Development and comparison of RNA-Sequencing pipelines for more accurate SNP identification: Practical example of functional SNP detection associated with feed efficiency in Nellore beef cattle.

S. Lam
University of Guelph
J. Zeidan
University of Guelph
F. Miglior
University of Guelph
A. Suárez-Vega
University of Guelph
I. Gómez-Redondo
Instituto Nacional de Investigacion y Tecnologia Agraria y Alimentaria
P. A. S. Fonseca
University of Guelph
L. L. Guan
University of Alberta
S. Waters
Teagasc Food Research Centre Ashtown
A. Cánovas (acanovas@uoguelph.ca)
University of Guelph https://orcid.org/0000-0002-0036-0757

Research article

Keywords: RNA-sequencing pipeline, optimization, SNPs, feed efficiency, Nellore beef cattle, beef production

Posted Date: January 28th, 2020

DOI: https://doi.org/10.21203/rs.2.22101/v1

Version of Record: A version of this preprint was published on October 8th, 2020. See the published version at https://doi.org/10.1186/s12864-020-07107-7.

Abstract

Optimization of an RNA-Sequencing (RNA-Seq) pipeline can maximize power and accuracy for identifying genetic variants, including SNPs, which may serve as genetic markers to select for feed efficiency, leading to economic benefits for beef production. This study determined an optimized pipeline for variant detection using a dataset with multiple samples and tissues. The RNA-Seq data (GEO Accession ID: PRJEB7696 and PRJEB15314) from muscle and liver tissue, respectively, from 12 Nellore beef steers selected from 585 steers with residual feed intake measures (RFI; n=6 low-RFI, n=6 high-RFI) were used. Three RNA-Seq pipelines were compared including multi-sample calling from i) non-merged samples; ii) merged samples by group for low-RFI and for high-RFI for each tissue, iii) merged samples by group and tissue for low- and high-RFI for both tissues. The RNA-Seq reads were aligned against the UMD3.1 bovine reference genome (release 94) assembly using STAR. Variants were called using BCFtools and variant effect prediction (VeP) and functional annotation (ToppGene) analyses were performed. Approaches were compared by comparing read depth, overlap of SNP detection results, and following SNP annotation for positional candidate genes. On average, total reads detected for Approach i) individual liver and muscle samples were 18,362,086.3 and 35,645,898.7, respectively. For Approach ii), total reads detected for each merged group of samples was 162,030,705, and for Approach iii) was 324,061,410, revealing the highest read depth. Additionally, Approach iii) encompassed the majority of localized positional genes detected by each approach, suggesting Approach iii) be applied to maximize detection power, read depth, and accuracy of SNP calling, therefore increasing confidence of variant detection and reducing false positive rate. Approach iii) was used to detect unique SNPs fixed within low- (12,145) and high-RFI (14,663) groups. Annotation of moderate to high functional impact SNPs revealed co-localized positional candidate genes for each RFI group (2,886 for low-RFI, 3,075 for high-RFI), which were significantly ($P<0.05$) associated with immune and metabolism pathways. The most optimized RNA-Seq pipeline allowed for more accurate identification of SNP, associated positional candidate genes, and associated metabolic pathways in muscle and liver tissues, providing insight on the genetic architecture of feed efficiency in beef cattle.

Introduction

High-throughput RNA-Sequencing (RNA-Seq) technology is a widely approach used to detect and quantify expressed transcripts, novel transcript discovery and analyze differential gene expression and alternative splicing in a biological sample (Mortazavi et al., 2008; Pightling et al., 2015; Wickramasinghe et al., 2014). In addition to these applications, RNA-Seq can detect functional genetic variants such as single nucleotide polymorphisms (SNPs), which are restricted to the expressed portion of the genome and represent a large amount of genetic variation in the genome (Cánovas et al., 2010; Chepelev et al., 2009). SNPs based genetic markers are useful due to their high abundance in the cattle genome and genetic stability (Nielsen, 2000; Heaton et al., 2001).

RNA-Seq experiments in livestock studies have identified significant SNPs in candidate genes associated with metabolic pathways that may play a role in the regulation of production traits (Cánovas et al., 2010;

Cánovas et al., 2012; Cánovas et al., 2013; Suárez-Vega et al., 2017; Cardoso et al., 2017; Dias et al., 2017). This has resulted in an improved understanding of the genetic architecture and a reduction in genome complexity of important production traits such as feed efficiency, health, fertility, and meat quality traits in beef cattle (Cánovas et al., 2010; Cánovas et al., 2013; Cánovas et al., 2014a; Fonseca et al., 2018; Weber et al., 2016). More specifically, the study of genetic variants that may serve as markers to select for feed efficiency or residual feed intake (RFI) may help lead to the genetic improvement of feed efficiency and result in economic and environmental benefits for beef production, as feed costs represent approximately 70% of livestock production expenses (FAO, 2006).

Although SNP identification for genetic markers has served as a powerful tool in genomics, the ability to better understand the relationship between genotype and phenotype relies on the accuracy of analysis to detect genomic variation. Studies have previously compared software for SNP identification and genotype and SNP calling methods using RNA-Seq data (Piskol et al., 2013; Quinn et al., 2013; Nielsen et al., 2011). With the availability of RNA-Seq data from multiple samples, animal groups, and tissues, the merging approaches to determine SNP calling workflows optimized for power and accuracy of SNP detection can be performed. However, the evaluation of RNA-Seq pipelines to identify variants across different phenotypic or genotypic groups that include samples from multiple tissues has not been evaluated and strategies for the use and merging of RNA-Seq data from multiple samples and tissues remain limited. Optimized RNA-Seq analysis approaches can then be applied for SNP discovery to detect SNPs that may serve as functional genetic markers and be used in selection strategies to improve economically relevant traits in livestock.

The aim of this study was to compare three RNA-Seq sample merging pipelines for SNP identification to determine the most optimized and accurate pipeline based on study experimental design and data analysis objectives. The approach considered as the most optimized and accurate approach for SNP detection using RNA-Seq data was then used to identify functional SNPs associated with feed efficiency in Nellore beef cattle to improve the understanding of the biology and metabolic pathways underlying genetic markers that may influence the function and regulation of feed efficiency in beef cattle. The specific objectives of this study were to 1) compare three RNA-Seq pipelines using samples from two divergent groups for feed efficiency (i.e., low- and high-RFI) and two tissues (i.e., liver and muscle) including multi-sample calling from: i) non-merged samples; ii) merged samples by group for low-RFI and merged samples for high-RFI for each tissue, iii) merged samples by group and tissue for low- and high-RFI for both tissues, 2) define the pipeline with maximized accuracy and power for SNP detection and apply it to identify unique SNPs fixed within high or low feed efficient Nellore beef steers, and determine associated genes and metabolic pathways.

Materials And Methods

RNA-Seq dataset

The datasets used in this study consist of RNA-Seq data obtained from liver and muscle tissue samples from Nellore beef steers deposited in the NCBI - Gene Expression Omnibus (GEO) public repository with PRJEB7696 and PRJEB15314 GEO accession numbers for both liver and muscle, respectively. Detailed animal management and sampling information was previously described by Tizioto et al., (2015) and by Tizioto et al., (2016). Briefly, 20 Nellore feedlot beef steers (n = 10 low-RFI and n = 10 high-RFI) at 21 months of age were used, which were selected from 585 steers that were calculated for feed efficiency through Best Linear Unbiased Prediction (BLUP) estimates for RFI (Tizioto et al., 2015; Tizioto et al., 2016). Tissue samples were collected from the longissimus thoracic muscle (Tizioto et al., 2016) and liver (Tizioto et al., 2015). RNA was extracted using Trizol method (Invitrogen®), and mRNA sample preparation was performed using TruSeq RNA Sample Preparation Kit ® (illumina, San Diego, CA). Cluster generation and sequencing was performed on the Illumina HiSeq 2000® which generated paired end-reads of 2 × 100 bp. The previous studies performed identification and annotation of differentially expressed genes across feed efficiency groups (Tizioto et al., 2015; Tizioto et al., 2016).

Comparison of animals from each study revealed four high-RFI individuals in the muscle group which did not match the high-RFI individuals in the liver group. This may be due to RFI re-ranking of animals, as liver tissue was sampled from only 83 of the 585 beef steers. Therefore, the four non-matching animals from the high-RFI muscle group (ERS1342436, ERS1342439, ERS1342440, ERS1342443) were removed from the study, and four corresponding individuals (ERS579404, ERS579406, ERS579407, ERS579411) from the high-RFI liver group were also removed. To maintain the same sample size across RFI groups, four animals were randomly selected to be removed from the low-RFI group for muscle (ERS1342447, ERS1342448, ERS1342450, ERS1342453) and liver (ERS579395, ERS579396, ERS579398, ERS579301). Therefore, this study used RNA-Seq data from 12 Nellore beef steers divergent for feed efficiency (n = 6 low-RFI and n = 6 high-RFI) from both muscle and liver tissue (Fig. 1).

RNA-Seq analysis workflow

Identification of SNPs was performed using the workflow shown in Fig. 2. Fastq files were downloaded using the SRA toolkit command fastq-dump with the option –split-files to download the data from each sample into two files, one for each of the paired ends. Quality of sequence reads were verified using FastQC (version 0.11.8; Andrews, 2014) to identify sequencing read artifacts including sites with low quality Phred scores, duplicated reads, uncalled bases (N sequences), and potential contamination (Cardoso et al., 2018; Cánovas et al., 2014b). Next, reads were trimmed to remove Illumina adapters and low quality bases at the start and end of reads (sites removed if Phred score < 30) using Trimmomatic (Version 0.38; Bolger et al., 2014). Additionally, reads with an average quality score below 20 within a sliding window of 5 nucleotides, and with length less than 75 bp were removed. Quality of sequence reads were re-evaluated post-trimming using FastQC (Andrews, 2014). The resulting trimmed reads from each sample were individually aligned to the bovine reference genome (Bos Taurus Assembly UMD3.1. release 94), using STAR (version 2.7.0.; Dobin et al., 2013) with the following filtering options: –outFilterMismatchNmax 999, allowing a maximum of 999 mismatches per pair, –outFilterMismatchNoverReadLmax 0.04, allowing alignment to be output only if its ratio of mismatches

to read length is less than 0.04, and –outFilterMultimapNmax 1, which allowed a max of 1 alignment per read (if exceeded, the read is considered unmapped). ReadGroups (RG) were then added to each sample and PCR duplicates were marked and removed using PICARD tools (Version 2.18.25.; http://broadinstitute.github.io/picard/). The RGs allowed for differentiation of samples by assigning the origin of the read (low- or high-RFI group) and assignment of SNPs to a specific genotype.

RNA-Seq read merging approaches

A diagram outlining the strategy used for merging BAM files is shown in Fig. 3. Three different RNA-Seq read merging approaches were compared including: i) non-merged samples, ii) merged samples by group for low-RFI and for high-RFI for each tissue, and iii) merged samples by group and tissue for low- and high-RFI for both tissues (liver and muscle). Each approach required the merging of aligned reads prior to variant calling (Fig. 2). BAM files were merged using SAMtools (Version 1.4; Li et al., 2011), which merged multiple BAM files and produced a single output file per strategy, leading to a total of six new BAM files containing merged aligned reads from: low-RFI muscle samples, low-RFI liver samples, high-RFI muscle samples, high-RFI liver samples, low-RFI muscle and liver samples, and high-RFI muscle and liver samples (Fig. 3).

Variant calling and filtering

Variant calling was performed for each read merging method to identify SNPs using the ‘mpileup’ and ‘call’ commands from BCFtools (Version 1.9-77-gd0cf724+; Danecek et al., 2018). This involved conversion of BAM files into genomic positions and variant calling, producing a BCF (Binary Variant Call Format) file containing variant information including genomic position, alternative allele detected, quality of SNP call, and other information (Danacek et al., 2011). The multi-sample calling (or joint calling) method, previously shown to enhance the sensitivity of SNP detection and accuracy of genotype calling over calling each sample independently in datasets with low sequencing depth, was employed for all read merging approaches.

The comparisons of approaches performed with VCF files of varying genotypes (GT) is summarized in Table 2 and Fig. 3. For approach i), BAM files (aligned reads) from the different animals and tissues were not merged and files for each tissue type were called for variants individually using multi-sample calling, resulting in one multi-sample VCF file for each tissue type, each containing 12 genotypes (1 GT for each animal). For approach ii), BAM files of animals in the same RFI group were merged for each tissue and files for each tissue type were called for variants using multi-sample calling, resulting in one VCF file for each tissue, each containing 2 genotypes (1 GT per RFI group). For approach iii), BAM files of animals in the same RFI group were merged for both tissues prior to multi-sample calling, resulting one VCF file that contained 2 genotypes (1 GT per RFI group). Variant filtering was performed using VCFtools to remove variants with a minimum read depth below 10 and a minimum of 2 supporting reads for the alternative allele as well as to filter SNPs within 3 bp surrounding a gap as described by Cánovas et al. (2010). BCFtools filter was used to remove variants with quality values below 30 (based on Phred scaled scores

for the assertion made in the alternative allele), filter SNPs within 5 bp of an INDEL, and filter any alternative allele with a lower frequency of 20% in the population.

Table 2
Summary of merging strategy for each approach and description of VCF file comparisons.

Approach	i) Non-merged samples	ii) Merged samples for low-RFI and merged samples for high-RFI for each tissue	iii) Merged samples for low- and high-RFI for both tissues
Samples per RFI group	6	6	12
RFI group	Non-merged	Merged	Merged
Tissues	Non-merged	Non-merged	Merged
Number of merged GT per VCF file for comparisons	To compare approach i) and ii)		
	12 GT	2 GT	
	To compare approach i) and iii)		
	24 GT		2 GT
	To compare approach ii) and iii)		
		4 GT	2 GT
Variant calling strategy	Multi-sample calling from multiple samples in one command line	Multi-sample calling from merged samples per RFI group for each tissue	Multi-sample calling from merged samples per RFI group for both tissues

RFI = residual feed intake; n: sample size; VCF = Variant Calling Format; GT = Genotype

Whole Transcriptome Coverage Analysis

The Whole Transcriptome Coverage Analysis tool of CLC Genomics Workbench 12.0.2 (https://www.qiagenbioinformatics.com/) was used to determine the total number of unfiltered reads across the whole transcriptome or each individual sample in Approach i) and for merged sample groups in Approach ii) and iii). This tool is used to identify regions in read mappings with unexpectedly low or high coverage, however, the summary of results of this analysis presents the 'Total Mapped Reads' across the whole transcriptome for each approach, counting both reads of a paired end sequence, while ignoring 'non-specific matches', or 'broken pairs' (CLC Genomics Workbench 12.0.2, https://www.qiagenbioinformatics.com/).

Identification and gene annotation of unique SNPs fixed within low- or high-RFI groups for each approach

Unique SNPs fixed within low- or high-RFI groups were identified using the VCF files containing filtered SNPs as previously described for each approach. These SNPs were filtered for moderate or high functional impact using Variant Effect Predictor (VeP) (McLaren et al., 2010). Variants with functional consequence within High, Moderate, or Modifier categories were selected for further analysis. In addition, in order to identify positional candidate genes, the genes localized with these SNPs were also determined using custom R scripts (Version 1.1.463.; R Core Team, 2013) and the VennDiagram R package (Cánovas et al., 2014).

Approach iii) was then used to identify unique SNPs fixed within low- or high-RFI groups using the VCF files containing filtered SNPs as previously described (Fig. 2). Therefore, 2 VCF were compared (1 low-RFI and 1 high-RFI VCF) with 1 GT each (representative of multiple samples). SnpSift (Version 4.0; Cingolani et al., 2012) filtering command was used to remove all variants present in one RFI group but missing in the other RFI group. The VCF file was then split using VCFtools vcf-subset to create one VCF file with only low-RFI variants and one VCF file with only high-RFI variants, which were then compared using the BCFtools isec command to determine the intersection of the files and create three files: SNPs exclusive to low-RFI VCF file, SNPs exclusive to high-RFI VCF file, and SNPs shared between both low- and high-RFI groups VCF file. Positional candidate genes in which these variants were localized were selected to determine associated metabolic pathways. Functional annotation was performed using ToppGene (Chen et al., 2009) to determine metabolic pathways significantly associated with the gene lists.

Results And Discussion

In this study, we compared three multi-sample calling approaches and their SNP calling results and applied the most optimized approach to perform a more accurate SNP discovery for genetic markers associated with feed efficiency in beef cattle. The number of total reads, total uniquely mapped reads, and percentage of uniquely mapped reads is given in Table 1. Overall, the number of uniquely mapped reads identified in muscle tissue (205,269,868) were observed to be greater than that detected in liver tissue (87,466,593) (Table 1). This may have resulted in a lower number of total SNPs detected in liver compared to muscle in both the non-merged and merged samples approaches (Table 3).

Table 1

Liver and muscle sample information including feed efficiency group, read mapping information.

Sample Accession Number by study	Feed efficiency group	Total Reads	# uniquely mapped reads	% uniquely mapped reads
Tissue (study accession #)				
Liver (PRJEB7696)				
ERS579394	high	7,338,446	6,975,563	95.06
ERS579397	high	7,079,855	6,714,099	94.83
ERS579399	high	8,325,791	7,903,077	94.92
ERS579400	high	7,716,404	7,305,248	94.67
ERS579402	high	7,354,529	6,978,182	94.88
ERS579403	high	7,114,744	6,742,752	94.77
ERS579405	low	6,363,266	6,042,120	94.95
ERS579408	low	6,596,129	6,272,955	95.10
ERS579409	low	7,923,682	7,497,828	94.63
ERS579410	low	7,132,356	6,642,434	93.13
ERS579412	low	6,913,752	6,537,053	94.55
ERS579413	low	7,607,639	7,209,026	94.76
Muscle (PRJEB15314)				
ERS1342445	high	17,227,810	16,261,921	94.39
ERS1342446	high	20,976,425	19,931,762	95.02
ERS1342449	high	20,926,000	17,475,676	83.51
ERS1342451	high	14,807,990	14,064,136	94.98
ERS1342452	high	15,773,643	10,430,495	66.13
ERS1342454	high	13,852,120	13,174,573	95.11
ERS1342435	low	18,443,978	17,618,969	95.53
ERS1342437	low	15,475,954	14,733,896	95.21
ERS1342438	low	17,516,264	16,728,409	95.50
ERS1342441	low	17,307,502	16,462,170	95.12

Sample Accession Number by study	Feed efficiency group	Total Reads	# uniquely mapped reads	% uniquely mapped reads
ERS1342442	low	14,900,733	14,179,347	95.16
ERS1342444	low	18,061,449	17,254,366	95.53

Table 3

Summary of total SNPs detected using bcftools for each approach scenario used for comparisons.

Approach	Tissue	n	Total SNPs before filtering	Total SNPs after filtering	Percentage of SNPs passing all filters (%)
i) a)	Liver	6	626,460	258,120	41.20
i) b)	Muscle	6	940,143	396,705	42.20
i) c)	Liver + Muscle	6	1,205,664	511,092	42.39
ii) a)	Liver	6	521,588	197,309	37.82
ii) b)	Muscle	6	770,685	296,169	38.43
ii) c)	Liver + Muscle	12	1,005,696	388,322	38.61
iii)	Liver + Muscle	12	1,048,370	416,216	39.70

i) Non-merged samples; ii) Merged samples for low-RFI and merged samples for high-RFI for each tissue; iii) Merged samples for low- and high-RFI for both tissues.

Table 3 displays all merging and non-merging approaches and includes total SNP identified before and after applying quality filters. On average, the percentage of SNPs that passed all quality filters for all approaches was 40.05%. The frequency of overlapping SNPs between liver (Table 3a)) and muscle (Table 3b)) tissues was relatively high, confirming that variants were expressed in these tissues and supporting our rationale that merging data by tissues could lead to increased accuracy of SNP detection by increasing read depth and reducing possible false positives. To compare the first two merging approaches (non-merging and merging by RFI group) to the third approach (merging by RFI group and tissues), the VCF files of the liver and muscles tissues for the first two approaches (i and ii) were merged using VCFtools. A higher number of total SNPs were identified in the non-merging method compared to all other merging approaches. This may be due to an increase in detection of rare variants (i.e., variants detected in a small subset of the animals).

Comparison of RNA-Seq merging approaches for more accurate SNP detection

To compare the overlap of the SNPs detected by the various approaches, we determined the total number and percentage of SNPs identified as shared or unique across the approaches being compared (Table 4). Results revealed that when comparing the approaches of non-merging (Approach i)) and merging samples by RFI group (Approach ii)) for liver, the majority of SNPs were shared (76.20%) between both approaches. A considerable number of SNPs detected by the non-merging approach (23.66%) were unique to this approach and were not detected by the merging samples by tissue approach, while very few SNPs (0.13%) were found to be uniquely detected by the merging samples by tissue approach. Similar results were detected when comparing these two approaches (i and ii) for the muscle tissue. Comparing the non-merging approach to that of merging tissues (i and iii) yielded similar results, with the majority of SNPs found shared (72.84%) or unique to the non-merging approach (22.42%). In addition, 4.74% of SNPs were detected uniquely by the merging tissues approach. Comparing the merging samples by tissue approach to that of merging tissues (ii and iii), a greater overlap between SNPs was detected (86.70%), with 3.41% of SNPs found unique to merging samples by tissue and 9.89% of SNPs found unique to merging tissues.

Table 4

Results of approach comparisons showing total SNP identified unique within approach and shared between both approaches.

	Approach comparisons			
	Approach i) v.s. Approach ii)			
	Liver Non-merged v.s. Liver Merged by RFI group			
	Individual	Shared	Merged samples	Total
Total number of SNPs	61,158	196,962	347	258,467
Percentage of SNPs	23.66	76.20	0.13	
	Approach i) v.s. Approach ii)			
	Muscle Non-merged v.s. Muscle Merged by RFI group			
	Individual	Shared	Merged samples	Total
Total number of SNPs	101,047	295,658	511	397,216
Percentage of SNPs	25.44	74.43	0.13	
	Approach i) v.s. Approach iii)			
	Liver and Muscle Non-merged v.s. Liver and Muscle Merged by RFI group and Tissues			
	Individual	Shared	Merged tissues	Total
Total number of SNPs	120,301	390,791	25,425	536,517
Percentage of SNPs	22.42	72.84	4.74	
	Approach ii) v.s. Approach iii)			
	Liver and Muscle Merged by RFI group v.s. Liver and Muscle Merged by RFI group and Tissues			
	Merged samples	Shared	Merged tissues	Total
Total number of SNPs	14,699	373,623	42,593	430,915

i) Non-merged samples; ii) Merged samples for low-RFI and merged samples for high-RFI for each tissue; iii) Merged samples for low- and high-RFI for both tissues.

	Approach comparisons		
Percentage of SNPs	3.41	86.70	9.88

i) Non-merged samples; ii) Merged samples for low-RFI and merged samples for high-RFI for each tissue; iii) Merged samples for low- and high-RFI for both tissues.

The SNPs that are uniquely detected by the non-merging method could represent SNPs that are present in a small subset of animals and hence not be representative of a specific RFI group. For SNPs with a low non-reference allele frequency, merging reads from multiple samples could lead to dilution of reads supporting the variant and consequent call as homozygous for reference (VanRaden et al., 2019). Alternatively, the Phred quality score of a SNP may be inflated when detected in a large number of samples and consequently, some of the SNPs uniquely detected by the non-merging method could have been removed by the quality filters in the merging methods suggesting possible false positives (Liao et al., 2017). On the other hand, the detection of SNPs that are unique to the merging methods suggests that merging samples and tissues improves SNP detection and Phred quality scores due to increased read depth and reducing potential false positives.

Comparison of RNA-Seq merging approaches based on read depth, annotated SNPs, and annotated genes fixed within RFI groups

The analysis resulted in the total number of mapped reads on the reference for each individual sample in Approach i) (non-merged samples), and for merged map reads of samples in Approach ii) (merged samples by RFI group) and Approach iii) (merged samples by RFI and tissue group) (Table 5). To identify the most optimized approach with highest read depth, the total reads mapped across the whole transcriptome for each approach were compared (Table 5). On average, the total reads for Approach i) individual liver samples and individual muscle samples were 18,362,086.3 and 35,645,898.7, respectively. For Approach ii), the average total number of reads for each merged group of samples was 162,030,705, and for Approach iii) was 324,061,410. Approach iii) had the highest read depth and coverage across the whole transcriptome, suggesting that this approach may have higher read depth to filter out false positives and more accurately detect SNPs and INDELs.

Table 5

Total reads across genome for each approach from CLC Whole Genome Coverage Analysis.

Sample Accession number or samples merged description	Feed efficiency group	Total Reads
Approach i): Non-merged		
Individual liver samples		
ERS579394	high	18,505,235
ERS579397	high	17,915,515
ERS579399	high	21,085,175
ERS579400	high	19,602,337
ERS579402	high	18,583,880
ERS579403	high	18,078,573
ERS579412	low	17,489,472
ERS579413	low	19,044,144
ERS579408	low	16,647,703
ERS579409	low	19,843,569
ERS579410	low	17,775,528
ERS579405	low	15,773,905
Individual muscle samples		
ERS1342445	high	35,720,847
ERS1342446	high	43,782,111
ERS1342449	high	43,403,981
ERS1342451	high	30,832,942
ERS1342452	high	33,478,392
ERS1342454	high	28,850,693
ERS1342435	low	38,307,806
ERS1342437	low	32,252,078
ERS1342438	low	36,486,302
ERS1342441	low	36,091,768

*Total reads - refers to total reads based on coverage across genome

Sample Accession number or samples merged description	Feed efficiency group	Total Reads
ERS1342442	low	31,067,715
ERS1342444	low	37,476,149
Approach ii): Merged by RFI group		
Liver samples for high feed efficiency merged	high	113,770,715
Liver samples for low feed efficiency merged	low	106,601,321
Muscle samples for high feed efficiency merged	high	216,068,966
Muscle samples for low feed efficiency merged	low	211,681,818
Approach iii): Merged by RFI group and tissue group		
Liver and Muscle samples for high feed efficiency merged	high	329,839,681
Liver and Muscle samples for low feed efficiency merged	low	318,283,139
*Total reads - refers to total reads based on coverage across genome		

The list of SNPs fixed within low- or high-RFI groups and filtered for moderate or high functional impact were compared across each approach (Fig. 4). The most optimized approach for detecting SNPs that best represented the feed efficiency trait was determined by comparing the SNP detection results from each approach. From this, SNPs fixed within low- or high-RFI groups, filtered for moderate, modifier, or high functional impact, were compared across each approach (Fig. 4), and the associated genes harboring these SNPs were also compared across approaches (Fig. 5). The total amount of SNPs after filtering, detected by each approach were 23,228, 22,957, and 27,429 for approach i), ii), and iii), respectively (Fig. 4). A large overlap of SNPs (14,807) between approach ii) and iii) was observed which is likely due to the merging of samples in these approaches. The total amount of genes unique to low- and high-RFI groups identified for each approach were 4,568, 4,804, and 3,938 for approach i), ii), and iii), respectively. We observed a higher average of SNPs detected per gene for approach iii) (1.832 SNPs/gene; SD = 1.564), compared to approach i) (1.775 SNPs/gene; SD = 1.499) and ii) (1.773 SNPs/gene; SD = 1.537), as approach iii) had more SNPs and less annotated genes ratio (Figs. 4 and 5). This may suggest that approach iii) reveals that more SNPs are influencing fewer genes.

As shown in the Venn Diagram in Fig. 5, the majority of genes (1,975) from each approach were identified as shared across all three approaches. Additionally, 1,296, 715, and 486 genes were identified as unique to approach i), ii), and iii), respectively. The number of genes shared between approach i) and ii) was 967, and between ii) and iii) was 1,147, and between i) and iii) was 330. When considering the total genes identified in each approach, approach iii) shares the most genes among both approach i) and ii) (50%), and 8.37% with approach i) and 29.13% with approach ii). The genes shared among all three approaches were of most importance and interest, as they were the genes most representative of the trait, and approach iii) had the most genes located in this mutual group. Additionally, approach iii) had the lowest

percent of genes unique to its own group (12.3%). This may suggest that approach iii) is most representative of the unique genes associated with feed efficiency, and was able to identify more positional candidate genes (486) that are likely to play a role in regulating feed efficiency, while potentially excluding false positive SNPs that may have been annotated for untrue genes identified by the other approaches. Notably, approach ii) and iii) had more shared variants (86.70%) compared to the other approach comparisons in Table 4, and approach ii) and iii) also have the most shared annotated genes (Fig. 5). Similarly, approach i) and iii) had the least shared variants (Table 4) and also have the least shared annotated genes (Fig. 5). This suggests that the comparison of unique and shared SNPs across approaches in Table 4 is representative or may be correlated with the annotated genes (for low- and high-RFI) compared between approaches in Fig. 5.

Table 6
Metabolic pathways significantly associated with genes in which SNPs from Low- and High-RFI were localized.

Pathway Name	p-value	FDR[i]	Genes associated with SNPs detected	Function
Low-RFI[ii]				
Members of the BCR[a] signaling	1.68×10^{-6}	5.12×10^{-3}	20	Immune
Oxytocin signaling	1.70×10^{-5}	1.87×10^{-2}	42	Reproduction and Metabolism
EPHA2[b] forward signaling	2.72×10^{-5}	1.87×10^{-2}	10	Growth and Metabolism
High-RFI				
Members of the BCR signaling	1.21×10^{-5}	1.52×10^{-2}	19	Immune
B cell activation	1.313×10^{-5}	1.52×10^{-2}	22	Immune
Regulation of RAC1[c] activity	1.48×10^{-5}	1.52×10^{-2}	17	Metabolism
[i] FDR = False Discovery Rate				
[ii] Top 3 pathways out of 111 significantly associated pathways				
[a] BCR = B Cell Receptor				
[b] EPHA2 = Ephrin type-A receptor 2				
[c] RAC1 = Ras-related C3 botulinum toxin substrate 1				

Unique SNPs fixed within low- or high-RFI groups and their associated candidate genes and metabolic pathways

Unique or shared SNPs identified across low- and high-RFI groups using approach iii) are shown in Fig. 6. In total, 415,624 SNPs were detected, with 13,145 SNPs unique in low-RFI, and 14,663 SNPs unique in high-RFI. In addition, 387,816 SNPs were identified as shared across both low- and high-RFI groups. The unique SNPs were filtered for variant impact information including only High, Moderate, or Modifier impact. With variant impact ‘High’, meaning the variant may have a high or disruptive impact in the

protein that would lead to protein truncation, loss of function, or tissue nonsense, mediated delay; 'Moderate', meaning a non-disruptive variant that may change protein effectiveness; or 'Modifier' non-coding variant, or variant affecting non-coding genes, where predictions are difficult or the impact is unknown (McLaren et al., 2010). Further investigation of the genes localized with these SNPs with more potential of impacting protein function were selected as they may be more likely to have a functional impact in metabolic pathways that may play a role in regulating feed efficiency. In total, 2,886 and 3,075 genes co-localized with SNPs were identified for the low- and high-RFI groups, respectively. The gene lists were significantly ($P < 0.01$) enriched for 111 and 3 functional pathways for low- and high-RFI groups, respectively. The 3 most significant pathways associated with each RFI group are displayed in Table 6. Low-RFI animals display a genetic architecture with fewer genes affecting more biological pathways, compared to high-RFI animals, explaining the large number of significantly enriched genes (111) for low-RFI group, compared to the high-RFI group (3).

When using RNA-Seq technology to detect structural genetic variants (Cánovas et al., 2010), the structural variants are expected to be detected from mRNA reads, which are not necessarily differentially expressed or highly expressed in a specific tissue. The following structural variants and co-localized genes are discussed in a means that the structural variants are unique or fixed within RFI groups and are located within or near the discussed genes that could be of interest to better understand feed efficiency, but not necessary in terms of the expression of the genes.

The three most significant pathways associated with low-RFI (more feed efficient) cattle, included immune response, fertility, and metabolism pathways, including the BCR signaling pathway, oxytocin signaling pathway, and EPHA2 forward pathway (Table 6). The BCR signaling pathway is associated with immune response and fertility; Olivieri et al. (2016) previously identified candidate genes related to feed efficient Nellore beef cattle which were associated with immune system function, including NLRP14 gene which was present in the low feed efficient group in this study, and CACNG7 which was present in high feed efficient group in this study. The regulation of NLRP14 has also been suggested to be associated with excessive accumulation of undifferentiated spermatogonia germ cells in cattle, implying its role in reproductive function (Cai et al., 2017). It may be possible that these genes which are co-localized with variants unique to low- and high-RFI steers are being expressed. The CACNG7 gene is associated with oxytocin signaling; this may explain the oxytocin enriched pathway in the current study which was associated with more feed efficient cattle, supporting a link between metabolic processes with feed efficiency. Oxytocin is a neuropeptide which has also been found to regulate brown adipocyte production, which functions for metabolic maintenance of temperature regulation (thermogenesis) and gluconeogenesis (Ding et al., 2018).

The EPHA2 forward signaling was also associated with more feed efficient cattle in this study. The EPHA2 is a gene known to be targeted by miR-26b, a microRNA (miRNA) in pituitary tissues of Zanbian cattle, and this miRNA is known to regulate cell proliferation, differentiation, apoptosis, and development (Yuan et al., 2015). In addition, EPHA2 receptor is a key modulator for a wide variety of cellular functions, such as embryonic development, tissue boundary formation, central nervous system function, bone

remodeling, vascular organization (Park et al., 2013), suggesting its role in metabolic maintenance and regulation. This may be relevant in the regulation of feed efficiency due to the strong correlation between feed efficiency and efficiency of metabolic energy use to meet physiological demands.

The three most significant pathways associated with high-RFI (less feed efficient) cattle, also included BCR signaling pathway (Table 6), which may suggest similar variants are present in both low- and high-RFI groups, but are acting in different ways to regulate feed efficiency in cattle. This is supported by the discovery in the study where many SNPs were detected in both low- and high-RFI groups with the same position but with different alleles. Another significantly associated pathway to less feed efficient cattle was B cell activation which is additionally related to immune response. It has been found that B cell activation pathway was enriched for differentially expressed genes in pigs selected for RFI (Schroyen et al., 2016). This may suggest the influence of feed efficiency on immune pathways and function in livestock. Regulation of RAC1 was another significantly associated pathways with less feed efficient cattle, which is known to have a role in glucose transport and skeletal muscle (Sylow et al., 2013), and could be associated with the regulation of gluconeogenesis, along with oxytocin as discussed previously.

Overall, the most significant pathways associated with the extreme feed efficiency groups were both related to metabolic, reproductive, and immune functions (Table 6). Notably, both feed efficiency groups found similar significant pathways. From the analysis, similar SNP were found with different alleles in both low- and high-RFI, suggesting that the same SNPs with different alleles are present in animals in both high and low feed efficiency groups, influencing genes differently, to make an animal more or less feed efficient.

Conclusion

In conclusion, this study demonstrates the different results obtained in SNP detection from using different sample merging pipelines for RNA-Seq analysis. We suggest that the most optimized RNA-Seq pipeline for SNP detention when analyzing multiple samples and tissues in divergent traits or phenotypic groups is to merge multiple samples and tissue data by group to increase aligned reads at each variant site which increases read depth and coverage, leading to improved accuracy and power of SNP and genotype calling and reduction of false positive detection. This pipeline can be used to identify SNPs associated with extreme phenotypes of interest, which can be annotated to identify genes that may regulate that trait. Overall, this work presents an optimized RNA-Seq pipeline to discover SNPs in coding regions to improve the detection of potential trait-associated variations using RNA-Seq data. Additionally, using the proposed optimized RNA-Seq pipeline, this study successfully identified functional candidate SNPs within genes involved in major metabolic pathways associated with high and low feed efficient Nellore beef cattle, suggesting the relationship between immune, reproductive, and metabolic function with feed efficiency in beef cattle, and confirming the successful use of the most optimized RNA-Seq pipeline described in this study.

Declarations

Ethics approval and consent to participate

This study used publicly available data and therefore did not involve the use of animals or individual's information.

Availability of data and materials

The datasets supporting the results of this article include RNA-Seq datasets which are available from the NCBI - Gene Expression Omnibus (GEO) public repository with PRJEB7696 [https://www.ncbi.nlm.nih.gov/bioproject/PRJEB7696/] and PRJEB15314 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJEB15314] GEO accession numbers for liver and muscle tissue, respectively.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SL analysed and interpreted the data regarding the comparison of the RNA-Seq pipelines. SL was a major contributor in writing the manuscript. JZ, ASV, IGR, PASF assisted with data analysis. All authors read and approved the final manuscript.

Funding

The authors acknowledge the funding for this study, which was supported by the Ontario Ministry of Agriculture, Food, and Rural Affairs (OMAFRA), Ontario Ministry of Research and Innovation, Agriculture and Agri-Food Canada (AAFC) and Genome Canada. This study (FDE.13.17) was also supported by the Sustainable Beef and Forage Science Cluster funded by the Canadian Beef Cattle Check-Off, Beef Cattle Research Council (BCRC), Alberta Beef Producers, Alberta Cattle Feeders' Association, Beef Farmers of Ontario, La Fédération des Productuers de bovins du Québec, and Agriculture and Agri-Food Canada's Canadian Agricultural Partnership.

Acknowledgements

The authors acknowledge financial support from the Ontario Ministry of Agriculture, Food, and Rural Affairs (OMAFRA), Ontario Ministry of Research and Innovation, Agriculture and Agri-Food Canada (AAFC) and Genome Canada. This study (FDE.13.17) was also supported by the Sustainable Beef and Forage Science Cluster funded by the Canadian Beef Cattle Check-Off, Beef Cattle Research Council (BCRC), Alberta Beef Producers, Alberta Cattle Feeders' Association, Beef Farmers of Ontario, La Fédération des Productuers de bovins du Québec, and Agriculture and Agri-Food Canada's Canadian Agricultural Partnership.

References

1. Andrews S. 2014. FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 2019 January 27.
2. Basarab JA, Beauchemin KA, Baron, VS, Ominski, KH, Guan LL, Miller SP, Crowley JJ. Reducing GHG emissions through genetic improvement for feed efficiency: Effects on economically important traits and enteric methane production. Animal. 2013;7:303–315.
3. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-2120. Broad Institute. Picard Tools. http://broadinstitute.github.io/picard/ (Retrieved October 15 2018).
4. Cai X, Yu S, Mipam TD, Yng F, Zhao W, Liu W, Cao SZ, Shen L, Zhao F, Sun L, Xu C, Wu S. Comparative analysis of testis transcriptomes associated with male infertility in cattle yak. Theriogenology. 2017;88: 28-42.
5. Cánovas A. Looking ahead: applying new genomic technologies to accelerate genetic improvement in beef cattle. Ceiba. 2016;54:41-49.
6. Cánovas A, Rincon G, Islas-Trejo A, Flores RJ, Laubscher A, Medrano JF. RNA sequencing to study gene expression and single nucleotide polymorphism variation associated with citrate content in cow milk. Journal of Dairy Science. 2013;96(4):2637-2648.
7. Cánovas A, Rincon G, Islas-Trejo A, Wickramasinghem S, Medrano JF. SNP discovery in the bovine milk transcriptome using RNA-Seq technology. Mammalian Genome. 2010;21(11-12):592-598.
8. Cánovas A, Reverter A, DeAtley KL, Ashley RL, Colgrave ML, Fortes MRS, Islas-Trejo A, Lehnert S, Porto-Neto L, Rincón G, Silver GA, Snelling WM, Medrano J.F, Thomas MG. Multi-Tissue Omics Analyses Reveal Molecular Regulatory Networks for Puberty in Composite Beef Cattle. PLoS One. 2014a.;9(7): e102551.
9. Cánovas A, Rincon G, Bevilacqua C, Islas-Trejo A, Brenaut P, Hovey RC, Boutinaud M, Morgenthaler C, WanFlompenberg MK, Martin P, Medrano JF. Comparison of five different RNA sources to examine the lactating bovine mammary gland transcriptome using RNA-Sequencing. Nature Scientific Reports. 2014b;4:5297.
10. Cánovas A, Pena RN, Gallardo D, Ramírez O, Amills M, Quintanilla R. Segregation of regulatory polymorphisms with effects on the gluteus medius transcriptome in a purebred pig population. PLoS One. 2012;7(4):e35583.
11. Cardoso, T. F., Cánovas, A., Canela-Xandri, O., González-Prendes, R., Amills, M., Quintanilla, R. 2017. RNA-seq based detection of differentially expressed genes in the skeletal muscle of Duroc pigs with distinct lipid profiles. Scientific Reports. 7:40005. doi: 10.1038/srep40005.
12. Cardoso TF, Quintanilla R, Castelló A, González-Prendes R, Amills M, Cánovas A. Differential expression of mRNA isoforms in the skeletal muscle of pigs with distinct growth and fatness profiles. BMC Genomics. 2018;19:145.

13. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Research. 2009;37: 305-311.
14. Chepelev I, Wei G, Tang Q, Zhao K. Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. Nucleic Acids Research. 2009;37:e106.
15. Cingolani P, Patel VM, Coon M, Nguyen T, Land SJ, Ruden DM, Lu X. Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. Frontiers in Genetics. 2012;3:35. doi: 10.3389/fgene.2012.00035.
16. Couldrey C, Keehan M, Johnson T, Tiplady K, Winkleman A, Littlejohn, MD, Scott A, Kemper KE, Hayes B, Davis SR, Spelman RJ. Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. Journal of Dairy Science. 2017;100(7):5472-5478.
17. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27:2156-2158.
18. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics 2017;33, 2037-2039.
19. Dias MM, Cánovas A, Mantilla-Rojas C, Riley DG, Luna-Nevarez P, Coleman SJ, Speidel SE, Enns RM, Islas-Trejo A, Medrano JF, Moore SS, Fortes MR, Nguyen LT, Venus B, Diaz IS, Souza FR, Fonseca LF, Baldi F, Albuquerque LG, Thomas MG, Oliveira HN. SNP detection using RNA-sequences of candidate genes associated with puberty in cattle. Genetics and Molecular Research. 2017;16(1).
20. Ding C, Leow MKS, Magkos F. Oxytocin in metabolic homeostasis: implications for obesity and diabetes management. Obesity Reviews. 2018;20: 22-40.
21. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15-21.
22. Fonseca PAS, Id-Lahoucine S, Reverter A, Medrano JF, Fortes MS, Casellas J, Miglior F, Brito, L, Carvalho MRS, Schenkel FS, Nguyen LT, Porto-Neto LR, Thomas MG, Cánovas A. Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle. PLoS One. 2018;13(10):e0205295.
23. Food and Agriculture Organization of the United Nations. 2006. Livestock's long shadow: environmental issues and options. Rome: FAO.
24. Gou B, Greenwood PL, Café LM, Zhou G, Zhang W, Dalrymple BP. Transcriptome analysis of cattle muscle identifies potential markers for skeletal muscle growth rate and major cell types. BMC Genomics. 2015;16:177.
25. Heaton MP, Grosse WM, Kappes SM, Keele JW, Chitko-McKown CG, Cundiff LV, Braun A, Little DP, Laegreid WW. Estimation of DNA sequence diversity in bovine cytokine genes. Mammalian Genome. 2001;12:32–37.
26. Kelly AK, Waters SM, McGee M, Fonseca RG, Carberry C, Kenny DA. mRNA expression of genes regulating oxidative phosphorylation in the muscle of beef cattle divergently ranked for residual feed

intake. Physiological Genomics. 2011;42: 12-13.

27. Li HA. Statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 2011;(21):2987-93.
28. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, Hess GT, Zappala Z, Strober BJ, Scott AJ, Li A, Ganna A, Bassik MC, Merker JD, GTEx Consortium, Hall IM, Battle A, Montgomery SB. The impact of rare variation on gene expression across tissues. Nature. 2017;550:239-243.
29. Liao P, Satten GA, Hu YA. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. Genetic Epidemiology. 2017;41:375-387.
30. McLaren W, Pritchard B, Rios D, Chen Y, Filicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP effect predictor, Bioinformatics. 2010;26: 2069-70.
31. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods. 2008;5:621–628.
32. Nielsen R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics. 2000;154:931–942.
33. Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from nest-generation sequencing data. Nature Reviews: Genetics. 2011;12:443-451.
34. Nookaew I, Papini, M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J. A comprehensive comparison of RNA-seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Research. 2009;40:10084–10097.
35. Olivieri BF, Mercadante ME, Cyrillo JN, Branco RH, Bonilha SF, de Albuquerque LG, Silva RM, Baldi F. Genomic Regions Associated with Feed Efficiency Indicator Traits in an Experimental Nellore Cattle Population. PLoS One. 2016;11(10):e0164390.
36. Park JE, Son AI, Zhou R. Roles of EphA2 in Development and Disease. Genes. 2013;4:334-357.
37. Pightling AW, Petronella N, Pagotto F. Choice of reference-guided sequence assembler and SNP caller for analysis of Listeria monocytogenes short-read sequence data greatly influences rates of error. BMC Research Notes. 2015;8:748.
38. Piskol R, Ramaswami G, Li JB. Reliable identification of genomic variants from RNA-seq data. American Journal of Human Genetics. 2013;93: 641-651.
39. Pootakham W, Shearman JR, Ruangareerate P, Sonthirod C, Sangsrakru D, Jomchai N, Yoocha T, Triwitayakorn K, Tragoonrung S, Tangphatsornyuang S. Large-Scale SNP Discovery through RNA Sequencing and SNP Genotyping by Targeted Enrichment Sequencing in Cassava (Manihot esculenta Crantz). PLoS One 2014;9(12):16028.
40. Quinn EM, Cormican P, Kenny EM, Hill M, Anney R, Gill M, Corvin AP, Morris DW. Development of strategies for SNP detection in RNA-seq data: application to lymphoblastoid cell lines and evaluation using 1000 Genomes data. PLoS One. 2013;8:e58815.

41. R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
42. Schroyen M, Feye KM, Nguyen YT, Rakhshandeh A, Gabler NK. Toward robust blood biomarkers for residual feed intake in pigs. Journal of Animal Science. 2016;94: 56-57.
43. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics. 2014;15:121–132.
44. Standish KA, Carland TM, Lockwood GK, Pfeiffer W, Tatineni M, Huang CC, Lamberth S, Cherkas Y, Brodmerkel C, Jaeger E, Smith L, Rajagopal G, Curran ME, Schork Group-based variant calling leveraging next-generation supercomputing for large-scale whole-genome sequencing studies. BMC Bioinformatics. 2015;16(1): 304.
45. Suárez-Vega A, Gutinrrez-Gil B, Klopp C, Tosser-Klopp G, Arranz JJ. Variant discovery in the sheep milk transcriptome using RNA sequencing. BMC Genomics. 2017;18(1):170.
46. Sylow L, Jensen TE, Kleinert M, Højlund K, Liens B, Wojtaszewski J, Prats C, Schjerling P, Richter EA. Rac1signaling is required for insulin-stimulated glucose uptake and is dysregulated in insulin-resistant murine and human skeletal muscle. Diabetes. 2013;62: 1865-75.
47. Tizioto PC, Coutinho LL, Decker JE, Schnabel RD, Rosa KO, Oliveira PSN, Souza MM, Mourão GB, Tullio RR, Chaves AS, Lanna DPD, Zerlotini-Neto A, Mudadu MA, Taylor JF, Regitano LCA. Global liver gene expression differences in Nelore steers with divergent residual feed intake phenotypes. BMC Genomics. 2015;16:242.
48. Tizioto PC, L Coutinho LL, Oliveira PSN, Cesar ASM, Diniz WJS, Lima AO, Rocha MI, Decker JE, Schnabel RD, Mourão GB, Tullio RR, Zerlotini A, Taylor JF, Regitano LCA. Gene expression differences in Longissimus muscle of Nelore steers genetically divergent for residual feed intake. Scientific Reports. 2019;6:39493.
49. VanRaden PM, Bickhart DM, O'Connell JR. Calling known variants and identifying new variants while rapidly aligning sequence data. Journal of Dairy Science. 2019;102:3216-3229.
50. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews. Genetics. 2019;10:57–63.
51. Wickramasinghe S, Cánovas A, Rincón G, Medrano JF. RNA-Sequencing: A tool to explore new frontiers in animal genetics. Livestock Science. 2014;166:206–216.
52. Weber KL, Welly BT, Van Eenennaam AL, Young AE, Porto-Neto LR, Reverter A, Rincon G. Identification of gene networks for residual feed intake in Angus cattle using genomic prediction and RNA-Seq analysis. PLoS One. 2016;11:e0152274.
53. Yuan B, Yu W, Dai, L, Gao Y, Ding Y, Yu X. Expression of microRNA-26b and identification of its target gene EphA2 in pituitary tissues in Yanbian cattle. Molecular Medicine Reports. 2015l;12: 5753-5761.

Figures

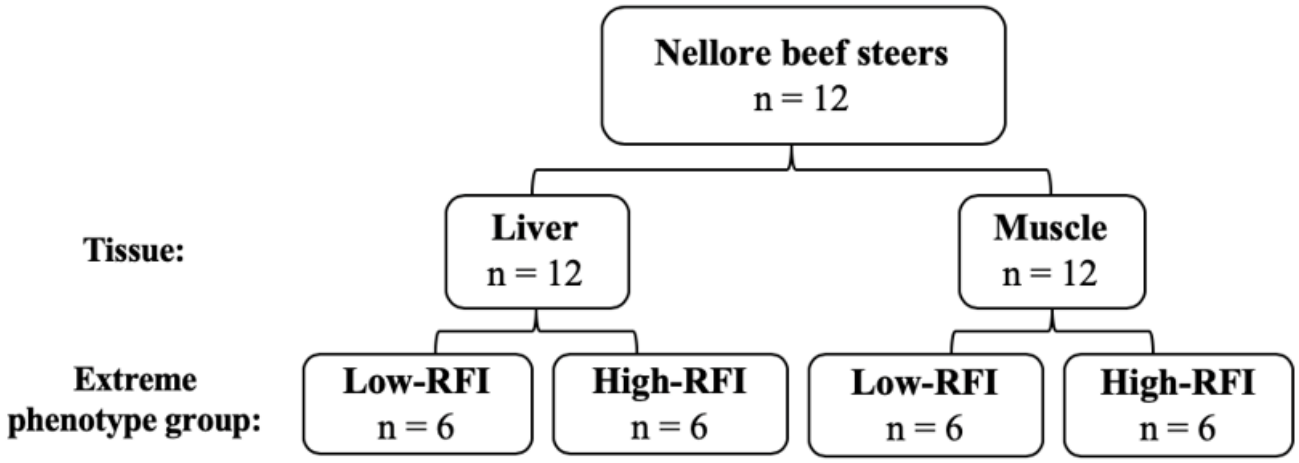

Figure 1

Population, tissue sample, and feed efficiency group structure using RNA-Seq data from two studies* RFI = residual feed intake; n = sample size. Initial population sizes were n = 20 (liver) and n = 20 (muscle); however, 4 individuals were removed from each high-RFI group due to non-matching animal IDs. Additionally, 4 individuals were randomly removed from each low-RFI group to maintain a consistent sample size of n = 6 for each RFI group. *GEO Accession ID: PRJEB7696 and PRJEB15314 for liver and muscle tissue data, respectively

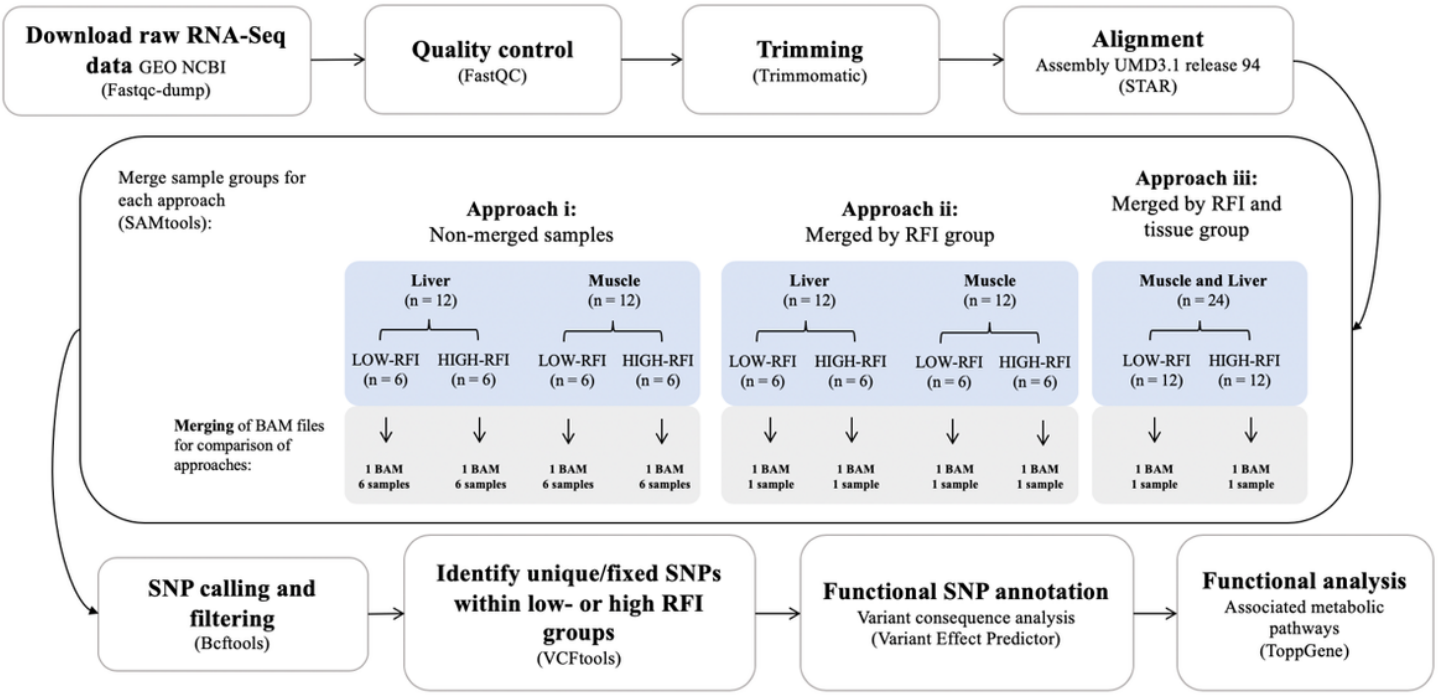

Figure 2

Workflow diagram to compare SNP calling approaches, identify functional SNP, and predict corresponding variant effects. i Samples were merged based on three different approaches: i) Non-

merged samples, meaning BAM files were not merged and each file was called individually using a multi-sample VCF file containing 6 GT (1 for each animal) for each group; ii) Merged samples for low-RFI and merged samples for high-RFI for each tissue, meaning BAM files of same group and tissue were merged. Using one VCF file for each RFI group per tissue, containing 1 GT; iii) Merged samples for low- and high-RFI for both tissues, meaning BAM files of same group were merged for tissues. Using one VCF file for each RFI group for both tissues, containing 1 GT. RFI = residual feed intake; GT = genotype; SNP = single nucleotide polymorphism; VCF = variant calling format.

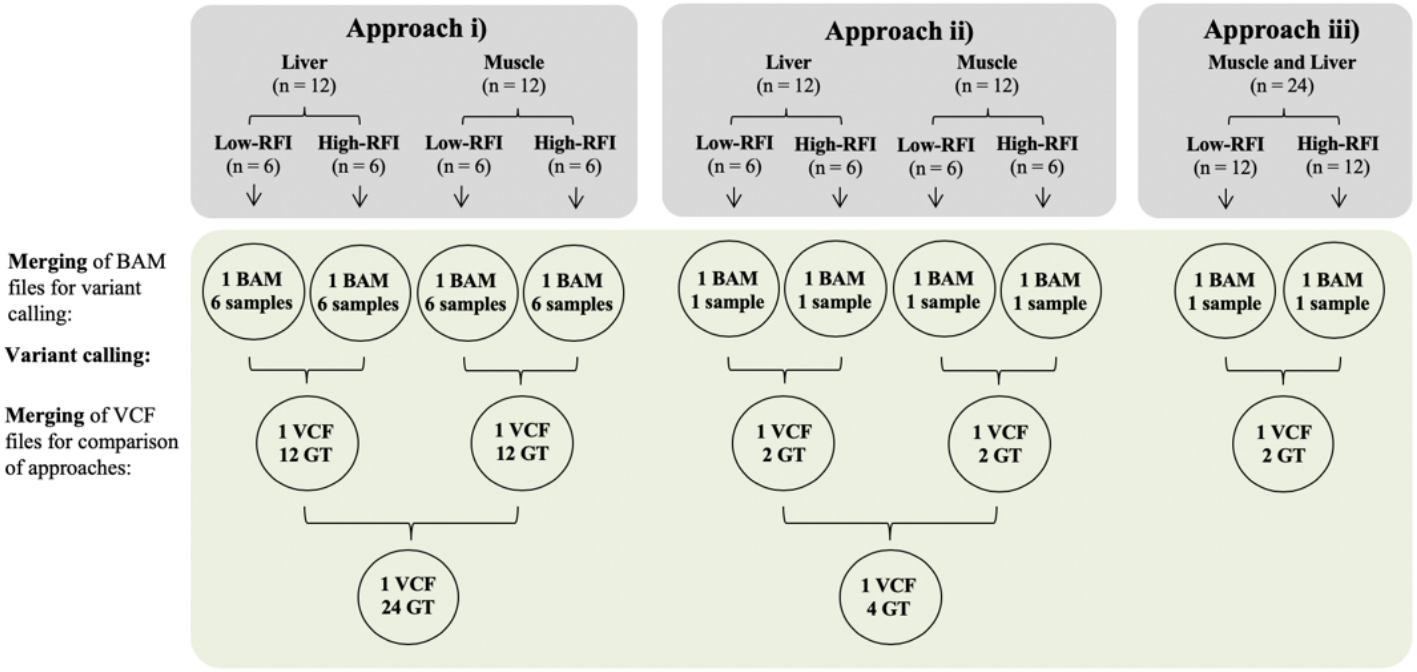

Figure 3

Workflow diagram of sample non-merging and merging strategies for each approach and for approach comparisons RFI = residual feed intake; n = sample size; BAM = Binary Alignment Map file; VCF = Variant Calling Format; GT = Genotype

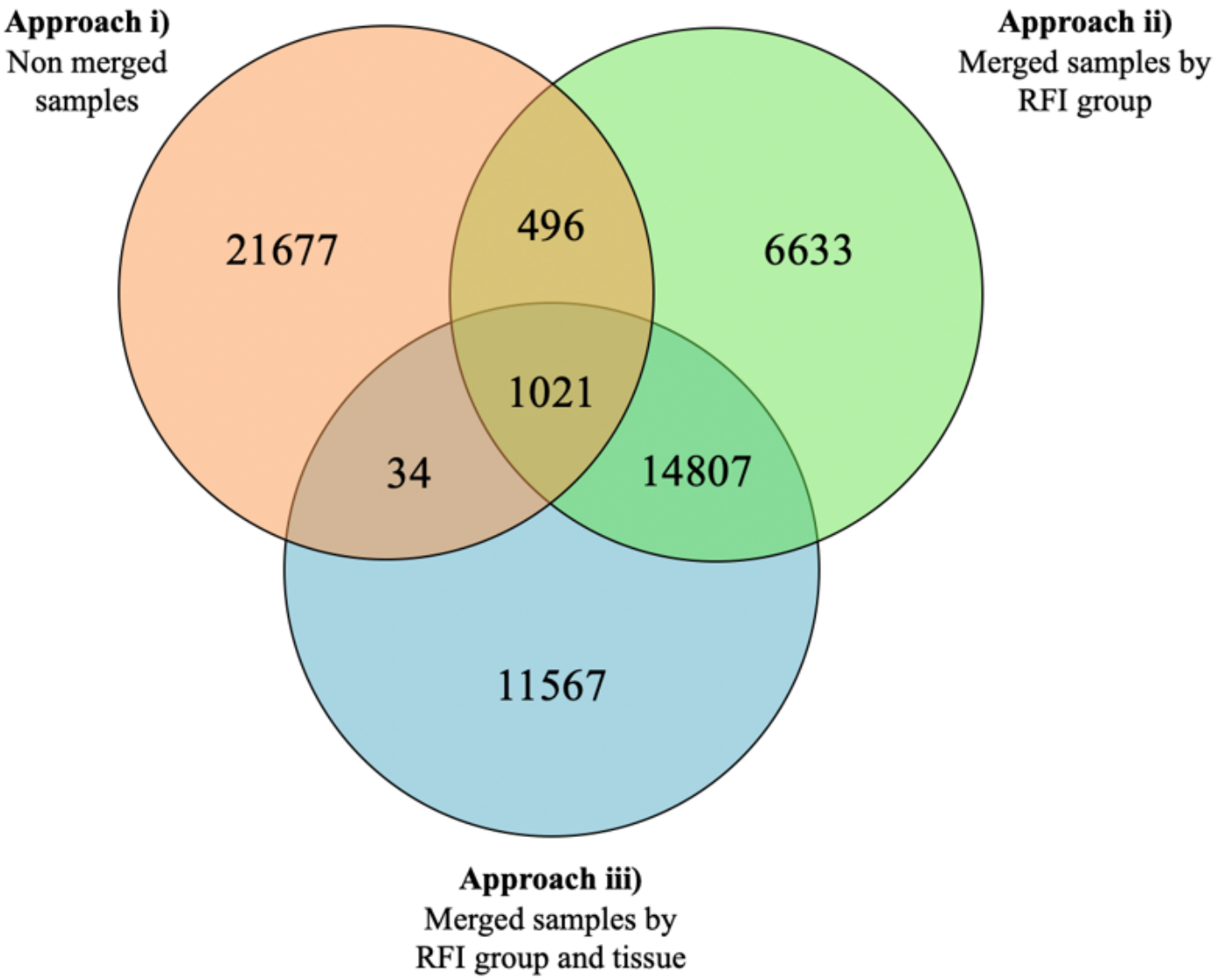

Figure 4

Common and unique SNPs which were fixed within low- or high-RFI groups across each approach. Approach i) = Non-merged samples Approach ii) = Merged samples for low-RFI and merged samples for high-RFI for each tissue Approach iii) = Merged samples for low- and high-RFI for both tissues

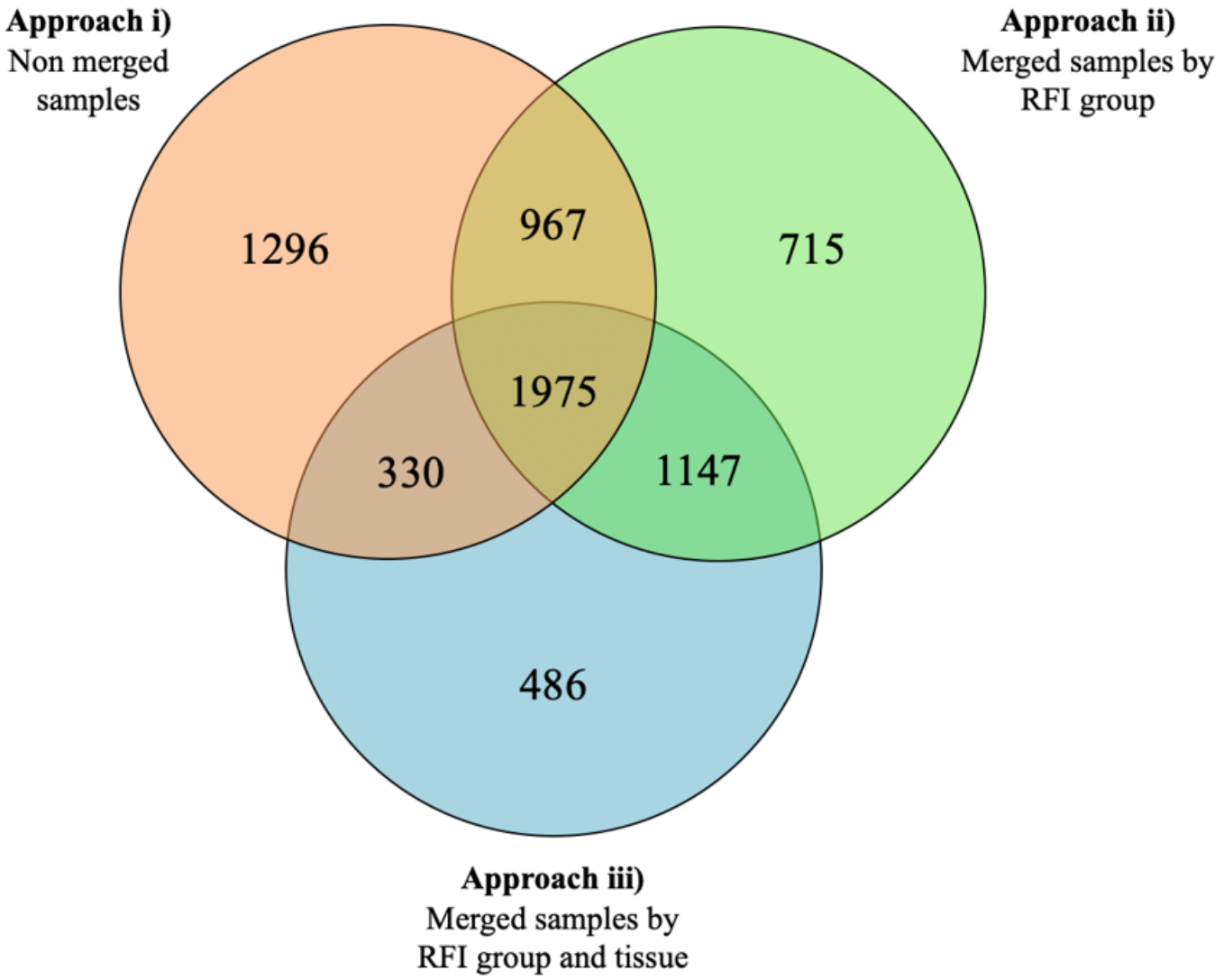

Figure 5

Common and unique genes co-localized with variants fixed within low- and high-RFI groups from each approach. Approach i) = Non-merged samples Approach ii) = Merged samples for low-RFI and merged samples for high-RFI for each tissue Approach iii) = Merged samples for low- and high-RFI for both tissues

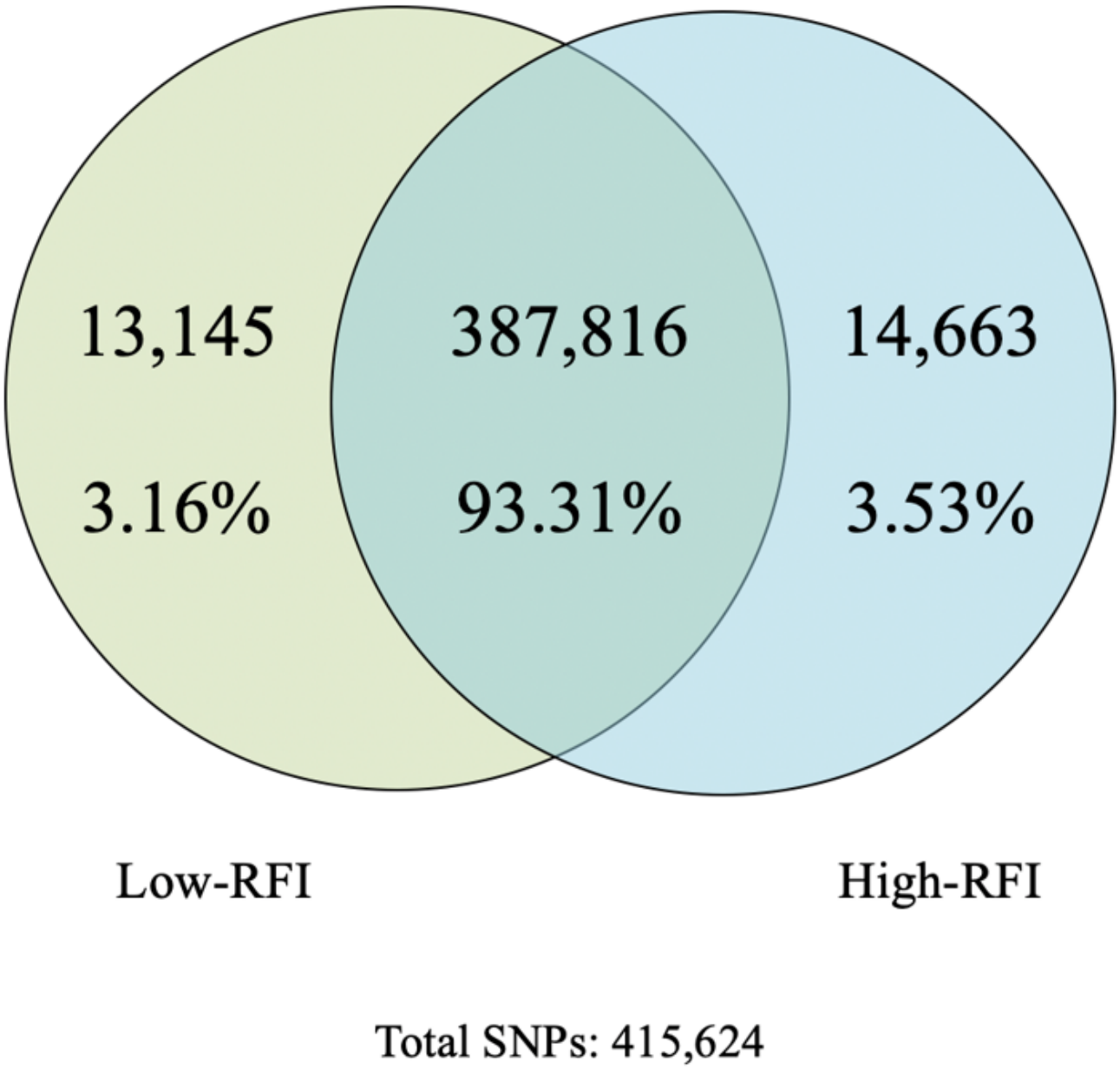

Figure 6

Unique or common SNPs between low- and high-RFI groups using the most optimized RNA-Seq pipeline*
*Approach iii) merged samples for low- and high-RFI for both tissues.