

RESEARCH

Open Access



Optimizing Cost-Effective gene expression phenotyping approaches in cattle using 3' mRNA sequencing

Ruwaa I. Mohamed¹, Taylor B. Ault-Seay², Sonia J. Moisés², Jonathan E. Beever^{1,2}, Agustín G. Ríos² and Troy N. Rowan^{1,2*}

Abstract

Background Genetic and genomic selection programs require large numbers of phenotypes observed for animals in shared environments. Direct measurements of phenotypes like meat quality, methane emission, and disease susceptibility are difficult and expensive to measure at scale but are critically important to livestock production. Our work leans on our understanding of the “Central Dogma” of molecular genetics to leverage molecular intermediates as cheaply-measured proxies of organism-level phenotypes. The rapidly declining cost of next-generation sequencing presents opportunities for population-level molecular phenotyping. While the cost of whole transcriptome sequencing has declined recently, its required sequencing depth still makes it an expensive choice for wide-scale molecular phenotyping. We aim to optimize 3' mRNA sequencing (3' mRNA-Seq) approaches for collecting cost-effective proxy molecular phenotypes for cattle from easy-to-collect tissue samples (i.e., whole blood). We used matched 3' mRNA-Seq samples for 15 Holstein male calves in a heat stress trial to identify the (1) best library preparation kit (Takara SMART-Seq v4 3' DE and Lexogen QuantSeq) and (2) optimal sequencing depth (0.5 to 20 million reads/sample) to capture gene expression phenotypes most cost-effectively.

Results Takara SMART-Seq v4 3' DE outperformed Lexogen QuantSeq libraries across all metrics: number of quality reads, expressed genes, informative genes, differentially expressed genes, and 3' biased intragenic variants. Serial downsampling analyses identified that as few as 8.0 million reads per sample could effectively capture most of the between-sample variation in gene expression. However, progressively more reads did provide marginal increases in recall across metrics. These 3' mRNA-Seq reads can also capture animal genotypes that could be used as the basis for downstream imputation. The 10 million read downsampled groups called an average of 109,700 SNPs and 11,367 INDELS, many of which segregate at moderate minor allele frequencies in the population.

Conclusion This work demonstrates that 3' mRNA-Seq with Takara SMART-Seq v4 3' DE can provide an incredibly cost-effective (< 25 USD/sample) approach to quantifying molecular phenotypes (gene expression) while discovering sufficient variation for use in genotype imputation. Ongoing work is evaluating the accuracy of imputation and the ability of much larger datasets to predict individual animal phenotypes.

Keywords Molecular phenotyping, Gene expression, High-throughput, Livestock, Transcriptomics

*Correspondence:
Troy N. Rowan
trowan@utk.edu

¹Genome Science and Technology Program, Bredesen Center, University of Tennessee, Knoxville, TN, USA

²Department of Animal Science, University of Tennessee Institute of Agriculture (UTIA), Knoxville, TN, USA



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Background

In livestock populations, selection decisions are made mainly based on statistical estimates of an animal's genetic merit in the form of an estimated breeding value (EBV) [1]. Using EBV-based selection has led to remarkable genetic gain over relatively short periods. These genetic evaluations rely on large numbers of phenotypes measured on individuals in shared environments throughout the population [2, 3]. In most cases where genetic predictions do not exist for an economically-relevant trait, it is due to a lack of phenotypic measurements [4]. Some phenotypes, such as methane emission, disease susceptibility, or metabolic efficiency, are exceedingly challenging to quantify at the scale needed for genetic evaluation [5, 6, 7]. Future improvements to these economically important and sustainability-related traits will rely on novel approaches to collecting indicator phenotypes [8].

One possible solution to this challenge is to use molecular measurements (e.g., gene expression, protein, or metabolite abundance) as high-dimensional proxies for economically significant phenotypes [9]. These molecular measurements could be used in place of or alongside organismal phenotypes in genetic evaluations or in more complex models that predict animal phenotypes rather than additive genetic merit [10, 11, 12, 13]. Proxy phenotypes collected from milk samples via mid-infrared (MIR) spectroscopy are used as indicators for multiple efficiency and health traits in the dairy industry [14]. A similar proxy does not yet exist for beef cattle. As with any complex trait, genetic predictions that leverage intermediate phenotypes will also require large numbers of samples to be useful. As such, tissue collection must be accessible, and molecular data generation must be cheap. At present, the cost of whole transcriptome sequencing and proteomics remains prohibitively expensive. To

equip future analyses with population-scale molecular phenotype data, we propose an optimized approach to quantifying gene expression via 3' mRNA sequencing (3' mRNA-Seq) in whole blood.

Whole transcript sequencing (mRNA-Seq) and 3' mRNA-Seq are two related transcriptomic approaches, each offering unique representations of an individual's gene expression landscape. While both approaches start by capturing mRNA of expressed genes, whole transcriptome sequencing libraries usually involve random fragmentation of entire genes, followed by fragment size selection for sequencing (Fig. 1A). As a result, sequencing reads are distributed evenly across full transcripts. This can result in a biased overrepresentation of longer genes/transcripts amongst reads because they result in more fragments-per-transcript compared to shorter genes/transcripts [15]. This biased overrepresentation of longer genes can negatively affect the detection of differentially expressed genes (DEGs) and downstream analyses of biological functions like gene ontology (GO) enrichments and pathways analysis [16]. This requires that whole RNA sequencing analyses normalize the fragment counts per gene based on gene length and sequencing depth.

Comparatively, 3' mRNA-Seq approaches do not fragment the captured transcripts and sequence multiple fragments from the same transcript. Instead, 3' mRNA-Seq involves a biased library preparation approach that generates reads from the 3' end of mRNA molecules and creates a cDNA library from only the fragment containing the poly-A tail (Fig. 1B, C). This results in an unbiased representation of long and short transcripts in the sequencing library [15]. As a result, gene expression analysis of 3' mRNA-Seq does not require normalization based on gene length. The most important benefit of 3' mRNA-Seq is that gene expression can be quantified with a fraction of the reads necessary in a full transcriptome

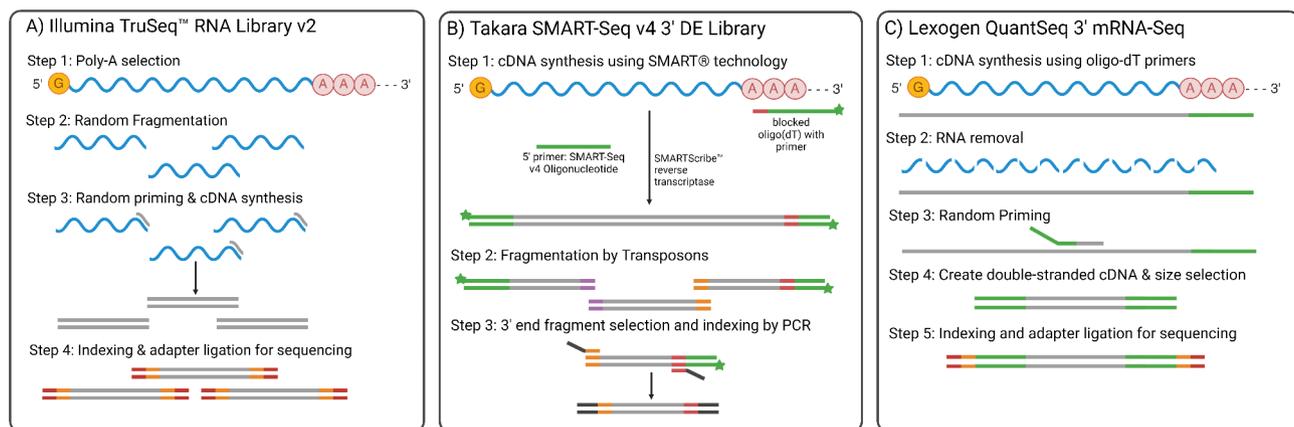


Fig. 1 Comparison of the library preparation approaches for whole transcriptome sequencing and 3' mRNA-Seq. **A)** Library preparation for Illumina TruSeq RNA library preparation used for whole transcriptome sequencing. **B)** Library preparation protocol for Takara SMART-Seq v4 DE approach used for 3' biased sequencing. **C)** Library preparation protocol for Lexogen QuantSeq 3' mRNA-Seq

study due to a reduction in the number of redundant reads that belong to the same transcript [15]. The main drawback of 3' mRNA-Seq is that reads cannot be used to identify novel transcripts or alternative splicing events. Further, it can only call 3' biased intragenic variants in cases where genomic variants are of interest. Multiple commercial library preparation approaches are available for 3' mRNA-Seq, including Tag-seq [17, 18], MustSeq [19], QuantSeq (Lexogen) [20], and SMART-Seq v4 3' DE (Takara). The Lexogen QuantSeq library is one of the most widely used 3' mRNA-Seq libraries across studies, used in over 1,600 publications in the past decade [21].

The literature is conflicting on whether 3' mRNA-Seq or whole transcriptome sequencing approaches are better equipped to detect DEGs [15]. Some side-by-side comparisons found that full transcriptomes were more effective at detecting DEGs [15], while others have identified 3' mRNA-Seq as the superior approach [19, 20, 22, 23]. However, the majority of comparisons suggest that the libraries are equally capable of detecting both DEGs and expressed genes [17, 18, 24, 25, 26, 27]. Finally, compared with whole transcriptome sequencing, 3' mRNA-Seq represents the major cost savings needed for widespread molecular phenotyping applications [15, 18].

This project aimed to identify best practices for carrying out cost-effective 3' mRNA-Seq for eventual application in the molecular phenotyping of livestock. We explored how different library preparation approaches and sequencing depths affect the quality and amount of information generated. Here, we compare two popular library preparation approaches for 3' mRNA-Seq: Lexogen QuantSeq and Takara SMART-Seq v4 3' DE. We also use an iterative downsampling approach to simulate the impacts of different sequencing depths. This allows us to identify the minimum depth needed to capture the optimal amount of variation in transcript quantities. Finally, we explore our ability to call variants from 3' biased sequencing reads. The ability to impute genotypes based on this reduced-representation gene expression data would be invaluable for many genotype-to-phenotype applications in plants and livestock. Coupled with developments in laboratory automation, the best practices identified by this work could make molecular phenotyping practical for livestock genetic evaluations.

Methods

Sample collection and RNA extraction

We opportunistically collected whole blood samples from 15 Holstein male calves housed under climate control rooms [28] in the East Tennessee Research and Education Center– JRTU undergoing an acute heat stress trial for an unrelated project (Yu et al., under revision). We handled all animals according to the University of Tennessee's Institutional Animal Care and Use Committee

Protocol 2851–0921. Fifteen whole blood samples were collected from the cattle at 6:30 before heat exposure, and 14 samples were collected at 18:30 immediately after 12 h of heat exposure from the same animals (Yu et al., under revision). Climate in the room was obtained following our previous work in heat-stressed calves [28]. 10 mL of blood was mixed with 30 mL of 1X NH₄Cl red blood cell lysis buffer and was centrifuged at 2000 ×g for 10 min. The supernatant was aspirated and cell pellets resuspended in 1.2 mL of Trizol. RNA was isolated according to the protocol detailed in Rio et al. [29]. RNA purification and genomic DNA removal were performed using the Zymo RNA Clean and Concentrator kit according to manufacturer protocol.

Library preparation and sequencing

Sequencing libraries were prepared from isolated RNA using two different kits: Takara SMART-Seq v4 3' DE (Takara) and Lexogen QuantSeq 3' (Lexogen) (Supplementary Table S1) as per the manufacturer's instructions. Final library quality and concentrations for pooling were evaluated using the Agilent TapeStation 4200 system. For each kit, libraries for all 27 samples were pooled to achieve equal concentrations (Supplementary Table 1). For the Takara pool, each sample had 5 ng of cDNA. For the Lexogen pool, each sample had 3.2 ng of cDNA. They were sequenced on a single SP flow cell on the Illumina Novaseq6000 (University of Tennessee Genomics Core - Knoxville, TN) with a 200-cycle v1.5 reagent kit. For Takara 3' libraries, Read 1 was 150 bp, and Read 2 consisted of 26 bp for demultiplexing. Whereas Lexogen 3' libraries were sequenced with single-end with 150 bp reads.

Sequence processing & gene expression quantification

Only forward reads were used for sequence analysis, as reverse reads contained only indices for demultiplexing for the Takara kit. For all samples, Trimmomatic (v.0.39) was used for trimming and filtering with the following parameters: (LEADING:5 TRAILING:5 SLIDINGWINDOW:5:20 MINLEN:30) to trim bases from each end of each read that has a quality ≤ 5, trim at the first occurrence of sliding window of size 5 with average quality ≤ 20, and filter out the reads that are shorter than 30 bp after the trimming [30]. The quality of reads before and after trimming & filtering was visualized and evaluated using FastQC (v.0.11.9) [31] and MultiQC (v.1.14) [32] tools. The STAR alignment software (v.2.7.10b) [33] was used to index the *Bos taurus* genome (ARS-UCD2.0; Jul 2023) [34] - which was obtained from NCBI - with sjdbOverhang of 149 bp. STAR was then used to map filtered reads to the indexed genome. The quantMode "GeneCounts" was used in STAR to quantify the number of reads mapped to the annotated genes.

Differential gene expression analysis & functional enrichment analysis

We evaluated the sensitivity of each test condition (library preparation method and downsampled read number) to detect expressed genes and differentially expressed genes. The number of expressed genes (count > 0) and informative genes (count > 10 in 50% of the samples) were calculated in R (R 4.2.1 “Funny-Looking Kid”) using the GeneCounts generated by the STAR tool. All expressed genes (i.e. genes with expression > 0 in at least one sample) were used for DEG analysis via the DESeq function from the DESeq2 R package (v.1.40.2) [35]. Time of sample (morning vs. night) was the only variable tested in the model. No genes were excluded before this step because DESeq function filters genes with low counts based on Negative Binomial Gamma-Poisson distribution and it’s recommended not to remove any expressed genes. The DEGs results were extracted with an FDR-corrected alpha threshold of 0.05 but no log fold change threshold was used. DEGs were plotted using the Bioconductor EnhancedVolcano library [36].

Variant calling

We utilized the GATK best practices for calling variants from RNA-sequencing data to identify genomic variation in areas covered by 3’ mRNA-Seq reads [37, 38]. STAR was used to remap the reads with a per-sample 2-pass mapping step. Samples from the same animal (from before and after heat exposure) were merged using SAMtools (v.1.20) [39]. We used the GATK tool (v.4.3.0.0) [38] and picard tools (v.2.27.4-0) [40] to: (1) Assign all reads from each animal to a single read group identified by the animal ID using the AddOrReplaceReadGroup tool, (2) mark duplicated reads in each of the merged samples, (3) split the reads that contain Ns in their cigar string using SplitNCigarReads tool, (4) Call variants within each of the merged samples using HaplotypeCaller with a confidence threshold of 20, (5) combine the variants from all samples across all chromosomes and known scaffolds in the genome using GenomicsDBImport, (6) perform joint variant calling from all samples using GenotypeGVCFs, (7) Separate SNPs and INDELS using SelectVariants, and (8) Hard-filter SNPs and INDELS, separately, with VariantFiltration tool with the following parameters for SNPs and INDELS (QD < 20.0, FS > 20.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, and SOR > 5.0). Even though the GATK recommends variant recalibration before calling variants with HaplotypeCaller tool rather than using hard filters, this step requires having one or more databases of known polymorphic sites from the population of interest. At the time of the analysis, reference polymorphic sites databases were not available for Holstein dairy cattle compared to the reference Hereford genome assembly (ARS-UCD2.0). The information fields of VCF files for

SNPs and INDELS were imported into R environment (R version 4.2.1 (2022-06-23)) and filtered based on either (1) total depth across samples and alleles and the number of reads supporting an alternate allele [DP_total >= 10 and ALT_AD_total > 0] or (2) only on the depth of alternative allele (AD) or alternative allele ratio (AR) per allele per sample followed by a filter on re-calculated depth from all samples [((AD >= 2) or (AR >= 0.1 and DP > 10)) and DP_total_recalculated >= 5]. The second filtering method is more stringent and filtered out more SNPs. To validate the variant calling results, we compared the called variants with known species variants (i.e. population variants) obtained from the ENSEMBL database (as of August 29, 2024).

Downsampling for sequencing-depth benchmarking

We used the Seqtk tool (v.1.4) [41] to generate ten random replicates of downsampled read sets simulating seven ascending sequencing depths from raw FASTQ files for each sample. We simulated this downsampling using ten random seeds (127, 2, 5, 7, 9, 11, 12, 81, 21, 47) at each of seven different sequencing depths: 0.5 M, 1 M, 2 M, 5 M, 7.5 M, 10 M, and 12 M reads/sample. Higher sequencing depth downsampling past 12 M reads was not possible for all samples. We performed the same analysis described above (alignment, gene expression counting, DEG analysis, and variant calling) to quantify the relative performance of each sequencing depth across replicates. The downsampling efficacy of the variant calling was also explored using the more stringent filtering criteria (based on depth per alternative allele per sample).

The results of the downsampling were quantified using precision, recall, and F-score where the whole dataset’s results were considered the true positive set. Precision is the measure of the positive predicted value (PPV), recall is the measure of sensitivity, and F-score is the harmonic mean of precision and recall [42]. The three metrics were calculated using the following equations:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F.score = \frac{TP}{TP + \frac{1}{2} \times (FP + FN)}$$

Where.

- True positive (TP) is the number of genes present in both the subsample and the full dataset.
- False positive (FP) is the number of genes present only in the subsample.

- False negative (FN) is the number of genes present only in the full dataset but not the subsample.

The plateau value, indicating the point where additional reads failed to uncover additional genes for each metric, was estimated by fitting the downsampling results from all ten replicates to a Self-Starting NLS Asymptotic Regression Model (SSasypm function from the stats package in R) with the following equation: $y = y_{max} + (y_0 - y_{max}) \cdot e^{-exp^{lrc} \cdot x}$, Where

- y_{max} is the asymptote value on the y-axis,
- y_0 is the response when $x = 0$, and
- lrc is the natural logarithm of the rate constant.

The plateau’s initial point (i.e., saturation sequencing depth) was calculated by setting the derivative of this model to 0.0001, which represents an increase of one gene from adding 10,000 extra raw reads per sample (i.e., marginal information gain).

$$x_{plateau} = - \frac{\ln \left(\frac{-0.0001}{y_0 - y_{max}} \right) - lrc}{e^{lrc}}$$

Results

Comparing Takara and Lexogen libraries

Sequencing depth was similar for both library preparation kits (p-value 0.9761; Welch two-sample t-test)

with a median of 18.6 million (range: 12.5 M – 35.6 M) and 16.4 million (range: 7.3 M – 95.7 M) raw reads per sample for the Takara and Lexogen libraries, respectively (Supplementary Figure S1). One sample sequenced with the Lexogen library had significantly more reads than all other samples (>5 standard deviations from the mean) because it was inadvertently loaded to a higher concentration. On average, quality control filters dropped significantly fewer reads for the Takara library (1.94% of the raw reads) compared with the Lexogen library (2.39% of the raw reads) (p-value < 0.0001; Welch two-sample t-test) (Fig. 2A). When mapped to the cattle reference genome, the reads from Lexogen libraries had a significantly higher mapping rate on average, with 87.95% of the raw reads (90.1% of the filtered reads) where uniquely mapped compared to Takara libraries with 75.56% of the raw reads (77.05% of the filtered reads) uniquely mapped (p-value < 0.0001; Welch two-sample t-test) (Fig. 2B, C). The low mapping rate was mainly driven by more reads in the Takara data being unmapped because they were too short (5–15% of filtered reads). More stringent filtering criteria would decrease the rate of unmapped reads from both sequencing libraries but at the cost of possible information loss.

Despite Lexogen libraries having a higher mapping rate, Takara libraries captured significantly more expressed genes than Lexogen libraries (p-value < 0.0001; Welch two-sample t-test). We considered “expressed genes” as genes to which at least a single read aligned. Takara

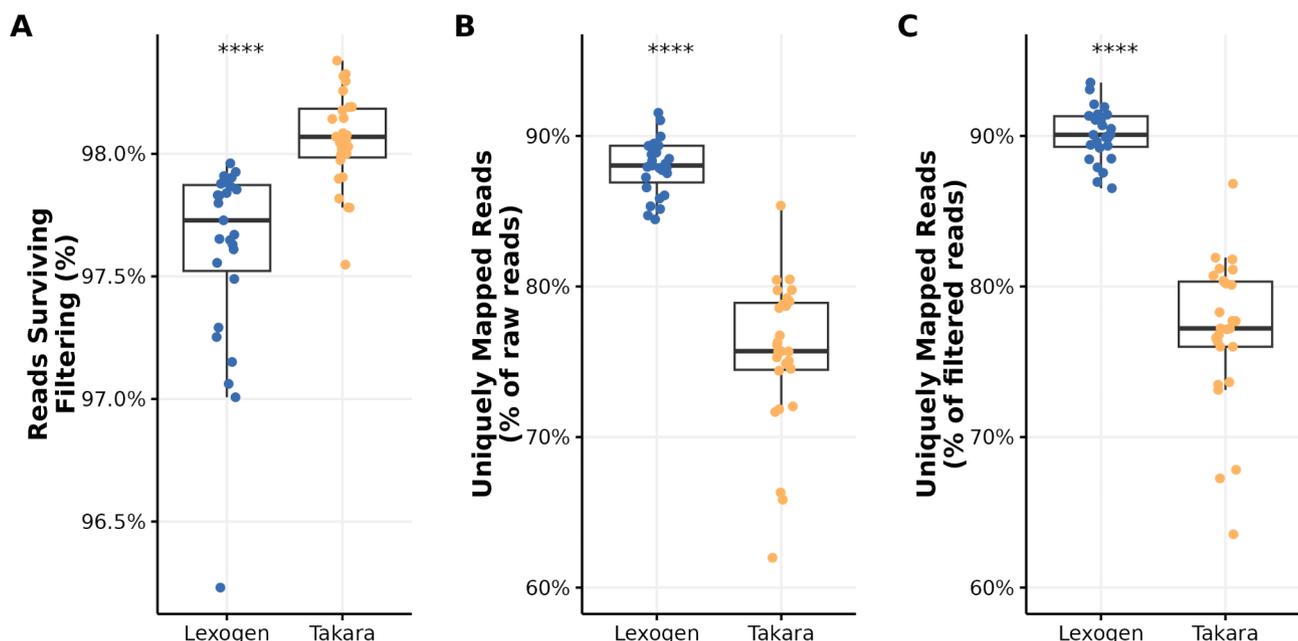


Fig. 2 Comparison of Takara and Lexogen filtering outcomes and mapping rates. **(A)** Boxplots represent the percentage of the raw reads that survived filtering from both sequencing libraries. **(B)** Boxplots representing the percentage of the raw reads that survived filtering and were uniquely mapped to the cattle reference genome. **(C)** Boxplots represent the percentage of the filtered genes that were uniquely mapped to the reference genome from both sequencing libraries. (****, p < 0.0001)

libraries detected an average of 16,957 expressed genes per sample (minimum = 15,921; maximum = 18,532), while Lexogen libraries averaged only 13,417 (minimum = 9,273; maximum = 16,033) (Fig. 3A). Using both libraries, the number of expressed genes from all animals (Fig. 3A black stars) is greater than the number

of expressed genes from any individual, with a total of 24,607 expressed genes from Takara libraries and 21,730 expressed genes from Lexogen libraries. Our definition of an expressed gene was quite liberal, meaning that many were captured only in one or a handful of individuals, which could mean they would not be as informative as

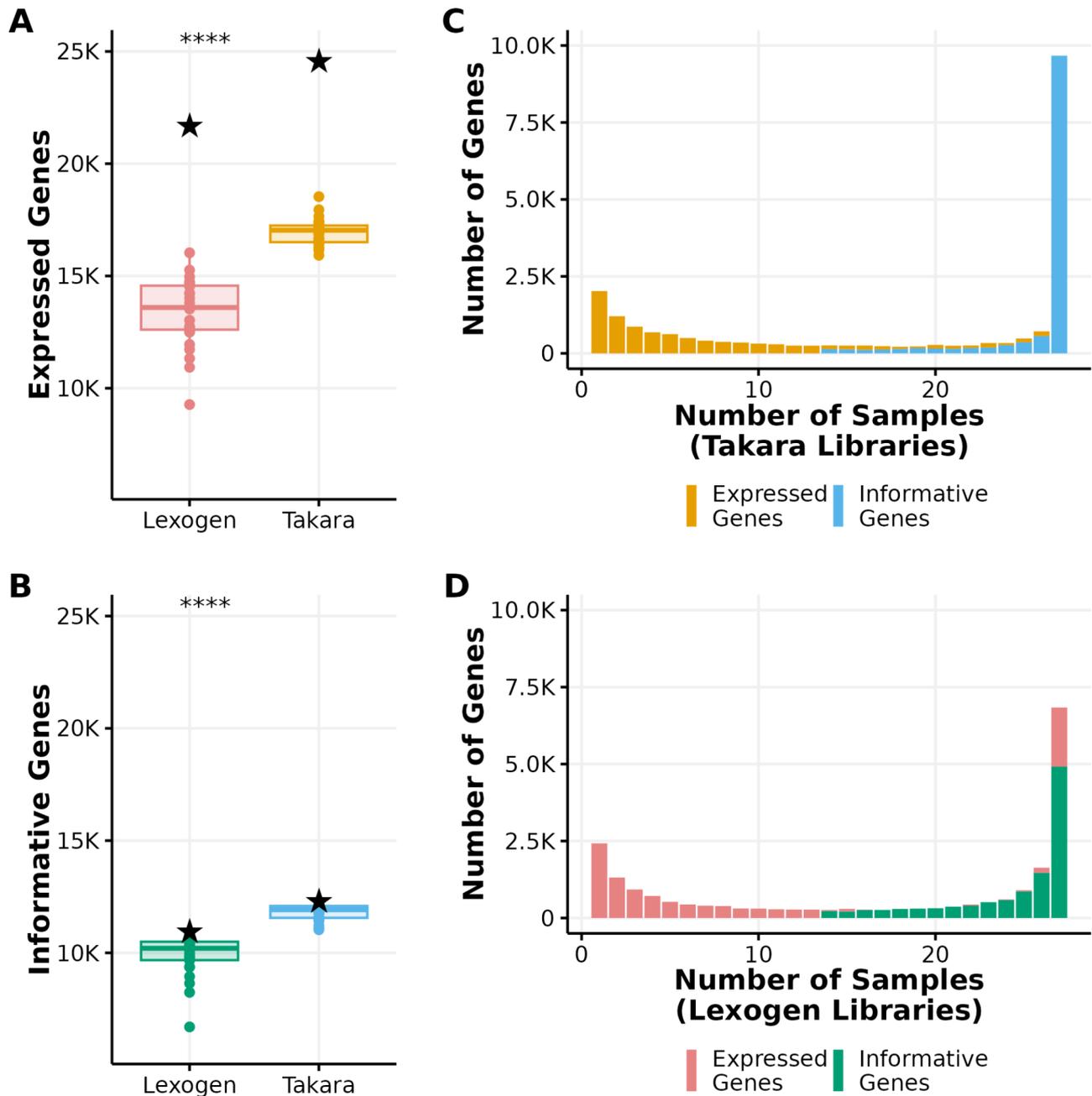


Fig. 3 Comparison of expressed and informative genes detected by Takara and Lexogen library preparation kits. **(A)** Boxplots representing the numbers of expressed genes (genes with at least one read mapped) for Takara and Lexogen libraries. Each dot represents a sample, and the black star represents the number of expressed genes identified in the full dataset for each library preparation method. **(B)** Boxplots representing the number of informative genes (genes with at least ten mapped reads in 50% of the samples) for Takara and Lexogen libraries. Each dot represents a sample, and the black star represents the number of informative genes present in at least 50% of the samples identified in the full dataset. **(C, D)** Distribution of the number of expressed genes and informative genes captured by *Takara libraries* **(C)** and *Lexogen libraries* **(D)** as a function of the number of individuals in which these genes were identified. (****: $p < 0.0001$). The colors in subfigures **(A)** and **(B)** match the colors in subfigures **(C)** and **(D)**, respectively

molecular phenotypes (Fig. 3C, D). We defined a more conservative set of “informative genes” as genes where at least ten reads mapped to the gene in at least 50% of the samples. Takara libraries captured significantly more informative genes per sample (mean = 11,824) than Lexogen libraries (mean = 9,904) (p -value < 0.0001; Welch two-sample t -test) (Fig. 3B). Takara libraries captured 12,351 informative genes from all samples, whereas Lexogen libraries captured only 10,997 informative genes (Fig. 3B black stars).

When we modeled the heat stress response between samples, Takara enabled us to identify more DEGs than Lexogen. Takara libraries identified 4,821 DEGs (3,142 up-regulated genes (13% of the expressed genes) and 1,679 down-regulated genes (6.8% of the expressed genes)), whereas Lexogen libraries captured only 1,285 DEGs (1,025 up-regulated genes (4.9% of the expressed genes) and 260 down-regulated genes (1.2% of the expressed genes)) (Fig. 4). Only 1,095 DEGs were consistently detected using both libraries (964 up-regulated genes and 131 down-regulated). Most of these DEGs had a small effect size as only 179 of the up-regulated genes had $|LFC| \geq 1$ from the Takara libraries and only 47 up-regulated genes and 2 down-regulated genes from Lexogen libraries had $|LFC| \geq 1$ (not shown). 94% of the upregulated genes and 50% of the downregulated genes captured by Lexogen were also captured by Takara libraries (Fig. 4C).

We called more genetic variants using reads from the Takara libraries than from Lexogen libraries. Also, the more stringent filtering criteria results in less variants, both SNPs and INDELs, compared to the less stringent criteria from both libraries. After variant filtration, we obtained 183,475 and 234,140 SNPs and 19,692 and 34,882 INDELs from Takara libraries using the stringent and less stringent filtering criteria, respectively. By comparison, Lexogen libraries identified 113,762 and 138,380 SNPs and 11,193 and 20,705 INDELs using the two filtering approaches (Fig. 5A). For Takara libraries using the more stringent filtering criteria, 147,643 SNPs (80.5% of the called Takara SNPs) were intragenic (overlapping with 14,953 annotated genes), with the majority (69.3%) of the SNPs overlapping with informative genes (Fig. 5A). This compares with 104,742 SNPs (92.1% of the called Lexogen SNPs) intragenic SNPs (overlapping with 9,842 annotated genes) from Lexogen libraries with the majority (85.2%) of the SNPs overlapping with informative genes (Fig. 5A). For the less stringent filtering criteria, 198,103 SNPs called from Takara libraries (84.6% of the SNPs) were intragenic (overlapping with 15,641 annotated genes), with the majority (75.6%) of the SNPs overlapping with informative genes (Fig. 5A). Compared with 127,085 SNPs (91.8% of the called Lexogen SNPs) intragenic SNPs (overlapping with 10,846 annotated genes)

from Lexogen libraries with the majority (84.9%) of the SNPs overlapping with informative genes (Fig. 5A).

When we compared the called variants with known population variants, 131,835 and 144,992 SNPs (61.9 and 71.9% of total SNPs) called from Takara libraries were known true population variants using the more and less stringent filtering, respectively (Fig. 5B). Lexogen libraries called far fewer known variants. These variants represented only 31.5 and 35% of the total number of called SNPs using more and less, stringent filtering (Fig. 5B). This makes Takara libraries much better suited for downstream imputation applications.

From both sequencing libraries, less than 0.5% of the INDELs were known true population variants in the ENSEMBL database. Interestingly, there is minimal overlap in the true population variants called by both sequencing libraries (12.3%), as the majority of known variants are called by Takara libraries (76%). From both libraries, a higher density of SNPs was called near the 3' end of the gene body, as we would expect given the library preparation approaches (Fig. 5C). However, the location of the reads from both libraries does not exactly match for all expressed genes which explains the incomplete overlap in the true variants called.

Identifying an optimal read depth for 3' mRNA-Seq in molecular phenotyping applications

After identifying Takara as the superior library preparation approach, we performed a series of read number-downsampling iterations to determine the optimal sequencing depth that balanced capturing expression variability while minimizing costs. The unique mapping rate and filtering rate were the same across all downsampling iterations, as it is driven by sample quality rather than read number. (Supplementary Figure S2 A, B). However, the number of expressed and informative genes per sample increased with sequencing depth (Supplementary Figure S2 C, D). The number of expressed genes and informative genes captured by all replicates increased exponentially as the sequencing depth increased (Fig. 6A, B). For expressed genes, asymptotic regression models plateau at 7,993,318 reads per sample, capturing 23,605 predicted expressed genes (turquoise dot in Fig. 6A). For informative genes, the model starts to plateau at 8,357,463 reads per sample, capturing 11,137 predicted informative genes (turquoise dot in Fig. 6B). The model begins to plateau at 6,565,915 reads per sample for DEGs, capturing 4,297 DEGs (turquoise dot in Fig. 6C) out of 4,821 DEGs captured by the whole dataset. The model predicted that using the full number of reads generated (12,601,460) would only capture an additional 235 DEGs.

The depth of sequencing also affected the number of variants that we were able to call from 3' mRNA-Seq reads. Unlike gene expression, the number of filtered

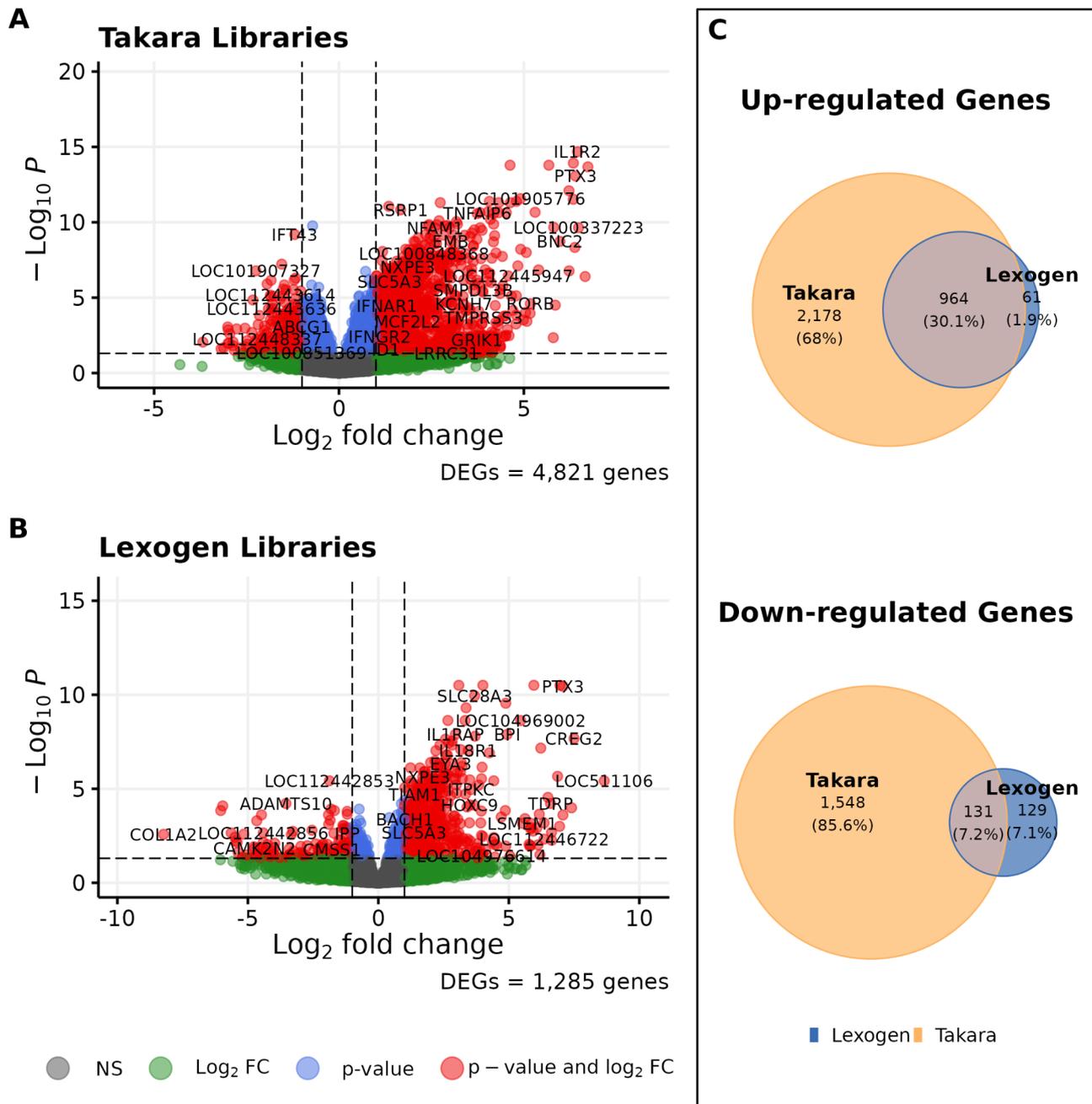


Fig. 4 Volcano plots representing the differentially expressed genes of heat-stressed calves using (A) Takara libraries and (B) Lexogen libraries. The horizontal dashed line represents the adjusted p-value threshold of 0.05; all genes above the line are statistically significant. The vertical dashed lines represent $|LFC| \geq 1$; all genes beyond the vertical lines have moderate-to-high effect size as a response to heat stress. Differentially expressed genes (red dots) are both statically significant and have moderate-to-high effect sizes. (C) Number of DEGs detected by both library kits

variants did not plateau at any sequencing depth, including at the full 12.5 M reads per sample (Fig. 6D, E). We could not call variants at higher sequencing depth as only nine samples had a sequencing depth ≥ 20 M, and only two samples had ≥ 30 M reads. As expected, our analysis showed that the number of samples is as important as the number of reads in accurately calling variants. Using our limited dataset, the asymptotic regression

model extrapolates the plateau to be around 83,685,731 reads per sample for calling SNPs and 45,168,229 reads per sample for calling INDELS; however, extrapolation of regression models beyond the range of observed data should be approached with caution, as the reliability of predictions in these regions is often questionable. Using the optimal sequencing depth identified for identifying expressed and informative genes 8 million reads

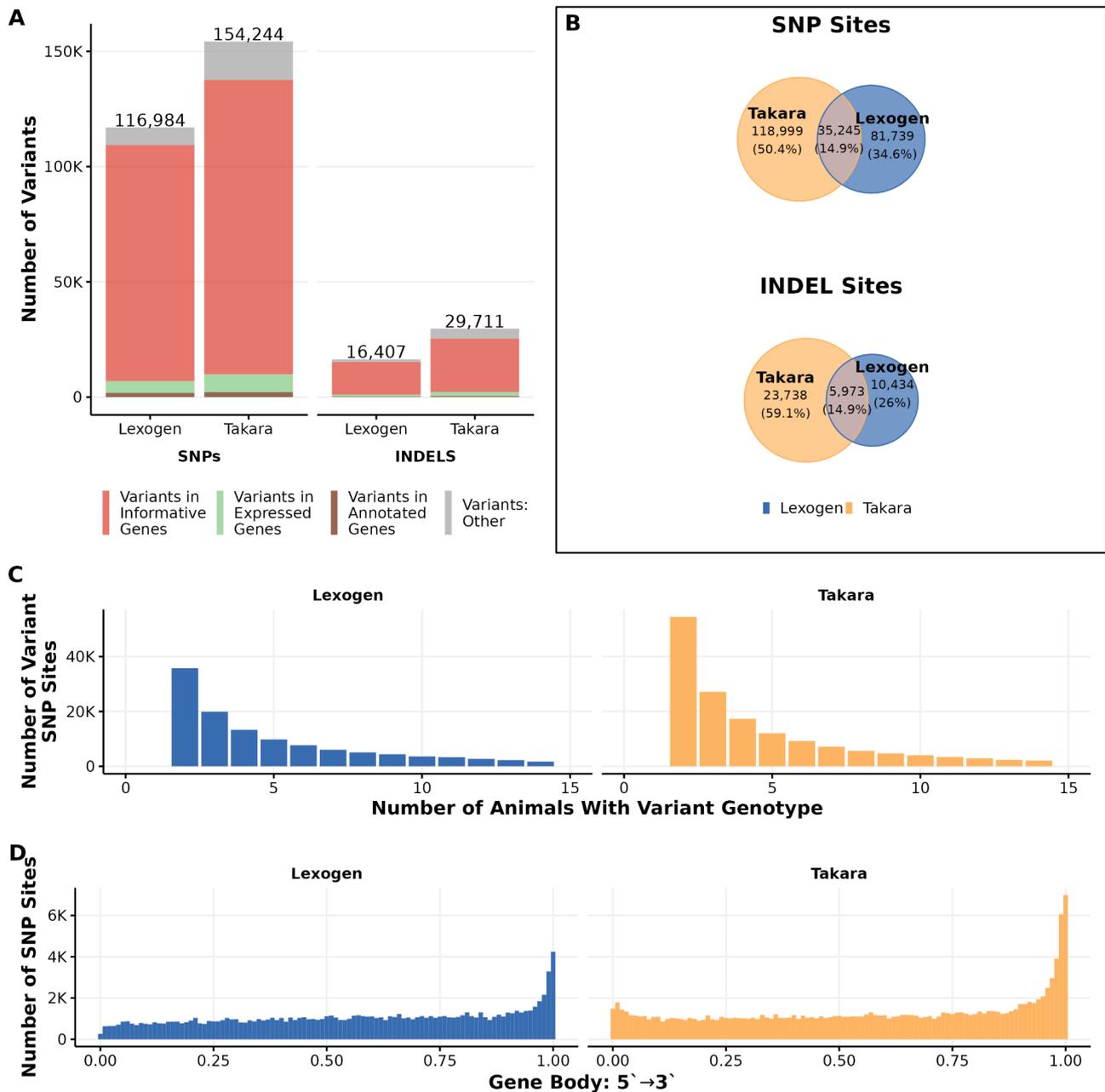


Fig. 5 Variant calling results from Takara and Lexogen libraries. **(A)** Bar chart depicting differences in the number and location of variant sites. **(B)** Venn diagrams show the number and percentage of the variants called from the reads of each library preparation approach that are true population variants based on the ENSEMBL known variants GVF file. The two Venn diagrams on the left are when the variants are filtered more stringently (Alternate Allele Depth ≥ 10), and the two Venn diagrams on the right are when SNPs are filtered less stringently (Depth > 10 and Alternate Allele Depth ≥ 1). **(C)** A histogram representing the distribution of called intragenic SNPs across the gene body from 5' \rightarrow 3' of the gene

per sample), we called 93,915 SNPs and 9,725 INDEL in the dataset. This represented 51.2% and 49.5% of total SNP and INDELS called with the complete dataset, respectively.

Discussion

Our study demonstrates the potential of using 3' mRNA sequencing (3' mRNA-Seq) as a cost-effective method for molecular phenotyping in cattle. We primarily focus on identifying superior library preparation kits and optimizing sequencing depth to balance information content and cost. Our findings indicate that the Takara SMART-Seq v4 3' DE library preparation kit outperforms the

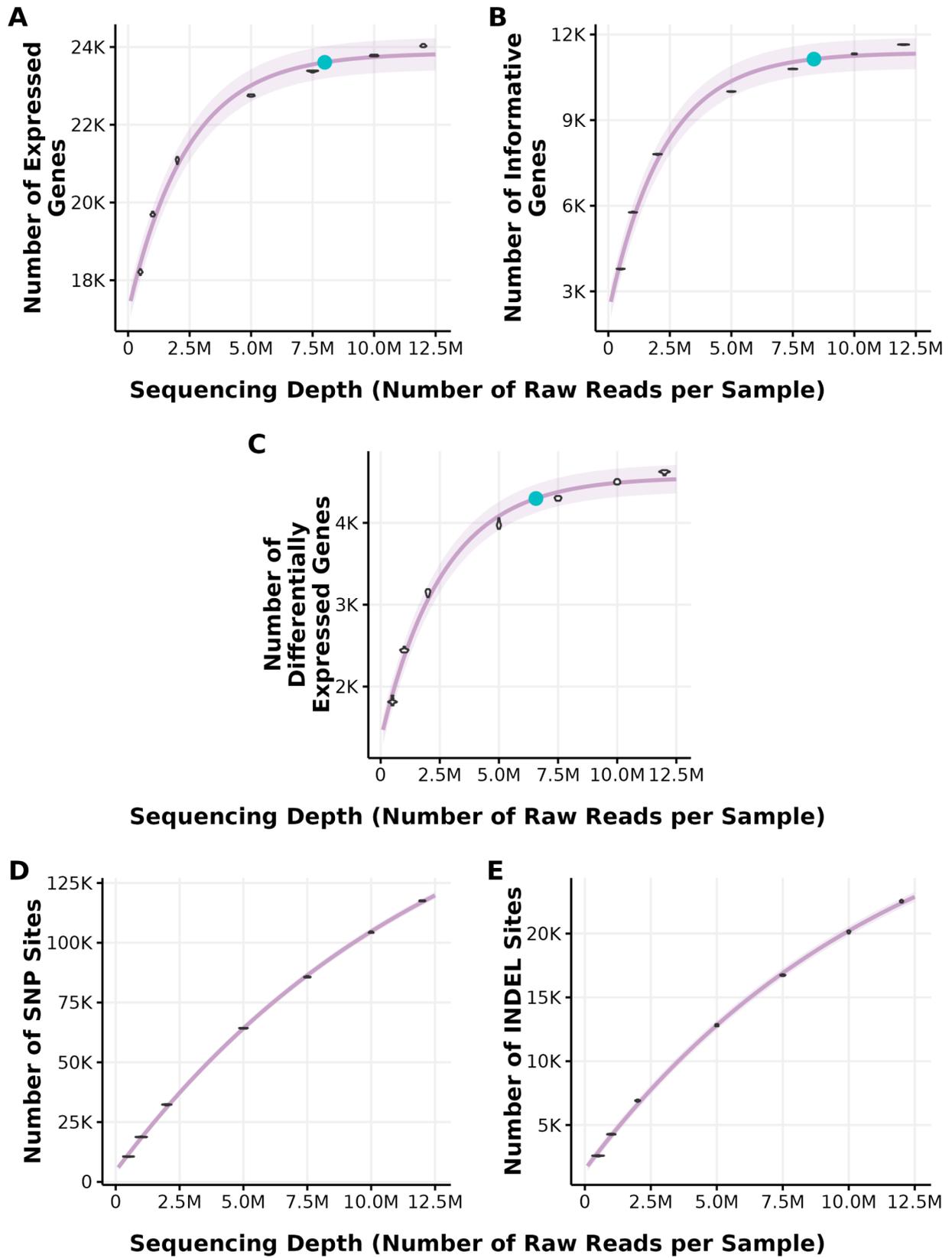


Fig. 6 (See legend on next page.)

(See figure on previous page.)

Fig. 6 The effect of sequencing depth on capturing expressed genes, informative genes, DEGs, and variants. Each sequencing depth downsample consisted of ten repetitions (a cluster of 10 points from the downsampling subsets). Lilac lines are non-linear asymptotic regression models; each parameter was fit to a regression model independently. Turquoise dots represent the start of the plateau, defined by a slope of 0.0001, representing a gain of 1 gene/variant when increasing the sequencing depth by 10,000 reads per sample. **(A)** Number of expressed genes as a function of sequencing depth per sample. **(B)** Number of informative genes (expressed genes with at least 10 reads mapped to in at least 50% of the samples) as a function of sequencing depth per sample. **(C)** Number of differentially expressed genes (DEGs) as a function of sequencing depth per sample. **(D)** Number of SNPs called as a function of sequencing depth per sample. **(E)** Number of INDELS called as a function of sequencing depth per sample

Lexogen QuantSeq kit across various metrics, including the number of quality reads, expressed genes detected, informative genes detected, DEGs identified, and intra-genic variants called. The lone area where the Lexogen approach outperformed Takara was in the proportion of uniquely mapped reads that survived filtering. There was a wider range of mapping rates in the Takara samples, which was mostly caused by a high percentage of reads too short to be mapped (5–15% of total filtered reads). We could not rule out the effect of different cDNA concentration pooled for sequencing on the quality and amount of the sequenced reads as Takara and Lexogen libraries were pooled with slightly different cDNA concentrations.

Overall, this cost-effective 3'-biased approach to sequencing can both increase sample sizes in differential gene expression analysis & eQTL studies and enable population-level molecular phenotyping. We anticipate that gene expression phenotypes may be useful as high-dimensional indicators for other hard-to-measure traits in cattle, such as methane emissions, metabolic efficiency, reproductive predisposition, or disease susceptibility. However, these 3' mRNA-Seq technologies have limitations such as the inability to capture alternative splicing events. Similar to total mRNA-Seq, 3' mRNA-Seq is also incapable of capturing intergenic polymorphic sites which reduces its power in eQTL analysis and similar analyses in the absence of truth set of population haplotypes.

In this work, we primarily evaluated 3' mRNA-Seq library preparation approaches and optimal sequencing depth by the number of informative genes detected. We defined informative genes as those with at least ten mapped counts that were detected in at least 50% of the samples. While detecting all expressed genes is important, this core set of genes will be necessary for extrapolating latent phenotypes across the wider population. We found that Takara libraries captured a significantly greater number of informative genes compared to Lexogen libraries. In the DEGs analysis, Takara libraries identified 74% more DEGs compared to Lexogen libraries. Of the total number of DEGs identified by both methods, the Lexogen libraries identified only 14% of DEGs that were not also identified by Takara. In addition to the comparative performance of the library preparation kits, our analysis of the sequencing depth revealed that the Takara SMART-Seq v4 3' DE library preparation kit's

sensitivity to capture expressed genes and informative genes from cattle whole blood saturates at a sequencing depth of around 8 million reads per sample. We observed that increasing the read depth beyond 8 million reads provided only marginal gains in the recall of gene expression metrics, suggesting a point of diminishing returns in terms of cost versus data richness. In contrast, *Xiong et al.* found that the sensitivity of the 3' mRNA-Seq library to detect expressed genes saturates at 2–3 million reads per sample from human primary cardiomyocyte cell lines [24]. That said, the optimal sequencing depth is likely species- and tissue-specific, reflecting transcription activity in different types of cells and organisms.

The ability to call variants from 3' mRNA-Seq reads also demonstrated a strong dependence on library preparation method, filtering stringency, sequencing depth, and number of samples. Reads from these samples would certainly be adequate for assigning parentage, and likely for performing accurate genotype imputation for intergenic regions when adequate reference panels are available. Further work that directly evaluates the impacts of these variables on imputation accuracy is forthcoming, but outside the scope of this paper. Our results in a small set of animals showed that while gene expression metrics plateaued at lower read depths, variant identification continued to benefit from increased sequencing depth without reaching an apparent plateau within the range tested. Further sequencing would likely result in greater proportions of gene bodies to be sequenced, resulting in more variants.

Our study highlights the importance of carefully optimizing sequencing protocols to balance cost and data quality. Using 3' mRNA-Seq with the Takara SMART-Seq v4 3' DE kit offers a promising approach for large-scale molecular phenotyping in livestock, providing sufficient data quality at a substantially lower cost than whole transcriptome sequencing. Using fractional quantities of reagents for library preparations could provide even further reductions in cost. This approach enables the capture of large-scale gene expression data and may also facilitate genotype imputation, thereby enhancing the utility of molecular data in genetic evaluations and breeding programs.

Conclusion

In conclusion, we show that 3' mRNA-Seq is a cost-efficient (<\$25/sample) approach to studying and representing complex traits in cattle through phenotyping by gene expression. In our samples, the Takara SMART-seq v4 library was superior to the Lexogen QuantSeq library in capturing expressed, informative, and DEGs, as well as calling sequence variants. We have also shown that 8 million reads per sample effectively capture most of the inter-sample variation in gene expression with a marginal increase in the number of expressed genes with increasing read depth.

Abbreviations

DEGs	Differentially Expressed Genes
FP	False Positive
FN	False Negative
GATK	Genome Analysis Toolkit
GO	Gene Ontology
INDEL	Insertions/Deletions
MAF	Minor Allele Frequency
SNP	Single Nucleotide Polymorphism
TP	True Positive

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-025-11571-4>.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5

Acknowledgements

The authors thank the UTIA Genomics Center for the Advancement of Agriculture for providing wet lab and computational support, specifically from Trevor Freeman. Computing for the project was performed on the ISAAC Next Generation cluster hosted by the University of Tennessee's High Performance Scientific Computing group.

Author contributions

T.R. and J.B. conceptualized the experiment. A.R., S.M., and T.S. performed the heat stress experiment and collected tissues. T.S., S.M., and J.B. generated sequencing data. R.M. and T.R. analyzed data and generated figures. R.M. and T.R. drafted main manuscript text. All authors reviewed the manuscript.

Funding

This work was funded in part by an AG2PI Round 2 Seed Grant. The AG2PI project is supported by USDA-NIFA awards 2020-70412-32615 and 2021-70412-35233. Additional funding from FFAR Grant No. 22-000087 supported the execution of this study. Funding bodies had no role in study design, data collection, analysis, or manuscript preparation.

Data availability

Gene expression reads and associated metadata have been deposited in the Gene Expression Omnibus as GEO project GSE272596.

Declarations

Ethics approval and consent to participate

This study was approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Tennessee, Knoxville, under approval number 2851-0921. All procedures were conducted in accordance with IACUC ethical guidelines.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 19 July 2024 / Accepted: 4 April 2025

Published online: 16 April 2025

References

- Jonas E, de Koning D-J. Genomic selection needs to be carefully assessed to Meet specific requirements in livestock breeding programs. *Front Genet.* 2015;6.
- Henderson CR. Selection Index and Expected Genetic Advance. *Statistical Genetics and Plant Breeding.* 1963;141–63.
- Henderson CR. Best linear unbiased Estimation and prediction under a selection model. *Biometrics.* 1975;31:423–47.
- Dekkers JCM, Mathur PK, Knol EF. Genetic improvement of the pig. In: *The Genetics of the Pig.* 2nd edition. CAB International; 2011. pp. 390–425.
- Hayes BJ, Donoghue KA, Reich CM, Mason BA, Bird-Gardiner T, Herd RM, et al. Genomic heritabilities and genomic estimated breeding values for methane traits in Angus cattle. *J Anim Sci.* 2016;94:902–8.
- Berry DP, Bermingham ML, Good M, More SJ. Genetics of animal health and disease in cattle. *Ir Vet J.* 2011;64:5.
- Pryce JE, Gonzalez-Recio O, Nieuwhof G, Wales WJ, Coffey MP, Hayes BJ, et al. Hot topic: definition and implementation of a breeding value for feed efficiency in dairy cows. *J Dairy Sci.* 2015;98:7340–50.
- Tuggle CK, Clarke JL, Murdoch BM, Lyons E, Scott NM, Beneš B, et al. Current challenges and future of agricultural genomes to phenomes in the USA. *Genome Biol.* 2024;25:8.
- Fontanesi L. Merging metabolomics, genetics, and genomics in livestock to dissect complex production traits. In: Kadarmideen HN, editor. *Systems biology in animal production and health.* Cham: Springer International Publishing; 2016. pp. 43–62.
- Runcie DE, Qu J, Cheng H, Crawford L. MegaLMM: Mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome Biol.* 2021;22:213.
- Legarra A, Christensen OF. Genomic evaluation methods to include intermediate correlated features such as high-throughput or omics phenotypes*. *JDS Commun.* 2023;4:55–60.
- Zhao T, Zeng J, Cheng H. Extend mixed models to multilayer neural networks for genomic prediction including intermediate omics data. *Genetics.* 2022;221.
- Li Z, Gao N, Martini JWR, Simianer H. Integrating gene expression data into genomic prediction. *Front Genet.* 2019;10:126.
- McParland S, Lewis E, Kennedy E, Moore SG, McCarthy B, O'Donovan M, et al. Mid-infrared spectrometry of milk as a predictor of energy intake and efficiency in lactating dairy cows. *J Dairy Sci.* 2014;97:5863–71.
- Ma F, Fuqua BK, Hasin Y, Yukhtman C, Vulpe CD, Lusk AJ, et al. A comparison between whole transcript and 3' RNA sequencing methods using Kapa and Lexogen library Preparation methods. *BMC Genomics.* 2019;20:9.
- Mandelbloum S, Manber Z, Elroy-Stein O, Elkon R. Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias. *PLOS Biol.* 2019;17.
- Meyer E, Aglyamova GV, Matz MV. Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. *Mol Ecol.* 2011;20:3599–616.
- Lohman BK, Weber JN, Bolnick DI. Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. *Mol Ecol Resour.* 2016;16:1315–21.
- Mai L, Qiu Y, Lian Z, Chen C, Wang L, Yin Y, et al. MustSeq, an alternative approach for multiplexable strand-specific 3' end sequencing of mRNA transcriptome confers high efficiency and practicality. *RNA Biol.* 2021;18:232–43.

20. Moll P, Ante M, Seitz A, Reda T. QuantSeq 3' mRNA sequencing for RNA quantification. *Nat Methods*. 2014;11:i–iii.
21. Ivin M. 10 Years of QuantSeq: applications of 3' mRNA-Seq in inspiring scientific discoveries. Lexogen GmbH. 2024. <https://www.lexogen.com/10-years-of-quantseq/>
22. Tandonnet S, Torres TT. Traditional versus 3' RNA-seq in a non-model species. *Genomics Data*. 2016;11:9–16.
23. Janjic A, Wange LE, Bagnoli JW, Geuder J, Nguyen P, Richter D, et al. Prime-seq, efficient and powerful bulk RNA sequencing. *Genome Biol*. 2022;23:88.
24. Xiong Y, Soumillon M, Wu J, Hansen J, Hu B, van Hasselt JGC, et al. A comparison of mRNA sequencing with random primed and 3'-Directed libraries. *Sci Rep*. 2017;7:14626.
25. Jarvis S, Birsa N, Secrier M, Fratta P, Plagnol V. A comparison of low read depth QuantSeq 3' sequencing to total RNA-Seq in FUS mutant mice. *Front Genet*. 2020;11.
26. Crow TM, Gill JA, Whitehead A, Runcie DE. 3' RNA sequencing does not increase power or reduce costs for gene expression analysis. 2022;2022.04.13.488225.
27. Song Y, Milon B, Ott S, Zhao X, Sadzewicz L, Shetty A, et al. A comparative analysis of library Prep approaches for sequencing low input transcriptome samples. *BMC Genomics*. 2018;19:696.
28. Rius AG, Kaufman JD, Li MM, Hanigan MD, Ipharraguerre IR. Physiological responses of Holstein calves to heat stress and dietary supplementation with a postbiotic from *Aspergillus oryzae*. *Sci Rep*. 2022;12:1587.
29. Rio DC, Ares M, Hannon GJ, Nilsen TW. Purification of RNA using trizol (TRI Reagent). *Cold Spring Harb Protoc*. 2010;2010.pdb.prot5439.
30. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30:2114–20.
31. Bittencourt SA AS, Bittencourt a. S. FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinf*. 2010.
32. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32:3047–8.
33. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
34. Vanvanhossou SFU, Yin T, Scheper C, Fries R, Dossa LH, König S. Unraveling admixture, inbreeding, and recent selection signatures in West African Indigenous cattle populations in Benin. *Front Genet*. 2021;12:657282.
35. Love MI, Huber W, Anders S. Moderated Estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
36. Blighe K, Rana S, Lewis M, EnhancedVolcano. Publication-ready volcano plots with enhanced colouring and labeling. R package version 1.22.0. <https://doi.org/10.18129/B9.bioc.EnhancedVolcano>
37. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8.
38. Van der Auwera GA, O'Connor BD. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. 1st edition. O'Reilly Media; 2020.
39. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of samtools and BCFtools. *GigaScience*. 2021;10:giab008.
40. Broad Institute. Picard Toolkit. *GitHub Repository*. 2019.
41. Li H. Seqtk. *GitHub Repository*. 2023.
42. Williams CR, Baccarella A, Parrish JZ, Kim CC. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*. 2017;18:38.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.