W262–W267 *Nucleic Acids Research, 2020, Vol. 48, Web Server issue* *Published online 2 June 2020*
doi: 10.1093/nar/gkaa452

mirnaQC: a webserver for comparative quality control of miRNA-seq data

Ernesto Aparicio-Puerta[1,2,3,4], Cristina Gómez-Martín[1,2], Stavros Giannoukakos[1,2], José María Medina[1,2], Juan Antonio Marchal[3,4,5] and Michael Hackenberg[1,2,3,4,*]

[1]Department of Genetics, Faculty of Science, University of Granada, 18071 Granada, Spain, [2]Bioinformatics Laboratory, Biotechnology Institute, Centro de Investigación Biomédica, PTS, Avda. del Conocimiento s/n, 18100-Granada. Spain, [3]Instituto de Investigación Biosanitaria ibs.GRANADA, University of Granada, 18071 Granada, Spain, [4]Excellence Research Unit "Modelling Nature" (MNat), University of Granada, 18071 Granada, Spain and [5]Department of Human Anatomy and Embryology, Institute of Biopathology and Regenerative Medicine, University of Granada, Granada, Spain

Received March 12, 2020; Revised April 25, 2020; Editorial Decision May 08, 2020; Accepted May 21, 2020

ABSTRACT

Although miRNA-seq is extensively used in many different fields, its quality control is frequently restricted to a PhredScore-based filter. Other important quality related aspects like microRNA yield, the fraction of putative degradation products (such as rRNA fragments) or the percentage of adapter-dimers are hard to assess using absolute thresholds. Here we present mirnaQC, a webserver that relies on 34 quality parameters to assist in miRNA-seq quality control. To improve their interpretability, quality attributes are ranked using a reference distribution obtained from over 36 000 publicly available miRNA-seq datasets. Accepted input formats include FASTQ and SRA accessions. The results page contains several sections that deal with putative technical artefacts related to library preparation, sequencing, contamination or yield. Different visualisations, including PCA and heatmaps, are available to help users identify underlying issues. Finally, we show the usefulness of this approach by analysing two publicly available datasets and discussing the different quality issues that can be detected using mirnaQC.

INTRODUCTION

Different aspects of miRNA-seq such as RNA extraction, storage conditions and sample processing together with the chosen library preparation protocol have a great impact on the obtained sequencing results (1,2). In any bioinformatics analysis of high-throughput sequencing data, quality control (QC) is the key step to reveal the existence of technical artefacts. Neglecting this step can lead to both false discoveries and failure to identify the existing biological signal. The processing of miRNA-seq data is no exception, and QC approaches should focus on measurable sample features that can be linked to quality aspects. Moreover, whenever possible, these quality parameters should hint or point out specific technical artefacts. This approach would offer the user the chance to take appropriate actions like excluding low-quality samples from the analysis or applying statistical models in order to correct for such technical variation when possible like in the case of batch effects (3).

Several sample attributes are generally calculated in sequencing experiments including the total number of sequenced reads, number of adapter-trimmed and filtered reads, percentage of mapped/unmapped reads and PhredScore to measure the quality of the sequencing. Besides these general statistics, many pipelines such as sRNAbench (4), mirTrace (5) or miRge (6) implement measurements that are specifically useful for miRNA-seq analysis like number of unique reads, percentage of miRNA-mapping reads, read length distribution and relative abundance of fragments from other RNA types (mostly tRNA and rRNA). Many of these parameters are clearly relevant for quality. Some indicate good quality samples when they hold high values (number of miRNA reads, total number of reads) while others (percentage of rRNA, adapter dimers) do it when they are low. Some of these features can be directly linked to a particular artefact, like a high percentage of adapter dimers which is normally caused by issues with adapters and/or input RNA concentrations (7). Other measurements, like smeared out read length distributions, can also be attributed to specific problems, in this case RNA degradation. However, most of them can be affected by several different artefacts thus it is frequently not possible to directly reveal specific technical issues when considering each quality feature individually. For instance, the yield in microRNAs can be influenced by any artefacts that impact the total read yield including contamination.

*To whom correspondence should be addressed. Tel: +34 958 249 695; Email: hackenberg@ugr.es

Regardless of the values good quality measurements should take, the context-free interpretation of sample features is generally not straightforward. For example, as an obvious source of unwanted fragments, rRNA presence should be minimized, but it is difficult and arbitrary to establish a specific threshold (5%, 10%, 20%) to discard samples. Therefore, rather than working with predefined or user-provided values that are hard to justify, a more agnostic approach would rely on relative values calculated from a background of comparable experiments (i.e. similar samples) which would in turn simplify the interpretation of the QC outcome.

A vast amount of publicly available data exists that can be exploited for purposes beyond their original goal (8). To generate a reference corpus of experiments that can be used to rank quality features, we downloaded over 36 000 raw sequencing datasets from the Sequence Read Archive (SRA) covering most model species. Samples were first processed using sRNAbench and then 34 quality features were extracted from each sample and subsequently organised into the reference set. Furthermore, sample metadata is used to tailor comparisons to more relevant sets of experiments (i.e. samples from the same species and/or processed with the same library preparation protocol).

In contrast with previously available software (5,6), mirnaQC calculates absolute and relative values for several quality-related features for a set of miRNA-seq samples. Input data can be uploaded as FASTQ files or provided as SRA run accessions that will subsequently be ranked making use of the reference corpus mentioned above. An apparent advantage of this approach is that fixed thresholds are no longer needed and decisions can be made based on background statistics. Users can explore mirnaQC results by means of interactive plots and tables that hold both absolute and relative values of the 34 quality attributes. The output report is structured into several categories trying to relate the quality attributes to the different possible technical artefacts. This approach can help to identify low quality samples or reveal issues in the sample processing which is extremely important for protocol optimisation.

mirnaQC SAMPLE FEATURES AND QUALITY MEASURES

The success of a small RNA sequencing run depends on many different factors including RNA quality, quantity and purity, an optimized library processing protocol and the sequencing itself. However, it is not always easy or even possible to directly relate features extracted from sequencing data to any technical artefacts. mirnaQC calculates and ranks several quality parameters conceived to hint problems in the different aspects involved in the preparation of miRNA sequencing libraries. Below we describe the different sections and, wherever possible, the putative artefacts or quality issues that can be derived from them.

Sequencing yield

This section focuses on the amount of reads and the fraction that can be assigned to known miRNAs. Generally, parameters in this category (percentage of valid reads, detected microRNAs) indicate high quality when they hold high values. Low numbers (especially for the percentage of valid input reads) can be related to problems in RNA processing or low input material. Some sources like exosomes extracted from bodily fluids however, are known to hold low levels of miRNA, thus high numbers should not be expected for all sample types even for high quality libraries.

Library quality

In this category we list the number of reads that are filtered out due to minimum length (15nt), the percentage of ribosomal RNA and the percentage of short reads (15–17 nt). Their presence may be attributed to degradation products from longer RNA molecules as no small RNAs are known in this length range.

High percentages of adapter-dimers (0, 1 or 2 nt fragments after trimming) normally indicate issues with the ratio of adapter to input RNA concentration. In practice, it is very difficult to completely avoid adapter-dimers, especially in low input samples such as blood. Nevertheless the percentile may still be useful as it might show potential for improvement.

Ultra-short reads are defined as fragments with lengths between 3nt and 14nt (both inclusive).

Library complexity

In general it is also interesting to assess the complexity of the sample since low complexity libraries provide very little information, even for otherwise high-quality datasets. Several measurements are provided to grasp the complexity at two levels that should be interpreted together:

- Sequencing library complexity: This is calculated as the ratio of the total number of reads to unique reads. Lower values suggest higher RNA diversity but it can also be caused by degradation.
- miRNA complexity: Frequently most microRNA reads correspond to few miRNA genes preventing lowly expressed miRNAs from being detected. Several measures are given to estimate complexity at this level: (i) percentage of miRNA expression assigned to the first, the first 5 and first 20 most expressed miRNAs, (ii) the number of miRNAs required to reach 50%, 75% and 95% of the total miRNA expression.

Putative contamination

The percentage of reads that could not be mapped to the species' genome is calculated. Contamination is subsequently estimated by mapping against a collection of bacterial and viral genomes.

Read length distribution

A narrow peak around 22 nucleotides in the read length distribution indicates good quality samples whereas degraded or poor RNA quality manifests in a broader distribution. Furthermore, it is clear that the 22nt peak should be present for miRNA assigned reads and RNA quality issues might exist if samples deviate from this.

We summarise the miRNA read length distribution in several ways: mean length, mode of the distribution, the fraction of reads with lengths 21, 22 or 23 nt, the standard deviation and the skewness of the distribution.

RNA composition

The relative abundance of other RNA molecules is automatically profiled using the sRNAtoolbox database (9). Most of these longer RNA species (rRNA,mRNA, lincRNA) are not known to be processed into smaller molecules that can be picked up by miRNA-seq. Their presence is a symptom of RNA degradation since smaller fragments are randomly generated and then sequenced. Among these, rRNA is typically used because it's the most abundant one.

Sequencing quality

Sequencing quality is calculated by means of FastQC (10). We determine the mean values of the different percentiles provided by the program over all positions of the read.

GENERATION OF THE miRNA-seq REFERENCE CORPUS (BACKGROUND KNOWLEDGE)

The vital part of the presented quality control tool mirnaQC is the comparison corpus of miRNA-seq data that is used to rank user's samples. Using OmicIDXR API we obtained a list of >3000 SRA studies that were annotated as 'miRNA-Seq' or 'ncRNA-Seq' (several 'RNA-Seq' were also included after checking they were in fact 'miRNA-Seq' datasets).

For each study we performed the following steps:

- Read the meta-data for a study generating one entry per experiment (SRX level)
- Download all SRR files that correspond to this SRX by means of fastq-dump (fastq.gz)
- Detect the library preparation protocol
- Analyse the small RNA sequencing data with sRNAbench using all available annotations from sRNAtoolboxDB
- Upload sRNAbench results to a MySQL database

In total we analysed 36 338 samples from 30 different species. We distinguish 8 different protocols: Illumina, Illumina_2 (3′ adapter sequence), Next England Biolabs (NEBnext), Qiagen_UMI, NextFlex, adapter trimmed, SoliD and all others (custom). Over 500 billion sequencing reads were analysed.

mirnaQC WORKFLOW AND IMPLEMENTATION

An overview of the mirnaQC workflow is displayed on Figure 1. The only required input is sequencing data in FASTQ format (or SRR accessions) although sample species and library protocol information is recommended if known. If the protocol or species are not provided by the user an automatic detection algorithm, trained with a set of manually curated samples from liqDB (11), will find the right input parameters. Condition or group information can also be provided (optional). All files belonging to a given group should be compressed into a single *.zip, tar.gz* or *.7z* file and then separately uploaded. The file names will be used as group labels, and this information will appear in some of the plots.

Input data is subsequently processed by sRNAbench in two steps: First reads are simultaneously mapped to the species genome and a collection of virus and bacterial genomes from sRNAtoolboxDB (9) allowing one mismatch. Preference is given to the reference genome in case of multiple mapping reads. In the second step reads are mapped to microRNA reference libraries (12,13), RNAcentral (14) and Ensembl annotations (15) for ncRNA and mRNA. Note that although samples are mapped to both reference libraries, miRBase and MirGeneDB, currently the miRNA related figures are extracted from the miRBase mappings.

Out of both sRNAbench output folders we extract a total of 34 quality attributes that are next compared to 5 different reference sets: (i) samples from the same kingdom (animals and plants), (ii) samples from the same species, (iii) samples from the same kingdom and protocol, (iv) samples from the same species and protocol and (v) low-input samples (defined as those obtained from bodily fluids). Each comparison can be browsed separately on the output page.

The processing pipeline is a java programme that includes sRNAbench, Bowtie (16) and a MySQL client to query the reference corpus. The web interface was developed using Python and Django and runs on Apache. The results report includes the six sections described in the *mirnaQC sample features and quality measures* with tables and styles from multiQC (17) and Plotly visualisations. Both absolute values and percentiles are displayed and highlighted using a quartile colour code (see Figure 2C).

WORKING EXAMPLE

To show the usefulness of this tool, we analysed two publically available studies. Basic statistics from the first dataset, one of the earliest large studies designed to detect cervical cancer (18), can be seen on Figure 2A. To help users identify potential issues, the quality parameters and their percentiles are displayed using a quartile-based colour code (from better to worse values: green, yellow, orange and red). Using this guide, several problems can be identified: with few exceptions, most parameters rank on the third (orange, Q3) and fourth quartiles (red, Q4). More specifically, miRNA 'peak' values show that a rather low percentage of microRNA reads have lengths between 21 and 23nt in the majority of samples. This means that although those reads can be assigned to miRNA reference sequences, they do not correspond to the canonical miRNA lengths. This hints an RNA processing issue that might still be tolerated if all samples are similarly affected, which can indicate either systematic artefacts or biological reasons.

It may also happen that not all samples are equally affected by a quality issue, which can be more problematic if two or more conditions are to be compared. mirnaQC allows users to assign samples to conditions in order to explore this possibility. Figure 2B shows a PCA plot of the expression values of the 50 most expressed miRNAs.

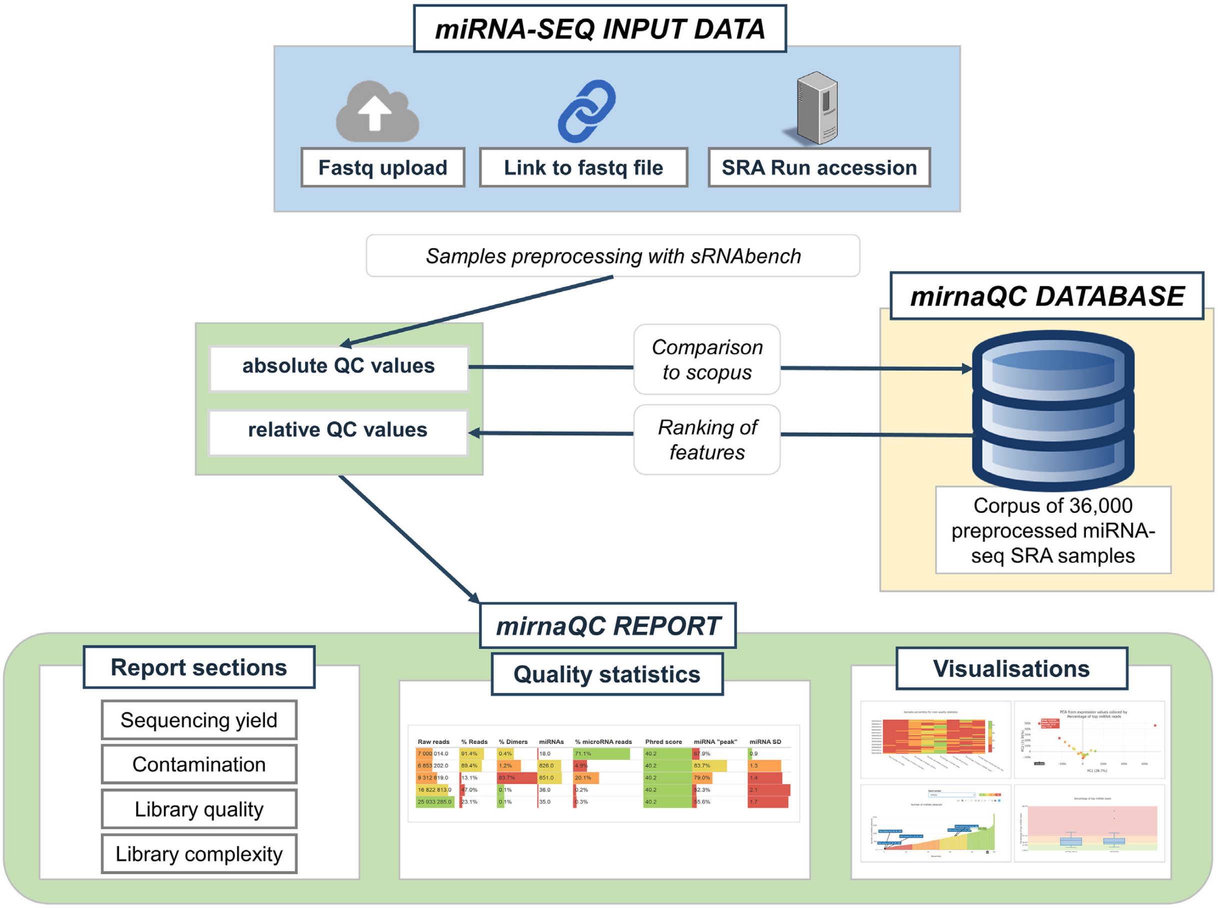

Figure 1. A schema of the front end and workflow of mirnaQC. Some features of the quality report are depicted at the bottom.

Users can decide which quality attribute should be used to colour the markers, in this case we used '% top miRNA' (the percentage of reads assigned to the most abundant microRNA). This graph shows that the two outlier samples are much less complex than the rest. Furthermore, because conditions are marked with different symbols (control-circles and carcinoma-squares), we know that these two samples belong to the same group. Keeping such samples in the analysis is not recommended since they will certainly bias the results. Figure 2E displays the distribution of this feature for both groups by means of boxplots. Here we can see that these two samples are outliers but otherwise both conditions show reasonably similar distributions for this parameter.

Users can also explore potential sources of contamination from reads mapped to viral and bacterial genomes. Figure 2D shows that all samples suffer from rather high percentages of contamination reads. All samples have more bacterial/viral reads than 75% of all animal samples in the reference set. This range of values indicates serious contamination with the possible exception of cervical cancer samples, where this might be caused by sample extraction or even have a role in the disease (19).

Finally, Figure 2E shows the library complexity of different tissues from *Takifugu rubripes* (20). While the top expressed microRNA in intestine picks up 47.1% of all reads (percentile 88.4), in ovary this figure drops to 8.3% which corresponds to percentile 2.6. In ovary, eight microRNA sum up over 50% of all miRNA expression while in intestine it takes only two to reach the same percentage which indicates a higher complexity of miRNA expression in germ cells. Furthermore, ovary and testis exhibit much lower percentages of miRNAs. This might be related to their larger repertoire of small RNAs in germ cells (21) which automatically would lead to a lower relative fraction of microRNAs in those samples.

CONCLUSION

We present a user-friendly web server for the comparative quality control of miRNA-seq data that can be useful in several scenarios: to identify low-quality samples that should be excluded from downstream analysis; to reveal systematic errors in order to improve the library preparation process, something especially relevant for pilot studies; and finally, to provide external quality validation for datasets so it can be used as a standard proof of quality.

mirnaQC provides several output tables and visualisations for a total of 34 quality attributes which allow users to rank their results against a large corpus of comparable

Sample Name	Raw reads	% Reads	% Dimers	% short reads	miRNAs	% microRNA reads	miRNA "peak"
SRR040008	1939877	47.9%	3.1%	35.9%	969	11.1%	52.8%
SRR040009	1020150	57.6%	1.8%	25.0%	372	23.3%	5.1%
SRR040012	1682141	68.1%	1.5%	11.0%	816	71.8%	78.7%
SRR040013	88476	32.1%	14.4%	25.3%	273	12.2%	51.2%
SRR040021	139633	47.5%	6.9%	43.3%	258	6.7%	25.7%
SRR040032	923339	57.1%	22.8%	16.6%	672	55.6%	57.7%
SRR040033	93985	49.9%	11.2%	22.6%	404	41.7%	80.8%
SRR040040	1240271	38.4%	9.5%	29.8%	805	24.1%	47.7%
SRR040041	267758	34.8%	3.5%	29.2%	397	13.0%	44.1%
SRR040042	1094061	39.4%	4.4%	13.8%	469	13.8%	80.5%
SRR040043	1712707	37.3%	0.8%	12.3%	459	12.4%	81.5%

Sample Name	% ref.genome	% unmapped	% bacteria	% virus
SRR040008	16.5%	74.5%	7.8%	1.2%
SRR040009	58.6%	19.1%	20.8%	1.4%
SRR040012	81.0%	10.2%	7.3%	1.5%
SRR040013	56.7%	4.0%	34.2%	5.1%
SRR040021	50.2%	1.9%	39.6%	8.4%
SRR040032	73.3%	7.4%	16.1%	3.1%
SRR040033	75.1%	3.8%	18.1%	3.0%
SRR040040	28.7%	51.0%	17.2%	3.1%
SRR040041	48.9%	4.9%	39.8%	6.4%
SRR040042	11.7%	82.9%	4.5%	0.8%
SRR040043	12.0%	82.1%	5.1%	0.8%

Sample Name	Reads per unique	% top miRNA	% top5 miRNA	% top20 miRNA	#miRNA p50	#miRNA p75	#miRNA p95	miRNAs per 1000 reads
SRR1777549_Intestine	13.7	47.1%	67.8%	88.3%	2	8	36	0.1
SRR1777552_Brain	13.4	17.6%	46.8%	81.2%	6	16	50	0.1
SRR1777555_Heart	14.9	15.8%	41.9%	79.7%	8	18	41	0.2
SRR1777556_Liver	15.7	23.9%	56.9%	82.4%	4	14	42	0.3
SRR1777560_Ovary	35.8	8.3%	37.5%	79.2%	8	17	43	0.6
SRR1777562_Testis	45.5	14.5%	43.6%	82.8%	7	16	34	4.1

Figure 2. Different examples of mirnaQC sections and visualisations. (**A–D**) display different quality aspects of a cancer study. (**E**) shows the output of different sample complexity measures for different tissue types and two columns from Basic Statistics section at the right.

samples. In this way, no absolute thresholds need to be applied and the user can evaluate their sequencing data based on percentiles. Future developments include new types of analysis and improved visualisations intended to detect confounding variables related to quality issues that can affect downstream steps. Additionally, a dockerized version of the tool will be made available so the pipeline can be run locally or in computing clusters.

DATA AVAILABILITY

https://arn.ugr.es/mirnaqc/.

ACKNOWLEDGEMENTS

The authors acknowledge the usage of the computational infrastructure of the Computational Epigenomics Lab of the University of Granada.

FUNDING

This work was supported by European Union [765492]; Spanish Government [AGL2017-88702-C2-2-R] to M.H.; Consejería de Economía, Conocimiento, Empresas y Universidad de la Junta de Andalucía and European Regional Development Funds (ERDF) [SOMM17-6109, UCE-PP2017-3] to J.A.M. and M.H.; Instituto de Salud Carlos III, ERDF funds [PIE16/00045] to J.A.M.; Chair 'Doctors Galera-Requena in cancer stem cell research' (to J.A.M.); Instituto de Salud Carlos III [IFI16/00041] to E.A. Funding for open access charges: Excellence Research Unit "Modelling Nature" (MNat) [SOMM17-6109].
Conflict of interest statement. None declared.

REFERENCES

1. Sorefan,K., Pais,H., Hall,A.E., Kozomara,A., Griffiths-Jones,S., Moulton,V. and Dalmay,T. (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, **3**, 4.
2. Tiberio,P., Callari,M., Angeloni,V., Daidone,M.G. and Appierto,V. (2015) Challenges in using circulating miRNAs as cancer biomarkers. *Biomed Res. Int.*, **2015**, 731479.
3. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.
4. Aparicio-Puerta,E., Lebrón,R., Rueda,A., Gómez-Martín,C., Giannoukakos,S., Jaspez,D., Medina,J.M., Zubkovic,A., Jurak,I., Fromm,B. *et al.* (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res.*, **47**, W530–W535.
5. Kang,W., Eldfjell,Y., Fromm,B., Estivill,X., Biryukova,I. and Friedländer,M.R. (2018) miRTrace reveals the organismal origins of microRNA sequencing data. *Genome Biol.*, **19**, 213.
6. Lu,Y., Baras,A.S. and Halushka,M.K. (2018) miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics*, **19**, 275.
7. Shore,S., Henderson,J.M., Lebedev,A., Salcedo,M.P., Zon,G., McCaffrey,A.P., Paul,N. and Hogrefe,R.I. (2016) Small RNA library preparation method for Next-Generation sequencing using chemical modifications to prevent adapter dimer formation. *PLoS One*, **11**, e0167009.
8. Plocik,A.M. and Graveley,B.R. (2013) New insights from existing sequence data: generating breakthroughs without a pipette. *Mol. Cell*, **49**, 605–617.
9. Rueda,A., Barturen,G., Lebrón,R., Gómez-Martín,C., Alganza,Á., Oliver,J.L. and Hackenberg,M. (2015) sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res.*, **43**, W467–W473.
10. Andrews,S., Krueger,F., Segonds-Pichon,A., Biggins,L., Krueger,C. and Wingett,S. (2012) *FastQC: a quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
11. Aparicio-Puerta,E., Jáspez,D., Lebrón,R., Koppers-Lalic,D., Marchal,J.A. and Hackenberg,M. (2019) liqDB: a small-RNAseq knowledge discovery database for liquid biopsy studies. *Nucleic Acids Res.*, **47**, D113–D120.
12. Fromm,B., Domanska,D., Høye,E., Ovchinnikov,V., Kang,W., Aparicio-Puerta,E., Johansen,M., Flatmark,K., Mathelier,A., Hovig,E. *et al.* (2020) MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.*, **48**, D132–D141.
13. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
14. The RNAcentral Consortium, Petrov,A.I., Kay,S.J.E., Kalvari,I., Howe,K.L., Gray,K.A., Bruford,E.A., Kersey,P.J., Cochrane,G., Finn,R.D. *et al.* (2017) RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic Acids Res.*, **45**, D128–D134.
15. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G. *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, **46**, D754–D761.
16. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
17. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
18. Witten,D., Tibshirani,R., Gu,S., Fire,A. and Lui,W.-O. (2010) Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol.*, **8**, 58.
19. Łaniewski,P., Barnes,D., Goulder,A., Cui,H., Roe,D.J., Chase,D.M. and Herbst-Kralovetz,M.M. (2018) Linking cervicovaginal immune signatures, HPV and microbiota composition in cervical carcinogenesis in non-Hispanic and Hispanic women. *Sci. Rep.*, **8**, 7593.
20. Wongwarangkana,C., Fujimori,K.E., Akiba,M., Kinoshita,S., Teruya,M., Nezuo,M., Masatoshi,T., Watabe,S. and Asakawa,S. (2015) Deep sequencing, profiling and detailed annotation of microRNAs in Takifugu rubripes. *BMC Genomics*, **16**, 457.
21. Saxe,J.P. and Lin,H. (2011) Small noncoding RNAs in the germline. *Cold Spring Harb. Perspect. Biol.*, **3**, a002717.