Supplementary Material to Wen *et al.*, "AQUARIUM: accurate quantification of circular isoforms using model-based strategy"

Supplementary Data

Supplementary Methods.

Supplementary Table S1. The public RNA-seq data used in this study

Supplementary Table S2. The source code and version of the computational tools used in this study

Supplementary Table S3. CircRNA BSJ sites that can be amplified by RT-qPCR using the primers

Supplementary Table S4. The BSJ location, reconstruction type, exon number, circRNA length, and the fraction of exons to the length of circRNAs for all reconstructed circRNAs in samples with RT-qPCR values

Supplementary Table S5. The estimated expression values by different computational tools and the experimental expression value by RT-qPCR for all circRNAs in 10 samples

Supplementary Table S6. Time and memory usage of the detection and quantification module in AQUARIUM implementation on three simulated RNA-seq datasets, which was performed on a DELL-T7600 workstation with 2.4G Intel Xeon CPUs and 128G memory

Supplementary Table S7. The experimentally determined expression values of circular isoforms by RT-qPCR and the estimated circRNA expressions by *CIRI-full* and *AQUARIUM* in the biological RNA-seq dataset *Hela-R2*

Supplementary Figure S1. Comparison of the quantification performance between the *AQUARIUM*, *Sailfish-cir*, *CIRIquant*, *CIRI2*, *CIRI-full*, and *CLEAR*. *X*-axis and *Y*-axis represent the circRNA expression measured by RT-qPCR and the estimated circRNA expression by each tool from RNA-seq data (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM* and *Sailfish-cir*, CPM for *CIRIquant*, FPB_{circ} for *CLEAR*, and number of BSJ reads for *CIRI2* and *CIRI-full*. The *r* and *P* were computed by *Pearson* correlation test between *X*- and *Y*-axes.

Supplementary Figure S2. Comparison of the quantification performance with varied sequencing depth. The quantification results generated from the simulated dataset *Hela-S1* by the *AQUARIUM*, *CIRIquant*, *CIRI2*, and *CLEAR* algorithms were included. *X*-axis and *Y*-axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM*, CPM for *CIRIquant*, and number of BSJ reads for *CIRI2* and *CLEAR*. The *n* demonstrates the total number of the identified circRNAs from simulated datasets by each tool. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.

Supplementary Figure S3. Comparison of the quantification performance between

AQUARIUM and *Sailfish-cir* with varied sequencing depth and reconstructed sequence concordance. CircRNAs that had C_{rs} values less than 0.2 were grouped as the nearly full circRNAs, while circRNAs with C_{rs} values larger than 0.2 were grouped as the partial circRNAs. The simulated dataset *Hela-S1* was used in this comparison. *X* -axis and *Y* -axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM* and *Sailfish-cir*. The *n* demonstrates the total number of the identified circRNAs from simulated datasets by each tool. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.

Supplementary Figure S4. Comparison of the quantification performance with varied reconstructed sequence concordance. CircRNAs that had C_{rs} values less than 0.2 were grouped as the nearly full circRNAs, while circRNAs with C_{rs} values larger than 0.2 were grouped as the partial circRNAs. The quantification results generated from one RNA-seq data (60M, PE 150) in the simulated dataset *Hela-S2* by the *AQUARIUM*, *CIRIquant*, *CIRI2*, and *CLEAR* algorithms were included. *X* -axis and *Y* -axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM*, CPM for *CIRIquant*, and number of BSJ reads for *CIRI2* and *CLEAR*. The *n* demonstrates the total number of the identified circRNAs from simulated datasets by each tool. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.

Supplementary Figure S5. Performance of *AQUARIUM* in the simulated dataset *Hela-S2* with different read length. *X*-axis and *Y*-axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The *n* demonstrates the total number of the identified circRNAs. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.

Supplementary Figure S6. Example view of two circRNAs with over-estimated expression values by count-based tools, including *CIRIquant*, *CIRI2*, *CIRI-full*, *CLEAR*, and *KNIFE*, since RNA sequencing reads were non-uniformly distributed along the circRNA transcripts. **(A)** Two circRNAs (green and red dots) were selected to show the different estimation accuracy by the *AQUARIUM*, *CIRIquant*, *CIRI2*, *CIRI-full*, *CLEAR*, and *KNIFE*. *X*-axis and *Y*-axis represent the circRNA expression measured by RT-qPCR and the estimated circRNA expression by each tool from RNA-seq data (log_2-transformed), respectively. The *r* and *P* were computed by *Pearson* correlation test between *X*- and *Y*-axes. **(B)** RNA sequencing read coverage, circRNA structure, and BSJ read coverage of two reconstructed circRNAs. The upper panel was related to the green dot in **(A)**, and the bottom panel was corresponding to the red dot in **(A)**.

Supplementary Figure S7. Comparison of circRNA isoform quantification between

AQUARIUM and *CIRI-full*. Each dot represents one circRNA isoform. TPM and BSJ number were used to represent the expression (log_2-transformed) quantified by *AQUARIUM* and *CIRI-full*, respectively. The *r* and *P* were computed by Pears*on* correlation test between the *X*- and *Y*-axes. **(A)** The expression quantified by *AQUARIUM* shows higher concordance with the RT-qPCR readouts of 12 circRNA isoforms compared with *CIRI-full*. **(B)** *AQUARIUM* outperforms *CIRI-full* on 109 circRNA isoforms identified from the simulated dataset *Hela-S1* with 30M PE150 reads.

Supplementary Methods

1. Biological rRNA-depleted RNA-seq data

To evaluate the performance of *AQUARIUM* in circRNA quantification, we downloaded three biological rRNA-depleted RNA-seq datasets. First, we downloaded 11 RNA-seq data from NCBI GEO (Barrett, et al., 2013) under accession number GSE64283. This dataset consisted of the RNA-seq data of a panel of 11 human fetal tissues at different developmental time points (*Fetal-R1*, **Supplementary Table S1**) (Szabo, et al., 2015). Second, we downloaded 4 rRNA-depleted RNA-seq data of human *Hela* cell line from Genome Sequence Archive (GSA) (Wang, et al., 2017) under accession number CRA001838 (*Hela-R2*, **Supplementary Table S1**) (Zhang, et al., 2020). We chose these two datasets since the expression values of some circRNAs at the BSJ level had been experimentally validated by RT-qPCR. Third, we downloaded SRR7309440 run of SRA project PRJNA475651 from NCBI sequence read archive (SRA) (Leinonen, et al., 2011) as well. This is a rRNA-depleted RNA-seq data of human *Hela* cell line with the circRNA expression at the isoform level, which was validated by RT-qPCR (*Hela-R3*, **Supplementary Table S1**) (Zheng, et al., 2019).

2. Simulated rRNA-depleted RNA-seq data

In order to evaluate the performance of *AQUARIUM* in quantifying circRNA expression, we also generated two simulated rRNA-depleted RNA-seq datasets using *Polyester* (Frazee, et al., 2015). In *Polyester* implementation, the sequences and expression abundances of both linear and circular transcripts were set as the values derived from the biological RNA-seq dataset *Hela-R3*. Specifically, we ran *AQUARIUM* to quantify both linear and circular transcripts in *Hela-R3* dataset using the default parameters. In total, the expression values of 2,435 circular and 37,566 linear transcripts were measured in this dataset. We then set *Polyester's* parameters of *fold_changes*, *readspertx* and *coverage* based on the expression values of each expressed transcript, the expected read length and the expected sequencing depth. To simulate circular transcripts, we repeated each circular transcript 10 times to construct BSJ sites. *Simulate_experiment_countmat()* function in *Polyester* was used to generate the RNA-seq reads for each transcript at the specified number. Finally, two simulated RNA-seq datasets, *Hela-S1* and *Hela-S2*, with different sequencing depth and read length were generated. Dataset *Hela-S1* had three RNA-seq data with paired-end reads at the same read length of 150bp (PE150), but varied in sequencing depth with the total number of reads at 15M, 30M and 60M, respectively. Dataset *Hela-S2* also had three RNA-seq data, but they had the same sequencing depth at 60M reads

with varied read length at PE100, PE150 and PE250.

3. Comparison of circRNA quantification at the BSJ level

We compared the performance of *AQUARIUM* with several existing tools, including *CIRIquant*, *CIRI2*, *CIRI-full* and *CLEAR*. *CIRIquant*, *CIRI2*, *CIRI-full*, and *CLEAR* are tools that quantify circRNA expressions at the BSJ level, which is different from the circRNA expression at the isoform level that *AQUARIUM* measures. Therefore, we performed the performance comparison of circRNA quantification between *AQUARIUM*, *CIRIquant*, *CIRI2*, *CIRI-full* and *CLEAR* at the BSJ level. We ran all these tools (see **Supplementary Table S2** for software sources) with default parameters on *Fetal-R1* and *Hela-R2* datasets and two simulated datasets (*Hela-S1* and *Hela-S2*), and computed the expression values of all circular transcripts. We used human reference genome (*GRCh38*) and *Ensembl* gene annotation (version 94) as the genomic references in these analyses.

For *Fetal-R1* dataset, we retained those circRNAs with expression data determined by RT-qPCR in the original publication according to the primers (**Supplementary Table S3** and **Supplementary Table S4**). Next, we parsed the estimated expression values of these circRNA BSJ sites from the output of each quantification tool (**Supplementary Table S5**). For *AQUARIUM*, we parsed the TPM (transcripts per million) as the expression value. The CPM (counts per million), the FPB_{circ} (fragments per billion mapped bases), and the number of BSJ reads were used for *CIRIquant*, *CLEAR*, and *CIRI2*, respectively. For *CIRI-full* and *KNIFE*, we also used the number of reads that support the BSJ as the expression value of that circRNA. Finally, the estimated expression values were compared against the CT values experimentally determined by RT-qPCR to evaluate the estimation performance of each tool. Since the comparison was performed at the BSJ level, the expression value of circRNA in *AQUARIUM* and *CIRI-full* was set as the sum of the expression values of all the circular isoforms with the same BSJ site. To compare the estimation performance among these tools, we performed the same analysis on *Hela-R2* dataset as well.

For those two simulated datasets *Hela-S1* and *Hela-S2*, we used the expression set in the *Polyester* simulation as the real expression values of circRNAs, and compared them against the expression quantified by each algorithm (see **Supplementary Table S6**) for time and memory usage in *AQUARIUM* implementation). As we performed in *Fetal-R1* dataset, we used TPM value as the measurement of circRNA expression for *AQUARIUM*. The CPM value was used for *CIRIquant*, while the number of BSJ reads were for *CLEAR* and *CIRI2*.

To explore whether the reconstructed circRNA sequences may affect *AQUARIUM*'s estimation accuracy, we calculated the concordance (C_{rs}) between the

reconstructed sequences (A) and the input transcripts in *Polyester* simulation (B) for each circRNA using edit distance (Navarro, 2001) and pairwise alignment.

C_{rs}(A, B) = 1 - *edit_distance*(A, B) / *length*(*pairwise_alignment*(A, B))

Using a cutoff of C_{rs} value at 0.2, we categorized circRNAs into two groups, and compared the estimation performance of all tools separately for circRNA group.

4. Comparison of circRNA quantification at isoform level

Different from the above computational tools that quantify circRNA expressions at the BSJ level, *CIRI-full* can estimate the expression of circular transcripts at the isoform level (Zheng, et al., 2019). Therefore, we also evaluated the performance of *AQUARIUM* in quantifying circRNA isoforms and compared it with *CIRI-full*. We ran both *CIRI-full* and AQUARIUM using the default parameters on two RNA-seq data (one biological RNA-seq data in *Hela-R3* and one simulated RNA-seq data with 30M PE150 reads in *Hela-S1*), and measured expressions of all the circular isoforms in these two datasets.

For the biological RNA-seq data in the *Hela-R3* dataset, we filtered circular isoforms with expression values are experimentally measured by RT-qPCR according to the BSJ site and isoform length (**Supplementary Table S7**). Next, we parsed the expression values of these circular isoforms estimated by *AQUARIUM* and *CIRI-full*. We used the TPM and the number of BSJ reads as the expression measurement for *AQUARIUM* and *CIRI-full*, respectively. Finally, we compared these values against the RT-qPCR values to evaluate their accuracy.

For the simulated RNA-seq data, we identified 109 alternatively spliced circular isoforms. Similarly, we parsed the expression values of these isoforms estimated by *AQUARIUM* and *CIRI-full*, and compared them against the expression values set in the *Polyester* simulation to evaluate the quantification accuracy.

5. CircRNA visualization

We used *CIRI-vis* (Zheng and Zhao, 2020) to visualize the read coverage along the circular transcript, read coverage around BSJ location, and its internal structure of each randomly selected reconstructed circRNA.

6. Statistical analysis

The correlation between the estimated expression values and the simulated or RT-qPCR determined expression values was used to assess the estimation accuracy of circRNA expression. The expression values were log_2-transformed for estimated or simulated expressions. The *cor.test()* function in the *R* platform (R Core Team, 2015) was used to calculate the *Pearson* correlation.

References

Barrett, T.*, et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* 2013;41(Database issue):D991-995.

Frazee, A.C.*, et al.* Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* 2015;31(17):2778-2784.

Leinonen, R.*, et al.* The Sequence Read Archive. *Nucleic Acids Research* 2011;39(suppl_1):D19-D21.

Navarro, G. A guided tour to approximate string matching. *Acm Comput Surv Csur* 2001;33(1):31-88.

R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.

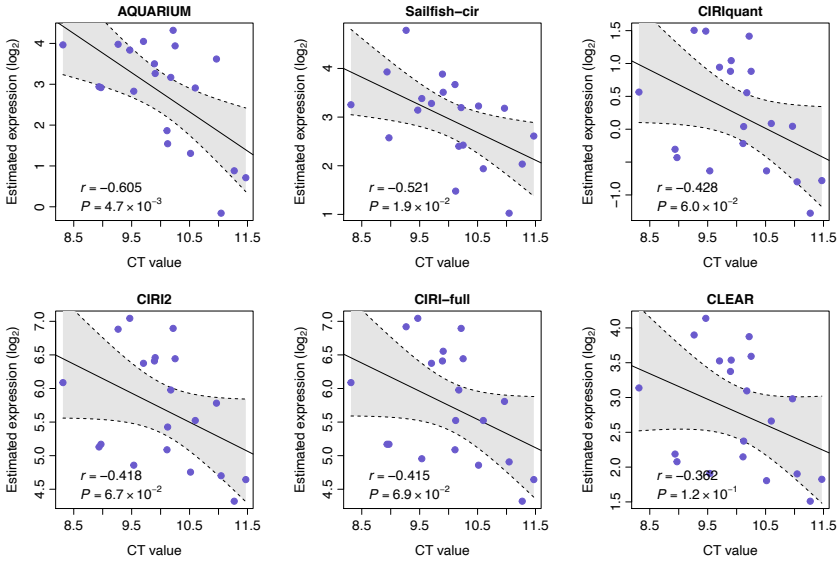
Szabo, L.*, et al.* Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome biology* 2015;16:126.

Wang, Y.*, et al.* GSA: Genome Sequence Archive. *Genom Proteom Bioinform* 2017;15(1):14-18.

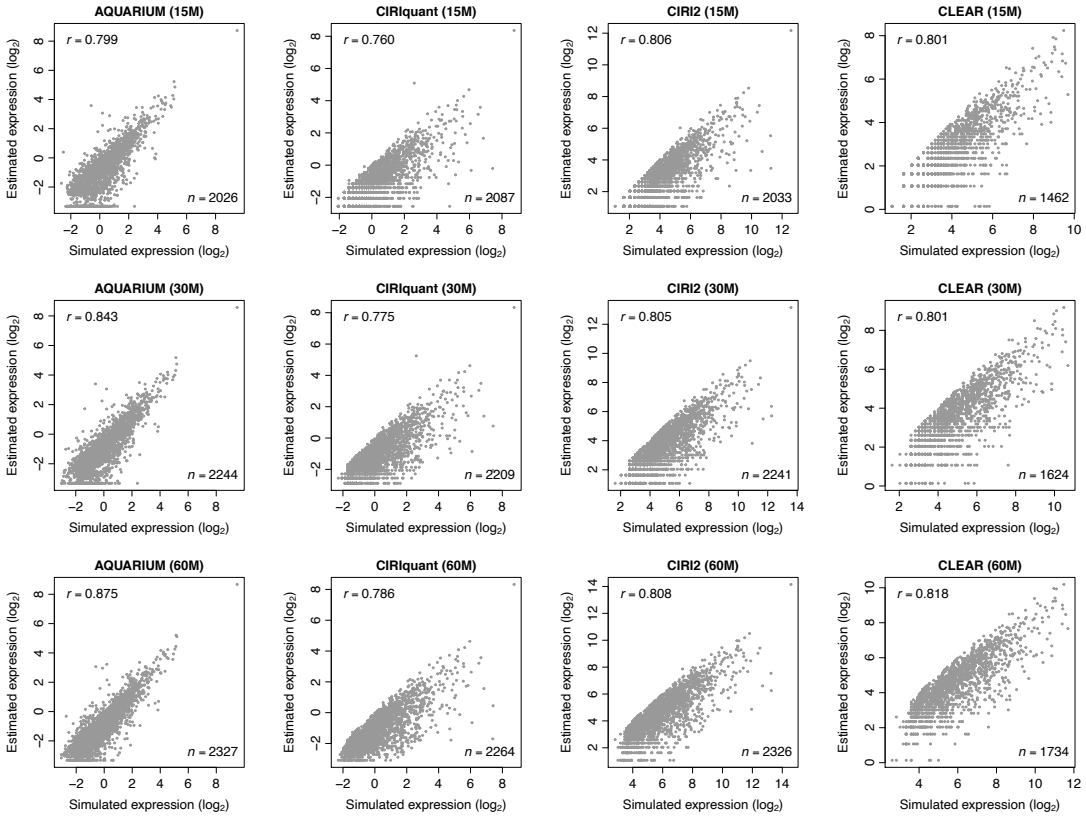
Zhang, J.*, et al.* Accurate quantification of circular RNAs identifies extensive circular isoform switching events. *Nature Communications* 2020;11(1):90.

Zheng, Y.*, et al.* Reconstruction of full-length circular RNAs enables isoform-level quantification. *Genome Med* 2019;11(1):2.

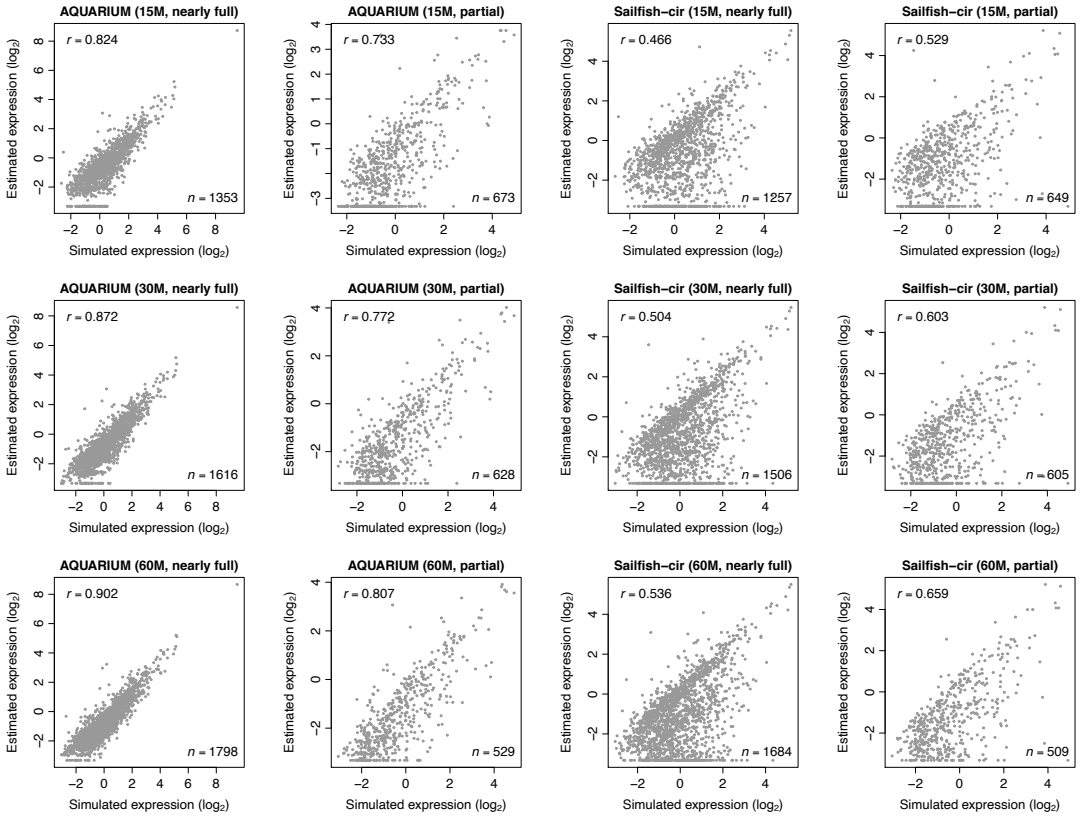
Zheng, Y. and Zhao, F. Visualization of circular RNAs and their internal splicing events from transcriptomic data. *Bioinformatics* 2020;36(9):2934-2935.



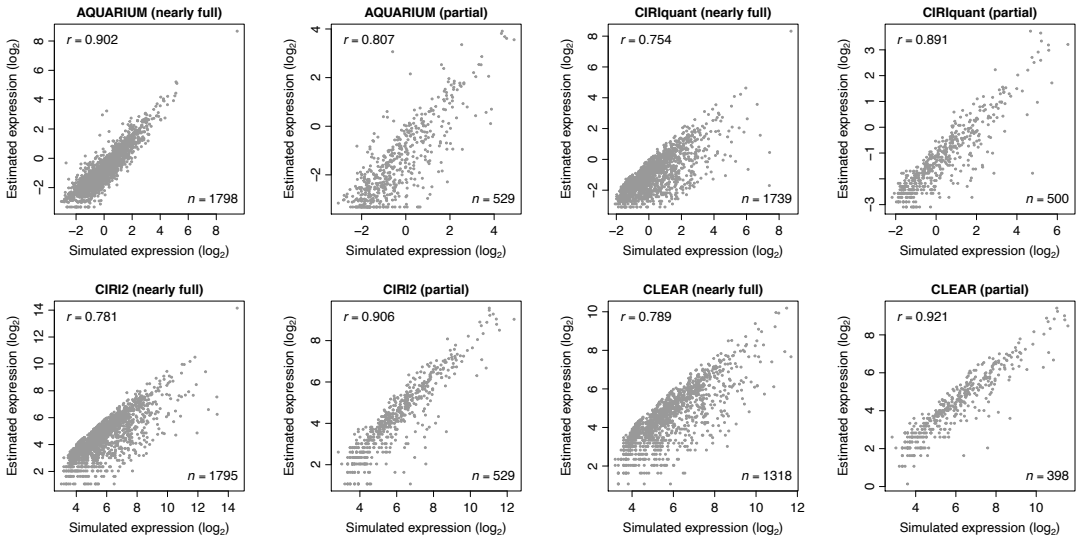
Supplementary Figure S1. Comparison of the quantification performance between the *AQUARIUM*, *Sailfish-cir*, *CIRIquant*, *CIRI2*, *CIRI-full*, and *CLEAR*. *X*-axis and *Y*-axis represent the circRNA expression measured by RT-qPCR and the estimated circRNA expression by each tool from RNA-seq data (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM* and *Sailfish-cir*, CPM for *CIRIquant*, FPB_{circ} for *CLEAR*, and number of BSJ reads for *CIRI2* and *CIRI-full*. The *r* and *P* were computed by *Pearson* correlation test between *X*- and *Y*-axes.



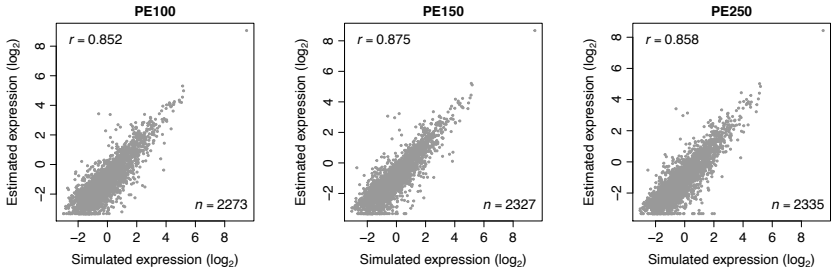
Supplementary Figure S2. Comparison of the quantification performance with varied sequencing depth. The quantification results generated from the simulated dataset *Hela-S1* by the *AQUARIUM*, *CIRIquant*, *CIRI2*, and *CLEAR* algorithms were included. *X* -axis and *Y* -axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM*, CPM for *CIRIquant*, and number of BSJ reads for *CIRI2* and *CLEAR*. The *n* demonstrates the total number of the identified circRNAs from simulated datasets by each tool. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.



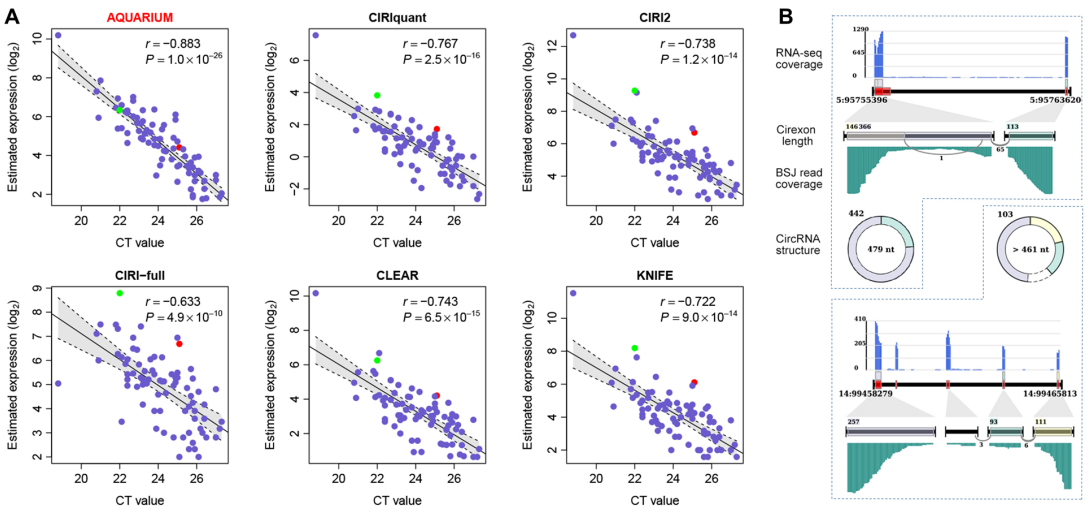
Supplementary Figure S3. Comparison of the quantification performance between *AQUARIUM* and *Sailfish-cir* with varied sequencing depth and reconstructed sequence concordance. CircRNAs that had C_{rs} values less than 0.2 were grouped as the nearly full circRNAs, while circRNAs with C_{rs} values larger than 0.2 were grouped as the partial circRNAs. The simulated dataset *Hela-S1* was used in this comparison. *X* -axis and *Y* -axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM* and *Sailfish-cir*. The *n* demonstrates the total number of the identified circRNAs from simulated datasets by each tool. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.



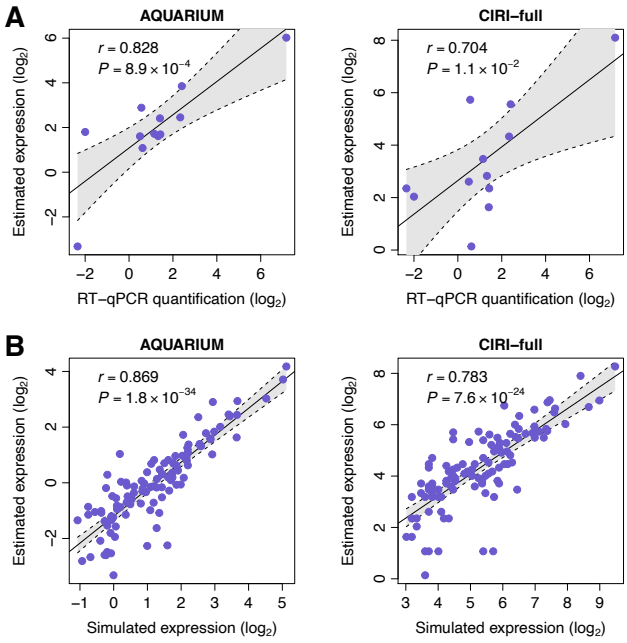
Supplementary Figure S4. Comparison of the quantification performance with varied reconstructed sequence concordance. CircRNAs that had C_{rs} values less than 0.2 were grouped as the nearly full circRNAs, while circRNAs with C_{rs} values larger than 0.2 were grouped as the partial circRNAs. The quantification results generated from one RNA-seq data (60M, PE 150) in the simulated dataset *Hela-S2* by the *AQUARIUM*, *CIRIquant*, *CIRI2*, and *CLEAR* algorithms were included. *X* -axis and *Y* -axis represent the simulated and estimated expression of circRNAs (log_2-transformed), respectively. The expression data were recorded as TPM for *AQUARIUM*, CPM for *CIRIquant*, and number of BSJ reads for *CIRI2* and *CLEAR*. The *n* demonstrates the total number of the identified circRNAs from simulated datasets by each tool. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.



Supplementary Figure S5. Performance of *AQUARIUM* in the simulated datasets *Hela-S2* with different read length. *X* -axis and *Y* -axis represent the simulated and estimated expression of circRNAs (log_2- transformed), respectively. The *n* demonstrates the total number of the identified circRNAs. The *r* was computed by *Pearson* correlation test between the simulated and estimated expression.



Supplementary Figure S6. Example view of two circRNAs with over-estimated expression values by count-based tools, including *CIRIquant*, *CIRI2*, *CIRI-full*, *CLEAR*, and *KNIFE*, since RNA sequencing reads were non-uniformly distributed along the circRNA transcripts. **(A)** Two circRNAs (green and red dots) were selected to show the different estimation accuracy by the *AQUARIUM*, *CIRIquant*, *CIRI2*, *CIRI-full*, *CLEAR*, and *KNIFE*. *X*-axis and *Y*-axis represent the circRNA expression measured by RT-qPCR and the estimated circRNA expression by each tool from RNA-seq data (log_2-transformed), respectively. The *r* and *P* were computed by *Pearson* correlation test between *X*- and *Y*-axes. **(B)** RNA sequencing read coverage, circRNA structure, and BSJ read coverage of two reconstructed circRNAs. The upper panel was related to the green dot in **(A)**, and the bottom panel was corresponding to the red dot in **(A)**.



Supplementary Figure S7. Comparison of circRNA isoform quantification between *AQUARIUM* and *CIRI-full*. Each dot represents one circRNA isoform. TPM and BSJ number were used to represent the expression (log_2-transformed) quantified by *AQUARIUM* and *CIRI-full*, respectively. The *r* and *P* were computed by *Pearson* correlation test between the *X* - and *Y* -axes. **(A)** The expression quantified by *AQUARIUM* shows higher concordance with the RT-qPCR readouts of 12 circRNA isoforms compared with *CIRI-full*. **(B)** *AQUARIUM* outperforms *CIRI-full* on 109 circRNA isoforms identified from the simulated dataset *Hela-S1* with 30M PE150 reads.