GigaScience

GALLO: An R package for Genomic Annotation and integration of multiple data source in livestock for positional candidate LOci

--Manuscript Draft--

Manuscript Number:	GIGA-D-20-00265R2
Full Title:	GALLO: An R package for Genomic Annotation and integration of multiple data source in livestock for positional candidate LOci
Article Type:	Technical Note
Funding Information:	Sustainable Beef and Forage Science Cluster funded by the Canadian Beef Cattle Check-Off, Beef Cattle Research Council, 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 (FDE.13.17) Dr Angela Cánovas
Abstract:	The development of high-throughput sequencing and genotyping methodologies and precision livestock farming allowed the identification of thousands of genomic regions associated with several complex traits. The integration of multiple sources of biological information is a crucial step to better understand patterns regulating the development of complex traits. Genomic Annotation in Livestock for positional candidate LOci (GALLO) is an R package, for the accurate annotation of genes and quantitative trati loci (QTLs) located in regions identified in the most common genomic analyses performed in livestock, such as Genome-Wide Association Studies and transcriptomics using RNA-Sequencing. Moreover, GALLO allows the graphical visualization of gene and QTL annotation results, data comparison among different grouping factors (e.g., methods, breeds, tissues, statistical models, studies, etc.), and QTL enrichment in different livestock species including cattle, pigs, sheep, chicken, etc. Consequently, GALLO is a useful package for annotation, identification of hidden patterns across datasets, datamining of previous reported associations, as well as the efficient scrutinization of the genetic architecture of complex traits in livestock.
Corresponding Author:	Pablo Augusto de Souza Fonseca University of Guelph Guelph, ON CANADA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Guelph
Corresponding Author's Secondary Institution:	
First Author:	Pablo Augusto de Souza Fonseca
First Author Secondary Information:	
Order of Authors:	Pablo Augusto de Souza Fonseca
	Aroa Suárez-Vega
	Gabiele Marras
	Angela Cánovas
Order of Authors Secondary Information:	
Response to Reviewers:	Reviewer reports: Reviewer #1: The authors have clarified the questions I had and included the clarification appropriately in the manuscript. L298-302: this should be in the discussion not in the conclusion. One reviewer

suggested to include more species than just livestock and the authors responded that it is possible as long as the species has a resource list like the AnimalQTL database. "We acknowledge this comment in the revised version of the manuscript and have included a sentence highlighting the applicability to other species (Lines 298-302)." This sentence is added to the conclusion, however, I urge discussing it in the discussion. It is a point of discussion and not a conclusion of the core manuscript. (If properly discussed it may be included in the conclusion.)

Answer: Thank you for your suggestion. We edited the current version of the manuscript and the information about applicability to other species was moved to the discussion section.

Some more textual errors arose in the newly written sections:
L277: remove one 'the'

Answer: Done.

L278: change 'sama' into 'same'

Answer: Done.

L282: change 'easy to be handle' into easy to handle of easy to be handled

Answer: Done.

L283: change 'have' into 'has'

Answer: Done.

Just a note for future revisions: For this comment & answer below (Reviewer 1) I didn't find the sections on the lines indicated, but elsewhere (141-153 & 277-285). Please make sure you refer to the correct line-numbers in the future to accommodate the reviewers.

The authors indicated that the R package is similar to BiomaRt, and gave performance differences in term of execution time of comparable commands. BiomaRt is a renowned package and was faster. It would be nice if the authors can indicate what benefits GALLO has over BiomaRt. Why was this package needed (e.g. what did you miss in biomaRt)?
Also it may be worthwhile to explicitly indicate why R is the appropriate language for this package. There are thing mentioned scattered over the paper, e.g. like visuals and no need for intermediate output files, please summarize them somewhere.

Answer: Thank you for the comment. The comparison between GALLO and other available tools is better discussed on lines 241-253 and 468-476 of the revised version of the manuscript.

Answer: Thank you very much for the comment. In the next submissions the authors will be awarded about this issue.

Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and	Yes

statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

GALLO: An R package for Genomic Annotation and integration of multiple data sources in Livestock for positional candidate LOci

Pablo A.S. Fonseca[1*], Aroa Suárez-Vega[1], Gabriele Marras[1,2], and Ángela Cánovas[1*]

[1]University of Guelph, Department of Animal Biosciences, Centre for Genetic Improvement of Livestock, Guelph, N1G 2W1, Ontario, Canada.

[2]The Semex Alliance, Guelph N1G 3Z2, Ontario, Canada

Contact:

PASF: pfonseca@uoguelph.ca - https://orcid.org/0000-0002-6917-7475

ASV: asuarezv@uoguelph.ca - https://orcid.org/0000-0002-7726-4288

GM: gmarras@uoguelph.ca - https://orcid.org/0000-0001-5115-370X

AC: acanovas@uoguelph.ca -https://orcid.org/0000-0002-0036-0757

[*] Corresponding author

Abstract

Background: The development of high-throughput sequencing and genotyping methodologies allowed the identification of thousands of genomic regions associated with several complex traits. The integration of multiple sources of biological information is a crucial step required to better understand patterns regulating the development of these traits. **Findings:** Genomic Annotation in Livestock for positional candidate LOci (GALLO) is an R package developed for the accurate annotation of genes and quantitative trait loci (QTLs) located in regions identified in common genomic analyses performed in livestock, such as Genome-Wide Association Studies and transcriptomics using RNA-Sequencing. Moreover, GALLO allows the graphical visualization of gene and QTL annotation results, data comparison among different grouping factors (e.g., methods, breeds, tissues, statistical models, studies, etc.), and QTL enrichment in different livestock species including cattle, pigs, sheep, and chickens, etc. **Conclusions:** Consequently, GALLO is a useful package for the annotation, identification of hidden patterns across datasets, datamining previously reported associations, as well as the efficient scrutinization of the genetic architecture of complex traits in livestock.

Keywords: Multi-omics integration; QTL annotation; Gene annotation; Datamining; QTL enrichment analysis; Livestock

Background

The identification of quantitative trait loci (QTLs), genomic regions linked to complex traits through association tests using genetic markers and phenotypic traits, is a crucial step in the improvement of genomic selection and economic profitability in livestock [1–4]. The development of high-throughput methodologies (e.g., Genome-Wide Association Studies, Transcriptomics, Metabolomics, Proteomics, etc.) for the study of the genetic architecture of complex traits allows for the identification of potential candidate genes associated with economically relevant traits in livestock. Taken together, these technologies can substantially improve the accuracy of detection of candidate regions associated with economically important traits across the genome in livestock species [5]. Consequently, the number of QTLs identified across the genome in livestock species increased substantially in the last few years. As of October 2020, the Animal QTLdb can retrieve information about QTLs previously identified in cattle (159,844), chickens (12,508), horses (2,451), pigs (30,871), rainbow trout (584) and sheep (3,411) [6]. The proper integration of results obtained from different methodologies and technologies is a crucial step for the accurate identification of the biological processes regulating complex traits as well as, the identification of potential functional candidate genes for each trait or those shared among traits [5,7–9]. The integration of both structural and functional data can help scrutinize the genetic architecture of economically relevant traits, and consequently, help to better understand complex biological patterns regulating the expression of these traits, such as pleiotropic effects, epistasis, and genetic hitchhiking, among others.

Despite the potential to improve the identification of functional candidate genes and/or QTLs through the integration of multiple data sources, the current process poses limitations in the pipelines and algorithms implemented in the tools available for livestock. Currently, there are several tools that implement functions for gene (i.e., Biomart and BEDTools) and QTL annotation (Animal QTLdb) [6,10,11]. However, these tools have limitations regarding the automatization process to analyze results from multiple candidate regions (Biomart web application and the R package and Animal QTLdb) or for the visualization of the results. Moreover, although automatization is possible, there is no direct link between the candidate regions and/or markers with the annotated genes and QTLs. Consequently, this gap forces the user to back solve the overlap between the input and output files in order to perform the proper association between the candidate region and/or markers and the annotated genes and/or positional co-localized QTLs. In addition, there is still a need for customized QTL enrichment analyses in the available software and databases. The Genomic Annotation in Livestock for positional candidate LOci (GALLO) is an R package designed to provide an automatized and a straightforward environment for gene and QTL annotation in multiple candidate regions, as well as the integration of data from multiple sources. Additionally, the QTL enrichment analysis can be performed directly by GALLO using the output obtained from the QTL annotation step. GALLO also provides a set of functions for graphical visualization of the annotation, comparison, integration and QTL enrichment results. In this context, the GALLO package was developed as an alternative tool: 1) to allow the integration and simultaneous annotation of multiple datasets for genes and QTLs; 2) to provide graphical visualization tools to visually integrate the annotation and similarity against datasets; 3) to perform QTL enrichment analysis for the positional candidate genomic regions and/or markers associated with economically relevant traits in livestock.

Implementation

The GALLO package was written in the R language [12]. The stable release is available as an R package on CRAN (https://cran.r-project.org/web/packages/GALLO/index.html). The code was extensively tested with several datasets from different sources and methodologies and reviewed to ensure it meets the packages high quality standards. Additionally, the vignettes were created to be comprehensive and to present practical examples in order to provide a user-friendly tutorial.

The GALLO package provides a useful set of functions that gives a straightforward approach to data integration, comparison, gene and QTL annotation, and visualization of several data sources and methodologies, such as variants from genome-wide association study (GWAS), RNA-Sequencing, whole-genome sequencing, etc. (Figure 1 and Table 1). The main advantage to perform an automated analysis from multiple datasets is the ability to handle the output using different subsets (traits, populations, models, etc.) in the same environment without generating multiple intermediate output files.

Case study – Candidate regions for scrotal circumference and fertility in cattle

The dataset used to present the basic usage and advantages of the GALLO package is composed by the markers significantly associated with scrotal circumference in the Canchim breed [13] and noncompensatory fertility in Holstein cattle [14]. These two studies were previously analyzed together in a systematic review regarding male fertility in cattle [8]. Therefore, the data used herein comprises a multi-study and multi-breed analysis. These candidate markers (527 single nucleotide polymorphisms (SNPs)) are available in Supplementary Table 1. In addition to the candidate markers, we presented as Supplementary Files 1 and 2, the annotation gff file containing the QTL database information for cattle (obtained from the Animal QTLdb; https://www.animalgenome.org/cgi-bin/QTLdb/BT/download?file=gffUMD_3.1) and the gtf file containing the genes annotated in the cattle genome obtained from Ensembl (ftp://ftp.ensembl.org/pub/release-94/gtf/bos_taurus/). The genomic coordinates of both files were based on the bovine reference genome version UMD 3.1 due to the original coordinates used to report the location of the candidate markers in the original studies. Here, the analysis performed follows the same logical order to the one presented in the GALLO vignette (https://rpubs.com/pablo_bio/GALLO_vignette). However, the dataset used in the user practical tutorial is a subset of the data presented here, aiming to reduce the computational demand for the user. The script with all the commands used to perform the analysis presented here are available in Supplementary File 3. All the tests were performed using a desktop with a processor Intel Core i5 2.4 GHz with 8 Gb of RAM memory.

Importing datasets and annotating genes and QTLs around candidate markers

The first step in the pipeline consists of importing the databases which will be used for the analysis with the *import_gff_gtf()* function. In our specific example, we imported both cattle gene annotation (gtf) and QTL (gff) databases. The *import_gff_gtf()* function receives the database file (db_file) and the file type (*file_type= "gff" or "gtf"*) as arguments and creates a dataframe with the respective information from each file. The system time taken to import the gtf and gff files were 0.045 and 0.311 seconds, respectively, indicating an efficient importing process. The file containing the candidate markers can be imported using any available function in the R environment such as *read.table()* and *read.csv()*.

The main function of GALLO, *find_genes_qtls_around_markers()*, performs the annotation of genes and/or co-localized QTLs within or nearby candidate markers or genomic regions (using the user's defined interval/window). This function uses the information provided in the .gtf file (for gene annotation) or .gff (for QTL annotation) to retrieve the requested information. The output combines the information available in the input file provided by the user with the information available for the genes and QTLs mapped in the candidate genomic regions. For example, for an input file composed of three genomic coordinates where four genes are annotated in each of the intervals determined by the user, the output file of *find_genes_qtls_around_markers()* will contain 12 rows. The minimum information necessary for the gene and QTL annotation procedures is a data frame with two columns containing the chromosome (CHR) and position in base pairs (BP) in the case of the candidate SNPs input file. In the case of the candidate haplotypes, windows, copy number variations (CNVs) or candidate regions; the input file is composed by three columns corresponding to the chromosome (CHR), the start position in base pairs (BP1) and the end position in base pairs (BP2). Data examples for the candidate markers and windows input files can be obtained using the *data("QTLmarkers")* and *data("QTLwindows")* commands in R. Additionally, examples of QTL and gene annotation results are accessible through the *data("gtfGenes")* and *data("gffQTLs")* commands, respectively. These outputs can be easily handled by summary functions in R, such as *table()*, to obtain information such as the total number of genes and QTLs, the number of genes and QTLs annotated per variants, etc. The gene annotation process was compared with the *getBM()* function from the biomaRt package. The gene annotation process on GALLO needed 0.424 seconds to completely annotate the genes in a 200 Kb interval (upstream and downstream) from candidate markers, while the biomaRt function required 0.019 seconds. The QTL annotation on GALLO was compared with the Bedtools -wao -C command, resulting in 0.851 and 0.12 seconds required for each approach, respectively. It is important to highlight that for both gene and QTL annotation using biomaRt and Bedtools, respectively, a posterior processing of the output file is required in order to match the candidate markers and the genes and QTLs mapped within the candidate intervals. On the other hand, the output file from *find_genes_qtls_around_markers()* function was designed to allow this match in an intuitive way, combining the rows of both candidate markers file and database files (gff and gtf). Additionally, GALLO allows the user to perform both annotations for genes and QTLs with a single software and programming language. Consequently, GALLO obtains a more elaborate and informative output without substantially compromising the computational demand required for the analysis. The output files obtained in the gene and QTL annotation are available on Supplementary Tables 2 and 3, respectively.

Comparing and visualizing the overlapping of genes and QTLs annotated within the candidate regions

The output file generated by the *find_genes_qtls_around_markers()* function can be used as an input file for the other set of GALLO functions. An advantage from the output of *find_genes_qtls_around_markers()* function is that any additional information present in the input file will be retained in the output file. Consequently, this information can be used to compare the retrieved information between groups of population, methodologies, statistical models, etc. For example, the functions *overlapping_among_groups()* and *plot_overlapping()* can be used to create matrices with the overlapping values among groups and to visualize this overlap. Figure 2 shows the genes and QTLs overlapping between the positional markers obtained in the two selected studies from the dataset of markers analyzed, Feugang et al. (2009) [14] and Buzanskas et al. (2017) [13]. It is important to highlight that the overlapping matrix informing the percentage of shared records is not symmetrical. The percentage of genes from study A shared with the study B, and vice-versa, are calculated as a function of the total number of genes in A or B, respectively. Briefly, this matrix is not symmetrical because GALLO calculates the percentage of records shared as a function of the total number of records for each group. For example, groups A and B shared 5 records, where group A has 10 records in total and group B has 5 records. Consequently, the percentage of shared records in A is 50% while the percentage of shared genes in B is 100%. In the current example, it is possible to note that only a small percentage of the positional candidate genes were shared between the studies. However, the analyses of overlapping QTLs (using the trait name as reference ID) indicated a higher similarity between the studies, 46% of the QTLs annotated in the candidate regions from Feugang et al. (2010) [14] were also present in Buzanskas et al. (2017) [13] and 93% of the QTLs annotated in the candidate regions from Buzanskas et al. (2017) were also present in Feugang et al. (2010) [13,14].

Understanding the QTL context of the candidate regions

A more precise investigation of the QTL representativeness and diversity can help to better understand the genomic context of the candidate regions. The recurrent association of particular genomic regions with multiple traits might suggest the presence of complex genetic mechanisms regulating that region, such as pleiotropy, epistasis, hitchhiking effect, among others [15,16]. The *plot_qtl_info()* function from GALLO allows for the graphical visualization of the summary of QTL types and traits annotated. The percentage of each QTL type for cattle (i.e., milk, meat and carcass, health, production, reproduction and exterior) annotated within the candidate regions is presented in a pie plot through the use of the argument *qtl_plot=”qtl_type”*, while the percentage of each trait associated with a specific QTL type can be plotted using the argument *qtl_plot=”qtl_name”* and informing the additional argument *qtl_class* (that must receive the name of the QTL class to be plotted). Figure 3 shows that for Feugang et al. (2009) [14] the two most frequent QTL types were Milk (50.42%) and Reproduction (16.97%), while for Buzanskas et al. (2017) [13] the most frequent QTL types were Reproduction (87.06%) and Meat and Carcass (5.03%). An in-depth analyses can be performed for each QTL type in order to observe the frequency of each trait associated with a specific QTL type. The most frequent traits related with

Reproduction QTLs were calving ease (>3%) and scrotal circumference (>60%) for Feugang et al. (2009) and Buzanskas et al. (2017) [13,14], respectively (Figure 3). The comparison between the frequency of traits related with Reproduction QTLs annotated in Feugang et al. (2009) and Buzanskas et al. (2017) [13,14] indicated that among the top 10 most frequent QTLs, calving ease, inhibin levels, stillbirth, interval to first estrus after calving, and birth index were shared between the studies. The combined analysis (not filtering by study) indicated that the Reproduction and Milk QTL types were the two most frequent classes with 76.99% and 10.62% of all QTL types, respectively. In addition, scrotal circumference, inhibin level and calving ease were the most frequent Reproduction QTL related traits in the combined analysis.

QTL enrichment analysis

In some cases, the biases produced with more research in certain areas/traits of higher relevance to animal production (such as milk production related traits in the QTL database for cattle) may result in a larger proportion of records for these traits in the QTL database. Consequently, the simple investigation of the proportion of each QTL type might not be totally useful. The GALLO package allows the user to perform a QTL enrichment analysis to test the significance of the QTL representativeness. The QTL enrichment analysis function in the GALLO package is based on a hypergeometric test approach, where the number of QTLs annotated within the candidate regions for each QTL type or trait, is compared with the observed number of QTLs in the reference database. Briefly, using an enrichment for individual traits in a chromosome-wide approach as an example, the number of traits per chromosome annotated within the candidate regions and the total number of each individual trait in the QTL database are computed. Subsequently, this information is integrated into a hypergeometric test in order to estimate if the number of observed records, for a specific trait, in a chromosome is larger than expected by chance. The *qtl_enrich()* function allows the user to perform the QTL enrichment analysis for both QTL types and traits (*qtl_type= "QTL_type"* or *"Name"*), for the whole genome or chromosome-wide (*enrich_type= "genome"* or *"chromosome"*) and for all the annotated chromosomes or a subset (*chr.subset= NULL* or the object with the subset of chromosomes). The use of a chromosome-wide enrichment analysis might help to detect specific regions across the genome with a high number of QTLs for a specific trait, i.e. BTA14 in cattle for milk production [17]. A total of 161 unique pairs of traits and chromosomes were tested for the enrichment using the annotated QTLs from both studies. The system time required to perform the enrichment analysis was 5.32 seconds, suggesting efficient processing. The top 10 enriched QTLs (False Discovery Rate (FDR) < 0.05) for the combined analysis is shown in Table 2 and the enrichment results for all the annotated QTLs is shown in Supplementary Table 4. Additionally, GALLO also allows the user to obtain a graphical visualization, in a bubble plot, of the enrichment results using the *QTLenrich_plot()* function. This function receives the enriched table obtained from *qtl_enrich()*, the name of the column with the trait names to be plotted and the name of the column with the p-values to be plotted as arguments. A total of 28 pairs of traits and chromosomes were found to be enriched in the combined analysis, with scrotal circumference (BTA 5, 18, 9, and 21), milk glycosylated kappa-casein percentage (BTA 6 and 16), inhibin level (BTA 5), triglyceride level (BTA 5), milk kappa-casein percentage (BTA 6) and milk iron content (BTA 23) in the list of top 10 most enriched traits. Figure 4 shows the top 5 enriched QTLs identified in this analysis.

Relationship between studies and enriched QTLs

An interesting functionality of GALLO is the graphical visualization of the relationship between groups using a chord plot. The *relationship_plot()* function receives as arguments a dataframe (it can use the gene or QTL annotation results, the QTL enrichment, or any other table with two groups of information to be compared), the two groups to be compared (arguments x and y) and the graphical arguments to set the size, color and gap between the sector in the chord plot. Figure 5 shows the chord plot obtained using a subset of the QTL annotation dataframe composed only by the top 10 enriched traits and the studies which these traits were annotated. This plot indicates that only inhibin levels and scrotal circumference on BTA5 are shared between Feugang et al. (2009) and Buzanskas et al. (2017) [13,14]. Additionally, milk glycosylated kappa-casein percentage (BTA 6 and 16), milk kappa-casein percentage (BTA 6) and milk iron content (BTA 23) were annotated only in Feugang et al. (2009) [14] and scrotal circumference (BTA 9, 18, 21) and triglyceride level (BTA 5) were annotated only in Buzanskas et al. (2017) [13]. Inhibin is produced by the Sertoli cells and can be used as a biomarker for sexual development [18]. In addition, the inhibin levels were already associated with both scrotal circumference and sperm quality traits in several studies, suggesting an important role in male fertility [19–23]. The results obtained here through the integration of the GWAS results from two independent studies followed by QTL annotation reinforces this association. Additionally, QTLs not associated with reproductive phenotypes were identified in the enrichment analysis, suggesting the presence of complex biological mechanisms such as a pleiotropic effect, epistasis and genetic hitchhiking effect. Previous studies have highlighted the possible role of genomic regions with these kinds of processes in the cattle genome [24,25]. An additional integration of the QTL annotation and enrichment analysis performed here with the gene annotation and prospection for functional candidate genes can be a powerful tool to better understand the genetic architecture and the relationship among complex traits.

Discussion

The GALLO package is composed of a group of functions designed to perform an efficient and direct downstream analysis for the gene and QTL annotation for candidate markers/SNPs, haplotypes, genomic windows, runs of homozygosity, CNVs, etc. The functions implemented in GALLO were designed to allow the integration of multiple datasets simultaneously. A brief summary of these functions is shown in Table 1. For example, GWAS results from multiple traits and/or populations or breeds can be analyzed together and compared or, individually analyzed in the downstream analysis. This can be easily performed by adding an extra column in the input file with the grouping factors to classify each dataset. These input files can be easily adapted from the output of commonly used softwares to analyze high-throughput genomic data, such as PLINK, BLUPF90, DESeq2, etc. [26–28]. In addition, GALLO provides a set of functions designed for the visualization of the annotation results, overlap among groups, relationship between groups (i.e., markers and candidate genes, datasets and QTLs, models and positional candidate genes, etc.), and QTL enrichment results. This set of functions provides the capability of integrating several results from multiple sources including different methodologies (GWAS, RNA-sequencing, proteomics, etc.), populations (breeds, time-points, etc.), traits or the different combination of these groups or others. Taken together, this set of functions provide to the possibility to perform all the steps of gene/QTL annotation, comparison and summary in the same environment. Additionally, the output obtained using GALLO was designed to allow a direct connection between the candidate genomic regions and the genes/QTLs which overlap those regions. Therefore, compared with outputs provided by other tools, such as biomaRt and Bedtools, the interpretation of the output provided by GALLO is straightforward and easy to handle. Finally, the QTL enrichment analysis available on GALLO is a useful and new approach that has the potential to better understand the relationship between candidate genomic regions and the target phenotype. It is important to highlight that despite the fact that GALLO was primarily designed for livestock species, the package can perform gene annotation and data comparison for any other species without any additional alterations to the input files. Regarding the QTL annotation and the respective graphical visualization, the user should provide the gff file from the QTL database in a format matching the gff files available on Animal QTLdb.

A summary of usage examples and output descriptions for all the functions available on GALLO can be found in the reference manual (Supplementary File 4). It is important to highlight that the two studies used as an example here are also part of the bovine QTL database. Consequently, the results obtained here for annotation and enrichment would be expected, once the candidate regions from the example file are present in the database used for the annotation. This approach was used as a proof of concept of the methodology and indicates a precise annotation of the candidate regions.

Conclusion

The integration of multiple datasets for gene and QTL annotation is one of the major bottlenecks for the automatization of functional analysis of the results obtained using high-throughput methodologies. The GALLO package provides a user-friendly and straightforward environment to perform gene and QTL annotation, visualization, data comparison and QTL enrichment for functional studies in livestock species. Consequently, the use of GALLO in the analyses of data generated from high-throughput methodologies may improve the identification of hidden patterns across datasets, datamining of previously reported associations, as well as efficiency in the scrutinization of the genetic architecture of complex traits in livestock.

Availability and requirements

Project name: Genomic Annotation in Livestock for positional candidate LOci (GALLO)

Project home page: https://github.com/pablobio/GALLO

Operating system(s): Platform independent

Programming language: R

Other requirements: Depends: R (>= 3.5.0)

License: GPL-3

SciCrunch: SCR_019212

Bio.tools: biotools:genomic_annotation_in_livestock_for_positional_candidate_loci_gallo

Availability of supporting data

All of the data analyzed in the present study can be accessed in the public repository hosting the R package (https://github.com/pablobio/GALLO). The input files and results used as examples in the manuscript preparation are available in the supplementary Tables 1-4. A manual including usage examples and output descriptions for all the functions available on GALLO can be found in the package vignette (https://cran.r-project.org/web/packages/GALLO/vignettes/GALLO.html). An archival copy of the code and supporting data is available via the GigaScience repository, GigaDB [28].

Declarations

List of abbreviations

BP: position in base pairs; BP1: start position in base pairs; BP2: end position in base pairs; CHR: Chromosome; CNV: Copy Number Variation; GALLO: Genomic Annotation in Livestock for positional candidate Loci; GWAS: Genome-Wide Association Study; QTL: Quantitative trait loci; SNP: Single Nucleotide Polymorphism.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by the Sustainable Beef and Forage Science Cluster (FDE.13.17) funded by the Canadian Beef Cattle Check-Off, Beef Cattle Research Council, 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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' contributions

PASF and AC were responsible for the conceptualization. PASF, ASV and AC were responsible for the data processing and review of the codes. PASF and ASV were responsible for data curation.

PASF and GM were responsible for the implementation of the bioinformatic pipeline, integration of datasets, and the coding. AC was responsible for funding acquisition.

Acknowledgements

Not applicable.

References

1. Ron M, Weller JI. From QTL to QTN identification in livestock - Winning by points rather than knock-out: A review. Anim. Genet. 2007.

2. Ernst CW, Steibel JP. Molecular advances in QTL discovery and application in pig breeding. Trends Genet. 2013.

3. Miglior F, Fleming A, Malchiodi F, Brito LF, Martin P, Baes CF. A 100-Year Review: Identification and genetic selection of economically important traits in dairy cattle. J Dairy Sci. 2017;

4. Pértille F, Guerrero-Bosagna C, Silva VH Da, Boschiero C, Nunes JDRDS, Ledur MC, et al. High-throughput and Cost-effective Chicken Genotyping Using Next-Generation Sequencing. Sci Rep. 2016;

5. Cánovas A, Reverter A, DeAtley KL, Ashley RL, Colgrave ML, Fortes MRS, et al. Multi-tissue omics analyses reveal molecular regulatory networks for puberty in composite beef cattle. PLoS One. 2014;

6. Hu ZL, Park CA, Reecy JM. Building a livestock genetic and genomic information knowledgebase through integrative developments of Animal QTLdb and CorrDB. Nucleic Acids Res. 2019;

7. De Souza Fonseca PA, Id-Lahoucine S, Reverter A, Medrano JF, Fortes MS, Casellas J, et al. Combining multi-OMICs information to identify key-regulator genes for pleiotropic effect on fertility and production traits in beef cattle. PLoS One. 2018;13:1–22.

8. Fonseca PA de S, dos Santos FC, Lam S, Suárez-Vega A, Miglior F, Schenkel FS, et al. Genetic mechanisms underlying spermatic and testicular traits within and among cattle breeds: Systematic review and prioritization of GWAS results. J Anim Sci. 2018;

9. Suárez-Vega A, Gutiérrez-Gil B, Benavides J, Perez V, Tosser-Klopp G, Klopp C, et al. Combining GWAS and RNA-Seq approaches for detection of the causal mutation for hereditary junctional epidermolysis bullosa in sheep. PLoS One. 2015;

10. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;

11. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;

12. R Core Team (2019). R: A language and environment for statistical computing. Accessed 1st April 2019. 2019;

13. Buzanskas ME, Grossi D do A, Ventura RV, Schenkel FS, Chud TCS, Stafuzza NB, et al. Candidate genes for male and female reproductive traits in Canchim beef cattle. J Anim Sci Biotechnol. 2017;

14. Feugang JM, Kaya A, Page GP, Chen L, Mehta T, Hirani K, et al. Two-stage genome-wide association study identifies integrin beta 5 as having potential role in bull fertility. BMC Genomics. 2009;

15. Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. Open Biol. 2017.

16. Id-Lahoucine S, Molina A, Cánovas A, Casellas J. Screening for epistatic selection signatures: A simulation study. Sci Rep. 2019;

17. Kühn C, Thaller G, Winter A, Bininda-Emonds ORP, Kaupe B, Erhardt G, et al. Evidence for multiple alleles at the DGAT1 locus better explains a quantitative trait locus with major effect on milk fat content in cattle. Genetics. 2004;

18. Phillips DJ. Activins, inhibins and follistatins in the large domestic species. Domest. Anim. Endocrinol. 2005.

19. Fortes MRS, Reverter A, Kelly M, Mcculloch R, Lehnert SA. Genome-wide association study for inhibin, luteinizing hormone, insulin-like growth factor 1, testicular size and semen traits in bovine species. Andrology. 2013;

20. Fortes MRS, Reverter A, Hawken RJ, Bolormaa S, Lehnert S a. Candidate genes associated with testicular development, sperm quality, and hormone levels of inhibin, luteinizing hormone, and insulin-like growth factor 1 in Brahman bulls. Biol Reprod [Internet]. 2012 [cited 2013 Sep 6];87:58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22811567

21. Bame JH, Dalton JC, Degelos SD, Good TEM, Ireland JLH, Jimenez-Krassel F, et al. Effect of Long-Term Immunization against Inhibin on Sperm Output in Bulls1. Biol Reprod. 1999;

22. Martin TL, Williams GL, Lunstra DD, Ireland JJ. Immunoneutralization of Inhibin Modifies Hormone Secretion and Sperm Production in Bulls1. Biol Reprod. 1991;

23. Sato T, Kudo T, Ikehara Y, Ogawa H, Hirano T, Kiyohara K, et al. Chondroitin sulfate N-acetylgalactosaminyltransferase 1 is necessary for normal endochondral ossification and aggrecan metabolism. J Biol Chem. 2011;

24. Bolormaa S, Pryce JE, Reverter A, Zhang Y, Barendse W, Kemper K, et al. A Multi-Trait, Meta-analysis for Detecting Pleiotropic Polymorphisms for Stature, Fatness and Reproduction in Beef Cattle. PLoS Genet. 2014;10.

25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.

26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;

27. Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, H LD. BLUPF90 and related programs (BGF90). Proc 7th world Congr Genet Appl to Livest Prod. 2002. p. 743–4.

28. Fonseca PAS, Suárez-Vega A, Marras G, Cánovas A. GALLO: An R package for Genomic Annotation and integration of multiple data source in livestock for positional candidate LOci GigaScience Database. 2020. http://dx.doi.org/10.5524/100834

Tables

Table 1: Description of the functions implemented in the GALLO package.

Function	Description	Output
Gene and QTL annotation		
import_gff_gtf	Import the gff and gtf files used for QTL and gene annotation, respectively	A dataframe composed by the information present in the gtf and gff files
find_genes_qtls_around_markers	Annotation of genes and QTLs around candidate regions	A data frame composed of the columns present in the input file and the genes or QTLs mapped within or around (if interval provided) the candidate regions
Data visualization		
overlapping_among_groups	Overlap between grouping factors (such as different traits, statistical models, populations, studies, stc.)	A list with three matrices: 1) A matrix with the number of overlapping data; 2) A matrix with the percentage of overlap; 3) A matrix with the combination of the two previous ones
plot_overlapping	Plot overlap between data and grouping factors	A heatmap with the overlap between groups
plot_qtl_info	Plot QTL information from the gene or QTL annotation output	A pie plot (if QTL class is chosen) or a bar plot (if trait name is chosen) for the annotated QTLs
relationship_plot	Plot the relationship among the candidate regions or	A chord plot linking a grouping factor (genomic regions, traits, populations, etc.) with the annotated genes or QTLs

	grouping factors with the annotated genes and QTLs	
QTL enrichment		
qtl_enrich	Performs a QTL enrichment analysis based on a Bootstrap simulation for each QTL class or trait	A data frame composed of the enrichment results for QTL classes or traits present in the input file. 1) QTL: The QTL class or trait used for the enrichment; 2) CHR: The chromosome for that specific QTL or trait (if the option "chromosome" is informed to the argument enrich_type); 3) N_QTLs: Number of observed QTLs or traits in the dataset; 4) N_QTLs_db: Number of each annotated QTL in the qTL database; 5) Total_annotated_QTLs: Total number of annotated QTLs; 6) Total_QTLs_db: Total number of QTLs in the QTL database; 7) pvalue: P-value for the enrichment analysis; 8) adj.pval: The adjusted p-value based on the multiple test correction selected by the user; 9) QTL_type= The QTL type for each annotated trait.
QTLenrich_plot	Creates a bubble plot with the QTL enrichment results	A plot with the QTL enrichment results

Table 2: Top 10 enriched QTLs for the combined analysis performed with the candidate regions from the two studies, Feugang et al. (2009) and Buzanskas et al. (2017), used in the example dataset.

QTL	CHR	# QTLs	# QTLs db	Total # QTLs	Total # QTLs db	p-value	FDR	QTL type
Scrotal circumference	5	132	134	347	5942	1.56E-171	4.98E-169	Reproduction
Scrotal circumference	18	11	13	41	2147	2.20E-18	3.52E-16	Reproduction
Scrotal circumference	9	11	14	30	1395	2.04E-17	2.18E-15	Reproduction
Milk glycosylated kappa-casein percentage	6	71	1607	204	12158	1.86E-15	1.49E-13	Milk
Inhibin level	5	47	285	347	5942	3.38E-11	2.16E-09	Reproduction
Scrotal circumference	21	4	5	12	3606	3.51E-10	1.87E-08	Reproduction
Milk kappa-casein percentage	6	76	2637	204	12158	2.39E-07	1.01E-05	Milk
Triglyceride level	5	6	7	347	5942	2.53E-07	1.01E-05	Health
Milk glycosylated kappa-casein percentage	16	7	44	21	1440	1.29E-06	4.58E-05	Milk
Milk iron content	23	4	8	19	1159	3.48E-06	0.000111329	Milk

Figure legends:

Figure 1: Workflow explaining the main functions implemented on GALLO. The grey rectangles represent the functions, while the rounded and sharp rectangles represent the main goal of that respective function and its input, respectively.

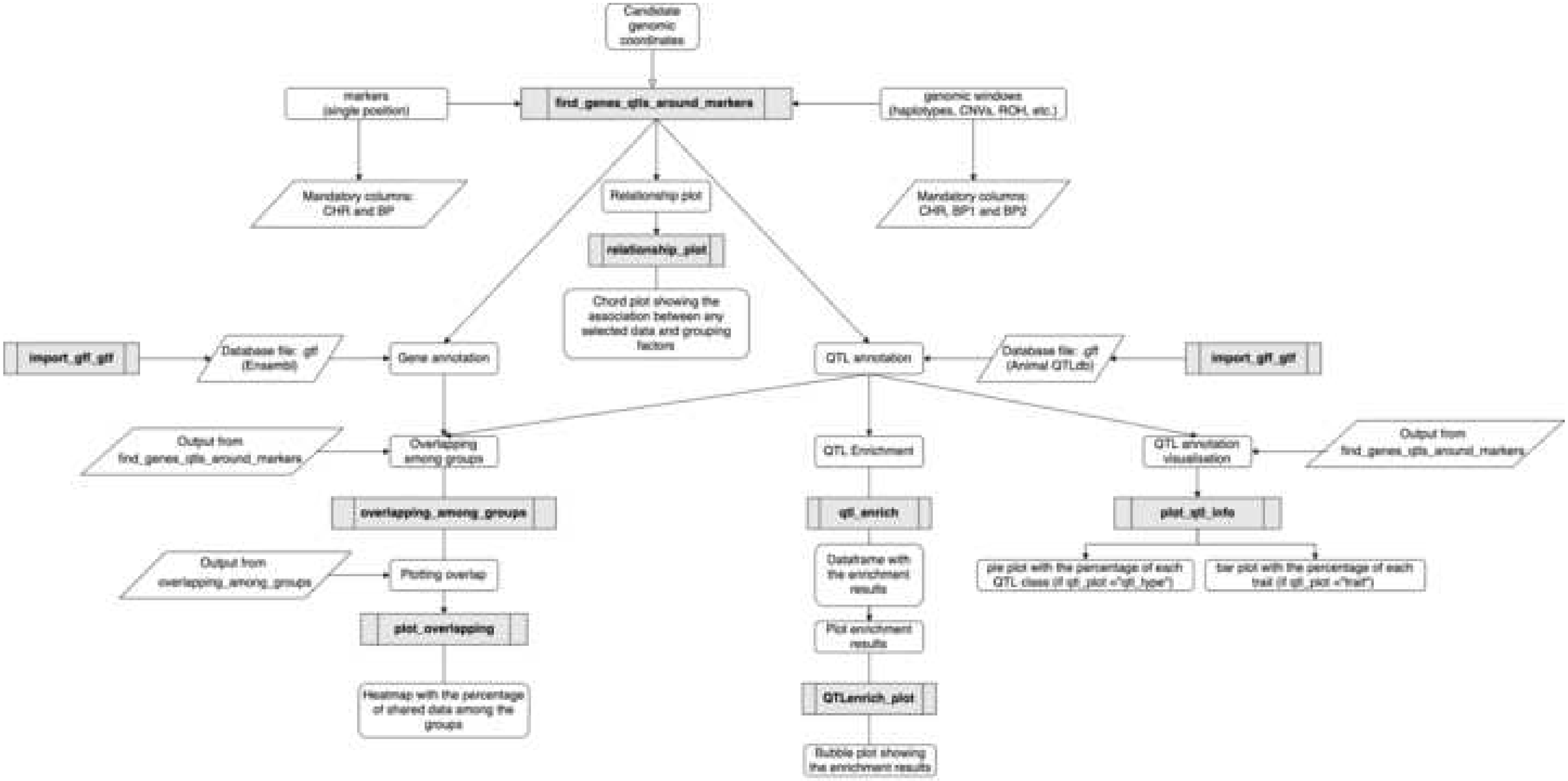
Figure 2: Overlapping between genes (A) and QTLs (B) annotated within the candidate regions (100 Kb downstream and upstream from the significant markers) from Feugang et al. (2009) and Buzanskas et al. (2017). The darker the color within the squares the higher the percentage of shared genes or QTLs.

Figure 3: Percentage of QTL type (pie plot) and trait related to Reproduction QTLs (barplots) for the QTL annotation results obtained for Feugang et al. (2009) (A), Buzanskas et al. (2017) (B) and the combined analysis (using both studies; C).

Figure 4: Bubble plot displaying the enrichment results for the top 5 enriched QTLs identified using the QTLs annotated within the candidate regions from Feugang et al. (2009) and Buzanskas et al. (2017). The darker the red shade in the circles, the more significant the enrichment. The area of the circles is proportional to the number of QTLs. The x-axis shows a richness factor obtained by the ratio of the number of QTLs annotated in the candidate regions and the total number of each QTL (and chromosome in the case of this plot) in the reference database.

Figure 5: Chord plot showing the relationship between the top 10 enriched QTLs (Scrotal circumference – SCRCIR, Inhibin level – INHIB, Triglyceride level – TRIGLY, Milk glycosylated kappa-casein percentage – MGKCASP, Milk iron content – MFE, Milk kappa-casein percentage - MKCASP) and the studies (Feugang et al. (2009) in purple and Buzanskas et al. (2017) in pink).

Figure 1


Candidate genomic coordinates
markers (single position)
find_genes_qtls_around_markers
genomic windows (haplotypes, CNVs, ROH, etc.)
Mandatory columns: CHR and BP
Relationship plot
Mandatory columns: CHR, BP1 and BP2
relationship_plot
Chord plot showing the association between any selected data and grouping factors
import_gff_gtf
Database file: gtf (Ensembl)
Gene annotation
QTL annotation
Database file: gff (Animal QTLdb)
import_gff_gtf
Output from find_genes_qtls_around_markers
Overlapping among groups
QTL Enrichment
QTL annotation visualisation
Output from find_genes_qtls_around_markers
overlapping_among_groups
qtl_enrich
plot_qtl_info
Output from overlapping_among_groups
Plotting overlap
Dataframe with the enrichment results
pie plot with the percentage of each QTL class (if qtl_plot ="qtl_type")
bar plot with the percentage of each trait (if qtl_plot ="trait")
plot_overlapping
Plot enrichment results
Heatmap with the percentage of shared data among the groups
QTLenrich_plot
Bubble plot showing the enrichment results

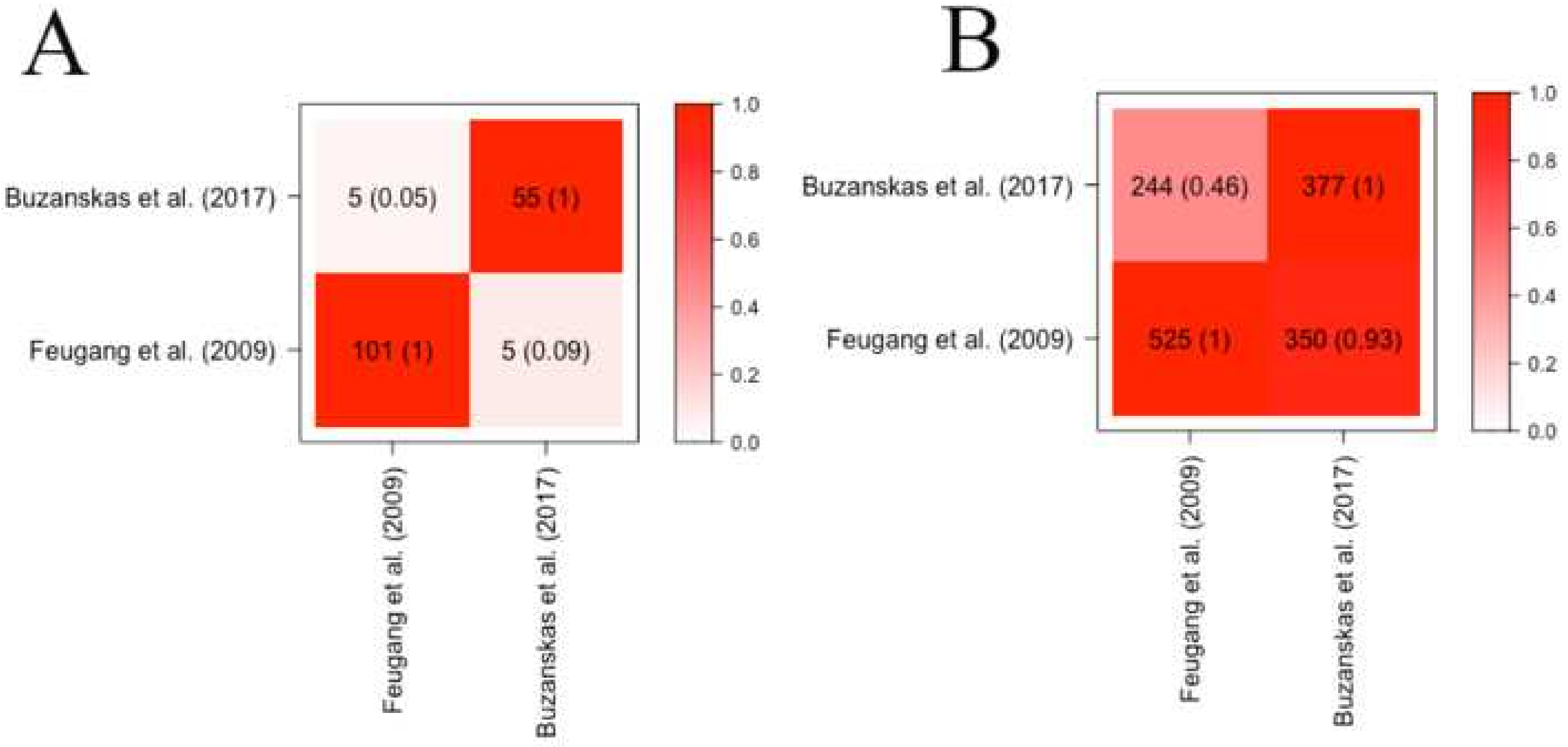

A
Buzanskas et al. (2017)
Feugang et al. (2009)
5 (0.05)
55 (1)
101 (1)
5 (0.09)
Feugang et al. (2009)
Buzanskas et al. (2017)
1.0
0.8
0.6
0.4
0.2
0.0
B
Buzanskas et al. (2017)
Feugang et al. (2009)
244 (0.46)
377 (1)
525 (1)
350 (0.93)
Feugang et al. (2009)
Buzanskas et al. (2017)
1.0
0.8
0.6
0.4
0.2
0.0

Figure 3

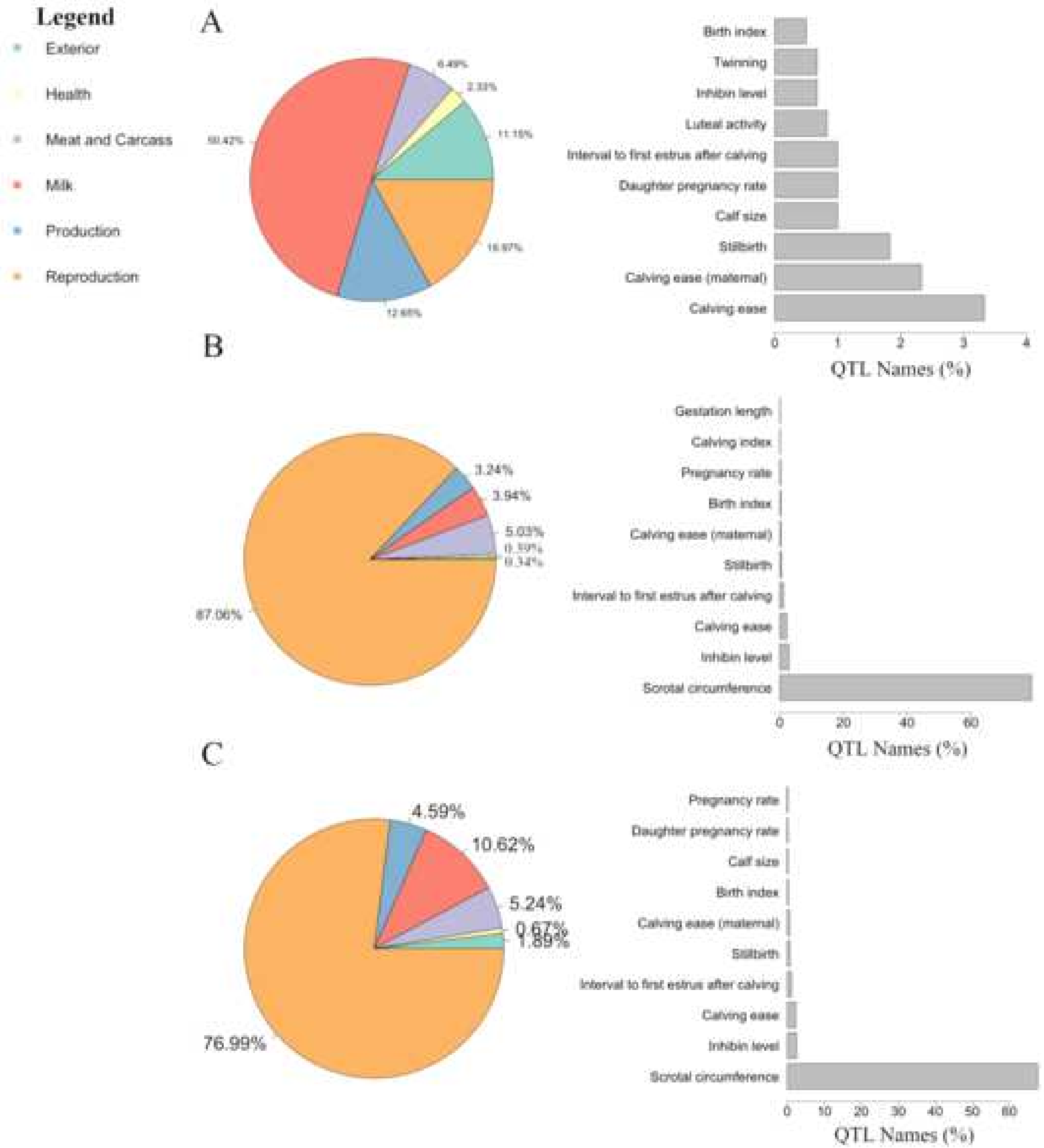
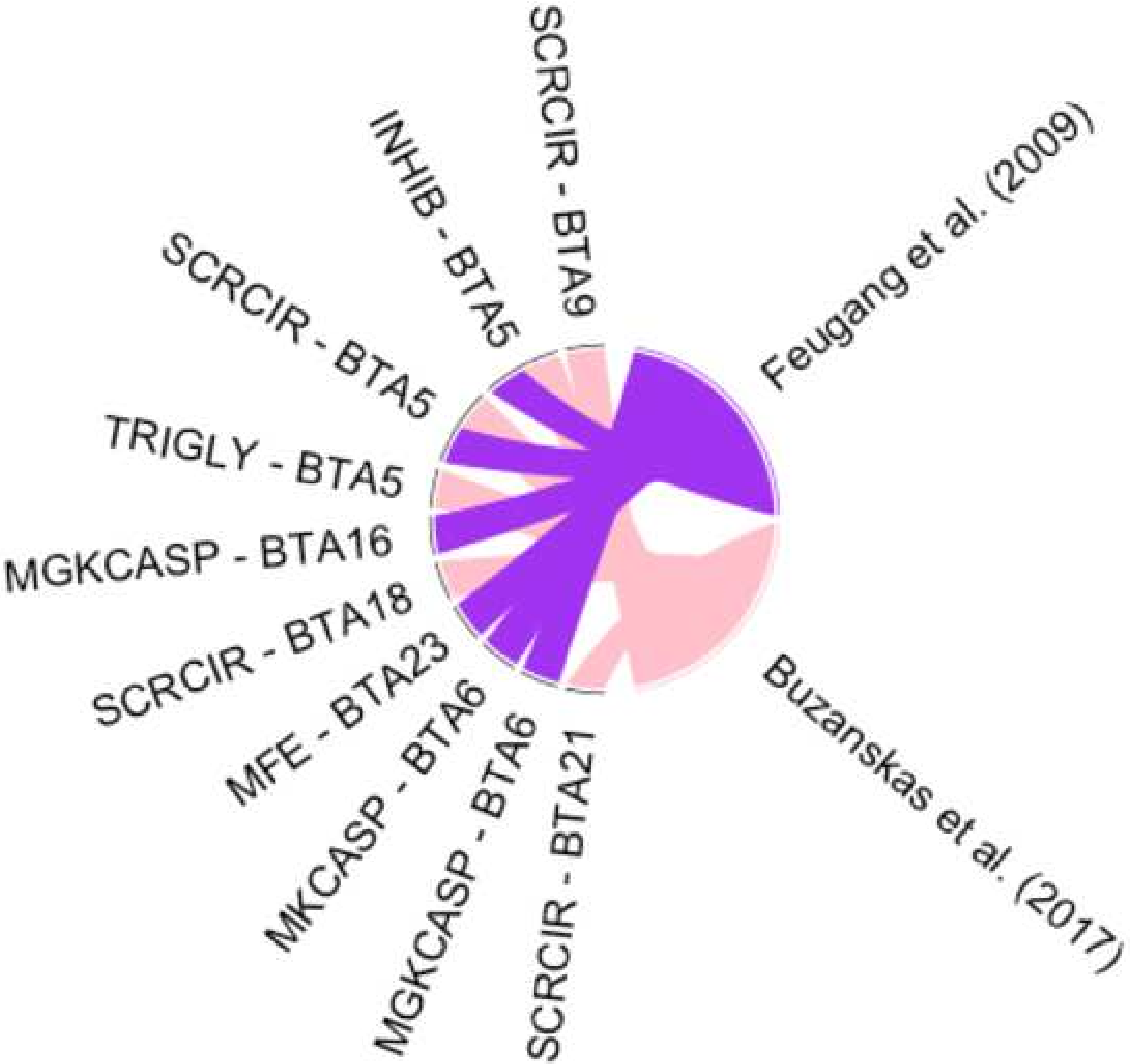
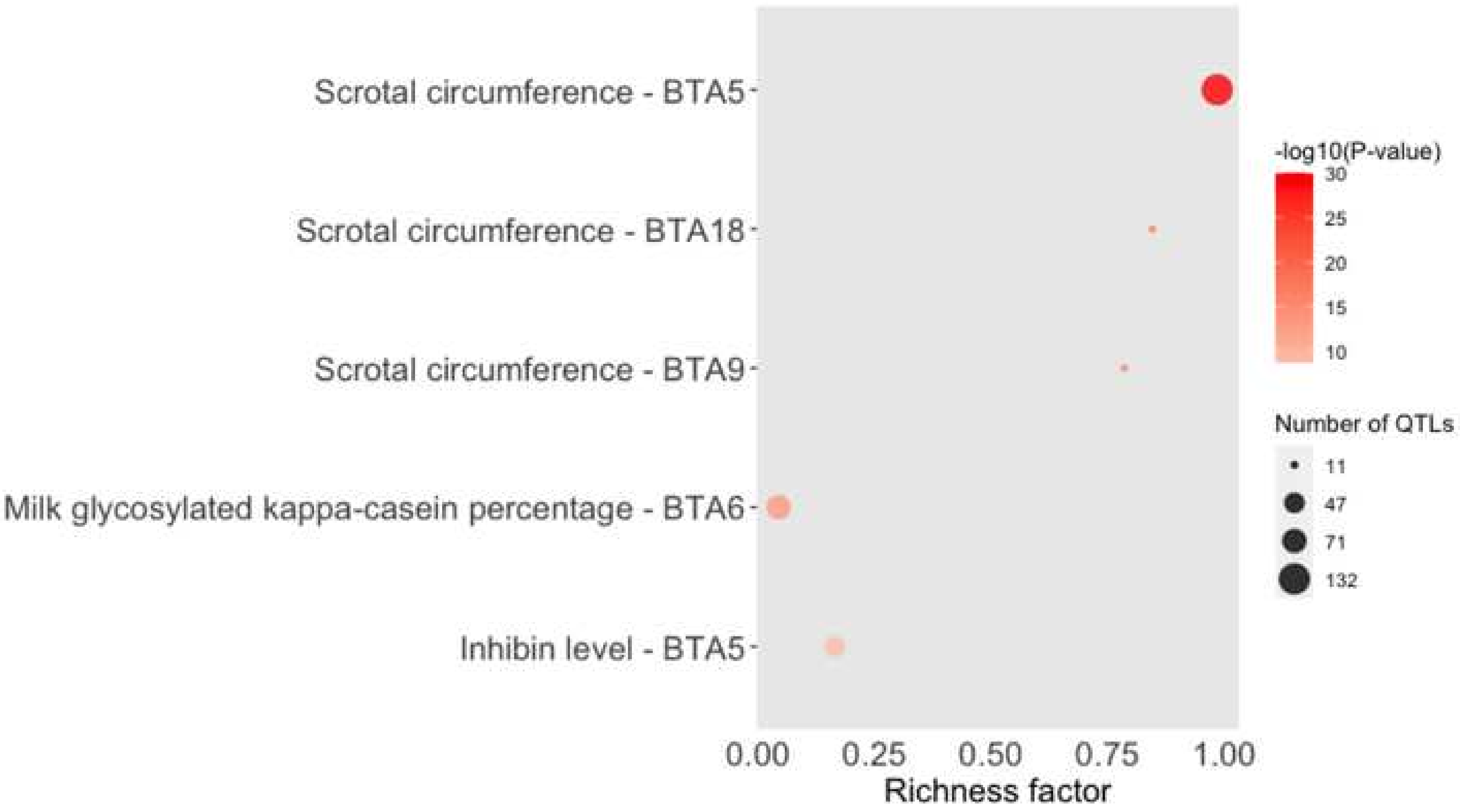

Legend
Exterior
Health
Meat and Carcass
Milk
Production
Reproduction
A
6.49%
2.33%
11.15%
16.97%
12.65%
50.42%
Birth index
Twinning
Inhibin level
Luteal activity
Interval to first estrus after calving
Daughter pregnancy rate
Calf size
Stillbirth
Calving ease (maternal)
Calving ease
0
1
2
3
4
QTL Names (%)
B
3.24%
3.94%
5.03%
0.39%
0.34%
87.06%
Gestation length
Calving index
Pregnancy rate
Birth index
Calving ease (maternal)
Stillbirth
Interval to first estrus after calving
Calving ease
Inhibin level
Scrotal circumference
0
20
40
60
QTL Names (%)
C
4.59%
10.62%
5.24%
0.67%
1.89%
76.99%
Pregnancy rate
Daughter pregnancy rate
Calf size
Birth index
Calving ease (maternal)
Stillbirth
Interval to first estrus after calving
Calving ease
Inhibin level
Scrotal circumference
0
10
20
30
40
50
60
QTL Names (%)

Figure 5


SCRCIR - BTA9
INHIB - BTA5
SCRCIR - BTA5
TRIGLY - BTA5
MGKCASP - BTA16
SCRCIR - BTA18
MFE - BTA23
MKCASP - BTA6
MGKCASP - BTA6
SCRCIR - BTA21
Feugang et al. (2009)
Buzanskas et al. (2017)


Scrotal circumference - BTA5
Scrotal circumference - BTA18
Scrotal circumference - BTA9
Milk glycosylated kappa-casein percentage - BTA6
Inhibin level - BTA5
0.00
0.25
0.50
0.75
1.00
Richness factor
-log10(P-value)
30
25
20
15
10
Number of QTLs
11
47
71
132

Supplementary Material

Guelph, September 1st, 2020

Dear Editorial Office,

We are pleased to re-submit the manuscript entitled "GALLO: An R package for Genomic Annotation and integration of multiple data source in livestock for positional candidate LOci" for consideration to publish it in the GigaScience. This is a resubmission of this manuscript after the inclusion of all the suggestion and considerations raised by the editor and the prior publication of the package in an official repository, in this case, the CRAN.

The present study introduces the applicability and the functionalities of GALLO package, developed in the R environment.

The identification of quantitative trait loci (QTLs) is a crucial step in the improvement of genomic selection and economic profitability in livestock. The development of high-throughput sequencing and genotyping methodologies and precision livestock farming allowed the identification of thousands of genomic regions associated with several complex traits. Consequently, the number of QTLs identified across the genome in livestock species increased substantially in the last years. Currently, in the Animal QTLdb it is possible to retrieve information about QTLs previously identified in cattle (127,191), chicken (11,340), horse (2,260), pig (29,865), rainbow trout (584) and sheep (3,001). The proper integration of the results obtained from different methodologies and technologies available is a crucial step for the accurate identification of the biological processes regulating the development of complex traits as well as the identification of potential functional candidate genes. However, currently, the integration of multiple data sources is not very straightforward due to limitations in the pipelines and algorithms implemented in the tools available for livestock. Moreover, although the automatization is possible, the direct link between the candidate regions and/or markers with the annotated genes and QTLs is missed. Consequently, this gap is forcing the user to back solve the overlap between the input and output files in order to perform the proper association between the candidate region and/or markers and the annotated genes and/or positional co-localized QTLs. In addition, nowadays there is still a lack of for customized QTL enrichment analyses in the available software and databases. Genomic Annotation in Livestock for positional candidate LOci (GALLO) is an R package, for the accurate annotation of genes and QTLs located in regions identified using the most common genomic analyses performed in livestock, such as Genome-Wide Association Studies and transcriptomics using RNA-Sequencing. Moreover, GALLO allows the graphical visualization of gene and QTL annotation results, data comparison among different grouping factors (e.g., methods, breeds, tissues, statistical models, studies, etc.), and QTL enrichment in different livestock species including cattle, pigs, sheep, chicken, etc. Consequently, GALLO is a useful package for annotation, identification of hidden patterns across datasets, datamining of previous reported associations, as well as the efficient scrutinization of the genetic architecture of complex traits in livestock.

We affirm that this manuscript has not been published elsewhere and is not under consideration by any other journal. All authors have approved the manuscript and agree with its submission to GigaScience.

The authors declare that they have no competing interests. With my best regards,

Angela Cánovas, PhD

Associate Professor Beef Genomics and Small Ruminants
University of Guelph
Department of Animal Biosciences
Centre for Genetics and Improvement of Livestock
Telephone: (519) 824-4129 ext. 56295
email: acanovas@uoguelph.ca