Plant Biotechnology Journal
Open Access

Plant Biotechnology Journal (2023) **21**, pp. 2241–2253 doi: 10.1111/pbi.14126

Affordable, accurate and unbiased RNA sequencing by manual library miniaturization: A case study in barley

Christopher Arlt[1], Thorsten Wachtmeister[2], Karl Köhrer[2] and Benjamin Stich[1,3,4,*,a]

[1]*Institute of Quantitative Genetics and Genomics of Plants, Heinrich Heine University Duesseldorf, Duesseldorf, Germany*
[2]*Genomics & Transcriptomics Laboratory, Biological and Medical Research Centre (BMFZ), Heinrich Heine University Duesseldorf, Duesseldorf, Germany*
[3]*Cluster of Excellence on Plant Sciences (CEPLAS), Duesseldorf, Germany*
[4]*Max Planck Institute for Plant Breeding Research, Cologne, Germany*

Received 16 November 2022;
revised 12 May 2023;
accepted 1 July 2023.
**Correspondence* (Tel: +49 38209 45201; Fax: +49 38209 45222; email Benjamin.Stich@julius-kuehn.de)
[a]Present address: Institute for Breeding Research on Agricultural Crops, Julius Kühn Institute (JKI) - Federal Research Centre for Cultivated Plants, Sanitz, Germany

Keywords: methods and techniques, RNA sequencing, library preparation, miniaturization, plant breeding, plant genetics.

Summary

We present an easy-to-reproduce manual miniaturized full-length RNA sequencing (RNAseq) library preparation workflow that does not require the upfront investment in expensive lab equipment or long setup times. With minimal adjustments to an established commercial protocol, we were able to manually miniaturize the RNAseq library preparation by a factor of up to 1:8. This led to cost savings for miniaturized library preparation of up to 86.1% compared to the gold standard. The resulting data were the basis of a rigorous quality control analysis that inspected: sequencing quality metrics, gene body coverage, raw read duplications, alignment statistics, read pair duplications, detected transcripts and sequence variants. We also included a deep dive data analysis identifying rRNA contamination and suggested ways to circumvent these. In the end, we could not find any indication of biases or inaccuracies caused by the RNAseq library miniaturization. The variance in detected transcripts was minimal and not influenced by the miniaturization level. Our results suggest that the workflow is highly reproducible and the sequence data suitable for downstream analyses such as differential gene expression analysis or variant calling.

Introduction

Next-generation sequencing (NGS) technologies have been evolving rapidly in the last two decades and continue to do so (Mardis, 2011; McCombie *et al.*, 2019). NGS can be used for a variety of different applications and is nowadays an integral part of many genetic research projects. This became possible in part by steadily decreasing sequencing costs (Wetterstrand, 2021).

This development not only enhanced the possibilities of whole genome sequencing (Auton *et al.*, 2015; Chung *et al.*, 2017; Harris and Wailan Alexander, 2021; Linderman *et al.*, 2016) but also mRNA sequencing projects (Li, 2021; Stark *et al.*, 2019). Because the transcriptome size is relatively consistent between species, the sequencing costs for species with large genomes benefit greatly by focusing on the protein-coding part of the genome. Additionally, targeting only the mRNA for sequencing is a useful complexity reduction when investigating the genotype–phenotype relationship (Jehl *et al.*, 2021; Piskol *et al.*, 2013; Shomroni *et al.*, 2022; Wang *et al.*, 2021). The relatively low complexity of mRNA libraries and the increased read output of large sequencing platforms expanded the multiplexing potential in RNA sequencing (RNAseq) projects allowing for 384+ samples to be pooled and sequenced in the same sequencing reaction. This results in a cost distribution shift, making the library preparation step the most expensive part of many RNAseq projects. The pressure to reduce the costs of this step is therefore rising and many approaches have been developed to do exactly that (Alpern *et al.*, 2019; Bagnoli *et al.*, 2018; Foley *et al.*, 2019; Hashimshony *et al.*, 2016; Hou *et al.*, 2015; Islam *et al.*, 2012; Kumar *et al.*, 2012; Pallares *et al.*, 2019; Picelli *et al.*, 2013; Shishkin *et al.*, 2015). One way to save costs during the preparation of RNAseq libraries is to switch from commercial protocols to previously published custom protocols. The latter often implement novel techniques to optimize the procedure and save costs. If these approaches are successful enough, commercial adaptations are developed, as was the case with the examples mentioned below.

The traditional protocols create RNAseq libraries using full-length mRNA molecules (Hou *et al.*, 2015; Islam *et al.*, 2012; Kumar *et al.*, 2012; Picelli *et al.*, 2013; Shishkin *et al.*, 2015). A more cost-efficient alternative is to create libraries of the 3′ or 5′ end of the mRNA exclusively (Foley *et al.*, 2019; Macosko *et al.*, 2015; Pallares *et al.*, 2019; Vahrenkamp *et al.*, 2019). Some protocols employ early multiplexing to further reduce hands-on time and costs within the prime end enriched library preparation methods (Alpern *et al.*, 2019; Bagnoli *et al.*, 2018; Hashimshony *et al.*, 2016; Soumillon *et al.*, 2014). In most full-length mRNA protocols, the multiplex bar code is part of the adapter sequence which is added to the library fragments late in the library preparation workflow and the sample pooling is conducted after amplification and clean-up. When utilizing early multiplexing, a unique bar code is added to the sequences in one of the initial steps of the protocol. This enables early multiplexing and reduces the number of samples handled during the remaining library preparation steps. While both strategies are a good way to reduce costs, they limit the application of the resulting data for further analyses for example, genomic variant calling or novel transcript identification. Additionally, early multiplexing strategies make it impossible to re-sequence individual samples.

A different approach is most commonly known as miniaturization. It involves the reduction of the utilized reagent volume during the library preparation using commercial protocols. Most

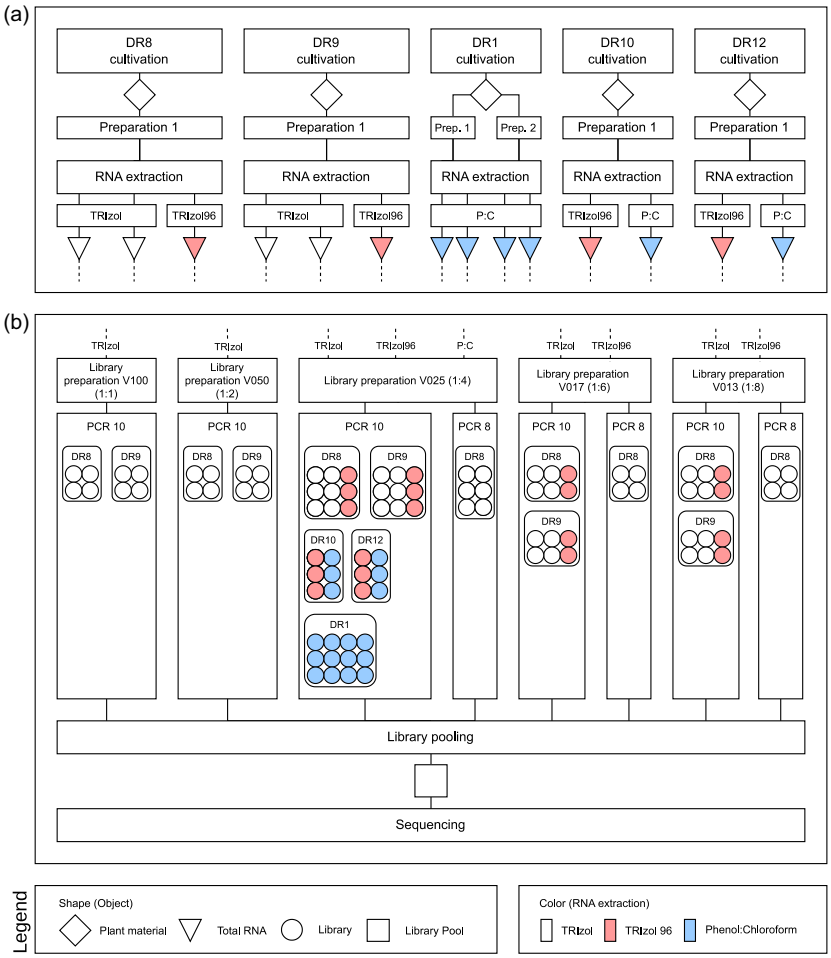

Figure 1 Overview of experimental design and miniaturization levels of the 96 samples. (a) Five different recombinant inbred lines (RIL) were cultivated and harvested (diamond). For all RILs except DR1, a single plant material preparation was used for one or more total RNA extractions using different methods (white: TRIzol, red: TRIzol-96, blue: Phenol:Chrloroform (P:C)), resulting in a total of 14 different total RNA samples (triangle). The DR1 plants were used to create coarse (Prep. 1) and fine (Prep. 2) ground plant material. (b) We tested the library preparation miniaturization levels 1:1 (V100, original), 1:2 (V050), 1:4 (V025), 1:6 (V017) and 1:8 (V013). The PCR cycles were reduced from 10 (PCR 10) to 8 (PCR 8) for a subset of the samples. For each RIL in each miniaturization and number of PCR cycles, 2–3 RNA extraction and library preparation replicates were created. In total, 96 RNA sequencing libraries were created (circles) and combined to a single library pool (square).

described workflows use full-length mRNA protocols combined with liquid handling automatization (e.g. Jaeger *et al.*, 2020; Mayday *et al.*, 2019; Mildrum *et al.*, 2020; Mora-Castilla *et al.*, 2016). Adding automatization to the miniaturization workflow has two major advantages. First, all automated preparation steps reduce hands-on time and therewith labour costs. Second, the inherent reduction in sample-to-sample variance by replacing the less error-prone hands-on steps increases the level of accuracy and precision (Tegally *et al.*, 2020). However, the investment costs of acquiring all the lab equipment required for an automated library preparation workflow are high and, thus, not feasible for many research groups. To our knowledge, with the exception of Li *et al.* (2019), who miniaturized the DNA library preparation of *E. coli* genomes, no studies are available on the capabilities of manual miniaturization.

While it is possible to reduce the costs of the library preparation step in many different ways as was outlined above, it is crucial that the quality of the resulting data sets is not impaired and has no negative impact on downstream analyses (Aigrain *et al.*, 2016; Alberti *et al.*, 2014; Dabney and Meyer, 2012; McNulty *et al.*, 2020; Romero *et al.*, 2014). However, to the best of our knowledge, no comprehensive characterization of library complexity and biases was performed for manual library preparation miniaturizations.

Here we present an easy-to-reproduce, manual miniaturized full-length mRNA sequencing library preparation workflow that does not require the upfront investment in expensive lab equipment. A miniaturization level of up to 1:8 was tested which reduced the library preparation costs significantly. In addition, we provide the results of a wide set of quality control analyses, evaluating the impact of the miniaturization on the resulting sequencing data.

Results

To evaluate the success of the library preparation miniaturization workflow, 96 samples were prepared without miniaturization (1:1, V100) and the miniaturization levels 1:2 using 50% of all reagents (V050), 1:4 using 25% of all reagents (V025), 1:6 using 17% of all reagents (V017) and 1:8 using 13% of all reagents (V013) (Figure 1). Five different genotypes from three different recombinant inbred line (RIL) populations were used to evaluate the miniaturization workflow. Two of the five RIL (DR8 and DR9) were included in all miniaturization levels to allow orthogonal comparisons. DR10 and DR12 were prepared using different RNA extraction methods and DR1 was used to analyse the potential impact of plant material coarseness on the workflow. The sequencing results were analysed with regards to library quality and its properties in common downstream analyses.

RNA extraction and library pool

All total RNA samples except the TRIzol-96 RNA extractions were evaluated using the Fragment Analyzer. The average RNA quality

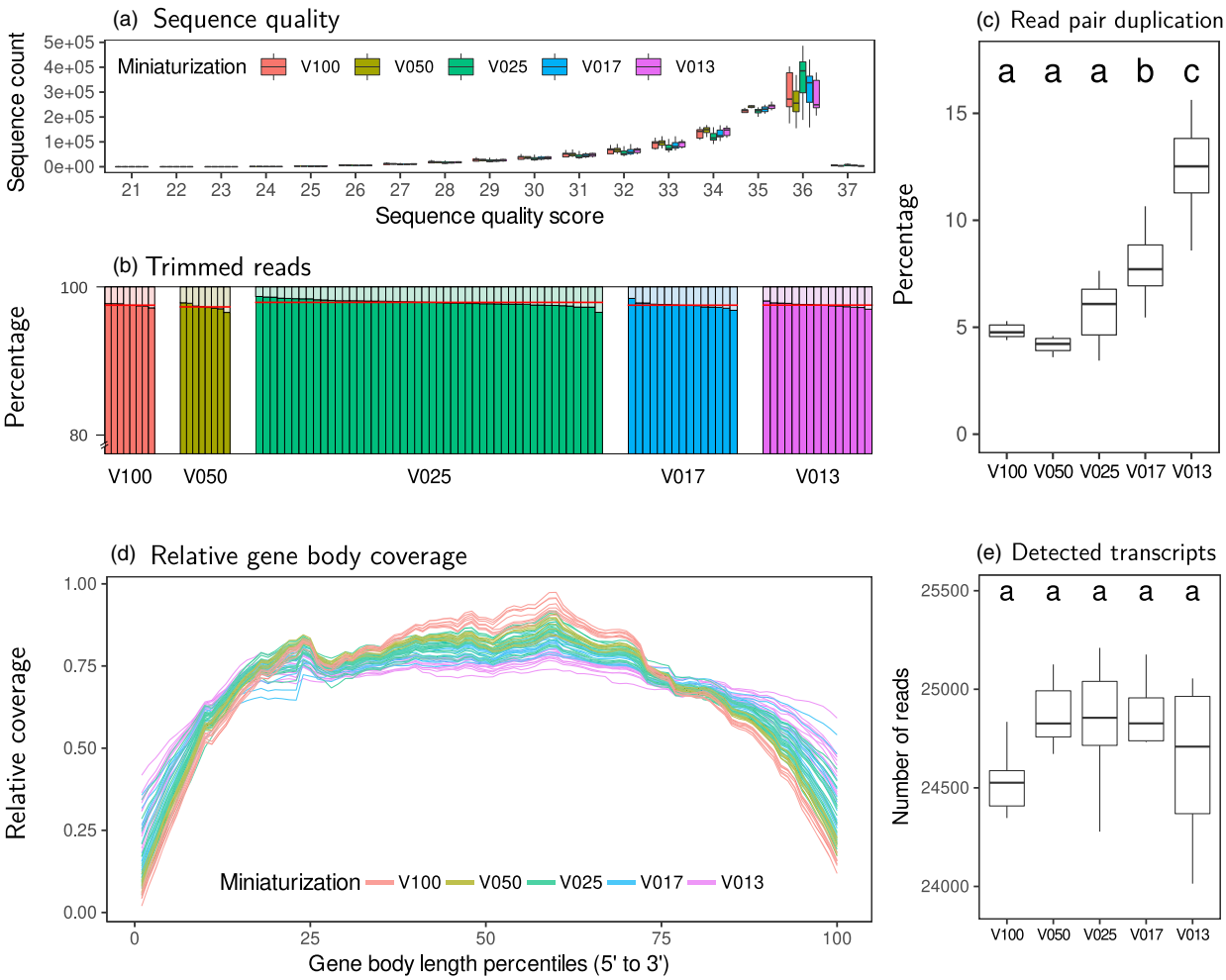

Figure 2 Overview of sequencing and library quality metrics. (a) Mean per base quality score for all 94 samples coloured by miniaturization level. (b) Percentage of reads after trimming (dark colour) per miniaturization level. The mean rate of remaining reads per miniaturization level shown as red line. (c) Total percentage of read pair duplicates by miniaturization. Miniaturization marked by the same letter are not significantly ($\alpha = 0.05$) different from each other. (d) Gene body coverage for the highly expressed genes (>90 transcript expression quantile) shown as percentage of reads located in each of the 100 gene segments starting at the 5′ end across all 94 samples coloured by miniaturization. (e) Number of unique transcripts detected by miniaturization.

number (RQN) score for the Phenol:Chloroform extraction method was the lowest with an average of 5.1 (coefficient of variation (CV) 0.13). The TRIzol extractions have an average RQN score of 7.4 (CV 0.04) (Figure S1). We also used the Fragment Analyzer to characterize the size distribution of the final library pool that was sequenced. While the fragments were with a peak at 392 bp smaller than aimed for, the size distribution itself was as expected (Figure S2).

Sequencing and library quality assessment

To evaluate the impact of the miniaturization on the sequencing process itself, we compared the mean per sequence quality score of all reads and did not find any difference between samples and miniaturizations in that regard (Figure 2a). The same was true for the per base *n* count, sequence length distribution, the per base sequence content and the adapter content (Figure S3). In addition, no negative trend in the trimming rates was observed between miniaturizations (Figure 2b). The mean read pair duplication rate was 7.1%. The highest read pair duplication ratio and significantly ($P < 0.001$) different from the remaining miniaturization levels in both RIL was observed for V013 (12.9%) (Figure 2c). For one of the two RIL, the miniaturization level V017 was significantly ($P < 0.001$) different from the rest. We assessed whether a general 5′ gene body coverage bias existed in our data set, but could not find one. The distribution did also not differ between miniaturizations (Figure 2d). Random RNA fragmentation was tested by comparing the nucleotide compositions around the fragmentation site for all miniaturization levels (Figure S4). Significant ($P < 0.001$) differences between un-miniaturized samples and miniaturized samples were observed. However, these were not consistent between RILs DR8 and DR9 with the exception of eight positions comparing the miniaturization levels V100 and V013.

The number of transcripts detected per sample ranged between 24 500 and 25 000 and did not significantly ($P > 0.05$) differ between miniaturizations (Figure 2e) when comparing random sub samples with 2 million reads. Pearson correlation coefficients of the read counts between pairs of miniaturization levels were calculated for DR8 and DR9. For both RIL, the highest similarity was observed between the V025 and

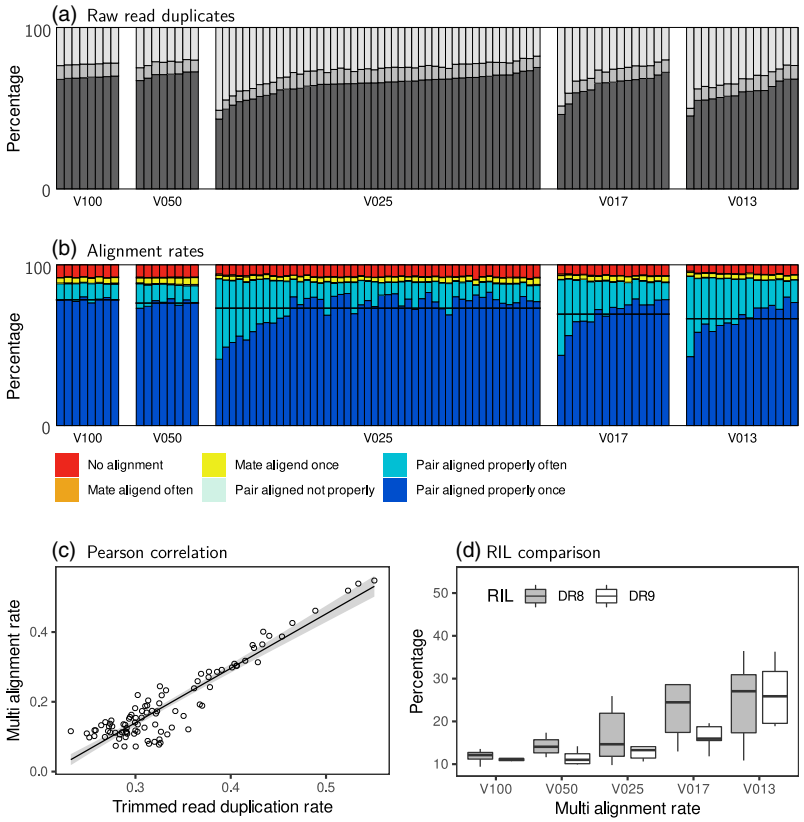

Figure 3 Correlation between read duplication rate and multi-alignment rate. (a) Trimmed read duplication rate of a subset with 1 million reads of all 94 samples coloured by uniqueness and grouped by miniaturization (Unique reads: bottom, dark grey; unique duplicated reads: middle, grey; remaining duplicated reads: top, light grey). (b) Alignment statistics for each sample grouped by miniaturization and coloured by alignment type. (c) Pearson correlation between multi-alignment rate and trimmed read duplication rate. (d) Comparison of multi-alignment rate between recombinant inbread lines (RIL) DR8 and DR9 for each miniaturization level.

V050 samples (DR8: $r = 0.9977$; DR9: $r = 0.9964$) (Table S1). The least similar sample groups were V100-V013 for DR8 ($r = 0.9613$) and V100-V017 for DR9 ($r = 0.9655$). We detected a high similarity between the read counts of replicates. The Pearson correlation coefficients for DR8 and DR9 were calculated for each miniaturization and replication type separately. The library replicates ranged from 0.999 to 0.977 and for RNA extraction replicates from 0.999 to 0.991 (Table S2).

While we focused on evaluating the impact of the miniaturization on the libraries, we also examined the impact of the number of PCR cycles, RNA extraction method and degree of plant material grinding. Therefore, we looked for differences in the rate of duplicated read pairs and the number of detected transcripts in these categories. Two of the six tests resulted in a significant ($P < 0.05$) difference in at least one group. First, the custom phenol:chloroform (HTP96) extraction method resulted in a significantly ($P < 0.001$) lower number of detected transcripts compared to the TRIzol method (0.77%). Secondly, the coarse grinding of plant material resulted in a significantly ($P < 0.001$) lower number of detected transcripts compared to the fine grinding of plant material (0.81%).

Increased variability in highly miniaturized samples

When comparing the rate of raw read duplications, we observed an increased variability between the samples within miniaturization levels above 1:4. Without miniaturization (V100) or a minimal miniaturization level (V050), the number of unique reads varied between 67.1% and 72.6% (Table S3). With higher miniaturization levels (V025, V017, V013) the variability between samples increased. For these samples, the rate of unique reads was between 45.2% and 75.3% (Figure 3a). The average rate of unique reads for DR8 and DR9 dropped from 69% (±0.68%) at V100 to 59% (±5.70%) for V013. Furthermore, the rate of uniquely aligned reads was highest for V100 (78% ± 1.22%) and lowest for V013 (66% ± 9.29%) and did therefore show the same trend as the rate of raw read duplications (Figure 3b). The overall alignment rate increased slightly with increasing miniaturization levels from V100 (92% ± 0.31%) to V013 (95% ±0.63%). The proportion of reads that were not properly aligned (once or multiple times) was similar across all miniaturizations (Table S4). The correlation coefficient between the raw read duplicates and read pair duplicates ($r = 0.52$, $P < 4.46e\text{-}08$) was lower than the correlation coefficient between raw read duplication rate and multi-alignment rate ($r = 0.93$, $P < 2.2e\text{-}16$) (Figure 3c). We also observed an impact of the genotype on the rate of multi-aligned reads (Figure 3d). In all miniaturizations, RIL DR9 had lower rates than DR8 but the differences were not significant ($P > 0.05$). The proportion of duplicated sequences with more than 10 identical duplicates was considerably increased for V025, V017 and V013 compared to V100 and V050 (Figure S5a). Additionally, for V017 and V013 the proportion of duplicated reads with more than 100 identical duplicates was further increased compared to the other miniaturizations.

Characterizing the multi-aligned reads

To further investigate the increased between-sample variability that was detected in V025, V017 and V013 miniaturizations, we created two data subsets. First, we subsetted the alignment results including only reads that were mapped multiple times during the alignment. Those read IDs were used to create the

subset of multi-aligned reads. Comparing the rate of raw read duplications between the subset of multi-aligned reads and the total data set showed that the mean duplication rate increased by 36% in the subset (Figure S6a). For a considerable portion of these duplicates, more than 10 reads of the same sequence were present (Figure S5b). The mean proportion of all duplicated reads included in the subset of multi-aligned reads was around 40%, compared to only 7% of all unique reads (Figure S6b).

The origin of the multi-aligned reads was analysed by investigating gene annotation, transposable elements (TEs) sequence overlap and rRNA contamination. A GO term enrichment analysis between the subset of multi-aligned reads and the total data set resulted in multiple significantly underrepresented genes related to TE activity (biological process: 'RNA-dependent DNA biosynthetic process'; molecular function: 'RNA-directed DNA polymerase activity', 'RNA–DNA hybrid ribonuclease activity'). The most overrepresented genes were related to photosynthesis (various biosynthetic process, cellular compartment and molecular function annotations) and transcription (biological process: 'regulation of transcription by RNA polymerase II'; cellular compartment: 'mediator complex') (Figure S7).

On average, more than 55% of the multi-aligned reads could not be assigned to an annotated transcript (no feature, NF). This was significantly higher ($P < 0.001$) than the 16% of reads in the total data set (Figure S8a). The rate of NF reads in the subset of multi-aligned reads did not correlate well with the multi-alignment rate (Figure S8b). While the proportion of NF reads that were multi-aligned varied (82.4%–20.2%), the number of uniquely aligned NF reads remained constant between all 94 samples (4.4% ± 1.0%) (Figure S8c).

The rate of TE reads between the subset of multi-aligned reads and the total data set was significantly increased ($P < 0.001$) (Figure S9a). The rate of TE reads was positively correlated with the multi-alignment rate (Figure S9b). Nevertheless, on average 17% of TE reads were not multi-aligned (Figure S9c). The variability between samples was highest for the TE reads that aligned multiple times (9.6% ± 7.1%).

Lastly, two different rRNA reference sequence libraries were created and the subset of multi-aligned reads was aligned against them. The *Hordeum vulgare* rRNA reference library showed the highest overall alignment rate with most read pairs aligning multiple times against the reference sequences (Figure S6c,d). For both rRNA reference sequence libraries, the subset of multi-aligned reads showed a significantly higher alignment rate ($P < 0.001$) than the total data set (Figure 4a). The Pearson correlation coefficient between the rRNA alignment rate and the multi-alignment rate was 0.999 for the subset of multi-aligned reads and 0.996 for the total data set (Figure 4b). While the proportion of multi-aligned reads that were of rRNA origin varied (93.9%–16.5%), the number of multi-aligned non-rRNA reads remained constant between all 94 samples (3.6% ± 0.5%) (Figure 4c). Additionally, only a small proportion of rRNA reads were uniquely aligned (1.9% ± 2.2%).

Variant calling and differential expression analyses

While the miniaturization did not considerably change the overall number of detected SNPs, the ratio between reference SNPs and alternative SNPs changed in many but not all miniaturization scenarios in favour of an increase in alternative SNPs in higher miniaturizations. However, the change was not consistent and no clear trend was observed when increasing the miniaturization level (Table S5).

We used a principal component analysis (PCA) to examine the data set's capability to cluster the samples based on genetic differences. When using read count data, the first two principal components explained 33.7% of the variance (Figure 5a). When using the SNP data set, the first two principal components explained 58.6% of the variance (Figure 5b). Based on both data sets, we could show that all samples of the same population clustered together. Additionally, the SNP data differentiated each of the five RILs. The samples which were prepared using the same miniaturization level did not cluster together across RIL (Figure S10).

The mean proportion of detected transcripts between miniaturizations for DR8 and DR9 in a 2 million read subset were very similar (30%–31%). V100 had the lowest proportion of detected transcripts with 30.0% which was significantly ($P = 0.042$) less then V050 (30.5%). We created lists of consensus transcripts separately for each miniaturization of DR8 and DR9. The resulting number of detected transcripts within replicates of RIL DR8 and DR9 for each miniaturization level varied between 15 428 (V100, DR9) and 16 930 (V025, DR8). Lastly, we characterized the overlap between the transcripts of each group resulting in 81.2% and 80.8% of all detected transcripts present in all miniaturizations for DR8 and DR9, respectively (Figure 6). For both RIL the second biggest group of transcripts was detected in all miniaturization levels except V100. Only 159 (DR8) and 143 (DR9) transcripts were exclusively present in V100.

Discussion

The workflow we describe in this manuscript reduces RNA library preparation costs by miniaturizing the process by up to 1:8 of the original reagent volume of the commercial kit. However, in order to ensure that the proposed modifications to the commercial protocol did not decrease the quality of the resulting library, an in-depth characterization was required. While many automated miniaturizations have been shown to have negligible impact on the library quality (Jaeger *et al*., 2020; Kong *et al*., 2019; Mildrum *et al*., 2020; Mora-Castilla *et al*., 2016), we used a procedure without automation. Therefore one has to investigate the practicality of the workflow considering the potentially increased pipetting error variance and potentially decreased reproducibility by manually handling <2 μL volumes.

Quality control: Library complexity and biases

In a first step, we had to make sure that our modifications to the protocol had no negative impact on the sequencing process. The per sequence quality scores, per base *n* content and sequencing length distribution of V100 did not significantly ($P > 0.05$) out-perform miniaturized samples (Figures 2a and S3). Additionally, the proportion of reads that were discarded by trimming was comparable between all samples regardless of the miniaturization level (Figure 2b). These observations led to the conclusion that the miniaturization did not have a negative impact on the sequencing process itself.

Next, we investigated additional quality metrics that characterize the library properties directly. Particularly, the library complexity and the potential library biases were investigated. Library complexity, or the number of unique molecules in solution, represents the potential of the given library to produce a complete picture of the genome or transcriptome and unravels potential problems during the library preparation, for example, if a significant number of unique molecules was lost. One way to

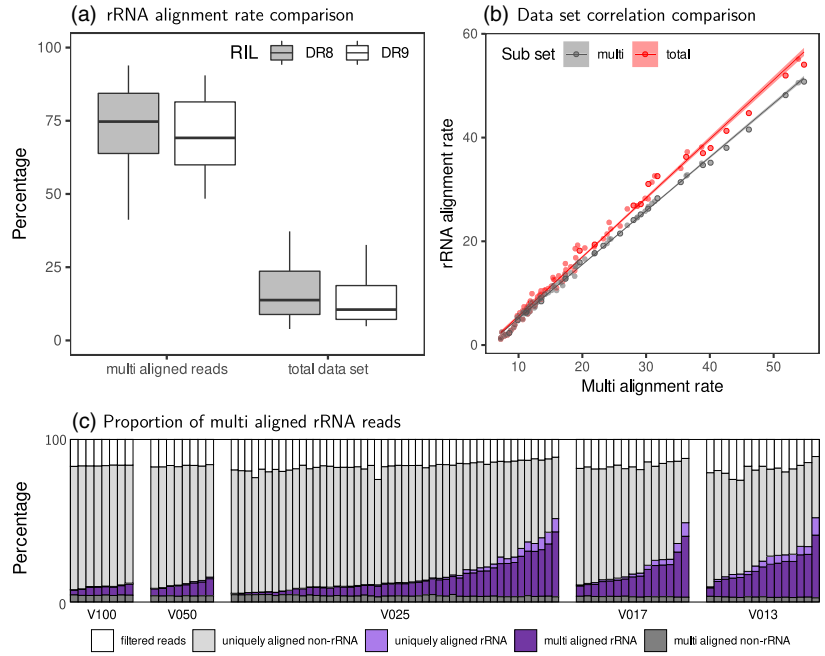

Figure 4 Investigation of the rRNA origin in the subset of multi-aligned reads. (a) Comparison of rRNA alignment rates using the *Horedum vulgare* rRNA reference library between the subset of multi-aligned reads and the total data set (recombinant inbred line (RIL) DR8 in grey and DR9 in white). (b) The correlation between the multi-alignment rate and the rRNA alignment rate for the subset of multi-aligned reads (multi, grey) and the total data set (total, red). (c) The proportion of filtered reads (white), uniquely aligned non-rRNA reads (light grey)/rRNA reads (light purple) and multi-aligned rRNA reads (dark purple)/non-rRNA reads (dark grey).

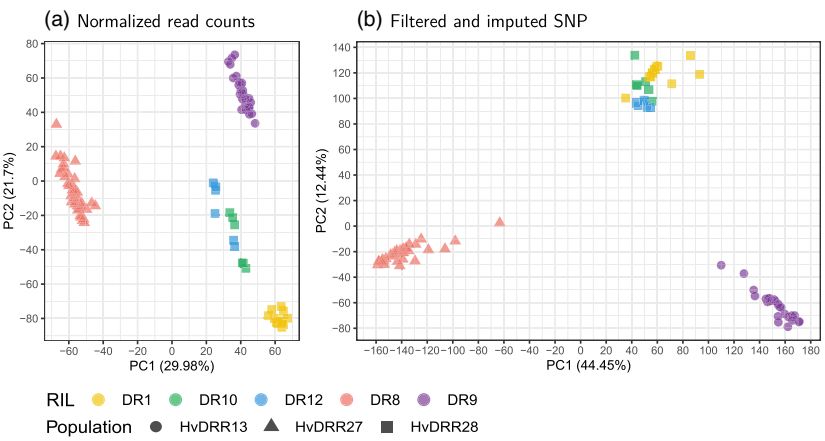

Figure 5 Principal component analysis based on RNA sequencing data of the five recombinant inbred lines (RIL). (a) Normalized read counts used as a basis. (b) Filtered and imputed SNPs used as a basis. PC 1 and PC 2 are the first and second principal component, respectively, and the number in parentheses refers to the proportion of variance explained by the principal components in percent.

assess library complexity is to compare read pair duplication rates (Alberti *et al.*, 2014). A reduced number of unique molecules statistically increases the number of PCR duplicates included in the sequencing pool and therefore increases the probability for read pair duplication in the resulting sequencing data and at the same time reduces the number of unique transcripts that can be discovered. Alternatively, when analysing RNA sequencing data, the number of detected genes can be evaluated (Mereu *et al.*, 2020). In our results, the mean read pair duplication was 7.1% with only the V013 miniaturizations showing significantly ($P < 0.001$) increased duplication rates for both RIL (Figure 2c). Therefore, V025 is an attractive miniaturization factor which makes most efficient use of the sequencing while at the same time resulting in a major cost decrease in the library preparation costs. Nevertheless, the overall read pair duplication rate observed for all miniaturization levels was comparable to those of previous studies (Bansal, 2017; Fu *et al.*, 2018; Parekh *et al.*, 2016). The number of unique detected transcripts did not differ significantly ($P > 0.05$) between all miniaturization levels (Figure 2e) and both V017 and V013 did not separate in our PCA using read count data. This indicated that the overall library complexity was not notably impacted and, thus, also our V017 and V013 libraries remain suitable for read count analyses despite the significantly increased mean read pair duplication for the V013 miniaturization. Both miniaturization factor V017 and V013 have in common that all reagents were diluted to match the V025 volume. We are not aware of any published research examining similar aspects and can therefore only speculate about a potential link between dilution and read pair duplication. If there is a relationship, the most likely reason would be a decreased effectiveness of the cDNA synthesis step in the library preparation prior to the PCR amplification. The decreased effectiveness could have led to a slight decrease in library complexity which in return could have caused an increase in duplicates created by the PCR amplification.

We evaluated non-random RNA fragmentation as a potential cause for a library bias and tested for it by looking at the nucleotide base ratios on both sides of the library fragmentation site. Only at eight positions, one or more bases were significantly ($P < 0.001$) changed between un-miniaturized and miniaturized samples for both RIL DR8 and DR9 (Figure S4). All eight positions were found in the comparison between V013 and V100. We

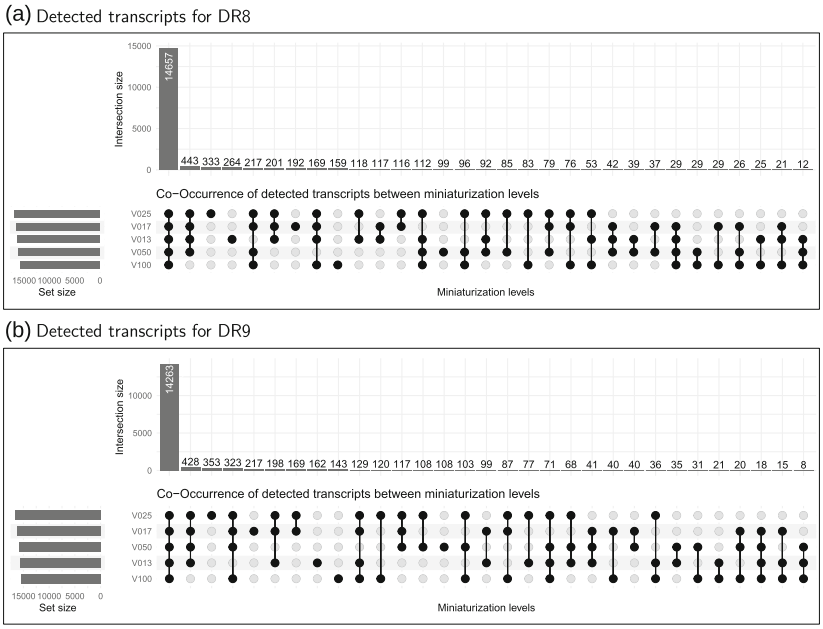

Figure 6 Co-occurrence of detected transcripts between miniaturization levels. Overlap of detected transcripts was calculated based on a subset of 2 million reads of the two recombinant inbred lines (RIL) DR8 (a) and DR9 (b), respectively.

suspect that the significant changes are caused by changes in sequence composition in samples with a high number of duplicates. However, these differences did not lead to a 5′ gene body coverage bias in any sample with no considerable difference between the different miniaturization levels (Figure 2d). All the above-mentioned results indicate that our modification to the library preparation workflow did not negatively impact the library complexity and did not introduce biases.

Quality control: Increased variability among replicates

Two aspects that were significantly ($P < 0.001$) changed by the miniaturization were (1) the proportion of raw read duplications and (2) the rate of multi-aligned reads. First, the increase in raw read duplications for the miniaturizations V025, V017 and V013 compared to V50 and V100 coincided with an increased variability among replicates within these miniaturization levels (Figure 3a). The number of raw read duplications can change based on differences in transcriptome composition between genotypes (Bansal, 2017). However, both RIL that were included in all examined miniaturization levels showed the same trend (Table S3) and, therefore the genotype was only partially able to explain the observations.

It is unclear if the remaining duplicated reads were introduced by sampling biological duplicates based on the redundancy of mRNA molecules or were technical duplicates created during PCR amplification. Expanding the workflow to include unique molecular identifiers (UMIs) would make the distinction between these two cases possible, but also increases the cost (Kivioja *et al.*, 2011). However, UMIs are also reported to increase the accuracy of transcript quantification, which could justify the added cost under certain circumstances (Fu *et al.*, 2018). Results of previous studies suggested that the number of usable reads can be increased by up to 40% by deduplication using UMIs (Collins *et al.*, 2015; Fu *et al.*, 2018; Girardot *et al.*, 2016). However, the selection of commercial kits that use adapters with UMI is limited when the number of multiplexed samples is bigger than 96. At the time of writing, we are not aware of any dual indexed UMI adapter kits that would have been compatible with our library preparation kit and have the capability of a 384-sample multiplex.

The raw read duplicates were further characterized by estimating the multiplication levels of the duplicated sequences. The number of unique duplicates, defined as the number of unique sequences that were duplicated, was consistent between all samples and miniaturization levels and did not reflect the variability in raw read duplicates (Figure S5a). This indicated that the sample-dependent copy number increase of a relatively small number of sequences was responsible for the observed variability.

Secondly, the number of properly paired unique alignments was reduced and the variability increased for the miniaturization levels V025 and above compared to V050 and V100 (Figure 3b). This observation cannot be fully explained by the difference between RILs (Figure 3d). In addition, the rate of multi-aligned reads and the rate of raw read duplications were strongly correlated on a sample-by-sample basis (Figure 3c). The connection between these two measures was not clear to us, which is why we further investigated their nature.

The origin of the increased variability

We began by thoroughly examining the relationship between the rate of multi-aligned reads and the rate of raw read duplications on a read-by-read basis, comparing the rate of raw read duplications between the subset of only multi-aligned reads and all reads in the total data set. The average proportion of duplicated reads was increased by 36% in the subset of multi-aligned reads and included up to 73% of all duplicates. This observation indicated that many multi-aligned reads are also duplicated reads (Figure S6a,b). To better understand the source of multi-aligned reads, we investigated their biological origin. One biological reason for the occurrence of multi-aligned reads are gene duplication events (Deschamps-Francoeur *et al.*, 2020). Therefore, we have examined genomic features that are classically connected to genome duplications and repetitiveness such as rRNA, histone gene family and transposable elements (Deschamps-Francoeur *et al.*, 2020; Magadum *et al.*, 2013; Rooney and Ward, 2005).

When checking the positional overlap of our reads with TEs in barley, the overlap in the subset of multi-aligned reads was higher than in the total data set (Figure S9a). This indicated that at least a part of the variance of the multi-alignment rate across samples can be explained by a sample-dependent increase in TEs. At the same time, a GO term enrichment analysis between the subset of multi-aligned reads and the total data set indicated a reduction in GO terms associated with retrotransposon activity. The same GO term enrichment analysis resulted in no significant ($P > 0.05$) changes in GO terms connected to the synthesis, modification or regulation of histones (Figure S7). Importantly, only a small proportion of the subset of multi-aligned reads was included in the analysis, because most of the reads could not be assigned to a gene (Figure S6a). This favours a non-gene-related explanation for the variance of multi-alignment rates like TE read or rRNA contamination.

From a technical perspective, the contamination with rRNA would be the most plausible explanation for the variance of multi-alignment rates. We tested for rRNA by using available rRNA annotation data in *Hordeum vulgare* to search for rRNA sequences in our multi-aligned read subset. The mean overall alignment rate for the read subset was considerably higher than for the total data set (Figure 4a). We could show that the increase in multi-alignments was strictly correlated with an increase in rRNA reads (Figure 4b). The rate of multi-aligned reads that could not be identified as rRNA was low (3.6%) and remained constant across all 94 samples (Figure 4c). Additionally, on average less than 2% of rRNA sequences were not aligning multiple times. These observations strongly suggested that the increase in multi-alignments as well as the increase in raw read duplications in the miniaturizations V025, V017 and V013 compared to V050 and V100 was caused by rRNA contamination.

The origin and consequences of rRNA contamination

We speculate that the rRNA contamination is caused by incomplete separation during the mRNA capture process at the beginning of the library preparation. This is the first and one of the most crucial magnetic bead separation steps, which are in our experience the most error-prone steps and very susceptible to the decrease in volume caused by the miniaturization (Figure S11).

Depending on the planned usage of the RNA sequencing data, an increased rRNA sequence content leads to a decrease in useful sequencing read output. This in return requires the number of sequenced reads to be increased, which inflates the costs. When accounting for the highest difference in multi-aligned read rates that was observed in our study for a single sample with 35%, the increased sequencing depth added 6 Euro to the overall costs. This was considerably lower than the overall cost reduction of the miniaturized library preparation by 21 Euro realized in our study (Table 1). Therefore, even accounting for the highest possible rRNA sequence content detected in this study, the cost-saving potential of miniaturized library preparation remains attractive.

To prevent rRNA contamination, the poly-A mRNA capture method in our workflow could be replaced by an rRNA depletion step, which was shown to have higher success in removing rRNA than poly-A capture methods (Kumar et al., 2017). This would presumably decrease the number of rRNA molecules and consequently the number of raw read duplications and multi alignments in our data set. However, the addition of an rRNA depletion step would greatly reduce the cost saving potential of the workflow and therefore only rarely be a reasonable alternative. Alternatively, for reaction volumes below 30 µL that include magnetic beads, a special low-elution 96-well magnet plate can be used to increase the magnetic bond and enable more reliable molecule selection. When this study was performed, a low-elution 96-well magnetic plate was not available to us and we were using a standard magnetic plate (96S Super Magnet, Alpaqua Engineering, USA). This most likely caused the mRNA capture and therefore the rRNA exclusion success to vary when handling smaller volumes in the higher miniaturizations.

Table 1 Summary of workflow costs. The relative costs comparison for the RNA isolation, RNA library preparation (RNA library prep.) and RNA sequencing between all miniaturization factors. The costs are relative to a standard workflow defined as: RNA extraction using RNeasy Plant Mini Kit (Qiagen, Germany) and library preparation using TruSeq RNA Library Prep Kit v2 (Illumina, USA). In comparison, the miniaturization workflow used: TRIzol-96 RNA isolation, VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina library preparation. Both workflows aimed to sequence 10 million reads using the DNBSEQ-G400 platform. The sequencing depth was increased to compensate for multi-aligned reads according to the maximum multi alignment rate for each miniaturization factor. The multi alignment rate of the standard workflow was set to 0%

	Miniaturization factors			
Workflow	V100	V050	V025	V013
RNA isolation	31.47%	15.74%	15.74%	15.74%
RNA library prep.	30.66%	15.33%	7.66%	3.83%
RNA sequencing	111.95%	115.32%	135.41%	147.07%
Total	44.39%	32.19%	29.70%	28.72%

Workflow modifications

The here described workflow is compatible with state-of-the-art liquid handling automatization solutions. The labour and time intensive magnetic bead separation steps would greatly benefit from the integration of an automated liquid handler designed for working with 96-well plates. The steps that would benefit are: the poly-A tail capture, size selection and all library fragment clean-ups. The other steps would require a liquid handling device that is able to accurately transfer small volumes (e.g. positive displacement instruments, acoustic droplet ejection instruments). This would potentially enable miniaturization factors beyond the ones examined in this study.

When using different types of plant material, the requirements to the workflow could change. We see a potential need to adjust the RNA extraction method depending on the material used. We showed that our miniaturization workflow produces high-quality data with multiple types of extraction methods. While the extraction methods can slightly alter the overall costs of the workflow, it does not affect the validity of the miniaturization as a cost saving measure. After successfully extraction of high-quality total RNA, our preliminary data do not indicate that the later steps of the workflow will be affected by the type of input material.

Data application: Sequence variation and read counts

We picked two common use cases to examine the suitability of the RNA sequencing data generated with our workflow. First, the potential of the data sets to identify sequence variants and second, the ability to describe differential expression using read

counts. The capabilities of the data set to call sequence variation and the potential influence the miniaturization had on it, were investigated by comparing the mean number of detected SNPs. The total number of SNP for DR8 and DR9, respectively did not change significantly ($P > 0.05$) between miniaturizations (Table S5). This finding was in agreement with the observation made in the PCA we conducted using filtered imputed SNP data (Figure 5). The results showed a high percentage of the variance between samples that matched the genetic structure of the samples, while the miniaturization was not shown to explain any variance and therefore could not be shown to have a systematic effect on the generated data (Figure S10). This makes the workflow capable of detecting sequence variances, which then could be used for example, in genome-wide association studies (Rodriguez *et al*., 2020).

When further examining the read counts, between 85%–92% of the consensus transcripts (expressed at least once per miniaturization) were found in all miniaturizations and the overlap between V100 and V025 was >97% (Figure 6). This illustrates the high consistency between miniaturizations which was further underpinned by Pearson correlation coefficients of at least 0.9613 for the RIL DR8 and DR9 among the miniaturization levels (Table S1). The correlation coefficient was even higher when comparing RNA extraction and library preparation replicates for both RIL (Table S2). These observations suggested that the miniaturization did not affect the ability to capture the transcriptome and all levels are capable of being used for comparative read count analyses for example, in differential gene expression (Cantalapiedra *et al*., 2017; Kintlová *et al*., 2021).

Conclusions

With minimal adjustments to an established commercial RNAseq library preparation protocol, we were able to manually miniaturize the library preparation by a factor of up to 1:8. This leads to cost savings of up to 54.5% compared to the same library preparation protocol without miniaturization and up to 86.1% compared to the gold standard. The rigorous quality control analysis of the resulting sequencing data and its application did not result in any indication of biases or inaccuracies caused by the library miniaturization. All libraries created in this study can be considered high quality and are ready to be used in a wide range of projects. The observed rRNA contamination did not affect the quality of the library itself and can be addressed by workflow adjustments, for example, using low elution volume magnetic plates. As shown by our cost projections even the potential efficiency decrease caused by the increase in multi-aligned reads did not change the fact that the method proposed here is an uncomplicated way to reduce the cost of RNAseq library preparation without a long set-up time or the need for scarcely available lab automation equipment. While in our study only a single commercial RNAseq library preparation kit was evaluated, we are confident that the general principle of miniaturization as well as observed unbiased results can be applied to a wide range of kits.

Materials and methods

Genetic material

Our study was based on five barley RIL from three HvDRR subpopulations (Casale *et al*., 2021). The HvDRR population was developed from pairwise crosses among 23 diverse parental inbreds (Weisweiler *et al*., 2019) using the double round robin (DRR) mating design (Stich, 2009).

Experimental design

The experimental design was set up such that it is possible to statistically test the effect of the miniaturization, RNA extraction method, number of PCR cycles and degree of plant tissue grinding on library complexity and biases (Figure 1). A total of 96 samples were examined. Five different genotypes were used: from HvDRR sub-population #28 line #33 (DR1), #46 (DR12) and #57 (DR10), from HvDRR sub-population #27 line #40 (DR8) and from HvDRR sub-population #13 line #29 (DR9). The libraries were prepared without miniaturization (1:1, V100), the miniaturization levels 1:2 using 50% of all reagents (V050), 1:4 using 25% of all reagents (V025), 1:6 using 17% of all reagents (V017) and 1:8 using 13% of all reagents (V013). Two of the five RIL (DR8 and DR9) were included in all miniaturization levels to allow orthogonal comparisons. Depending on the miniaturization level, two or three RNA extraction replicates and library preparation replicates each was present. RNA extraction replicates were distinct RNA extractions and library preparations using the same plant material. Library preparation replicates were distinct library preparations using the same extracted RNA. For the miniaturization levels 1:4, 1:6 and 1:8, the library of line DR8 was amplified using 8 PCR cycles and 10 PCR cycles. For the remaining miniaturization levels, only 10 PCR cycles were used. DR1 libraries were only prepared using 1:4 miniaturization but with plant material being either coarsely or finely ground as the basis for RNA isolation. Finally, 1:4 miniaturization libraries for the lines DR10 and DR12 were prepared using different RNA extraction methods.

Plant cultivation

A total of 15 seeds from each of the five RIL were sterilized with sodium hypochlorite (13%) for 10 min. All seeds of a single RIL were placed in a rectangular (12 × 12 cm) Petri dish between two sheets of filter paper (12 × 12 cm) supplied with tap water. The seeds were lined up in the lower half of the Petri dish on top of the water-soaked first filter paper. A second water-soaked filter paper was placed above the seeds, starting 1 cm above the seeds so that the lower third of the paper was reaching out of the Petri dish (Figure S12). This ensured that both filter paper sheets do not dry out over time. The Petri dishes were then stacked and placed in a vertical orientation in a plant tray (40 × 60 cm). This way, the space requirements per RIL were minimized and when filling the tray with water (approx. 3 L) the seedlings can grow for more than 7 days without further maintenance. The seedlings were cultivated for 7 days in a reach-in growth chamber under the following conditions: 70% relative humidity, 16 h of light (6:00–22:00), 22 degrees (day)/20 degrees (night) and light intensity of 400 μmol m^{-1} s^{-1}. The time of day for the cultivation start and the harvest was similar (within 2 h) for all samples. The RIL DR8, DR9 and DR10 were cultivated at the same time, while DR1 and DR12 were each cultivated at different dates.

Plant material processing

The whole seedlings of one Petri dish were harvested by transfer into a collection tube and immediately freezing them in liquid nitrogen. Afterward, they were ground by using steel beads and a paint shaker until the plant material was powdery (fine). Additionally, half of the plant material from DR1 was only

ground until it was flaky (coarse). Afterward, 50 mg or 100 mg of plant material was transferred into different collection tubes depending on the extraction method and stored at −80 degrees until RNA extraction started.

RNA extraction

Three different total RNA extraction methods were applied. First, a custom Phenol:Chloroform extraction optimized for high throughput extraction in 96-well plates (Box *et al*. (2011); Phenol:Chloroform, P:C). Second, a TRIzol reagent (Thermo Fisher Scientific) based RNA extraction following the manufactures instructions (TRIzol). For the third RNA extraction, TRIzol reagent (Thermo Fisher Scientific) was used in a 96-well format with an adapted protocol (TRIzol-96). The input plant material and all reagents for the TRIzol-96 extraction were halved compared to the standard protocol. The final washing step in 75% ethanol was repeated one time to assure that all the remaining phenol was removed. All other steps were executed as proposed by the manufacturer. The first two methods used 100 mg and the third method 50 mg of frozen fresh plant material as input. The total RNA concentration was quantified using a NanoPhotometer NP 80 (Implen, Germany). All samples, except the TRIzol-96 extractions, were evaluated using the Fragment Analyzer (Agilent).

Library preparation

The mRNA was selected based on a poly-A tail mRNA capture method (Vazyme, China) using 1 μg total RNA as input. The full-length mRNA library was constructed using the VAHTS Universal V6 RNA-seq Library Prep Kit for Illumina (Vazyme, China). We miniaturized the kits by reducing the reagent volume to 50% (V050), 25% (V025), 17% (V017) and 13% (V013) of the original. The reaction volume for V017 and V013 were kept at 25% of the original volume to avoid pipetting volumes below 1 μL. The remaining reaction volume was filled up with RNase-free water, which resulted in a dilution for the miniaturizations V017 (1:1.5) and V013 (1:2). Size selection and clean-up were performed using magnetic DNA Clean Beads (Vazyme, China). In that step, the reaction volume and reagents were reduced to the respective miniaturization level with the exception of V017 and V013 for which the V025 reaction volume and reagents were used. Apart from these changes, the manufacturer's protocol was followed aiming for 250–450 bp long inserts. The 96 separate libraries were prepared in a 96-well plate with each miniaturization level occupying two to four columns. The order of the columns was randomized starting with the library preparation. The costs of our experimental workflow were compared to a gold standard which includes RNA extraction using RNeasy Plant Mini Kit (Qiagen, Germany) and library preparation using TruSeq RNA Library Prep Kit v2 (Illumina, USA).

Sequencing, read processing and alignment

The sequencing was performed by BGI on the DNBSEQ-G400 platform. All 96 samples were pooled and a total of 1.42 billion 150 bp paired-end reads were sequenced with an average of 14.8 million read pairs per sample. Two samples did have less than 2 million reads sequenced and were excluded from all further analyses. Both samples were from RIL DR8, RNA extraction replicate #1 and miniaturization levels V013 and V017, respectively. Various quality statistics of the raw sequencing reads were calculated using FastQC (Andrews, 2019) and afterward trimmed with trimmomatic (ILLUMINACLIP:TruSeq3-PE:2:30:10:1:TRUE SLIDINGWINDOW:4:15 LEADING:3 TRAILING:3 MINLEN:36) (Bolger *et al*., 2014). The trimmed reads were then aligned to the Morex V3 reference sequence (Mascher, 2019) using Hisat2 (−no-softclip −max-seeds 1000) (Kim *et al*., 2019).

Alignment against rRNA reference libraries

In an effort to learn about the origin of multi-aligned reads, a read subset was created including only reads flagged as multi-mapped in the primary alignment against the reference sequence Morex. The total data set and the subset of multi-aligned reads were then aligned against two different rRNA reference libraries to estimate the percentage of reads originating from rRNA.

The rRNA reference libraries were created using HISAT2 without exon and splice site information. The sequences were searched for and downloaded as .fasta files from the RNA Central Expert Database using the following search criteria: (1) *Hordeum vulgare* subsp. *vulgare* rRNA (search term: taxonomy: "112509" AND rna_type: "rRNA") (1399 sequences) and (2) Ensembl Plant database rRNA (search term: rna_type: "rRNA" AND expert_db: "Ensembl Plants") (14 880 sequences). The much larger Ensembl Plant library includes rRNA sequences from many different plant species (e.g. *Arabidopsis thaliana*, *Oryza sativa* Japonica, *Triticum aestivum*, *Hordeum vulgare*) was used to evaluate the *Hordeum vulgare* library capability to create a comprehensive rRNA sequences alignment.

Variant calling and SNP analyses

Variant calling was performed using bcftools mileup (filter: -q 20 -Q 20) and call functions (Li *et al*., 2009). The variants were filtered based on the QUAL score (≥10), the median read depth per sample (≥5) and the total depth per variant (≥30). The raw variant call data was imputed using Beagle 5.4 (Browning *et al*., 2021) based on standard settings without a reference sequence. In the resulting SNP data set, all monomorphic and triallelic SNP and all SNP with more than 30% heterozygosity were removed. The remaining heterozygote SNP were set to NA and afterward median imputed. The SNP count comparisons of filtered and imputed variants used mean SNP counts of each miniaturization and were based on a 2 million read subset.

Read count analysis

The sorted and filtered alignments were then used to determine the read count per gene with the help of htseq count (−mode union) (Anders *et al*., 2015). The read counts were filtered and the Trimmed Mean of the M-values (TMM) method was used to apply a between-sample normalization using the R package edgeR (Robinson *et al*., 2009). The mean Pearson correlation was calculated for all pairwise library replicate combinations within each RNA extraction replicate. Each DR8 and DR9 library replicate was averaged for each miniaturization level. For calculating the correlation between the RNA extraction replicates, the mean of the read counts across the library preparation replicates was used. Afterward, the mean of the correlations was calculated for both DR8 and DR9. Pearson correlations were also calculated between miniaturizations using the mean read counts of all available replicates. Read counts were not only calculated for the total data set but also for the subset of multi-aligned reads. Here we had to allow for non-unique alignments to be included using the '–nonunique all' option of the htseq count function. The number of detected transcripts was calculated based on raw read counts

of 2 million read subsets of all samples. For the estimation of the number of consensus transcripts, transcripts with read counts below 10 were set to 0 in all samples. Afterward, the number of transcripts present was counted for all combinations of miniaturization levels.

Gene body coverage

In order to estimate the relative gene body coverage, a non-random subset of gene-associated reads was evaluated based on their relative position within the gene. We divided each transcript into 100 equally sized windows and counted the number of overlapping reads for each window. Each read was allowed to be counted multiple times. This analysis was performed for all expressed transcripts and afterward, the mean number of reads per sample for each window was calculated. The means were adjusted to accommodate for varying numbers of total reads per sample and rescaled to the range [0, 1] using general minimum and maximum read counts. Because of limitations of the Rsamtools R-library, only transcripts in the first 536 870 912 bases of each chromosome were included in the analysis.

Additional data analyses

To evaluate, if miniaturization leads to non-random fragmentation, the rate of each of the four nucleotide bases was calculated for the first nine bases before and after a fragmentation site. The first base of each forward read was defined as the first base after a fragmentation site. Unless the read start was equal to a transcript start. The first nine bases after the fragmentation site were therefore the first nine bases of a forward read and the nine bases before a fragmentation site were the last nine bases of a reverse read. A GO term enrichment analysis was conducted between the total data set and the subset of multi-aligned reads. Statistical differences were calculated using the Fisher exact test. The *P*-values were adjusted for multiple testing using the Benjamini–Hochberg procedure.

The rate of TE reads in the total data set and the subset of multi-aligned reads was estimated by calculating the rate of reads that overlap with genome positions annotated as TEs. Significant differences between groups (e.g. miniaturization levels) were assessed with *post hoc* Tukey's honestly significant difference tests. A PCA of the filtered, imputed variants across all samples was performed. A similar analysis was conducted on the TMM-normalized read counts. All metrics ascertained during the general sequencing data processing were aggregated using multiQC (Ewels *et al*., 2016). The significance threshold for all statistical tests in this study was set to 0.05.

Author contributions

The study was conceptualized and designed by C.A and B.S.; K.K. and T.W advised the laboratory experiments performed by C.A; The data was analysed and interpreted by C.A and B.S; The manuscript was written by C.A. and edited by B.S., C.A., K.K. and T.W; All authors read and approved the final manuscript.

Acknowledgements

Computational infrastructure and support were provided by the Centre for Information and Media Technology at Heinrich-Heine-University Duesseldorf. This research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2048/1, Project ID:390686111). Open Access funding enabled and organized by Projekt DEAL.

Conflict of interest

The authors declare no conflict of interest.

References

Aigrain, L., Yong, G. and Quail, M.A. (2016) Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays - a systematic comparison of DNA library preparation kits for illumina sequencing. *BMC Genom.* **17**, 458.

Alberti, A., Belser, C., Engelen, S., Bertrand, L., Orvain, C., Brinas, L., Cruaud, C. *et al*. (2014) Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genom.* **15**, 912.

Alpern, D., Gardeux, V., Russeil, J., Mangeat, B., Antonio, C., Meireles-Filho, A., Breysse, R. *et al*. (2019) BRB-seq: ultra-affordable high-throughput transcriptomics enabled by bulk RNA barcoding and sequencing. *Genome Biol.* **20**, 12.

Anders, S., Pyl, P.T. and Huber, W. (2015) Htseq-a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Andrews, S. (2019) *Fastqc*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Bentley, D.R., Chakravarti, A., Clark, A.G. *et al*. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Bagnoli, J.W., Ziegenhain, C., Janjic, A., Wange, L.E., Vieth, B., Parekh, S., Geuder, J. *et al*. (2018) Sensitive and powerful single-cell RNA sequencing using mcSCRB-seq. *Nat. Commun.* **9**, 12.

Bansal, V. (2017) A computational method for estimating the PCR duplication rate in dna and RNA-seq experiments. *BMC Bioinform.* **18**, 113–123.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Box, M.S., Coustham, V., Dean, C. and Mylne, J.S. (2011) Protocol: A simple phenol-based method for 96-well extraction of high quality RNA from arabidopsis. *Plant Methods* **7**, 7.

Browning, B.L., X.T., Zhou, Y. and Browning, S.R. (2021) Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.* **108**, 1880–1890.

Cantalapiedra, C.P., García-Pereira, M.J., Gracia, M.P., Igartua, E., Casas, A.M. and Contreras-Moreira, B. (2017) Large differences in gene expression responses to drought and heat stress between elite barley cultivar scarlett and a spanish landrace. *Front. Plant Sci.* **8**, 647.

Casale, F., Van Inghelandt, D., Weisweiler, M., Li, J. and Stich, B. (2021) Genomic prediction of the recombination rate variation in barley – a route to highly recombinogenic genotypes. *Plant Biotechnol. J.* **20**, 676–690.

Chung, Y.S., Choi, S.C., Jun, T.H. and Kim, C. (2017) Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. *Hortic. Environ. Biotechnol.* **58**, 425–431.

Collins, J.E., Wali, N., Sealy, I.M., Morris, J.A., White, R.J., Leonard, S.R., Jackson, D.K. *et al*. (2015) High-throughput and quantitative genome-wide messenger RNA sequencing for molecular phenotyping. *BMC Genom.* **16**, 1–13.

Dabney, J. and Meyer, M. (2012) Length and gc-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern dna sequencing libraries. *Biotechniques*, **52**, 87–94.

Deschamps-Francoeur, G., Simoneau, J. and Scott, M.S. (2020) Handling multi-mapped reads in RNA-seq. *Computational and Structural. Biotechnol. J.* **18**, 1569–1576.

Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016) Multiqc: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048.

Foley, J.W., Zhu, C., Jolivet, P., Zhu, S.X., Lu, P., Meaney, M.J. and West, R.B. (2019) Gene expression profiling of single cells from archival tissue with laser-capture microdissection and smart-3seq. *Genome Res.* **29**, 1816–1825.

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.

Girardot, C., Scholtalbers, J., Sauer, S., Shu Yi, S. and Furlong, E.E.M. (2016) Je, a versatile suite to handle multiplexed NGS libraries with unique molecular identifiers. *BMC Bioinform.* **17**, 1–6.

Harris, P.N.A. and Wailan Alexander, M. (2021) Beyond the core genome: Tracking plasmids in outbreaks of multidrug-resistant bacteria. *Clin. Infect. Dis.* **72**, 421–422.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D. *et al*. (2016) CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-seq. *Genome Biol.* **17**, 77.

Hou, Z., Jiang, P., Swanson, S.A., Elwell, A.L., Nguyen, B.K.S., Bolin, J.M., Stewart, R. *et al*. (2015) A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci. Rep.* **5**, 9570.

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.B., Lönnerberg, P. and Linnarsson, S. (2012) Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. *Nat. Protoc.* **7**, 813–828.

Jaeger, B.N., Yángüez, E., Gesuita, L., Denoth-Lippuner, A., Kruse, M., Karayannis, T. and Jessberger, S. (2020) Miniaturization of smart-seq2 for single-cell and single-nucleus RNA sequencing. *STAR Protocols* **1**, 100081.

Jehl, F., Degalez, F., Bernard, M., Lecerf, F., Lagoutte, L., Désert, C., Coulée, M. *et al*. (2021) RNA-seq data for reliable snp detection and genotype calling: Interest for coding variant characterization and cis-regulation analysis by allele-specific expression in livestock species. *Front. Genet.* **12**, 1104.

Kim, D., Paggi, J.M., Park, C., Bennett, C. and Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and hisat-genotype. *Nat. Biotechnol.* **37**, 907–915.

Kintlová, M., Vrána, J., Hobza, R., Blavet, N. and Hudzieczek, V. (2021) Transcriptome response to cadmium exposure in barley (hordeum vulgare l.). *Front. Plant Sci.* **12**, 1359.

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.

Kong, S.L., Li, H., Tai, J.A., Courtois, E.T., Poh, H.M., Lau, D.P., Haw, Y.X. *et al*. (2019) Concurrent single-cell RNA and targeted DNA sequencing on an automated platform for comeasurement of genomic and transcriptomic signatures. *Clin. Chem.* **65**, 272–281.

Kumar, R., Ichihashi, Y., Kimura, S., Chitwood, D.H., Headland, L.R., Peng, J., Maloof, J.N. *et al*. (2012) A high-throughput method for illumina RNA-seq library preparation. *Front. Plant Sci.* **3**, 202.

Li, H. (2021) Single-cell RNA sequencing in drosophila: Technologies and applications. *Wiley Interdiscip. Rev. Dev. Biol.* **10**, e396.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. *et al*. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Li, H., Kun, W., Ruan, C., Pan, J., Wang, Y. and Long, H. (2019) Cost-reduction strategies in massive genomics experiments. *Mar. Life sci. Technol.* **1**, 15–21.

Linderman, M.D., Nielsen, D.E. and Green, R.C. (2016) Personal genome sequencing in ostensibly healthy individuals and the peopleseq consortium. *J. Pers. Med.* **6**, 14.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I. *et al*. (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, **161**, 1202–1214.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. and Ravikesavan, R. (2013) Gene duplication as a major force in evolution. *J. Genet.* **92**, 155–161.

Mardis, E.R. (2011) A decade's perspective on dna sequencing technology. *Nature*, **470**, 198–203.

Mascher, M. (2019) *Pseudomolecules and annotation of the second version of the reference genome sequence assembly of barley cv. morex [morex v2]*. https://doi.ipk-gatersleben.de:443/DOI/83e8e186-dc4b-47f7-a820-28ad37cb176b/d1067eba-1d08-42e2-85ec-66bfd5112cd8/2

Mayday, M.Y., Khan, L.M., Chow, E.D., Zinter, M.S. and DeRisi, J.L. (2019) Miniaturization and optimization of 384-well compatible RNA sequencing library preparation. *PloS One*, **14**, e0206194.

McCombie, W.R., McPherson, J.D. and Mardis, E.R. (2019) Next-generation sequencing technologies. *Cold Spring Harb. Perspect. Med.* **162**, e59.

McNulty, S.N., Mann, P.R., Robinson, J.A., Duncavage, E.J. and Pfeifer, J.D. (2020) Impact of reducing DNA input on next-generation sequencing library complexity and variant detection. *J. Mol. Diagn.* **22**, 720–727.

Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Varela, A.Á., Batlle, E. *et al*. (2020) Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755.

Mildrum, S., Hendricks, A., Stortchevoi, A., Kamelamela, N., Butty, V.L. and Levine, S.S. (2020) High-throughput minitaturized RNA-seq library preparation. *J. Biomol. Tech.* **31**, 151–156.

Mora-Castilla, S., Cuong To, Vaezeslami, S., Morey, R., Srinivasan, S., Dumdie, J.N., Cook-Andersen, H. *et al*. (2016) Miniaturization technologies for efficient single-cell library preparation for next-generation sequencing. *J. Lab. Autom.* **21**, 557–567.

Pallares, L.F., Picard, S. and Ayroles, J.F. (2019) Tm3′seq: A tagmentation-mediated 3′ sequencing approach for improving scalability of rnaseq experiments. *G3: Genes—Genomes—Genetics*, **10**, 143–150.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. (2016) The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 1–11.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G. and Sandberg, R. (2013) Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1100.

Piskol, R., Ramaswami, G. and Li, J.B. (2013) Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651.

Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009) edger: A bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Rodriguez, M., Scintu, A., Posadinu, C.M., Yimin, X., Nguyen, C.V., Sun, H., Bitocchi, E. *et al*. (2020) Gwas based on RNA-seq snps and high-throughput phenotyping combined with climatic data highlights the reservoir of valuable genetic diversity in regional tomato landraces. *Gene*, **11**, 1–25.

Romero, I.G., Pai, A.A., Tung, J., Gilad, Y., Romero, I.G., Pai, A.A., Tung, J. *et al*. (2014) RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 42.

Rooney, A.P. and Ward, T.J. (2005) Evolution of a large ribosomal RNA multigene family in filamentous fungi: Birth and death of a concerted evolution paradigm. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 5084–5089.

Shishkin, A.A., Giannoukos, G., Kucukural, A., Ciulla, D., Busby, M., Surka, C., Chen, J. *et al*. (2015) Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat. Methods*, **12**, 323–325.

Shomroni, O., Sitte, M., Schmidt, J., Parbin, S., Ludewig, F., Yigit, G., Zelarayan, L.C. *et al*. (2022) novel single-cell RNA-sequencing approach and its applicability connecting genotype to phenotype in ageing disease. *Sci. Rep.* **12**, 1–14.

Soumillon, M., Cacchiarelli, D., Semrau, S., van Oudenaarden, A. and Mikkelsen, T.S. (2014) *Characterization of directed differentiation by high-throughput single-cell RNA-seq. bioRxiv*. 003236.

Stark, R., Grzelak, M. and Hadfield, J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656.

Stich, B. (2009) Comparison of mating designs for establishing nested association mapping populations in maize and arabidopsis thaliana. *Genetics*, **183**, 1525–1534.

Tegally, H., San, J.E., Giandhari, J. and de Oliveira, T. (2020) Unlocking the efficiency of genomics laboratories with robotic liquid-handling. *BMC Genom.* **21**, 12.

Vahrenkamp, J.M., Szczotka, K., Dodson, M.K., Jarboe, E.A., Soisson, A.P. and Gertz, J. (2019) Ffpecap-seq: A method for sequencing capped rnas in formalin-fixed paraffin-embedded samples. *Genome Res.* **29**, 1826–1835.

Wang, T., Liu, Y., Ruan, J., Dong, X., Wang, Y. and Peng, J. (2021) A pipeline for RNA-seq based eQTL analysis with automated quality control procedures. *BMC Bioinform.* **22**, 1–18.

Weisweiler, M., De Montaigu, A., Ries, D., Pfeifer, M. and Stich, B. (2019) Transcriptomic and presence/absence variation in the barley genome assessed from multi-tissue mrna sequencing and their power to predict phenotypic traits. *BMC Genom.* **20**, 10.

Wetterstrand, K.A. (2021) *The cost of sequencing a human genome*. https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Supporting information

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

Figure S1. Fragment Analyzer results for total RNA.
Figure S2. Fragment Analyzer results for the library pool.
Figure S3. Sequencing read quality control.
Figure S4. Calculation of nucleotide base distribution around fragmentation sites.
Figure S5. Overview of different multiplication levels.
Figure S6. Duplication rates and rRNA alignment.
Figure S7. GO term enrichment between the subset of multi-aligned reads and the total data set.
Figure S8. Investigation of the no feature (NF) read origin in the subset of multi aligned reads.
Figure S9. Investigation of the transposable element (TE) read origin in the subset of multi-aligned reads.
Figure S10. Principal component analysis based on RNA sequencing data of the five recombinant inbred lines (RIL).
Figure S11. Evaluation of the impact of miniaturization on magnetic bead selection steps
Figure S12. Illustration of the cultivation system.
Table S1. Correlation of read counts between miniaturizations.
Table S2. Correlation of read counts between library preparation and RNA extraction replicates.
Table S3. Mean and standard deviation (SD) of the raw read duplication rate.
Table S4. Mean and standard deviation (SD) of the alignment statistics.
Table S5. Comparison of SNP counts between miniaturizations.

