RENATA DE FÁTIMA BRETANHA ROCHA

PHENOTYPIC AND GENOMIC ANALYSIS OF THE NUMBER OF OOCYTES AND EMBRYOS OF THE GIR BREED

Thesis submitted to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

Adviser: Simone Eliza Facioni Guimarães

Co-advisers: João Cláudio do Carmo Panetto

Pamela Itajara Otto

VIÇOSA - MINAS GERAIS

2024

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
R672p
2023

Rocha, Renata de Fátima Bretanha, 1994-
Phenotypic and genomic analysis of the number of oocytes and embryos of the Gir breed / Renata de Fátima Bretanha Rocha. – Viçosa, MG, 2023.
1 tese eletrônica (345 f.): il. (algumas color.).

Texto em inglês e português.
Inclui apêndices.
Orientador: Simone Eliza Facioni Guimarães.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Zootecnia, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2024.043
Modo de acesso: World Wide Web.

1. Gir (Bovino). 2. Bovinos de leite - Melhoramento genético. 3. Bovinos de leite - Seleção. 4. Polimorfismo de nucleotídeo único. 5. Marcadores genéticos. 6. Depressão por endogamia. I. Guimarães, Simone Eliza Facioni, 1966-. II. Universidade Federal de Viçosa. Departamento de Zootecnia. Programa de Pós-Graduação em Zootecnia. III. Título.

CDD 22. ed. 636.21420821

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

RENATA DE FÁTIMA BRETANHA ROCHA

PHENOTYPIC AND GENOMIC ANALYSIS OF THE NUMBER OF OOCYTES AND EMBRYOS OF THE GIR BREED

Thesis submitted to the Animal Science Graduate Program of the Universidade Federal de Viçosa, in partial fulfillment of the requirements for the degree of *Doctor Scientiae*.

APPROVED: November 27, 2023.

Assent:

Documento assinado digitalmente
gov.br
RENATA DE FATIMA BRETANHA ROCHA
Data: 09/02/2024 11:57:35-0300
Verifique em https://validar.iti.gov.br

Renata de Fátima Bretanha Rocha

Author

Documento assinado digitalmente
gov.br
SIMONE ELIZA FACIONI GUIMARAES
Data: 06/02/2024 09:25:11-0300
Verifique em https://validar.iti.gov.br

Simone Eliza Facioni Guimarães

Adviser

ACKNOWLEDGENTS

Firstly, I would like to thank God for all the blessings received;

Thanks to my parents (*in memoriam*), José Airton, and especially my mother Maria Terezinha, who has always been very picky and cared to give a good education to her daughters, encouraging and doing everything possible to be present and helpful through this journey. I wish they were here to celebrate this moment;

My sisters Káren and Sarita, my niece Maria Júlia and my nephew Matheus Henrique, who always supported me and were patient even from a distance.

All my friends for all those years of friendship and yet tolerating me;

To the Federal University of Viçosa and Animal Science Department for support and the opportunity;

To the Wageningen University & Research and WUR Animal Breeding and Genomics Group, for allowing me to do part of my doctorate degree at the University, providing all support during my training period;

To National Council for Scientific and Technological Development (Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq) – processes [142600/2019-9] and [200147/2022-6] and Coordination and Improvement of Higher Level Personnel (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES) - Finance Code 001, for financial support with scholarships;

My adviser, Professor Simone Eliza Facioni Guimarães for support, teachings and orientation;

My co-advisers João Panetto e Pamela Otto for the advices, help with the analyses and confidence;

My co-advisers Mario Calus for all help, good discussions, and tips;

To Embrapa Dairy Cattle Research Center, Juiz de Fora – MG, for the partnership and for providing the data needed to carry out this study.

ABSTRACT

ROCHA, Renata de Fátima Bretanha, D.Sc., Universidade Federal de Viçosa, November, 2023. **Phenotypic and genomic analysis of the number of oocytes and embryos of the Gir breed**. Advisor: Simone Eliza Facioni Guimarães. Co-advisors: João Cláudio do Carmo Panetto and Pamela Itajara Otto.

Gir Dairy cattle, a zebu breed (*Bos primigenius indicus*), are predominant in Brazilian herds due to their tolerance to tropical climate conditions and good milk production. Zebu animals are also more efficient in retrieving oocytes and, consequently, in producing embryos in vitro than taurine cattle (*Bos primigenius taurus*). In this context, our objectives were (i): to obtain genetic parameters for the number of total (TO) and viable oocytes (VO) and the number of embryos (EMBR) of the Gir breed; (ii): to verify the profile of runs of homozygosity (ROH) and inbreeding depression over TO, VO, and EMBR of the Gir breed and identify signatures of selection, genes, and enriched regions between Gir subpopulations sorted by breeding value for these traits; (iii): to identify genomic regions, genes, and biological processes associated with TO, VO, and EMBR in Gir dairy cattle; (iv): to verify the inheritance of genomic regions and genes associated with TO, VO, and EMBR based on a daughter design approach; (v): to identify the parental origin of the alleles, through genomic *imprinting*, affecting TO, VO, and EMBR of the Dairy Gir. As results, heritability was moderate for TO and VO, and low for EMBR; repeatability estimates were moderate for all traits; genetic correlations were high; and phenotypic correlations ranged from moderate to high among all traits. All these results suggested that the selection of Gir donors for total oocyte production can be done preferably at younger ages, which might lead to an increase in the number of viable oocytes and embryos obtained and hasten genetic progress. The ROH profile indicates ancient inbreeding in this Gir population. General ROH-based inbreeding had a negative effect on the studied traits, which is important to consider when selecting animals for reproduction in order to avoid inbreeding depression events. From GWAS analyses, we have that ARNT, EGR1, HIF1A, AHR and PAX2 genes are good markers for the production of oocytes and embryos in Gir cattle. Results for a daughter design approach indicated that genomic inheritance is similar between families, especially between TO and VO traits, which is consistent with previous results of high genetic correlation between these traits. Also,

genomic regions located on BTA7 between 82974837 and 83997563 base pairs proved to be the most common across families, where genes with oocyte maturation and embryo development related-functions were identified. Parental origin analysis results showed complete paternal imprinting effects affecting TO. Our results contribute new knowledge about oocyte and embryo production for Brazilian livestock, highlighting the importance of studying reproductive traits in cattle. Our findings contribute to the genetic improvement of the Gir breed and direct future research into gene expression and/or omics that may address the traits studied in this work.

Keywords: Dairy cattle. Family analysis. GWAS. Imprinting. ROH. Signatures of selection

RESUMO

ROCHA, Renata de Fátima Bretanha, D.Sc., Universidade Federal de Viçosa, novembro de 2023. **Análise fenotípica e genômica do número de oócitos e embriões da raça Gir**. Orientadora: Simone Eliza Facioni Guimarães. Coorientadores: João Cláudio do Carmo Panetto, Pamela Itajara Otto.

O gado Gir Leiteiro, raça zebuína (*Bos primigenius indicus*), é predominante nos rebanhos brasileiros por sua tolerância a condições de clima tropical e boa produção de leite. Os animais zebuínos também são mais eficientes na recuperação de oócitos e, consequentemente, na produção de embriões *in vitro* do que os bovinos taurinos (*Bos primigenius taurus*). Nesse contexto, nossos objetivos foram (i): obter os parâmetros genéticos para número de oócitos totais (TO) e viáveis (VO) e número de embriões (EMBR) na raça Gir; (ii): verificar o perfil de corridas do homozigose (ROH) e depressão endogâmica sobre TO, VO e EMBR na raça Gir e identificar assinaturas de seleção, genes e regiões enriquecidas entre subpopulações de Gir classificadas por valor genético para essas características; (iii): identificar regiões genômicas, genes e processos biológicos associados a TO, VO e EMBR em bovinos Gir leiteiro; (iv): verificar a herança de regiões genômicas e genes associados a TO, VO e EMBR com base na abordagem *daughter design*; (v): identificar a origem parental dos alelos, através de *imprintings* genômicos, afetando TO, VO e EMBR na raça Gir. Como resultados, a herdabilidade foi moderada para TO e VO e baixa para EMBR; as estimativas de repetibilidade foram moderadas para todas as características; as correlações genéticas foram altas e as correlações fenotípicas variaram de moderadas a altas entre todas as características. Todos esses resultados sugerem que a seleção de doadoras Gir para produção total de oócitos pode ser feita preferencialmente em idades mais jovens, podendo levar ao aumento do número de oócitos viáveis e embriões obtidos e acelerar o progresso genético. O perfil de ROH indica a existência de endogamia antiga nesta população de Gir. A endogamia baseada em ROH geral tem um efeito negativo nas características estudadas, o que é importante considerar ao selecionar animais para reprodução, a fim de evitar eventos de depressão por endogamia. Pelas análises de GWAS, temos que os genes ARNT, EGR1, HIF1A, AHR e PAX2 são bons marcadores para a produção de oócitos e embriões em bovinos Gir. Os resultados da abordagem *daughter design* indicaram que a herança de marcadores é semelhante entre as

famílias, especialmente entre as características TO e VO, o que é consistente com resultados anteriores de alta correlação genética entre essas características. Além disso, as regiões genômicas localizadas no BTA7 entre 82974837 e 83997563 pares de bases provaram ser as mais comumente herdadas entre as famílias, onde foram identificados genes com funções relacionadas à maturação do oócito e desenvolvimento do embrião. Os resultados da análise da origem parental mostraram efeito completo de *imprinting* paterno afetando TO. Nossos resultados contribuem com novos conhecimentos sobre a produção de oócitos e embriões para a pecuária brasileira, destacando a importância do estudo de características reprodutivas em bovinos. Nossos achados contribuem para o melhoramento genético da raça Gir e direcionam futuras pesquisas sobre expressão gênica e/ou ômicas que possam abordar as características estudadas neste trabalho.

Palavras-chave: Análise de famílias. Assinatura de seleção. Gado leiteiro. GWAS. Imprinting. ROH

SUMMARY

1 GENERAL INTRODUCTION

1.1 The Gir breed

The Gir breed, imported from India to Brazil around 1906 (SANTIAGO, 1986), stood out in milk production in the tropics in such a way that in 1985, the Brazilian Agricultural Research Corporation (EMBRAPA) Dairy Cattle and Brazilian Association of Dairy Gir Breeders created the Dairy Gir National Breeding Program (in Portuguese: Programa Nacional de Melhoramento do Gir Leiteiro - PNMGL) aiming to promote genetic improvement of the Gir breed through the identification and selection of genetically superior animals for milk production, reproduction and conformation traits (PANETTO et al., 2021).

From its introduction in the country, the Gir breed became widely used for milk production and for crosses with European breeds to obtain heterosis, improve productive efficiency, and adapt to different production systems (MCMANUS et al., 2008; REIS et al., 2020). Because of their tolerance to stress conditions, resistance to several tropical diseases (GAUR; KAUSHIK; GARG, 2003), and good milk production (PANETTO et al., 2021) the Gir breed has become prevalent in Brazilian herds. More than a decade ago, Malhado et al. (2010) suggested an increase in the management of the mating process to prevent inbreeding and to maintain the genetic variability of the Gir breed. In addition, Santana Jr et al. (2014) suggested that the Gir population in Brazil is highly structured and can face a breeding program with high selection intensity.

Ever since, several studies have been carried out to explore aspects of the Gir breed such as inbreeding in production and reproduction traits (REIS FILHO et al., 2015), genetic parameters especially for milk production (GONZÁLEZ-HERRERA et al., 2022; HORTOLANI et al., 2022; NASCIMENTO RANGEL et al., 2018; OLIVEIRA et al., 2017; PEREIRA et al., 2019; WENCESLAU et al., 2000), and conformation and management traits (CARVALHO et al., 2021).

In 2001, the PNMGL started collecting biological material (blood and semen) of Gir cattle to build a DNA dataset of this breed, which was a starting point for studies aiming to evaluate molecular traits in this breed (PANETTO et al., 2021). With an increase in genomic data availability, several studies exploring the genomic architecture of the Gir breed began to emerge (BOISON et al., 2017; BRAGA et al., 2023; GARCIA et al., 2023; MAIORANO et

al., 2018; MILANESI et al., 2022; OLIVEIRA et al., 2023). As can be seen from the literature, most of the genomic studies in Gir cattle are recent (less than a decade ago).

It is fair to assume that there are a lot of ongoing projects exploring genomics and different types of traits in the Gir breed, which are still to be published, and each day new data allows for new discoveries. For example, Oliveira et al. (2023) were able to identify differences in genomic estimated breeding values (GEBV) for milk yield (305MY) up to 800 kg, not in live animals, but in embryos obtained on the same farm. In addition to Oliveira et al. (2023), other studies exploring embryo-related traits and their potential to be used in the Gir breeding program, including the ones carried out in the current work (PEREIRA, 2019; ROCHA et al., 2022, 2023; VIZONÁ et al., 2020), were possible due to an increased use of assisted reproductive technologies and production of embryos *in vitro*.

1.2 Assisted reproductive technologies

Reproductive biotechnologies have emerged to replace the conventional techniques used in livestock, in order to face the growing challenges for the productivity, reproduction, and health of animals, in the face of environmental conditions (CHOUDHARY et al., 2016). One of these biotechnologies is the Embryo *In Vitro* Production (IVP), which combines the technologies of oocyte collection, *in vitro* fertilization, and embryo transfer (MOORE; HASLER, 2017). About a decade ago, it was already known that the IVP technology stood out among the reproductive traits due to its economic impact, since it accelerates genetic improvement, reduces the generation interval, and promotes an increase in the number of products/cow/year, as it allows the use of pre-pubescent heifers, cows in early pregnancy, and old cows (MARTINS, 2010). In 2021, a milestone was reached for this technology, since for the first time, over one million IVP cattle embryos were transferred worldwide (VIANA, 2022). In recent years, IVP has made advances in terms of theoretical and practical procedures, which has allowed for an increase in its efficiency (MELLO et al., 2016a). Currently, the two countries with the largest commercial herds of cattle in the world (Brazil and the USA) are those with the highest production of embryos *in vitro*, with most of this production coming from dairy breeds, and Brazil is the leader in IVP embryo transfers worldwide (VIANA, 2022).

However, as it was mentioned, there are many steps until the embryos are obtained, which means that IVP is influenced by several factors (MELLO et al., 2016b). Among these

factors, it can be mentioned donor-related factors, such as category (primiparous or multiparous), age, estrous cycle, nutrition, sanity, breed, physiology, genetic factors, ovarian size, and environment-related factors such as breeder, seasonality, and temperature (MELLO et al., 2016b; OLIVEIRA et al., 2016; PEREIRA et al., 2017; SENEDA et al., 2002).

Carneiro et al. (2019) indicated environmental factors such as the season of the year, for example, with a direct influence on the morphological quality of aspirated oocytes, and in regions with a warmer climate, in the winter period, oocytes have better quality, which will directly influence in successful *in vitro* embryo production. Likewise, (MELLO et al., 2016c), studying Sindhi (*Bos primigenius indicus*) animals, indicated that in the rainy season, it is possible to obtain a higher proportion of viable oocytes, which was also verified in donors younger than six years.

Stroebech et al. (2015) mention that, among several factors that can affect the number of oocytes and embryos obtained, the milk yield can influence the results of Ovum Pick-Up (OPU) and IVP. In agreement, Lopez et al. (2004) found a negative correlation between milk production and estrus duration, as they identified that cows with higher milk production have lower estradiol concentration and larger follicle size.

In general, zebu (*Bos primigenius indicus*) females have more follicular waves and a greater number of follicles per wave than taurine (*Bos primigenius taurus*) females, in addition to a greater amount of small follicles (< 4 mm), which increases the efficiency of oocyte retrieval, allowing greater success for embryo *in vitro* production in comparison to taurine animals (PONTES et al., 2011; ROTAR; SOUZA, 2019). This way, zebu breeds have been highlighted in the use of reproductive biotechnologies, for example, in the production of clones by nuclear somatic cell transfer (DRUM; SARTORI; FRANÇA E MELO, 2019).

In the literature, it is mentioned that the number of oocytes obtained per follicular aspiration session (or OPU session) is a good parameter to assist in the selection of good donors, favoring the process of obtaining embryos *in vitro* (OLIVEIRA et al., 2016; PEREIRA et al., 2017; PEREZ et al., 2016). To understand the genetic structure of these traits in Gir cattle and suggest the best way to apply selection, genetic and genomic studies become necessary.

1.3 Genetic analyses

In order to understand the behavior of a trait in the population and its relationship with other traits, genetic parameters can be estimated, which are known as heritability, correlation, and repeatability. Genetic parameters estimation helps to obtain breeding values and to predict response to selection in each population, allowing the choice of the most appropriate breeding method (GUIMARÃES, 2016).

Heritability is a technical term in genetics, defined as a ratio of variances, which helps to verify whether observed variation in a trait is due to environmental or genetic factors (VISSCHER; HILL; WRAY, 2008). In this work, heritability is calculated as it is defined in genetic studies, that is, the proportion of total variance in a population for a particular trait that is attributable to variation in additive genetic values — termed the narrow-sense heritability (or just heritability, h^2) (FALCONER; MACKAY, 1996; VISSCHER; HILL; WRAY, 2008). Heritability ranges from 0 to 1 being high when it is > 0.4 (VISSCHER; HILL; WRAY, 2008). High heritability is commonly observed for carcass traits (MORAES et al., 2019). Moderate heritability is found in general for production traits like 305-day milk yield, fat, protein, lactose, total solid, day weight, and some reproductive traits such as age at first calving and scrotal perimeter (CARRARA et al., 2022), but usually, reproductive traits present low heritability as well as type traits, such as angularity, fore udder attachment, rear udder height, fore teats placement, teat length, and udder depth (ISMAEL et al., 2021). Heritability remains key to the response to selection and new estimation methods using marker data highlight the relevance of heritability in the genomics era (VISSCHER; HILL; WRAY, 2008).

Correlations can be measured as a phenotypic or genetic parameter (VAN RHEENEN et al., 2019). From both, genetic correlation is the most useful for genetic selection in livestock species (FALCONER; MACKAY, 1996). The main causes of genetic correlation are the pleiotropic effects of genes on numerous traits and/or linkage disequilibrium among several loci (FALCONER; MACKAY, 1996; VAN RHEENEN et al., 2019). Its values range from -1 to 1, indicating traits can be positively or negatively related and according to the studied traits, it can be favorable or detrimental (VAN RHEENEN et al., 2019).

The repeatability is the correlation between repeated phenotypic values. Repeated measures are those traits that can be measured more than once in animals at different points in time or space (MOTA, 2010). Some economically important phenotypes with repeated measures were studied to find their repeatability in cattle populations, such as milk production and milk constituents (COSTA et al., 2019), anogenital distance (RAJESH et al., 2022), feed efficiency (FISCHER; DAI; KALSCHEUR, 2022), and number of oocytes and embryos

(MERTON et al., 2009; PARKER GADDIS et al., 2017). A high repeatability indicates that the first phenotypic measure is a good predictor of the next measure (MOTA, 2010).

Few studies explored those genetic parameters for the traits number of oocytes and number of embryos in cattle, being most of them performed in taurine breeds (JATON et al., 2016, 2017, 2020; PARKER GADDIS et al., 2017; PINHEIRO, 2019). In Zebu cattle, Pinheiro (2019) found genetic variability and high and positive correlation for total oocyte and embryo production even with moderate heritability in Nellore cattle and indicated that focus should be given to environmental effects to increase the efficiency of *in vitro* embryo production.

1.4 Runs of Homozygosity (ROH)

Natural and artificial selection have been used over the years to achieve more desirable phenotypes in livestock species (BRITO et al., 2017). Runs of homozygosity (ROH), as a result of selection processes, are contiguous lengths of homozygous genotypes that are present in an individual due to parents transmitting identical haplotypes to their offspring (BRITO et al., 2017; PURFIELD et al., 2012). As pointed out by Purfield et al. (2012), the recent practices of intense selection of sires, and the use of breeding biotechnologies such as artificial insemination and embryo transfer in some breeds may affect the levels of homozygosity.

Mating of related animals results in the formation of homozygous haplotypes in the progeny genome, which forms large ROHs. However, with subsequent mating in the following generations between unrelated animals, the crossing-over process allows “to break” these homozygous regions. Therefore, the ROH size tends to reduce in the herd in each generation (REBELATO; CAETANO, 2018), since crossing over can. As patterns of ROH may be decomposed, the ancestry of an individual and its population can be evaluated by the extent and frequency of ROH in the following way: the shorter the segment, the more ancient the relatedness of individuals, which can be seen in pedigrees with a lack of information (PURFIELD et al., 2012), and the longer the ROH segment, the more likely the inbreeding is recent in the population (KIRIN et al., 2010). Even in outbred individuals, uncommon long runs of homozygosity may persist due to linkage disequilibrium (LD), recombination rates at certain genomic locations, and rare mutations (GIBSON; MORTON; COLLINS, 2006).

Also, in the lack of pedigree information, even from relatively few samples, the extent of a genome under ROH can be used to infer aspects of recent population history (PERIPOLLI et al., 2018; PURFIELD et al., 2012; SHI et al., 2020). ROH was explored in many species such as chicken (TALEBI et al., 2020), pigs (SHI et al., 2020), small ruminants (BERTOLINI et al., 2018; NOSRATI et al., 2021), and large ruminants (NASCIMENTO et al., 2021; NEVES et al., 2015; ŠIDLOVÁ et al., 2015), for example, to understand the genetic diversity and demography, identify candidate genes and/or to estimate inbreeding.

1.5 Signatures of Selection

Signatures of selection are unique patterns left in the genome of individuals who have undergone natural and/or artificial selection (QANBARI; SIMIANER, 2014) and can be identified by high ROH frequencies in a population (REBELATO; CAETANO, 2018). With the search for selection signatures, we can mainly assess high and unusual allelic frequencies of loci adjacent to favorable mutations and identify genes associated with important traits in livestock, in addition to exploring the genetic diversity of a population (BRITO et al., 2017). Signatures of selection can be identified due to their lower genetic variability and distinctive geographic patterns of linkage disequilibrium (PÉREZ O'BRIEN et al., 2014). The first-time report for signatures of selection derived from whole-genome re-sequencing data in Brazilian locally adapted cattle breeds was carried out by Peripolli et al. (2020) for Gir, Caracu Caldeano, Crioulo Lageano, and Pantaneiro breeds.

The fixation index (Fst), integrated haplotype score (iHS), and cross-population extend haplotype homozygosity (XP-EHH) are approaches used to identify signatures of selection (MAIORANO et al., 2018). In a review of signatures of selection in cattle carried out by Randhawa et al. (2016), Fst seems to be the most used approach by different researchers.

1.6 Genome-Wide Association Studies (GWAS)

Genome-Wide Association Study (GWAS) research is being carried out in the search for genetic variants associated with the phenotypic expression of several traits found in animal populations (BERTOLINI et al., 2018; JATON et al. 2020, ROCHA et al., 2023).

GWAS studies are based on linkage disequilibrium (LD) between molecular markers and causal variants (BOLORMAA et al., 2013), and the higher the density of the marker panels, the higher the LD levels tend to be (RINCON et al., 2011). Thus, the use of GWAS aims to identify markers, known as Single Nucleotide Polymorphisms (SNPs), associated with the trait of interest and to explore regions of the genome where these SNPs are inserted, in order to find genes and /or QTLs related to the phenotypic expression of these traits (UTSUNOMIYA et al., 2013). Furthermore, post-GWAS analyses aim to identify the function of genes in metabolic pathways and interaction with other genes of similar character contribute to a better understanding of the biological action of these genes or identified genomic regions.

1.7 Family analyses

For many years, family analysis has been carried out to map and detect QTL (Quantitative Trait Loci) in dairy cattle populations by applying the "granddaughter design" or "daughter design" (WELLER; KASHI; SOLLER, 1990). Silva et al. (2011) was the first work to detect QTLs associated with traits related to milk production in Gir cattle families. The family analysis can also contribute to the identification of families with higher or better production of oocytes and embryos in dairy Gir cattle.

Ever since these approaches have been used to find quantitative trait loci (**QTL**) and single nucleotide polymorphism (**SNP**) associated with economic traits in several cattle breeds (JIANG et al., 2010; OLIVEIRA JÚNIOR et al., 2019; SILVA et al., 2011). Daughter design is carried out considering genotypic information for sires and their daughters, with phenotypes collected from the daughters, aiming to identify quantitative effects associated with marker alleles as a significant component of variance between daughter subgroups receiving alternative marker alleles from their sires (WELLER et al., 1990). According to Weller et al. (1990), this method is helpful in identifying genes with relatively small quantitative effects, which is the case for polygenic traits.

1.8 Parental origin analysis

With the genomic regions identified in the GWAS and in the analysis of families and, later, their functions and ways of action explored in the post-GWAS analyses, another analysis that can be performed is the search for the origin of genes. The analysis of parental origin allows us to scrutinize the different behavior of genes according to their parental origin (PATERSON; NAIMARK; PETRONIS, 1999). The most understood form of gene expression difference is genomic imprinting, which normally involves the silencing of a gene originating from one parent and the expression of its homolog originating from the other parent (HITCHCOCK; GARDNER, 2019) and is the most well-understood common form of epigenetic regulation of gene expression in mammals (O'DOHERTY et al., 2015). One aspect of this type of analysis is that it can be performed with genotyped parents, even if their phenotype is unknown, as is often the case in livestock (BLUNK et al., 2019).

In recent years, this type of analysis has been extensively explored in the areas of plants (HORNSLIEN; MILLER; GRINI, 2019), humans (PILVAR et al., 2019), cattle (BLUNK et al., 2019), pigs (NEUGEBAUER; LUTHER; REINSCH, 2010), among other species. In Holsteins, Jiang et al. (2017) found significant contributions of dominance and imprinting effects to production and reproduction traits and also indicated that imprinting effects contribute a greater proportion to reproduction traits than production traits. However, work to identify the parental origin of genes affecting reproductive traits in dairy cattle is scarce.

2 OBJECTIVES

2.1 Main objectives

The current work aims to bring new genetic and genomic information on the traits number of total oocytes (TO), number of viable oocytes (VO) and the number of embryos (EMBR) of dairy Gir cattle. In addition, to contribute with new insights to the Dairy Gir National Breeding Program (in Portuguese: Programa Nacional de Melhoramento do Gir Leiteiro - PNMGL).

2.2 Specific objectives

- (i): to obtain genetic parameters for TO, VO, and EMBR of the Gir breed though repeatability and random regression models;
- (ii): to verify the profile of ROH and inbreeding depression in TO, VO, and EMBR in Gir Indicine cattle. Also, to identify signatures of selection, genes, and enriched regions between Gir subpopulations sorted by breeding value for these traits;
- (iii): to identify QTLs, genes, and biological processes associated with TO, VO, and EMBR in Gir dairy cattle;
- (iv): to verify the inheritance of genomic regions and genes associated with TO, VO, and EMBR based on a daughter design approach;
- (v): to identify the parental origin of the alleles, through genomic imprintings, affecting TO, VO, and EMBR of the Dairy Gir breed.

3 REFERENCES

BERTOLINI, F. et al. Genome-wide patterns of homozygosity provide clues about the population history and adaptation of goats. **Genetics Selection Evolution**, v. 50, n. 1, p. 59, 19 dez. 2018. DOI: 10.1186/s12711-018-0424-8

BLUNK, I. et al. Scanning the genomes of parents for imprinted loci acting in their un-genotyped progeny. **Scientific Reports**, v. 9, n. 1, p. 654, 24 dez. 2019. DOI: 10.1038/s41598-018-36939-3

BOISON, S. A. et al. Accuracy of genomic predictions in Gyr (Bos indicus) dairy cattle. **Journal of Dairy Science**, v. 100, n. 7, p. 5479–5490, jul. 2017. DOI: 10.3168/jds.2016-11811

BOLORMAA, S. et al. Detection of quantitative trait loci in Bos indicus and Bos taurus cattle using genome-wide association studies. **Genetics Selection Evolution**, v. 45, n. 1, p. 1–12, 2013. DOI: 10.1186/1297-9686-45-43

BRAGA, L. G. et al. Identification of copy number variations in the genome of Dairy Gir cattle. **PLOS ONE**, v. 18, n. 4, p. e0284085, 10 abr. 2023. DOI: 10.1371/journal.pone.0284085

BRITO, L. F. et al. Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. **BMC Genomics**, v. 18, n. 1, 2017. DOI: 10.1186/s12864-017-3610-0

CARNEIRO, I. de M. B. et al. Oócitos bovinos: influência das estações do ano e maturação in vitro em meio enriquecido com quercetina. **Magistra**, v. 30, p. 134–142, 2019.

CARRARA, E. R. et al. Genetic study of quantitative traits supports the use of Guzerá as dual-purpose cattle. **Animal Bioscience**, v. 35, n. 7, p. 955–963, 1 jul. 2022. DOI: 10.5713/ab.21.0458

CARVALHO, N. S. et al. Genetic parameters and genetic trends of conformation and management traits in Dairy Gir cattle. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 73, n. 4, p. 938–948, ago. 2021. DOI: 10.1590/1678-4162-12341

CHOUDHARY, K. K. et al. Advances in reproductive biotechnologies. **Veterinary World**, v. 9, n. 4, p. 388–395, abr. 2016. DOI: 10.14202/vetworld.2016.388-395

COSTA, A. et al. Heritability and repeatability of milk lactose and its relationships with traditional milk traits, somatic cell score and freezing point in Holstein cows. **Animal**, v. 13, n. 5, p. 909–916, 2019. DOI: 10.1017/S1751731118002094

DRUM, J. N.; SARTORI, R.; FRANÇA E MELO, L. DE. New perspectives in assisted reproductive technologies in cattle. **Revista Ecuatoriana de Ciencia Animal**, v. 3, n. 2, p. 85–106, 14 out. 2019.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics**. London, UK: Longman Group Ltd., 1996.

FISCHER, A.; DAI, X.; KALSCHEUR, K. F. Feed efficiency of lactating Holstein cows is repeatable within diet but less reproducible when changing dietary starch and forage concentrations. **animal**, v. 16, n. 8, p. 100599, ago. 2022. DOI: 10.1016/j.animal.2022.100599

GARCIA, A. O. et al. Pedigree reconstruction and population structure using SNP markers in Gir cattle. **Journal of Applied Genetics**, v. 64, n. 2, p. 329–340, 16 maio 2023. DOI: 10.1007/s13353-023-00747-x

GAUR, G. K.; KAUSHIK, S. N.; GARG, R. C. The Gir cattle breed of India - characteristics and present status. **Animal Genetic Resources Information**, v. 33, p. 21–29, 1 abr. 2003. DOI: 10.1017/S1014233900001607

GIBSON, J.; MORTON, N. E.; COLLINS, A. Extended tracts of homozygosity in outbred human populations. **Human Molecular Genetics**, v. 15, n. 5, p. 789–795, 2006. DOI: 10.1093/hmg/ddi493

GONZÁLEZ-HERRERA, L. G. et al. Genetic evaluation of lactation persistency in the Gyr breed by using a two-trait random regression model. **Animal Production Science**, v. 62, n. 3, p. 216, 2022. DOI: 10.1071/AN20267

GUIMARÃES, N. C. **Parâmetros genéticos de caracteres quantitativos relacionados à produtividade de rebanhos selecionados da raça Nelore**. 2016. Dissertação (Mestrado em Zootecnia) – Escola de Veterinária e Zootecnia da Universidade, Universidade Federal de Goiás, Goiânia, GO, 2016.

HITCHCOCK, T. J.; GARDNER, A. Parent-of-origin specific gene expression and dispersal. **Current Opinion in Behavioral Sciences**, v. 25, p. 36–43, fev. 2019. DOI: 10.1016/j.cobeha.2018.06.007

HORNSLIEN, K. S.; MILLER, J. R.; GRINI, P. E. Regulation of Parent-of-Origin Allelic Expression in the Endosperm. **Plant Physiology**, v. 180, n. 3, p. 1498–1519, jul. 2019. DOI: 10.1104/pp.19.00320

HORTOLANI, B. et al. Genetic parameters for body weight and milk production of dairy Gyr herds. **Tropical Animal Health and Production**, v. 54, n. 1, p. 84, 28 fev. 2022. DOI: 10.1007/s11250-022-03088-9

ISMAEL, H. et al. Estimation of heritability and genetic correlations between milk yield and linear type traits in primiparous Holstein-Friesian cows. **Revista Brasileira de Zootecnia**, v. 50, 14 abr. 2021. DOI: 10.37496/rbz5020200121

JATON, C. et al. Genetic analysis of superovulatory response of Holstein cows in Canada. **Journal of Dairy Science**, v. 99, n. 5, p. 3612–3623, maio 2016. DOI: 10.3168/jds.2015-10349

JATON, C. et al. Genetic analysis for quality of frozen embryos produced by Holstein cattle donors in Canada. **Journal of Dairy Science**, v. 100, n. 9, p. 7320–7329, set. 2017. DOI: 10.3168/jds.2017-12851

JATON, C. et al. Genetic and genomic analyses of embryo production in dairy cattle. **Reproduction, Fertility and Development**, v. 32, n. 2, p. 50, 2020. DOI: 10.1071/RD19275

JIANG, J. et al. Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. **BMC Genomics**, v. 18, n. 1, p. 425, 30 dez. 2017. DOI: 10.1186/s12864-017-3821-4

JIANG, L. et al. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. **PLoS ONE**, v. 5, n. 10, p. e13661, 27 out. 2010. DOI: 10.1371/journal.pone.0013661

KIRIN, M. et al. Genomic Runs of Homozygosity Record Population History and Consanguinity. **PLoS ONE**, v. 5, n. 11, p. e13996, 15 nov. 2010. DOI: 10.1371/journal.pone.0013996

LOPEZ, H.; SATTER, L. D.; WILTBANK, M. C. Relationship between level of milk production and estrous behavior of lactating dairy cows. **Animal Reproduction Science**, v. 81, n. 3–4, p. 209–223, abr. 2004. DOI: 10.1016/j.anireprosci.2003.10.009

MAIORANO, A. M. et al. Assessing genetic architecture and signatures of selection of dual purpose Gir cattle populations using genomic information. **PLOS ONE**, v. 13, n. 8, p. e0200694, 2 ago. 2018. DOI: 10.1371/journal.pone.0200694

MALHADO, C. H. M. et al. History of registered Gyr breed in Brazilian Northeast: population

structure and genetic improvement of growth traits. **Ciência Rural**, v. 40, n. 6, p. 1385–1391, jun. 2010. DOI: 10.1590/S0103-84782010000600023

MARTINS, C. F. **O impacto da transferência de embriões e da fecundação in vitro na produção de bovinos no BrasilEmbrapa Cerrados**, 2010. Embrapa Cerrados: INFOTECA-E. Available in: https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/875061. Acessed in: 04 fev. 2020.

MCMANUS, C. et al. Características produtivas e reprodutivas de vacas Holandesas e mestiças Holandês × Gir no Planalto Central. **Revista Brasileira de Zootecnia**, v. 37, n. 5, p. 819–823, maio 2008. DOI: 10.1590/S1516-35982008000500006

MELLO, R. R. C. et al. Produção in vitro (PIV) de embriões em bovinos. **Revista Brasileira de Reprodução Animal**, v. 40, n. 2, 2016a.

MELLO, R. R. C. et al. Fatores ligados à doadora que influenciam na produção de embriões in vitro (PIVE). **Revista Brasileira de Reprodução Animal**, v. 40, n. 2, p. 51–57, 2016b.

MELLO, R. R. C. et al. Parâmetros da produção in vitro de embriões da raça Sindi. **Pesquisa Agropecuária Brasileira**, v. 51, n. 10, p. 1773–1779, out. 2016c. DOI: 10.1590/s0100-204x2016001000009

MERTON, J. S. et al. Genetic parameters for oocyte number and embryo production within a bovine ovum pick-up–in vitro production embryo-production program. **Theriogenology**, v. 72, n. 7, p. 885–893, out. 2009. DOI: 10.1016/j.theriogenology.2009.06.003

MILANESI, M. et al. Genomic breed composition of pure registered Brazilian Gir. **Tropical Animal Health and Production**, v. 54, n. 5, 2022. DOI: 10.1007/s11250-022-03299-0

MOORE, S. G.; HASLER, J. F. A 100-Year Review: Reproductive technologies in dairy science. **Journal of Dairy Science**, v. 100, n. 12, p. 10314–10331, dez. 2017. DOI: 10.3168/jds.2017-13138

MORAES, G. F. et al. Selection for feed efficiency does not change the selection for growth and

carcass traits in Nellore cattle. **Journal of Animal Breeding and Genetics**, v. 136, n. 6, p. 464–473, 22 nov. 2019. DOI: 10.1111/jbg.12423

MOTA, M. D. S. Parâmetros genéticos – Repetibilidade. **PUBVET Publicações em Medicina Veterinária e Zootecnia. Londrina, PR.**, v. 4, n. 17, 2010.

NASCIMENTO, A. V. et al. Inbreeding coefficients and runs of homozygosity islands in Brazilian water buffalo. **Journal of Dairy Science**, v. 104, n. 2, p. 1917–1927, fev. 2021. DOI: 10.3168/jds.2020-18397

NASCIMENTO RANGEL, A. H. do et al. Estimativa de tendências genéticas e fenotípicas para características leiteiras de animais das raças gir e guzerá. **Acta Scientiarum - Animal Sciences**, v. 40, 2018. DOI: 10.4025/actascianimsci.v40i1.36685

NEUGEBAUER, N.; LUTHER, H.; REINSCH, N. Parent-of-origin effects cause genetic variation in pig performance traits. **Animal**, v. 4, n. 5, p. 672–681, 2010. DOI: 10.1017/S1751731109991625

NEVES, H. H. R. et al. Preliminary study to determine extent of linkage disequilibrium and estimates of autozygosity in Brazilian Gyr dairy cattle. **Archivos de Zootecnia**, v. 64, n. 246, p. 99–108, 10 jun. 2015. DOI: 10.21071/az.v64i246.383

NOSRATI, M. et al. The pattern of runs of homozygosity and genomic inbreeding in world-wide sheep populations. **Genomics**, v. 113, n. 3, p. 1407–1415, maio 2021. DOI: 10.1016/j.ygeno.2021.03.005

O'DOHERTY, A. M. et al. Genomic imprinting effects on complex traits in domesticated animal species. **Frontiers in Genetics**, v. 6, n. APR, 24 abr. 2015. DOI: 10.3389/fgene.2015.00156

OLIVEIRA, C. S. et al. Imputation accuracy for genomic selection using embryo biopsy samples in Gir. **Reproductive Biology**, v. 23, n. 2, p. 100765, jun. 2023. DOI: 10.1016/j.repbio.2023.100765

OLIVEIRA, H. R. de et al. Bayesian Models combining Legendre and B-spline polynomials for

genetic analysis of multiple lactations in Gyr cattle. **Livestock Science**, v. 201, p. 78–84, jul. 2017. DOI: 10.1016/j.livsci.2017.05.007

OLIVEIRA, L. M. de P. de et al. Recuperação de Oócitos em Bovinos da Raça Gir e Girolando. **Revista Eletrônica Biociências, Biotecnologia e Saúde**. Curitiba - PR, v. 6, n. 15, mai/ago 2016. Available in: https://interin.utp.br/index.php/GR1/article/view/1695

OLIVEIRA JÚNIOR, G. A. et al. Fine mapping of genomic regions associated with female fertility in Nellore beef cattle based on sequence variants from segregating sires. **Journal of Animal Science and Biotechnology**, v. 10, n. 1, p. 97, 16 dez. 2019. DOI: 10.1186/s40104-019-0403-0

PANETTO, J. C. do C. et al. **Programa Nacional de Melhoramento do Gir Leiteiro - Sumário Brasileiro de Touros - 4ª Avaliação Genômica de Touros - Resultado do Teste de Progênie.** Embrapa Gado de Leite. 2021. Available in: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/1132113/programa-nacional-de-melhoramento-do-gir-leiteiro---sumario-brasileiro-de-touros-4-avaliacao-genomica-de-touros---resultado-do-teste-de-progenie---maio-2021

PARKER GADDIS, K. L. et al. Evaluation of genetic components in traits related to superovulation, in vitro fertilization, and embryo transfer in Holstein cattle. **Journal of Dairy Science**, v. 100, n. 4, p. 2877–2891, abr. 2017. DOI: 10.3168/jds.2016-11907

PATERSON, A. D.; NAIMARK, D. M. J.; PETRONIS, A. The Analysis of Parental Origin of Alleles May Detect Susceptibility Loci for Complex Disorders. **Human Heredity**, v. 49, n. 4, p. 197–204, 1999. DOI: 10.1159/000022875

PEREIRA, J. F. S. et al. Recuperação de oócitos em bovinos da raça Guzerá e Holandesa. **Revista Eletrônica Biociências, Biotecnologia e Saúde.** Curitiba - PR, v. 10, n. 19, set/dez 2017. Available in: https://interin.utp.br/index.php/GR1/article/view/1549

PEREIRA, M. A. **Análise genética de características de produção in vitro de embriões de doadoras da raça Gir Leiteiro**. PhD thesis, Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agr´arias e Veterinárias, Jaboticabal, 2019. Available at

https://repositorio.unesp.br/ handle/11449/191802

PEREIRA, R. J. et al. Test-day or 305-day milk yield for genetic evaluation of Gir cattle. **Pesquisa Agropecuária Brasileira**, v. 54, 2019. DOI: 10.1590/s1678-3921.pab2019.v54.00325

PEREZ, B. C. et al. Genetic analysis of oocyte and embryo production traits in Guzerá breed donors and their associations with age at first calving. **Genetics and Molecular Research**, v. 15, n. 2, p. 1–9, 2016. DOI: 10.4238/gmr.15027583

PÉREZ O'BRIEN, A. M. et al. Assessing signatures of selection through variation in linkage disequilibrium between taurine and indicine cattle. **Genetics Selection Evolution**, v. 46, n. 1, p. 19, 4 dez. 2014. DOI: 10.1186/1297-9686-46-19

PERIPOLLI, E. et al. Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. **BMC Genomics**, v. 19, n. 1, p. 1–13, 2018. DOI: 10.1186/s12864-017-4365-3

PERIPOLLI, E. et al. Genome-wide detection of signatures of selection in indicine and Brazilian locally adapted taurine cattle breeds using whole-genome re-sequencing data. **BMC Genomics**, v. 21, n. 1, p. 624, 11 dez. 2020. DOI: 10.1186/s12864-020-07035-6

PILVAR, D. et al. Parent-of-origin-specific allelic expression in the human placenta is limited to established imprinted loci and it is stably maintained across pregnancy. **Clinical Epigenetics**, v. 11, n. 1, p. 94, 26 dez. 2019. DOI: 10.1186/s13148-019-0692-3

PINHEIRO, A. K. **Parâmetros produtivos e genéticos da produção in vitro de embriões em bovinos Nelore no Acre**. Masters dissertation, Universidade Federal do Acre, Programa de P´os-Graduação em Sanidade e Produção Animal Sustent´avel na Amazônia Ocidental, Rio Branco. Available at http://www.alice.cnptia.embrapa.br/alice/handle/doc/1119026

PONTES, J. H. F. et al. Ovum pick up, in vitro embryo production, and pregnancy rates from a large-scale commercial program using Nelore cattle (Bos indicus) donors. **Theriogenology**, v. 75, n. 9, p. 1640–1646, jun. 2011. DOI: 10.1016/j.theriogenology.2010.12.026

PURFIELD, D. C. et al. Runs of homozygosity and population history in cattle. **BMC Genetics**, v. 13, n. 1, p. 70, 2012. DOI: 10.1186/1471-2156-13-70

QANBARI, S.; SIMIANER, H. Mapping signatures of positive selection in the genome of livestock. **Livestock Science**, v. 166, n. 1, 2014. DOI: 10.1016/j.livsci.2014.05.003

RAJESH, I. et al. Repeatability of anogenital distance measurements from birth to maturity and at different physiological states in female Holstein cattle. **Journal of Dairy Science**, v. 105, n. 3, p. 2699–2707, mar. 2022. DOI: 10.3168/jds.2021-21419

RANDHAWA, I. A. S. et al. A Meta-Assembly of Selection Signatures in Cattle. **PLOS ONE**, v. 11, n. 4, p. e0153013, 5 abr. 2016. DOI: 10.1371/journal.pone.0153013

REBELATO, A. B.; CAETANO, A. R. Runs of homozygosity for autozygosity estimation and genomic analysis in production animals. **Pesquisa Agropecuária Brasileira**, v. 53, n. 9, p. 975–984, set. 2018. DOI: 10.1590/s0100-204x2018000900001

REIS, A. F. et al. Caracterização fenotípica de vacas F1 Holandês x Zebu de diferentes bases maternas. **Arquivo Brasileiro de Medicina Veterinária e Zootecnia**, v. 72, n. 4, p. 1458–1468, ago. 2020. DOI: 10.1590/1678-4162-11759

REIS FILHO, J. C. et al. Inbreeding on productive and reproductive traits of dairy Gyr cattle. **Revista Brasileira de Zootecnia**, v. 44, n. 5, p. 174–179, maio 2015. DOI: 10.1590/S1806-92902015000500002

RINCON, G. et al. Hot topic: performance of bovine high-density genotyping platforms in Holsteins and Jerseys. **Journal of Dairy Science**, v. 94, n. 12, p. 6116–6121, 2011. DOI: 10.3168/jds.2011-4764

ROCHA, R. de F. B. et al. Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed. **Animal Production Science**, v. 62, n. 17, p. 1661–1670, 19 jul. 2022. DOI: 10.1071/AN21588

ROCHA, R. de F. B. et al. Runs of homozygosity and signatures of selection for number of oocytes

and embryos in the Gir Indicine cattle. **Mammalian Genome**, v. 34, n. 3, p. 482-496., 31 mar. 2023. DOI: 10.1007/s00335-023-09989-w

ROCHA, R. de F. B. et al. Single-step genome-wide association studies and post-GWAS analyses for the number of oocytes and embryos in Gir cattle. **Mammalian Genome**, v.34, n. 3, p.497-508. DOI: 10.1007/s00335-023-10009-0

ROTAR, L. N.; SOUZA, J. F. Morphological Characteristics for Oocyte-Cumulus Complexes of Bos taurus and Bos indicus Cattle Used for a Variety of Purposes. **Russian Agricultural Sciences**, v. 45, n. 4, p. 399–402, 2 jul. 2019. DOI: 10.3103/S1068367419040141

SANTANA JR, M. L. et al. History, structure, and genetic diversity of Brazilian Gir cattle. **Livestock Science**, v. 163, n. 1, p. 26–33, maio 2014. DOI: 10.1016/j.livsci.2014.02.007

SANTIAGO, A. A. **O Zebu na Índia, no Brasil e no mundo**. Campinas: Instituto Campineiro de Ensino Agrícola, 1986. 745p.

SENEDA, M. M. et al. Aspectos técnicos e biológicos da obtenção de oócitos bovinos: revisão de literatura. **Semina: Ciências Agrárias**, v. 23, n. 1, p. 101-110, 2002.

SHI, L. et al. Estimation of inbreeding and identification of regions under heavy selection based on runs of homozygosity in a Large White pig population. **Journal of Animal Science and Biotechnology**, v. 11, n. 1, p. 46, 28 dez. 2020. DOI: 10.1186/s40104-020-00447-0

ŠIDLOVÁ, V. et al. Genomic variability among cattle populations based on runs of homozygosity. **Poljoprivreda**, v. 21, n. 1 sup, p. 44–47, 18 set. 2015. DOI: 10.18047/poljo.21.1.sup.9

SILVA, A. A. et al. Quantitative trait loci affecting milk production traits on bovine chromosome 6 in zebuine Gyr breed. **Journal of Dairy Science**, v. 94, n. 2, p. 971–980, fev. 2011. DOI: 10.3168/jds.2009-2970

STROEBECH, L. et al. In vitro production of bovine embryos: revisiting oocyte development and application of systems biology. **Anim. Reprod**, v. 12, n. 3, p. 465–472, 2015.

TALEBI, R. et al. Runs of Homozygosity in Modern Chicken Revealed by Sequence Data. **G3 Genes|Genomes|Genetics**, v. 10, n. 12, p. 4615–4623, 1 dez. 2020. DOI: 10.1534/g3.120.401860

UTSUNOMIYA, Y. T. et al. Genome-wide association study for birth weight in Nellore cattle points to previously described orthologous genes affecting human and bovine height. **BMC Genetics**, v. 14, 2013. DOI: 10.1186/1471-2156-14-52

VAN RHEENEN, W. et al. Genetic correlations of polygenic disease traits: from theory to practice. **Nature Reviews Genetics**, v. 20, n. 10, p. 567–581, 6 out. 2019. DOI: 10.1038/s41576-019-0137-z

VIANA, J. 2021 Statistics of embryo production and transfer in domestic farm animals. A new milestone has been reached: Transfers of IVP embryos were over one million worldwide. **Embryo Technology Newsletter**, v. 40, n. 4, p. 22–40, 2022.

VISSCHER, P. M.; HILL, W. G.; WRAY, N. R. Heritability in the genomics era — concepts and misconceptions. **Nature Reviews Genetics**, v. 9, n. 4, p. 255–266, 4 abr. 2008. DOI: 10.1038/nrg2322

VIZONÁ, R. G. et al. Genetic analysis of in-vitro embryo production traits in Dairy Gir cattle. **Theriogenology**, v. 148, p. 149–161, maio 2020. DOI: 10.1016/j.theriogenology.2020.02.014

WELLER, J. I.; KASHI, Y.; SOLLER, M. Power of Daughter and Granddaughter Designs for Determining Linkage Between Marker Loci and Quantitative Trait Loci in Dairy Cattle. **Journal of Dairy Science**, v. 73, n. 9, p. 2525–2537, set. 1990. DOI: 10.3168/jds.S0022-0302(90)78938-2

WENCESLAU, A. A. et al. Estimação de parâmetros genéticos de medidas de conformação, produção de leite e idade ao primeiro parto em vacas da raça Gir Leiteiro. **Revista Brasileira de Zootecnia**, v. 29, n. 1, p. 153–158, fev. 2000. DOI: 10.1590/S1516-35982000000100021

4 ARTICLES

Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed*

*Published at Animal Production Science (https://doi.org/10.1071/AN21588)

Renata de Fátima Bretanha Rocha[A], Pamela Itajara Otto[B], Marcos Vinicius Gualberto Barbosa da Silva[C], Marta Fonseca Martins[C], Marco Antonio Machado[C], Renata Veroneze[A], Felipe Damasceno Leandro[D], Stela Naetzold Pereira[B], Simone Eliza Facioni Guimarães[A], João Cláudio do Carmo Panetto[C, E]

[A]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, MG, 36570-900, Brazil, e-mail: renata.bretanha@ufv.br, renata.veroneze@ufv.br, sfacioni@ufv.br

[B]Department of Animal Science, Universidade Federal de Santa Maria, Santa Maria, RS, 97105-900, Brazil, e-mail: pamela.otto@ufsm.br, snaetzold@gmail.com

[C]EMBRAPA - Dairy Cattle National Research Center, Juiz de Fora, MG, 36038-330, Brazil, e-mail: joao.panetto@embrapa.br, marcos.vb.silva@embrapa.br, marta.martins@embrapa.br, marco.machado@embrapa.br

[D]Universidade José do Rosário Vellano / UNIFENAS, Alfenas, MG, 37132-440, Brazil, e-mail: felipe.leandro@aluno.unifenas.br

[E]Corresponding author: joao.panetto@embrapa.br

Abstract

Context. The greater production of oocytes and embryos from Gir donors contributes to greater fertility and increasing genetic progress. ***Aims.*** Obtaining genetic parameters for oocyte and embryo production in the Gir breed. ***Methods.*** Repeatability and random regression models were applied to data consisting of 17,526 Ovum Pick Up (OPU) information from 1,641 Gir donors from five different herds. Single and multi-trait analyses were carried out with the application of both models for the traits: number of viable oocytes (VO), number of total oocytes (TO) and number of embryos (EMBR) using the BLUPF90 family programs. Legendre polynomials of second order were used in the random regression model. ***Key results.*** Considering the repeatability model, the additive genetic variance ranged from 0.06 to 0.13 and the permanent environment variance ranged from 0.05 to 0.08 for all evaluated traits. The residual variance ranged from 0.30 to 0.45. The heritability estimates were 0.10 for EMBR, 0.24 for TO, and 0.25 for VO. Repeatability estimates were moderate ranging from 0.20 to 0.40, and genetic correlation estimates were always above 0.80. Phenotypic correlation was high only between VO and TO (0.95), and moderate in the other cases. Random regression model results were consistent with those from the repeatability model. The heritability values remained similar throughout the donors' ages, with moderate values for VO and TO, and low values for EMBR. Genetic correlations among ages for each trait were moderate to high. Also, the genetic correlations between traits within each age were high, with values always above 0.7. ***Conclusions.*** Selection of Gir donors for total oocyte production at any time, between 1 and 16 years of age, might lead to an increase in the number of viable oocytes and embryos obtained, but it's preferable at younger ages to improve genetic progress. Non-genetic variation plays an important role for all traits, but they are more relevant for number of embryos. ***Implications.*** The use of repeatability model to estimate genetic parameters of oocytes and embryos resulted in similar results compared to random regression model.

Keywords: dairy cattle; genetic correlation; heritability; ovum pick up; phenotypic correlation; repeatability; reproduction; zebu cattle.

Introduction

The Gir cattle (*Bos primigenius indicus)* is resistant to heat and parasites, besides, it is one of the most efficient indicine breeds in milk production in tropical climate conditions like Brazil (Santana Jr. et al. 2014). Moreover, this breed also has good reproductive capacity, making it widely used in crosses with *Bos taurus* breeds, putting together productive and reproductive performance for dairy production (McManus et al. 2008; Wenceslau et al. 2000). Although heat stress can affect estrus detection and cause the death of embryos in the early stages of pregnancy (Berman 2011), *Bos indicus* breeds have double the oocyte production compared to *Bos taurus* cattle, proving their potential to be used for *in vitro* embryo production (Rotar and Souza 2019). Being a *Bos primigenius indicus*, the Gir breed has been highlighted in the use of reproductive biotechnologies, usually obtaining higher oocyte quality when compared to *Bos primigenius taurus*, which have been allowing better in vitro embryo production performances than taurine animals (Drum et al. 2019).

The techniques used in embryo *in vitro* production evolved in recent years presenting efficient results for producers, contributing to increased herd sizes and genetic gains (Becher et al. 2018; Perez et al. 2019; Rotar and Souza 2019; Stroebech et al. 2015). According to Viana (2019), European and American countries contribute significantly with embryo industry data and, among several countries working with these technologies, *in vitro* production is more frequent in North and South America. In the United States, for example, the evaluations of oocyte and embryo quality, as well as the optimization of gametes and the effective use of sexed semen have contributed to genetic breeding strategies in dairy cattle (Sanches et al. 2019). In Brazil, although the use of these technologies has firstly occurred in beef breeds, the improvement in embryo *in vitro* production (PIV) efficiency and the availability of sexed semen allowed the expansion of embryo production also for dairy breeds (Gonçalves and Viana 2019). Due to the great advance of technologies for evaluating these traits over the years, it is important to explore genetic parameters to better understand the action of additive genetic and environmental effects on the variation of these traits. Genetic parameters for heat-stress resistance traits, milk yield, and reproductive traits in cattle raised in tropical regions have been studied for a long time (Davis 1993) until recent days (Otto et al. 2019; 2020), to obtain more accurate estimates of responses to selection. Genetic parameters are essential information to define the design of animal breeding programs, as they allow us to know the proportion of phenotypic variance attributable to genetic variance of quantitative traits. Genetic variability in traits related to milk production, age at first calving and conformation

has been reported in Gir cattle, showing their potential for selection programs (Wenceslau et al. 2000). The same can be explored for other traits to indicate better breeding strategies for this breed.

Regarding oocyte and embryo production in cattle, most part of the genetic parameters studies have been carried out in the Holstein breed (Asada and Terawaki 2002; Eriksson et al. 2007; Jaton et al. 2016; König et al. 2007; Merton et al. 2009; Parker Gaddis et al. 2017; Tonhati et al. 1999), and few studies for other breeds such as Swedish Red cattle (Eriksson et al. 2007), Guzerat (Perez et al. 2016), Nelore (Pinheiro 2019) and Gir (Pereira 2019; Vizoná et al. 2020). Random regression analysis for the number of oocytes and embryos were not found in literature. Considering the lack of studies for these traits in tropical dairy cattle, more research is needed to understand the genetic components and to support strategies for breeding programs. This study aimed to estimate genetic parameters for oocyte and embryo production in the Gir breed, using repeatability and random regression models.

Material and methods

Phenotypic data

After quality control, the final data set contained 17,526 follicular aspirations from 1,641 Gir donors, daughters of 127 sires and 771 dams. The pedigree file, including 4,679 animals, was used to determine the relationships covering up to fifteen generations. The number of Ovum Pick Up (OPU) sessions ranged from one to 70 per donor, with an average of 10. A total of 238 bulls were used for the *in vitro* fertilization processes. The frequency of use of these bulls ranged from one to 1,789 samples, with an average of 73 samples per bull.

The dataset used in this work contained information of OPU and embryo *in vitro* production (PIV) from five different farms, with animals originated from many different lineages of the breed, constituting a good sample of the Brazilian Dairy Gir population. The following traits were included: number of viable oocytes (VO), number of total oocytes (TO) and number of embryos (EMBR). The EMBR trait refers to the total number of embryos recovered, despite of quality or stage of these embryos. Phenotype quality control of the data was performed, with the exclusion of records within the following conditions: donors without correct identification, pool of donor samples, other breeds than Gir, duplicate data and samples in which the number of embryos was greater than the number of total oocytes. Samples out of the age range between 1 and 16 years were excluded, because they were represented by a very few numbers of observations. The OPU sessions in all analyzed herds were conducted between January 2005 and June 2020 in the Minas Gerais State of Brazil. Eight companies

responsible for OPU and *in vitro* fertilization (IVF) processes were listed in the dataset received from farms, each company using their own protocols; in this way, we considered eight protocols in total for the analysis, varying among farms or sometimes within farms. OPU seasons were classified as: (1) January to March; (2) April to June; (3) July to September; (4) October to December. Animals were separated into 138 contemporary groups (CG) composed according to year, season, farm and OPU-IVF protocol, considering a minimum of three animals in each group; lower numbers implicated in data exclusion. Conventional or sexed semen straws from Gir (n = 35) or Holstein (n = 92) bulls were used individually or combined in pools with a maximum of two bulls. Also, a minimum of three observations per cow was considered. With analysis of variance, we identified the significant effect ($P<0.05$) of CG and age (linear and quadratic) for all traits, aspiration interval only for total oocytes and viable oocytes and, for embryos, the effects of bull and bull breed were significant. The type of semen (conventional or sexed) was not significant ($P = 0.45$), so it was not included in the model. Several models were tested and compared (Supplementary file S1) to choose the best model according to the significance statistics results for the effects.

Repeatability model

Estimates of variance components, heritability and repeatability were obtained with the use of single-trait analyses. Covariances and genetic correlations were obtained with the use of multi-trait analyses for VO, TO, and EMBR. Phenotypic correlations were obtained through the Pearson method with the R software (R Core Team 2020). Heritability, repeatability, phenotypic and genetic correlations were calculated according to Falconer and Mackay (1996). All estimates were obtained by the restricted maximum likelihood (REML) method using BLUPF90 family programs (Misztal et al. 2002), namely RENUMF90 and REMLF90 (version 1.82), with animal models. To obtain a normal distribution of the residuals, the traits were transformed using the logarithmic scale: ln (X + 1), where ln is the natural logarithm and X is the raw data (oocyte and embryos). The normal distribution was verified by the Anderson-Darling test (Stephens 1986; Thode Jr. 2002). The model proposed to evaluate the traits was:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Wp} + \boldsymbol{e}$$

where, $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{a}$, $\boldsymbol{p}$, and $\boldsymbol{e}$ are the vectors of observations; fixed effects; additive genetic random effects; permanent environment random effects and residual effects, respectively; $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are the incidence matrices of fixed and random effects, respectively. For VO and TO, only CG was considered a fixed effect. For EMBR, the fixed effects of bull and bull breed were

additionally considered. Donor's age in days (linear and quadratic components) was included as covariates for all traits. Aspiration interval (linear component) was included for VO and TO. The effects of animal, permanent environment and residual were assumed to be random. In the single-trait model, it was assumed that $\boldsymbol{a} \sim N(0, \boldsymbol{A}\sigma_a^2)$, $\boldsymbol{p} \sim N(0, \boldsymbol{I}\sigma_p^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_a^2, σ_p^2 and σ_e^2 are the additive genetic, permanent environmental and residual variances, respectively; **A** is the relationship matrix and **I** is the identity matrix. In the multi-trait model, it was assumed that $\boldsymbol{a} \sim N(0, G_0 \otimes \boldsymbol{A})$, $\boldsymbol{p} \sim N(0, P_0 \otimes \boldsymbol{I})$ and $\boldsymbol{e} \sim N(0, R_0 \otimes \boldsymbol{I})$, in which G_0, P_0, and R_0 are the additive genetic, permanent environmental, and residual (co)variance matrices, respectively.

Random regression model

This model is similar to the repeatability model, using logarithmic scale transformed data and also included a fixed regression on the longitudinal variable from which the random regression (RR) would deviate. According to Akaike Information Criterion (AIC) the best fit model for the random regression analysis was obtained using a second-order Legendre polynomial (AIC = –559,83), when compared to 1st order (AIC = –266,32), 3rd order (AIC = –490,29) and 4th order (AIC = –294,04). The multi-trait model used to evaluate the traits was basically the same as the repeatability model:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Wp} + \boldsymbol{e}$$

in which the difference is that $\boldsymbol{a}$ is the vector of random regression coefficient on the Legendre polynomial for additive genetic effects; $\boldsymbol{p}$ is the random regression coefficient on Legendre polynomial for the permanent environmental effect. Homogeneity of variances within each age was assumed. All effects included in the model supposed the same distribution described in the repeatability model, assuming here a quadratic Legendre polynomial function.

Results

Descriptive statistics

Table 1 shows the descriptive statistics analysis for the OPU interval, donors' age and the traits: VO, TO, and EMBR. The longest interval between aspirations was 2,570 days and the donors' age during OPU sections ranged from one to 16 years. The number of total oocytes obtained during these years reached almost 300K, with an average of 15.82 oocytes per donor and a maximum of 182 in one aspiration section. Almost 52,000 embryos were produced,

with an average of 3.7 embryos per sample and a maximum value of 46. Table 2 shows the descriptive statistics analysis for cows, VO, TO and EMBR.

Table 1. Descriptive statistics[a] on aspiration interval, donors' age and structures[b]

Traits	N	Total	Mean	SD	MIN[c]	MAX	CV
Aspiration interval (days)	17,526	--	97.01	171.10	0	2570	176.37
Age (years)	17,526	--	6.27	3.19	1	16	50.83
Viable oocytes	17,524	277,292	15.82	12.89	0	182	81.46
Total oocytes	17,526	379,599	21.66	16.10	0	197	74.33
Embryos	17,498	51,506	2.94	3.70	0	46	125.66

[a]N: number of data samples; Total: total sum of structures (oocytes and embryos); SD: standard deviation; MIN: Minimum; MAX: Maximum; CV: coefficient of variation.

[b]Variable not log transformed: LN (X +1).

[c]Minimum for aspiration interval means two aspirations on the same cow on the same day with semen from different bulls.

Table 2. Descriptive statistics by age for cows and structures[a]

	Number	Number of structures			Mean			Standard deviation		
Age	of cows	VO	TO	EMBR	VO	TO	EMBR	VO	TO	EMBR
1	137	214	214	214	23.58	31.79	2.86	15.70	18.69	3.99
2	489	1122	1122	1122	22.07	28.89	3.25	14.71	16.84	3.81
3	736	2094	2094	2094	19.18	26.34	3.09	13.31	16.04	3.85
4	799	2546	2546	2541	17.84	24.28	2.92	14.21	17.33	3.69
5	774	2390	2390	2387	17.11	23.51	3.12	13.51	17.07	4.13
6	663	2243	2243	2238	16.58	22.89	3.24	12.76	16.63	3.95
7	543	1840	1841	1837	15.47	21.42	3.08	11.51	14.83	3.65
8	403	1322	1322	1321	13.77	18.33	2.76	10.35	12.47	3.63
9	309	962	962	958	13.12	18.15	2.77	10.93	13.58	3.31
10	221	770	771	768	11.36	15.80	2.88	9.54	12.13	3.46
11	178	674	674	673	9.70	13.89	2.63	7.92	10.79	3.08
12	111	437	437	435	8.07	11.55	2.32	7.68	10.05	2.71
13	71	283	283	283	7.95	11.22	2.35	7.25	10.23	2.98
14	59	239	239	239	6.11	8.82	1.87	5.56	7.74	2.31

15	46	227	227	227	5.81	8.26	1.67	6.22	8.52	2.43
16	38	161	161	161	4.20	6.14	1.36	3.64	5.07	1.55

[a]VO: number of viable oocytes; TO: number of total oocytes; EMBR: number of embryos. Variable not log transformed: LN (X +1).

Repeatability model

In Table 3 we have the variance components and genetic parameters resulting from the single-trait analysis for TO, VO, and EMBR. The additive genetic, permanent environment and residual variance estimates ranged respectively from 0.06 to 0.13, 0.05 to 0.08 and from 0.30 to 0.45. The heritability estimates were 0.24, 0.25 and 0.10 and the repeatability estimates were 0.39, 0.40 and 0.20 for TO, VO, and EMBR, respectively. The values of genetic and phenotypic correlations are shown in Table 4. The genetic correlations were high between all pairs of traits, with values above 0.80. The phenotypic correlation was high between VO and TO (0.95), and moderate for the others cases (below 0.40).

Table 3. Variance components and genetic parameters[a] estimated by single-trait analysis for viable oocytes (VO), total oocytes (TO) and embryos (EMBR)

Traits	$\hat{\sigma}_a$	$\hat{\sigma}_p$	$\hat{\sigma}_e$	h²	r
VO	0.13±0.02	0.08±0.01	0.31±0.004	0.25±0.03	0.40±0.01
TO	0.12±0.02	0.07±0.01	0.30±0.003	0.24±0.03	0.39±0.01
EMBR	0.06±0.01	0.05±0.01	0.45±0.005	0.10±0.02	0.20±0.01

[a]$\hat{\sigma}_a$: additive genetic variance; $\hat{\sigma}_p$: permanent environment variance; $\hat{\sigma}_e$: residual variance; h²: heritability; r: repeatability.

Table 4. Genetic (above the diagonal) and phenotypic correlations (bellow the diagonal) for viable oocytes (VO), total oocytes (TO) and embryos (EMBR)

Traits	VO	TO	EMBR
VO	-	1.00	0.82
TO	0.95	-	0.82
EMBR	0.35	0.37	-

Random regression model

The results from the random regression model were consistent with the repeatability model results and genetic values for the animals in each trait were also obtained (Supplementary file S2). The greatest number of oocytes and embryos was obtained between in the age range from two and eight years (> 1,000 units by age). The heritability estimates remained almost the same throughout ages, with moderate values for VO and TO, and low values for EMBR (Fig. 1). The heritability values ranged from 0.26 to 0.40 for VO, 0.24 to 0.40 for TO, and 0.08 to 0.20 for EMBR, with the age of cows ranging from one to 16 years.

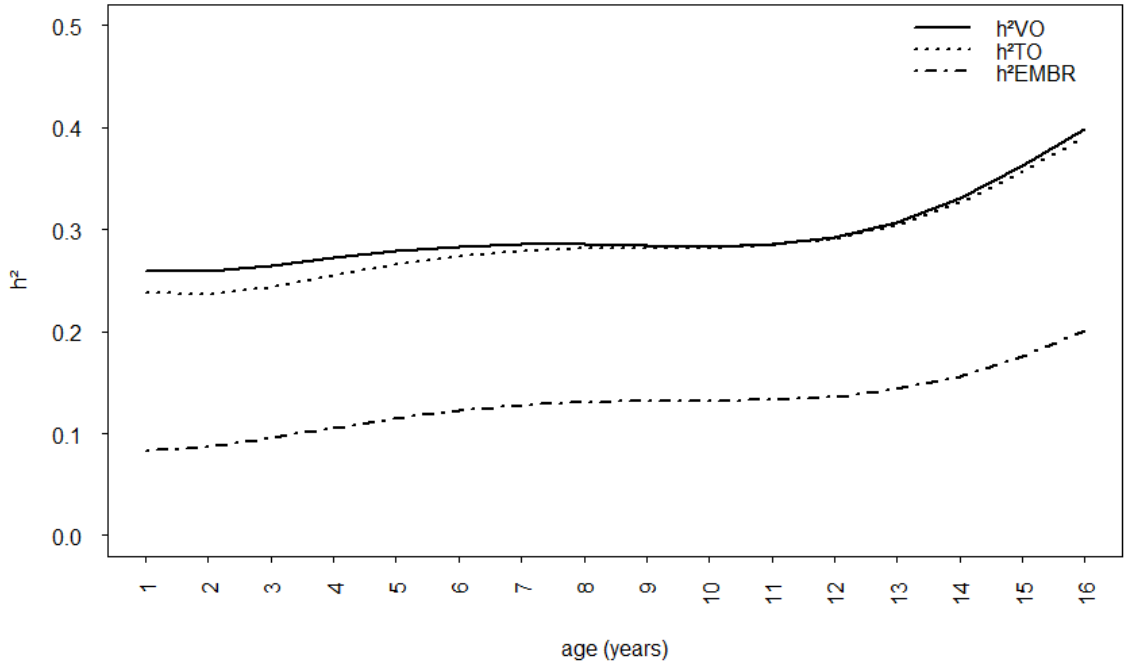


Fig. 1. Heritability (h^2) values for viable oocytes (VO), total oocytes (TO) and embryos (EMBR) throughout ages calculated by random regression model.

Additive genetic and permanent environmental variances across ages for each trait were estimated by random regression models (Fig. 2). For additive genetic variance, values ranged from 0.13 to 0.34 for VO, 0.10 to 0.35 for TO and 0.05 to 0.15 for EMBR and for permanent environmental variance the estimates ranged from 0.05 to 0.20 for VO, 0.05 to 0.26 for TO and 0.04 to 0.13 for EMBR, respectively.

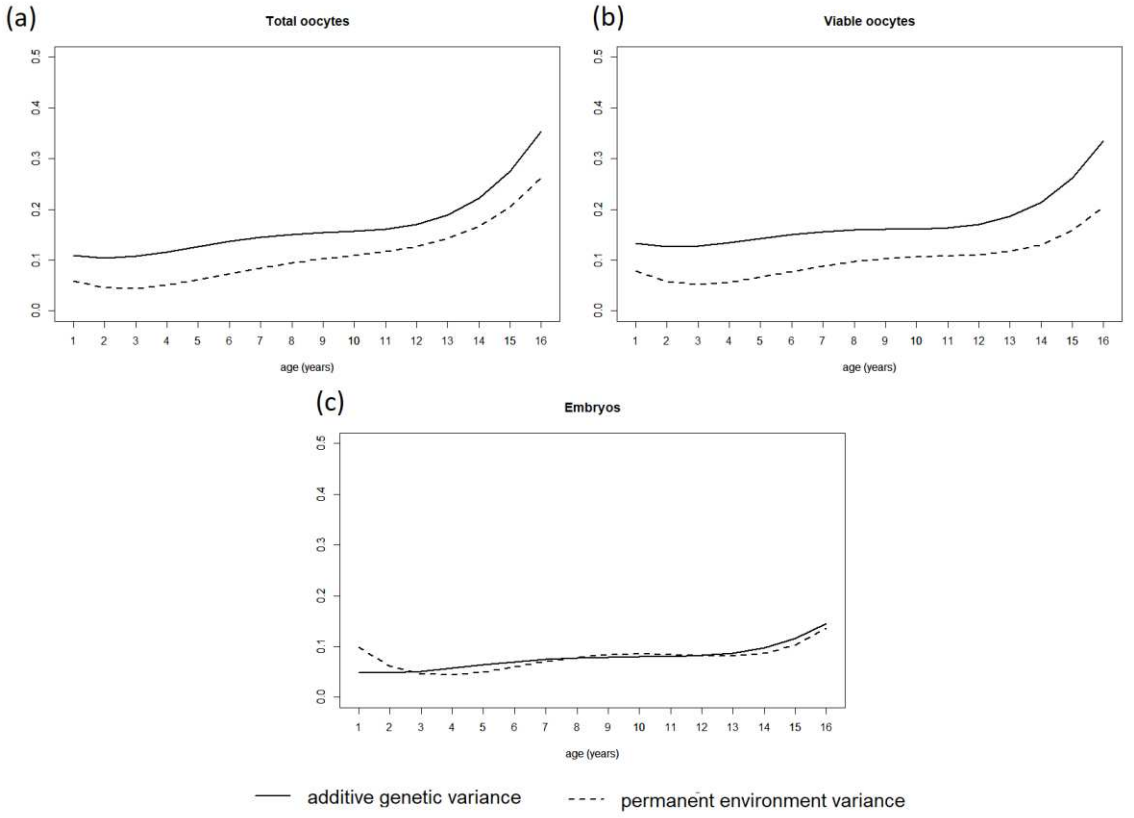


Fig. 2. Additive genetic and permanent environmental variance estimates for total oocytes (a), viable oocytes (b) and embryos (c) throughout ages (1 to 16 years old cows) calculated by random regression model.

Genetic correlation of repeated measures across ages for each trait was estimated with random regression models (Fig. 3) and resulted in moderate to high estimates among ages for each trait, with a minimum value of 0.43 for VO, 0.41 for TO and 0.47 for EMBR.

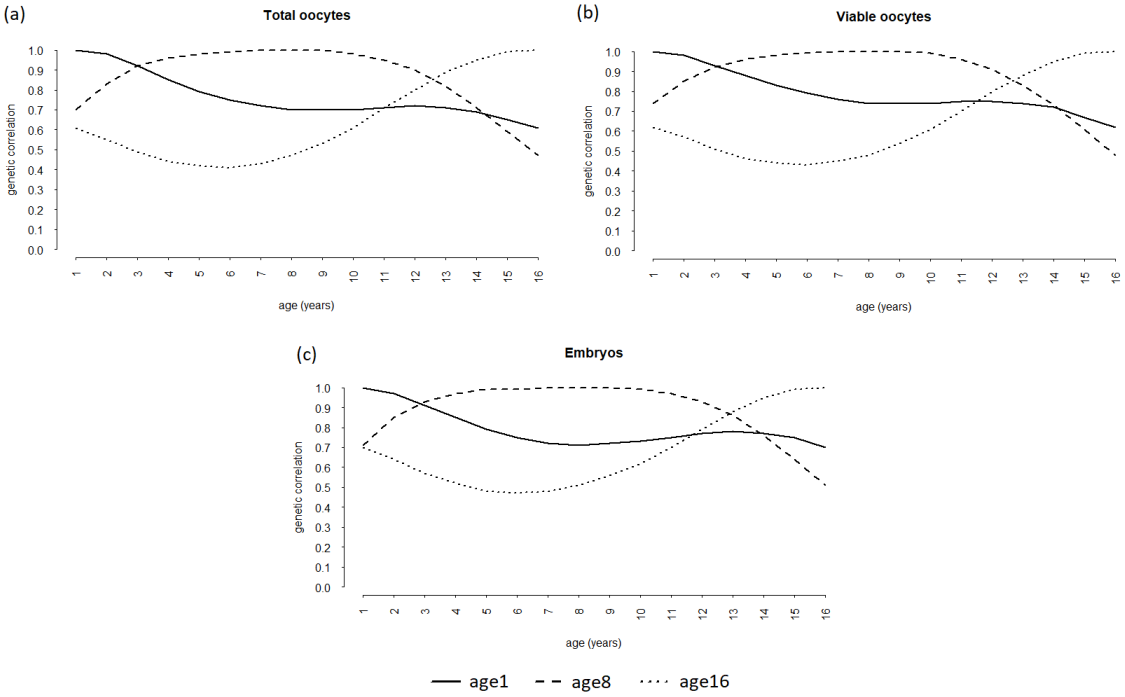


Fig. 3. Genetic correlations of repeated measures throughout ages calculated by random regression model with the ages of one year (age1, continuous), eight years (age8, dashed), or 16 years (age16, dotted) for total oocytes (a), viable oocytes (b) and embryos (c), individually.

Also, the genetic correlations between traits in each age were obtained (Fig. 4). These results were similar to the genetic correlation between traits estimated in the repeatability model. In all ages, the genetic correlation values between traits were positive, high (>0.70) and favorable. Despite the donors' age, the genetic correlation between total oocytes and viable oocytes was close to 1.

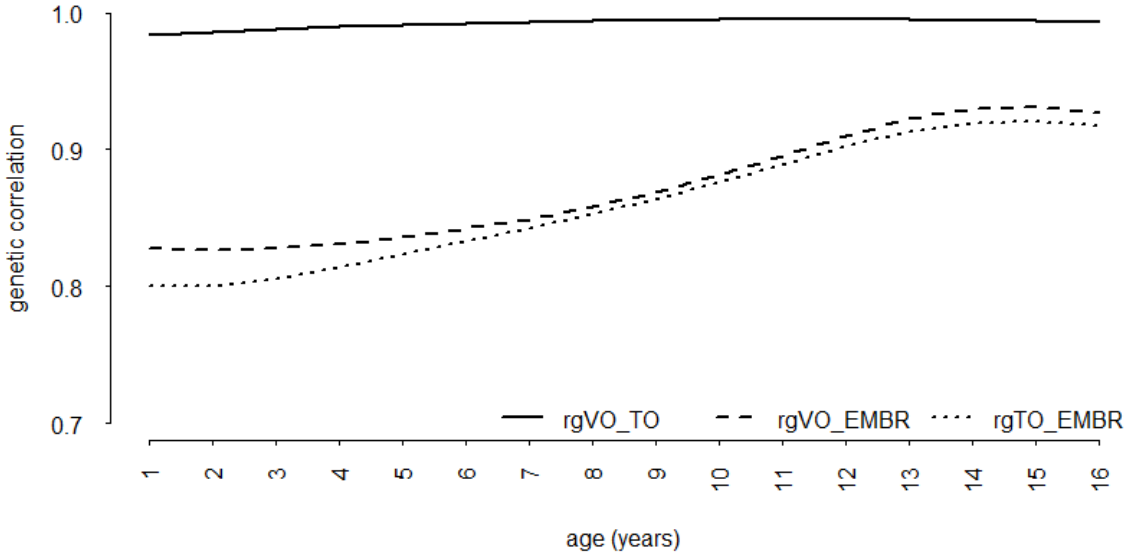


Fig. 4. Genetic correlation (r_g) between traits calculated by random regression model: viable oocytes (VO), total oocytes (TO) and Embryos (EMBR) in each age.

Discussion

Considering the importance of reproductive efficiency for the success in livestock production systems, the better understanding of the genetic parameters for the reproductive traits, such as number of oocytes and embryos may facilitate the selection of donors (Watanabe et al. 2017). Pinheiro (2019), for example, studied 1,292 samples for number of total oocytes and number of embryos from 571 Nellore donors, in 10 different herds. This author found an additive genetic variance of 0.06 and a residual variance of 0.10 for both traits, using repeatability models. In addition, heritabilities of 0.38 for number of total oocytes and 0.34 for number of total embryos were estimated. In our study, the additive genetic variation estimated with the repeatability model for number of total embryos was similar to the results of Pinheiro (2019) and higher in the case of total oocytes, although residual variances were greater and heritabilities were lower for those traits. Although our data were collected from only five different herds, it is worth mentioning that these herds are amongst the largest ones and included various different lineages of the breed, fully representing the Brazilian National Dairy Gir breeding program. On top of that, the environmental variation among them seems to be large as a result of a great residual variance, since such farms are located in regions bearing different management and environment conditions (data not shown). Besides, the differences in heritability values found in our work may be also due to breed differences, compared to the data evaluated by Pinheiro (2019).

For low heritability traits, as most of those related to reproduction, a small proportion of the phenotypic variation is due to genetics, however, it does not mean the additive genetic variance is not relevant (Visscher et al. 2008). Genetic progress through selection may be achieved in both productive and reproductive traits (Reis Filho et al. 2015), even if small additive genetic variation is observed, since genetic change is permanent and lasting, if there is continuous selection for the trait. Thus, the existence of additive genetic variation found in our work, in either repeatability or random regression models, indicates that selection of Gir donors for the production of oocytes and embryos is feasible.

Jaton et al. (2020) found in their literature review, heritability for number of embryos in Holstein cattle ranging from 0.03 to 0.21, similarly to a previous work of the same group (Jaton et al. 2016), which had been performed using repeatability models. Although we are working with a *Bos indicus* breed, the heritability for number of embryos is in the same range reported by Jaton et al. (2020), either with the repeatability or with the random regression model. Pereira (2019) studied 12,491 information of ovum pick-up from 5,496 Dairy Gir donors and found heritability of 0.12 to the number of viable oocytes and 0.03 for the number of viable embryos, which is also in the same range reported by Jaton et al. (2020). These results demonstrate that despite the breed, analytical method or environmental factors, embryo production presents low heritability estimates, being mainly affected by non-genetic factors. Beyond genetic and environmental influences over the donor in the process related to ovulation, aspiration and OPU, genetic and environmental effects related to sire, to semen preparation and laboratorial process have also to be considered.

Perez et al. (2016) studied 5,173 data points on number of viable oocytes, number of grade I oocytes, cleaved embryos, viable embryos, and the percentages of them, from 1,080 donor cows from the Guzerat breed, using a repeatability model and the variables transformed by $\log_e(X+1)$, the same used in our study. The heritability estimates in their study were 0.19 and 0.14, and repeatabilities were 0.32 and 0.21 for the number of viable oocytes and the number of viable embryos, respectively; in the present study, heritability and repeatability were slightly higher for viable oocytes and similar for embryos. Although their result of additive genetic variance was similar, the permanent environmental variance was greater and residual variance was lower compared to our study. The results found by Perez et al. (2016) indicate that the variation for both traits was of non-genetic origin. According to previous studies, *in vitro* production of embryos can be affected by several factors, such as climate and season (Chinchilla-vargas et al. 2018; Mello et al. 2016), nutrition, hormonal stimulation, number of performed OPU sections, reproductive effects, and laboratory process (Becher et al. 2018).

This justifies our repeatability model results with elevated values for the residual variances, which are related to the external environment of the animal. However, factors directly linked to the donor can also impact these traits, such as the genotype itself and the donor status with respect to its genetic merit, age, estrous cycle, and health (Mello et al. 2016). Also, the low variation of heritability and additive genetic variance according to the age of the cow, as shown by our results with the random regression analyses, indicates that the use of repeatability model is suitable for the estimation of genetic parameters for these traits; on top of that, repeatability models demand less computational effort than random regression models. However, it is important to highlight that the type of result and information provided by each model is different. Heritability estimates that were low and almost constant for the number of embryos indicate that this trait is more affected by environmental factors regardless the donors' age. Moreover, moderate, and likewise, constant heritability for number of viable and total oocytes, indicates that these traits are preferable for selection especially considering the high correlation with embryos. This way, selection of younger animals for these traits would decrease generation interval and hence increase genetic progress through identification of animals that produce viable oocytes and embryos at an early age, allowing them to be sampled for a longer period of time. The increase observed in heritabilities estimated for older animals, especially for those above 11 or 12 years of age, on the other hand, should not be considered for conclusions within the scope of the present study, because of the decreased number of data available for this age range, as seen in Table 2.

Other studies have also found moderate to high repeatability estimates for the number of oocytes and embryos or similar traits (Pereira 2019; Perez et al. 2016; Vizoná et al. 2020). Our repeatability results were moderate (Table 3), corroborating with our results using the random regression model, where genetic correlation values between repeated measurements at different ages for each trait ranged from moderate to high (Figure 3). This means that the closer the ages, the more the measurements were similar, and the more distant, the smaller the similarity is. So, the correlation between repeated measures changes from high to moderate with an increase in age difference, especially related to embryos and oocyte parameters.

In the present work, the genetic correlations between traits were higher than the phenotypic correlations, using the repeatability model (Table 4), consistent with the study of Pinheiro (2019). As mentioned earlier in the text, several environmental effects modulate the expression of the phenotype, demonstrating that the phenotype does not entirely express the genetic relationship between different traits. Sodini et al. (2018) suggest the genetic correlations of an independent sample help to predict the phenotypic correlations of another

independent sample within the same population. Our work also showed, using the random regression model (Figure 4), that there are high positive and favorable correlations between number of viable oocytes, number of total oocytes and number of embryos, regardless of donor´s age. So, the selection for viable oocytes will increase the number of total oocytes and embryos in the population. In our work, either using repeatability or random regression model, we found that the genetic correlation between total oocytes and viable oocytes was close to 1. Jaton et al. (2016), using a repeatability model described high genetic correlation (0.97) between number of embryos and number of viable embryos. Our study demonstrated with a random regression model that these high values are similar during all life span. Still, the genetic correlation values between total or viable oocytes and embryos are always over 0.7; in this way, the selection for these traits can be done in the first years of age.
Vizoná et al. (2020) indicated that *in vitro* embryo production can be used as a criterion to select Gir donors. However, in our work we found high genetic correlation between the number of oocytes (total and viable) and embryos, indicating that the number of oocytes is preferable to be used as a criterion to select Gir donors. When two traits have high genetic correlation, the genes shared between them are generally co-inherited (Sodini et al. 2018). The genetic and phenotypic correlations between viable and total oocytes were very high; this is expected considering they are the same structure/trait, and the correlations of these with the number of embryos were similar between them. Viable and total oocytes heritability was moderate, and also not different from each other. These results indicate that both ways of measuring the numbers of oocytes, total or viable, can be used with similar expected results in selective processes. Estimates obtained for genetic correlations between viable oocytes and embryos, or between total oocytes and embryos were high and similar, indicated that selection for oocyte numbers should result in a favorable correlated response for embryo numbers. The moderate estimates obtained for the phenotypic correlations between these traits, together with the low heritability for number of embryos, indicate that there are important non-genetic factors influencing number of embryos. Information regarding genetic parameters for the number of oocytes and embryos may be used in individual genetic selection to improve genetic progress for reproductive traits. This genetic progress can be expected with selection for Gir donors at younger ages, decreasing the generation interval. Further studies evaluating the genetic relationship between the number of oocytes or embryos and other economic traits in Gir breed are suggested, in order to understand genetic inheritance for the reproductive traits and define effective ways to select animals considering a consistent and permanent genetic progress in the breed.

Conclusion

Repeatability models can be used to estimate genetic parameters of oocytes and embryos as it requires less computational effort than the random regression models and the parameter estimates throughout ages do not vary substantially. Heritability values were basically invariable through ages. Due to the existence of genetic variation for the studied traits and the fact that number of embryos is more affected by non-genetic factors, the selection of donors for these traits, can be done preferably based on the number of oocytes (total or viable) at younger ages in order to increase genetic progress. Correlations between heritability estimates ranged from moderate values between distant ages to higher values between closer ages, so repeated measures are expected to be more similar at closer ages than at distant ages for all traits evaluated. Selection of Gir donors for higher total oocyte production, due to the positive correlated response, is expected to generate increasing numbers of embryos. Moreover, much of the variation in number of embryos come from non-genetic origin, indicating that environmental, technical and management aspects also need attention in assisted reproduction techniques, such as OPU and *in vitro* fertilization.

Data availability

The data sets analyzed during the current study are not publicly available because databases belong to the private commercial farms.

Conflicts of interest

The authors declare no conflicts of interest.

Declaration of Funding

Coordination and Improvement of Higher Level Personnel (CAPES), National Council for Scientific and Technological Development (CNPq) – process number 402935/2021-7, Ministério da Ciência Tecnologia e Inovação (MCTI) and Brazilian National Institute of Science and Technology in Animal Science (INCT-CA) provided financial support towards this study.

Acknowledgments

We thank the farms that provided the data for this study.

References

Asada Y, Terawaki Y (2002) Heritability and Repeatability of Superovulatory Responses in Holstein Population in Hokkaido, Japan. Asian-Australasian J Anim Sci 15:944–948. https://doi.org/10.5713/ajas.2002.944

Becher B, Neto A, Oliveira W, et al (2018) Fatores que afetam a produção in vitro de embriões (PIVE) em bovinos. Enciclopédia Biosf 15:554–570. https://doi.org/10.18677/EnciBio_2018B48

Berman A (2011) Invited review: Are adaptations present to support dairy cattle productivity in warm climates? J Dairy Sci 94:2147–2158. https://doi.org/10.3168/jds.2010-3962

Chinchilla-vargas J, Jahnke MM, Dohlman TM, et al (2018) Climatic factors a ff ecting quantity and quality grade of in vivo derived embryos of cattle. Anim Reprod Sci 192:53–60. https://doi.org/10.1016/j.anireprosci.2018.02.012

Davis G (1993) Genetic parameters for tropical beef cattle in northern Australia: a review. Aust J Agric Res 44:179. https://doi.org/10.1071/AR9930179

Drum JN, Sartori R, França e Melo L de (2019) New perspectives in assisted reproductive technologies in cattle. Rev Ecuatoriana Cienc Anim 3:85–106

Eriksson S, Häggström M, Stålhammar H (2007) ´Genetic parameters for superovulatory response in Swedish Red Cattle and Swedish Holstein heifers. In: ´58th Annual Meeting of the European Association for Animal Production 1–6'. (EAAP)

Falconer DS, Mackay TFC (1996) ´Introduction to quantitative genetics.´ (Longman Group Ltd., London, UK)

Gonçalves RLR, Viana JHM (2019) Situação atual da produção de embriões bovinos no Brasil e no mundo. Revista Brasileira de Reprodução Animal 43:156–159.

Jaton C, Koeck A, Sargolzaei M, et al (2016) Genetic analysis of superovulatory response of Holstein cows in Canada. J Dairy Sci 99:3612–3623. https://doi.org/10.3168/jds.2015-10349

Jaton C, Schenkel FS, Chud TCS, et al (2020) Genetic and genomic analyses of embryo production in dairy cattle. Reprod Fertil Dev 32:50. https://doi.org/10.1071/RD19275

König S, Bosselmann F, von Borstel UU, Simianer H (2007) Genetic Analysis of Traits Affecting the Success of Embryo Transfer in Dairy Cattle. J Dairy Sci 90:3945–3954. https://doi.org/10.3168/jds.2007-0089

McManus C, Teixeira R de A, Dias LT, et al (2008) Características produtivas e reprodutivas de vacas Holandesas e mestiças Holandês × Gir no Planalto Central. Rev Bras Zootec 37:819–823. https://doi.org/10.1590/S1516-35982008000500006

Mello RRC, Ferreira JE, Sousa SLG de, et al (2016) Fatores ligados à doadora que influenciam na

produção de embriões in vitro (PIVE). Revista Brasileira de Reprodução Animal 40:51–57.

Merton JS, Ask B, Onkundi DC, et al (2009) Genetic parameters for oocyte number and embryo production within a bovine ovum pick-up–in vitro production embryo-production program. Theriogenology 72:885–893. https://doi.org/10.1016/j.theriogenology.2009.06.003

Misztal I, Tsuruta S, Strabel T, Auvray B, Druet T, Lee DH (2002) BLUPF90 and related programs (BGF90). In: Proceedings of 7th World Congress on Genetics Applied to Livestock Production. Montpellier, France: (WCGALP), pp. 19–23.

Otto PI, Guimarães SEF, Calus MPL, et al (2020) Single-step genome-wide association studies (GWAS) and post-GWAS analyses to identify genomic regions and candidate genes for milk yield in Brazilian Girolando cattle. J Dairy Sci 103:10347–10360. https://doi.org/10.3168/jds.2019-17890

Otto PI, Guimarães SEF, Verardo LL, et al (2019) Genome-wide association studies for heat stress response in Bos taurus × Bos indicus crossbred cattle. J Dairy Sci 102:8148–8158. https://doi.org/10.3168/jds.2018-15305

Parker Gaddis KL, Dikmen S, Null DJ, et al (2017) Evaluation of genetic components in traits related to superovulation, in vitro fertilization, and embryo transfer in Holstein cattle. J Dairy Sci 100:2877–2891. https://doi.org/10.3168/jds.2016-11907

Pereira MA (2019) ´Análise genética de características de produção in vitro de embriões de doadoras da raça Gir Leiteiro´ Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Available in: https://repositorio.unesp.br/handle/11449/191802

Perez BC, Balieiro JCC, Oliveira Junior GA, et al (2019) State of inbreeding and genetic trends for estimated breeding values in IVF embryos and oocyte donors in the Brazilian Guzerá cattle. Theriogenology 125:71–78. https://doi.org/10.1016/j.theriogenology.2018.10.025

Perez BC, Peixoto MGCD, Bruneli FT, et al (2016) Genetic analysis of oocyte and embryo production traits in Guzerá breed donors and their associations with age at first calving. Genet Mol Res 15:1–9. https://doi.org/10.4238/gmr.15027583

Pinheiro AK (2019) ´Parâmetros produtivos e genéticos da produção in vitro de embriões em bovinos Nelore no Acre´ Dissertação (Mestrado) – Universidade Federal do Acre, Programa de Pós-Graduação em Sanidade e Produção Animal Sustentável na Amazônia Ocidental, Rio Branco. Available in: http://www.alice.cnptia.embrapa.br/alice/handle/doc/1119026

R Core Team (2020) ´R: A language and environment for statistical computing´ R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Reis Filho JC, Verneque R da S, Torres R de A, et al (2015) Inbreeding on productive and

reproductive traits of dairy Gyr cattle. Rev Bras Zootec 44:174–179. https://doi.org/10.1590/S1806-92902015000500002

Rotar LN, Souza JF (2019) Morphological Characteristics for Oocyte-Cumulus Complexes of Bos taurus and Bos indicus Cattle Used for a Variety of Purposes. Russ Agric Sci 45:399–402. https://doi.org/10.3103/S1068367419040141

Sanches BV, Zangirolamo AF, Seneda MM (2019) Intensive use of IVF by large-scale dairy programs. Anim Reprod 16:394–401. https://doi.org/10.21451/1984-3143-AR2019-0058

Santana ML, Pereira RJ, Bignardi AB, et al (2014) History, structure, and genetic diversity of Brazilian Gir cattle. Livest Sci 163:26–33. https://doi.org/10.1016/j.livsci.2014.02.007

Sodini SM, Kemper KE, Wray NR, Trzaskowski M (2018) Comparison of Genotypic and Phenotypic Correlations: Cheverud's Conjecture in Humans. Genetics 209:941–948. https://doi.org/10.1534/genetics.117.300630

Stephens MA (1986) Tests based on EDF statistics. In ´Goodness-of-Fit Techniques´. (Eds RB D'Agostino, MA Stephens), Marcel Dekker, New York: pp. 97-193.

Stroebech L, Mazzoni G, Pedersen HS, et al (2015) In vitro production of bovine embryos: revisiting oocyte development and application of systems biology. Animal Reproduction 12:465–472.

Thode Jr., HC (2002) ´Testing for Normality.´ (Marcel Dekker, New York.)

Tonhati H, Lôbo RB, Oliveira HN (1999) Repeatability and heritability of response to superovulation in holstein cows. Theriogenology 51:1151–1156. https://doi.org/10.1016/S0093-691X(99)80018-1

Viana J (2019) 2018 Statistics of embryo production and transfer in domestic farm animals. Embryo Technology Newsletter-IETS 36:1–26.

Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era — concepts and misconceptions. Nat Rev Genet 9:255–266. https://doi.org/10.1038/nrg2322

Vizoná RG, da Costa Perez B, Campolina Diniz Peixoto MG, et al (2020) Genetic analysis of in-vitro embryo production traits in Dairy Gir cattle. Theriogenology 148:149–161. https://doi.org/10.1016/j.theriogenology.2020.02.014

Watanabe YF, Souza HA, Mingoti RD, et al (2017) Number of oocytes retrieved per donor during OPU and its relationship with in vitro embryo production and field fertility following embryo transfer. Anim Reprod 14:635–644. https://doi.org/10.21451/1984-3143-AR1008

Wenceslau AA, Lopes PS, Teodoro RL, et al (2000) Estimação de parâmetros genéticos de medidas de conformação, produção de leite e idade ao primeiro parto em vacas da raça Gir Leiteiro. Rev Bras Zootec 29:153–158. https://doi.org/10.1590/S1516-35982000000100021

Runs of homozygosity and signatures of selection for number of oocytes and embryos in the Gir Indicine cattle*

* Published at Mammalian Genome (https://doi.org/10.1007/s00335-023-09989-w)

Renata de Fátima Bretanha Rocha[1], Arielly Oliveira Garcia[1], Pamela Itajara Otto[2], Marcos Vinícius Barbosa da Silva[3], Marta Fonseca Martins[3], Marco Antônio Machado[3], João Claudio do Carmo Panetto[3], Simone Eliza Facioni Guimarães[1*]

[1]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil, e-mail: renata.bretanha@ufv.com, arielly.garcia@ufv.br, sfacioni@ufv.br
[2]Department of Animal Science, Universidade Federal de Santa Maria, Santa Maria, Rio Grande do Sul, Brazil, e-mail: pamela.otto@ufsm.br
[3]EMBRAPA - Dairy Cattle National Research Center, Juiz de Fora, Minas Gerais, Brazil, e-mail: marcos.vb.silva@embrapa.br, marta.martins@embrapa.br, marco.machado@embrapa.br, joao.panetto@embrapa.br
[*]Corresponding author: Simone Eliza Facioni Guimarães, sfacioni@ufv.br

ORCID

Renata de Fátima Bretanha Rocha https://orcid.org/0000-0002-0552-7042
Arielly Oliveira Garcia https://orcid.org/0000-0001-6959-8117
Pamela Itajara Otto https://orcid.org/0000-0003-1702-4626
Marcos Vinícius Barbosa da Silva https://orcid.org/0000-0001-5449-1413
Marta Fonseca Martins https://orcid.org/0000-0002-4260-5329
Simone Eliza Facioni Guimarães https://orcid.org/0000-0003-3704-8131
João Claudio do Carmo Panetto https://orcid.org/0000-0002-9198-9728

Abstract

Runs of homozygosity (ROH) and signatures of selection are the results of selection processes in livestock species that have been shown to affect several traits in cattle. The aim of the current work was to verify the profile of ROH and inbreeding depression in the number of total (TO) and viable oocytes (VO) and the number of embryos (EMBR) in Gir Indicine cattle. In addition, we aim to identify signatures of selection, genes and enriched regions between Gir subpopulations sorted by breeding value for these traits. The genotype file contained 2,093 animals and 420,718 SNP markers. Breeding values used to sort Gir animals

were previously obtained. ROH and signature of selection analyses were performed using PLINK software, followed by ROH-based (F_{ROH}) and pedigree-based inbreeding (F_{ped}) and a search for genes and their functions. An average of 50 ± 8.59 ROHs were found per animal. ROHs were separated into classes according to size, ranging from 1 to 2 Mb ($ROH_{1–2\ Mb}$: 58.17%), representing ancient inbreeding, $ROH_{2–4\ Mb}$ (22.74%), $ROH_{4\text{-}8\ Mb}$ (11.34%), $ROH_{8\text{-}16\ Mb}$ (5.51%) and $ROH_{>16\ Mb}$ (2.24%). Combining our results, we conclude that the increase in general F_{ROH} and F_{ped} significantly decreases TO and VO; however, in different chromosomes traits can increase or decrease with F_{ROH}. In the analysis for signatures of selection, we identified 15 genes from 47 significant genomic regions, indicating differences in populations with high and low breeding value for the three traits.

Keywords: FST, inbreeding, in vitro embryo production, zebu cattle

Introduction

Natural and artificial selection have been used over the years to achieve more desirable phenotypes in livestock species (Brito et al. 2017). Runs of homozygosity (**ROH**), as a result of selection processes are contiguous lengths of homozygous genotypes that are present in an individual due to parents transmitting identical haplotypes to their offspring (Broman and Weber 1999, Purfield et al. 2012; Brito et al. 2017). In addition to natural/artificial selection, ROH size and distribution across the genome depend on several factors, such as genetic drift, founder effect and effective population size (Rebelato and Caetano 2018). As patterns of ROH may be decomposed, the ancestry of an individual and its population can be evaluated by the extent and frequency of ROHs in the following way: the shorter the segment, the more ancient is the relatedness of individuals (Purfield et al. 2012), and the longer the ROH segment, the more likely the inbreeding is recent in the population (Gibson et a. 2006; Kirin et al. 2010). These findings agree with Rebelato and Caetano (2018), to whom ROH sizes tend to decrease in the herd generation after generation.

ROH are used to explore genetic diversity and demography, identify candidate genes and/or estimate inbreeding in many species, such as chickens (Talebi et al. 2020), pigs (Shi et al. 2020), small ruminants (Bertolini et al. 2018; Nosrati et al. 2021) and large ruminants (Nascimento et al. 2021; Neves et al. 2015; Šidlová et al. 2015). ROHs have also been used as a potential inbreeding measure in livestock (Bosse et al. 2012; Marras et al. 2015), including the Gir breed (Peripolli et al. 2018; Toro Ospina et al. 2020).

According to Ma et al. (2019), the selection of dairy breeds over time focused on production traits, which decreased reproductive performance. Moreover, the negative correlation between milk production and reproductive traits, such as oocytes and embryos, is also described in the literature (Leroy et al. 2008; Ayasan et al. 2011). When exploring ROH analysis for Gir dairy cattle, Peripolli et al. (2018) found genes and ROHs associated with milk production and heat adaptation, but nothing was mentioned about reproductive traits. In addition, the analysis of ROH is useful to monitor the rate of inbreeding in a given population as well as the negative effects of inbreeding. In a breeding program, ROH analysis and monitoring might enhance the response to selection, pinpointing genomic regions associated with the target traits (Shi et al. 2020).

Genomic regions harboring candidate genes can be identified through signatures of selection, which are unique patterns in the genome of individuals who have undergone natural and/or artificial selection (Qanbari and Simianer 2014). Inside these genomic conserved regions, genes affecting phenotypic variation in livestock can be identified, in addition to exploring the genetic diversity of a population (Brito et al. 2017). Methodologies to identify signatures of selection have been studied for over 20 years. As an example, Akey et al. (2002) explored signatures of selection using high-density SNP panels with Wright's Fixation Index (F_{ST}, "Wright's Fixation Index"), which is ideal to compare populations (Wright 1978). Randhawa et al. (2016) carried out a review on signatures of selection identified in cattle and found that FST is the most commonly used. This methodology was also used to identify signatures of selection between beef and dairy Gir cattle (Maiorano et al. 2018). Randhawa et al. (2016) reviewed selection signatures found for adaptation, appearance and production traits in cattle and suggested that conclusions are limited in Zebu breeds, as there are few studies compared to European breeds.

Zebu breeds such as Gir cattle (*Bos primigenius indicus*) have higher oocyte quality than taurine cattle (*Bos primigenius taurus*), which allows for greater success in reproductive biotechnologies such as in vitro embryo production (Drum et al. 2019; Rotar and Souza 2019). Given the importance of the Gir breed in the tropics, work has been carried out to explore its productive capacity (do Nascimento Rangel et al. 2018; Pereira et al. 2019; Hortolani et al. 2022) as well as its reproductive performance (González-Herrera et al. 2022). However, citations exploring the production of oocytes and embryos using genomic approaches are still lacking in Zebu cattle.

In this way, our hypothesis in the current manuscript is that ROH and ROH-based inbreeding in Gir cattle can be positively or negatively related to the number of aspirated

oocytes and embryos produced in vitro. Also, we hypothesized that ROH-based genomic regions differing between subpopulations sorted by breeding values for the studied traits can be identified, as well as putative candidate genes. The aim of the current work is to verify the profile of runs of homozygosity and ROH-based inbreeding effects in the number of total and viable aspirated oocytes and the number of embryos produced in vitro in Gir Indicine cattle. In addition, we aim to identify signatures of selection and genes among animals grouped according to the breeding value for these traits and discuss their enriched regions.

Material and methods

Genotypic data

A premade genotype file with 2,093 animals and 420,718 SNP markers was provided by Brazilian Agricultural Research Corporation (EMBRAPA) – Center for Research in Dairy Cattle. This premade file is a result of animals genotyped with different density SNP panels, including Illumina BovineSNP50 BeadChip v2 (50K), Illumina BovineHD BeadChip (777K), GGP Indicus (34K), Z-Chip (30K) and GGP Indicus (50K). All were extracted and/or imputed for 777K, resulting in 420,718 SNPs.

Phenotypic data

A dataset with 17,526 follicular aspirations from 1,641 Gir donors was provided by five farms. These donors are daughters of 127 sires and 771 dams. These herds are among the largest herds and include various lineages of the Gir breed, fully representing the Brazilian National Dairy Gir breeding program. The traits evaluated were the number of viable oocytes (VO), number of total oocytes (TO) and number of embryos (EMBR). The data were used to verify inbreeding depression (see *Inbreeding effect on traits* section).

The phenotypic dataset along with a pedigree file, including 4,679 animals, containing up to fifteen generations previously described by Rocha et al. (2022), were previously used to estimate breeding values (BVs) for the animals in these three traits. A brief description of the data can be seen in Table 1. Additional information of the data such as heritability, repeatability, genetic and phenotypic correlation can be seen in Rocha et al. (2022). Briefly, a normal distribution of the residuals was obtained for all traits using the logarithmic scale ln (X + 1), where ln is the natural logarithm and X is the raw data (oocyte and embryos). The normal distribution was verified by the Anderson–Darling test (Stephens 1986; Thode 2002). BVs for all traits were predicted using BLUPF90 family programs (Misztal et al. 2002) with animal models. BVs were used in the current work to sort and group Gir subpopulations to run the signature of selection analysis.

Table 2. Descriptive statistics[A] on aspiration interval, donors' age and structures[B]

Traits	N	Total	Mean	SD	MIN[C]	MAX	CV
Aspiration interval (days)	17,526	--	97.01	171.10	0	2570	176.37
Age (years)	17,526	--	6.27	3.19	1	16	50.83
Viable oocytes	17,524	277,292	15.82	12.89	0	182	81.46
Total oocytes	17,526	379,599	21.66	16.10	0	197	74.33
Embryos	17,498	51,506	2.94	3.70	0	46	125.66

[A]N: number of data samples; Total: total sum of structures (oocytes and embryos); SD: standard deviation; MIN: Minimum; MAX: Maximum; CV: coefficient of variation.

[B]Variable not log transformed: LN (X +1).

[C]Minimum for aspiration interval means two aspirations on the same cow on the same day with semen from different bulls.

Runs of homozygosity (ROH)

The ROH analysis was performed with the genotype file with 2093 animals and 420,718 SNPs and their pedigree (sire and dam), with a total of 2625 animals. All parameters used to define ROHs were based on the literature (Purfield et al. 2012; Peripolli et al. 2018) and testing. First, our genotype file is a result of genotyped animals with different density SNP panels, which were extracted and/or imputed for 777K. Therefore, we separated the genotype file into three files: 1) all animals; 2) only animals originally genotyped with the Illumina BovineHD BeadChip; and 3) only animals genotyped with density chips other than the Illumina BovineHD BeadChip. In this way, we ran around 30 analyses for each file testing different parameters to obtain ROH with file 1. The same analyses were tested for files 2 and 3. Our aim with this testing was to define the parameters to obtain ROH and check the proportion of homozygosity and heterozygosity between files 1, 2 and 3, as the addition of heterozygous genotypes may occur in the imputation processes.

In this way, homozygous segments in each individual's genome were detected by the PLINK 1.9 program (Chang et al. 2015) using the following parameters: (i) sliding window of 50 SNPs across the genome; (ii) proportion of homozygous overlapping windows was 0.05; (iii) minimum number of consecutive SNPs included in an ROH was 50; (iv) minimum length of an ROH was set to 1 Mb; (v) maximum gap between consecutive homozygous SNPs was 1000 kb; (vi) density of one SNP per 100 kb; and (vii) maximum of 1 SNP with missing genotypes and up to one heterozygous genotype were allowed in an ROH. For quality control,

the parameters considered were (viii) Hardy-Weinberg equilibrium (HWe) exact test p-value of 10^{-6}; (ix) minor allele frequency (MAF) of 0.05; (x) call rates for variants of 0.02; and (xi) call rates for samples of 0.1. Variants with values lower than those established for HWe and MAF resulted in variant exclusion, as well as higher values of call rate for variants. In the same way, samples with higher values than the one established for call rate for samples were excluded. After quality control, 380,942 SNPs and all 2,093 animals remained. ROH were obtained in each animal and classified by size: ROH1–2 Mb, ROH2–4 Mb, ROH4–8 Mb, ROH8–16 Mb, and ROH>16 Mb. All ROH classes were present in generations 7 to 14, except ROH>16 Mb, which was only present in generations 7 to 13.

Numbers of ROHs per chromosome were obtained by calculating the average number of ROHs for all animals in each chromosome. The average size of ROHs per chromosome was obtained by calculating the average size of ROH per animal, and then the average per chromosome. The total size of ROHs per chromosome was obtained by calculating the sum of ROH sizes per animal and then the average per chromosome. Genome coverage, which is the proportion of the autosomal genome covered by ROHs, was also obtained. Genome coverage by each ROH class was calculated by multiplying the average number of ROHs per animal by the average ROH length, then dividing it by the total size of the bovine genome, 2489.37 Mb, and finally multiplying by 100 to obtain the percentage value. Genome coverage by chromosome was calculated by dividing the average length of ROH obtained for a given chromosome by the total size of this chromosome. This result was multiplied by 100 to obtain the percentage value. The total size of the bovine genome and chromosome lengths were obtained according to the consensus map ARS-UCD1.2 (Rosen et al. 2020).

Inbreeding

After obtaining runs of homozygosity, inbreeding coefficients based on ROH (F_{ROH}) were estimated for all individuals with the detectRUNS package (Biscarini et al. 2018) using the Froh_inbreeding() and Froh_inbreedingClass() functions and considering parameter class = 2. These analyses were performed in R software (R version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria), following McQuillan et al. (2008):

$$F_{ROH} = \frac{\sum_{j=1}^{n} L_{ROH}}{L_{aut}}$$

where L_{ROH} is the length of ROH and L_{aut} is the total size of the autosomes covered by markers. L_{aut} was taken to be 2489.37 Mb (2,489,385,779 base pairs), based on the consensus map of the bovine genome, ARS-UCD1.2 (Rosen et al. 2020). In addition, F_{ROH} was

calculated by chromosome and based on the ROH distribution of five minimum different lengths: F_{ROH}1–2 Mb, F_{ROH}2–4 Mb, F_{ROH}4–8 Mb, F_{ROH} 8–16 Mb, and F_{ROH} > 16 Mb.

In addition, pedigree-based inbreeding (F_{ped}) was obtained considering the pedigree file with 4,679 animals using the "pedigree" package (Coster 2012) in R software (R version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria).

Inbreeding effect on traits

Considering F_{ROH} and pedigree-based inbreeding per animal, a simple linear regression was used to verify the effect of the inbreeding coefficient on the raw number of structures (number of oocytes and embryos not log transformed) to verify how the original average number of structures varied with inbreeding increase. The following model was used:

$$y = \beta 0 + \beta 1x + e$$

where y is the number of raw structures, x is the F_{ROH}, β0 represents the intercept, β1 is the regression coefficient and e is the error. Linear regression was also obtained for each trait by chromosome. A Bonferroni correction test was performed for general inbreeding, adjusting the P-value for the number of tests [$P< 0.05/(3*(29+F_{ROH}))$], where 3 is the number of traits, 29 is the number of chromosomes, and F_{ROH} is the general inbreeding]. This test was also performed for each individual chromosome and F_{ped}, adjusting the P-value for the number of tests [$P < 0.05/3$], where 3 is the number of traits.

Signatures of selection

Breeding values (BV) for VO, TO and EMBR were previously obtained. Animals were ordered from high to low BV for each trait. The breeding values' mean and standard deviation for each trait were calculated. Animals were separated into group 1 (high BV) if they had a breeding value above the mean plus one standard deviation and group 2 (low BV) if their breeding value was below the mean minus one standard deviation. Group 1 had 312 animals for VO, 313 for TO, and 319 for EMBR, and group 2 had 332 animals for VO, 332 for TO and 330 for EMBR. For each trait, both groups were compared in the search for signatures of selection. A similar approach was used by Nani and Peñagaricano (2020). In group 1, there were 312 common animals between VO and TO, 216 between VO and EMBR, and 216 between TO and EMBR. In total, 216 animals were common among all three traits. In group 2, there were 332 common animals between VO and TO, 243 between VO and EMBR, and 243 between TO and EMBR. In total, 243 animals were common among all three traits. The same group sorting was performed considering only animals with accuracy of BV higher than 0.75. Group 1 had 76 animals for VO, 84 for TO, and 48 for EMBR, and group 2 had 90 animals for VO, 100 for TO and 50 for EMBR. In group 1, there were 76 common animals

between VO and TO, 32 between VO and EMBR, and 32 between TO and EMBR. In total, 32 animals were common among all three traits. In group 2, there were 86 common animals between VO and TO, 40 between VO and EMBR, and 38 between TO and EMBR. In total, 38 animals were common among all three traits (Supplementary File). Due to the small number of animals in each group considering accuracy of BV higher than 0.75, more than 2000 signatures of selection were identified, thus these results were not used for further analyses. Signature selection analyses were performed using Wright's FST command on PLINK software to obtain F_{ST} estimates for each autosomal variant (Weir and Cockerham, 1984).

The signatures of selection were identified when F_{ST} values were above 0.15, since as described by Wright (1978), the range from 0.15 to 0.25 indicates moderately great differentiation. This methodology has been used in cattle (Makina et al. 2015; Zhao et al. 2015). A total of 47 genomic regions were found, which were used to find positional genes located within a 20,000 base pair (bp) window (10 kb up- and downstream) of the SNP. Genes were retrieved from the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) based on the ARS-UCD1.2 assembly of the bovine genome. The literature was reviewed to identify whether the gene functions were related to VO, TO and EMBR. In addition, all annotated genes, except LOCs (uncharacterized regions in the genome), were further investigated based on the key words "oocyte" and "embryo" (one at a time) through VarElect NGS Phenotyper (Stelzer et al. 2016). The VarElect tool (https://varelect.genecards.org/) has been used in the literature to prioritize variant genes based on the disease/phenotype of interest (Suárez-Vega et al. 2018; Gutierrez-Quintana et al. 2021). To make such inferences, VarElect uses algorithms seeking associations between genes and phenotypes based on shared pathways, interaction networks, parology relations and other sources.

For the identification of enriched transcription factors with the TFM-explorer program (http://bioinfo.lifl.fr/TFM/TFME/), promoter sequences (FASTA format) 3000 bp upstream and 300 bp downstream from the gene transcription start site (Otto et al. 2019) based on the bovine genome were used. To identify the most likely candidate genes related to all traits, Cytoscape software (Shannon et al. 2003) was used to obtain biological processes from the Biological Networks Gene Ontology tool (BiNGO) (Maere et al. 2005). From all genes identified, LOC's were excluded for biological processes enrichment analyses, which were performed using the ClueGO application in Cytoscape (Bindea et al. 2009), considering a Bonferroni correction test.

Results

Runs of homozygosity (ROH)

A total of 105,327 ROH were found among the 2,093 Gir animals analyzed in this study. An average of 50 ± 8.59 ROHs were found per animal, with a maximum number of 95 and a minimum of 26 ROHs. The average length of ROHs per individual was 3.13 ± 1.01 Mb, with a minimum of 1 Mb and a maximum of 89.46 Mb. ROHs ranging from 1 to 4 Mb were present in all animals and represented 80.91% of all ROHs (Table 2). The proportion of ROHs larger than 16 Mb was 2.24%, with an average length of 24.66 Mb, and they were present in 1,132 animals. ROHs larger than 16 Mb covered 2.07% of the genome, the largest proportion compared to the other classes.

Table 2. Descriptions[A] of runs of homozygosity (ROH) number and length (in Mb) by ROH length class ($ROH_{1-2\ Mb}$, $ROH_{2-4\ Mb}$, $ROH_{4-8\ Mb}$, $ROH_{8-16\ Mb}$, and $ROH_{>16\ Mb}$).

Class	NROH	Percent (%)	LROH (Mb)	Number of animals	SROH	Genome coverage
$ROH_{1-2\ Mb}$	61,269	58.17	1.35	2,093	29.27	1.59%
$ROH_{2-4\ Mb}$	23,949	22.74	2.78	2,093	11.44	1.28%
$ROH_{4-8\ Mb}$	11,942	11.34	5.53	2,081	5.74	1.27%
$ROH_{8-16\ Mb}$	5,805	5.51	11.05	1,847	3.14	1.40%
$ROH_{>16\ Mb}$	2,362	2.24	24.66	1,132	2.09	2.07%

[A]NROH = total number of ROH; LROH = average length of ROH; SROH = average number of ROH per animal.

The average number and size, and total size of runs of homozygosity (ROH) per animal for each chromosome are presented in Figure 1 (A, B, and C) and Supplementary Table S1. The highest number of ROHs was identified on BTA5 (3.153), followed by BTA2 (3.004) and BTA1 (2.975), while the lowest was identified on BTA28, with 0.687 ROHs detected, BTA27 (0.715), and BTA25 (0.763). With respect to the size of the ROHs, BTA28 had the largest ROHs (4.042 Mb), followed by BTA14 (3.861 Mb) and BTA20 (3.699 Mb), and the smallest ROHs were present on BTA12 (2.699 Mb), BTA18 (2.783 Mb) and BTA17 (2.856 Mb).

The average percentage of chromosome coverage by runs of homozygosity is shown in Figure 1 (D) and Supplementary Table S1. BTA28 was the chromosome with the highest

coverage by ROHs (8.88%), followed by BTA25 (7.98%) and BTA27 (6.79%). The chromosomes with the lowest coverage were BTA1 (1.81%), BTA2 (2.50%), and BTA8 (2.59%).

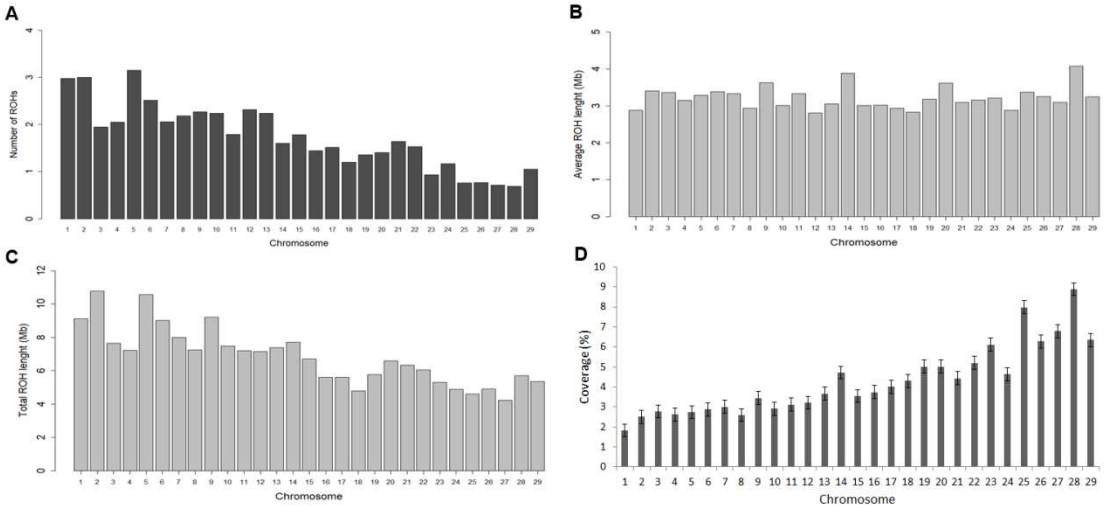


Fig. 1. Average number (A), average size in Mb (B), total size in Mb (C) of runs of homozygosity (ROH) per chromosome, and average percentage of chromosome coverage by ROH (D) with error bars indicating standard error

ROH-based inbreeding (F_{ROH})

From detectRUNS package, ROH-based inbreeding coefficient (F_{ROH}) is obtained for chromosome-wide, i.e., for each chromosome. F_{ROH} ranged from 0.05 on BTA1 to 0.11 on BTA28 (Figure 2; Supplementary Table S1, sheet 1). F_{ROH} was also obtained genome-wide, i.e., for each animal, and is also available (Supplementary Table S1, sheet 2). F_{ROH} was also obtained for every five minimum different lengths: 0.06 for F_{ROH}1–2 Mb, 0.04 for F_{ROH}2–4 Mb, 0.03 for F_{ROH}4–8 Mb, 0.02 for F_{ROH} 8–16 Mb, and 0.02 for F_{ROH} > 16 Mb.

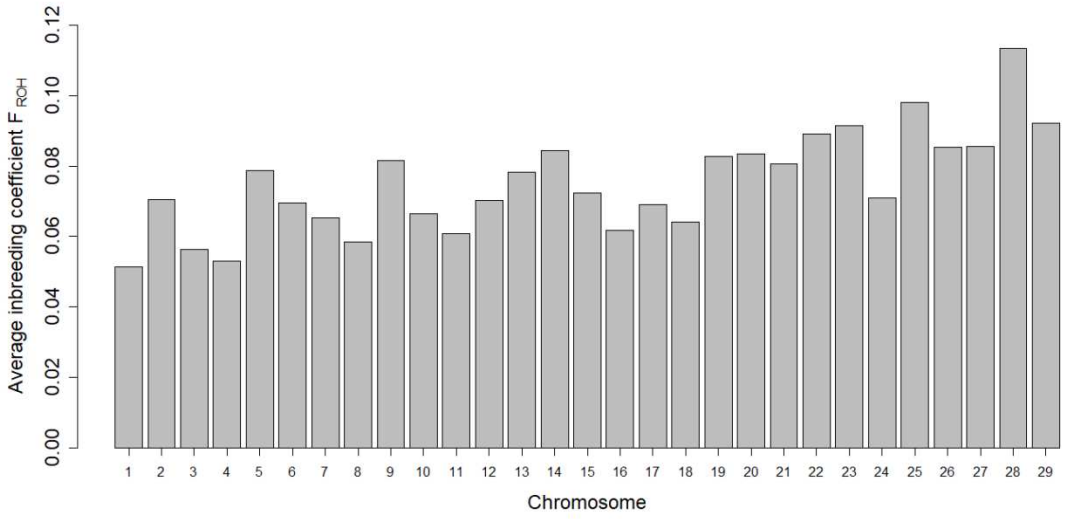


Fig. 2. Inbreeding coefficient based on runs of homozygosity (F_{ROH}) by chromosome

Inbreeding effect on traits

The number of VO and TO significantly decreased ($P < 0.001$) with the increase in F_{ROH} and F_{ped} (Figure 3). More details about intercept (β0), linear coefficient (β1), R^2, adjusted R^2 and p-values are available in Supplementary Table S2.

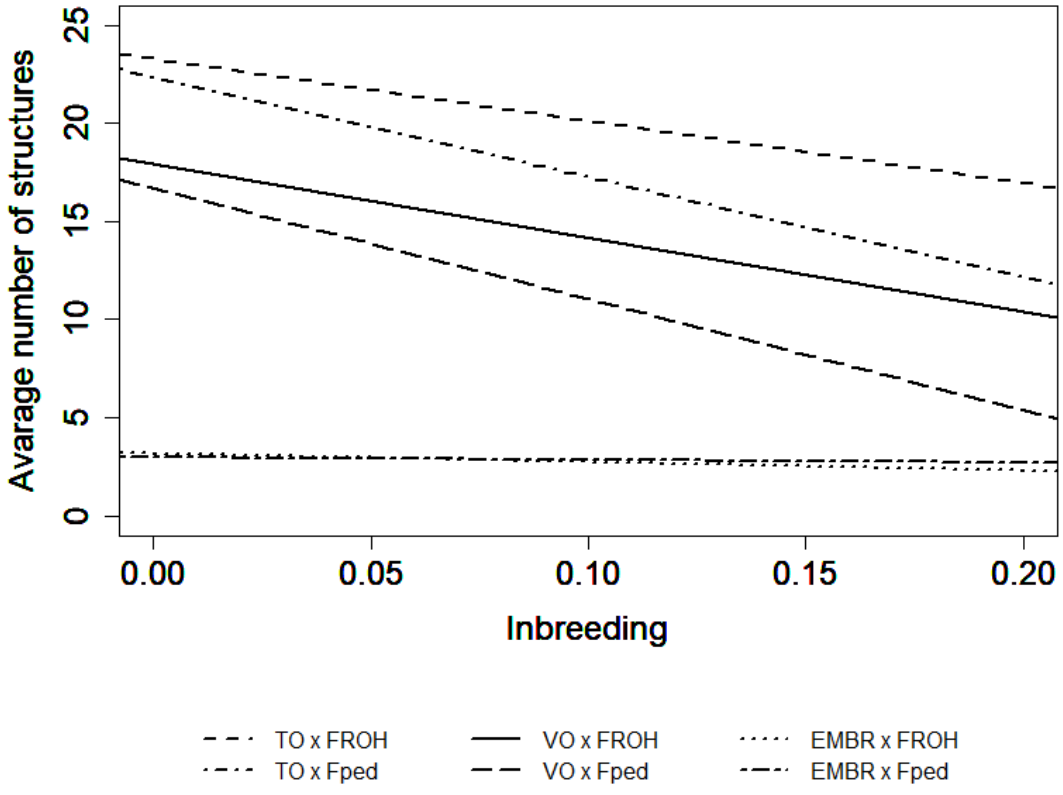


Fig. 3. Variation in the average number of total oocytes (TO), viable oocytes (VO) and embryos (EMBR) considering overall ROH-base (F_{ROH}) and pedigree-based inbreeding (F_{ped})

Despite the overall decrease in the average raw number of structures for the three traits, for each chromosome, the variation in F_{ROH} had a different effect on the traits (Figure 4). Chromosomes BTA4, BTA5, BTA11, BTA15, BTA16, BTA18, BTA19, BTA20, BTA22 and BTA26 had a significant decrease, according to the Bonferroni correction test, in the raw number of structures for all traits (see Supplementary Table S2 for more details on P-values). With the increase in F_{ROH}, the number of embryos decreased on BTA1 and BTA3 and increased on BTA6 and BTA24. The number of viable oocytes decreased on BTA1, BTA10 and BTA23, and along with TO, it also decreased on BTA17, BTA21, BTA27, BTA28 and BTA29. All the linear coefficients and p-values, such as other information related to these linear regressions, are presented in Supplementary Table S2. The total length of ROH (L_{ROH}), total size of the autosomes covered by markers (L_{aut}) and inbreeding coefficients based on ROH (F_{ROH}) for each animal in each *Bos taurus* autosome (BTA) were obtained using R software (R version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria), and are available in Supplementary Table S1.3. Also, scatterplots for all traits with the increase in FROH are available in Supplementary File.

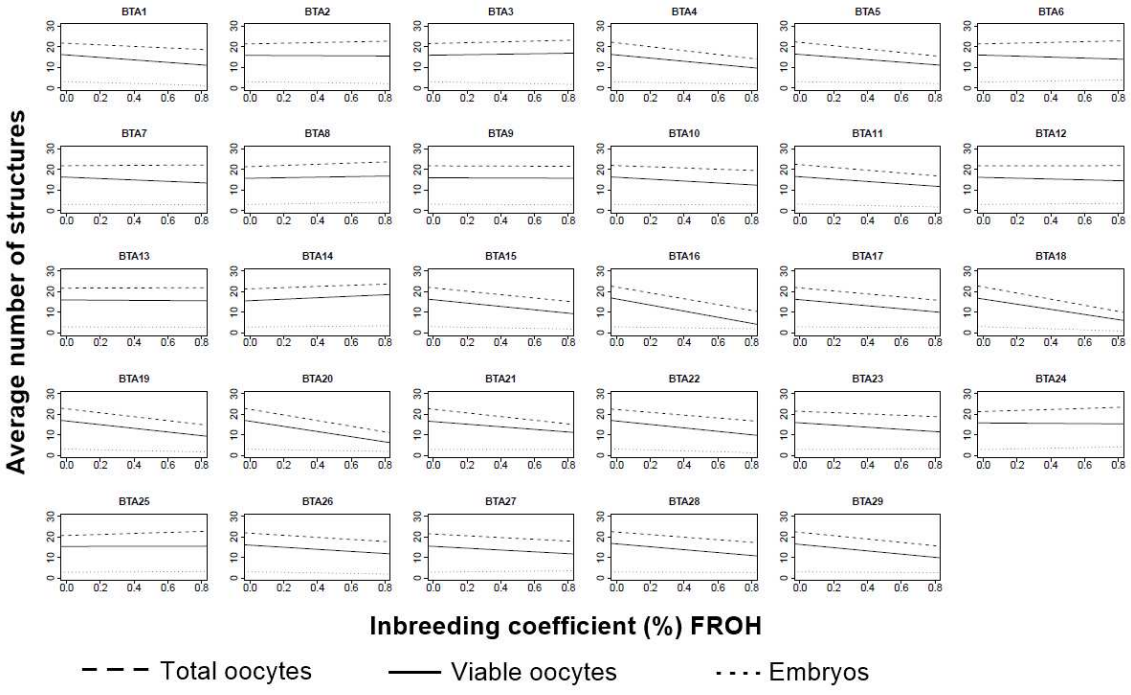


Fig. 4. Variation in the average number of structures considering F_{ROH} by chromosome

Signatures of selection

Manhattan plots with F_{ST} values for TO, VO and EMBR are presented in Figures 5, 6, and 7, respectively. F_{ST} values higher than 0.15 indicate moderately great differentiation between the groups of high and low breeding values (obtained for the traits TO, VO and EMBR with logarithmic transformation). The same 11 significant SNPs were identified for total and viable oocytes, and 40 significant SNPs were identified for embryos. Two positions on BTA11, one on BTA14 and one on BTA21 were coincident among all traits.

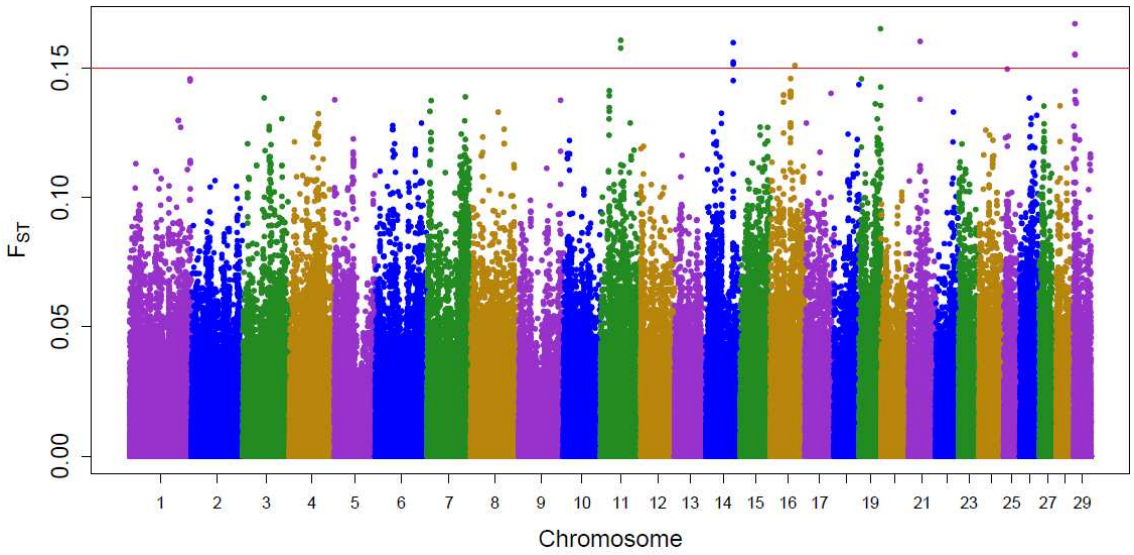


Fig. 5. Manhattan plot of the chromosomal distribution of F_{ST} in Gir cattle for the number of total oocytes. The red horizontal line represents the threshold $F_{ST} = 0.15$; above this, selection signatures are assigned.

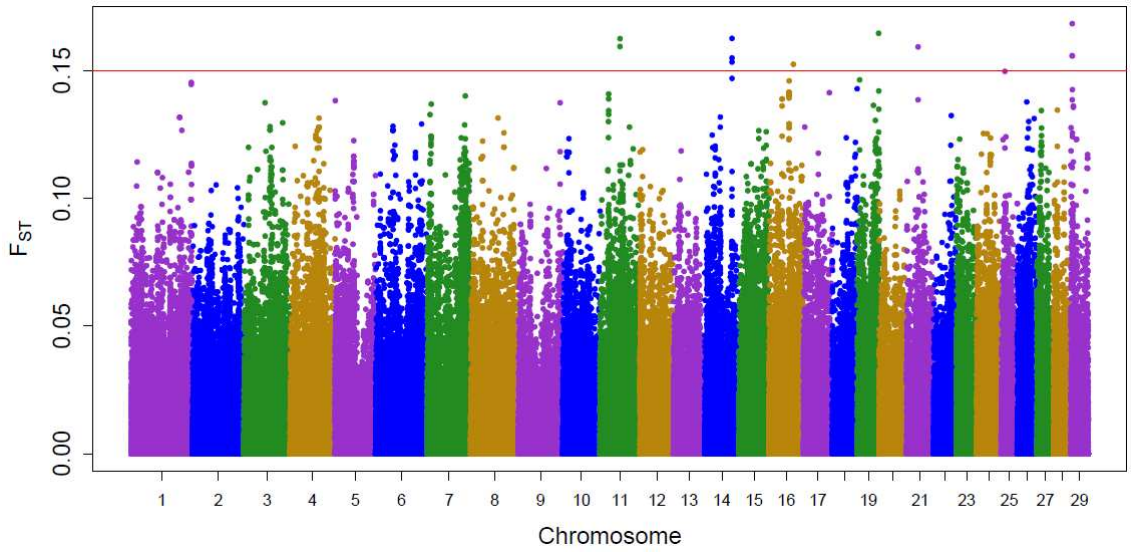


Fig. 6. Manhattan plot of the chromosomal distribution of F_{ST} in Gir cattle for the number of viable oocytes. The red horizontal line represents the threshold F_{ST} = 0.15; above this, selection signatures are assigned

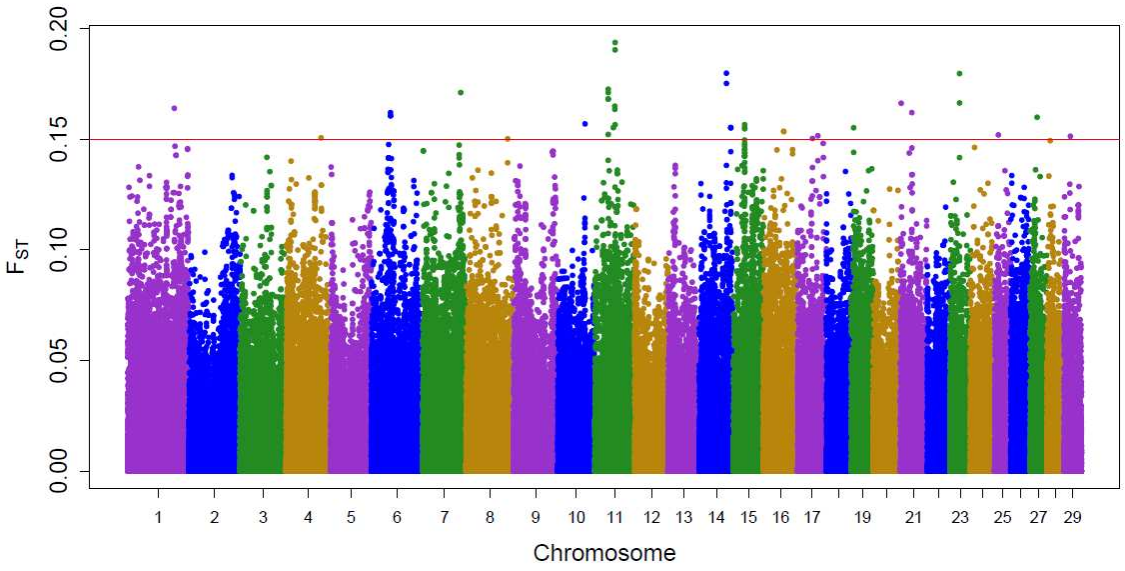


Fig. 7. Manhattan plot of the chromosomal distribution of F_{ST} in Gir cattle for the number of embryos. The red horizontal line represents the threshold F_{ST} = 0.15; above this, selection signatures are assigned

Gene search

From all 47 SNPs found with F_{ST} values higher than 0.15, 15 genes were identified (BTA): *TMEM245* (8), *USP39* (11), *DNAH6* (11), *LOC112448798* (11), *LOC112441622* (15), *SORL1* (15), *ABL2* (16), *KNTC1* (17), *RSRC2* (17), *LOC101907107* (19), *PEAK1* (21), *GSTA1* (23), *GSTA5* (23), *LOC112443730* (23) and *TENM4* (29) in a range of 10 kb upstream and 10 kb downstream of the SNP position (Supplementary Table S3). Excluding uncharacterized regions in the genome (LOCs), 11 protein-coding genes were found. According to the VarElect online tool, *KNTC1* showed a direct match to "oocyte" with a score of 0.27. This score is an indication of the strength of the connection between the gene and the traits, ranging from 1 to 200. The ratio of each score is calculated, and total scores are normalized to a range of 0.05 - 0.8. All other 10 genes were indirectly matched to oocytes. For "embryo", the genes (scores) directly matched were *ABL2* (0.67), *TENM4* (0.43), *SORL1* (0.24), *KNTC1* (0.21), *DNAH6* (0.14) and *USP39* (0.07). The other five genes were indirectly matched to embryos by protein-protein interactions.

Gene ontology

Eleven protein-coding genes found in the gene search were used to search for transcription factors (TFs) with the TFM-explorer program. Thirteen TFs were identified: *NFE2L1*, *MAFG*, *HNF1B*, *NFATC2*, *CTCF*, *NR3C1*, *LHX3*, *FOXD3*, *CEBPA*, *REL*, *ARID3A*, *HLTF* and *SRF* (Supplementary Table S3). These TFs were then used to search biological processes and to determine which gene ontology (GO) terms were significantly overrepresented. In this way, 127 biological processes were identified (Supplementary Table S4).

The gene set with the eleven protein-coding genes found in signatures of selection results was used to perform the biological process enrichment analysis. For that, a GO network highlighting biological roles and gene interactions was generated (Table 3). These are related to embryonic development, blastocyst formation, genetic imprinting and other processes related to cellular and oocyte development and early prenatal life.

Table 3. Enriched biological processes and their respective gene groups related to total oocytes, viable oocytes and embryos in Gir cattle.

Function	Group Genes
Central nervous system myelin formation	*TENM4*
Glutathione metabolic process	*GSTA2*
Mitotic cell cycle checkpoint	*KNTC1*
Protein autophosphorylation	*ABL2|PEAK1*
Regulation of choline O-acetyltransferase activity	*SORL1*
Spliceosomal complex assembly	*USP39*

Discussion

In the current article, we use genomic analysis to describe not only the state of art for the Gir Indicine cattle level of inbreeding, but we also use such information to build knowledge about the effects of such inbreeding in the production of oocytes and embryos, which is ultimately needed for successful genetic improvement of the most relevant Indicine dairy breed. We emphasize that the goal of selection signature analysis is not to look for association between genomic regions and any studied traits but to uncover genomic regions that differ between populations with high and low breeding values inside the Gir breed for number of oocytes and number of embryos. After the identification of such genomic regions through selection signature analysis, the search for genes harbored inside them was performed. Although no association studies were performed between the identified genes and the studied traits,

disentangling the genetic architecture behind the phenotypes might expand the range of information about these less explored traits. Genome-wide association studies to identify the actual association between genomic regions and the number of oocytes and embryos in this population will be carried out further.

Runs of Homozygosity

A high frequency of $ROH_{1\text{-}2}$ Mb compared to other ROH classes was detected, and this may reflect a strong ancient relationship within the breed, which is attributed to the use of few sires and families for breeding, hence leading to loss of genetic variability (Toro Ospina et al. 2020). Peripolli et al. (2018) also studied runs of homozygosity in Gir cattle and found 161,362 homozygous segments in 2908 individuals, in which most ROHs were shorter segments (ROH1–2 Mb). They also found an average number of ROH per animal of 55.12 ± 10.37 and a mean ROH length of 3.17 Mb. Considering the number of individuals used in their study and considering that the total number of ROHs obtained depends on the total number of animals and parameters to define an ROH, these results from Peripolli et al. (2018) are similar to those described in our work.

Inbreeding depression

Performing functional annotation analyses of ROH-enriched regions and with recently formed ROH (> 8 Mb) in Retinta cattle, Goszczynski et al. (2018) found gene clusters related to male and female reproductive functions, such as pregnancy-associated proteins, immune reaction and flagellum assembly, which is related to the assembly of cilium, a specialized eukaryotic organelle that consists of a filiform extrusion of the cell surface, and its impaired assembly may produce an abnormal movement of sperm flagellum. This may explain in part the reduced fertility in inbred animals. In our work, we have also described a reduction in the number of oocytes (total and viable) with increasing overall ROH-based (F_{ROH}) and pedigree-based inbreeding (F_{ped}).

Moreover, Goszczynski et al. (2018) also showed that animals harboring the same inbreeding coefficient may present different phenotypes, indicating that inbreeding will not necessarily be harmful to the trait of interest. In our study, we observed that the inbreeding coefficient in different chromosomes may affect the number of oocytes and embryos in different ways. Looking at the inbreeding level per chromosome and the significant positions found in the selection signature when comparing the groups with high and low breeding values for the number of oocytes and embryos, BTA11, BTA14, BTA16, BTA19, BTA21 and BTA29 had significant positions for total and viable oocytes, but only BTA14 showed a significant increase in the number of viable oocytes with an increase in inbreeding, while all

the others presented a decrease in the number of viable and total oocytes. BTA14 was the second chromosome with the greater size of ROHs, meaning that a part of recent inbreeding may be positively affecting the production of viable oocytes. Although no gene was identified in our search on BTA14 that could be related to an increased number of viable oocytes, the genes identified on BTA11, BTA16, BTA19, BTA21 and BTA29 probably have a negative influence on the number of oocytes.

Signatures of selection

In this work, we use 0.15 as an empirical threshold to identify signatures of selection since it is widely used in the literature (Makina et al. 2015; Wright 1978; Zhao et al. 2015). Not all approaches can accurately identify genes that are under positive selection because the proportion of true positives in such genes is unknown, so gene ontology (GO) categorization is often used to identify likely candidates (Pavlidis et al. 2012). In this work, as we found signatures of selection differing between groups with high and low breeding values for the number of total and viable oocytes and the number of embryos, our focus in the literature search was to verify whether the genes found could play a role in these traits. As suggested by Pavlidis et al. (2012), the pleiotropy of many genes, their complex relationships, and the wealth of information available in the literature may lead us to explain the action of positive selection on an arbitrary gene. However, many genes found in our work play important roles in biological processes related to reproductive traits, and as this is one of the first works for this type of trait, it provides insights and may open opportunities for new studies to validate these genomic regions, such as studies in other breeds, genome-wide association studies and gene expression.

The selection of groups based on accuracy above 0.75 promoted the selection of reliable estimated breeding values (BV), but reduced the number of animals available for the search of distinct genomic regions between groups. Due to this small number of animals, the number of signatures of selection found between groups increased and it became less assertive. This was seen by the identification of more than 2000 signatures of selection from these groups considering accuracy of BV higher than 0.75. As it gives many potentially biased results due to the small number of animals, results of groups with accuracy higher than 0.75 were not used for gene ontology analyses. This bias due to the small number of animals could be due to the fact that the Gir breed is an endogamic breed originated decades ago from animals imported from India (Santiago 1986). Our aim in this study was to have an overview about the genomic regions that currently influence such important traits (TO, VO, EMBR), especially with the growing use of embryo in vitro production (Drum et al. 2019; Rotar and

Souza 2019). Therefore, signatures of selection identified between groups considering all BVs were used for better overall picture in the genomic organization and the influence of endogamy in the Gir Breed.

Interaction: signatures x inbreeding x ROH

In our work, we found the *TENM4* gene on BTA29, which is the chromosome with the third highest inbreeding level. Additionally, a significant decrease in the number of oocytes was observed on BTA29. In the literature, the *TENM4* gene is related to the regulation of the nervous system and heart development (Karimi et al. 2021; Manzari et al. 2019). As *TENM4* plays a role in initial development and the inbreeding level of BTA29 negatively affects the number of oocytes, monitoring the inbreeding level becomes important to avoid impaired embryo development. When related to reproductive traits, the *DNAH6* gene on *BTA11* does not have many results in the literature for cattle, but in humans, Henarejos-Castillo et al. (2021) found novel variants predictive of ovarian failure, two of which are related to microtubule activity.

Our results showed that the inbreeding coefficient of BTA16 may be related to a decreased number of oocytes. Fonseca et al. (2020) compared highly fertile heifers and subfertile heifers and found differential expression of *ABL2* (BTA16) in the top 10 hub genes in one of the highly fertile modules of several cattle breeds (Angus, Limousin, Brangus, Simmental, Hereford, and Polled Hereford). Therefore, this gene seems to be important for fertility in cattle, and the inbreeding level negatively affects fertility traits, showing the importance of inbreeding monitoring, especially for reproductive traits.

Although genes such as *PEAK1* have not yet been related to reproductive traits, they may have undergone genetic drift over the years during selection for other traits. For example, the *PEAK1* gene (BTA21) was involved in the ATP binding process in a search for candidate genes related to carcass traits in Hanwoo cattle (Srikanth et al. 2020) and adaptative immune responses in West African cattle (Goyache et al. 2021). Exploring selection signatures based on ROH in Slovak spotted cattle, Kasarda et al. (2021) found *PEAK1* to be one of the candidate genes located in regions under selection pressure involved in cell development. Additionally, the *PEAK1* gene was described by Wang et al. (2018) as a tyrosine kinase and scaffold protein that facilitates cell movement and growth through integrin-mediated extracellular matrix signals, being highly conserved in vertebrates and important for angiogenesis.

Regarding the genes already mentioned, the *TMEM245* gene (BTA8) is one of the differentially expressed genes in blastocysts and is functionally involved in methylation and

acetylation in Holstein cows (Kalo et al. 2019). The *SORL1* gene (BTA15) was involved in conceptuses recovering from cows diagnosed with nonuterine diseases by Ribeiro et al. (2016). These authors explained that inflammatory diseases before breeding may reduce oocyte fertilization and morula development. Therefore, ROHs formed through our selection process over the years may harbor genes influencing fertility traits in different ways.

In our work, we identified the *KNTC1* and *RSRC2* genes on BTA17 when comparing breeding values for the number of embryos. Additionally, on BTA17, we observed a significant decrease in the number of oocytes with an increase in the inbreeding coefficient. Fonseca et al. (2020) mention the *KNTC1* gene (BTA17) as one of the target genes of an upstream regulator comparing cows with high and low fertility. Although there are few publications about *RSRC2* on BTA17 of cattle, Shaw et al. (2013) found that some transcripts encoding variants such as arginine/serine-rich_coiled-coil_2 (RSRC2) were shared exclusively between oocytes and 4-cell embryos, representing a maternal message expressed by the oocyte but removed after the onset of embryonic genome activation. As the *KNTC1* and *RSRC2* genes seem to be related to female fertility, it makes sense that the inbreeding level in females may affect oocyte production. This may explain why the *KNTC1* gene was directly matched with both oocyte and embryo terms in the VarElect online tool.

Fertility traits are complex and controlled by a large number of genes, each with small individual effects (Wiggans et al. 2011). This was observed in this study mostly for the number of embryos, in which we found 40 significant SNPs in the signature of selection analyses. In a meta-assembly of genomic regions associated with female reproductive efficiency in dairy and beef cattle breeds, Khatkar et al. (2014) reviewed and found multiple QTLs and SNP associations on every bovine chromosome. According to these authors, the highest meta-QTL scores were on BTA1, 2, 3, 6, 7 and 9, whereas the highest meta-GWAS scores were found on BTA1, 5, 13 and 16. Some of these chromosomes are coincident with those described in our work. In a genome-wide association study in Holstein breed, Parker Gaddis et al. (2017) found the largest proportion of variance explained by a 10-SNP window on chromosomes 8 and 14 for total number of structures collected by Ovum-Pick-Up and number of viable embryos, respectively. In our work, BTA8 and BTA14 also stood out for the number of oocytes (total and viable) and the number of embryos.

In relation to the identified genes, overexpression of several genes, including *ABL2* for inner cell mass compared to trophectoderm, was also found in the horse blastocyst transcriptome, and the gene ontology biological processes enriched analyses showed a relation with protein amino acid phosphorylation and actin filament-based process (Iqbal et al.

2014). Those studies corroborate our findings, not only for *ABL2* but also for the *PEAK1* gene and their relation with protein autophosphorylation enriched function.

In a search for the main dysregulated miRNAs in the uterine media, utero-tubal junction and isthmus exposed to seminal plasma, the *TMEM245* gene was one of several predicted targets of differentially expressed miRNAs in explant media from the uterus and utero-tubal junction of sows (Barranco et al. 2020). In a transcriptome analysis carried out in four brain regions of mice to determine estrous cycle-specific changes, *TMEM245* was an example of a differentially expressed gene in the hippocampus with higher expression in estrus than in diestrus (DiCarlo et al. 2017). Hou et al. (2017) listed *TMEM245* among mouse orthologs of human puberty-associated genes with expression related to puberty in hypothalamic-pituitary-gonadal axis tissues (hypothalamus, pituitary, ovary and testis). These findings might explain the importance of *TMEM245* related to central nervous system myelin formation function in the enriched biological processes that we found in this work.

The expression of glutathione S-transferase subunits α1 and α2 (GSTA1, GSTA2) is tissue- and cell-specific and has already been related to steroidogenically active cells, which are hormonally regulated by gonadotropins in bovine ovarian follicles (Rabahi et al. 1999). As a gene family member from *GSTA2*, glutathione-S-transferase A1 (GSTA1) function is related to protecting cells from reactive oxygen species, and its expression level positively correlates with oocyte quality (Salhab et al. 2013). Mezera et al. (2021) studied the effect of prostaglandin F2α pulses on the bovine luteal transcriptome during spontaneous luteal regression and found that glutathione s-transferases were mentioned as enriched terms. Therefore, the findings in the current work relating the enriched glutathione metabolic process with *GSTA2* may be due to its relation with oxidative stress response regulation.

Bogliotti et al. (2020) studied transcript profiling of bovine embryos and found several significantly enriched GO terms for transcripts whose relative abundance decreased from metaphase II eggs to the 8-to-16-cell stage, including *KNTC1*, which is related to the biological process of the cell cycle. In beef cows (75% Red Angus, 25% MARC III), McFee et al. (2021) found that *KNTC1*'s gene function is related to mitosis and that this gene is less expressed in granulosa cell nuclei with a high androstenedione concentration ([A4] ≥40 ng/mL) than follicles with an average/control concentration (A4 ≤20 ng/mL) and average estradiol. These findings in the literature are consistent with the results on the *KNTC1* gene being related to the mitotic cell cycle checkpoint in gene ontology analyses.

In our work, the *SORL1* gene was found to be involved in the regulation of choline O-acetyltransferase activity, which is a general term able to influence several functions and

processes in the organism. For example, the *SORL1* gene has been identified as a candidate gene for lipid metabolism in Nellore cattle (Carreño et al. 2019). On the other hand, the *SORL1* gene is related to cholesterol metabolism being upregulated in granulosa cells collected from chicken follicles (Zhu et al. 2019). When related to reproduction, the *SORL1* transcript was in the top 10 downregulated genes of conceptuses recovered from cows diagnosed with nonuterine diseases before artificial insemination when compared with those that were not diagnosed with nonuterine disease (Ribeiro et al. 2016).

Ubiquitin-specific peptidase 39 (*USP39*) functions in pre-mRNA splicing, and its knockdown was suggested as a novel alternative to targeted therapy of medullary thyroid carcinoma (An et al. 2015). This is consistent with our result that the *USP39* gene is involved in spliceosomal complex assembly, a process responsible for catalyzing the splicing of nuclear mRNA. However, information about this gene in bovine species in the literature is still scarce.

Conclusion

The runs of homozygosity (ROH) profile indicates an ancient genomic relationship in this Gir population.

General ROH-based inbreeding (F_{ROH}) has a negative effect on the number of oocytes (total and viable) and embryos of Gir cattle. This indicates that more attention should be paid when selecting animals for reproduction to avoid inbreeding depression events. However, the effect of F_{ROH} on the studied traits may be positive or negative depending on which chromosome the ROH are located. These findings provide great insight into how genomic regions can be beneficial or detrimental when affecting traits. Further association studies may consider this paper when checking where the identified genomic regions or genes associated with those traits are located. Signatures of selection were identified, in which eleven protein-coding genes were located. Some of these protein-coding genes are already related in the literature to reproductive traits or to the same enriched biological processes identified in our study. This indicates that even inside a breed, selection marks for reproductive traits are present. Although the number of oocytes and embryos are not considered selection traits yet by the Brazilian National Dairy Gir breeding program, the selection process for other traits may have led to signatures affecting the studied traits. To uncover the actual effect of these genes in oocytes and embryos, studies on gene expression, such as transcriptional landscapes, are recommended.

Acknowledgments We thank the farmers who provided data for this study.

Author contributions JCdCP, MAM, MVBdS, SEFG: conceptualization of the study. RFBR, SEFG: drafted the manuscript. RFBR: performed the experiments and analysis. AOG, PIO: statistical analyses support. MVGBdS, MFM, MAM, JCdCP: bioinformatics technical support. All authors read and approved the final manuscript.

Funding Coordination and Improvement of Higher Level Personnel (CAPES), National Council for Scientific and Technological Development (CNPq), Ministério da Ciência Tecnologia e Inovação (MCTI) and Brazilian National Institute of Science and Technology in Animal Science (INCT-CA) provided financial support towards this study.

Data availability statement The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Statements and Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval Not applicable. Farmers who provided the data adherence to a high standard of veterinary care.

Consent for publication Not applicable.

References

Akey JM, Zhang G, Zhang K, et al (2002) Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Res 12:1805–1814. https://doi.org/10.1101/gr.631202

An Y, Yang S, Guo K, et al (2015) Reduced USP39 expression inhibits malignant proliferation of medullary thyroid carcinoma in vitro. World J Surg Oncol 13. https://doi.org/10.1186/s12957-015-0669-4

Ayasan T, Hizli H, Çamlidag A, et al (2011) Determination of relationship between milk production and the quality of embryo of donor cows. Indian J Anim Sci 81:912–914

Barranco I, Padilla L, Martinez CA, et al (2020) Seminal plasma modulates mirna expression by sow genital tract lining explants. Biomolecules 10:933. https://doi.org/10.3390/biom10060933

Bertolini F, Cardoso TF, Marras G, et al (2018) Genome-wide patterns of homozygosity

provide clues about the population history and adaptation of goats. Genet Sel Evol 50. https://doi.org/10.1186/s12711-018-0424-8

Bindea G, Mlecnik B, Hackl H, et al (2009) ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25:1091–1093. https://doi.org/10.1093/bioinformatics/btp101

Biscarini F, Cozzi P, Gaspa G, Marras G (2018) detectRUNS: an R package to detect runs of homozygosity and heterozygosity in diploid genomes. R package version 0.9.6. https://CRAN.R-project.org/package=detectRUNS.

Bogliotti YS, Chung N, Paulson EE, et al (2020) Transcript profiling of bovine embryos implicates specific transcription factors in the maternal-to-embryo transition. Biol Reprod 102:671–679. https://doi.org/10.1093/biolre/ioz209

Bosse M, Megens H-J, Madsen O, et al (2012) Regions of Homozygosity in the Porcine Genome: Consequence of Demography and the Recombination Landscape. PLoS Genet 8:e1003100. https://doi.org/10.1371/journal.pgen.1003100

Brito LF, Kijas JW, Ventura RV, et al (2017) Genetic diversity and signatures of selection in various goat breeds revealed by genome-wide SNP markers. BMC Genomics 18. https://doi.org/10.1186/s12864-017-3610-0

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet 65:1493–1500. https://doi.org/10.1086/302661

Carreño LOD, Da Conceição Pessoa M, Espigolan R, et al (2019) Genome Association Study for Visual Scores in Nellore Cattle Measured at Weaning. BMC Genomics 20. https://doi.org/10.1186/s12864-019-5520-9

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:s13742–015–0047–8. https://doi.org/10.1186/s13742-015-0047-8

Coster A (2012) pedigree: Pedigree functions. R package version 1.4. https://CRAN.R-project.org/package=pedigree.

DiCarlo LM, Vied C, Nowakowski RS (2017) The stability of the transcriptome during the estrous cycle in four regions of the mouse brain. J Comp Neurol 525:3360– 3387. https://doi.org/10.1002/cne.24282

Drum JN, Sartori R, França e Melo L de (2019) New perspectives in assisted reproductive technologies in cattle. Rev Ecuatoriana Cienc Anim 3:85–106

Fonseca PAS, Suárez-Vega A, Cánovas A (2020) Weighted gene correlation network meta-

analysis reveals functional candidate genes associated with high-and sub-fertile reproductive performance in beef cattle. Genes (Basel) 11:543. https://doi.org/10.3390/genes11050543

Gibson J, Morton NE, Collins A (2006) Extended tracts of homozygosity in outbred human populations. Hum Mol Genet 15:789-795. https://doi.org/10.1093/hmg/ddi493

González-Herrera LG, Pereira RJ, El Faro L, Albuquerque LG (2022) Genetic evaluation of lactation persistency in the Gyr breed by using a two-trait random regression model. Anim Prod Sci 62:216-224. https://doi.org/10.1071/AN20267

Goszczynski D, Molina A, Terán E, et al (2018) Runs of homozygosity in a selected cattle population with extremely inbred bulls: Descriptive and functional analyses revealed highly variable patterns. PLoS One 13:e0200069. https://doi.org/10.1371/journal.pone.0200069

Goyache F, Pérez-Pardal L, Fernández I, et al (2021) Ancient autozygous segments subject to positive selection suggest adaptive immune responses in West African cattle. Gene 803:145899. https://doi.org/10.1016/j.gene.2021.145899

Gutierrez-Quintana R, Mellersh C, Wessmann A, et al (2021) Hereditary sensory and autonomic neuropathy in a family of mixed breed dogs associated with a novel *RETREG1* variant. J Vet Intern Med 35:2306–2314. https://doi.org/10.1111/jvim.16242

Henarejos-Castillo I, Aleman A, Martinez-Montoro B, et al (2021) Machine learning-based approach highlights the use of a genomic variant profile for precision medicine in ovarian failure. J Pers Med 11:609. https://doi.org/10.3390/jpm11070609

Hortolani B, Bernardes PA, Filho AEV, et al (2022) Genetic parameters for body weight and milk production of dairy Gyr herds. Trop Anim Health Prod 54. https://doi.org/10.1007/s11250-022-03088-9

Hou H, Uusküla-Reimand L, Makarem M, et al (2017) Gene expression profiling of puberty-associated genes reveals abundant tissue and sex-specific changes across postnatal development. Hum Mol Genet 26:3585–3599. https://doi.org/10.1093/hmg/ddx246

Iqbal K, Chitwood JL, Meyers-Brown GA, et al (2014) RNA-seq transcriptome profiling of equine inner cell mass and trophectoderm. Biol Reprod 90:61. https://doi.org/10.1095/biolreprod.113.113928

Kalo D, Vitorino Carvalho A, Archilla C, et al (2019) Mono(2-ethylhexyl)phthalate (MEHP)induces transcriptomic alterations in oocytes and their derived blastocysts. Toxicology 421:59–73. https://doi.org/10.1016/j.tox.2019.04.016

Karimi K, Farid AH, Myles S, Miar Y (2021) Detection of selection signatures for response to

Aleutian mink disease virus infection in American mink. Sci Rep 11. https://doi.org/10.1038/s41598-021-82522-8

Kasarda R, Moravčíková N, Olšanská B, et al (2021) The evaluation of genomic diversity and selection signals in the autochthonous Slovak Spotted cattle. Czech J Anim Sci 66:251–261. https://doi.org/10.17221/265/2020-CJAS

Khatkar MS, Randhawa IAS, Raadsma HW (2014) Meta-assembly of genomic regions and variants associated with female reproductive efficiency in cattle. Livest Sci 166:144–157. https://doi.org/10.1016/j.livsci.2014.05.015

Kirin M, McQuillan R, Franklin CS, et al (2010) Genomic runs of homozygosity record population history and consanguinity. PLoS One 5:e13996. https://doi.org/10.1371/journal.pone.0013996

Leroy J, Opsomer G, Van Soom A, et al (2008) Reduced Fertility in High-yielding Dairy Cows: Are the Oocyte and Embryo in Danger? Part I The Importance of Negative Energy Balance and Altered Corpus Luteum Function to the Reduction of Oocyte and Embryo Quality in High-yielding Dairy Cows*. Reprod Domest Anim 43:612–622. https://doi.org/10.1111/j.1439-0531.2007.00960.x

Ma L, Cole JB, Da Y, VanRaden PM (2019) Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. J Dairy Sci 102:3735–3743. https://doi.org/10.3168/jds.2018-15269

Maere S, Heymans K, Kuiper M (2005) BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21:3448–3449. https://doi.org/10.1093/bioinformatics/bti551

Maiorano AM, Lourenco DL, Tsuruta S, et al (2018) Assessing genetic architecture and signatures of selection of dual purpose Gir cattle populations using genomic information. PLoS One 13:e0200694. https://doi.org/10.1371/journal.pone.0200694

Makina SO, Muchadeyi FC, Van Marle-Köster E, et al (2015) Genome-wide scan for selection signatures in six cattle breeds in South Africa. Genet Sel Evol 47. https://doi.org/10.1186/s12711-015-0173-x

Manzari Z, Mehrabani-Yeganeh H, Nejati-Javaremi A, et al (2019) Detecting selection signatures in three Iranian sheep breeds. Anim Genet 50:298–302. https://doi.org/10.1111/age.12772

Marras G, Gaspa G, Sorbolini S, et al (2015) Analysis of runs of homozygosity and their relationship with inbreeding in five cattle breeds farmed in Italy. Anim Genet 46:110–121. https://doi.org/10.1111/age.12259

McFee RM, Romereim SM, Snider AP, et al (2021) A high-androgen microenvironment inhibits granulosa cell proliferation and alters cell identity. Mol Cell Endocrinol 531:111288. https://doi.org/10.1016/j.mce.2021.111288

McQuillan R, Leutenegger AL, Abdel-Rahman R, et al (2008) Runs of Homozygosity in European Populations. Am J Hum Genet 83:359–372. https://doi.org/10.1016/j.ajhg.2008.08.007

Mezera MA, Li W, Liu L, et al (2021) Effect of natural pre-luteolytic prostaglandin F2αpulses on the bovine luteal transcriptome during spontaneous luteal regression†. Biol Reprod 105:1016–1029. https://doi.org/10.1093/biolre/ioab123

Misztal I, Tsuruta S, Strabel T, et al (2002) BLUPF90 and related programs (BGF90).

Nani JP, Peñagaricano F (2020) Whole-genome homozygosity mapping reveals candidate regions affecting bull fertility in US Holstein cattle. BMC Genomics 21. https://doi.org/10.1186/s12864-020-6758-y

Nascimento AV, Cardoso DF, Santos DJA, et al (2021) Inbreeding coefficients and runs of homozygosity islands in Brazilian water buffalo. J Dairy Sci 104:1917–1927. https://doi.org/10.3168/jds.2020-18397

Nascimento Rangel AH do, Araújo TPM, da Costa Lima GF, et al (2018) Estimativa de tendências genéticas e fenotípicas para características leiteiras de animais das raças gir e guzerá. Acta Sci - Anim Sci 40. https://doi.org/10.4025/actascianimsci.v40i1.36685

NCBI - National Center for Biotechnology Information (2022) Accessed 15 Feb 2022. https://www.ncbi.nlm.nih.gov/.

Neves HHR, Desidério JA, Pimentel ECG, et al (2015) Preliminary study to determine extent of linkage disequilibrium and estimates of autozygosity in Brazilian Gyr dairy cattle. Arch Zootec 64:99–108. https://doi.org/10.21071/az.v64i246.383

Nosrati M, Asadollahpour Nanaei H, Javanmard A, Esmailizadeh A (2021) The pattern of runs of homozygosity and genomic inbreeding in world-wide sheep populations. Genomics 113:1407–1415. https://doi.org/10.1016/j.ygeno.2021.03.005

Otto PI, Guimarães SEF, Verardo LL, et al (2019) Genome-wide association studies for heat stress response in Bos taurus × Bos indicus crossbred cattle. J Dairy Sci 102:8148–8158. https://doi.org/10.3168/jds.2018-15305

Parker Gaddis KL, Dikmen S, Null DJ, et al (2017) Evaluation of genetic components in traits related to superovulation, in vitro fertilization, and embryo transfer in Holstein cattle. J Dairy Sci 100:2877–2891. https://doi.org/10.3168/jds.2016-11907

Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A Critical Assessment of Storytelling:

Gene Ontology Categories and the Importance of Validating Genomic Scans. Mol Biol Evol 29:3237–3248. https://doi.org/10.1093/molbev/mss136

Pereira RJ, Ayres DR, Santana Junior ML, et al (2019) Test-day or 305-day milk yield for genetic evaluation of Gir cattle. Pesqui Agropecuária Bras 54:e00325. https://doi.org/10.1590/s1678-3921.pab2019.v54.00325

Peripolli E, Stafuzza NB, Munari DP, et al (2018) Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. BMC Genomics 19. https://doi.org/10.1186/s12864-017-4365-3

Purfield DC, Berry DP, McParland S, Bradley DG (2012) Runs of homozygosity and population history in cattle. BMC Genet 13. https://doi.org/10.1186/1471-2156-13-70

Qanbari S, Simianer H (2014) Mapping signatures of positive selection in the genome of livestock. Livest Sci 166:133–143. https://doi.org/10.1016/j.livsci.2014.05.003

R Core Team.´R: A language and environment for statistical computing´ (2022) R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rabahi F, Brûlé S, Sirois J, et al (1999) High expression of bovine α glutathione S-transferase (GSTA1, GSTA2) subunits is mainly associated with steroidogenically active cells and regulated by gonadotropins in bovine ovarian follicles. Endocrinology 140:3507–3517. https://doi.org/10.1210/endo.140.8.6886

Randhawa IAS, Khatkar MS, Thomson PC, Raadsma HW (2016) A MetaAssembly of Selection Signatures in Cattle. PLoS One 11:e0153013. https://doi.org/10.1371/journal.pone.0153013

Rebelato AB, Caetano AR (2018) Runs of homozygosity for autozygosity estimation and genomic analysis in production animals. Pesqui Agropecuária Bras 53:975–984. https://doi.org/10.1590/s0100-204x2018000900001

Ribeiro ES, Gomes G, Greco LF, et al (2016) Carryover effect of postpartum inflammatory diseases on developmental biology and fertility in lactating dairy cows. J Dairy Sci 99:2201–2220. https://doi.org/10.3168/jds.2015-10337

Rocha RFB, Otto PI, da Silva MVGB, et al (2022) Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed. Anim Prod Sci 62:1661-1670 https://doi.org/10.1071/AN21588

Rosen BD, Bickhart DM, Schnabel RD, et al (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9:giaa021. https://doi.org/10.1093/gigascience/giaa021

Rotar LN, Souza JF (2019) Morphological Characteristics for Oocyte-Cumulus Complexes of

Bos taurus and Bos indicus Cattle Used for a Variety of Purposes. Russ Agric Sci 45:399–402. https://doi.org/10.3103/S1068367419040141

Salhab M, Dhorne-Pollet S, Auclair S, et al (2013) In vitro maturation of oocytes alters gene expression and signaling pathways in bovine cumulus cells. Mol Reprod Dev 80:166–182. https://doi.org/10.1002/mrd.22148

Santiago AA (1986) O Zebu na Índia, no Brasil e no mundo. Instituto Campineiro de ensino agrícola: Campinas, São Paulo.

Shannon P, Markiel A, Ozier O, et al (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303

Shaw L, Sneddon SF, Zeef L, et al (2013) Global Gene Expression Profiling of Individual Human Oocytes and Embryos Demonstrates Heterogeneity in Early Development. PLoS One 8:e64192. https://doi.org/10.1371/journal.pone.0064192

Shi L, Wang L, Liu J, et al (2020) Estimation of inbreeding and identification of regions under heavy selection based on runs of homozygosity in a Large White pig population. J Anim Sci Biotechnol 11. https://doi.org/10.1186/s40104-020-00447-0

Šidlová V, Kasarda R, Moravčíková N, et al (2015) Genomic variability among cattle populations based on runs of homozygosity. Poljoprivreda 21:44–47. https://doi.org/10.18047/poljo.21.1.sup.9

Srikanth K, Lee SH, Chung KY, et al (2020) A gene-set enrichment and protein–protein interaction network-based gwas with regulatory snps identifies candidate genes and pathways associated with carcass traits in hanwoo cattle. Genes 11:316. https://doi.org/10.3390/genes11030316

Stelzer G, Plaschkes I, Oz-Levi D, et al (2016) VarElect: the phenotype-based variation prioritizer of the GeneCards Suite. BMC Genomics 17:195–206. https://doi.org/10.1186/s12864-016-2722-2

Stephens MA (1986) Tests based on EDF statistics. In: Stephens, MA (1974) EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 69: 730-737. https://doi.org/10.1080/01621459.1974.10480196

Suárez-Vega A, Arranz JJ, Pérez V, et al (2018) Early adipose deposits in sheep: comparative analysis of the perirenal fat transcriptome of Assaf and Churra suckling lambs. Anim Genet 49:605–617. https://doi.org/10.1111/age.12725

Talebi R, Szmatoła T, Mészáros G, Qanbari S (2020) Runs of Homozygosity in Modern Chicken Revealed by Sequence Data. G3 Genes|Genomes|Genetics 10:4615–4623.

https://doi.org/10.1534/g3.120.401860

TFM-Explorer Transcription Factor Matrix Explorer – Bonsai Bioinformatics (2022) Accessed 17 Feb (2022) https://bioinfo.lifl.fr/tfm-explorer/tfm-explorer.php.

Thode HC (2002) Testing For Normality. Marcel Dekker, New York, USA

Toro Ospina AM, Silva Faria RA, Vercesi Filho AE, et al (2020) Genome-wide identification of runs of homozygosity islands in the Gyr breed (*Bos indicus*). Reprod Domest Anim 55:333–342. https://doi.org/10.1111/rda.13639

Wang H, Lapek J, Fujimura K, et al (2018) Pseudopodium-enriched atypical kinase 1 mediates angiogenesis by modulating GATA2-dependent VEGFR2 transcription. Cell Discov 4. https://doi.org/10.1038/s41421-018-0024-3

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. Evolution 38:1358–1370.

Wiggans GR, VanRaden PM, Cooper TA (2011) The genomic evaluation system in the United States: Past, present, future. J Dairy Sci 94:3202–3211. https://doi.org/10.3168/jds.2010-3866

Wright S (1978) Variability Within and Among Natural Populations. In: Wright S (ed) Evolution and the Genetics of Populations, 4th vol. University Chicago Press, Chicago, USA.

Zhao F, McParland S, Kearney F, et al (2015) Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet Sel Evol 47. https://doi.org/10.1186/s12711-015-0127-3

Zhu G, Fang C, Li J, et al (2019) Transcriptomic Diversification of Granulosa Cells during Follicular Development in Chicken. Sci Rep 9. https://doi.org/10.1038/s41598-019-41132-1

Single-step genome-wide association studies and post-GWAS analyses for the number of oocytes and embryos in Gir cattle*

*Published at Mammalian Genome (https://doi.org/10.1007/s00335-023-10009-0)

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

[1]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, MG, 36570-900, Brazil, e-mail: renata.bretanha@ufv.br, arielly.garcia@ufv.br, mateus.g.santos@ufv.br, sfacioni@ufv.br

[2]Department of Animal Science, Universidade Federal de Santa Maria, Santa Maria, RS, 97105-900, Brazil, e-mail: pamela.otto@ufsm.br

[3]EMBRAPA - Dairy Cattle, Juiz de Fora, MG, 36038-330, Brazil, e-mail: joao.panetto@embrapa.br, marcos.vb.silva@embrapa.br, marta.martins@embrapa.br, marco.machado@embrapa.br

[*]Corresponding author: Simone Eliza Facioni Guimarães, sfacioni@ufv.br

ORCID

Renata de Fátima Bretanha Rocha https://orcid.org/0000-0002-0552-7042

Arielly Oliveira Garcia https://orcid.org/0000-0001-6959-8117

Pamela Itajara Otto https://orcid.org/0000-0003-1702-4626

Mateus Guimarães dos Santos https://orcid.org/0000-0002-6475-9951

Marcos Vinícius Barbosa da Silva https://orcid.org/0000-0001-5449-1413

Marta Fonseca Martins https://orcid.org/0000-0002-4260-5329

Simone Eliza Facioni Guimarães https://orcid.org/0000-0003-3704-8131

João Claudio do Carmo Panetto https://orcid.org/0000-0002-9198-9728

Abstract

Genome-Wide Association Studies (GWAS) are used for identification of quantitate trait loci (QTL) and genes associated with several traits. We aimed to identify genomic regions, genes, and biological processes associated with number of total and viable oocytes, and number of embryos in Gir dairy cattle. A dataset with 17,526 follicular aspirations, including the following traits: number of viable oocytes (VO), number of total oocytes (TO) and number of

embryos (EMBR) from 1,641 Gir donors was provided by five different stock farms. A genotype file with 2,093 animals and 395,524 SNP markers was used to perform a single-step GWAS analysis for each trait. The top 10 windows with the highest percentage of additive genetic variance explained by 100 adjacent SNPs were selected. The genomic regions identified in our work were overlapped with QTLs from QTL database on chromosomes 1, 2, 5, 6, 7, 8, 9, 13, 17, 18, 20, 21, 22, 24 and 29. These QTLs were classified as External, Health, Meat and carcass, Production or Reproduction traits, and about 38% were related to Reproduction. In total, 117 genes were identified, of which 111 were protein-coding genes. Exclusively associations were observed for 42 genes with EMBR, and 1 with TO. Also, 42 genes were in common between VO and TO, 28 between VO and EMBR and four genes were in common among all traits. In conclusion, great part of the identified genes plays a functional role in initial embryo development or general cell functions. The protein coding genes *ARNT*, *EGR1*, *HIF1A*, *AHR* and *PAX2* are good markers for the production of oocytes and embryos in Gir cattle.

Keywords: embryo in vitro production, genes, QTL, reproduction, zebu cattle

Introduction

Given the importance of Gir cattle breed in Brazil, the National Program for the Improvement of Dairy Gir (in Portuguese: Programa Nacional de Melhoramento do Gir Leiteiro - PNMGL) was created and has been conducted by Brazilian Agricultural Research Corporation (EMBRAPA) – Center for Research in Dairy Cattle and Brazilian Association of Dairy Gir Breeders (in Portuguese: Associação Brasileira dos Criadores de Gir Leiteiro – ABCGIL) for almost 40 years (Panetto et al. 2021). Ever since, several works have been carried out to explore Gir breed productive capacity and contribute with PNMGL and literature information (Carvalho et al. 2021; González-Herrera et al. 2022; do Nascimento Rangel et al. 2018; Pereira et al. 2019). Previous works using the same Gir population used in this work explored genetic parameters, runs of homozygosity and signatures of selection for reproductive traits in this breed (Rocha et al. 2022; Rocha et al; 2023). However, references exploring genomic information about reproductive traits in Gir cattle are scarce. So far, the genomic investigation in Gir breed showed that quantitative traits associated to fertility, milk production, beef quality, and growth were involved in the differentiation process of dual purpose populations (Maiorano et al. 2018). Despite reproductive traits usually present only a small proportion of

phenotypic variance due to additive actions of individual genes, it does not mean that reproductive traits are not under genetic control (Ortega 2018).

An example of a reproductive technology that has been increasing in cattle production in the past years, is in vitro embryo production (IVP), which combines oocyte collection, in vitro fertilization and embryo transfer (Moore and Hasler 2017). Brazil ranks second in embryo production, and is the leader in IVP embryo transfers worldwide (Viana 2022). About a decade ago, it was already known that the IVP technology stood out among the reproductive traits due to its economic impact, since it accelerates genetic improvement, reduces the generation interval and promotes an increase in the number of products/cow/year, as it allows the use of pre-pubescent heifers, cows in early pregnancy, cows with acquired subfertility and senile cows (Martins 2010). Currently, a milestone was reached for this technology, since for the first time in 2021, over one million IVP cattle embryos were transferred worldwide (Viana 2022). This shows the importance of studying oocyte and embryo production traits in cattle.

The Genome-Wide Association Studies (**GWAS**) are used to find genomic regions that contribute to genetic variation of a trait, and most reproductive traits are controlled by many genes. As a *Bos taurus indicus*, the Gir breed has been highlighted in the use of reproductive biotechnologies, such as higher oocyte quality when compared to *Bos taurus taurus*, which allows greater success of in vitro embryo production than taurine animals (Drum et al. 2019). In this way, the identification of genetic variants for such traits through GWAS can provide clues to understand fertility regulation and genetic architecture in Gir cattle (Ortega 2018; Ma et al. 2019). Usually, GWAS for fertility traits in dairy cattle results in several genomic regions identified due to the pleiotropic effect of genes affecting these polygenic traits (Jaton et al. 2018; Parker Gaddis et al. 2016; Parker Gaddis et al. 2017; Sbardella 2021). Ma et al. (2019) suggested that the continuous data collection of fertility traits would enable more powerful estimations of dairy fertility. We hypothesize that genomic regions associated with number of oocytes and embryos can be identified in Gir. Considering the most importance of Gir breed for dairy production in Brazil and that the studied traits have high economic impact in livestock production accelerating genetic improvement, we aimed to identify genomic regions, genes, and biological processes associated with number of total and viable oocytes, and number of embryos in Gir dairy cattle.

Material and Methods

Phenotypic Data

A dataset with 17,526 follicular aspirations, including the following traits: number of viable oocytes (VO), number of total oocytes (TO) and number of embryos (EMBR) from 1,641 Gir donors was provided by five farms. These donors are daughters of 127 sires and 771 dams. This dataset along with a pedigree file, including 4,679 animals, covering up to fifteen generations were used to estimate genetic values for the animals in these three traits previously to this work (Rocha et al. 2022). The Ovum Pick Up (OPU) sessions were conducted between January 2005 and June 2020 at the Minas Gerais State in Brazil, where the farms are located. Eight companies responsible for OPU and *in vitro* fertilisation (IVF) processes were listed in the dataset received from farms, each company using their own protocols; in this way, we considered eight protocols in total for the analysis, varying among farms or sometimes within farms. The number of OPU sessions ranged from 1 to 70 per donor, with an average of 10. A total of 238 bulls were used for the in vitro fertilization processes. The frequency of use of these bulls ranged from 1 to 1789 samples, with an average of 73 samples per bull. A brief description of the data can be seen in Table 1.

Table 3. Descriptive statistics[A] on aspiration interval, donors' age and structures[B]

Traits	N	Total	Mean	SD	MIN[C]	MAX	CV
Aspiration interval (days)	17,526	--	97.01	171.10	0	2570	176.37
Ovum Pick Up Age (years)	17,526	--	6.27	3.19	1	16	50.83
Viable oocytes	17,524	277,292	15.82	12.89	0	182	81.46
Total oocytes	17,526	379,599	21.66	16.10	0	197	74.33
Embryos	17,498	51,506	2.94	3.70	0	46	125.66

[A]N: number of data samples; Total: total sum of structures (oocytes and embryos); SD: standard deviation; MIN: Minimum; MAX: Maximum; CV: coefficient of variation.

[B]Variable not log transformed: LN (X +1).

[C]Minimum for aspiration interval means two aspirations on the same cow on the same day with semen from different bulls. Source: Rocha et al. (2022).

Genotyping and Quality Control

The genotype file with 2,093 animals and 420,718 SNP markers was previously described by Rocha et al. (2023). This premade genotyped file was provided by Brazilian Agricultural Research Corporation (EMBRAPA) – Center for Research in Dairy Cattle, and it is a result of animals genotyped with different density SNP panels, including Illumina BovineSNP50 BeadChip v2 (50K), Illumina BovineHD BeadChip (777K), GGP Indicus (34K), Z-Chip

(30K) and GGP Indicus (50K). Finally, all were extracted and/or imputed for 777K, resulting in 420,718 SNPs. EMBRAPA carried out the imputation process and provided the genomic information of 2,093 animals that were present in the pedigree file used in this research with 4,679 animals. The remaining animals of the pedigree did not have genomic information.

Genome Wide Association Study analyses

From the dataset with 2,093 animals, 170 were males and 1,923 were females, of which 1,202 females had phenotypic data. For GWAS analysis, quality control was performed for this dataset with 420,718 SNPs, considering the following parameters for variant exclusion: SNP with minor allele frequency ≤ 0.05, maximum difference between observed and expected allele frequency for Hardy Weinberg Equilibrium of 0.15, parent-progeny conflict and SNP heritability (h^2), in which markers with estimated h^2<0.98 were discarded. After quality control analysis, 395,524 SNPs and 2,091 genotyped animals remained for further analysis.

The single-step GWAS was conducted using the BLUPF90 software family (Misztal et al. 2002). The variance components were previously obtained (Rocha et al. 2022). Briefly, animals were separated into 138 contemporary groups (CG) composed according to OPU year and season, farm and OPU-IVF protocol, considering a minimum of three animals in each CG; lower numbers implicated in data exclusion. Cows without a minimum of three observations were excluded. As the traits did not present normal distribution, they were transformed using the logarithmic scale: ln (X + 1), where ln is the natural logarithm and X is the raw data (oocyte and embryos), verified by the Anderson-Darling test (Stephens 1986; Thode 2002). The following matrix model was used for all traits:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Wp} + \boldsymbol{e}$$

where, $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{a}$, $\boldsymbol{p}$, and $\boldsymbol{e}$ are the vectors of observations; fixed effects; additive genetic random effects; permanent environment random effects and residual effects, respectively; $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are the incidence matrices of fixed, additive genetic and permanent environment, respectively. The effects of animal, permanent environment and residual were assumed to be random. In the single-trait model, it was assumed that $\boldsymbol{a} \sim N(0, \boldsymbol{H}\sigma_a^2)$, $\boldsymbol{p} \sim N(0, \boldsymbol{I}\sigma_p^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_a^2, σ_p^2 and σ_e^2 are the additive genetic, permanent environmental and residual variances, respectively; **H** is a matrix combining genomic and pedigree information, as proposed by Aguilar et al. (2010) and **I** is the identity matrix. The statistical model applied for VO and TO can be described as follows:

$$Y_{ijkl} = \mu + CG_i + Age_j + Age^2{}_k + AI_l + e_{ijklm}$$

and for EMBR:

$$Y_{ijklm} = \mu + CG_i + Age_j + Age^2{}_k + B_l + BB_m + e_{ijklmn}$$

where $\boldsymbol{Y}$ is the dependent variable, $\boldsymbol{\mu}$ is the mean, $\boldsymbol{CG_i}$ is the fixed effect of contemporary group; $\boldsymbol{Age_j}$ and $\boldsymbol{Age^2{}_k}$ are the donor's age when OPU was carried out and respective squared age (covariates); $\boldsymbol{AI_l}$ is the aspiration interval (covariate); $\boldsymbol{B_l}$ is the fixed effects of the bull used in the in vitro fertilization process; $\boldsymbol{BB_m}$ is the effect of bull breed (Gir, n = 35; Holstein, n = 92); and $\boldsymbol{e_{ijklmn}}$ is the residual assumed to be independent and to have a normal distribution.

The SNP effects were calculated using postGSf90 software (http://nce.ads.uga.edu/; Aguilar et al. 2010), and the GWAS results were reported as the percentage of genetic variance explained by a window of 100 adjacent SNPs. Manhattan plots with the percentage of genetic variance were created using 'qqman' package (Turner 2018) in R software (R Core Team. 2022, R version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria).

QTL regions, Gene Search and Gene Ontology analysis

The top 10 windows explaining the highest percentage of additive genetic variance (Otto et al. 2020; Tiezzi et al. 2015; Vargas et al. 2018) were selected and overlapped with Quantitative Trait Loci (QTL) regions previously described in Animal QTL database (https://www.animalgenome.org/QTLdb). Within top 10 windows, putative candidate genes were identified based on initial and final coordinates of each selected window on the ARS-UCD1.2 assembly of the bovine genome (Rosen et al. 2020), using the National Center for Biotechnology Information (NCBI 2022) Map Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/?org=bos-taurus). R software (R Core Team. 2022, R version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria) was used to perform QTL and candidate gene search.

A Venn diagram was built to show the relation of all genes identified among all three traits using 'VennDiagram' (Chen 2022) and 'RColorBrewer' packages (Neuwirth 2022) in R software (R Core Team. 2022, R version 4.1.2; R Foundation for Statistical Computing, Vienna, Austria). All 111 protein coding genes identified were used to perform the biological processes enrichment analysis to identify the most likely candidate genes for the number of total and viable oocytes and number of embryos. This analysis was performed using the ClueGO application (Bindea et al. 2009) in Cytoscape (Shannon et al. 2003), considering a Bonferroni correction test. For the identification of enriched transcriptional factors (TF), promoter sequences (FASTA format) with 3000 bp upstream and 300 bp downstream from genes transcription start site (Verardo et al. 2016) were submitted to the Transcription Factor

Matrix Explorer (TFM-explorer 2022, http://bioinfo.lifl.fr/TFM/TFME/). To determine significantly overrepresented functional GO terms, TF list was analyzed in Cytoscape software, using the Biological Networks Gene Ontology tool (BiNGO; Maere et al. 2005). The most likely candidate genes related to the studied traits were identified through gene-TF Network Analyzer tool in Cytoscape.

In addition, all annotated genes were further investigated through VarElect NGS Phenotyper (Stelzer et al. 2016). The genes found for the number of embryos and the number of oocytes (total and viable) were investigated based on the keywords "embryo" and "oocyte", respectively. Also, all transcriptional factors were investigated based on both keywords. VarElect tool (https://varelect.genecards.org/) has been used in literature to prioritize variant genes based on disease/phenotype of interest (Gutierrez-Quintana et al. 2021; Rocha et al. 2023; Suárez-Vega et al. 2018). To do such inferences, VarElect uses algorithms seeking for association between genes and phenotypes based on shared pathways, interaction networks, relation of paralog genes and other sources. VarElect tool gives a score indicating the strength of the connection between the gene and the traits, ranging from 1 to 200. The ratio of each score is calculated and total scores are normalized to a range of 0.05 - 0.8.

Results

The percentage of genetic variance explained by each window of 100 adjacent SNPs are displayed in Manhattan plots for number of viable oocytes (Figure 1), number of total oocytes (Figure 2) and number of embryos (Figure 3). The additive genetic variance explained by the 10 windows was 3.30% for VO and 3.54% for TO and EMBR.

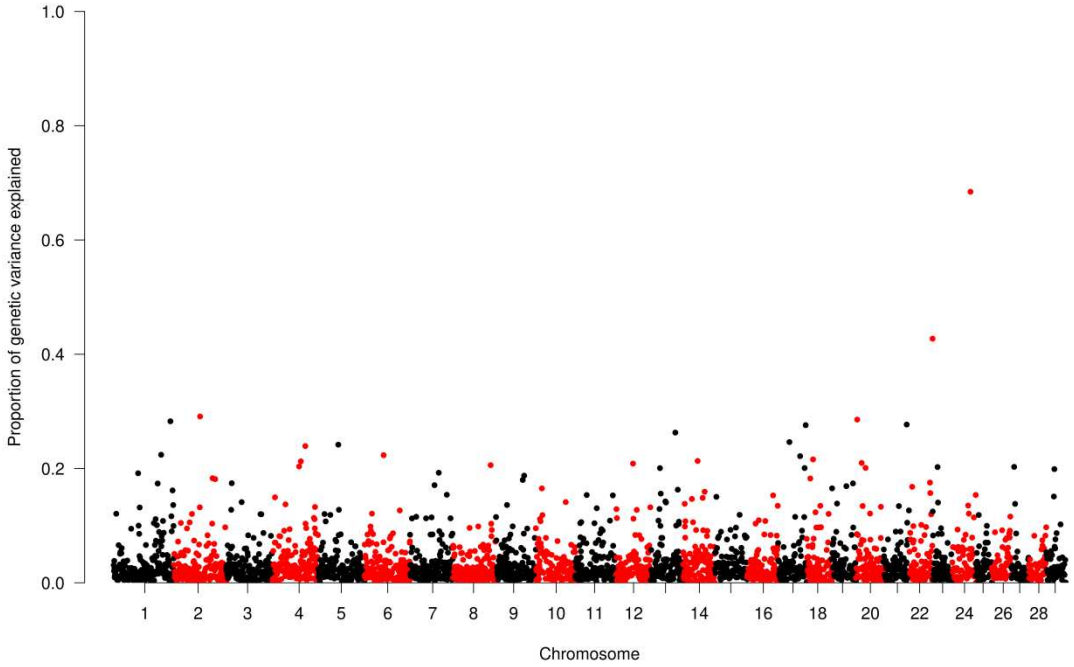


Fig. 1 Manhattan plot of proportion of additive genetic variance (%) explained by 100 adjacent SNPs in Gir dairy cattle evaluated for number of viable oocytes. Each dot denotes a window

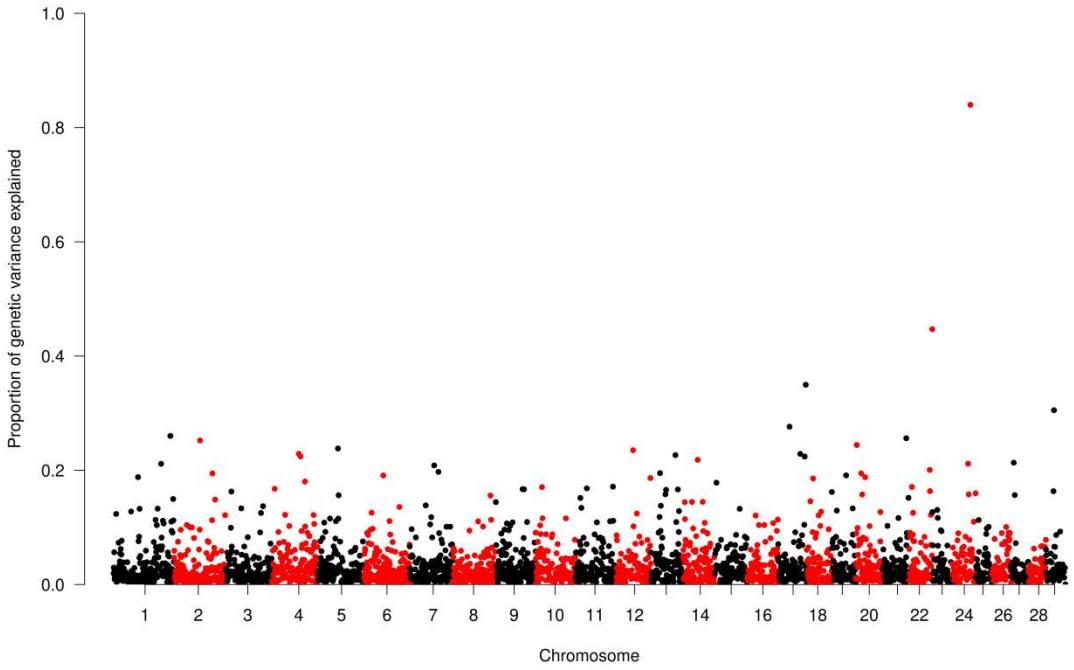


Fig. 2 Manhattan plot of proportion of additive genetic variance (%) explained by 100 adjacent SNPs in the Gir dairy cattle evaluated for the number of total oocytes. Each dot denotes a window

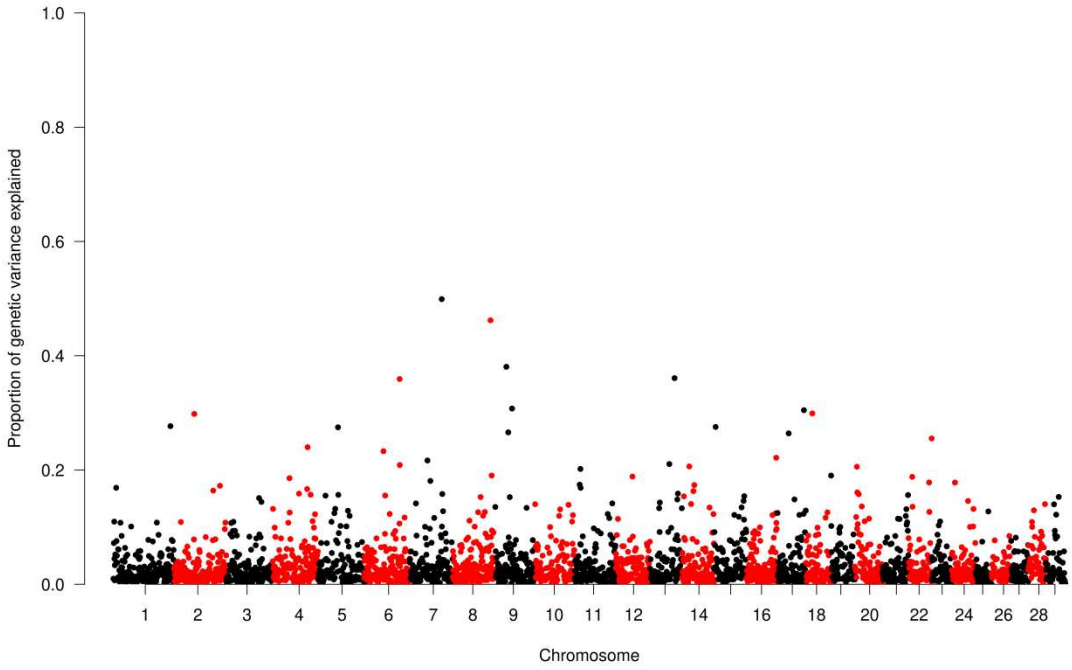


Fig. 3 Manhattan plot of proportion of additive genetic variance (%) explained by 100 adjacent SNPs in the Gir dairy cattle evaluated for the number of embryos. Each dot denotes a window

The genomic regions from the top 10 windows identified on chromosomes 1, 2, 5, 6, 7, 8, 9, 13, 17, 18, 20, 21, 22, 24 and 29 were overlapped with Quantitative Trait Loci from QTL database (Figure 4; Online Resource ESM1; sheet 1). From QTL database, these QTLs were classified as External, Health, Meat and carcass, Production or Reproduction traits. None of them were directly related to oocyte or embryo. However, about 38% of QTLs found are related to Reproduction traits, such as sperm counts, stillbirth, interval to first estrus after calving, birth index, gestation length, calving ease, daughter pregnancy rate, inseminations per conception, age at first calving, non-return rate, scrotal circumference, inhibin level, first service conception, perinatal mortality, temperament, interval from first to last insemination, sperm motility and conception rate.

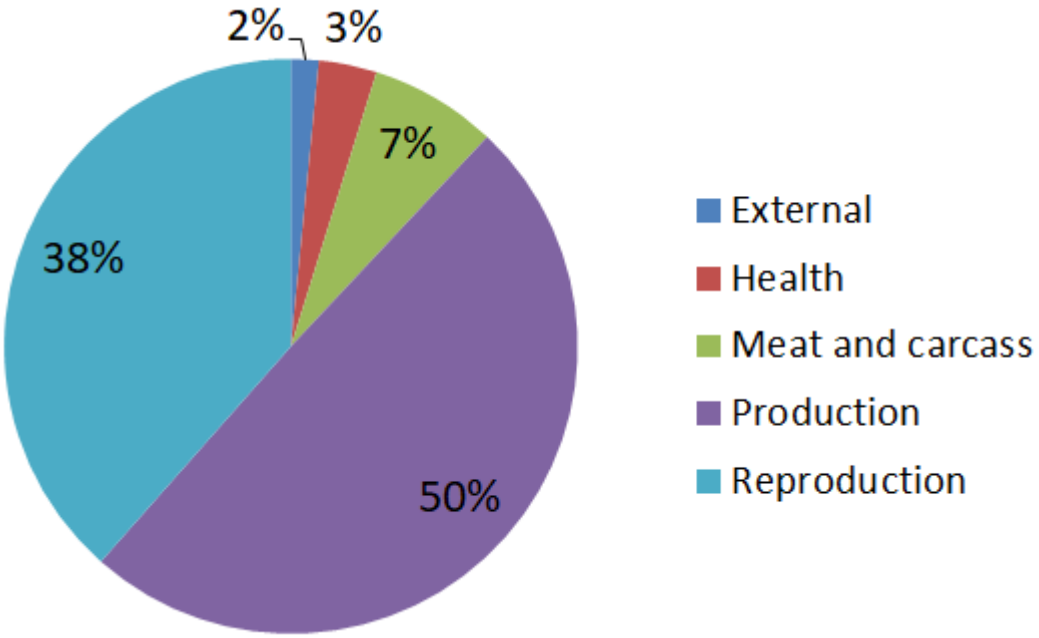


Fig. 4 Pie plot for the Quantitative Trait Loci from QTL database overlapped with the genomic regions identified in the top 10 windows

In addition, 117 genes were identified, two of which were open reading frames (ORFs), four were microRNAs (*MIR130B*, *MIR301B*, *MIR499*, *MIR12063*) and 111 were protein coding genes. From these, 42 genes were exclusively associated with EMBR, 1 with TO and no gene was exclusively associated with VO, 42 genes were in common between VO and TO, 28 between VO and EMBR and four genes were in common among all traits, which were *CHAF1B*, *CLDN14*, *DOP1B*, and *MORC3* (Figure 5; Online Resource ESM1; sheet 2).

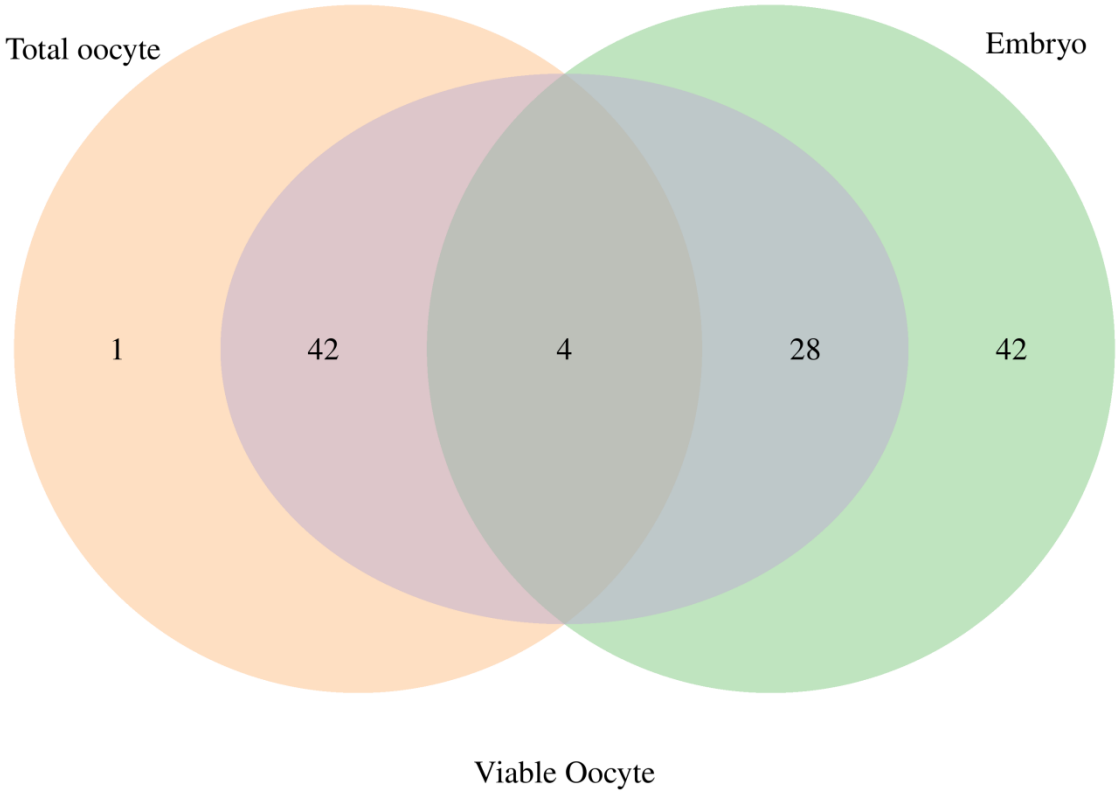


Fig. 5 Venn Diagram of the number of genes found for total oocytes, viable oocytes and embryos

From ClueGO application, 857 biological processes were found related to the 111 protein coding genes (Online Resource ESM1; sheet 3). Great part of the identified genes plays a functional role in initial embryo development or general cell functions. Nine significant biological processes according to Bonferroni correction test ($P < 0.05$) were related to cellular catabolic processes. Considering the promoter sequences of our 111 protein coding genes, 16 transcription factors (TFs) were identified (Table 1). These transcription factors were analyzed in Cytoscape software, using BiNGO application to identify biological networks gene ontology (Online Resource ESM1; sheet 4). From this, five TFs connected to biological networks most related to the studied traits, number of total and viable oocytes and number of embryos, were identified: *ARNT*, *EGR1*, *HIF1A*, *AHR* and *PAX2*. These TFs were reviewed in literature to verify their function related to oocytes or embryos. After that, 81 candidate genes associated with these five TFs (Figure 6) were identified and reviewed in literature to relate their functions to the studied traits.

From VarElect results, 19 of 71 genes were directly related to "oocyte" keyword: *UBTD2*, *HIC2*, *UBE2L3*, *NCOA6*, *SLC7A4*, *LARP1B*, *PLK4*, *CRKL*, *CHAF1B*, *P2RX6*,

STK10, *FBXW11*, *SMAD7*, *SNAP29*, *CTIF*, *TUBA3E*, *HSPA4L*, *ITCH* and *SLC25A31*, in this order from the highest to the lowest score. From 70 genes, 39 were directly related to "embryo" keyword: *GJA1*, *GDF5*, *BVES*, *EDIL3*, *LIN28B*, *PTPN3*, *HAPLN1*, *VCAN*, *TCN2*, *INPP5J*, *PROCR*, *NCOA6*, *DOP1B*, *SIAH1*, *PES1*, *MAP1LC3A*, *RBM12*, *MORC2*, *MORC3*, *TBC1D32*, *POPDC3*, *ITCH*, *GSS*, *MMP24*, *UQCC1*, *EIF6*, *ABCC11*, *SMTN*, *SEC14L6*, *HACE1*, *PREP*, *EPB41L4B*, *N4BP1*, *RBM39*, *ACSS2*, *DYNLRB1*, *SEC14L2*, *SF3A1* and *SEC14L3*, in this order from the highest to the lowest score. From all transcriptional factors, 11 were directly related to both terms in this score order: *EGR1*, *PAX2*, *HIF1A*, *YY1*, *KLF4*, *REL*, *RELA*, *AHR*, *NF-KAPPAB*, *ARNT*, and *MAX*, while four were directly related exclusively with "embryo" keyword: *MAFB*, *INSM1*, *FEV* and *EBF1*.

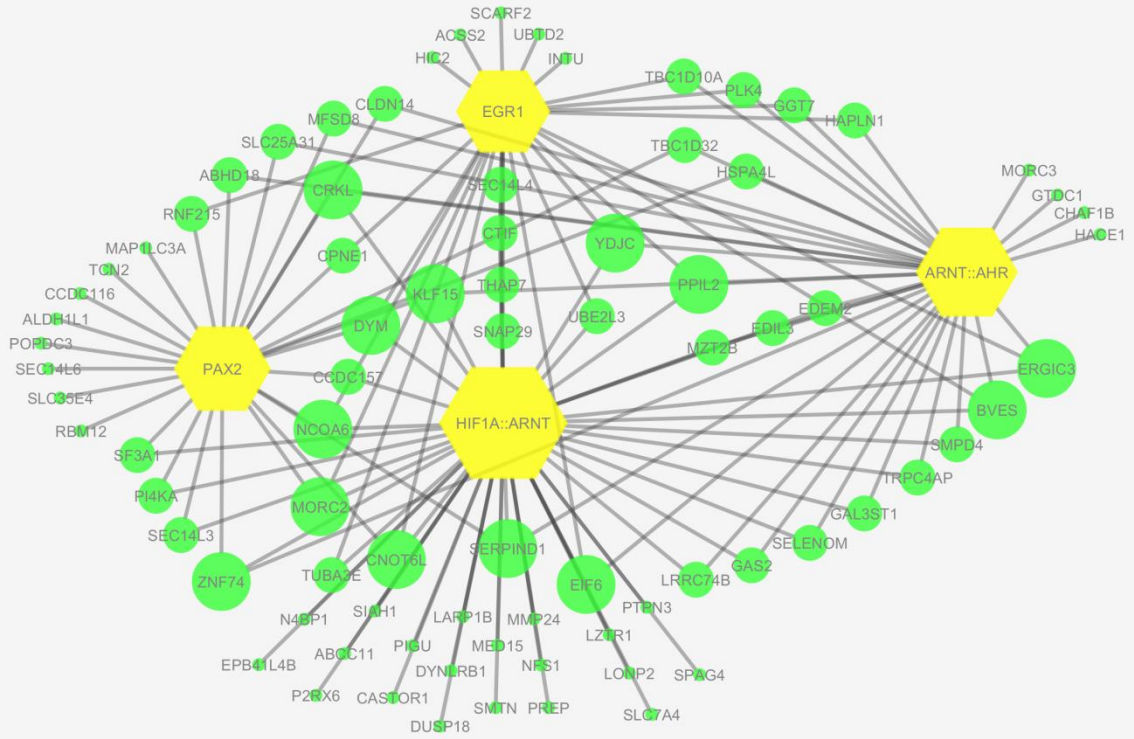


Fig. 6 Gene-transcription factor (TF) network: Genes located in the top 10 windows for the number of total and viable oocytes and number of embryos (green circle nodes) and their associated TF (yellow hexagon nodes). Node size corresponds to network analyses (Cytoscape; Shannon et al., 2003), in which larger nodes denotes a higher edge density associated with the number of TF binding sites

Discussion

In our work, we found Quantitative Trait Loci (QTL) overlapped with the top 10 windows with the highest percentage of additive genetic variance explained for the number of viable

oocytes, number of total oocytes, and number of embryos, which were related to different types of traits. As we studied polygenic traits, that is, traits affected by many low effect genes, it was expected to find several genomic regions and that these regions may affect more than one trait. Focusing on reproductive-related traits, some QTLs were related to interval to age at first calving, birth index, daughter pregnancy rate, first estrus after calving, daughter pregnancy rate, inseminations per conception, interval from first to last insemination depend on estrus detection, and a correlation between these traits with oocytes and embryos was searched in literature. Perez et al. (2016) found low correlation between age at first calving with number of viable oocytes (-0.003) and number of transferable embryos (-0.006) in Guzerá cattle breed, which can be understood from the physiological point of view in which the reproductive system of the animal is still in development. Cumulative live birth rates were associated to a higher number of oocytes retrieved (Fanton et al. 2023). A low positive correlation (0.12) was found between the number of retrieved oocytes and pregnancy rates (Hsu et al. 2016). Usually, estrus detection depends, among other factors, on the number of ovulation after calving, and it is the first step in getting a cow pregnant (Roelofs and van Erp-van der Kooij 2015). In addition to female fertility performance, male fertility traits are important for conception and then, embryo production. In this case, sperms per count and sperm motility represent some male reproductive traits found in QTL search. Such traits may be determinant of conception success or failure (Berry et al. 2019).

Searching deeper on QTL regions, identification of genomic variants and genes associated with reproductive traits can improve the knowledge about fertility regulation (Ortega 2018). In our work, genes were found on several chromosomes and, at first, focus will be given to interesting results found for the four genes in common among all three traits or genes exclusively associated with one trait. *CHAF1B* gene, for example, was associated with all three traits: number of total oocytes, viable oocytes and embryos and it was directly related to the “oocyte” term in VarElect. From its related biological processes obtained by ClueGO analyses, *CHAF1B* plays a role in nucleosome assembly and organization. Another gene, *MORC3*, also associated with all three traits and directly related to the “oocyte” term in VarElect. *MORC3* is related to cell aging and proliferation, protein localization and stability and post-embryonic development biological processes. In initial embryo development there is a transition between maternal and onset of embryo gene expression know as zygotic/embryonic genome activation, that is, when maternal mRNAs are degraded and zygotic transcription begins (Jukam et al. 2017). Both *CHAF1B* and *MORC3* were mentioned by Graf et al. (2014) to be expressed as primary transcripts with intronic sequences

determining the onset of embryonic gene expression. Biological processes related with *CHAF1B* and *MORC3* genes were mentioned as important functions during embryonic genome activation, such as nucleosome assembly and protein localization as consequences of embryo cell division (Jukam et al. 2017). Given the importance in this transition to maternal embryo expression, it explains why *CHAF1B* and *MORC3* genes were associated with the number of oocytes and embryos in our work.

On the other hand, *GAS2* is a gene exclusively associated with the number of total oocytes. Also, *GAS2* is related to several biological processes in our work, including response to extracellular stimulus, response to nutrient levels, regulation of microtubule-based process, macromolecular complex disassembly and negative regulation of hematopoiesis. These processes are related to programmed cell death or apoptosis, since apoptotic signaling begins with the attachment of extracellular ligands causing a variety of biological events such as proliferation/homeostasis, development, differentiation and elimination of harmful cells (Goldar et al. 2015). On GeneCards (https://www.genecards.org/), an integrative database that provides information on all annotated and predicted human genes, it is mentioned that *GAS2* may play a role in apoptosis by acting as a cell death substrate for caspases. Indeed it makes sense that gene expression related to apoptosis would be observed in total oocytes, since not all oocytes obtained in follicular aspiration maintain quality integrity to be used for in vitro fertilization, hence not being observed in viable oocytes. It is also acceptable that an apoptosis related-gene would not be associated with number of embryos since embryo development requires biological processes related to cell development and maintenance, not cell death.

In the search for genomic regions associated with oocytes and embryos, Jaton et al. (2018) found significant SNPs on chromosome 11 with pleiotropic effect on the number of embryos and viable embryos in Holstein cattle. In our work, we did not find high proportion of additive genetic variance on BTA11, but we found common genes between the number of total and viable oocytes and the number of embryos, which means that those genes may have a pleiotropic effect on the studied traits. This is consistent to our previous study, where we found high positive genetic correlation between the number of total and viable oocytes and the number of embryos (Rocha et al. 2022). Still in Holstein cattle, Parker Gaddis et al. (2017) found large proportion of variance for number of good quality (grade 1) embryos on BTA14 and for total structures recovered (i.e., total number of unfertilized oocytes and embryos) on BTA8, followed by BTA5, BTA10 and BTA13. In our work, the largest proportion of additive genetic variance for the number of total and viable oocytes was on BTA24 and for embryos on BTA7. This is a good representation of our previous work where a high positive

genetic correlation (1.00) was identified for the number of viable oocytes and the number of total oocytes (Rocha et al. 2022). This means that genes on BTA24 may play important role in cell viability for maintenance of viable oocytes. Also, our results indicate that genes present on chromosome 7 may have a greater influence in embryo developmental process.

In dairy Gir cattle, Pereira (2019) found windows with higher genetic variance explained (> 1%) for the number of viable oocytes on chromosomes 1, 2, 3, 4, 5, 8, 13, 14, 17, e 27 and for the number of viable embryos on chromosomes 2, 6, 9, 14, 18 e 25. Our results corroborate in part with Pereira (2019), such that in our results chromosomes 1, 2, 5, 6, 8, 9, 13, 17 and 18 were present in the windows with the higher proportion of genetic variance in at least one of the three traits studied. Given the small difference in the traits between our study (number of total oocytes, viable oocytes and embryos) and Pereira (2019) (number of viable oocytes and number of viable embryos), it can be seen that genetic variance for this type of traits is scattered among chromosomes, which may be due to its polygenic nature. In addition, *SMTN* and *INPP5J* genes were associated with number of viable oocytes by Pereira (2019), while the same genes were associated with number of embryos in our work. Repression of *INPP5J* expression may result in cell proliferation of human ovarian cancer cells (Zhu et al. 2015). Also, it is known that *SMAD* transcription factor is essential for oviduct and uterus maintenance and integrity, but knockout mouse for these transcription factors increase *SMTN* expression in oviducts, indicating that this gene may have a supportive role during early pregnancy in mice (Rodriguez et al. 2016). In this way, *SMTN* and *INPP5J* genes may play an secondary role in oocyte and embryo production.

About transcription factors (TFs) found in this work, focus will be given to the five TFs most related to the studied traits. In our work, *AHR* was associated with anatomical structure, system and organ development. Its function was confirmed by Gialitakis et al. (2017) who mentioned that *AHR* plays a significant role in cell cycle progression and fetal development, because its interaction with genes involved in pluripotency. In our work, *HIF1A* was associated with embryonic hematopoiesis, morphogenesis process, neural crest cell-related process and heart formation. Turhan et al. (2021) showed that *HIF1alpha* expression is important for cumulus expansion and expression, since it regulates expression of functional protein markers in bovine cumulus cells during in vitro maturation and subsequent embryo development. Also, based on our findings, it seems that *AHR* and *HIF1A* interact with *ARNT* gene. According to Munakata et al. (2016), *HIF1A* is the alpha and *ARNT* the beta subunits of *HIF1* gene, a basic helix-loop-helix transcription factor, whose associated genes tend to be upregulated during follicle development. *EGR1* transcriptional factor was related to

hematopoiesis and organ, system and anatomical structure development. Although the exact function of the gene is not clear, Ulloa et al. (2015) found higher expression of *EGR1* in in vivo-produced embryos than in their in vitro counterparts and suggested that *EGR1* expression could serve as a marker of embryo quality.

In our work, *PAX2* transcriptional factor was involved in several processes including morphogenesis, mesenchymal to epithelial transition and general embryonic development. In embryonic genome activation, *PAX2a* plays important roles since it is expressed as a maternal gene in oocytes, needed for oogenesis and early development (Pachoensuk et al. 2020). This means that most transcriptional factors found in our results have consistent regulatory functions associated with the oocyte and embryo production, or at least, with basic cell functions, needed in several biological processes.

Although microRNAs found in the GWAS were not used in gene ontology analysis, interesting results for some of these microRNAs were found in the literature that are worth mentioning. It already has been shown that functional modulation of miRNA-130b is involved in bovine granulosa and cumulus cells function, oocyte maturation and preimplantation embryo development (Sinha et al. 2017). *MiRNA 130* and *miRNA 301B* were associated with follicular growth and development in cattle (Zielak-Steciwko and Evans 2016). Interestingly, in a review of placenta-derived miRNAs, (Xu et al. 2021) mentioned that *miRNA 499* derived from bovine placental exosome inhibits the activation of *NF-kB* via *Lin28B*/*let-7* axis, attenuating inflammatory responses and creating an immune-tolerant microenvironment in the uterus. *NF-kB* was one of the 16 transcriptional factors found in our research and *LIN28B* was one of the genes found on BTA9 related to the number of embryos. On GeneCards it is explained that overexpression of *LIN28B* gene in primary tumors is linked to the repression of let-7 family of microRNAs, which contributes to maintain the pluripotent state of embryonic stem cells. Functions associated with these five transcriptional factors described and their interaction between themselves, the genes, and microRNAs validated their importance in oocyte maturation and the general development of the embryo, indicating that they are good markers for the production of oocytes and embryos.

Considering all genes found in our work, 39 genes in common with Graf et al. (2014), who evaluated the transcriptome of bovine oocyte maturation and early embryonic development and identified specific genes according to the timing of embryonic activation. From these common genes, *ARHGAP15* was mentioned as being expressed in blastocyst transcripts, but not present in oocytes, that is, its transcripts arise after fertilization, which is in accordance with our results, since *ARHGAP15* was associated exclusively with number of

embryos. Finally, some genes found in our work were also mentioned by Graf et al. (2014) to be expressed as primary transcripts with intronic sequences determining the onset of embryonic gene expression, such as *EIF6*, *MORC3* and *RBM39* expressed in 4-cell embryos, *ACSS2*, *CHAF1B*, *DUSP18*, *DYM*, *DYNLRB1*, *EDEM2*, *ERGIC3*, *GAS2*, *GJA1*, *GTDC1*, *INTU*, *ITCH*, *LONP2*, *MED15*, *MFSD8*, *MORC2*, *MRPL1*, *PLK4*, *PREP*, *SF3A1*, *SIAH1*, *SPAG4*, *TBC1D10A* and *UBTD2* expressed in 8-cell embryos, *N4BP1*, *PIGU*, *PTPN3* and *SLC25A31* expressed in 16-cell embryos, and *EPB41L4B*, *GSS*, *NFS1* and *SEC14L2* expressed in blastocysts. As the above-mentioned genes are important in the onset of embryonic gene expression, it may explain the results found in VarElect, in which some of these genes were directly related only with "oocyte", or only with "embryo" or with both terms.

Assou et al. (2013) found 9,805 genes expressed in both cumulus cells from patients stimulated with highly purified human menopausal gonadotropin (HP-hMG) or recombinant FSH (rFSH), that are part of the most widely used protocols for controlled ovarian stimulation and in vitro fertilization in humans. In our work, 56 genes were in common with Assou et al. (2013) results. As they are expressed in human periovulatory follicles, they may play a role in ovulation and may be related to embryo development. Using single-cell transcriptomic RNA sequencing and differential expression analysis in human oocytes, Li et al. (2020) discovered several differentially expressed genes in oocytes that failed to mature completely when compared with normally mature oocytes. Some genes cited by Li et al. (2020) were also found in our work, including *MAP1LC3A*, *NCOA6* and *PTPN3* as down-regulated genes and *CEP250*, *CTIF*, *DYNLRB1*, *LZTR1*, *OSBP2*, *RBM39*, *UQCC1* and *VCAN* as up-regulated in maturation-deficient oocytes. Li et al. (2020) also found 2.329 genes differentially spliced into more than 1 alternative isoform, in which 12 genes were in common with our findings: *ABHD18*, *CHAF1B*, *CPNE1*, *DYNLRB1*, *FBXW11*, *GTDC1*, *LARP1B*, *LONP2*, *MFSD8*, *PIGU*, *RBM39* and *UBE2L3*. In addition, some of these genes were directly correlated with oocyte and/or embryo in VarElect. Great part of the genes discovered in our work, related to the number of total and viable oocytes and embryos are also related to ovulation and embryo development or other reproductive traits in several species. Our work provide a variety of genes that may be explored in further research to better understand the exact function of these genes on oocytes and embryo production and development.

Considering now the most likely candidate genes identified through gene-TF Network Analyzer tool, *ALDH1L1* was related to 10-formyltetrahydrofolate catabolic process. In humans, it is known that neural tube formation is the origin of the brain and spinal cord and

folate supplementation prevents neural tube defects (Sadler 2005). Anthony and Heintz (2007) suggested that folate might modulate proliferation within the early neural tube on *ALDH1L1*-positive regulated cells. In our work, several enriched biological pathways were associated with catabolic processes. Despite catabolic process is a basic metabolic function, we can see that in this case, the catabolic process related to the *ALDH1L1* gene is crucial at the onset of embryo's development.

Following up the most likely candidate genes associated with the number of oocytes and embryos in our work, *TCN2* codifies for transcobalamin 2 and its higher abundance in human placenta suggests a potential role in vitamin B12 transport to the fetus (Layden et al. 2016). *CNOT6L* and *CRKL* were found in our work and in Assou et al. (2013), who showed that these genes may play a role in ovulation and embryo development. *TBC1D32* was associated exclusively with number of embryos in our work and among its related biological processes was determination of left/right symmetry. Anton et al. (2018) associated *TBC1D32* gene with fertility breeding value in Hungarian Simmental cattle, with a possible role in left–right symmetry and embryonic digit morphogenesis, in which a digit is one of the terminal divisions of an appendage, such as a finger or toe. Furthermore, *PI4KA* gene was related to viral replication complex formation and maintenance in our work and, according to GeneCards (https://www.genecards.org/), this gene plays a role in base cell metabolism functions, such as transferase and kinase activity. *PI4KA* was upregulated in Japanese Black (Wagyu) embryos after heat shock and it was considered by Sakatani et al. (2013) an important gene for embryonic development. Also, *PI4KA* was a differentially expressed gene correlated with post-partum metabolite/hormone patterns in the oviduct of cyclic Holstein dairy cows (Valour et al. 2013). Therefore, genes that have important base cell functions may play a direct or indirect role in embryo development. Further studies are suggest to explore the genetic and genomic relation between the number of oocytes or embryos and other important reproductive traits in dairy cattle, especially the ones found in our QTL results, to provide better understanding over the general relation among fertility traits.

Conclusions

Most part of genes found in the genome wide association analysis has roles in initial embryo development or general cell functions. Five transcriptional factors found, *ARNT*, *EGR1*, *HIF1A*, *AHR* and *PAX2* are good markers for the production of oocytes and embryos. Further studies are needed to verify the expression of these genes in dairy Gir cattle to better understand their role in oocyte and embryo production. Genes identified in this work may

play a role in basic regulatory cell processes, indicating that they also affect other economic important traits in livestock. As genomic regions found in our work were related to several types of QTLs, we also suggest further studies to explore the relation between the number of oocytes/embryos and production traits.

Acknowledgments We thank the farmers who provided data for this study.

Author contributions JCdCP, MAM, MVBdS, SEFG: conceptualization of the study. RFBR, SEFG: drafted the manuscript. RFBR: performed the experiments and analysis. AOG, MGS, PIO: statistical analyses support. MVBdS, MFM, MAM, JCdCP: bioinformatics technical support. All authors read and approved the final manuscript.

Funding Coordination and Improvement of Higher Level Personnel (CAPES), National Council for Scientific and Technological Development (CNPq) – regarding process 142600/2019-9, Ministério da Ciência Tecnologia e Inovação (MCTI) and Brazilian National Institute of Science and Technology in Animal Science (INCT-CA) provided financial support towards this study.

Data availability statement The datasets generated and analyzed during this study are available from the corresponding author on reasonable request.

Statements and Declarations

Conflict of interest The authors declare no conflict of interest.

Ethics approval Not applicable. Farmers who provided the data adherence to a high standard of veterinary care.

Consent for publication Not applicable.

References

Aguilar I, Misztal I, Johnson DL, et al (2010) Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 93:743–752. https://doi.org/10.3168/jds.2009-2730

Animal QTL Database – QTLdb (2022) Accessed 10 Abr 2022. http://www.animalgenome.org/QTLdb.

Anthony TE, Heintz N (2007) The folate metabolic enzyme ALDH1L1 is restricted to the

midline of the early CNS, suggesting a role in human neural tube defects. J Comp Neurol 500:368–383. https://doi.org/10.1002/cne.21179

Anton I, Húth B, Füller I, et al (2018) Effect of single-nucleotide polymorphisms on the breeding value of fertility and breeding value of beef in Hungarian Simmental cattle. Acta Vet Hung 66:215–225. https://doi.org/10.1556/004.2018.020

Assou S, Haouzi D, Dechaud H, et al (2013) Comparative Gene Expression Profiling in Human Cumulus Cells according to Ovarian Gonadotropin Treatments. Biomed Res Int 2013:1–13. https://doi.org/10.1155/2013/354582

Berry DP, Eivers B, Dunne G, McParland S (2019) Genetics of bull semen characteristics in a multi-breed cattle population. Theriogenology 123:202–208. https://doi.org/10.1016/j.theriogenology.2018.10.006

Bindea G, Mlecnik B, Hackl H, et al (2009) ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics 25:. https://doi.org/10.1093/bioinformatics/btp101

Carvalho NS, Daltro DS, Machado JD, et al (2021) Genetic parameters and genetic trends of conformation and management traits in Dairy Gir cattle. Arq Bras Med Vet Zootec 73:938-948. https://doi.org/10.1590/1678-4162-12341

Chen H (2022) VennDiagram: Generate High-Resolution Venn and Euler Plots. R package version 1.7.3. https://CRAN.R-project.org/package=VennDiagram.

Drum JN, Sartori R, França e Melo L de (2019) New perspectives in assisted reproductive technologies in cattle. Rev Ecuatoriana Cienc Anim 3:85–106

Fanton M, Cho JH, Baker VL, et al (2023) A higher number of oocytes retrieved is associated with an increase in fertilized oocytes, blastocysts, and cumulative live birth rates. Fertil Steril 119:762-769. https://doi.org/10.1016/j.fertnstert.2023.01.001

GeneCards: The Human Gene Database (2022) Accessed 20 Apr 2022. https://www.genecards.org/.

Gialitakis M, Tolaini M, Li Y, et al (2017) Activation of the Aryl Hydrocarbon Receptor Interferes with Early Embryonic Development. Stem Cell Reports 9:1377–1386. https://doi.org/10.1016/j.stemcr.2017.09.025

Goldar S, Khaniani MS, Derakhshan SM, Baradaran B (2015) Molecular Mechanisms of Apoptosis and Roles in Cancer Development and Treatment. Asian Pacific J Cancer Prev 16:2129–2144. https://doi.org/10.7314/APJCP.2015.16.6.2129

González-Herrera LG, Pereira RJ, El Faro L, Albuquerque LG (2022) Genetic evaluation of lactation persistency in the Gyr breed by using a two-trait random regression model.

Anim Prod Sci 62:216. https://doi.org/10.1071/AN20267

Graf A, Krebs S, Zakhartchenko V, et al (2014) Fine mapping of genome activation in bovine embryos by RNA sequencing. Proc Natl Acad Sci 111:4139–4144. https://doi.org/10.1073/pnas.1321569111

Hsu MI, Wang CW, Chen CH, Tzeng CR (2016) Impact of the number of retrieved oocytes on pregnancy outcome in in vitro fertilization. Taiwan J Obstet Gynecol 55:821-825. https://doi.org/10.1016/j.tjog.2015.05.009

Jaton C, Schenkel FS, Sargolzaei M, et al (2018) Genome-wide association study and in silico functional analysis of the number of embryos produced by Holstein donors. J Dairy Sci 101:7248–7257. https://doi.org/10.3168/jds.2017-13848

Jukam D, Shariati SAM, Skotheim JM (2017) Zygotic Genome Activation in Vertebrates. Dev Cell 42:316–332. https://doi.org/10.1016/j.devcel.2017.07.026

Layden AJ, O'Brien KO, Pressman EK, et al (2016) Vitamin B12 and placental expression of transcobalamin in pregnant adolescents. Placenta 45:1–7. https://doi.org/10.1016/j.placenta.2016.06.011

Li J, Lu M, Zhang P, et al (2020) Aberrant spliceosome expression and altered alternative splicing events correlate with maturation deficiency in human oocytes. Cell Cycle 19:2182–2194. https://doi.org/10.1080/15384101.2020.1799295

Ma L, Cole JB, Da Y, VanRaden PM (2019) Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. J Dairy Sci 102:3735–3743. https://doi.org/10.3168/jds.2018-15269

Maere S, Heymans K, Kuiper M (2005) BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. Bioinformatics 21:. https://doi.org/10.1093/bioinformatics/bti551

Maiorano AM, Lourenco DL, Tsuruta S, et al (2018) Assessing genetic architecture and signatures of selection of dual purpose Gir cattle populations using genomic information. PLoS One 13: e0200694. https://doi.org/10.1371/journal.pone.0200694

Martins CF (2010) O impacto da transferência de embriões e da fecundação in vitro na produção de bovinos no Brasil. Embrapa Cerrados: INFOTECA-E. Available in: https://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/875061. Accessed 17 Mai 2022.

Misztal I, Tsuruta S, Strabel T, et al (2002) BLUPF90 and related programs (BGF90). Commun. No. 28-07

Moore SG, Hasler JF (2017) A 100-Year Review: Reproductive technologies in dairy science.

J Dairy Sci 100:10314–10331. https://doi.org/10.3168/jds.2017-13138

Munakata Y, Kawahara-Miki R, Shiratsuki S, et al (2016) Gene expression patterns in granulosa cells and oocytes at various stages of follicle development as well as in *in vitro* grown oocyte-and-granulosa cell complexes. J Reprod Dev 62:359–366. https://doi.org/10.1262/jrd.2016-022

do Nascimento Rangel AH, Araújo TPM, da Costa Lima GF, et al (2018) Estimativa de tendências genéticas e fenotípicas para características leiteiras de animais das raças gir e guzerá. Acta Sci - Anim Sci 40:. https://doi.org/10.4025/actascianimsci.v40i1.36685

Neuwirth E (2022) RColorBrewer: ColorBrewer Palettes. R package version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer.

NCBI - National Center for Biotechnology Information (2022) Accessed 24 Abr 2022. https://www.ncbi.nlm.nih.gov/.

Ortega MS (2018) Identification of genes associated with reproductive function in dairy cattle. Anim Reprod 15:923–932. https://doi.org/10.21451/1984-3143-AR2018-0018

Otto PI, Guimarães SEF, Calus MPL, et al (2020) Single-step genome-wide association studies (GWAS) and post-GWAS analyses to identify genomic regions and candidate genes for milk yield in Brazilian Girolando cattle. J Dairy Sci 103:10347–10360. https://doi.org/10.3168/jds.2019-17890

Pachoensuk T, Fukuyo T, Wanlada K, et al (2020) Pax2a is expressed in oocytes and is responsible for early development and oogenesis in zebrafish. Biochem Biophys Res Commun 533:592–599. https://doi.org/10.1016/j.bbrc.2020.09.059

Panetto JCC, Silva MVGB, Verneque RS, et al (2021) Programa Nacional de Melhoramento do Gir Leiteiro - Sumário Brasileiro de Touros; 4ª Avaliação genômica de touros - Resultado do Teste de Progênie. Embrapa Gado de Leite.

Parker Gaddis KL, Null DJ, Cole JB (2016) Explorations in genome-wide association studies and network analyses with dairy cattle fertility traits. J Dairy Sci 99:6420–6435 http://dx.doi.org/10.3168/jds.2015-10444

Parker Gaddis KL, Dikmen S, Null DJ, et al (2017) Evaluation of genetic components in traits related to superovulation, in vitro fertilization, and embryo transfer in Holstein cattle. J Dairy Sci 100:2877–2891. https://doi.org/10.3168/jds.2016-11907

Pereira MA (2019) Análise genética de características de produção in vitro de embriões de doadoras da raça Gir Leiteiro. UNESP.

Perez BC, Peixoto MGCD, Bruneli FT, Ramos PVB, Balieiro JCC (2016) Genetic analysis of oocyte and embryo production traits in Guzerá breed donors and their associations with

age at first calving. Genet Mol Res 15:15027583. http://dx.doi.org/10.4238/gmr.15027583

R Core Team (2022) ´R: A language and environment for statistical computing´ R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rocha R de FB, Garcia AO, Otto PI, et al (2023) Runs of homozygosity and signatures of selection for number of oocytes and embryos in the Gir Indicine cattle. Mamm Genome. https://doi.org/10.1007/s00335-023-09989-w

Rocha RdFB, Otto PI, da Silva MVGB, et al (2022) Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed. Anim Prod Sci 62:1661–1670. https://doi.org/10.1071/AN21588

Rodriguez A, Tripurani SK, Burton JC, et al (2016) SMAD Signaling Is Required for Structural Integrity of the Female Reproductive Tract and Uterine Function During Early Pregnancy in Mice. Biol Reprod 95:44–44. https://doi.org/10.1095/biolreprod.116.139477

Roelofs JB, van Erp-van der Kooij E (2015) Estrus detection tools and their applicability in cattle: recent and perspectival situation. Anim Reprod 12.

Rosen BD, Bickhart DM, Schnabel RD, et al (2020) De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9:. https://doi.org/10.1093/gigascience/giaa021

Sadler TW (2005) Embryology of neural tube development. Am J Med Genet Part C Semin Med Genet 135C:2–8. https://doi.org/10.1002/ajmg.c.30049

Sakatani M, Bonilla L, Dobbs KB, et al (2013) Changes in the transcriptome of morula-stage bovine embryos caused by heat shock: relationship to developmental acquisition of thermotolerance. Reprod Biol Endocrinol 11:3. https://doi.org/10.1186/1477-7827-11-3

Sbardella AP, Watanabe RN, da Costa RM, et al (2021) Genome-Wide Association Study Provides Insights into Important Genes for Reproductive Traits in Nelore Cattle. Animals (Basel) 11:1386. https://doi.org/10.3390/ani11051386

Shannon P, Markiel A, Ozier O, et al (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res 13:2498–2504. https://doi.org/10.1101/gr.1239303

Sinha PB, Tesfaye D, Rings F, et al (2017) MicroRNA-130b is involved in bovine granulosa and cumulus cells function, oocyte maturation and blastocyst formation. J Ovarian Res 10:37. https://doi.org/10.1186/s13048-017-0336-1

Stelzer, G, Plaschkes, I, Oz-Levi, D, et al (2016) VarElect: The phenotype-based variation

prioritizer of the GeneCards Suite. BMC Genomics. 17:195–206. https://doi.org/10.1186/s12864-016-2722-2

Stephens MA (1986) Tests based on EDF statistics. In: Stephens, MA (1974) EDF statistics for goodness of fit and some comparisons. J Am Stat Assoc 69: 730-737. https://doi.org/10.1080/01621459.1974.10480196

TFM-Explorer (2022) Transcription Factor Matrix Explorer – Bonsai Bioinformatics. Accessed 24 Abr 2022. https://bioinfo.lifl.fr/tfm-explorer/tfm-explorer.php.

Thode HC (2002) Testing For Normality. Marcel Dekker, New York, USA

Tiezzi F, Parker-Gaddis KL, Cole JB, et al (2015) A Genome-Wide Association Study for Clinical Mastitis in First Parity US Holstein Cows Using Single-Step Approach and Genomic Matrix Re-Weighting Procedure. PLoS One 10:e0114919. https://doi.org/10.1371/journal.pone.0114919

Turhan A, Pereira MT, Schuler G, et al (2021) Hypoxia-inducible factor (HIF1alpha) inhibition modulates cumulus cell function and affects bovine oocyte maturation in vitro†. Biol Reprod 104:479–491. https://doi.org/10.1093/biolre/ioaa196

Turner SD (2018) qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. J Open Source Softw 3:731. https://doi.org/10.21105/joss.00731

Ulloa SMB, Heinzmann J, Herrmann D, et al (2015) Effects of different oocyte retrieval and *in vitro* maturation systems on bovine embryo development and quality. Zygote 23:367–377. https://doi.org/10.1017/S0967199413000658

Valour D, Hue I, Degrelle S, et al (2013) Pre- and Post-Partum Mild Underfeeding Influences Gene Expression in the Reproductive Tract of Cyclic Dairy Cows. Reprod Domest Anim 48:484–499. https://doi.org/10.1111/rda.12113

Vargas G, Neves HHR, Camargo GMF, et al (2018) Genome-wide association study and functional analysis of feet and leg conformation traits in Nellore cattle. J Anim Sci 96:1617–1627. https://doi.org/10.1093/jas/sky079

Verardo LL, Silva FF, Lopes MS, et al (2016) Revealing new candidate genes for reproductive traits in pigs: Combining Bayesian GWAS and functional pathways. Genet Sel Evol 48:. https://doi.org/10.1186/s12711-016-0189-x

Viana JHM (2022) 2021 Statistics of embryo production and transfer in domestic farm animals. A new milestone has been reached: Transfers of IVP embryos were over one million worldwide. Embryo Technology Newsletter 40.

Xu P, Ma Y, Wu H, Wang Y-L (2021) Placenta-Derived MicroRNAs in the Pathophysiology of Human Pregnancy. Front Cell Dev Biol 9:. https://doi.org/10.3389/fcell.2021.646326

Zhu T, Yuan J, Wang Y, et al (2015) MiR-661 contributed to cell proliferation of human ovarian cancer cells by repressing INPP5J expression. Biomed Pharmacother 75:123–128. https://doi.org/10.1016/j.biopha.2015.07.023

Zielak-Steciwko AE, Evans ACO (2016) Genomic portrait of ovarian follicle growth regulation in cattle. Reprod Biol 16:197–202. https://doi.org/10.1016/j.repbio.2016.07.003

Running head: Family analyses in Gir cattle families

Inheritance of genomic regions and genes associated with number of oocytes and embryos in Gir cattle through daughter design*

* Published at Journal of Dairy Science (https://doi.org/10.3168/jds.2023-24111)

R. F. B. Rocha,[1,2] A. O. Garcia,[1] M. G. dos Santos,[1] P. I. Otto,[3] M. V. B. da Silva,[4] M. F. Martins,[4] M. A. Machado,[4] J. C. C. Panetto,[4] M. P. L. Calus,[2] S. F. Guimarães,[1*]

[1]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, MG, 36570-900, Brazil, e-mail: renata.bretanha@ufv.br, arielly.garcia@ufv.br, mateus.g.santos@ufv.br, sfacioni@ufv.br

[2]Animal Breeding and Genomics, Wageningen University & Research, 6700 AH Wageningen, the Netherlands, e-mail: mario.calus@wur.nl

[3]Department of Animal Science, Universidade Federal de Santa Maria, Santa Maria, RS, 97105-900, Brazil, e-mail: pamela.otto@ufsm.br

[4]EMBRAPA - Dairy Cattle, Juiz de Fora, MG, 36038-330, Brazil, e-mail: joao.panetto@embrapa.br, marcos.vb.silva@embrapa.br, marta.martins@embrapa.br, marco.machado@embrapa.br

[*]Corresponding author: Simone Eliza Facioni Guimarães, sfacioni@ufv.br

Interpretative Summary

Genome Wide Association Studies (**GWAS**) help to identify genes that affect traits of economic interest. Even more in-depth, studies of the inheritance of genes in the larger Gir families can help to highlight genes that are commonly inherited by daughters and are related to oocyte and embryo production. This type of study helps to highlight specific regions of the genome. It is possible to explore the genetic inheritance of genes associated with traits of livestock interest through studies called “daughter design”.

Abstract

Over the past decades, daughter designs, including genotyped sires and their genotyped daughters, have been used as an approach to identify quantitative trait loci (QTL) related with economic traits. The aim of this study was to identify genomic regions inherited by Gir sire families and genes associated with number of viable oocytes (VO), total number of oocytes (TO) and number of embryos (EMBR) based on a daughter design approach. In total, 15 Gir sire families were selected. The number of daughters per family ranged from 26 to 395, which were genotyped with different SNP panels and imputed to the Illumina BovineHD BeadChip (777K) and had phenotypes for oocyte and embryo production. Daughters had phenotypic data for VO, TO, and EMBR. The search for QTL was performed through Genome-Wide Association Study (GWAS) based on genomic best linear unbiased prediction (gBLUP). QTL were found for each trait among and within families based on the top 10 genomic windows with the greatest genetic variance. For EMBR, genomic windows identified among families were located on BTA4, BTA5, BTA6, BTA7, BTA8, BTA13, BTA16 and BTA17, and they were most frequent on BTA7 within families. For VO, genomic windows were located on BTA2, BTA4, BTA5, BTA7, BTA17, BTA21, BTA22, BTA23 and BTA27 among families, being most frequent on BTA8 within families. For TO, top 10 genomic windows were identified on BTA2, BTA4, BTA5, BTA7, BTA17, BTA21, BTA22, BTA26 and BTA27, being most frequent on BTA7 and BTA8 within families. Considering all results, the greatest number of genomic windows was found on BTA7, where *VCAN*, *XRCC4*, *TRNAC-ACA*, *HAPLN1* and *EDIL3* genes were identified in the common regions. In conclusion, 15 Gir sire families with 26 to 395 daughters per family with phenotypes for oocyte and embryo production helped to identify the inheritance of several genomic regions, especially on BTA7, where *EDIL3*, *HAPLN1* and *VCAN* candidate genes were associated with number of oocytes and embryos in Gir cattle families.

Keywords: bovine; family analysis; reproductive traits; zebu dairy cattle

Introduction

Artificial selection for reproductive traits promotes a slow response to selection due to the high polygenic nature and low heritability of this type of trait, as they depend on the action of many genes and are influenced by environmental factors (Fathoni et al. 2022, Rocha et al., 2022). In recent years, collection of genomic information of animals started to be used to identify genetic variants that affect several traits in livestock, in addition to including this molecular information in genetic evaluations, promoting a faster response to selection. As an example, evaluations for molecular traits in dairy Gir in Brazil allowed the inclusion of genes related to milk production and hereditary conditions in its breeding program, enabling genomic selection and increasing the accuracy of the genetic values for Gir sires (Panetto et al., 2021). Over the years, the search for genetic variants associated with innumerous traits has grown in livestock farming as well as the methodologies applied to identify them (Grisart et al., 2002, Otto et al., 2020, Rocha et al., 2023).

Advances in DNA-based marker technology have enabled the development of new methodologies, including family analyses based on daughter and granddaughter designs, proposed as first by Weller et al. in 1990. These methodologies have been instrumental in discovering genomic regions associated with important traits in livestock species. Over the past decades, these approaches were used to find quantitative trait loci (**QTL**) and single nucleotide polymorphism (**SNP**) associated with economic traits in several cattle breeds (Jiang et al., 2010; Silva et al., 2011; Oliveira Júnior et al., 2019).

In a daughter design approach, genotypic information for sires and their daughters is considered, with phenotypes collected from the daughters, aiming to identify quantitative effects associated with marker alleles as a significant component of variance between daughter subgroups receiving alternative marker alleles from their sires (Weller et al., 1990). According to Weller et al. (1990), this method is helpful to identify genes with relatively small quantitative effects, which is the case for polygenic traits.

Some examples of the use of both approaches (daughter and granddaughter design) are in Grisart et al. (2002), who identified a nonconservative K232A substitution in the DGAT1 gene with a major effect on milkfat content and other milk traits in Holstein-Friesian cattle. In another example, daughter design employed with Genome-Wide Association Studies (GWAS) in a Holstein cattle population revealed 105 genome-wise significant SNPs for milk production traits, in which the majority (86 out of 105) of the significant SNPs was located

within the reported QTL regions, and others, in particular two SNPs, ARS-BFGL-NGS-4939 and BFGL-NGS-118998, were located 160 base pairs apart to the DGAT1 gene (Jiang et al., 2010). Using a similar approach to the daughter design in Nellore cattle breed, Oliveira Júnior et al. (2019) found QTLs on BTA 5 and 14, in which potential causal mutations were associated with reproduction-related biological processes such as oocyte maturation, embryo development, placenta development and response to reproductive hormones.

The Gir cattle breed, one of the most important *Bos taurus indicus* used to maintain pure Gir or for crossbreeding with Holstein cattle to generate the composite Girolando breed (Milanesi et al., 2022). Given the importance of Gir cattle breed for the dairy production system in tropical countries, a daughter design was used to search QTL affecting milk production traits (Silva et al., 2011). The Gir breed has a greater oocyte quality compared to *Bos taurus taurus* breeds, which allows greater success of in vitro embryo production (Drum et al., 2019). Identification of genetic variants related to such traits are important to provide clues to understand fertility regulation and genetic architecture in Gir cattle (Ma et al., 2019; Rocha et al., 2023). However, few studies using the daughter design approach for Gir cattle were reported in literature, and none so far for the embryo production related traits in this breed. Considering the importance of female fertility to livestock production systems, the aim of this study is to identify genomic regions inherited by Gir sire families and genes associated with number of viable oocytes (**VO**), total number of oocytes (**TO**) and number of embryos (**EMBR**) based on daughter design approach.

The hypothesis was that common genomic regions inherited across Gir cattle families could be identified and, in these regions, genes related to oocytes and embryo production could be found.

Material and Methods

Population structure

In this study, the daughter design methodology proposed by Weller et al. (1990) was applied. This methodology is carried out considering genotypic information for sires and their daughters, with phenotypes collected from the daughters. A pedigree file with 4,679 Gir animals was provided by Brazilian Agricultural Research Corporation (EMBRAPA), Dairy Cattle Center. From this pedigree file, 2,093 animals had genomic data, of which 169 were sires. As exclusion criteria for the daughter design, sire families with less than 20 phenotyped and genotyped daughters were not considered for the analyses (Silva et al., 2011). In this study, 15 families of the most important sires of the Gir cattle breed were evaluated. The

number of daughters per sire family ranged from 26 to 395, totalizing 1,474 daughters in all families (Table 1). These daughters had phenotypic data for number of viable oocytes (VO), total number of oocytes (TO) and number of embryos (EMBR). Table 2 shows the descriptive statistics of VO, TO and EMBR for 1,474 daughters of all 15 families. To be included, cows were required to have a minimum of three observations for all traits. The description for the phenotypic data was previously described by a recent study from our lab (Rocha et al., 2022). Briefly, the Ovum Pick-Up (**OPU**) sessions in all analyzed herds were conducted between January 2005 and June 2020 in herds located in Minas Gerais State (Brazil). Eight companies responsible for OPU and in vitro fertilization (**IVF**) processes were listed in the dataset, each company using their own protocols; in this way, we considered eight protocols in total for the analysis, varying among farms or sometimes within farms. OPU seasons were classified as: (1) January–March; (2) April–June; (3) July–September; (4) October–December. Animals were separated into contemporary groups (CG) by intersecting year and season of oocyte aspiration, farm and OPU-IVF protocol. Records belonging to CG with fewer than three animals were removed.

Table 1. Families, number and percentage of daughters, number of SNPs[1] after quality control per family used in the analyses

Family	Daughters		SNP Markers
	Number	%	
1	395	26.80	312976
2	367	24.90	324081
3	110	7.46	324191
4	79	5.36	311240
5	73	4.95	317841
6	62	4.21	310612
7	56	3.80	317460
8	55	3.73	312781
9	51	3.46	307673
10	49	3.32	295100
11	46	3.12	317975
12	38	2.58	318550
13	37	2.51	310847

14	30	2.04	290609
15	26	1.76	296773
Total	1474	100	386236

[1]SNP: Single Nucleotide Polymorphism

Table 2. Descriptive statistics[1] on Ovum Pick-Up (OPU) interval, donors' age and structures.

	N	Total	Mean	SD	MIN	MAX	CV
OPU Interval (days)	11,916	-	93.52	153.96	0	2,466	164.63
Donor's age (years)	11,916	-	6.30	3.16	1	16	50.24
Viable oocytes	11,915	191,057	16.04	12.78	0	180	79.72
Total oocytes	11,916	258,863	21.72	15.79	0	180	72.66
Embryos	11,903	34,568	2.90	3.70	0	46	127.23

[1]N: number of data samples; Total: total sum of structures (oocytes and embryos); SD: standard deviation; MIN: minimum; MAX: maximum; CV: coefficient of variation; Variables not transformed to the logarithmic scale; a minimum of 0 for OPU interval means two OPU's on the same cow on the same day, which were further fertilized with semen from different bulls.

Genotypic information

A genotype file with 2,093 animals and 420,718 SNPs was provided by EMBRAPA, Dairy Cattle Center. This file is the result of an imputation process where different density SNP panels and its respective number of animals (Illumina BovineSNP50 BeadChip v2 (50K; n = 288), Illumina BovineHD BeadChip (777K; n = 171), GGP Indicus (34K; n = 948), Z-Chip (30K; n = 459) e GGP Indicus (50K; n = 227)) were extracted and/or imputed to an HD v. 2 panel using FIMPUTE 2.2 software (Sargolzaei et al., 2014). The quality control parameters considered in the imputation process as exclusion criteria for variants included SNP with minor allele frequency ≤ 0.05, an absolute difference between observed and expected allele frequency of the heterozygote genotype for Hardy Weinberg Equilibrium of > 0.15, GenCall score ≤ 0.70, call rate ≤ 0.98 and SNP-SNP correlation > 0.995. Samples with call rate < 0.90 were also excluded.

Genome Wide Association Study (GWAS)

The search for Quantitative Trait Loci (QTL) among and within families was performed through GWAS analyses. From the provided data with 2,093 animals and 420,718 SNPs, genotypes of the 15 sires and their 1,474 daughters were selected. Thereafter, quality control

was applied to the genotype data of each family and all families together. The following parameters for variant exclusion were considered: SNP with minor allele frequency ≤ 0.05, an absolute difference between observed and expected allele frequency for Hardy Weinberg Equilibrium > 0.15, and SNP heritability (h^2), in which markers with estimated $h^2<0.98$ were discarded (Forneris et al., 2015). By default, PreGSf90 program used for molecular quality control also verifies parent-progeny Mendelian conflicts and eliminates progenies with conflicts (Misztal et al., 2002). The number of SNP remaining within and among families after quality control is presented in Table 1.

Statistical analyses

The GWAS based on genomic best linear unbiased prediction (**gBLUP**) within and among families was conducted using the BLUPF90 software family (Misztal et al., 2002). As the considered traits (VO, TO, and EMBR) did not have a normal distribution of the residuals, they were transformed using the logarithmic scale: ln (X + 1), where ln is the natural logarithm and X is the original data. After this transformation, the normal distribution of the data was verified by the Anderson–Darling test (Stephens, 1986; Thode, 2002). The following model was used:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Za} + \boldsymbol{Wp} + \boldsymbol{e}$$

where, $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{a}$, $\boldsymbol{p}$, and $\boldsymbol{e}$ are the vectors of observations; fixed effects; additive genetic random effects; permanent environment random effects and residual effects, respectively; $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are the incidence matrices of fixed, additive genetic and permanent environment, respectively. For VO and TO, CG was considered as fixed effect, while for EMBR, bull and bull breed used in the insemination process (Gir or Holstein) were additionally considered as fixed effects. Donor's age in days (linear and quadratic components) was included as covariates for all traits. OPU interval (linear component) was included as a covariate for VO and TO. The additive genetic, permanent environment and residual effects were assumed to be random. In the single-trait model, it was assumed that $\boldsymbol{a} \sim N(0, \boldsymbol{G}\sigma_a^2)$, $\boldsymbol{p} \sim N(0, \boldsymbol{I}\sigma_p^2)$ and $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$, where σ_a^2, σ_p^2 and σ_e^2 are the additive genetic, permanent environmental and residual variances, respectively; **I** is the identity matrix; and **G** is the genomic relationship matrix. GWAS results were reported as the percentage of genetic variance explained by a sliding genomic window of 100 adjacent SNPs for TO, VO, EMBR (logarithmic scale).

QTL and gene search

To identify regions possibly harboring QTLs, for each trait (TO, VO, EMBR), the top 10 genomic windows explaining the greatest percentage of additive genetic variance were

selected both within each family, and among all families (Tiezzi et al., 2015; Vargas et al., 2018; Otto et al., 2020). The genomic windows of all families were overlapped to identify the most common inherited genomic regions across families. Considering these regions, the lower and the upper positions in base pairs were noted. Gene search was carried out within this interval using the National Center for Biotechnology Information (NCBI, 2023, https://www.ncbi.nlm.nih.gov/) to identify candidate genes associated with all traits.

Results

The top 10 genomic windows with the greatest genetic variance among and within families are shown in Figure 1. For the number of embryos (EMBR), the top 10 genomic windows were located on BTA4, BTA5, BTA6, BTA7, BTA8, BTA13, BTA16 and BTA17 among families. For the results in top 10 genomic windows within each family, BTA7 was the most common for EMBR, present in 8 of 15 families. For the number of viable oocytes (VO), genomic windows in top 10 were located on BTA2, BTA4, BTA5, BTA7, BTA17, BTA21, BTA22, BTA23 and BTA27 among families. BTA8 was the most common in the top 10 genomic windows within families for VO, and was also present in 8 of 15 families. For the total number of oocytes (TO), genomic windows in the top 10 were present on BTA2, BTA4, BTA5, BTA7, BTA17, BTA21, BTA22, BTA26 and BTA27. Within families, BTA7 and BTA8 were the most common chromosomes, both present in 7 of 15 families. VO and TO traits shared similar chromosome regions among and within families.

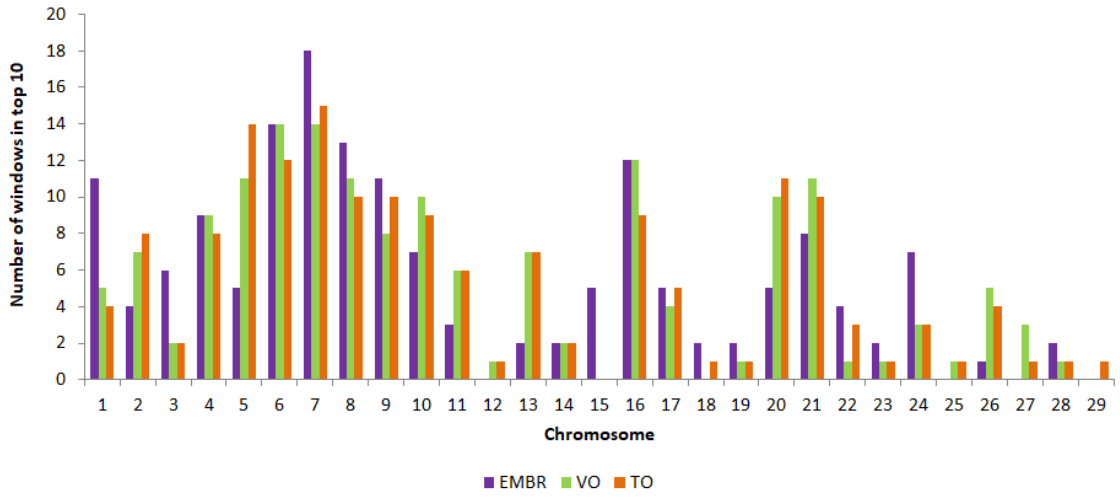


Figure 1. Number of genomic windows per chromosome among and within families associated with number of embryos (EMBR), number of viable oocytes (VO) and total number of oocytes (TO) present in the top 10 with the greatest genetic variance.

Considering the top 10 genomic windows within 15 families for three traits there were, in total, 450 genomic windows identified. In addition, animals from all 15 families were included in a single analysis, of which top 10 genomic windows were also selected for all three traits, adding 30 more genomic windows to the results. In total, 480 genomic windows were identified, of which 227 overlapped in different chromosomes. For example, unique genomic windows (with same initial and final positions in base pairs (bp)) were identified for all traits on BTA4 in positions 69914886 to 70357035 bp and 92208844 to 92628843 bp. Other examples of overlapping genomic windows (not necessarily with the same initial and final position in bp) were two genomic windows on BTA5, three on BTA6, three on BTA7, one on BTA13, one on BTA20 and two on BTA24 for all traits. Considering VO and TO traits, 110 unique genomic windows were found for both traits.

Considering all top 10 genomic windows for each trait within 15 families, 9% of these genomic windows were present on BTA7, that is, 42 of 450 genomic windows. Besides being the most frequent chromosome across the results, common genomic windows were found for all traits among and within some of the families on BTA7 (Table 3; Figure 2). With a gene search, *tRNAC-ACA*, *VCAN*, *XRCC4*, *HAPLN1* and *EDIL3* genes were identified in this common region (82974837 to 83997563 bp).

Table 3. Common regions on BTA7 for number of embryos (EMBR), number of viable oocytes (VO) and total number of oocytes (TO) within Gir cattle families

Initial position (bp[1])	Final position (bp)	Genetic variance[2]	Family	Trait
82974837	83407120	0.741	3rd family	TO
83007119	83440120	0.656	3rd family	VO
83113911	83520654	0.675	3rd family	EMBR
83409266	83955917	0.606	14th family	TO
83415663	83966202	0.729	14th family	VO
83422105	83955917	1.325	2nd family	EMBR
83422105	83966202	0.626	4th family	VO
83422105	83966202	0.568	4th family	TO
83440120	83997563	0.629	7th family	EMBR
83440120	83997563	0.666	7th family	VO

83440120	83997563	0.518	7th family	TO
83441451	83966202	0.471	12th family	EMBR

[1]bp: base pairs

[2]Percentage of genetic variance explained by each genomic window of 100 adjacent SNPs

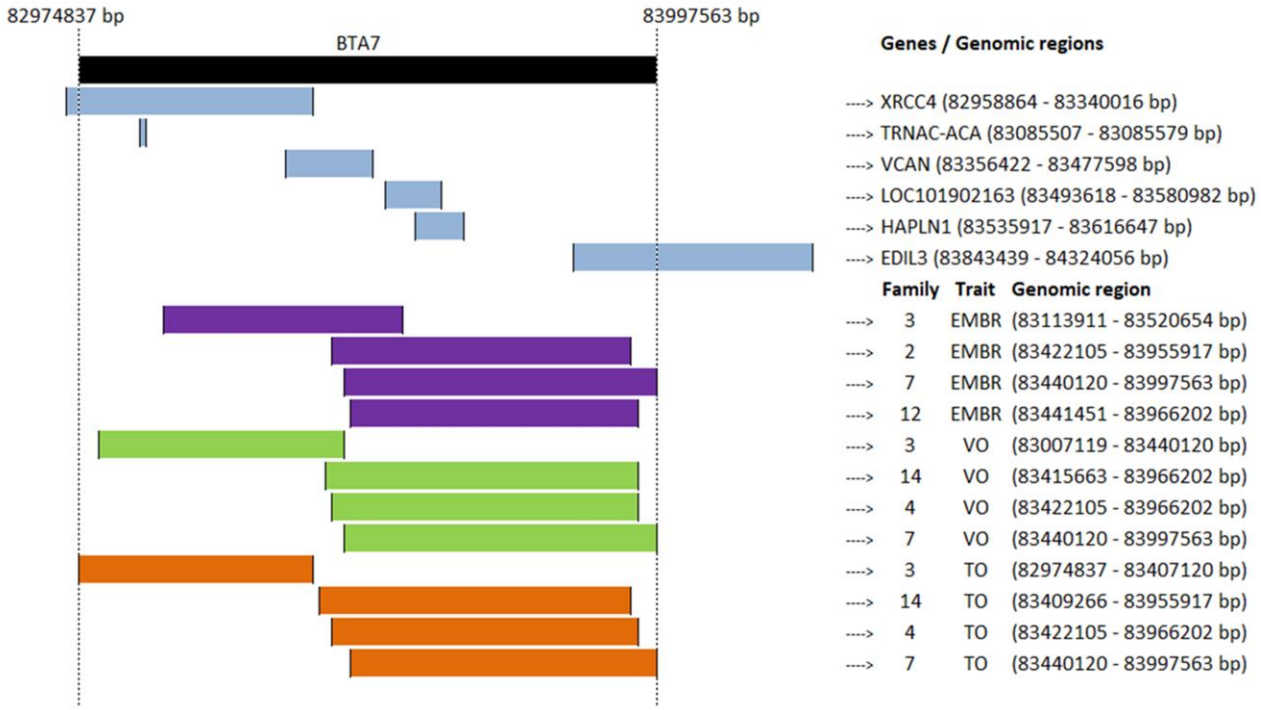


Figure 2. Common regions represented in base pairs (bp) on BTA7 for number of embryos (EMBR), number of viable oocytes (VO) and total number of oocytes (TO) within Gir cattle families (https://www.ncbi.nlm.nih.gov/) accessed on August 15th, 2023

Discussion

In the current work, we aimed to identify QTL regions and genes associated with VO, TO and EMBR based on a daughter design approach with GWAS. Considering the results for all three traits, BTA7 had the greatest number of genomic windows. In the overlapping genomic windows identified on BTA7 (82974837 to 83997563 bp) that are common to 40% of the families (6 out of 15 families), *tRNAC-ACA*, *VCAN*, *XRCC4*, *HAPLN1* and *EDIL3* genes were identified.

Considering the trait number of embryos, eight out of 29 chromosomes were present in the top 10 genomic windows among families. In a previous study, we found that the number of embryos can be affected by many genes, being therefore, a polygenic trait (Rocha et al., 2023). The identification of several genomic regions in different chromosomes in the current

work corroborates this. In a recent study from our lab (Rocha et al., 2022), we found high genetic correlation between VO and TO (1.00), VO and EMBR (0.82) and TO and EMBR (0.82), which may explain why most of the genomic windows identified for VO and TO in the current work are similar in chromosomes and base pairs (bp) positions. It is worth mentioning that the population used in this study is a smaller genotyped sample from the data used in our previous study (Rocha et al., 2022), of which variance components were obtained and used in the current study. The results obtained in our previous study indicated that it is possible to select for VO, TO and EMBR in Gir animals and encouraged us to conduct the present study looking for possible genomic markers related to these traits, which can be incorporated into genetic breeding programs and genomic selection.

On the Animal Quantitative Trait Loci (QTL) Database website (Animal QTLdb, 2023 https://www.animalgenome.org/QTLdb) only five published studies, all in Holstein, are cited for the traits number of embryos and number of transferable embryos, but none mentions the number of oocytes. This demonstrates the scarcity of QTLs found for the number of oocytes and embryos, emphasizing the importance of this study, even more, in Zebu breeds.

Jaton et al. (2018), studying Holstein cattle, mapped QTLs on BTA2, BTA11, BTA25 and BTA29 for number of embryos and, in addition to other studies, QTLs were mapped on BTA2, BTA11 and BTA15 for number of transferable embryos and on BTA2 and BTA11 for number of embryos degenerated. In our study, 12 genomic windows were identified on BTA2, and 10 on BTA11 for EMBR, VO, and TO within families, indicating that regions of these chromosomes may be a source of genetic influence on bovine female fertility.

Kadarmideen and Mazzoni (2019) identified expression quantitative trait loci (eQTLs) targeting candidate genes for oocytes and embryos in Holstein cattle, and from those significant eQTLs for oocyte, two similar regions were also identified in our study, on BTA16 for number of viable oocytes and BTA18 for total number of oocytes. Cai et al. (2019) found seven QTLs associated with female fertility index in Nordic Holstein cattle and identified significant SNPs in six chromosomes. One of the SNPs identified by Cai et al. (2019) on BTA17 (71,393,345 bp) indicated *LIF* gene as a candidate gene for female fertility, since it is required for blastocyst implantation and establishment of pregnancy. Similarly in our work, the same region was identified in one of the top 10 genomic windows associated with the number of embryos.

Although many regions were identified in our study for the number of embryos and oocytes, special attention will be given to the region on BTA7 (82974837-83997563 bp), where most of the genomic windows were identified within families. The *VCAN* gene was one of the

genes found in these common regions within families on BTA7, and according to GeneCards (2023, https://www.genecards.org/) this gene may take part in the regulation of cell motility, growth and differentiation. Expression levels of the *VCAN* gene were related to cumulus cells expansion process, being associated with the early embryo morphology score (Shen et al., 2020) and clinical pregnancy in humans (Ocampo et al., 2018), and associated with oocyte quality in cattle (Kussano et al., 2021). The gene *HAPLN1* is one of the cumulus cell expansion-related genes in pigs, and just like the *VCAN* gene, it is critical for oocyte maturation and high-quality embryos production (Mito et al., 2013; Park et al., 2018). Association between the expression of *VCAN* and *HAPLN1* genes and oocyte maturation found in literature may explain why these genes were identified in the most common BTA7 regions within families for the number of oocytes and embryos in our work, so these genes can be considered good candidates for improving the production of oocytes and embryos in the Gir breed. Gene expression of the *VCAN* and *HAPLN1* genes can be explored in the future to confirm their role in oocyte maturation and embryo quality in Gir cattle. Another gene found on BTA7 was *XRCC4*, which encodes a protein responsible for DNA double-strand breaks repair. This was verified by *XRCC4* transcript levels in porcine aged oocytes (Lin et al., 2021). However, DNA repair as basic function of body cells can also be observed in other types of tissue. Ruis et al., 2020 mentions that the absence of *XRCC4* in human cells influences most of DNA repair and recombination.

Another interesting gene found in our work was *EDIL3*, which encodes an integrin ligand protein and may be involved in regulation of vascular morphogenesis of remodeling in embryonic development, according to GeneCards (2023, https://www.genecards.org/). It is interesting to mention that chromosomal locations of *EDIL3*/*MFGE8* genes revealed a nested syntenous relationship with *HAPLN1*/*HAPLN3* and *VCAN*/*ACAN*, genes involved in vertebrate calcification, and *EDIL3* overexpression in eggshell tissue occurs at the onset of shell calcification in chicken (Stapane et al., 2019). The relationship mentioned by Stapane et al. (2019) between the genes *EDIL3*, *HAPLN1* and *VCAN* in chicken, which were also identified in our work, indeed suggests that these genes make contribution to embryo development. Willforss et al. (2021), studying Holstein bull fertility through protein markers in seminal plasma, mentions a negative correlation of EDIL3 expression with an increase in non-return rates to estrus (i.e. the proportion of cows not coming back to estrous after insemination and therefore deemed to be pregnant) and suggested that EDIL3 gene has a regulatory role on the immune response of the female genital tract. As well as Willforss et al. (2021), we also suggest a further investigation not only for the *EDIL3* gene but also for

HAPLN1 and *VCAN* genes. Exploring the expression of *EDIL3*, *HAPLN1*, and *VCAN* genes may give a better idea on how these genes regulate the oocyte and embryo production. Limitations regarding the daughter design are related to the number of phenotyped and genotyped daughters for each sire family. In addition, not only the sires, but also the dams have an important contribution, but this methodology becomes more limited regarding the evaluation of dam families, since it is easier to have enough number of daughters per sire than to have enough daughters per dam. In our data, only one dam reached the threshold of having at least 20 phenotyped and genotyped daughters that were used to select sires included in our study. For most cattle breeds, it is more common to genotype the sires, since sires can have a greater number of progeny than dams, and so, sires genomic information become more frequently used in genomic selection. In this study it was possible to obtain sufficient genomic information thanks to the Brazilian Dairy Gir Breeding Program, which collects genomic data of this breed since 2001 (Panetto et al. 2021). However, for other breeds or even species, application of this methodology might be limited by the availability of sufficient genomic data.

Conclusion

The approach used in this work allowed uncovering many genomic regions in different chromosomes for the number of total and viable oocytes and for the number of embryos, which is consistent with these traits being polygenic. The results suggest that the largest genomic regions affecting the number of total and viable oocytes may be the same in different families from the Gir breed, which is consistent with earlier observations that these traits are highly genetically correlated. The identified genomic regions located on BTA7, between 82974837 and 83997563 bp, can be highlighted as the most interesting results, since it is where the most common regions were found among and within families. Several genes, i.e., *EDIL3*, *HAPLN1* and *VCAN*, associated with oocyte maturation and embryo development related-functions in the literature, were here identified as possible candidate genes associated with the number of oocytes and embryos.

Acknowledgments

We thank the farms and Brazilian Agriculture Research Corporation (EMBRAPA) Dairy Cattle Research Center (Juiz de Fora) for providing data for this study. Coordination and Improvement of Higher Level Personnel (CAPES), National Council for Scientific and Technological Development (CNPq), Ministério da Ciência Tecnologia e Inovação (MCTI)

and Brazilian National Institute of Science and Technology in Animal Science (INCT-CA) provided financial support towards this study. The authors have not stated any conflicts of interest.

References

Animal QTLdb. 2023. Animal QTL Database (QTLdb). Accessed Feb. 10, 2023. https://www.animalgenome.org/cgi-bin/QTLdb/index

Cai, Z., B. Guldbrandtsen, M.S. Lund, and G. Sahana. 2019. Prioritizing candidate genes for fertility in dairy cows using gene-based analysis, functional annotation and differential gene expression. BMC Genomics 20:255. doi:10.1186/s12864-019-5638-9.

Drum, J.N., R. Sartori, and L. de França e Melo. 2019. New perspectives in assisted reproductive technologies in cattle. Rev. Ecuatoriana Cienc. Anim. 3:85–106.

Fathoni, A., Boonkum, W., Chankitisakul, V., Duangjinda, M. 2022. An Appropriate Genetic Approach for Improving Reproductive Traits in Crossbred Thai-Holstein Cattle under Heat Stress Conditions. Vet Sci. 9:163. doi: 10.3390/vetsci9040163.

Forneris, N.S., Legarra, A., Vitezica, Z.G., Tsuruta, S., Aguilar, I., Misztal, I., Cantet, R.J. 2015 Quality control of genotypes using heritability estimates of gene content at the marker. Genetics 199:675-681. doi: 10.1534/genetics.114.173559.

GeneCards. 2023. GeneCards: The Human Gene Database. Accessed Feb. 15, 2023. https://www.genecards.org/

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional Candidate Cloning of a QTL in Dairy Cattle: Identification of a Missense Mutation in the Bovine DGAT1 Gene with Major Effect on Milk Yield and Composition. Genome Res. 12:222–231. doi:10.1101/gr.224202.

Jaton, C., F.S. Schenkel, M. Sargolzaei, A. Cánova, F. Malchiodi, C.A. Price, C. Baes, and F. Miglior. 2018. Genome-wide association study and in silico functional analysis of the number of embryos produced by Holstein donors. J. Dairy Sci. 101:7248–7257. doi:10.3168/jds.2017-13848.

Jiang, L., J. Liu, D. Sun, P. Ma, X. Ding, Y. Yu, and Q. Zhang. 2010. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. PLoS One 5:e13661. doi:10.1371/journal.pone.0013661.

Kadarmideen, H.N., and G. Mazzoni. 2019. Transcriptomics–genomics data integration and expression quantitative trait loci analyses in oocyte donors and embryo recipients for

improving invitro production of dairy cattle embryos. Reprod. Fertil. Dev. 31:55. doi:10.1071/RD18338.

Kussano, N.R., L. de O. Leme, and M.A.N. Dode. 2021. Protein source in maturation media affects gene expression in cumulus cells and embryo development in cattle. Anim. Biotechnol. 1–14. doi:10.1080/10495398.2021.2019755.

Lin, T., L. Sun, J.E. Lee, S.Y. Kim, and D. Il Jin. 2021. DNA damage repair is suppressed in porcine aged oocytes. J. Anim. Sci. Technol. 63:984–997. doi:10.5187/jast.2021.e90.

Ma, L., J.B. Cole, Y. Da, and P.M. VanRaden. 2019. Symposium review: Genetics, genome-wide association study, and genetic improvement of dairy fertility traits. J. Dairy Sci. 102:3735–3743. doi:10.3168/jds.2018-15269.

Milanesi, M., A.M.P. Obrien, A.T. Utsunomiya, L.F. Feres, T.S. Sonstegard, and J.F. Garcia. 2022. Genomic breed composition of pure registered Brazilian Gir. Trop. Anim. Health Prod. 54. doi:10.1007/s11250-022-03299-0.

Misztal, I., S. Tsuruta, T. Strabel, B. Auvray, T. Druet, and D.H. Lee. 2002. BLUPF90 and related programs (BGF90). Montpellier, France.

Mito, T., K. Yoshioka, M. Noguchi, S. Yamashita, and H. Hoshi. 2013. Recombinant human follicle-stimulating hormone and transforming growth factor-alpha enhance in vitro maturation of porcine oocytes. Mol. Reprod. Dev. 80:549–560. doi:10.1002/mrd.22190.

NCBI. 2023. National Center for Biotechnology Information. Accessed Accessed Feb. 14, 2023. https://www.ncbi.nlm.nih.gov/

Ocampo, A., J. Pedraza, G. Ortiz, E. Hernández-Pérez, L. Porchia, and E. López-Bayghen. 2018. Assessment of Prostaglandin-Endoperoxide Synthase 2 and Versican gene expression profile from the cumulus cells: association with better in vitro fertilization outcomes. J. Ovarian Res. 11:84. doi:10.1186/s13048-018-0456-2.

Oliveira Júnior, G.A., D.J.A. Santos, A.S.M. Cesar, S.A. Boison, R. V. Ventura, B.C. Perez, J.F. Garcia, J.B.S. Ferraz, and D.J. Garrick. 2019. Fine mapping of genomic regions associated with female fertility in Nellore beef cattle based on sequence variants from segregating sires. J. Anim. Sci. Biotechnol. 10. doi:10.1186/s40104-019-0403-0.

Otto, P.I., S.E.F. Guimarães, M.P.L. Calus, J. Vandenplas, M.A. Machado, J.C.C. Panetto, and M.V.G.B. da Silva. 2020. Single-step genome-wide association studies (GWAS) and post-GWAS analyses to identify genomic regions and candidate genes for milk yield in Brazilian Girolando cattle. J. Dairy Sci. 103:10347–10360. doi:10.3168/jds.2019-17890.

Panetto, J.C, do C., Silva, M.V.G.B., Verneque, R. da S., Machado, M.A., Fernandes, A.R., Machado, C.H.C., Martins, M.F., Reis, D.R. de L., Borges, C.A.V., Oliveira, J.C. de,

Ventura, H.T., Pereira, M.A., Garcia, A.O., Leandro, F.D. 2021. Programa Nacional de Melhoramento do Gir Leiteiro - Sumário Brasileiro de Touros - 4a Avaliação Genômica de Touros - Resultado do Teste de Progênie - Juiz de Fora.

Park, Y.-G., S.-E. Lee, Y.-J. Son, S.-G. Jeong, M.-Y. Shin, W.-J. Kim, E.-Y. Kim, and S.-P. Park. 2018. Antioxidant β-cryptoxanthin enhances porcine oocyte maturation and subsequent embryo development in vitro. Reprod. Fertil. Dev. 30:1204. doi:10.1071/RD17444.

Rocha, R. de F.B., A.O. Garcia, P.I. Otto, M.G. dos Santos, M.V.B. da Silva, M.F. Martins, M.A. Machado, J.C. do C. Panetto, and S.E.F. Guimarães. 2023. Single-step genome-wide association studies and post-GWAS analyses for the number of oocytes and embryos in Gir cattle. Mamm. Genome 34:497–508. doi:10.1007/s00335-023-10009-0.

Rocha, R.D.F.B., P.I. Otto, M.V.G.B. da Silva, M.F. Martins, M.A. Machado, R. Veroneze, F.D. Leandro, S.N. Pereira, S.E.F. Guimarães, and J.C.D.C. Panetto. 2022. Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed. Anim. Prod. Sci. 62:1661–1670. doi:10.1071/AN21588.

Ruis, B., A. Molan, T. Takasugi, and E.A. Hendrickson. 2020. Absence of XRCC4 and its paralogs in human cells reveal differences in outcomes for DNA repair and V(D)J recombination. DNA Repair (Amst). 85:102738. doi:10.1016/j.dnarep.2019.102738.

Sargolzaei, M., J.P. Chesnais, and F.S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15. doi:10.1186/1471-2164-15-478.

Shen, Q., M. Chen, X. Zhao, Y. Liu, X. Ren, and L. Zhang. 2020. Versican expression level in cumulus cells is associated with human oocyte developmental competence. Syst. Biol. Reprod. Med. 66:176–184. doi:10.1080/19396368.2020.1725685.

Silva, A.A., A.L.S. Azevedo, R.S. Verneque, K. Gasparini, M.G.C.D. Peixoto, M.V.G.B. da Silva, P.S. Lopes, S.E.F. Guimarães, and M.A. Machado. 2011. Quantitative trait loci affecting milk production traits on bovine chromosome 6 in zebuine Gyr breed. J. Dairy Sci. 94:971–980. doi:10.3168/jds.2009-2970.

Stapane, L., N. Le Roy, M.T. Hincke, and J. Gautron. 2019. The glycoproteins EDIL3 and MFGE8 regulate vesicle-mediated eggshell calcification in a new model for avian biomineralization. J. Biol. Chem. 294:14526–14545. doi:10.1074/jbc.RA119.009799.

Stephens, M.A. 1986. Tests based on EDF statistics. R.B. D'Agostino and M.A. Stephens, ed. Marcel Dekker, New York.

Thode, H.C. 2002. Testing For Normality. 1st Edition. Marcel-Dekker, New York.

Tiezzi, F., K.L. Parker-Gaddis, J.B. Cole, J.S. Clay, and C. Maltecca. 2015. A Genome-Wide Association Study for Clinical Mastitis in First Parity US Holstein Cows Using Single-Step Approach and Genomic Matrix Re-Weighting Procedure. PLoS One 10:e0114919. doi:10.1371/journal.pone.0114919.

Vargas, G., H.H.R. Neves, G.M.F. Camargo, V. Cardoso, D.P. Munari, and R. Carvalheiro. 2018. Genome-wide association study and functional analysis of feet and leg conformation traits in Nellore cattle. J. Anim. Sci. 96:1617–1627. doi:10.1093/jas/sky079.

Weller, J.I., Y. Kashi, and M. Soller. 1990. Power of Daughter and Granddaughter Designs for Determining Linkage Between Marker Loci and Quantitative Trait Loci in Dairy Cattle. J. Dairy Sci. 73:2525–2537. doi:10.3168/jds.S0022-0302(90)78938-2.

Willforss, J., J.M. Morrell, S. Resjö, T. Hallap, P. Padrik, V. Siino, D.J. de Koning, E. Andreasson, F. Levander, and P. Humblot. 2021. Stable bull fertility protein markers in seminal plasma. J. Proteomics 236:104135. doi:10.1016/j.jprot.2021.104135.

ORCIDS

Renata de Fátima Bretanha Rocha https://orcid.org/0000-0002-0552-7042

Arielly Oliveira Garcia https://orcid.org/0000-0001-6959-8117

Mateus Guimarães dos Santos https://orcid.org/0000-0002-6475-9951

Pamela Itajara Otto https://orcid.org/0000-0003-1702-4626

Marcos Vinícius Barbosa da Silva https://orcid.org/0000-0001-5449-1413

Marta Fonseca Martins https://orcid.org/0000-0002-4260-5329

Marco Antonio Machado https://orcid.org/0000-0002-5868-851X

João Claudio do Carmo Panetto https://orcid.org/0000-0002-9198-9728

Mario P.L. Calus https://orcid.org/0000-0002-3213-704X

Simone Eliza Facioni Guimarães https://orcid.org/0000-0003-3704-8131

Parent-of-origin effects for the number of oocytes and embryos in Gir cattle*

* Published at Livestock Science (https://doi.org/10.1016/j.livsci.2024.105412)

Renata de Fátima Bretanha Rocha [a,b], Arielly Oliveira Garcia [a], Mateus Guimarães dos Santos [a], Pamela Itajara Otto [c], Marcos Vinícius Barbosa da Silva [d], Marta Fonseca Martins [d], Marco Antônio Machado [d], João Claudio do Carmo Panetto [d], Mario P. L. Calus [b], Jeremie Vandenplas [b], Simone Eliza Facioni Guimarães [a]*

[a]Department of Animal Science, Universidade Federal de Viçosa, Viçosa, MG, 36570-900, Brazil, e-mail: renata.bretanha@ufv.br, arielly.garcia@ufv.br, mateus.g.santos@ufv.br, sfacioni@ufv.br

[b]Animal Breeding and Genomics, Wageningen University & Research, 6700 AH Wageningen, the Netherlands, e-mail: mario.calus@wur.nl, jeremie.vandenplas@wur.nl

[c]Department of Animal Science, Universidade Federal de Santa Maria, Santa Maria, RS, 97105-900, Brazil, e-mail: pamela.otto@ufsm.br

[d]EMBRAPA - Dairy Cattle, Juiz de Fora, MG, 36038-330, Brazil, e-mail: joao.panetto@embrapa.br, marcos.vb.silva@embrapa.br, marta.martins@embrapa.br, marco.machado@embrapa.br

*Corresponding author: Simone Eliza Facioni Guimarães, sfacioni@ufv.br

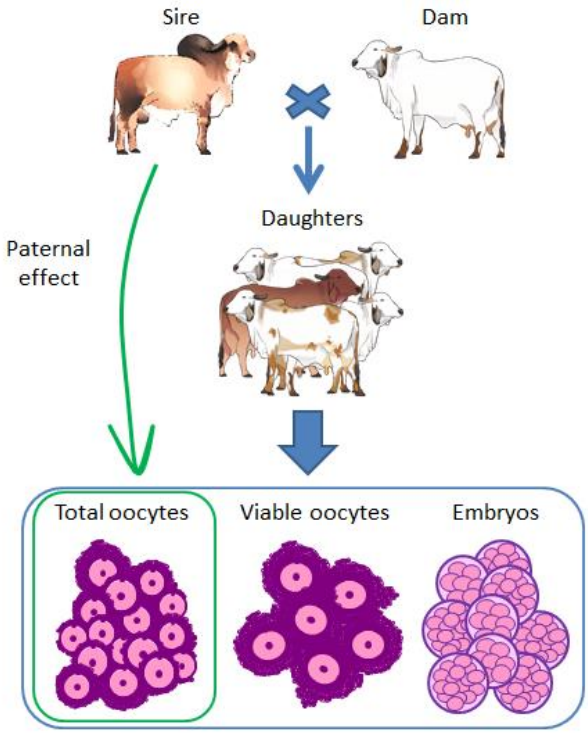
Abstract

Imprinting is a phenomenon that alters the expression of genes according to the parental origin of their alleles. A quantitative form to evaluate the imprinting effect is known as parent-of-origin effect. Our aim with this work is to identify parent-of-origin effects that influence the number of oocytes and embryos in Gir dairy cattle. A dataset with 17,526 Ovum Pick Up observations from 1,641 Gir donors was used to estimate parent-of-origin effects for the traits number of total oocytes (TO), number of viable oocytes (VO) and number of embryos (EM). To identify parent-of-origin effects, dam and sire gametic effects were included, individually or together, in an animal model for TO, VO and EM traits. For TO, inclusion of paternal origin effects in the model was significant ($P < 0.05$), and explained 6% of the total phenotypic variance. For VO and EM no significant parent-of-origin effects were found for either parental line. In conclusion, paternal effects appear to influence the total oocyte production in the Gir cattle breed.

Keywords: dairy cattle; gametic matrix; paternal effect; reproductive traits

Graphical abstract

Paternal imprinting effect on the number of total oocytes in Gir cattle



1. Introduction

Parent-of-origin effects occur when the expression level of an allele depends on parental inheritance (Lawson et al., 2013). Although several factors can cause parent-of-origin effects, genomic imprinting, which arises due to DNA-modifications during the gametogenesis process, is the best characterized (Blunk et al., 2017a, 2019). Imprinting is an epigenetic process that involves complete or partial inactivation of a gene inherited from one parent while its homologous gene inherited from the other parent is expressed (Hitchcock and Gardner, 2019). Although parent-of-origin effects plays a critical role in a variety of biological processes, including agricultural important traits, its effects are not commonly included in animal breeding programs (Blunk et al., 2017b; Lawson et al., 2013). Tier and Meyer (2012) mentioned that the lack of inclusion of parent-of-origin effects in breeding programs may result in biased estimated breeding values and genetic parameters. Given the importance of these effects, in the past years, analyses to identify them have been widely explored in several species, such as Arabidopsis (Hornslien et al., 2019), humans (Pilvar et al., 2019), sheep (Amiri Roudbar et al., 2018, 2017), goat (Mokhtari et al., 2022), horses (Perdomo-González et al., 2023), cattle (Blunk et al., 2019; Tier and Meyer, 2012), and swine (Neugebauer et al., 2010a). Using a genomic approach, Jiang et al. (2017) found significant contributions of imprinting effects for production and reproduction traits in the Holstein breed and indicated that imprinting effects contribute proportionally more to reproduction traits than to production traits. Using a quantitative approach, Amiri Roudbar et al. (2018) found parent-of-origin effect for reproductive traits in sheep. Some of the studies mentioned above worked from a quantitative perspective, identifying parent-of-origin effects based on phenotype data and statistical models (Amiri Roudbar et al., 2018, 2017; Mokhtari et al., 2022; Neugebauer et al., 2010a; Perdomo-González et al., 2023; Tier and Meyer, 2012), while other studies used a genomic approach, identifying imprinting patterns in genomic regions (Blunk et al., 2019; Hornslien et al., 2019; Jiang et al., 2017; Pilvar et al., 2019).

In the review made by Li and Li (2019) it is notable that the study of embryo development had a significant contribution to the discovery of genomic imprinting. They mention that, according to nuclear transfer experiments, it is highly probable that genomic imprinting of parental genomes plays a crucial role in ensuring complete embryonic development. Likewise, maternal and paternal imprinting was also discovered to affect oocyte growth in mice (Hiura et al., 2006; Joh et al., 2018). The importance of studying embryo and oocyte-related traits, especially in cattle, comes from the rising embryo industry, which

reached over one million in vitro-produced (IVP) bovine embryos worldwide in 2021 (Viana, 2022). Furthermore, the successful embryogenesis process depends on oocyte growth and development (Walker and Biase, 2020), making it important to also study oocyte-related traits and its viability.

In recent years, studies have emerged exploring genetic and genomic aspects of traits related to oocyte and embryo production in cattle breeds (Jaton et al., 2020; Parker Gaddis et al., 2017; Perez et al., 2016; Rocha et al., 2022). For example, the Gir cattle, a breed originating from India, is also known for its resilience against high temperatures and parasites, and stands out as an exceptionally productive indicine breed for milk production in tropical regions such as Brazil (Panetto et al., 2021). Bos indicus breeds produce twice the oocyte production in comparison to Bos taurus cattle, demonstrating their potential for in vitro embryo production (Rotar and Souza, 2019). The Gir breed, as a Bos primigenius indicus, has been recognized for its utilization in reproductive technologies, achieving superior oocyte quality when compared to Bos primigenius taurus. This distinction has enabled the Gir breed to outperform taurine animals in terms of in vitro embryo production.

In our previous study (Rocha et al., 2022), we estimated variance components and genetic parameters for the number of viable oocytes, the number of total oocytes and the number of embryos, using an animal model based on repeatability and random regression approaches. However, considering the importance of oocyte and embryo-related traits for livestock production, there is a lack of studies estimating parent-of-origin effects related to the number of oocytes and embryos in cattle. Therefore, complementing our previous research (Rocha et al., 2022), our hypothesis in the current study is that parent-of-origin effects may affect oocyte and embryo production in Gir dairy cattle. Our aim with this study is to identify parent-of-origin effects that influence the number of oocytes and embryos in Gir dairy cattle.

2. Material and Methods

2.1. Data structure

A dataset with 17,526 Ovum Pick Up (OPU) observations from 1,641 Gir donors previously described by Rocha et al. (2022) was used. The dataset came from five commercial farms including the following traits: the number of viable oocytes (VO), the number of total oocytes (TO), and the number of embryos (EM). The collected oocytes were sent to the lab for quantification of viable and non-viable oocytes. Viable oocytes were fertilized in the in vitro fertilization (IVF) process. The EM trait is related to the total number of embryos produced after IVF, regardless of their quality or stage of development. For quality control,

the following criteria were applied for sample exclusion: donors without correct identification, pooled donor samples, donors from other breeds, duplicate data and samples in which the number of embryos was greater than the number of total oocytes. Donors out of the age range between 1 and 16 years were excluded, because they were represented by very few observations. Only records from cows with at least three observations were considered. Descriptive statistics for the traits after quality control are available in Table 1.

Table 4. Descriptive statistics[A] for the number of viable oocytes (VO), number of total oocytes (TO), and number of embryos (EM).

Traits	N	Mean	SD	MIN	MAX	CV
VO	17,524	15.82	12.89	0	182	81.46
TO	17,526	21.66	16.10	0	197	74.33
EM	17,498	2.94	3.70	0	46	125.66

[A]**N**: number of data samples; **SD**: standard deviation; **MIN**: minimum; **MAX**: maximum; **CV**: coefficient of variation. Source: Adapted from Rocha et al. (2022) to include only the traits.

The OPU sessions were performed from 2005 to 2020 in four seasons: (1) January–March; (2) April–June; (3) July–September; (4) October–December. The farms hired different companies to carry out the processes ranging from Ovum Pick Up to in vitro fertilization (OPU-IVF). In some cases, one farm hired more than one company to carry out these processes. Each company has its own protocol for OPU, IVF and other laboratory processes. In total, eight OPU-IVF protocols were applied varying among farms and in a few cases within farms. The number of OPU sessions ranged from 3 to 70 per donor, with a mean (standard deviation) of 10.68 (8.91). The interval (in days) between OPU sessions for each animal had a mean (standard deviation) of 97.01 (171.10). A minimum of 0 days for OPU interval indicated that two samples of the OPU session from the same cow on the same day were obtained and sent to the lab to be fertilized with semen from different bulls. The maximum OPU interval was 2,570 days. Donor's age ranged from 1 to 16 years, with a mean (standard deviation) of 6.27 (3.19).

Animals were separated into 138 contemporary groups (CG) composed according to year, season, farm and OPU/In Vitro Fertilization protocol. A total of 238 bulls (Gir = 84 or Holstein = 154) were used for the in vitro fertilization processes. The frequency of use of these bulls ranged from 1 to 1789 samples, with an average of 73 samples per bull.

2.2. Gametic relationship matrix

The pedigree of the 1,641 donors included 127 sires and 771 dams, with a total of 4,679 animals covering up to 15 generations. Some of these Gir sires have semen samples available for purchase through commercial sire summaries, so several farms may have access to some of the same sires for reproduction purposes. A pedigree-based inbreeding was calculated with the "pedigree" package (Coster, 2012) using the R software (R Core Team, 2023 – R version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria). Average inbreeding for the whole population was 0.015 and average inbreeding for the donors was 0.017. Additional details for the pedigree are available in Table 2.

Table 2. Descriptive statistics of the pedigree used in the analyses.

Item	Value
Total number of individuals	4,679
Total number of donors	1,641
Average inbreeding for the whole population	0.015
Average inbreeding for the donors	0.017
Percentage of individuals with known sire	88.89%
Percentage of individuals with known dam	88.27%
Percentage of individuals with known sire and dam	87.67%
Percentage of donors with known sire	100%
Percentage of donors with known dam	99.57%
Percentage of donors with known sire and dam	99.57%

To estimate the parent-of-origin effects, the gametic relationship matrix approach was used. This matrix was obtained from the gametic pedigree, which in turn was obtained from the animals' original pedigree file (Tier and Meyer, 2012). Then, the inverse of the genetic relationship matrix (Schaeffer et al., 1989) was obtained using the GGRinv program (Reinsch, 2020).

2.3. Parent-of-origin effects

Traits were transformed using the logarithmic scale: ln (X + 1), where ln is the natural logarithm and X is the raw data (oocyte and embryos). The normal distribution of the residuals was verified by the Anderson–Darling test (Stephens, 1986; Thode, 2002). The animal model was defined in one of our previous studies (Rocha et al. 2022). In brief, analysis

of variance was carried out to test for the significance of the effects (Table 3). After that, several models were tested for each trait and compared using Log Likelihood and Akaike Information Criterion values to choose the model with the best fit.

Table 3. Analyses of variance results (P-values) indicating significant effects for the number of viable oocytes (VO), the number of total oocytes (TO), and the number of embryos (EM) in the animal model.

Effects	**Type of effect**	**P-values**		
		VO	**TO**	**EM**
Contemporary groups	Fixed	0.000	0.000	0.000
Bull	Fixed	--[A]	--	0.000
Bull breed	Fixed	--	--	0.039
OPU interval (days)	Linear covariate	0.000	0.000	0.868
Donor's age at OPU (years)	Linear covariate	0.087	0.386	0.000
Donor's age at OPU (years)	Quadratic covariate	0.000	0.000	0.000

[A]Not included in the model for the given trait.

In the current study, paternal and maternal gametic effects were included in the animal model previously defined. Single-trait models were used to estimate paternal and/or maternal effects on the traits:

Animal model (Am) – no parental effect:

$$y = X\beta + Z_a a + \wp + e$$

Animal + Dam gametic effect model (ADm):

$$y = X\beta + Z_a a + Z_d g_d + \wp + e$$

Animal + Sire gametic effect model (ASm):

$$y = X\beta + Z_a a + Z_s g_s + \wp + e$$

Animal + Dam gametic effect + Sire gametic effect model (ADSm):

$$y = X\beta + Z_a a + Z_d g_d + Z_s g_s + \wp + e$$

where, $\boldsymbol{y}$, $\boldsymbol{\beta}$, $\boldsymbol{a}$, g_d, g_s, $\boldsymbol{p}$, and $\boldsymbol{e}$ are the vectors of observations, fixed effects, additive genetic random effects, dam gametic random effects, sire gametic random effects, permanent environmental random effects and residual effects, respectively; $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{W}$ are the incidence matrices of fixed and random effects, respectively. For VO and TO, the fixed effect CG was

considered. For EM, the fixed effects of CG, bull and bull breed used in *in vitro* fertilization were considered. The bull used in the IVF process as well as the breed of this bull were tested for EM, but not for TO and VO, as IVF takes place after oocyte counting. Donor's age in days (linear and quadratic components) was included as a covariate for all traits. OPU interval (linear component) was included for VO and TO.

The animal, permanent environmental and residual effects were assumed to be random, i.e. $\boldsymbol{a} \sim N(0, A\sigma_a^2)$, $\boldsymbol{p} \sim N(0, I\sigma_p^2)$, $\boldsymbol{g} \sim N(0, G\sigma_g^2)$, and $\boldsymbol{e} \sim N(0, I\sigma_e^2)$, where σ_a^2, σ_p^2, σ_g^2 and σ_e^2 are the additive genetic, permanent environmental, gametic and residual variances, respectively; $\boldsymbol{A}$ is the relationship matrix, $\boldsymbol{I}$ is the identity matrix, and $\boldsymbol{G}$ is the gametic relationship matrix. ASREML software v.4.1 (Gilmour et al., 2015) was used to simultaneously estimate the variance components, and the random and fixed effects using Residual Maximum Likelihood. The ASREML output also provided Log Likelihood (LogL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) values. Delta Log Likelihood values for ADm, ASm and ADSm were calculated relative to the Am model. Chi-square test statistics were obtained as twice Delta Log Likelihood values. Significance of parent-of-origin effects were verified based p-values for the LogL ratio test, calculated from the chi-square test statistics with the 'pchisq' function in R software (R Core Team, 2023, R version 4.2.2; R Foundation for Statistical Computing, Vienna, Austria) considering 1 degree of freedom.

3. Results

Table 4 contains the estimated variance components for TO, VO, and EM, considering animal models with or without the inclusion of parental effects. Based on AIC and BIC values, the best fit for VO was the animal model (Am). For VO, the Am model estimated an additive genetic variance of 0.13, a permanent environmental variance of 0.08, and a residual variance of 0.31. For TO, AIC values indicated the ASm model as the best fit, BIC values designed Am as the best fit. Nevertheless, ASm and ADSm model for TO had very similar LogL values. For TO, ASm estimated an additive genetic variance of 0.12, a permanent environmental variance of 0.05, and a residual variance of 0.30. For EM, Log Likelihood ratio test and AIC values defined ASm as the best fit, which estimated an additive genetic variance of 0.05, a permanent environmental variance of 0.05 and a residual variance of 0.45.

The estimated total phenotypic variance for VO ranged from 0.52 to 0.53, for TO ranged from 0.49 to 0.50, and was 0.56 for EM in all models. A significant paternal effect was

found for TO. For VO and EM no parental effects were found for either parental line ($P > 0.05$). For VO and EM, maternal and paternal variances were close to zero. For TO, the paternal variance was greater than maternal variance. Also, for TO, the paternal variance represented 6% of the total phenotypic variance in the ASm and 8% in the ADSm.

Table 4. Models[A] and variance components[B] obtained from the analyses for the number of viable oocytes (VO), number of total oocytes (TO), and number of embryos (EM).

Trait	Model	LogL	χ^2	P-val	σ_p^2 (se)	σ_a^2 (se)	σ_{pe}^2 (se)	σ_e^2 (se)	σ_d^2 (se)	σ_s^2 (se)	AIC	BIC
VO	Am	-337.16	0.00	--	0.52 (0.01)	0.13 (0.02)	0.08 (0.01)	0.31 (0.00)	--	--	680.32	703.60
	ADm	-337.11	0.11	0.74	0.52 (0.01)	0.13 (0.02)	0.08 (0.01)	0.31 (0.00)	0.00 (0.01)	--	682.21	713.26
	ASm	-336.53	1.26	0.26	0.53 (0.01)	0.13 (0.02)	0.07 (0.01)	0.31 (0.00)	--	0.01 (0.01)	681.07	712.11
	ADSm	-336.53	1.26	0.26	0.53 (0.01)	0.13 (0.02)	0.07 (0.01)	0.31 (0.00)	0.00 (0.00)	0.01 (0.01)	683.07	721.87
TO	Am	53.33	0.00	--	0.49 (0.01)	0.12 (0.02)	0.07 (0.01)	0.30 (0.00)	--	--	-100.66	-77.38
	ADm	54.02	1.37	0.24	0.49 (0.01)	0.11 (0.02)	0.07 (0.01)	0.30 (0.00)	0.01 (0.01)	--	-100.04	-68.99
	ASm	55.83	5.01	0.03	0.50 (0.02)	0.12 (0.02)	0.05 (0.02)	0.30 (0.00)	--	0.03 (0.02)	-103.67	-72.63
	ADSm	55.95	5.25	0.02	0.50 (0.02)	0.10 (0.03)	0.05 (0.02)	0.30 (0.00)	0.01 (0.02)	0.04 (0.02)	-101.91	-63.10
EM	Am	-3272.37	0.00	--	0.56 (0.01)	0.06 (0.01)	0.05 (0.01)	0.45 (0.01)	--	--	6550.74	6573.99
	ADm	-3272.37	0.00	1.00	0.56 (0.01)	0.06 (0.01)	0.05 (0.01)	0.45 (0.01)	0.00 (0.00)	--	6552.74	6583.73
	ASm	-3271.19	2.36	0.12	0.56 (0.01)	0.05 (0.01)	0.05 (0.01)	0.45 (0.01)	--	0.01 (0.01)	6550.39	6581.38
	ADSm	-3271.19	2.36	0.12	0.56 (0.01)	0.05 (0.01)	0.05 (0.01)	0.45 (0.01)	0.00 (0.00)	0.01 (0.01)	6552.39	6591.12

[A]Animal model (Am); Animal + Dam gametic effect model (ADm); Animal + Sire gametic effect model (ASm); Animal + Dam gametic effect + Sire gametic effect model (ADSm)

[B]LogL: Final Log Likelihood; χ^2: Chi-square statistic test; σ_p^2: phenotypic variance; σ_a^2: additive genetic variance; σ_{pe}^2: permanent environmental variance; σ_e^2: residual variance; σ_d^2: Dam variance; σ_s^2: sire variance; se: standard error; AIC: Akaike Information Criterion; BIC: Bayesian Information Criterion.

4. Discussion

In this study, we aimed to identify parent-of-origin effects that influence the number of oocytes and embryos in Gir dairy cattle and found a significant paternal effect for the number of total oocytes. It is important to reiterate that although imprinting effects are often referred to as parent-of-origin effects, they are not synonymous, that is, imprinting is an epigenetic phenomenon that alters the expression of genes according to the parental origin of their alleles, being one of possible the causes for parent-of-origin effects (Blunk et al., 2017b; Tier and Meyer, 2012).

Possible processes underlying parent-of-origin effects starts at the pre-implantation phase of embryos, where methylation imprinting can occur in differentially methylated regions of the germ cell lines. Methylation can occur, for example, by the replacement of protamines by histones in the paternal genome, active demethylation of the paternal genome and subsequent passive demethylation of both parental genomes (Li and Sasaki, 2011). Also, during embryo development towards fetal formation, the oocyte stock of mammalian females is formed, meaning that when females are born, they already carry all oocytes that will be available for use during adulthood (Seneda et al., 2021). Therefore, it is possible that the methylation imprinting process occurring during embryo development can influence the number of oocytes produced by each fetus. This might explain the significant paternal effect that we found for the number of total oocytes.

In addition, it has been reported that reproductive technology protocols may induce epigenetic changes in imprinted genes (Mangiavacchi et al., 2021). In the current study, part of the donors was produced by the use of reproductive technologies, such as in vitro fertilization. Mangiavacchi et al. (2021) mention that embryos generated by assisted reproduction technologies may show morphological abnormalities at different stages of embryonic development as well as in the postnatal period associated with hypergrowth, which can indicate epigenetic failures. This suggests that specific epigenetic events may have occurred during donors' formation, and we speculate that this may have affected their oocytes production.

It is known that during mouse embryo development, imprints are acquired in a sex-specific manner in the mature germline, with the establishment of paternal imprints occurring prenatally and maternal imprints occurring postnatally (Plasschaert and Bartolomei, 2014). The zygote receives the paternal and maternal epigenetic imprints that were formed in the germline through fertilization, and these imprints are preserved throughout growth and adulthood (Li and Sasaki, 2011). In cattle, DNA-methylated regions identified in the paternal

germline were related to the fertility of bulls and their daughters (Costes et al., 2022; Y. Zhang et al., 2023a, 2023b). DNA methylation patterns from bulls used in artificial insemination were found in genomic regions covering genes related to spermatogenesis and early embryonic development (Costes et al., 2022; Štiavnická et al., 2022). Zhang et al. (2023a) found that most of the methylated regions in bovine sperm samples were distributed on the X and Y chromosomes, demonstrating that the sex chromosomes play essential roles in bull fertility. In another study, Zhang et al. (2023b) selected 12 bulls classified with high or low daughter fertility to study their enzymatic methylation and found methylation profiles in sperm samples distributed more on the X chromosome than on the autosome, associated with daughter fertility. The daughter fertility index studied by Zhang et al. (2023b) is a combination of reproductive traits, such as age at first service, the 56-day non-return rate for heifers/cows, the interval from calving to the first service, and the interval from insemination to conception (heifer/cow). This indicates that paternal epigenetic effects are associated with daughter fertility.

Nevertheless, regarding imprinting genes and parent-of-origin effects reported for cattle, only one work was found for parent-of-origin effect in the Catalogue of Parent of Origin Effects (Morison et al., 2005, www.otago.ac.nz/IGC), which is related to milk production (Kuehn et al., 2007). This demonstrates the need for more studies exploring imprinting and parent-of-origin effects in bovine and other species. In the current study, parent-of-origin effects for the number of total oocytes, viable oocytes and embryos in dairy Gir cattle were explored using a gametic relationship approach (Tier and Meyer, 2012). In our previous work (Rocha et al., 2022), we had the same population data sample, but used repeatability and random regression models on BLUPF90 family programs (Misztal et al., 2002). The difference from our previous study to this one is the addition of parental gametic effects, where parent-of-origin effects can be identified, which is not possible using only the animal model. The identification of a parental effect can influence the selection of animals, that is, the focus can be given to the sire or the dam at the time of selection for a given trait. In the current study, ASREML software (Gilmour et al., 2015) was used since it was easier to include the gametic relationship matrices.

Despite the use of different programs, variance components were similar between studies. The additive genetic variance found by Rocha et al. (2022) using BLUPF90 family programs was 0.13 for VO, 0.12 for TO and 0.06 for EM. In the current work, using ASREML software, the additive genetic variance was 0.13 for VO regardless the model, it ranged from 0.10 to 0.12 for TO among models and it was 0.05 or 0.06 for EM. The

permanent environmental variance found by Rocha et al. (2022) for VO was 0.08 and in this work it was 0.07 or 0.08 depending on the model. For TO, Rocha et al. (2022) found a permanent environmental variance of 0.07 and in this work it was 0.05 or 0.07. For EM, the permanent environmental variance found was 0.05 in both studies. Values of residual variance were 0.31 for VO, 0.30 for TO and 0.45 for EM in both studies. The existence of genetic variation for oocytes and embryos allow for selection for these traits in the Gir breed. Also, greater values of residual variance than additive genetic variance were expected considering that the data came from five farms and different Ovum Pick Up/in vitro fertilization processes were used.

Regarding the parent-of-origin variances, most parent-of-origin studies were performed for meat and carcass traits in pigs (de Vries et al., 1994; Neugebauer et al., 2010a) and cattle (Blunk et al., 2017a, 2017c, 2017b; Engellandt and Tier, 2002; Inoue et al., 2021; Neugebauer et al., 2010b; Okamoto et al., 2019; Tier and Meyer, 2012). In addition to production traits, Amiri Roudbar et al. (2018) applied the gametic relationship matrix approach to study reproduction traits in sheep and found maternal imprinting effects for total litter weight at weaning per ewe lambing. Tier and Meyer (2012) mention that most models applied for genomic selection consider additive genetic effect with an equal contribution from both parents, which undervalue loci modified by their parents for heterozygous offspring and could lead to an inefficient selection.

To the best of our knowledge, there are currently no other parent-of-origin studies for oocyte and embryo related-traits in cattle. Although the number of oocytes and embryos are not yet included in genetic evaluations of the Dairy Gir National Breeding Program, Gir cattle producers collect this information from Gir donors in order to assess the reproductive quality of the herd and increase oocyte and embryo productivity. The present study identifies the existence of parent-of-origin effects on the number of oocytes and embryos, which has never been explored in the Gir breed. The results of this work may help to reinforce the existence of imprinting effects on traits of economic impact in livestock farming, since these effects are little explored in animal breeding programs. These results may also contribute to future research to verify the expression of parental alleles on the production of oocytes. Furthermore, this study opens the possibility of including them in breeding programs due to their economic impact. It is worth noting that our study is a first step to identify the presence or absence of parent-of-origin effects on the number of oocytes and embryos. Future studies using molecular approaches might contribute with new insights; for example, identifying imprinted genomic regions from each parental line.

5. Conclusion

We found a paternal effect for the number of total oocytes in Gir cattle. The paternal effect explained 6% of the total phenotypic variance. Considering the importance of reproductive traits for livestock production and taking into account that we found a paternal effect for the number of total oocytes emphasizes the need for studies to evaluate the parent-of-origin effects of alleles for reproductive traits in cattle. Furthermore, statistical models considering only additive effects may underestimate parental genomic imprinting modifications from the animals under genetic evaluation.

CRediT authorship contribution statement

Renata de Fátima Bretanha Rocha: Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Arielly Oliveira Garcia**: Investigation, Visualization. **Mateus Guimarães dos Santos**: Investigation, Visualization. **Pamela Itajara Otto**: Methodology, Investigation, Visualization. **Marcos Vinícius Gualberto Barbosa da Silva**: Resources, Data curation, Visualization. **Marta Fonseca Martins**: Resources, Data curation, Visualization. **Marco Antônio Machado**: Resources, Data curation, Visualization. **João Claudio do Carmo Panetto**: Resources, Data curation, Visualization. **Mario P. L. Calus**: Methodology, Software, Validation, Formal analysis, Investigation. **Jeremie Vandenplas**: Methodology, Software, Validation, Formal analysis, Investigation. **Simone Eliza Facioni Guimarães**: Methodology, Investigation, Writing – review & editing, Supervision, Project administration.

Declaration of Competing Interest

None.

Acknowledgments

We thank the farms and Brazilian Agriculture Research Corporation (EMBRAPA) Dairy Cattle Research Center (Juiz de Fora) for providing data for this study. Coordination and Improvement of Higher Level Personnel (CAPES - 88881.844754/2023-01), National Council for Scientific and Technological Development (CNPq) – processes [402935/2021-7], [142600/2019-9] e [200147/2022-6] and Brazilian National Institute of Science and Technology in Animal Science (INCT-CA) provided financial support towards this study.

References

Amiri Roudbar, M., Abdollahi-Arpanahi, R., Ayatollahi Mehrgardi, A., Mohammadabadi, M., Taheri Yeganeh, A., Rosa, G.J.M., 2018. Estimation of the variance due to parent-of-origin effects for productive and reproductive traits in Lori-Bakhtiari sheep. Small Rumin. Res. 160, 95–102. https://doi.org/10.1016/j.smallrumres.2018.01.022

Amiri Roudbar, M., Mohammadabadi, M., Mehrgardi, A.A., Abdollahi-Arpanahi, R., 2017. Estimates of variance components due to parent-of-origin effects for body weight in Iran-Black sheep. Small Rumin. Res. 149, 1–5. https://doi.org/10.1016/j.smallrumres.2017.01.002

Blunk, I., Mayer, M., Hamann, H., Reinsch, N., 2019. Scanning the genomes of parents for imprinted loci acting in their un-genotyped progeny. Sci. Rep. 9, 654. https://doi.org/10.1038/s41598-018-36939-3

Blunk, I., Mayer, M., Hamann, H., Reinsch, N., 2017a. A parsimonious model for an analysis of parent-of-origin effects on beef traits in dual-purpose Simmental. Adv. Anim. Biosci. 8, s76–s78. https://doi.org/10.1017/S2040470017001765

Blunk, I., Mayer, M., Hamann, H., Reinsch, N., 2017b. A new model for parent-of-origin effect analyses applied to Brown Swiss cattle slaughterhouse data. Animal 11, 1096–1106. https://doi.org/10.1017/S1751731116002391

Blunk, I., Mayer, M., Hamann, H., Reinsch, N., 2017c. Parsimonious model for analyzing parent-of-origin effects related to beef traits in dual-purpose Simmental. J. Anim. Sci. 95, 559. https://doi.org/10.2527/jas2016.0997

Coster, A., 2012. pedigree: Pedigree functions. R package version 1.4. https://CRAN.R-project.org/package=pedigree.

Costes, V., Chaulot-Talmon, A., Sellem, E., Perrier, J.-P., Aubert-Frambourg, A., Jouneau, L., Pontlevoy, C., Hozé, C., Fritz, S., Boussaha, M., Le Danvic, C., Sanchez, M.-P., Boichard, D., Schibler, L., Jammes, H., Jaffrézic, F., Kiefer, H., 2022. Predicting male fertility from the sperm methylome: application to 120 bulls with hundreds of artificial insemination records. Clin. Epigenetics 14, 54. https://doi.org/10.1186/s13148-022-01275-x

de Vries, A.G., Kerr, R., Tier, B., Long, T., Meuwissen, T.H.E., 1994. Gametic imprinting effects on rate and composition of pig growth. Theor. Appl. Genet. 88, 1037–1042. https://doi.org/10.1007/BF00220813

Engellandt, T., Tier, B., 2002. Genetic variances due to imprinted genes in cattle. J. Anim. Breed. Genet. 119, 154–165. https://doi.org/10.1046/j.1439-0388.2002.00323.x

Gilmour, A., Gogel, B., Cullis, B., Welham, S., Thompson, R., 2015. ASReml user guide

release 4.1: Structural specification VSN International Ltd., Hemel Hempstead, UK.

Hitchcock, T.J., Gardner, A., 2019. Parent-of-origin specific gene expression and dispersal. Curr. Opin. Behav. Sci. 25, 36–43. https://doi.org/10.1016/j.cobeha.2018.06.007

Hiura, H., Obata, Y., Komiyama, J., Shirai, M., Kono, T., 2006. Oocyte growth-dependent progression of maternal imprinting in mice. Genes to Cells 11, 353–361. https://doi.org/10.1111/j.1365-2443.2006.00943.x

Hornslien, K.S., Miller, J.R., Grini, P.E., 2019. Regulation of Parent-of-Origin Allelic Expression in the Endosperm. Plant Physiol. 180, 1498–1519. https://doi.org/10.1104/pp.19.00320

Inoue, K., Nishio, M., Shoji, N., Hirooka, H., 2021. Effects of parent-of-origin models with different pedigree information on beef carcass traits and fatty acid composition in Japanese Black cattle. J. Anim. Breed. Genet. 138, 45–55. https://doi.org/10.1111/jbg.12493

Jaton, C., Schenkel, F.S., Chud, T.C.S., Malchiodi, F., Sargolzaei, M., Price, C.A., Canovàs, A., Baes, C., Miglior, F., 2020. Genetic and genomic analyses of embryo production in dairy cattle. Reprod. Fertil. Dev. 32, 50. https://doi.org/10.1071/RD19275

Jiang, J., Shen, B., O'Connell, J.R., VanRaden, P.M., Cole, J.B., Ma, L., 2017. Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. BMC Genomics 18, 425. https://doi.org/10.1186/s12864-017-3821-4

Joh, K., Matsuhisa, F., Kitajima, S., Nishioka, K., Higashimoto, K., Yatsuki, H., Kono, T., Koseki, H., Soejima, H., 2018. Growing oocyte-specific transcription-dependent de novo DNA methylation at the imprinted Zrsr1-DMR. Epigenetics Chromatin 11, 28. https://doi.org/10.1186/s13072-018-0200-6

Kuehn, C., Edel, C., Weikard, R., Thaller, G., 2007. Dominance and parent-of-origin effects of coding and non-coding alleles at the acylCoA-diacylglycerol-acyltransferase (DGAT1) gene on milk production traits in German Holstein cows. BMC Genet. 8, 62. https://doi.org/10.1186/1471-2156-8-62

Lawson, H.A., Cheverud, J.M., Wolf, J.B., 2013. Genomic imprinting and parent-of-origin effects on complex traits. Nat. Rev. Genet. 14, 609–617. https://doi.org/10.1038/nrg3543

Li, Y., Li, J., 2019. Technical advances contribute to the study of genomic imprinting. PLOS Genet. 15, e1008151. https://doi.org/10.1371/journal.pgen.1008151

Li, Y., Sasaki, H., 2011. Genomic imprinting in mammals: its life cycle, molecular mechanisms and reprogramming. Cell Res. 21, 466–473. https://doi.org/10.1038/cr.2011.15

Mangiavacchi, P.M., Caldas-Bussiere, M.C., Mendonça, M. da S., Dias, A.J.B., Rios, Á.F.L., 2021. Multi-locus imprinting disturbances of Beckwith-Wiedemann and Large offspring syndrome/Abnormal offspring syndrome: A brief review. Theriogenology 173, 193–201. https://doi.org/10.1016/j.theriogenology.2021.08.005

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D.H., 2002. BLUPF90 and related programs (BGF90). Montpellier, France.

Mokhtari, M., Barazandeh, A., Roudbari, Z., Bahrampour, J., Ghafouri-Kesbi, F., Roudbar, M.A., 2022. Genetic analysis of parent-of-origin effects on growth traits and yearling greasy fleece weight in Raeini Cashmere goat. Small Rumin. Res. 216, 106813. https://doi.org/10.1016/j.smallrumres.2022.106813

Morison, I.M., Ramsay, J.P., Spencer, H.G., 2005. A census of mammalian imprinting. Trends Genet. 21, 457–465. https://doi.org/10.1016/j.tig.2005.06.008

Neugebauer, N., Luther, H., Reinsch, N., 2010a. Parent-of-origin effects cause genetic variation in pig performance traits. Animal 4, 672–681. https://doi.org/10.1017/S1751731109991625

Neugebauer, N., Räder, I., Schild, H.J., Zimmer, D., Reinsch, N., 2010b. Evidence for parent-of-origin effects on genetic variability of beef traits1. J. Anim. Sci. 88, 523–532. https://doi.org/10.2527/jas.2009-2026

Okamoto, K., Oishi, K., Nakamura, R., Abe, A., Inoue, K., Kumagai, H., Hirooka, H., 2019. Parent-of-origin effects on carcass traits in Japanese Black cattle. J. Anim. Breed. Genet. 136, 190–198. https://doi.org/10.1111/jbg.12382

Panetto, J.C. do C., Silva, M.V.G.B., Verneque, R. da S., Machado, M.A., Fernandes, A.R., Machado, C.H.C., Martins, M.F., Faza, D.R. de L.R., Borges, C.A.V., Oliveira, J.C. de, Ventura, H.T., Pereira, M.A., Garcia, A.O., Leandro, F.D., 2021. Programa Nacional de Melhoramento do Gir Leiteiro - Sumário Brasileiro de Touros - 4a Avaliação Genômica de Touros - Resultado do Teste de Progênie - Maio 2021. Juiz de Fora.

Parker Gaddis, K.L., Dikmen, S., Null, D.J., Cole, J.B., Hansen, P.J., 2017. Evaluation of genetic components in traits related to superovulation, in vitro fertilization, and embryo transfer in Holstein cattle. J. Dairy Sci. 100, 2877–2891. https://doi.org/10.3168/jds.2016-11907

Perdomo-González, D.I., Varona, L., Molina, A., Laseca, N., Valera, M., 2023. Quantitative analysis of parent-of-origin effect in reproductive and morphological selection criteria in the Pura Raza Española horse. J. Anim. Breed. Genet. 140, 596–606. https://doi.org/10.1111/jbg.12811

Perez, B.C., Peixoto, M.G.C.D., Bruneli, F.T., Ramos, P.V.B., Balieiro, J.C.C., 2016. Genetic analysis of oocyte and embryo production traits in Guzerá breed donors and their associations with age at first calving. Genet. Mol. Res. 15, 1–9. https://doi.org/10.4238/gmr.15027583

Pilvar, D., Reiman, M., Pilvar, A., Laan, M., 2019. Parent-of-origin-specific allelic expression in the human placenta is limited to established imprinted loci and it is stably maintained across pregnancy. Clin. Epigenetics 11, 94. https://doi.org/10.1186/s13148-019-0692-3

Plasschaert, R.N., Bartolomei, M.S., 2014. Genomic imprinting in development, growth, behavior and stem cells. Development 141, 1805–1813. https://doi.org/10.1242/dev.101428

R Core Team, 2023. 'R: a language and environment for statistical computing.' (R Foundation for Statistical Computing: Vienna, Austria) Available at https://www.R-project.org/.

Reinsch, N., 2020. GGRinv – a Fortran program for computing the inverse of the generalized gametic relationship matrix from pedigrees. Institute of Genetics and Biometry, Leibniz-Institute for Farm Animal Biology, Dummerstorf, Germany.

Rocha, R.F.B., Otto, P.I., da Silva, M.V.G.B., Martins, M.F., Machado, M.A., Veroneze, R., Leandro, F.D., Pereira, S.N., Guimarães, S.E.F., Panetto, J.C.D.C., 2022. Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed. Anim. Prod. Sci. 62, 1661–1670. https://doi.org/10.1071/AN21588

Rotar, L.N., Souza, J.F., 2019. Morphological Characteristics for Oocyte-Cumulus Complexes of Bos taurus and Bos indicus Cattle Used for a Variety of Purposes. Russ. Agric. Sci. 45, 399–402. https://doi.org/10.3103/S1068367419040141

Schaeffer, L.R., Kennedy, B.W., Gibson, J.P., 1989. The Inverse of the Gametic Relationship Matrix. J. Dairy Sci. 72, 1266–1272. https://doi.org/10.3168/jds.S0022-0302(89)79231-6

Seneda, M.M., Bergamo, L.Z., González, S.M., Zangirolamo, A.F., Morotti, F., 2021. Oogênese e Foliculogênese em Bovinos. Rev. Bras. Reprodução Anim. 45, 323–328. https://doi.org/10.21451/1809-3000.RBRA2021.042

Stephens, M.A., 1986. Tests based on EDF statistics, in: D'Agostino, R.B., Stephens, M.A. (Eds.), Goodness-of-Fit Techniques. Marcel Dekker, New York, pp. 97–193.

Štiavnická, M., Chaulot-Talmon, A., Perrier, J.-P., Hošek, P., Kenny, D.A., Lonergan, P., Kiefer, H., Fair, S., 2022. Sperm DNA methylation patterns at discrete CpGs and genes involved in embryonic development are related to bull fertility. BMC Genomics 23, 379.

https://doi.org/10.1186/s12864-022-08614-5

Thode, H.C., 2002. Testing For Normality, 1st Edition. ed, Testing For Normality. Marcel-Dekker, New York. https://doi.org/10.1201/9780203910894

Tier, B., Meyer, K., 2012. Analysing quantitative parent-of-origin effects with examples from ultrasonic measures of body composition In Australian beef cattle. J. Anim. Breed. Genet. 129, 359–368. https://doi.org/10.1111/j.1439-0388.2012.00996.x

Viana, J., 2022. 2021 Statistics of embryo production and transfer in domestic farm animals. A new milestone has been reached: Transfers of IVP embryos were over one million worldwide. Embryo Technol. Newsl. 40, 22–40.

Walker, B.N., Biase, F.H., 2020. The blueprint of RNA storages relative to oocyte developmental competence in cattle (Bos taurus). Biol. Reprod. 102, 784–794. https://doi.org/10.1093/biolre/ioaa015

Zhang, Y., Bruna de Lima, C., Labrecque, R., Sirard, M.-A., 2023a. Whole-genome DNA methylation analysis of the sperm in relation to bull fertility. Reproduction 165, 557–568. https://doi.org/10.1530/REP-22-0283

Zhang, Y., Plessis, C., Prunier, J., Martin, H., Labrecque, R., Sirard, M.A., 2023b. DNA methylation profiles in bovine sperm are associated with daughter fertility. Epigenetics 18. https://doi.org/10.1080/15592294.2023.2280889

Artigo técnico

A genética envolvida na produção de oócitos e embriões em Gir Leiteiro*

*Publicado na revista Balde Branco (ANO 59, NÚMERO 703, DEZEMBRO 2023/JANEIRO 2024, https://baldebranco.com.br/).

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], Mario P. L. Calus[4] e 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
[4]Animal Breeding and Genomics, Wageningen University & Research, 6700 AH Wageningen, the Netherlands

Dada a importância da raça Gir para a produção leiteira nacional, é esperada uma tendência de aumento no número desses animais nos rebanhos. Uma técnica que recebe destaque no aumento do tamanho dos rebanhos e seus respectivos ganhos genéticos é a produção de embriões *in vitro* (Rotar and Souza 2019). No ano de 2021 ocorreu um marco histórico, quando mais de um milhão de embriões bovinos produzidos *in vitro* foram transferidos a nível mundial. Neste cenário, o Brasil se destacou como o segundo maior produtor de embriões, ficando atrás apenas dos Estados Unidos (Viana 2022).

Para este estudo, informações sobre o número de oócitos totais, número de oócitos viáveis e número de embriões de doadoras da raça Gir foram fornecidos por cinco fazendas de Minas Gerais. Informações de pedigree e dados genômicos foram fornecidas pela Embrapa Gado de Leite.

Produção de oócitos e embriões na raça Gir

A produção de embriões *in vitro* (PIV) é uma técnica reprodutiva que promove o aumento no ganho genético nos rebanhos, pois permitem melhor aproveitamento de matrizes de alto mérito genético, por meio da produção de maior número de bezerros que seriam produzidos

naturalmente (Becher et al. 2018). Embora existam diversos fatores que podem afetar a PIV em bovinos, as raças *Bos primigenius indicus*, como a Gir, produzem o dobro de oócitos em comparação às raças *Bos primigenius taurus*, como a Holandesa, mostrando potencial para serem utilizadas para produção de embriões *in vitro* (Rotar and Souza 2019).

Os parâmetros genéticos como a herdabilidade, repetibilidade e correlação são importantes para os programas de melhoramento e para a seleção dos animais. Na população Gir avaliada, a herdabilidade para número de oócitos totais e viáveis indicou potencial genético para melhoria destas características e a correlação genética entre as características indicou que a seleção pode ser feita para o número de oócitos totais (Rocha et al. 2022). Os resultados de repetibilidade indicaram que doadoras podem ser selecionadas em qualquer idade entre um e 16 anos. Contudo, é aconselhável realizar a seleção em idades mais jovens, com o propósito de reduzir o intervalo de geração e promover o ganho genético.

Endogamia

O acasalamento entre indivíduos aparentados dentro de uma população resulta na produção de animais conhecidos como endogâmicos. Este tipo de acasalamento pode reduzir o vigor e o desempenho produtivo e reprodutivo dos animais, como consequências de um evento conhecido como depressão endogâmica. Uma das formas de avaliar a endogamia na população é por meio da avaliação de corridas de homozigose (ROHs), que são segmentos contínuos de genótipos homozigotos (Purfield et al. 2012). Neste estudo, verificou-se que existe depressão endogâmica na raça Gir para a produção de oócitos, indicando que, com o aumento da endogamia é esperada uma queda no número de oócitos totais obtidos e, consequentemente, no número de oócitos viáveis (Rocha et al. 2023a).

Quando ocorre seleção natural ou artificial (aquela feita por intermédio humano) nos animais dentro de uma raça, surgem as assinaturas de seleção, que são padrões únicos no genoma dos indivíduos. Por meio delas podem ser identificados diversos genes que afetam fenótipos de interesse econômico na pecuária (Brito et al. 2017). Neste estudo, duas subpopulações da raça Gir foram formadas no presente trabalho, uma com alto valor genético para número de oócitos e embriões, enquanto a outra possuía baixos valores. Contrastando os dois grupos genéticos, foi possível identificar 11 genes de interesse, indicando que dentro de uma raça podem existir marcas genéticas distintas para características reprodutivas (Rocha et al. 2023a). O resumo gráfico com os resultados deste estudo está representado na Figura 1.

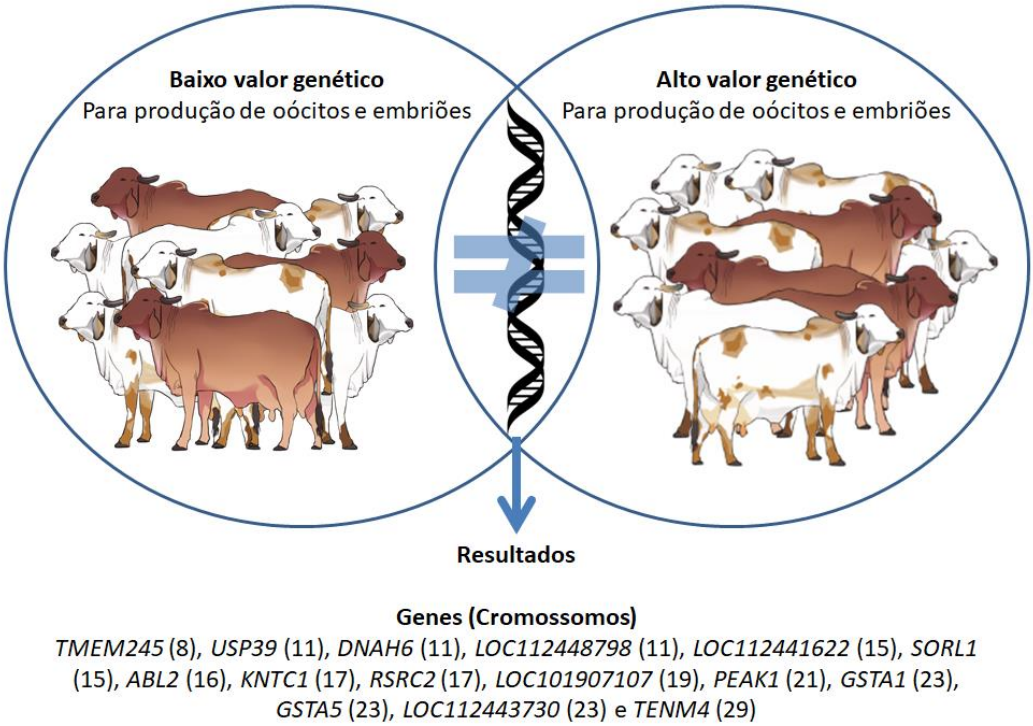


Figura 1: Resumo gráfico sobre o estudo das assinaturas de seleção. Fonte: Do autor.

Genes associados à produção de oócitos e embriões

Com o avanço das tecnologias de coleta de informações genômicas em bovinos, surgem novos estudos que auxiliam na identificação de genes associados a doenças e/ou características de interesse econômico na pecuária. A busca por essas variantes genéticas é conhecida como estudos de associação genômica ampla (GWAS). No presente estudo, 111 genes associados ao número de oócitos e embriões de doadoras Gir foram identificados (Rocha et al. 2023b). Para que os genes sejam expressos no organismo é preciso ocorrer a ativação dos mesmos, realizada por meio dos fatores de transcrição. No presente trabalho, cinco fatores de transcrição relacionados aos 111 genes candidatos foram identificados, sendo eles: ARNT, EGR1, HIF1A, AHR e PAX2. O resumo gráfico com os resultados dos genes associados à produção de oócitos e embriões está esquematizada na Figura 2.

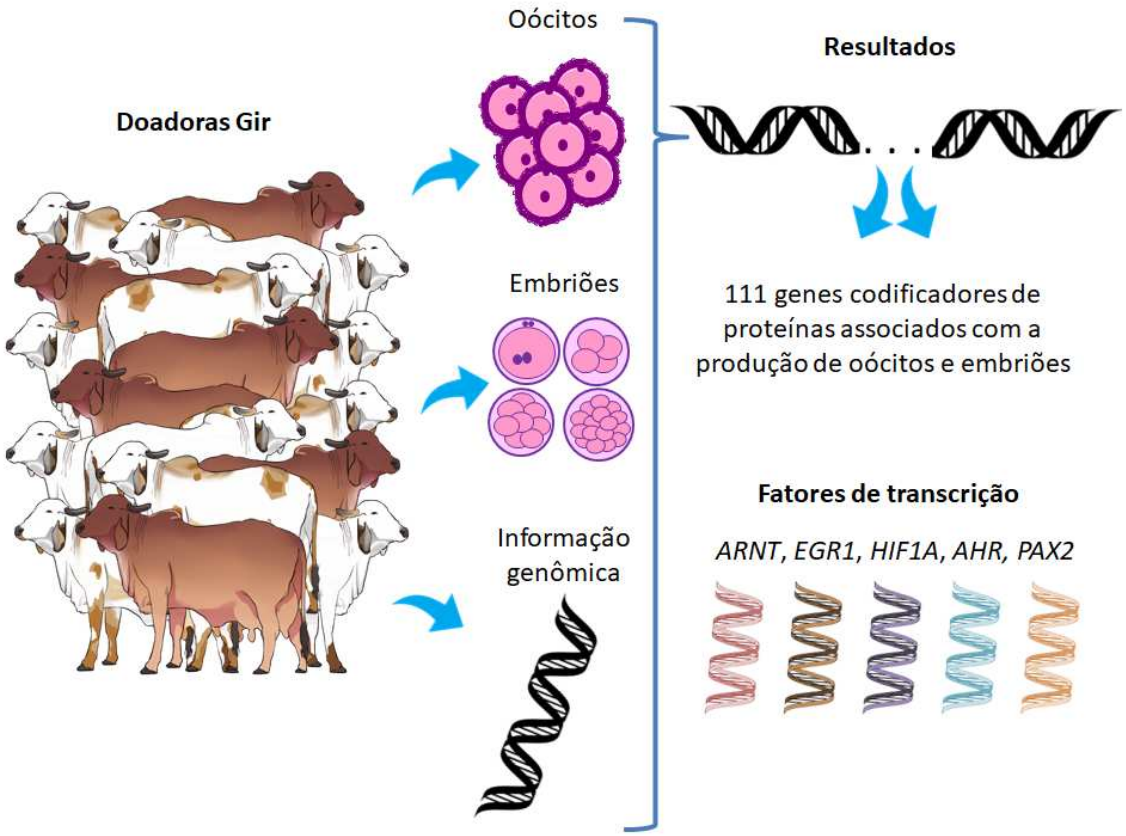


Figura 2: Resumo gráfico dos genes associados à produção de oócitos e embriões. Fonte: Do autor.

Herança genética

Como foi dito anteriormente, os estudos de GWAS auxiliam na identificação de genes que afetam características de interesse econômico. De forma ainda mais aprofundada, os estudos da herança dos genes podem ajudar a destacar genes que são comumente herdados pelas filhas dos touros. É possível explorar a herança genética de genes associados a características de interesse pecuário por meio de estudos chamados “daughter design”.

No presente trabalho, foi aplicado o estudo de “daughter desing” para 15 famílias de touros da raça Gir. Os resultados encontrados indicaram a presença de uma região particularmente frequente no cromossomo sete (BTA7) dos bovinos, que é mais comumente herdada entre as famílias e está relacionada com a produção de oócitos e embriões. Nesta região foram destacados os genes VCAN, HAPLN1 e EDIL3, que apresentam importantes funções relacionadas às características estudas. O resumo gráfico com os resultados de herança genética está representado na Figura 3.

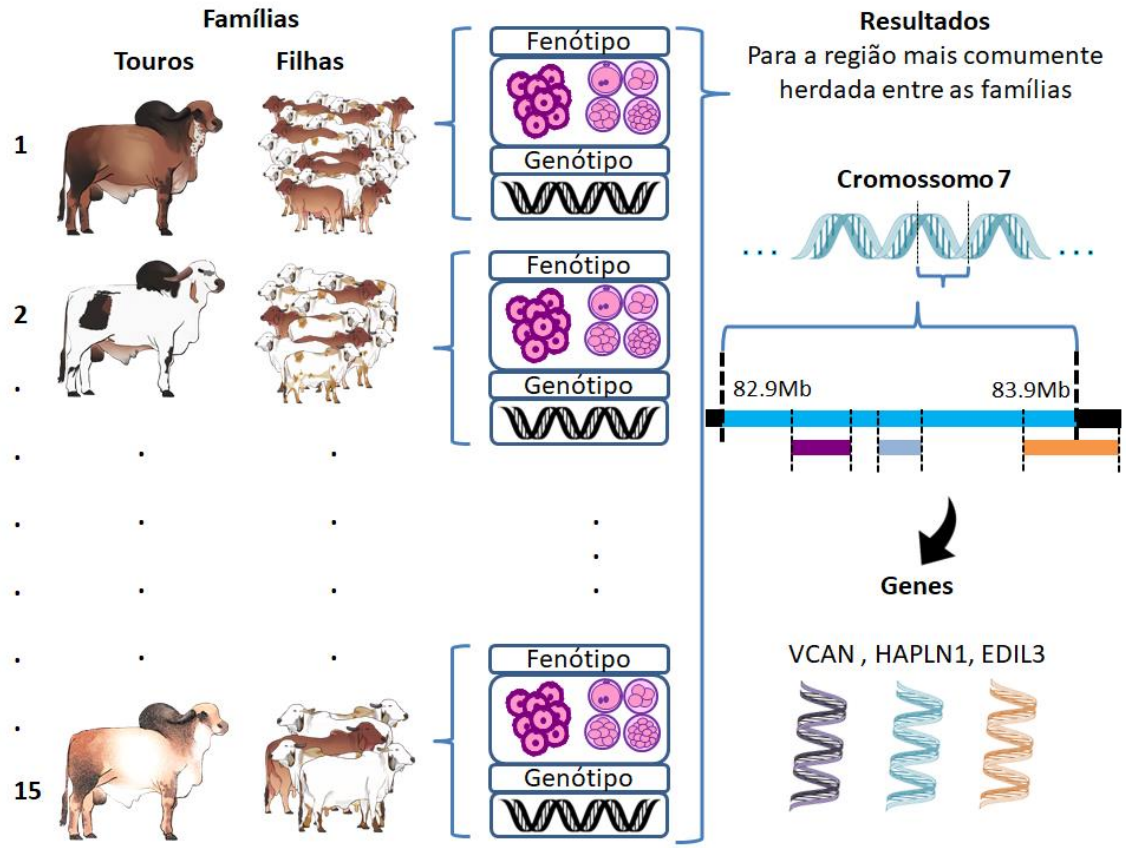


Figura 3: Resumo gráfico da herança genética com relação à produção de oócitos e embriões. Fonte: Do autor.

Efeito de imprinting

Sabe-se que o genoma de um animal é formato por metade do gameta paterno e metade do gameta materno. No entanto, nem sempre os genes herdados de ambos os pais são expressos da mesma forma e na mesma proporção, pois existem fenômenos conhecidos como efeitos de origem parental, sendo o mais conhecido deles o imprinting genômico. O imprinting genômico nada mais é do que o silenciamento de um gene herdado de um dos pais, enquanto que o gene herdado do outro é expresso (Hitchcock and Gardner 2019).

A exclusão destes efeitos em programas de melhoramento pode resultar em viés na predição dos valores genéticos (Tier and Meyer 2012). No presente trabalho, foi identificado efeito de imprinting paterno sobre o número de oócitos totais produzidos pela doadora.

Referências

Becher B, Neto A, Oliveira W, et al (2018) Fatores que afetam a produção in vitro de embriões (PIVE) em bovinos. Enciclopédia Biosf 15:554–570. https://doi.org/10.18677/EnciBio_2018B48

Brito LF, Kijas JW, Ventura R V., et al (2017) Genetic diversity and signatures of selection in

various goat breeds revealed by genome-wide SNP markers. BMC Genomics 18:1–20. https://doi.org/10.1186/s12864-017-3610-0

Hitchcock TJ, Gardner A (2019) Parent-of-origin specific gene expression and dispersal. Curr Opin Behav Sci 25:36–43. https://doi.org/10.1016/j.cobeha.2018.06.007

Purfield DC, Berry DP, McParland S, Bradley DG (2012) Runs of homozygosity and population history in cattle. BMC Genet 13:70. https://doi.org/10.1186/1471-2156-13-70

Rocha R de FB, Garcia AO, Otto PI, et al (2023a) Runs of homozygosity and signatures of selection for number of oocytes and embryos in the Gir Indicine cattle. Mamm Genome. https://doi.org/10.1007/s00335-023-09989-w

Rocha R de FB, Garcia AO, Otto PI, et al (2023b) Single-step genome-wide association studies and post-GWAS analyses for the number of oocytes and embryos in Gir cattle. Mamm Genome 34:497–508. https://doi.org/10.1007/s00335-023-10009-0

Rocha R de FB, Otto PI, da Silva MVGB, et al (2022) Repeatability and random regression models to estimate genetic parameters for oocyte and embryo production in the Gir breed. Anim Prod Sci 62:1661–1670. https://doi.org/10.1071/AN21588

Rotar LN, Souza JF (2019) Morphological Characteristics for Oocyte-Cumulus Complexes of Bos taurus and Bos indicus Cattle Used for a Variety of Purposes. Russ Agric Sci 45:399–402. https://doi.org/10.3103/S1068367419040141

Tier B, Meyer K (2012) Analysing quantitative parent-of-origin effects with examples from ultrasonic measures of body composition In Australian beef cattle. J Anim Breed Genet 129:359–368. https://doi.org/10.1111/j.1439-0388.2012.00996.x

Viana J (2022) 2021 Statistics of embryo production and transfer in domestic farm animals. A new milestone has been reached: Transfers of IVP embryos were over one million worldwide. Embryo Technol Newsl 40:22–40

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).

5 GENERAL CONCLUSIONS

In the field of genetic improvement of dairy cattle, the Gir breed stands out for its milk production in the tropics. The efficiency of the breed is verified by the Dairy Gir National Breeding Program (in portuguese: Programa Nacional de Melhoramento do Gir Leiteiro, PNMGL). To contribute to the improvement of the Gir breed, this work identified important aspects of this breed.

We concluded that repeatability models can be used to estimate the genetic parameters of oocytes and embryos as they require less computational effort than random regression models. As the Dairy Gir National Breeding Program work with a large amount of data from different farms, this information can help target analyses and save time. Nevertheless, in the search for more detailed results for repeated measures, random regression models can be used, since these models can provide more results than the repeatability model.

As a contribution to the farmers, we recommend selecting the trait number of total oocytes in young donors (over 1-year-old) in order to obtain an improvement in the number of viable oocytes and embryos. Our recommendation for selecting the number of total oocytes is in comparison to the number of viable oocytes and number of embryos. However, several production, reproduction, and type traits are important in the dairy industry. Therefore other studies should be consulted to verify the need for selection for other reproductive traits, not only for the number of total oocytes. Likewise, taking into account the importance of milk production traits, it is recommended that these are selected first, and then within the selected animals with greater milk production, animals can be selected for greater oocyte production. Likewise, when we only consider the database with the traits evaluated in this study, we recommend the selection of young donors (from one year of age onwards). However, it is important to emphasize that as selection must also be made for other reproductive traits, producers must consider the donors' physiology to avoid invasive manipulation in these young animals.

Producers who can benefit from this study are those who use the technologies involved in *in vitro* embryo production and if they are interested, these traits can contribute to an increase in herd size, as well as an increase in genetic gain.

We identified an ancient genomic relationship in this Gir population, indicating that the PNMGL work contributed to managing the inbreeding of the Gir breed in the past years. Managing inbreeding is important to avoid inbreeding depression, something we found that

can occur with the studied traits, that is, increasing levels of inbreeding cause a decrease in the number of oocytes and embryos.

We found several genes affecting the number of oocytes and embryos in Gir cattle, indicating that this is a polygenic trait. To shorten the list, we indicated five transcriptional factors as good markers for the production of oocytes and embryos. These findings might contribute to future research on dairy cattle reproductive traits since they give target genes to explore their expression. It is important to highlight that the regions identified in signatures of selection analyses indicate differences in selection between two populations, while the genomic regions identified in the Single-Step Genome-Wide Studies are directly associated with the traits studied, and are recommended for exploration in gene expression studies. We also identified common inherited genomic regions among families of Gir bulls through the "daughter design" approach. The most common region inherited across families allowed the identification of some genes that, like the genes mentioned in single-step GWAS, can be explored in gene expression studies, as they are associated with the number of oocytes and embryos. The difference between the genes highlighted in the conclusions of GWAS and the Family Analyses approach is that in the GWAS approach, they are transcription factors, genes responsible for activating or deactivating several genes that can influence the number of oocytes and embryos produced. In the Family Analyses approach, genes in linkage disequilibrium with a marker associated with the number of oocytes and embryos are highlighted.

The existence of an effect of parental origin through paternal imprinting on the total number of oocytes obtained was identified. Now that the existence of this effect has been identified, how this imprinting occurs and which genes are affected by imprinting can be explored in future studies. With the greater exploration of works in this area, perhaps the PNMGL can, in the future, consider the inclusion of imprinting effects in the genetic evaluation of the animals in order to verify if this affects the genetic values.

The technical article brings the results in a clearer and more summarized way to be presented to producers and people with less technical knowledge of statistical and genomic analyses, but who are also interested in the results found.

6 APPENDICES

Foram elaborados quatro tutoriais (passo a passo) de como realizar algumas das análises feitas nesta tese. Estes tutoriais não são definitivos, mas servem de base para pesquisadores iniciantes e experientes que desejam explorar novos temas. Os tutoriais foram publicados em português no site do Grupo de Discussão em Genética e Melhoramento Animal do departamento de Zootecnia da Universidade Federal de Viçosa: https://www.gdma.ufv.br/materiais-para-download/

Apêndice 6.1: TUTORIAL Análise de corridas de homozigose e assinaturas de seleção no programa PLINK

ISBN – 978-65-5668-138-2
DOI – http://dx.doi.org/10.26626/9786556681382.2023B0001

Apêndice 6.2: TUTORIAL Busca por genes e Quantitative Trait Loci (QTL) sobrepostos em regiões do genoma

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

Apêndice 6.3: TUTORIAL VarElect: The Next Generation Sequencing Phenotyper

ISBN – 978-65-5668-140-5
DOI – http://dx.doi.org/10.26626/9786556681405.2023B0001

Apêndice 6.4: TUTORIAL Análises de ontologia gênica (Gene Ontology - GO)

ISBN – 978-65-5668-141-2
DOI – http://dx.doi.org/10.26626/9786556681412.2023B0001

UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS AGRÁRIAS
DEPARTAMENTO DE ZOOTECNIA

TUTORIAL

Análise de corridas de homozigose e assinaturas de seleção no programa PLINK

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

Análise de corridas de homozigose e assinaturas de seleção no programa PLINK

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: 978-65-5668-138-2
DOI: http://dx.doi.org/10.26626/9786556681382.2023B0001

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

Este tutorial é para ajudar a fazer uma análise de corridas de homozigose (ROH) e assinaturas de seleção pela metodologia FST usando o software PLINK, por exemplo, para quem e não sabe por onde começar.

Este tutorial não é definitivo! Estude sobre o assunto para saber qual o tipo de análise se adequa melhor à sua pesquisa e como ela pode ser realizada!

SUMÁRIO

1. Arquivos necessários para a análise de ROH

- Arquivo de mapa
- Arquivo de pedigree

Os arquivos de mapa e fenótipos/genótipos são diferentes dos usados em GWAS.

Este tutorial mostra a formatação dos arquivos necessários para a análise de ROH a partir de arquivos pré-existentes para uma análise nos programas do Misztal. As formatações são feitas no software R e no LINUX.

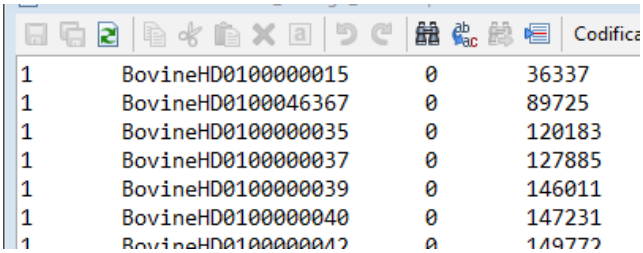
Em seguida tem o passo a passo para realizar a análise de ROH, o cálculo da endogamia baseada em ROH (FROH) e, também, para realizar a análise de FST.

1.1 O arquivo de mapa

As colunas do arquivo de mapa tem que ser todas separadas por espaço ou todas separadas por TAB. Por padrão, cada linha do arquivo de mapa descreve um único marcador e deve conter exatamente quatro colunas:

- Cromossomo
- rs# ou identificador/nome do SNP
- Distância genética em morgans
- Posição em pares de bases (unidades de pb)

Neste caso a coluna de distância genética em morgans está zerada, pois não preciso dessa informação para esta análise.



1	BovineHD0100000015	0	36337
1	BovineHD0100046367	0	89725
1	BovineHD0100000035	0	120183
1	BovineHD0100000037	0	127885
1	BovineHD0100000039	0	146011
1	BovineHD0100000040	0	147231

Os arquivos de mapa e de pedigree devem ter o mesmo nome, mas o de mapa deve ter a extensão .map e o de pedigree .ped. Exemplo:

- arquivo123.map
- arquivo123.ped

1.2 O arquivo de pedigree

As colunas do arquivo de pedigree tem que ser todas separadas por espaço ou todas separadas por TAB. As primeiras seis colunas são obrigatórias:

- Family ID (identificação da família – se não tiver, deixar 0)
- Individual ID (identificação do indivíduo)
- Paternal ID (identificação do pai – se não tiver, deixar 0)
- Maternal ID (identificação da mãe – se não tiver, deixar 0)
- Sexo (1=macho; 2=fêmea; outro número=desconhecido)
- Fenótipo (para essa análise não precisa, então deixar é 0)
- Em seguida vem as colunas com as letras dos genótipos (vai ter uma coluna para cada SNP, então é um arquivo muito grande)

Exemplo:

0	AB52502	Tgir-824	Gir-2697	2	0	A	B	A	B	A	B	A	A	B
0	AB52503	Tgir-879	RRP6537	2	0	A	B	A	A	A	B	A	A	B
0	AB52505	Tgir-1037	Gir-4774	2	0	A	B	A	B	A	B	A	A	B
0	AB52508	Tgir-1399	Gir-2705	2	0	A	B	A	A	B	B	A	B	B
0	AB52511	Tgir-1009	Gir-716	2	0	A	B	B	B	A	B	A	A	A
0	AB52512	Tgir-1009	Ne21RRP6419	2	0	A	B	A	B	A	B	A	A	B
0	AB52514	Tgir-1295	Gir-716	2	0	A	A	A	B	A	A	B	A	B
0	AB52516	Tgir-1437	RRP6395	2	0	A	B	A	B	A	B	A	A	A
0	AB52517	Tgir-879	Gir-2805	2	0	A	B	A	B	A	B	A	A	B

1.3 O programa PLINK

Com a versão 1.07 do PLINK, pode ser feita a análise de ROH. Com a versão 1.09 do PLINK, além da análise de ROH também pode ser feita a análise de FST

- Mais informações em: https://zzz.bwh.harvard.edu/plink/data.shtml
- Ou só pesquisar: “PLINK: Whole genome data analysis toolset”

Referências

- Para a versão 1.07 do PLINK:

Purcell S, Neale B, Todd-Brown K, et al (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 81:559–575. https://doi.org/10.1086/519795

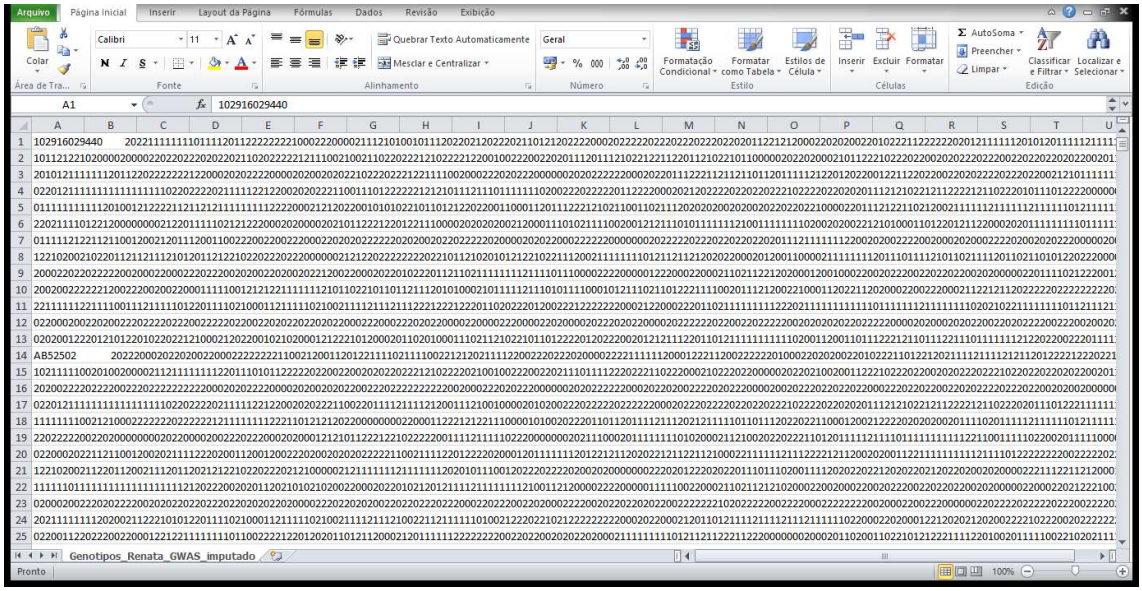
- Para a versão 1.09 do PLINK:

Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4:s13742–015–0047–8. https://doi.org/10.1186/s13742-015-0047-8

2. Como montar os arquivos .ped e .map a partir de um arquivo de genótipos usados nos programas do Misztal

Como estes arquivos são grandes, pode ser necessário usar um servidor LINUX. No servidor pode-se instalar o software R e usá-lo para montar os arquivos ou usar o próprio sistema LINUX para montar os arquivos.

Arquivo de genótipos no formato para programas do Misztal: genotipos.csv



O arquivo de genótipos acima contém duas colunas: na primeira tem a identificação dos animais e na segunda tem os genótipos (formato para serem usados em programas do Misztal para análise de GWAS e seleção genômica, por exemplo).
O arquivo deste exemplo contém 420.718 marcadores SNPs nas colunas e 2.093 animais nas linhas. Como o arquivo é grande, não é viável manipular os dados pela planilha no Excel. Arquivos maiores que este pode ser que nem abram neste formato.

2.1 Transformando o arquivo de dados .csv para .txt

a. *Se o arquivo de genótipos for pequeno*

Alguns arquivos de genótipos para exercícios em sala de aula, por exemplo, podem conter pouca quantidade de animais e marcadores, apenas para fins didáticos. É possível copiar o conteúdo para o Notepad++ e salvar o arquivo.

b. *No software R*

geno<-read.csv(genotipos.csv) # importando os dados

library(gdata) # abrindo a biblioteca para rodar o comando abaixo

write.fwf(geno, file = ‘./genotipos.txt’, rownames=F, colnames=F, quote=F, sep = “ “, justify=’left’) # salvando o arquivo no formato para os programas do Misztal

c. *No LINUX*

O comando abaixo pega o arquivos em .csv e salva em .txt

tr ',' '\n' < genotipos.csv > genotipos.txt

2.2 Formatando arquivo de pedigree/genótipo

O arquivo de genótipos do PLINK é uma junção do pedigree e dos genótipos nos animais, além de mais algumas informações, como visto no tópico 1.2. Então a partir do arquivo de pedigree e de genótipos no formato dos programas do Misztal, podemos formar o arquivo para ser usado no PLINK.

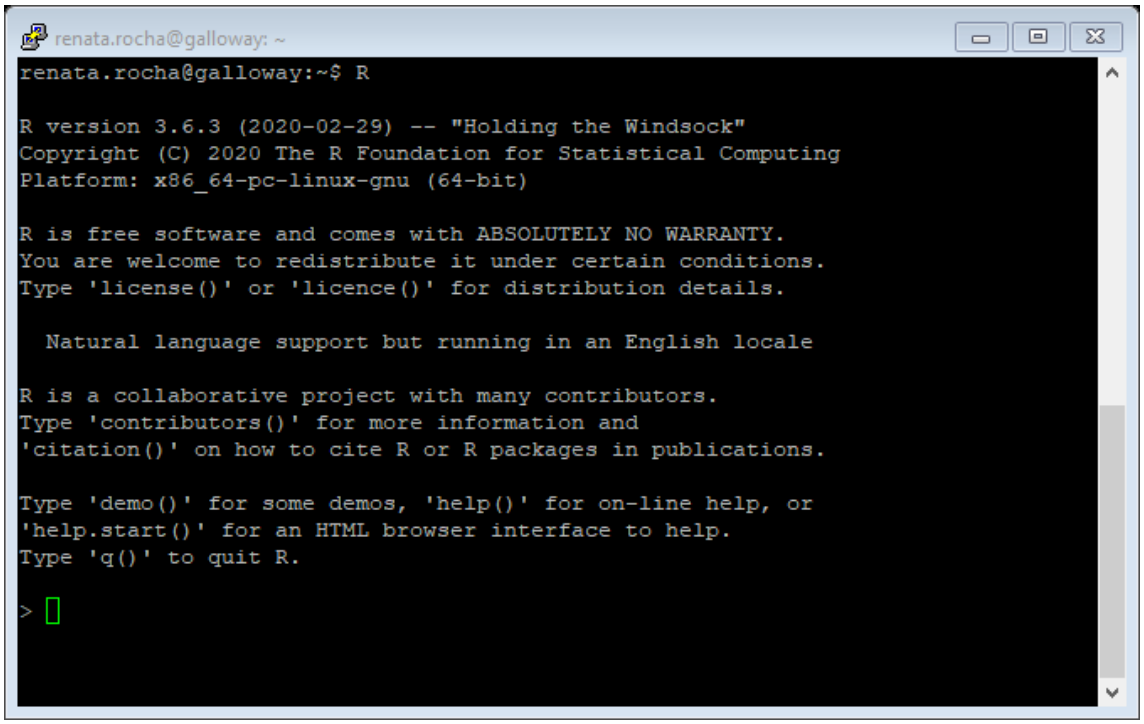
A seguir são apresentadas duas formas (a. No software R) e (b. Parte no software R e outra parte no LINUX) para obter o arquivo de pedigree/genótipo para PLINK.

a. *No software R*

Neste caso, o servidor já tem o software R instalado. A forma de abrir o programa no servidor do exemplo é apenas digitando:

R

Clicar na tecla ‘ENTER’



Como mostrado na imagem acima o servidor vai direcionar para dentro do software R.

Em seguida, usar os comandos do R para montar o arquivo:

rm(list=ls()) #Limpando o *environment* (conteúdo do R)
options(stringsAsFactors=F) # Este é um argumento para a função ‘data.frame()' no R. É uma lógica que indica se as *strings* em um quadro de dados devem ser tratadas como variáveis de fator ou apenas como *strings* simples.
setwd("/home/pasta_com_arquivos") # Direcionando para a pasta
ped<-read.table("ped.txt",h=F,sep="\t") # Importando arquivo de pedigree
dim(ped) # Verificando a dimensão do arquivo: número de linhas e de colunas
colnames(ped)<-c("ID","PAI","MAE","SEX") # Atribuindo nomes nas colunas

A função read.table() demoraria muito pra importar o arquivo de genótipos, por isso, usar o seguinte:
library('data.table')
genotipos <-fread("genotipos.txt",h=F) # Importando arquivo de genótipos
colnames(genotipos)<-c("ID","GEN") # Atribuindo nomes nas colunas do arquivo
dim(genotipos) # Verificando a dimensão do arquivo: número de linhas e de colunas

```
ped1<-merge(ped, genotipos[,1], by = intersect("ID", "ID")) # Pegando o pedigree só dos animais que tem genótipo
head(ped1) #Visualizando as primeiras 6 linhas do arquivo

### Recodificando os genótipos – nesta parte estamos transformando os números em letras e separando todos por espaço simples
# Se estiver rodando a análise no R do Windows, pode deixar as letras dentro do comando gsub separadas por TAB.
# Assim: gsub("0", "A	A", x)
# Se estiver rodando a análise no R no servidor, ao copiar e colar a linha de comando lá o espaçamento do TAB não aparecer, então tem que deixar separado por espaço simples.

geno_total<-as.data.frame(genotipos[,2])
total1<-data.frame(lapply(geno_total, function(x) {gsub("0", "A A", x)}))
total2<-data.frame(lapply(total1, function(x) {gsub("1", "A B", x)}))
total3<-data.frame(lapply(total2, function(x) {gsub("2", "B B", x)}))
total4<-data.frame(lapply(total3, function(x) {gsub("AA", "A A", x)}))
total5<-data.frame(lapply(total4, function(x) {gsub("BB", "B B", x)}))
total6<-data.frame(lapply(total5, function(x) {gsub("AB", "A B", x)}))
total7<-data.frame(lapply(total6, function(x) {gsub("BA", "B A", x)}))
dim(total7) # As linhas representam o número de animais, mas vai ter apenas uma coluna.

### Formando o arquivo para o PLINK
seq0<-rep(0,nrow(geno_total)) # Essa sequência de zeros ficará no lugar das colunas de família e de ano de nascimento
arquivo123_esp<-cbind(seq0, ped1, seq0, total7) # Juntando as colunas
# No arquivo formado terão as seguintes colunas nesta sequência:
# Coluna zerada que seria a identificação da família
# Identificação do indivíduo
# Identificação do pai
# Identificação da mãe
# Sexo
```

Coluna zerada (que seria o fenótipo)
As colunas seguintes são as letras dos genótipos

PARA SALVAR

Se estiver rodando a análise no R do Windows:
write.table(arquivo123_esp," arquivo123_esp.ped", quote = F, row.names = F,
col.names = F, sep = "\t")
# Este arquivo será salvo com as colunas separadas por TAB
# Assim temos o arquivo de pedigree+genótipos para o LINUX!!! (com apenas as colunas de genótipos separadas por espaço)

Se estiver rodando a análise no R no servidor:
write.table(arquivo123_esp," arquivo123_esp.ped",quote=F,row.names=F,col.names=F)
# Este arquivo será salvo com todas as colunas separadas por espaço simples
# Agora para pegar esse arquivo e separar as colunas por TAB:
# 1ª Opção -> Recomendada (mais rápida): sair do R no servidor
q() #comando para sair do R
Save workspace image? [y/n/c]:
n

No LINUX, podemos separar as colunas do arquivo por TAB com o seguinte comando:
sed -e 's/ /\t/g' arquivo123_esp.ped > arquivo123.ped
Assim temos o arquivo de pedigree+genótipos para o LINUX!!!

# 2ª Opção -> Pode ser usada para arquivos pequenos, por exemplo, porque quanto maior o arquivo mais pode demorar.
# No próprio R, importar o arquivo novamente (Ao importar novamente, os genótipos já vêm divididos em colunas)
# Não é necessário fazer nenhuma edição no arquivo
setwd("/home/pasta_com_arquivos")

```
library('data.table')
arquivo123<-fread("arquivo123_esp.ped",h=F) # Importando o arquivo
# O comando acima demora um pouco (~ 5 a 10min) neste exemplo.
# No comando para salvar, usamos sep="\t" que separa todas as colunas por TAB
write.table(arquivo123," arquivo123.ped",quote=F,row.names=F,col.names=F,sep="\t")
# Assim temos o arquivo de pedigree/genótipos para o LINUX!!!
```

b. Parte no software R e outra parte no LINUX

```
### Essa primeira parte feita no software R é usada apenas para separar o pedigree dos animais que tem genótipo.
## Se já tiver o pedigree desses animais separados, pode pular a parte do script do R.

## Precisamos ter os arquivos:
# ped.txt (com animal, pai, mae e sexo) -> Se não tiver info de sexo e não precisar para a análise basta criar uma coluna de zeros
# ids.txt -> Arquivo que tem uma coluna com as identificações dos animais do arquivo de genótipos
```

Novamente, no servidor, digitar:

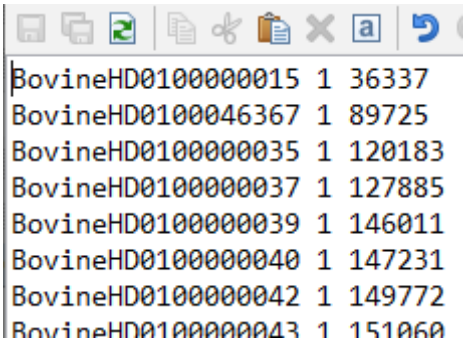
```
R
rm(list=ls()) #limpa todo o conteúdo/arquivos que existem no R
options(stringsAsFactors=F)

setwd("/home/pasta_com_arquivos") # Direcionando para a pasta onde estão os arquivos
ped<-read.table("ped.txt",h=F,sep="\t") # Importando arquivo de pedigree
colnames(ped)<-c("ID","PAI","MAE","SEX") # atribuindo os nomes de cada coluna do arquivo de pedigree
ids<-read.table("ids.txt",h=F) # Importando as identificações dos animais do arquivo de genotipos
colnames(ids)<-c("ID") # atribuindo o nome da coluna do arquivo de identificação
ped_roh<-merge(ids, ped, by=intersect("ID","ID")) # pegando os animais em comum do arquivo de identificações (genótipos) e do pedigree
```

```
dim(ped_roh) # mostra o número de linhas (número de animais) e 4 colunas (id, pai, mae e sexo)
write.table(ped_roh, "ped_roh.ped", quote=F, row.names=F, col.names=F, sep="\t")
# Criando arquivo com sequência de zeros para usar no LINUX
seq0<-as.data.frame(rep(0,nrow(ped_roh)))
write.table(seq0,"seq0.txt",quote=F, row.names=F, col.names=F) # salvando o arquivo
q() #sair do R

## No LINUX ##
# Cuidado ao copiar comandos com TAB e colar no servidor - ele não cola o espaçamento do TAB
# Fazer cada comando por vez, porque cada comando é um pouco lento.
awk '{ print $2 }' genotipos.txt > snps.ped #pegando a coluna só com os genótipos
sed -i 's/0/A A/g' snps.ped
sed -i 's/1/A B/g' snps.ped
sed -i 's/2/B B/g' snps.ped
sed -i 's/AA/A A/g' snps.ped
sed -i 's/BB/B B/g' snps.ped
sed -i 's/AB/A B/g' snps.ped
sed -i 's/BA/B A/g' snps.ped
## Formando o arquivo de genótipos para PLINK ##
#Juntando seq0, ped_roh, zeros, snps
# Esse arquivo seq0 foi criado no script do R (acima)
# Tem que ter certeza que os animais no arquivo de ped estão na mesma ordem do arquivo de genótipos!
paste seq0.txt ped_roh.ped seq0.txt snps.ped > arquivo123_esp.ped

# Por padrão, a função paste() separa as linhas de cada coluna do arquivo com TAB:
# https://www.vivaolinux.com.br/dica/O-comando-paste
## Atenção: uma parte dos genótipos continua sendo uma coluna com as letras separadas por espaço.
# Então, para substituir os espaços por TAB:
sed -e 's/ /\t/g' arquivo123_esp.ped > arquivo123.ped
```

2.3 Formatando arquivo de mapa no software R

Como visto no tópico 1.1, o arquivo de mapa deve ter um formato específico para análises no PLINK.

Supondo que temos um arquivo de mapa prévio com o seguinte formato: três colunas → a primeira tem o nome do SNP, a segunda tem o cromossomo e a terceira tem a posição:



```
BovineHD0100000015 1 36337
BovineHD0100046367 1 89725
BovineHD0100000035 1 120183
BovineHD0100000037 1 127885
BovineHD0100000039 1 146011
BovineHD0100000040 1 147231
BovineHD0100000042 1 149772
BovineHD0100000043 1 151060
```

Este arquivo pode ser transformado em um arquivo de mapa específico para análise no PLINK com o seguinte script para o software R (pode ser no Windows ou no LINUX, porque esse arquivo é menor).

- No software R:

```
rm(list=ls())
options(stringsAsFactors=F) # tem que ter esse comando, se não os nomes dos SNPs viram números na hora de formar o mapa novo
setwd("/home/pasta_com_arquivos") # Direcionando para a pasta
mapa<-read.table("snpmap.txt",h=F) # Importando arquivo de mapa
dim(mapa) # Verificando a dimensão do arquivo de mapa - número de linhas (SNPs) e colunas (três neste caso: nome do SNP, cromossomo e posição)
head(mapa) # Visualizando as seis primeiras linhas do arquivo de mapa
coluna0<-rep(0,nrow(mapa)) # Criando uma coluna só com valores zero, com o mesmo número de linhas do arquivo de mapa.
length(coluna0) # Comprimento da coluna de zeros
snpnames<-mapa[,1] # Criando um vetor que contém apenas o nome dos SNPs
# Em seguida, criamos um arquivo com as colunas do arquivo de mapa para o PLINK:
1ª col = cromossomo, 2º col = nomes dos SNPs, 3º col = coluna de zeros (esse seria preenchida pela posição dos marcadores de Morgans, mas essa informação não é
```

necessária nesta análise, portanto zeramos a coluna toda) e 4ª = posição dos marcadores em pares de bases.

```
mapa_novo<-cbind(mapa[,2],snpnames, coluna0,mapa[,3])
head(mapa_novo) # Visualizando as primeiras linhas do arquivo de mapa criado para o PLINK
#Em seguida, salvar o arquivo criado:
write.table(mapa_novo,"arquivo123.map", quote=F, row.names=F, col.names=F, sep="\t") # Salvando o arquivo de mapa para o PLINK
```

3. Análise de ROH no PLINK

Para a análise de ROH, usamos o programa PLINK para formar arquivos binários. Podemos rodar a análise sem transformar para arquivo binário, neste caso, é só tirar o "b" de "bfile" no comando, mas tendo os arquivos binários a análise é mais rápida.

Em uma pasta no servidor, deixar os seguintes arquivos:

- Arquivo de mapa (Ex.: arquivo123.map)
- Arquivo de pedigree (Ex.: arquivo123.ped)
- Programa PLINK

Os arquivos .map e .ped tem que ter o mesmo nome, pois na linha de comando da análise só escrevemos "arquivo123" sem identificar o ".map" ou ".ped". O programa lê os dois de uma vez.

3.1 Rodando a análise

- No servidor, direcionar para a pasta com os arquivos:

  cd meusdados/pasta_arquivos

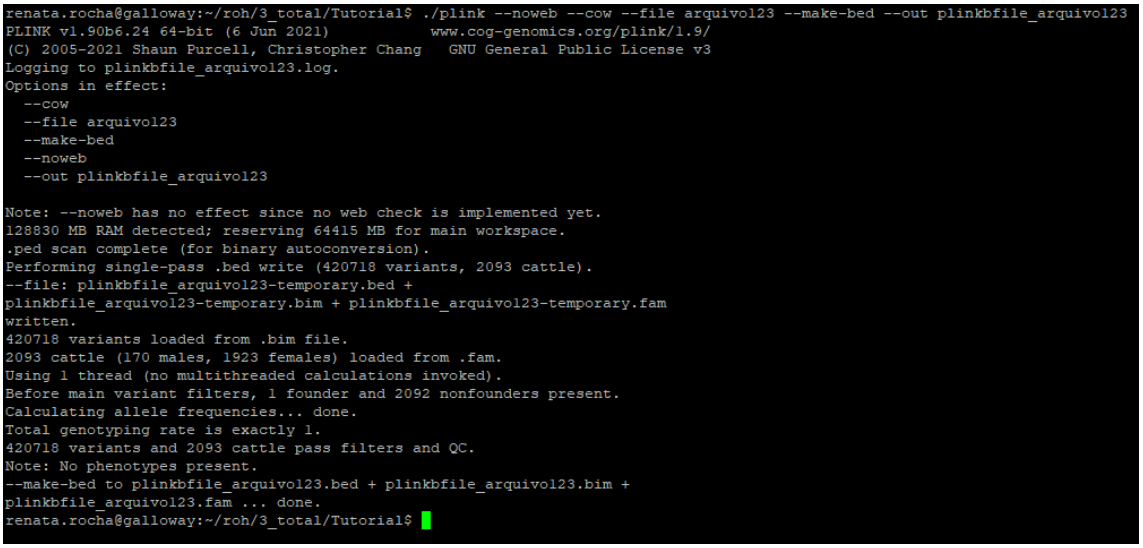
  Dá ENTER
- Usar o comando seguinte para abrir espaço no servidor:

  ulimit -s unlimited

  Dá ENTER
- Usar o comando seguinte no servidor para criar o arquivo binário:

  ./plink --noweb --cow --file arquivo123 --make-bed --out plinkbfile_arquivo123

  Dá ENTER



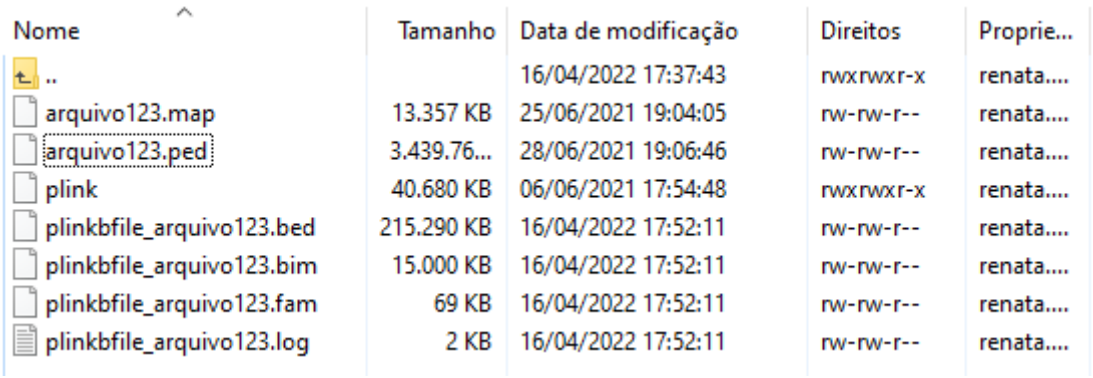
- Depois que a análise acima termina, vai soltar as seguintes saídas na pasta:

➢ plinkbfile_arquivo123.beb

➢ plinkbfile_arquivo123.bim

➢ plinkbfile_arquivo123.fam

➢ plinkbfile_arquivo123.log



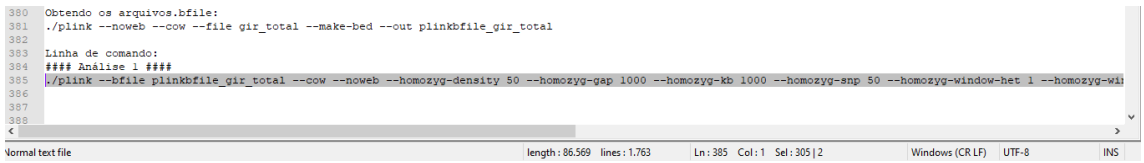
Nome	Tamanho	Data de modificação	Direitos	Proprie...
..		16/04/2022 17:37:43	rwxrwxr-x	renata....
arquivo123.map	13.357 KB	25/06/2021 19:04:05	rw-rw-r--	renata....
arquivo123.ped	3.439.76...	28/06/2021 19:06:46	rw-rw-r--	renata....
plink	40.680 KB	06/06/2021 17:54:48	rwxrwxr-x	renata....
plinkbfile_arquivo123.bed	215.290 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
plinkbfile_arquivo123.bim	15.000 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
plinkbfile_arquivo123.fam	69 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
plinkbfile_arquivo123.log	2 KB	16/04/2022 17:52:11	rw-rw-r--	renata....

- Em seguida, usar a linha de comando abaixo para rodar a análise de ROH:

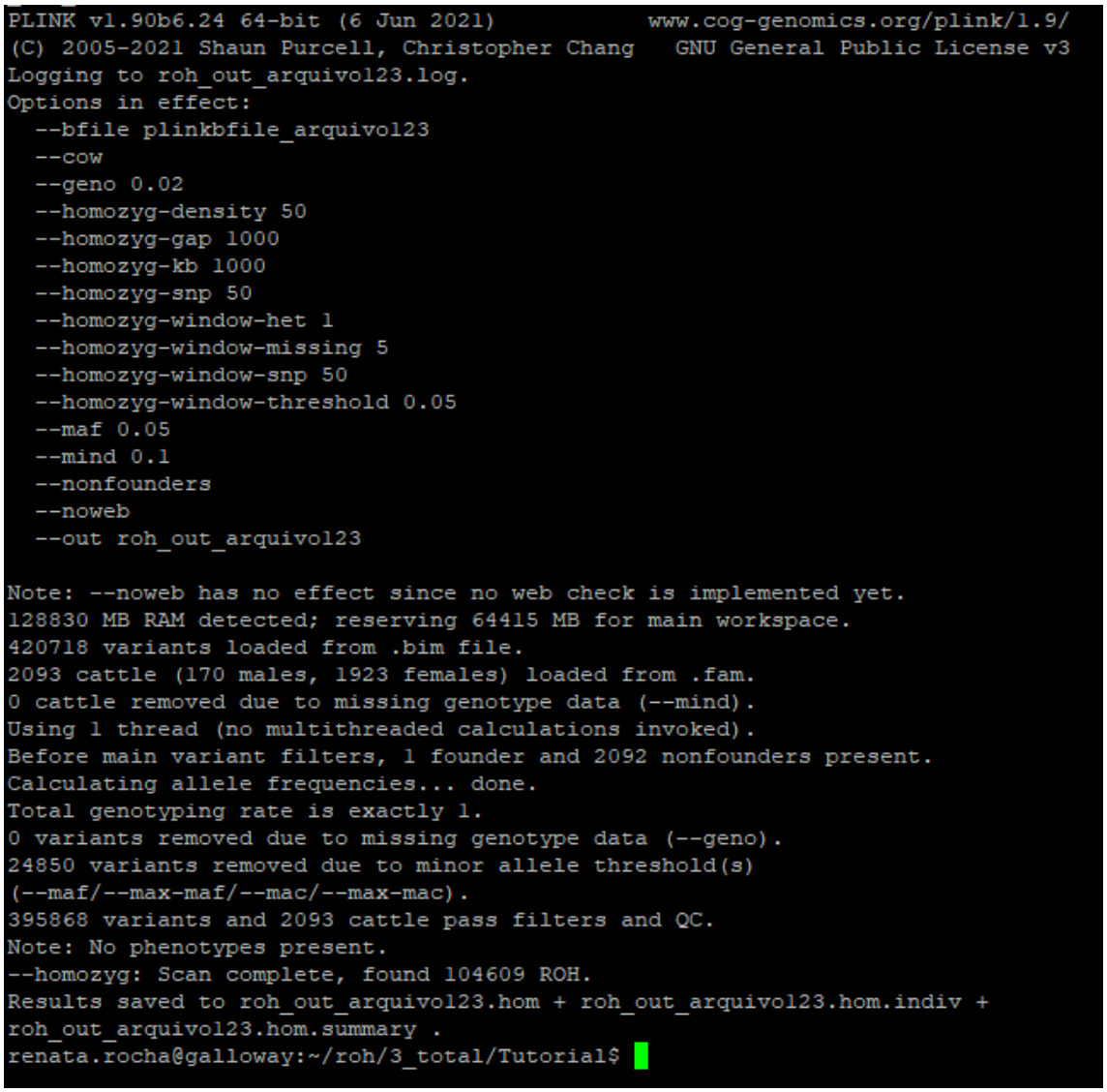
  ./plink --bfile plinkbfile_arquivo123 --cow --noweb --homozyg-density 50 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-window-missing 5 --homozyg-window-snp 50 --homozyg-window-threshold 0.05 --nonfounders --geno 0.02 --maf 0.05 --mind 0.1 --out roh_out_arquivo123

  Dá ENTER

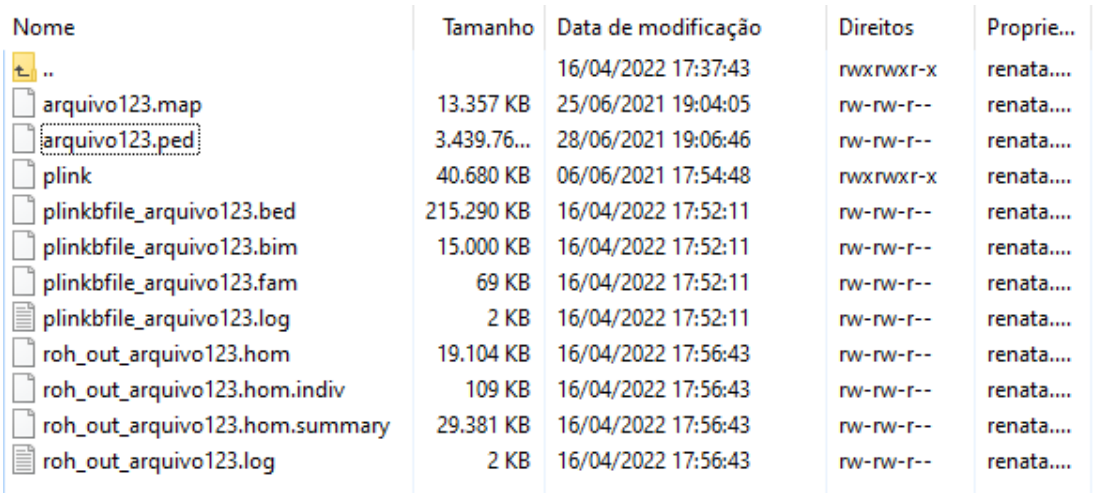
  - **Detalhes importantes:**
  - **Linha de comando**: aqui neste arquivo a linha de comando fica "quebrada" por causa do limite de espaço, mas a linha de comando não pode estar "quebrada" quando é colada no Linux, pois dá erro e a análise não é finalizada como deveria.
  - **Sugestão**: Colar a linha de comando no Notepad++ e daí copiar ela inteira para colar no servidor.



```
Obtendo os arquivos.bfile:
./plink --noweb --cow --file gir_total --make-bed --out plinkbfile_gir_total

Linha de comando:
#### Análise 1 ####
./plink --bfile plinkbfile_gir_total --cow --noweb --homozyg-density 50 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-wi
```



- Depois que a análise termina, solta as seguintes saídas:
- roh_out_arquivo123.hom
- roh_out_arquivo123.hom.indiv
- roh_out_arquivo123.hom.summary
- roh_out_arquivo123.log



Nome	Tamanho	Data de modificação	Direitos	Proprie...
..		16/04/2022 17:37:43	rwxrwxr-x	renata....
arquivo123.map	13.357 KB	25/06/2021 19:04:05	rw-rw-r--	renata....
arquivo123.ped	3.439.76...	28/06/2021 19:06:46	rw-rw-r--	renata....
plink	40.680 KB	06/06/2021 17:54:48	rwxrwxr-x	renata....
plinkbfile_arquivo123.bed	215.290 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
plinkbfile_arquivo123.bim	15.000 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
plinkbfile_arquivo123.fam	69 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
plinkbfile_arquivo123.log	2 KB	16/04/2022 17:52:11	rw-rw-r--	renata....
roh_out_arquivo123.hom	19.104 KB	16/04/2022 17:56:43	rw-rw-r--	renata....
roh_out_arquivo123.hom.indiv	109 KB	16/04/2022 17:56:43	rw-rw-r--	renata....
roh_out_arquivo123.hom.summary	29.381 KB	16/04/2022 17:56:43	rw-rw-r--	renata....
roh_out_arquivo123.log	2 KB	16/04/2022 17:56:43	rw-rw-r--	renata....

FLAGS relacionadas à ROH:

- --homozyg-density ## No default: uma ROH deve ter pelo menos um SNP por 50 kb em média.
- --homozyg-gap ## No default: se dois SNPs consecutivos estiverem separados por mais de 1000 kb, eles não podem estar no mesmo ROH.
- --homozyg-snp ## No default: apenas ROH contendo pelo menos 100 SNPs e de comprimento total ≥ 1000 kb são anotadas.
- --homozyg-kb ## Podemos alterar esses mínimos com --homozyg-snp e --homozyg-kb, respectivamente.
- --homozyg-window-het ## No default: uma ocorrência de janela de varredura pode conter no máximo 1 chamada heterozigótica.
- --homozyg-window-missing ## No default: uma ocorrência de janela de varredura pode conter no máximo 5 chamadas perdidas.
- --homozyg-window-snp ## No default: a janela de varredura contém 50 SNPs.
- --homozyg-window-threshold ## No default: para que um SNP seja elegível para inclusão em um ROH, a taxa de acerto de todas as janelas de varredura contendo o SNP deve ser de pelo menos 0,05.
- --nonfounders ## Somente fundadores são normalmente considerados por esses filtros (use a flag para mudar isso).

  FLAGS relacionadas ao controle de qualidade (CQ) do arquivo de genótipos.
- --geno ## No default: filtra todos os marcadores com taxas de chamadas ausentes excedendo o valor fornecido (padrão 0,1) para serem removidas.

- --maf ## No default: filtra todas marcadores com frequência de alelo menor abaixo do limite fornecido (padrão 0,01).
- --mind ## Idem --geno só que para amostras.
- --hwe ## Exclui amostras (individuos) e/ou marcadores com base no critério definido para o equilíbrio de Hardy-Weinberg
- --hardy## Faz uma estatística resumida das taxas de HWE

3.2 Resultados da análise

Mais detalhes em: https://www.cog-genomics.org/plink/1.9/formats#hom

- roh_out_arquivo123.log

  Este arquivo tem um resumo dos parâmetros usados na análise:

```
PLINK v1.90b6.24 64-bit (6 Jun 2021)
Options in effect:
  --bfile plinkbfile_arquivo123
  --cow
  --geno 0.02
  --homozyg-density 50
  --homozyg-gap 1000
  --homozyg-kb 1000
  --homozyg-snp 50
  --homozyg-window-het 1
  --homozyg-window-missing 5
  --homozyg-window-snp 50
  --homozyg-window-threshold 0.05
  --maf 0.05
  --mind 0.1
  --nonfounders
  --noweb
  --out roh_out_arquivo123

Hostname: galloway
Working directory: /home/renata.rocha/roh/3_total/Tutorial
Start time: Sat Apr 16 20:56:36 2022
```

  E também resumo dos resultados da análise:

- ...
- 420718 é o número de marcadores/SNPs no arquivo original
- 2093 é o número de indivíduos (170 machos e 1923 fêmeas)
- 0 indivíduos removidos devido a genótipo ausente.
- ...
- 0 marcadores removidos devido a genótipo ausente. O arquivo imputado original neste exemplo realmente não tinha genótipos imputados.
- 24850 marcadores removidos devido à MAF
- 395868 marcadores e 2093 indivíduos passaram no filtro do controle de qualidade
- --homozyg: varredura completa, encontrou 104609 ROHs

```
Note: --noweb has no effect since no web check is implemented yet.
Random number seed: 1650142596
128830 MB RAM detected; reserving 64415 MB for main workspace.
420718 variants loaded from .bim file.
2093 cattle (170 males, 1923 females) loaded from .fam.
0 cattle removed due to missing genotype data (--mind).
Using 1 thread (no multithreaded calculations invoked).
Before main variant filters, 1 founder and 2092 nonfounders present.
Calculating allele frequencies... done.
Total genotyping rate is exactly 1.
0 variants removed due to missing genotype data (--geno).
24850 variants removed due to minor allele threshold(s)
(--maf/--max-maf/--mac/--max-mac).
395868 variants and 2093 cattle pass filters and QC.
Note: No phenotypes present.
--homozyg: Scan complete, found 104609 ROH.
Results saved to roh_out_arquivo123.hom + roh_out_arquivo123.hom.indiv +
roh_out_arquivo123.hom.summary .

End time: Sat Apr 16 20:56:43 2022
```

- roh_out_arquivo123.hom

  - Produzido quando uma bandeira da família --homozyg está presente. Acompanhado por pelo menos um arquivo .hom.indiv e um arquivo .hom.summary

  - Cada linha do arquivo .hom representa uma ROH encontrada! Então o total de linhas (excluindo o cabeçalho) indica o total de ROHs encontradas. O número de ROHs encontradas pode alterar dependendo do número de animais e dos parâmetros usados na análise!

  - O arquivo .hom tem as seguintes colunas:

- FID # Identificação da família -> Estará zerada neste caso, porque não era uma informação necessária.
- IID # Identificação do indivíduo. Um mesmo indivíduo pode aparecer várias vezes, pois ele tem várias ROHs.
- PHE # Valor fenotípico -> Neste caso não trabalhamos com o valor fenotípico, por isso está zerada.
- CHR #Cromossomo em que está aquela ROH
- SNP1 # Identificação do primeiro SNPs na ROH
- SNP2 # Identificação do último SNPs na ROH
- POS1 # Posição do primeiro SNP em pares de bases

- POS2 # Posição do último SNP em pares de bases
- KB # Comprimento da ROH em kilobases
- NSNP # Número de SNPs nessa ROH
- DENSITY # Densidade de SNP inverso em Kb/SNP
- PHOM# Proporção de chamadas (marcadores?) homozigóticas
- PHET # Proporção de chamadas (marcadores?) heterozigóticas

  Dica: Se o arquivo de genótipos tiver animais com genótipos imputados, seria interessante comparar o PHET resultante de uma análise com todos os genótipos com o PHET resultante de uma análise que tem apenas animais originalmente genotipados em HD, porque o processo de imputação "adiciona" heterozigotos. Assim, pode-se verificar se existe alguma influência do processo de imputação na análise de ROH.

FID	IID	PHE	CHR	SNP1	SNP2	POS1	POS2	KB	NSNP	DENSITY	PHOM	PHET
0	102916029440	-9.000	1	BovineHD0100006914	BovineHD0100012022	23381349	42095937	18714.589	2949	6.346	1.000	0.000
0	102916029440	-9.000	1	BovineHD0100047134	BovineHD0100029888	102458103	105162097	2703.995	466	5.803	0.994	0.006
0	102916029440	-9.000	1	BovineHD0100041142	BovineHD0100043043	143141556	148784184	5642.629	1056	5.343	0.998	0.002
0	102916029440	-9.000	1	BovineHD0100043081	BovineHD0100044179	148857503	152002893	3145.391	410	7.672	1.000	0.000
0	102916029440	-9.000	1	BovineHD0100044217	BovineHD0100045574	152126033	156040374	3914.342	708	5.529	0.999	0.001
0	102916029440	-9.000	2	BovineHD0200019851	BovineHD0200020565	68732542	71627667	2895.126	199	14.548	1.000	0.000
0	102916029440	-9.000	2	BovineHD0200022016	BovineHD0200025272	76593223	89069133	12475.911	1465	8.516	0.999	0.001
0	102916029440	-9.000	2	BovineHD0200031011	BovineHD0200032436	107756558	112727237	4970.680	608	8.175	0.998	0.002
0	102916029440	-9.000	2	BovineHD0200033469	BovineHD0200033810	116160522	117331836	1171.315	155	7.557	1.000	0.000
0	102916029440	-9.000	3	BovineHD0300008694	BovineHD0300010084	27353121	32151804	4798.684	964	4.978	0.999	0.001
0	102916029440	-9.000	3	BovineHD0300030376	BovineHD0300030777	105981690	107048733	1067.044	181	5.895	0.994	0.006
0	102916029440	-9.000	4	BovineHD0400003894	Hapmap44709-BTA-70915	13098702	15575489	2476.788	439	5.642	0.998	0.002
0	102916029440	-9.000	4	BovineHD0400005756	BovineHD0400006069	19227607	20249307	1021.701	180	5.676	1.000	0.000
0	102916029440	-9.000	4	BovineHD0400009568	BovineHD0400010275	33749851	36697411	2947.561	459	6.422	1.000	0.000
0	102916029440	-9.000	4	BovineHD0400018018	BovineHD0400018470	65727477	67206589	1479.113	162	9.130	0.994	0.006
0	102916029440	-9.000	4	BovineHD0400034641	Hapmap32572-BTA-142704	118313905	120555019	2241.115	249	9.000	0.988	0.012
0	102916029440	-9.000	5	BovineHD0500010967	BovineHD0500011316	38339336	39502190	1162.855	147	7.911	0.993	0.007
0	102916029440	-9.000	5	BovineHD0500013540	BovineHD0500014334	47015593	49821847	2806.255	261	10.752	0.992	0.008
0	102916029440	-9.000	5	BovineHD0500017708	BovineHD0500018068	63361467	64604457	1242.991	198	6.278	1.000	0.000
0	102916029440	-9.000	5	BovineHD0500030954	BovineHD0500035330	107529728	120963715	13433.988	1618	8.303	0.999	0.001
0	102916029440	-9.000	6	BovineHD0600010245	BovineHD0500010487	36759019	37885679	1126.661	144	7.824	0.986	0.014
0	102916029440	-9.000	7	BovineHD0700015341	BovineHD0700015761	53361738	54507678	1145.941	173	6.624	1.000	0.000
0	102916029440	-9.000	8	BovineHD0800000328	BovineHD0800003381	975443	10479876	9504.434	1975	4.812	1.000	0.000
0	102916029440	-9.000	9	BovineHD0900004946	BovineHD0900005513	18250661	20272219	2021.559	399	5.067	0.997	0.003
0	102916029440	-9.000	9	BovineHD0900018186	BovineHD0900018563	66064158	67172083	1107.926	216	5.129	1.000	0.000
0	102916029440	-9.000	9	BovineHD0900029583	BovineHD0900030799	101651200	104924541	3273.342	573	5.713	1.000	0.000
0	102916029440	-9.000	12	BovineHD1200007319	BovineHD1200007831	24396971	25955944	1558.974	189	8.249	0.995	0.005
0	102916029440	-9.000	12	BovineHD1200008340	BovineHD4100009484	28035606	29685420	1649.815	74	22.295	0.973	0.027
0	102916029440	-9.000	12	BovineHD1200015688	BTA-23785-no-rs	56835613	57900635	1065.023	118	9.026	0.983	0.017
0	102916029440	-9.000	13	BovineHD1300023645	BovineHD1300024087	81578587	82896251	1317.665	175	7.530	0.994	0.006
0	102916029440	-9.000	15	BovineHD1500006023	BovineHD1500006528	23467684	24799473	1331.790	296	4.499	1.000	0.000
0	102916029440	-9.000	15	BovineHD1500007739	BovineHD1500008973	28783351	33106808	4323.458	645	6.703	0.997	0.003
0	102916029440	-9.000	15	BovineHD1500022754	BovineHD1500024224	78218321	82853210	4634.890	688	6.737	1.000	0.000

- roh_out_arquivo123.hom.indiv

  - Produzido quando uma bandeira da família --homozyg está presente.

  - Cada linha do arquivo .hom.indiv representa um indivíduo. O total de linhas (exceto o cabeçalho) representa o número de indivíduos.

  - O arquivo .hom.indiv tem as seguintes colunas:

- FID # Identificação da família
- IID # Identificação do indivíduo
- PHE # Valor fenotípico
- NSEG # Número de ROHs encontradas no indivíduo
- KB # Comprimento total de ROHs (kb) → Soma de todos os comprimentos de ROHs do indivíduo

- KBAVG # Comprimento médio de ROHs (kb) → Média de todos os comprimentos de ROHs do indivíduo

FID	IID	PHE	NSEG	KB	KBAVG
0	102916029440	-9	51	198256	3887.37
0	AB52502	-9	52	213103	4098.13
0	AB52503	-9	42	102783	2447.22
0	AB52505	-9	57	202055	3544.82
0	AB52508	-9	57	262986	4613.79
0	AB52511	-9	64	285553	4461.76
0	AB52512	-9	61	346225	5675.82
0	AB52514	-9	64	375508	5867.32
0	AB52516	-9	47	108046	2298.85
0	AB52517	-9	47	97550.3	2075.54

- roh_out_arquivo123.hom.summary

  - Produzido quando uma bandeira da família --homozyg está presente.

  - Cada linha do arquivo .hom.summary representa um marcador. O total de linhas (exceto o cabeçalho) representa o número de marcadores.

  - O arquivo .hom.summary tem as seguintes colunas:

- CHR # Cromossomo
- SNP # Identificação do SNP
- BP # Posição do SNP/marcador em pares de bases
- AFF # Número de casos com ROHs incluindo esse SNP/marcador
- UNAFF # Número de não-casos com ROHs incluindo esse SNP/marcador

  Observe que as amostras com fenótipos ausentes são contadas na coluna 'UNAFF'. Se o fenótipo for quantitativo, todos serão contados em 'UNAFF'.

CHR	SNP	BP	AFF	UNAFF
1	BovineHD0100046367	89725	0	46
1	BovineHD0100000035	120183	0	46
1	BovineHD0100000039	146011	0	46
1	BovineHD0100000040	147231	0	46
1	BovineHD0100000042	149772	0	46
1	BovineHD0100000043	151060	0	46
1	BovineHD0100000044	152374	0	46
1	BovineHD0100000048	158820	0	46
1	BovineHD0100000049	160007	0	46
1	BovineHD0100000052	164683	0	46
1	BovineHD0100000054	168997	0	46
1	BovineHD0100000057	183040	0	46

3.3 Como cada análise é descrita no artigo

Tomando a seguinte análise como exemplo:

--homozyg-density 100 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-window-missing 5 --homozyg-window-snp 50 --homozyg-window-threshold 0.05 --geno 0.02 --mind 0.1 --maf 0.05 --hwe 0.15 --nonfounders --out roh_out_arquivo123

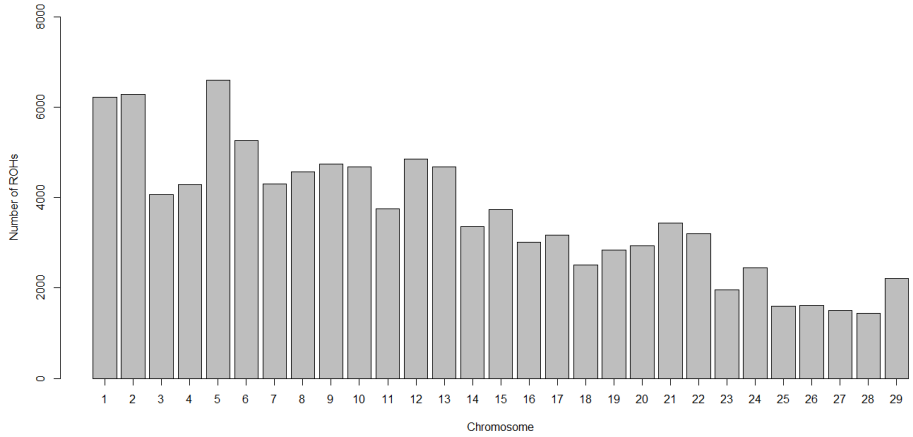
A descrição abaixo segue **respectivamente** cada uma das FLAGS do exemplo acima. (Esta é apenas uma sugestão!):

"*A density of one SNP per 100 kb was used. The maximum gap between consecutive homozygous SNPs was 1000 kb. The minimum length of a ROH was set to 1 Mb. The minimum number of consecutive SNPs included in a ROH was 50. Up to one heterozygous genotype were allowed in a ROH. A maximum of 5 SNPs with missing genotypes and a sliding window of 50 SNPs across the genome were used. The proportion of homozygous overlapping windows was 0.05. For the quality control, the analysis considered call rate of 0.98 for genotype and 0.90 for samples, minor allele frequency (MAF) of 0,05 and Hardy Weinberg equilibrium of 0,15.*"

3.4 Número e Tamanho de ROHs por cromossomo e classe

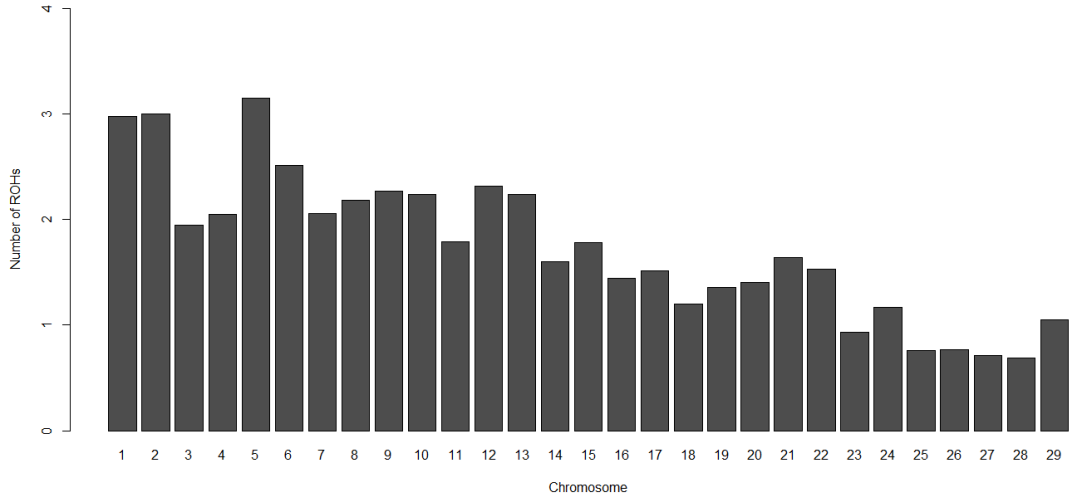
Com o arquivo .hom e o script "1_ROH_graficos_classes.R" (anexa no final) podem ser feitos alguns exemplos práticos das análises, como a montagem de gráficos com o número e tamanho de ROHs por animal para incluir na publicação do artigo ou montar uma tabela com as informações de ROH por classe de tamanho. Os resultados gráficos podem ser apresentados como o da Figura 1:

Figura 1. Exemplo do numero total de ROHs por cromossomo



Alguns artigos publicam o número total de ROHs por cromossomo (Figura 1). Outros artigos publicam o gráfico com as médias calculadas por animal e cromossomo (Figura 2). Para montar esses gráficos, verificar o script do R (1_ROH_graficos_classes.R).

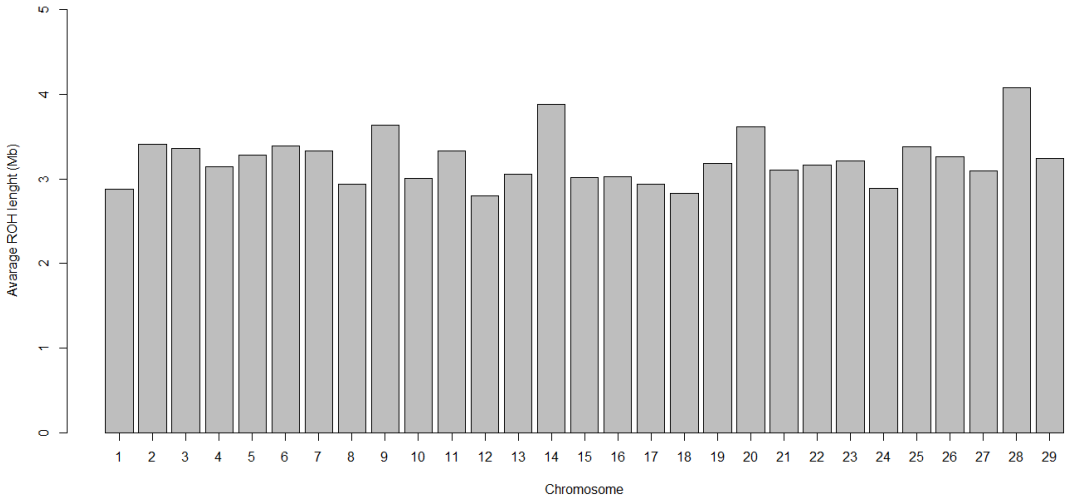
Figura 2. Exemplo do número médio de ROHs por indivíduo e cromossomo



Os gráficos das Figuras 1 e 2 não mudam “em estrutura”, só mudam em escala (eixo Y). O autor pode escolher qual acha ser mais adequado.

Com o script do R (1_ROH_graficos_classes.R) também podemos montar o gráfico do tamanho médio de ROHs por cromossomo.

Figura 3. Exemplo do tamanho médio de ROHs por cromossomo.



Script do R (1_ROH_graficos_classes.R) → **Seção**: Proporção de ROHs por tamanho
Muitos artigos exploram as ROHs com relação ao seu tamanho, pois as ROHs maiores (mais longas) indicam endogamia mais recente. O que alguns pesquisadores fazem é dividir as ROHs em cinco classes de tamanho: de 0 ou 1 a 2 Mb, 2 a 4 Mb, 4 a 8 Mb, 8 a 16 Mb e acima de 16 Mb.
Assim, o arquivo .hom e o script do R (1_ROH_graficos_classes.R) também pode ser usado para montar resultados para essas classes de ROH.

Tabela 1. Exemplo de parâmetros avaliados por classe de ROH.

Class	NROH	Percent (%)	LROH (Mb)	Number of animals	SROH	Genome coverage
$ROH_{1–2\ Mb}$	61,269	58.17	1.35	2,093	29.27	0.05%
$ROH_{2–4\ Mb}$	23,949	22.74	2.78	2,093	11.44	0.11%
$ROH_{4–8\ Mb}$	11,942	11.34	5.53	2,081	5.74	0.22%
$ROH_{8–16\ Mb}$	5,805	5.51	11.05	1,847	3.14	0.44%
$ROH_{>16\ Mb}$	2,362	2.24	24.66	1,132	2.09	0.99%

[1]NROH = número de ROHs; LROH = comprimento médio de ROHs; SROH = número médio de ROHs por animal.

A coluna Genome Coverage mostra a proporção do genoma que é coberta por essa classe do ROH. Considerando que o tamanho total do genoma bovino (espécie deste exemplo) é de 2489,37 Mb, a forma de se calcular a cobertura do genoma é dividir o

LROH pelo comprimento total do genoma, assim, (1,35 / 2489,35)*100 = 0,05% na classe $ROH_{1–2\ Mb}$ e assim por diante.

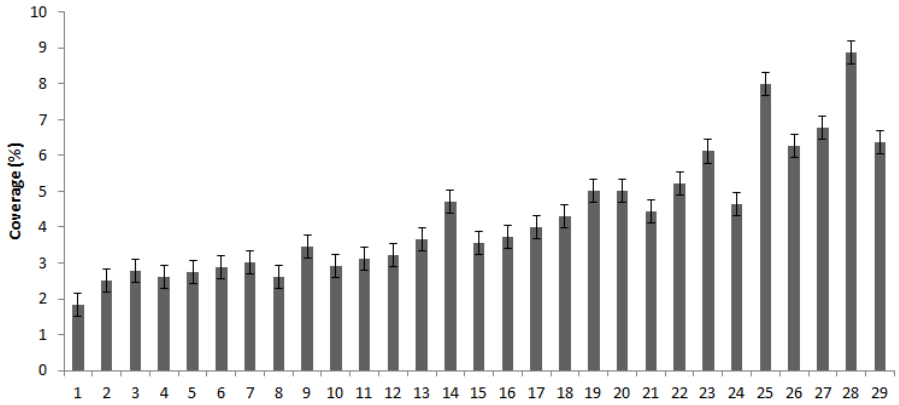
3.5 Genome Coverage por cromossomo

A informação de tamanho médio das ROHs por cromossomo, usada para construir a Figura 3, também pode ser usada juntamente com a informação do tamanho total em Megabases (Mb) de cada cromossomo para montar um gráfico com a proporção do genoma que é coberto por ROHs.

Como saber o tamanho de cada cromossomo?

Caso não tenha a informação do tamanho de cada cromossomo da espécie com a qual está trabalhando, isso pode ser pesquisado no site NCBI. O processo é relativamente rápido. Verificar tutorial "Como saber o tamanho de cada cromossomo.PDF".
Para construir um gráfico de Genome Coverage por cromossomo como na Figura 4 foi utilizado uma planilha do Excel. Ao final deste tutorial podem ser encontrados os arquivos em Rmarkdown para os scripts do R e também o esquema para montar a figura 4 no Excel (2_Genome_Coverage_xlsx).

Figura 4. Porcentagem média de cobertura dos cromossomos que é coberta por corridas de homozigose. As barras de erro indicam erro padrão



4. Calculando a endogamia baseada em ROH (FROH)

Para calcular a endogamia com base em ROH (F_{ROH}), neste caso, foi usado o pacote detectRUNS (Biscarini et al., 2019) no software R (R version 4.0.2; R Foundation for Statistical Computing, Vienna, Austria). O pacote detectRUNS é aplicado para genomas diploides.

A forma de se calcular a F_{ROH} é pela seguinte fórmula (McQuillan et al., 2008):

$$F_{ROH} = \frac{\sum_{j=1}^{n} L_{ROH}}{L_{aut}}$$

onde L_{ROH} é a soma de ROH por animal acima de um certo critério de comprimento e L_{aut} é o tamanho total dos cromossomos autossômicos, coberto por marcadores. L_{aut} para o genoma amplo aqui neste exemplo foi tomado como 2489.37Mb em comprimento, baseado nas posições do mapa ARS-UCD1.2 da montagem do genoma bovino (Rosen et al., 2020).

A endogamia pode ser obtida para cada cromossomo, para cada animal e, também, por animal e classe de ROH (de 1 a 2 Mb, 2 a 4 Mb, 4 a 8 Mb, 8 a 16 Mb e acima de 16 Mb).

O script do R (3_Calculo_FROH.R) tem a forma de se calcular a endogamia baseada em ROH com o pacote detectRUNS. Para calcular a endogamia será necessário o arquivo de mapa (arquivo123.map) e o arquivo de saída com os resultados (.hom) usado na análise de ROH do PLINK.

4.1 FROH por cromossomo e por classe

Para montar os gráficos do coeficiente de endogamia (FROH) por cromossomo (Figura 5) e por classe de ROH (Figura 6), usar o script do R (4_FROH_graficos.R).

Figura 5. Coeficiente de endogamia (F_{ROH}) por cromossomo

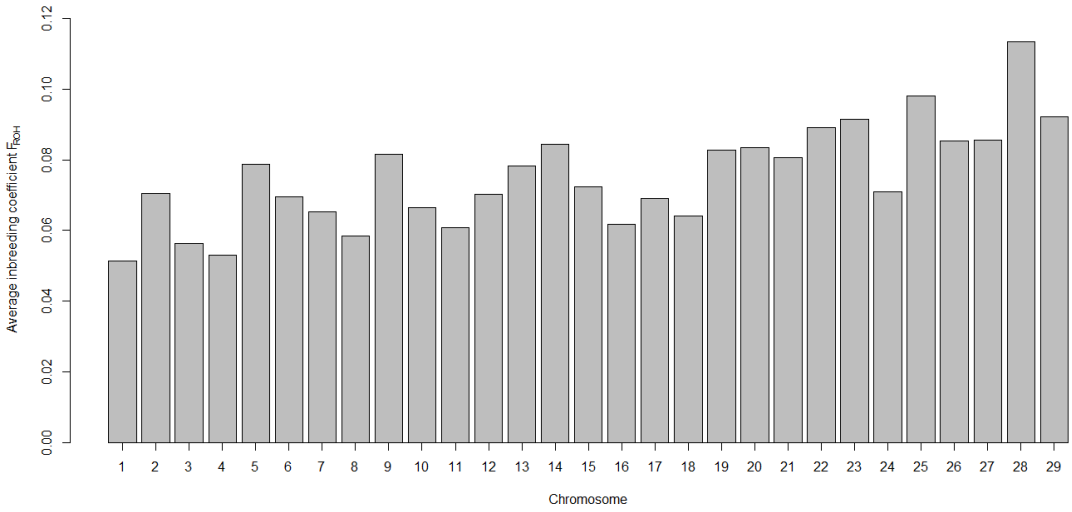
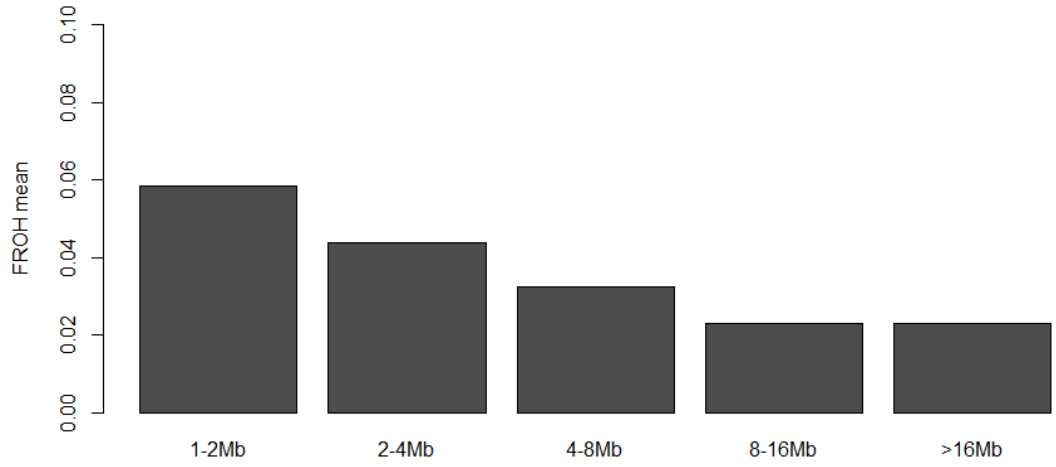


Figura 6. Coeficiente de endogamia (F_{ROH}) por classe de ROH



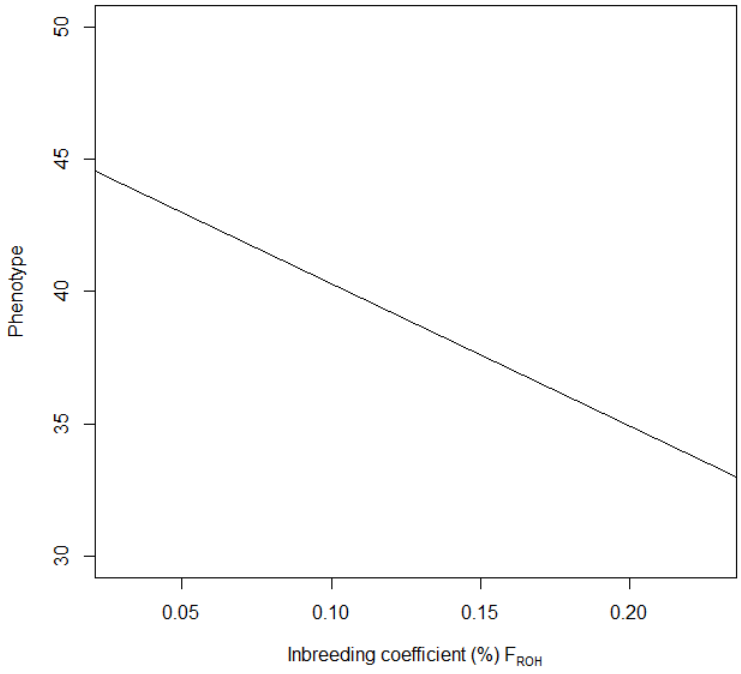
4.2 Variação de uma característica com a FROH

Caso tenha um fenótipo e queira verificar como a endogamia está interferindo nesta característica, verificar o script do R: 5_Trait_FROH.R

Para montar um gráfico com a relação entre o fenótipo e a endogamia, será preciso o arquivo de fenótipos com pelo menos uma coluna com a identificação do animal e uma coluna com os valores da característica. Também precisa do arquivo Froh_GW.txt que

contém uma coluna de identificação do animal e uma coluna com o valor da FROH por animal, gerado anteriormente com o script do R (3_Calculo_FROH.R).
Montando o gráfico da variação da característica em relação ao coeficiente de endogamia (F_{ROH}), podemos observar se a F_{ROH} afeta a característica de forma positiva ou negativa. Com o exemplo hipotético (dados2.txt) podemos observar na Figura 7 que com o aumento da endogamia temos uma queda no fenótipo (depressão endogâmica).

Figura 7. Exemplo hipotético de variação de uma característica com o coeficiente de endogamia (F_{ROH})



Para afirmar que a queda ou aumento das características é significativo, precisamos verificar o coeficiente linear e nível de significância. O script do R (5_Trait_FROH.R) também pode ser usado para salvar vários resultados da regressão da característica em função do coeficiente de endogamia (F_{ROH}).

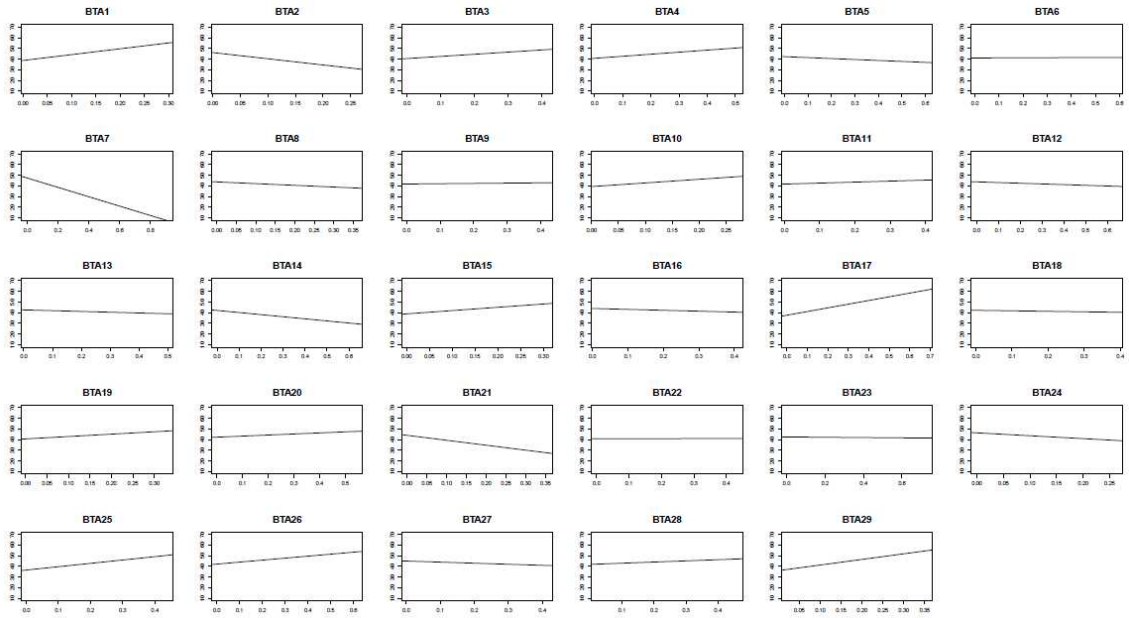
Tabela 2. Exemplo de resultados da regressão da característica em função da FROH

Item	Resultado
Intercepto	45.6974910287186
Coeficiente_linear	-53.9241104581322
Desvio_padrao_residual	17.4429714950595
Graus_de_liberdade	48
R^2	0.0122640135781325
R^2_ajustado	-0.00831381947232313

Estatística_F	0.595981780397474
p-value	0.443899685018646

Também é possível verificar como as características variam com a FROH em cada cromossomo (5_Trait_FROH.R → Seção 'Variação da característica com a F_{ROH} e cada cromossomo'). Com a endogamia por cromossomo podemos observar que a característica varia de forma positiva, negativa ou não sofre variação dependendo do cromossomo (Figura 8).

Figura 8. Exemplo hipotético de variação de uma característica com o coeficiente de endogamia (F_{ROH}) por cromossomo



Os cromossomos autossômicos da espécie *Bos taurus* são identificados pela sigla BTA: *Bos taurus autossome*.

O script do R (5_Trait_FROH.R) também pode ser usado para salvar os resultados da regressão da característica em função do coeficiente de endogamia (F_{ROH}) por cromossomo.

Referências

Biscarini, F., P. Cozzi, G. Gaspa, and G. Marras. 2019. detectRUNS: an R package to detect runs of homozygosity and heterozygosity in diploid genomes. Univ. Guelph. https://CRAN.R-project.org/package=detectRUNS.

McQuillan, R., A.L. Leutenegger, R. Abdel-Rahman, C.S. Franklin, M. Pericic, L.

Barac-Lauc, N. Smolej-Narancic, B. Janicijevic, O. Polasek, A. Tenesa, A.K. MacLeod, S.M. Farrington, P. Rudan, C. Hayward, V. Vitart, I. Rudan, S.H. Wild, M.G. Dunlop, A.F. Wright, H. Campbell, and J.F. Wilson. 2008. Runs of Homozygosity in European Populations. Am. J. Hum. Genet. 83. doi:10.1016/j.ajhg.2008.08.007.

R Core Team. 2022. ´R: A language and environment for statistical computing´ R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Rosen, B.D., D.M. Bickhart, R.D. Schnabel, S. Koren, C.G. Elsik, E. Tseng, T.N. Rowan, W.Y. Low, A. Zimin, C. Couldrey, R. Hall, W. Li, A. Rhie, J. Ghurye, S.D. McKay, F. Thibaud-Nissen, J. Hoffman, B.M. Murdoch, W.M. Snelling, T.G. McDaneld, J.A. Hammond, J.C. Schwartz, W. Nandolo, D.E. Hagen, C. Dreischer, S.J. Schultheiss, S.G. Schroeder, A.M. Phillippy, J.B. Cole, C.P. Van Tassell, G. Liu, T.P.L. Smith, and J.F. Medrano. 2020. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience 9. doi:10.1093/gigascience/giaa021.

5. Análise de FST – assinatura de seleção

Para encontrar assinaturas de seleção por meio do índice de fixação de Wright (FST), a análise pode ser feita na mesma linha de comando quando fazemos a busca por corridas de homozigose no PLINK.

- Para a análise de assinatura de seleção por FST usando o PLINK, precisamos de:
  - Arquivo adicional com as colunas: FID (família), ID (identificação do animal), cluster (grupo). Exemplo:

FID	ID	cluster
0	1	1
0	2	1
0	3	1
0	4	2
0	5	2
0	6	2
0	7	2
0	8	3
0	9	3
0	10	3

A coluna de família está zerada, pois não será usada nesta análise. A coluna de Id identifica o indivíduo e a coluna cluster indica a qual grupo esse animal pertence. O grupo pode ser referente a raças (1, 2 e 3), a indivíduos com alto e baixo valor genético para uma característica (1: alto valor genético e 2: baixo valor genético) ou outro tipo de classificação – isso vai depender da metodologia da pesquisa.

  - Parâmetros adicionados na linha de comando: --within, --fst

./plink --bfile plinkbfile_arquivo123 --cow --noweb --homozyg-density 50 --homozyg-gap 1000 --homozyg-kb 1000 --homozyg-snp 50 --homozyg-window-het 1 --homozyg-window-missing 5 --homozyg-window-snp 50 --homozyg-window-threshold 0.05 --nonfounders --geno 0.02 --maf 0.05 --mind 0.1 --within arquivo_com_grupos.ped --fst --out roh_out_arquivo123

Resultados

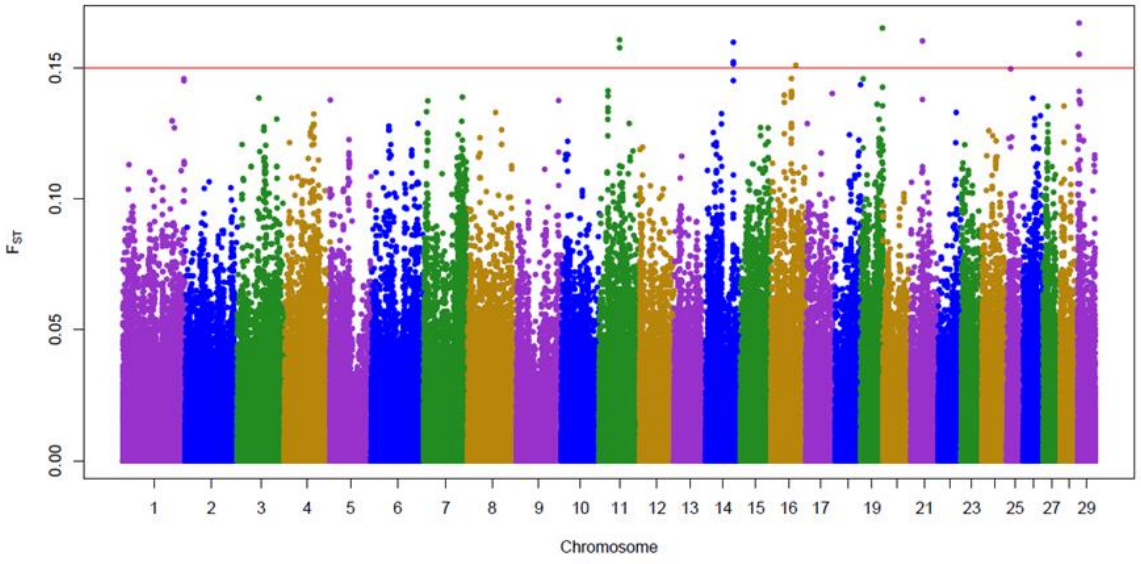
Além dos resultados da análise de ROH, um arquivo .fst será resultante dessa análise. Este arquivo contém as colunas:

- CHR # Cromossomo
- SNP # Identificação do marcador

- POS # Posição do marcador em pares de bases
- NMISS # Número total de indivíduos usados na análise
- FST # Valor de FST para cada marcador

O script do R (6_ManhattanPlot_FST.R) pode ser usado para montar um Manhattan Plot com os resultados do arquivo .fst, como no exemplo abaixo:

Figura 9. Exemplo de Manhattan Plot que pode ser construído para identificar assinaturas de seleção pela metodologia FST



Como o threshold (nível de significância) foi definido?

- Fst é uma medida de diferenciação entre duas populações.
- Valores variando de 0 (nenhuma diferença entre as populações) a 1 (diferenças fixas entre as populações).
- Valores de Fst entre 0 e 0,05 indicam pouca diferenciação genética
- Valores de Fst entre 0,05 e 0,15 indicam diferenciação genética moderada
- Valores de Fst entre 0,15 e 0,25 indicam grande diferenciação genética
- Valores de Fst acima de 0,25 indicam um grau muito grande de diferenciação genética

Fonte: Wright, S. 1978. Evolution and the Genetics of Populations. Vol. 4. Variability Within and Among Natural Populations. University Chicago Press, Chicago, USA.

- A forma de definir o threshold pode depender do pesquisador.

- O threshold estabelecido neste caso foi de 0,15, porque não havia valores acima de 0,25.

Com o script **6_ManhattanPlot_FST.R** também podemos verificar e salvar as regiões (posições dos marcadores e cromossomos) acima do threshold estabelecido, no caso, FST > 0,15.

O próximo passo seria uma análise de Ontologia Gênica:
- Busca de genes nestas regiões genômicas
- Pesquisa por processos biológicos relacionados a estes genes

1_ROH_graficos_classes.R

Particular

2023-10-04

```
#######################################
#### TUTORIAL  - GRAFICOS DE ROHs ####
#######################################

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

# direcionar para o diretorio onde estao os arquivos
setwd("D:\\PessoalD\\Doutorado\\Cap 2 ROH\\Tutorial\\1_ROH_FROH")

saida_hom<-read.table("roh_out_gir_total.hom",h=T) # Lendo o arquivo .hom
dim(saida_hom) # N?mero de linhas e colunas (dimens?o) do arquivo
```

```
## [1] 105327     13
```

```
head(saida_hom) # Visualizando as primeiras 6 linhas do arquivo
```

```
##   FID          IID PHE CHR               SNP1               SNP2      POS1      POS2        KB NSNP
DENSITY  PHOM  PHET
## 1   0 102916029440  -9   1 BovineHD0100006914 BovineHD0100012022  23381349  42095937 18714.589 2816
6.646 1.000 0.000
## 2   0 102916029440  -9   1 BovineHD0100047134 BovineHD0100029888 102458103 105162097  2703.995  377
7.172 0.995 0.005
## 3   0 102916029440  -9   1 BovineHD0100041142 BovineHD0100043043 143141556 148784184  5642.629  975
5.787 0.998 0.002
## 4   0 102916029440  -9   1 BovineHD0100043081 BovineHD0100044197 148857503 152070453  3212.951  397
8.093 0.997 0.003
## 5   0 102916029440  -9   1 BovineHD0100044217 BovineHD0100045574 152126033 156040374  3914.342  685
5.714 0.999 0.001
## 6   0 102916029440  -9   2 BovineHD0200019851 BovineHD0200020565  68732542  71627667  2895.126  181
15.995 1.000 0.000
```

```
saida_hom$MB<-saida_hom$KB/1000 # Criando uma coluna com o tamanho das ROHs em Mbases


#### Numero de ROHs por individuo ####
roh_animais<-as.data.frame(table(saida_hom$IID)) # Numero de ROHs por individuo
# Esse numero de ROHs por individuo é o mesmo que a coluna NSEG do arquivo .hom.indiv
head(roh_animais)
```

```
##           Var1 Freq
## 1 102916029440   51
## 2      AB52502   52
## 3      AB52503   41
## 4      AB52505   55
## 5      AB52508   58
## 6      AB52511   64
```

```
mean(roh_animais$Freq) # Numero medio de ROHs por individuo
```

```
## [1] 50.32346
```

```
sd(roh_animais$Freq) # Desvio padrao do numero medio de ROHs por individuo
```

```
## [1] 8.591984
```

```
min(roh_animais$Freq) # Numero minimo de ROHs por individuo
```

```
## [1] 26
```

```
max(roh_animais$Freq) # Numero maximo de ROHs por individuo
```

```
## [1] 95
```

```
#### Numero de ROHs por cromossomo ####
N_ROH_CHR<-as.data.frame(table(saida_hom$CHR))
head(N_ROH_CHR) # Numero total de ROHs por cromossomo
```

```
##   Var1 Freq
## 1    1 6227
## 2    2 6289
## 3    3 4071
## 4    4 4284
## 5    5 6600
## 6    6 5260
```

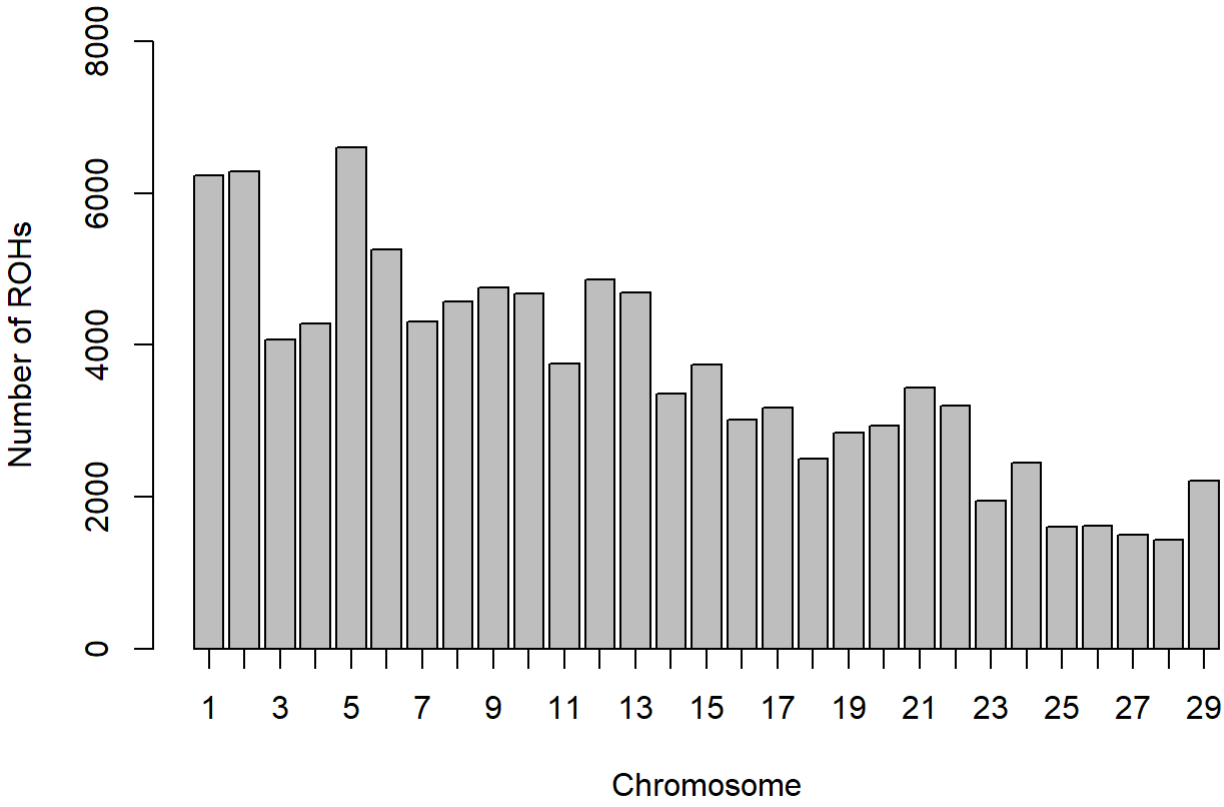
```
# Se quiser, pode fazer o grafico com esse arquivo
# Se for usar e quiser salvar esse arquivo, tirar o # da linha de comando abaixo.
#write.table(N_ROH_CHR, "Resultado_numero_ROH_por_CHR.txt",quote=F,row.names=F,colnames=T)

#### Grafico do numero total de ROHs por cromossomo ####
# Exemplo do numero total de ROHs por cromossomo
max(N_ROH_CHR$Freq) # Verificando qual o valor maximo para estabelecer o ylim no comando abaixo
```

```
## [1] 6600
```

```
w<-barplot(N_ROH_CHR$Freq,ylim=c(0,8000),ylab="Number of ROHs",xlab="Chromosome")
w
```

```
##       [,1]
##  [1,]  0.7
##  [2,]  1.9
##  [3,]  3.1
##  [4,]  4.3
##  [5,]  5.5
##  [6,]  6.7
##  [7,]  7.9
##  [8,]  9.1
##  [9,] 10.3
## [10,] 11.5
## [11,] 12.7
## [12,] 13.9
## [13,] 15.1
## [14,] 16.3
## [15,] 17.5
## [16,] 18.7
## [17,] 19.9
## [18,] 21.1
## [19,] 22.3
## [20,] 23.5
## [21,] 24.7
## [22,] 25.9
## [23,] 27.1
## [24,] 28.3
## [25,] 29.5
## [26,] 30.7
## [27,] 31.9
## [28,] 33.1
## [29,] 34.3
```

```
axis(1, at=w,labels=1:29)
```



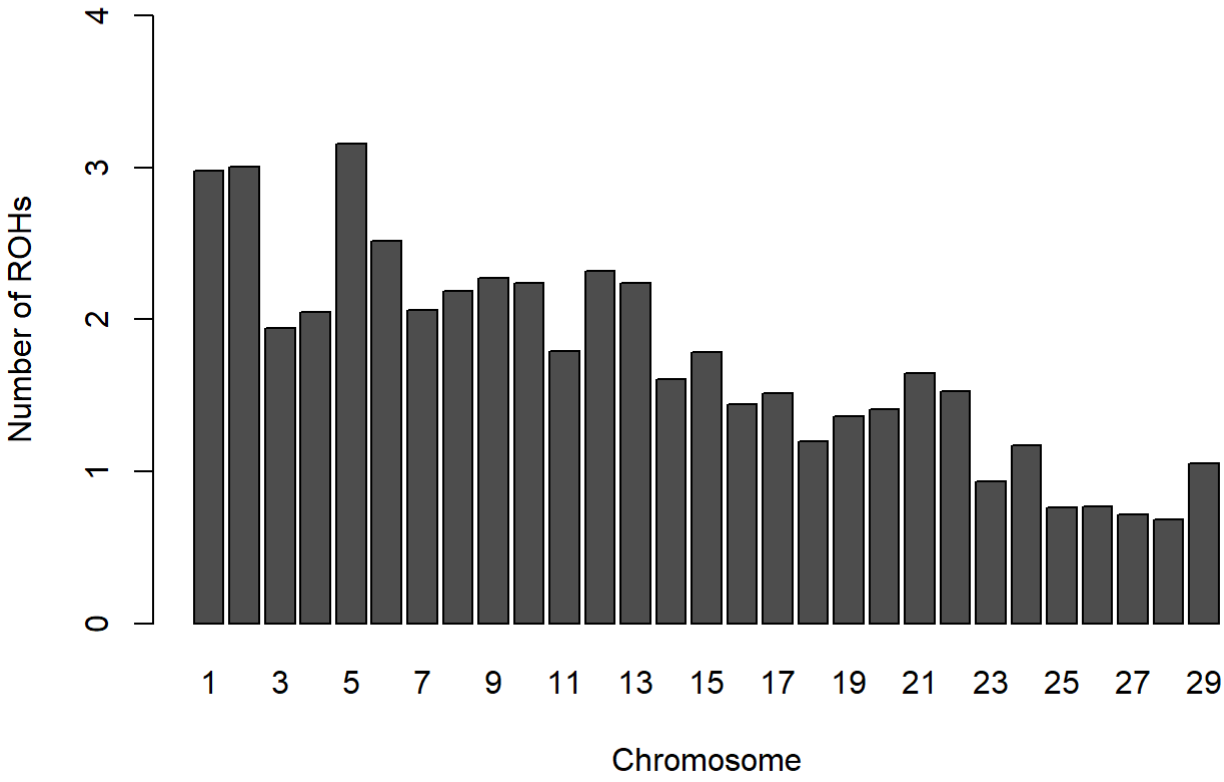
```
# Se for usar esse grafico, SALVAR! Export -> Save as Image ... ou ... Arquivo -> salvar como...

#### Numero de ROHs por individuo e cromossomo ####
N_ROH_IID_CHR<-table(saida_hom$IID, saida_hom$CHR)
N_ROH_IID_CHR1<-as.data.frame(colMeans(N_ROH_IID_CHR))
N_ROH_IID_CHR1
```

```
##    colMeans(N_ROH_IID_CHR)
## 1                2.9751553
## 2                3.0047778
## 3                1.9450549
## 4                2.0468227
## 5                3.1533684
## 6                2.5131390
## 7                2.0592451
## 8                2.1839465
## 9                2.2704252
## 10               2.2365026
## 11               1.7940755
## 12               2.3210702
## 13               2.2398471
## 14               1.6053512
## 15               1.7849976
## 16               1.4409938
## 17               1.5150502
## 18               1.1978022
## 19               1.3602484
## 20               1.4056378
## 21               1.6445294
## 22               1.5284281
## 23               0.9331104
## 24               1.1720019
## 25               0.7634974
## 26               0.7716197
## 27               0.7152413
## 28               0.6865743
## 29               1.0549451
```

```
# Se for usar e quiser salvar esse arquivo, tirar o # da linha de comando abaixo.
#write.table(N_ROH_IID_CHR1, "Resultado_numero_ROH_por_indiv_CHR.txt",quote=F,row.names=F,colnames=T)


#### Grafico do numero de ROHs por individuo e cromossomo ####
# Exemplo do numero medio de ROHs por individuo e cromossomo
max(N_ROH_IID_CHR1)
```

```
## [1] 3.153368
```

```
barplot(t(N_ROH_IID_CHR1),ylim=c(0,4),ylab="Number of ROHs",xlab="Chromosome")
```




```
# Se for usar esse grafico, SALVAR!



#### Comprimento medio de ROHs por cromossomo (Mb) ####
size_CHR<-aggregate(MB ~ CHR, saida_hom, mean)
size_CHR
```

```
##    CHR       MB
## 1    1 2.882119
## 2    2 3.409471
## 3    3 3.362215
## 4    4 3.144614
## 5    5 3.282147
## 6    6 3.385661
## 7    7 3.328516
## 8    8 2.940057
## 9    9 3.630920
## 10  10 3.009317
## 11  11 3.332975
## 12  12 2.803641
## 13  13 3.056790
## 14  14 3.880058
## 15  15 3.012148
## 16  16 3.025738
## 17  17 2.933941
## 18  18 2.827001
## 19  19 3.183279
## 20  20 3.612372
## 21  21 3.099087
## 22  22 3.161019
## 23  23 3.209078
## 24  24 2.883651
## 25  25 3.377691
## 26  26 3.258615
## 27  27 3.095501
## 28  28 4.078299
## 29  29 3.245302
```

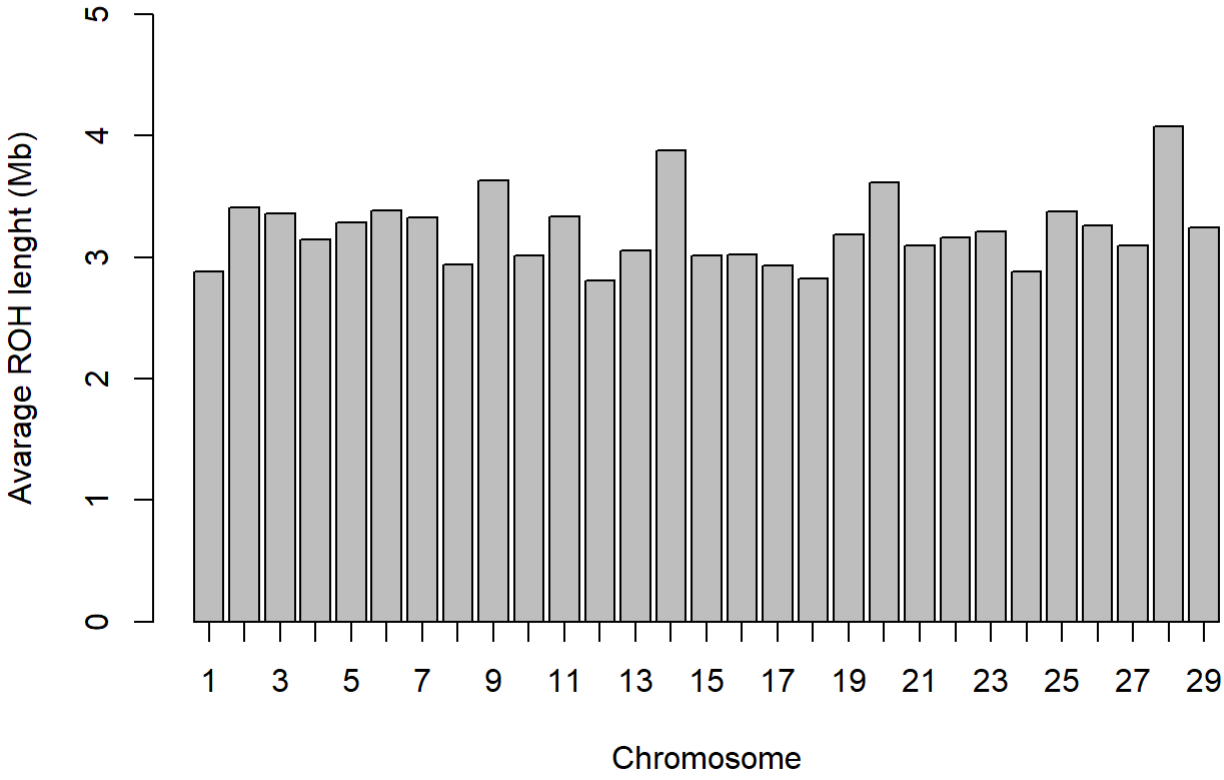
```
# Se for usar e quiser salvar esse arquivo, tirar o # da linha de comando abaixo.
#write.table(size_CHR, "Resultado_tamanho_ROH_por_CHR.txt",quote=F,row.names=F,colnames=T)

#### Grafico do comprimento medio de ROHs por cromossomo (Mb) ####
# Tamanho medio de ROHs por cromossomo
max(size_CHR$MB)
```

```
## [1] 4.078299
```

```
y<-barplot(size_CHR$MB,ylim=c(0,5),ylab="Avarage ROH lenght (Mb)",xlab="Chromosome")
y
```

```
##       [,1]
##  [1,]  0.7
##  [2,]  1.9
##  [3,]  3.1
##  [4,]  4.3
##  [5,]  5.5
##  [6,]  6.7
##  [7,]  7.9
##  [8,]  9.1
##  [9,] 10.3
## [10,] 11.5
## [11,] 12.7
## [12,] 13.9
## [13,] 15.1
## [14,] 16.3
## [15,] 17.5
## [16,] 18.7
## [17,] 19.9
## [18,] 21.1
## [19,] 22.3
## [20,] 23.5
## [21,] 24.7
## [22,] 25.9
## [23,] 27.1
## [24,] 28.3
## [25,] 29.5
## [26,] 30.7
## [27,] 31.9
## [28,] 33.1
## [29,] 34.3
```

```
axis(1, at=y,labels=1:29)
```



```
# Se for usar esse grafico, SALVAR!



########################################
#### Proporcao de ROHs por tamanho ####
#### ROH de 1 a 2 Mb, 2 a 4 Mb, ... ###
########################################



#### Separando as ROHs de acordo com o tamanho
head(saida_hom)
```

```
##   FID          IID PHE CHR               SNP1               SNP2      POS1      POS2        KB NSNP
DENSITY  PHOM  PHET        MB
## 1   0 102916029440  -9   1 BovineHD0100006914 BovineHD0100012022  23381349  42095937 18714.589 2816
6.646 1.000 0.000 18.714589
## 2   0 102916029440  -9   1 BovineHD0100047134 BovineHD0100029888 102458103 105162097  2703.995  377
7.172 0.995 0.005  2.703995
## 3   0 102916029440  -9   1 BovineHD0100041142 BovineHD0100043043 143141556 148784184  5642.629  975
5.787 0.998 0.002  5.642629
## 4   0 102916029440  -9   1 BovineHD0100043081 BovineHD0100044197 148857503 152070453  3212.951  397
8.093 0.997 0.003  3.212951
## 5   0 102916029440  -9   1 BovineHD0100044217 BovineHD0100045574 152126033 156040374  3914.342  685
5.714 0.999 0.001  3.914342
## 6   0 102916029440  -9   2 BovineHD0200019851 BovineHD0200020565  68732542  71627667  2895.126  181
15.995 1.000 0.000  2.895126
```

```
mb_1_2<-saida_hom[saida_hom$MB<2,] # Arquivo com ROHs de 1 a 2 Mb
Porc1_2Mb<-(nrow(mb_1_2)*100)/nrow(saida_hom) # Proporcao de ROHs de 1 a 2 Mb
mb_2_4<-saida_hom[saida_hom$MB>2 & saida_hom$MB<4,] # Arquivo com ROHs de 2 a 4 Mb
Porc2_4Mb<-(nrow(mb_2_4)*100)/nrow(saida_hom) # Proporcao de ROHs de 2 a 4 Mb
mb_4_8<-saida_hom[saida_hom$MB>4 & saida_hom$MB<8,] # Arquivo com ROHs de 4 a 8 Mb
Porc4_8Mb<-(nrow(mb_4_8)*100)/nrow(saida_hom) # Proporcao de ROHs de 4 a 8 Mb
mb_8_16<-saida_hom[saida_hom$MB>8 & saida_hom$MB<16,] # Arquivo com ROHs de 8 a 16 Mb
Porc8_16Mb<-(nrow(mb_8_16)*100)/nrow(saida_hom) # Proporcao de ROHs de 8 a 16 Mb
mb_16<-saida_hom[saida_hom$MB>16,] # Arquivo com ROHs acima de 16 Mb
Porc16Mb<-(nrow(mb_16)*100)/nrow(saida_hom) # Proporcao de ROHs acima de 16 Mb

###### CRIANDO AS COLUNAS DA TABELA #####

# Nomes das classes
classes<-rbind("ROH1-2 Mb","ROH2-4 Mb","ROH4-8 Mb","ROH8-16 Mb","ROH>16 Mb")
# Numero de ROHs total por classe
N_ROH<-rbind(nrow(mb_1_2),nrow(mb_2_4),nrow(mb_4_8),nrow(mb_8_16),nrow(mb_16))
# Proporcao dos ROHs nas classes
Porc_Mb<-rbind(Porc1_2Mb,Porc2_4Mb,Porc4_8Mb,Porc8_16Mb,Porc16Mb)
# Tamanho das ROHs por classe (em Megabases)
size<-rbind(mean(mb_1_2$MB),mean(mb_2_4$MB),mean(mb_4_8$MB),mean(mb_8_16$MB),mean(mb_16$MB))
# Numero de animais que tem estas classes
N_IID<-rbind(nrow(as.data.frame(table(mb_1_2$IID))),nrow(as.data.frame(table(mb_2_4$IID))),
             nrow(as.data.frame(table(mb_4_8$IID))),nrow(as.data.frame(table(mb_8_16$IID))),
             nrow(as.data.frame(table(mb_16$IID))))
# Numero medio de ROHs por individuo
N_ROH_IID<-rbind(mean(as.data.frame(table(mb_1_2$IID))[,2]),
                 mean(as.data.frame(table(mb_2_4$IID))[,2]),
                 mean(as.data.frame(table(mb_4_8$IID))[,2]),
                 mean(as.data.frame(table(mb_8_16$IID))[,2]),
                 mean(as.data.frame(table(mb_16$IID))[,2]))

###### CRIANDO A TABELA #####
# Tutorial - Tabela 1. Exemplo de parametros avaliados por classe de ROH
tabela_ROH<-cbind(classes,N_ROH,Porc_Mb,size,N_IID,N_ROH_IID)
tabela_ROH
```

```
##            [,1]         [,2]    [,3]               [,4]               [,5]   [,6]
## Porc1_2Mb  "ROH1-2 Mb"  "61269" "58.1702697314079" "1.34880904263167" "2093" "29.2732919254658"
## Porc2_4Mb  "ROH2-4 Mb"  "23949" "22.7377595488336" "2.77562034448202" "2093" "11.4424271380793"
## Porc4_8Mb  "ROH4-8 Mb"  "11942" "11.3380234887541" "5.52765317082566" "2081" "5.73858721768381"
## Porc8_16Mb "ROH8-16 Mb" "5805"  "5.51140733145347" "11.0534292037898" "1847" "3.1429344883595"
## Porc16Mb   "ROH>16 Mb"  "2362"  "2.24253989955092" "24.6646769665538" "1132" "2.08657243816254"
```

2_Genome_Coverage_xlsx

Dados

Exemplo para calcular a proporção do genoma coberta por ROH em cada cromossomo				
Cromossomo	**Tamanho do Cromossomo em Mb**	**Tamanho médio das ROHs**	**Fórmula[1]**	**Proporção[2]**
BTA1	158,53	2,88	(Tamanho médio das ROHs *100) / Tamanho do Cromossomo em Mb	1,82
BTA2	136,23	3,41		2,50
BTA3	121,01	3,36		2,78
BTA4	120	3,14		2,62
BTA5	120,09	3,28		2,73
BTA6	117,81	3,39		2,87
BTA7	110,68	3,33		3,01
BTA8	113,32	2,94		2,59
BTA9	105,45	3,63		3,44
BTA10	103,31	3,01		2,91
BTA11	106,98	3,33		3,12
. . .	. . .	. . .	. . .	. . .
BTA29	51,1	3,25	6,35	6,35

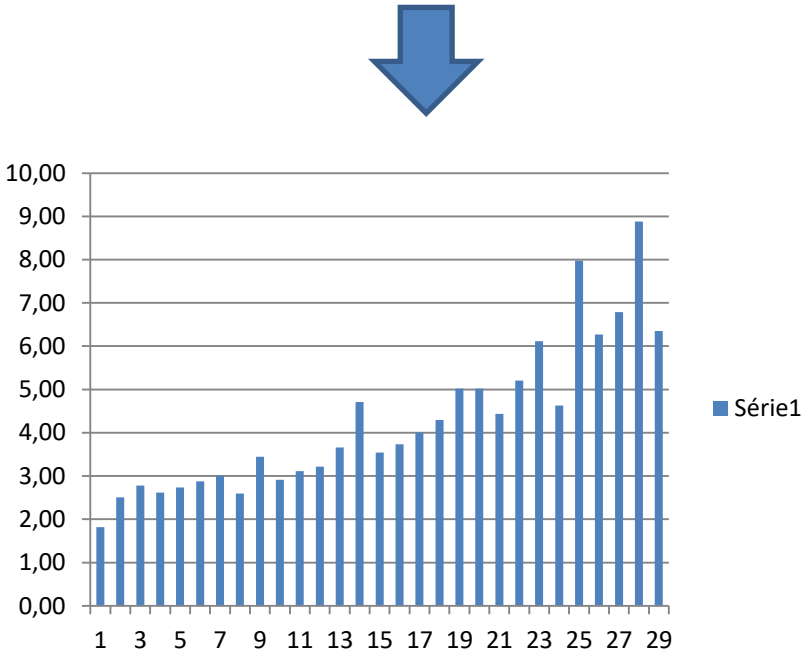
[1]A coluna "fórmula" contém a fórmula para calcular a proporção do cromossomo que é coberta por ROHs
[2]A coluna "Proporção" contém os valores sem a fórmula. Assim essa coluna pode ser usada para montar o gráfico

No Excel

Exemplo para calcular a proporção do genoma coberta por ROH em cada cromossomo				
Cromossomo	Tamanho do Cromossomo em Mb	Tamanho médio das ROHs	Fórmula[1]	Proporção[2]
BTA1	158,53	2,88	1,82	1,82
BTA2	136,23	3,41	2,50	2,50
BTA3	121,01	3,36	2,78	2,78
BTA4	120	3,14	2,62	2,62
BTA5	120,09	3,28	2,73	2,73
BTA6	117,81	3,39	2,87	2,87
BTA7	110,68	3,33	3,01	3,01
BTA8	113,32	2,94	2,59	2,59
BTA9	105,45	3,63	3,44	3,44
BTA10	103,31	3,01	2,91	2,91
BTA11	106,98	3,33	3,12	3,12
BTA12	87,22	2,80	3,21	3,21
BTA13	83,47	3,06	3,66	3,66
BTA14	82,4	3,88	4,71	4,71
BTA15	85,01	3,01	3,54	3,54
BTA16	81,01	3,03	3,74	3,74
BTA17	73,17	2,93	4,01	4,01
BTA18	65,82	2,83	4,30	4,30
BTA19	63,45	3,18	5,02	5,02
BTA20	71,97	3,61	5,02	5,02
BTA21	69,86	3,10	4,44	4,44
BTA22	60,77	3,16	5,20	5,20
BTA23	52,5	3,21	6,11	6,11
BTA24	62,32	2,88	4,63	4,63
BTA25	42,35	3,38	7,98	7,98
BTA26	51,99	3,26	6,27	6,27
BTA27	45,61	3,10	6,79	6,79
BTA28	45,94	4,08	8,88	8,88
BTA29	51,1	3,25	6,35	6,35

- Selecionar a coluna de proporção

- Guia ‘Inserir’
- Selecionar ‘Gráfico de colunas’

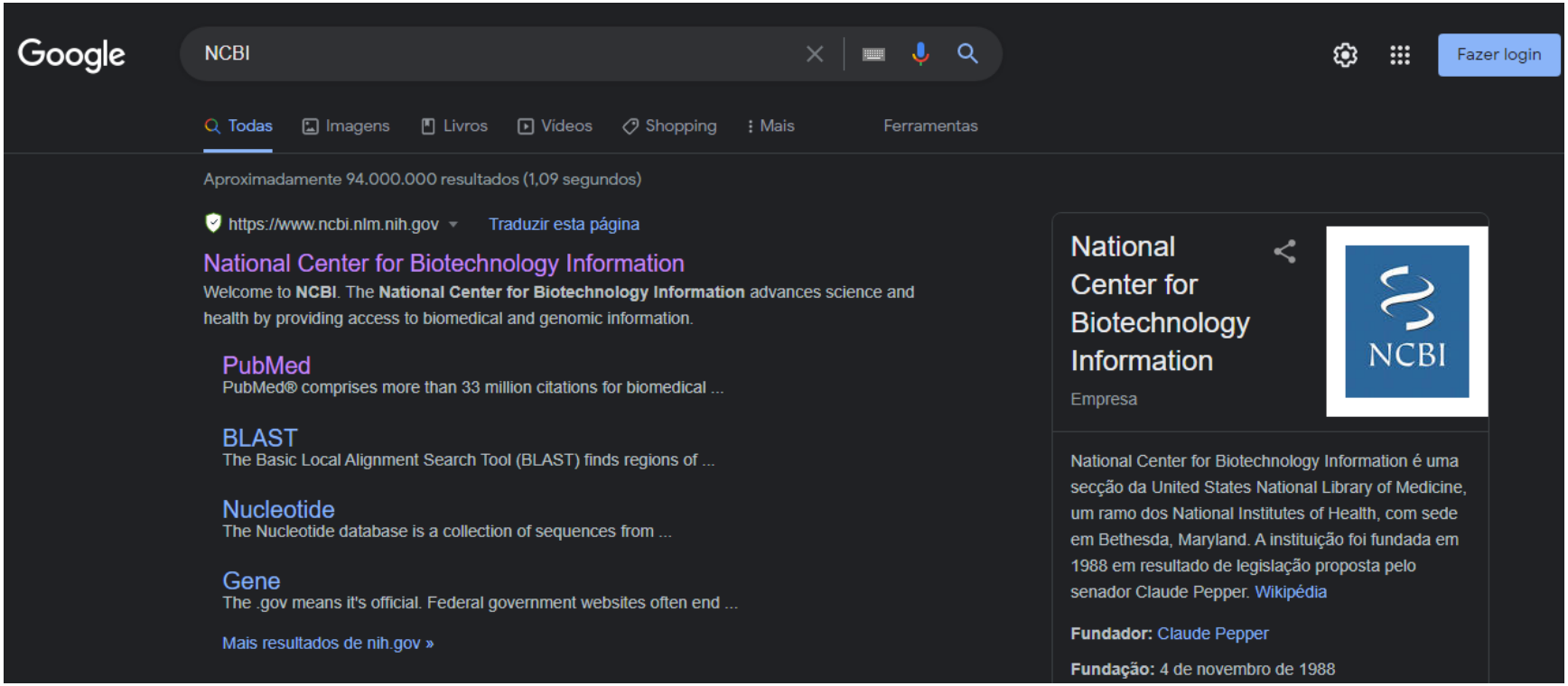


Demais edições podem ser feitas na guia de Ferramentas de Gráfico

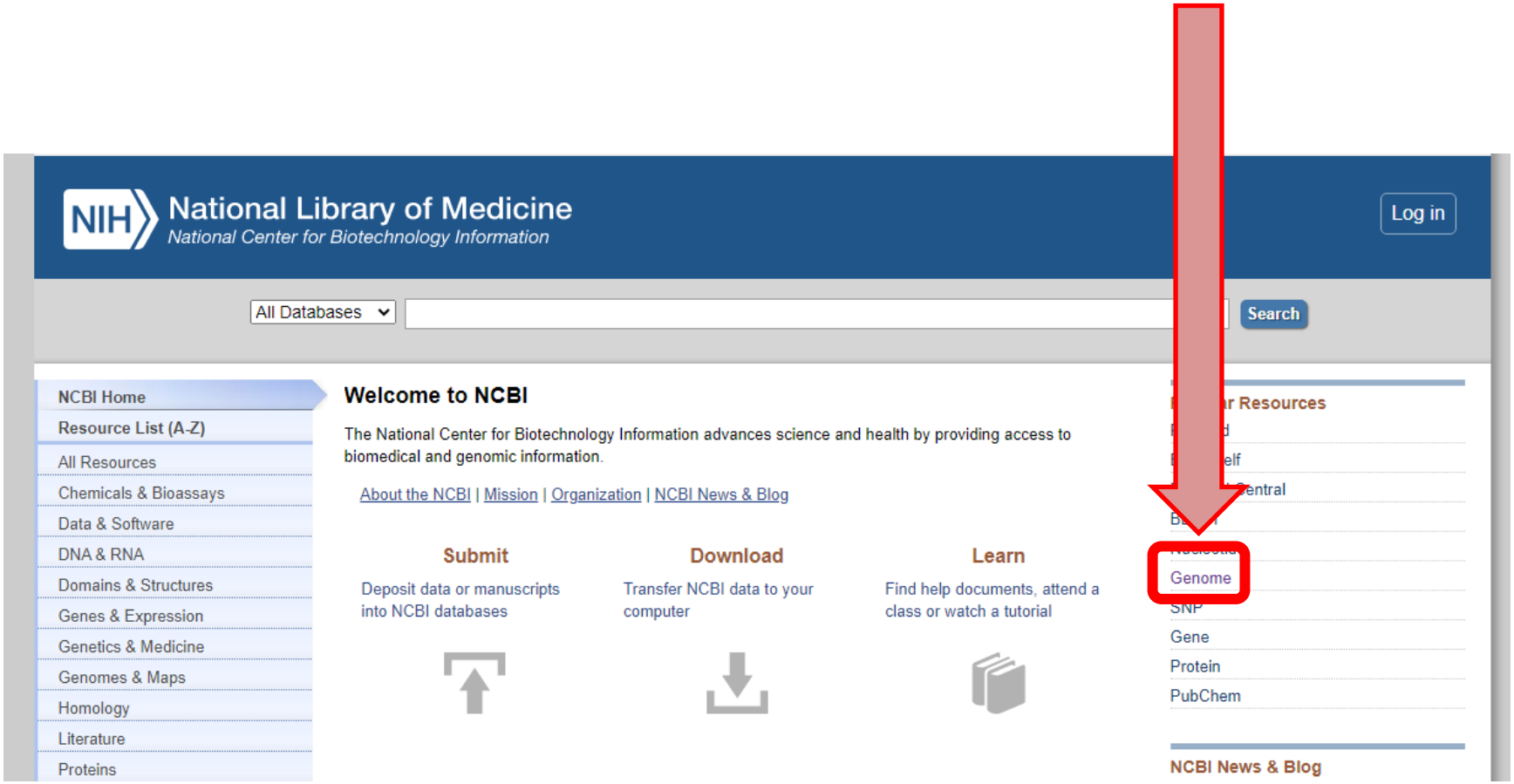
Como saber o tamanho de cada cromossomo?

Caso não tenha a informação do tamanho de cada cromossomo da espécie com a qual está trabalhando, isso pode ser pesquisado no site NCBI

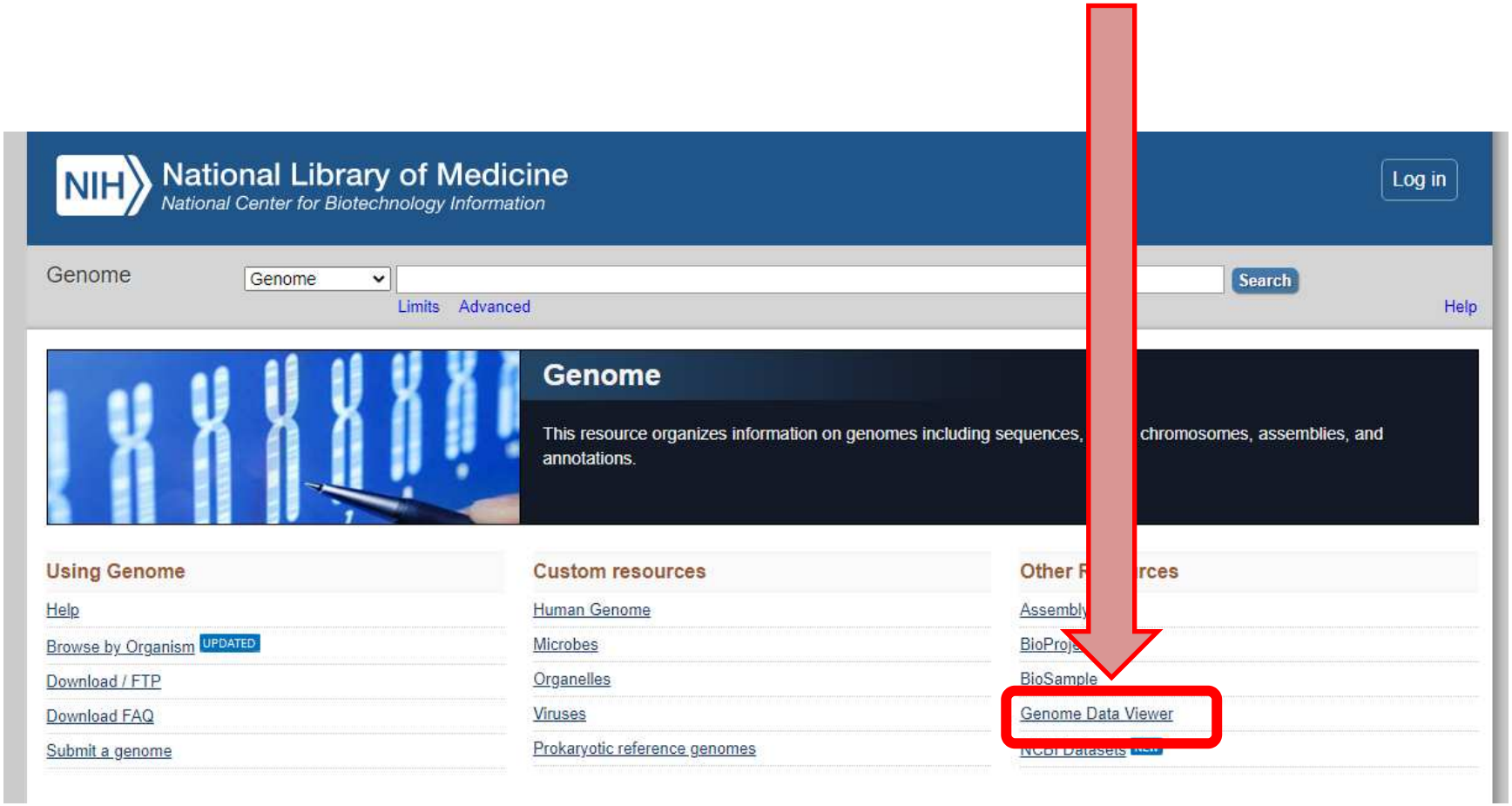
- Digitar NCBI na guia de pesquisa do google
- Clicar no primeiro link



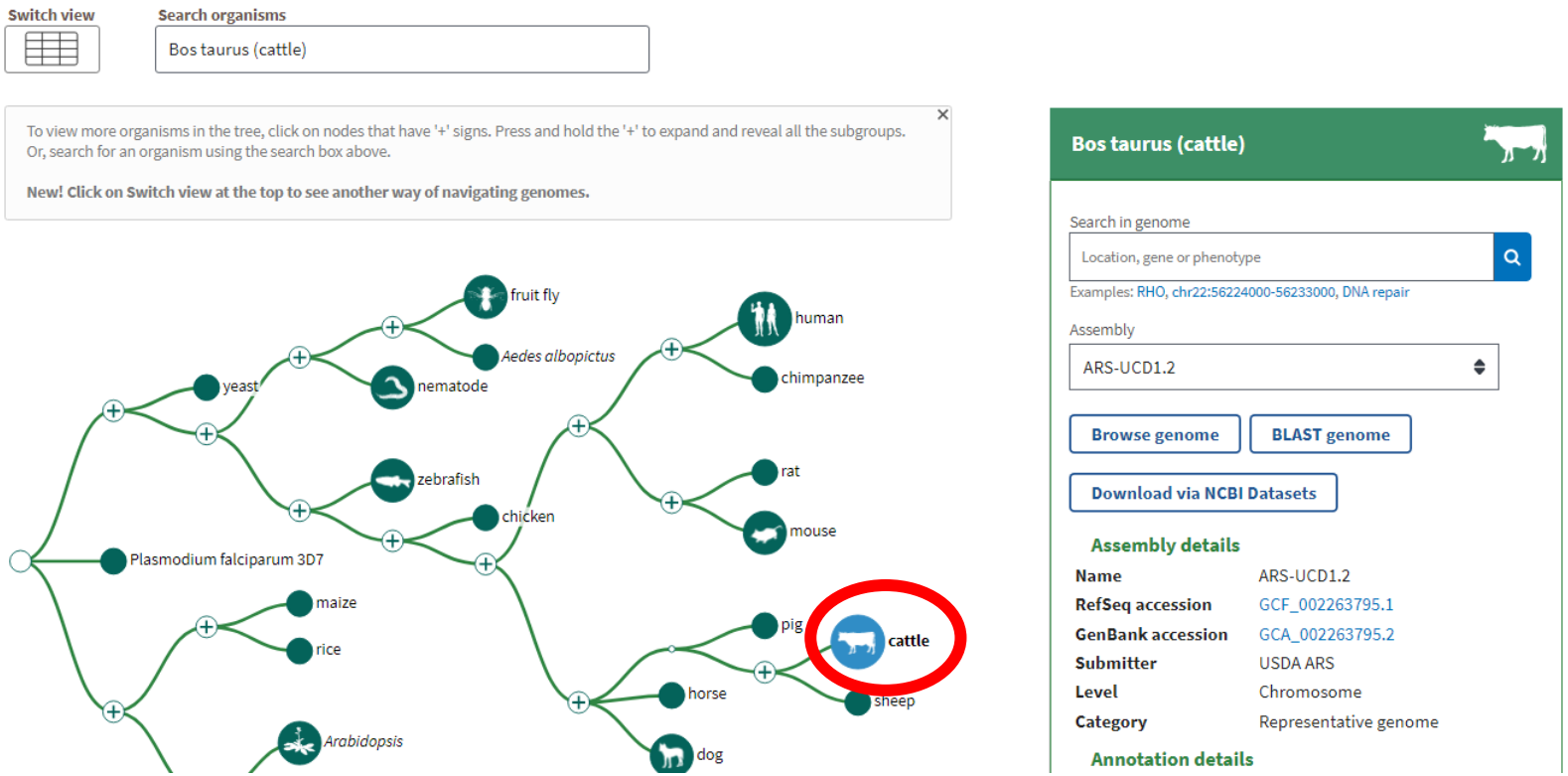
- Clicar em Genome no menu no canto direito



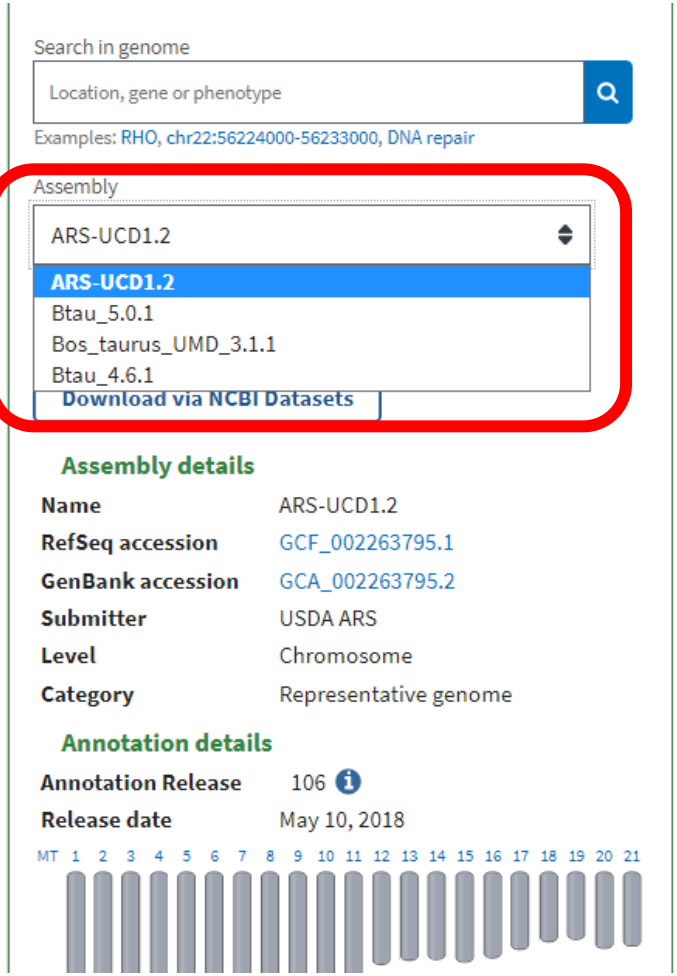
- Clicar em Genome Data Viewer



- Selecionar a espécie de interesse (neste caso foi a espécie *Bos taurus*)
- Caso a espécie da sua pesquisa não esteja visível, clique nos “+” no diagrama para ver mais espécies



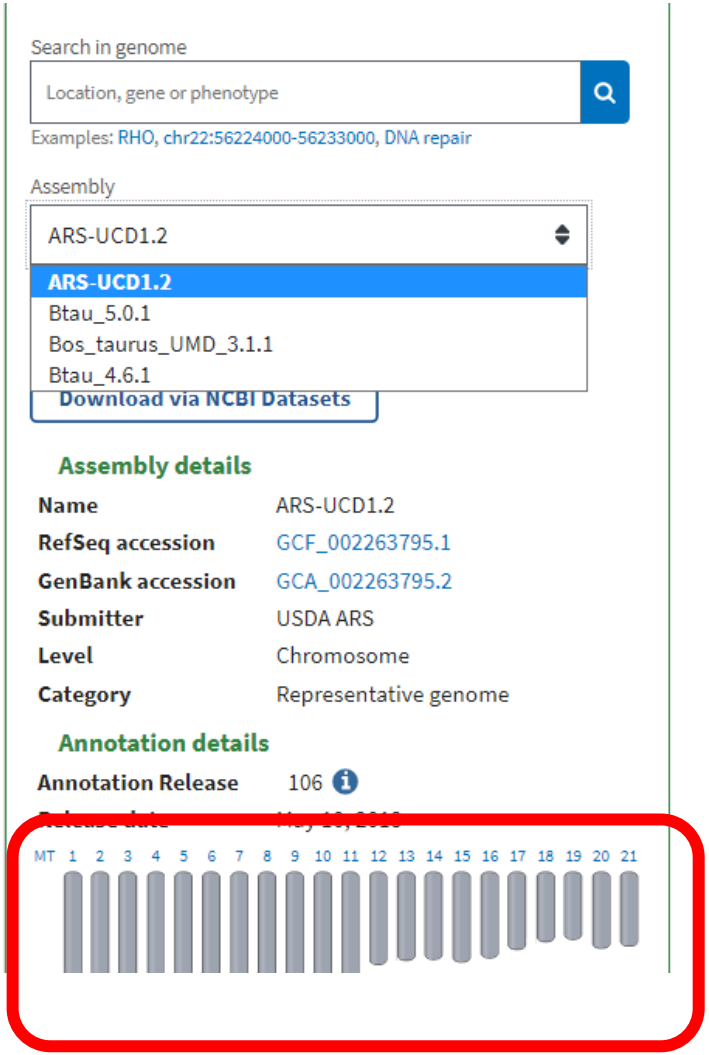
No menu à direita, é importante selecionar corretamente a Assembly



- Se você estiver trabalhando com animais genotipados, a fazenda/empresa que genotipou os animais pode te passar essa informação.

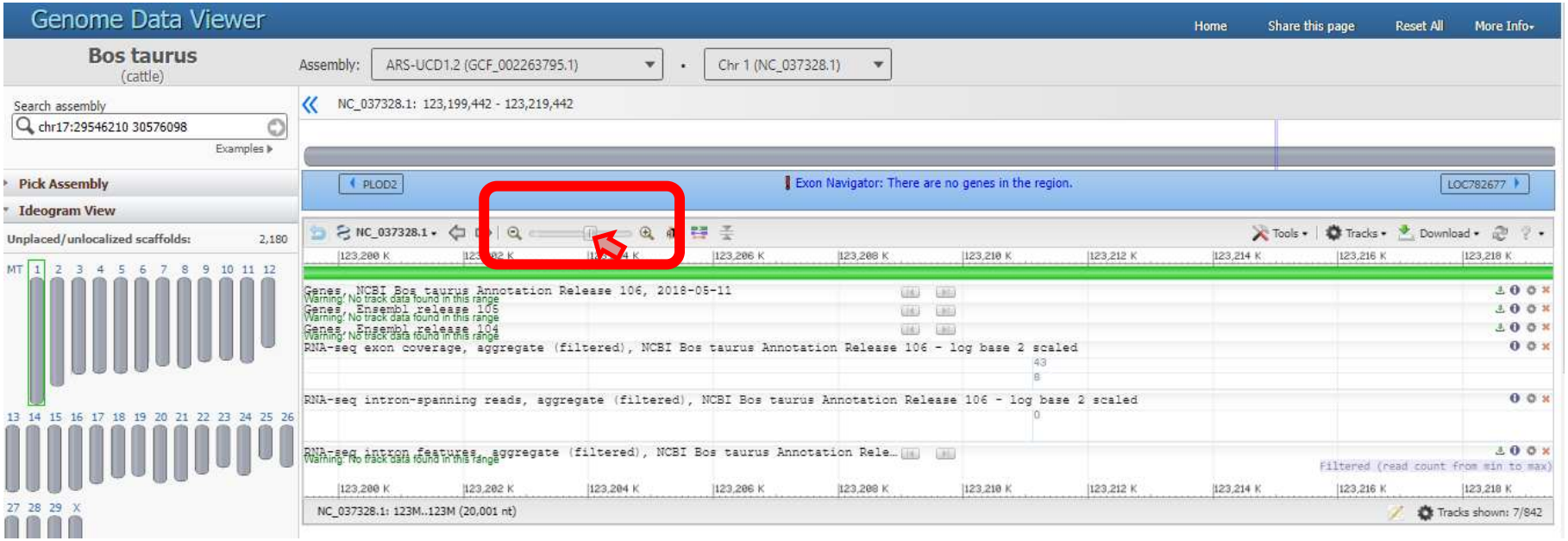
- Neste caso, a opção é ARS-UCD1.2

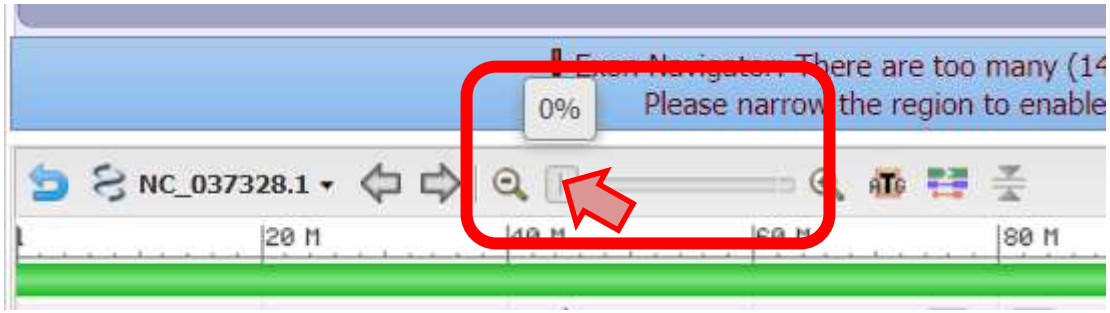
Ainda no menu à direita



- Os cromossomos da espécie estarão listados abaixo
- Clicar no Cromossomo 1
- Vai carregar uma nova página

- Na nova página, encontre a opção da lupa e deixe a opção de zoom em 0%

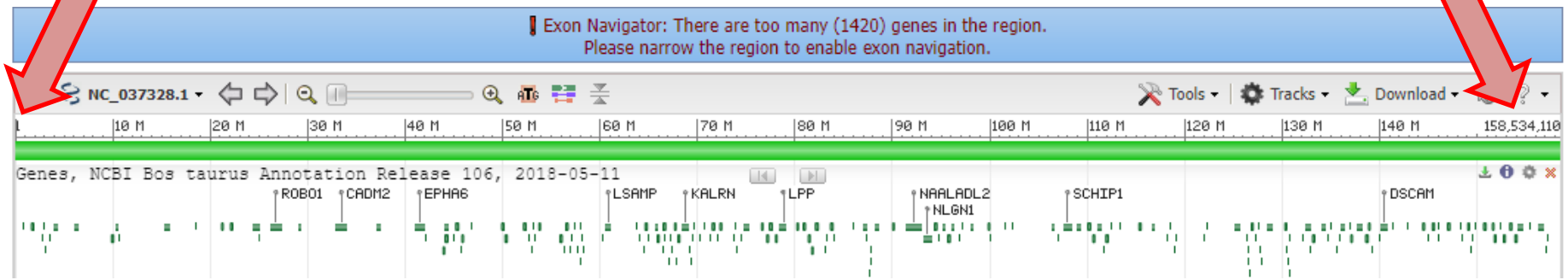




- Assim, o cromossomo será mostrado em toda a sua extensão, de 0 até seu tamanho final em Mb

0 Mb

158,53 Mb

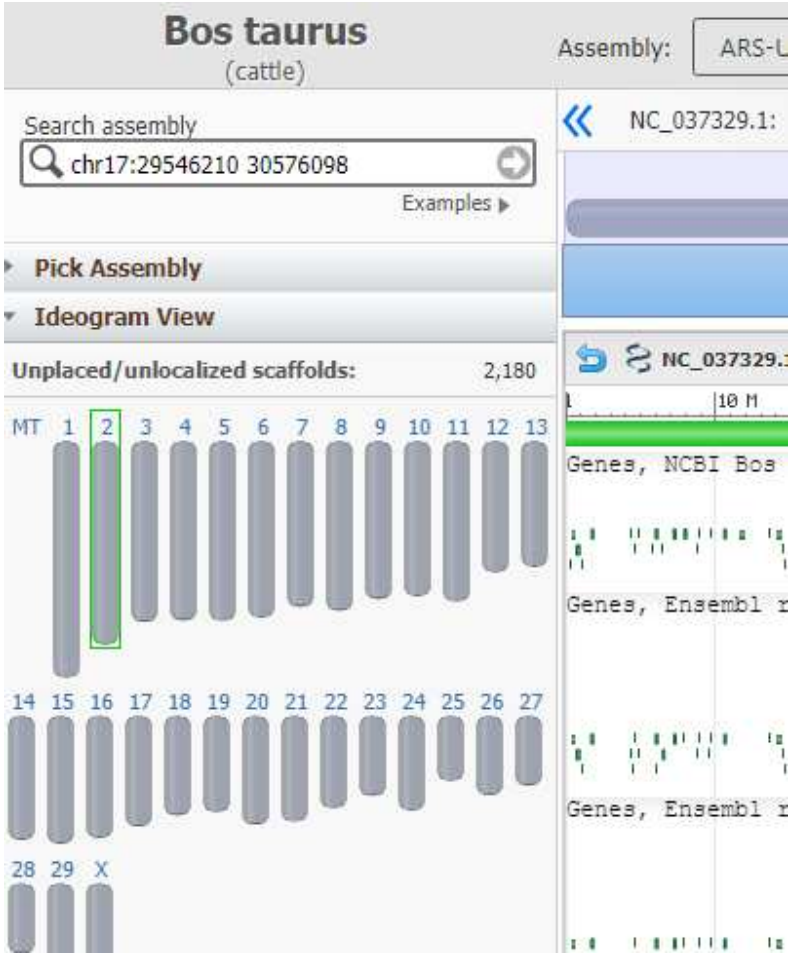


Então, o tamanho/comprimento do cromossomo 1 dos bovinos (BTA1), na montagem ARS-UCD1.2 é de 158,534,110 Megabases (Mb)

BTA: *Bos taurus autosome*

(Os cromossomos autossômicos da espécie *Bos taurus* são identificados pela sigla BTA)

- No canto esquerdo clicar nos outros cromossomos para ver e anotar os tamanhos



Cromossomo	Tamanho do Cromossomo (Mb)
BTA1	158,53
BTA2	136,23
BTA3	121,01
BTA4	120
BTA5	120,09
BTA6	117,81
BTA7	110,68
BTA8	113,32
BTA9	105,45
BTA10	103,31
BTA11	106,98
BTA12	87,22
BTA13	83,47
BTA14	82,4
BTA15	85,01
BTA16	81,01
BTA17	73,17
BTA18	65,82
BTA19	63,45
BTA20	71,97
BTA21	69,86
BTA22	60,77
BTA23	52,5
BTA24	62,32
BTA25	42,35
BTA26	51,99
BTA27	45,61
BTA28	45,94
BTA29	51,1

3_CALCULO_FROH.R

```
########### Calculo da endogamia por ROH (FROH) pelo pacote detectRUNS #
# Para cálculo de endogamia (com base no genoma, em chr ou por comprimentos), sobreposição de ROH e estatísticas da saída:

# detectRUNS is a R package for the detection of runs of homozygosity (ROH/ROHom) and of heterozygosity (ROHet, a.k.a. “heterozygosity-rich regions”) in diploid genomes.

# Mais informações em: https://cran.r-project.org/web/packages/detectRUNS/index.html
###########

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

setwd("direcionar/para/o/diretorio") # Copiar e colar aqui entre os "" o caminho da pasta no qual estão os arquivos

# Instalando o pacote! Remover a # do comando abaixo e rodar para instalar o pacote
#install.packages("detectRUNS")
library(detectRUNS) # Abrindo o pacote

# Arquivo de mapa com: chr, name, pos (cM), pos (bp)
mapFile <- "diretorio/arquivo123.map" #Só funciona com o caminho (NÃO LER O ARQUIVO COM 'READ.TABLE'!)

# Arquivo.hom contendo: group, id, chrom, nSNP, from, to, lengthBps
runsFile <- read.table("roh_out_arquivo123.hom", h=T)
head(runsFile)
dim(runsFile)
runsDF <- runsFile[, c("FID", "IID", "CHR", "NSNP", "POS1", "POS2", "KB")]
dim(runsDF)
colnames(runsDF) <- c("group", "id", "chrom", "nSNP", "from", "to", "lengthBps")
head(runsDF)
## OBS: a saída.hom do plink dá os comprimentos de ROH em KB, e o detectRUNS lê em Bp, tem que multiplicar por 1000 ou só tirar os "." como é feito no comando abaixo
runsDF$lengthBps <- as.numeric(gsub("\\.", "", runsDF$lengthBps))
head(runsDF)

# Calculando a endogamia (Froh) por cromossomo, por animal (genoma completo) e por animal e classe
Froh_chr <- Froh_inbreeding(runs = runsDF, mapFile =  mapFile, genome_wide = F) # Por cromossomo
Froh_GW <- Froh_inbreeding(runsDF, mapFile, genome_wide = T) # Por genoma completo
```

```
Froh_class <- Froh_inbreedingClass(runs = runsDF, mapFile = mapFile, Class = 2) # Por classe
(default: 0-2, 2-4, 4-8, 8-16, >16)

# Salvando os resultados
write.table(Froh_chr, file= "Froh_chr.txt", quote= F, row.names= F)
write.table(Froh_GW, file= "Froh_GW.txt", quote= F, row.names= F)
write.table(Froh_class, file= "Froh_class.txt", quote= F, row.names= F)
```

4_FROH_graficos.R

Particular

2023-10-04

```
##########################
#### Graficos de FROH ####
##########################

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

# Direcionar para o diretorio
setwd("D:\\PessoalD\\Doutorado\\Cap 2 ROH\\Tutorial\\1_ROH_FROH")

#### FROH por cromossomo ####

froh_chr<-read.table("Froh_chr.txt",h=T) # Importando FROH por cromossomo
dim(froh_chr) # Dimensao: numero de linhas (animais) e colunas
```

```
## [1] 2093   31
```

```
froh_chr[1:5,1:5] # Visualizando as primeiras 5 linhas e cinco colunas
```

```
##             id group      Chr_1       Chr_2        Chr_3
## 1 102916029440     0 0.21594607 0.124578646 4.884054e-02
## 2      AB52502     0 0.04430106 0.138408670 4.151394e-01
## 3      AB52503     0 0.02855563 0.059688670 9.169098e-05
## 4      AB52505     0 0.03985968 0.042222508 6.958319e-02
## 5      AB52508     0 0.22184243 0.008756868 1.789065e-02
```

```
mean_chr<-colMeans(froh_chr[,-c(1:2)],na.rm=T) # Calculando a m?dia de FROH por cromossomo
mean_chr # Visualizando
```

```
##      Chr_1      Chr_2      Chr_3      Chr_4      Chr_5      Chr_6
## 0.05145600 0.07040493 0.05635015 0.05295715 0.07875464 0.06962446
##      Chr_7      Chr_8      Chr_9     Chr_10     Chr_11     Chr_12
## 0.06529484 0.05833241 0.08163632 0.06654122 0.06084093 0.07023695
##     Chr_13     Chr_14     Chr_15     Chr_16     Chr_17     Chr_18
## 0.07829137 0.08437571 0.07228278 0.06183384 0.06897990 0.06401125
##     Chr_19     Chr_20     Chr_21     Chr_22     Chr_23     Chr_24
## 0.08280667 0.08343621 0.08066486 0.08904874 0.09155900 0.07103638
##     Chr_25     Chr_26     Chr_27     Chr_28     Chr_29
## 0.09805656 0.08545276 0.08569222 0.11352964 0.09209414
```

```
min(mean_chr) # Valor minimo de FROH
```

```
## [1] 0.051456
```

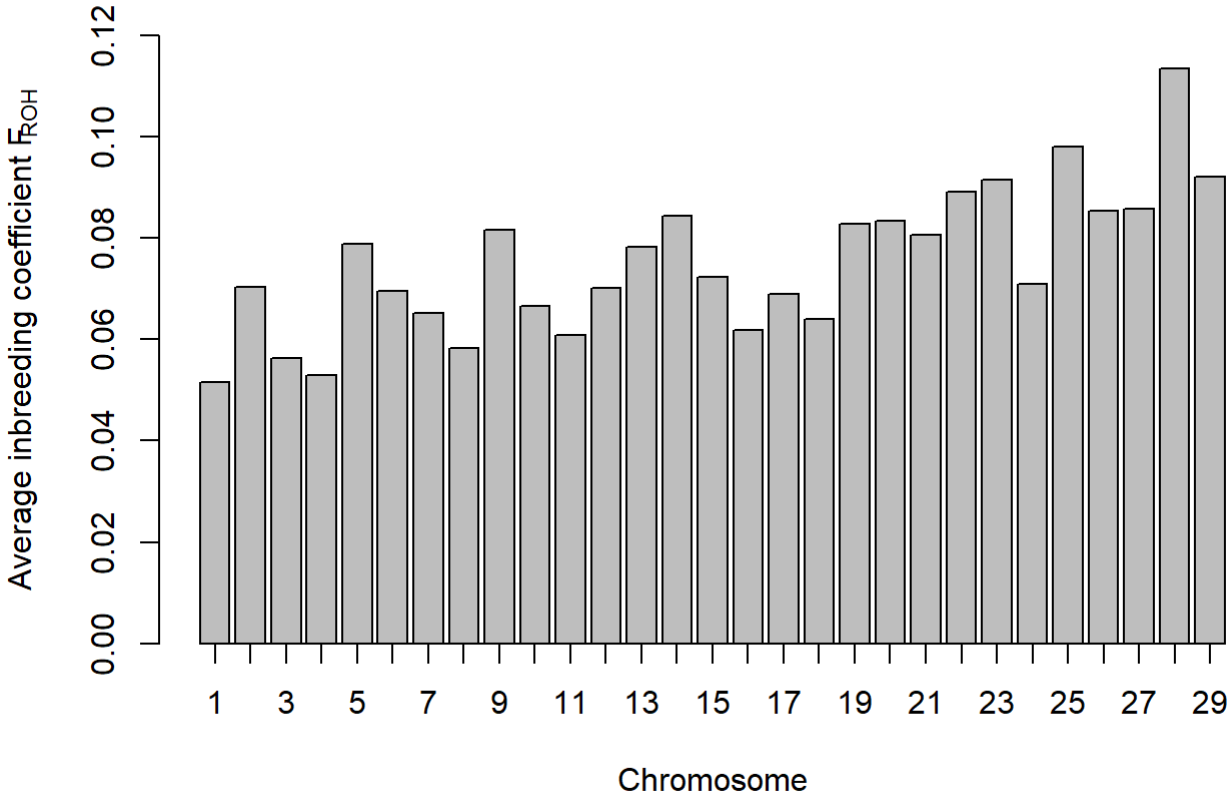
```
max(mean_chr) # Valor maximo de FROH
```

```
## [1] 0.1135296
```

```
# Montando o grafico
# Coeficiente de endogamia (FROH) por cromossomo
b<-barplot(mean_chr,xaxt="n",ylim=c(0,0.12),xlab="Chromosome",ylab=expression("Average inbreeding coeff
icient F"[ROH]))
b
```

```
##       [,1]
##  [1,]  0.7
##  [2,]  1.9
##  [3,]  3.1
##  [4,]  4.3
##  [5,]  5.5
##  [6,]  6.7
##  [7,]  7.9
##  [8,]  9.1
##  [9,] 10.3
## [10,] 11.5
## [11,] 12.7
## [12,] 13.9
## [13,] 15.1
## [14,] 16.3
## [15,] 17.5
## [16,] 18.7
## [17,] 19.9
## [18,] 21.1
## [19,] 22.3
## [20,] 23.5
## [21,] 24.7
## [22,] 25.9
## [23,] 27.1
## [24,] 28.3
## [25,] 29.5
## [26,] 30.7
## [27,] 31.9
## [28,] 33.1
## [29,] 34.3
```

```
axis(1, at=b,labels=1:29)
```




```
# Export -> Save as Image ... ou ... Arquivo -> salvar como...



#### FROH por classe de ROH ####

class_file<-read.table("Froh_class.txt",h=T) # importando o arquivo
dim(class_file) # Dimensao: numero de linhas (animais) e colunas
```

```
## [1] 2093   12
```

```
head(class_file) # Visualizando as primeiras 6 linhas
```

```
##            id group Sum_Class_0 Froh_Class_0 Sum_Class_2 Froh_Class_2
## 1 102916029440     0   173311902   0.06902866   146001659   0.05815122
## 2      AB52502     0   197876710   0.07881262   168752066   0.06721252
## 3      AB52503     0    83782838   0.03336999    54984053   0.02189968
## 4      AB52505     0   169283907   0.06742435   132805787   0.05289542
## 5      AB52508     0   217854221   0.08676949   185706670   0.07396539
## 6      AB52511     0   269212052   0.10722488   229258915   0.09131188
##   Sum_Class_4 Froh_Class_4 Sum_Class_8 Froh_Class_8 Sum_Class_16
## 1   112704275   0.04488916    62502126  0.024894067     18714589
## 2   127883341   0.05093485    86003771  0.034254573     38329988
## 3    39570388   0.01576055    10940141  0.004357365           NA
## 4   108538021   0.04322977    68867537  0.027429356     17451685
## 5   157682990   0.06280380   118285313  0.047112038     57685111
## 6   197730995   0.07875458   151389952  0.060297335     67165263
##   Froh_Class_16
## 1   0.007453862
## 2   0.015266509
## 3            NA
## 4   0.006950858
## 5   0.022975491
## 6   0.026751355
```

```
colnames(class_file) # Verificando os nomes das colunas
```

```
##  [1] "id"            "group"         "Sum_Class_0"   "Froh_Class_0"
##  [5] "Sum_Class_2"   "Froh_Class_2"  "Sum_Class_4"   "Froh_Class_4"
##  [9] "Sum_Class_8"   "Froh_Class_8"  "Sum_Class_16"  "Froh_Class_16"
```

```
#As colunas com "Sum_..." indicam a soma dos comprimentos de ROHs para cada animal e cada classe
#As colunas com "Froh_..." indicam a m?dia de Froh para cada animal em cada classe

# Criando uma arquivo so com as medias de Froh
froh_class<-cbind(class_file$Froh_Class_0,class_file$Froh_Class_2,
                  class_file$Froh_Class_4,class_file$Froh_Class_8,class_file$Froh_Class_16)
# Atribuindo os nomes as coluans do novo arquivo
colnames(froh_class)<-c("Froh_Class_0","Froh_Class_2","Froh_Class_4","Froh_Class_8","Froh_Class_16")
dim(froh_class) # Dimensao: Numero de linhas (animais) e colunas (classes)
```

```
## [1] 2093    5
```

```
head(froh_class) # Visualizando as primeiras 6 linhas do arquivo
```

```
##      Froh_Class_0 Froh_Class_2 Froh_Class_4 Froh_Class_8 Froh_Class_16
## [1,]   0.06902866   0.05815122   0.04488916  0.024894067   0.007453862
## [2,]   0.07881262   0.06721252   0.05093485  0.034254573   0.015266509
## [3,]   0.03336999   0.02189968   0.01576055  0.004357365            NA
## [4,]   0.06742435   0.05289542   0.04322977  0.027429356   0.006950858
## [5,]   0.08676949   0.07396539   0.06280380  0.047112038   0.022975491
## [6,]   0.10722488   0.09131188   0.07875458  0.060297335   0.026751355
```

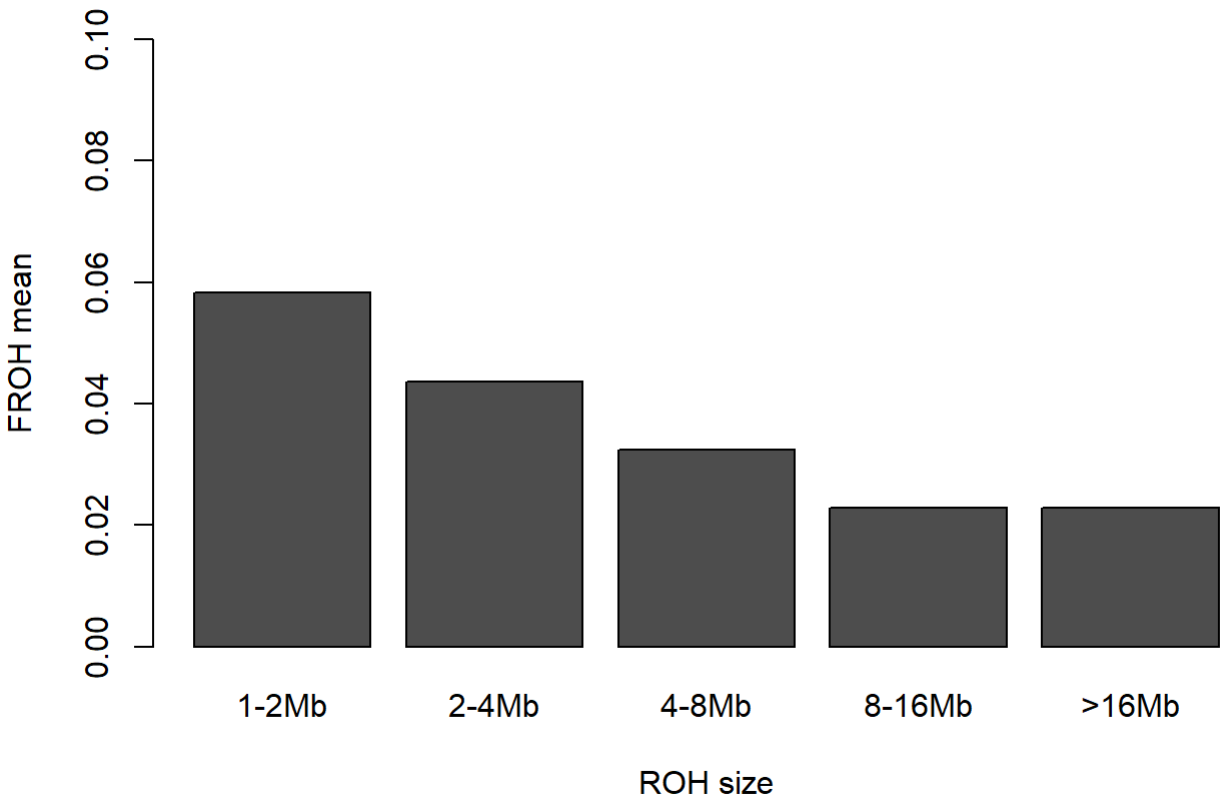
```
#Alguns animais nao tem ROH em algumas classes. Ex.: animal 3 nao tem ROH na classe 5 (acima de >16Kb)
# Por isso, usar a funcao "na.rm = ..."
media<-rbind(mean(froh_class[,1]),mean(froh_class[,2]),mean(froh_class[,3],na.rm=T),mean(froh_class[,
4],na.rm=T),mean(froh_class[,4],na.rm=T))
media<-as.matrix(media)
row.names(media)<-c("1-2Mb","2-4Mb","4-8Mb","8-16Mb",">16Mb")
media
```

```
##              [,1]
## 1-2Mb  0.05832895
## 2-4Mb  0.04366051
## 4-8Mb  0.03239830
## 8-16Mb 0.02288662
## >16Mb  0.02288662
```

```
max(media)
```

```
## [1] 0.05832895
```

```
# Montando o grafico
# Coeficiente de endogamia (FROH) por classe de ROH
barplot(t(media),ylim=c(0,0.1),ylab="FROH mean", xlab="ROH size")
```



5_Trait_FROH.R

Particular

2023-10-04

```
################################################
#### Variacao da caracteristica com a FROH ####
################################################

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

setwd("D:\\PessoalD\\Doutorado\\Cap 2 ROH\\Tutorial\\1_ROH_FROH")

## Importando o arquivo com a endogamia (FROH) por animal ##
# Esse arquivo foi obtido com o script 3_Calculo_FROH
Froh_GW<-read.table("Froh_GW.txt",h=T) # Importando
dim(Froh_GW) # Numero de linhas (animais) e colunas
```

```
## [1] 2093    4
```

```
head(Froh_GW) # Visuzalizando as primeiras 6 linhas do arquivo
```

```
##   id group       sum Froh_genome
## 1  1     0 173311902  0.06902866
## 2  2     0 197876710  0.07881262
## 3  3     0  83782838  0.03336999
## 4  4     0 169283907  0.06742434
## 5  5     0 217854221  0.08676949
## 6  6     0 269212052  0.10722488
```

```
## Importando o arquivo com os dados fenotipicos ####
dados=read.table("dados.txt",h=T) # Se o arquivo tiver cabecalho, usar "h=T", se nao, usar "h=F"
# Neste caso o arquivo tem duas colunas, identificacao do animal e fenotipo
head(dados)
```

```
##   id fenotipo
## 1  1       46
## 2  2       49
## 3  3       13
## 4  4       67
## 5  5       62
## 6  6       48
```

```
# Para o proximo passo vamos unir os dois arquivos: endogamia e fenotipo
# Vamos unir pela coluna da identificacao do animal
# Entao, o cabecalho da coluna de identificacao do animal deve ser o mesmo nos dois arquivos, por exemplo "id"
# Caso seja diferente, podemos usar o comando abaixo para mudar o nome das colunas
# ESSA PARTE NAO E NECESSARIA SE O NOME DAS COLUNAS DOS ANIMAIS JA E IGUAL ENTRE OS ARQUIVOS
# Esses comandos so funcionam se o arquivo for um data.frame
#str(dados) # Para verificar se e um data frame
#dados<-as.data.frame(dados) # Se nao for, esse comando transforma em data.frame
names(dados)[1]<-"id" # A primeira coluna desse arquivo tem os animais e sera nomeada "id"

# Para este exemplo, vamos selecionar do arquivo de Froh a mesma quantidade que temos no arquivo de dados usado como exemplo
# Na pratica, com ambos os arquivos completos com todos os animais, nao é necessario selecionar
# Isto é apenas para este exemplo
dim(dados) # tem 50 animais
```

```
## [1] 50  2
```

```
Froh_GW1<-Froh_GW[1:50,] #Vamos selecionar os 50 animais no arquivo de Froh
names(Froh_GW1)[1]<-"id" # A primeira coluna desse arquivo tem os animais e sera nomeada "id"
head(Froh_GW1)
```

```
##   id group       sum Froh_genome
## 1  1     0 173311902  0.06902866
## 2  2     0 197876710  0.07881262
## 3  3     0  83782838  0.03336999
## 4  4     0 169283907  0.06742434
## 5  5     0 217854221  0.08676949
## 6  6     0 269212052  0.10722488
```

```
# Com o comando abaixo, unir os dois arquivos
dados1<-merge(Froh_GW1,dados,by=intersect("id","id")) # unindo os dados pela coluna de id
dim(dados1) # Numero de linhas (animais) comuns entre os dois arquivos (que tem fenotipo e endogamia) e numero de colunas
```

```
## [1] 50  5
```

```
dados2<-dados1[order(dados1$Froh_genome),] # Ordenando pelo FROH
head(dados2) # Visualizando as primeiras 6 linhas do arquivo
```

```
##    id group       sum Froh_genome fenotipo
## 22 22     0  72635934  0.02893028       31
## 3   3     0  83782838  0.03336999       13
## 27 27     0  91294535  0.03636184       47
## 41 41     0  91448555  0.03642318       20
## 10 10     0  98526528  0.03924228       53
## 9   9     0  98865763  0.03937740       28
```

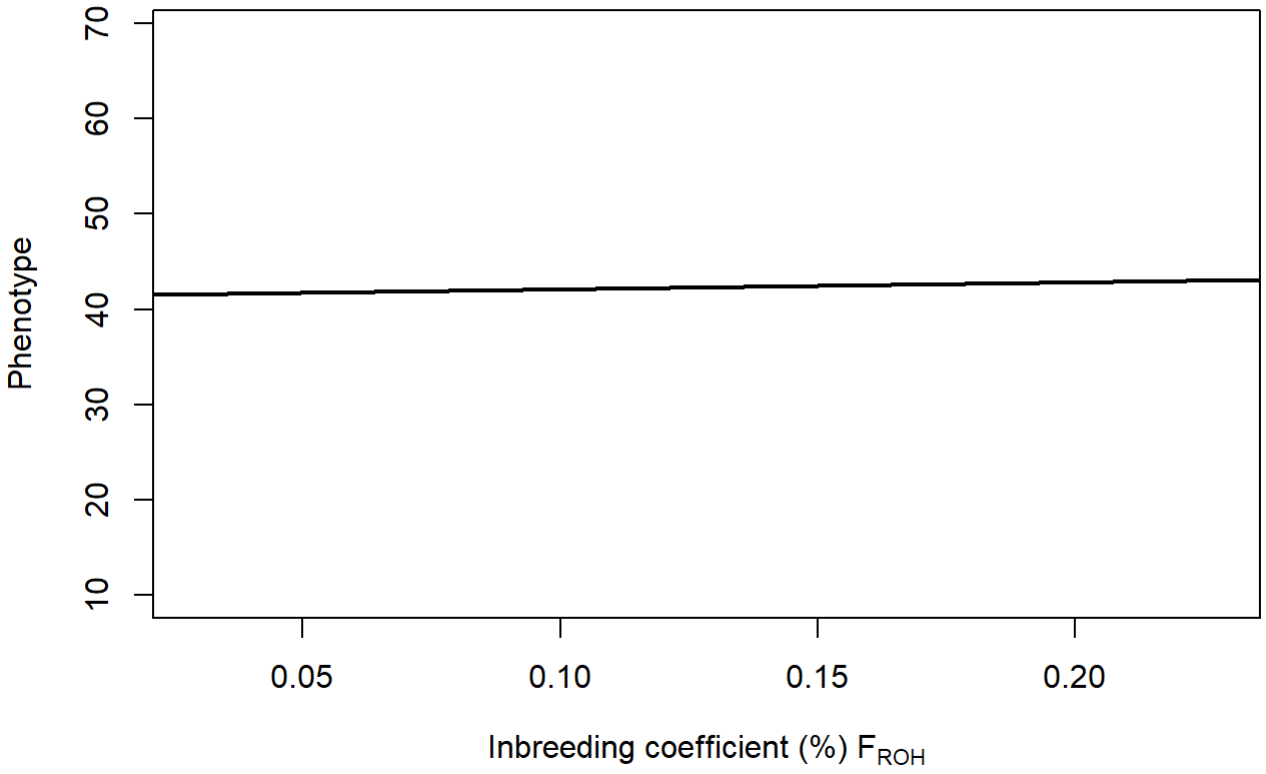
```
#### Grafico da relacao entre caracteristica e endogamia  ####

# Essa primeira parte e apenas para visualizar a linha de tendencia
colnames(dados2) # Verificando o nome das colunas
```

```
## [1] "id"          "group"       "sum"         "Froh_genome" "fenotipo"
```

```
regressao=lm(fenotipo~Froh_genome,data=dados2) # Criando um arquivo com a regressao da coluna 'fenotip
o' em relacao a coluna 'Froh_genome'
max(dados2$fenotipo) #Para saber o limite maximo do eixo y
```

```
## [1] 69
```

```
# Exemplo hipotetico de variacao de uma caracteristica com o coeficiente de endogamia (FROH)
plot(dados2$Froh_genome, dados2$fenotipo,type="n", xlab=expression("Inbreeding coefficient (%) F"[RO
H]),ylab="Phenotype")
abline(coef(regressao),lty=1,lwd=2)
```




```
# Neste exemplo, o grafico quase nao apresenta variacao
# Como se a endogamia nao afetasse o fenotipo (nem melhorando, nem piorando)
# Isso se deve ao fato de que para este exemplo foram usados poucos dados (50 fenotipos)



##########################################################
############## Resultados da regressao ###################
# Coeficiente linear, nivel de significancia e outros #
##########################################################

# Abaixo tem a descricao dos parametros que podem ser obtidos da analise
summary(regressao)$coefficients[1] #Retorna o primeiro coeff. = Intercept
```

```
## [1] 41.42539
```

```
summary(regressao)$coefficients[2] #Retorna o segundo coeff. = Linear
```

```
## [1] 7.031833
```

```
sigma(regressao) #Retorna o Residual standard error
```

```
## [1] 17.5491
```

```
summary(regressao)$fstatistic[3] #Retorna os graus deliberdade
```

```
## dendf
##    48
```

```
summary(regressao)$r.squared #Retorna o R-squared
```

```
## [1] 0.0002077465
```

```
summary(regressao)$adj.r.squared #Retorna o AdjustedR-squared
```

```
## [1] -0.02062126
```

```
summary(regressao)$fstatistic[1] #Retorna a estatistica F
```

```
##       value
## 0.009973904
```

```
summary(regressao)$coefficients[8] #Retorna o p-value do coeff. linear
```

```
## [1] 0.9208639
```

```
# Criando um arquivo com todos estes resultados
geral_results<-rbind((summary(regressao)$coefficients[1]),(summary(regressao)$coefficients[2]),
                     (sigma(regressao)),(summary(regressao)$fstatistic[3]),(summary(regressao)$r.square
d),
                     (summary(regressao)$adj.r.squared),(summary(regressao)$fstatistic[1]),(summary(reg
ressao)$coefficients[8]))
row.names(geral_results)<-c("Intercepto","Coeficiente_linear","Desvio_padrao_residual",
                            "Graus_de_liberdade","R-squared", "R-squared_ajustado",
                            "F-statistics","p-value")
geral_results
```

```
##                              dendf
## Intercepto             41.4253883864
## Coeficiente_linear      7.0318325175
## Desvio_padrao_residual 17.5491027351
## Graus_de_liberdade     48.0000000000
## R-squared               0.0002077465
## R-squared_ajustado     -0.0206212588
## F-statistics            0.0099739037
## p-value                 0.9208639387
```

```
# Salvando os resultados
write.table(geral_results,"resultados_regressao_geral.txt",row.names=T,col.names=F,quote=F,sep="\t")



#################################################################
# Variacao da caracteristica com a FROH e cada cromossomo ####
#################################################################


#### Grafico caracteristicas x FROH x cromossomo  ####

# Importando arquivo de endogamiapor cromossomo
froh_chr<-read.table("Froh_chr.txt",h=T)
froh_chr[1:5,1:5] # visualizando as primeiras linhas e colunas
```

```
##   id group      Chr_1       Chr_2      Chr_3
## 1  1     0 0.21594607 0.124578646 0.04884054
## 2  2     0 0.04430106 0.138408670 0.41513943
## 3  3     0 0.02855563 0.059688670 0.00009170
## 4  4     0 0.03985968 0.042222508 0.06958319
## 5  5     0 0.22184243 0.008756868 0.01789065
```

```
# Unindo o arquivo de endogamia por cromossomo e fenotipo
dados3<-merge(froh_chr,dados,by=intersect("id","id"))
dim(dados3) # Numero de linhas (animais) comuns entre os dois arquivos (que tem fenotipo e endogamia) e numero de colunas
```

```
## [1] 50 32
```

```
dados3[1:5,1:5] # visualizando as primeiras linhas e colunas
```

```
##   id group      Chr_1       Chr_2      Chr_3
## 1  1     0 0.21594607 0.124578646 0.04884054
## 2  2     0 0.04430106 0.138408670 0.41513943
## 3  3     0 0.02855563 0.059688670 0.00009170
## 4  4     0 0.03985968 0.042222508 0.06958319
## 5  5     0 0.22184243 0.008756868 0.01789065
```

```
colnames(dados3)
```

```
##  [1] "id"       "group"    "Chr_1"    "Chr_2"    "Chr_3"    "Chr_4"
##  [7] "Chr_5"    "Chr_6"    "Chr_7"    "Chr_8"    "Chr_9"    "Chr_10"
## [13] "Chr_11"   "Chr_12"   "Chr_13"   "Chr_14"   "Chr_15"   "Chr_16"
## [19] "Chr_17"   "Chr_18"   "Chr_19"   "Chr_20"   "Chr_21"   "Chr_22"
## [25] "Chr_23"   "Chr_24"   "Chr_25"   "Chr_26"   "Chr_27"   "Chr_28"
## [31] "Chr_29"   "fenotipo"
```

```
head(dados)
```

```
##   id fenotipo
## 1  1       46
## 2  2       49
## 3  3       13
## 4  4       67
## 5  5       62
## 6  6       48
```

```
head(froh_chr)
```

```
##   id group      Chr_1       Chr_2      Chr_3      Chr_4      Chr_5
## 1  1     0 0.21594607 0.124578646 0.04884054 0.08391452 0.15386698
## 2  2     0 0.04430106 0.138408670 0.41513943 0.09691464 0.04634005
## 3  3     0 0.02855563 0.059688670 0.00009170 0.01020341 0.07271159
## 4  4     0 0.03985968 0.042222508 0.06958319 0.14694807         NA
## 5  5     0 0.22184243 0.008756868 0.01789065 0.23282822 0.17773044
## 6  6     0 0.06301693 0.009213533 0.16547786 0.02671315 0.12968760
##        Chr_6       Chr_7      Chr_8       Chr_9     Chr_10     Chr_11
## 1 0.02014213 0.010174486 0.08387195 0.060867198         NA         NA
## 2 0.06007924 0.023873805 0.01992945 0.051140670 0.06189710 0.18233622
## 3 0.02721340          NA 0.08896271 0.028351651 0.02772251 0.01508080
## 4 0.05013045 0.077458780 0.02013321 0.009824868 0.05518680 0.04244231
## 5 0.01795546 0.009302063 0.04905976 0.050422096 0.15202194 0.06846282
## 6 0.30023230 0.010841819 0.09649080 0.071054343 0.17750809 0.01456146
##       Chr_12     Chr_13     Chr_14     Chr_15     Chr_16     Chr_17
## 1 0.04689745 0.01564366         NA 0.14449982 0.05301023 0.14434868
## 2 0.01168672 0.04001750 0.02466714 0.03068339         NA 0.10039012
## 3 0.05219868 0.01867367 0.17575187 0.02102333         NA 0.04753043
## 4 0.05391799 0.12082356 0.12009246 0.01324843 0.01740586         NA
## 5 0.02326157 0.15956870 0.03960716 0.02718649 0.10279618         NA
## 6 0.01209757 0.04902911 0.16329017 0.19613649 0.01306104 0.02177142
##       Chr_18     Chr_19      Chr_20      Chr_21      Chr_22     Chr_23
## 1 0.03267384         NA 0.029294298 0.000167925 0.001515385 0.08459218
## 2 0.01667810 0.03180139 0.182570542 0.041957728 0.093116791 0.08821302
## 3 0.04587168 0.12649567 0.009237833          NA 0.002363702         NA
## 4         NA 0.07202315 0.124747523          NA 0.027250368 0.03176840
## 5 0.04572153 0.10821356 0.029294298          NA 0.185221382 0.23624033
## 6 0.33123595 0.14377281 0.251036966          NA 0.135560026         NA
##       Chr_24    Chr_25      Chr_26     Chr_27     Chr_28     Chr_29
## 1 0.14722293 0.1955959 0.083423964 0.06188069         NA 0.02484815
## 2 0.05642749        NA          NA         NA         NA 0.21048636
## 3         NA        NA 0.002269424         NA         NA 0.03087911
## 4 0.01992549 0.0906491 0.444015687 0.03930256 0.30995865 0.26718011
## 5 0.06157190        NA 0.002226531 0.41290121 0.04753927 0.04069383
## 6 0.26356830 0.1626485 0.322633909 0.04944903 0.06786912 0.03657717
```

```
#### PLOT PDF Caracteristica x FROH x CHR ####

# Exemplo hipotetico de variacao de uma caracteristica com o coeficiente de endogamia (FROH) por cromos
somo
pdf("Exemplo_caract_FROH_CHR.pdf",height=27.8,width=50,pointsize=25)
par(mfrow=c(5,6))

#### Chr_1
regressao1<-lm(fenotipo~Chr_1,dados3)
plot(dados3$Chr_1,dados3$fenotipo,main="BTA1", ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao1),lty=1,lwd=2)

#### Chr_2
regressao2<-lm(fenotipo~Chr_2,dados3)
plot(dados3$Chr_2,dados3$fenotipo,main="BTA2",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao2),lty=1,lwd=2)

#### Chr_3
regressao3<-lm(fenotipo~Chr_3,dados3)
plot(dados3$Chr_3,dados3$fenotipo,main="BTA3",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao3),lty=1,lwd=2)

#### Chr_4
regressao4<-lm(fenotipo~Chr_4,dados3)
plot(dados3$Chr_4,dados3$fenotipo,main="BTA4",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao4),lty=1,lwd=2)

#### Chr_5
regressao5<-lm(fenotipo~Chr_5,dados3)
plot(dados3$Chr_5,dados3$fenotipo,main="BTA5",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao5),lty=1,lwd=2)

#### Chr_6
regressao6<-lm(fenotipo~Chr_6,dados3)
plot(dados3$Chr_6,dados3$fenotipo,main="BTA6",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao6),lty=1,lwd=2)

#### Chr_7
regressao7<-lm(fenotipo~Chr_7,dados3)
plot(dados3$Chr_7,dados3$fenotipo,main="BTA7",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao7),lty=1,lwd=2)

#### Chr_8
regressao8<-lm(fenotipo~Chr_8,dados3)
plot(dados3$Chr_8,dados3$fenotipo,main="BTA8",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao8),lty=1,lwd=2)

#### Chr_9
regressao9<-lm(fenotipo~Chr_9,dados3)
plot(dados3$Chr_9,dados3$fenotipo,main="BTA9",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao9),lty=1,lwd=2)

#### Chr_10
regressao10<-lm(fenotipo~Chr_10,dados3)
plot(dados3$Chr_10,dados3$fenotipo,main="BTA10",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao10),lty=1,lwd=2)

#### Chr_11
regressao11<-lm(fenotipo~Chr_11,dados3)
```

```
plot(dados3$Chr_11,dados3$fenotipo,main="BTA11",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao11),lty=1,lwd=2)

#### Chr_12
regressao12<-lm(fenotipo~Chr_12,dados3)
plot(dados3$Chr_12,dados3$fenotipo,main="BTA12",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao12),lty=1,lwd=2)

#### Chr_13
regressao13<-lm(fenotipo~Chr_13,dados3)
plot(dados3$Chr_13,dados3$fenotipo,main="BTA13",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao13),lty=1,lwd=2)

#### Chr_14
regressao14<-lm(fenotipo~Chr_14,dados3)
plot(dados3$Chr_14,dados3$fenotipo,main="BTA14",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao14),lty=1,lwd=2)

#### Chr_15
regressao15<-lm(fenotipo~Chr_15,dados3)
plot(dados3$Chr_15,dados3$fenotipo,main="BTA15",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao15),lty=1,lwd=2)

#### Chr_16
regressao16<-lm(fenotipo~Chr_16,dados3)
plot(dados3$Chr_16,dados3$fenotipo,main="BTA16",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao16),lty=1,lwd=2)

#### Chr_17
regressao17<-lm(fenotipo~Chr_17,dados3)
plot(dados3$Chr_17,dados3$fenotipo,main="BTA17",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao17),lty=1,lwd=2)

#### Chr_18
regressao18<-lm(fenotipo~Chr_18,dados3)
plot(dados3$Chr_18,dados3$fenotipo,main="BTA18",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao18),lty=1,lwd=2)

#### Chr_19
regressao19<-lm(fenotipo~Chr_19,dados3)
plot(dados3$Chr_19,dados3$fenotipo,main="BTA19",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao19),lty=1,lwd=2)

#### Chr_20
regressao20<-lm(fenotipo~Chr_20,dados3)
plot(dados3$Chr_20,dados3$fenotipo,main="BTA20",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao20),lty=1,lwd=2)

#### Chr_21
regressao21<-lm(fenotipo~Chr_21,dados3)
plot(dados3$Chr_21,dados3$fenotipo,main="BTA21",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao21),lty=1,lwd=2)

#### Chr_22
regressao22<-lm(fenotipo~Chr_22,dados3)
plot(dados3$Chr_22,dados3$fenotipo,main="BTA22",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao22),lty=1,lwd=2)

#### Chr_23
regressao23<-lm(fenotipo~Chr_23,dados3)
```

```
plot(dados3$Chr_23,dados3$fenotipo,main="BTA23",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao23),lty=1,lwd=2)

#### Chr_24
regressao24<-lm(fenotipo~Chr_24,dados3)
plot(dados3$Chr_24,dados3$fenotipo,main="BTA24",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao24),lty=1,lwd=2)

#### Chr_25
regressao25<-lm(fenotipo~Chr_25,dados3)
plot(dados3$Chr_25,dados3$fenotipo,main="BTA25",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao25),lty=1,lwd=2)

#### Chr_26
regressao26<-lm(fenotipo~Chr_26,dados3)
plot(dados3$Chr_26,dados3$fenotipo,main="BTA26",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao26),lty=1,lwd=2)

#### Chr_27
regressao27<-lm(fenotipo~Chr_27,dados3)
plot(dados3$Chr_27,dados3$fenotipo,main="BTA27",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao27),lty=1,lwd=2)

#### Chr_28
regressao28<-lm(fenotipo~Chr_28,dados3)
plot(dados3$Chr_28,dados3$fenotipo,main="BTA28",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao28),lty=1,lwd=2)

#### Chr_29
regressao29<-lm(fenotipo~Chr_29,dados3)
plot(dados3$Chr_29,dados3$fenotipo,main="BTA29",ylim=c(10,70),type="n", xlab="",ylab="",cex.main=1.8)
abline(coef(regressao29),lty=1,lwd=2)

#par(xpd=TRUE)
#legend("top",inset=c(0,-0.1),horiz=T,c("regressao"),lty=1, bty="n",lwd=2)


dev.off()
```

```
## png
##   2
```

```
#########################################################


# Se ja conferiu o script para todos os cromossomo, segue a dica:
# Clicar na setinha que fica do lado esquerdo (ao lado do titulo): PLOT PDF regressao x FROH x CHR
# Isso vai reduzir todos os comandos a apenas uma linha
# Fica mais facil para selecionar a linha e rodar tudo de uma vez
# E mais facil do que ficar selecionando o script todo ate o ultimo CHR




#########################################################
####### Resultados da regressao por cromossomo ########
# Coeficiente linear, nivel de significancia e outros #
#########################################################


#### Resultados por cromossomo ####
chr1<-rbind((summary(regressao1)$coefficients[1]),(summary(regressao1)$coefficients[2]),
            (sigma(regressao1)),(summary(regressao1)$fstatistic[3]),(summary(regressao1)$r.squared),
            (summary(regressao1)$adj.r.squared),(summary(regressao1)$fstatistic[1]),(summary(regressao
1)$coefficients[8]))

chr2<-rbind((summary(regressao2)$coefficients[1]),(summary(regressao2)$coefficients[2]),
            (sigma(regressao2)),(summary(regressao2)$fstatistic[3]),(summary(regressao2)$r.squared),
            (summary(regressao2)$adj.r.squared),(summary(regressao2)$fstatistic[1]),(summary(regressao
2)$coefficients[8]))

chr3<-rbind((summary(regressao3)$coefficients[1]),(summary(regressao3)$coefficients[2]),
            (sigma(regressao3)),(summary(regressao3)$fstatistic[3]),(summary(regressao3)$r.squared),
            (summary(regressao3)$adj.r.squared),(summary(regressao3)$fstatistic[1]),(summary(regressao
3)$coefficients[8]))

chr4<-rbind((summary(regressao4)$coefficients[1]),(summary(regressao4)$coefficients[2]),
            (sigma(regressao4)),(summary(regressao4)$fstatistic[3]),(summary(regressao4)$r.squared),
            (summary(regressao4)$adj.r.squared),(summary(regressao4)$fstatistic[1]),(summary(regressao
4)$coefficients[8]))

chr5<-rbind((summary(regressao5)$coefficients[1]),(summary(regressao5)$coefficients[2]),
            (sigma(regressao5)),(summary(regressao5)$fstatistic[3]),(summary(regressao5)$r.squared),
            (summary(regressao5)$adj.r.squared),(summary(regressao5)$fstatistic[1]),(summary(regressao
5)$coefficients[8]))

chr6<-rbind((summary(regressao6)$coefficients[1]),(summary(regressao6)$coefficients[2]),
            (sigma(regressao6)),(summary(regressao6)$fstatistic[3]),(summary(regressao6)$r.squared),
            (summary(regressao6)$adj.r.squared),(summary(regressao6)$fstatistic[1]),(summary(regressao
6)$coefficients[8]))

chr7<-rbind((summary(regressao7)$coefficients[1]),(summary(regressao7)$coefficients[2]),
            (sigma(regressao7)),(summary(regressao7)$fstatistic[3]),(summary(regressao7)$r.squared),
            (summary(regressao7)$adj.r.squared),(summary(regressao7)$fstatistic[1]),(summary(regressao
7)$coefficients[8]))

chr8<-rbind((summary(regressao8)$coefficients[1]),(summary(regressao8)$coefficients[2]),
            (sigma(regressao8)),(summary(regressao8)$fstatistic[3]),(summary(regressao8)$r.squared),
            (summary(regressao8)$adj.r.squared),(summary(regressao8)$fstatistic[1]),(summary(regressao
8)$coefficients[8]))
```

```
chr9<-rbind((summary(regressao9)$coefficients[1]),(summary(regressao9)$coefficients[2]),
            (sigma(regressao9)),(summary(regressao9)$fstatistic[3]),(summary(regressao9)$r.squared),
            (summary(regressao9)$adj.r.squared),(summary(regressao9)$fstatistic[1]),(summary(regressao
9)$coefficients[8]))

chr10<-rbind((summary(regressao10)$coefficients[1]),(summary(regressao10)$coefficients[2]),
             (sigma(regressao10)),(summary(regressao10)$fstatistic[3]),(summary(regressao10)$r.square
d),
             (summary(regressao10)$adj.r.squared),(summary(regressao10)$fstatistic[1]),(summary(regress
ao10)$coefficients[8]))

chr11<-rbind((summary(regressao11)$coefficients[1]),(summary(regressao11)$coefficients[2]),
             (sigma(regressao11)),(summary(regressao11)$fstatistic[3]),(summary(regressao11)$r.square
d),
             (summary(regressao11)$adj.r.squared),(summary(regressao11)$fstatistic[1]),(summary(regress
ao11)$coefficients[8]))

chr12<-rbind((summary(regressao12)$coefficients[1]),(summary(regressao12)$coefficients[2]),
             (sigma(regressao12)),(summary(regressao12)$fstatistic[3]),(summary(regressao12)$r.square
d),
             (summary(regressao12)$adj.r.squared),(summary(regressao12)$fstatistic[1]),(summary(regress
ao12)$coefficients[8]))

chr13<-rbind((summary(regressao13)$coefficients[1]),(summary(regressao13)$coefficients[2]),
             (sigma(regressao13)),(summary(regressao13)$fstatistic[3]),(summary(regressao13)$r.square
d),
             (summary(regressao13)$adj.r.squared),(summary(regressao13)$fstatistic[1]),(summary(regress
ao13)$coefficients[8]))

chr14<-rbind((summary(regressao14)$coefficients[1]),(summary(regressao14)$coefficients[2]),
             (sigma(regressao14)),(summary(regressao14)$fstatistic[3]),(summary(regressao14)$r.square
d),
             (summary(regressao14)$adj.r.squared),(summary(regressao14)$fstatistic[1]),(summary(regress
ao14)$coefficients[8]))

chr15<-rbind((summary(regressao15)$coefficients[1]),(summary(regressao15)$coefficients[2]),
             (sigma(regressao15)),(summary(regressao15)$fstatistic[3]),(summary(regressao15)$r.square
d),
             (summary(regressao15)$adj.r.squared),(summary(regressao15)$fstatistic[1]),(summary(regress
ao15)$coefficients[8]))

chr16<-rbind((summary(regressao16)$coefficients[1]),(summary(regressao16)$coefficients[2]),
             (sigma(regressao16)),(summary(regressao16)$fstatistic[3]),(summary(regressao16)$r.square
d),
             (summary(regressao16)$adj.r.squared),(summary(regressao16)$fstatistic[1]),(summary(regress
ao16)$coefficients[8]))

chr17<-rbind((summary(regressao17)$coefficients[1]),(summary(regressao17)$coefficients[2]),
             (sigma(regressao17)),(summary(regressao17)$fstatistic[3]),(summary(regressao17)$r.square
d),
             (summary(regressao17)$adj.r.squared),(summary(regressao17)$fstatistic[1]),(summary(regress
ao17)$coefficients[8]))

chr18<-rbind((summary(regressao18)$coefficients[1]),(summary(regressao18)$coefficients[2]),
             (sigma(regressao18)),(summary(regressao18)$fstatistic[3]),(summary(regressao18)$r.square
d),
             (summary(regressao18)$adj.r.squared),(summary(regressao18)$fstatistic[1]),(summary(regress
ao18)$coefficients[8]))

chr19<-rbind((summary(regressao19)$coefficients[1]),(summary(regressao19)$coefficients[2]),
```

```
            (sigma(regressao19)),(summary(regressao19)$fstatistic[3]),(summary(regressao19)$r.square
d),
            (summary(regressao19)$adj.r.squared),(summary(regressao19)$fstatistic[1]),(summary(regress
ao19)$coefficients[8]))

chr20<-rbind((summary(regressao20)$coefficients[1]),(summary(regressao20)$coefficients[2]),
            (sigma(regressao20)),(summary(regressao20)$fstatistic[3]),(summary(regressao20)$r.square
d),
            (summary(regressao20)$adj.r.squared),(summary(regressao20)$fstatistic[1]),(summary(regress
ao20)$coefficients[8]))

chr21<-rbind((summary(regressao21)$coefficients[1]),(summary(regressao21)$coefficients[2]),
            (sigma(regressao21)),(summary(regressao21)$fstatistic[3]),(summary(regressao21)$r.square
d),
            (summary(regressao21)$adj.r.squared),(summary(regressao21)$fstatistic[1]),(summary(regress
ao21)$coefficients[8]))

chr22<-rbind((summary(regressao22)$coefficients[1]),(summary(regressao22)$coefficients[2]),
            (sigma(regressao22)),(summary(regressao22)$fstatistic[3]),(summary(regressao22)$r.square
d),
            (summary(regressao22)$adj.r.squared),(summary(regressao22)$fstatistic[1]),(summary(regress
ao22)$coefficients[8]))

chr23<-rbind((summary(regressao23)$coefficients[1]),(summary(regressao23)$coefficients[2]),
            (sigma(regressao23)),(summary(regressao23)$fstatistic[3]),(summary(regressao23)$r.square
d),
            (summary(regressao23)$adj.r.squared),(summary(regressao23)$fstatistic[1]),(summary(regress
ao23)$coefficients[8]))

chr24<-rbind((summary(regressao24)$coefficients[1]),(summary(regressao24)$coefficients[2]),
            (sigma(regressao24)),(summary(regressao24)$fstatistic[3]),(summary(regressao24)$r.square
d),
            (summary(regressao24)$adj.r.squared),(summary(regressao24)$fstatistic[1]),(summary(regress
ao24)$coefficients[8]))

chr25<-rbind((summary(regressao25)$coefficients[1]),(summary(regressao25)$coefficients[2]),
            (sigma(regressao25)),(summary(regressao25)$fstatistic[3]),(summary(regressao25)$r.square
d),
            (summary(regressao25)$adj.r.squared),(summary(regressao25)$fstatistic[1]),(summary(regress
ao25)$coefficients[8]))

chr26<-rbind((summary(regressao26)$coefficients[1]),(summary(regressao26)$coefficients[2]),
            (sigma(regressao26)),(summary(regressao26)$fstatistic[3]),(summary(regressao26)$r.square
d),
            (summary(regressao26)$adj.r.squared),(summary(regressao26)$fstatistic[1]),(summary(regress
ao26)$coefficients[8]))

chr27<-rbind((summary(regressao27)$coefficients[1]),(summary(regressao27)$coefficients[2]),
            (sigma(regressao27)),(summary(regressao27)$fstatistic[3]),(summary(regressao27)$r.square
d),
            (summary(regressao27)$adj.r.squared),(summary(regressao27)$fstatistic[1]),(summary(regress
ao27)$coefficients[8]))

chr28<-rbind((summary(regressao28)$coefficients[1]),(summary(regressao28)$coefficients[2]),
            (sigma(regressao28)),(summary(regressao28)$fstatistic[3]),(summary(regressao28)$r.square
d),
            (summary(regressao28)$adj.r.squared),(summary(regressao28)$fstatistic[1]),(summary(regress
ao28)$coefficients[8]))

chr29<-rbind((summary(regressao29)$coefficients[1]),(summary(regressao29)$coefficients[2]),
```

```
            (sigma(regressao29)),(summary(regressao29)$fstatistic[3]),(summary(regressao29)$r.square
d),
            (summary(regressao29)$adj.r.squared),(summary(regressao29)$fstatistic[1]),(summary(regress
ao29)$coefficients[8]))

#### Unindo os resultados ####
results<-rbind(chr1,chr2,chr3,chr4,chr5,chr6,chr7,chr8,chr9,chr10,
               chr11,chr12,chr13,chr14,chr15,chr16,chr17,chr18,chr19,
               chr20,chr21,chr22,chr23,chr24,chr25,chr26,chr27,chr28,chr29)
rownames(results)<-rep(c("Intercepto","Coeficiente_linear","Desvio_padrao_residual",
                         "Graus_de_liberdade","R-squared", "R-squared_ajustado",
                         "F-statistics","p-value"),29)
dim(results)
```

```
## [1] 232   1
```

```
cromossomo<-c(rep(1,8),rep(2,8),rep(3,8),rep(4,8),rep(5,8),rep(6,8),rep(7,8),rep(8,8),
              rep(9,8),rep(10,8),rep(11,8),rep(12,8),rep(13,8),rep(14,8),rep(15,8),
              rep(16,8),rep(17,8),rep(18,8),rep(19,8),rep(20,8),rep(21,8),rep(22,8),
              rep(23,8),rep(24,8),rep(25,8),rep(26,8),rep(27,8),rep(28,8),rep(29,8))
cromossomo
```

```
##   [1]  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3
##  [24]  3  4  4  4  4  4  4  4  4  5  5  5  5  5  5  5  5  6  6  6  6  6  6
##  [47]  6  6  7  7  7  7  7  7  7  7  8  8  8  8  8  8  8  8  9  9  9  9  9
##  [70]  9  9  9 10 10 10 10 10 10 10 10 11 11 11 11 11 11 11 11 12 12 12 12
##  [93] 12 12 12 12 13 13 13 13 13 13 13 13 14 14 14 14 14 14 14 14 15 15 15
## [116] 15 15 15 15 15 16 16 16 16 16 16 16 16 17 17 17 17 17 17 17 17 18 18
## [139] 18 18 18 18 18 18 19 19 19 19 19 19 19 19 20 20 20 20 20 20 20 20 21
## [162] 21 21 21 21 21 21 21 22 22 22 22 22 22 22 22 23 23 23 23 23 23 23 23
## [185] 24 24 24 24 24 24 24 24 25 25 25 25 25 25 25 25 26 26 26 26 26 26 26
## [208] 26 27 27 27 27 27 27 27 27 28 28 28 28 28 28 28 28 29 29 29 29 29 29
## [231] 29 29
```

```
length(cromossomo)
```

```
## [1] 232
```

```
results<-cbind(results,cromossomo)
head(results)
```

```
##                              dendf cromossomo
## Intercepto             38.70465258          1
## Coeficiente_linear     54.80739557          1
## Desvio_padrao_residual 16.86296721          1
## Graus_de_liberdade     47.00000000          1
## R-squared               0.05561729          1
## R-squared_ajustado      0.03552404          1
```

```
# Salvando
write.table(results,"resultados_regressao_cromossomo.txt",row.names=T,col.names=F,quote=F,sep="\t")
```

6_ManhattanPlot_FST.R

Particular

2023-10-04

```
#############################################################################
##
#### Montando o grafico de Manhattan Plot baseado no criterio de FST ####
##
#############################################################################

rm(list = ls())

# R Code to create Manhattan plots - FST
setwd("D:\\PessoalD\\Doutorado\\Cap 2 ROH\\Tutorial\\1_ROH_FROH")
# Mudar o nome do arquivo abaixo se for necessario
zzz<-read.table("roh_out_arquivo123.fst",skip=1,h=F)
head(zzz);tail(zzz)
```

```
##   V1                 V2     V3  V4           V5
## 1  1 BovineHD0100046367  89725 645 -6.51297e-05
## 2  1 BovineHD0100000035 120183 645  2.31457e-03
## 3  1 BovineHD0100000039 146011 645 -1.09317e-03
## 4  1 BovineHD0100000040 147231 645 -8.27694e-04
## 5  1 BovineHD0100000042 149772 645 -8.09954e-04
## 6  1 BovineHD0100000043 151060 645  1.74824e-02
```

```
##        V1                  V2       V3  V4          V5
## 380937 29 BovineHD2900014962 51452986 645  0.02607780
## 380938 29 BovineHD2900014964 51468335 645  0.00437249
## 380939 29 BovineHD2900014966 51473944 645 -0.00131219
## 380940 29 BovineHD2900014968 51476452 645 -0.00154037
## 380941 29 BovineHD2900015102 51499351 645 -0.00122083
## 380942 29 ARS-BFGL-NGS-29340 51502868 645  0.02055440
```

```
dim(zzz)
```

```
## [1] 380942      5
```

```
zzz$V2<-seq(1,nrow(zzz)) #Atribuindo numeros aos SNPS

#As estimativas de FST tem alguns valores negativos que devem ser considerados como 0 (zero)
zzz$V5[zzz$V5<0]<-0
head(zzz);tail(zzz)
```

```
##   V1 V2     V3  V4         V5
## 1  1  1  89725 645 0.00000000
## 2  1  2 120183 645 0.00231457
## 3  1  3 146011 645 0.00000000
## 4  1  4 147231 645 0.00000000
## 5  1  5 149772 645 0.00000000
## 6  1  6 151060 645 0.01748240
```

```
##        V1     V2       V3  V4         V5
## 380937 29 380937 51452986 645 0.02607780
## 380938 29 380938 51468335 645 0.00437249
## 380939 29 380939 51473944 645 0.00000000
## 380940 29 380940 51476452 645 0.00000000
## 380941 29 380941 51499351 645 0.00000000
## 380942 29 380942 51502868 645 0.02055440
```

```
max(zzz$V5)
```

```
## [1] 0.167133
```

```
n=nrow(zzz)
x=zzz[,2] # Coluna de SNPs
y<-zzz[,5] # Estimativas de FST
chr1=zzz[,1] # Cromossomo
chr<-NULL
pos=NULL
for(i in unique(zzz$V1)){
  zz=zzz[zzz$V1==i,]
  key=as.numeric(row.names(zz))
  medio=round(nrow(zz)/2,0)
  z=key[medio]
  pos=c(pos,z)
}
chrn=unique(zzz$V1)
one   = which(chr1%%4==0)
two   = which(chr1%%4==1)
three = which(chr1%%4==2)
four  = which(chr1%%4==3)
chr[one]="darkgoldenrod"
chr[two]="darkorchid"
chr[three]="blue"
chr[four]="forestgreen"
pdf(file="ManhattanPlot_FST.pdf",family="sans",height=27.8,width=50,pointsize=50,bg="white")
plot(x,y,main="Manhattan plot for the 'Phenotype'",xlab="Chromosome",ylab=expression("F"[ST]),xlim=c(1,
n),ylim=c(0,max(y)),xaxt="n",pch=20,col=chr) #yaxt="n",
abline(h=0.15,col="red") # Linha de threshold
axis(1,at=pos,labels=chrn) #,cex.axis=1.5
dev.off()
```

```
## png
##   2
```

```
#### Verificando SNPs acima de 0,15 ou 0,25 (valor de FST)
max(zzz$V5)
```

```
## [1] 0.167133
```

```
novo=zzz[zzz$V5>0.15,]
dim(novo)
```

```
## [1] 11  5
```

```
novo
```

```
##        V1     V2       V3  V4       V5
## 194566 11 194566 49352635 645 0.160739
## 194567 11 194567 49370261 645 0.157701
## 239133 14 239133 68984866 645 0.151459
## 239143 14 239143 69035621 645 0.159822
## 239145 14 239145 69042539 645 0.152323
## 263590 16 263590 60560912 645 0.150992
## 297514 19 297514 61970518 645 0.165203
## 313218 21 313218 32479895 645 0.160308
## 374570 29 374570  4131935 645 0.167133
## 374580 29 374580  4189542 645 0.155218
## 374581 29 374581  4205711 645 0.155218
```

```
assinaturas<-as.data.frame(table(novo$V1))
write.table(assinaturas, "Assinaturas_de_selecao.txt",quote=F,row.names=F,col.names=T)
```

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).

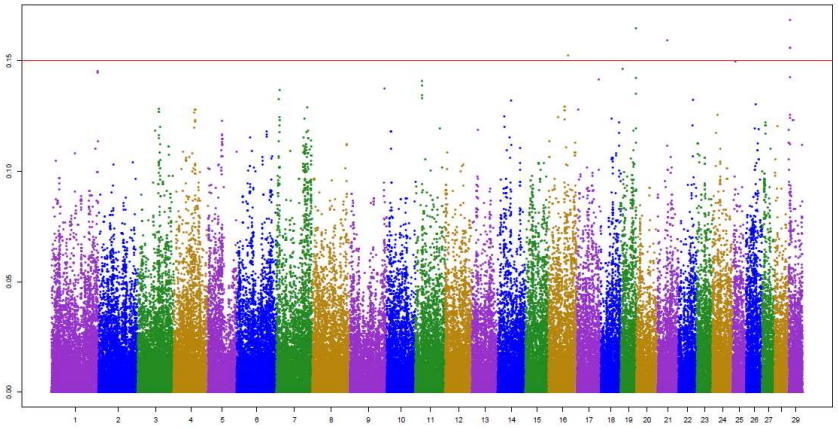
VIÇOSA
2023

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.

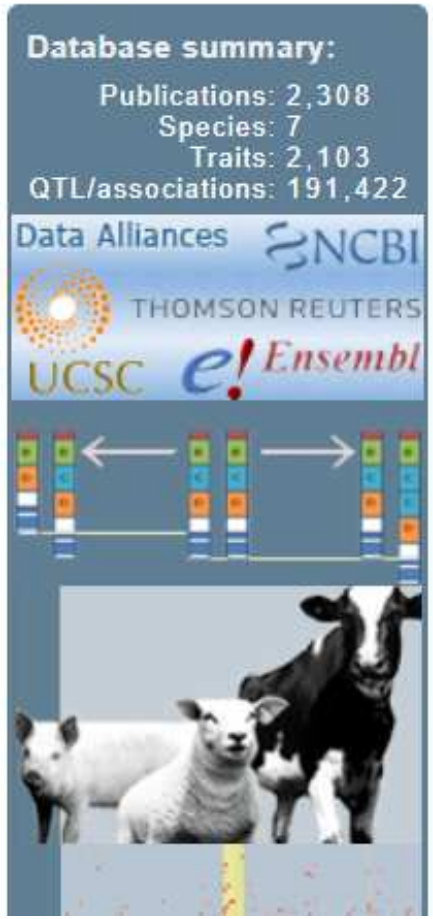
Many scientific journals require or recommend that any original QTL/association data be deposited into a public database before a paper may be accepted for publication. We provide user/curator accounts for direct data submission, and in return we supply users with a data summary link to facilitate the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

UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS AGRÁRIAS
DEPARTAMENTO DE ZOOTECNIA

TUTORIAL
VarElect: The Next Generation Sequencing Phenotyper

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

VarElect: The Next Generation Sequencing Phenotyper

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: 978-65-5668-140-5
DOI: http://dx.doi.org/10.26626/9786556681405.2023B0001

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).

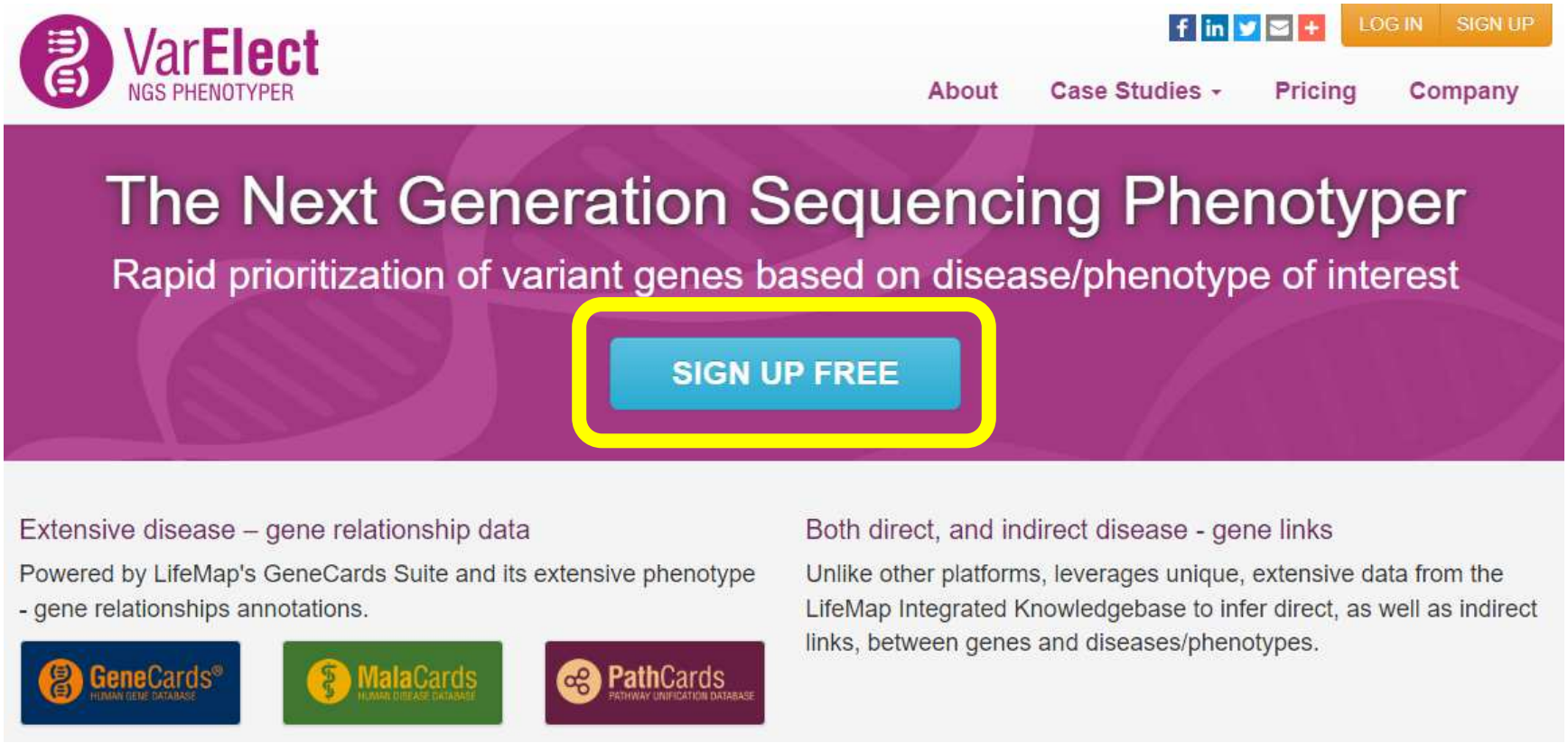
VarElect

Com uma quantidade muito grande de genes identificados associados com um(a) fenótipo/doença, o VarElect ajuda a reduzir essa lista buscando genes candidatos mais prováveis de estarem associados com a característica estudada.

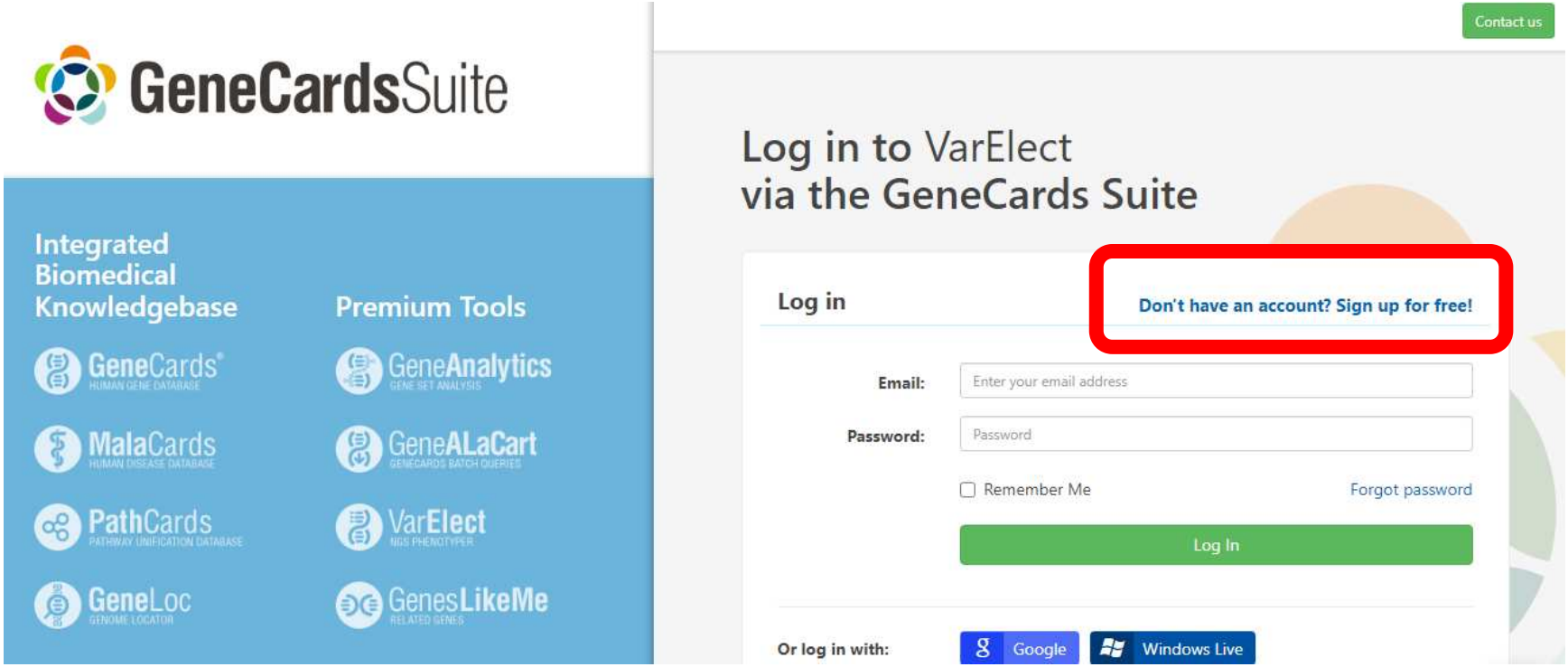
Mais informações no site: https://varelect.genecards.org/

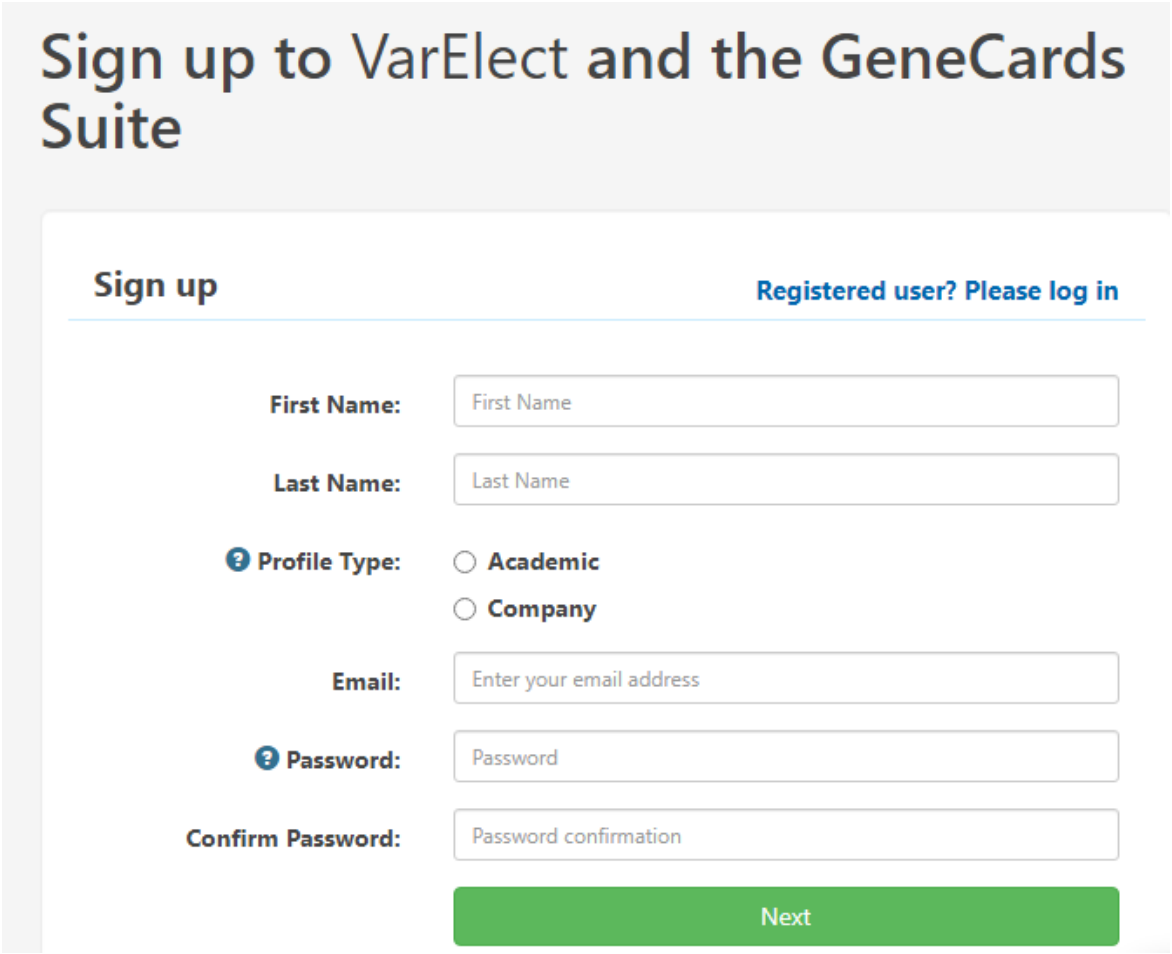
Vídeo com instruções (em inglês): https://youtu.be/rH1L4dGIS4g

VarElect

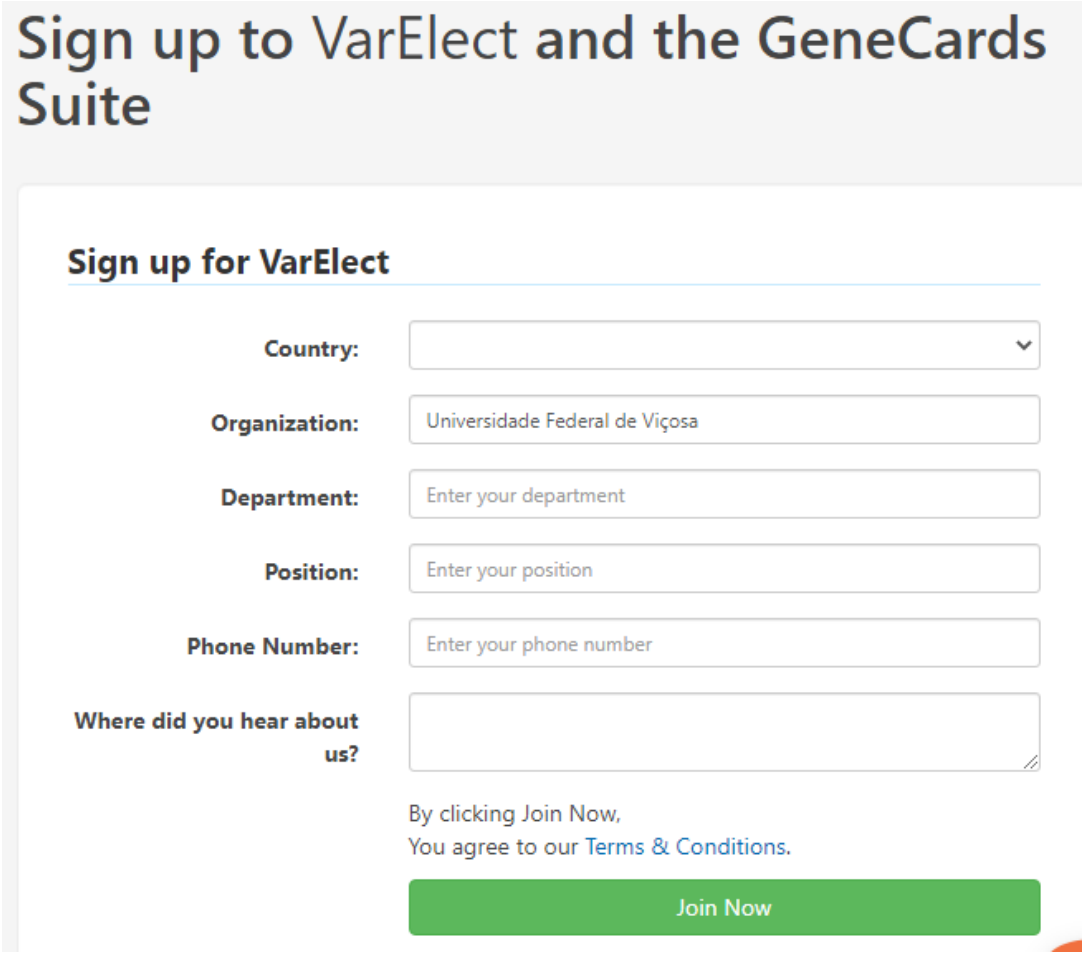

VarElect
NGS PHENOTYPER
LOG IN
SIGN UP
About
Case Studies
Pricing
Company
The Next Generation Sequencing Phenotyper
Rapid prioritization of variant genes based on disease/phenotype of interest
SIGN UP FREE
Extensive disease – gene relationship data
Powered by LifeMap's GeneCards Suite and its extensive phenotype - gene relationships annotations.
GeneCards
MalaCards
PathCards
Both direct, and indirect disease - gene links
Unlike other platforms, leverages unique, extensive data from the LifeMap Integrated Knowledgebase to infer direct, as well as indirect links, between genes and diseases/phenotypes.

Fazer login ou criar uma conta



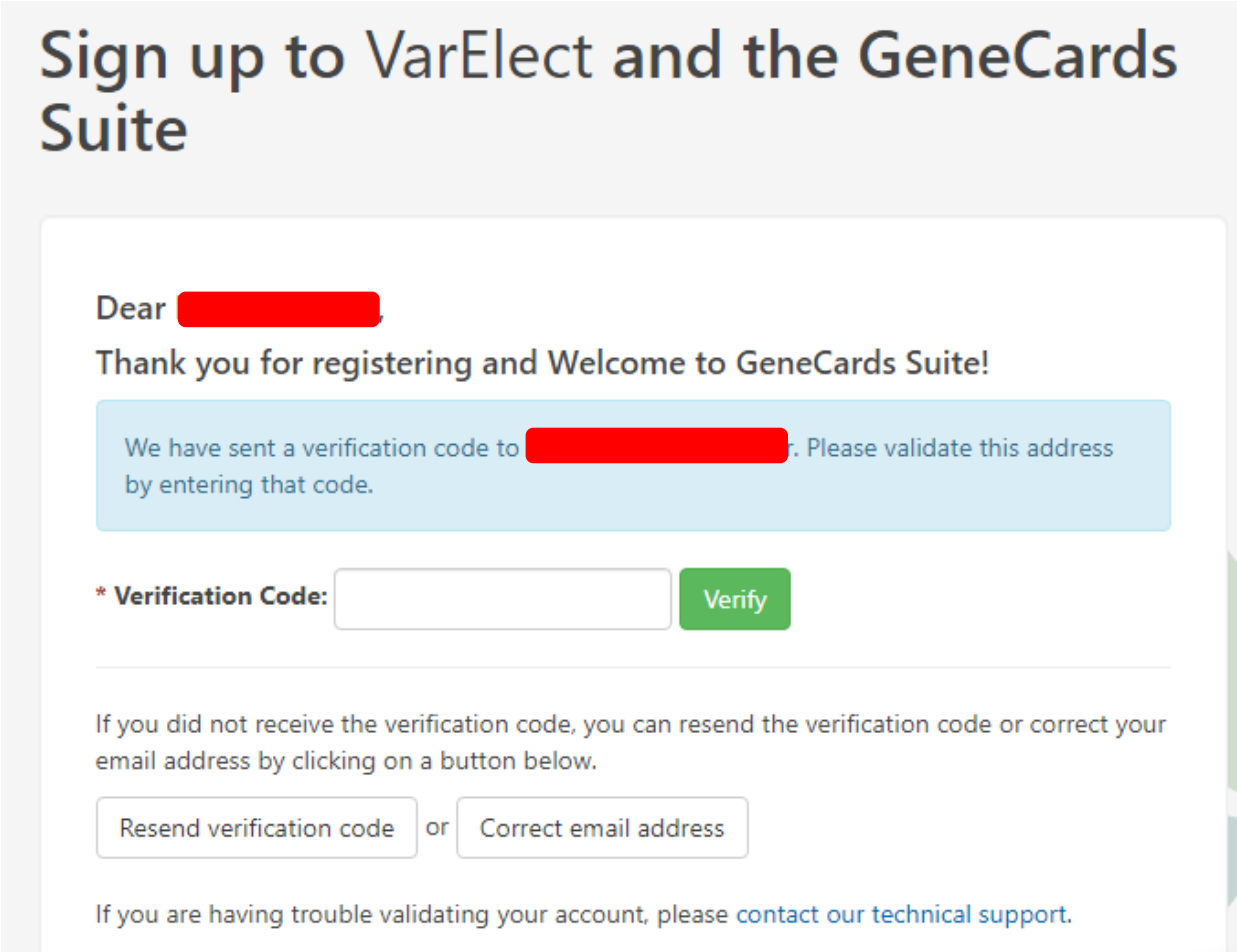


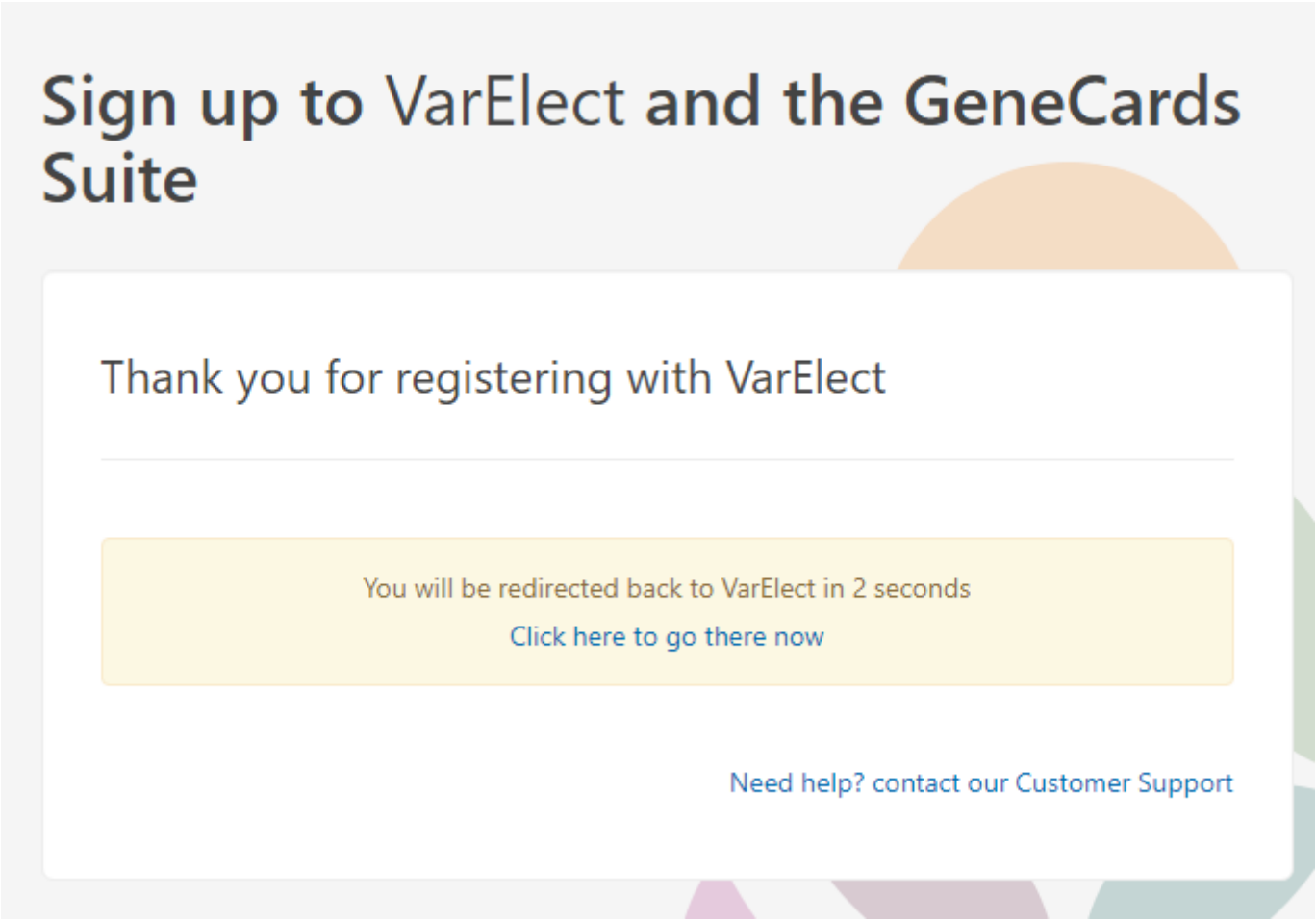
Para professores e alunos, cadastrar usando o e-mail institucional



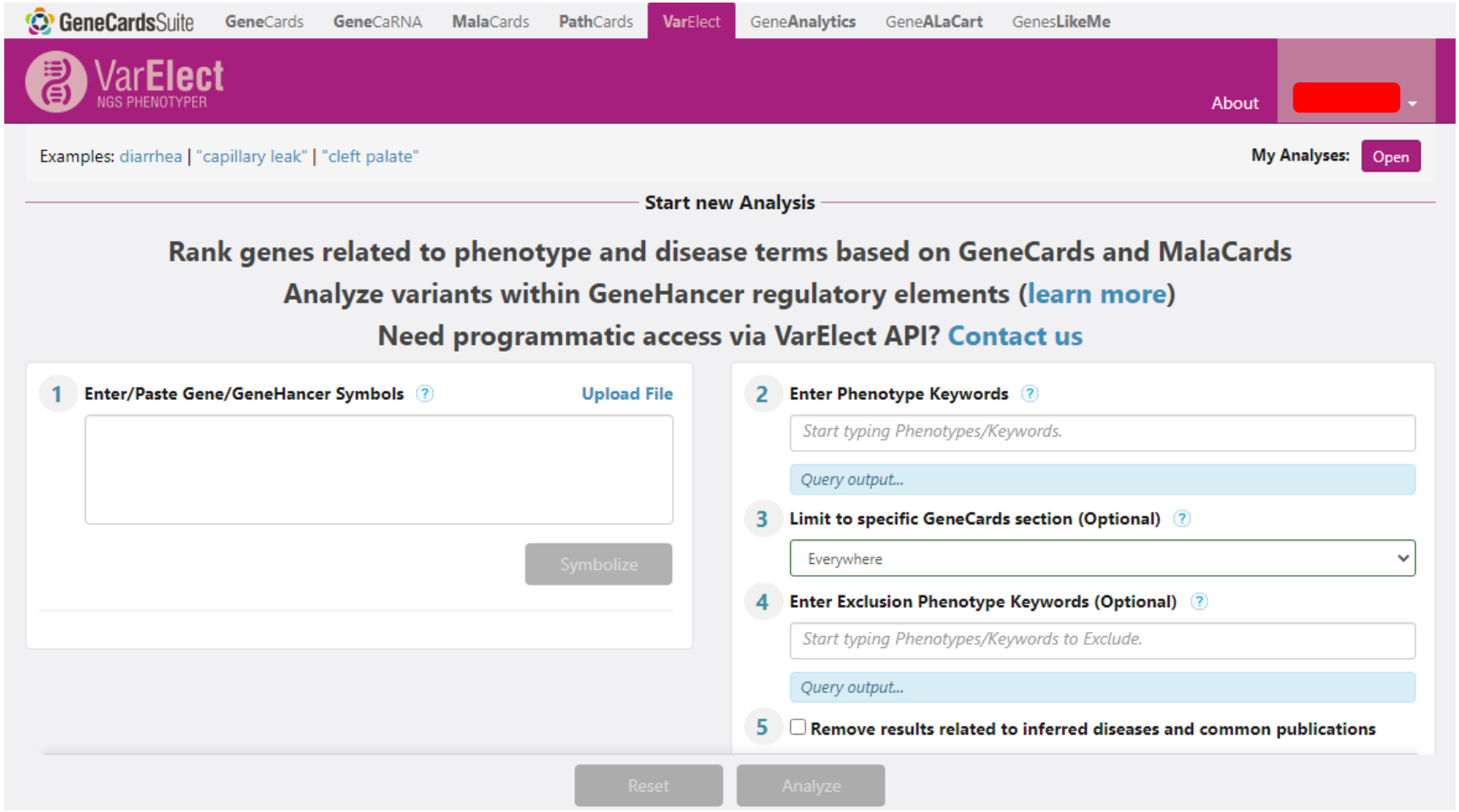
Telefone não é de preenchimento obrigatório

Um código será enviado para seu e-mail

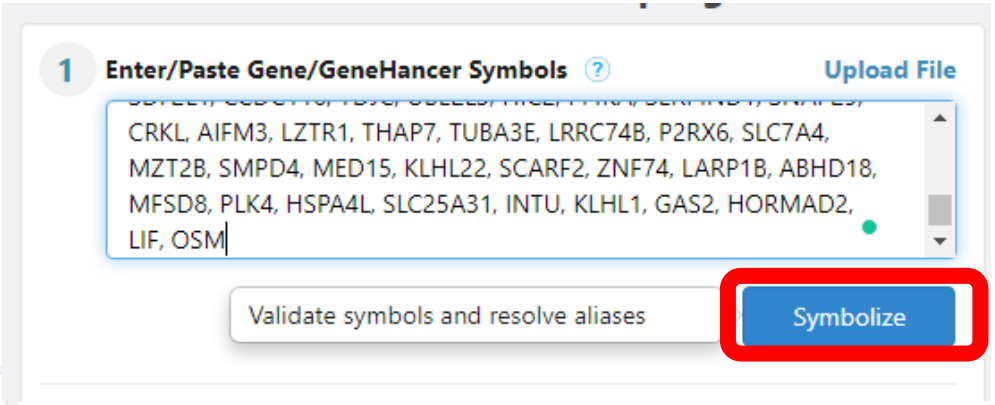


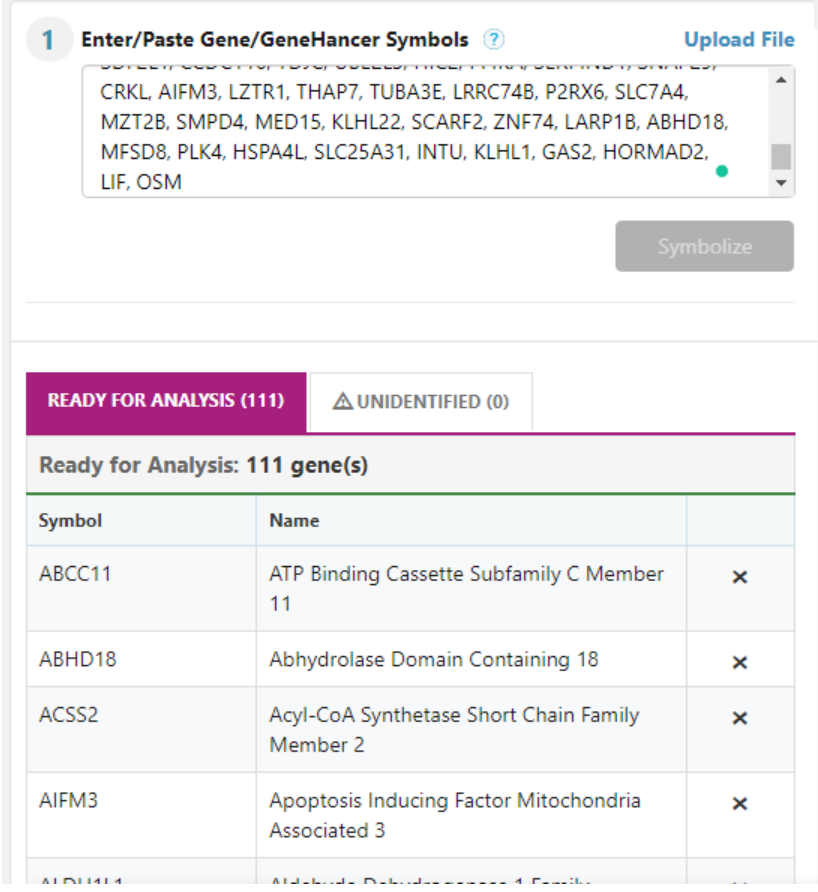

Sign up to VarElect and the GeneCards Suite
Thank you for registering with VarElect
You will be redirected back to VarElect in 2 seconds
Click here to go there now
Need help? contact our Customer Support

Página inicial do VarElect



- Inserir a lista de genes no campo (1)
- Clicar em "*Symbolize*"





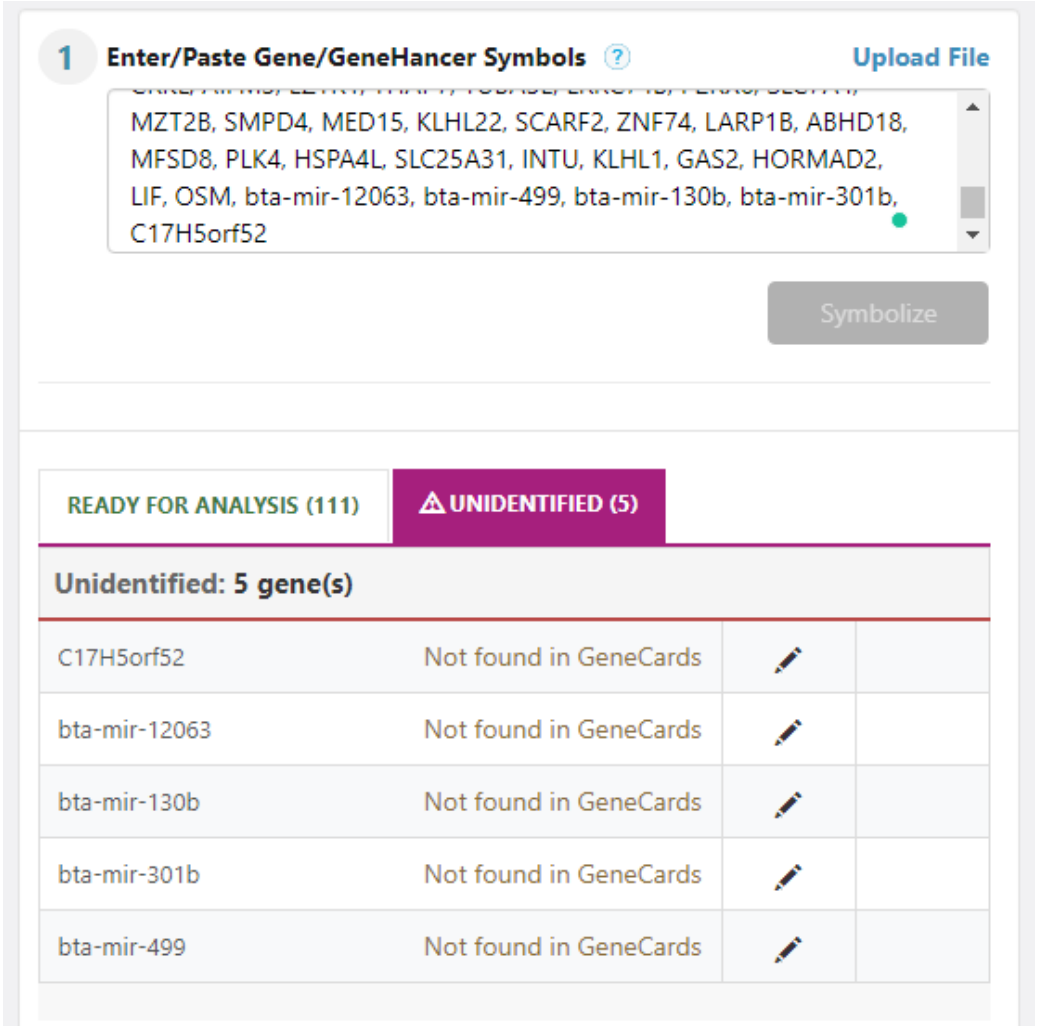
Symbol	Name	
ABCC11	ATP Binding Cassette Subfamily C Member 11	×
ABHD18	Abhydrolase Domain Containing 18	×
ACSS2	Acyl-CoA Synthetase Short Chain Family Member 2	×
AIFM3	Apoptosis Inducing Factor Mitochondria Associated 3	×

→ Depois de clicar em "*Symbolize*" ele vai mostrar os genes que foram identificados em "*READY FOR ANALYSIS*"

→ Se algum gene não foi identificado, estará na coluna "*UNIDENTIFIED*"

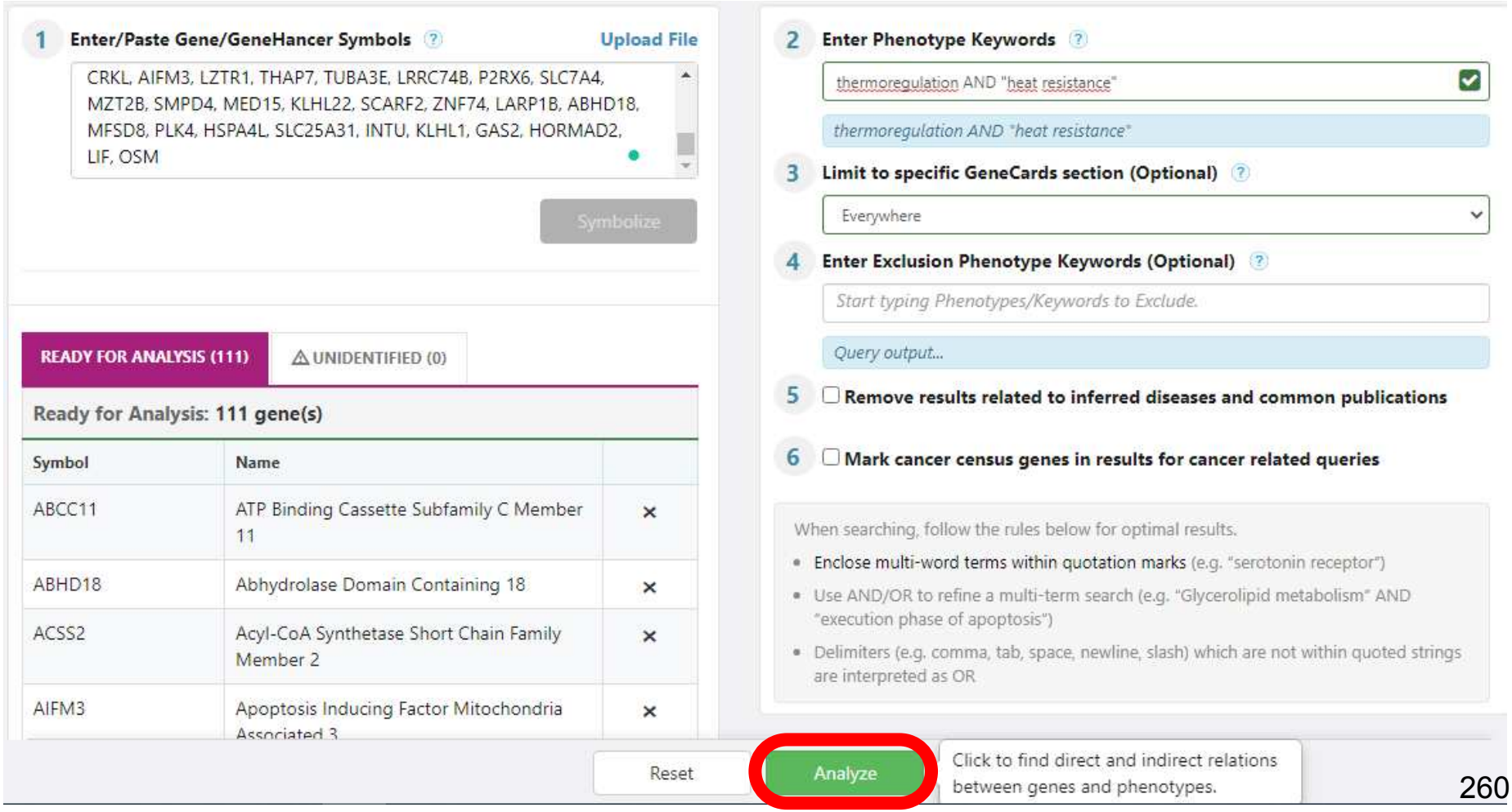
→ Somente à exemplo, abaixo estão genes que não seriam identificados

→ No caso, são microRNAs e ORF (*open reading frame*), uma fase de leitura aberta não identificada

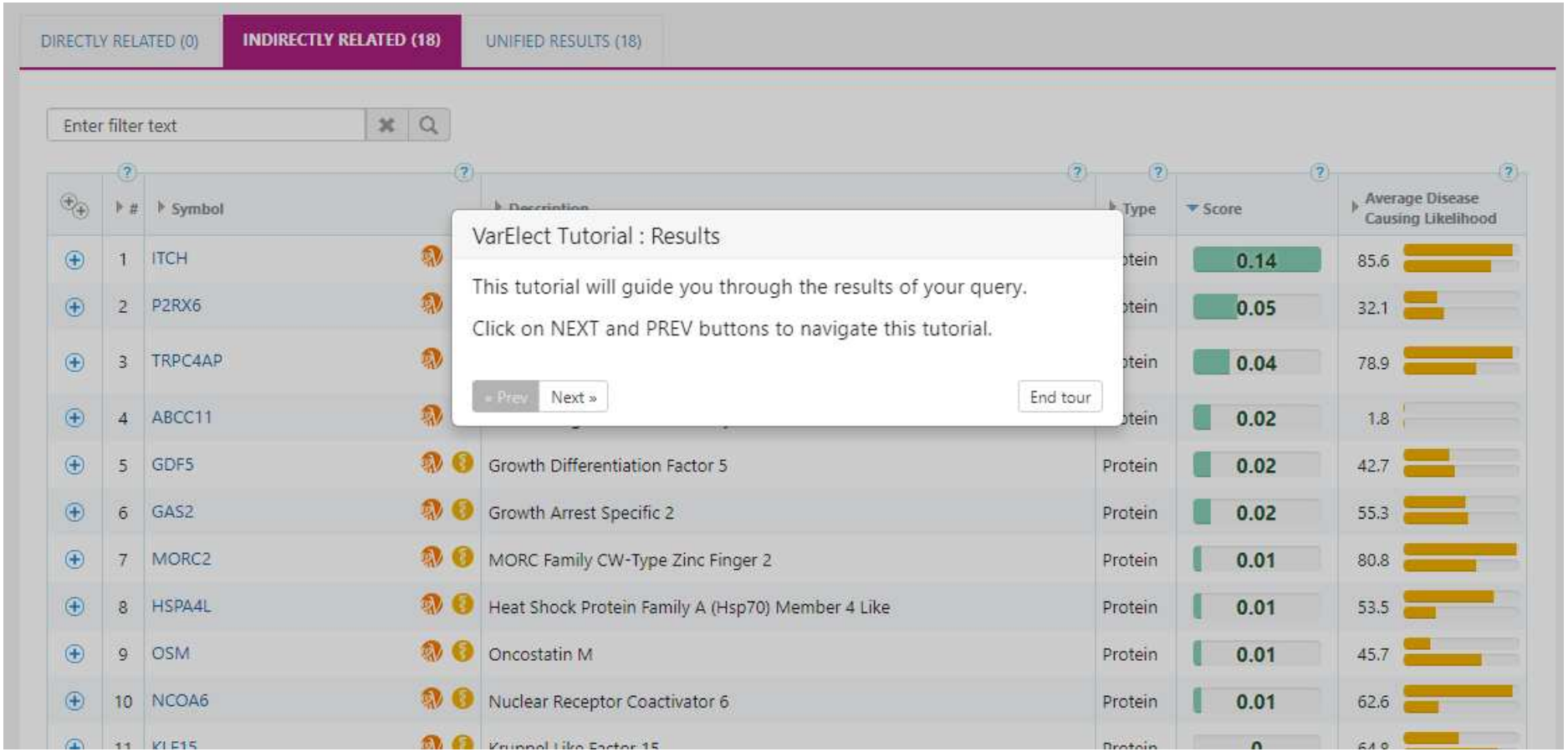


- Inserir o/s fenótipo/s no campo (2)
- Verificar mais detalhes nas (?), se necessário
- Os campos 3, 4, 5 e 6 são opcionais

- Clicar em "*Analyse*"

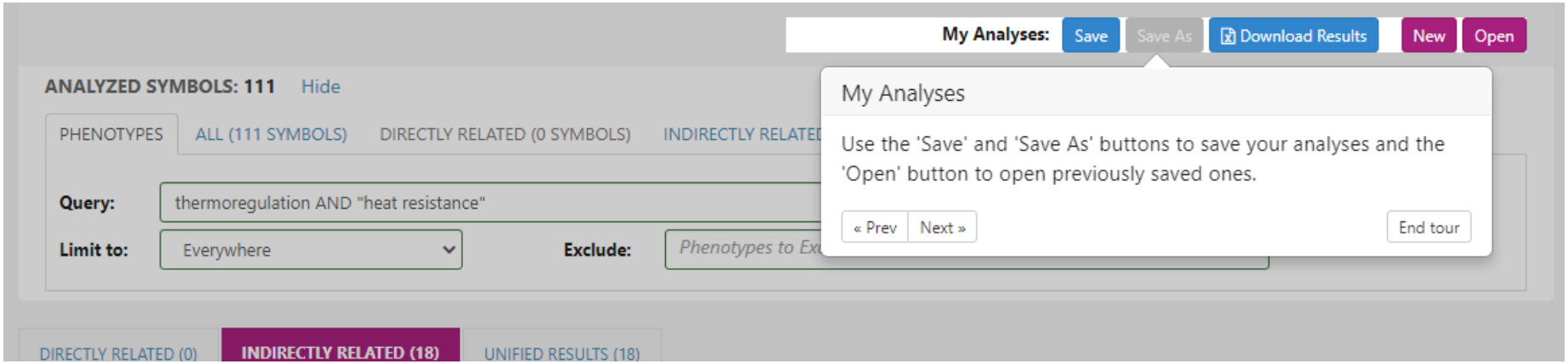


- A princípio, o próprio site passa um tutorial para você verificar os resultados da análise

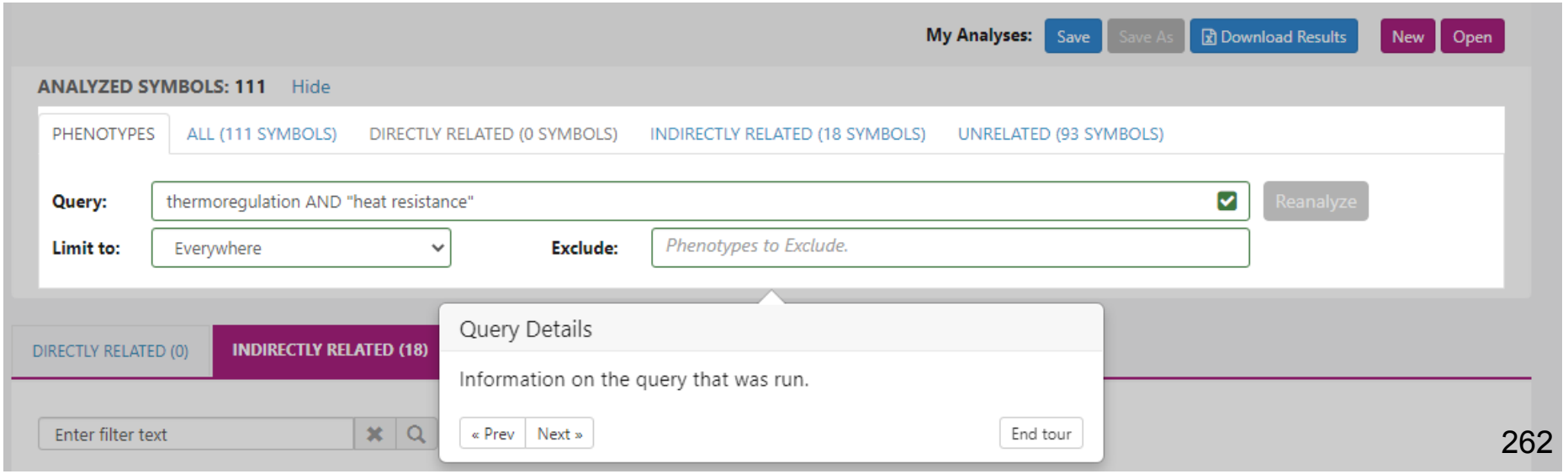


- “*My analyses*”:

- Use “*Save*” e “*Save As*” para salvar os resultados das análises
- Use “Open” para abrir análises que já foram salvas

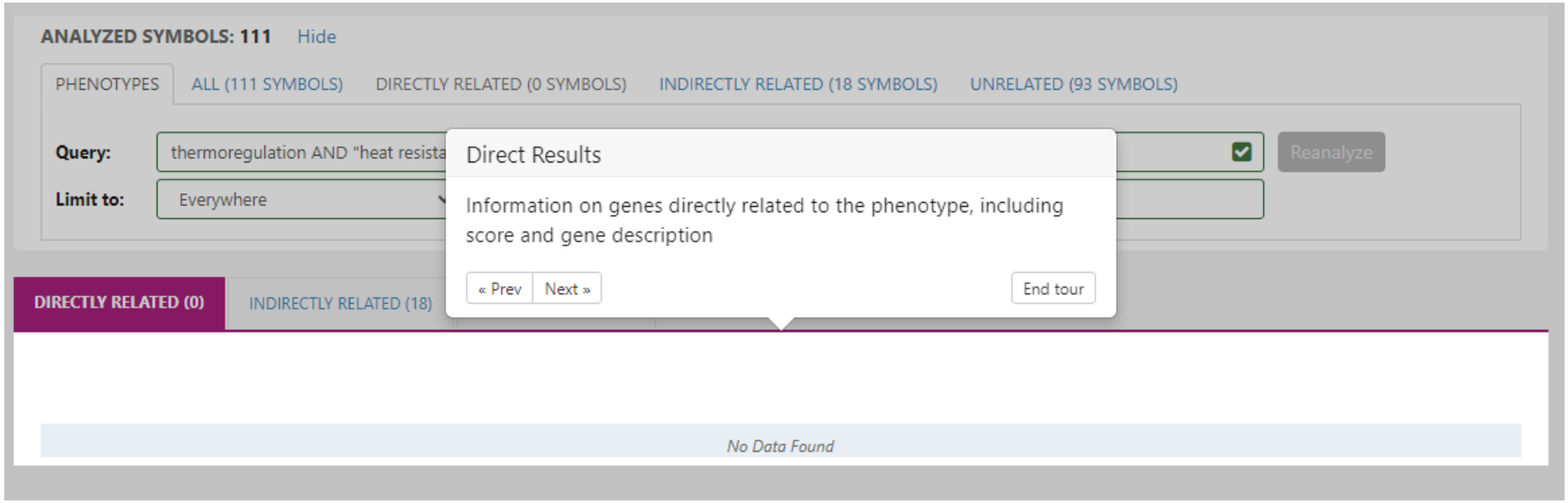


- Nesta parte tem os detalhes da análise: fenótipos incluídos e excluídos e mais...



- Em "*DIRECTLY RELATED*" terá a lista de genes **diretamente** relacionados com o fenótipo estudado

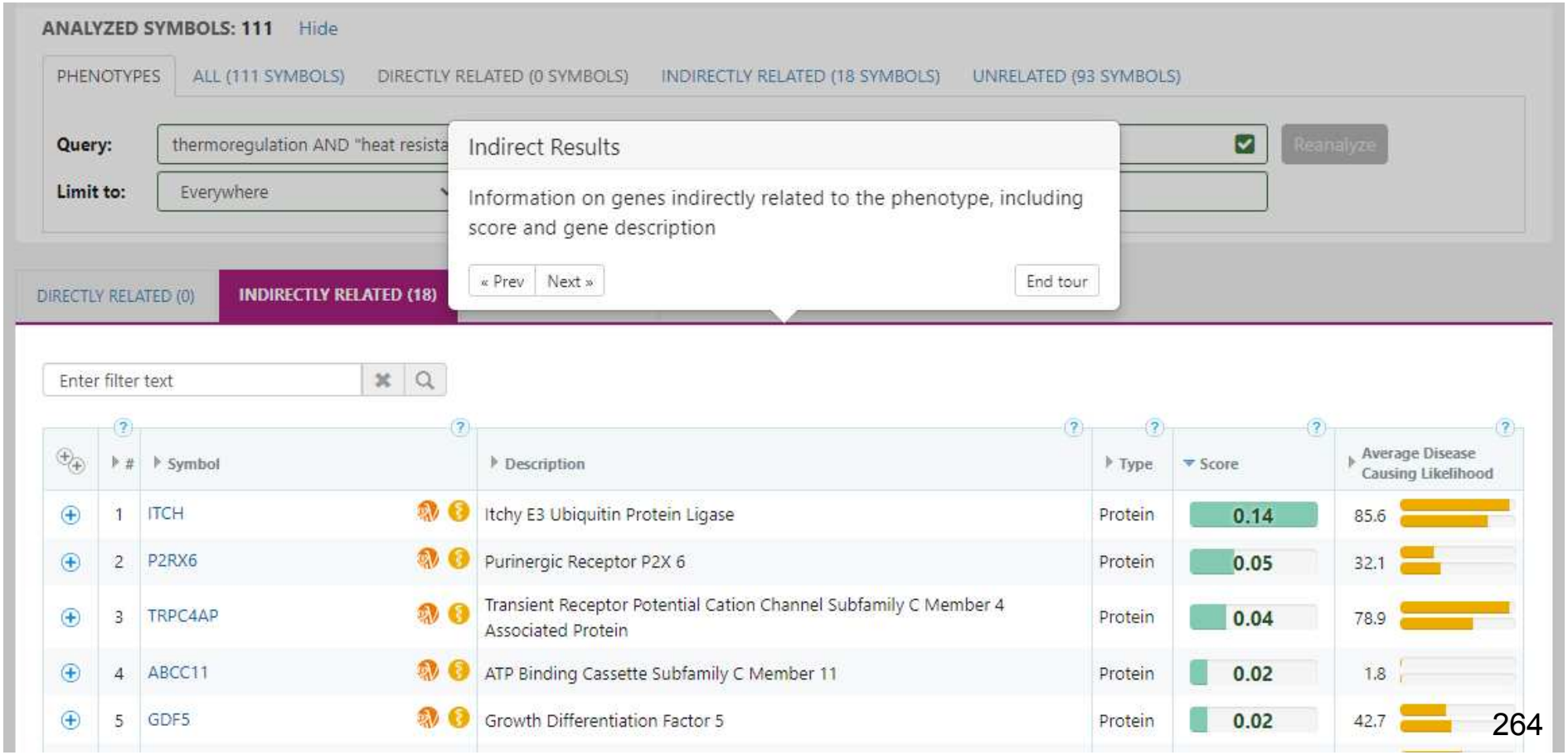
*Como este é só um exemplo aleatório, nenhum gene diretamente relacionado com o/s fenótipo/s foi encontrado



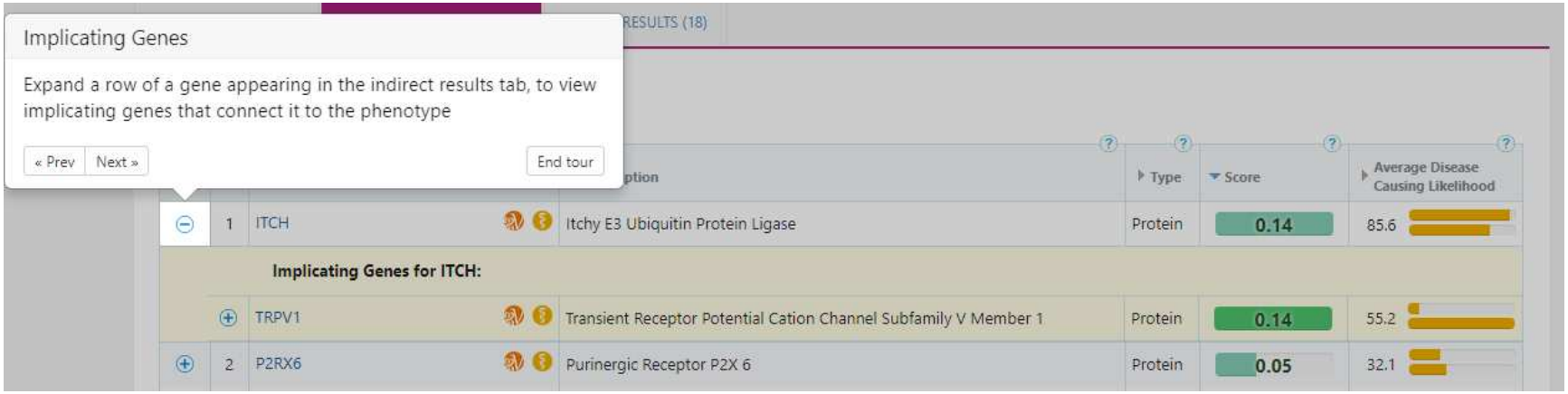
- Em "*INDIRECTLY RELATED*" terá a lista de genes **indiretamente** relacionados com a característica estudada
- Esta parte é especialmente relevante quando nenhuma ligação direta foi encontrada

- A associação indireta entre estes genes e o fenótipo pode ser por meio de:

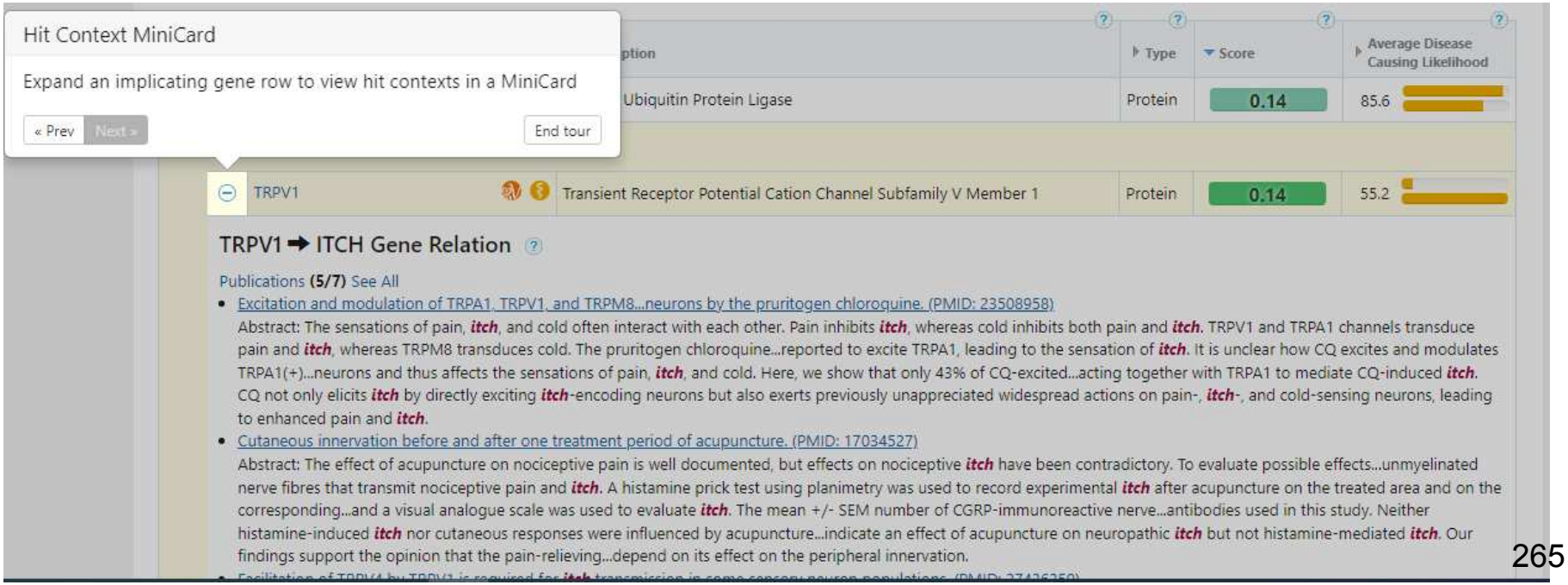
➢ Vias compartilhadas (*shared pathways*), interação de proteína (*protein interaction*), relação de genes parálogos (*paralog relation*) e/ou publicações mútuas (*mutual publication*)



- Ao clicar no sinal de + ao lado esquerdo do nome de cada gene, a linha será expandida mostrando detalhes da relação do gene com a característica



- É possível expandir e encontrar mais detalhes de publicações que mencionaram o gene

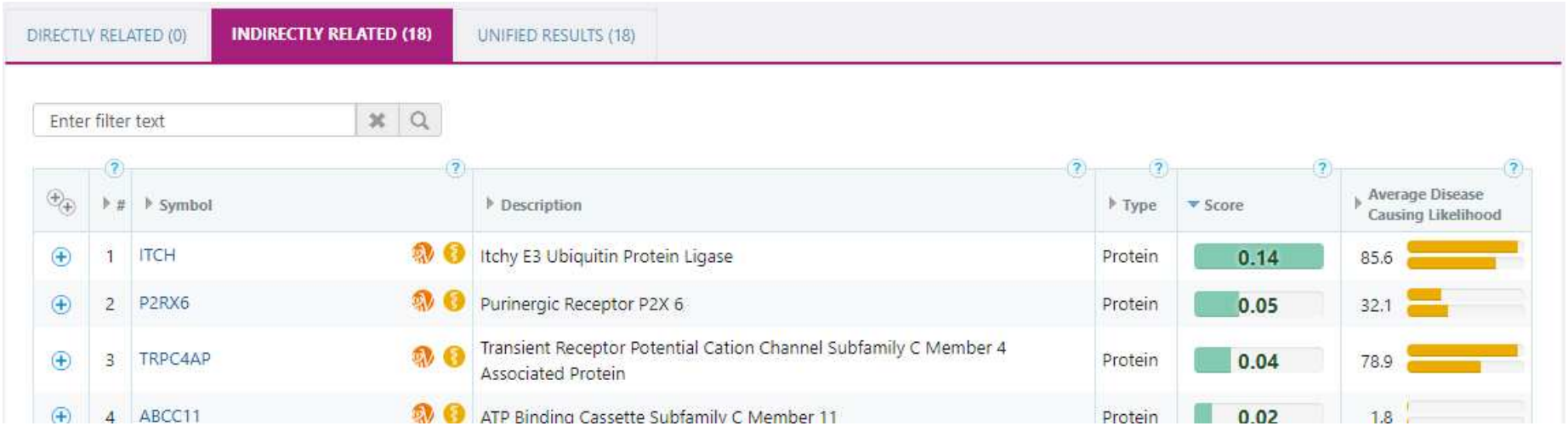


Dos 111 genes incluídos na análise, 18 foram selecionados

DIRECTLY RELATED (0) | **INDIRECTLY RELATED (18)** | UNIFIED RESULTS (18)

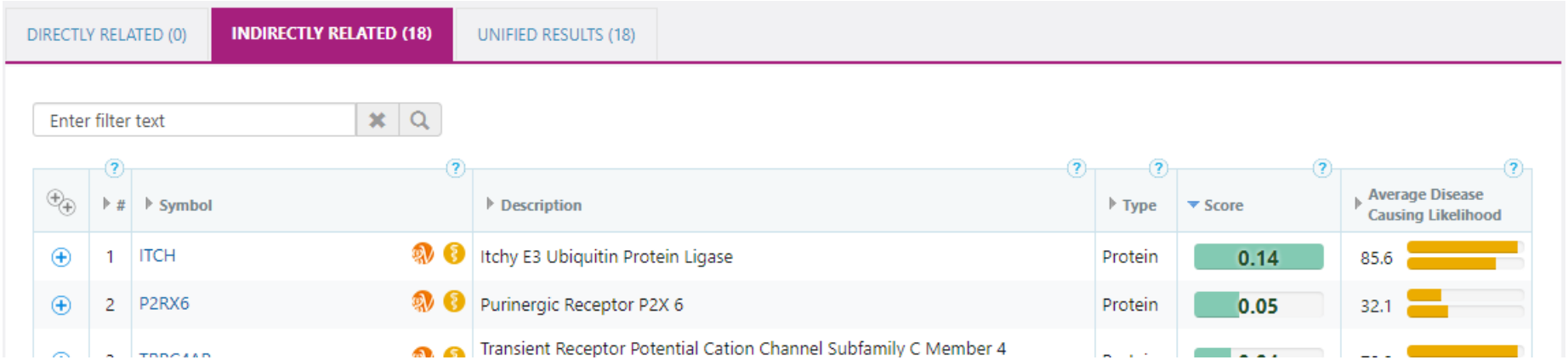
Enter filter text

#	Symbol	Description	Type	Score	Average Disease Causing Likelihood
1	ITCH	Itchy E3 Ubiquitin Protein Ligase	Protein	0.14	85.6
2	P2RX6	Purinergic Receptor P2X 6	Protein	0.05	32.1
3	TRPC4AP	Transient Receptor Potential Cation Channel Subfamily C Member 4 Associated Protein	Protein	0.04	78.9
4	ABCC11	ATP Binding Cassette Subfamily C Member 11	Protein	0.02	1.8
5	GDF5	Growth Differentiation Factor 5	Protein	0.02	42.7
6	GAS2	Growth Arrest Specific 2	Protein	0.02	55.3
7	MORC2	MORC Family CW-Type Zinc Finger 2	Protein	0.01	80.8
8	HSPA4L	Heat Shock Protein Family A (Hsp70) Member 4 Like	Protein	0.01	53.5
9	OSM	Oncostatin M	Protein	0.01	45.7
10	NCOA6	Nuclear Receptor Coactivator 6	Protein	0.01	62.6
11	KLF15	Kruppel Like Factor 15	Protein	0	64.8
12	MYH7B	Myosin Heavy Chain 7B	Protein	0	32.0
13	MAP1LC3A	Microtubule Associated Protein 1 Light Chain 3 Alpha	Protein	0	75.6
14	CLDN14	Claudin 14	Protein	0	53.4
15	HAPLN1	Hyaluronan And Proteoglycan Link Protein 1	Protein	0	55.3
16	LIF	LIF Interleukin 6 Family Cytokine	Protein	0	55.2
17	VCAN	Versican	Protein	0	16.7
18	CRKL	CRK Like Proto-Oncogene, Adaptor Protein	Protein	0	81.6



- No cabeçalho da tabela temos:

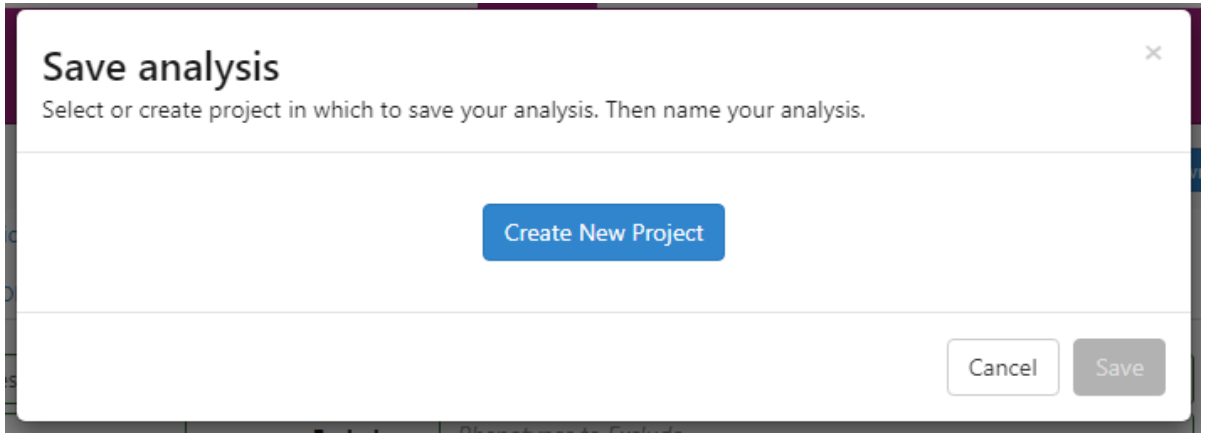
- ***Symbol***: o símbolo do gene
- ***Description***: a descrição/nome do gene
- ***Type***: Tipo de gene, exemplo codificador de proteína (*protein-coding*), Pseudogene, gene RNA, Família de gene (gene cluster), Locus genético (*genetic locus*) ou não categorizado (*uncategorized*).
- ***Score***: * próximo slide*
- ***Average Disease Causing Likelihood***: A coluna Probabilidade de Causar Doenças reflete o princípio de que uma variante em um gene com alta intolerância a mutações tem maior probabilidade de causar doenças. RVIS é o escore de intolerância à variação residual (Petrovski et al) e GDI é o Índice de Danos Genéticos (Itan et al). As probabilidades de causar doenças são 100% menos o percentil RVIS ou 100% menos o percentil GDI (barras laranja), com a média de ambos mostrada numericamente. ND - não determinado, NA - não aplicável.



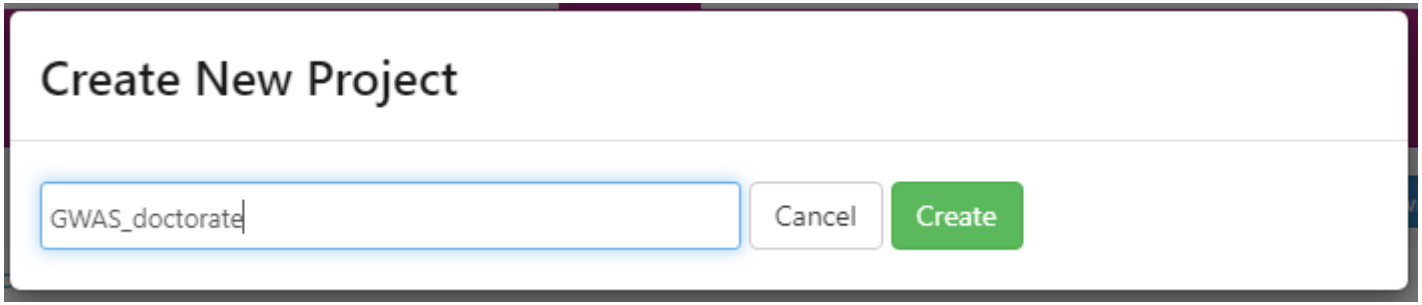
- ***Score***: Essa pontuação é uma indicação da força da conexão entre o gene e o(s) fenótipo(s).
  - As pontuações geralmente variam de 1 a 200.
  - O objetivo principal da pontuação é permitir classificar e priorizar a lista de genes consultados por relevância para o(s) fenótipo(s) consultado(s).
  - As pontuações de diferentes análises não podem e não devem ser comparadas, uma vez que as pontuações de relevância produzidas para cada análise são relativas dentro da análise e não absolutas.
  - A barra verde representa a proporção de cada pontuação/pontuação máxima da análise atual.
  - Ao analisar os elementos reguladores do GeneHancer, a pontuação do fenótipo é ajustada da seguinte forma: Cada associação gene-GeneHancer tem uma pontuação total, calculada pela multiplicação da pontuação de confiança GeneHancer pela pontuação da associação GeneHancer-Gene. As pontuações totais são normalizadas para um intervalo de 0,05 - 0,8 , e cada pontuação de gene-fenótipo do alvo do gene do elemento regulador é multiplicada pela pontuação total de GeneHancer normalizada.

Para salvar os resultados da análise

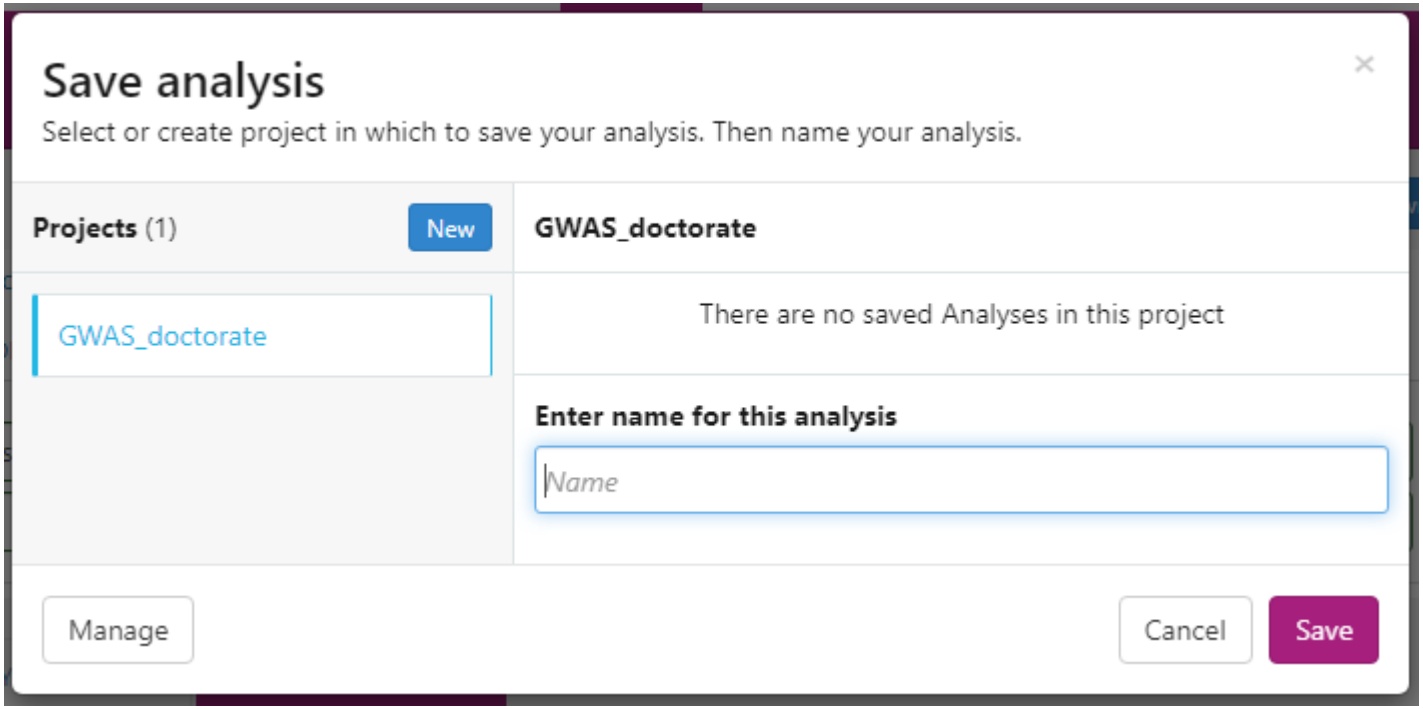
- Ir em "*My analyses*", clicar em "*Save*"
- Clicar em *"Create New Project"* para criar um novo projeto, ou se já tiver um, basta apenas selecioná-lo



- Após descrever o nome do projeto, clicar em "*Create*"



- Definir um nome para a análise realizada e clicar em "*Save*"



- As descrições originais sobre o ***Score*** e ***Average Disease Causing Likelihood*** estão nas (?) acima de cada item.

- Para mais detalhes e informações, verificar o site: https://varelect.genecards.org/

- Links úteis:
- https://varelect.genecards.org/about/#enhancers
- https://youtu.be/rH1L4dGIS4g

UNIVERSIDADE FEDERAL DE VIÇOSA
CENTRO DE CIÊNCIAS AGRÁRIAS
DEPARTAMENTO DE ZOOTECNIA

TUTORIAL
Análises de ontologia gênica (Gene Ontology - GO)

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

Análises de ontologia gênica (Gene Ontology - GO)

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: 978-65-5668-141-2
DOI: http://dx.doi.org/10.26626/9786556681412.2023B0001

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).

Tutorial

Análises GO no Cytoscape

Mais detalhes do uso do software podem ser consultados no Manual do Cytoscape

(Shannon et al. 2003; DOI: 10.1101/gr.1239303)

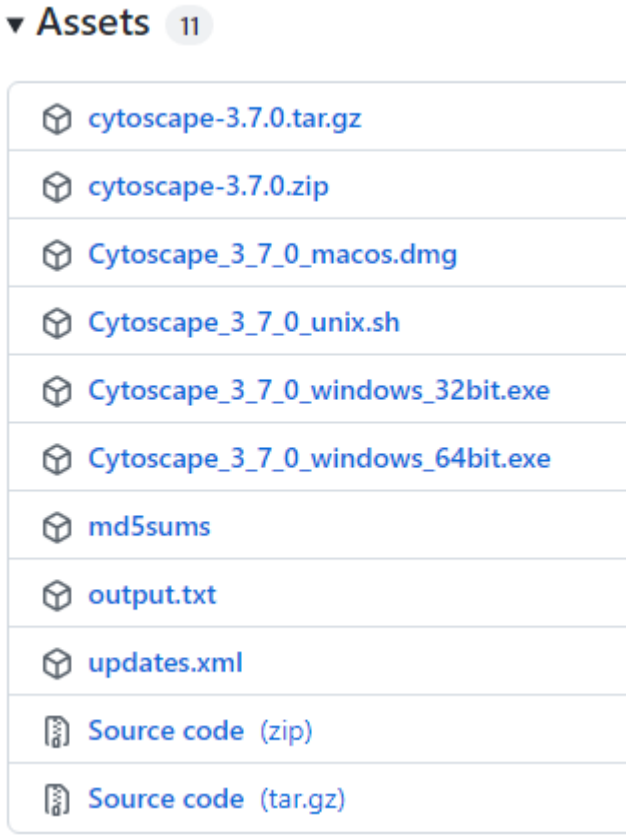
Para instalar o Cytoscape é preciso ter o Java instalado

Java: https://www.java.com/pt-BR/download/manual.jsp

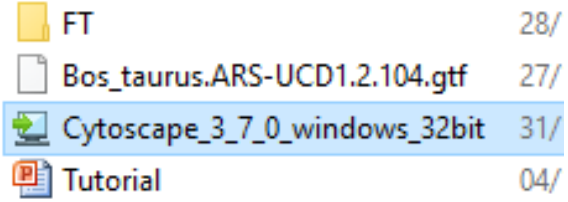
Cytoscape 3.7.0:
https://github.com/cytoscape/cytoscape/releases/3.7.0/

INSTALANDO CYTOSCAPE 3.7.0

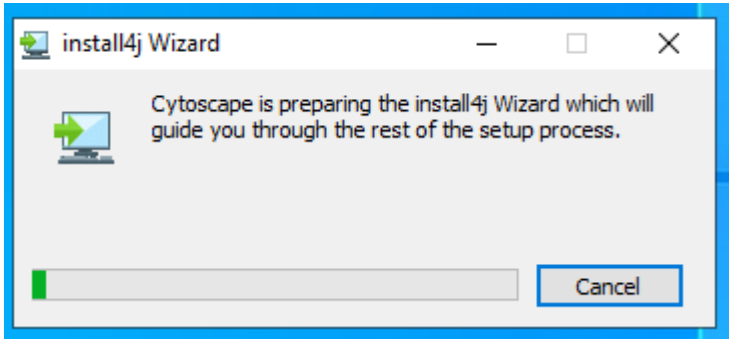
1º - No link do slide anterior, baixar a versão do Cytoscape condizente com seu pc



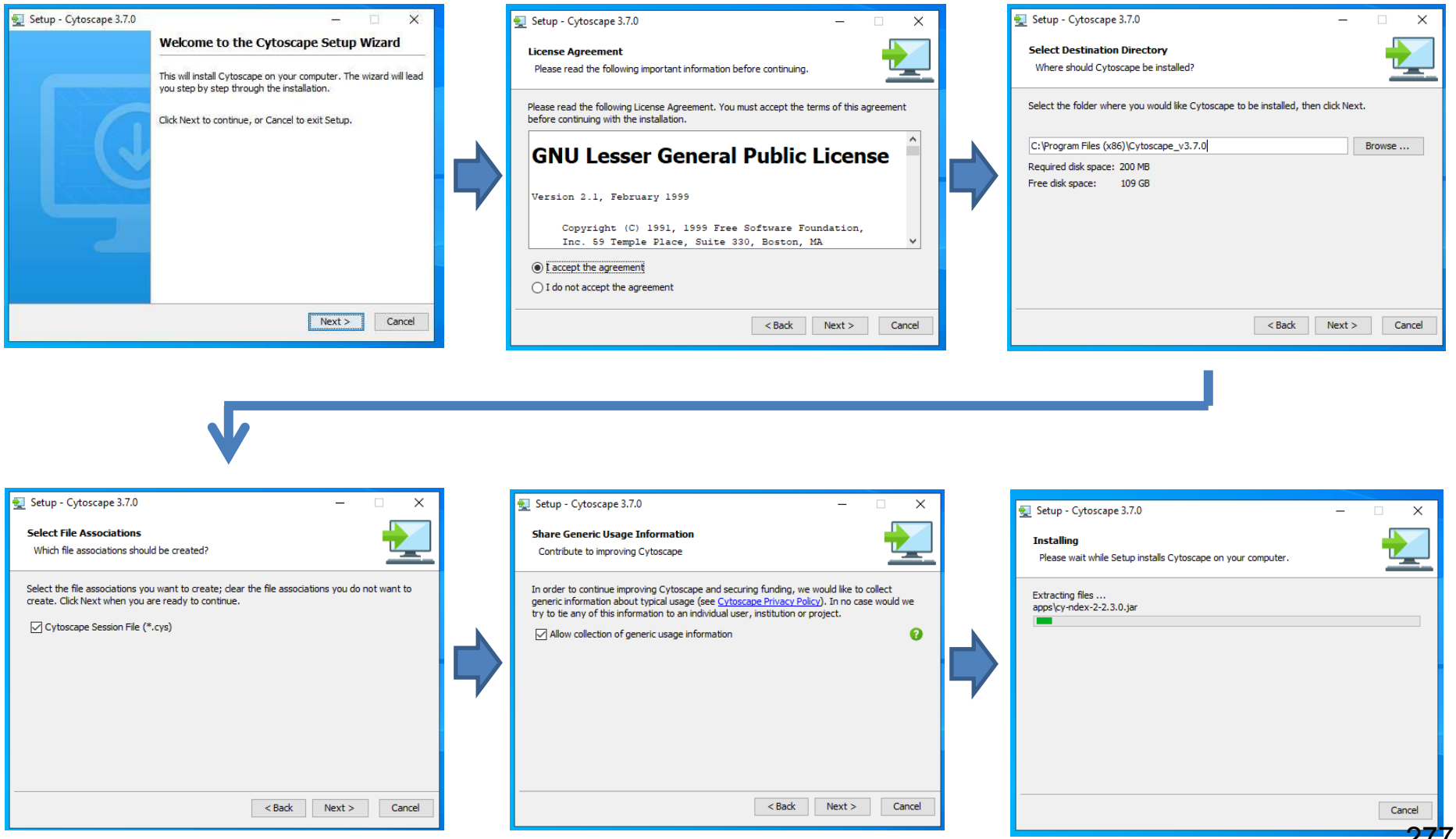
2º - Clicar duas vezes do ícone baixado



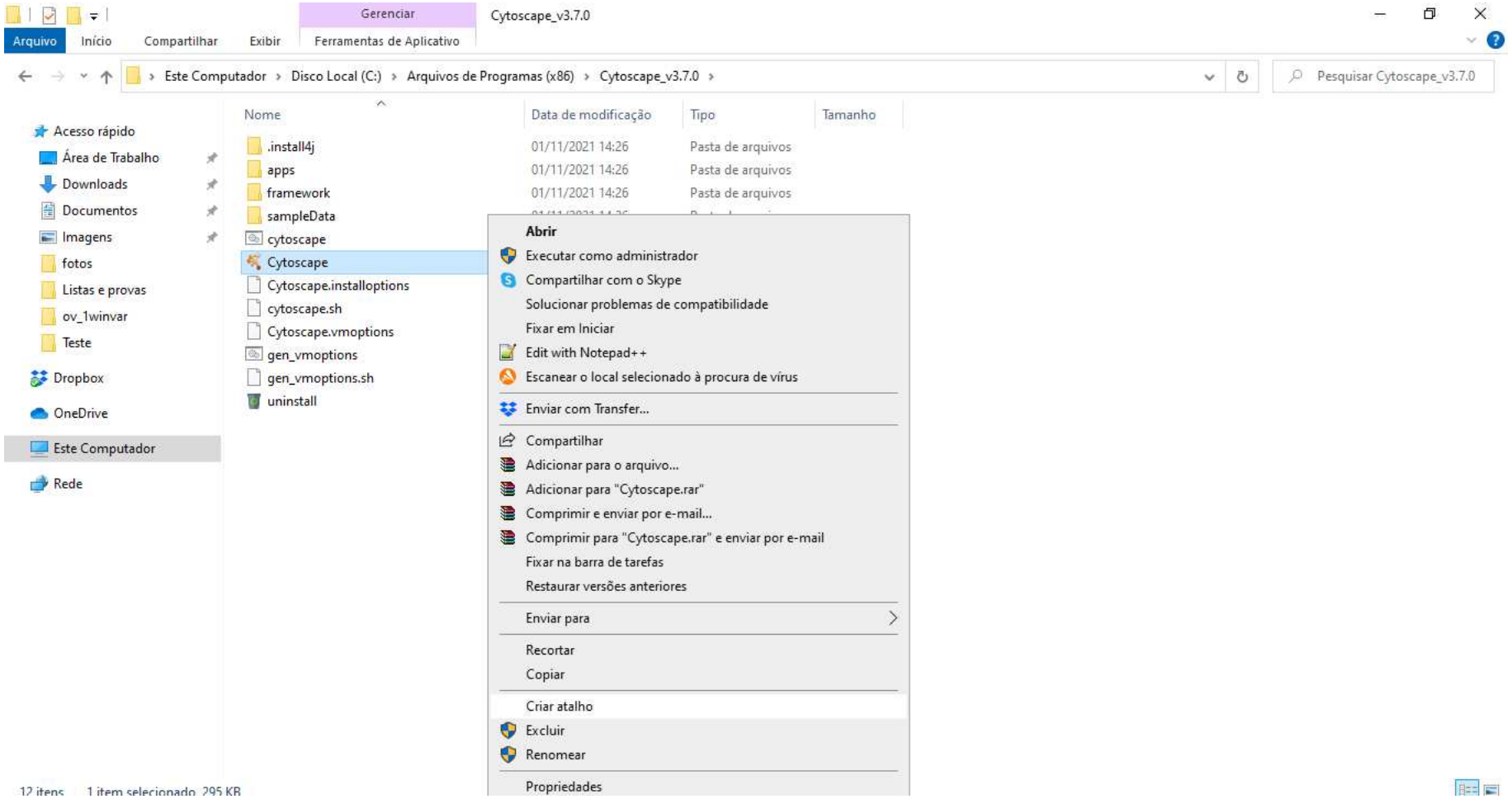
3º - Se não tiver o Java, vai aparecer uma tela perguntando se pode instalar o Java antes de instalar o Cytoscape
- Ou pode ir no site do Java, baixar e instalar
- Se já tiver o Java no seu PC, vai aparecer a seguinte tela, é só aguardar



Não alterar nenhum padrão durante a instalação, apenas clicar em Next



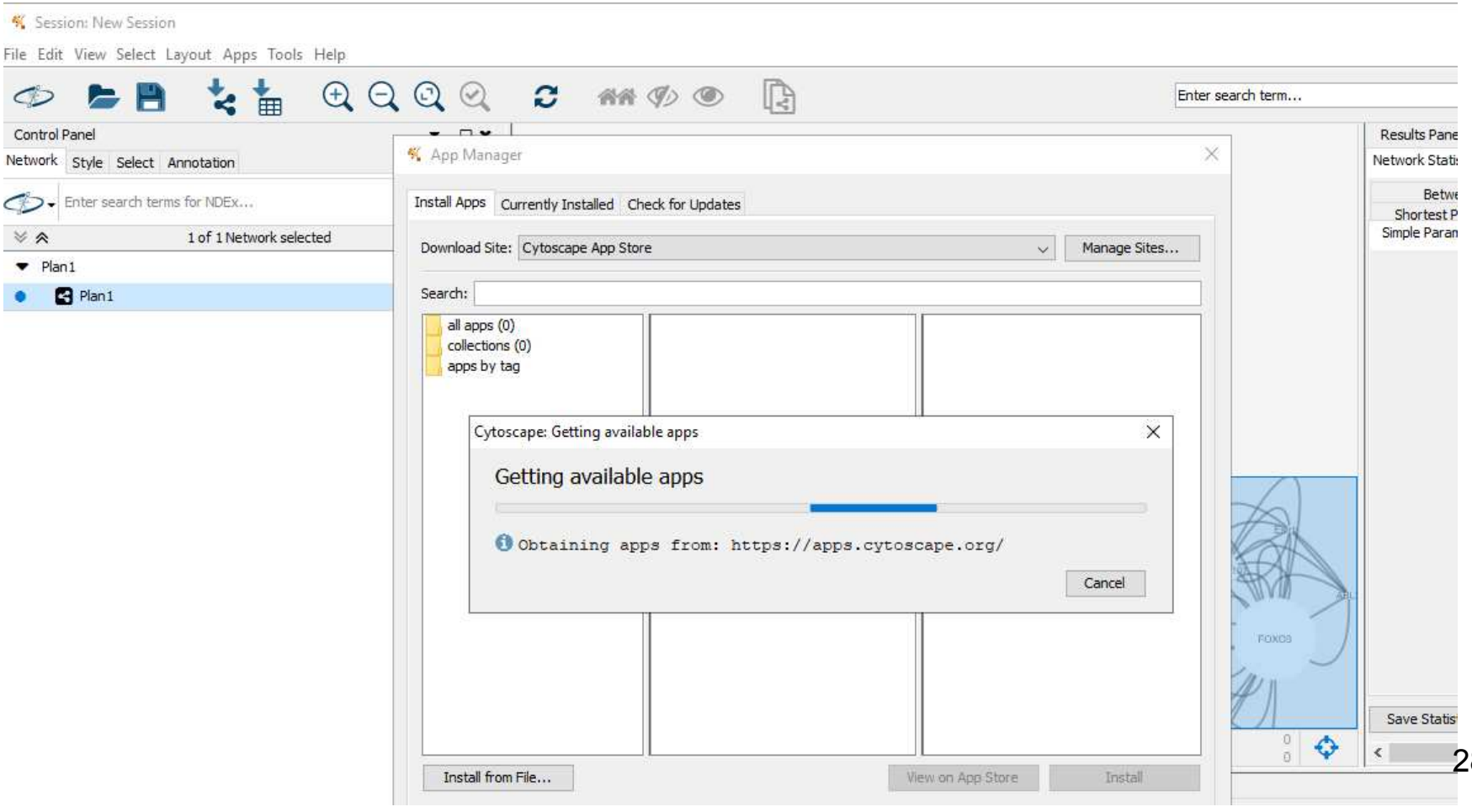
- Clicar em *Finish*
- O atalho para o Cytoscape ainda não vai aparecer na área de trabalho
- Para verificar se o programa instalou corretamente e para colocar um atalho do Cytoscape na área de trabalho, verificar o slide seguinte



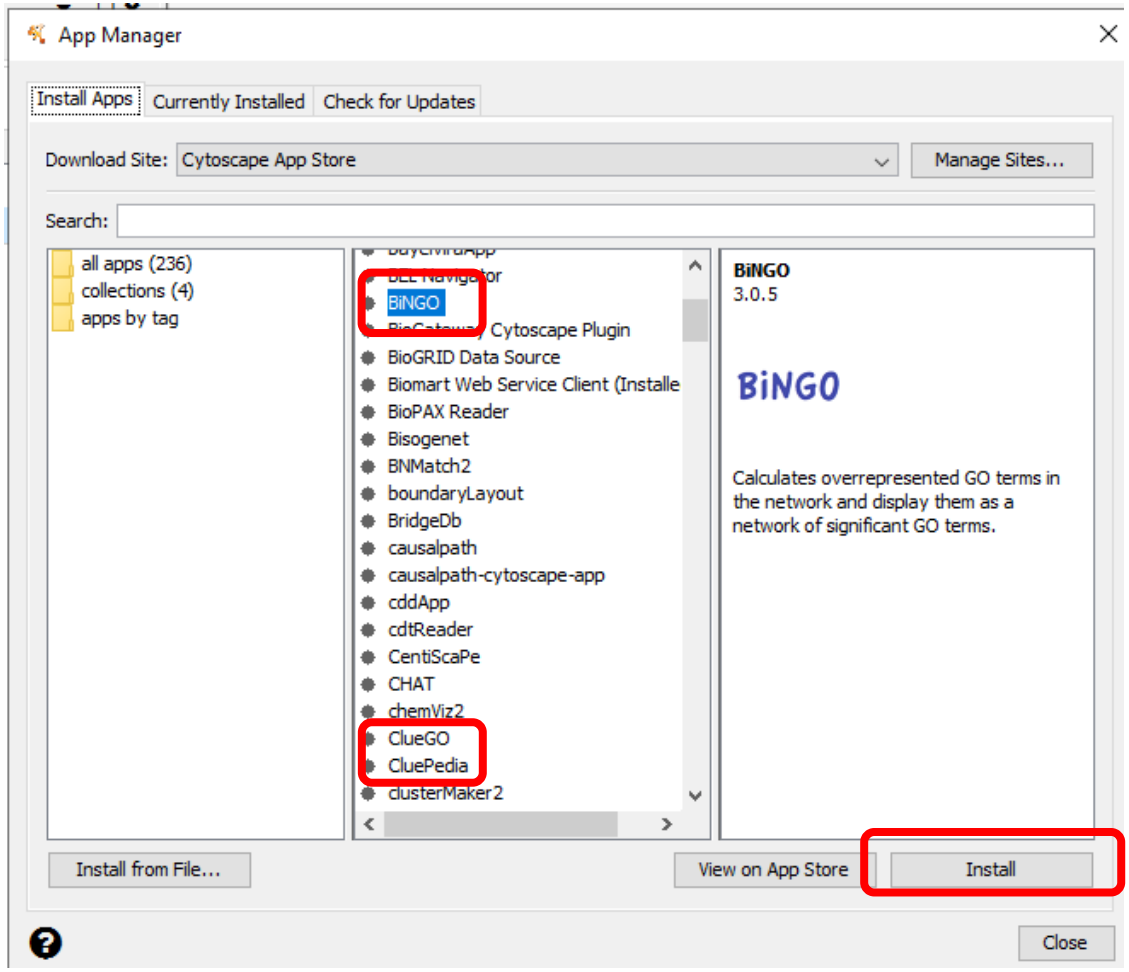
- Verificar se está na pasta direcionada na instalação
- Para verificar se foi instalado, clicar duas vezes no ícone Cytoscape com o desenho
- Se preferir, clicar com o botão direito sobre o mesmo ícone e clicar em ‘criar atalho’
- Vai aparecer uma caixa de mensagens dizendo que o atalho não pode ser criado dentro desta pasta e perguntando se quer criar esse atalho na área de trabalho
- Dê OK!

Instalando os Aplicativos BiNGO, ClueGO e CluePedia

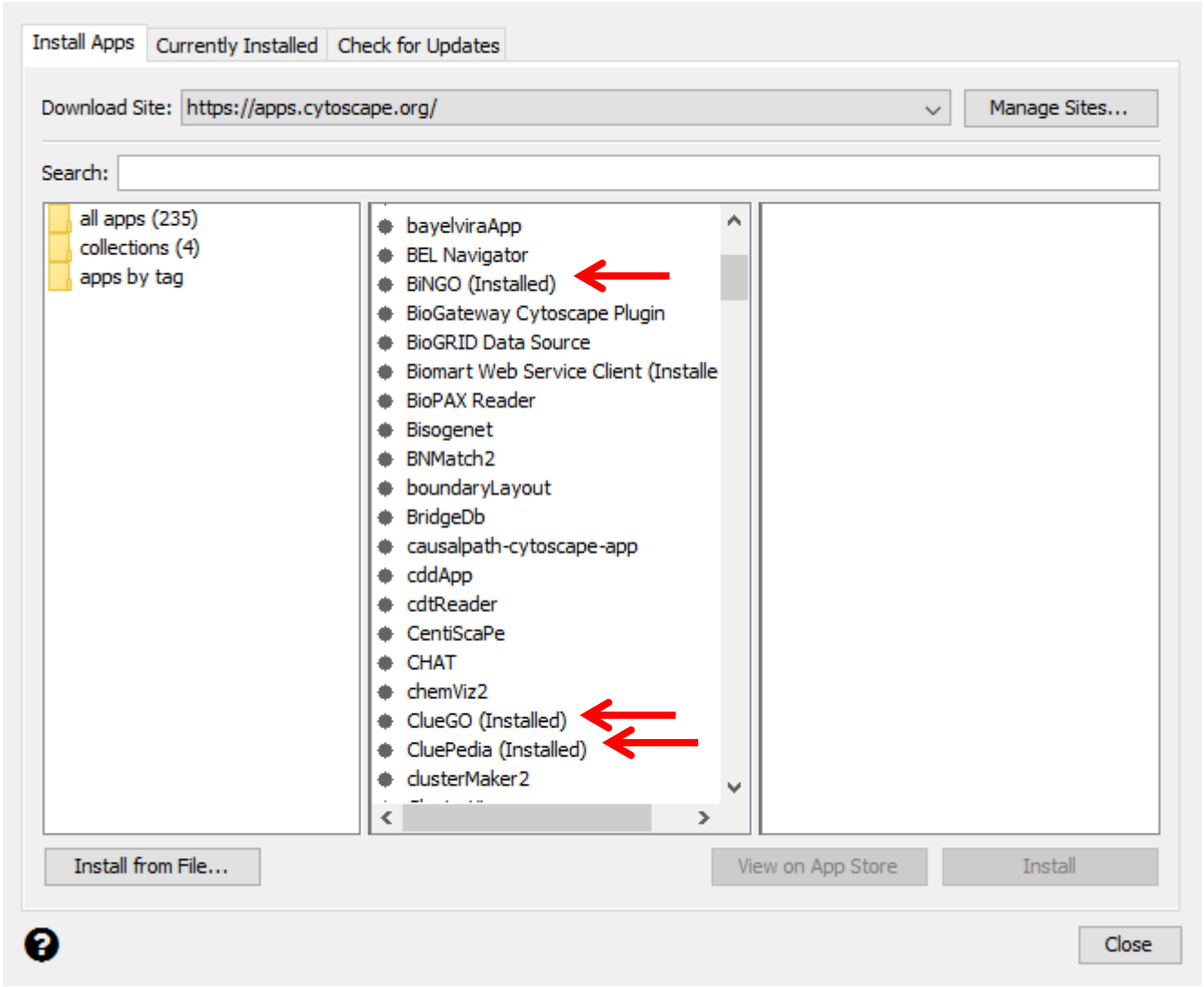
- Clicar em Apps → App Manager...



- Vai aparecer a tela com as opções dos aplicativos
- Clicar em BiNGO e clicar em Install
- Depois que terminar, fazer o mesmo com ClueGO e CluePedia, um de cada vez



Depois que está instalado aparece assim:



Mais detalhes dos aplicativos podem ser conferidos na documentação original do BiNGO (Maere et al., 2015; DOI: 10.1093/bioinformatics/bti551) e do ClueGO (Bindea et al. 2009; DOI: 10.1093/bioinformatics/btp101)

Análises Gene Ontology (GO)

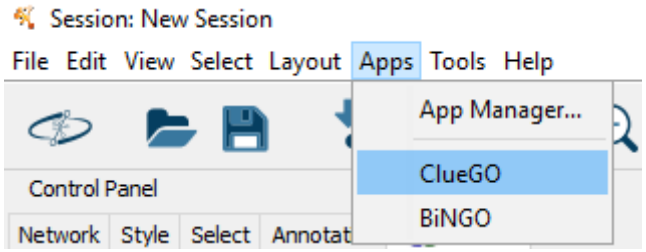
- Os seguintes genes serão usados de exemplo neste tutorial da análise de Gene Ontology:

- ABL2
- TMEM245
- LOC101907107
- PEAK1

Obtendo os processos biológicos associados aos genes

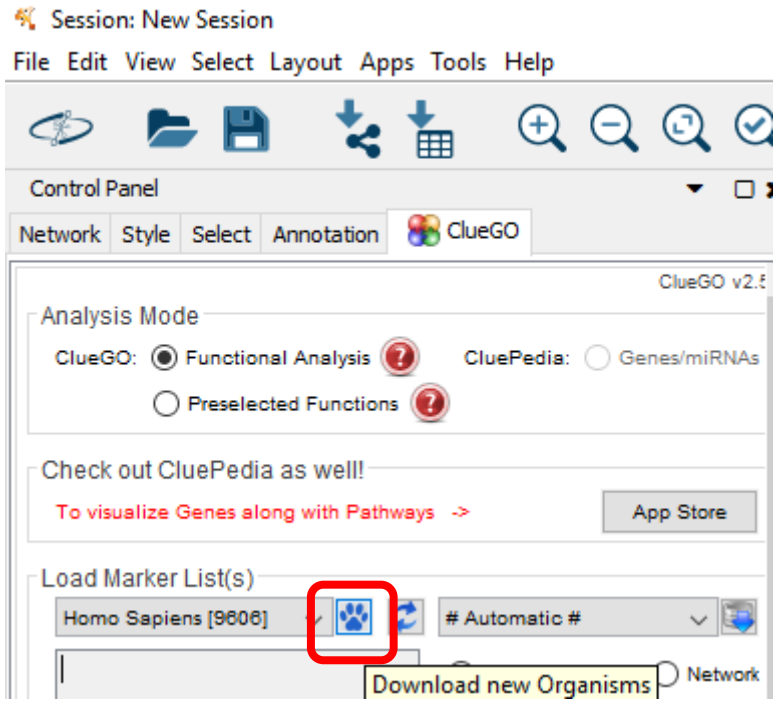
CLUEGO

ClueGO



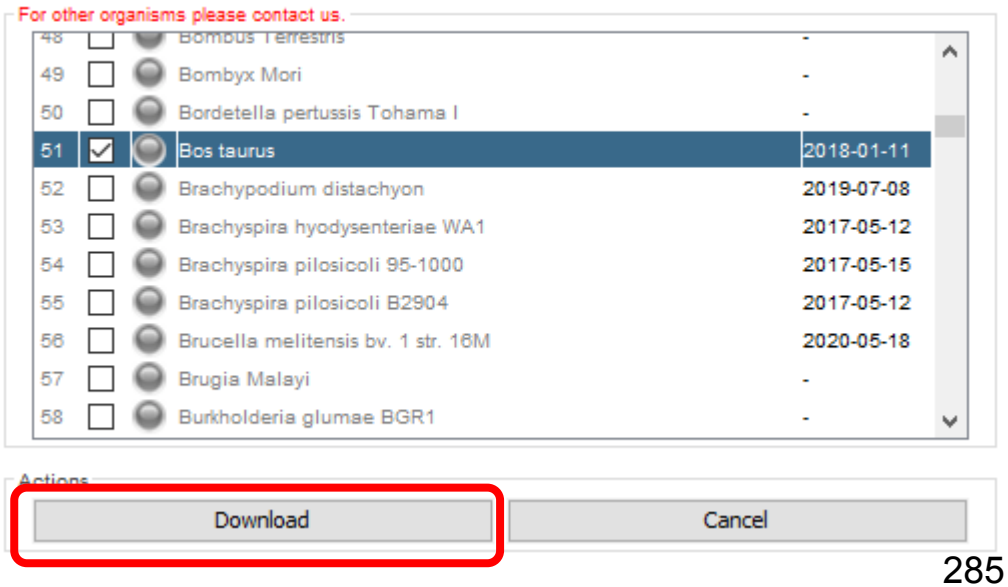
*Para usar o ClueGO é preciso ter uma licença. Para obter uma, usar o e-mail institucional, porque eles não enviam licença para e-mail pessoal. Essa licença pode ser usada por até 5 ou 6 pessoas

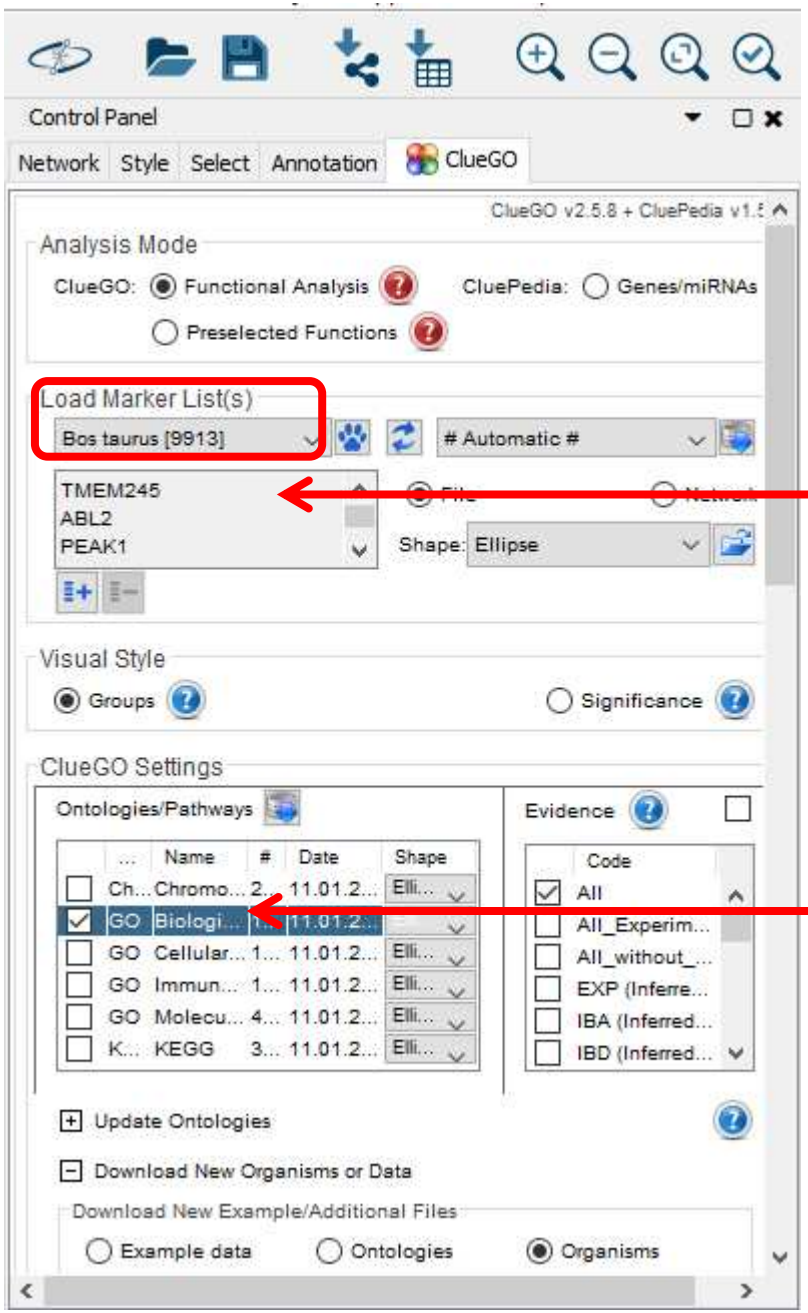
1º - Clicar na patinha



2º - Selecionar *Bos taurus* e fazer download

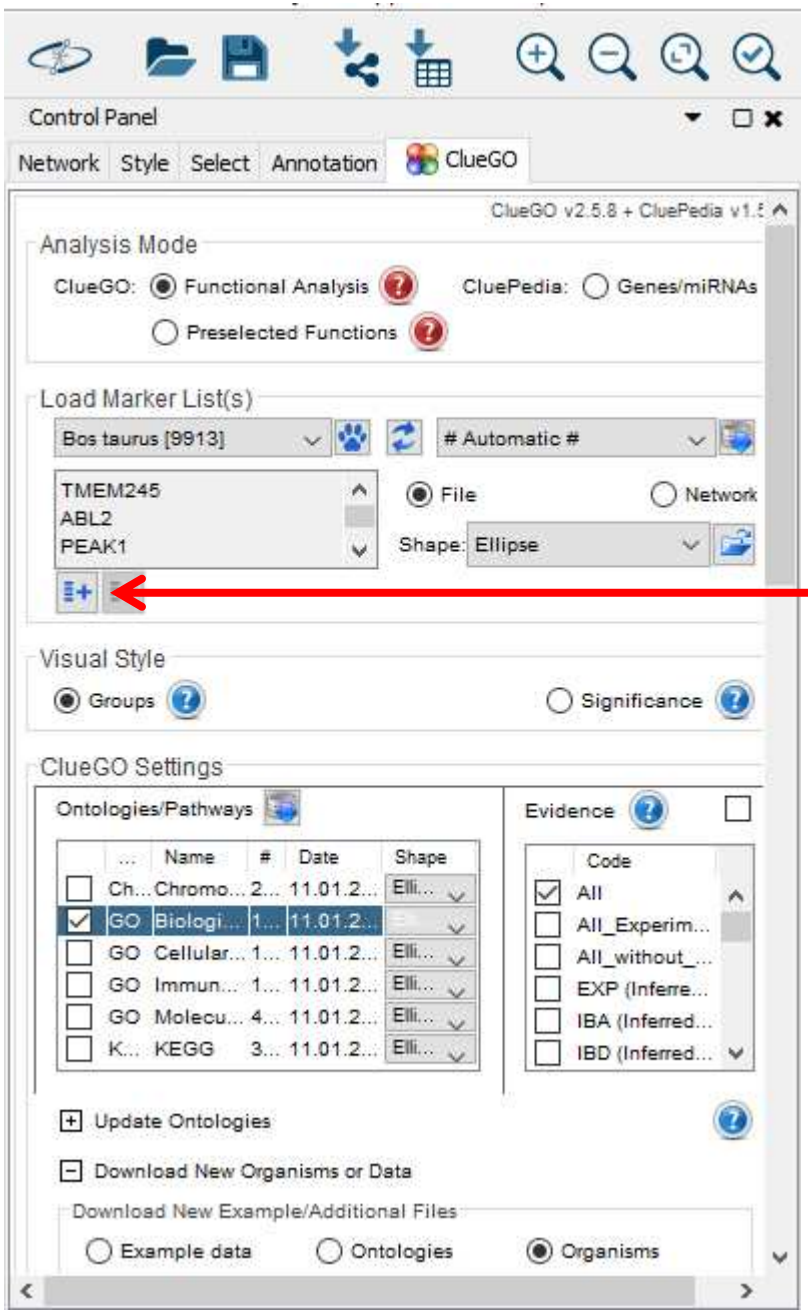
Se já tiver a espécie de interesse instalada pode apenas selecioná-la na setinha pra baixo





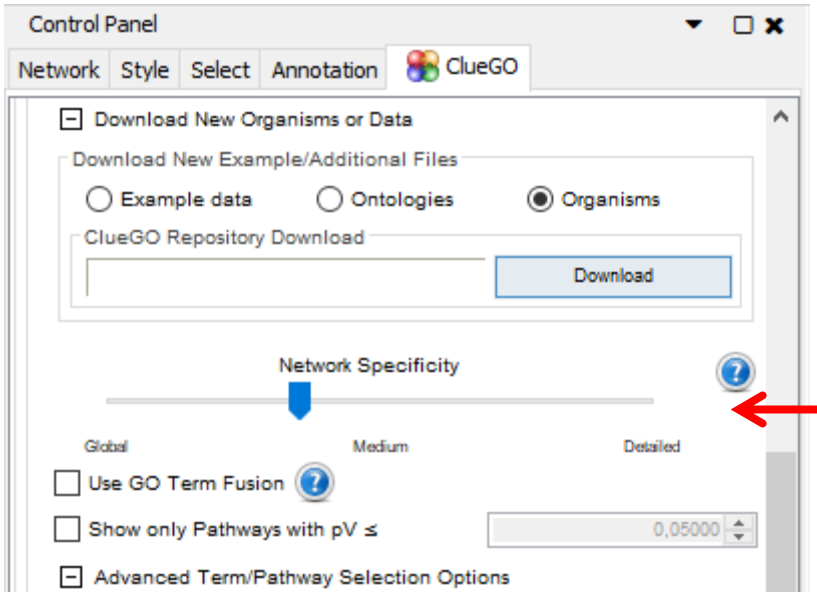
- Digitar o nome dos **genes** (Se tiver muitos genes, copiar de um arquivo pronto ou importar)

- Selecionar a opção de processos biológicos



- Nesta parte, se tiver diferentes grupos de genes, por exemplo, para características diferentes, clicar no + e digitar os nomes dos genes de cada grupo em cada caixa.

- Se não tiver, pode ignorar essa parte

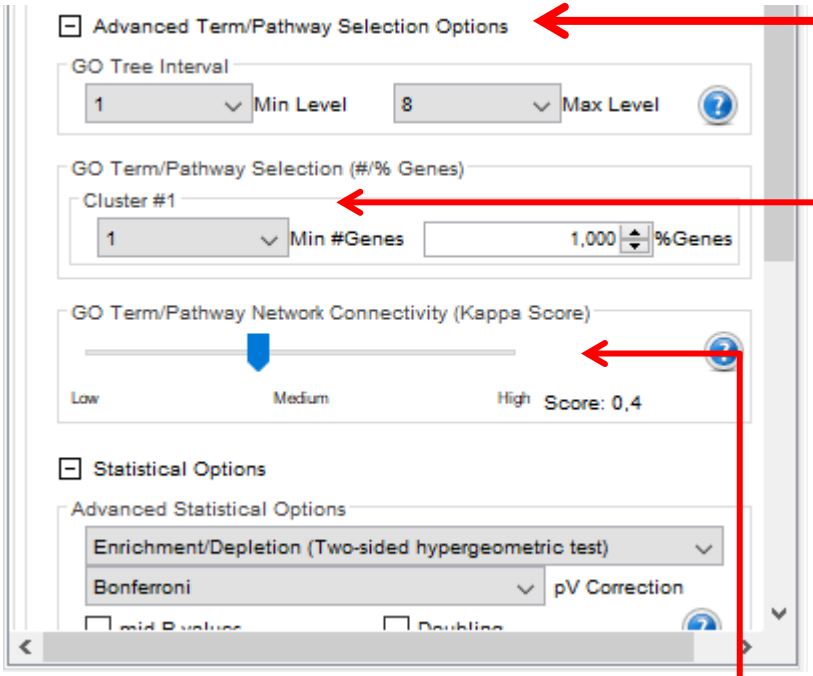


- Em **Network Specificity**

➢ Quando se tem muitos genes, podemos arrastar a setinha para deixar uma análise mais detalhada / específica

➢ Quando temos poucos genes, podemos fazer uma análise mais geral

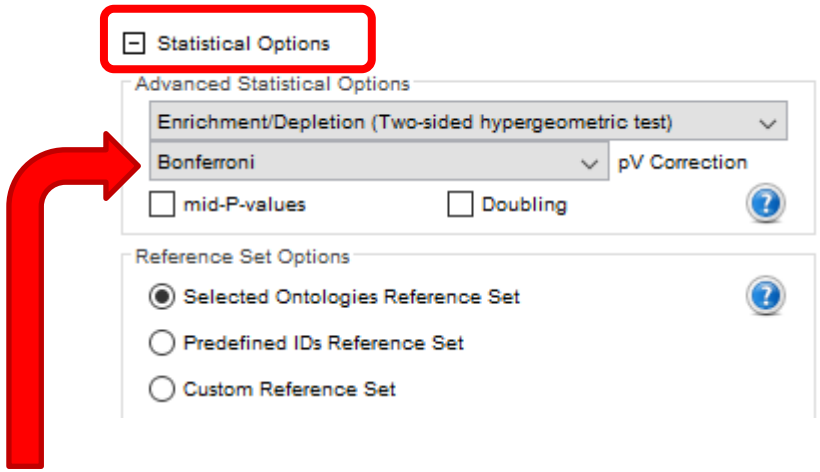
Ele não gera um valor para essa Network, mas é bom lembrarmos como esse parâmetro foi definido (se mais global, mais detalhado ou médio, ou outra opção nesse intervalo)



- Em **Advanced Term/Pathway Selection Options**
- GO Tree Interval
  - ➢ Mín = 1 e Máx = 8
- GO Term/Pathway Selection (#% Genes)
- Mín = 1
- Não alterar o Kappa Score, pois o mínimo padrão é 0,4 e isso já fica na análise

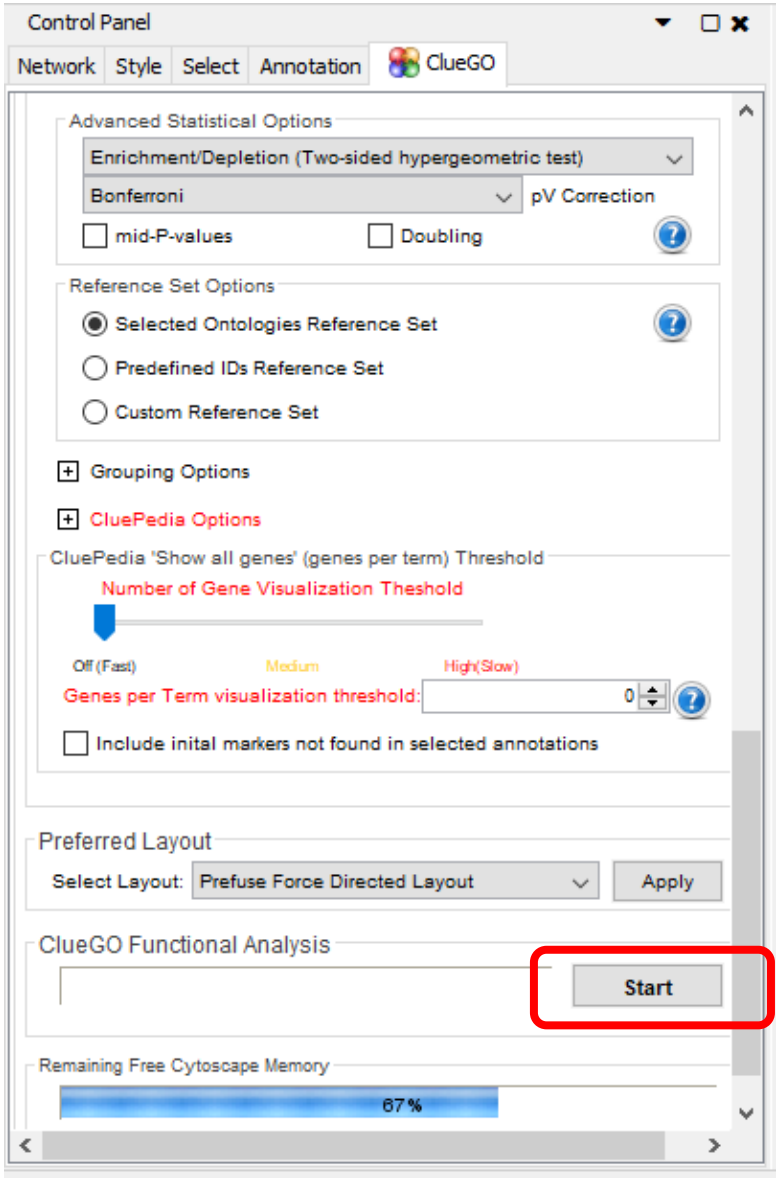
Estes parâmetros, assim como a Network Specificity (slide anterior), podem ser alterados para encontrar os resultados mais ajustados para o nosso objetivo, por isso essa parte pode ser realizada mais de uma vez para irmos comparando qual resultado fica melhor

Clicar em Statistical Options

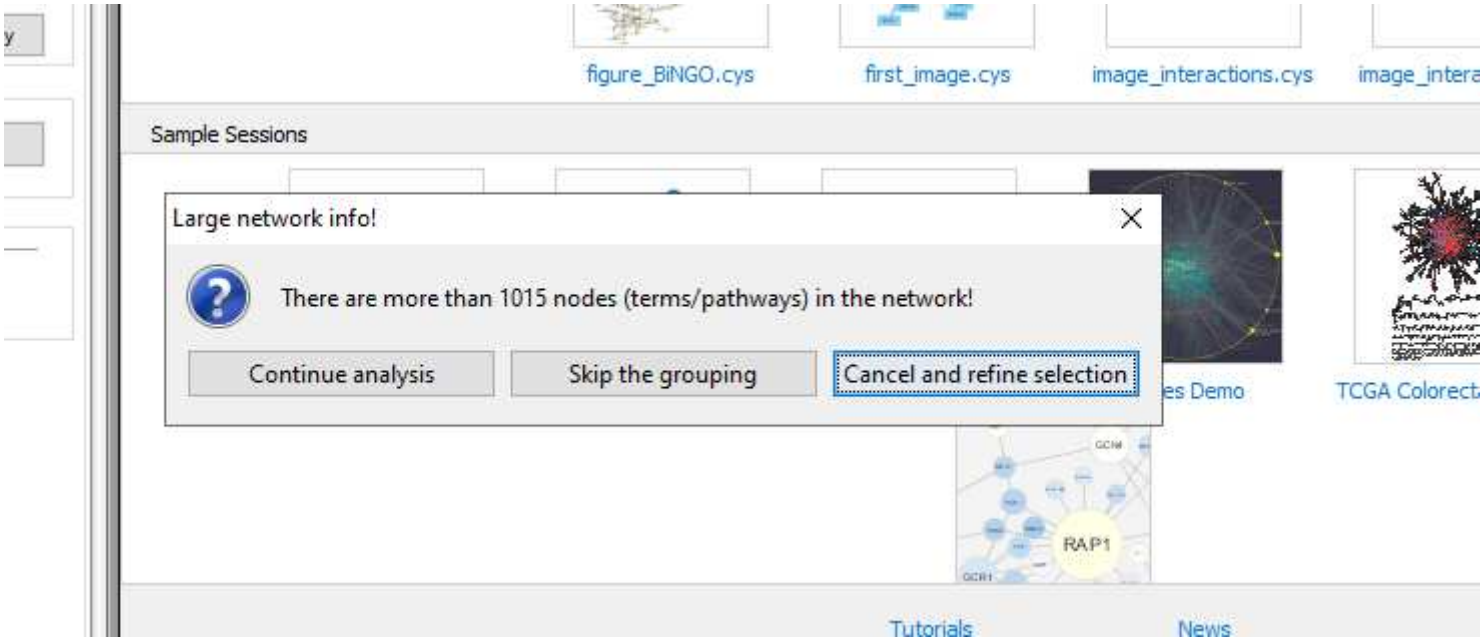


- Podemos escolher o teste estatístico mais apropriado
- Neste caso, a escolha foi o teste de Bonferroni
- Não alterar mais nenhum parâmetro

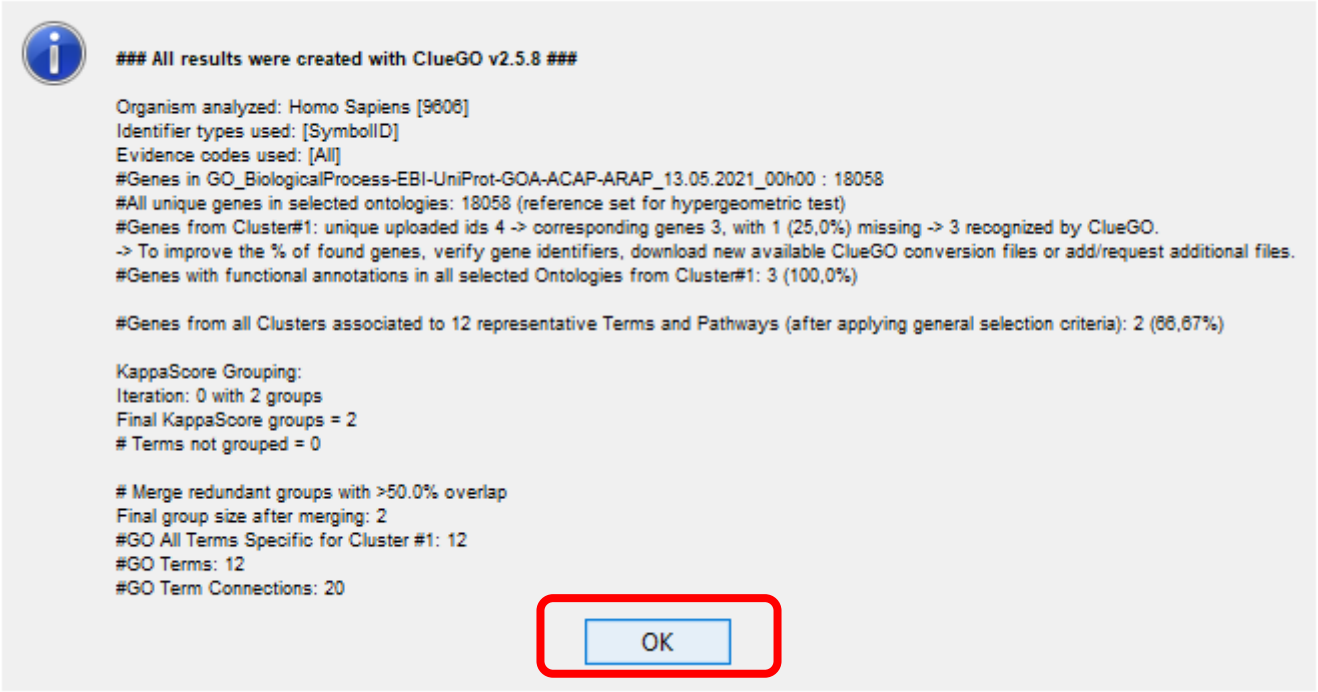
➢ Clicar em Start



Como isto é apenas um exemplo, temos poucos genes, mas na pesquisa real pode ser que tenha muitos genes. Neste caso, o Cytoscape pode dar esse aviso:



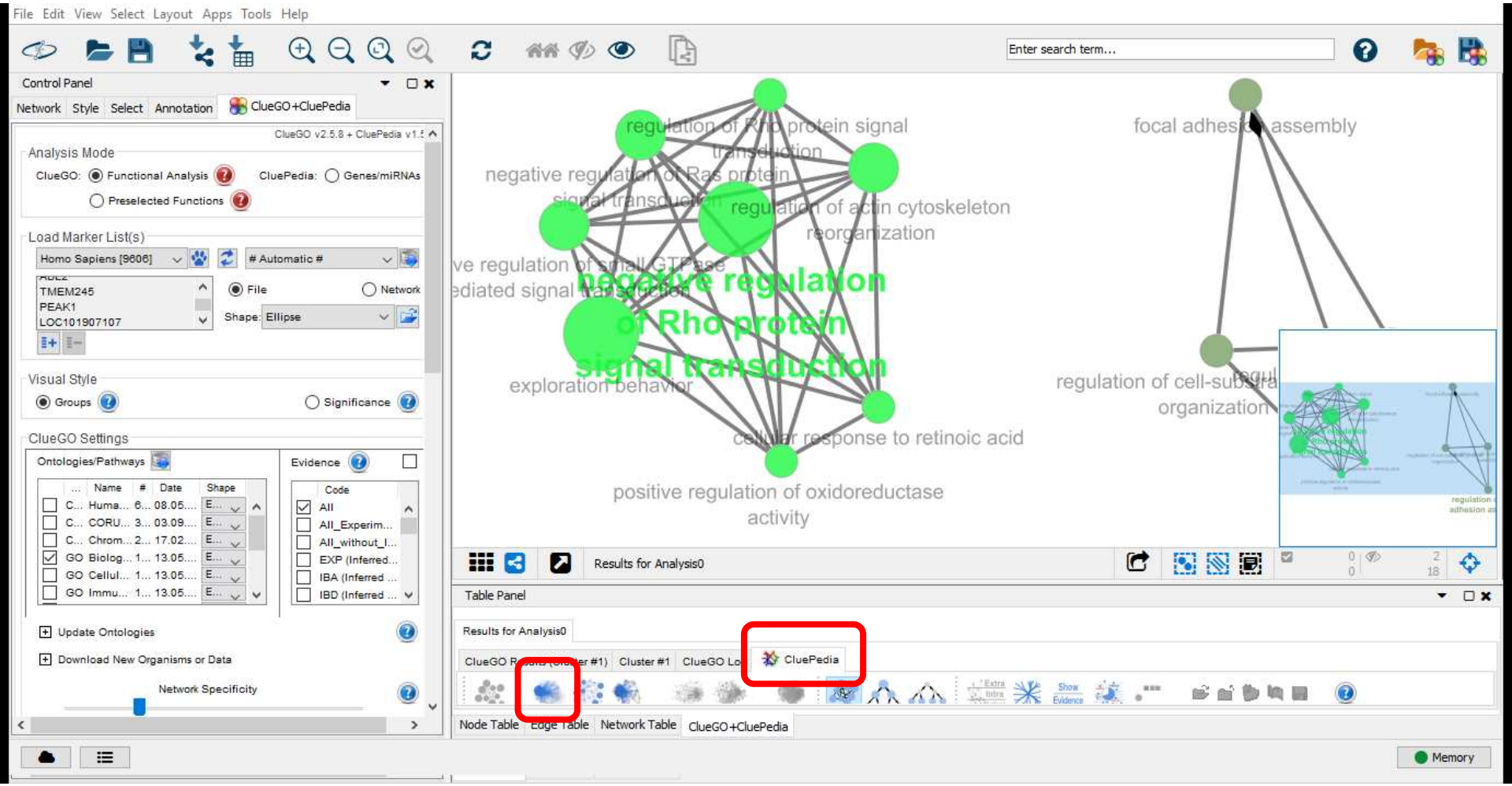
Se a análise prosseguir normalmente, vai mostrar a descrição dos resultados



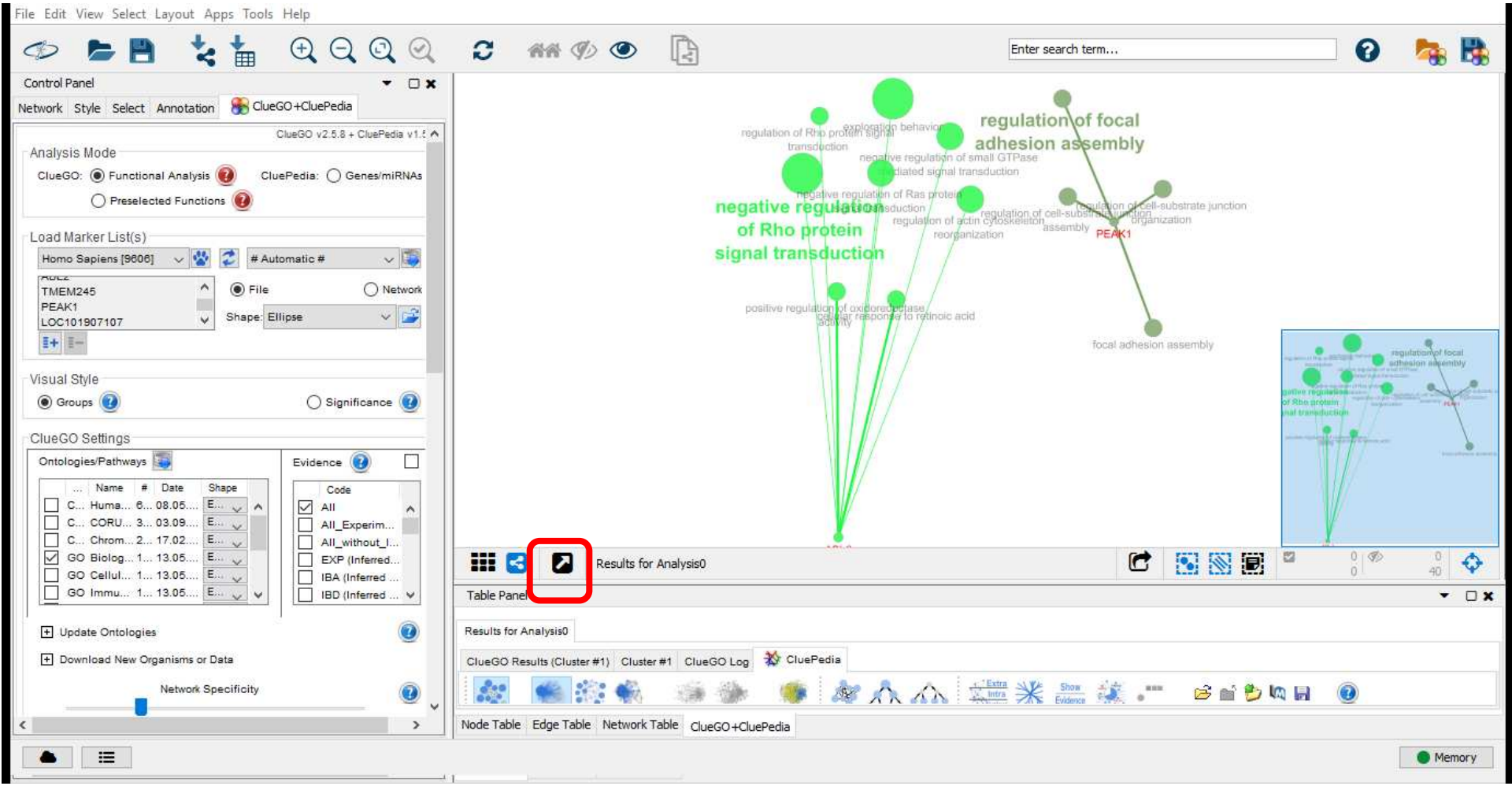
Nesta etapa o Cytoscape também indica quantos dos genes usados foram reconhecidos.
Geralmente microRNAs, RNAs e outras regiões não codificadoras como U2, U6, LOCs, etc não são identificadas pelo programa.

Na parte inferior:

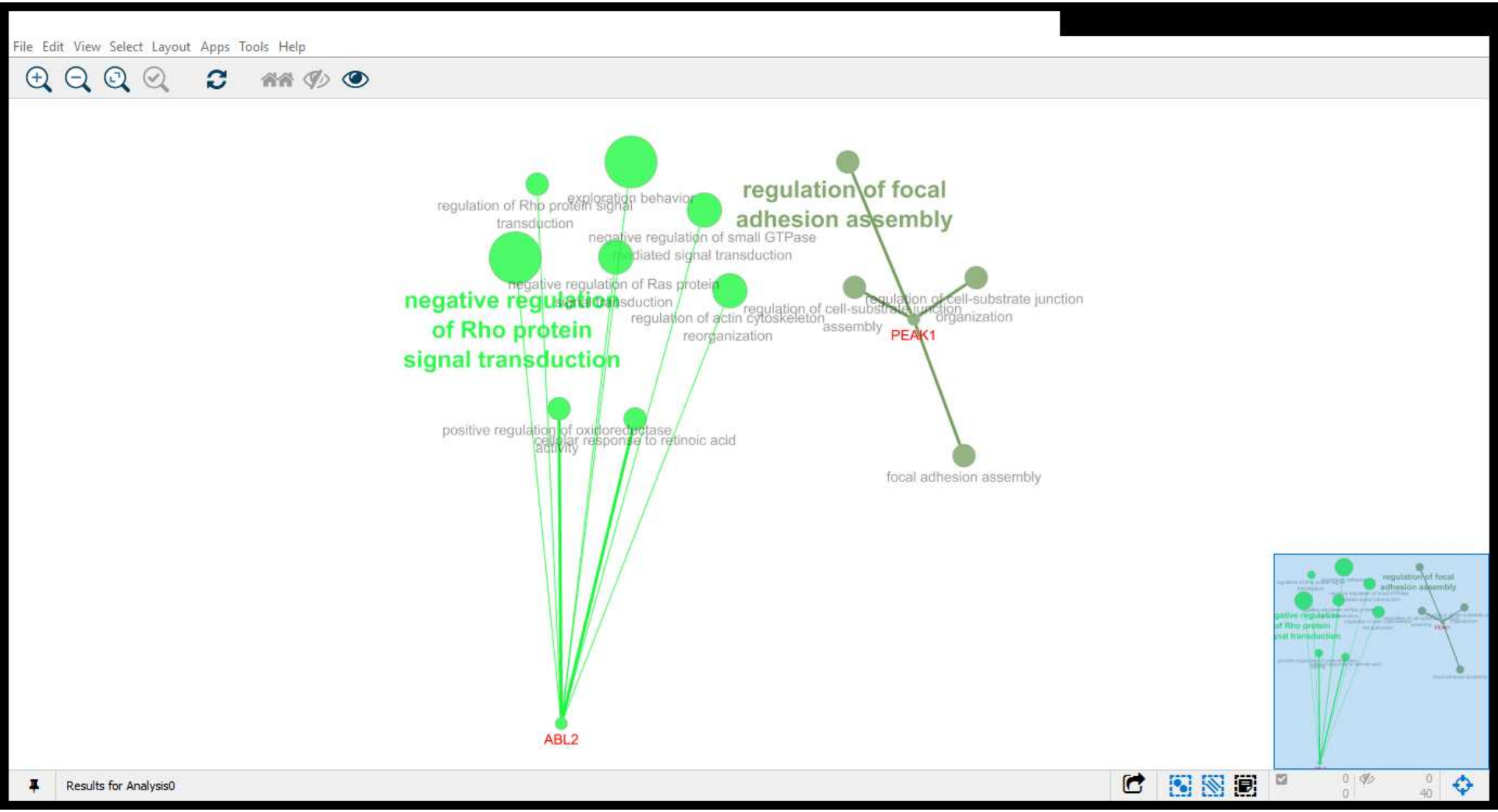
- Clicar em CluePedia
- Clicar na 2ª opção de quadrinho: ‘Show all genes from all Pathways/Terms’



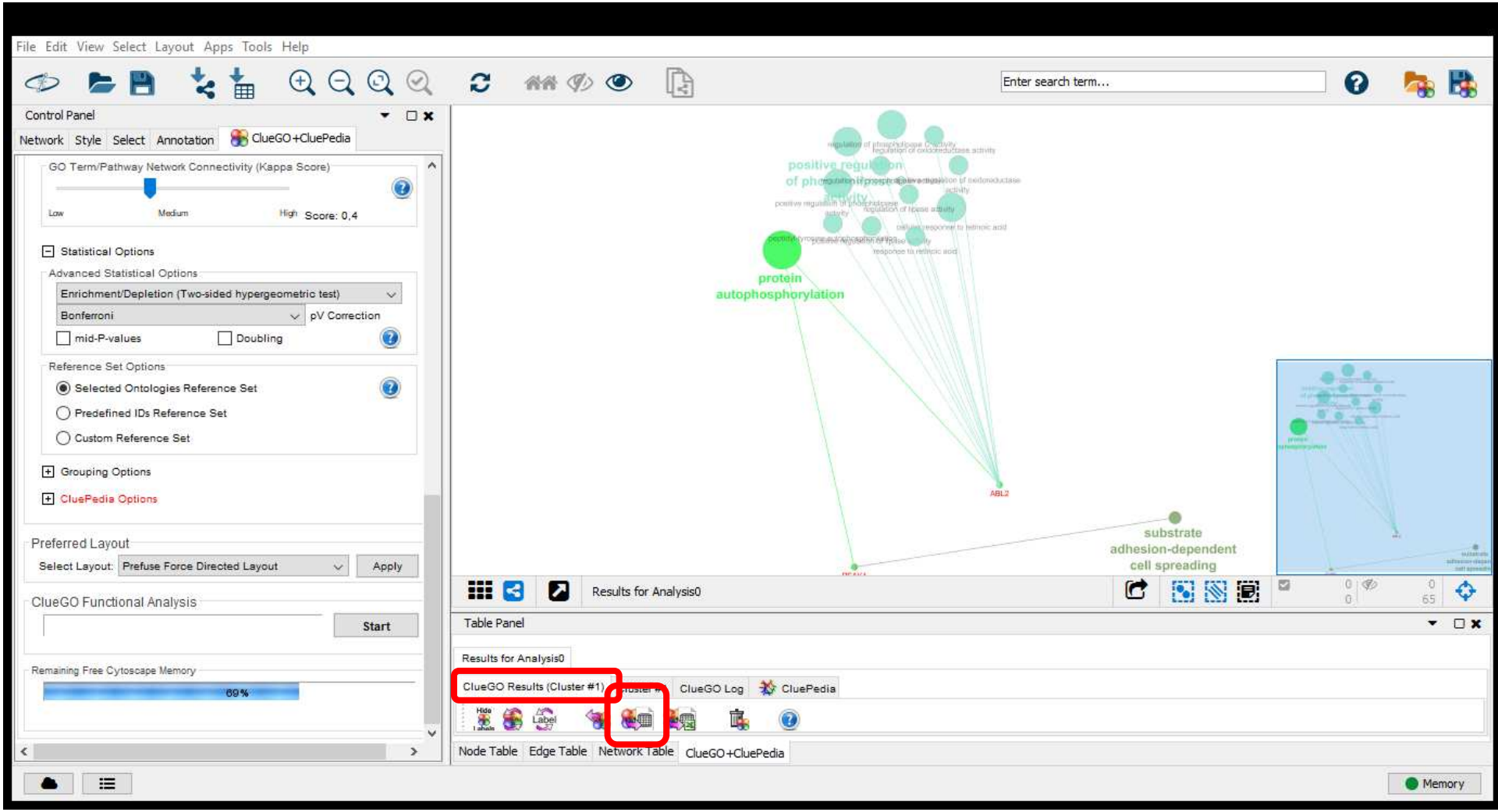
Clicar na setinha (Detach View) na parte inferior para expandir a tela das interação



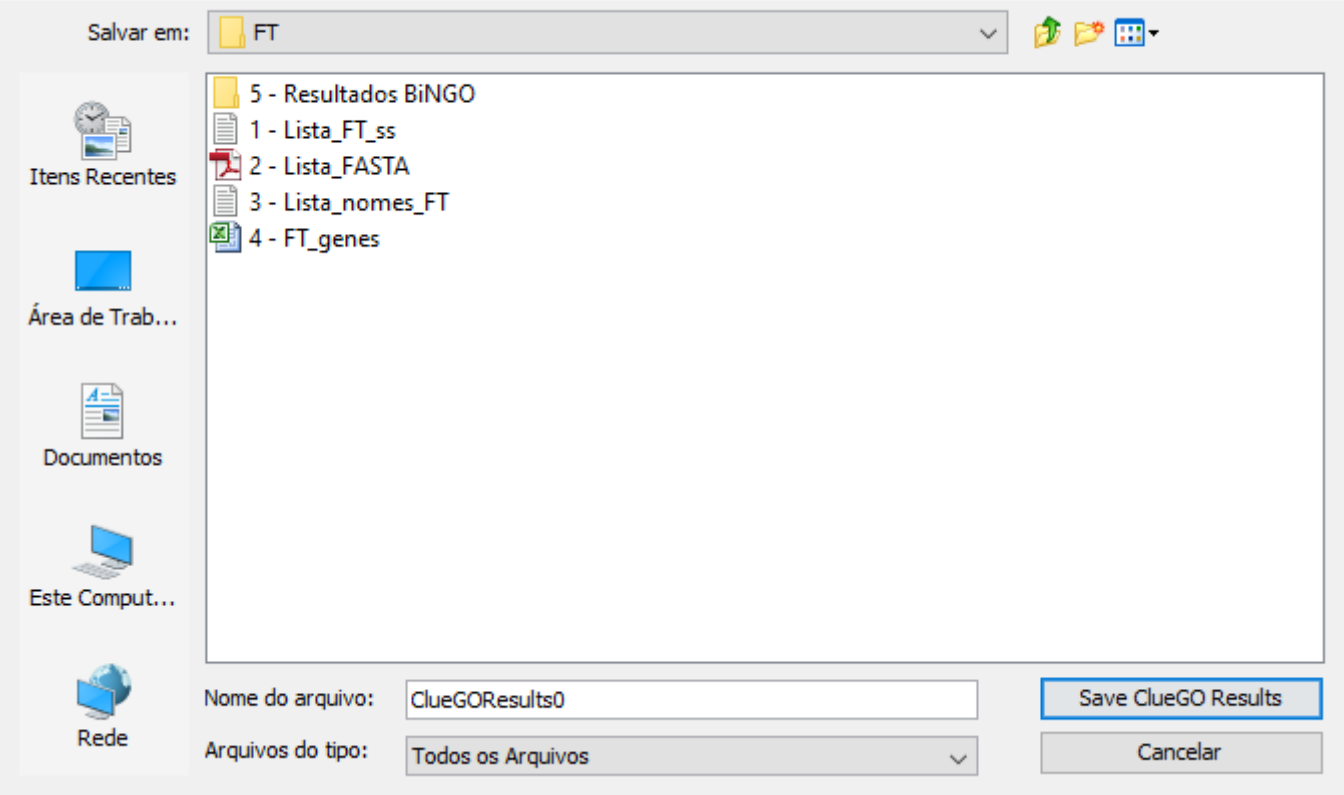
Nesta tela dá para dar zoom nos processos biológicos



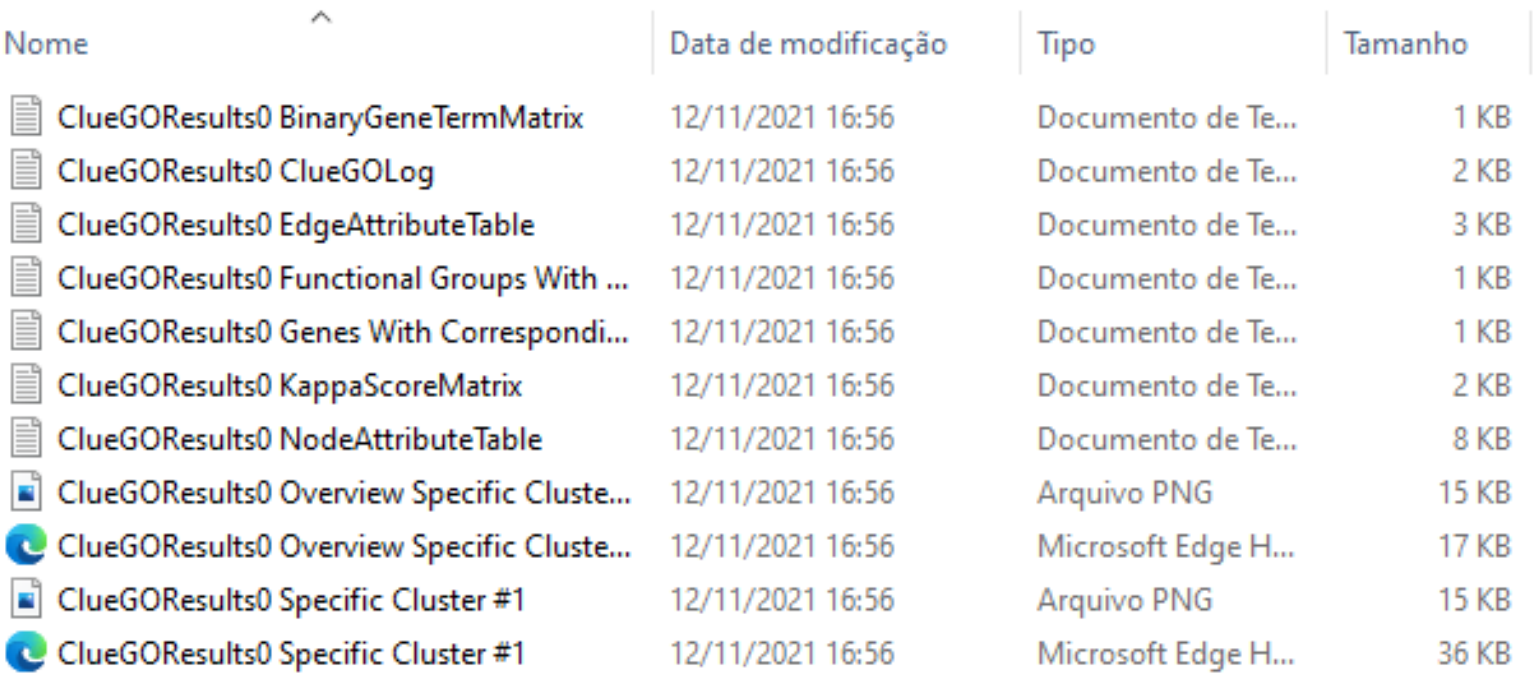
- Clicar em ClueGO Results (Cluster #1)
- Clicar no quadrinho ‘Save ClueGO Result Tables’



Direcionar os resultados para a pasta



Os resultados / output na pasta



Nome	Data de modificação	Tipo	Tamanho
ClueGOResults0 BinaryGeneTermMatrix	12/11/2021 16:56	Documento de Te...	1 KB
ClueGOResults0 ClueGOLog	12/11/2021 16:56	Documento de Te...	2 KB
ClueGOResults0 EdgeAttributeTable	12/11/2021 16:56	Documento de Te...	3 KB
ClueGOResults0 Functional Groups With ...	12/11/2021 16:56	Documento de Te...	1 KB
ClueGOResults0 Genes With Correspondi...	12/11/2021 16:56	Documento de Te...	1 KB
ClueGOResults0 KappaScoreMatrix	12/11/2021 16:56	Documento de Te...	2 KB
ClueGOResults0 NodeAttributeTable	12/11/2021 16:56	Documento de Te...	8 KB
ClueGOResults0 Overview Specific Cluste...	12/11/2021 16:56	Arquivo PNG	15 KB
ClueGOResults0 Overview Specific Cluste...	12/11/2021 16:56	Microsoft Edge H...	17 KB
ClueGOResults0 Specific Cluster #1	12/11/2021 16:56	Arquivo PNG	15 KB
ClueGOResults0 Specific Cluster #1	12/11/2021 16:56	Microsoft Edge H...	36 KB

Os arquivos interessantes:

- ClueGOResults0 BinaryGeneTermMatrix
- ➢ Esse arquivo mostra os processos biológicos (PB) nas linhas e os genes nas colunas e monta uma matriz associando os genes com cada PB

- ClueGOResults0 Functional Groups With Genes
- ➢ Esse arquivo tem os resultados dos processos biológicos agrupados por grupos funcionais de genes

- ClueGOResults0 NodeAttributeTable
- ➢ Este arquivo tem os PB e o P-valor e P-valor ajustado pelo teste de Bonferroni indicando quais genes e PB são significativos e quais não são

ClueGOResults0 NodeAttributeTable

- Este arquivo contém os gene, os processos biológicos, o P-valor, o P-valor ajustado para Bonferroni e outras informações...

- Para selecionar os genes geralmente usa-se o p-valor ajustado que chamamos de FDR pra evitar falso positivo
- Isso é usado quando existem muitas variáveis, nesse caso, os genes
- Por mais que as análises de redes tenham menos genes, é interessante padronizar

- “The Bonferroni correction adjusts probability (p) values because of the increased risk of a type I error when making multiple statistical tests.” (Armstrong, 2014. DOI: 10.1111/opo.12131)

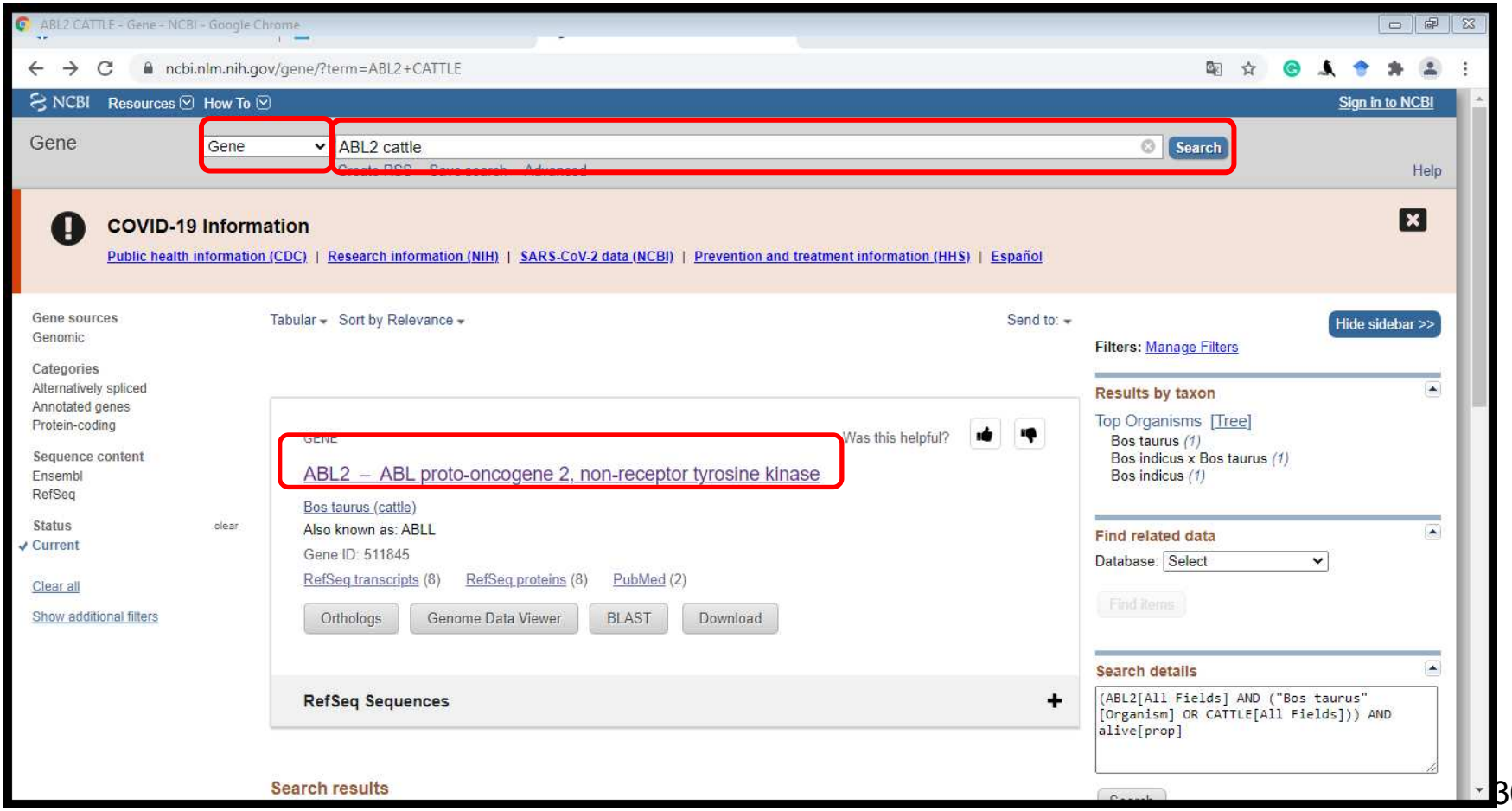
Busca das sequências promotoras no site NCBI

Busca dos fatores de transcrição no site TFM-Explorer

FATORES DE TRANSCRIÇÃO

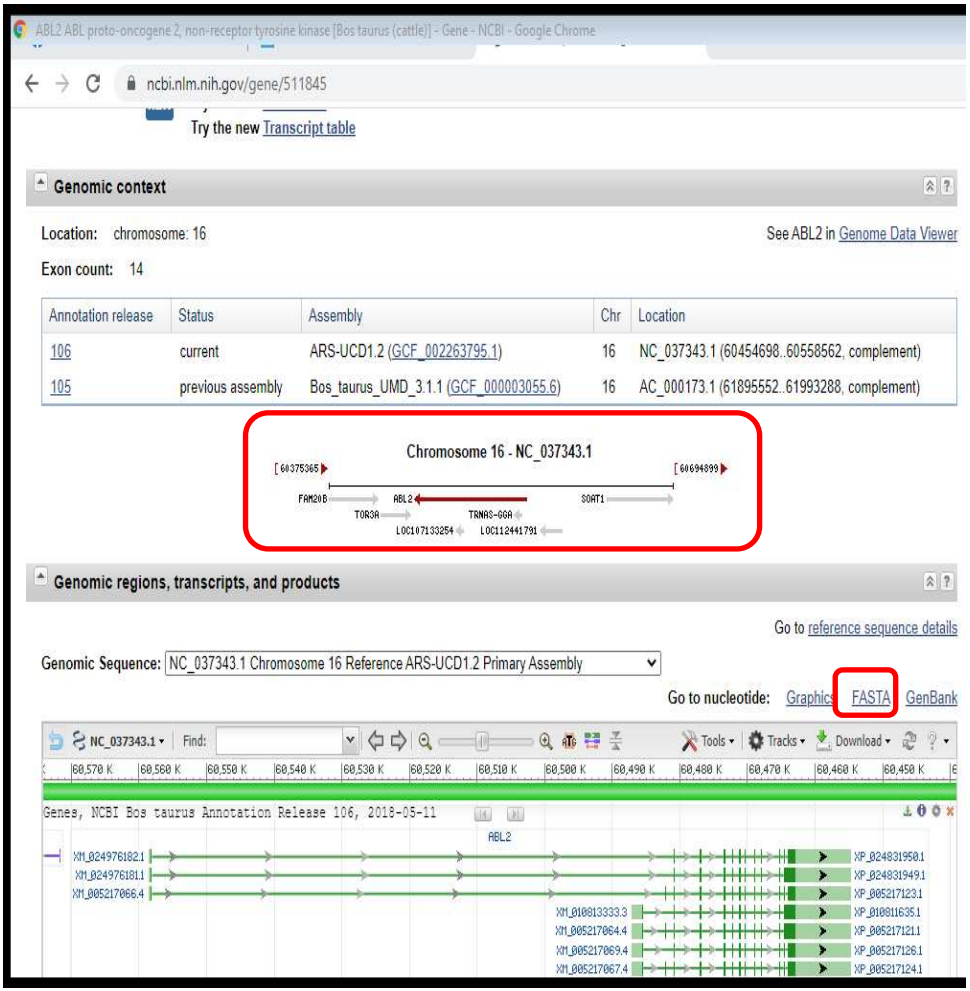
1º passo

- No site NCBI: https://www.ncbi.nlm.nih.gov/
- Pesquisar o nome do gene no NCBI para a espécie estudada
- Clicar na opção do gene identificada para a espécie



2º passo

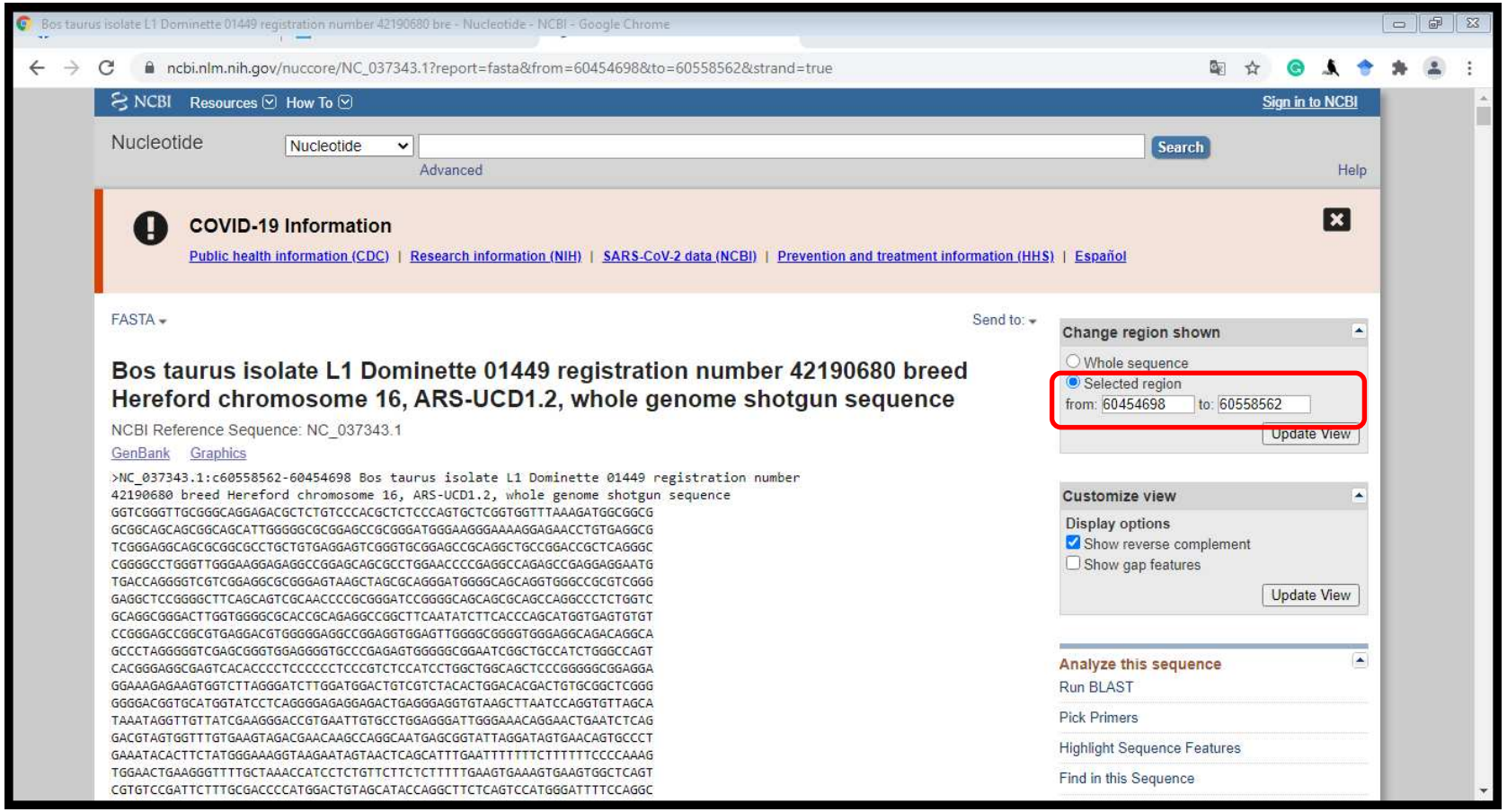
Vai carregar uma nova página



- Verificar se a sequência do gene é Foward ou Reverse – isso é importante para definir a posição inicial do gene
- A sequência é Foward se a setinha estiver seguindo da esquerda para a direita
- A sequência é Reverse se a setinha estiver seguindo da direita para a esquerda (imagem ao lado)
- Clicar em FASTA

3º passo

- Na nova página, verificar as posições iniciais dos genes nos quadrinhos “from” e “to” no canto direito
- ➢ Se o gene é Foward, a posição inicial dele está no quadrinho “from”, significa que o gene começa nesta posição
- ➢ Se o gene é Reverse, a posição inicial dele está no quadrinho “to”, significa que o gene começa nesta posição

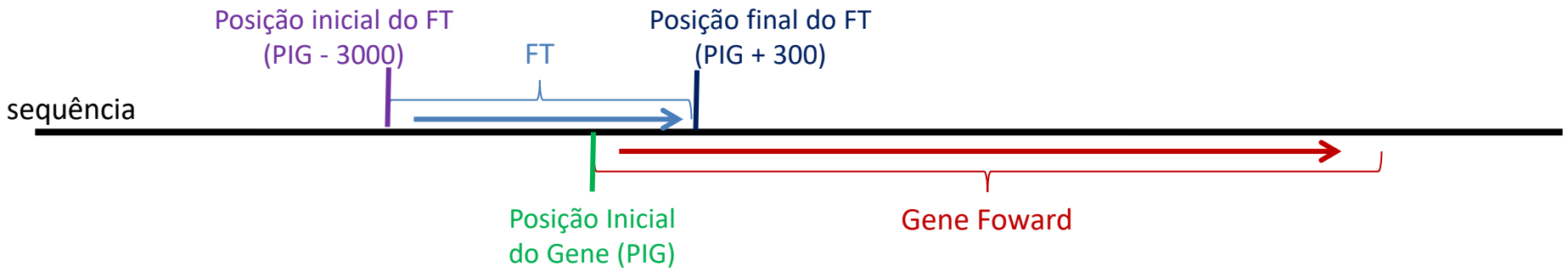


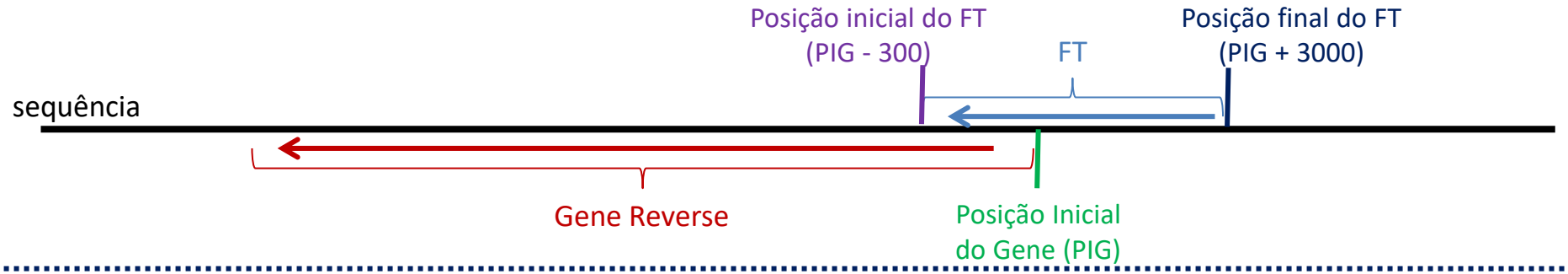
4º passo

Anotar/digitar as posições iniciais de cada gene e salvar em um arquivo (por precaução)

Para as análises seguintes, precisaremos do fator de transcrição (FT) de cada gene.
Como os FT se situam em uma posição anterior ao início do genes, precisamos calcular as posições inicias e finais dos fatores de transcrição de cada gene

Para genes Foward: assumimos que o FT começa 3000 pb antes e termina 300pb depois
Para genes Reverse: assumimos que o FT começa 3000 pb depois e termina 300pb antes





Bp: base pairs – pares de bases; O tamanho do FT entre 3000 e 300bp pode ser diferente em outros trabalhos

4º passo

- Na tabela do Excel abaixo é possível calcular as posições iniciais e finais dos fatores de transcrição

- Isso também pode ser feito na calculadora ou criar uma planilha no Excel à parte

- Na planilha abaixo, clicar duas vezes na tabela
- Pegar o valor da posição inicial do gene no 3º passo (por isso é importante saber se é Foward ou Reverse)
- Colar na cédula verde (abaixo da cédula com título 'Posição inicial do GENE')
- Dar ENTER
- Ao lado vão aparecer as posições iniciais e finais do FT

Para genes Foward		
Posição incial do GENE	**Posição inicial do FT**	**Posição final do FT**
0	-3000	300

Para genes Reverse		
Posição incial do GENE	**Posição inicial do FT**	**Posição final do FT**
60558562	60558262	60561562

Na tabela, temos o exemplo do gene ABL2

- É um gene Reverse
- Sua posição inicial é 60558562
- Assim

✓ A posição inicial do FT é 60558262

✓ A posição final do FT é 60561562

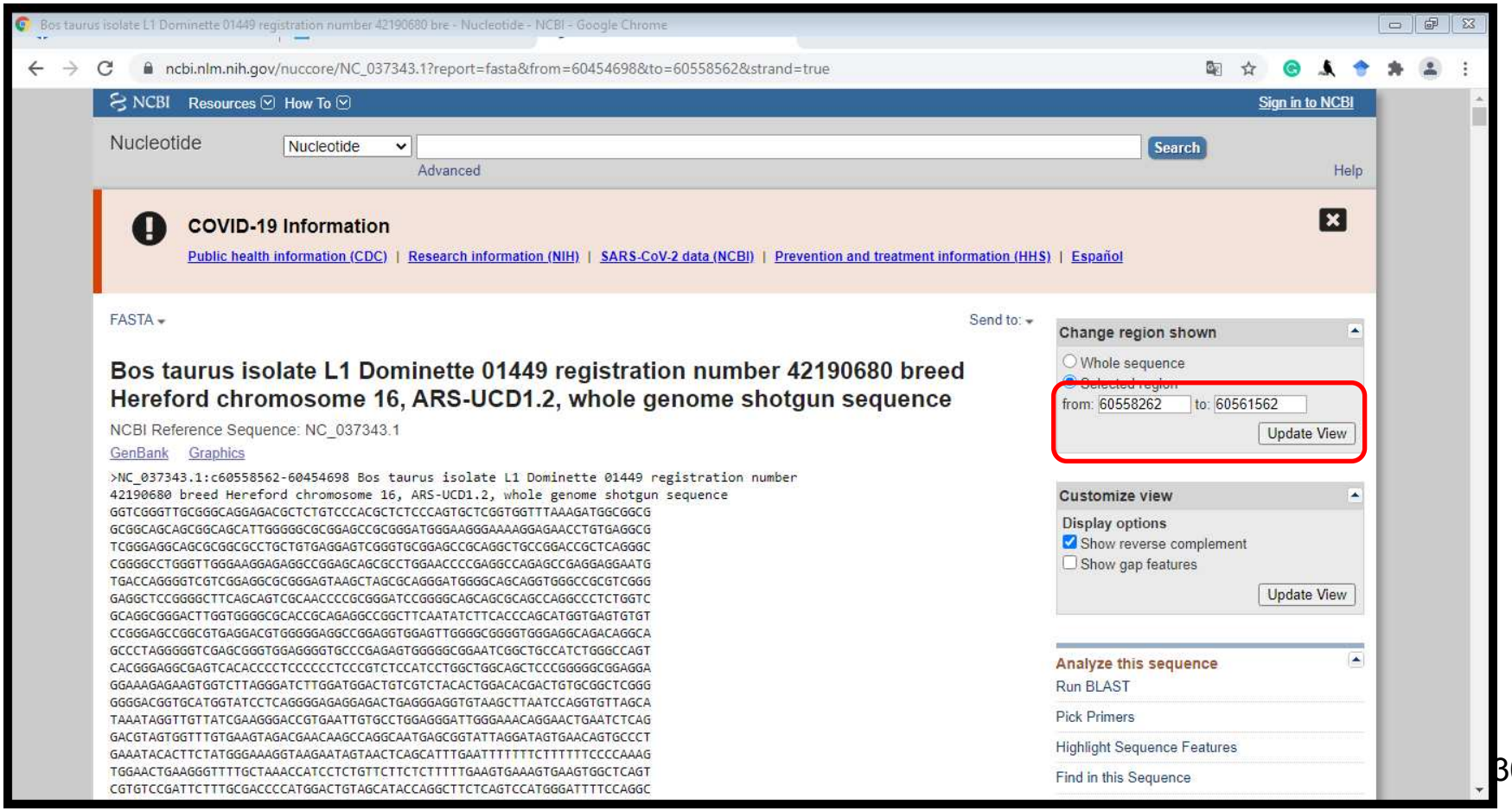
- É bom deixar essas informações anotadas em uma planilha

- É importante salvar essas informações, principalmente considerando que este é um exemplo pequeno com poucos genes, mas podem ser identificados mais de 40, 60, 80 genes, etc..

Gene	Sentido	Pos. Original	Nova pos. inicial	Nova pos final
ABL2	Reverse	60558562	60558262	60561562
TMEM245	...	...	...	...
LOC101907107	...	...	...	...
PEAK1	...	...	...	...

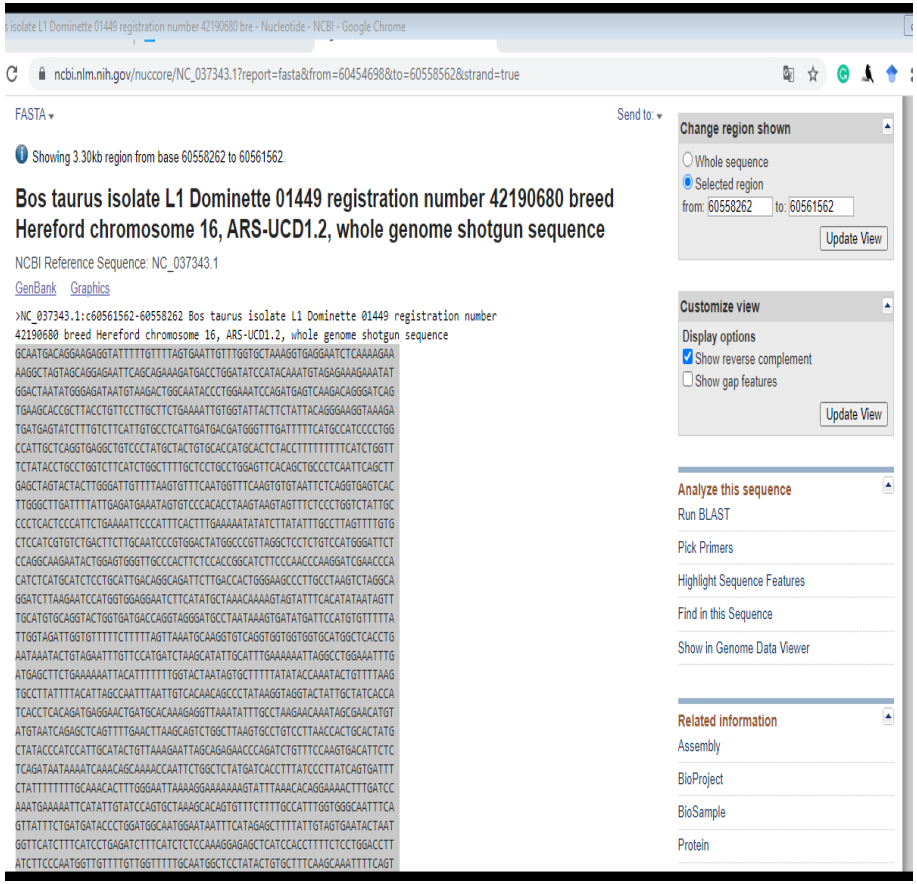
5º passo

- Voltar ao site do NCBI
- Atualizar as posições iniciais e finais nos quadrinhos
- A posição inicial vai no quadrinho ‘from’ e a posição final no ‘to’ (independente se é Foward ou Reverse, porque o calculo acima já estabelece o início e fim da sequência)
- clicar em Update View

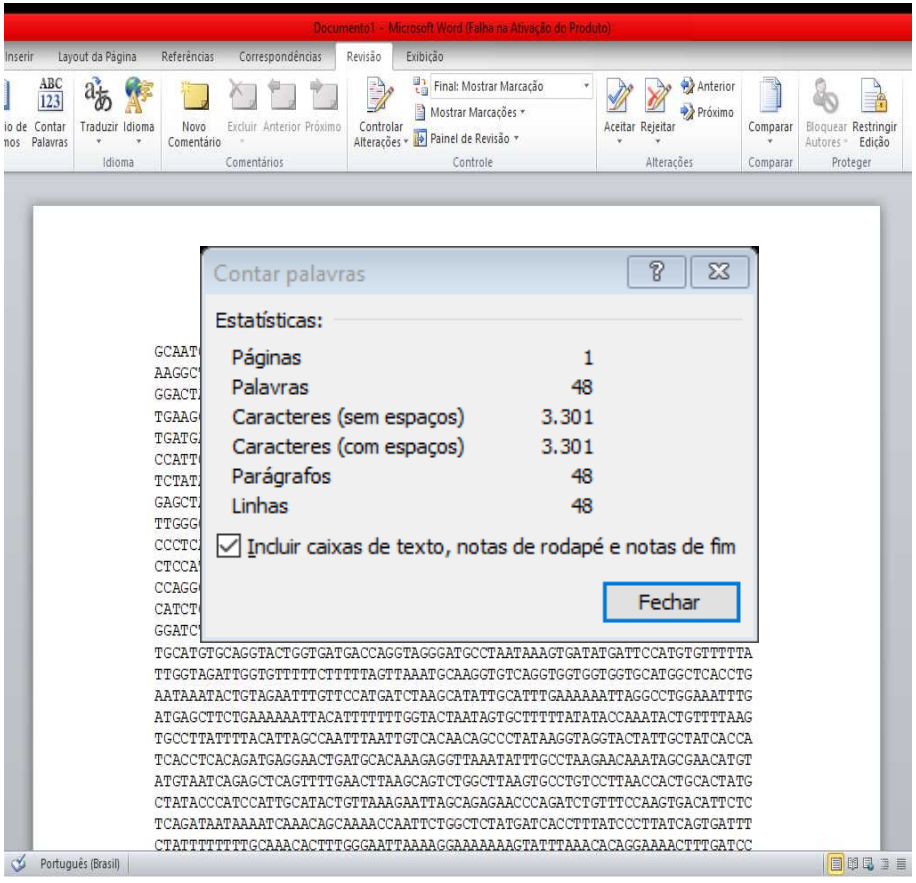


6º passo

- O site vai gerar uma nova sequência com 3.301 bases
- Esta é a sequência do FT
- Copiar a sequência da primeira até última base



- Colar a sequência em um arquivo no word
- Revisão -> contar Palavras
- Verificar se os caracteres dão 3.301



Colar sequências no bloco de notas

7º passo

O bloco de notas ficará assim:

>Nomedogene1|hg19

ACT..... (toda a sequência)

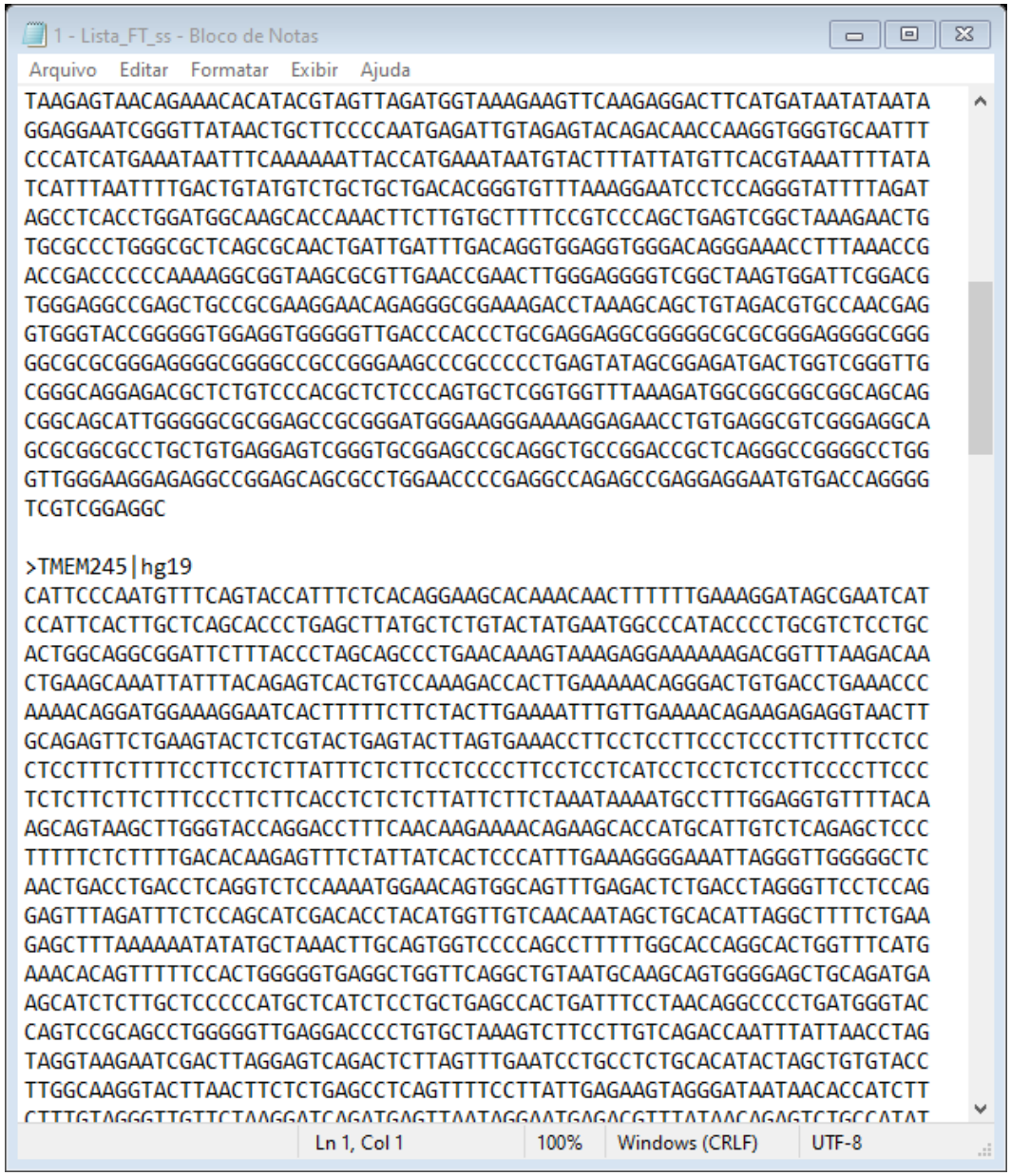
>Nomedogene2|hg19

ACT..... (toda a sequência)

Sempre lembrar de deixar uma linha em branco de intervalo entre o final de uma sequência e o início do nome do gene seguinte (exemplo imagem ao lado)

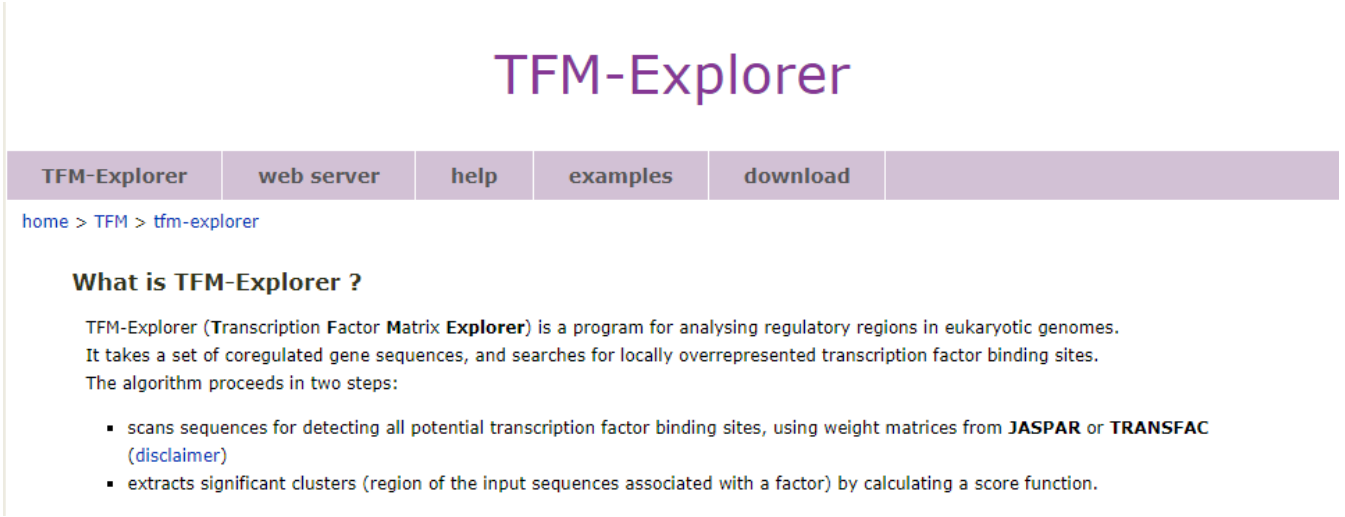
- As sequências são formadas com a s letras A, C, T e G.
- Quando tiver um “N” na sequência é *missing* (a base nitrogenada daquela posição não foi identificada)
- Tem que substituir o “N” por “.”

*Se algum gene tiver o nome com a letra N, tomar cuidado para não substituir/alterar o nome dos genes



- No fim, salvar a planilha com essas informações, por precaução

Gene	Sentido	Pos. inicial Original	Nova pos. inicial	Nova pos final
ABL2	Reverse	60558562	60558262	60561562
TMEM245	Reverse	98686213	98685913	98689213
LOC101907107	Reverse	61982951	61982651	61985951
PEAK1	Reverse	32554681	32554381	32557681

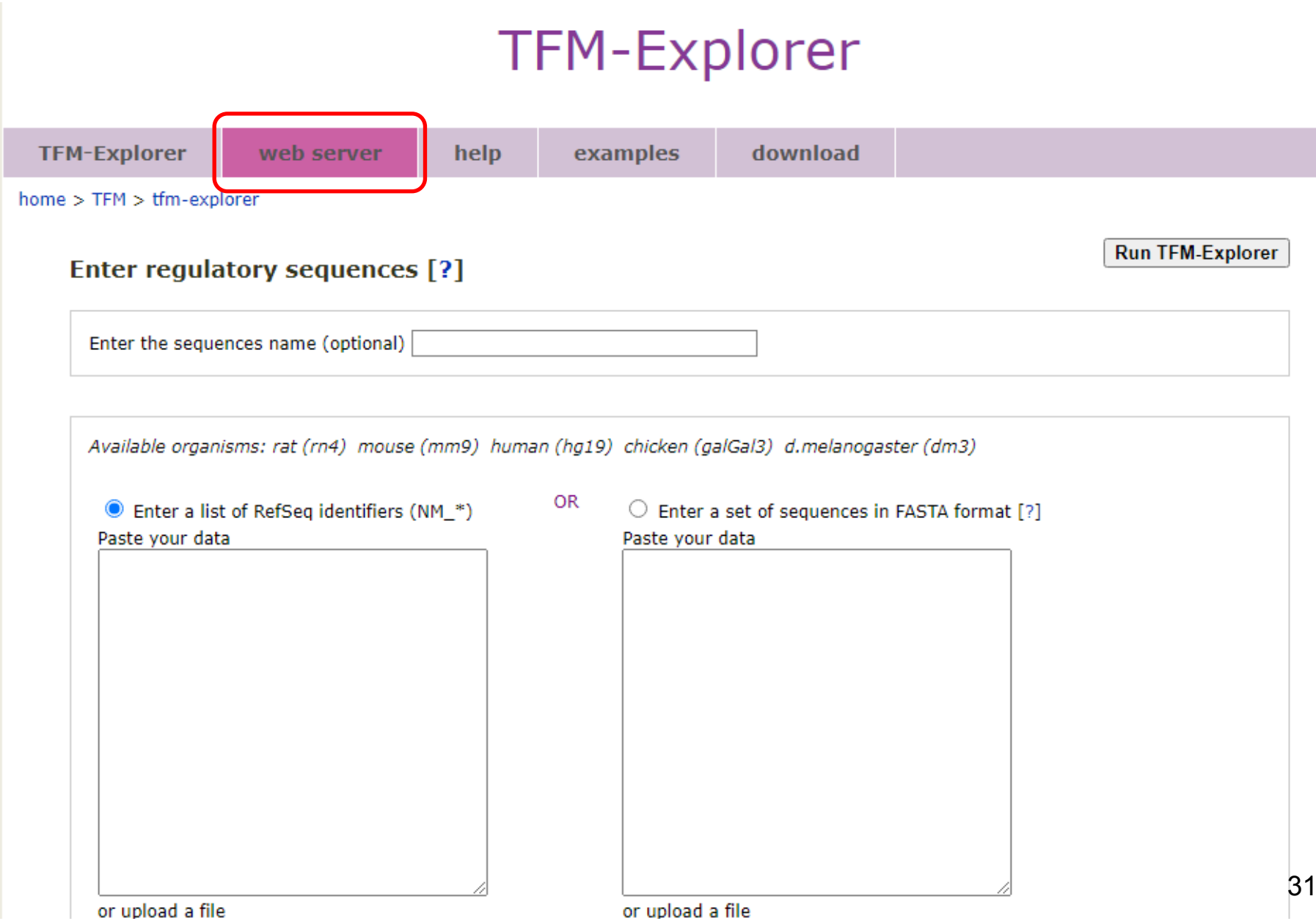


TFM EXPLORER – WEB BROWSER

https://bioinfo.lifl.fr/tfm-explorer/tfm-explorer.php

PESQUISAR OS NOMES DOS FATORES DE TRANSCRIÇÃO – 8º PASSO

No site TFM – Explorer, clicar em web server



Enter the sequences name (optional)

Deixar em branco

Available organisms: rat (rn4) mouse (mm9) human (hg19) chicken (galGal3) d.melanogaster (dm3)

Enter a list of RefSeq identifiers (NM_*)

Paste your data

or upload a file

Escolher arquivo Nenhum arquivo selecionado

OR

Enter a set of sequences in FASTA format [?]

Paste your data

or upload a file

Escolher arquivo 1 - Lista_FT_ss.txt

Anexar arquivo .txt montado na etapa anterior

Example

Deixar -3000 e +300

Location [?] -3000 +300

(boundaries of the sequences in regard of the TSS. authorized values : -10000 to 5000)

Select transcription factor binding profiles [?]

Available weight matrices are public TRANSFAC (6.0) matrices and JASPAR 2009 vertebrate matrices

- ○ Use public TRANSFAC matrices
  - only vertebrates
  - only insects
  - all TRANSFAC
- ◉ Use all JASPAR vertebrate matrices
- ○ Use all available TRANSFAC and JASPAR matrices
- ○ Select weight matrices in the list below
  - B$CRP_C
  - F$ABAA_01
  - F$ABF_C
  - F$ABF1_01
  - F$ADR1_01
- ○ Upload a file containing a list of weight matrice identifiers [?]

  Escolher arquivo | Nenhum arquivo selecionado

Não alterar nada nesta parte

Adjust parameters [?]

Number of clusters to display [?] (max 25)	25
Maximum P-value [?]	0.0001
Ratio (density of clusters) [?]	3

Diminuir o máximo P- value só se tiver muitos (?) FT
Se não, deixar como está

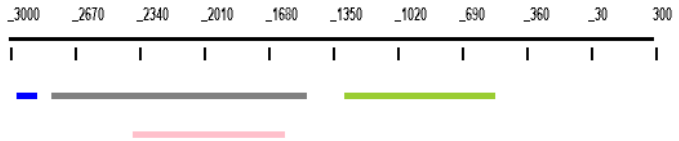
Enter your E-mail address *(optional)*:

Reset | Run TFM-Explorer

You request has been successfully submitted to **TFM-EXPLORER**
Your ID is **Nov_02_2021_00_03_57_9363**

Results for job Nov_02_2021_00_03_57_9363 [?]

4 clusters found. The figure below represents the clusters on the input sequences, each of them being described in the table. Click on a cluster for detailed information.

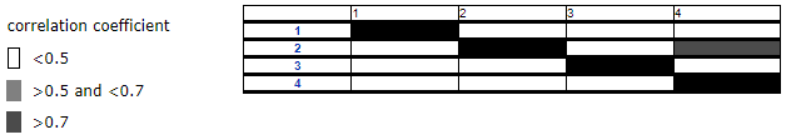


Click on a line to get more detailed information for a cluster [?].

Rank	Factor	Matrix ID	Location	Sequences	P-Value	Correlated with	Legend	Select
1	Esrrb	MA0141.1	+[-2957:-2851]	2	3.39e-05			
2	FEV	MA0156.1	+[-2360:-1575]	3	4.67e-05	4		
3	FOXO3	MA0157.1	+[-2779:-1467]	4	4.94e-05			
4	ELK4	MA0076.1	+[-1283:-0508]	3	8.90e-05	2		

View transcription factor binding sites associated to selected clusters [?] view

Summary of pairwise correlations between clusters [?]



Download results [?] : complete result files Nov_02_2021_00_03_57_9363.zip
summary in text format Nov_02_2021_00_03_57_9363.txt

Sequences Name	None
Number of sequences	4
Number of matrices	130
Database used	JASPAR
Region	-3000 : 300
Date	Tue Nov 2 00:04:23 2021
Minimum window size	30
Maximum window size	1500
Maximum number of windows to show	25
Minimal P-value	0.0001
Ratio	3.0
URL of results	http://bioinfo.lifl.fr/cgi-bin/tfm-explorer/tfme.py?command=result&run_id=Nov_02_2021_00_03_57_9363

Na nova página que carrega, tem a opção:
Download results [?]: ....
Clicar na opção gerada em 'summary in text format'

Vai carregar uma nova guia com os resultados (próximo slide)

bioinfo.lifl.fr/tfm-explorer/result/Nov_02_2021_00_03_57_9363/Nov_02_2021_00_03_57_9363.txt

```
;TFM Explorer
;Date              : Tue Nov  2 00:04:23 2021
;Scanned location  : -3000:300
;Scanned sequences : None (4 sequences)
;Scanned matrices  : (130 matrices)
;Parameters        : minsize= 30, maxsize= 1500, ratio=3.0, top=25
;
; headers
;    1    rank              window rank
;    2    matrix            matrix name
;    3    tf                transcription factor name
;    4    info              information of content of the matrix
;    5    gc                gc content of the matrix (G+C)%
;    6    location          location of the window (relatively to TSS)
;    7    pvalue            pvalue of the window

1   MA0141.1        Esrrb   12.8061908542   0.523864734014  -2957:-2851     3.38522472321e-05
          list of hits
          ABL2      -2957   -2945   +       6.40    TGCTAAAGGTGA
          ABL2      -2898   -2886   -       5.74    TGACCTGGATAT
          LOC101907107      -2952   -2940   -       5.96    TGACCCTGTCCT
          LOC101907107      -2918   -2906   -       6.49    TGACCCTGCCCT
          LOC101907107      -2866   -2854   -       5.96    TGACCCTGTCCT
          LOC101907107      -2851   -2839   -       7.50    TGACCCTGGCCT

2   MA0156.1        FEV     12.1208832256   0.442307695746  -2360:-1575     4.67270375297e-05
          list of hits
          TMEM245 -2323     -2315   +       5.26    GGGGAAAT
          TMEM245 -1977     -1969   -       7.43    ATTTCCTA
          LOC101907107      -2360   -2352   -       7.02    ATTTCCGT
          LOC101907107      -2274   -2266   +       7.96    AAGGAAGT
          LOC101907107      -2096   -2088   +       6.34    TCGGAAAT
          LOC101907107      -2052   -2044   -       7.43    ATTTCCTC
          LOC101907107      -1602   -1594   -       5.11    ACTTCCCA
          PEAK1     -1991   -1983   -       5.26    ATTTCCCC
          PEAK1     -1798   -1790   -       5.26    ATTTCCCA
          PEAK1     -1765   -1757   -       7.43    ATTTCCTA
          PEAK1     -1686   -1678   -       7.28    ACTTCCTC
          PEAK1     -1617   -1609   -       5.26    ATTTCCCC
          PEAK1     -1575   -1567   +       7.28    GAGGAAGT
```

genes

genes

Estão circulados em destaque os nomes dos fatores de transcrição

Para cada FT tem a lista de genes e as informações associadas à cada um.

O nome de cada FT e dos genes associados à cada um serão necessários para o próximo passo.

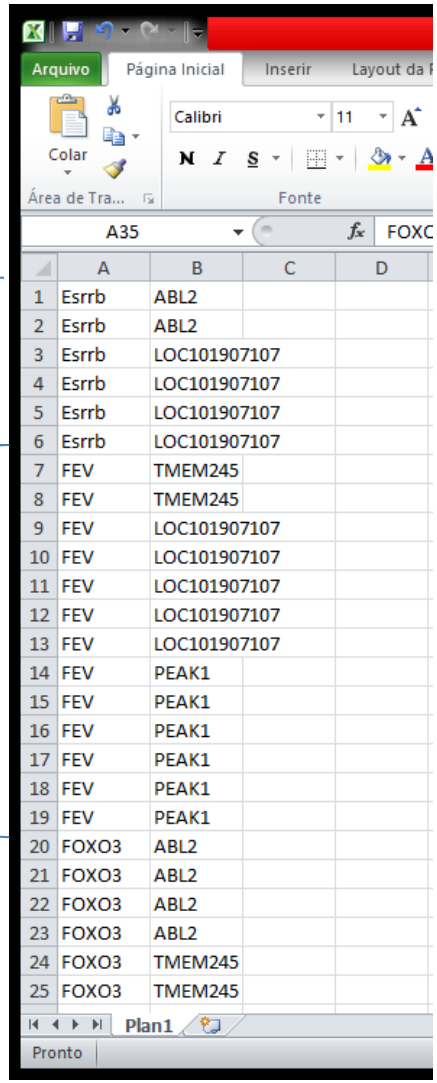
9º passo

Montar uma planilha no Excel com duas colunas:
1ª coluna: Nome do fator de transcrição (FT)
2ª coluna: Nomes dos genes associados ao FT
Não importa se tiver que repetir os nomes
- Não colocar cabeçalho na planilha

```
1   MA0141.1        Esrrb    12.8061
          list of hits
          ABL2      -2957    -2945
          ABL2      -2898    -2886
          LOC101907107  -2952
          LOC101907107  -2918
          LOC101907107  -2866
          LOC101907107  -2851

2   MA0156.1        FEV      12.1208
          list of hits
          TMEM245 -2323    -2315
          TMEM245 -1977    -1969
          LOC101907107  -2360
          LOC101907107  -2274
          LOC101907107  -2096
          LOC101907107  -2052
          LOC101907107  -1602
          PEAK1     -1991    -1983
          PEAK1     -1798    -1790
          PEAK1     -1765    -1757
          PEAK1     -1686    -1678
          PEAK1     -1617    -1609
          PEAK1     -1575    -1567
```

genes

genes



	A	B	C	D
1	Esrrb	ABL2		
2	Esrrb	ABL2		
3	Esrrb	LOC101907107		
4	Esrrb	LOC101907107		
5	Esrrb	LOC101907107		
6	Esrrb	LOC101907107		
7	FEV	TMEM245		
8	FEV	TMEM245		
9	FEV	LOC101907107		
10	FEV	LOC101907107		
11	FEV	LOC101907107		
12	FEV	LOC101907107		
13	FEV	LOC101907107		
14	FEV	PEAK1		
15	FEV	PEAK1		
16	FEV	PEAK1		
17	FEV	PEAK1		
18	FEV	PEAK1		
19	FEV	PEAK1		
20	FOXO3	ABL2		
21	FOXO3	ABL2		
22	FOXO3	ABL2		
23	FOXO3	ABL2		
24	FOXO3	TMEM245		
25	FOXO3	TMEM245		

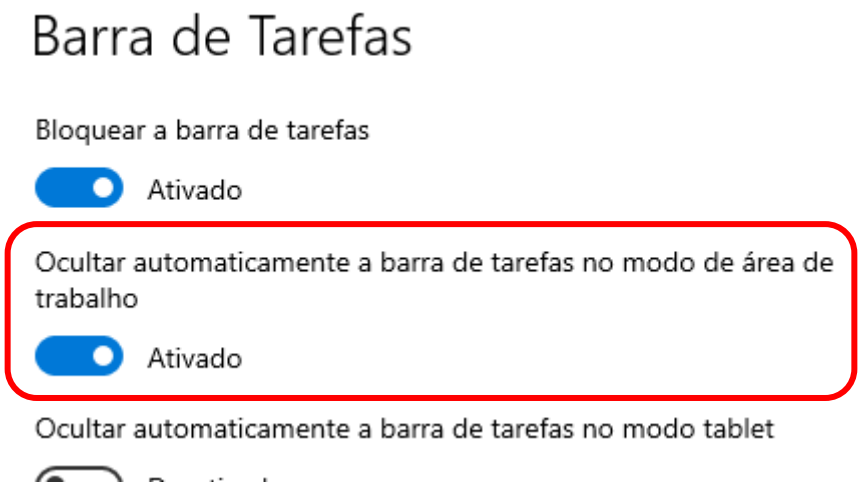
Obtendo as vias enriquecidas

BINGO

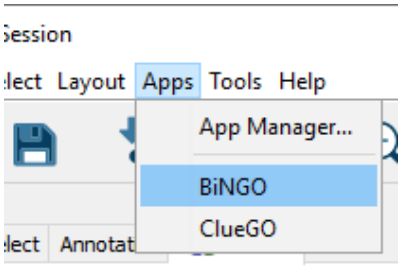
Para rodar o BiNGO, precisei ocultar a barra de tarefas do computador, ou então não daria para ver a tela toda da análise

Para ocultar a barra de tarefas:

- Clicar com o botão direito na barra de tarefas ou pesquisar
- Configurações da barra de tarefas
- Ativar o modo de ocultar a barra de tarefas da área de trabalho



BiNGO



Clicar em Apps → BiNGO

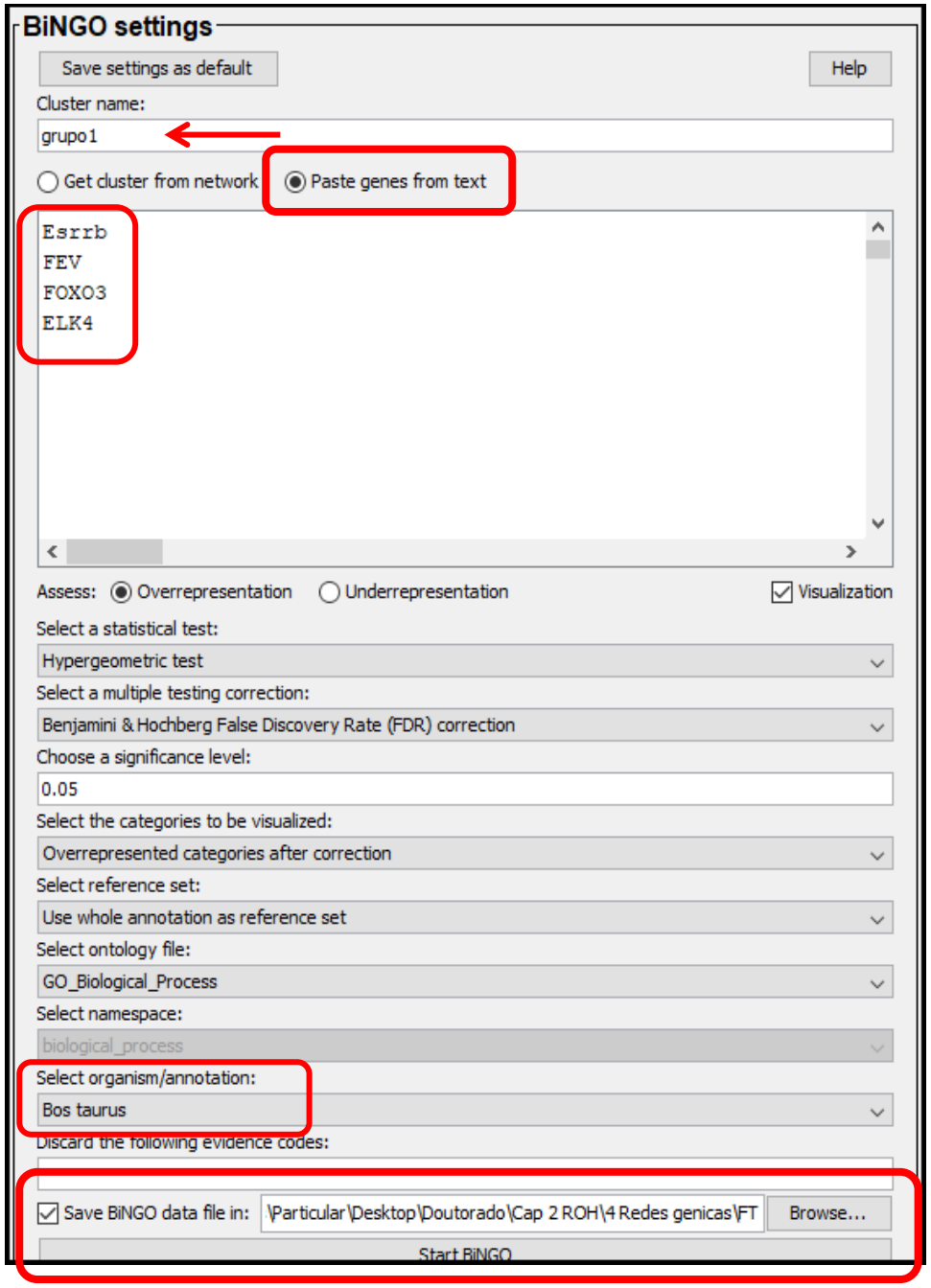
Às vezes tem que ocultar a barra de tarefas da área de trabalho pra poder visualizar toda a tela que aparece

Alterar apenas os parâmetros destacados

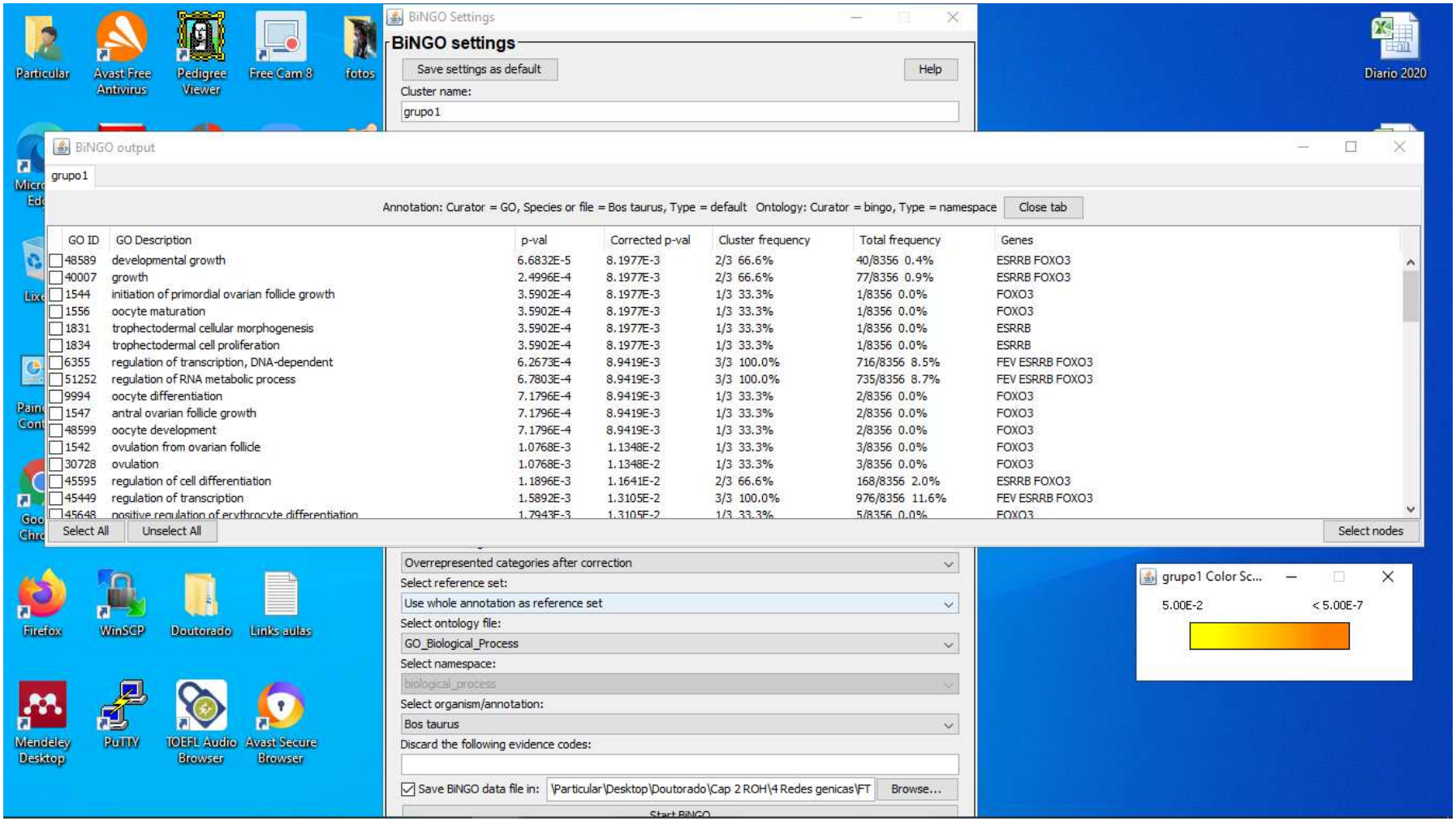
- Colocar algum nome em 'Cluster name'
- Selecionar 'Paste genes from text'
- Colocar os nomes dos FT na caixa de texto
- Em 'Select organism/annotation' colocar *Bos taurus* ou a espécie estudada

Nessa parte, se não tiver a espécie de interesse, colocar Homo Sapiens + descrição

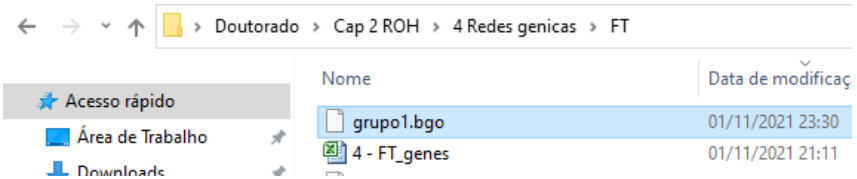
- Marcar a opção 'Save BiNGO data file in'
- Browse (pesquisar a pasta para salvar as saídas)
- Clicar em Start BiNGO



Vai aparecer a tela com os resultados dos processos biológicos



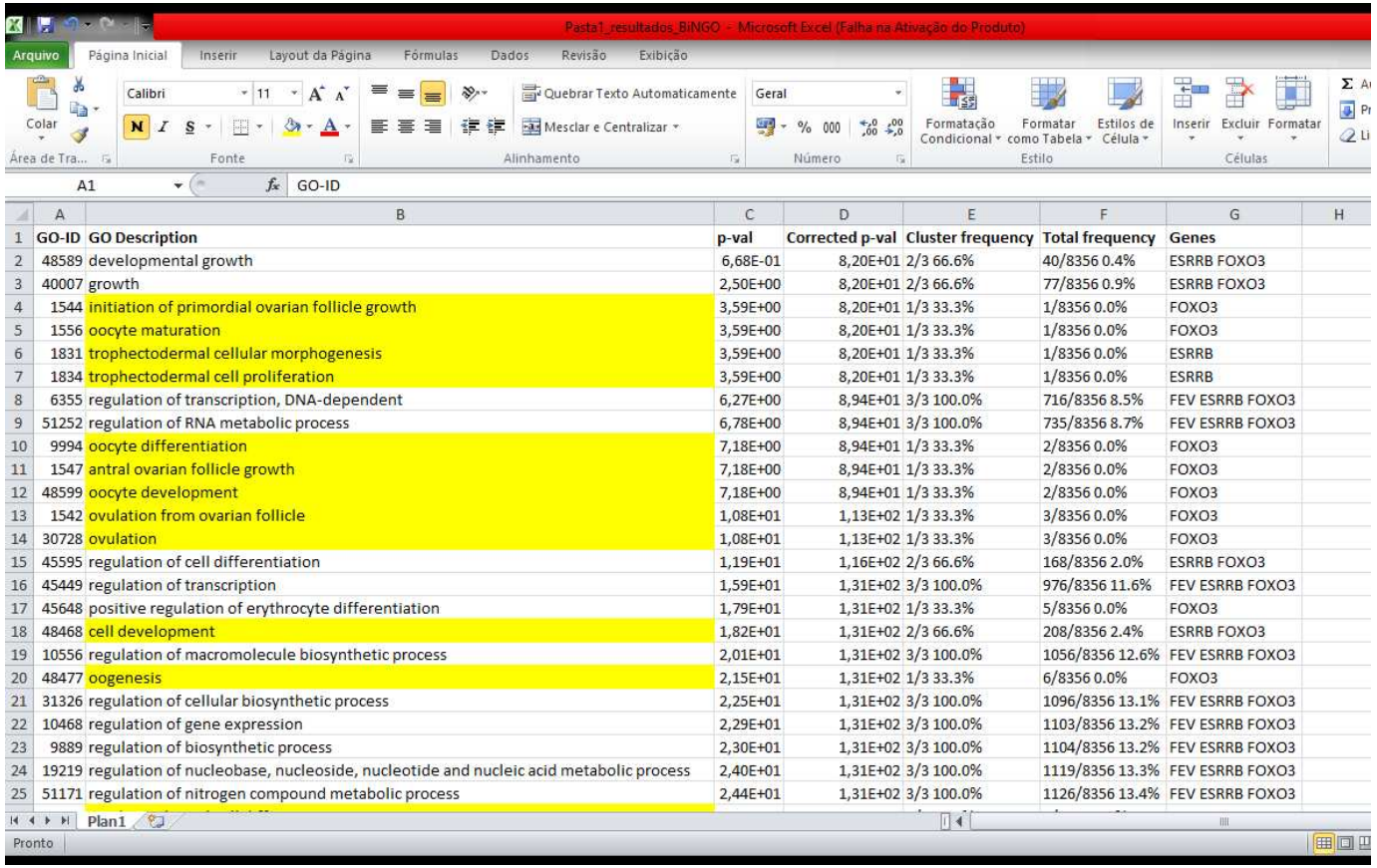
Os resultados do BiNGO serão salvos na pasta no formato .bgo



- Abrir esse arquivo com o Notepad++
- Selecionar os resultados dos processos biológicos e passar para uma planilha do Excel

```
File created with BiNGO v3.0.5  (c) on 01/11/2021 at 23:30:46

ontology: namespace
curator:  bingo

Selected ontology file : bundle://113.0:1/data/GO_Biological_Process
Selected annotation file : bundle://113.0:1/data/B_taurus_default
Discarded evidence codes :
Overrepresentation
Selected statistical test : Hypergeometric test
Selected correction : Benjamini & Hochberg False Discovery Rate (FDR) correction
Selected significance level : 0.05
Testing option : Use whole annotation as reference set
The selected cluster :
ELK4    FEV ESRRB   FOXO3

No annotations were retrieved for the following entities:
ELK4

GO-ID   p-value corr p-value    x   n   X   N   Description Genes in test set
48589   6.6832E-5   8.1977E-3   2   40  3   8356    developmental growth    ESRRB|FOXO3
40007   2.4996E-4   8.1977E-3   2   77  3   8356    growth  ESRRB|FOXO3
1544    3.5902E-4   8.1977E-3   1   1   3   8356    initiation of primordial ovarian follicle growth    FOXO3
1556    3.5902E-4   8.1977E-3   1   1   3   8356    oocyte maturation   FOXO3
1831    3.5902E-4   8.1977E-3   1   1   3   8356    trophectodermal cellular morphogenesis  ESRRB
1834    3.5902E-4   8.1977E-3   1   1   3   8356    trophectodermal cell proliferation  ESRRB
6355    6.2673E-4   8.9419E-3   3   716 3   8356    regulation of transcription, DNA-dependent  FEV|ESRRB|FOXO3
51252   6.7803E-4   8.9419E-3   3   735 3   8356    regulation of RNA metabolic process FEV|ESRRB|FOXO3
9994    7.1796E-4   8.9419E-3   1   2   3   8356    oocyte differentiation  FOXO3
1547    7.1796E-4   8.9419E-3   1   2   3   8356    antral ovarian follicle growth  FOXO3
48599   7.1796E-4   8.9419E-3   1   2   3   8356    oocyte development  FOXO3
1542    1.0768E-3   1.1348E-2   1   3   3   8356    ovulation from ovarian follicle FOXO3
30728   1.0768E-3   1.1348E-2   1   3   3   8356    ovulation   FOXO3
45595   1.1896E-3   1.1641E-2   2   168 3   8356    regulation of cell differentiation  ESRRB|FOXO3
45449   1.5892E-3   1.3105E-2   3   976 3   8356    regulation of transcription FEV|ESRRB|FOXO3
45648   1.7943E-3   1.3105E-2   1   5   3   8356    positive regulation of erythrocyte differentiation  FOXO3
48468   1.8197E-3   1.3105E-2   2   208 3   8356    cell development    ESRRB|FOXO3
10556   2.0133E-3   1.3105E-2   3   1056    3   8356    regulation of macromolecule biosynthetic process    FEV|ESRRB|FOXO3
```

Arquivo Excel



Pasta1_resultados_BiNGO - Microsoft Excel (Falha na Ativação do Produto)

Arquivo | Página Inicial | Inserir | Layout da Página | Fórmulas | Dados | Revisão | Exibição

A1 | GO-ID

	A	B	C	D	E	F	G	H
1	GO-ID	GO Description	p-val	Corrected p-val	Cluster frequency	Total frequency	Genes	
2	48589	developmental growth	6,68E-01	8,20E+01	2/3 66.6%	40/8356 0.4%	ESRRB FOXO3	
3	40007	growth	2,50E+00	8,20E+01	2/3 66.6%	77/8356 0.9%	ESRRB FOXO3	
4	1544	initiation of primordial ovarian follicle growth	3,59E+00	8,20E+01	1/3 33.3%	1/8356 0.0%	FOXO3	
5	1556	oocyte maturation	3,59E+00	8,20E+01	1/3 33.3%	1/8356 0.0%	FOXO3	
6	1831	trophectodermal cellular morphogenesis	3,59E+00	8,20E+01	1/3 33.3%	1/8356 0.0%	ESRRB	
7	1834	trophectodermal cell proliferation	3,59E+00	8,20E+01	1/3 33.3%	1/8356 0.0%	ESRRB	
8	6355	regulation of transcription, DNA-dependent	6,27E+00	8,94E+01	3/3 100.0%	716/8356 8.5%	FEV ESRRB FOXO3	
9	51252	regulation of RNA metabolic process	6,78E+00	8,94E+01	3/3 100.0%	735/8356 8.7%	FEV ESRRB FOXO3	
10	9994	oocyte differentiation	7,18E+00	8,94E+01	1/3 33.3%	2/8356 0.0%	FOXO3	
11	1547	antral ovarian follicle growth	7,18E+00	8,94E+01	1/3 33.3%	2/8356 0.0%	FOXO3	
12	48599	oocyte development	7,18E+00	8,94E+01	1/3 33.3%	2/8356 0.0%	FOXO3	
13	1542	ovulation from ovarian follicle	1,08E+01	1,13E+02	1/3 33.3%	3/8356 0.0%	FOXO3	
14	30728	ovulation	1,08E+01	1,13E+02	1/3 33.3%	3/8356 0.0%	FOXO3	
15	45595	regulation of cell differentiation	1,19E+01	1,16E+02	2/3 66.6%	168/8356 2.0%	ESRRB FOXO3	
16	45449	regulation of transcription	1,59E+01	1,31E+02	3/3 100.0%	976/8356 11.6%	FEV ESRRB FOXO3	
17	45648	positive regulation of erythrocyte differentiation	1,79E+01	1,31E+02	1/3 33.3%	5/8356 0.0%	FOXO3	
18	48468	cell development	1,82E+01	1,31E+02	2/3 66.6%	208/8356 2.4%	ESRRB FOXO3	
19	10556	regulation of macromolecule biosynthetic process	2,01E+01	1,31E+02	3/3 100.0%	1056/8356 12.6%	FEV ESRRB FOXO3	
20	48477	oogenesis	2,15E+01	1,31E+02	1/3 33.3%	6/8356 0.0%	FOXO3	
21	31326	regulation of cellular biosynthetic process	2,25E+01	1,31E+02	3/3 100.0%	1096/8356 13.1%	FEV ESRRB FOXO3	
22	10468	regulation of gene expression	2,29E+01	1,31E+02	3/3 100.0%	1103/8356 13.2%	FEV ESRRB FOXO3	
23	9889	regulation of biosynthetic process	2,30E+01	1,31E+02	3/3 100.0%	1104/8356 13.2%	FEV ESRRB FOXO3	
24	19219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2,40E+01	1,31E+02	3/3 100.0%	1119/8356 13.3%	FEV ESRRB FOXO3	
25	51171	regulation of nitrogen compound metabolic process	2,44E+01	1,31E+02	3/3 100.0%	1126/8356 13.4%	FEV ESRRB FOXO3	

Plan1 | Pronto

- Avaliar os processos biológicos (PB) dos FT

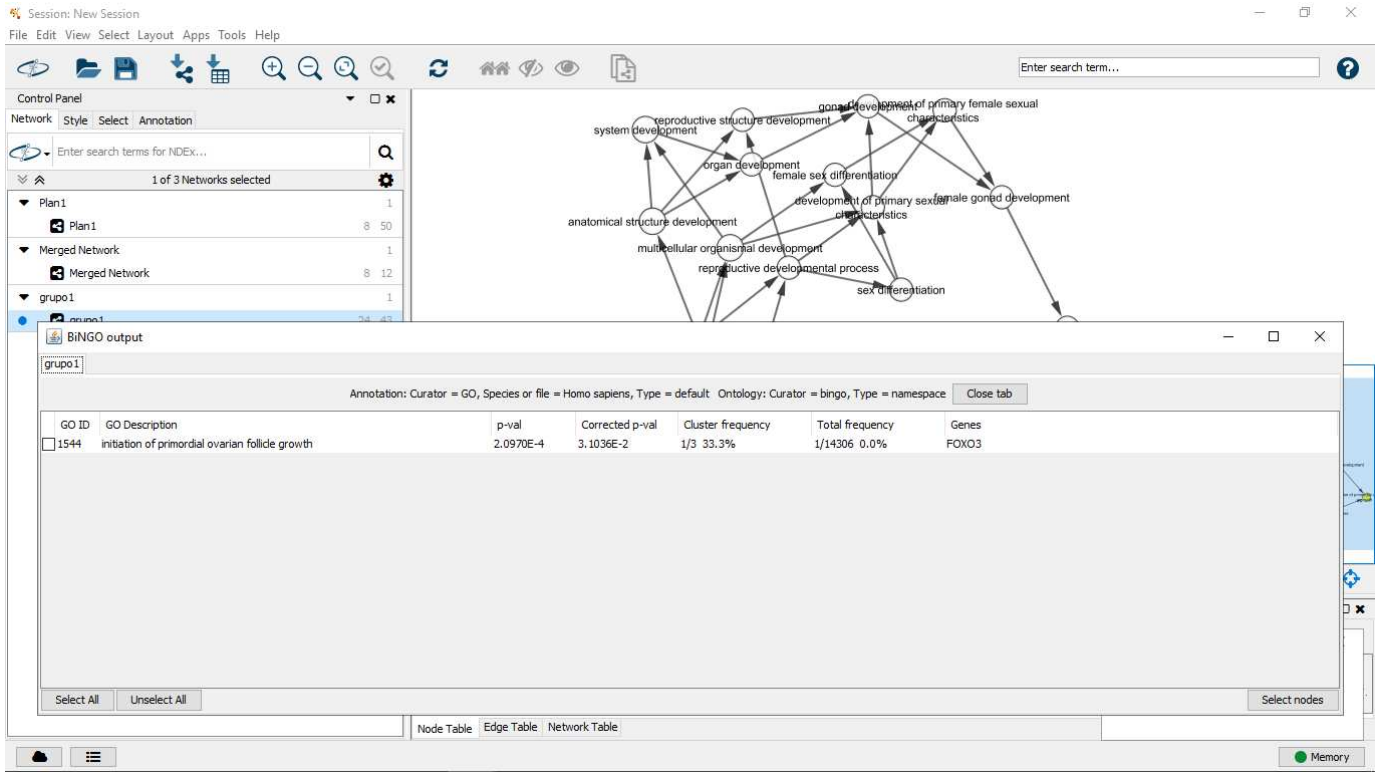
➢ Selecionar os FT com os PB mais associados à(s) característica(s) avaliada(s)

➢ Salvar os FT selecionados em outra planilha

Dica:

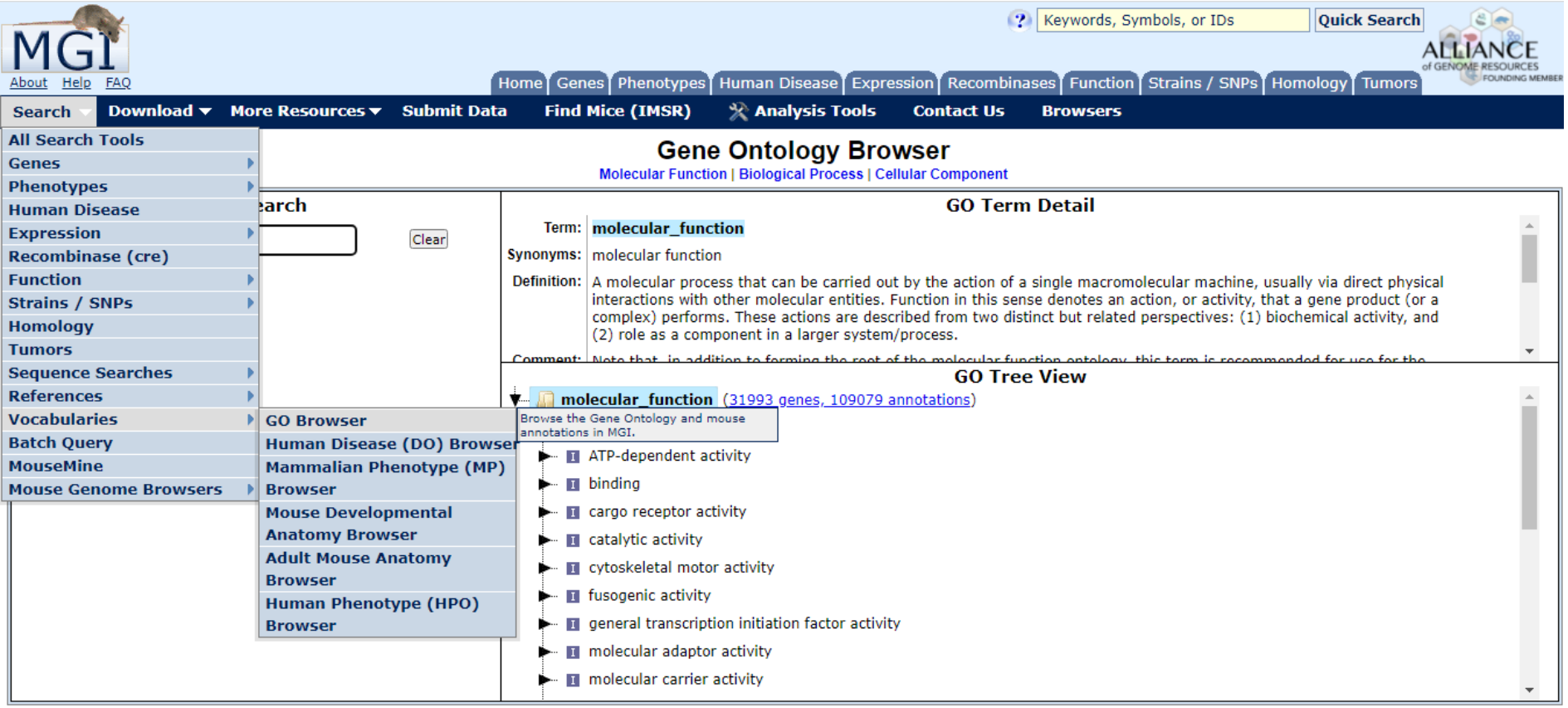
- No BiNGO também podemos fazer essa mesma análise considerando a espécie Homo sapiens ao invés da espécie animal estudada, porque geralmente tem mais resultados para humanos

- Neste caso deu apenas um resultado e foi bem específico para as características estudadas (Número de oócitos e embriões)



Dica: se houver dúvida sobre algum processo biológico

- Este site contém algumas definições para alguns dos processos biológicos:
- Mouse Genome Informatics (MGI): http://www.informatics.jax.org/
- Para pesquisar o PB: Search -> Vocabularies -> GO Browser



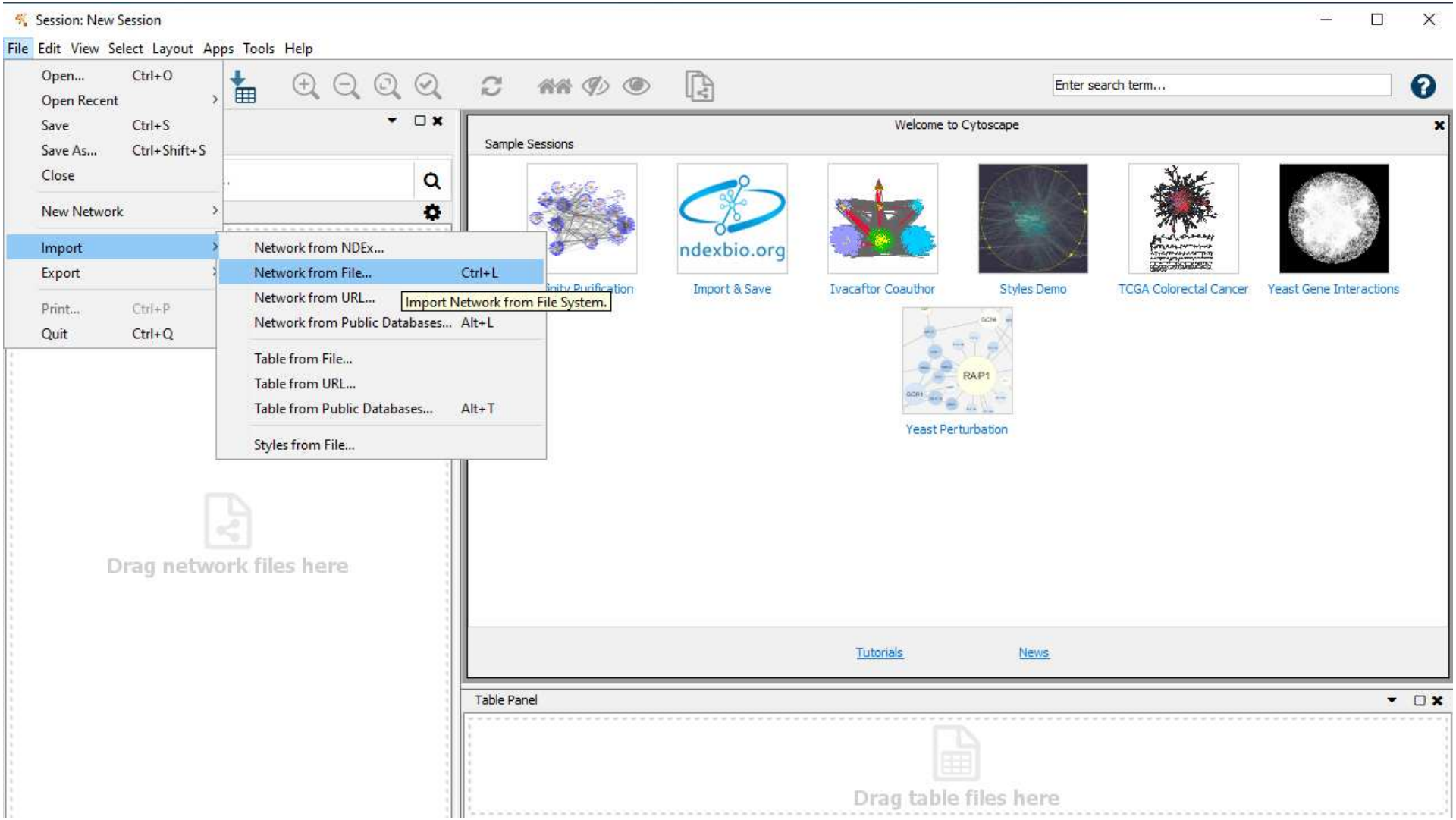
Ao final desta etapa

- Fazer revisão bibliográfica dos genes e fatores de transcrição:
  - Buscar associação dos genes/FT com as características estudas
  - Buscar associação dos genes/FT com os processos biológicos das vias enriquecidas

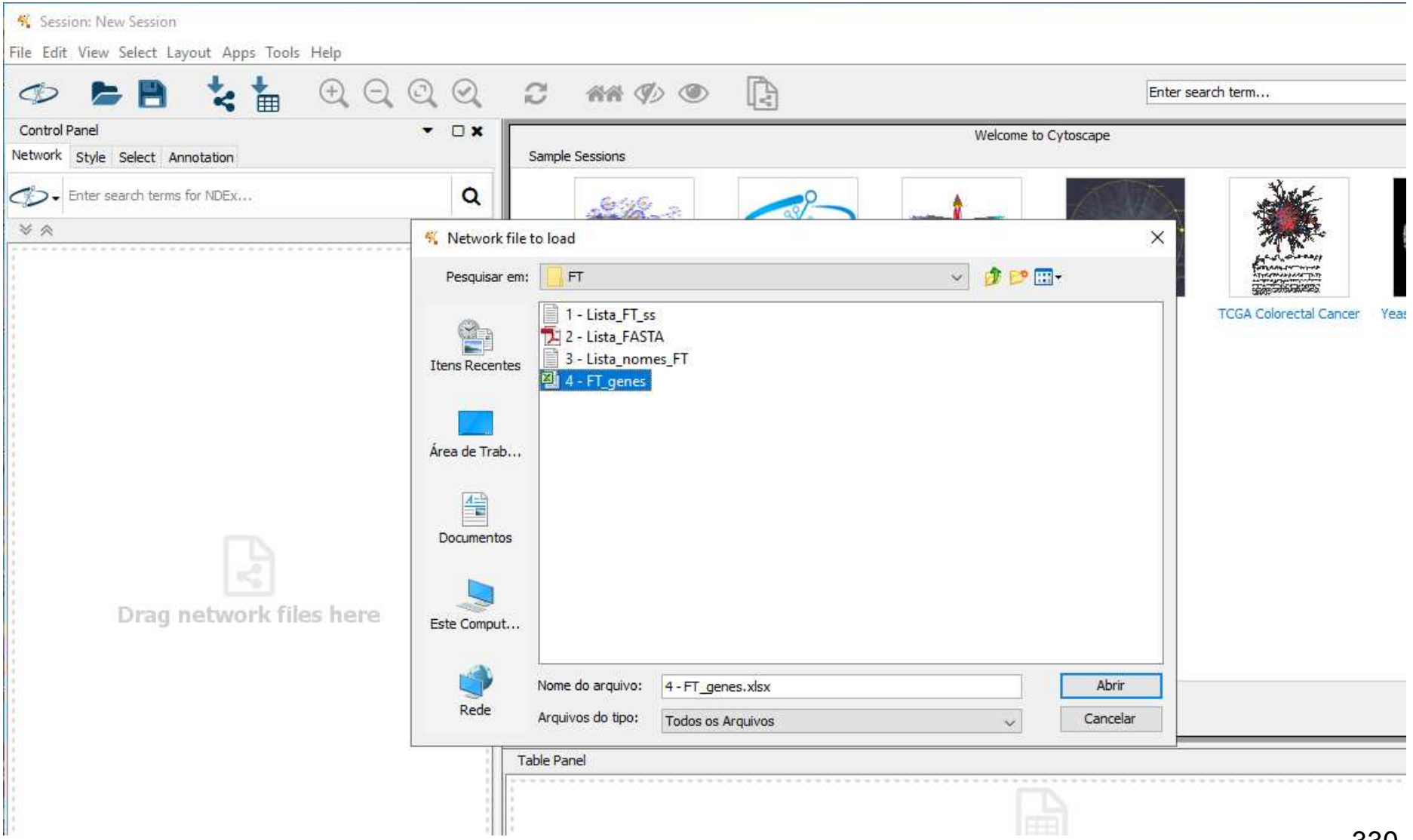
- Na planilha do Excel obtida no 9º passo, selecionar apenas os conjuntos de genes-FT que se destacaram na revisão bibliográfica – descartar os genes/FT não relevantes para a pesquisa

NETWORK ANALYZER

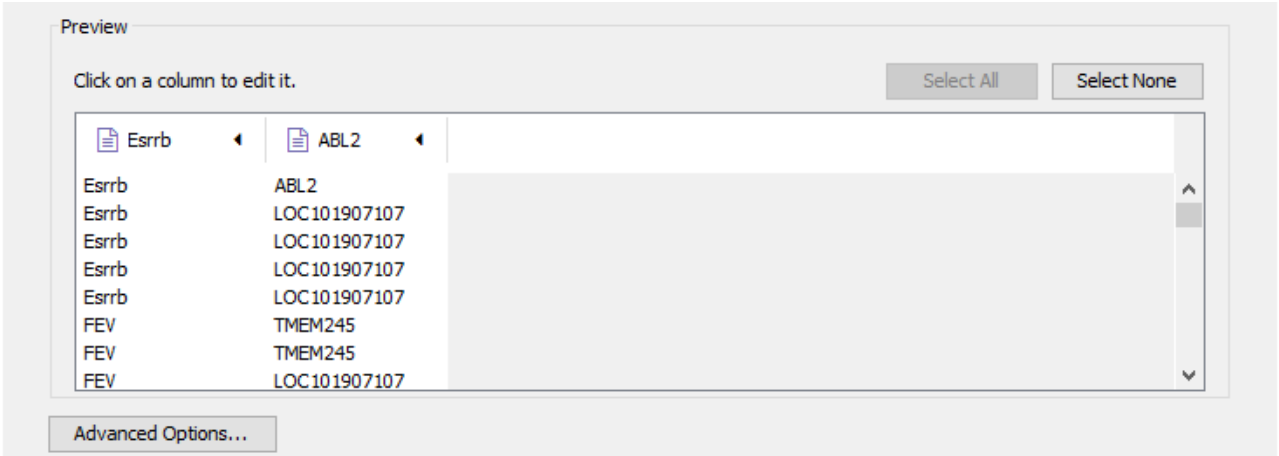
Com o programa aberto, clicar em File → Import → Network from File...

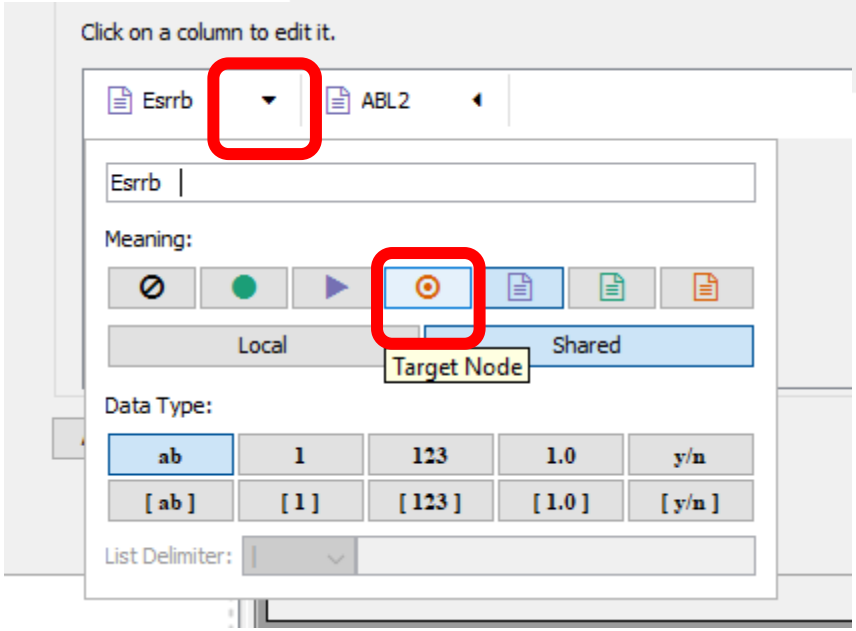


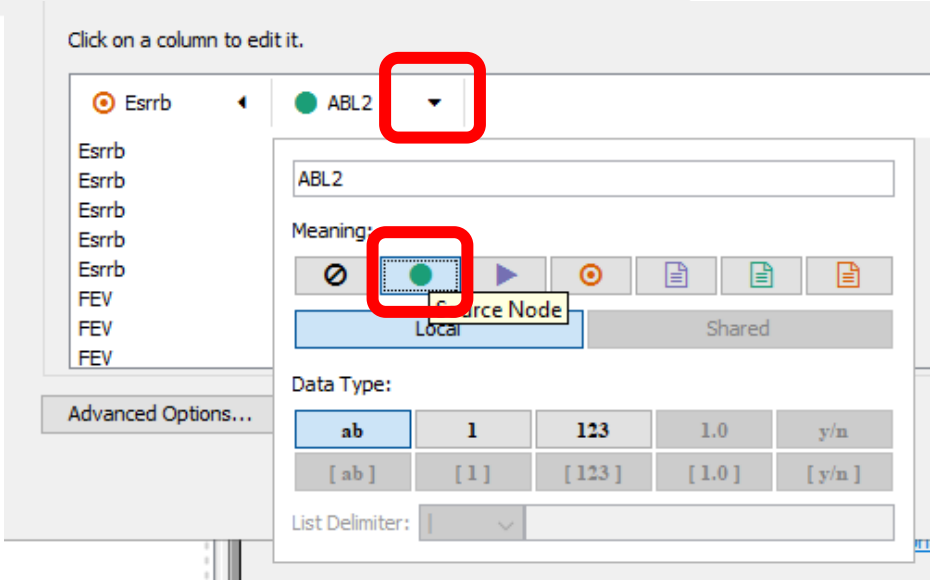
Pesquisar a planilha que foi salva no 9º passo (agora com genes e FT mais relevantes – após revisão bibliográfica)
Selecionar o arquivo e clicar em abrir



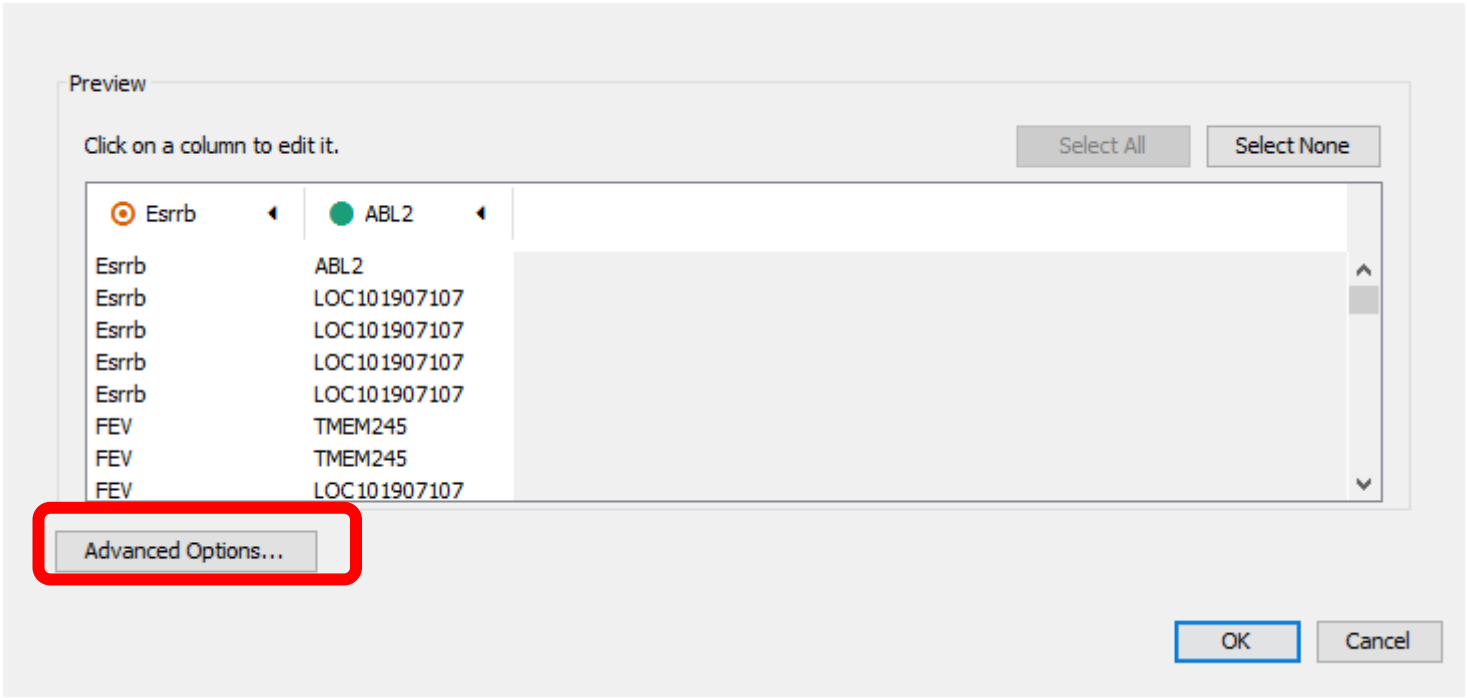
Vai aparecer a seguinte tela:
Na primeira coluna, clicar na setinha e clicar em Target Node (Alvo)
Na segunda coluna, clicar na setinha e clicar em Source Node (gene original)

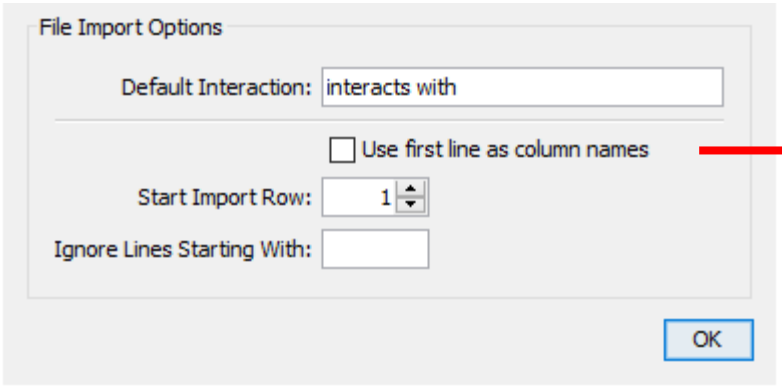






Clicar em Advanced Options...

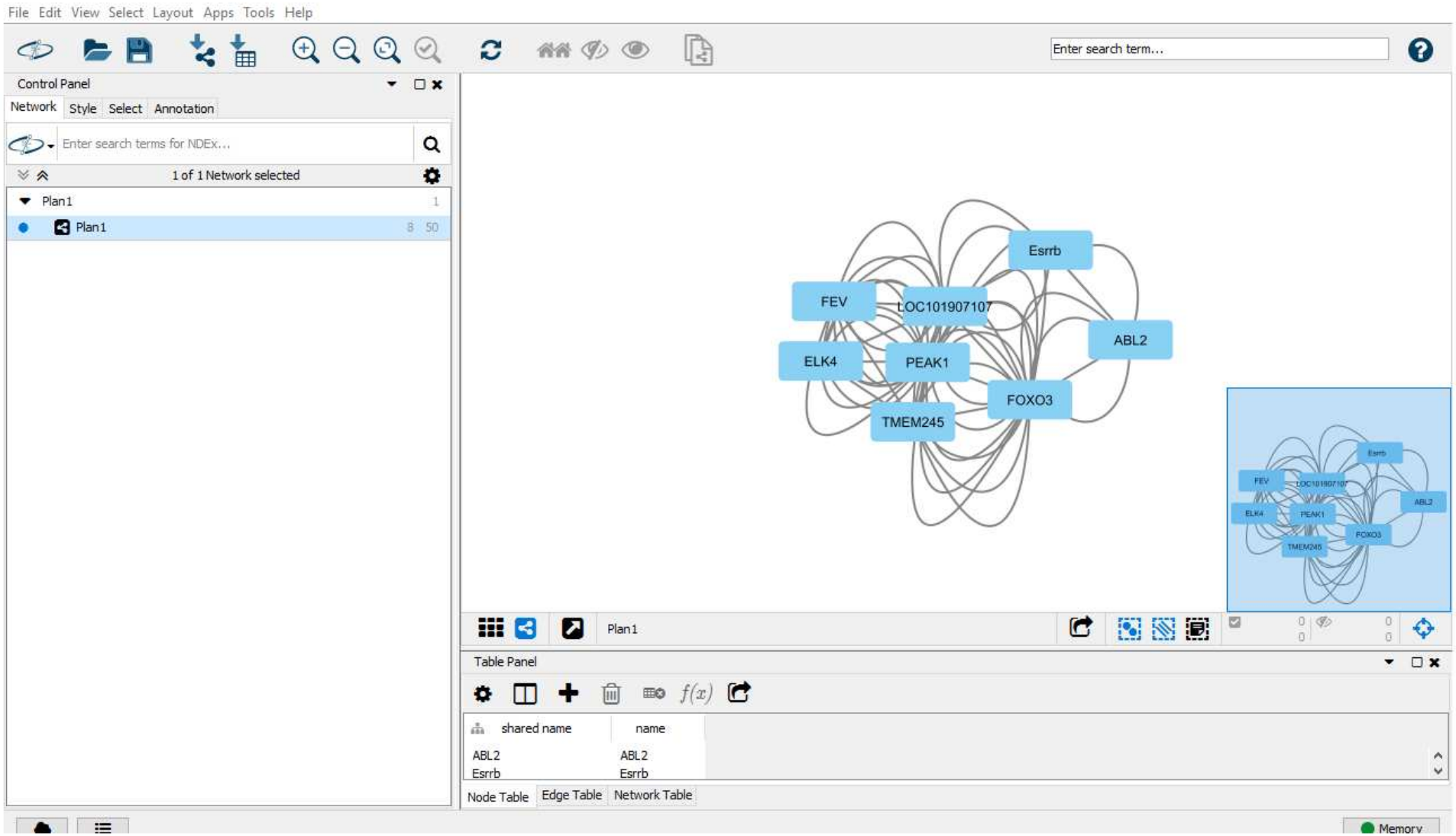




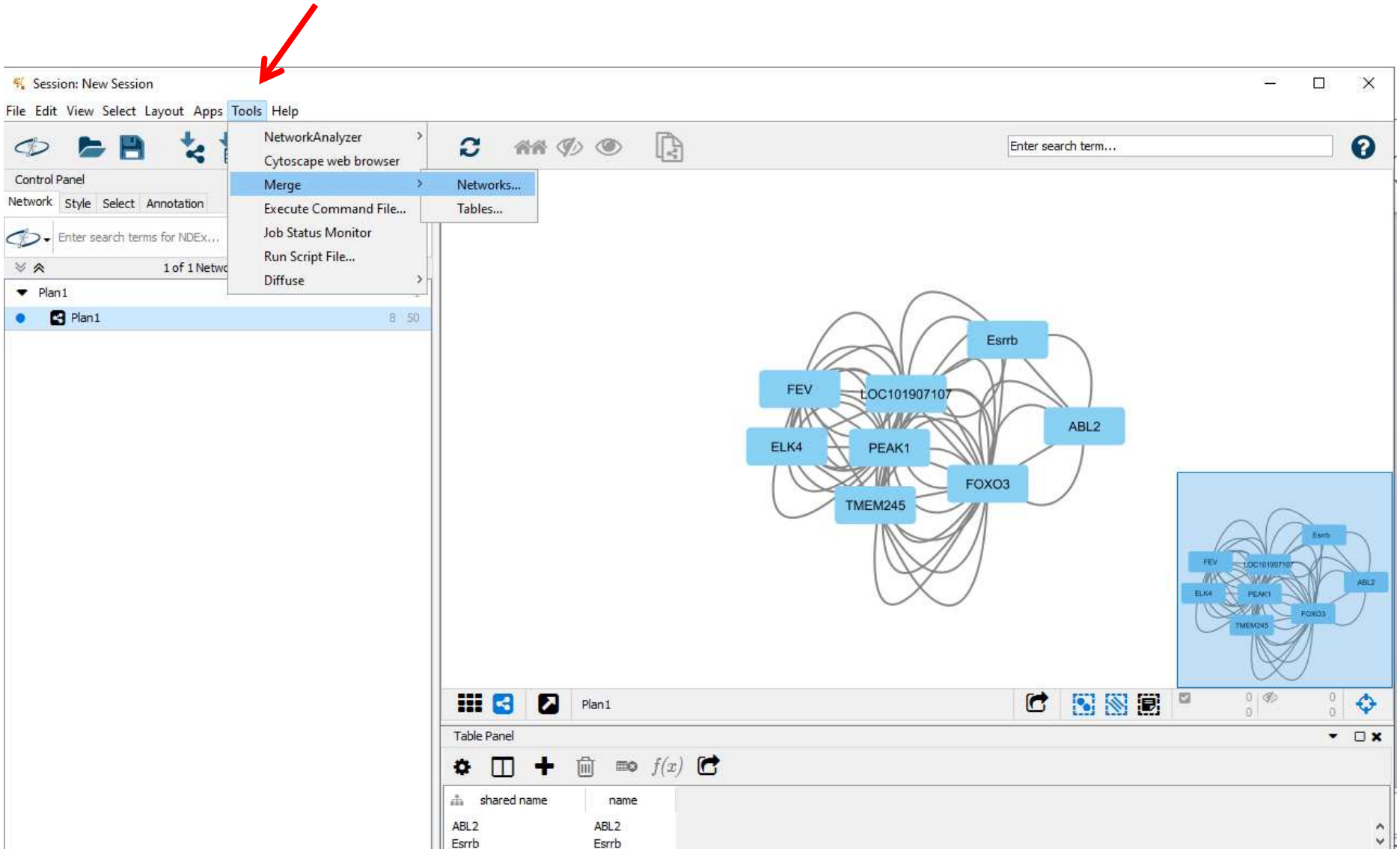
- Se esta opção estiver selecionada, remover a seleção, pois os dados da planilha não tem cabeçalho
- Se tiver deixado cabeçalho na planilha no 9º passo, pode deixar essa opção marcada

Depois dar OK nas duas caixas abertas

Vai aparecer a seguinte tela com as interações entre genes e FTs



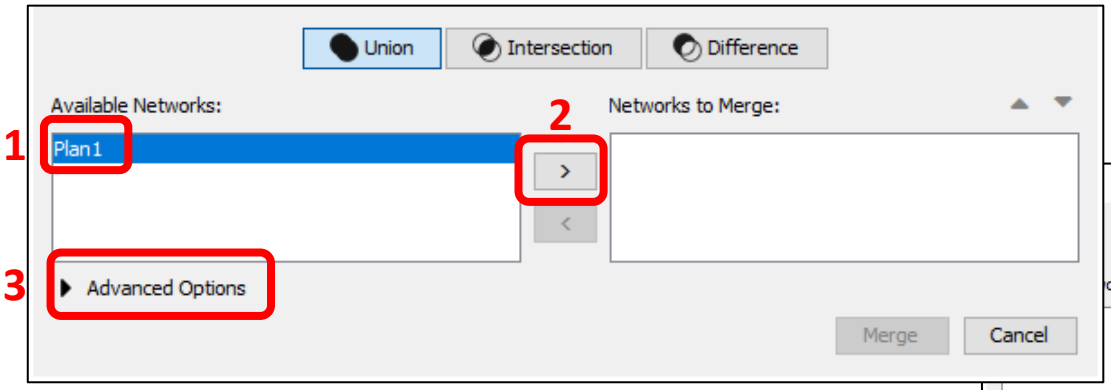
No próximo passo:
Clicar em Tools → Merge → Networks...

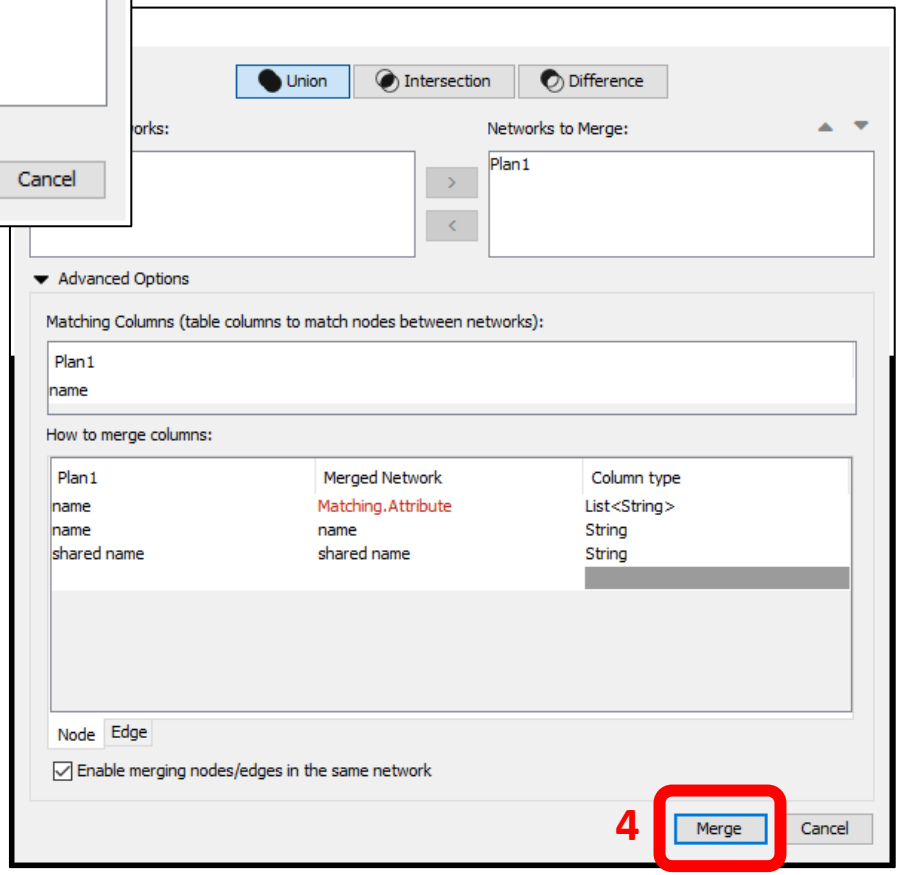


No próximo passo:

1 – Clicar na opção disponível na caixa ‘Available Networks’

2 – Clicar na setinha para a direita para passar para ‘Networks to Merge’

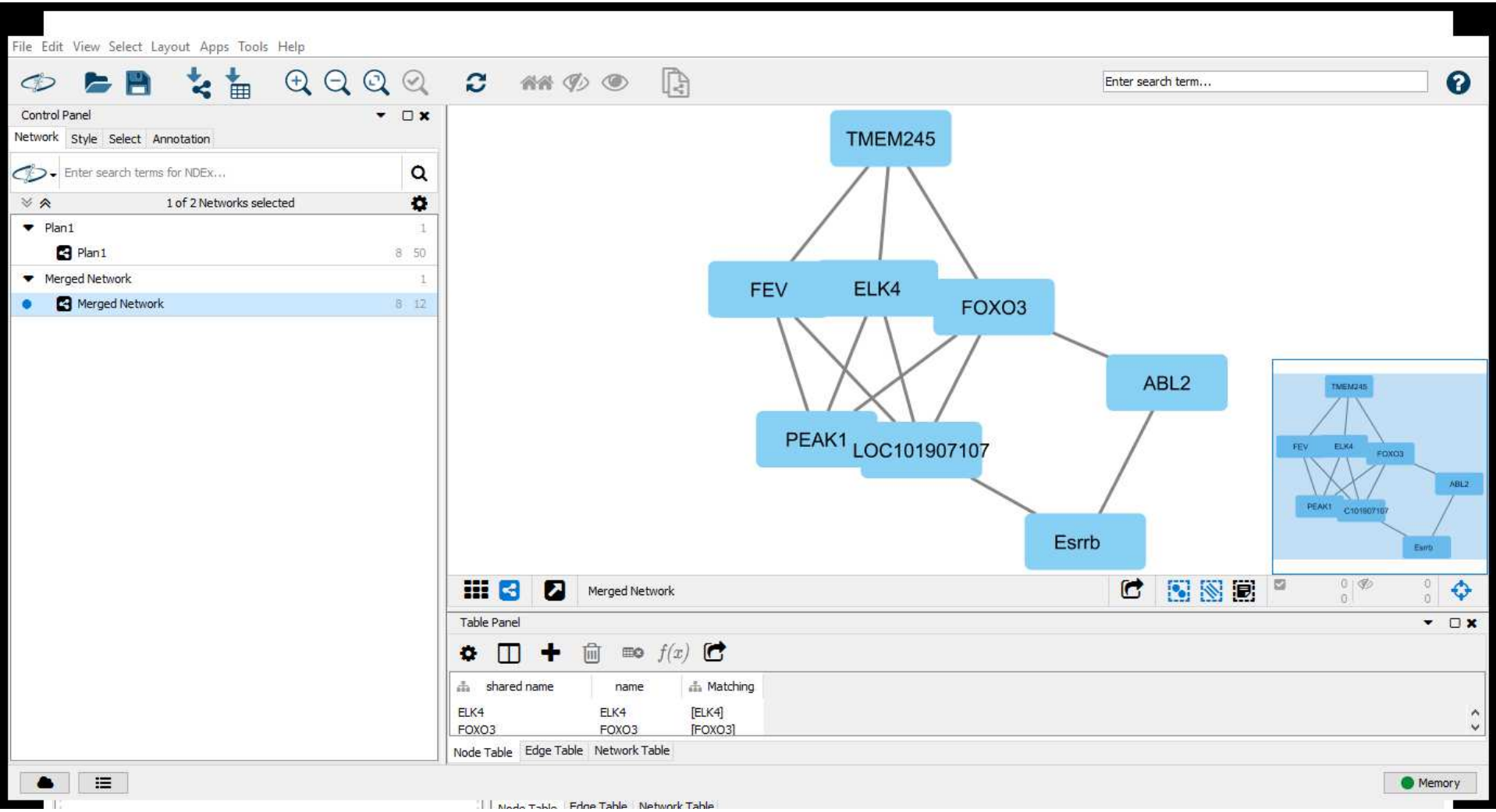




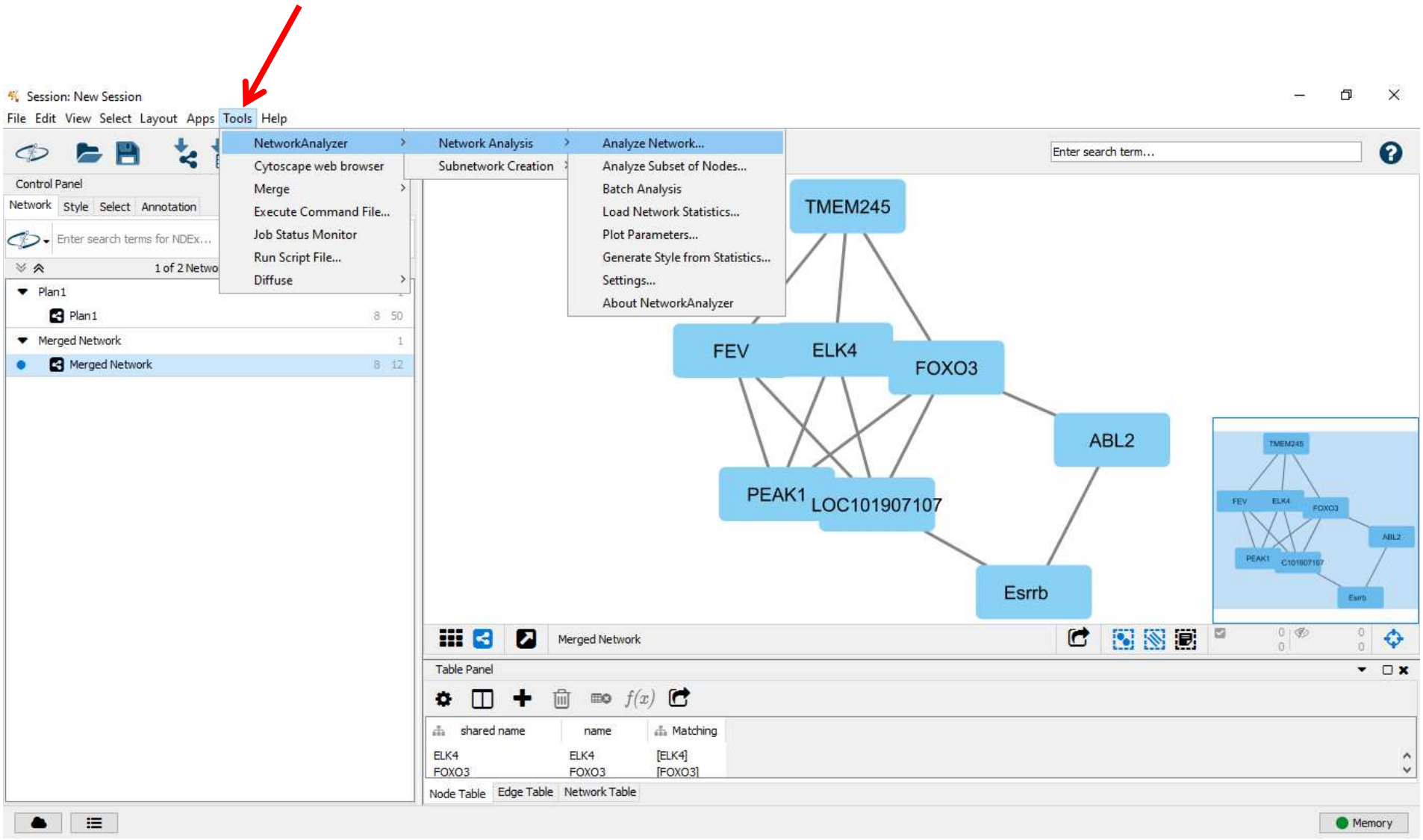
3 – Clicar em Advanced Options

Não alterar mais nenhum parâmetro

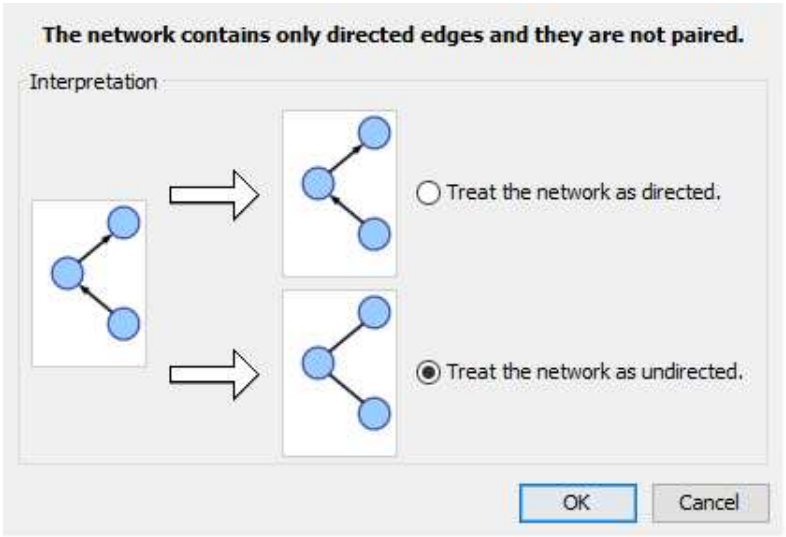
4 – Clicar em Merge


File Edit View Select Layout Apps Tools Help
Enter search term...
Control Panel
Network Style Select Annotation
Enter search terms for NDEx...
1 of 2 Networks selected
Plan1
Plan1
Merged Network
Merged Network
TMEM245
FEV
ELK4
FOXO3
ABL2
PEAK1
LOC101907107
Esrrb
Merged Network
Table Panel
shared name
name
Matching
ELK4
FOXO3
ELK4
FOXO3
[ELK4]
[FOXO3]
Node Table
Edge Table
Network Table
Memory

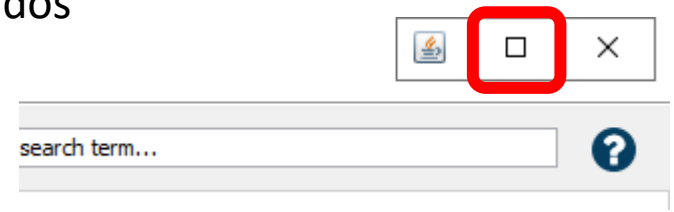
No próximo passo:
- Clicar em Tools → NetworkAnalyzer → Network Analysis → Analyze Network...

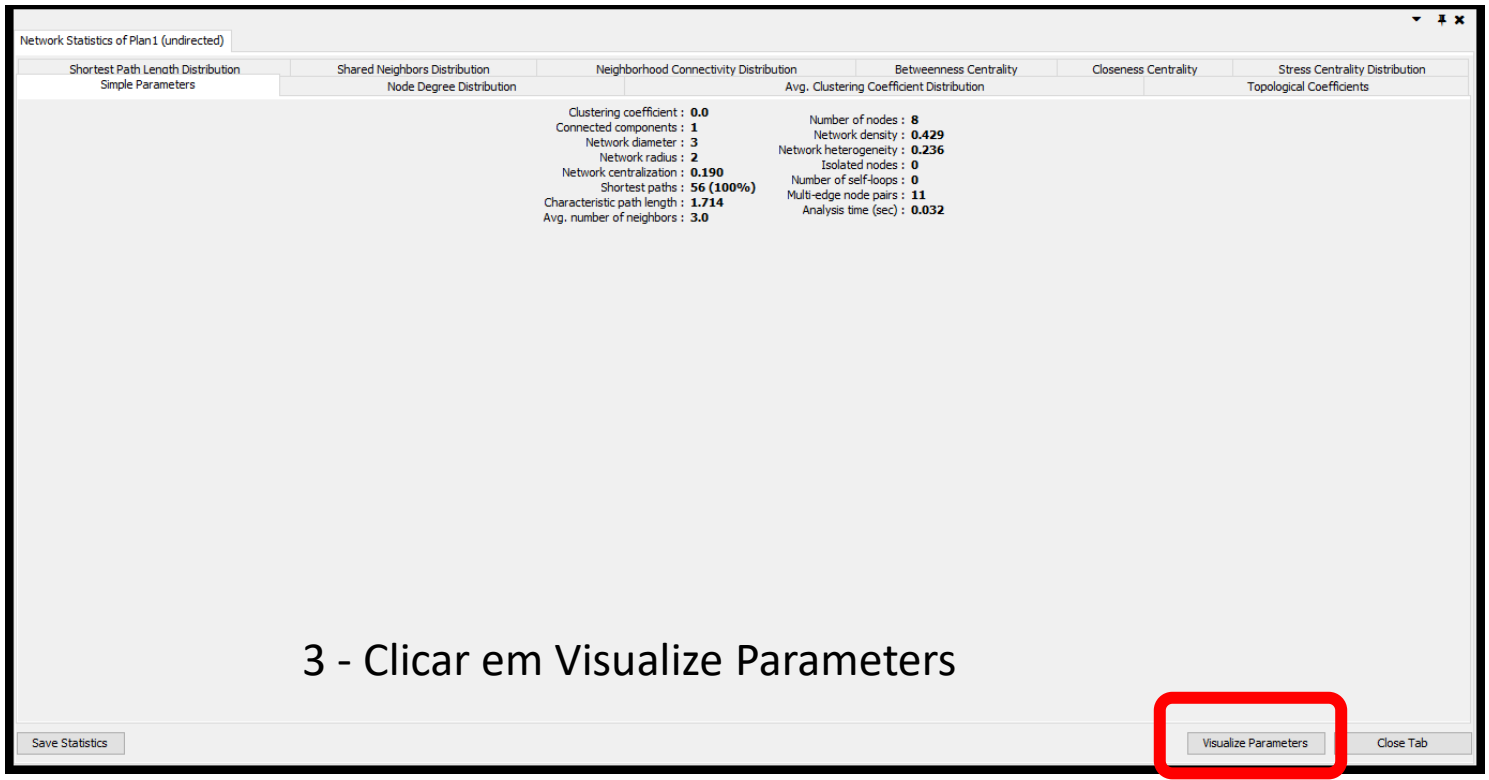


1 – Selecionar ‘Treat the network as undirected’



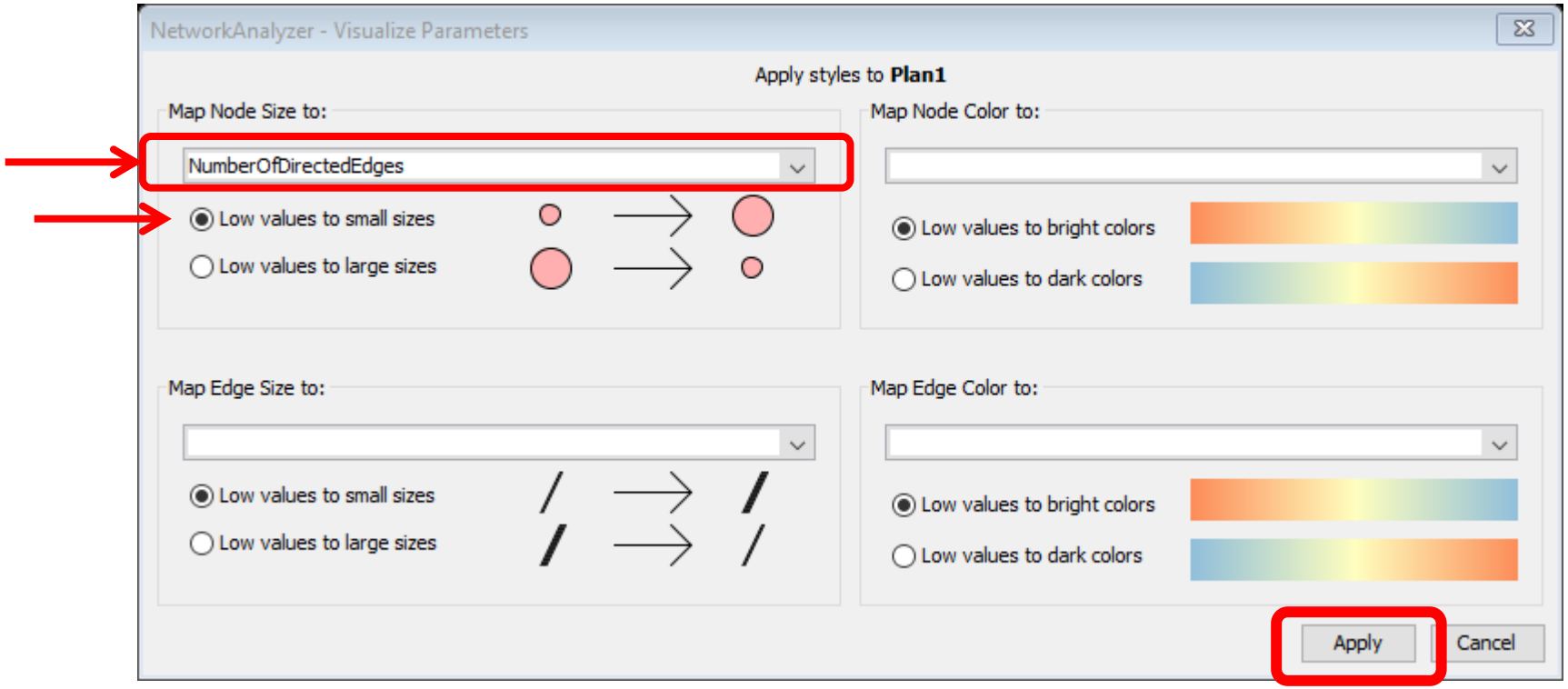
2 - Depois de clicar em OK, vai aparecer uns ícones pequenos no canto superior direito. Clicar no quadradinho para ampliar a tela de resultados



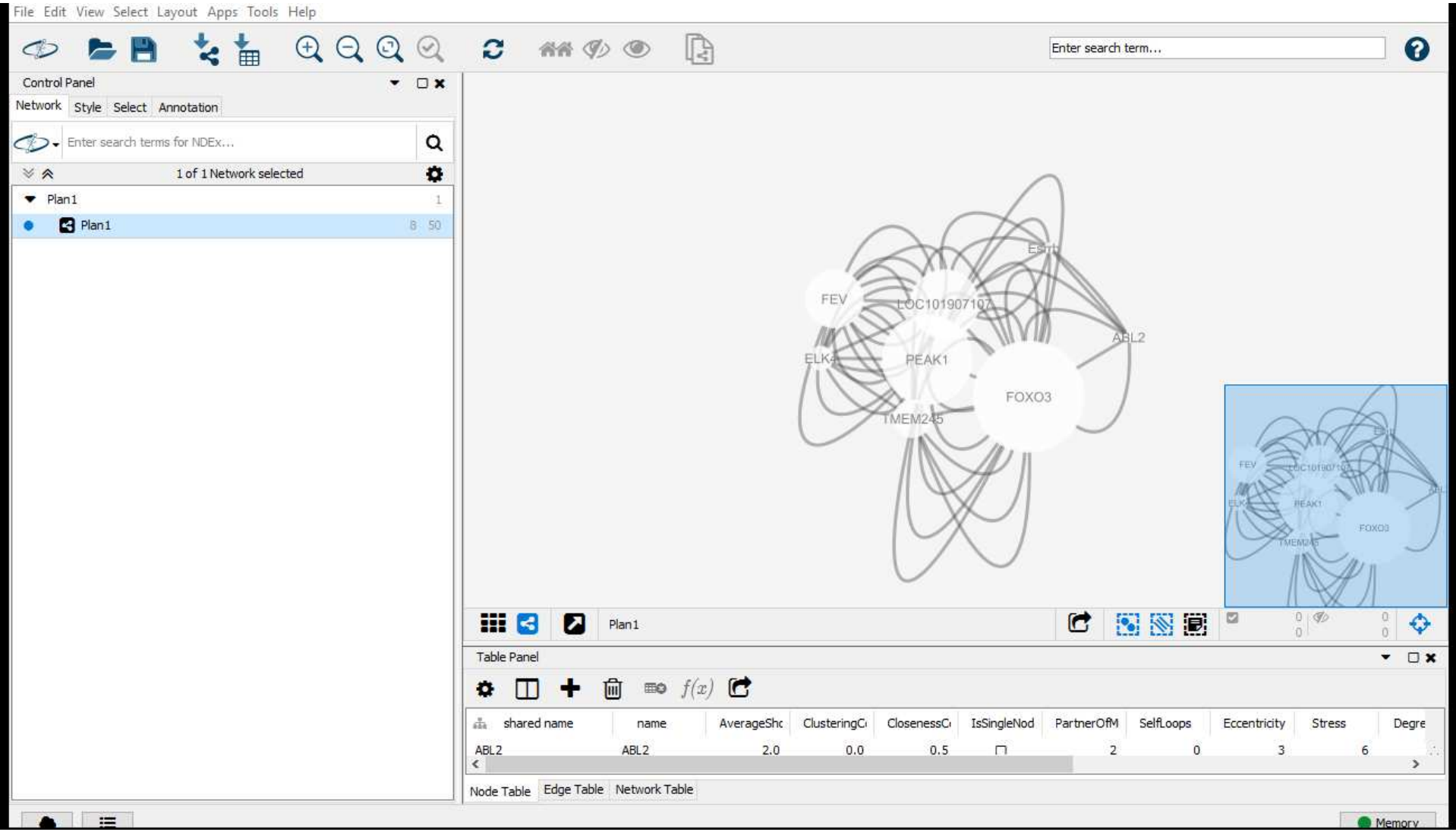


3 - Clicar em Visualize Parameters

- Na parte ‘Map Node Size to:’
- Escolher a opção: ‘NumberofDirectedEdges’
- Selecionar ‘Low values do small sizes’
- Manter as outras como estão
- Clicar em ‘Apply’



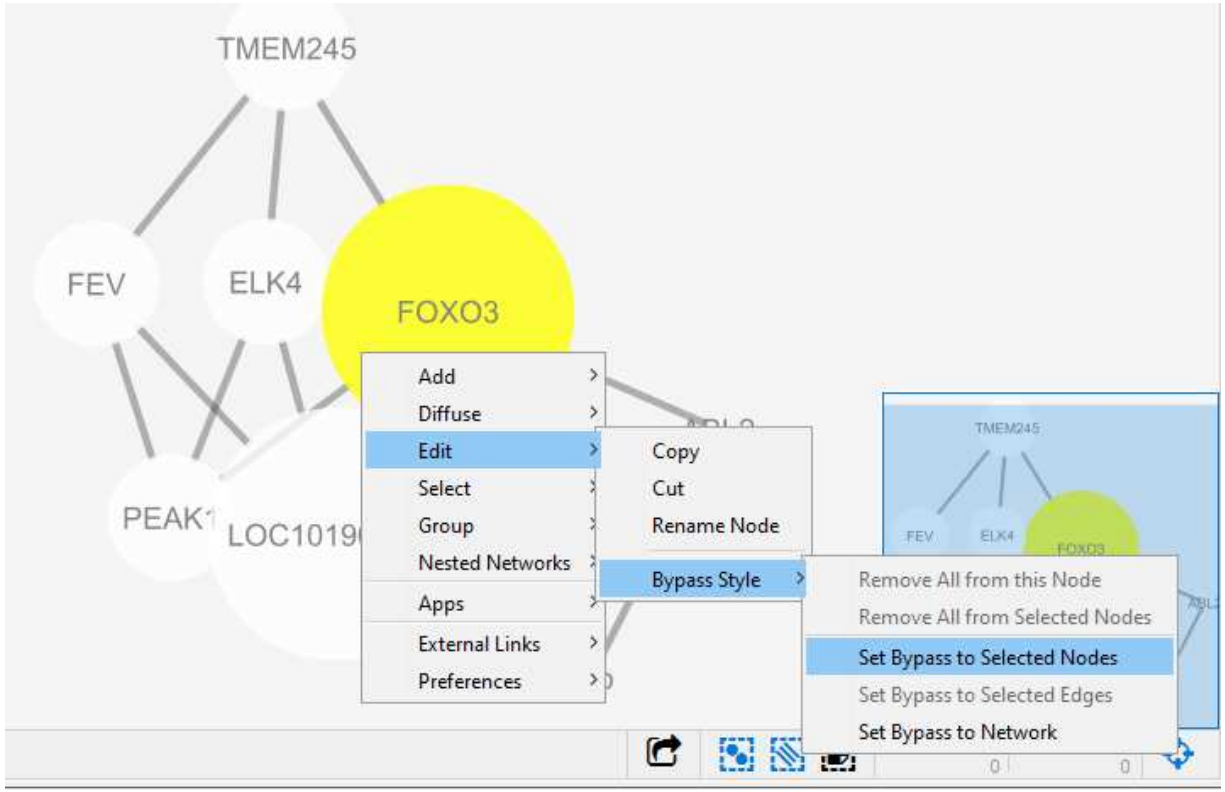
Vai aparecer a seguinte tela

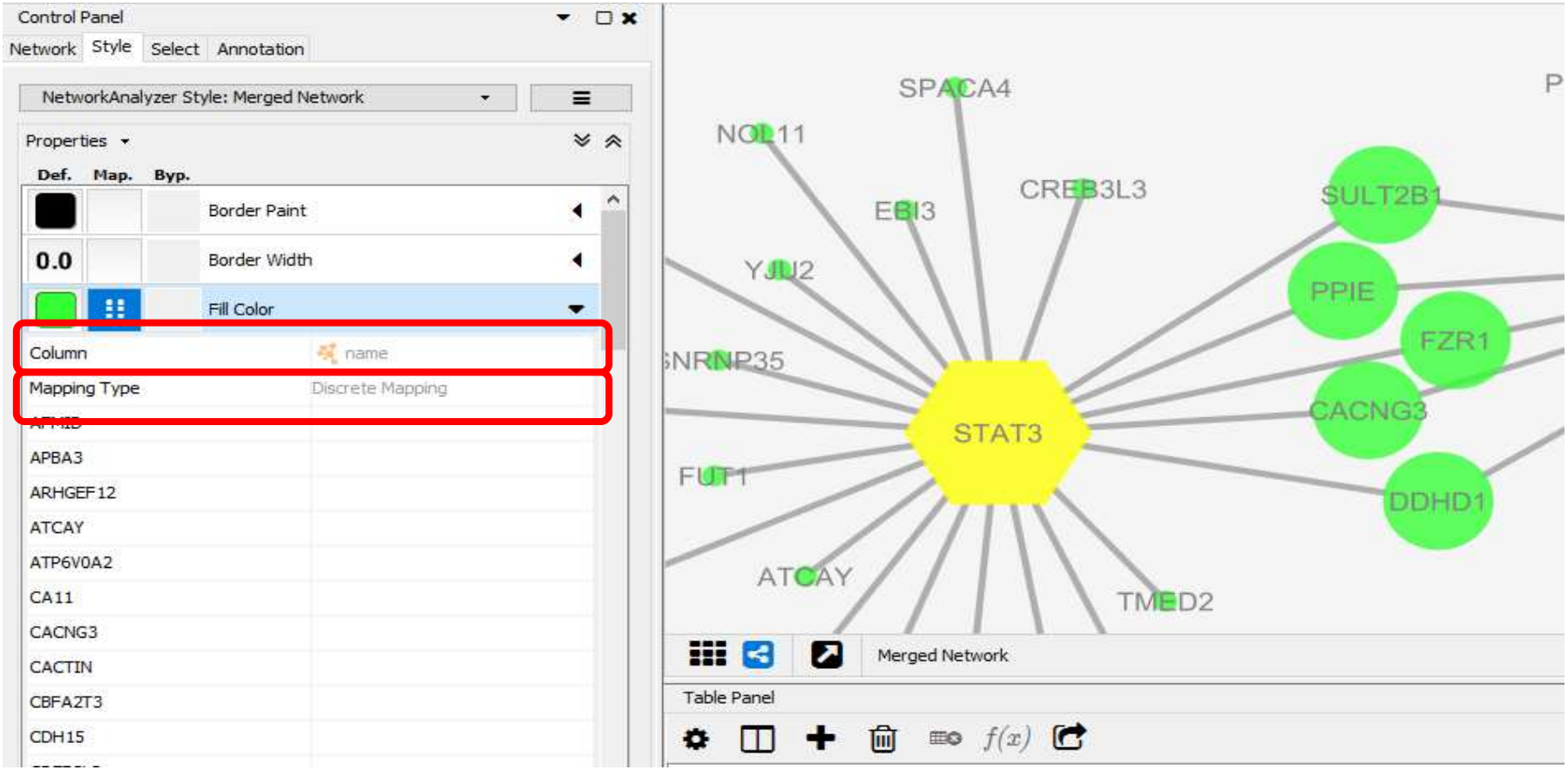


Se quiser editar as cores para identificar os genes e fatores de transcrição

- Ao clicar nas bolinhas com cada gene ou fator de transcrição, elas ficam com cor

- Para alterar a cor dos genes e fatores de transcrição:
- Clicar com o botão direito em cima da bolinha → Edit → Bypass Style → Set Bypass to Selected Nodes



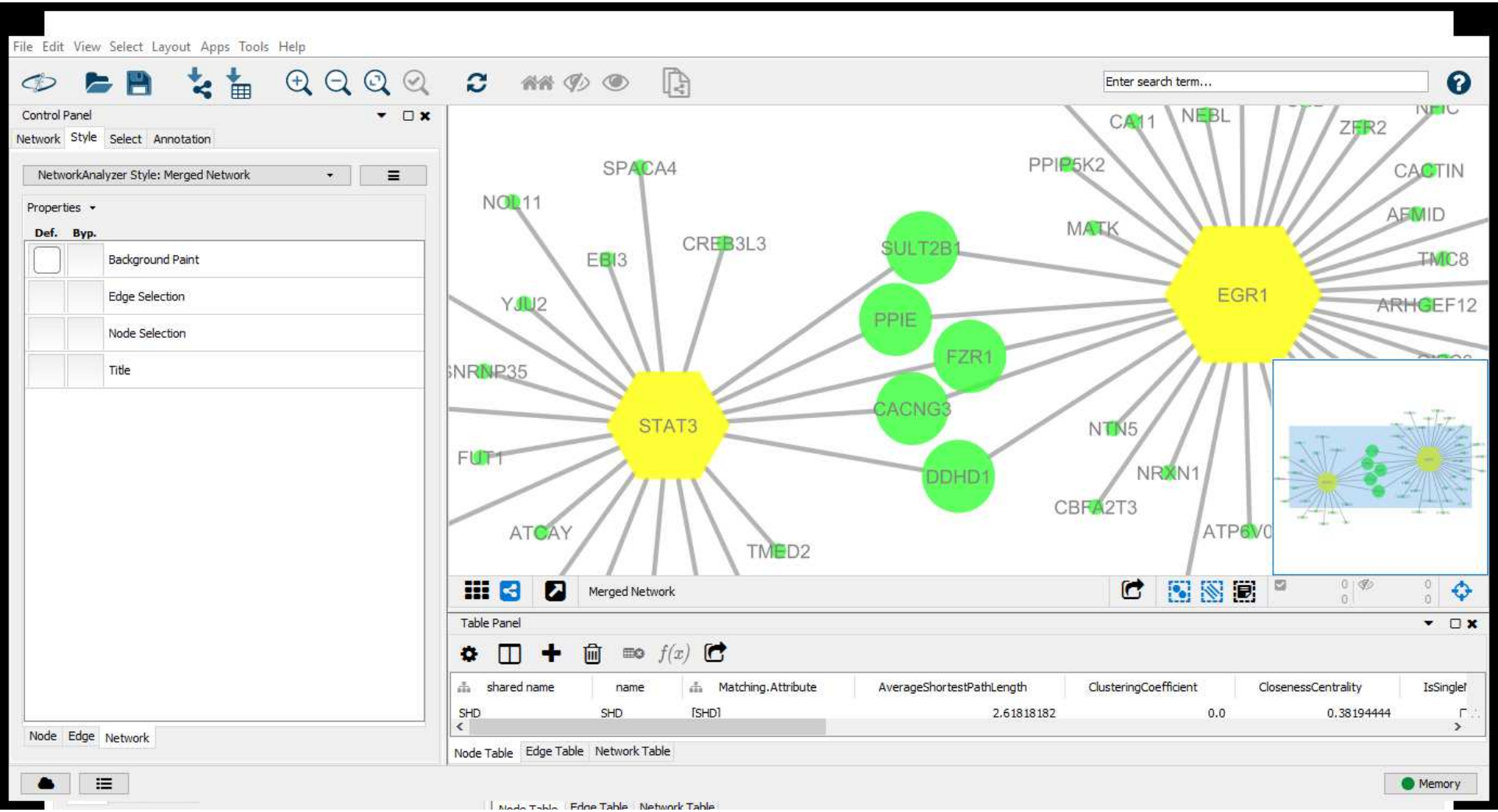


Para preencher as cores:
Se quiser todos da mesma cor é só selecionar a cor no primeiro quadrinho (onde está verde na imagem. Para preencher um por um:
Clicar na setinha ao lado de 'Fill Color' → Ao lado de 'Column' selecionar 'name' → Ao lado de 'Mapping Type' selecionar 'Discrete Mapping'

Clicar na segunda coluna (ao lado do nome de cada gene) → Clicar nos 3 pontinhos → Selecionar a cor (o mesmo vale para outras alterações, ex.: forma = Shape)

- Para mudar a cor de fundo:
- Clicar com o botão direito na imagem
- Edit → Bypass Style → Set Bypass to Network

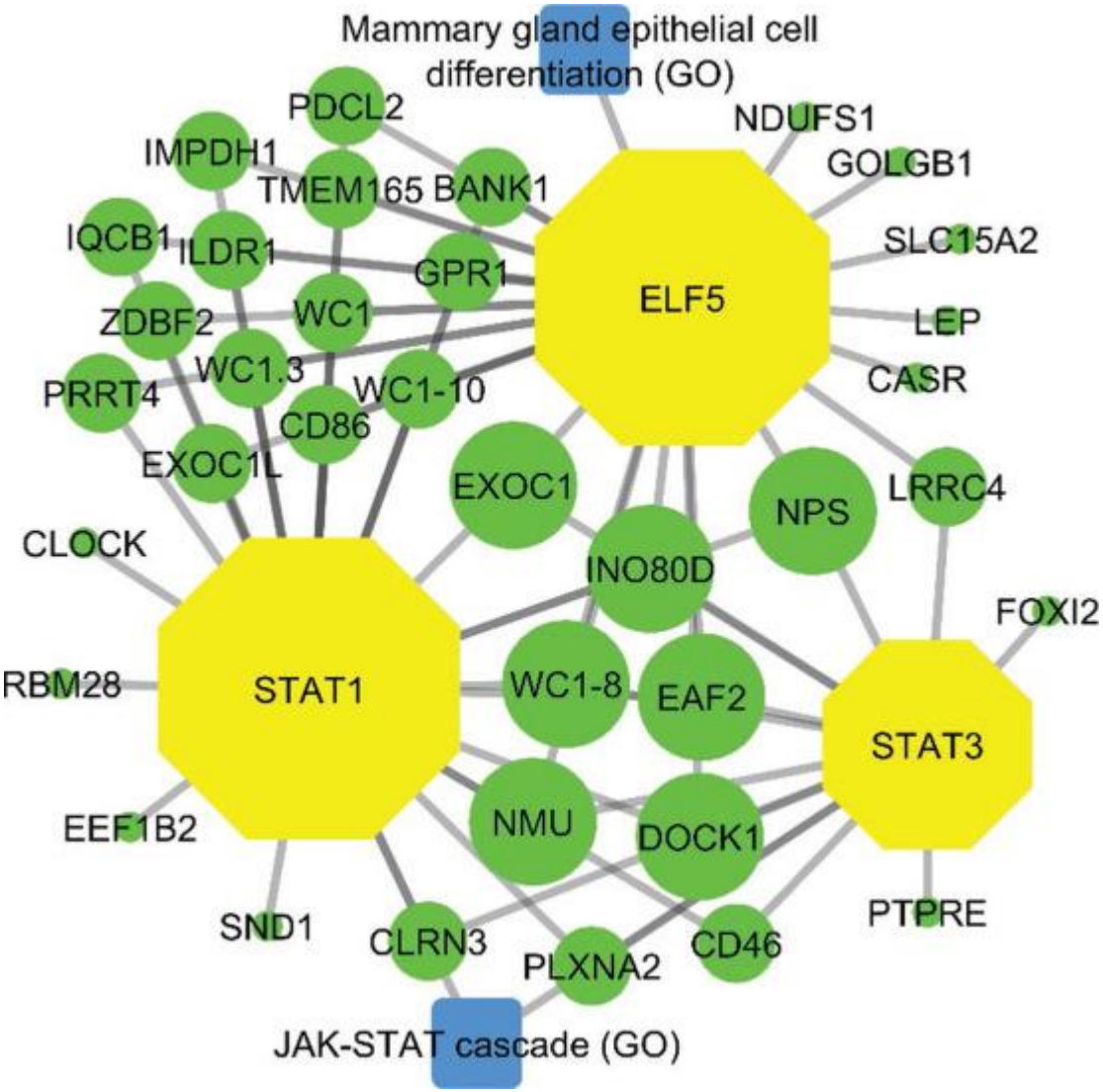
- No menu à esquerda:
- Em ‘Background Paint’ selecionar a cor desejada



Exemplo:

Figure 3. Gene-transcription factor (TF) network: Genes located in the top 10 windows for 305-d milk yield (kg; green circle nodes) and their associated TF (yellow octagon nodes). Node size corresponds to network analyses (Cytoscape; Shannon et al., 2003), in which larger nodes denotes a higher edge density associated with the number of TF binding sites. Blue square nodes show the gene ontology (GO) biological processes related to TF.

Otto et al. (2020) https://doi.org/10.3168/jds.2019-17890



Considerações finais

- Sobre os genes identificados, principalmente os que tem processo biológicos relacionados às características estudadas, seria interessante realizar estudos futuros avaliando a expressão gênica