Supplemental Material

TransMeta simultaneously assembles multi-sample RNA-seq reads

Ting Yu[1 †] Xiaoyu Zhao[1, 2, †] Guojun Li[1,3,*]

[1]Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, 266237, China

[2]School of Mathematics, Shandong University, Jinan, Shandong, 250100, China

[3]School of Mathematical Science, Liaocheng University, Liaocheng 252000, China

[*] To whom correspondence should be addressed

[†]These authors contributed equally to this work

Contents of Supplemental Material

1. Supplemental Notes

1.1 Parameter setup for the compared assemblers

We used HISAT2 and STAR to produce all alignments for the RNA-seq samples, where STAR was run with its default settings and HISAT2 was run with the option ***--dta*** (in the HISAT2 manual it means *"reports alignments tailored for transcript assemblers"*). The default parameters are used for the single-sample assemblers StringTie2 and Scallop to produce transcripts for each sample. To draw the precision-recall curves for each assembler at meta-assembly level, we used a wide range of minimum coverage thresholds to get the results. Thus, TransMeta was set the parameter *--min_meta_trans_cov* to be 0, 0.5, 1, 2(default), 5, 10, 20, 30, 50, 100, 150. PsiCLASS was set the parameter *–d* for its vote-transcripts procedure to be 0, 0.5, 1(default), 2, 3, 5, 10, 30, 50, 100, 200. TACO was set the parameter -- *isoform-frac* to be 0.01, 0.05(default), 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and StringTie2-Merge was set the parameter *–F* to be 0.01, 1(default), 1.5, 2, 2.5, 3, 3.5,

4, 5, 6. Detailed versions and running command for all the tools are described below. All the assemblers were tested on a server with 768GB of RAM and 32core of CPU. Besides, it is worth mentioning that the gtfmerge tool was downloaded from https://github.com/Kingsford-Group/rnaseqtools.

HISAT2	2.0.5	hisat2-build *GenomeFile.fasta* ***Index*** **-p** *20* hisat2 **-x** *Index* **-1** *fastq1* **-2** *fastq2* -S *SamFile* **-p** *20* **--dta**
STAR	2.5.3a	**STAR --runThreadN** *20* **--runMode** *genomeGenerate* **--genomeDir** *Index* **--genomeFastaFiles** *GenomeFile.fasta* **STAR --outSAMstrandField** *intronMotif* **--runThreadN** *10* **--genomeDir** *Index* **--readFilesIn** *fastq1 fastq2*
TransMeta	v.1.0	TransMeta **-B** *BamList* **-s** *strandness* **-p** *25* **-o** *TransMetaDir* **--min_meta_trans_cov** *coverage-threshold*
PsiClass	v1.0.1	psiclass **--lb** *BamList* **-p** *25* vote-transcripts **--lg** *psiclass-GtfList* **-d** *coverage-threshold*
StringTie	2.1.4	stringtie *bamfile* **-o** *GtfFile* stringtie **--merge** *GtflList* **-F** *coverage-threshold*
Scallop	v.0.10.4	scallop **-i** *bamfile* **-o** *GtfFile* **--library_type** *strandness*
TACO	0.7.3	taco_run *GtfList* **-p** 10 **-o** *TACO-outdir* **--gtf-expr-attr** *RPKM* **--isoform-frac** *filtering-thresholds*
RSEM	1.3.3	rsem-prepare-reference **--gtf** *transcripts.gtf genome.fa ref_name* **--bowtie2** rsem-calculate-expression **--paired-end** *reads_1.fq reads_2.fq ref_name* *sample_name.stat* **--bowtie2** rsem-simulate-reads *ref_name* *sample_name.stat.stat/sample_name.stat.model* *sample_name.stat.isoforms.results 0.25 read_number simulated_reads*
kallisto	0.43.1	kallisto **index -i** *kallisto_index transcritpt.fa* kallisto **quant -i** *kallisto_index* **-t** *20* **-o** *kallisto_results reads_1.fastq reads_2.fastq*
gtfmerge	-	gtfmerge **union** *gtf-list GTFMerge.gtf* **-t** *10*

1.2 Reference genome and transcripts used in this study

In order to evaluate the performances of the assemblers on real datasets, we downloaded the reference genome and transcriptome of the species Homo sapiens (version: GRCh38/hg38) from the UCSC Genome Browser at http://genome.ucsc.edu/cgi-bin/hgTables. As an alternative, the reference genome and transcriptome used in this research could also be available at https://sourceforge.net/projects/transmeta/files/RealData/.

1.3 Description for the simulated datasets

We simulated 3 datasets (S1, S2 and S3) in this research, with different sample size, different read lengths, and different sequence depths, by using the RSEM-simulator. The parameters (e.g., abundances, fragment and read length distributions, and sequencing error model parameters) of the RSEM simulation were learned from real data sets. We used different real datasets in this study to produce the simulated datasets. To be specific, the NCBI SRA accession code of 4 samples used for simulating data S1 were SRR7807483, SRR7807484, SRR7807485, and SRR7807486; the 5 samples used for simulating data S2 were SRR8315677, SRR8315678, SRR8315679, SRR8315680, and SRR8315682; and the 3 samples used for simulating data S3 were the 3 samples in the real dataset R7. For each of the samples, we produced 5 simulated RNA-seq data sets, therefore there were 20 samples in S1, 25 in S2, and 15 in S3. In the simulation process of RSEM, the first step is to estimate the abundances of a reference transcripts via using a real dataset. Then based on the estimated abundances and other sequencing model parameters, RSEM simulated the RNA-seq reads. Therefore, for each simulated RNA-seq sample, we explicitly know the expressed transcripts and the abundance of each expressed transcript. Subsequently, we set all the expressed transcripts from all the simulated samples as the ground truth for the whole RNA-seq simulation experiment (i.e. the ground truth for benchmarking the assemblers at the meta-assembly level), and for each transcript in the ground truth, we set the sum of this transcript's abundances from each sample as its expression level. Then, we first sorted the transcripts in the ground truth according to their expression levels, and all the expressed transcripts were equally divided into three parts, which corresponded to low, middle, and high expressed ones. The alignments of all the 60 simulated samples produced by HISAT2 and STAR. Particularly, as an illustration, we uploaded the HISAT2 alignments of the 20 samples in S1 to the website

https://sourceforge.net/projects/transmeta/files/SimulatedData/, as well as the ground truth for each sample and for the whole experiment. And，we also uploaded the assemble results of the tested assemblers to this website. The readers can easily and freely obtain the files.

1.4 Description for the real datasets

Raw sequence data for the real datasets R1-R6 can be obtained from NCBI BioProject under the accession PRJNA575230, PRJNA531468, PRJEB35202, PRJNA509906, PRJNA489891 and PRJNA30709 respectively. Specifically, in this research, R1 contains all the 73 runs from PRJNA575230; R2 contains all the 48 runs from PRJNA531468; R3 contains all the 35 runs from PRJEB35202; R4 contains 21 runs from PRJNA509906 with SRA accession code from SRR8315677 to SRR8315697; R5 contains all the 12 runs from PRJNA489891; and R6 contains 6 runs from PRJNA30709 with SRA accession code from SRR534319 to SRR534324. And sequence data for R7 (MCF-7) can be downloaded from UCSC genome browser at the website http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeCaltechRnaSeq. The GEUVADIS population variation project were publicly available from ArrayExpress with accession E-GEUV-6. Moreover, the assembled results of each assembler for datasets R1-R5 in the main text could be freely available at https://sourceforge.net/projects/transmeta/files/**.**

2. Supplemental Methods

Details for the constrained program.

The constrained optimization problem tries to link the in-coming and out-going splice junctions by optimally finding pairs of vectors respectively associated with in-coming and out-going splice junctions (edges) base on cosine similarity. Suppose that there are m in-coming splice junctions $\{j_1, j_2, \dots, j_m\}$and n out-going splice junctions $\{j'_1, j'_2, \dots, j'_n\}$ for a given exon, and denote CP as the set of consensus paired paths. In the constrained program, the variable M indicates that both the in-coming and out-going junctions for an exon will be partitioned into $M\,(1 < M \leq \min(m, n))$ parts, denoted as $P_1, P_2, .., P_M$ and $P'_1, P'_2, .., P'_M$. And P_i and P_i' are one-to-one correspondence, which indicates that a splice junction in P_i (P_i') can only be connected to a junction in P_i' (P_i) in the following path searching procedure. Since we expect that the corresponding parts P_i and P_i'

possess higher similarity, the objective is then taken to maximize the minimum one in $\{\mathrm{cs}(P_1, P'_1), cs(P_2, P'_2), \dots, cs(P_M, P'_M)\}$, with the constraints 1) $\bigcup_{i=1}^{M} P_i = \{j_1, j_2, \dots, j_m\}$ ensuring that each in-coming junction should belong to one part P_i; 2) $\bigcup_{i=1}^{M} P'_i = \{j'_1, j'_2, \dots, j'_n\}$ ensuring that each out-going junction should belong to one part P'_i; 3) $P_i \cap P_k = \emptyset$ & $P'_i \cap P'_k = \emptyset \ if \ (i \neq k)$ ensuring that a junction can only belong to one part; 4) $cs(P_i, P'_i) > cs(P_i, P'_k), i = 1,2, \dots, M \ and \ i \neq k$ meaning that the similarity between the corresponding parts should be higher than others; 5) $cs(P_i, P'_i) \geq \rho, 0 < \rho < 1$, where ρ is a parameter, meaning that the cosine similarity between the corresponding parts should be larger than ρ; 6) $\exists k \in [1, M] \ such \ that \ j \in P_k \ and \ j' \in P'_k, if(j, j') \subset cp, cp \in CP$ meaning that if there exist a consensus paired path continuously contains *j* and *j'*, then *j* and *j'* should be in a certain pair of the corresponding parts P_k and P_k'. Therefore, the constrained program takes both the paired-end information and the vector weights (i.e. the sequence depth information of each sample) into account, which leads to accurately combing the splice junctions for an exon.

3. Supplemental Figures

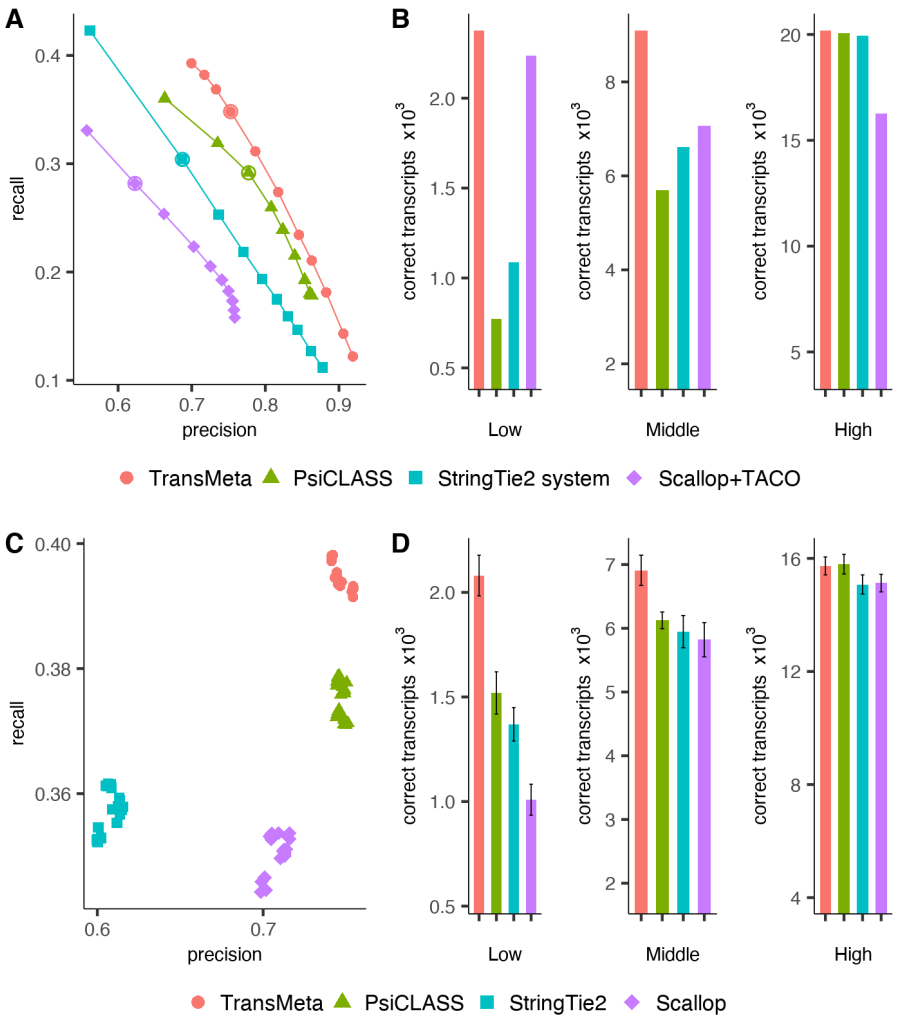

Figure S1: Performance evaluations on the simulated dataset S1 under the STAR alignments. **A)** Precision-recall curves of the assemblers. The points on the curve of an assembler correspond to the filtering thresholds of the assembler, and the circled one to the default value. **B)** Distributions of numbers of correctly recovered transcripts against the tools that were separately counted according to whether their expression is low, middle, or high at the meta-assembly level. **C)** The precisions and recalls of the assemblers on different samples at the individual samples level under their default settings. Different colors correspond to different assemblers, and each point corresponds to a specific sample. **D)** Distributions of averaged numbers of correctly recovered transcripts against the tools that were separately counted according to whether their expression is low, middle, or high at the individual samples level. The error bars show the standard deviation across the samples.

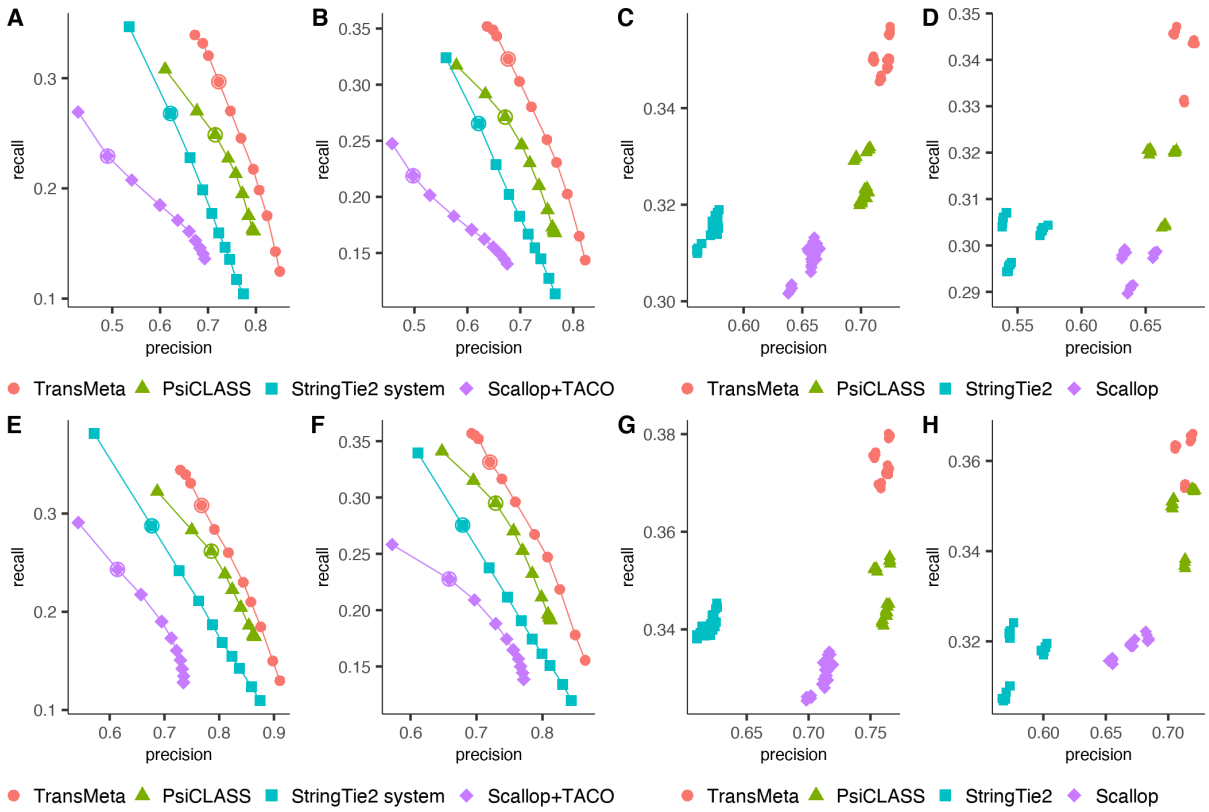

Figure S2: When employing the HISAT2 alignment, precision-recall curve of each assembler at the meta-assembly level **A)** on the dataset S2 and **B)** on the dataset S3; precisions and recalls of each assembler at the individual sample level **C)** on the dataset S2 and **D)** on the dataset S3. When employing the STAR alignments, precision-recall curve of each assembler at the meta-assembly level **E)** on the datasets S2 and **F)** on the dataset S3; precisions and recalls of each assembler at the individual sample level **G)** on the dataset S2 and **H)** on the dataset S3.

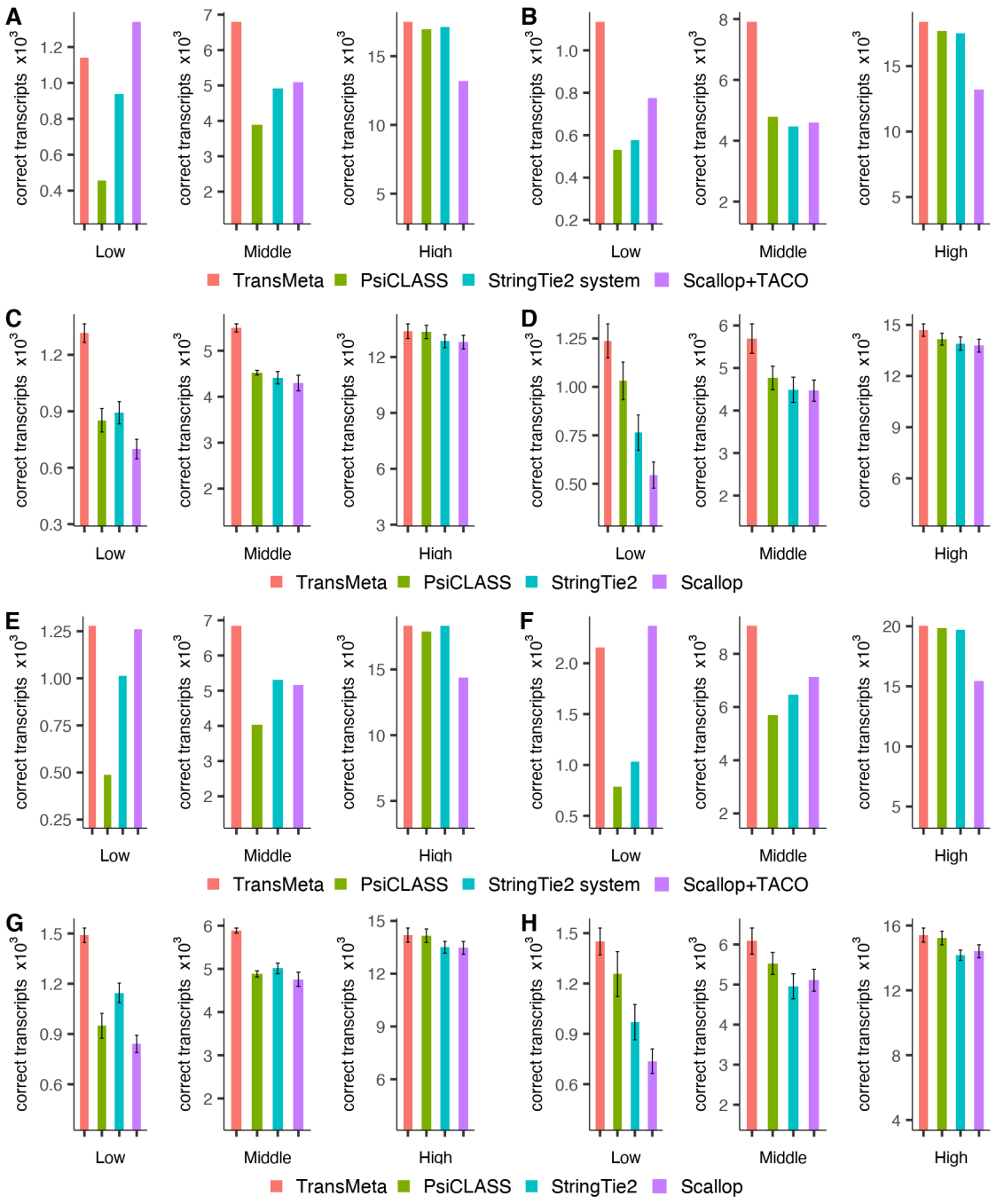

Figure S3: Under the HISAT2 alignments, correctly recovered transcripts of the assemblers based on different expression levels at the meta-assembly level **A**) on the dataset S2 and **B)** on the dataset S3; at the individual sample level **C)** on the dataset S2, and **D)** on the dataset S3. Under the STAR alignments, correctly recovered transcripts of the assemblers based on different expression levels at the meta-assembly level **E**) on the dataset S2 and **F)** on the dataset S3; at the individual sample level **G)** on the dataset S2, and **H)** on the dataset S3. The error bars show the standard deviation across the samples.

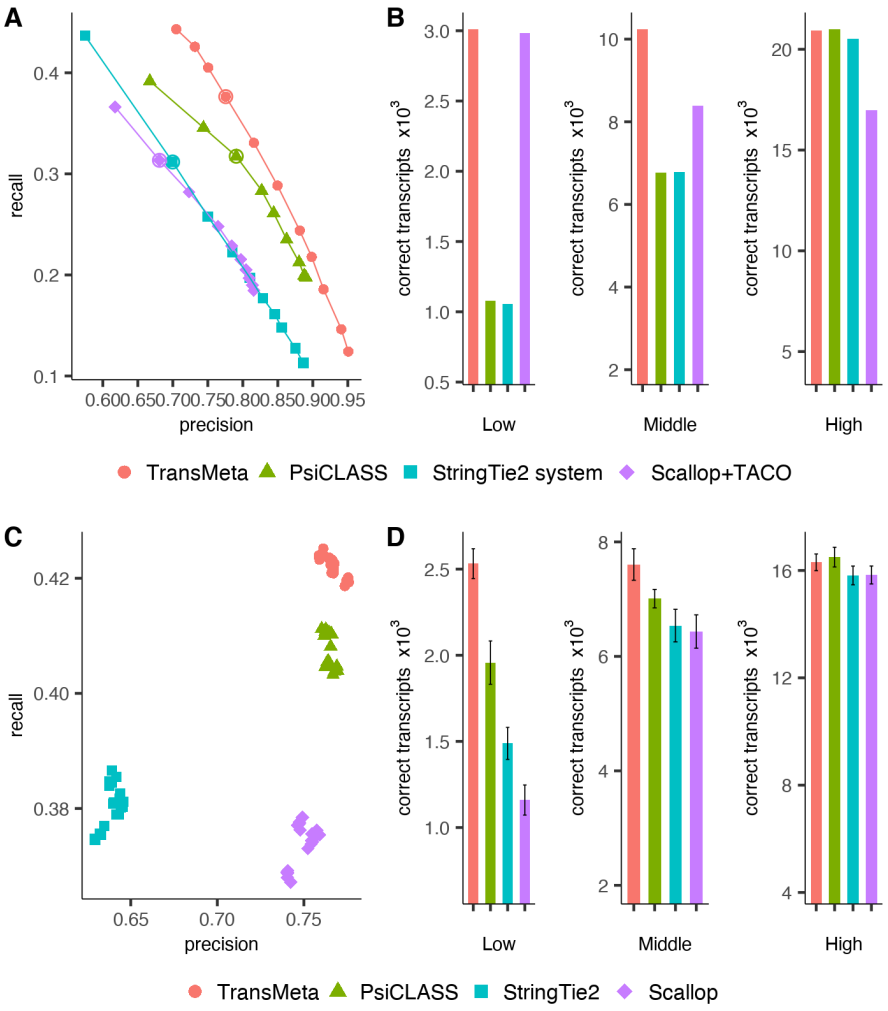

Figure S4: Performance evaluations when employing *--ss* and *--exon* options to build the index for the HISAT2 alignments. **A)** Precision-recall curves of the assemblers. The points on the curve of an assembler correspond to the filtering thresholds of the assembler, and the circled one to the default value. **B)** Distributions of numbers of correctly recovered transcripts against the tools that were separately counted according to whether their expression is low, middle, or high at the meta-assembly level. **C)** The precisions and recalls of the assemblers on different samples at the individual samples level under their default settings. Different colors correspond to different assemblers, and each point corresponds to a specific sample. **D)** Distributions of averaged numbers of correctly recovered transcripts against the tools that were separately counted according to whether their expression is low, middle, or high at the individual samples level. The error bars show the standard deviation across the samples.

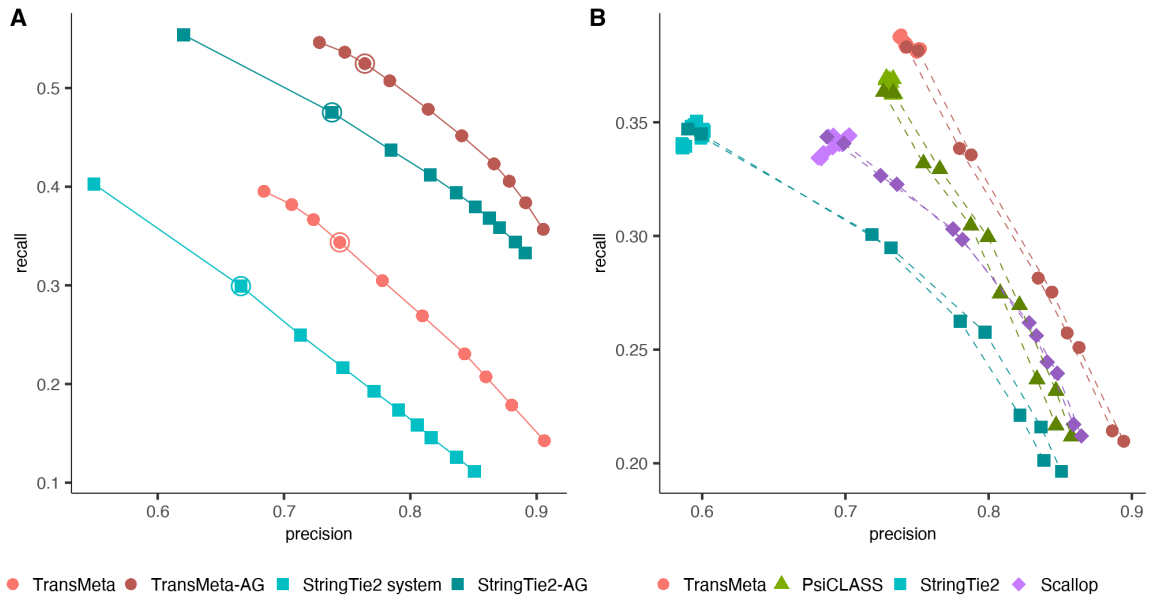

Figure S5. When employing the HISAT2 alignments**, A)** precision-recall curves for the annotation guided mode of TransMeta and StringTie2 system on the simulated dataset S1 at the meta-assembly level. The points on each curve correspond to the filtering thresholds of each assembler, and the circled one to the default value. TransMeta-AG means TransMeta was run with its option *-g*, and StringTie2-AG means StringTie2 *--merge* was run with its option *-G*. The reference transcripts provided to the assemblers were produced in the following way, we first randomly selected 25% the transcripts from the ground truth, then we added about 10% noise (transcripts that are not in the ground truth) to the selected ones as the final reference transcripts that provided to TransMeta and StringTie2; **B)** precision-recall curves of the assemblers on the simulated dataset S1 at the individual sample level. From the dataset S1, we selected two samples to draw the precision-recall curve with a range of filtering thresholds for each assembler.

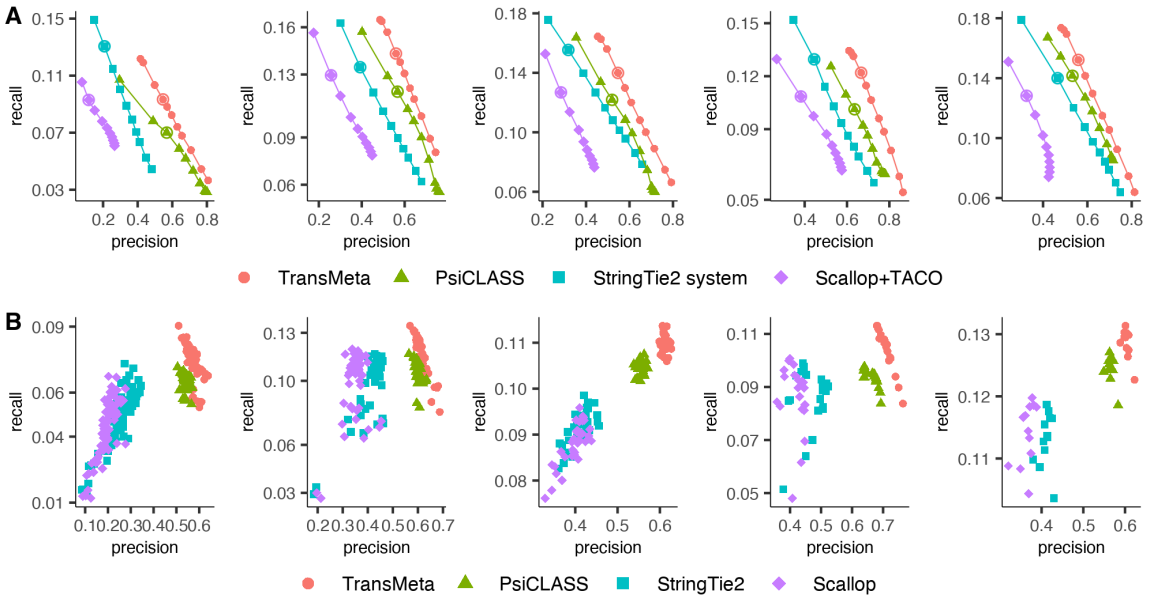

Figure S6. Performance evaluations on the five real datasets R1-R5 under the STAR alignments. **A)** Precision-recall curves of the assemblers. The points on each curve correspond to the filtering thresholds of each assembler, and the circled one to the default value. **B)** The precisions and recalls of the assemblers on different samples at the individual samples level under their default settings. Different colors correspond to different assemblers, and each point corresponds to a specific sample.

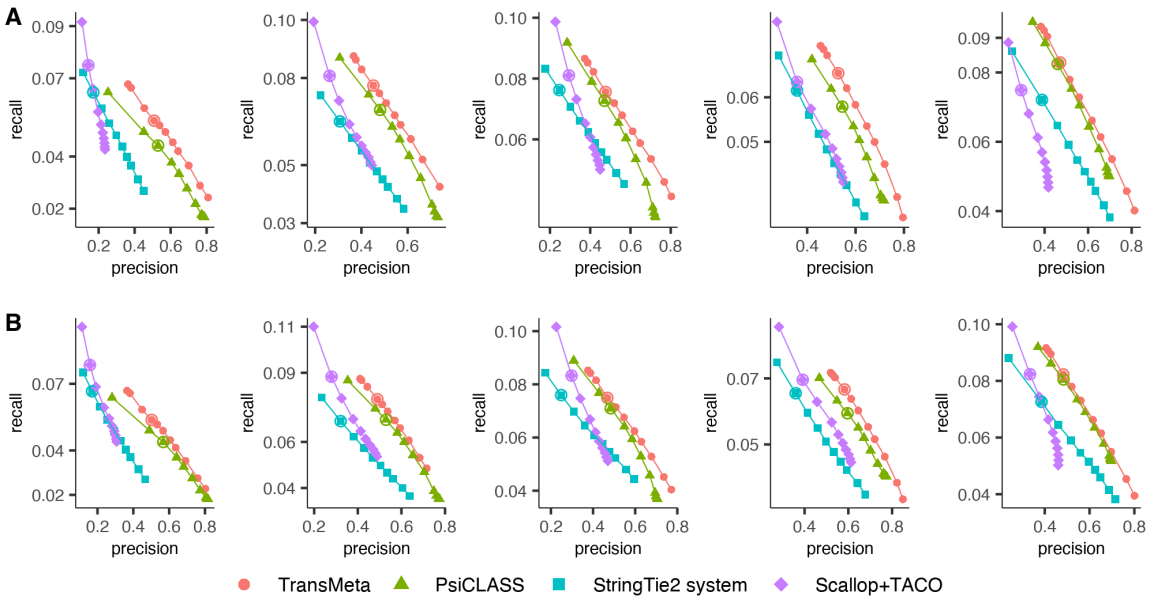

Figure S7 Precision-recall curves of assemblers on the five real datasets R1-R5 at the meta-assembly level according to the ground truth of GENECODE annotations **A)** under the HISAT2 alignments, and **B)** under the STAR alignments. The points on each curve correspond to the filtering thresholds of each assembler, and the circled one to the default value.

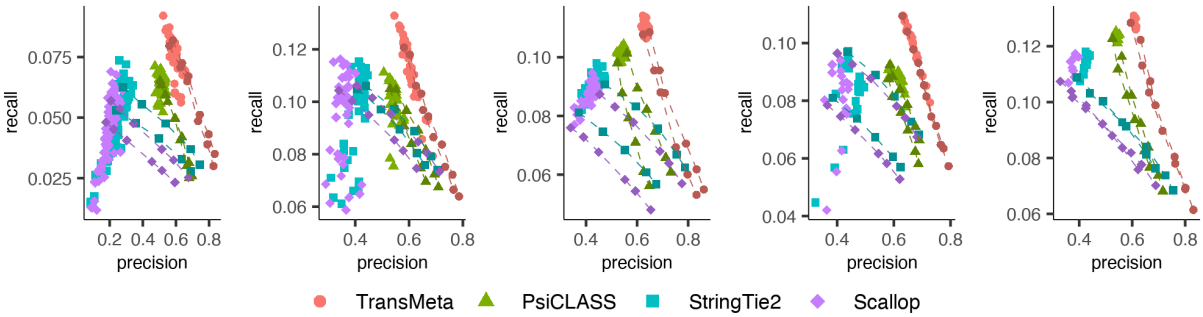

Figure S8 When employing the HISAT2 alignments**,** precision-recall curves for the assemblers on the five real datasets(R1-R5) at the individual sample level. From each of the datasets R1-R5, we selected two samples to draw the precision-recall curve with a range of filtering thresholds for each assembler.

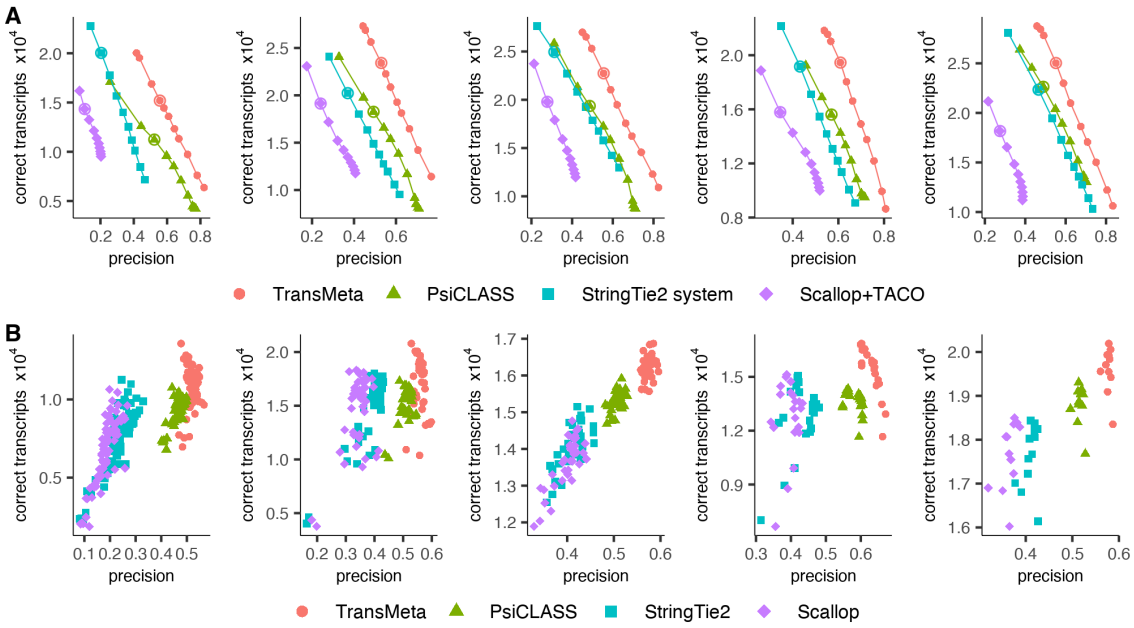

Figure S9. Based on the ground truth that estimated by kallisto, when employing the HISAT2 alignments, **A)** precision-correct number curves of the assemblers on the five real datasets R1-R5 at the meta-assembly level. The points on each curve correspond to the filtering thresholds of each assembler, and the circled one to the default value; **B)** precisions and correctly assembled transcripts of the assemblers on the five real datasets R1-R5 at the individual sample level under the default settings. Each point corresponds to specific a sample.

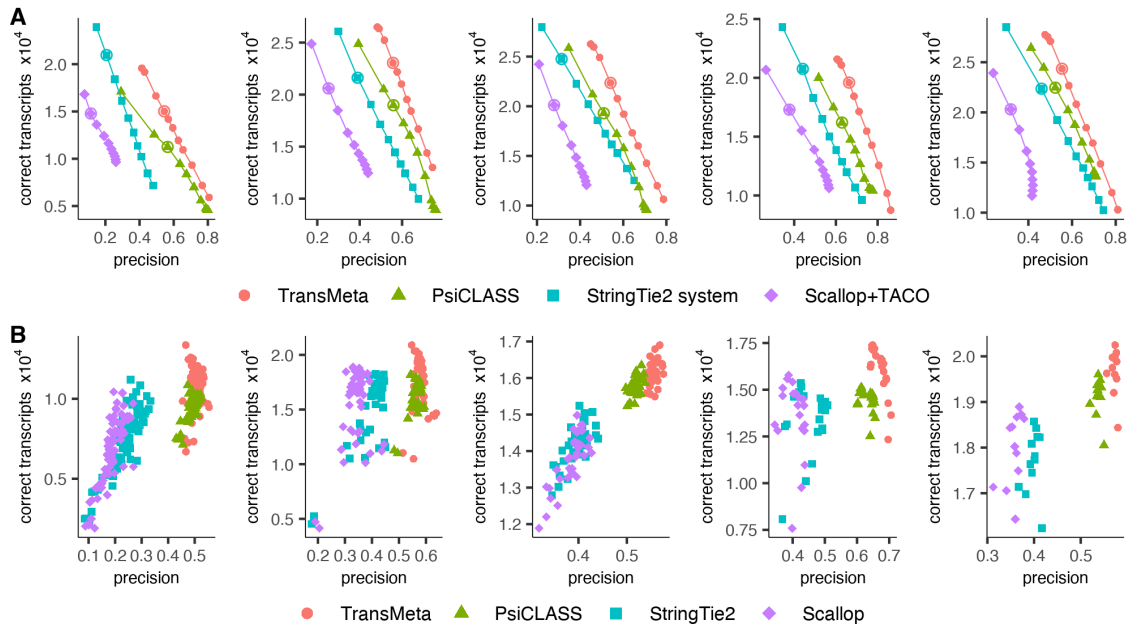

Figure S10. Based on the ground truth that estimated by kallisto, when employing the STAR alignments**, A)** precision-correct number curves of assemblers on the five real datasets R1-R5 at the meta-assembly level. The points on each curve correspond to the filtering thresholds of each assembler, and the circled one to the default value; **B)** precisions and correctly assembled transcripts of assemblers on the five real datasets R1-R5 at the individual sample level under the default settings. Each point corresponds to a specific sample.

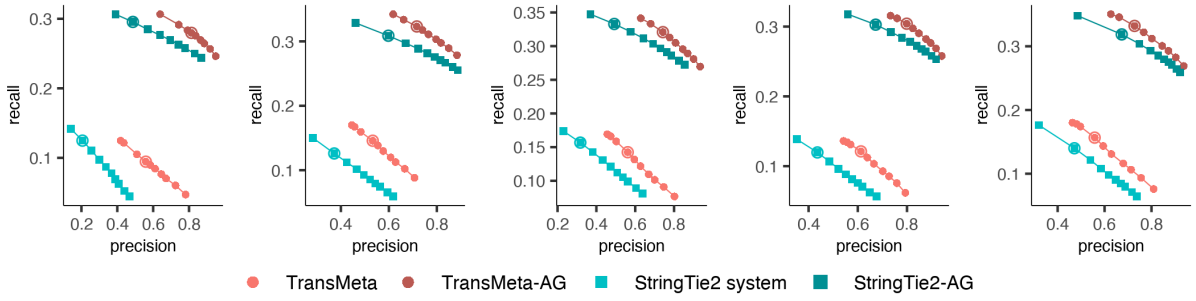

Figure S11 When employing the HISAT2 alignments**,** precision-recall curves for the annotation guided mode of TransMeta and StringTie2 system on the five real datasets R1-R5 at the meta-assembly level. The points on each curve correspond to the filtering thresholds of each assembler, and the circled one to the default value. TransMeta-AG means TransMeta was run with its option *-g*, and StringTie2-AG means StringTie2 *--merge* was run with its option *-G*. The reference transcripts provided to the assemblers were produced in the following way, we randomly selected 25% the reference transcripts from the NCBI_RefSeq annotations as the input for TransMeta and StringTie2.

4. Supplemental Tables

Table S1. Correctly recovered transcripts, candidates, precision, recall and F-score of the assemblers on the simulated dataset S1 at the meta-assembly level under the HISAT2 alignments. The abbreviations ST is for StringTie2, and SC for Scallop, and assembler-0 means the results for running the assembler with the minimum filtering threshold, and assembler-default indicates the results for the default settings.

Assemblers	Correct transcripts	Candidates	precision	recall	F-score
TransMeta-0	36119	52797	0.6841	0.3954	0.5011
TransMeta-default	31389	42179	0.7442	0.3436	0.4701
PsiCLASS-0	32328	51628	0.6262	0.3539	0.4522
PsiCLASS-default	26501	35410	0.7484	0.2901	0.4181
ST+STmerge-0	36782	66929	0.5496	0.4026	0.4648
ST+STmerge-default	27334	41051	0.6659	0.2992	0.4129
SC+TACO-0	29106	61637	0.4722	0.3186	0.3805
SC+TACO-default	25090	47338	0.5300	0.2746	0.3618
gtfmerge+ST	47858	195278	0.2451	0.5239	0.3339
gtfmerge+SC	42574	126783	0.3358	0.4660	0.3903

Table S2. Correctly recovered transcripts, candidates, precision, recall and F-score of the assemblers on the real dataset R1 at the meta-assembly level under the HISAT2 alignments. The abbreviations ST is for StringTie2, and SC for Scallop, and assembler-0 means the results for running the assembler with the minimum filtering threshold, and assembler-default indicates the results for the default settings.

Assemblers	Correct transcripts	Candidates	precision	recall	F-score
TransMeta-0	20143	48123	0.4186	0.1247	0.1922
TransMeta-default	15257	27290	0.5591	0.0945	0.1616
PsiCLASS-0	17292	66145	0.2614	0.1071	0.1519
PsiCLASS-default	11303	21473	0.5264	0.0700	0.1235
ST+STmerge-0	22912	162639	0.1409	0.1419	0.1414
ST+STmerge-default	20150	97915	0.2058	0.1248	0.1553
SC+TACO-0	16197	219191	0.0739	0.1003	0.0851
SC+TACO-default	14337	135559	0.1058	0.0888	0.0965
gtfmerge+ST	33884	879155	0.0385	0.2098	0.0651
gtfmerge+SC	28583	718865	0.0398	0.1770	0.0649

Table S3. Correctly recovered transcripts, candidates, precision, recall and F-score of the assemblers on the real dataset R2 at the meta-assembly level under the HISAT2 alignments. The abbreviations ST is for StringTie2, and SC for Scallop, and assembler-0 means the results for running the assembler with the minimum filtering threshold, and assembler-default indicates the results for the default settings.

Assemblers	Correct transcripts	Candidates	precision	recall	F-score
TransMeta-0	27467	61646	0.4456	0.1701	0.2462
TransMeta-default	23538	44183	0.5327	0.1457	0.2289
PsiCLASS-0	24540	73239	0.3351	0.1519	0.2091
PsiCLASS-default	18521	37038	0.5001	0.1147	0.1866
ST+STmerge-0	24252	85673	0.2831	0.1501	0.1962
ST+STmerge-default	20344	54646	0.3723	0.1260	0.1882
SC+TACO-0	23432	132436	0.1769	0.1451	0.1594
SC+TACO-default	19441	79403	0.2448	0.1204	0.1614
gtfmerge+ST	42235	530924	0.0795	0.2615	0.1220
gtfmerge+SC	36798	417102	0.0882	0.2278	0.1272

Table S4. Correctly recovered transcripts, candidates, precision, recall and F-score of the assemblers on the real dataset R3 at the meta-assembly level under the HISAT2 alignments. The abbreviations ST is for StringTie2, and SC for Scallop, and assembler-0 means the results for running the assembler with the minimum filtering threshold, and assembler-default indicates the results for the default settings.

Assemblers	Correct transcripts	Candidates	precision	recall	F-score
TransMeta-0	26994	59914	0.4505	0.1671	0.2438
TransMeta-default	22741	40976	0.5550	0.1408	0.2246
PsiCLASS-0	26473	82978	0.3190	0.1639	0.2166
PsiCLASS-default	19775	39798	0.4969	0.1224	0.1965
ST+STmerge-0	28059	121558	0.2308	0.1737	0.1982
ST+STmerge-default	25316	79770	0.3174	0.1567	0.2098
SC+TACO-0	24182	112049	0.2158	0.1497	0.1768
SC+TACO-default	20164	71114	0.2835	0.1248	0.1734
gtfmerge+ST	42578	417259	0.1020	0.2636	0.1471
gtfmerge+SC	35343	279982	0.1262	0.2188	0.1601

Table S5. Correctly recovered transcripts, candidates, precision, recall and F-score of the assemblers on the real dataset R4 at the meta-assembly level under the HISAT2 alignments. The abbreviations ST is for StringTie2, and SC for Scallop, and assembler-0 means the results for running the assembler with the minimum filtering threshold, and assembler-default indicates the results for the default settings.

Assemblers	Correct transcripts	Candidates	precision	recall	F-score
TransMeta-0	21979	40554	0.5420	0.1361	0.2175
TransMeta-default	19571	31937	0.6128	0.1212	0.2023
PsiCLASS-0	19533	42117	0.4638	0.1209	0.1918
PsiCLASS-default	15779	27398	0.5759	0.0977	0.1670
ST+STmerge-0	22365	63491	0.3523	0.1385	0.1988
ST+STmerge-default	19324	44325	0.4360	0.1196	0.1878
SC+TACO-0	19132	72245	0.2648	0.1185	0.1637
SC+TACO-default	16008	45730	0.3501	0.0991	0.1545
gtfmerge+ST	31116	230797	0.1348	0.1926	0.1586
gtfmerge+SC	28287	206246	0.1372	0.1751	0.1538

Table S6. Correctly recovered transcripts, candidates, precision, recall and F-score of the assemblers on the real dataset R5 at the meta-assembly level under the HISAT2 alignments. The abbreviations ST is for StringTie2, and SC for Scallop, and assembler-0 means the results for running the assembler with the minimum filtering threshold, and assembler-default indicates the results for the default settings.

Assemblers	Correct transcripts	Candidates	precision	recall	F-score
TransMeta-0	29092	62779	0.4634	0.1801	0.2594
TransMeta-default	25297	45353	0.5578	0.1566	0.2446
PsiCLASS-0	26998	70739	0.3817	0.1672	0.2325
PsiCLASS-default	23067	46237	0.4989	0.1428	0.2221
ST+STmerge-0	28481	89058	0.3198	0.1763	0.2273
ST+STmerge-default	22603	47951	0.4714	0.1399	0.2158
SC+TACO-0	21649	97810	0.2213	0.1340	0.1670
SC+TACO-default	18557	66018	0.2811	0.1149	0.1631
gtfmerge+ST	39578	245046	0.1615	0.2450	0.1947
gtfmerge+SC	36155	240244	0.1505	0.2238	0.1800

Table S7 Comparison of running time and memory usage for the assemblers on the five real datasets under the HISAT2 and STAR alignments**.**

			TransMeta	Psiclass	StringTie2	Scallop
R1	HISAT2	running time(min)	133	228	188	251
		Maximum memory(MB)	11463	11975	402	5017
	STAR	running time(min)	142	91	192	233
		Maximum memory(MB)	3757	10192	245	5222
R2	HISAT2	running time(min)	140	143	292	380
		Maximum memory(MB)	29937	6808	3481	32870
	STAR	running time(min)	141	76	227	360
		Maximum memory(MB)	6924	6309	465	32563
R3	HISAT2	running time(min)	35	22	52	102
		Maximum memory(MB)	10188	4914	74	3891
	STAR	running time(min)	35	13	51	96
		Maximum memory(MB)	7532	3571	122	3993
R4	HISAT2	running time(min)	31	16	65	88
		Maximum memory(MB)	12966	2515	604	6246
	STAR	running time(min)	26	13	116	81
		Maximum memory(MB)	3140	3200	157	5632
R5	HISAT2	running time(min)	35	61	52	105
		Maximum memory(MB)	8010	2533	275	14745
	STAR	running time(min)	36	13	426	103
		Maximum memory(MB)	3691	3396	142	15052