UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS AGRÁRIAS
DEPARTAMENTO DE ZOOTECNIA

TUTORIAL

Busca por genes e Quantitative Trait Loci (QTL) sobrepostos em regiões do genoma

Renata de Fátima Bretanha Rocha
Pamela Itajara Otto
Arielly Oliveira Garcia
Mateus Guimarães dos Santos
Marcos Vinicius Gualberto Barbosa da Silva
Marta Fonseca Martins
Marco Antonio Machado
João Cláudio do Carmo Panetto
Simone Eliza Facioni Guimarães

Viçosa

2023

TUTORIAL

Busca por genes e Quantitative Trait Loci (QTL) sobrepostos em regiões do genoma

Renata de Fátima Bretanha Rocha[1], Pamela Itajara Otto[2], Arielly Oliveira Garcia[1], Mateus Guimarães dos Santos[1], Marcos Vinicius Gualberto Barbosa da Silva[3], Marta Fonseca Martins[3], Marco Antonio Machado[3], João Cláudio do Carmo Panetto[3], Simone Eliza Facioni Guimarães[1]

[1]Departmento de Zootecnia, Universidade Federal de Viçosa, Viçosa, MG 36570-900, Brasil.
[2]Departmento de Zootecnia, Universidade Federal de Santa Maria, Santa Maria, RS 97105-900, Brasil.
[3]EMBRAPA – Gado de Leite, Juiz de Fora, MG 36038-330, Brasil

ISBN: 978-65-5668-139-9
DOI: http://dx.doi.org/10.26626/9786556681399.2023B0001

Agradecimentos
Agradecemos às fazendas e à Empresa Brasileira de Pesquisa Agropecuária (Embrapa) – Gado de Leite, Juiz de Fora – MG, que forneceram os dados para este estudo.

Financiamento
Este estudo recebeu o apoio financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil (CNPQ) - Processos 402935/2021-7, 142600/2019-9 e 200147/2022-6, da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)/PROEX 88887.844747/2023-00 e do Instituto Nacional de Ciência e Tecnologia de Ciência Animal (INCT-CA).

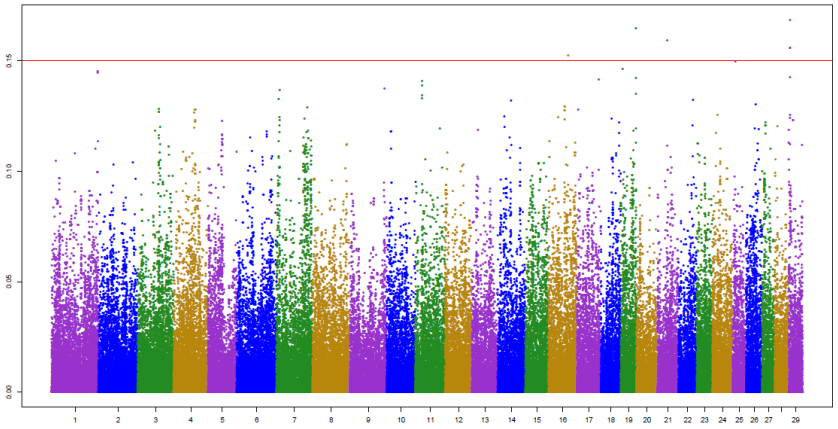
VIÇOSA
2023

Tutorial

Definindo a janela de busca

Resultados do GWAS

- Supondo que de uma análise de GWAS (ou qualquer outra) foram identificadas as seguintes regiões:



Chr	Position
4	99653824
8	98679541
16	60560912
19	5814828
19	61970518
21	32479895
29	4131935

O próximo passo é definir o intervalo de busca de genes

- Por exemplo:
- Para cada marcador/posição identificada, estabelecer um intervalo de até 20.000bp (janela de 20Kb), ou seja, 10 Kb acima ou abaixo dessa posição

- Se temos um marcador identificado na posição **99653824** no cromossomo 4

- Para pesquisar um gene na região desse marcador, vamos estabelecer o intervalo de 10 mil pares de bases antes e depois dessa região:
- **99653824** - 10.000 = 99643824 bases antes do marcador
- **99653824** + 10.000 = 99663824 bases depois do marcador

- Então o intervalo de busca dos genes será uma janela entre 99643824 e 99663824

- Existem outras opções para definir a janela de busca.
- A janela é estabelecida pelo pesquisador.

A exemplo, outras metodologias:

- Intervalo de até 250.000bp (250Kb) acima ou abaixo dessa posição, por Santos et al. (2017) doi: 10.1371/journal.pone.0169163

- Intervalo de até 10.000bp (10Kb) acima ou abaixo dessa posição, por Jaton et al. (2018) DOI: 10.3168/jds.2017-13848

➢ Estes são só alguns exemplos, os intervalos podem ser expandidos ou restringidos dependendo do que os autores acharem mais conveniente.

Montar um arquivo com as janelas de busca

Chr	Posição do marcador	Janela de busca	
		Posição inicial	Posição final
4	99653824	99643824	99663824
8	98679541	98669541	98689541
16	60560912	60550912	60570912
19	5814828	5804828	5824828
19	61970518	61960518	61980518
21	32479895	32469895	32489895
29	4131935	4121935	4141935

Busca de genes e QTLs

- Uma opção é usar os scripts do R sugeridos abaixo. Eles estão na pasta do tutorial de busca de genes e QTLs:

- Script_pesquisando_QTL.R

- Script_pesquisando_genes.R

Antes de rodar os scripts

- *Primeiramente é necessário fazer download do arquivo QTLdb_cattleARS_UCD1.bed.txt no site do Animal QTLdb*

- *Ver orientações a seguir:*

Baixando_infoQTL_QTLdb.pdf

QTL Mapping

Release 41
(Apr 26, 2020)

AnimalQTLdb

The Animal Quantitative Trait Loci (QTL) Database (Animal QTLdb) strives to collect all publicly available trait mapping data, *i.e.* QTL (phenotype/expression, eQTL), candidate gene and association data (GWAS), and copy number variations (CNV) mapped to livestock animal genomes, in order to facilitate locating and comparing discoveries within and between species. New data and database tools are continually developed to align various trait mapping data to map-based genome features such as annotated genes.

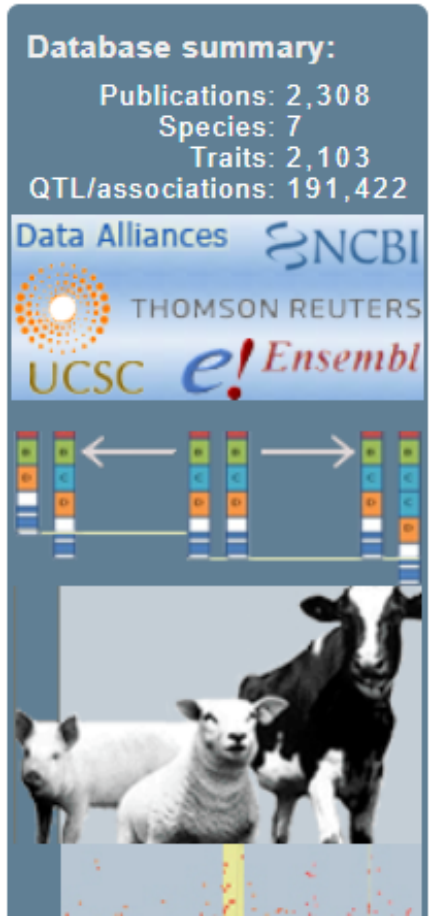
Many scientific journals require or recommend that any original QTL/association data be deposited into a public database before a paper may be accepted for publication. We provide user/curator accounts for direct data submission, and in return we supply users with a data summary link to facilitate the manuscript review process.

Cattle QTL

There are **142,261** QTL from **1,001** publications curated into the database. Those QTL represent **646** different traits (see **data summary** for details).

Catfish QTL

There are **0** QTL from **0** publications curated into the database. Those QTL represent **0** different traits (see **data summary** for details).



https://www.animalgenome.org/QTLdb/

QTL Mapping

Fazer o download e descompactar o arquivo

Animal_QTLdb_release46_cattle ARS_UCD1.bed.gz

QTL_ARS-UCD_1.2.bed

Animal QTLdb | Browse | Search | View Maps | F A Q | Release 41 (Apr 26, 2020)

CattleQTLdb

general search | search

This Cattle Quantitative Trait Locus (QTL) Database (**Cattle QTLdb**) contains cattle QTL and association data curated from published data. The database is designed to facilitate the process for users to compare, confirm, and locate the most plausible location for genes responsible for quantitative traits important to cattle production. We have been striving our best to curate all available data, and adding tools to the QTLdb for users to accomplish many data meta-analysis and comparison tasks.

The current release of the Cattle QTLdb contains **142,261** QTLs/associations from **1,001** publications. Those QTLs / associations represent **646** different traits (see **data summary** for more recent updates). The released data have also been submitted to the NCBI Gene, Ensembl, and UCSC databases, where the QTL information can be retrieved and analyzed using the respective tools on these sites. New tools and functions are continually added. Please see the **release history** or FAQ for new updates.

Information in the **Cattle QTLdb** can be accessed in the following ways:

1. Search and Analysis: Tools for search by chromosomes, traits, breeds, publications, candidate genes, etc. and tools for data analysis primed with search NEW
2. Data Browse:
   - By chromosome locations
   - By cattle trait hierarchies
3. View Maps:
   - Whole genome map linked to each chromosome
   - Alignments with genome features in GBrowse.
   - Alignments with genome features in JBrowse.
4. Downloads:
   - All data by cM
   - All data by bp (on ARS-UCD1.2 in bed format)
   - All data by bp (on ARS-UCD1.2 in gff format)
   - All data by bp (on ARS-UCD1.2 in sam format)
   - All data by bp (on Btau4.6 in bed format)
   - All data by bp (on Btau4.6 in gff format)

QTL on - Gbrowse - Jbrowse

Frequently Asked Questions

Video Tutorials

New Data - Curation - Batch load

Script_pesquisando_QTL.R

Particular

2023-10-04

```
######################################################################
##### Script para pesquisar QTL relatados no QTL database ############
######################################################################

rm(list=ls())
options(stringsAsFactors=F)

# Upload cattleQTLdb
### OBS: e necessario primeiro baixar o arquivo QTLdb_cattleARS_UCD1.bed.txt no site do QTLdb - ver ori
entacoes em Baixando_infoQTL_QTLdb.pdf

# Definindo o diretorio
setwd("D:\\PessoalD\\Doutorado\\Cap 2 ROH\\Tutorial\\2_Busca_genes_QTLs")

# Importando o arquivo QTLdb_cattleARS_UCD1.bed
qtl<-read.table("QTLdb_cattleARS_UCD1.bed",skip=25,header=F,sep="\t",quote="",comment.char="",
                col.names=c('chr','start.pos','end.pos','QTL','V5','V6','V7','V8','V9','V10','V11','V1
2'))
head(qtl)
```

```
##     chr start.pos end.pos                                  QTL V5 V6 V7 V8 V9 V10 V11 V12
## 1 Chr.X        NA      NA          Cold tolerance QTL (31201) NA  + NA NA  .   .   .   .
## 2 Chr.X        NA      NA          Cold tolerance QTL (31202) NA  + NA NA  .   .   .   .
## 3 Chr.X        NA      NA     Body weight (birth) QTL (31612) NA  + NA NA  .   .   .   .
## 4 Chr.X        NA      NA Milk protein percentage QTL (36343) NA  + NA NA  .   .   .   .
## 5 Chr.X        NA      NA      Somatic cell score QTL (64624) NA  + NA NA  .   .   .   .
## 6 Chr.X        NA      NA           Maturity rate QTL (65660) NA  + NA NA  .   .   .   .
```

```
str(qtl)
```

```
## 'data.frame':    176904 obs. of  12 variables:
##  $ chr      : chr  "Chr.X" "Chr.X" "Chr.X" "Chr.X" ...
##  $ start.pos: int  NA NA NA NA NA NA NA NA NA NA ...
##  $ end.pos  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ QTL      : chr  "Cold tolerance QTL (31201)" "Cold tolerance QTL (31202)" "Body weight (birth) QT
L (31612)" "Milk protein percentage QTL (36343)" ...
##  $ V5       : logi  NA NA NA NA NA NA ...
##  $ V6       : chr  "+" "+" "+" "+" ...
##  $ V7       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ V8       : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ V9       : chr  "." "." "." "." ...
##  $ V10      : chr  "." "." "." "." ...
##  $ V11      : chr  "." "." "." "." ...
##  $ V12      : chr  "." "." "." "." ...
```

```
table(qtl$chr)
```

```
##
##  Chr.1 Chr.10 Chr.11 Chr.12 Chr.13 Chr.14 Chr.15 Chr.16 Chr.17 Chr.18 Chr.19  Chr.2 Chr.20 Chr.21 Ch
r.22 Chr.23 Chr.24 Chr.25 Chr.26 Chr.27 Chr.28 Chr.29
##   4174   3410   4678   2125   3636  18958   2199   2405   5992   3185   4246   4650   5485   4179
1712   2015   1254   3597  14545   1775   1167   5657
##  Chr.3 Chr.30  Chr.4  Chr.5  Chr.6  Chr.7  Chr.8  Chr.9  Chr.X
##   4430      8   5456   8098  23469   4094   2155   2364  25786
```

```
# Este é apenas um exemplo com tamanhos de janelas e posições em pares de bases distintas


#### Pesquisando os QTL, uma janela por vez - alterar as coordenadas a cada janela avaliada
## QTL 1
Chr<-3
Chr<-paste0('Chr.',Chr)

list_win_1<-subset(qtl,chr==Chr & ((start.pos>=99643824 & start.pos<=99663824) |
                                   (end.pos>=99643824 &   end.pos<=99663824)),
                   select=c(chr,start.pos,end.pos,QTL))
dim(list_win_1)
```

```
## [1] 0 4
```

```
head(list_win_1)
```

```
## [1] chr       start.pos end.pos   QTL
## <0 linhas> (ou row.names de comprimento 0)
```

```
# Nenhum QTL vai ser encontrado sobreposto a este intervalo

## QTL 2
Chr<-13
Chr<-paste0('Chr.',Chr)

list_win_2<-subset(qtl,chr==Chr & ((start.pos>=63758262 & start.pos<=64910218) |
                                   (end.pos>=63758262 &   end.pos<=64910218)),
                   select=c(chr,start.pos,end.pos,QTL))
dim(list_win_2)
```

```
## [1] 223   4
```

```
head(list_win_2)
```

```
##          chr start.pos  end.pos                                                                  QTL
## 96231 Chr.13  31422984 64193574 Percentage decrease in PCV up to day 150 after challenge QTL (10524)
## 96232 Chr.13  31422984 64193574 Percentage decrease in PCV up to day 100 after challenge QTL (10525)
## 96233 Chr.13  31422984 64193574                                    Parasite detection rate QTL (10526)
## 96519 Chr.13  41141597 63892345                                       Feed conversion ratio QTL (5287)
## 96824 Chr.13  49782125 64193574                                          Somatic cell score QTL (4983)
## 97597 Chr.13  63777329 63777333                                Milk caprylic acid content QTL (163773)
```

```
# Alguns QTLs vao ser encontrados sobrepostos a este intervalo


## QTL 3
Chr<-12
Chr<-paste0('Chr.',Chr)

list_win_3<-subset(qtl,chr==Chr & ((start.pos>=60550912 & start.pos<=60570912) |
                                     (end.pos>=60550912 &   end.pos<=60570912)),
                   select=c(chr,start.pos,end.pos,QTL))
dim(list_win_3)
```

```
## [1] 0 4
```

```
head(list_win_3)
```

```
## [1] chr       start.pos end.pos   QTL
## <0 linhas> (ou row.names de comprimento 0)
```

```
# Nenhum QTL vai ser encontrado sobreposto a este intervalo


## QTL 4
Chr<-1
Chr<-paste0('Chr.',Chr)

list_win_4<-subset(qtl,chr==Chr & ((start.pos>=148476747 & start.pos<=148892436) |
                                     (end.pos>=148476747 &   end.pos<=148892436)),
                   select=c(chr,start.pos,end.pos,QTL))
dim(list_win_4)
```

```
## [1] 13  4
```

```
head(list_win_4)
```

```
##         chr start.pos   end.pos                          QTL
## 29748 Chr.1 148550655 148550659   sperm counts QTL (120268)
## 29749 Chr.1 148550655 148550659    sperm counts QTL (65823)
## 29750 Chr.1 148694672 148694676 Lean meat yield QTL (36679)
## 29751 Chr.1 148757106 148757110 Lean meat yield QTL (36674)
## 29752 Chr.1 148788139 148788143 Lean meat yield QTL (36673)
## 29753 Chr.1 148819573 148819577 Lean meat yield QTL (36678)
```

```
# Alguns QTLs vao ser encontrados sobrepostos a este intervalo


## QTL 5 ... e assim por diante


#### Juntando lista de QTL de todas as janelas
# Algumas janelas nao terao QTLs identificado, exemplo:
## As regioes do 'QTL1' e 'QTL3' tem um intervalo de apenas 20000 pares de bases, muito curto para enco
ntrar QTLs


# No comando abaixo, incluir somente as listas que tem QTLs, neste exemplo, seria o 'QTL 2' e 'QTL 4'
qtl.list<-rbind(list_win_2,list_win_4)
dim(qtl.list)
```

```
## [1] 236   4
```

```
head(qtl.list)
```

```
##          chr start.pos  end.pos                                                                  QTL
## 96231 Chr.13  31422984 64193574 Percentage decrease in PCV up to day 150 after challenge QTL (10524)
## 96232 Chr.13  31422984 64193574 Percentage decrease in PCV up to day 100 after challenge QTL (10525)
## 96233 Chr.13  31422984 64193574                                    Parasite detection rate QTL (10526)
## 96519 Chr.13  41141597 63892345                                       Feed conversion ratio QTL (5287)
## 96824 Chr.13  49782125 64193574                                          Somatic cell score QTL (4983)
## 97597 Chr.13  63777329 63777333                                Milk caprylic acid content QTL (163773)
```

```
### Salvando lista de QTLs encontrados
write.table(qtl.list,file="qtl.list.txt",quote=F,sep="\t",row.names=F,col.names=T)
```

Script_pesquisando_genes.R

Particular

2023-10-04

```
#############################################################
##### Script para pesquisar os genes da base Ensembl #######
#############################################################

rm(list=ls())
options(stringsAsFactors=F)

# instalar pacote biomaRt
# Remover as "#" das 2 linhas seguintes para rodar o pacote
#if (!requireNamespace("BiocManager", quietly = TRUE))
#  install.packages("BiocManager")
BiocManager::install("biomaRt")
```

```
## 'getOption("repos")' replaces Bioconductor standard repositories, see 'help("repositories", package
= "BiocManager")' for details.
## Replacement repositories:
##     CRAN: https://cran.rstudio.com/
```

```
## Bioconductor version 3.16 (BiocManager 1.30.22), R 4.2.2 (2022-10-31 ucrt)
```

```
## Warning: package(s) not installed when version(s) same as or greater than current; use `force = TRUE
` to re-install: 'biomaRt'
```

```
## Installation paths not writeable, unable to update packages
##   path: C:/Program Files/R/R-4.2.2/library
##   packages:
##     boot, class, codetools, foreign, KernSmooth, lattice, MASS, Matrix, mgcv, nlme, nnet, spatial, s
urvival
```

```
# carregar pacote biomaRt
library("biomaRt")

listMarts()                                          # listing databases
```

```
##                biomart                version
## 1 ENSEMBL_MART_ENSEMBL      Ensembl Genes 110
## 2   ENSEMBL_MART_MOUSE      Mouse strains 110
## 3     ENSEMBL_MART_SNP  Ensembl Variation 110
## 4 ENSEMBL_MART_FUNCGEN Ensembl Regulation 110
```

```
ensembl <- useMart("ENSEMBL_MART_ENSEMBL")            # loading genes database
ensembl <- useDataset("btaurus_gene_ensembl", ensembl) # loading bos taurus database (ARS-UCD1.2)
attributes <- c("ensembl_gene_id","chromosome_name","start_position","end_position","external_gene_nam
e")
filter<-"chromosomal_region"

#### Inserir coordenadas selecionadas nas janelas com maior variância
## As janelas podem ser selecionadas de diferentes modos:
## 1 - Top 10 janelas com maior varExp
## 2 - Traçar um limiar e selecionar as janelas acima deste limiar
## 3 - Selecionar janelas que explicam juntas x% da varExp da característica
## 4 - Montar novas janelas (ex: +-200Kb) com base nos SNP com maior varExp
##### A forma como selecionar as janelas deve ser definida antes de se iniciar a escrita do artigo sobr
e genes e QTL encontrado nas análises
#
### Selecionando a posicao inicial e a posicao final da janela de busca

### Inserindo coordenadas para busca de genes
value1_1<-c("4:99643824:99663824")
value1_2<-c("8:98669541:98689541")
value1_3<-c("16:60550912:60570912")
# e assim por diante

### Fazendo a pesquisa dos genes por janela (não alterar os comando, somente incluir mais janelas se pr
ecisar)
gene.list1_1<-getBM(attributes=attributes, filters=filter, values=value1_1, mart=ensembl)
gene.list1_2<-getBM(attributes=attributes, filters=filter, values=value1_2, mart=ensembl)
gene.list1_3<-getBM(attributes=attributes, filters=filter, values=value1_3, mart=ensembl)

dim(gene.list1_2)
```

```
## [1] 3 5
```

```
head(gene.list1_2)
```

```
##      ensembl_gene_id chromosome_name start_position end_position external_gene_name
## 1 ENSBTAG00000038794               8       98619608     98686264            TMEM245
## 2 ENSBTAG00000050794               8       98676196     98676272
## 3 ENSBTAG00000047947               8       98676196     98676495        Metazoa_SRP
```

```
### Juntando lista de genes de todas as janelas
gene.list<-rbind(gene.list1_1,gene.list1_2,gene.list1_3)

### Salvando lista de genes encontrado
write.table(gene.list,file="gene_list.txt",quote=F,sep="\t",row.names=F,col.names=T)
```