIRV0.1−Rep1_2
ALC: 0.17
MIRV0.1−Rep1_3
ALC: 0.17
MIRV0.1−Rep2_3
ALC: 0.13
MIRV0.625−Rep1
ALC: 0.11
MIRV0.625−Rep1
ALC: 0.1
MIRV0.625−Rep2
ALC: 0.12
rank percentile
1.00
0.75
0.50
0.25
0.00 0.25 0.50 0.75 1.00
rescaled log2 ratio


NUC0.3−Rep1_2
ALC: 0.13
NUC0.3−Rep1_3
ALC: 0.13
NUC0.3−Rep2_3
ALC: 0.1
NUC0.9−Rep1_2
ALC: 0.08
NUC0.9−Rep1_3
ALC: 0.14
NUC0.9−Rep2_3
ALC: 0.12
CIRC0.25−Rep1_
ALC: 0.14
CIRC0.25−Rep1_
ALC: 0.15
CIRC0.25−Rep2_
ALC: 0.17
CIRC5−Rep1_2
ALC: 0.07
CIRC5−Rep1_3
ALC: 0.07
CIRC5−Rep2_3
ALC: 0.09
rank percentile
1.00
0.75
0.50
0.25
0.00 0.25 0.50 0.75 1.00
rescaled log2 ratio



4.8.2 Overview ALC

4.9 Overview

Based on robust z-scores (for each performance metric: higher robust z-score is better)

4.9.1 Correlation between metrics

- Spearman correlation of robust z-scores (+hierarchical clustering)
- Overall, quite high correlation.
  - Yield is a bit less correlated and efficiency is a clear outlier, however, these are theoretical metrics: how well would the kit perform regardless of input and/or eluate volume restrictions (see [Yield] and [Effiency])
- Diversity/abundance related metrics show a high correlation
- The two precision metrics (ALC and filter threshold) highlight a different aspect of precision: ALC shows how similar gene counts between replicates are (see ALC), while the threshold gives an idea of the amount of noise - from which count threshold onward can you "trust" count values (see Filter threshold)

4.9.2 Comparison of kits

Many kits do quite well despite a rather low plasma input volumes (plasma input volume, in ml, is each time attached to the abbreviation of the kit)

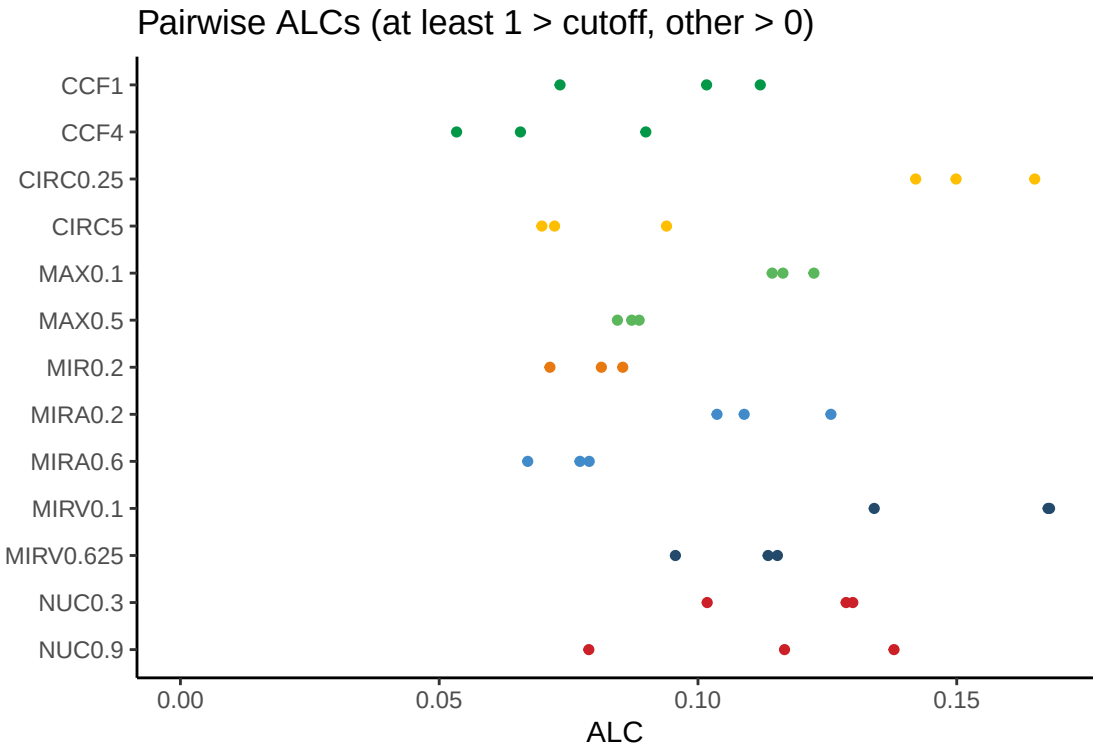


Figure 19: Precision based on ALCs (area-left-of-curve). The lower the ALC, the better (less difference between replicates). (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

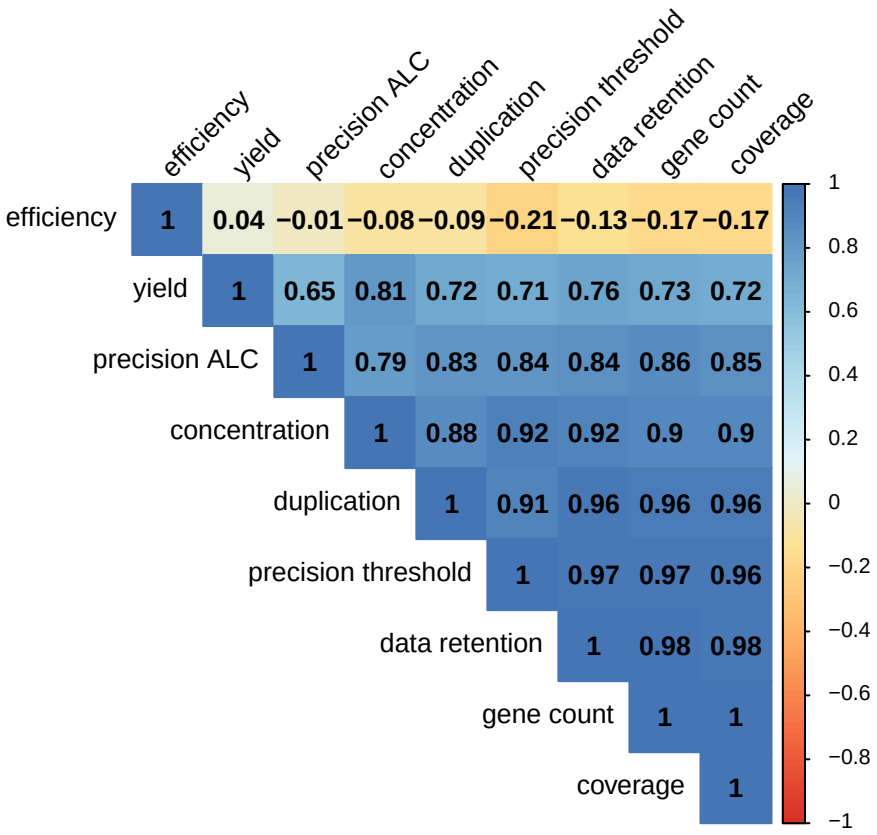


Figure 20: Correlation among metrics based on robust z-scores. Spearman correlation with complete hierarchical clustering. (concentration: relative endogenous miRNA concentration based LP spikes; data retention: % of counts left after applying precision threshold; efficiency: yield corrected for plasma input volume ; gene count: number of protein coding genes after applying precision threshold; precision ALC: precision based on area-left-of-curve; precision threshold: count threshold to filter out 95% of single positives between replicates; yield: concentration corrected for eluate volume; coverage: % of transcriptome that is covered at least once; duplication: % duplicated reads)

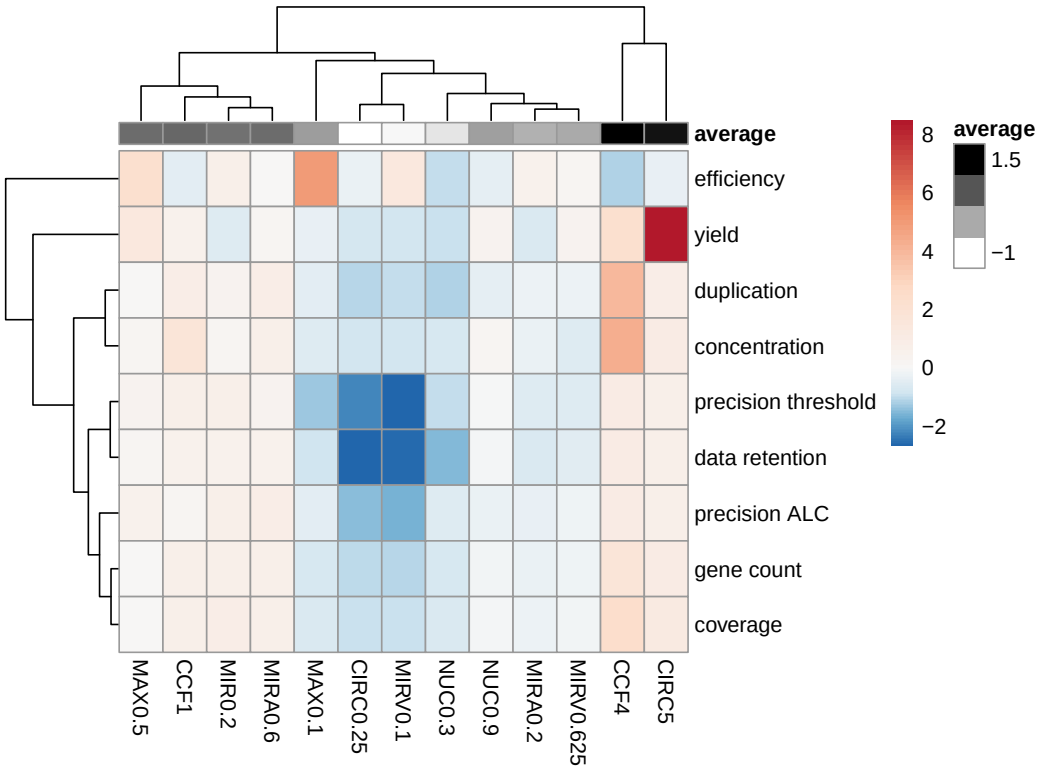


Figure 21: Comparison of kit performance based on robust z-scores. Higher means a better performance. Complete hierarchical clustering based on Euclidean distance. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input; concentration: relative endogenous miRNA concentration based LP spikes; data retention: % of counts left after applying precision threshold; efficiency: yield corrected for plasma input volume ; miR count: number of miRNAs after applying precision threshold; precision ALC: precision based on area-left-of-curve; precision threshold: count threshold to filter out 95% of single positives between replicates; yield: concentration corrected for eluate volume; average: average z score over all metrics)

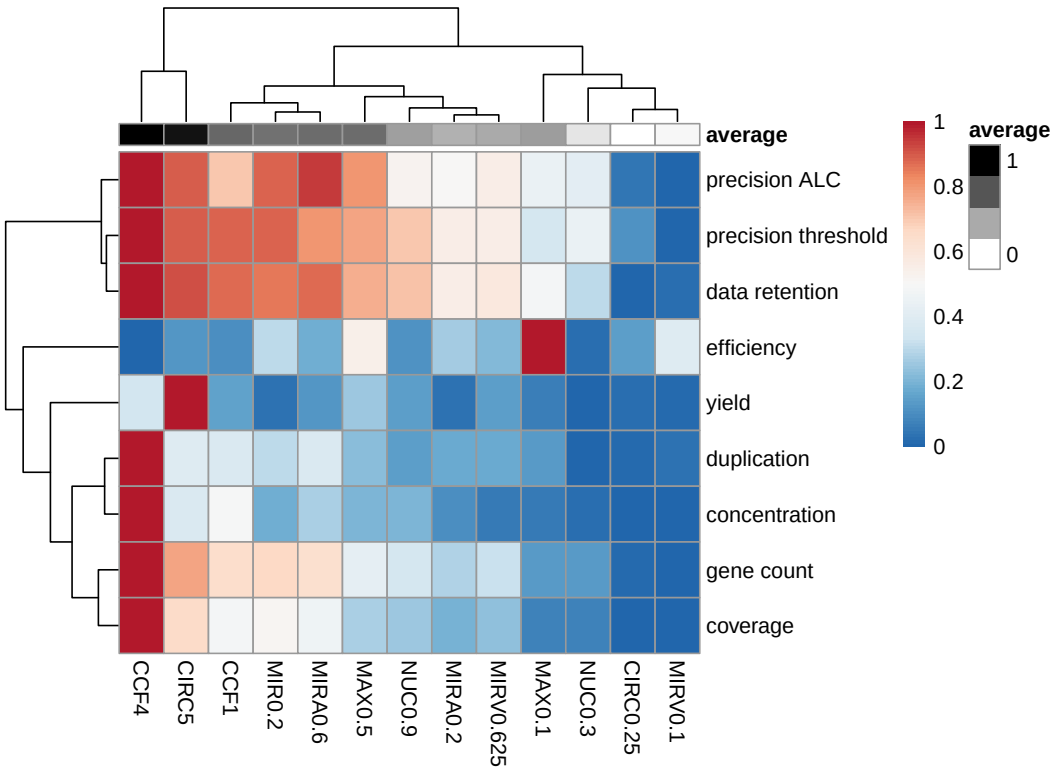


Figure 22: Comparison of kit performance based on robust z-scores - rescaled to [0,1] to stress difference within a metric. Higher means a better performance. Complete hierarchical clustering based on Euclidean distance. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input; concentration: relative endogenous miRNA concentration based LP spikes; data retention: % of counts left after applying precision threshold; efficiency: yield corrected for plasma input volume ; miR count: number of miRNAs after applying precision threshold; precision ALC: precision based on area-left-of-curve; precision threshold: count threshold to filter out 95% of single positives between replicates; yield: concentration corrected for eluate volume; average: average z score over all metrics)

5 Selection for phase 2

Selection of two kits for phase 2 of the study is based on robust z-score transformed metric for sensitivity (# detected genes, see Number of genes) and for reproducibility (area-left-of-curve, see ALC). Higher z-score = better

We looked at both metrics but in close calls, the sensitivity was given a higher weight. Moreover, we wanted to include at least one kit which is able to purify RNA in case you have less than 1ml of plasma as it is not always possible to collect or use multiple ml.

Selection: QIAamp (**CCF4**) & miRNeasy serum/plasma (**MIR0.2**)

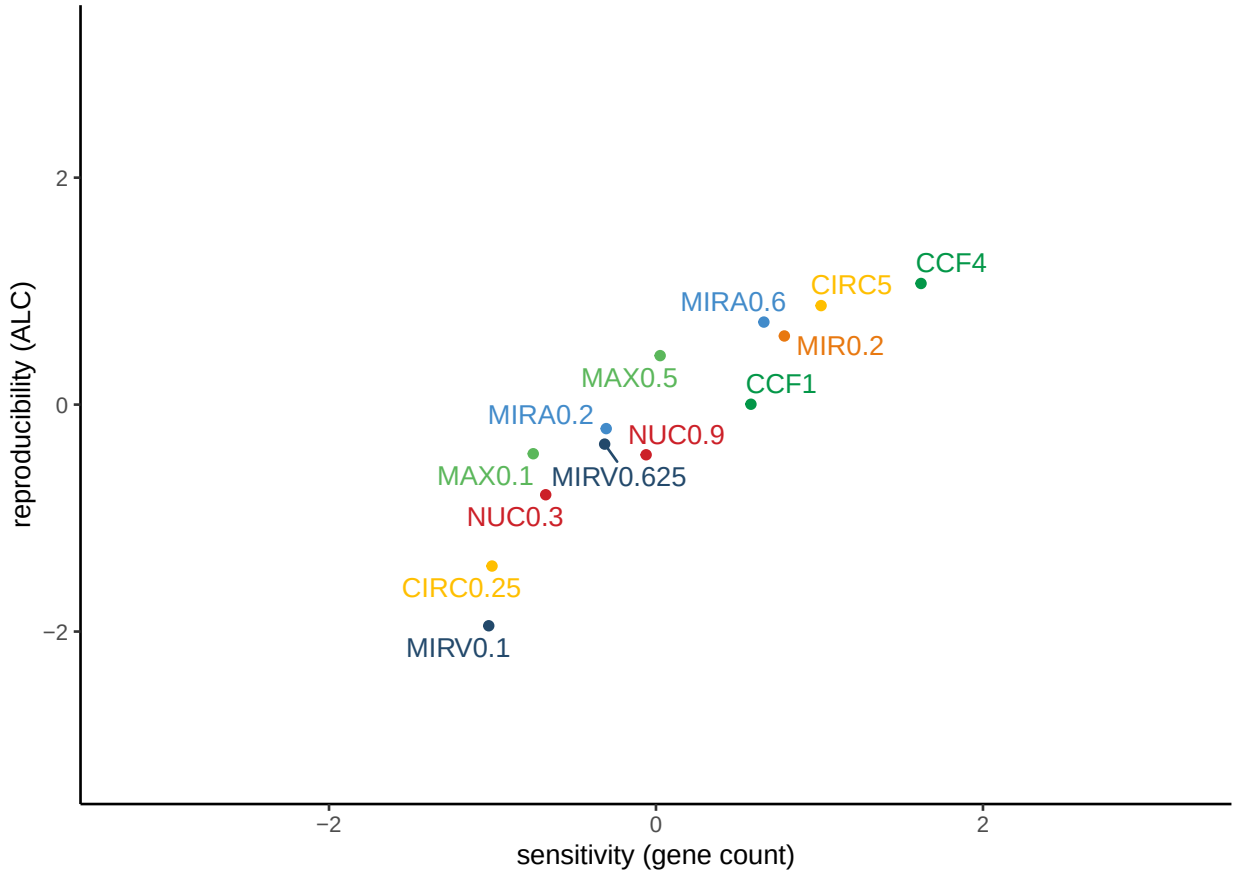


Figure 23: Robust z-scores (median per RNA isolation kit) for sensitivity (number of miRNAs) (x) and precision (ALC, area-left-of-curve) (y). CCF and MIR0.2 kits are selected for phase II. (CCF1: QIAamp ccfDNA/RNA kit, 1 ml input; CCF4: QIAamp ccfDNA/RNA kit, 4 ml input; CIRC0.25: plasma/serum circulating and exosomal RNA purification kit, 0.25 ml input; CIRC5: plasma/serum circulating and exosomal RNA purification kit, 5 ml input; MAX0.1: Maxwell RSC miRNA plasma and exosome kit, 0.1 ml input; MAX0.5: Maxwell RSC miRNA plasma and exosome kit, 0.5 ml input; MIR0.2: miRNeasy serum/plasma kit, 0.2 ml input; MIRA0.2: miRNeasy serum/plasma advanced kit, 0.2 ml input; MIRA0.6: miRNeasy serum/plasma advanced kit, 0.6 ml input; MIRV0.1: mirVana PARIS kit, 0.1 ml input; MIRV0.625: mirVana PARIS kit, 0.625 ml input; NUC0.3: NucleoSpin miRNA plasma kit, 0.3 ml; NUC0.9: NucleoSpin miRNA plasma kit, 0.9 ml input)

6 Spikes

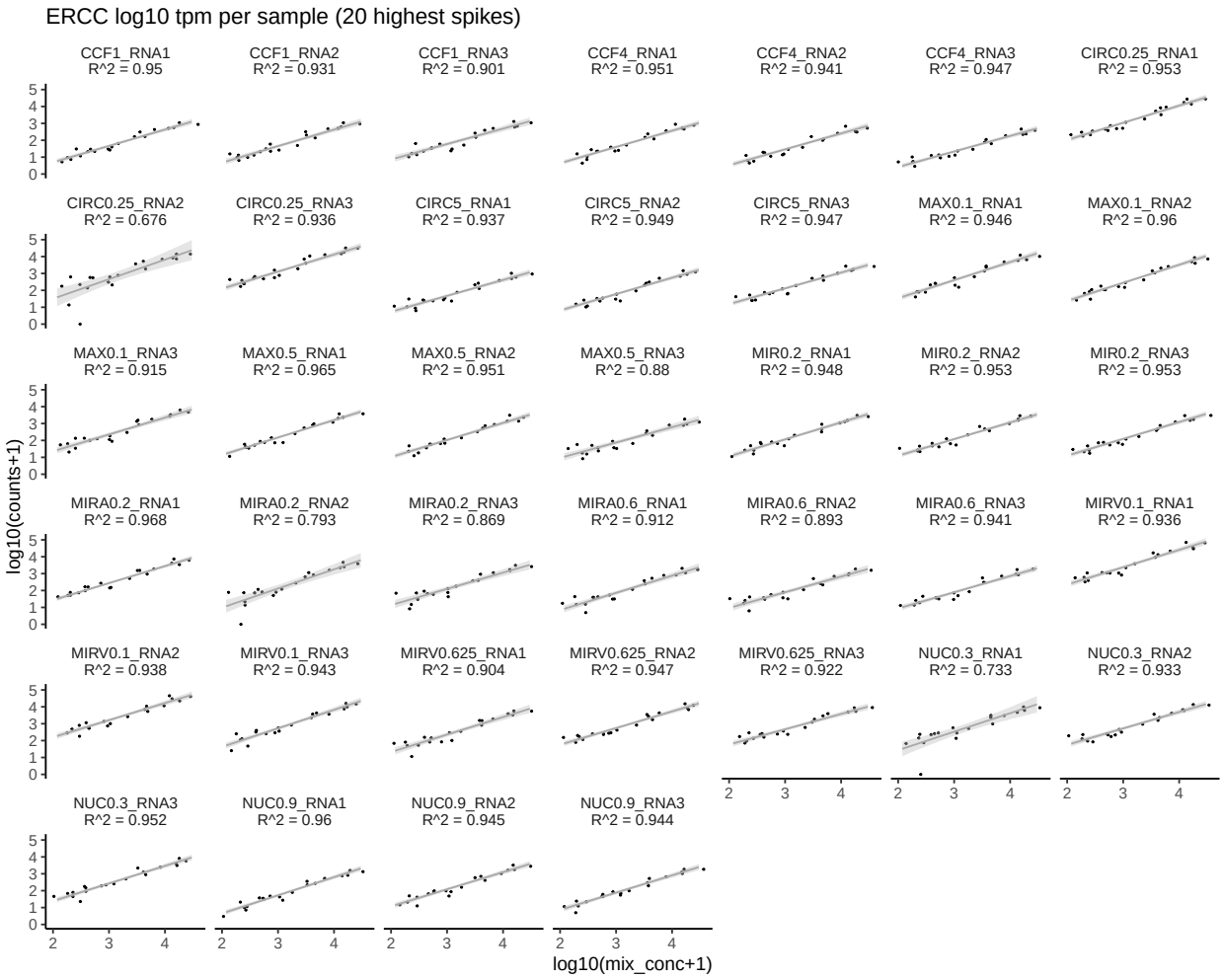
- ERCC spikes were added in the same amount each time to 12 ul of eluate (after RNA purification) -> indicative for RNA concentration in the eluate, and thus for RNA yield (if corrected for elution volume) see RNA concentration and RNA yield
  - More ERCC indicates less endogenous RNA in eluate
- Sequin could show differences in RNA concentration in plasma before RNA purification. However, we always start from the same plasma (from only 1 donorand samee collection tube) so these spikes are less relevant in this part of the project.

6.1 ERCC

- **ERCC spikes were added after RNA purification**

6.1.1 Linear models

- Linear model for each sample based on the expression of the spikes.
  - Linear models based on the 20 highest spikes (according to rowmeans)) because these spikes are picked up in (almost) every sample (for lower spike concentrations, some spikes drop out)
  - However, we could also use the 95% SP cut-off described below to select the exact number of spikes to use for this plot
- Plot shows that there is indeed a good fit
  - **Conclusion: We can get quantitative information from our experiment (however, analysis is only on 20 highest spikes)**



6.1.2 Recovery of spikes

- (mean) percentage of ERCC spikes detected in all triplicates
- remark: multiple ERCC spikes can have the same concentration thus if only 1 spike with certain concentration is not detected in 1 of the 3 replicates, the % can still be $> 66\%$
- x-axis: concentration of spike in mix
- y-axis: to what extent are ERCC spikes picked up in all replicates of one kit?
- the higher the spike concentration, the higher the percentage of replicates in which they are detected should be
- **What can we learn from these curves?**
  - **Interesting to see from which concentration spikes begin to drop out and/or if there is a problem with certain spikes**
  - A lot of spikes are consistently expressed
  - Some kits seem to pick up spikes with lower conc in the mix better than others with higher conc
  - Spikes with a concentration in the mix from 100 attomoles/microL onwards are picked up in almost all the samples

  - However, there seems to be a problem with some spikes that have a concentration of 469 and 938 attomoles/microL in the mix. Perhaps a problem with probe?
  - We could use this to define a cut-off e.g. how many counts required to pick up at least 95% of spikes from a certain concentration onward? This will differ for every sample as ERCC counts will be lower if there is more endogenous RNA, but everything will shift
  - In our experiment, we use a different cut-off based on eliminating 95% of single positives (but you need to have replicates for this)

- **Conclusion: If you do not have any replicates, you may determine a cut-off from ERCC spikes**