



Boosting the power of transcriptomics by developing an efficient gene expression profiling approach

Jing Wang^{1,†} , Jun Xu^{1,2,†}, Xiaohan Yang^{1,2,†}, Song Xu^{1,2}, Ming Zhang^{1,2} and Fei Lu^{1,2,3,*} 

¹State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, The Innovative Academy of Seed Design, Chinese Academy of Sciences, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

³CAS-JIC Centre of Excellence for Plant and Microbial Science (CEPAMS), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China

Received 24 June 2021;
revised 9 August 2021;
accepted 29 August 2021.

*Correspondence (Tel 86 10 64803401; fax 86 10 64803401; email flu@genetics.ac.cn)

[†]These authors contributed equally to this work.

Keywords: 3'RNA-seq, SiPAS, transcriptomics, gene expression profiling, plant genomics.

Summary

Recent advances in plant genomics are scaling up gene expression profiling from the individual level to the population level, making transcriptomics a more powerful tool while deciphering the genome function. This study developed an efficient 3'RNA-seq method, Simplified Poly(A) Anchored Sequencing (SiPAS), to perform large-scale experiments of gene expression quantification. Aside from being cost-effective, by conducting a comprehensive performance assessment of SiPAS in hexaploid wheat, we demonstrated that SiPAS is highly sensitive, accurate, and reproducible while quantifying gene expression. Our method is anticipated to boost studies of population transcriptomics in plants and improve our understanding of genome biology.

Introduction

RNA sequencing (RNA-seq) is a keystone technology for modern biological research, shifting many genomic studies from a solely genomic level to a multi-omic level and thus effectively improving our understanding of genome biology (Stark *et al.*, 2019). Over the last few years, vast amounts of genomic data have been generated in many plant species. For example, hundreds to thousands of accessions were whole-genome sequenced to construct genetic variation maps of wheat (Zhou *et al.*, 2020), maize (Bukowski *et al.*, 2018), rice (Wang *et al.*, 2018), cassava (Ramu *et al.*, 2017), potato (Hardigan *et al.*, 2017), and soybean (Liu *et al.*, 2020), etc. Likewise, high-quality pan-genomes were also rapidly assembled in many important crops (Gao *et al.*, 2019; Hifford *et al.*, 2021; Jayakodi *et al.*, 2020; Lu *et al.*, 2015; Walkowiak *et al.*, 2020; Zhao *et al.*, 2018). The gigantic size of genomic data is creating a vacuum where a large quantity of transcriptomic data needs to be filled to help decode the function of the genome. Highly efficient RNA-seq technologies are increasingly demanded in biological research.

The emergence of 3'RNA-seq is a giant leap of RNA-seq technologies (Miyoshi *et al.*, 2008). Although the 3'RNA-seq is not capable of detecting alternative splicing isoforms when compared with conventional RNA-seq approaches, it provides alternative benefits of being cost-effective to gene expression quantification (Corley *et al.*, 2019; Tandonnet and Torres, 2017). Recently, active methodological development has been made to the 3'RNA-seq technology. Major improvements include increasing multiplexity using sample barcoding (Alpern *et al.*, 2019; Bush *et al.*, 2017; Kamitani *et al.*, 2019; Pallares

et al., 2020; Sholder *et al.*, 2020; Tzfadia *et al.*, 2018; Ye *et al.*, 2018), and reducing the cost by simplifying procedures of library preparation (Bush *et al.*, 2017; Kamitani *et al.*, 2019; Pallares *et al.*, 2020; Sholder *et al.*, 2020; Ye *et al.*, 2018). Certainly, these protocols have achieved great success; however, it is worth noting that these 3'RNA-seq methods usually use custom sequencing format and they have not been sufficiently optimized for the standard paired-end 150/250 bp (PE150 or PE250) sequencing. Although custom sequencing format [e.g. shorter and uneven read length at read 1 (R1) and read 2 (R2)] can lower sequencing cost at the benchtop scale, one crucial, often overlooked fact is that an increasing amount of the sequencing projects have been outsourced from research facilities to commercial sequencing companies. At the production scale, these companies often provide their service using the standard sequencing format because it can offset the overall cost dramatically. In other words, 3'RNA-seq using standard PE sequencing is more cost-effective, and probably performs better if the approach was optimized properly.

In this study, we developed an efficient gene expression profiling approach, Simplified Poly(A) Anchored Sequencing (SiPAS), by combing the advantages of reported 3'RNA-seq methods and optimizing the use of standard PE150 sequencing format. Through testing SiPAS for its performance in bread wheat (*Triticum aestivum* ssp. *aestivum*, $2n = 6x = 42$, genome size = 16 Gb; Appels *et al.*, 2018), we presented evidence showing that SiPAS achieves a high level of sensitivity, accuracy, and reproducibility. It is anticipated that SiPAS will boost studies of population transcriptomics of crops as well as many other plants.

Results

The design of SiPAS

Illumina paired-end sequencing allows users to sequence both ends of a template fragment and generates reads at both ends, designated as read1 (R1, ligated to P5 adapter) and read2 (R2, ligated to P7 adapter). Reported 3'RNA-seq methods apply customized and shortened paired-end sequencing (Read length: $R1 < R2 < 150$ bp) to reduce the sequencing cost, where R1 (poly(T) end) is used for barcoding and R2 (non-poly(T) end) is used for read mapping (Bush *et al.*, 2017). Given the standard PE150 sequencing format, we hypothesized that 3'RNA-seq can be further improved from three aspects. First, PE150 sequencing itself may increase the detecting power of gene expression because longer reads can lead to significant gains in read mapping accuracy (Smith *et al.*, 2008). Second, switching the sequencing adapters, specifically, using R1 (non-poly(T) end) for mapping and R2 (poly(T) end) for barcoding may be useful because R1 has a higher base quality than R2 in Illumina sequencing, and base quality has a positive effect on read mapping accuracy (Canzar and Salzberg, 2017). Third, unique molecular identifiers (UMIs) have been used actively in single-cell RNA-seq to monitor PCR amplification artefacts, which originate from the same cDNA but are sequenced multiple times due to the PCR amplification bias (Fu *et al.*, 2011, 2018; Islam *et al.*, 2014; Kivioja *et al.*, 2011; Smith *et al.*, 2017); we hypothesized that UMIs can be valuable to bulk 3'RNA-seq as well.

By combining the technical advantages of the reported 3'RNA-seq methods, we established a framework workflow of SiPAS (Figure 1a, see Methods). According to the hypotheses mentioned above, we conducted both simulation analysis (Figure S1) and wet-lab protocol tests (T1, T2, T3 and T4 shown in Figure 1b) to explore the potential improvements of 3'RNA-seq with the goal to achieve an optimal design of SiPAS.

Simulation analysis of read mapping

Both read length and base quality are key to read mapping accuracy, which is fundamental to the performance of gene expression profiling. To examine how read length affects read mapping of RNA-seq, we simulated a data set of 100 000 reads from transcript sequences derived from the wheat reference genome (IWGSC RefSeq v1.0; Appels *et al.*, 2018). These simulated reads had different lengths ranging from 50 to 150 bp. By comparing the original position and the mapping position of individual reads (Figure S1b), despite the positive effect of read length on both precision and recall, the read mapping precision appeared to be fairly high and consistent—the precision values were all greater than 0.999. In contrast, the recall values were varying substantially from 0.75 to 0.95 (Figure 2a). Similarly, another data set of 100 000 reads was simulated with different base quality scores (from 25 to 37) to examine the effect of base quality on read mapping (Figure S2). The results showed that the precision values were high and consistent as well (>0.997), but the recall values increased with base quality in a relatively big range spanning from 0.87 to 0.89. The simulation analysis indicates that both read length and base quality impact mostly on mapping sensitivity rather than mapping precision, and read length has a larger effect on mapping sensitivity than base quality does. These results also demonstrate that the read mapping precision, or specificity, is high and almost unaffected by either read length or base quality, as long as the reads are uniquely mapped to the genome.

Read mapping of protocol tests

The simulation analysis predicts that higher base quality will improve read mapping sensitivity and increase the number of uniquely mapped reads (Figure 2b); accordingly, we conducted the four protocol tests using Illumina sequencing to evaluate how switching adapters affect the base quality and unique mapping in real data, for which RNA-seq experiments of wheat samples were performed. Wheat leaves sampled at 10 am were used for RNA-seq tests with 12 technological replicates. A high proportion of uniquely mapped reads is considered to be superior because only those reads are used to quantify gene expression. By switching the adapters, R1 becomes the non-poly(T) end of reads which is used for mapping. As expected, the results showed that T2 and T4, in which the adapters were switched, exhibited the highest base quality score in the non-poly(T) end of reads (Figure 2c). Read alignment (150 bp, $n = 5$ M) using the single-end mode showed that T2 and T4 increased the proportion of uniquely mapped reads by 10.37% when compared with T1 and T3 (Figure 2d).

Although switching adapters increased the base quality of non-poly(T) end of reads for T2 and T4, it was notable that the base quality of the poly(T) end was decreased substantially (Figure 2c), which is likely due to the combined effect of poly(T) and the low quality of R2. The low-quality R2 sequence of 150 bp length set a dilemma for read mapping—according to the simulation analysis, on the one hand, low base quality may reduce mapping sensitivity. On the other hand, paired-end reads of 300 bp length can increase mapping sensitivity (Figure S3). To assess the overall effect of R^2 , we performed read alignment of the paired-end mode using both ends of reads ($n = 5$ M). The results showed that the proportion of uniquely mapped reads rose in all four tests. For T2 and T4, the uniquely mapped reads increased by 2.71% and 2.34%, reaching 84.33% and 84.29%, respectively (Figure 2d), which is consistent with that read length has a larger effect on read mapping sensitivity than base quality does as shown in the simulation analysis (Figure 2a,b). The slightly higher percentage of uniquely mapped reads in T2 than T4 is probably due to the relatively longer effective read length in the poly(T) end (Figure S4). Given the higher proportion of uniquely mapped reads of paired-end alignment, we use paired-end mode for read mapping in the following analyses.

Gene expression quantification of protocol tests

Accurate and stable quantification of gene expression is crucial to RNA-seq applications. We then investigated the effect of UMI on correcting for PCR amplification bias in bulk 3'RNA-seq. Also, we compared the four protocol tests in respect of accuracy and reproducibility of gene expression quantification.

UMIs were anchored to RNA molecules in T3 and T4 (Figure 1b), in which we evaluated the effectiveness of UMI by comparing the read count and the UMI count. By examining the 12 replicates of individual protocols, the results showed that the mean Pearson correlation coefficient (r) between the read count and UMI count was all greater than 0.999 in T3 and T4. No outliers were found to be responsible for the high level of similarity (Figure 3a,b). Meanwhile, a similar number of genes can be detected by using either read count or UMI count to quantify gene expression (Figure S5). Both lines of evidence suggest that UMI is optional for bulk 3'RNA-seq when there is a wealth of RNA molecules going through low PCR cycles (e.g. more than 0.5 μ g total RNA per sample and 12 PCR cycles).

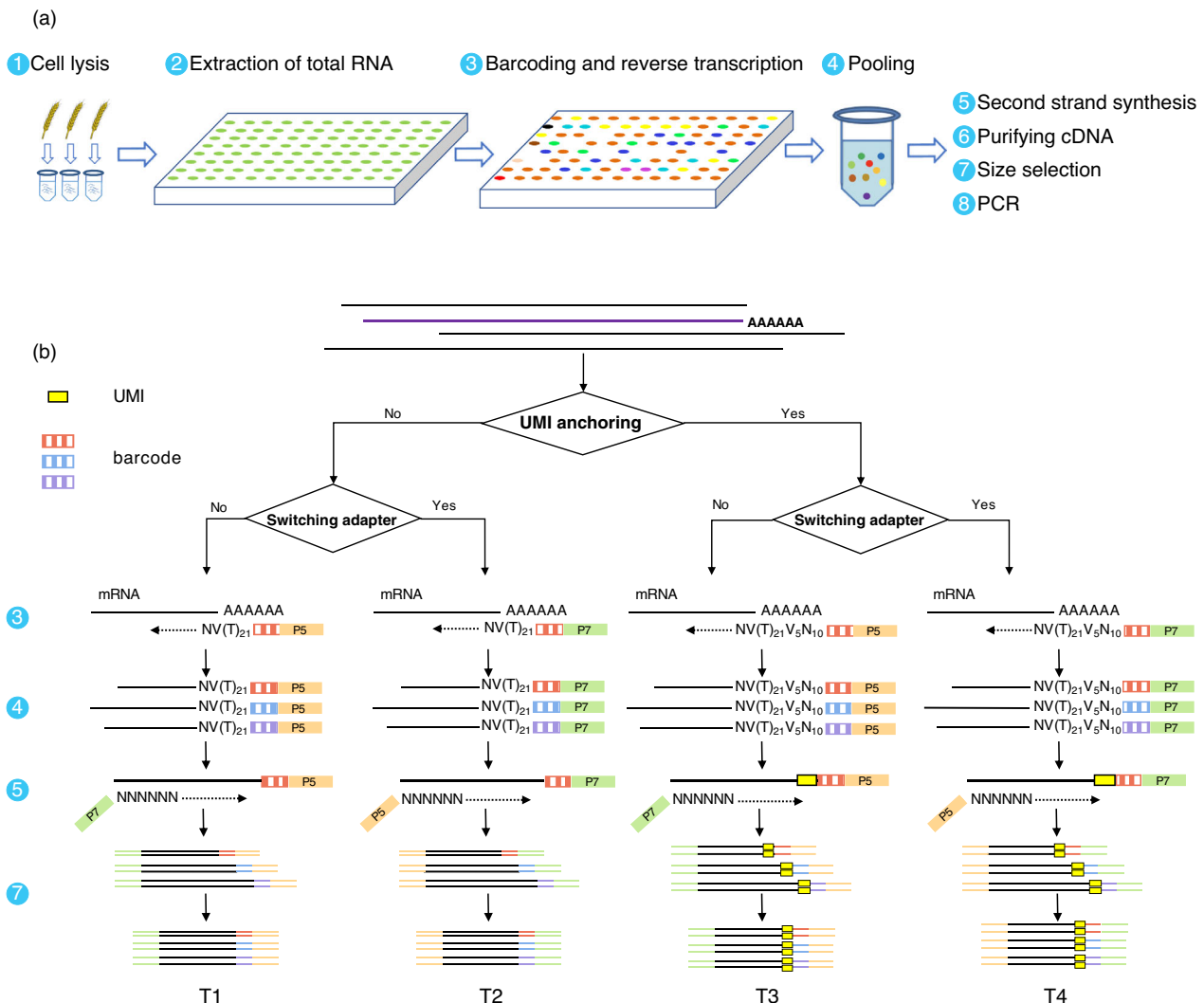


Figure 1 The design of SiPAS. (a) Framework workflow of SiPAS. For step 1, cell lysis in single tube is performed to break down the cell wall completely. For steps of 2 and 3, lysate is transferred into 96-well plates, followed by total RNA extraction, as well as barcoding and reverse transcription of total RNA using designed primers. For steps of 4–8, samples in one plate are pooled for second-strand synthesis, purifying cDNA, size selection, and PCR amplification for sequencing. (b) Optimization design of SiPAS. Four protocol tests (T1, T2, T3 and T4) were designed to assess the effect of switching adapters and using UMIs. In T1, barcodes are ligated to P5 adapter and UMIs are not used. In T2, barcodes are ligated to P7 adapter and UMIs are not used. In T3, barcodes are ligated to P5 adapter and UMIs are used. The optimal design of SiPAS can be obtained through the comparison of the four tests. In Illumina paired-end sequencing, R1 are reads with P5 adapter and R2 are reads with P7 adapter.

We used a set of standard RNA controls, External RNA Controls Consortium reference (ERCC; Lemire *et al.*, 2011; Pine *et al.*, 2016) as the 'true values' to evaluate the accuracy of gene expression quantification. ERCC has 92 molecules with known masses, which can be used to compare the concentration of individual molecules before and after RNA-seq experiments. For the purpose of comparison, we performed RNA-seq using TruSeq (Chao *et al.*, 2019; Palomares *et al.*, 2019; Sarantopoulou *et al.*, 2019) with 3 replicates on the same leaf sample used in the protocol tests. The results showed that TruSeq achieved the highest accuracy—Pearson's r between the original concentration of ERCC molecules and their predicted expression levels was higher than 3'RNA-seq approaches across different sequencing depth. In the four protocol tests, T2 outperformed the other three tests and showed slightly lower performance than TruSeq. The gap of Pearson's r between T2 and TruSeq was 0.019 on average

(Figure 3c). In addition to the accuracy, we also evaluated the reproducibility of the four protocol tests by calculating Pearson's r of the expression level of all wheat genes ($n = 107\,891$) between replicates of RNA-seq tests. T2 showed a slightly lower reproducibility than TruSeq with a gap of 0.015 in terms of Pearson's r , but outperformed the other three tests (Figure 3d).

Taken together, by simply switching adapters, T2 performed better than the rest of the protocol tests and achieved high sensitivity, accuracy, and reproducibility. Hence, we choose T2 protocol as the optimal design of SiPAS.

Performance comparison between SiPAS and TruSeq

As Illumina TruSeq has long been considered as the gold standard approach of gene expression profiling, we use the full-length transcriptome profiling method TruSeq (Chao *et al.*, 2019; Palomares *et al.*, 2019; Sarantopoulou *et al.*, 2019) to benchmark the

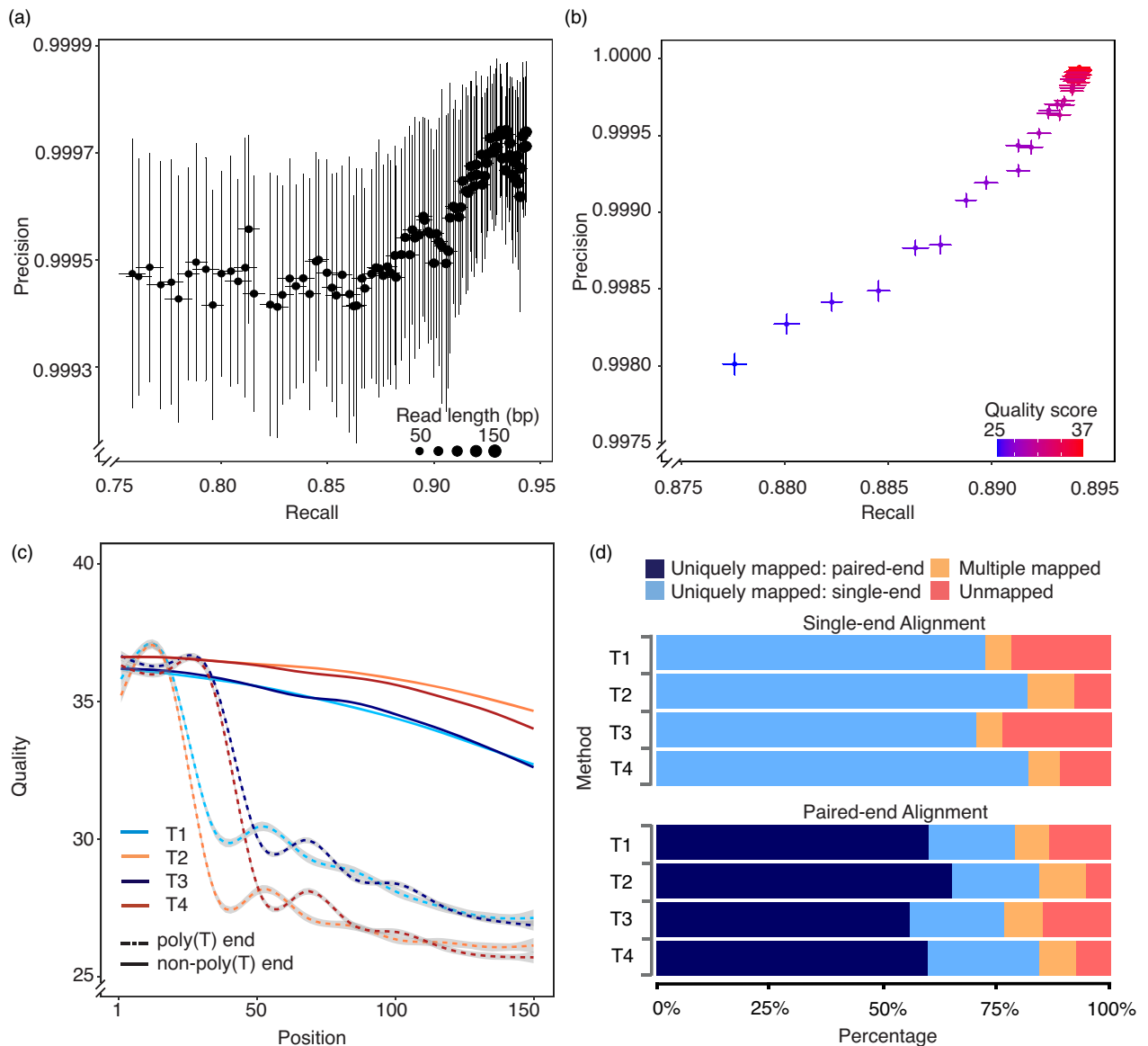


Figure 2 Read mapping performance of simulated reads and the four protocol tests. (a) Precision–recall plot of simulated data with different read length. Point represents mean and bar around point represents standard deviation (SD) of 100 replicates. The size of point corresponds to the read length. (b) Precision–recall plot of simulated data with different read quality score. The colour represents the mean quality score of reads and the bar around point represents SD of 100 replicates. (c) Read quality score of poly(T) end (dashed) and non-poly(T) end (solid) in the four tests. The shadow means 95% confidence interval. (d) Proportion of read alignment by category. Read mapping results from single-end alignment mode and paired-end alignment are shown.

newly developed SiPAS. Both TruSeq and SiPAS libraries were constructed using wheat leaves sampled at 10 am and 10 pm. Although SiPAS showed slightly lower accuracy and reproducibility than TruSeq (Figure 3c,d), the accordance between SiPAS and TruSeq increased with sequencing depth (Figure 4a). Pearson's r of gene expression level measured by the two approaches shifted from 0.84 to 0.91 when read number of individual samples increased from 1 to 12 M (Figure 4a). Given the clear diminishing return of accuracy and reproducibility (Figure 3c,d), we chose a sequencing depth of 5 M reads per sample in wheat to balance the benefit and cost of SiPAS, at the sequencing depth of which we observed a high level of agreement between TruSeq and SiPAS (Figure 4b).

Differentially expressed gene (DEG) analysis is one of the most common applications of RNA-seq. Both TruSeq and SiPAS libraries

were constructed using wheat leaves sampled at 10 am and 10 pm to identify DEGs. To allow a fair comparison, we used a sequencing depth of 5 M/replicate in both TruSeq and SiPAS. Principal component analysis (PCA) of gene expression showed that replicates from am and pm were clearly separated (Figure 4c). The two clusters (am and pm) from SiPAS were highly consistent with TruSeq. It is worth noting that the PC1, which represents the biological difference between leaf samples of am and pm, explains 78% of the total variance. Contrarily, the PC2, representing the technological difference between SiPAS and TruSeq, explains 18% of the total variance. These results suggest that SiPAS is well qualified to capture biological differences in DEG analysis.

Identifying DEGs in leaf tissue between am and pm was then performed based on three replicates from the two RNA-seq

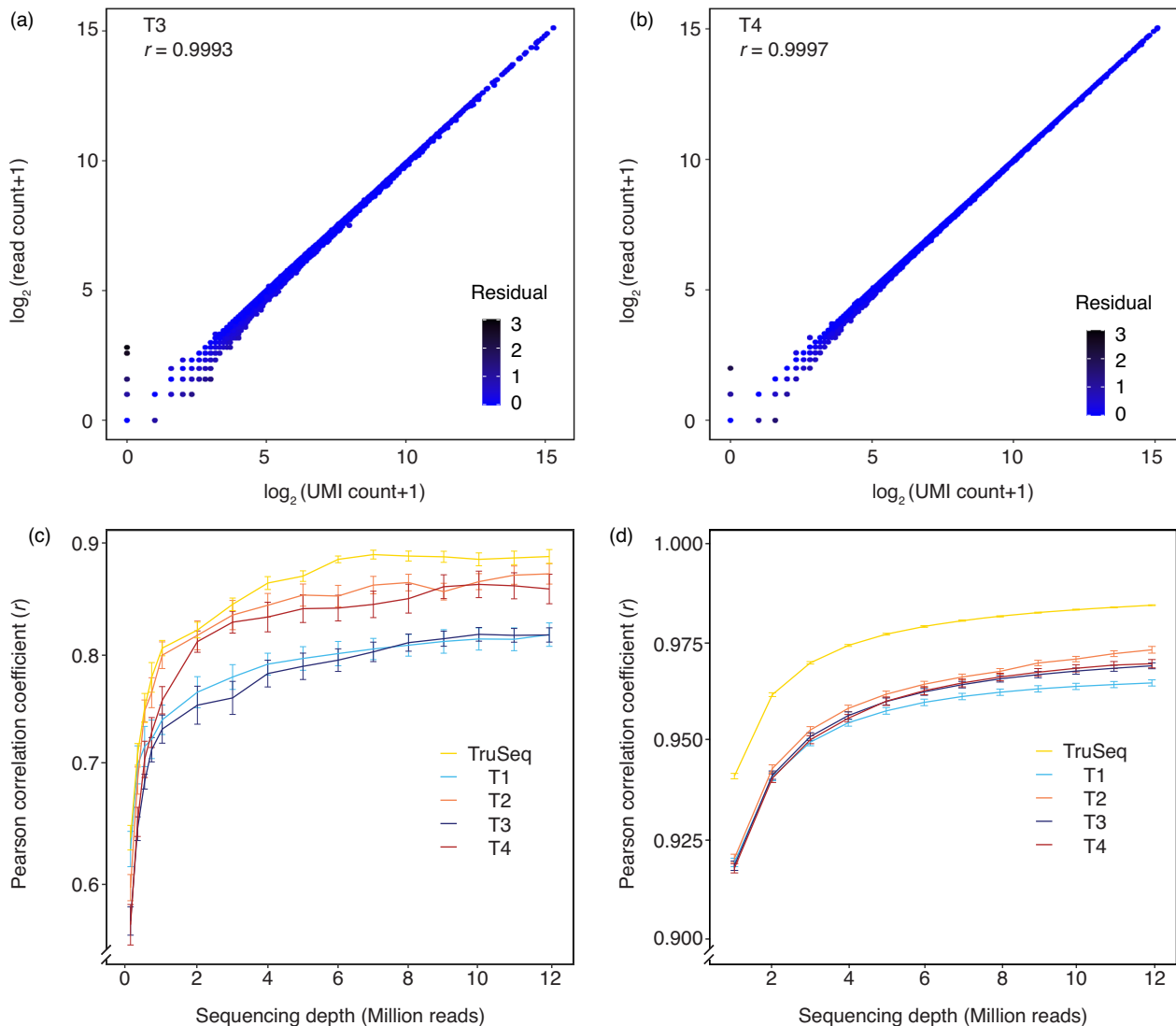


Figure 3 Gene expression quantification of protocol tests. (a) and (b) Evaluation in UMI correction for accurate reads counting in bulk RNA-seq in T3 and T4. The scatter plots show the correlation between read count and UMI count of a randomly chosen replicate from 12 replicates. Both count values are added a pseudo-count of 1 and take the logarithm. (c) Pearson correlation coefficient (r) between gene expression levels and corresponding transcripts concentrations at different subsampling reads number (TPM for TruSeq and CPM for four wet-lab protocol tests). (d) Pearson's r of wheat gene expression levels between technical replicates at different sequencing depth.

methods. By applying the same thresholds, where the absolute value of fold change of expression was greater than 2 and the false discovery rate was less than 0.05, we identified similar numbers of DEGs—a total of 6588 and 5940 DEGs were detected by TruSeq and SiPAS, respectively. A large number of DEGs ($n = 5340$) were shared between the two data sets. Pearson's r of the fold change of identified DEGs between SiPAS and TruSeq was up to 0.95, indicating that SiPAS is interchangeable with TruSeq for DEG analysis (Figure 4d).

Performance of SiPAS for degraded RNA

RNA molecules are fragile and susceptible to degradation. Therefore, RNA-Seq methods with a high tolerance for RNA degradation are favoured in high throughput transcriptomic studies. The integrity of RNA molecules, which is measured by RNA integrity number (RIN), reflects the extent of RNA degradation. To evaluate the tolerance of SiPAS to degraded RNA

molecules, we used the Mg^{++} cations to randomly fragment RNA and mimic the RNA degradation process. Compared with intact RNA (no treatment) with a RIN value of 7.4, two fragmented samples had RIN values of 6.8 and 2.3, respectively (Figure 5a). Gene expression quantification of fragmented samples showed that SiPAS was fairly robust to RNA degradation—RIN had a negligible effect on both reproducibility (Figure 5b) and accuracy (Figure 5c) of gene expression profiling using SiPAS. The high tolerance of RNA degradation ensures the ease of use of SiPAS for high-throughput RNA-Seq experiments.

Discussion

An in-depth understanding of genome function is fundamental to the precision breeding of crops. While thousands of individuals have been whole-genome sequenced for major crops (Bukowski *et al.*, 2018; Wang *et al.*, 2018; Zhou *et al.*, 2020), the fine

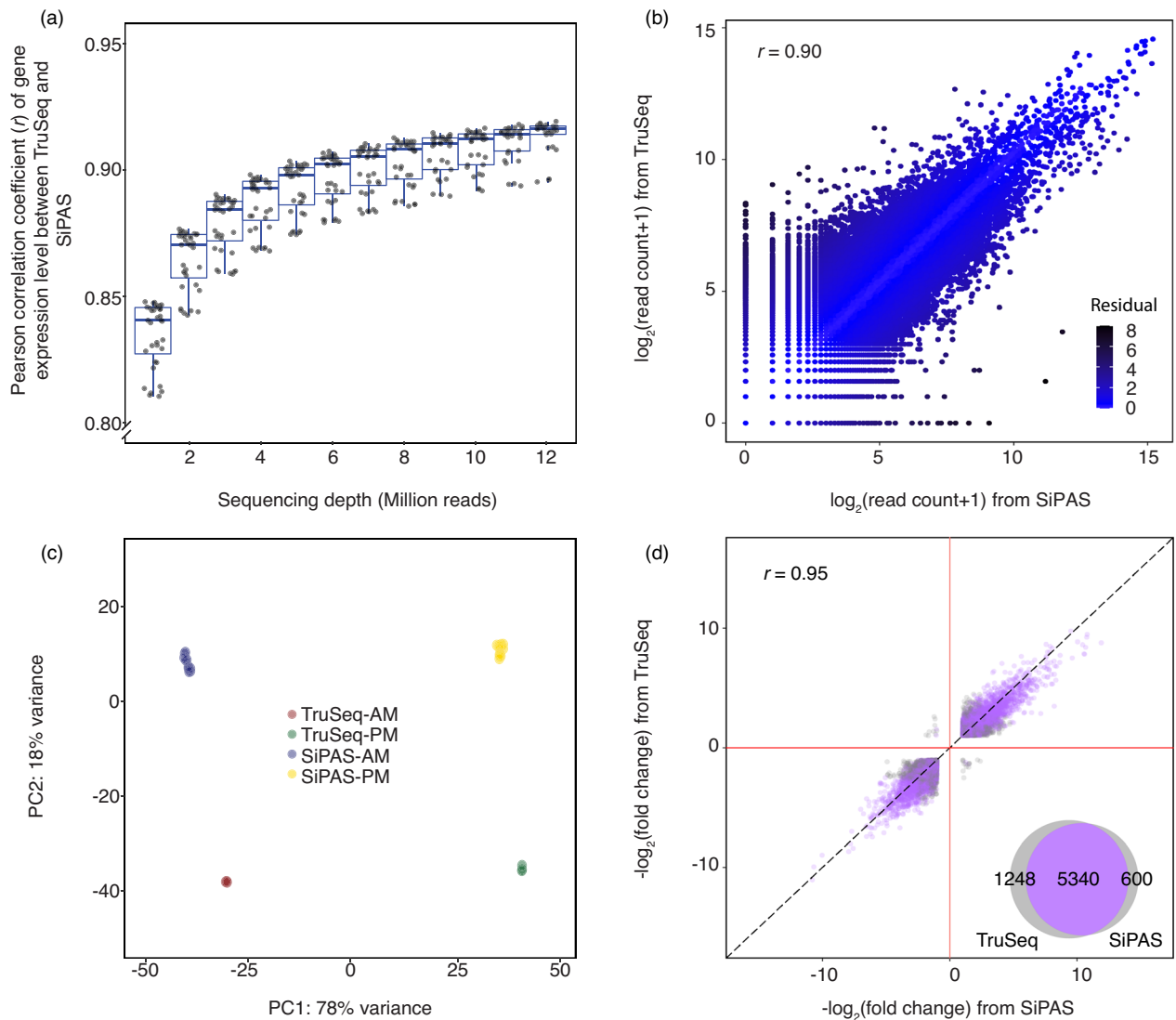


Figure 4 Comparison of SiPAS (T2) with TruSeq. (a) Correlation of gene expression levels between SiPAS and TruSeq at different sequencing depth ranging from 1 to 12 M. (b) The correlation of SiPAS and TruSeq in gene expression detection under 5 M reads. (c) PCA plot for samples constructed by SiPAS and TruSeq at 10 am and 10 pm using 5 M reads. The TruSeq and SiPAS contained 3 and 12 replicates in each condition, respectively. (d) Comparison of differentially expressed genes detection in SiPAS and TruSeq with q value < 0.05 and $|\text{Fold Change}| > 2$. Three replicates were used in 5 M read for each method of two conditions.

genotype-phenotype map is still far from complete. As a newly formed research field, population transcriptomics has shown remarkable power in addressing the genetic basis of fitness loss (Kremling *et al.*, 2018) and environmental adaptation (Groen *et al.*, 2020), as well as deciphering the regulatory machinery of gene expression (Washburn *et al.*, 2019). There is no doubt that transcriptomics at the population level will have a much broader application in plant genetics.

SiPAS, as an improved 3'RNA-seq method, provides multiple strengths to advance population transcriptomic studies in plants. First, SiPAS is effortless and cost-effective. The workflow is simplified by starting with total RNA instead of mRNA capture and bypassing RNA fragmentation without using Tn5 or Mg⁺⁺. Meanwhile, SiPAS is optimized and well suited for the standard sequencing format of Illumina (PE150). Benefiting from the simplified and standardized library construction process, the cost

of SiPAS is substantially reduced to \$2/sample (Table S2). Second, SiPAS is highly effective in quantifying gene expression. By switching P5 and P7 adapters, the read end used for alignment achieves higher base quality, resulting in increased read mapping sensitivity, and a high level of accuracy and reproducibility of gene expression quantification. Notably, for 107 891 genes in the wheat genome, only five million reads achieved Pearson's r of 0.96 between the gene expression level of two technical replicates. This suggests that SiPAS may not require technical replicates for transcriptomic analysis when the sample size is large. Third, SiPAS is robust to RNA degradation (Figure 5). This is because the 3'RNA is generally more stable than the rest of RNA sequences (Sigurgeirsson *et al.*, 2014). The high tolerance to RNA degradation lessens the variability during sample preparation and guarantees a fair comparison of gene expression between samples.

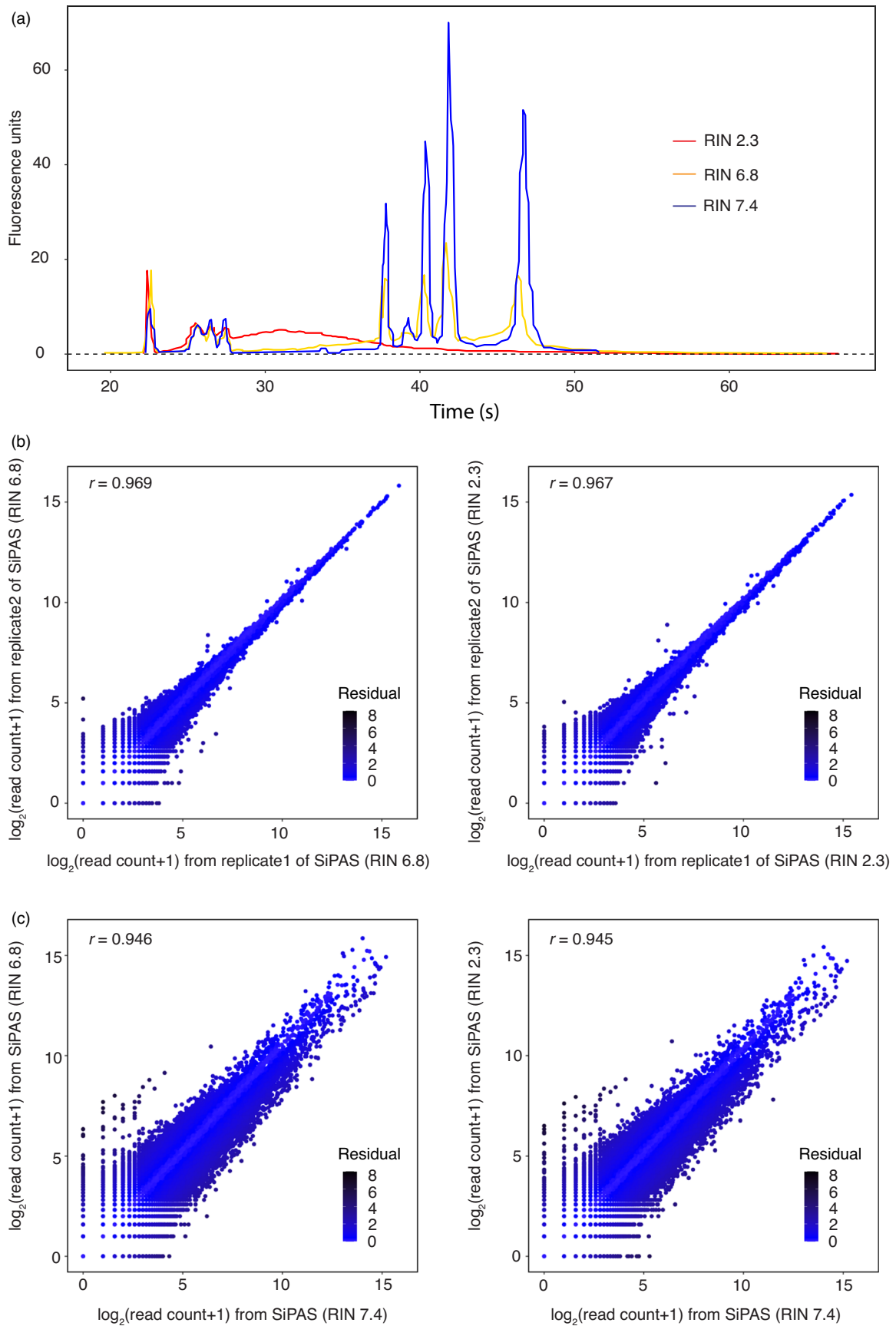


Figure 5 The performance of SiPAS on degraded RNA library. (a) RIN values obtained from Agilent 2100 Bioanalyzer System in different degrees of degraded RNA library. (b) Correlation of gene expression levels between technical replicates of degraded RNA library. (c) Correlation of gene expression levels before and after RNA degradation.

We did observe a slightly lower performance of SiPAS than TruSeq in terms of accuracy and reproducibility for gene expression quantification (Figure 3c,d). This is likely due to the fact that TruSeq has longer effective read length and higher base quality, because the barcodes and poly(T) actually reduce the read length of SiPAS; at the same time, the base quality of R2 for SiPAS decreased by the poly(T). It is also worth noting that, compared with the full-length RNA-Seq method, the accuracy of gene expression quantification of 3'RNA-Seq is more likely to be affected by the quality of gene/transcript annotation of the reference genome. However, 3'RNA-Seq methods, including SiPAS, will achieve their best performance when high-quality transcripts annotation is available for the species being studied. Overall, with the strengths of being cost-effective and labour-saving, and equivalent performance with TruSeq, SiPAS promises the ease of use of this method in large-scale population transcriptomic studies. We anticipate that SiPAS will be used in many species and contribute to an in-depth understanding of plant genomes.

Methods

Material cultivation

Chinese Spring, an accession of bread wheat (*Triticum aestivum*), was grown within growth pouches supplied by Hoagland's solution in growth chambers with 16 : 8 h of day: night length at constant 22 °C for 14 days. At 10 am (day) and 10 pm (night) on the 14th day, plant shoots at the three-leaf stage (Zadok stage 13) were selected for RNA extraction.

Workflow of SiPAS

Cell lysis

First, samples were immediately frozen in liquid nitrogen and grounded into fine powder. Then, 400 µL TRIZOL reagent (Invitrogen, # 15596018) was added into each tube and thoroughly mixed with the sample.

Extraction of total RNA

Direct-zol-96 RNA Kits (Zymo, #R2054) were used to extract the total RNA. This kit provided a mature process for high-quality RNA extraction (<https://www.zymoresearch.com/collections/direct-zol-rna-kits/products/direct-zol-96-rna-kits>).

Barcoding and reverse transcription

RNA was quantified using RNA Quantifluor (Qubit Fluorometer) and diluted to a concentration of 100 ng/µL. We transferred 5 µL of 100 ng/µL total RNA and 1 µL of 1 : 100 dilution of ERCC Ex-Fold Spike-Ins (Invitrogen, # 4456739) to each well. And then SiPAS reverse transcription (RT) primers were dispensed into each well. The RT primers for the four tested protocols are as follows, P5 adapter sequence (GTTCAGAGTTCTACAGTCCGACGATC) plus (barcode)(T)₂₁VN for T1, P5 adapter sequence plus (barcode)N₁₀V₅(T)₂₁VN for T3, P7 adapter sequence (GCCTTGGCACCCGAGAATTCCA) plus (barcode)(T)₂₁VN for T2, P7 adapter sequence plus (barcode)N₁₀V₅(T)₂₁VN for T4. The barcode sequence is supplemented in Table S2. To anneal the RT primers to poly(A) tail, the plate was heated to 94 °C for 2 min and placed immediately on ice for 5 min. RT mixture (containing 7 µL ProtoScript II Reaction Mix (2×) and 1 µL ProtoScript II Enzyme Mix (10×), NEB, #E6560L) was then dispensed into each well. The plate was cultivated on thermocycler with a programme

beginning with 25 °C for 5 min, and followed by 42 °C for 2 h and 80 °C for 5 min to end the reaction. To remove excess primers, 1 µL of 4× Exonuclease I and 4 µL Exonuclease I buffer (NEB, #M0293L) were added to each well and then incubated at 25 °C for an hour. For each well, 20 µL mixture (equal volume of 1 M NaOH and 0.5 M EDTA) was added and incubated at 65 °C for 15 min to hydrolyse RNA.

Pooling

For each plate, 10 µL of products from each well is pooled together. Then, the pooled samples were purified with QIAGEN MinElute kit (QIAGEN, #28004) according to the manufacturer's instructions and eluted with 17 µL nuclease-free water.

Second strand synthesis

For second-strand synthesis, 1 µL of 10 mM dNTP mixture (NEB, #N0447S) and 1 µL of 100 mM second strand synthesis primer (GCCTTGGCACCCGAGAATTCCANNNNNN for T1 and T3, GTTCAGAGTTCTACAGTCCGACGATCANNNNNN for T2 and T4) were added to the pooled cDNA. The mixture was then heated to 70 °C for 2 min and immediately placed on ice for 5 min. 2 µL of NEB Buffer 2 (NEB, #B7002S) and 1 µL of Klenow large fragment DNA polymerase (NEB, # M0210L) were added. Then, the plate was incubated at 37 °C for 30 min. The reaction was stopped by the addition of EDTA to a final concentration of 50 µM. AMPure XP beads (Beckman Counter, #A63881) were added to the reaction product at a 1 : 1 bead-to-sample ratio to perform clean-up.

Size selection

Size selection was performed to pooling target DNA using AMPure XP beads (insert DNA length is about 350 bp).

PCR

The double-stranded cDNA pool was amplified by PCR using NEBNext Ultra II Q5 Master Mix (NEB, #M0544L) with 0.5 µM Illumina RP1 primer and Illumina RPIx primer (where x is a number indicating the Illumina index). The thermocycling protocol began with 98 °C for 30 s followed by 12 cycles of 98 °C for 15 s, 62 °C for 15 s, and 72 °C for 60 s, and then incubated at 72 °C for 7 min. Three replicates of Illumina TruSeq library (Illumina, #20020594) for am or pm respectively were constructed starting with the same total RNA and ERCC Ex-Fold Spike-Ins with SiPAS libraries following the manufacturer's instructions. The libraries were sequenced using the Illumina NovoSeq platform with PE150 sequencing format.

Simulation analysis

We simulated two data set of reads to evaluate the effect of read length and base quality on the accuracy of read mapping. Each of the data sets consisted of 100 000 reads from transcript sequences derived from the wheat reference genome (Appels *et al.*, 2018). Reads in the data sets were set to have 1% sequence difference with the reference transcripts to mimic natural genetic diversity. For the first data set, these simulated reads had different read length ranging from 50 to 150 bp. For the second data set, these reads had the same read length of 150 bp, but different mean base quality from 25 to 37. By comparing the mapping position and known position of the simulated reads, we were able to access how read length and base quality affect read mapping accuracy.

Bioinformatics of gene expression quantification

We developed a pipeline, SiPAS-Profiler, to perform fastq file parsing, read alignment, and read counting for SiPAS (<https://github.com/PlantGeneticsLab/SiPAS-Profiler>). The detailed bioinformatic process of SiPAS-Profiler is as follows.

Parsing fastq files

SiPAS is a highly multiplexing approach. Each sample has its unique barcode which would be used for parsing samples. The barcodes are at the beginning of R2. A total of 12 barcodes with eight bases for each were applied as identifiers of samples through experiments. We built a HashMap of barcode-and-sample in SiPAS-Profiler to parse fastq files.

Read alignment

The splice-aware STAR (Dobin *et al.*, 2013) aligner was used to align reads against the wheat reference genome IWGSC 1.0 (Appels *et al.*, 2018), allowing a read to map to at most 10 locations (`--outFilterMultimapNmax 10`) with at most 10% mismatches (`--outFilterMismatchNoverLmax 0.1`), while filtering out all non-canonical intron motifs (`--outFilterIntronMotifs RemoveNoncanonicalUnannotated`). The minimum mapped read length was set to 80 (`--outFilterMatchNmin 80`).

Counting reads

HTSeq (Anders *et al.*, 2015) was used for read counting. Default settings of intersection-nonempty mode from HTSeq32 v.0.11.1 were used to obtain gene-level read counts from the resulting BAM files.

PCA

A total of 30 gene expression data sets (12 replicates of SiPAS and three replicates of TruSeq in both am and pm conditions) were used for the PCA. We performed PCA using the method of Sparse PCA implemented in DESeq2. Gene expression levels of individual genes are predictors.

DEG analysis

The gene expression data from HTSeq were directly used for clustering using DESeq2 (Love *et al.*, 2014). The genes with $|\text{Fold Change}| > 2$ between am and pm conditions in both methods were selected for analysis. Then, the q value < 0.05 was used to as the threshold of significance.

Accession numbers

The raw RNA-seq data reported in this article have been deposited in the Genome Sequence Archive (<https://ngdc.cncb.ac.cn/gsa/>) under accession numbers CRA004293.

Acknowledgements

This research was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA24020201 and XDA24040102) and the National Natural Science Foundation of China (31970631 and 31921005). We thank Changbin Yin, Aoyue Bi, Daxing Xu, Xuebo Zhao, Zhiliang Zhang, Lipeng Kang, and Jijin Zhang for their suggestions.

Conflict of interest

The authors declare no conflict of interests.

Author contributions

Experimental methodology for SiPAS was developed by J.W. J.W. and X.Y. performed wet experiments. Data processing, analysis, and supporting pipeline were completed by J.X. and X.Y. J.W. wrote the manuscript. F.L., X.Y., J.X., and J.W. revised the manuscript. S.X. assisted simulation analysis, and M.Z. assisted RNA extraction. F.L. designed and supervised the research.

References

- Alpern, D., Gardeux, V., Russeil, J., Mangeat, B., Meireles-Filho, A.C.A., Breyse, R., Hacker, D. *et al.* (2019) BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* **20**, 71.
- Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
- Appels, R., Eversole, K., Feuillet, C., Keller, B., Rogers, J., Stein, N., IWGSC whole-genome assembly principal investigators *et al.* (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.
- Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B.O. *et al.* (2018) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, 1–12.
- Bush, E.C., Ray, F., Alvarez, M.J., Realubit, R., Li, H., Karan, C., Califano, A. *et al.* (2017) PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens. *Nat. Commun.* **8**, 105.
- Canzar, S. and Salzberg, S.L. (2017) Short read mapping: an algorithmic tour. *Proc. IEEE*, **105**, 436–458.
- Chao, H.-P., Chen, Y., Takata, Y., Tomida, M.W., Lin, K., Kirk, J.S., Simper, M.S. *et al.* (2019) Systematic evaluation of RNA-Seq preparation protocol performance. *BMC Genom.* **20**, 1–20.
- Corley, S.M., Troy, N.M., Bosco, A. and Wilkins, M.R. (2019) QuantSeq: 3' sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. *Sci. Rep.* **9**, 18895.
- Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
- Fu, G.K., Hu, J., Wang, P.H. and Fodor, S.P.A. (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl Acad. Sci. USA*, **108**, 9026–9031.
- Fu, Y., Wu, P.H., Beane, T., Zamore, P.D. and Weng, Z. (2018) Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genom.* **19**, 1–14.
- Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A. *et al.* (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.* **51**, 1044–1051.
- Groen, S.C., Čalić, I., Joly-Lopez, Z., Platts, A.E., Choi, J.Y., Natividad, M., Dorph, K. *et al.* (2020) The strength and pattern of natural selection on gene expression in rice. *Nature*, **578**, 572–576.
- Hardigan, M.A., Laimbeer, F.P.E., Newton, L., Crisovan, E., Hamilton, J.P., Vaillancourt, B., Wiegert-Rininger, K. *et al.* (2017) Genome diversity of tuber-bearing Solanum uncovers complex evolutionary history and targets of domestication in the cultivated potato. *Proc. Natl. Acad. Sci. USA*, **114**, E9999–E10008.
- Hufford, M.B., Seetharam, A.S., Woodhouse, M.R., Chougule, K.M., Ou, S., Liu, J., Ricci, W.A. *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655–662.
- Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnberg, P. *et al.* (2014) Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*, **11**, 163–166.
- Jayakodi, M., Padmarasu, S., Haberer, G., Bonthala, V.S., Gundlach, H., Monat, C., Lux, T. *et al.* (2020) The barley pan-genome reveals the hidden legacy of mutation breeding. *Nature*, **588**, 284–289.
- Kamitani, M., Kashima, M., Tezuka, A. and Nagano, A.J. (2019) Lasy-Seq: a high-throughput library preparation method for RNA-Seq and its application in the analysis of plant responses to fluctuating temperatures. *Sci. Rep.* **9**, 1–14.

- Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*, **9**, 72–74.
- Kremling, K.A.G., Chen, S.-Y., Su, M.-H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F. et al. (2018) Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, **555**, 520–523.
- Lemire, A., Lea, K., Batten, D., Jian Gu, S., Whitley, P., Bramlett, K. and Qu, L. (2011) Development of ERCC RNA spike-in control mixes. *J. Biomol. Technol.* **22(Suppl)**, S46.
- Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.A. et al. (2020) Pan-genome of wild and cultivated soybeans. *Cell*, **182**, 162–176.e13.
- Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.
- Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Y. et al. (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun.* **6**, 6914.
- Miyoshi, T., Kanoh, J., Saito, M. and Ishikawa, F. (2008) Fission yeast pot1-Tpp1 protects telomeres and regulates telomere length. *Science*, **320**, 1341–1344.
- Pallares, L.F., Picard, S. and Ayroles, J.F. (2020) Tm3' Seq: a tagmentation-mediated 3' sequencing approach for improving scalability of RNAseq experiments. *G3 Genes Genomes Genet.* **10**, 143–150.
- Palomares, M.-A., Dalmasso, C., Bonnet, E., Derbois, C., Brohard-Julien, S., Ambroise, C., Battail, C. et al. (2019) Systematic analysis of TruSeq, SMARTer and SMARTer ultra-low RNA-seq kits for standard, low and ultra-low quantity samples. *Sci. Rep.* **9**, 1–12.
- Pine, P.S., Munro, S.A., Parsons, J.R., McDaniel, J., Lucas, A.B., Lozach, J., Myers, T.G. et al. (2016) Evaluation of the external RNA controls consortium (ERCC) reference material using a modified Latin square design. *BMC Biotechnol.* **16**, 1–15.
- Ramu, P., Esuma, W., Kawuki, R., Rabbi, I.Y., Egesi, C., Bredeson, J.V., Bart, R.S. et al. (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nat. Genet.* **49**, 959–963.
- Sarantopoulou, D., Tang, S.Y., Ricciotti, E., Lahens, N.F., Lekkas, D., Schug, J., Guo, X.S. et al. (2019) Comparative evaluation of RNA-Seq library preparation methods for strand-specificity and low input. *Sci. Rep.* **9**, 1–10.
- Sholder, G., Lanz, T.A.A., Moccia, R., Quan, J., Aparicio-Prat, E., Stanton, R. and Xi, H.S.S. (2020) 3'Pool-seq: an optimized cost-efficient and scalable method of whole-transcriptome gene expression profiling. *BMC Genom.* **21**, 64.
- Sigurgeirsson, B., Emanuelsson, O. and Lundeberg, J. (2014) Sequencing degraded RNA addressed by 3' tag counting. *PLoS One*, **9**(3), e91851.
- Smith, A.D., Xuan, Z. and Zhang, M.Q. (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinform.* **9**, 128.
- Smith, T., Heger, A. and Sudbery, I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* **27**, 491–499.
- Stark, R., Grzelak, M. and Hadfield, J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656.
- Tandonnet, S. and Torres, T.T. (2017) Traditional versus 3' RNA-seq in a non-model species. *Genomics Data*, **11**, 9–16.
- Tzfadia, O., Bocobza, S., Defoort, J., Almekias-Siegl, E., Panda, S., Levy, M., Storme, V. et al. (2018) The 'TranSeq' 3'-end sequencing method for high-throughput transcriptomics and gene space refinement in plant genomes. *Plant J.* **96**, 223–232.
- Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M.T., Brinton, J., Ramirez-Gonzalez, R.H. et al. (2020) Multiple wheat genomes reveal global variation in modern breeding. *Nature*, **588**, 277–283.
- Wang, W., Mauleon, R., Hu, Z., Chebotarov, D., Tai, S., Wu, Z., Li, M. et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.
- Washburn, J.D., Mejia-Guerra, M.K., Ramstein, G., Kremling, K.A., Valluru, R., Buckler, E.S. and Wang, H. (2019) Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl Acad. Sci. USA*, **116**, 5542–5549.
- Ye, C., Ho, D.J., Neri, M., Yang, C., Kulkarni, T., Randhawa, R., Henault, M. et al. (2018) DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun.* **9**, 4307.
- Zhao, Q., Feng, Q., Lu, H., Li, Y., Wang, A., Tian, Q., Zhan, Q. et al. (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat. Genet.* **50**, 278–284.
- Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D. et al. (2020) Triticum population sequencing provides insights into wheat adaptation. *Nat. Genet.* **52**, 1412–1422.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1 Workflow of simulation.

Figure S2 Quality score of simulated reads.

Figure S3 Precision and recall of read mapping in SE and PE modes.

Figure S4 Effective read length distribution of reads from four protocol tests.

Figure S5 Comparison of read count and UMI count for gene expression detection across different expression levels.

Table S1 Barcode of SiPAS RT primers

Table S2 Cost of SiPAS library preparation