Platform-integrated mRNA Isoform Quantification

Jiao Sun, Jae-Woong Chang, Teng Zhang, Jeongsik Yong, Rui Kuang and Wei Zhang

1 Scatter plots of gene expression and isoform expression estimated by different platforms

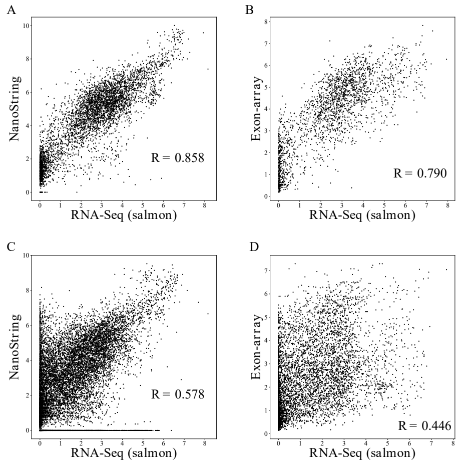

Figure 1: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. Salmon [3] was applied for isoform/gene expression quantification with RNA-Seq data.

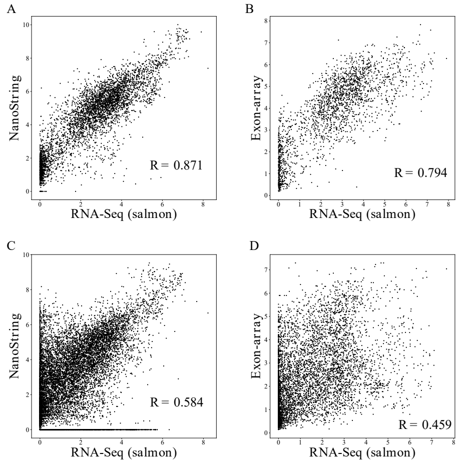

Figure 2: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. Salmon [3] with sequence-specific bias correction was applied for isoform/gene expression quantification with RNA-Seq data.

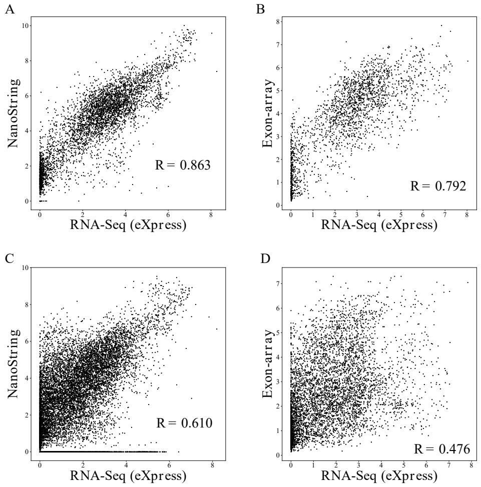

Figure 3: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. eXpress [4] was applied for isoform/gene expression quantification with RNA-Seq data.

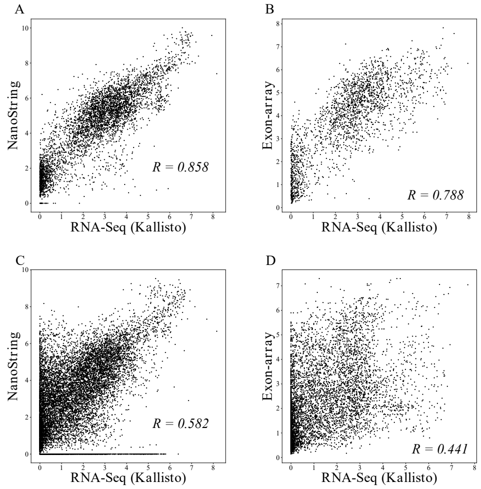

Figure 4: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. Kalliso [1] was applied for isoform/gene expression quantification with RNA-Seq data.

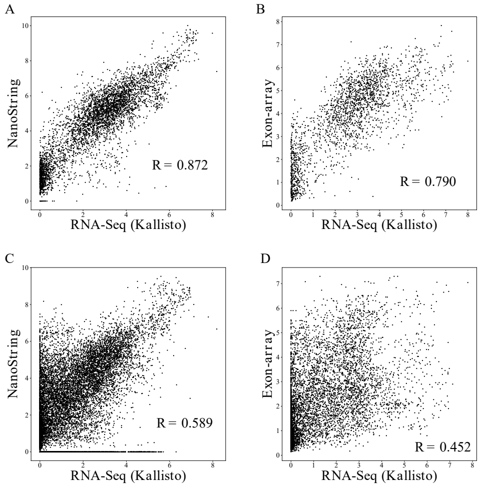

Figure 5: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. Kalliso [1] with sequence-specific bias correction was applied for isoform/gene expression quantification with RNA-Seq data.

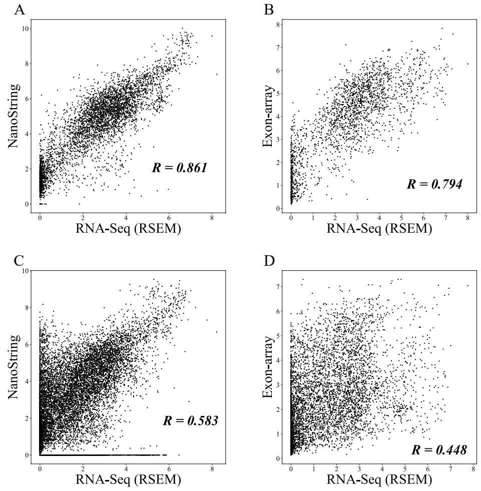

Figure 6: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. RSEM [2] was applied for isoform/gene expression quantification with RNA-Seq data.

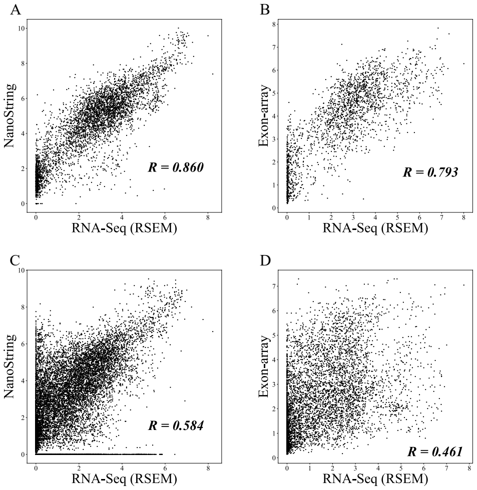

Figure 7: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. RSEM [2] with sequence-specific bias correction was applied for isoform/gene expression quantification with RNA-Seq data.

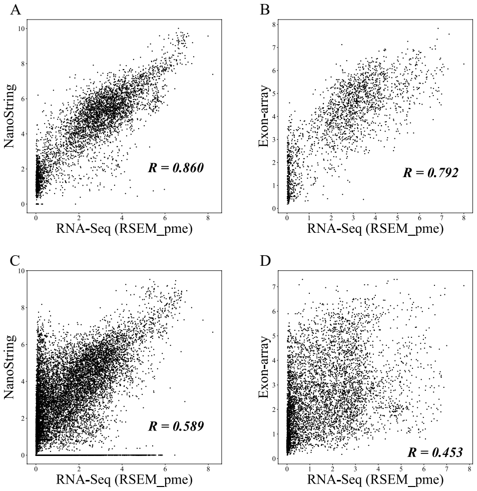

Figure 8: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. RSEM [2] posterior mean estimation was applied for isoform/gene expression quantification with RNA-Seq data.

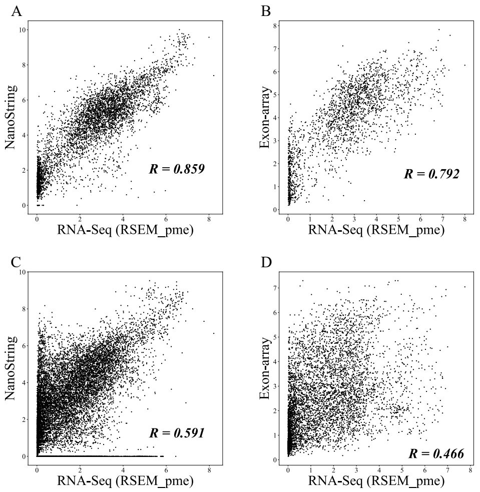

Figure 9: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. RSEM [2] posterior mean estimation with sequence-specific bias correction was applied for isoform/gene expression quantification with RNA-Seq data.

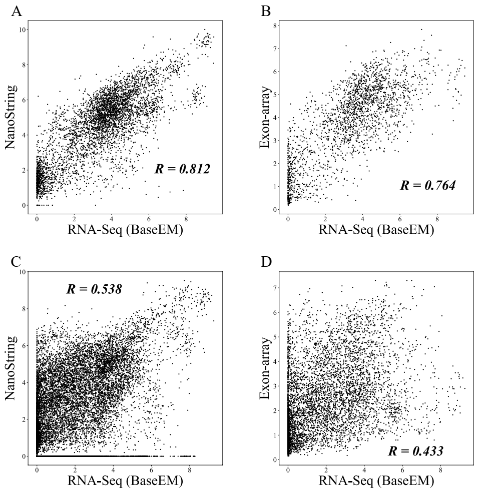

Figure 10: A and B show the correlation of gene expressions between RNA-Seq and NanoString/Exon-array. C and D show the correlation of isoform expressions between RNA-Seq and NanoString/Exon-array. BaseEM was applied for isoform/gene expression quantification with RNA-Seq data.

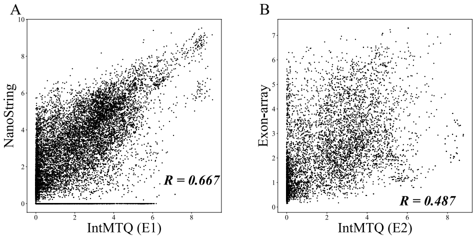

Figure 11: A and B show the correlation of isoform expressions between IntMTQ and NanoString/Exon-array.

2 Data summurization

Source & Cell Line Name		**NanoString**	**RNA-Seq(CCLE)**	**Exon-array(GEO)**	**RT-qPCR**
Ovary	NIH:OVCAR-3	1	1	1 (GSM736089)	NA
	A2780	1	1	1 (GSM1291129)	NA
	Hey-A8	1	1	NA	NA
	SK-OV-3	1	1	1 (GSM736094)	NA
	RMG-I	1	1	NA	NA
	OVCAR-4	1	1	1 (GSM736090)	NA
	Caov-3	1	1	1 (GSM1291130)	NA
	OVSAHO	1	1	NA	NA
	ES-2	1	1	1 (GSM1291140)	NA
	TOV-21G	1	1	1 (GSM1291153)	NA
	KURAMOCHI	1	1	NA	NA
	OVCAR-8	1	1	1 (GSM736092)	NA
Lung	DMS53	1	1	NA	NA
	NCI-H1299	1	1	NA	NA
	NCI-H460	1	1	1 (GSM736074)	NA
	Calu-1	1	1	NA	NA
	Calu-3	1	1	NA	NA
	SK-MES-1	1	1	NA	NA
	A549	1	1	1 (GSM736067)	1
	NCI-H358	1	1	NA	NA
	HCC-H827	1	1	NA	NA
Colon	HCT116	1	1	1 (GSM736062)	1
	SW480	1	1	NA	NA
	HT-29	1	1	1 (GSM736064)	1
	KM12C	1	1	1 (GSM736065)	NA
	HCT-15	1	1	1 (GSM736063)	1
	SW620	1	1	1 (GSM736066)	NA
	Lovo	1	1	NA	NA
Breast	MCF-7	1	1	1 (GSM419264)	1
	BT-549	1	1	1 (GSM419258)	1
	MDA-MB-231	1	1	1 (GSM419268)	1
	T47D	1	1	1 (GSM419291)	1
	SK-BR-3	1	1	1 (GSM419279)	NA
	Hs578T	1	1	1 (GSM419263)	NA
	MDA-MB-436	1	1	1 (GSM419273)	NA
	HCC1937	1	1	1 (GSM419262)	NA
Pancreas	Capan-1	1	1	NA	NA
	MIA-Paca2	1	1	NA	NA
	PANC-1	1	1	1 (GSM472938)	NA
	BxPC-3	1	1	NA	NA
Prostate	DU145	1	1	1 (GSM736095)	1
	PC-3	1	1	1 (GSM736096)	1
Stomach	AGS	1	1	1 (GSM831348)	1
Urinary bladder	J82	1	1	NA	NA
Connective tissue	HT-1080	1	1	1 (GSM969711)	1
Liver	HepG2	1	1	1 (GSM472906)	NA

Table 1: Cell lines in E1 experiment.

Source & Cell Line Name		NanoString	RNA-Seq(CCLE)	Exon-array(GEO)	RT-qPCR
Ovary	A2780	1	1	1 (GSM1291129)	NA
	SK-OV-3	1	1	1 (GSM736094)	NA
	OVCAR-4	1	1	1 (GSM736090)	NA
	Caov-3	1	1	1 (GSM1291130)	NA
	ES-2	1	1	1 (GSM1291140)	NA
	TOV-21G	1	1	1 (GSM1291153)	NA
	OVCAR-8	1	1	1 (GSM736092)	NA
Lung	NCI-H460	1	1	1 (GSM736074)	NA
	A549	1	1	1 (GSM736067)	1
Colon	HCT116	1	1	1 (GSM736062)	1
	HT-29	1	1	1 (GSM736064)	1
	KM12C	1	1	1 (GSM736065)	NA
	HCT-15	1	1	1 (GSM736063)	1
	SW620	1	1	1 (GSM736066)	NA
Breast	MCF-7	1	1	1 (GSM419264)	1
	BT-549	1	1	1 (GSM419258)	1
	MDA-MB-231	1	1	1 (GSM419268)	1
	T47D	1	1	1 (GSM419291)	1
	SK-BR-3	1	1	1 (GSM419279)	NA
	Hs578T	1	1	1 (GSM419263)	NA
	MDA-MB-436	1	1	1 (GSM419273)	NA
	HCC1937	1	1	1 (GSM419262)	NA
Prostate	DU145	1	1	1 (GSM736095)	1
	PC-3	1	1	1 (GSM736096)	1
Pancreas	PANC-1	1	1	1 (GSM472938)	NA
Stomach	AGS	1	1	1 (GSM831348)	1
Connective tissue	HT-1080	1	1	1 (GSM969711)	1
Liver	HepG2	1	1	1 (GSM472906)	NA

Table 2: Cell lines in E2 and E3 experiment.

3 Designed primer sequences for RT-qPCR experiment

Primer sequences to measure the expression for each transcript are the following:
hLRIG3 iso1,2 reverse CTCATGGAACTTGCCTTGATGA
hLRIG3 iso1 forward TTGTTCTCCCTCTGCTTGCT
hLRIG3 iso2 forward CGTCTTCCCGAGCCACTC
hNOTHC2 iso1,2 forward ACCTTGTGAACCATTTCAAGTGC
hNOTHC2 iso1 reverse GGCACAGTCATCAATGTTCTCT
hNOTHC2 iso2 reverse GACAATGCCCTGGATGGAAAA
hTPM4 iso1 forward AATATTCCGAGGACCTGAAGGA
hTPM4 iso1 reverse ATGCGTCGGTTGAGGGC
hTPM4 iso2 forward CGGTGAAACGCAAGATCCAG
hTPM4 iso2 reverse ATCACCTTCAGCTTTCTCGC
hCD79A iso1 forward GGAGGGCAACGAGTCATACC
hCD79A iso1 reverse GATTCGGTTCTTGGTGCCCT
hCD79A iso2 forward TCCTCCATGGCAACTACACG
hCD79A iso2 reverse ATTCGGTTCTTGGTGCCCTC
hBCL2 iso1 forward GCTTTTGTTTTGAGTTACTGGGGT
hBCL2 iso1 reverse AGAGCCATGGAAGGTAAAAGTATGA
hBCL2 iso2 forward TTGGTGATGTGAGTCTGGGC
hBCL2 iso2 reverse TTTATTTCGCCGGCTCCACA
hARID1A iso1 forward CCACCAAGCATGCAGAATCA
hARID1A iso1 reverse TGCAGGAATGGAGACTTGCT
hARID1A iso2 forward AGCCTGTGTTGAAGCAGAGGAG
hARID1A iso2 reverse GAGACCAGACTTGAGGGACATC
hBCR iso1 forward ACAGCTGAGCCAAACTGGAA
hBCR iso1 reverse TCCTCCTTGGGGATCTTCGT

hBCR iso2 forward CAAACTGGAACGAGCTGGACC
hBCR iso2 reverse CCCTGCTGTTGAACTTGACCG

4 Commands for TopHat2 alignment and baseline methods

1. TopHat2
   - transcriptome: tophat -p 20 -o out/ TopHat_hg19_refseq/hg19 *_1.fastq *_2.fastq
2. RSEM
   - no bias correction: rsem-calculate-expression -p 20 –paired-end –bowtie2 –calc-pme *_1.fastq.gz *_2.fastq.gz hg19_refseq out
   - bias correction: rsem-calculate-expression -p 20 –paired-end –bowtie2 –calc-pme –estimate-rspd *_1.fastq.gz *_2.fastq.gz hg19_refseq out

   Please note that the paired-end fastq files are used as input of RSEM to perform transcriptome alignment and there is only 1.27% difference based on the overall alignment rates reported by TopHat2 and RSEM (Bowtie2) on the 46 cell lines RNA-seq data. Although both RSEM and eXpress use transcriptome alignments as input, the "sorted.bam" file used by eXpress can not be the input for RSEM due to the different requirements of the input format.
3. Kallisto
   - no bias correction: kallisto quant -i hg19_refseq.idx -o out/ *_1.fastq *_2.fastq
   - bias correction: kallisto quant -i hg19_refseq.idx -o out/ –bias *_1.fastq *_2.fastq
4. eXpress
   - sort BAM file: samtools sort -n -o sorted.bam input.bam
   - no bias correction: express –no-bias-correct -o out hg19_refseq.fasta sorted.bam
   - bias correction: express -o out hg19_refseq.fasta sorted.bam
5. Salmon
   - no bias correction: salmon quant -i hg19_refseq_index -l A -1 *_1.fastq -2 *_2.fastq –validateMappings -o out/
   - bias correction: salmon quant -i hg19_refseq_index -l A -1 *_1.fastq -2 *_2.fastq –validateMappings –seqBias -o out/

5 Parameter tuning

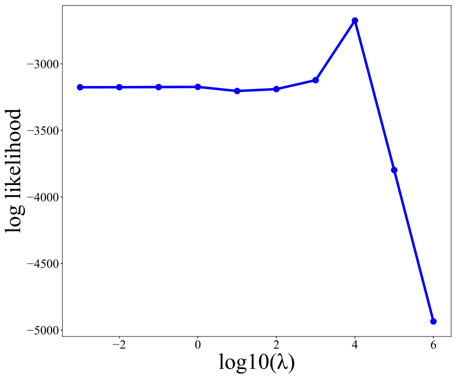

Figure 12: The plot of the log-likelihood function $\log\left(\mathcal{L}(\boldsymbol{P}_{(\lambda)};\boldsymbol{r})\right)$ with different λ values on experiment one (E1).

6 l_1-norm, l_2-norm, and infinite norm

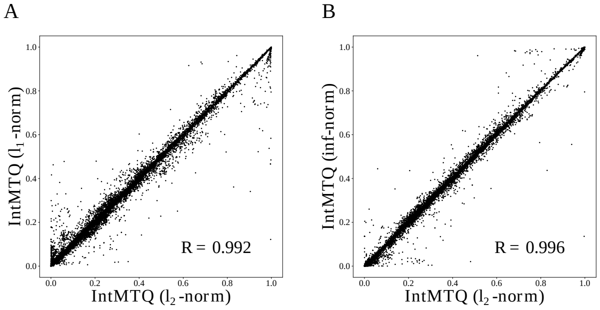

Figure 13: Scatter plots of isoform proportions estimated by IntMTQ with different norms. A shows the correlation of isoform proportions estimated by IntMTQ with l_1 and l_2 norms. B shows the correlation of isoform proportions estimated by IntMTQ with l_2 and infinite norms.

7 Robustness of IntMTQ

We performed following experiments to test the robustness of IntMTQ with different factors that could affect the isoform quantification results and accuracy: (1) *Number of isoforms*: we categorized the genes in experiment one (E1) into three groups based on the number of isoforms. The first group only contains the genes with two isoforms, the second group contains the genes with three isoforms, and the third group contains the genes with more than three isoforms. Then, we performed cancer cell line clustering follow the same strategy in section 3.4 for each group of genes. The clustering results are shown in the table below. In most cases, IntMTQ performs better than the other two baseline methods. It is also interesting to observe that the gene contains more isoforms, the isoforms in these genes have better discriminative power to cluster the cell lines into the correct groups.

(2) *Gene expression level*: We categorized the isoforms in E1 into two groups based on the expression level. The top 50% of the genes are considered as high expressed genes and the bottom 50% of the genes are considered as low expressed genes. Then, we performed cancer cell line clustering follow the same strategy in section 3.4 for both groups of genes. The clustering results

Category	Metrics	IntMTQ	BaseEM	Kallisto
c1 (2 isoforms)	ARI	**0.120**	0.115	0.048
	NMI	0.303	**0.304**	0.197
c2 (3 isoforms)	ARI	**0.138**	0.087	0.134
	NMI	0.283	0.270	**0.319**
c3 (more than 3 isoforms)	ARI	**0.233**	0.121	0.119
	NMI	**0.427**	0.318	0.346

Table 3: **Results of hierarchical clustering on four cancer types.** The genes are categorized into three groups based on different number of isoforms, and the clustering performances are evaluated separately for each group. The best result across the three methods are bold.

are shown in the table below. In three out of four cases, IntMTQ performs better than the other two baseline methods. Based on the small sets of isoforms and cell lines in E1, we observe that the isoforms in the low expressed gene have better discriminative power to cluster the cell lines into the correct groups.

Category	Metrics	IntMTQ	BaseEM	Kallisto
c1 (low)	ARI	**0.246**	0.171	0.161
	NMI	**0.419**	0.343	0.370
c2 (high)	ARI	0.111	**0.115**	0.074
	NMI	**0.242**	0.236	0.211

Table 4: **Results of hierarchical clustering on four cancer types.** The genes are categorized into two groups based on gene expression levels, and the clustering performances are evaluated separately for each group. The best result across the three methods are bold.

(3) *Sequence depth*: This small experiment was performed on RT-qPCR validated seven genes and twelve cell lines. To access the impact of the sequence depth, we randomly sampled RNA-Seq reads aligned to the genes from 10% to 90% and applied IntMTQ and BaseEM to estimate the isoform expression. This process was repeated 1000 times and the average root mean square errors between the estimated isoform expression and RT-qPCR results are plotted in the figure below. For both IntMTQ and BaseEM, the isoform quantification accuracy is significantly improved as more RNA-Seq reads are sequenced, and using 40% of the reads can reach similar quantification accuracy compare to using all the reads. Though IntMTQ get more accurate quantification results compared to BaseEM, we agree that IntMTQ is also sensitive to the sequence depth based on the experimental results.

References

[1] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5):525, 2016.

[2] B. Li and C. N. Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[3] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417, 2017.

[4] A. Roberts and L. Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, 10(1):71, 2013.

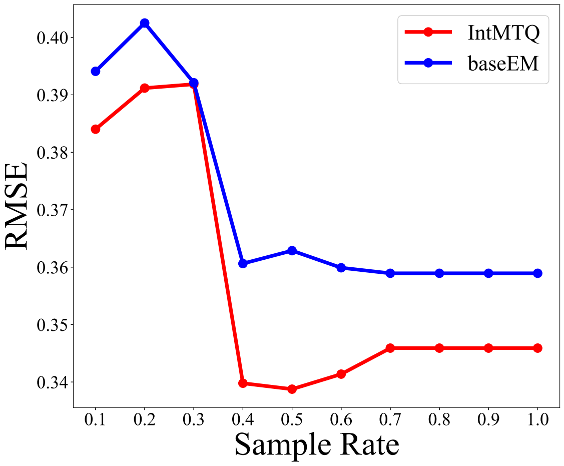

Figure 14: Average root mean square errors between BaseEM/IntMTQ and RT-qPCR results under different sequence sampling rate.