



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

DISSERTAÇÃO DE MESTRADO

Identificação e análise funcional *in silico* de variantes nos genes de proteínas do leite das raças Guzerá e Gir

ORIENTADA: Carolina Guimarães Ramos Matosinho

ORIENTADORA: Maria Raquel Santos Carvalho

CO-ORIENTADORA: Izinara Rosse da Cruz

BELO HORIZONTE

FEVEREIRO – 2016

Carolina Guimarães Ramos Matosinho

Identificação e análise funcional *in silico* de variantes nos genes de proteínas do leite das raças Guzerá e Gir

Dissertação submetida ao programa de Pós-Graduação em Genética da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do grau de Mestre em Genética

Área de Concentração: Genômica e Bioinformática

Orientadora: Maria Raquel Santos Carvalho

Co-orientadora: Izinara Rosse da Cruz

BELO HORIZONTE

INSTITUTO DE CIÊNCIAS BIOLÓGICAS - UFMG

FEVEREIRO – 2016

- 043 Matosinho, Carolina Guimarães Ramos.
Identificação e análise funcional in silico de variantes nos genes de proteínas do leite das raças Guzerá e Gir [manuscrito] / Carolina Guimarães Ramos Matosinho. – 2016.
83 f. : il. ; 29,5 cm.
- Orientadora: Profa. Dra. Maria Raquel Santos Carvalho. Co-orientadora: Profa. Dra. Izinara Rosse Cruz.
- Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.
1. Genômica. 2. Bioinformática. 3. Bovinos. 4. Gir (Zebu). 5. Guzera (Zebu). 6. Leite. 7. Caseínas. I. Carvalho, Maria Raquel Santos. II. Cruz, Izinara Rosse. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 575



ATA DA DEFESA DE DISSERTAÇÃO

Carolina Guimarães Ramos Matosinho

227
226/2016
entrada
1º/2014
CPF:
105.313.906-32

Às quatorze horas do dia 17 de fevereiro de 2016, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Dissertação, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: "**Identificação e análise funcional in silico de variantes nos genes de proteínas do leite das raças Guzará e Gir**", requisito para obtenção do grau de Mestre em **Genética**. Abrindo a sessão, a Presidente da Comissão, **Dra. Maria Raquel Santos Carvalho**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dra. Maria Raquel Santos Carvalho	UFMG	34965173019	Aprovada
Dra. Izinara Rosse da Cruz	UFMG	07894285633	Aprovada
Dr. Gabriel da Rocha Fernandes	Fiocruz	05270621622	Aprovada
Dra. Glória Regina Franco	UFMG	623.387.496-34	APROVADA

Pelas indicações, a candidata foi considerada: Aprovada
O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 17 de fevereiro de 2016.

Dra. Maria Raquel Santos Carvalho - Orientadora Maria Raquel Santos Carvalho
Dra. Izinara Rosse da Cruz - Coorientadora Izinara Rosse da Cruz
Dr. Gabriel da Rocha Fernandes Gabriel da Rocha Fernandes
Dra. Glória Regina Franco Glória Regina Franco

DEDICATÓRIA

Dedico esse trabalho à minha mãe, Jacqueline, ao meu pai, Eustáquio, aos meus avôs, Gui e Manoel, ao meu namorado, Pablo, e a todos os meus familiares que foram essenciais à realização dessa conquista.

AGRADECIMENTOS

À Deus, por todas as bênçãos recebidas, por me guiar e proteger meus passos;

À minha orientadora, Prof.^a Dr.^a Maria Raquel Santos Carvalho, pela oportunidade, confiança, atenção, disponibilidade e pelos ensinamentos;

À minha co-orientadora, Dr.^a Izinara Rosse da Cruz, pelos ensinamentos, confiança, atenção, disponibilidade, contribuições e pela oportunidade;

Aos meus amigos do LGHM, Marlene, Gabi, Pablo, Fernanda, Aline, Maíra, Flávia, Carol, Aninha, Daniel, Luiza e aos Lucas, pela ajuda, carinho e apoio no dia a dia;

A todos os professores, colegas, amigos e ao pessoal da secretaria da Pós-Graduação em Genética da UFMG pelos ensinamentos que contribuíram para o meu aprendizado;

Ao Prof. Peter Laspina, da Viamundi, pela ajuda nas correções do Inglês dos artigos, resumos e apresentações;

À toda minha família por compartilhar os momentos difíceis, pelo apoio e nunca me deixarem desistir. Aos meus pais pelo incentivo, apoio, paciência e carinho. Aos meus avós pelo apoio, carinho, paciência e incentivo e principalmente a minha vó pela companhia nas madrugadas. Aos meus tios por investirem na minha formação. Aos meus primos que me apoiam. Às minhas madrinhas, Regina e Ângela, pelo apoio e incentivo;

Ao meu namorado Pablo Cristian por compartilhar os momentos difíceis, pela compreensão, aconselhamento, paciência, incentivo e por nunca me deixar desistir. À família do meu namorado pelo apoio, carinho e paciência;

Às minhas amigas, Tailine, Camila, Dalila, Lygia, Tathi, Marcela, Sara e Lays, pelo apoio, carinho e aconselhamentos;

Aos meus amigos do curso de inglês, Guilherme, Camila, Ana, Gessa, Nathalia, Alice, Luciana, Anabela e Kelly, pelo apoio, carinho e aconselhamentos;

Ao pessoal do Grupo de Genômica e Biologia Computacional, Juliana, Francislon, Julliane, Laura, Anderson, Fausto, Guilherme, que me receberam de braços abertos. À Juliana Assis pela ajuda e ensinamentos;

Aos membros da banca por terem aceitado o convite;

À CAPES, FAPEMIG, CNPq e ao Programa de Pós-Graduação em Genética – UFMG, pelo apoio financeiro;

Enfim, agradeço a todos que de alguma forma me ajudaram nessa conquista.

RESUMO

O leite é um alimento de elevado valor nutritivo e de grande importância na alimentação humana, fonte de vitaminas, minerais (particularmente cálcio), carboidratos, lipídios e proteínas. O leite é composto por dois grupos de proteínas: uma fração sólida, formada predominantemente pelas caseínas, e uma fração solúvel, formada pelas proteínas do soro ou *whey proteins*. As caseínas são importantes na produção do queijo e são a base para os produtos lácteos. As proteínas do soro trazem muitos benefícios para a saúde humana, participando do reparo celular, gerando energia para atividades físicas e proteção para o sistema cardiovascular. O rebanho bovino brasileiro é formado por representantes de raças taurinas (*Bos taurus*), zebuínas (*Bos indicus*) e seus mestiços. Os zebuínos constituem a maior parte do rebanho nacional com mais de 165 milhões de animais entre bovinos de corte e de leite. Dentre os zebuínos leiteiros, as raças Guzerá e Gir são de grande importância para o cenário da agropecuária nacional, contribuindo como produtoras de leite e carne e como fornecedoras de material genético para cruzamentos. A composição do leite varia de acordo com a raça, alimentação e características individuais. Diferentes raças, ao longo de sua evolução, acumularam diferentes conjuntos de variantes genéticas. Algumas destas variantes podem ser funcionais, explicando diferenças individuais ou inter-raciais na composição do leite. Algumas variantes funcionais já foram investigadas, porém, nenhum estudo investigou essas variantes em escala genômica nas raças Guzerá e Gir. Nesse contexto, a presente dissertação teve como objetivos: 1. Identificar, através de análises *in silico*, variantes nos genes que codificam proteínas do leite, em três touros da raça Guzerá e três touros da raça Gir, submetidos a sequenciamento completo do genoma; 2. Prever, novamente através de análises *in silico*, o potencial para repercussão funcional destas variantes, ou seja, identificar quais deveriam ser alvo de investigação em estudos futuros, na busca por variantes, que possam ser responsáveis pelas diferenças na concentração das proteínas do leite entre essas raças e entre as raças zebuínas quando comparadas às taurinas. Como resultado, foram encontradas 70 variantes no *cluster* das caseínas [α S1-caseína (*CSN1S1*), β -caseína (*CSN2*), α S2-caseína (*CSN1S2*) e κ -caseína (*CSN3*)] e nos genes α -lactalbumina (*LALBA*), β -lactoglobulina (*LGB*) e lactoferrina (*LTF*). Desse total, 37 variantes foram descritas pela primeira vez. Para as análises funcionais, foram selecionadas as 37 variantes novas e uma variante em região codante já descrita. Evidências para a funcionalidade foram encontradas para 20 destas 38 variantes: duas são mutações missense, duas variantes afetam sítios de reconhecimento de *splicing* (destas, uma altera também sítio de reconhecimento de miRNA), 15 afetam uma posição evolutivamente conservada (destas, duas alteram também sítios de reconhecimento de miRNA) e uma altera sítio de reconhecimento de miRNA. As mutações *LTF* Thr546Asn e *LALBA* Ile60Val afetam posições nucleotídicas evolutivamente conservadas. Impactos sobre a estrutura 3D da proteína em função da mutação não foram detectados. A mutação no gene *LTF*, na posição BTA2:53548773, cria um sítio de *splicing* alternativo, que levaria a um stop codon prematuro, produzindo uma isoforma de 550 aminoácidos. Assim, com este estudo foi possível identificar 70 variantes nos genes que codificam proteínas do leite nas raças Guzerá e Gir e selecionar 20 variantes potencialmente funcionais para estudos futuros.

Palavras-chave: mutação, bovino, leite, *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*, *LALBA*, *LGB*, *LTF*, Gir, Guzerá

ABSTRACT

Milk is a food of high nutritional value and of great importance in human nutrition, a source of vitamins, minerals (especially calcium), carbohydrates, lipids and proteins. Milk consists of two groups of proteins: a solid fraction, formed predominantly by caseins, and a soluble fraction, formed by whey proteins. Caseins are important in cheese production and are a basis for dairy products. Whey proteins bring many benefits to human health, participating in cell repair, generating energy for physical activities and protection for the cardiovascular system. The Brazilian cattle herd is made up of representatives of Taurine (*Bos taurus*), Zebu (*Bos indicus*) and their crossbred breeds. Zebu cattle produce most of the national herd with more than 165 million animals, including beef and dairy cattle. Among dairy zebu cattle, the Guzerá and Gir breeds are of great importance for the national agricultural scenario, contributing as producers of milk and meat and as suppliers of genetic material for crossbreeding. The composition of the milk varies according to breed, diet and individual characteristics. Different breeds, throughout their evolution, have accumulated different sets of genetic variants. Some of these variants can be functional, explaining individual or interbreed differences in milk composition. Some functional variants have already been investigated, however, no study has investigated these variants on a genomic scale in the Guzerá and Gir breeds. In this context, the present dissertation had as objectives: 1. To identify, through in silico analyzes, variants in the genes that encode milk proteins, in three Guzerá bulls and three Gir bulls, submitted to whole genome sequencing; 2. Predict, using in silico analysis, the potential for functional repercussion of these variants, that is, to identify which ones should be the target of investigation in future studies, in the search for variants, that would respond for the differences in the concentration of milk proteins between these breeds and between Zebu breeds when compared to taurines. As a result, 70 variants were found in the casein cluster [α S1-casein (*CSN1S1*), β -casein (*CSN2*), α S2-casein (*CSN1S2*) and κ -casein (*CSN3*)] and in the α -lactalbumin (*LALBA*) genes, β -lactoglobulin (*LGB*) and lactoferrin (*LTF*). Of that total, 37 variants were described for the first time. For functional analysis, 37 new variants and one variant in the coding region already described were selected. Evidence for functionality were found for 20 of these 38 variants: two are missense mutations, two variants affect splicing recognition sites (of these, one also changes a miRNA recognition site), 15 affect an evolutionarily conserved position (of these, two also change miRNA recognition sites) and one altered miRNA recognition site. *LTF* Thr546Asn and *LALBA* Ile60Val mutations affect evolutionarily conserved nucleotide positions. Impacts on the 3D structure of the protein due to these mutations were not detected. The mutation in the *LTF* gene, at position BTA2:53548773, creates an alternative splicing site, which would lead to a premature stop codon, producing an isoform of 550 amino acids. Thus, with this study it was possible to identify 70 variants in the genes that encode milk proteins in the Guzerá and Gir breeds and select 20 potentially functional variants for future studies.

Key-words: mutation, bovine, milk, *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3*, *LALBA*, *LGB*, *LTF*, Gyr, Guzerá

LISTA FIGURAS

Figura 1 - Composição do cluster das caseínas	16
Figura 2 - Localização do gene LALBA	17
Figura 3 – Localização do gene LGB	18
Figura 4 – Localização do gene LTF	19
Figura 5 – Estrutura da micela das caseínas	20
Figura 6 – Estrutura terciária da proteína beta-lactoglobulina	21
Figura 7 – Estrutura terciária da proteína lactoferrina	22
Figura 8 – Estrutura da ligação com o ferro da lactoferrina bovina	23
Figura 9 – Estrutura terciária da proteína alfa-lactalbumina	24
Figura 10 - Elementos em cis que atuam no splicing	28
Figura 11 - Splicing: motivos conservados no sítio doador (5'ss), ponto de ramificação (branching site) e sítio acceptor (3'ss)	29
Figura 12 - Mecanismos de splicing alternativos	30
Figura 13 - Esquema geral da modelagem de proteínas por homologia	32
Figura 14 - Etapas do protocolo básico da predição de estrutura 3D utilizado pelo programa Modeller	33
Figura 15 - Cobertura em extensão observada para os genes de interesse	43
Figura 16 - Cobertura em profundidade observada para os genes de interesse	44
Figura 17 - Análise de conservação evolutiva para o SNV BTA22:53546045 no gene LTF	51
Figura 18 – Modelos preditos para a proteína LTF	53
Figura 19 – Sítios de glicosilação no gene LTF	54
Figura 20 - Análise de conservação evolutiva para o SNV BTA5:31348386 no gene LALBA	55
Figura 21 – Modelos preditos para a proteína LALBA, selvagem e com a mutação Ile60Val	56
Figura 22 - Análise de conservação evolutiva para o SNV BTA6:87280796 no gene CSN1S2	57
Figura 23 - Posição da SNV BTA6:87280796 no gene <i>CSN1S2</i>	58
Figura 24 - Sítios de splicing alternativo para o gene LTF	59
Figura 25 - Sequência da proteína com o splicing alternativo para a variante BTA22:53534325	59
Figura 26 - Sequência da proteína com o splicing alternativo para a variante BTA22:53538773	60
Figura 27 – Modelos preditos para a proteína LTF, na presença da SNV BTA22:53534325	61
Figura 28 – Modelos preditos para a proteína LTF, na presença da BTA22:53548773	62
Figura 29 – ECRs no gene LALBA	64

LISTA DE TABELAS

Tabela 1 - SNPs identificados por Kang et. al., 2008	26
Tabela 2 - Descrição das amostras utilizadas nesta dissertação	38
Tabela 3 - Espécies utilizadas para as análises de conservação evolutiva e os GI (número de acesso NCBI) de cada uma delas	41
Tabela 4 - Cobertura em profundidade e cobertura em extensão nos genes sobre investigação	45
Tabela 5 - Variações, SNVs e INDELs, encontradas no presente estudo	46
Tabela 6 - Sumário das variantes descobertas com potencial funcional: efeitos sobre sítios de splicing, sítios de reconhecimento de miRNA e posições conservadas evolutivamente	49
Tabela 7 - Resultados da modelagem para os cinco melhores modelos 3D da LTF	52
Tabela 8 - Resultados da modelagem para os cinco melhores modelos 3D	56
Tabela 9 - Resultados da modelagem para os cinco melhores modelos 3D	61

LISTA DE ABREVIATURAS

ABCGil - Associação Brasileira dos Criadores de Gir-leiteiro

ACGB - Associação dos Criadores de Guzerá do Brasil

ASSP – Alternative Splice Site Predictor

BCMs – β -casomorfínas

BTA - *Bos taurus*

CBMG² - Centro Brasileiro de Melhoramento do Guzerá

CSN1S1 – α S1Caseína

CSN1S2 – α S2 Caseína

CSN2 – β -Caseína

CSN3 – κ -Caseína

DOPE – Discrete Optimized Protein Energy

INDELs – Inserções e Deleções

LALBA – α -lactoalbumina

LGB – β -lactoalbumina

LGHM – Laboratório de Genética Humana e Médica

LTF – lactoferrina

PAEP- progesterone-associated endometrial protein, β -lactoalbumina

SNVs – Single Nucleotide Variations

UFMG – Universidade Federal de Minas Gerais

SUMÁRIO

1 INTRODUÇÃO	13
1.1 Os Bovinos no Brasil	13
1.2 Genes envolvidos na lactogênese	14
1.3 Estrutura das Proteínas	19
1.3.1 Caseínas	19
1.3.2 Beta-lactoglobulina	20
1.3.3 Lactoferrina	21
1.3.4 Alfa-lactalbumina	23
1.4 Marcadores moleculares para o melhoramento genético animal	24
1.4.1 SNPs	24
1.4.2 INDELs	25
1.4.3 Polimorfismos nos genes das proteínas do leite	25
1.5 Análises funcionais in silico	26
1.5.1 Sítios de reconhecimento de ligação de miRNAs	26
1.5.2 Sítios de reconhecimento de splicing alternativo	27
1.5.3 Análises de conservação evolutiva	31
1.5.4 Modelagem por homologia	31
1.6 Variantes em genes das proteínas do leite de vaca associadas a doenças humanas	33
1.7 Justificativa e relevância	34
1.8 Objetivos	35
1.8.1 Objetivo geral	35
1.8.2 Objetivos específicos	35
2 MATERIAIS E MÉTODOS	37
2.1 A amostra	37
2.2 Montagem dos genomas	38
2.3 Seleção dos genes	39
2.4 Detecção das variações	39
2.5 Seleção das variantes potencialmente funcionais	40
3 RESULTADOS	43
3.1 Cobertura em extensão e em profundidade obtida para os genes de interesse	43
3.2 Localização, tipo e impacto das variantes genéticas	46

3.3 Análises de impacto das variantes missense	50
3.3.1 Mutação no gene LTF Thr546Asn	50
3.3.2 Mutação no gene LALBA Ile60Val	54
3.4 Análises de impacto das demais variantes	57
3.4.1 Mutação na 3'-UTR do gene CSN1S2	57
3.4.2 Mutação em sítios de splicing no gene LTF	58
3.4.3 SNVs que afetam sítios de reconhecimento a miRNA	62
3.4.4 Conservação evolutiva	62
3.4.5 SNVs em ECRs	63
4 DISCUSSÃO	65
5 CONCLUSÕES	68
6 PERSPECTIVAS	69
6 REFERÊNCIAS BIBLIOGRÁFICAS	70
7 ANEXO 1 – Produção	77
Manuscrito em preparação	77
Resumos publicados em anais de congressos	77
Apresentação de trabalho	77
Premiação	77
Organização de evento	78
8 APÊNDICE	79

1 INTRODUÇÃO

1.1 Os Bovinos no Brasil

Em 2014, o rebanho brasileiro era composto por 212,3 milhões de animais, sendo o segundo maior rebanho do mundo e o quinto maior produtor mundial de leite de vaca produzindo, aproximadamente, 35 bilhões de quilos de leite por ano (IBGE 2015).

O rebanho brasileiro é formado por representantes de raças taurinas (*Bos taurus*), zebuínas (*Bos indicus*) e seus mestiços. Os zebuínos constituem a maior parte do rebanho nacional com mais de 165 milhões de animais, entre bovinos de corte e leite. Isto se deve a sua melhor adaptação ao clima tropical, maior resistência a parasitas e maior capacidade de aproveitamento de alimentos de baixo valor nutritivo (Jorge, Issa et al. 2006, DIAS and HORIZONTE 2008). Dentre os zebuínos leiteiros, as raças Guzerá e Gir são de grande importância para o cenário da agropecuária nacional, contribuindo como produtoras de leite e carne e como fornecedoras de material genético para cruzamentos (Paneto, Ferraz et al. 2008)

A raça Guzerá foi a primeira raça zebuína introduzida no Brasil. Representantes da raça foram trazidos da Índia no final do século XIX, importados por criadores de Recife, Salvador e Rio de Janeiro, para a produção de carne e leite (Silva, Packer et al. 2006; ACGB, acesso em: 20 de novembro de 2014). Atualmente, existem rebanhos por todo o país com predomínio na região Nordeste, povoando desde áreas férteis litorâneas, passando pelo agreste, até o sertão semi-árido (ACGB Acesso em: 20 de novembro de 2014). Essa raça é caracterizada pela sua rusticidade, dispensa suplementação alimentar, baixa susceptibilidade à mastite e boa habilidade materna. Além disto, se caracteriza por chifres desenvolvidos e simétricos, curvando-se para cima, com formato de lira.

No Brasil, a maioria do gado Guzerá foi selecionado para carne. Uma parcela destes animais foi submetida também à seleção para leite, originando animais de dupla aptidão (Penna, Peixoto et al. 2006; Faria, Vercesi Filho et al. 2004; Leão, Pivatto et al. 2013; ACGB, acesso em: 20 de novembro de 2014). O Brasil é um dos maiores centros criadores de Guzerá do mundo e um grande fornecedor de material genético da raça para outros países. Além de importante recurso genético como raça pura, o Guzerá tem sido uma importante opção nos cruzamentos para produção de mestiços leiteiros (Penna, Peixoto et al. 2006). As linhagens leiteiras também são de grande porte, já tendo sido registrado fêmeas com produção leiteira

de 4.500 kg na lactação (Penna, Peixoto et al. 2006; Leão, Pivatto et al. 2013; ACGB, acesso em: 20 de novembro de 2014).

A raça Gir é originária do Parque Nacional da floresta de Gir, na Índia Ocidental, no sul da península Kathiawar. A primeira vaca Gir foi introduzida no Brasil em 1906, na região do Triângulo Mineiro e muitos touros foram importados entre 1914 e 1921, sendo mantidos em Minas Gerais e São Paulo (Queiroz e Lôbo 1993; Leão, Pivatto et al. 2013). A raça é caracterizada pelos chifres laterais frequentemente retorcidos, cupim de tamanho volumoso nos machos e facilidade de manejo. Ocupa o segundo lugar no controle leiteiro oficial no Brasil, sendo considerada uma raça altamente adaptada às condições brasileiras com uma produção de cerca de 3.233kg/305 dias de lactação (Leão, Pivatto et al. 2013).

As raças Guzerá e Gir fazem parte do Programa Nacional de Melhoramento Genético de Zebuínos (ABCZ e EMBRAPA). Estes programas têm como objetivos integrar modernas ferramentas de melhoramento animal levando a uma seleção confiável e rápida por meio da identificação e seleção de touros geneticamente melhores para as características de manejo e produção (Peixoto, Tomita Bruneli et al. 2014; VERNEQUE, PANETTO et al. 2014)

1.2 Genes envolvidos na lactogênese

A lactogênese é o início da formação da secreção láctea e compreende o período que vai desde o começo ou recomeço da atividade mamária até a abundante secreção do leite. Este processo inclui dois estágios. O estágio I corresponde ao período compreendido entre a retomada da atividade da glândula mamária até o parto. É caracterizado pela diferenciação mamária, proliferação, progressiva expressão de proteínas do leite e secreção do pré-colostro. O estágio II se inicia no momento do parto e se estende até o final da lactação. Neste estágio, ocorre a formação e secreção do colostro e, subsequentemente, do leite (Gao, Lin et al. 2013). Portanto, durante a lactogênese, vários genes são diferencialmente regulados (*super- e/ou subexpressão*), para que ocorra o desenvolvimento da glândula mamária e a produção do colostro e do leite, assim como para a determinação da composição do leite (Gao, Lin et al. 2013).

O leite de vaca é um alimento de elevado valor nutritivo e de grande importância na alimentação humana, sendo fonte de vitaminas, minerais, carboidratos, lipídios e proteínas. A composição do leite varia de acordo com a raça, alimentação, saúde, período de lactação,

idade, características individuais, entre outras (González, Dürr et al. 2001, Venturini, Sarcinelli et al. 2007).

O leite produzido por vacas zebuínas apresenta maior teor de lactose e porcentagem de gordura e proteínas do que o leite de algumas raças taurinas (González, Dürr et al. 2001, Hax 2009). Desta forma, entender as variantes genéticas por trás dessas diferenças na composição do leite poderia auxiliar na produção de um leite mais saudável e menos alergênico para a dieta humana (FONTANELI 2001; González, Dürr et al. 2001).

Vários estudos de expressão gênica vêm sendo realizados com o intuito de identificar quais são os genes envolvidos nos estágios da lactogênese (Ikonen, Ojala et al. 1999, Azevedo, Nascimento et al. 2008; Gao, Lin et al. 2013).

No presente trabalho, foram analisados os genes diferencialmente expressos nos dois estágios da lactação, identificados por Gao e colaboradores (2013): *CSN1S1* (α S1-caseína), *CSN2* (β -caseína), *CSN1S2* (α S2-caseína), *CSN3* (κ -caseína) e *LALBA* (α -lactoalbumina). Além desses, *LGB* (β -lactoalbumina) e *LTF* (lactoferrina) também foram analisados. O gene *LGB* foi selecionado por ter associação com a produção, composição e propriedades de coagulação do leite (Cecchinato, Ribeca et al. 2012); *LTF* foi selecionado pela sua atividade antibacteriana, antiviral, antifúngica, além de ser marcador para susceptibilidade de mastite (Kang, Li et al. 2008).

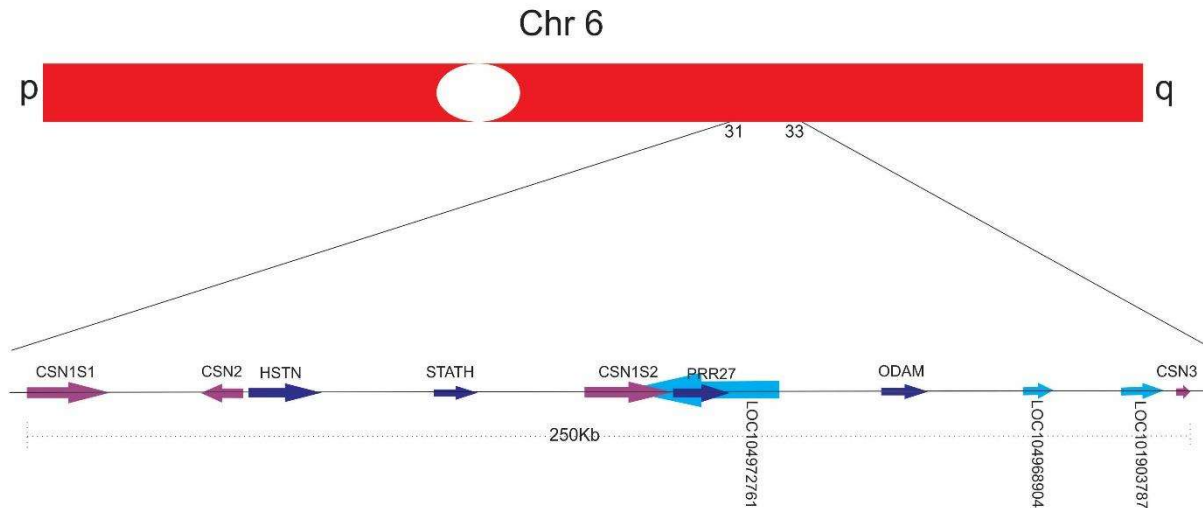
O leite é composto por dois grupos de proteínas: as caseínas e as proteínas do soro (*whey proteins*) (Haraguchi, ABREU et al. 2006, Kamiński, Cieślińska et al. 2007), α S1-caseína (*CSN1S1*), β -caseína (*CSN2*), α S2-caseína (*CSN1S2*) e κ -caseína (*CSN3*), constituindo 38, 36, 10 e 13% do total de proteínas no leite, respectivamente (Silva and Malcata 2005; Sulimova, Azari et al. 2007; Azevedo, Nascimento et al. 2008; Nilsen, Olsen et al. 2009). As proteínas do soro constituem aproximadamente 20% das proteínas do leite e são constituídas por α -lactoalbumina (*LALBA*), β -lactoalbumina (*LGB*), albumina do soro bovino (*BSA*), imunoglobulinas (*IGs*), glicomacropéptidos (*GMP*), lactoferrina (*LTF*) e outras proteínas de menor abundância (Haraguchi, ABREU et al. 2006; Kamiński, Cieślińska et al. 2007).

Os genes das caseínas formam um *cluster* em BTA6q31-33 (no cromossoma 6 de taurinos (BTA), região 6q31-33), na seguinte ordem *CSN1S1*, *CSN2*, *CSN1S2* e *CSN3*, e abrange uma extensão total de 250 kb, sendo que o gene *CSN3* está a uma distância de 70kb dos outros (Sulimova, Azari et al. 2007; Trakovická, Moravčíková et al. 2012).

Entre os genes *CSN2* e *CSN1S2* tem os genes *histatherin* (*HSTN*) e *statherin* (*STATH*). Entre os genes *CSN1S2* e *CSN3* se localizam os genes *proline-rich protein 27 precursor* (*PRR27*)

e *odontogenic, ameloblast associated (ODAM)*, além das LOC104972761, LOC104968904 e LOC101903787, como pode ser observado na Figura 1.

Figura 1 - Composição do cluster das caseínas

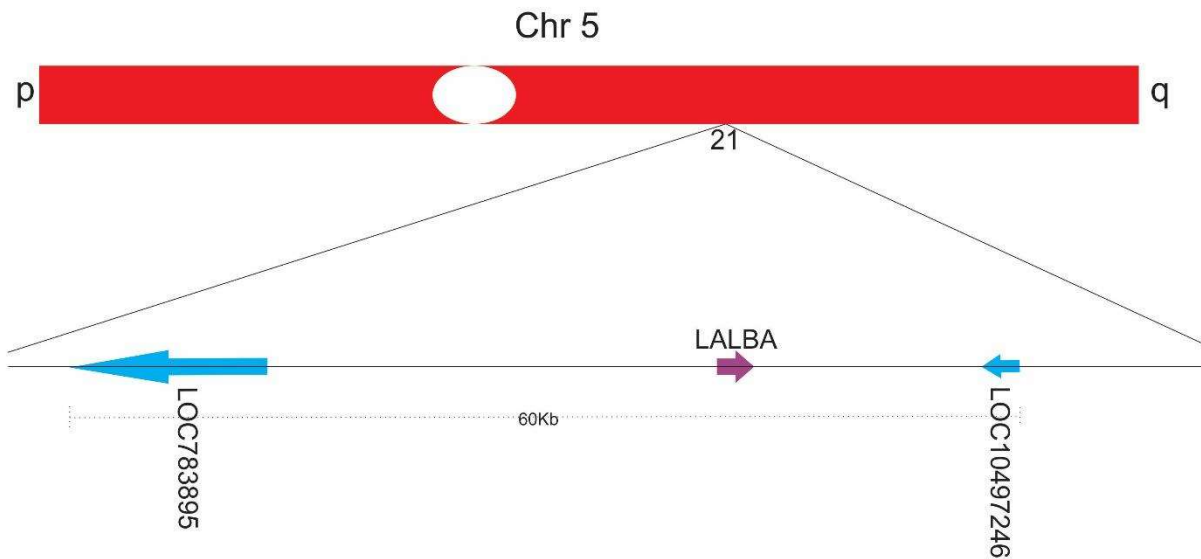


O *cluster* se localiza no BTA6q31-33. Em roxo, podemos ver os genes das caseínas que seguem a ordem *CSN1S1*, *CSN2*, *CSN1S2* e *CSN3*. Em azul escuro os genes *HSTN*, *STATH*, *PRR27* e *ODAM* e em azul claro as LOC100298118, LOC101903787 e LOC104968905.

Fonte: Produzida pela autora, usando imagem do Genome Browser da UCSC, acesso em, 2015.

O *LALBA* está localizado no BTA5q21 e possui 3.090 pb. *LALBA* é uma das subunidades da lactose-sintetase (a outra é a galactosil-transferase), enzima responsável pela síntese da lactose, que é o principal açúcar do leite, além de constituir um dos seus fatores osmóticos (González, Dürr et al. 2001; Haraguchi, ABREU et al. 2006; Kamiński, Cieślińska et al. 2007; VĂTĂSESCU, Georgescu et al. 2008; FREITAS 2010). Na região *upstream* e *downstream* desse gene se encontram as LOC783895 e LOC10497246, respectivamente, como pode ser observado na Figura 2.

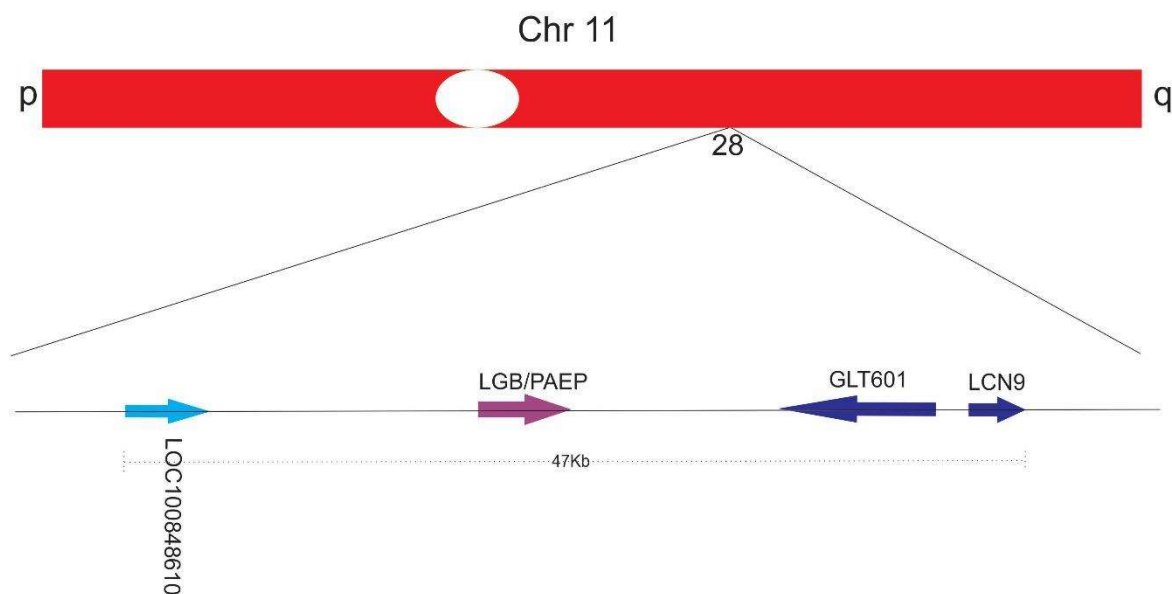
Figura 2 - Localização do gene *LALBA*



O gene se localiza no cromossomo 5, região q21. Em roxo podemos ver o gene *LALBA*. Em azul claro as *LOC783895* e *LOC10497246*.

Fonte: Produzida pela autora, usando imagem do Genome Browser da UCSC, acesso em 2015.

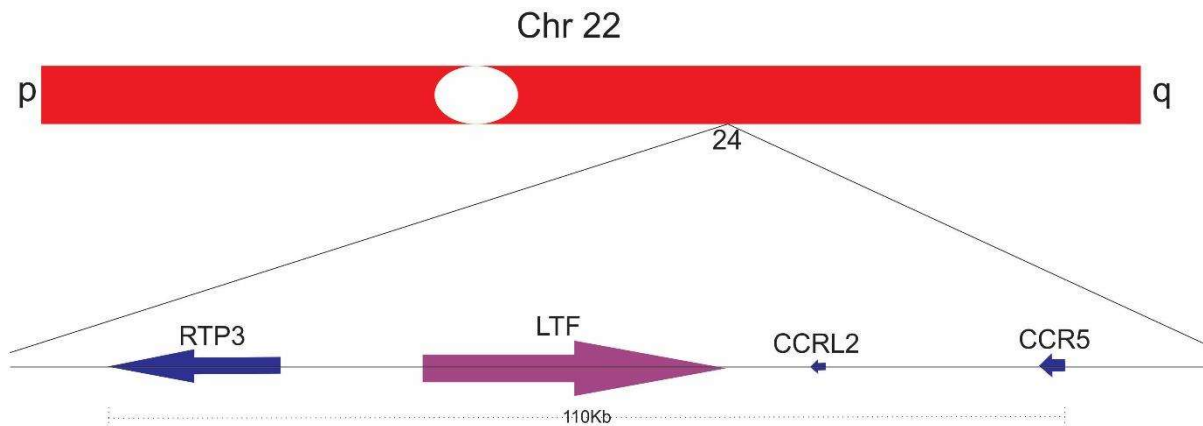
O *LGB*, também chamado *PAEP* ou *BGL*, está localizado em BTA11q28 (no cromossomo 11 em BTA, região q28) e possui 4.718pb (NCBI). Próximos a esse gene estão localizados a *LOC100848610* e dois genes *glycosyltransferase 6 domain containing 1* (*GLT6D1*) e *lipocalin 9* (*LCN9*), como pode ser observado na Figura 3. É interessante observar que estes outros genes também se expressão na glândula mamária em lactação.

Figura 3 – Localização do gene *LGB*

O gene se localiza no cromossomo 11, região q28. Em roxo podemos ver o gene *LGB*, em azul escuro os genes *GLT6D1* e *LCN9*, e em azul claro a *LOC100848610*.

Fonte: Produzida pela autora, usando imagem do Genome Browser da UCSC, acesso em 2015.

O gene *LTF* está localizado em BTA22q24 (no cromossoma 22 em BTA, região q24) e consiste de 17 exons (NCBI; Dinesh, Verma et al. 2015). Trata-se de uma proteína bioativa presente em vários fluídos (leite, lágrimas, saliva, plasma seminal e secreção nasal) (Zheng, Ather et al. 2005; González-Chávez, Arévalo-Gallegos et al. 2009; Fang, Zhang et al. 2014; Dinesh, Verma et al. 2015). Esta proteína possui várias atividades, como, por exemplo, atividade antimicrobiana; anti-inflamatória; anticarcinogênico; antibacteriana; antiviral, tendo uma forte atividade contra HIV em estudos *in vitro*; antifúngica; antiparasitária. Além disto, tem sido associada à resistência à mastite (González-Chávez, Arévalo-Gallegos et al. 2009; Dinesh, Verma et al. 2015). Na vizinhança desse gene estão localizados três genes que são *receptor transporter protein 3 (RTP3)*, *C-C motif chemokine receptor like 2 (CCRL2)* e *C-C motif chemokine receptor 5 (CCR5)*, como pode ser observado na Figura 4. É interessante notar que o receptor que permite a internalização do vírus HIV (*CCR5*) se localiza nesta mesma região.

Figura 4 – Localização do gene *LTF*

O gene se localiza no cromossomo 22, região q24. Em roxo podemos ver o gene *LTF* e em azul escuro os genes *RTP3*, *CCRL2* e *CCR5*.

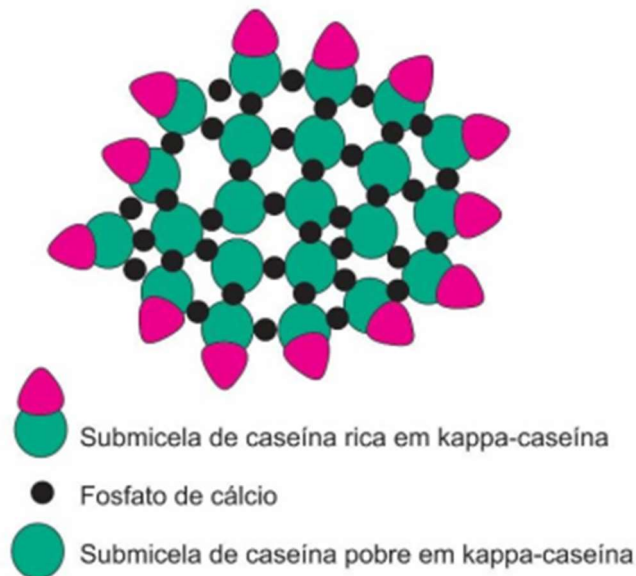
Fonte: Produzida pela autora, usando imagem do Genome Browser da UCSC, acesso em 2015.

1.3 Estrutura das Proteínas

1.3.1 Caseínas

As caseínas pertencem à família de fosfoproteínas e no leite estão presentes em formas de micelas, que transportam uma grande porção de minerais. As caseínas são divididas em dois grupos, cálcio sensível e cálcio não sensível. Vários modelos da estrutura das micelas das caseínas foram propostos e podemos observar o mais frequente na Figura 5 (Horne 2002; Horne 2006).

Figura 5 – Estrutura da micela das caseínas



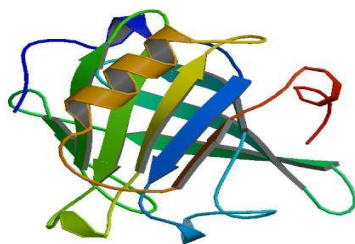
Nesse modelo as caseínas são formadas subunidades de 15 a 20 moléculas por interação hidrofóbica. A parte rica em CSN3 forma a superfície do micela e a parte pobre ou sem CSN3 fica localizado dentro do micela.

Fonte: Adaptado de Horne 2006.

1.3.2 Beta-lactoglobulina

A beta-lactoglobulina é uma proteína globular e pertence a família das lipocalinas. A cadeia da proteína é composta por 162 aminoácidos e tem peso molecular de 18,4 kDa (Figura 6) (FREITAS, De Almeida 2013). Possui um domínio de semelhante a calicina, que está localizado entre os aminoácidos 21 a 176 (Finn, Bateman et al. 2013; Mitchell, Chang et al. 2014). Os peptídeos derivados da beta-lactoglobulina tem como atividade biológica: hipocolesterolêmica, imunoestimulante, antimicrobiana, antioxidante e anti-hipertensiva (De Almeida 2013).

Figura 6 – Estrutura terciária da proteína beta-lactoglobulina



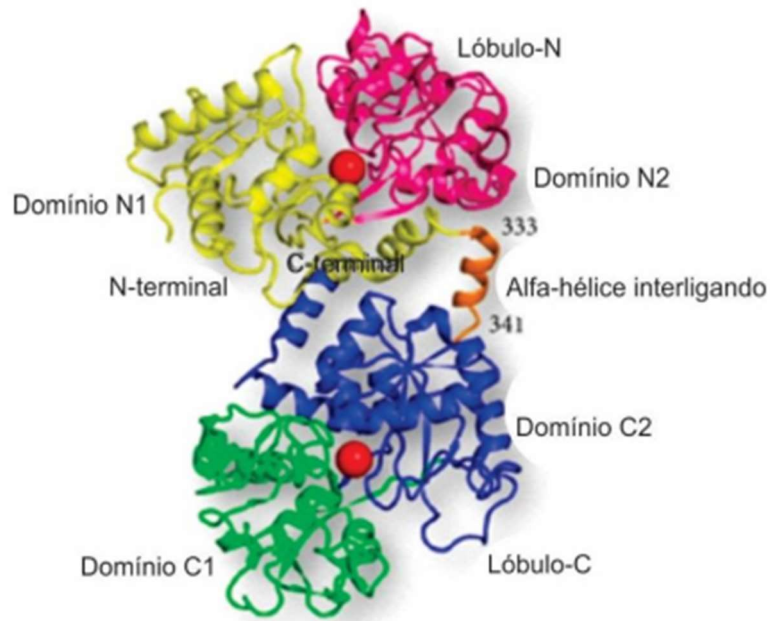
É uma proteína globular composta de 162 aminoácidos, com apenas uma alfa-hélice, e que pode ser encontrada na forma de dímeros ou de monômeros. Sua estrutura apresenta mudanças de acordo com a faixa de pH e temperatura.

Fonte: PDB e Freitas, 2010.

1.3.3 Lactoferrina

A lactoferrina é uma proteína globular composta por dois lóbulos (N- e C-terminal) de tamanho similar e com 40% de similaridade entre eles. A cadeia da proteína é composta por 708 aminoácidos e tem peso molecular de 40 kDa, sendo que do aminoácido 1 ao 19 há um peptídeo sinal e nos aminoácidos 36 a 60 possui um domínio lactoferricina-B (UNIPROT Consortium 2014). Essa proteína está presente em fluidos secretores como o leite, lágrimas, saliva, plasma seminal, e secreções nasais (Zheng, Ather et al. 2005; O'Halloran, Bahar et al. 2009; Fang, Zhang et al. 2014). Um modelo da estrutura dessa proteína pode ser observado na Figura 7. Ela possui dois domínios de transferrina, o primeiro está localizado entre os aminoácidos 25 a 363, e o segundo domínio está localizado entre os aminoácidos 364 a 697 (Finn, Bateman et al. 2013; Mitchell, Chang et al. 2014).

Figura 7 – Estrutura terciária da proteína lactoferrina

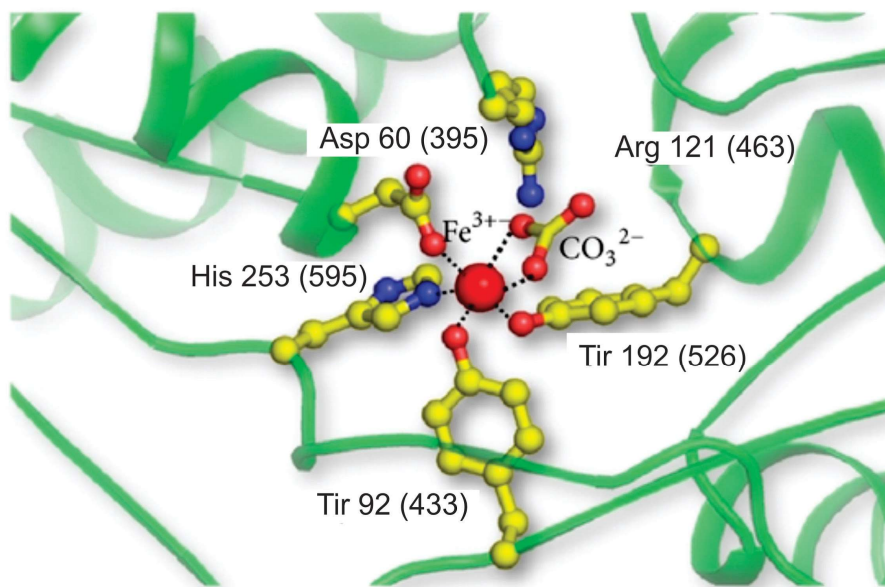


Nesse modelo podemos observar os dois lóbulos da proteína, em amarelo e rosa o N-terminal e em azul e verde o C-terminal. Em laranja podemos ver a alfa-hélice que conecta os dois lóbulos e em vermelho as representações dos íons ferro.

Fonte: Adaptado de Sharma et al., 2013.

A ligação com o íon ferro é composta por quatro resíduos: duas tirosinas, um aspartato e uma histidina. Na parte N-terminal se localizam os aminoácidos 60, 92, 192 e 253 e na parte C-terminal os aminoácidos 395, 433, 526 e 595. Essa ligação é coordenada pelo ferro e o ânion carbonato (Figura 8) (Sharma, Sinha et al. 2013).

Figura 8 – Estrutura da ligação com o ferro da lactoferrina bovina



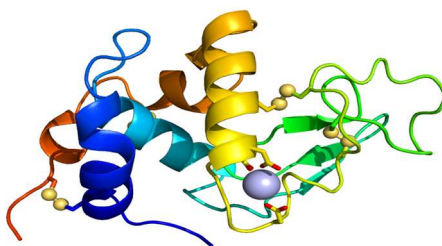
O íon ferro é mostrado em vermelho (círculo grande), os aminoácidos que participam da ligação com o ferro em amarelo. O ânion carbonato estão indicados em amarelo e vermelho (círculos pequenos). Os círculos azuis representam os íons nitrogênio dos aminoácidos. Cada molécula de lactoferrina possui dois sítios de ligação ao ferro, um no N-terminal e outro no C-terminal. Para simplificar a representação, os dois sítios são mostrados juntos: para cada aminoácido, que participa da ligação, é mostrado, fora dos parênteses, a posição na parte N-terminal e, dentro dos parênteses, a posição na parte C-terminal.

Fonte: Adaptado de Sharma, Sinha et al. 2013

1.3.4 Alfa-lactalbumina

A alfa-lactalbumina é uma proteína globular composta por dois lóbulos, uma região rica em alfa-hélices e uma região com duas pequenas folhas-beta, uma pequena alfa-helix e um sítio de ligação ao íon cálcio, como podemos observar na Figura 9 (Hendrix, Griko et al. 1996; Chaudhuri and Chattopadhyay 2014; Hamborg, Rose et al. 2014; Nadeem, Sanborn et al. 2015). A cadeia da proteína é composta por 142 aminoácidos com peso molecular de 14.2 kDa, sendo que do aminoácido 1 ao 19 está localizado um peptídeo sinal (Consortium 2014). Possui um domínio de lisozima que está localizado entre os aminoácidos 21 a 141 (Finn, Bateman et al. 2013; Mitchell, Chang et al. 2014).

Figura 9 – Estrutura terciária da proteína alfa-lactalbumina



Em lilás está representado o íon cálcio e os círculos amarelos estão representadas as ligações dissulfeto. As folhas-beta são mostradas em verde, e as alfa-hélices em azul escuro, azul claro, amarelo e laranja.

Fonte: Jøhnke and Petersen 2012.

1.4 Marcadores moleculares para o melhoramento genético animal

Marcadores moleculares são diferentes classes de variações presentes no DNA ou na proteína. Quando a variante mais frequente apresenta uma frequência de até 99%, são classificados como polimorfismos genéticos. Os polimorfismos genéticos identificados podem ser utilizados para caracterização genética e são associados com características de interesse agropecuário. Sendo importantes tanto para programas de conservação e de melhoramento animal (FREITAS XXX).

1.4.1 SNPs

Os SNPs (do Inglês, *single nucleotide polymorphisms*) são mutações em apenas uma base da cadeia nitrogenada, podendo ocorrer transições ou transversões. As transições são quando ocorre a troca de uma purina por uma purina ($A \leftrightarrow G$) ou de uma pirimidina por uma pirimidina ($T \leftrightarrow C$). Já as transversões são quando ocorre a troca de uma purina por uma pirimidina (A ou $G \leftrightarrow T$ ou C). Uma posição nucleotídica polimórfica pode ter de dois a quatro alelos. Entretanto, quanto mais alelos em uma posição nucleotídica, maior a chance de que se trate de erro de sequenciamento. Consequentemente, para as análises são usados geralmente apenas os SNPs bi-alélicos (Caetano 2009). SNPs podem ocorrer e ser funcionais em qualquer região do genoma, não apenas no entorno dos genes. A maioria deles é

evolutivamente neutra, mas podem ser usados como etiquetas para uma região, permitindo a detecção do efeito de variantes genéticas não avaliadas.

1.4.2 INDELs

INDELs são pequenas inserções ou deleções que ocorrem no genoma, cujo comprimento pode variar de 1 a 10.000 bases. A quantidade de INDELs e o tamanho detectado varia entre as espécies e os indivíduos. Assim como SNPs e variações estruturais maiores, INDELs são muito importantes, pois podem causar alterações fenotípicas (Mills, Luttig et al. 2006; Mullaney, Mills et al. 2010). Por exemplo, em um estudo foram identificadas 262 INDELs em 36 doenças humanas, sendo que mais da metade delas não são múltiplas de 3 bp, alterando assim janelas de leituras (*frameshift mutations*) e, conseqüentemente, a sequência das proteínas (Mills, Luttig et al. 2006).

1.4.3 Polimorfismos nos genes das proteínas do leite

Os genes *CSN1S1* e *CSN1S2* têm poucas variantes alélicas já descritas e, estudos de polimorfismos de proteínas em *CSN2* mostram que, em bovinos, existem pelo menos 11 variantes alélicas, sendo os alelos A1, A2 e B os mais frequentes. Muitas variantes alélicas foram encontradas também na proteína *CSN3*, sendo os alelos A e B os mais frequentes (Sulimova, Azari et al. 2007; FREITAS 2010; Trakovická, Moravčíková et al. 2012).

No gene da *LALBA*, polimorfismos foram identificados na região codificadora. Um deles envolve a substituição de uma adenina por uma guanina, gerando as variantes A e B. Na posição 10 da proteína, a variante A da *LALBA* contém uma glutamina e a B, uma arginina. A mais comum dessas variantes, tanto no Gir quanto no Guzerá, é a variante B (FREITAS 2010).

No gene *LGB*, 11 polimorfismos já foram identificados, sendo que as variantes A e B são as mais comuns (Ganai, Bovenhuis et al. 2009). No gene *LTF*, Kang e colaboradores, 2008 identificaram 11 SNPs (Tabela 1) missense associados a efeitos genéticos no gene *LTF* (Kang et. al., 2008).

Tabela 1 - SNPs identificados por Kang et. al., 2008

Gene	Variations
<i>LTF</i>	Val45Ile
<i>LTF</i>	Asp66Pro
<i>LTF</i>	Ala67Gly
<i>LTF</i>	Leu164Pro
<i>LTF</i>	Gln165Pro
<i>LTF</i>	Phe297Ser
<i>LTF</i>	Gly340Ala
<i>LTF</i>	Ser436Thr
<i>LTF</i>	His439Tyr
<i>LTF</i>	Ala514Arg
<i>LTF</i>	His632Arg

Em sua grande maioria, estas variantes foram descritas em raças taurinas. Muitas delas existem e são polimórficas também em raças zebuínas, mas variantes próprias das raças zebuínas têm sido menos investigadas.

1.5 Análises funcionais in silico

1.5.1 Sítios de reconhecimento de ligação de miRNAs

Os miRNAs foram descobertos em 1993, porém apenas em 2001 foram reconhecidos como sendo uma grande classe de pequenos RNAs regulatórios. Os miRNAs são classificados como: (1) miRNAs intergênicos, se localizam entre dois genes e tem a expressão regulada por promotores próprios; (2) miRNAs intrônicos, se localizam dentro de íntrons de genes e geralmente usam os mesmos promotores do gene ao qual pertencem; e, (3) miRNAs exônicos, que são menos frequentes, geralmente se localizam no final dos éxons e geralmente usam os mesmos promotores do gene onde estão localizados (Pereira 2015).

Além dessas localizações, miRNA em região promotora podem como inibidor da expressão do gene (Young e Corey, 2011). Além disso, esses autores mostraram que miRNAs podem também reconhecer ncRNAs em promotores de genes overlapping e regular a transcrição (Young e Corey 2011).

Existem diferentes mecanismos de ação de miRNAs, como: (1) inibição da tradução; (2) desestabilização do RNA - alvo; (3) silenciamento transcricional; (4) promoção da transcrição; e, (5) mecanismos de aumento da eficiência da tradução (Pereira 2015).

Por exemplo, Zhang e colaboradores, em 2016, investigaram quais mecanismos estariam por trás do efeito do miRNA let-7d no silenciamento do gene *DRD3* e encontraram uma região altamente complementar entre o miRNA e a região promotora do gene que possivelmente determinaria a repressão da transcrição do gene por um pequeno RNA duplo. As análises dos genes mostraram que essa complementaridade ocorre na região *upstream* ao gene *DRD3*, sugerindo que regiões complementares a sequências de miRNAs na região promotora do gene podem induzir o silenciamento do gene durante a transcrição em células mamárias (Zhang, Cheng et al. 2016).

Os mecanismos de “inibição da tradução” ocorrem devido a vários fatores: (1) competição pelo 5'CAP; (2) inibição da montagem dos ribossomos; (3) deadenilação seguida por inibição da iniciação da tradução; (4) desmontagem prematura de ribossomos; (5) redução da velocidade de elongação; e, (6) proteólise durante a fase de elongação (Pereira 2015).

Mecanismos de “desestabilização do RNA alvo” pode ocorrer devido a: (1) Clivagem do RNA-alvo; e, (2) deadenilação seguida pela remoção do 5'CAP (Pereira 2015).

Apesar de serem menos conhecidos, os mecanismos de silenciamento transcricional podem ocorrer devido a processos diferentes, como: (1) metilação do DNA; e, (2) modificações da cromatina (Pereira 2015).

Diante da importância funcional dos miRNAs no processo transcricional e pós-transcricional, particularmente na retirada dos éxons e na geração de isoformas do RNA, no presente trabalho foi investigada a existência de alterações em sítios de reconhecimento de ligação de miRNAs, nos genes das proteínas do leite de vaca.

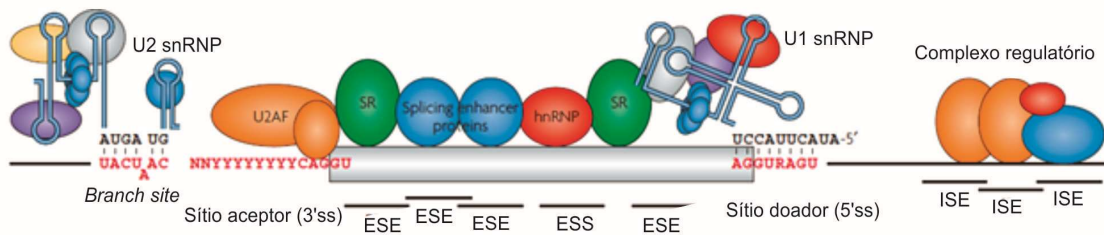
1.5.2 Sítios de reconhecimento de *splicing* alternativo

O *splicing* alternativo é um processo onde são produzidos diferentes mRNAs maduros, levando à produção de isoformas de proteínas com funções diversas e até antagônicas

(Garcia-Blanco et al., 2004; Wang e Cooper, 2007; Wang e Burge, 2008). Os elementos em cis, que atuam no *splicing* são os sítios doadores (5'ss), sítios aceptores (3'ss), o sítio de ramificação (*branching site*), acentuadores de *splicing* exônico (ESE), silenciadores de *splicing* exônico (ESS), acentuadores de *splicing* intrônicos (ISE), e silenciadores de *splicing* intrônicos (ISS) (FIG. 10) (Garcia-Blanco et al., 2004; Wang e Cooper, 2007; Wang e Burge, 2008).

No processo de retirada dos íntrons, os sítios doadores (5'ss) e o *branching site* são ligados por pequenas ribonucleoproteínas (snRNPs) U1 e U2, respectivamente. Mutações no pré-mRNA, que afetem essa ligação, diminuem a eficiência do reconhecimento do exon. Os exons e os íntrons contêm elementos de *enhancer* (acentuadores), que irão regular o *splicing*, podendo potencializar ou reprimir (Garcia-Blanco et al., 2004; Wang e Cooper, 2007; Wang e Burge, 2008).

Figura 10 - Elementos em cis que atuam no *splicing*

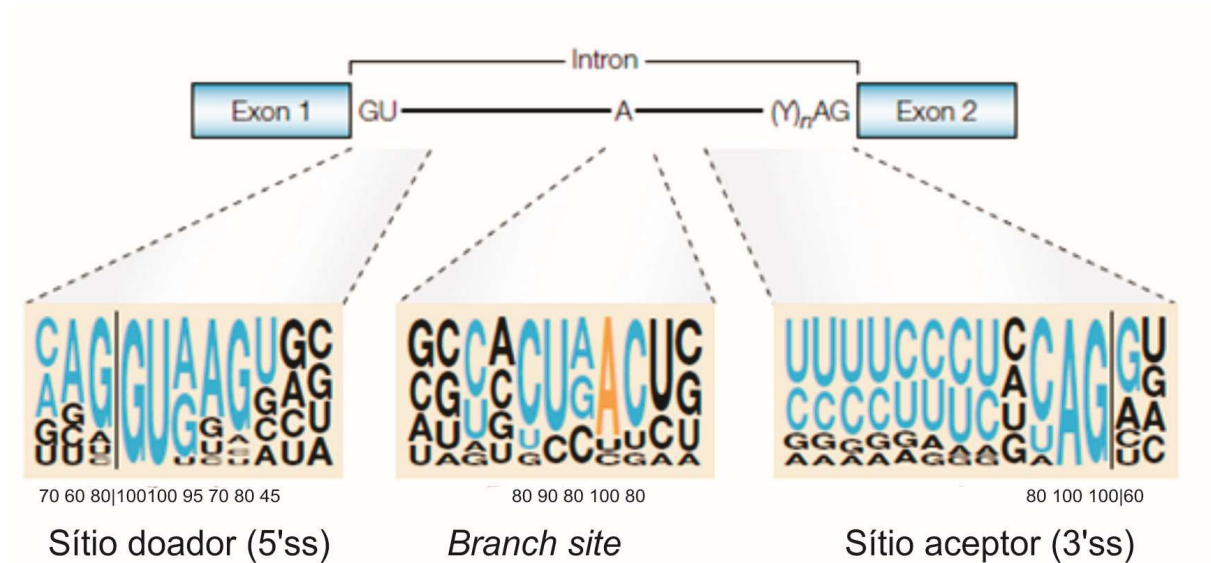


Nota: Sítios doadores (5'ss), sítios aceptores (3'ss), branch site, acentuadores de *splicing* exônico (ESE), silenciador de *splicing* exônico (ESS), acentuadores de *splicing* intrônicos (ISE), silenciador de *splicing* intrônicos (ISS), pequenas ribonucleoproteínas (snRNPs), ribonucleoproteínas heterogêneas nucleares (hnRNP).

Fonte: Adaptado de Wang e Cooper, 2007

Existe uma alta similaridade entre os sítios de doadores, sítio de ramificação, e aceptores de *splicing* em mamíferos, roedores, primatas e outros vertebrados (Shapiro e Senapathy, 1987; Senapathy et al., 1990). Na FIG. 11 são mostrados os *motifs* conservados do sítio doador (5'ss), ponto de ramificação (branch site) e sítio aceptor (3'ss). Os dinucleotídeos GU e AG são quase invariáveis, a sequência de polipirâmidas (Y) precedendo o AG no sítio aceptor é bem conservada e o nucleotídeo A no branch site (Cartegni et al., 2002).

Figura 11 - *Splicing*: motivos conservados no sítio doador (5'ss), ponto de ramificação (*branching site*) e sítio aceptor (3'ss)

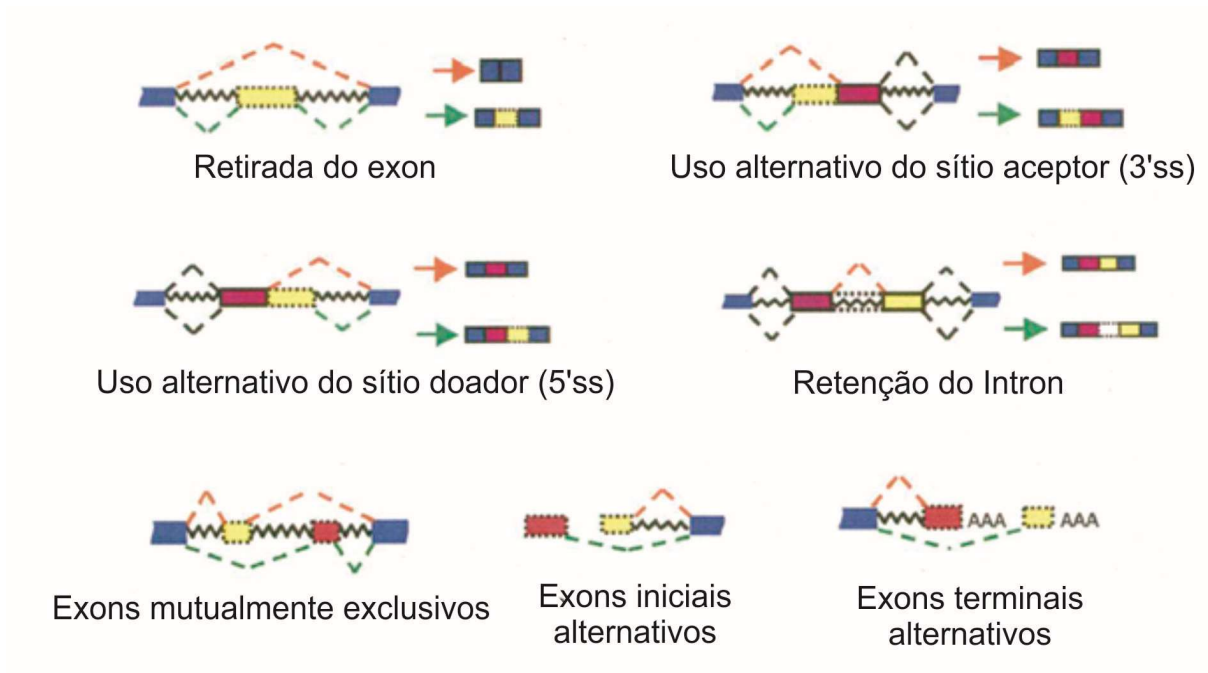


Nota: Em azul são mostrados os nucleotídeos que fazem parte do consenso clássico, exceto o ponto de ramificação (*branching site*), que é mostrado em laranja. Em baixo são mostrados as frequências de conservação desses nucleotídeos.

Fonte: Adaptado de Cartegni, Chew et al. 2002

Os mecanismos de *splicing* alternativo podem ser classificados como: retenção de intron, uso alternativo dos sítios de *splicing* 5'ss/3'ss, retirada de exon e exons mutualmente exclusivos (FIG. 12) (Durham, 2007; Wang e Burge, 2008).

Figura 12 - Mecanismos de *splicing* alternativos



Nota: Na retirada do exon, um exon, que está presente em uma sequência, não está na outra. No uso alternativo dos sítios de *splicing* 5'ss/3'ss, uma parte do intron é incluído na sequência. Na retenção de intron, uma parte de um intron, que não está presente normalmente na sequência do RNAm, passa a ser incluído. Nos exons mutuamente exclusivos, as isoformas que incluem um exon, não incluem o outro e vice-versa.

Fonte: Adaptado de Wang and Burge 2008.

Vários estudos têm mostrado que variantes podem alterar sítios de *splicing*, gerando isoformas potencialmente funcionais (Bongiorni et al., 2016; Liu et al., 2016; Wohlrres-Viana et al., 2016; Zhang et al., 2017). Bongiorni e colaboradores, em 2016, estudando maciez de carne em bovinos, encontraram genes associados a características de produção da carne, que sofrem *splicing* alternativo devido a variantes genéticas (Bongiorni et al., 2016). Liu e colaboradores, em 2016, estudando qualidade de sêmen em bovinos, encontraram SNPs que possivelmente estão envolvidos com características morfológicas normais e da espermatogênese, que sofrem *splicing* alternativo afetando sítio de ligação a miRNA (Liu et al., 2016).

Diante da importância funcional do *splicing* alternativo e do *branching site*, no presente trabalho foi verificado a existência de alterações em sítios de reconhecimento de *splicing* alternativo e em *branch site*.

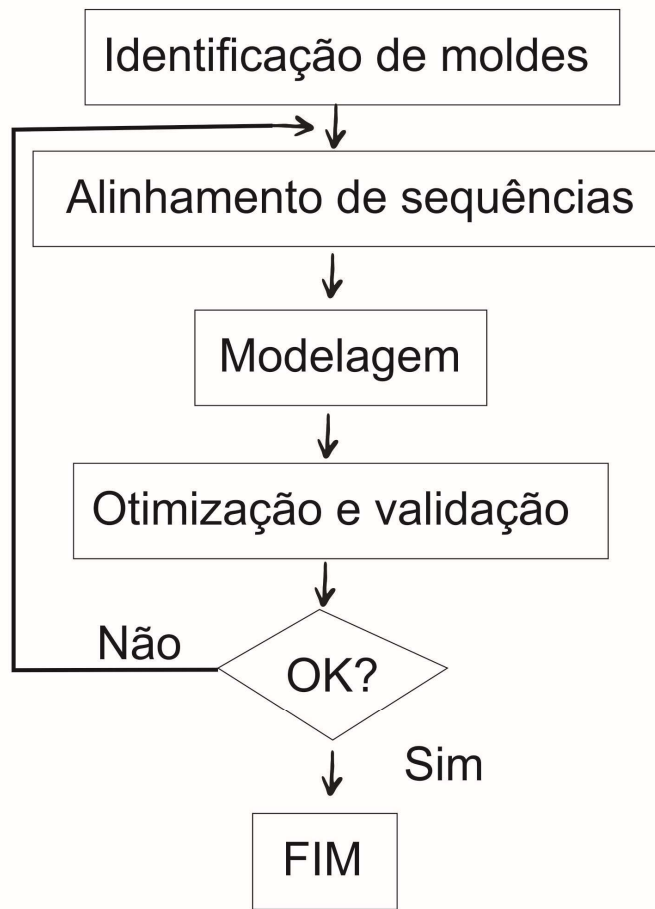
1.5.3 Análises de conservação evolutiva

Estudos sobre a conservação de sequências fora da região transcricional podem indicar a presença de elementos biológicos significativos, como por exemplo, envolvendo regulação distal de genes.

1.5.4 Modelagem por homologia

Modelagem por homologia ou comparativa é uma metodologia, que permite prever a estrutura da proteína, utilizando-se como moldes proteínas com sequências e estruturas secundárias similares à proteína-alvo. A Figura 13 mostra o esquema do processo, existindo vários métodos e programas para serem utilizados em cada um dos passos (Santos Filho and Alencastro 2003; Cavasotto and Phatak 2009; Sodero 2011).

Figura 13 - Esquema geral da modelagem de proteínas por homologia

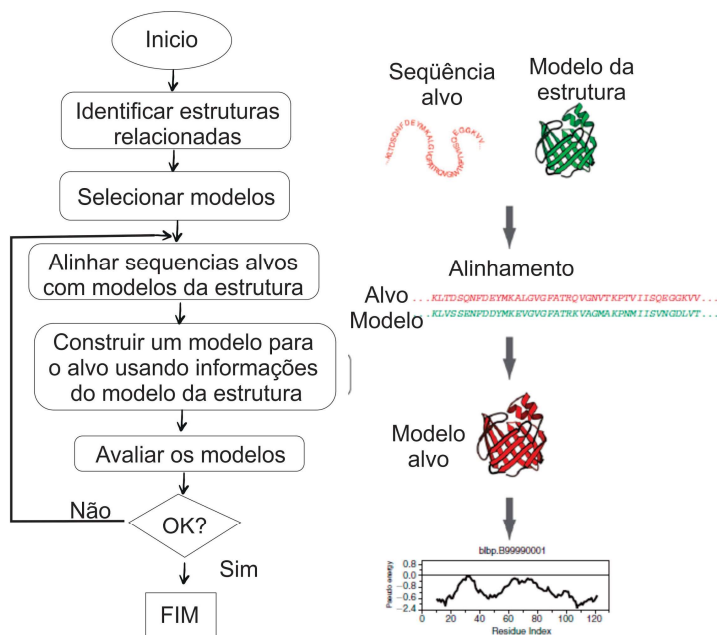


O primeiro passo é a identificação de moldes, que pode ser a mesma proteína ou uma similar. Logo após é feito o alinhamento de sequências para verificar a similaridade, a partir daí é feita a construção da cadeia principal e a modelagem das alças. Com o modelo primitivo é feita a validação dele e a otimização se necessário. Se a modelagem foi boa você terá o seu modelo final, porém de não ficou boa será necessário retornar ao segundo passo ou talvez ao primeiro passo e escolher outro molde.

Fonte: Adaptado de Santos Filho and Alencastro 2003.

Um dos programas mais utilizados para a modelagem comparativa para a predição de estruturas 3D de proteínas é o Modeller. Neste programa, a predição acontece de acordo com os seguintes passos (Figura 14): (1) identificação de estruturas relacionadas; (2) seleção de moldes; (3) alinhamento da sequência alvo com o molde escolhido; (4) Construção do modelo alvo usando informação da estrutura do molde escolhido; e, (5) validação do modelo (Webb and Sali 2014).

Figura 14 - Etapas do protocolo básico da predição de estrutura 3D utilizado pelo programa Modeller



O programa começa identificando as estruturas relacionadas que foram fornecidas pelo usuário, no segundo passo ele realiza a seleção do molde. Em seguida ele faz o alinhamento da sequência alvo com o molde escolhido e a construção do modelo alvo usando informação da estrutura do molde escolhido. Como último passo ele realiza a validação do modelo.

Fonte: Adaptado de Webb and Sali 2014.

O Modeller possui vários diferenciais em relação a outros programas existentes, como por exemplo, poder ser usado em Windows, Linux e Mac, sendo compatível com sistemas de 32 e 64 bit; usa a linguagem Python com uma sintaxe fácil; não necessita de drivers ou softwares adicionais e nem bibliotecas para ser utilizado; se necessário pode-se alterar o script para incluir energia de ligação, reparar átomos e realizar a minimização da estrutura; e é possível fazer mutações de pontos únicos em estruturas de PDB (Sakhteman and Zare 2016). Esse programa será utilizado para a predição de estrutura 3D no presente trabalho.

1.6 Variantes em genes das proteínas do leite de vaca associadas a doenças humanas

Existem diversos estudos correlacionando a frequência de determinadas variantes alélicas em proteínas do leite de vaca com a frequência de determinadas doenças humanas (Silva and Malcata 2005; Kamiński, Cieślińska et al. 2007; Woodford 2011). Algumas dessas proteínas são ativadoras de opióides, outras tem efeitos no sistema cardiovascular.

Dentre os alelos mais frequentes em *CSN2*, A1 e A2 diferem em um aminoácido na posição 67, histidina ou prolina, respectivamente. Estudos mostram que a variante alélica A1 em *CSN2* pode ser um fator de risco para a isquemia cardíaca e diabetes mellitus tipo 1 (Kamiński, Cieślińska et al. 2007, Woodford 2011, Dinesh, Verma et al. 2015). Evidências epidemiológicas sugerem que a variante A2 em *CSN2* seja melhor para a saúde humana do que a variante A1. A frequência dessas variantes difere entre as raças bovinas (Kamiński, Cieślińska et al. 2007), destacando a importância das análises raça-específica.

As beta-casomorfina (BCMs) são peptídeos produzidos a partir da clivagem da *CSN2*, sendo o BCM7 o primeiro a ser isolado. Tem sido sugerido que o BCM7 seja um fator de risco para autismo, doenças autoimunes e alergias leves, além de diabetes mellitus tipo 1, aterosclerose, síndrome da morte súbita infantil (SIDS) e isquemia cardíaca (Kamiński, Cieślińska et al. 2007; Woodford 2011).

Já a proteína *CSN3* possui mecanismos de coagulação do leite similares com os mecanismos da coagulação do sangue, baseada em uma interação entre fibrinogênio e trombina (Silva and Malcata 2005).

Todos estes estudos descrevem a identificação de variantes nos genes envolvidos na composição do leite, porém nenhuma análise em escala genômica foi realizada para as raças Guzerá e Gir, o que foi realizado no presente trabalho.

1.7 Justificativa e relevância

Com o avanço das técnicas de Biologia Molecular, marcadores moleculares para características da produção leiteira vêm sendo identificados e são considerados promissores para a estratégia de seleção. No entanto, grande parte dos marcadores moleculares identificados são baseados em variantes taurinas e muitos desses marcadores não são polimórficos em zebuínos. Assim, levando-se em consideração as características adaptativas dos zebuínos em relação aos taurinos, identificar as variantes raça-específicas, responsáveis por essas características é fundamental para o melhor direcionamento dos programas de melhoramento genético.

Neste contexto, foram desenvolvidos e financiados pela FAPEMIG, dois projetos que visaram a realizar o sequenciamento de touros da raça Gir e Guzerá, com o intuito de disponibilizar polimorfismos mais eficazes para os programas de melhoramento genético destas raças.

A partir desses projetos, as primeiras versões dos genomas de touros da raça Gir e Guzerá foram montadas e diversas variantes genéticas foram identificadas (manuscrito em fase de publicação), disponibilizando-se assim, uma grande quantidade de dados para análises posteriores.

Neste contexto, nosso grupo de pesquisa, Laboratório de Genética Humana e Médica do ICB/UFMG, coordenado pela Prof^a. Dr^a. Maria Raquel Carvalho, em parceria com a EMBRAPA Gado de Leite, sob coordenação do Dr. Marcos Vinícius G. B. da Silva, e o Grupo de Genômica e Biologia Computacional da FIOCRUZ/MG, sob coordenação do Dr. Guilherme Oliveira, vem desenvolvendo projetos, identificando variantes raça-específicas, que possam estar associadas às diferenças na produção e composição leiteira dos zebuínos. Desta forma, entender os eventos moleculares, que dirigem a lactogênese em vacas leiteiras, irá contribuir para o entendimento dos mecanismos por trás do desenvolvimento da glândula e poderá contribuir para o desenvolvimento de novas tecnologias para o manejo e melhoramento genético dos bovinos leiteiros (Gao, Lin et al. 2013).

1.8 Objetivos

1.8.1 Objetivo geral

A presente dissertação teve como objetivos:

1. Identificar, através de análises *in silico*, variantes nos genes, que codificam proteínas do leite, em três touros da raça Guzerá e três touros da raça Gir, submetidos a sequenciamento completo do genoma;
2. Prever, novamente através de análises *in silico*, o potencial para repercussão funcional destas variantes, ou seja, identificar quais deveriam ser alvo de investigação em estudos futuros, na busca por variantes, que possam ser responsáveis pelas diferenças na concentração das proteínas do leite entre essas duas raças e entre estas raças zebuínas quando comparadas às taurinas.

1.8.2 Objetivos específicos

A presente dissertação teve como objetivos específicos:

1. Selecionar variantes genéticas para análise pormenorizada de impacto funcional, através da aplicação de filtro baseado em critérios de qualidade e pela confirmação visual nas variantes genéticas identificadas;
2. Prever, através de análise *in silico*, criação/abolição de sítios de reconhecimento de splicing alternativo, alteração de sítios de reconhecimento de ligação à miRNA e análises de conservação evolutiva;
3. Prever modelos 3D das estruturas das proteínas, que tiverem variantes missense.

2 MATERIAIS E MÉTODOS

2.1 A amostra

O presente trabalho utilizará dados de sequenciamento, que foram realizados para três touros da raça Gir-leiteiro e três touros da raça Guzerá, durante a tese de doutorado da Dr^a. Izinara Rosse, conforme podemos observar na Tabela 2 (Rosse 2015).

As amostras 1 e 4 foram selecionadas de acordo com os seguintes critérios: menor índice de endogamia, maior número de descendentes, participação nos programas oficiais de melhoramento genético desenvolvidos pela EMBRAPA-CNPGL e serem de propriedade de empresas públicas (EPAMIG ou EMBRAPA). As amostras 2, 3, 5 e 6 foram selecionadas com foco na diversidade genética do indivíduo, assim, os três representantes de cada raça pertencem a diferentes rebanhos do país (Rosse 2015).

Todo o processo de montagem de bibliotecas e o sequenciamento das amostras 1 e 4 foram desenvolvidos pela equipe coordenada pelo Dr. Guilherme Oliveira, da FIOCRUZ-Minas. Já a montagem das bibliotecas e o sequenciamento das amostras 2, 3, 5 e 6 realizadas na ESALQ sob supervisão do Dr. Luiz Coutinho (Tabela 2) (Rosse 2015).

No trabalho de Rosse, 2015, foi realizado o sequenciamento genômico das amostras 1 e 4, foi necessária a construção de bibliotecas do tipo *mate-paired*, que contém um inserto, de tamanho conhecido, entre um fragmento e outro, o que ajuda na etapa de montagem das sequências. Duas bibliotecas tipo *mate-paired* foram construídas para cada raça, com insertos de 1-2 kb e 3-4 kb. Todos os *kits* necessários foram fornecidos pela Life Technologies (Life Technologies, Califórnia, Estados Unidos). O método de sequenciamento utilizado pela Plataforma SOLiD é o de ligação de oligos marcados com fluoróforos diferentes na superfície das *beads* (esferas magnéticas) (Mardis 2008).

Neste sequenciamento, foram geradas sequências de 50 pb de comprimento e a quantidade de sequências, que permaneceram após o filtro de qualidade podem ser observadas na Tabela 2 (Rosse 2015).

Em relação às amostras 2, 3, 5 e 6, foram construídas bibliotecas *paired-end*. Nesse tipo de biblioteca, o DNA genômico é fragmentado por uma transposase, ligado aos adaptadores,

desnaturado e imobilizado em uma superfície complementar aos adaptadores (Ansorge 2009).

Neste sequenciamento, foram geradas sequências de 100 pb de comprimento e a quantidade de sequências que permaneceram após o filtro de qualidade está apresentada na Tabela 2 (Rosse 2015).

A análise das sequências foi realizada pelos três grupos (ICB/UFMG, EMBRAPA e CEBio/FIOCRUZ).

Tabela 2 - Descrição das amostras utilizadas nesta dissertação

Raça	Amostra	Plataforma de Sequenciamento	Local do Sequenciamento	Tipo de biblioteca	Tamanho do inserto (Kb)	N. reads filtradas	Cobertura esperada do Genoma
Gir-leiteiro	1	SOLiD v4	FIOCRUZ-Minas	mate-paired	1-2	635.210.959	12X
			FIOCRUZ-Minas	mate-paired	3-4	342.786.110	6X
	2	HiSeq2500	ESALQ	paired-end	-	372.226.063	14X
	3	HiSeq2500	ESALQ	paired-end	-	361.612.147	13X
Guzerá	4	SOLiD v4	FIOCRUZ-Minas	mate-paired	1-2	582.533.747	11X
			FIOCRUZ-Minas	mate-paired	3-4	748.941.592	14X
	5	HiSeq2500	ESALQ	paired-end	-	314.142.888	11X
	6	HiSeq2500	ESALQ	paired-end	-	400.217.538	15X

Fonte: Rosse, 2015

2.2 Montagem dos genomas

As seqüências oriundas do sequenciador SOLiD v4 foram filtradas e corrigidas através do csfasta_quality_filter e SAET, respectivamente. Enquanto as sequências da plataforma HiSeq 2500 foram filtradas utilizando-se a ferramenta trimmomatic (Bolger, Lohse et al. 2014). Passaram pelo filtro de qualidade as sequências cuja média de qualidade PHRED foi igual ou superior a 20 e 30, para as sequências SOLiD e HiSeq, respectivamente (Rosse 2015).

As sequências foram mapeadas contra a referência de *Bos taurus* (UMD 3.1), disponível publicamente (Zimin, Delcher et al. 2009), usando-se o software BWA (com o parâmetro "aln" e "mem" para as sequências SOLiD e HiSeq, respectivamente (Li and Durbin 2009). Toda essa etapa foi realizada durante o doutoramento de Izinara Rosse da Cruz (Rosse 2015).

2.3 Seleção dos genes

Os genes selecionados foram *CSN1S1* (α S1-caseína), *CSN2* (β -caseína), *CSN1S2* (α S2-caseína), *CSN3* (κ -caseína), *LALBA* (α -lactoalbumina), *LGB* (β -lactoalbumina) e *LTF* (lactoferrina). Os genes *CSN1S1*, *CSN2*, *CSN1S2*, *CSN3* e *LALBA* foram selecionados por possuírem variantes funcionais já descritas e serem diferencialmente expressos durante a lactogênese (Gao, Lin et al. 2013). *LGB* foi selecionado por ter associação com a produção, composição e propriedades de coagulação do leite (Cecchinato, Ribeca et al. 2012). *LTF* foi selecionado pela sua atividade antibacteriana, antiviral, antifúngica, além de ser marcador para susceptibilidade de mastite (Kang, Li et al. 2008).

2.4 Detecção das variações

O arquivo de resultados do mapeamento foi utilizado para a detecção dos SNVs e INDELs, usando o pacote SAMtools (Li, Li et al. 2008).

Após a identificação das variações, as mesmas foram submetidas a um filtro de qualidade com o intuito de reduzir o número de falso-positivos. Os critérios do filtro de qualidade são os seguintes: (1) variantes com cobertura em extensão mínima de 50%. e, (2) variantes com profundidade de cobertura mínimo de 4 sequências de cobertura em profundidade (Choi, Liao et al. 2013, adaptado de Rosse 2015).

Logo após, as variações foram confirmadas visualmente usando o software IGV (Thorvaldsdóttir, Robinson et al. 2013) e foram submetidas a um filtro de qualidade, onde as variantes com menos de duas reads iguais ao novo alelo foram excluídas das análises subsequentes.

Dessa maneira, foram consideradas para as análises as variantes acima de 80% de cobertura em extensão e 5 X de cobertura em profundidade, que foram compartilhadas pelos seis indivíduos sequenciados e confirmadas visualmente pelo software IGV.

2.5 Seleção das variantes potencialmente funcionais

Para o presente trabalho, foram selecionadas as variantes exônicas e as variantes novas que haviam sido anotadas como pertencentes aos genes de interesse do presente estudo.

As variantes foram analisadas quanto a criação/abolição de sítios de reconhecimento de *splicing* através do software Alternative Splice Site Predictor (ASSP) (Wang and Marín 2006); à alteração de sítios de reconhecimento de ligação à miRNA pelo algoritmo presente na base de dados miRBase (Kozomara and Griffiths-Jones 2014); e, análises de conservação evolutiva para as variantes potencialmente funcionais, utilizando-se o Mulan (Ovcharenko, Loots et al. 2005) e o ECR Browser (Ovcharenko, Nobrega et al. 2004).

Para as variantes, que afetam sítios de *splicing* alternativo, a predição das sequências de isoformas alternativas foi realizada através do programa ExPASy translate (Artimo, Jonnalagedda et al. 2012). Essas isoformas foram procuradas em bancos de dados de proteínas através do blastp (NCBI) e a estrutura 3D foi gerada, como será descrito posteriormente. Para variantes em regiões *upstream*, que afetam sítios de *splicing* alternativo, foi checado se estão presentes em regiões promotoras usando o Promoter 2.0 Prediction Server (Knudsen 1999).

Para as análises evolutivas, uma sequência foi selecionada para cada um dos genes em cada uma das espécies utilizadas, sendo nove mamíferos e uma ave (*Gallus gallus*), como grupo externo. As sequências foram extraídas do banco de dados do NCBI, das seguintes espécies: Homem, Camundongo, Galo, Cavalo, Rato, Porco, Ovelha, Cabra e Búfalo (Tabela 3), além das sequências de *Bos taurus* e das sequências das amostras estudadas nesta dissertação. Foi feito um alinhamento múltiplo usando o programa Mulan .

Tabela 3 - Espécies utilizadas para as análises de conservação evolutiva e os GI (número de acesso NCBI) de cada uma delas

Espécie	GI	GI	GI	GI	GI
	<i>LTF</i>	<i>LALBA</i>	<i>CSN1S1</i>	<i>CSN1S2</i>	<i>CSN3</i>
<i>Homo sapiens</i>	568815595	568815586	568815594		568815594
<i>Mus musculus</i>	372099101	372099095	372099105		372099105
<i>Gallus gallus</i>	358485503	358485511			
<i>Equus caballus</i>	194246364	194246387	194246382	194246382	194246382
<i>Rattus norvegicus</i>	666183917	666183999	666183623		666183623
<i>Sus scrofa</i>	347618781	347618789	347618786	347618786	347618786
<i>Ovis aries</i>	417531901	417531961	417531944	417531944	417531944
<i>Capra hircus</i>	541128965	541128982	541128981	541128981	541128981
<i>Bubalus bubalis</i>	551720832	551721652	551720848	551720848	551720848

A presença das variantes em regiões conservadas evolutivamente (ECRs) em regiões afetadas foi verificada usando o programa ECR Browser (Ovcharenko, Nobrega et al. 2004) incluindo doze diferentes espécies: Humano (*Homo sapiens*), Camundongo (*Mus musculus*), Chimpanzé (*Pan troglodytes*), Rhesus (*Macaca mulatta*), Boi (*Bos taurus*), Galo (*Gallus gallus*), Gambá (*Didelphis marsupialis*), Cachorro (*Canis lupus familiaris*), Sapo (*Bufo bufo*), Zebrafish (*Danio rerio*), Fugu (*Tetraodontidae*), e Tetraodon (*Tetraodon biocellatus*).

O impacto das variantes exônicas foi avaliado também em relação ao impacto da substituição do aminoácido na proteína ou sobre o domínio da proteína, e foi feito a modelagem da estrutura 3D das proteínas. Os possíveis impactos de codon usage na tradução da nova isoforma da proteína foram avaliados usando-se a ferramenta Sequence Manipulation Suite: Codon Usage (Stothard 2000) e Codon Usage Database (Nakamura, Gojobori et al. 2000).

Para a predição dos modelos 3D das estruturas das proteínas foi utilizado o programa Modeller 9.15 [34] e foi realizado apenas para os genes *LALBA* e *LTF* que apresentaram variantes missenses e variantes que alteram sítios de *splicing* alternativo. Para o gene *LALBA* foi utilizado cinco moldes (1F6R, 1FKQ, 1HFY, 1HFZ e 1HMK) e para o gene *LTF* foi utilizado quatro moldes (1B1X, 1BLF, 1LCF e 2BJJ) ambos depositados no PDB (Berman, Westbrook et al. 2000). Para verificar a similaridade entre os moldes depositados no PDB e as sequências contendo as variantes observadas no presente trabalho foi realizado um alinhamento múltiplo com o MultAlin (Corpet 1988).

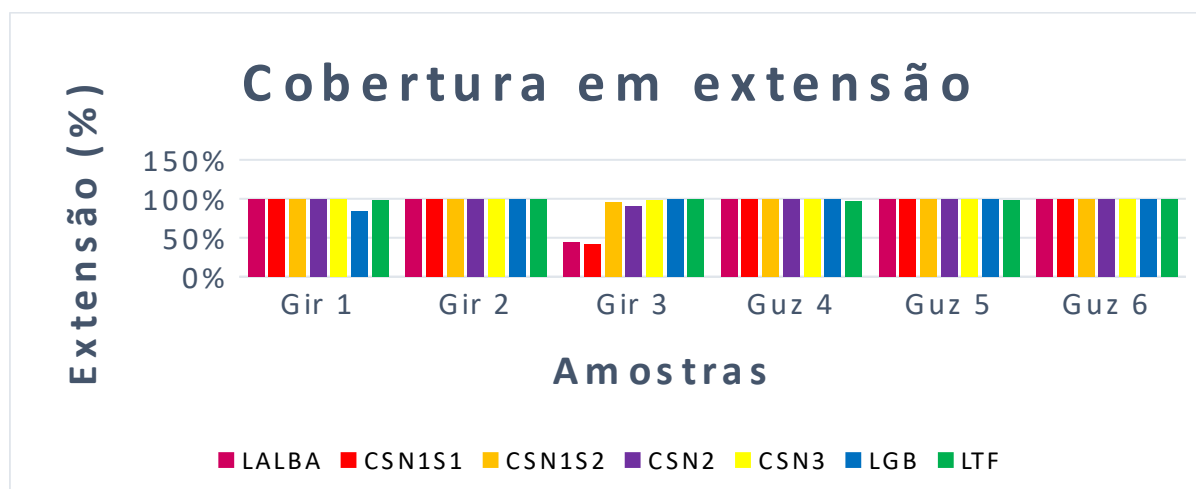
Foram gerados mil modelos para a proteína normal e mil modelos para a proteína com a variante para cada variante. Em seguida, os cinco modelos da proteína normal e das proteínas com as variantes com maiores valores de GA341 (esse método usa a porcentagem de identidade da sequência entre o molde e o modelo de parâmetro) e os menores valores de DOPE (*Discrete Optimized Protein Energy*) foram selecionados para a validação. Para a validação foi utilizado o software PROCHECK (Laskowski, MacArthur et al. 1993), disponível no banco de dados do PDBsum Generate (Laskowski, Hutchinson et al. 1997), e também gerado o gráfico de Ramachandran. Esse gráfico mostra o número total de resíduos, o número de resíduos de glicina, número de resíduos de prolina, o número de resíduos de não-glicina e não-prolina, número de resíduos finais (excluído Gly e Pro), resíduos em regiões mais favorecidas, resíduos em regiões permitidas adicional, resíduos em regiões permitidas generosamente e resíduos em regiões proibidas. Foi gerado um gráfico de energia e calculado os valores de z-scores usando o ProSa-web (*Protein Structure Analysis*) (Wiederstein and Sippl 2007). A validação foi calculada dividindo-se a porcentagem dos resíduos em regiões mais favorecidas pelo z-score e o maior valor indica o melhor modelo a ser usado. Assim, nós selecionamos os melhores modelos para cada proteína e comparamos as proteínas normais com as proteínas com variantes realizando a sobreposição entre os modelos, usando o programa TM-align (Zhang and Skolnick 2005) e o software PyMOL (*Molecular Graphics System*) (Schrodinger 2015). As substituições de aminoácidos foram marcadas para ajudar na visualização de mudanças na conformação das proteínas.

3 RESULTADOS

3.1 Cobertura em extensão e em profundidade obtida para os genes de interesse

Comparando com o genoma de referência de *Bos taurus* UMD3.1, e considerando apenas a região dos genes selecionados para esse estudo, a média da cobertura em extensão obtida foi 95.85% e a média da cobertura em profundidade obtida foi 13.39X (Figura 15 e 16) (Tabela 4). Baixa cobertura foi observada para os genes *CSN1S1* e *LALBA* na amostra Gir 3 (para um dos touros Gir sequenciados). Essas sequências foram excluídas das análises subsequentes. Após essa exclusão, a média das coberturas observadas foram superiores a 80% (84-100) e 5X (5.2-34.7), respectivamente.

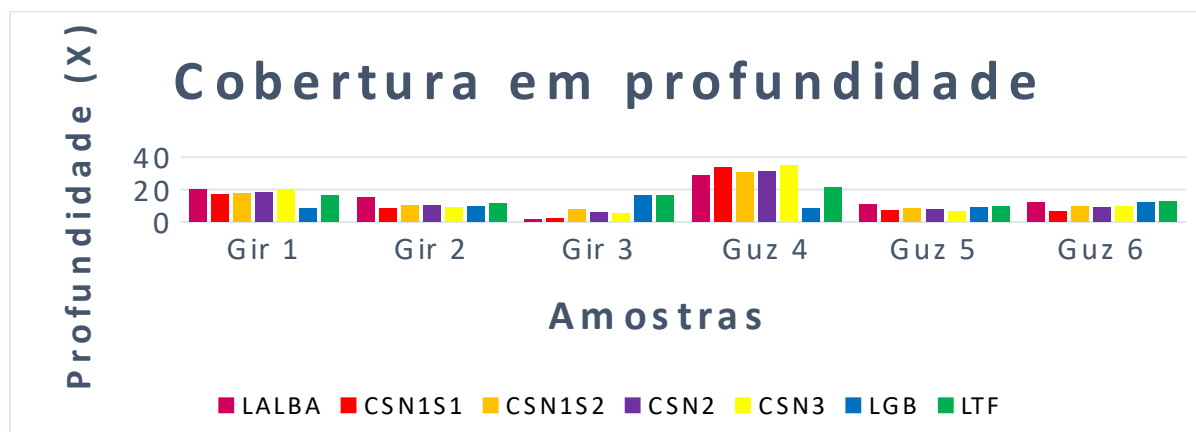
Figura 15 - Cobertura em extensão observada para os genes de interesse



No eixo do X, podemos observar as amostras e, no eixo Y a cobertura em extensão em porcentagem. Cada cor representa um gene, rosa o gene *LALBA*, vermelho o gene *CSN1S1*, laranja o gene *CSN1S2*, roxo o gene *CSN2*, amarelo o gene *CSN3*, azul o gene *LTF* e verde o gene *LGB*. Para os genes *LALBA* e *CSN1S1* os valores de cobertura em extensão e em profundidade para a amostra Gir 3, foram excluídas das análises devido ao fato de não atenderem aos critérios de qualidade. Após a remoção desses genes, todos os valores de cobertura em extensão foram superiores a 80%, evidenciando a confiabilidade dos resultados.

Fonte: Produzida pela autora em 2015.

Figura 16 - Cobertura em profundidade observada para os genes de interesse



No eixo do X, podemos observar as amostras e, no eixo do Y, a cobertura em profundidade em vezes. Cada cor representa um gene, rosa o gene *LALBA*, vermelho o gene *CSN1S1*, laranja o gene *CSN1S2*, roxo o gene *CSN2*, amarelo o gene *CSN3*, azul o gene *LTF* e verde o gene *LGB*. Os dados dos genes *LALBA* e *CSN1S1* da amostra Gir foram excluídos das análises subsequentes, por não atenderem aos critérios de qualidade. Após a remoção desses genes, todos os valores de cobertura em extensão foram superiores a 5X, evidenciando a confiabilidade dos resultados.

Fonte: Produzida pela autora em 2015.

Como foi mencionado, existem diferenças na concentração de proteínas no leite de taurinos e zebuínos. As raças taurinas, por terem sido submetidas a programas de melhoramento genético por mais tempo, têm um número maior de variantes genéticas descritas. Dentre os zebuínos, o Gir e o Guzerá são importantes para a produção de leite e participam de estudos de melhoramento genético. Porém, até o presente momento, nenhum estudo havia sido realizado com intuito de identificar e prever, através de análises *in silico* em escala genômica, variantes genéticas nos genes que codificam proteínas do leite e seu potencial para repercussão funcional. Neste contexto, na presente dissertação, foram realizadas análises *in silico* para identificar as variantes genéticas nos principais genes codificadores de proteínas do leite em dois dos principais zebuínos no Brasil.

Tabela 4 - Cobertura em profundidade e cobertura em extensão nos genes sobre investigação

Amostras	<i>CSNIS1</i>		<i>CSNIS2</i>		<i>CSN2</i>		<i>CSN3</i>		<i>LALBA</i>		<i>LGB</i>		<i>LTF</i>	
	Cobertura em profundidade (X)	Cobertura em extensão (%)	Cobertura em profundidade (X)	Cobertura em extensão (%)	Cobertura em profundidade (X)	Cobertura em extensão (%)	Cobertura em profundidade (X)	Cobertura em extensão (%)	Cobertura em profundidade (X)	Cobertura em extensão (%)	Cobertura em profundidade (X)	Cobertura em extensão (%)	Cobertura em profundidade (X)	Cobertura em extensão (%)
1	17.21	99,80	17.37	99,90	18.37	100	19.24	100	19.81	100	8.58	84,55	16.08	98,42
2	8.19	99,75	10.13	99,79	10.44	99,84	9.00	99,96	14.84	100	9.82	99,98	11.24	98,92
3	1.91*	41,31*	7.51	95,11	5.90	89,83	5.17	98,55	1.50*	44,02*	16.24	100	16.43	99,45
4	33.78	99,99	30.41	99,87	30.93	100	34.69	100	28.62	100	8.03	83,70	21.00	97,34
5	7.16	99,53	8.49	99,72	7.57	99,99	6.18	100	10.64	100	8.95	99,98	9.52	98,45
6	6.69	99,91	9.68	99,99	8.80	99,06	9.36	100	11.86	100	12.18	100	12.81	98,81
Média	12.49	90,05	13.93	99,06	13.67	98,12	13.94	99,75	14.54	90,67	10.63	94,70	14.51	98,57

*valores com cobertura abaixo do filtro de qualidade. Amostras 1 a 3: Gir; amostras 4 a 6:

Guzerá.

3.2 Localização, tipo e impacto das variantes genéticas

Foram encontrados 64 SNVs nos sete genes em estudo. Entre essas variantes, 32 (50%) são variantes intrônicas; 21 (32,8%) estão localizadas na região *upstream* do gene; sete (10,9%) na região *downstream* do gene; uma (1,6%) na região intergênica; duas (3,1%) são variantes missense; e, uma (1,6%) está na região 3'UTR (Tabela 5). A maioria das SNVs foram encontradas no gene *LTF*, sendo 23 variantes intrônicas; uma variante exônica; quatro na região *upstream* do gene; e, cinco na região *downstream* do gene. Nos genes *CSN2* e *LGB* foi encontrada apenas uma SNV em cada, localizadas na região *upstream* desses genes.

Foram encontradas 6 INDELs nesses sete genes. Entre essas variantes, 1 (16,7%) está na região *downstream* do gene, 1 (16,7%) na região *upstream* do gene, 1 (16,7%) na região intergênica e 3 (50%) na região intrônica. Sendo que eles foram encontrados nos genes *CSN3*, *LALBA* e *LTF*. Não foram encontrados INDELs nos genes *CSN1S1*, *CSN1S2*, *CSN2* e *LGB* (Tabela 5).

Vale ressaltar que 32 (50%) SNVs e cinco (83%) INDELs são novas, ou seja, não foram identificadas em nenhum dos bancos de dados disponíveis publicamente.

Tabela 5 - Variações, SNVs e INDELs, encontradas no presente estudo

Gene	Tipo	Localização	Posição cromossômica	Alelos#	dbSNP ID
<i>CSN1S1</i>	SNV	<i>Upstream</i>	BTA6:87137341	T→G	rs137341845
	SNV	<i>Upstream</i>	BTA6:87139047	C→T	rs110660686
	SNV	Intrônica	BTA6:87153410	A→G	NOVA
	SNV	Intrônica	BTA6:87155005	A→G	rs135213517
	SNV	Intrônica	BTA6:87158264	G→A	rs133257014
	SNV	<i>Downstream</i>	BTA6:87163461	T→G	NOVA
<i>CSN1S2</i>	SNV	Intrônica	BTA6:87272568	G→T	NOVA
	SNV	Intrônica	BTA6:87276023	C→T	rs109185641
	SNV	3'UTR	BTA6:87280796	T→C	NOVA
<i>CSN2</i>	SNV	<i>Upstream</i>	BTA6:87188870	T→A	rs109955918
<i>CSN3</i>	SNV	<i>Upstream</i>	BTA6:87374090	T→C	NOVA

Nota: Intervalos de *upstream* e *downstream* foram definidos contando 5kb a partir do nucleotídeo +1 e do primeiro nucleotídeo depois do stop codon, respectivamente; # novos alelos são mostrados a direita; NOVA= identificadas no presente estudo.

Tabela 5 - Variações, SNVs e INDELs, encontradas no presente estudo (continuação)

Gene	Tipo	Localização	Posição cromossômica	Alelos#	dbSNP ID
	SNV	<i>Upstream</i>	BTA6:87374308	A→C	rs109198841
	SNV	Intergenic	BTA6:87382646	T→G	NOVA
	INDEL	Intergenic	BTA6:87389773	TG→TGG	NOVA
<i>LALBA</i>	SNV	<i>Upstream</i>	BTA5:31343070	G→A	rs135508437
	SNV	<i>Upstream</i>	BTA5:31343277	T→C	NOVA
	SNV	<i>Upstream</i>	BTA5:31343847	T→C	NOVA
	SNV	<i>Upstream</i>	BTA5:31346231	A→T	NOVA
	SNV	<i>Upstream</i>	BTA5:31347399	G→C	NOVA
	SNV	<i>Upstream</i>	BTA5:31347426	G→A	NOVA
	SNV	<i>Upstream</i>	BTA5:31347456	T→C	NOVA
	SNV	<i>Upstream</i>	BTA5:31347479	A→G	NOVA
	SNV	<i>Upstream</i>	BTA5:31347538	C→A	NOVA
	SNV	<i>Upstream</i>	BTA5:31347611	A→G	NOVA
	SNV	<i>Upstream</i>	BTA5:31347772	T→C	NOVA
	SNV	Intrônica	BTA5:31348166	C→T	NOVA
	SNV	<i>Missense</i>	BTA5:31348386	A→G	NOVA
	SNV	Intrônica	BTA5:31349118	A→G	NOVA
	SNV	Intrônica	BTA5:31349177	A→C	NOVA
	SNV	Intrônica	BTA5:31349185	A→G	rs110248183
	SNV	<i>Downstream</i>	BTA5:31350957	C→T	NOVA
	INDEL	<i>Downstream</i>	BTA5:31353933	GC→G	NOVA
<i>LGB/PAEP</i>	SNV	<i>Upstream</i>	BTA11:103299655	T→C	rs110922198
<i>LTF</i>	SNV	<i>Upstream</i>	BTA22:53517110	T→G	rs110552427
	SNV	<i>Upstream</i>	BTA22:53520209	C→T	rs134217828
	SNV	<i>Upstream</i>	BTA22:53520314	A→G	rs137054020
	SNV	<i>Upstream</i>	BTA22:53520861	G→A	rs211642876
	SNV	Intrônica	BTA22:53529171	G→A	rs133581561
	SNV	Intrônica	BTA22:53531650	G→C	rs41565709
	SNV	Intrônica	BTA22:53534325	A→T,G	NOVA
	SNV	Intrônica	BTA22:53536899	T→C	rs137274059
	SNV	Intrônica	BTA22:53538011	T→C	rs136145466
	SNV	Intrônica	BTA22:53539635	T→G	rs135706891
	SNV	Intrônica	BTA22:53540102	T→C	rs137737245
	SNV	Intrônica	BTA22:53541291	T→C	rs135734723
	SNV	Intrônica	BTA22:53541952	G→A	NOVA
	SNV	Intrônica	BTA22:53542142	G→A	rs133585886
	SNV	Intrônica	BTA22:53542738	A→G	rs109046761
	SNV	Intrônica	BTA22:53543092	A→G	rs136830243

Tabela 5 - Variações, SNVs e INDELs, encontradas no presente estudo (continuação)

Gene	Tipo	Localização	Posição cromossômica	Alelos#	dbSNP ID
	SNV	<i>Missense</i>	BTA22:53546045	C→A	rs208566369
	SNV	Intrônica	BTA22:53547182	T→C	NOVA
	SNV	Intrônica	BTA22:53547194	A→C	NOVA
	SNV	Intrônica	BTA22:53547203	A→G	NOVA
	SNV	Intrônica	BTA22:53547260	G→A	rs209704014
	SNV	Intrônica	BTA22:53548175	T→C	rs135319050
	SNV	Intrônica	BTA22:53548745	A→G	NOVA
	SNV	Intrônica	BTA22:53548773	G→A	NOVA
	SNV	Intrônica	BTA22:53549225	A→C	NOVA
	SNV	Intrônica	BTA22:53552302	A→T	NOVA
	SNV	Intrônica	BTA22:53555036	C→T	rs136566018
	SNV	Intrônica	BTA22:53555434	T→G	rs136149065
	SNV	<i>Downstream</i>	BTA22:53558828	C→A	rs134817148
	SNV	<i>Downstream</i>	BTA22:53560203	T→A	rs135699289
	SNV	<i>Downstream</i>	BTA22:53560259	C→T	NOVA
	SNV	<i>Downstream</i>	BTA22:53560817	G→A	NOVA
	SNV	<i>Downstream</i>	BTA22:53560833	G→A	rs136896438
	INDEL	<i>Upstream</i>	BTA22:53519590	TC→TCC	NOVA
	INDEL	Intrônica	BTA22:53526379	ATT→AT	NOVA
	INDEL	Intrônica	BTA22:53539360	TC→TCCC	rs135392628
	INDEL	Intrônica	BTA22:53552429	C→CT	NOVA

Nota: Intervalos de *upstream* e *downstream* foram definidos contando 5kb a partir do nucleotídeo +1 e do primeiro nucleotídeo depois do stop codon, respectivamente; # novos alelos são mostrados a direita; NOVA= identificadas no presente estudo. As posições cromossômicas se referem à montagem UMD3.1 do genoma bovino.

Dentre as 70 variantes identificadas, 20 foram consideradas potencialmente funcionais (Tabela 6). Todas elas se localizam em ECRs e apenas três não são posições nucleotídicas evolutivamente conservadas. Estas, entretanto, tem outras evidências de funcionalidade, ou seja, afetam *splicing*, sítios de reconhecimento de miRNAs ou ambos. Todas as variantes nessa tabela são novas exceto para BTA22:53546045, dbSNP ID é rs208566369, que foi mantida nas análises porque é uma variante *missense*, que não foi estudada antes.

Tabela 6 - Sumário das variantes descobertas com potencial funcional: efeitos sobre sítios de *splicing*, sítios de reconhecimento de miRNA e posições conservadas evolutivamente

Gene	Localização	Posição cromossômica	Alelos #	Posição evolutivamente conservada	Efeitos em sítios de <i>splicing</i>	Efeitos em miRNAs
<i>CSN1S1</i>	Intrônica	BTA6:87153410	A→G	+		
<i>CSN1S2</i>	3'UTR	BTA6:87280796	T→C			Abole um sítio de reconhecimento de bta-miR-340
<i>CSN3</i>	<i>Upstream</i>	BTA6:87374090	T→C	+		
<i>LALBA</i>	<i>Upstream</i>	BTA5:31343847	T→C	+		
<i>LALBA</i>	<i>Upstream</i>	BTA5:31346231	A→T	+		
<i>LALBA</i>	<i>Upstream</i>	BTA5:31347426	G→A	+		
<i>LALBA</i>	<i>Upstream</i>	BTA5:31347538	C→A	+		
<i>LALBA</i>	<i>Upstream</i>	BTA5:31347611	A→G	+		
<i>LALBA</i>	Missense	BTA5:31348386	A→G	+		
<i>LALBA</i>	Intrônica	BTA5:31349118	A→G	+		
<i>LALBA</i>	Intrônica	BTA5:31349177	A→C	+		Abole um sítio em bta-miR-2444 e cria um sítio em bta-miR-1814b
<i>LALBA</i>	<i>Downstream</i>	BTA5:31350957	C→T	+		
<i>LALBA</i>	<i>Downstream</i>	BTA5:31353933	GC→G	+		
<i>LTF</i>	Intrônica	BTA22:53534325	A→T, G		Abole um sítio	
<i>LTF</i>	Missense	BTA22:53546045	C→A	+		
<i>LTF</i>	Intrônica	BTA22:53547194	A→C	+		
<i>LTF</i>	Intrônicas	BTA22:53548773	G→A		Cria um sítio	Creates one site in bta-miR-2429
<i>LTF</i>	Intrônica	BTA22:53552302	A→T	+		Abole um sítio em bta-miR-376e e cria dois sítios em bta-miR-376a e bta-miR-376d
<i>LTF</i>	<i>Downstream</i>	BTA22:53560259	C→T	+		
<i>LTF</i>	<i>Downstream</i>	BTA22:53560817	G→A	+		

Nota: Intervalos de *upstream* e *downstream* foram definidos contando 5kb a partir do nucleotídeo +1 e do primeiro nucleotídeo depois do stop codon, respectivamente. Todas as variantes nessa tabela são novas exceto para BTA22:53546045, dbSNP ID é rs208566369. + indica que esse SNV afeta uma posição conservada evolutivamente.

Na era pós-genômica, a cada novo projeto de sequenciamento de genoma completo, milhares de variantes são identificadas. Nestes casos, realizar estudos funcionais para todas essas variantes genéticas descobertas seria laborioso e de custo elevado. Desta forma, no presente trabalho foi delineada uma estratégia para selecionar as variantes potencialmente funcionais através de análises *in silico*. A partir dessas análises foi possível selecionar um número viável

de variantes genéticas, que podem ser estudadas mais a fundo através de estudos de associação e funcionais.

Vários estudos têm demonstrado, que as proteínas do leite são muito polimórficas. Até o ano de 2009, foram identificadas no gênero *Bos*, incluindo *Bos indicus* nove variantes genéticas no gene *CSN1S2*, doze variantes no gene *CSN2*, quatro variantes no gene *CSN1S2*, quatorze variantes no gene *CSN3*, onze variantes no gene *LGB* e três variantes no gene *LALBA*. O presente trabalho foi o primeiro a fazer uma análise desses genes em uma escala genômica para as raças Gir e Guzerá e identificou duas variantes novas nos genes *CSN1S2*, *CSN1S1*, *CSN3*, dezesseis variantes novas no gene *LALBA* e quatorze variantes novas no gene *LTF*.

Considerando o restritivo filtro de qualidade aplicado, onde a variante tem que estar presente em todas as seis amostras sequenciadas, as variantes selecionadas para análises subsequentes presentes nas raças Guzerá e Gir podem representar alelos típicos de zebuínos. Desta forma, será interessante avaliar a frequência destes alelos em amostras populacionais das raças Gir e Guzerá.

3.3 Análises de impacto das variantes missense

Duas SNVs missense foram identificadas no presente estudo, uma no gene *LTF* (Thr546Asn) e outra no gene *LALBA* (Ile60Val). A SNV no gene *LTF* já era conhecida, porém nenhum estudo investigando a repercussão funcional desta mutação foi descrito.

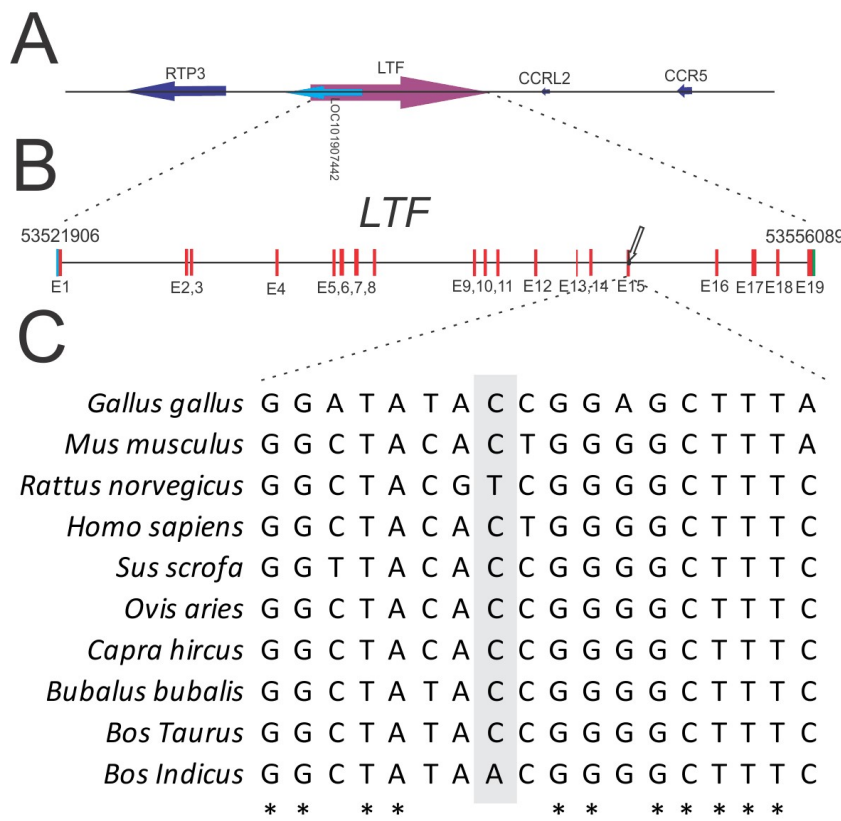
3.3.1 Mutação no gene LTF Thr546Asn

A mutação no gene *LTF* Thr546Asn já foi descrita no dbSNP sob o ID rs208566369 e afeta uma posição nucleotídica conservada evolutivamente. Os resultados da análise de conservação evolutiva para o SNV BTA22:53546045 são mostrados na figura 17, o alinhamento múltiplo de uma região maior é mostrado na figura suplementar 3 no Apêndice.

OBSERVAÇÃO - A SNV BTA22:53546045 está presente no segundo domínio da proteína, na posição 546. Na literatura, foi descrito que na posição 545 de LTF há uma asparagina, que é um sítio de glicosilação. Porém, na sequência disponível no NCBI e nas sequências do presente estudo há uma tirosina nessa posição. Assim, considerando-se as posições 545 e 546, há pelo menos quatro sequências: uma sequência 545Asn546Thr, uma 545Asn/546Asn,

uma 545Tyr546Thr e uma 545Tyr/546Asn. Nas análises realizadas aqui tanto na predição da estrutura 3D quanto na predição da glicosilação, foram usadas as sequências contendo 545Tyr546Thr, como referência, e 545Tyr/546Asn, como variante.

Figura 17 - Análise de conservação evolutiva para o SNV BTA22:53546045 no gene *LTF*



Em A, o segmento do cromossomo ao redor do gene *LTF* é mostrado. Em B, a posição da SNV, no exon 15 do gene *LTF* é mostrada. Em C, o resultado do alinhamento múltiplo evidenciando a conservação evolutiva. As espécies incluídas na comparação evolutiva apresentam o alelo C na posição, exceto para *Rattus norvegicus* e *Bos indicus*, que apresentam os alelos T e A, respectivamente. O alelo divergente está presente em *Bos indicus*.

Fonte: Produzida pela autora em 2015.

Na Tab. 7, são mostrados os resultados da predição da estrutura 3D para a mutação LTF (Thr546Asn), que envolve a substituição de um códon ACC (Thr) por um AAC (Asn), ambos aminoácidos polares e neutros. Dos quatro moldes usados no Modeller para a predição dos modelos tridimensionais para esta proteína, o melhor molde foi o 1BLF. Os valores de GA431 obtidos foram 1,00000 para os 1.000 modelos gerados tanto para a sequência da proteína LTF original, quanto para a sequência com a mutação (Tabela 7).

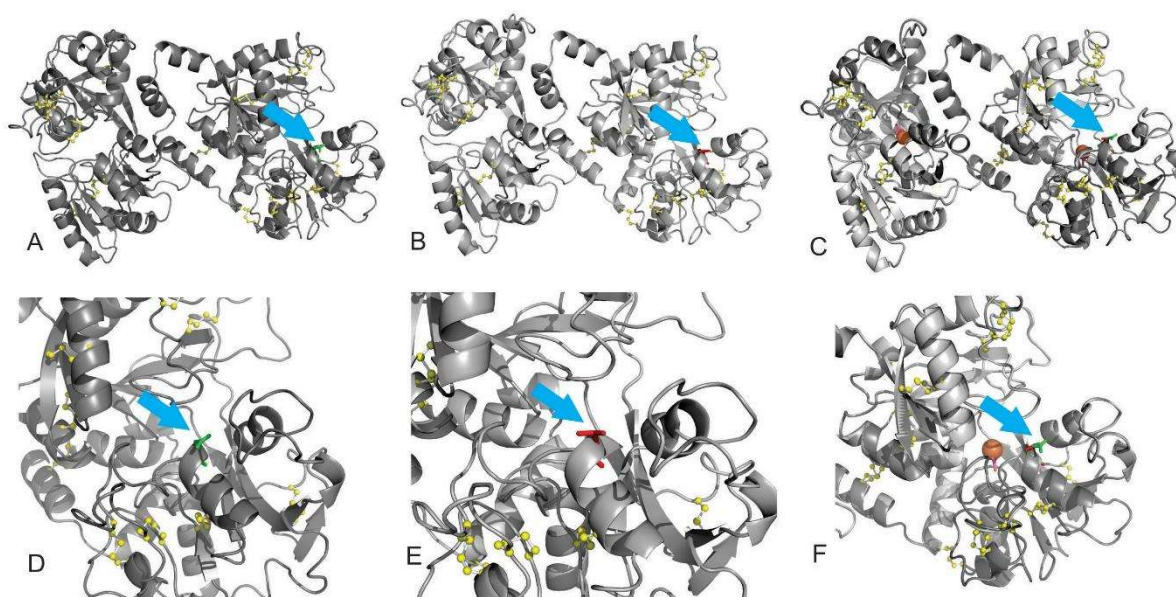
Tabela 7 - Resultados da modelagem para os cinco melhores modelos 3D da LTF

Proteína	Arquivo	Molpdf	DOPE score	GA341 score	z-score	Validação
LTF referência	B.99990987	4139,92676	-78264,21094	1,00000	-14,55	-6.23*
	B.99990894	3841,93555	-78265,56250	1,00000	-14,43	-6.25
	B.99990746	3789,22412	-78343,26563	1,00000	-14,50	-6.25
	B.99990097	3963,81592	-78346,00000	1,00000	-14,42	-6.29
	B.99990534	4889,62598	-78349,73438	1,00000	-14,56	-6.26
LTF variante	B.99990717	3876,30200	-78105,63281	1,00000	-14,48	-6.22
	B.99990753	3827,21997	-78191,98438	1,00000	-14,40	-6.27
	B.99990892	4010,10132	-78275,69531	1,00000	-14,51	-6.28
	B.99990369	3995,76660	-78295,79688	1,00000	-14,35	-6.26
	B.99990418	3789,77856	-78302,03906	1,00000	-14,58	-6.16*

Nota: *Melhores modelos. Para a referência foi usado o molde 1BFL da PBD, acesso em 2015.

O melhor modelo 3D para a proteína com a mutação foi sobreposto com o melhor modelo 3D da proteína original (Figura 18). As posições da substituição de aminoácidos estão marcadas de vermelho e verde e indicados com a seta azul. Esse resíduo se encontra numa alfa-helix, na parte C-terminal da proteína.

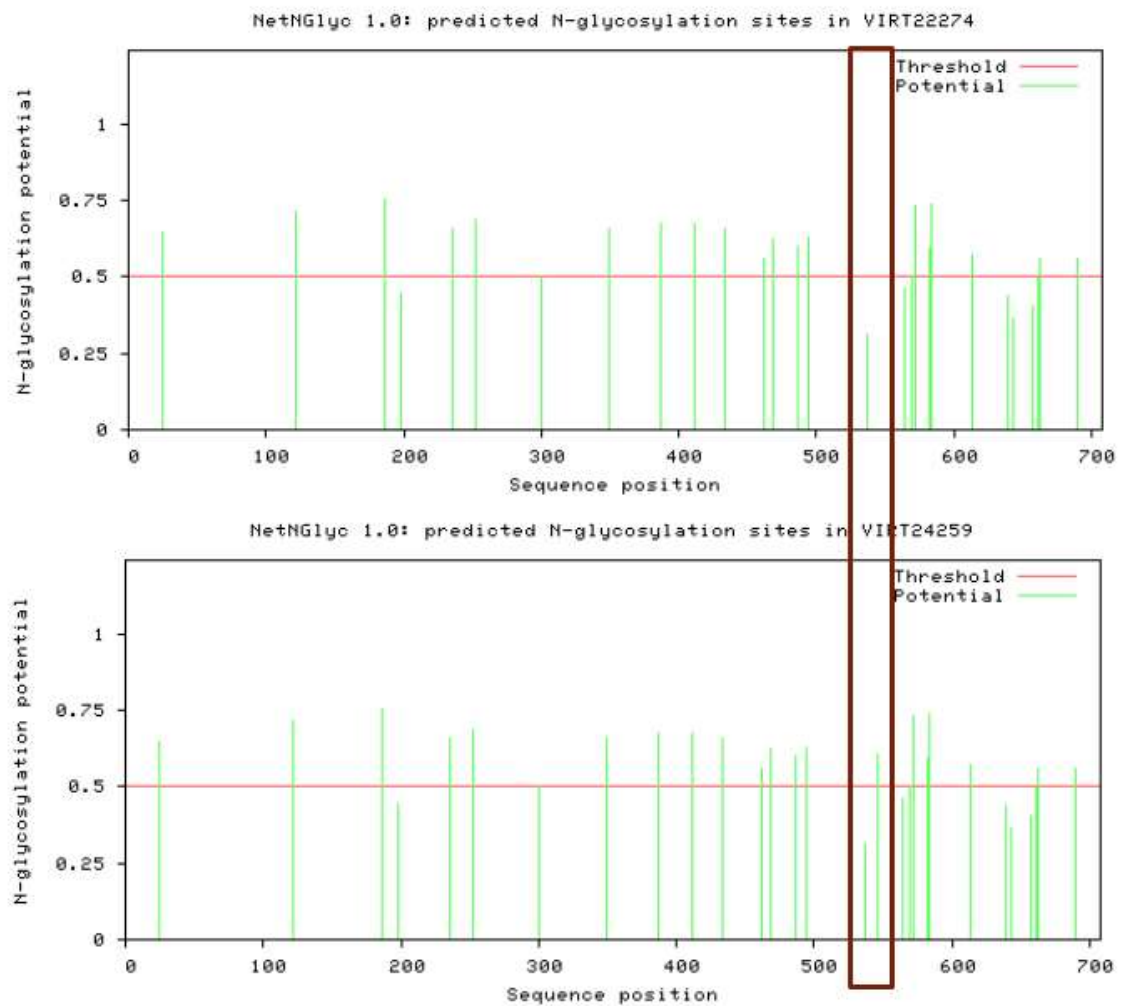
Figura 18 – Modelos preditos para a proteína LTF



Em A, o modelo 3D da sequência com o alelo previamente já conhecido, contendo uma treonina na posição 546, mostrada em verde e indicada pela seta azul. Os círculos amarelos representam as ligações dissulfeto. Em B, o modelo 3D da sequência contendo uma asparagina na posição 546, é mostrada em vermelho indicado pela seta azul. Em C, os modelos 3D sobrepostos mostrando a posição do ferro (círculo marrom) e do carbonato (em rosa). Em D, E e F, os modelos 3D com zoom na região da variação. Apesar de esta SNV *missense* estar localizado em uma α -helix, o modelo da estrutura 3D da proteína mutada não apresenta nenhuma distorção óbvia. Fonte: Produzida pela autora em 2015.

Para verificar os sítios de glicosilação no gene *LTF*, nós realizamos a predição de sítios de glicosilação em humanos (Gupta 2004) e foi possível observar, que a inserção de uma asparagina, como acontece nas sequências do presente estudo, leva à criação de um novo sítio de glicosilação (Figura 19), sugerindo que essa variação tenha repercussão funcional. A avaliação da estrutura 3D e da glicosilação sofre impacto destas diferenças observadas na literatura e na base de dados.

Figura 19 – Sítios de glicosilação no gene *LTF*

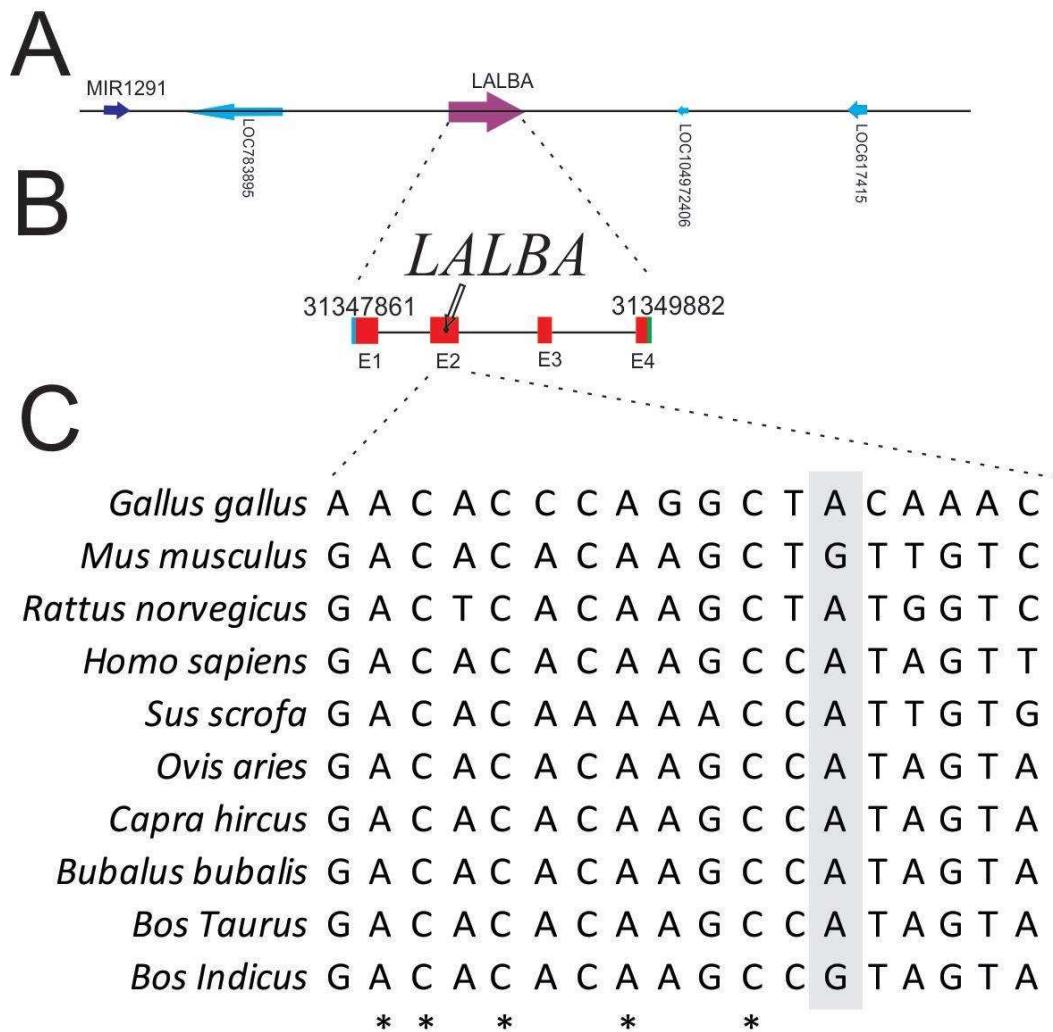


Em cima, os sítios de glicosilação da proteína de referência. Na parte de baixo, os sítios de glicosilação da proteína com o SNV. Na parte marcada, podemos observar a formação de um novo sítio de glicosilação. Essa análise foi feita usando-se o NetNGlyc 1.0 Server (Gupta 2004).

3.3.2 Mutação no gene *LALBA* Ile60Val

A mutação no gene *LALBA* Ile60Val é nova e afeta uma posição conservada evolutivamente. Os resultados da análise de conservação evolutiva para o SNV BTA5.31348386 são mostrados na figura 20, o alinhamento múltiplo de uma região maior é mostrado na figura suplementar 2 no Apêndice.

Figura 20 - Análise de conservação evolutiva para o SNV BTA5:31348386 no gene *LALBA*



Em A, o segmento do cromossoma ao redor do gene *LALBA* é mostrado. Em B, a posição do SNV, no exon 2 do gene *LALBA* é mostrada. Em C, o resultado do alinhamento múltiplo, evidenciando a conservação evolutiva. As espécies incluídas na comparação evolutiva apresentam o alelo A na posição, exceto por *Mus musculus* e *Bos indicus*, que apresentam o alelo G.

Fonte: Produzida pela autora em 2015.

Essa mutação missense envolve a substituição de um códon AUA (Ile) por um GUA (Val), ambos aminoácidos são apolares e alifáticos, e tendem a formar interações hidrofóbicas, estabilizando a estrutura da proteína. Dos quatro moldes usados no Modeller para a predição dos modelos para esse gene, o melhor molde foi o 1F6R. Os valores de GA431 obtidos foram 1.00000 para os 1.000 modelos gerados, tanto para a sequência da proteína original do gene *LALBA* quanto para a sequência com a mutação (Tabela 8).

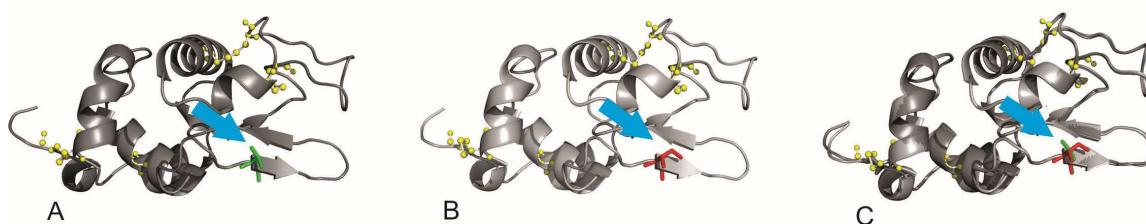
Tabela 8 - Resultados da modelagem para os cinco melhores modelos 3D

Proteína	Arquivo	Molpdf	DOPE score	GA341 score	z-score	Validação
LALBA referência	B.99990112	901,15851	-12414,32031	1,00000	-6,01	-15.11
	B.99990481	844,52271	-12538,43652	1,00000	-5,77	-15.74
	B.99990159	860,30646	-12570,94531	1,00000	-5,96	-14.78*
	B.99990556	797,74188	-12606,90332	1,00000	-5,97	-15.21
	B.99990648	882,27399	-12636,26172	1,00000	-5,93	-15.46
LALBA variante	B.99990666	790,40204	-12458,63477	1,00000	-5,81	-15.96
	B.99990614	804,94427	-12495,43555	1,00000	-5,82	-16.08
	B.99990239	953,33942	-12505,79297	1,00000	-5,91	-15.21
	B.99990717	788,54932	-12515,17480	1,00000	-5,85	-15.37
	B.99990503	734,79504	-12516,64063	1,00000	-6,05	-14.86*

Nota: *Melhores modelos. Para a referência foi usado o molde 1F6R da PBD, acesso em 2015.

O melhor modelo 3D para a proteína com a mutação foi sobreposto com o melhor modelo 3D da proteína original (Figura 21). As posições da substituição dos aminoácidos estão marcadas de vermelho. Esse resíduo se encontra numa folha-beta, na parte N-terminal da proteína. O SNV BTA5:31348386 causa a mutação Ile60Val, que está presente no domínio da proteína.

Figura 21 – Modelos preditos para a proteína LALBA, selvagem e com a mutação Ile60Val



Em A, o modelo 3D da sequência com o alelo selvagem, com a isoleucina na posição 60, mostrada na seta em verde. Em B, o modelo 3D da sequência mutada, contendo a valina na posição 60, mostrada em vermelho. Em C, os modelos 3D sobrepostos. Os círculos amarelos representam as ligações dissulfeto. Apesar desse SNV missense estar localizado em uma folha-beta, o modelo da estrutura 3D da proteína mutada não apresenta nenhuma distorção óbvia à inspeção visual.

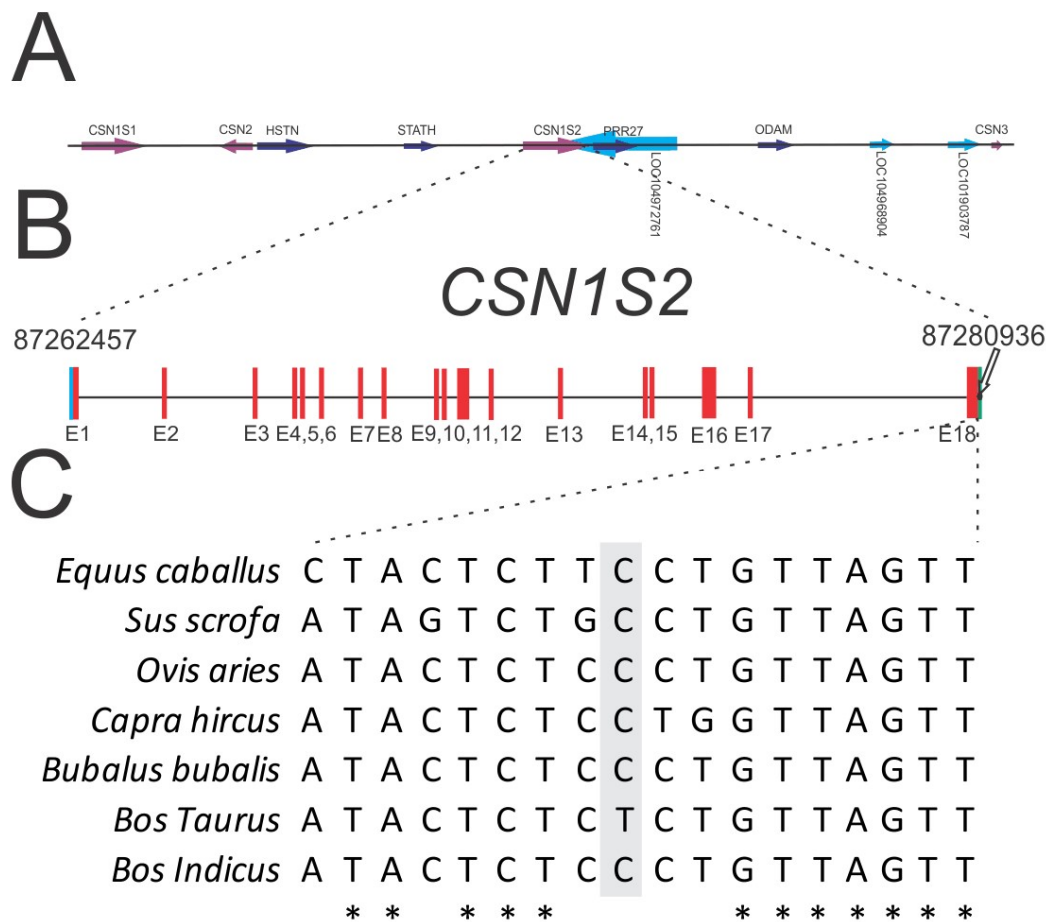
Fonte: Produzida pela autora em 2015.

3.4 Análises de impacto das demais variantes

3.4.1 Mutação na 3'-UTR do gene CSN1S2

A SNV BTA6:87280796 no gene *CSN1S2* abole um sítio de reconhecimento do bta-miR-340. A localização cromossômica do gene e da mutação e os resultados da conservação evolutiva da posição nucleotídica para o SNV BTA6:87280796 são mostrados na figura 22; os resultados do alinhamento múltiplo de uma região maior são mostrados na Fig. Suplementar XXX, no Apêndice. A localização desta variante na 3'-UTR é mostrada na Figura 23. Apesar de não afetar uma posição conservada evolutivamente, esse SNV abole um sítio de reconhecimento de miRNA e pode ser funcional, como discutido abaixo.

Figura 22 - Análise de conservação evolutiva para o SNV BTA6:87280796 no gene *CSN1S2*



Em A, o segmento do cromossoma ao redor do gene *CSN1S2* é mostrado. Em B, a posição do SNV, na região 3'UTR do gene *CSN1S2* é mostrada. Em C, o resultado do alinhamento múltiplo evidenciando a conservação evolutiva. As espécies incluídas na comparação evolutiva apresentam o alelo C na posição, exceto para *Bos taurus*, que apresenta o alelo T. O alelo divergente está presente em *Bos taurus*.

Fonte: Produzida pela autora em 2015.

Figura 23 - Posição da SNV BTA6:87280796 no gene *CSN1S2*

```

ATATTCCATTGCCTGGACTACTTGTCTTCCTTTTAGGAAACGAGGACAAAGCAAACATGAAGTTCTTCAT
CTTTACCTGCCTTTTGGCTGTTGCCCTTGCAAAGAATACGATGGAACATGTCTCCTCCAGTGAGGAATCT
ATCATCTCCCAGGAAACATATAAGCAGGAAAAGAATATGGCCATTAATCCCAGCAAGGAGAACCTTTGCT
CCACATTCTGCAAGGAAGTTGTAAGGAACGCAAATGAAGAGGAATATTCTATCGGCTCATCTAGTGAGGA
ATCTGCTGAAGTTGCCACTGAGGAAGTTAAGATTACTGTGGACGATAAGCACTACCAGAAAGCACTGAAT
GAAATCAATCAGTTTTATCAGAAGTTCCCCCAGTATCTCCAGTATCTGTATCAAGGTCCAATTGTTTTGA
ACCCATGGGATCAGGTTAAGAGGAATGCTGTTCCATTACTCCCACTCTGAACAGAGAGCAGCTCTCCAC
CAGTGAGGAAAATTCAAAGAAGACCGTTGACATGGAATCAACAGAAGTATTCATAAGAAAATAAACTG
ACTGAAGAAGAAAAGAATCGCCTAAATTTTCTGAAAAAAATCAGCCAGCGTTACCAGAAATTCGCCTTGC
CCCAGTATCTCAAGACTGTTTATCAGCATCAGAAAGCTATGAAGCCATGGATTCAACCTAAGACAAAGGT
TATTCCTATGTGAGGTACCTTTTAGATTCTGGAATTAAGTCTTCTACCTGGTTATGGTTGGACTGGAA
AATCTATCTTCTACAGTTTCATATCTACCACTTTACTTCATACAACCAGCATGTTTGGAGTGGTGCATCA
ATAAGACAAATCGCAGAGTATTTCTGAATTATTTAACTCTCCTGTTAGTTTATATTGAATATACTGG
GTTTATATGGTGGTTTCATACAACTTAGCTGATGATTATTTAAATCCTTTCCCTACCTTTCTGAGCT
ATAGAAATATTTTCTTCCCTCTGAGAATAAAATTTCAAGTGC-poly(A) tail

```

Pink - stop codon

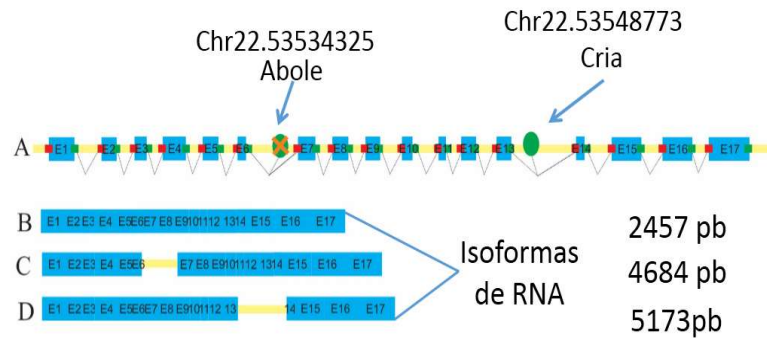
Red - variation rs109261203

Bold - AU-rich element (ARE)

Parte da sequência de *CSN1S2*, indicando em pink o codon de parada da tradução, a variante em vermelho e em negro, elementos ricos em AU (ARE). Sublinhado está o sítio de miRNA abolido.

3.4.2 Mutação em sítios de *splicing* no gene *LTF*

Das 38 variantes analisadas, duas afetam sítios de reconhecimento de *splicing*. A variante intrônica BTA22:53534325 abole um sítio aceptor de *splicing* e a variante intrônica BTA22:53548773 cria um sítio doador de *splicing*. Na figura 24, os sítios de *splicing* alternativo para o gene *LTF*. A primeira variante está localizada entre os éxons 6 e 7 e a segunda entre os éxons 13 e 14. Nas letras B, C e D são mostradas as isoformas de mRNA formadas a partir de cada alelo.

Figura 24 - Sítios de *splicing* alternativo para o gene *LTF*

Em A, os exons são mostrados em azul, introns em amarelo, sítios de *splicing*: bordo 5' do intron, em verde, e bordo 3' do intron, em vermelho. A primeira variante está localizada entre os éxons 6 e 7 e a segunda entre os éxons 13 e 14. Nas letras B, C e D são mostradas as isoformas de mRNA. Na letra B, a isoforma da referência com 2.457 pares de base. Na letra C, a isoforma que seria formada se o sítio do SNV BTA22:53534325 não fosse abolido, com 4.684 pares de bases. Na letra D, a isoforma que seria formada se o sítio do SNV BTA22:53548773 fosse usado, com 5.173 pares de bases (adaptado de Vogel and Motulsky 2013).

Para verificar se estes SNVs são potencialmente funcionais, nós previmos as sequências de mRNA e de proteína geradas por estas mutações e buscamos por estas sequências em bases de dados públicas e nenhuma delas foi identificada (Fig. 25 e 26). Além disto, foi feita a predição dos modelos 3D das proteínas (Tabela 9) e foi observado que esses SNVs induzem a formação de stop códons prematuros.

Figura 25 - Sequência da proteína com o *splicing* alternativo para a variante BTA22:53534325

LTF referência

MKLFVPALLSLGALGLCLAAPRKNVRWCTISQPEWFKCRRWQWRMKKLGAPSITCVRRFA
ALECIRAIAEKKADAVTLDGGMVFEAGRDPYKLRPVAAEIYGTKEPQTHYYAVAVVKKG
SNFQLDQLQGRKSCHTGLGRSAGWIIPMGILRPYLSWTESLEPLQGAVAKFFSASCVP
DRQAYPNLCQLCKGEGENQCACSSREPYFGYSGAFKCLQDGAGDVAFVKETTTFV**ENLPEK**
ADRDQYELLCLNNSRAPVDAFKECHLAQVPSHAVVARSVDGKEDLIWKLLSKAQEKFGKN
KSRSFQLFGSPPGQRDLLFKDSALGFLRIPSKVDSALYLGSRYLTLTKNLRETAEEVKAR
YTRVVWCAVGPEEQKKCQQWSQQSGQNVTCATASTDDCIVLVLKGEADALNLDGGYIYT
AGKCGLPVLAENRKSSKHSSLDVLRPTGYLAVAVVKKANEGLTWNLSLKDKKSCHTAV
DRTAGWNIPMGLIVNQTGSCAFDEFFSQSCAPGADPKSRLCALCAGDDQGLDKCVPNSKE
KYYGYTGAFRCLAEDVGDVAFVKNDTVWENTNGESTADWAKNLNREDFRLLCLDGT
TEAQSCHLAVAPNHAVVSRSDRAAHVKQVLLHQQALFGKNGKNCPCDKFCLFKSETKNLLF
NDNTECLAKLGGRPITYEYLGTEYVTAIANLKKCSTSPLEACAFLTR

BTA 22:53534325

MKLFVPALLSLGALGLCLAAPRKNVRWCTISQPEWFKCRRWQWRMKKLGAPSITCVRRFA
ALECIRAIAEKKADAVTLDGGMVFEAGRDPYKLRPVAAEIYGTKEPQTHYYAVAVVKKG
SNFQLDQLQGRKSCHTGLGRSAGWIIPMGILRPYLSWTESLEPLQGAVAKFFSASCVP
DRQAYPNLCQLCKGEGENQCACSSREPYFGYSGAFKCLQDGAGDVAFVKETTTFV**GKGRRE**
ATGTASFALLSFHFIVNPSSLI

Nessa figura são mostrados as sequências da proteína do LTF de referência e a isoforma da proteína que seria formada se o sítio do SNV BTA22:53534325 não fosse abolido. O trecho mostrado em amarelo são os

aminoácidos que seriam perdidos em função desta variante genética. E o trecho em rosa são os aminoácidos que seriam inseridos em função desta variante genética antes da criação de um stop codon prematuro.

Figura 26 - Sequência da proteína com o *splicing* alternativo para a variante BTA22:53538773

LTF referência

MKLFVPALLSLGALGLCLAAPRKNVRWCTISQPEWFKCRRWQWRMKKLGAPSITCVRRAF
ALECIRAIAEKKADAVTLDGGMVFEAGRDPYKLRPVAAEIYGTKEPQTHYYAVAVVKKG
SNFQLDQLQGRKSCHTGLGRSAGWIIPMGILRPYLSWTESLEPLQGAVAKFFSASCVP
DRQAYPNLCQLCKGEGENQCACSSREPYFGYSGAFKCLQDGAGDVAFAVKETTVFENLPEK
ADRDQYELLCLNNSRAPVDAFKECHLAQVPSHAVVARSDGKEDLIWKLLSKAQEKFGKN
KRSRFLFGSPPGQRDLLFKDSALGFLRIPSKVDSALYLGSRYLTTLNKLNRETAEEVKAR
YTRVVWCAVGPEEQKKCQQWSQQSGQNVTCATASTDDCIVLVKGEADALNDGGYIYT
AGKCGLPVLAENRKSSKHSSLDCLVRPTEGYLAVAVVKKANEGLTWNLSLKDCKKSCHTAV
DRTAGWNIPMGLIVNQTGSCAFDEFFSQSCAPGADPKSRLCALCAGDDQGLDKCVPNSKE
KYYGYTGAFR**CLAEDVGDVAFVKN**DTVWENTNGESTADWAKNLNRED**FRLLCLD**GTRKPV
TEAQSCHLAVAPNHAVVSRSDRAAHVKQVLLHQQALFGKNGKNCPKDFCLFKSETKNLLF
NDNTECLAKLGGRPTYEEYLGTEYVTAIANLKKCSTSPLEACAFLTR

BTA 22:53538773

MKLFVPALLSLGALGLCLAAPRKNVRWCTISQPEWFKCRRWQWRMKKLGAPSITCVRRAF
ALECIRAIAEKKADAVTLDGGMVFEAGRDPYKLRPVAAEIYGTKEPQTHYYAVAVVKKG
SNFQLDQLQGRKSCHTGLGRSAGWIIPMGILRPYLSWTESLEPLQGAVAKFFSASCVP
DRQAYPNLCQLCKGEGENQCACSSREPYFGYSGAFKCLQDGAGDVAFAVKETTVFENLPEK
ADRDQYELLCLNNSRAPVDAFKECHLAQVPSHAVVARSDGKEDLIWKLLSKAQEKFGKN
KRSRFLFGSPPGQRDLLFKDSALGFLRIPSKVDSALYLGSRYLTTLNKLNRETAEEVKAR
YTRVVWCAVGPEEQKKCQQWSQQSGQNVTCATASTDDCIVLVKGEADALNDGGYIYT
AGKCGLPVLAENRKSSKHSSLDCLVRPTEGYLAVAVVKKANEGLTWNLSLKDCKKSCHTAV
DRTAGWNIPMGLIVNQTGSCAFDEFFSQSCAPGADPKSRLCALCAGDDQGLDKCVPNSKE
KYYGYTGAFR

Nessa figura são mostrados as sequências da proteína do LTF de referência e a isoforma da proteína que seria formada se o sítio do SNV BTA22:53548773 fosse usado. Observe que a proteína ficaria encurtada, perdendo o trecho de aminoácidos mostrado em amarelo, em função da criação de um stop codon prematuro.

Dos quatro moldes usados no Modeller para a predição dos modelos para esta proteína, o melhor molde foi o 1BLF. Os valores de GA431 obtidos foram 1,00000 para os 1.000 modelos gerados tanto para a sequência selvagem da proteína LTF, quanto para as sequências das isoformas geradas pelas mutações (Tabela 9).

Tabela 9 - Resultados da modelagem para os cinco melhores modelos 3D

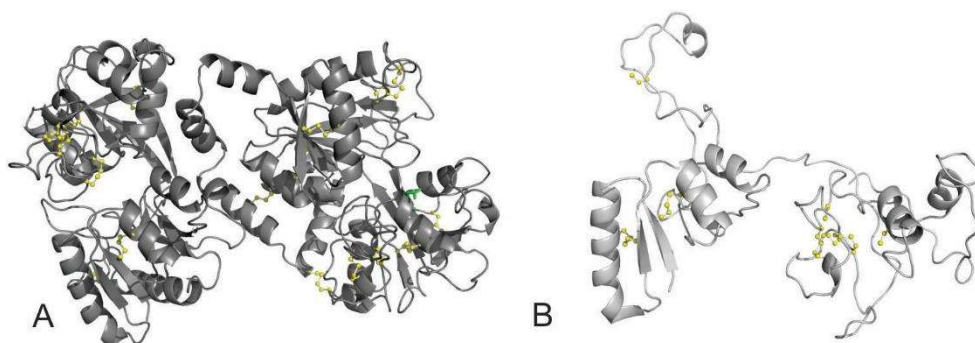
Proteína	Arquivo	Molpdf	DOPE score	GA341 score	z-score	Validação
LTF SNP BTA22:53534325	B.99990163	179,177.246	-16997,94727	1,00000	-5,14	-16.98
	B.99990184	181,513.538	-17138,00781	1,00000	-5,23	-16.33
	B.99990796	195,435.669	-17166,91602	1,00000	-5,42	-15.48*
	B.99990890	184,566.150	-17185,71484	1,00000	-5,23	-16.33
	B.99990908	181,531.091	-17197,84961	1,00000	-4,98	-16.46
LTF SNP BTA22:53548773	B.99990076	355,943.140	-51521,87500	1,00000	-9,66	-8.98
	B.99990144	357,304.785	-51593,57031	1,00000	-9,80	-8.92
	B.99990441	331,663.623	-51797,88672	1,00000	-9,64	- 8.97
	B.99990097	359,422.559	-51834,56250	1,00000	-9,77	-8.83
	B.99990180	334,642.603	-51879,35547	1,00000	-9,95	-8.76*

Nota: *Melhores modelos. Para a referência foi usado o molde 1BFL da PBD, acesso em 2015.

Os melhores modelos 3D para as proteínas com as mutações foram sobrepostos com o melhor modelo 3D da proteína original (Figura 27 e Figura 28). Na figura 27, os efeitos do SNV BTA22:53534325 na proteína, quando o sítio é abolido. Na figura 28, os efeitos do SNV BTA22:53548773 na proteína, quando o sítio é ativo.

A proteína do LTF possui 708 aminoácidos. A presença do SNV BTA22:53534325 resultaria em uma isoforma com 263 aminoácidos (Figura 27). A presença do SNV BTA22:53548773 resultaria em uma isoforma com 550 aminoácidos (Figura 28).

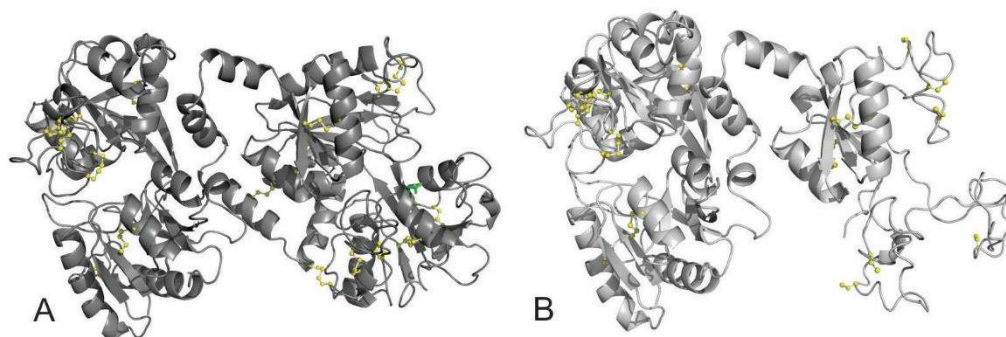
Figura 27 – Modelos preditos para a proteína LTF, na presença da SNV BTA22:53534325



Em A, o modelo 3D da sequência de referência. Os círculos amarelos representam as ligações dissulfeto. Em B, o modelo 3D da sequência contendo o novo alelo do SNV. Lembrando que a proteína do LTF usualmente tem 708 aminoácidos. Essa variação produz uma variante da proteína com 263 aminoácidos. Do primeiro aminoácido ao 234, a sequência é igual a de referência. A partir desse aminoácido, a sequência difere com a inserção de um stop códon prematuro.

Fonte: Produzida pela autora em 2016.

Figura 28 – Modelos preditos para a proteína LTF, na presença da BTA22:53548773



Em A, o modelo 3D da sequência de referência. Nas bolas em amarelo está representado as ligações dissulfeto. Em B, o modelo 3D da sequência contendo o novo alelo do SNV. Lembrando que a proteína do LTF usualmente tem 708 aminoácidos. Essa variação produz uma variante da proteína com 550 aminoácidos. Sendo todos os aminoácidos iguais a sequência de referência terminando com a inserção de um stop códon prematuro.

Fonte: Produzida pela autora em 2016.

Nota: Subsequentemente a essas análises, fizemos a busca de sítios de reconhecimento de *splicing* usando o programa *Human Splicing Finder* (HSF3.0) e estes efeitos sobre sítios de reconhecimento de *splicing* não se confirmaram.

3.4.3 SNVs que afetam sítios de reconhecimento a miRNA

Quatro SNVs (BTA6:87280796, BTA5:31349177, BTA22:53548773 e BTA22:53552302) alteram sete sítios de reconhecimento à miRNAs, abolindo três sítios e criando quatro novos sítios (Tabela 7). Esses sete sítios são: bta-miR-340, bta-miR-376a, bta-miR-376d, bta-miR-376e, bta-miR-1814b, bta-miR-2429 e bta-miR-2444.

3.4.4 Conservação evolutiva

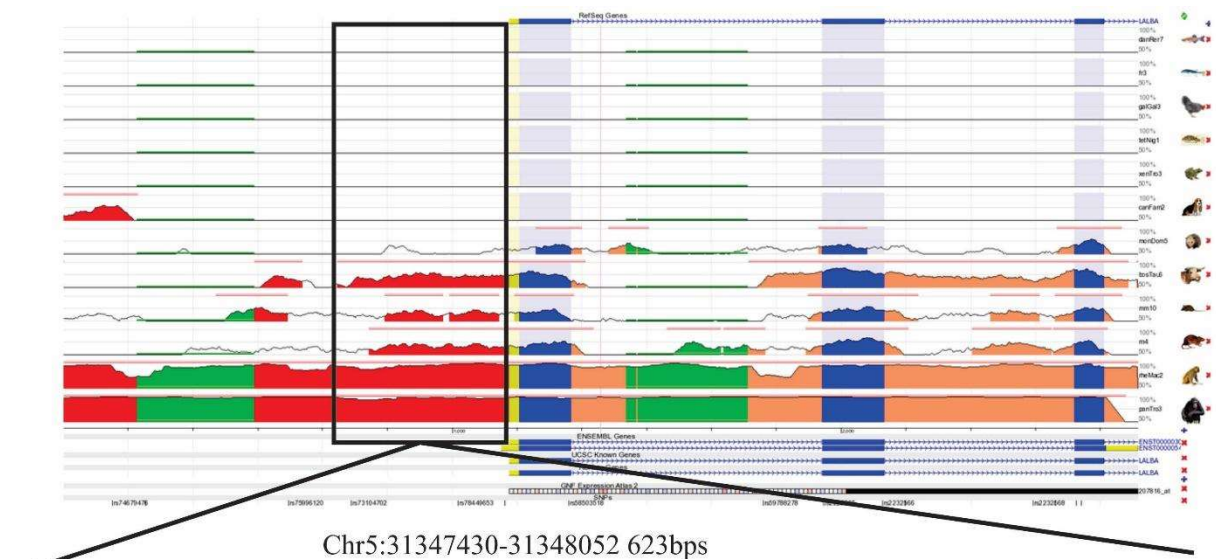
Duas diferentes estratégias foram usadas para avaliar a conservação evolutiva de posições/regiões de nucleotídeos afetados pelas SNVs. Inicialmente, a análise foi conduzida usando alinhamento múltiplo. Adicionalmente, foi realizada a pesquisa por ECR.

Considerando o resultado do alinhamento múltiplo, todas as 38 variantes (SNVs e INDELs) foram analisadas em relação a mudança em posições conservadas evolutivamente. Em 21 (55,26%) dessas posições, o alelo taurino é o divergente e em 17 (44,74%) dessas posições, o alelo zebuino é o divergente.

São essas dezessete variantes (BTA6:87153410; BTA6:87374090; BTA5:31343847; BTA5:31346231; BTA5:31347426; BTA5:31347538; BTA5:31347611; BTA5:31348386; BTA5:31349118; BTA5:31349177; BTA5:31350957; BTA5:31353933; BTA22:53546045; BTA22:53547194; BTA22:53552302; BTA22:53560259 e BTA22:53560817) que afetam uma posição conservada evolutivamente (Figuras Suplementares 1, 2 e 3 presentes no Apêndice).

3.4.5 SNVs em ECRs

Em relação ao outro método, foi utilizado o ECR Browser para realizar as análises de ECR e foram obtidas cinco SNVs em regiões de ECR, todas localizadas no gene *LALBA* (Figura 29). Duas (BTA5:31347538 e BTA5:31347611) dessas cinco SNVs afetam uma posição conservada evolutivamente. Para as demais não foi possível detectar sua potencialidade funcional.

Figura 29 – ECRs no gene *LALBA*

TGACATATGATGATGTACAGTCCTTTTCCATATTCTGTATGTCTCTAAGAGGAAGGAGGA
 GTTGGCCGTGGACCCTTTGTGCATTTTCTGATTGCTTCACTTGTATTAACCTGAGGCCC
 CCTTTGTTCTGAAATAGGTTGGGCACATCTTGCTTCCTAGAACCAACACTACCAGAAAC
 ACATAAATAAAGCCAAATGGGAAACAGGATCATGTTTGTAACACTCTTTGGGCAGGTAA
 CAATACCTAGTATGGACTAGAGATTCTGGGGAGGAAAGGAAAAGTGGGGGTGAAATTACTG
 AAGGAAGCTCAATGTTTCTTTGTTGGTTTTACTGGCCTCTCTTGTGCATCCTCTTCTCTGGA
 TGTAAGGCTTGATGCCAGGGCCCCCTAAGGCTTTTCCACAAATAAAGGAGGTGAGCAGT
 GTGGTGACCCCATTTTCAGAATCTTGGGGGGTAACCAAAAATGATGTCCTTTGTCTCTCTGC
 TCCTGGTAGGCATCCTATTCCATGCCACCCAGGCTGAACAGTTAACAATAATGTGAGGTGT
 TCCGGGAGCTGAAAGACTTGAAGGGCTACGGAGGTGTCAGTTTGCCTGAATGTGAGTTCC
 CTGCTATTTTGTCTTTGTCCCATA

SNVs estão localizadas no intervalo BTA5:31347430-31348052. Na parte de cima da figura, podemos observar na cor azul, os exons codantes; em verde, os transposons e repetições simples; em laranja, regiões intrônicas; em vermelho, regiões intergênicas; em amarelo, as UTRs; e, grifado de rosa, as ECRs. A citosina e adenina em vermelho localizados em BTA5:31347538 e BTA5:31347611, afetam posições conservadas evolutivamente. Para as timinas e a adenina em azul (BTA5:31347772, BTA5:31347456 e BTA5:31347479) não foi encontrado evidência de repercussão funcional nas outras análises.

Fonte: Produzida pela autora em 2015.

4 DISCUSSÃO

O melhoramento genético em bovinos apresentou grande desenvolvimento nos últimos anos, em função do desenvolvimento de grandes plataformas de genotipagem de SNPs e, mais recentemente, de sequenciamento de nova geração. Além disso, novos métodos estatísticos, como a seleção genômica, deram novo impulso ao melhoramento genético. A eficácia destes métodos depende de diversos aspectos. Os mais importantes são a qualidade da fenotipagem e o tamanho efetivo da população. Desta forma, raças diferentes terão impacto diferente destas metodologias (Wiggans et al. 2017). Raças com grandes contingentes populacionais tem apresentado grande progresso em termos de melhoramento genético usando seleção genômica, para outras, o processo de seleção convencional é o mais apropriado.

É importante ter em mente que a pesquisa básica, que leva ao entendimento destes processos não pode parar. Além disto, é necessário que se adapte a demandas locais. Neste contexto, estamos investigando os processos metabólicos subjacentes a produção leiteira em duas das raças zebuínas mais importantes do cenário de produção leiteira brasileiro, Gir e Guzerá. A abordagem escolhida, sequenciamento de genomas completos, descoberta de variantes potencialmente funcionais, para aplicação em estudos de associação é linear e tem ótimas perspectivas de sucesso.

No presente estudo, foi desenvolvida a predição de impacto funcional de variantes genéticas em genes que codificam as principais proteínas do leite. Em função do grande número de variantes descobertas, usou-se um filtro com uma sequência de critérios de exclusão, incluindo conservação evolutiva (região e posição nucleotídica), impacto na regulação da transcrição e na sequência do RNA (*splicing*, modificações pós-síntese, ligação de miRNAs), na regulação da tradução, na sequência da proteína e nas modificações pós-síntese de proteína, entre outros fenômenos investigados.

Foram encontrados 64 SNVs e seis INDELs nos sete genes em estudo. Destas, duas variantes são *missense*. Uma no gene *LTF*, Thr546Asn, e uma no gene *LALBA*, Ile60Val, ambas afetando posições conservadas evolutivamente. Apesar dessas variantes implicarem em trocas de aminoácidos, ambas as predições dos modelos 3D não mostram nenhuma distorção óbvia. Entretanto, a modelagem 3D de proteínas sofre com a falta de dados sobre estrutura quaternária de muitas destas moléculas. Os efeitos funcionais dessas variantes podem estar ligados a estabilidade da proteína ou a efeitos sobre a modificação a modificações pós-síntese, à regulação da meia vida da proteína, à sua habilidade de formar multímeros. Assim, estas duas variantes são ótimas candidatas para estudos funcionais.

Além dessas variantes, foi encontrada uma variante no gene *CSNIS2* na região 3'-UTR, abolindo um sítio de reconhecimento do bta-mir-340. Além dessa variante, foram encontradas mais três variantes alterando sítios de reconhecimento de microRNA. A literatura é escassa para a maioria desses miRNA, porém alguns serão discutidos. Estudos funcionais têm fornecido suporte de funcionalidade para alguns miRNAs que são afetados por SNVs descritos no presente estudo. Por exemplo, três desses miRNAs são diferencialmente expressos em bovinos: miR-340 é menos expresso em placenta (Liu, Jin et al. 2014); miR-376e é altamente expresso em gordura intramuscular, em um estudo realizado procurando os microRNAs envolvidos na deposição de gordura na carne bovina (Wang, Zheng et al. 2013); e, neste mesmo estudo, foi observado que o mir-2429 é menos expresso em gordura intramuscular (Wang, Zheng et al. 2013). Apesar de nenhum estudo ter sido encontrado relacionando esses microRNAs com as proteínas do leite, esses estudos mostram que eles estão presentes no controle da expressão gênica e essas variantes devem ser investigadas no futuro.

Foram encontradas também duas variantes alterando sítios de reconhecimento de *splicing* alternativo. As isoformas que seriam geradas por estas variantes não foram encontradas em bancos de dados. Em bovinos, foi descrito na literatura que SNPs em sítios de *splicing* afetam a expressão do gene *CD46* e *NCF4*, quando envolvidos na susceptibilidade a mastite (Wang, Zhong et al. 2014; Ju, Wang et al. 2015). Nossos resultados sugerem que o SNV Chr22:53534325 pode ter repercussão funcional, porque abole um sítio de *splicing* críptico. Entretanto, essa repercussão seria funcional se essa isoforma realmente existisse. Portanto, como as isoformas resultantes da mutação não foram encontradas em bancos de dados de RNA ou proteína, não encontramos suporte para a repercussão funcional desse SNV.

Em relação ao SNV BTA22:53538773, nossos resultados sugerem que ele pode ter repercussão funcional, levando a geração de uma nova isoforma com 550pb. Essa isoforma apresenta 75% de similaridade entre a sequência da proteína LTF disponível em bancos de dados e o segundo domínio de transferrina incompleto. Apesar de não ter sido encontrada essa isoforma em bancos de dados, o SNV BTA22:53538773 constitui um bom exemplo de variante, que com base na repercussão funcional *in silico*, deve ser incluída em futuros estudos de associação. É importante considerar que no que toca a isoforma, particularmente as tecido-específicas, a informação disponível ainda é muito incompleta.

Em relação a conservação evolutiva foram usadas duas metodologias diferentes para avaliar a conservação evolutiva: o alinhamento múltiplo e a pesquisa em ECR. O alinhamento múltiplo foi usado para caracterizar a conservação evolutiva em uma posição de nucleotídeo

afetado por uma variante ou em pequenos segmentos próximos. A pesquisa em ECRs permite identificar as regiões conservadas, até fora dos genes. Apenas cinco variantes no gene *LALBA*, estão presentes em regiões de ECRs. Até o presente estudo, não foi encontrado na literatura associação entre SNPs em ECRs no gene *LALBA*. Variantes em ECRs podem induzir modificações na proteína, alterar sítios de *splicing*, mudar sítios de reconhecimento de miRNA e afetar os padrões de expressões. Portanto, as cinco SNVs no gene *LALBA* presentes em ECRs são candidatas interessantes para investigações futuras.

É importante considerar que este estudo tem limitações. A aplicação de uma sequência de filtros de exclusão certamente faz com que se percam variantes potencialmente funcionais. Isto certamente aconteceu neste estudo e não temos ferramentas para estimar a magnitude desta perda. As variantes retidas apresentaram evidências de funcionalidade em mais de um nível, pois em sua maioria colocalizaram com regiões/posições nucleotídicas evolutivamente conservadas. Além disto, geralmente outro critério foi identificado. Conforme esperado, a inferência de impacto é mais clara para mutações *missense*. Para outras evidências, como efeito sobre o reconhecimento do DNA ou do RNA por miRNA ou efeitos sobre sítios de *splicing* alternativo, faltam informações necessárias, como bases de dados de expressão em diferentes tipos celulares e estágios de desenvolvimento.

Apesar destas limitações, a estratégia adotada aqui se mostrou eficiente na identificação de variantes genéticas com maior potencial de repercussão funcional, para inclusão em estudos subsequentes.

5 CONCLUSÕES

Dentre os zebuínos leiteiros, as raças Guzerá e Gir são de grande importância para o cenário da pecuária nacional, contribuindo como produtoras de leite e carne e como fornecedoras de material genético para cruzamentos. O leite produzido por essas raças possui maior concentração de proteínas do que o leite produzido por representantes das raças taurinas, evidenciando uma importante ferramenta para a produção de leite mais saudável para a dieta humana. Neste contexto, no presente trabalho foram realizadas análises funcionais *in silico* de variantes na sequência de DNA dos principais genes das proteínas do leite nas raças Guzerá e Gir. Essas análises permitiram identificar 70 variantes (SNPs e INDELs) nos genes do *cluster* das caseínas (*CSN1S1*, *CSN2*, *CSN1S2* e *CSN3*), no gene *LALBA*, no gene *LGB* e no gene *LTF*. Desse total de 70 variantes, 20 foram consideradas potencialmente funcionais por alterarem sítios de *splicing* e de reconhecimento de miRNA, afetarem posições conservadas evolutivamente e/ou afetarem a estrutura terciária da proteína. Entre esses, 3 merecem destaque e deveriam ser estudadas primeiramente em estudos funcionais futuros, as SNVs BTA6:87280796, BTA5:31348386 e BTA22:53546045. As SNVs BTA5:31348386 e BTA22:53546045 estão localizadas na região exônica, são missense e ambas alteram uma posição conservada evolutivamente. A SNV BTA6:87280796 está localizada na região 3'UTR e altera um sítio de ligação ao bta-miR-340. Desta forma, o presente projeto identificou variações que podem ser responsáveis pelas principais diferenças na composição e concentração de proteínas no leite entre zebuínos e taurinos. Além disso, ele demonstra que análises de bioinformática podem ajudar a identificar e selecionar, entre milhares de variantes normalmente detectadas após o sequenciamento do genoma completo, aquelas que devem ser consideradas primeiro nos estudos de associação e análises funcionais.

6 PERSPECTIVAS

Como perspectivas do presente trabalho, pretende-se inferir os haplótipos das caseínas para as raças estudadas. Além disso, serão realizados estudos de associação entre as variantes potencialmente funcionais identificadas no presente estudo e a concentração de proteínas no leite.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ABCG. Associação dos Criadores de Guzerá do Brasil. Disponível em: <http://www.guzera.org.br/imagens/submenu/arquivo06797.pdf>. Acesso em: 20 de novembro. 2014.

ANSORGE, Wilhelm J. Next-generation DNA sequencing techniques. **New biotechnology**, v. 25, n. 4, p. 195-203, 2009. <https://doi.org/10.1016/j.nbt.2008.12.009>

ARTIMO, P. et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res*, v. 40, n. W1, p. W597-W603, 2012. <https://doi.org/10.1093/nar/gks400>.

AZEVEDO, A. et al. Genetic polymorphism of the kappa-casein gene in Brazilian cattle. *Genet Mol Res*, v. 7, n. 3, p. 623-630, 2008. <https://doi.org/10.4238/vol7-3gmr428>.

BERMAN, H. M. et al. The protein data bank. *Nucleic Acids Res*, v. 28, n. 1, p. 235-242, 2000. <https://doi.org/10.1093/nar/30.1.245>.

BOLGER A. M.; LOHSE M.; USADEL B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014. <https://doi.org/10.1093/bioinformatics/btu170>.

CAETANO, Alexandre Rodrigues. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, p. 64-71, 2009. <https://doi.org/10.1590/S1516-35982009001300008>

Campolina, D. P. et al. Programa Nacional de Melhoramento do Guzerá para Leite: resultados do Teste de Progenie, do Programa de Melhoramento Genético de Zebuínos da ABCZ e do Núcleo MOET. 2014.

CAVASOTTO, Claudio N.; PHATAK, Sharangdhar S. Homology modeling in drug discovery: current trends and applications. **Drug discovery today**, v. 14, n. 13-14, p. 676-683, 2009. <https://doi.org/10.1016/j.drudis.2009.04.006>

CECCHINATO, Alessio et al. Effects of β -lactoglobulin, stearoyl-coenzyme A desaturase 1, and sterol regulatory element binding protein gene allelic variants on milk production, composition, acidity, and coagulation properties of Brown Swiss cows. **Journal of dairy science**, v. 95, n. 1, p. 450-454, 2012. <https://doi.org/10.3168/jds.2011-4581>

CHAUDHURI, Arunima; CHATTOPADHYAY, Amitabha. Lipid binding specificity of bovine α -lactalbumin: a multidimensional approach. **Biochimica et Biophysica Acta (BBA)-Biomembranes**, v. 1838, n. 8, p. 2078-2086, 2014. <https://doi.org/10.1016/j.bbamem.2014.04.027>

CHOI, J.W. et al. Massively parallel sequencing of Chikso (Korean brindle cattle) to discover genome-wide SNPs and InDels. *Mol Cells*, v.36, n.3, p.203-11, 2013 <https://doi.org/10.1007/s10059-013-2347-0>.

UNIPROT CONSORTIUM. UniProt: a hub for protein information. **Nucleic acids research**, v. 43, n. D1, p. D204-D212, 2014. <https://doi.org/10.1093/nar/gku989>

CORPET, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, v. 16, n. 22, p. 10881-10890, 1988. <https://doi.org/10.1093/nar/16.22.10881>.

DE ALMEIDA, C. C.; JUNIOR CONTE C. A; SILVA, A. C. O.; ALVARES, T. S. Proteína do soro do leite: composição e suas propriedades funcionais. *Centro Científico Conhecer*, v. 9, n. 16, p. 1840-1854, 2013. Disponível em: <https://www.conhecer.org.br/enciclop/2013a/agrarias/proteina%20do%20soro.pdf>.

DIAS, J. C. Comportamento sexual, concentrações periféricas de testosterona e perfil cromatográfico de proteínas do sêmen com afinidade à heparina, em touros jovens da raça Guzerá. 2008. Tese (Doutorado em Zootecnia) -Universidade Federal de Minas Gerais, Belo Horizonte, 2008. <http://hdl.handle.net/1843/31797>.

DINESH, K. et al. Identification of polymorphism in exons 7 and 12 of lactoferrin gene and its association with incidence of clinical mastitis in Murrah buffalo. *Trop Anim Health Prod*, v. 47, n. 4, p. 643-647, 2015. <https://doi.org/10.1007/s11250-015-0765-z>.

FANG, B.; ZHANG M.; TIAN M. et al. Bovine lactoferrin binds oleic acid to form an anti-tumor complex similar to HAMLET. *Biochim Biophys Acta*, v. 1841, n. 4, p. 535-543, 2014. <https://doi.org/10.1016/j.bbali.2013.12.008>.

FARIA, FÁBIO JOSÉ CARVALHO et al. Variabilidade genética da raça Guzerá. **Proceedings of the V Simpósio da Sociedade Brasileira de Melhoramento Animal**, 2004.

FINN R.D.; BATEMAN A.; CLEMENTS J. et al. Pfam: the protein families database. *Nucleic Acids Res*, v.42, D222-30. <https://doi.org/10.1093/nar/gkt1223>.

FONTANELI, R. 2001. Fatores que afetam a composição e as características físicoquímicas do leite. Disponível em: www6.ufrgs.br/favet/lacvet/restrito/pdf/quimica_leite.pdf Acesso em 7 de novembro de 2015.

FREITAS, S. F. A. Estudo do Polimorfismo da Kappa-Caseína e Alfa-Lactoalbumina em Bovinos, Girolando, do Brasil, e Siboney, de Cuba. 2010.

GANAI, N. et al. Novel polymorphisms in the bovine β -lactoglobulin gene and their effects on β -lactoglobulin protein concentration in milk. *Anim Genet*, v. 40, n. 2, p. 127-133, 2009. <https://doi.org/10.1111/j.1365-2052.2008.01806.x>.

Gao Y, Lin X, Shi K, Yan Z, Wang Z. Bovine mammary gene expression profiling during the onset of lactation. *PLoS One*. 2013 Aug 21;8(8):e70393. doi: 10.1371/journal.pone.0070393. PMID: 23990904; PMCID: PMC3749150.

GONZÁLEZ-CHÁVEZ, S. A. et al. Lactoferrin: structure, function and applications. *Int J Antimicrob Agents*, v. 33, n. 4, p. 301. e1-301. e8, 2009. <https://doi.org/10.1016/j.ijantimicag.2008.07.020>.

FONTANELLI, Roberto Serena. Uso do leite para monitorar a nutrição e o metabolismo de vacas leiteiras. 2001.

Gupta, R. J., E.; and Brunak, S. Prediction of N-glycosylation sites in human proteins. In preparation. 2004.

Hamborg M, Rose F, Jorgensen L, Bjorklund K, Pedersen HB, Christensen D, Foged C. Elucidating the mechanisms of protein antigen adsorption to the CAF/NAF liposomal vaccine adjuvant systems: effect of charge, fluidity and antigen-to-lipid ratio. *Biochim Biophys Acta*. 2014 Aug;1838(8):2001-10. doi: 10.1016/j.bbamem.2014.04.013. Epub 2014 Apr 25. PMID: 24769435.

HARAGUCHI, Fabiano Kenji; ABREU, Wilson César de; PAULA, Heberth de. Proteínas do soro do leite: composição, propriedades nutricionais, aplicações no esporte e benefícios para a saúde humana. **Revista de nutrição**, v. 19, p. 479-488, 2006. <https://doi.org/10.1590/S1415-52732006000400007>

Hax, L. T. F., E.K.; Pereira, R.A.; Schneider, A.; Bianchi, I. e Corrêa, M.N. Diferenças entre vacas zebuínas (*Bos indicus*) e taurinas (*Bos taurus*) no desempenho reprodutivo pós-parto. Artigo Técnico, Cultivar.2009.

HENDRIX, T. M. et al. Energetics of structural domains in alpha-lactalbumin. *Protein Sci*, v. 5, n. 5, p. 923, 1996. <http://dx.doi.org/10.1002/pro.5560050514>.

HORNE, D. S. Casein micelle structure: models and muddles. *Curr Opin Colloid Interface Sci*, v. 11, n. 2, p. 148-153, 2006. <https://doi.org/10.1016/j.cocis.2005.11.004>.

HORNE, D. S. Casein structure, self-assembly and gelation. *Curr Opin Colloid Interface Sci*, v. 7, n. 5, p. 456-461, 2002. [https://doi.org/10.1016/S1359-0294\(02\)00082-1](https://doi.org/10.1016/S1359-0294(02)00082-1).

IBGE. 2015. Instituto Brasileiro de Geografia e estatística. Banco de Dados **Disponível em:** Acessado em: 04, Fev, 2016.

Ikonen T, Ojala M, Ruottinen O. Associations between milk protein polymorphism and first lactation milk production traits in Finnish Ayrshire cows. *J Dairy Sci*. 1999 May;82(5):1026-33. doi: 10.3168/jds.S0022-0302(99)75323-3. PMID: 10342242.

JØHNKE, M.; PETERSEN, T. E. The alpha-lactalbumin/oleic acid complex and its cytotoxic activity. *Milk Protein*, p. 97-953, 2012. <https://doi.org/10.5772/48184>.

JORGE, W. et al. Caracterização genética de bovinos por meio de estudos de cromossomo Y e do DNA mitocondrial. 2006.

JU, Z.; WANG, C.; WANG, X. et al. Role of an SNP in Alternative Splicing of Bovine NCF4 and Mastitis Susceptibility. *PloS one*, v. 10, n. 11, p. e0143705, 2015. <https://doi.org/10.1371/journal.pone.0143705>.

Kamiński S, Cieslińska A, Kostyra E. Polymorphism of bovine beta-casein and its potential effect on human health. *J Appl Genet*. 2007;48(3):189-98. doi: 10.1007/BF03195213. PMID: 17666771.

KANG, J. F. et al. Bioinformatics analysis of lactoferrin gene for several species. *Biochem Genet*, v. 46, n. 5-6, p. 312-322, 2008. <https://doi.org/10.1007/s10528-008-9147-9>.

Knudsen S. Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics*. 1999 May;15(5):356-61. doi: 10.1093/bioinformatics/15.5.356. PMID: 10366655.

KOZOMARA, A.; GRIFFITHS-JONES, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, v. 42, n. D1, p. D68-D73, 2014. <https://doi.org/10.1093/nar/gkt118>.

LASKOWSKI, R. A. et al. PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem Sci*, v. 22, n. 12, p. 488-490, 1997. [https://doi.org/10.1016/s0968-0004\(97\)01140-7](https://doi.org/10.1016/s0968-0004(97)01140-7).

LASKOWSKI, R. A. et al. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr*, v. 26, n. 2, p. 283-291, 1993. <https://doi.org/10.1107/S0021889892009944>.

LEÃO, G. F. M.; PIVATTO, D. R. D.; CARNIEL, H. et al. Melhoramento genético em zebuínos leiteiros—uma revisão. *Agropecuária Científica no Semiárido*, v. 9, n. 4, p. 09- 14, 2013. Disponível em: <http://www.fazendaestreladosulmg.com.br/resources/Melhoramento%20gen%C3%A9ti%20co%20em%20zebu%C3%ADnos%20leiteiros%20%E2%80%93%20uma%20revis%C3%A3o..pdf>.

LI, H.; DURBIN, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, v. 26, n. 5, p. 589-95, 2010. <https://doi.org/10.1093/bioinformatics/btp698>.

Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008 Mar 1;24(5):713-4. doi: 10.1093/bioinformatics/btn025. Epub 2008 Jan 28. PMID: 18227114.

LIU, F-J.; JIN, L-J.; MA, X-G. et al. Differentially expressed microRNAs and affected signaling pathways in placentae of transgenic cloned cattle. *Theriogenology*, v. 82, n. 2, p. 338-346. e3, 2014. <https://doi.org/10.1016/j.theriogenology.2014.04.010>.

Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*. 2008;9:387-402. doi: 10.1146/annurev.genom.9.081307.164359. PMID: 18576944.

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*. 2006 Sep;16(9):1182-90. doi: 10.1101/gr.4565806. Epub 2006 Aug 10. PMID: 16902084; PMCID: PMC1557762.

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJ, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res*. 2015 Jan;43(Database issue):D213-21. doi: 10.1093/nar/gku1243. Epub 2014 Nov 26. PMID: 25428371; PMCID: PMC4383996.

Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet*. 2010 Oct 15;19(R2):R131-6. doi: 10.1093/hmg/ddq400. Epub 2010 Sep 21. PMID: 20858594; PMCID: PMC2953750.

NADEEM, A.; SANBORN, J.; GETTEL, D. L. et al. Protein receptor-independent plasma membrane remodeling by HAMLET: a tumoricidal protein-lipid complex. *Sci Rep*, v. 5, n. 16432, 2015. <https://doi.org/10.1038/srep16432>.

NAKAMURA, Y. et al. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res*, v. 28, n. 1, p. 292-292, 2000. <https://doi.org/10.1093/nar/28.1.292>.

NCBI. National Center for Biotechnology Information. U.S. National Library of Medicine. Disponível em: <https://www.nlm.nih.gov/>. Acesso em: 10 de janeiro de 2016.

NILSEN, H.; OLSEN, H. G.; HAYES, B. et al. Casein haplotypes and their association with milk production traits in Norwegian Red cattle. *Genet Sel Evol*, v. 41, n. 1, p. 24, 2009. <https://doi.org/10.1186/1297-9686-41-24>.

O'HALLORAN, F. et al. Characterisation of single nucleotide polymorphisms identified in the bovine lactoferrin gene sequences across a range of dairy cow breeds. *Biochimie*, v. 91, n. 1, p. 68-75, 2009. <https://doi.org/10.1016/j.biochi.2008.05.011>.

OVCHARENKO, I. et al. ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res*, v. 32, n. suppl 2, p. W280-W286, 2004. <https://doi.org/10.1093/nar/gkh355>.

OVCHARENKO, I.; LOOTS, G. G.; GIARDINE, B. M. et al. Mulan: multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res*, v. 15, n. 1, p. 184-194, 2005. <https://doi.org/10.1101/gr.3007205>.

PANETO, J. C. C.; FERRAZ, J. B. C.; BALIEIRO, J. C. C. et al. Bos indicus or Bos taurus mitochondrial DNA-comparison of productive and reproductive breeding values in a Guzerat dairy herd. *Genet Mol Res*, v. 7, n. 3, p. 592-602, 2008. <https://doi.org/10.4238/vol7-3gmr449>.

PENNA, V. et al. Situação atual e potencialidades da raça Guzerá na pecuária leiteira nacional. Aspectos técnicos, econômicos, sociais e ambientais da atividade leiteira, v. 1, p. 103-110, 2005. Disponível em: <http://cbmegguzera.com.br/artigostecnicos/artigostecnicospdf/Situa%C3%A7%C3%A3o%20atual%20e%20peculiaridades%20da%20ra%C3%A7a%20Guzer%C3%A1%20na%20pecu%C3%A1ria%20leiteira%20nacional.pdf>.

Pereira, T. C. Introdução ao mundo dos microRNAs. Ribeirão Preto: Sociedade Brasileira de Genética. 2015.

QUEIROZ, S.; LÔBO, R. Genetic relationship, inbreeding and generation interval in registered Gir cattle in Brazil. *J Anim Breed Genet*, v. 110, n. 1-6, p. 228-233, 1993. <https://doi.org/10.1111/j.1439-0388.1993.tb00734.x>.

ROSSE, I. C. et al. Genoma do Guzerá e do Gir: Identificação e análises funcionais de polimorfismos nos principais genes envolvidos no metabolismo de lipídios da glândula mamária de zebuínos. 2015. Tese (Doutorado em Genética) - Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

SAKHTEMAN, A.; ZARE, B. Modelface: an application programming interface (API) for homology modeling studies using Modeller software. *Iran J Pharm Res*, v. 15, n. 4, p. 801-807, 2016. <https://pubmed.ncbi.nlm.nih.gov/28243276/>.

SANTOS FILHO, Osvaldo Andrade; ALENCASTRO, Ricardo Bicca de. Modelagem de proteínas por homologia. **Química Nova**, v. 26, p. 253-259, 2003. <https://doi.org/10.1590/S0100-40422003000200019>

SCHRÖDINGER, LLC. The PyMOL Molecular Graphics System, Version 1.8. November 2015. Disponível em: <https://pymol.org/2/>.

SHARMA, S. et al. C-lobe of lactoferrin: the whole story of the half-molecule. *Biochem Res Int*, v. 2013, 2013. <https://doi.org/10.1155/2013/271641>.

SILVA, Itiberê Saldanha et al. Avaliação de modelos para estimação de componentes de variância e parâmetros genéticos para características de crescimento de bovinos da raça Guzerá. **Revista Brasileira de Zootecnia**, v. 35, p. 1943-1950, 2006. <https://doi.org/10.1590/S1516-35982006000700009>

SILVA, S. V.; MALCATA, F. X. Caseins as source of bioactive peptides. *Int Dairy J*, v. 15, n. 1, p. 1-15, 2005. <https://doi.org/10.1016/j.idairyj.2004.04.009>.

SODERO, Ana Carolina Rennó et al. **Modelagem molecular de Aspartil Proteases de Schistosoma mansoni (SmAPs) para o desenvolvimento de potenciais moléculas esquistossomicidas**. 2011. Tese de Doutorado.

STOTHARD, P. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques*, v. 28, n. 6, p. 1102-1104, 2000. <https://doi.org/10.2144/00286ir01>.

SULIMOVA, G. E. et al. κ -casein gene (CSN3) allelic polymorphism in Russian cattle breeds and its information value as a genetic marker. *Russ J Genet*, v. 43, n. 1, p. 73-79, 2007. <https://doi.org/10.1134/S1022795407010115>.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013 Mar;14(2):178-92. doi: 10.1093/bib/bbs017. Epub 2012 Apr 19. PMID: 22517427; PMCID: PMC3603213.

TRAKOVICKÁ, A. et al. Kappa-casein gene polymorphism (CSN3) and its effect on milk production traits. *Acta Fytotechnica et Zootechnica*, v. 15, n. 3, 2012. http://www.acta.fapz.uniag.sk/journal/index.php/on_line/article/view/39.

VĂTĂSESCU, R. et al. Alpha-lactalbumin genotypes identification in Romanian Black Spotted cattle breed. *Scientific Papers Animal Science and Biotechnologies*, v. 41, n. 1, p. 169-173, 2008. ISSN 2344-4576.

VENTURINI, K. S. et al. Características do leite. *Boletim Técnico (Programa Institucional de Extensão) - Universidade Federal do Espírito Santo, PIE-UFES*, v. 1007, n. 6, 2007. http://agais.com/telomc/b01007_caracteristicas_leite.pdf.

VERNEQUE, R. da S. et al. Programa Nacional de Melhoramento do Gir Leiteiro-Sumário Brasileiro de Touros-Resultado do teste de progênie-5ª prova de pré-seleção de touros. Maio/2014. 2014.

VOGEL, Friedrich; MOTULSKY, Arno G. **Vogel and motulsky's human genetics: Problems and approaches**. Springer Science & Business Media, 2013.

Wang H, Zheng Y, Wang G, Li H. Identification of microRNA and bioinformatics target gene analysis in beef cattle intramuscular fat and subcutaneous fat. *Mol Biosyst.* 2013 Aug;9(8):2154-62. doi: 10.1039/c3mb70084d. Epub 2013 May 31. PMID: 23728155.

Wang M, Marín A. Characterization and prediction of alternative splice sites. *Gene.* 2006 Feb 1;366(2):219-27. doi: 10.1016/j.gene.2005.07.015. Epub 2005 Oct 13. PMID: 16226402.

Wang X, Zhong J, Gao Y, Ju Z, Huang J. A SNP in intron 8 of CD46 causes a novel transcript associated with mastitis in Holsteins. *BMC Genomics.* 2014 Jul 28;15(1):630. doi: 10.1186/1471-2164-15-630. PMID: 25070150; PMCID: PMC4124149.

WEBB, B.; SALI, A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics*, v.54, n. 1, p. 5.6. 1-5.6. 37, 2016. <https://doi.org/10.1002/cpbi.3>.

Wiggans GR, Cole JB, Hubbard SM, Sonstegard TS. Genomic Selection in Dairy Cattle: The USDA Experience. *Annu Rev Anim Biosci.* 2017 Feb 8;5:309-327. doi: 10.1146/annurev-animal-021815-111422. Epub 2016 Nov 16. Review. PubMed PMID: 27860491.

WIEDERSTEIN, M.; SIPPL, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res*, v. 35, n. suppl 2, p. W407-W410, 2007. <https://doi.org/10.1093/nar/gkm290>.

WOODFORD, K. B. Milk protein and human health: A1 versus A2 β -casein. General practitioners conference, Sydney, 2011. <https://www.researchgate.net › milk-gpcemay2011>.

YOUNGER, S. T.; COREY, D. R. Transcriptional gene silencing in mammalian cells by miRNA mimics that target gene promoters. *Nucleic Acids Res*, v. 39, n. 13, p. 5682– 5691, 2011. <https://doi.org/10.1093/nar/gkr155>.

ZHANG, Y.; CHENG, C.; SHI, W. et al. Transcriptional gene silencing of dopamine D3 receptor caused by let-7d mimics in immortalized renal proximal tubule cells of rats. *Gene*, v. 580, n. 2, p. 89-95, 2016. <https://doi.org/10.1016/j.gene.2015.12.071>

Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* 2005 Apr 22;33(7):2302-9. doi: 10.1093/nar/gki524. PMID: 15849316; PMCID: PMC1084323.

Zheng J, Ather JL, Sonstegard TS, Kerr DE. Characterization of the infection-responsive bovine lactoferrin promoter. *Gene.* 2005 Jun 20;353(1):107-17. doi: 10.1016/j.gene.2005.04.016. PMID: 15935571.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL. A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biol.* 2009;10(4):R42. doi: 10.1186/gb-2009-10-4-r42. Epub 2009 Apr 24. PMID: 19393038; PMCID: PMC2688933.

7 ANEXO 1 – Produção

Manuscrito em preparação

1. MATOSINHO, C. G. R.; Rosse, I. C.; FONSECA, P. A. S.; Assis, J.G.; OLIVEIRA, F. S.; ARAUJO, F.; SALIM, A.; LOPES, B. C.; ARBEX, W. A.; MACHADO, M. A.; PEIXOTO, M. G. C. D.; VERNEQUE, R. S.; MARTINS, M. F.; COIMBRA, R. S.; SILVA, M. V. G. B.; OLIVEIRA, G.; CARVALHO, M. R. S. . Identification and functional analysis *in silico* of variations in milk protein genes of Guzerá and Gir cattle.

Resumos publicados em anais de congressos

1. MATOSINHO, C. G. R.; SANTOS, F. C.; FONSECA, P. A. S.; Rossi, I.C.; CARVALHO, M. R. S.; PIRES, M. F. A.; BRUNELI, F. A. T.; SANTOS, G. G.; PEIXOTO, M. G. C. D. . Data-mining for selecting potential functional candidate genes for temperament of Guzerá cattle in a region of chromosome 1 identified by GWAS. In: ISCB-Latin American x-Meeting on Bioinformatics with BSB & SoiBio, 2014, Belo Horizonte. ISCB-LA / X-meeting / BSB / SoiBio, 2014.

2. MATOSINHO, C. G. R.; SANTOS, F. C.; PEIXOTO, M. G. C. D.; FONSECA, P. A. S.; PIRES, M. F. A.; Rossi, I.C.; BRUNELI, F. A. T.; SANTOS, G. G.; CARVALHO, M. R. S. . Selecting potential functional candidate genes for temperament of Guzerá cattle in a region of chromosome 1. In: First ISCB Latin American Student Council Symposium, 2014, Belo Horizonte. Program Booklet ISCB Student Council LA - SCS, 2014.

3. MATOSINHO, C. G. R.; FRANCO, T. R. L. R.; CAMPOLINA, S. S. . Chromosome number analysis of *Cyprinus Carpio* L. collected from fishfarming in Contagem. In: V Encontro de Genética de Minas Gerais, 2014, Belo Horizonte. Anais V Encontro de Genética de Minas Gerais, 2014.

Apresentação de trabalho

1. MATOSINHO, C. G. R.; Rosse, I. C.; FONSECA, P. A. S.; Assis, J.G.; OLIVEIRA, F. S.; ARAUJO, F.; SALIM, A.; LOPES, B. C.; ARBEX, W. A.; MACHADO, M. A.; PEIXOTO, M. G. C. D.; VERNEQUE, R. S.; MARTINS, M. F.; COIMBRA, R. S.; SILVA, M. V. G. B.; OLIVEIRA, G.; CARVALHO, M. R. S. . Identification and functional analysis *in silico* of polymorphisms in milk protein genes of Guzerá and Gir cattle. 2015.

Premiação

1. Menção honrosa pela participação no Prêmio Paulo Sodero Martins com o trabalho "Identification and functional analysis *in silico* of polymorphisms in milk protein genes of Guzerá and Gir cattle", 61º Congresso Brasileiro de Genética.

Organização de evento

1. CARVALHO, M. R. S.; GODARD, A. L. B.; MIYOSHI, A.; **MATOSINHO, C. G. R.**; SANTOS, F. C.; FONSECA, P. A. S.; Rossi, I.C.; MARTINS, A. P. V.; SILVA, B. C.; REPOLES, B. M.; LIMA, C. P.; SILVA, D. A. S. E.; COELHO, F. S.; SALAZAR, G. C.; MEDEIROS, J. D.; SANTANA, K. T. O.; CARVALHO, L. M.; ALVIM, L. B.; MIRANDA, M.; SILVA, M. C. S.; ALMEIDA, M. P.; ARAUJO, N. P.; TARTAGLIA, N. R.; MAGALHAES, T. A. V Encontro de Genética de Minas Gerais. 2014.

8 APÊNDICE

Figura Suplementar 1 - Conservação evolutiva avaliada usando alinhamento múltiplo

Chr6.87153410

[illegible]

Chr6.87374090

[illegible]

Chr5.31343847

<i>Ovis aries</i>	T C T T G C T A T T C T T T G G A A C T C T G C A T T C A G A T G C T T A T A T C T T T C C T T T T C T C C T T T G C T	
<i>Bubalus bubalis</i>	C C T T G C T A T T C T T T G G A A C T C T G A A T T C A G A T G C T T A T G T C T T T C C T T T T C T C C T T T G C T	
<i>Bos indicus</i>	T C T T G C T A T T C T T T G G A A C T C A G C A T T C A G A C G C T C A T A T C T T T C C T T T T C T C C T T T G C T	
<i>Bos taurus</i>	T C T T G C T A T T C T T T G G A A C T C A G C A T T C A G A T G C T C A T A T C T T T C C T T T T C T C C T T T G C T	1015
	* * * * *	*

Chr5.31346231

[illegible]

Chr5.31347426

[illegible]

Chr5.31347538

[illegible]

Resultado do alinhamento múltiplo de conservação evolutiva para as SNVs BTA6:87153410 no gene *CSN1S1*; BTA6:87374090 no gene *CSN3*; BTA5:31343847, BTA5:31346231, BTA5:31347426 e BTA5:31347538 no gene *LALBA*. O alelo divergente está presente no *Bos indicus*.

Figura Suplementar 2 - Conservação evolutiva avaliada usando alinhamento múltiplo

Chr5.31347611

Capra hircus	T	T	G	C	T	T	C	T	A	G	A	A	C	C	A	A	C	A	C	T	A	C	C	A	G	A	A	A	C	A	A	C	A	T	A	A	A	G	C	A	A	A	C	G	G	A	A	A	C	A	G	G	A				
Ovis aries	T	T	G	C	T	T	C	T	A	G	A	A	C	C	A	A	C	A	C	T	A	C	C	A	G	A	A	A	C	A	C	A	T	A	A	A	T	A	A	A	G	C	A	A	A	T	G	G	A	A	A	C	A	G	G	A	
Bubalus bubalis	T	T	G	C	T	T	C	T	A	G	A	A	C	C	A	A	C	A	C	T	A	C	C	A	G	A	A	A	C	A	C	A	T	A	A	A	T	A	A	A	G	C	A	A	A	T	G	G	A	A	A	C	A	G	G	A	
Bos indicus	T	T	G	C	T	T	C	T	A	G	A	A	C	C	A	A	C	A	C	T	A	C	C	A	G	A	A	A	C	A	C	A	T	A	A	A	T	A	A	A	G	C	A	A	A	T	G	G	A	A	A	C	A	G	G	A	
Bos taurus	T	T	G	C	T	T	C	T	A	G	A	A	C	C	A	A	C	A	C	T	A	C	C	A	G	A	A	A	C	A	C	A	T	A	A	A	T	A	A	A	G	C	A	A	A	T	G	G	A	A	A	C	A	G	G	A	1679
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*					

Chr5.31348386

<i>Capra hircus</i>	A	T	T	T	C	A	T	A	C	C	A	G	T	G	G	T	T	A	T	G	A	C	A	C	A	A	G	C	C	A	T	A	G	T	A	C	A	A	A	C	A	A	T	G	A	C	A	G	C	A	C	A	G	A	A	T	A	
<i>Ovis aries</i>	A	T	T	T	C	A	T	A	C	C	A	G	T	G	G	T	T	A	T	G	A	C	A	C	A	A	G	C	C	A	T	A	G	T	A	C	A	A	A	C	A	A	T	G	A	C	A	G	C	A	C	A	G	A	A	T	A	
<i>Bubalus bubalis</i>	A	T	T	T	C	A	T	A	C	C	A	G	T	G	G	T	T	A	T	G	A	C	A	C	A	A	G	C	C	A	T	A	G	T	A	C	A	A	A	C	A	A	T	G	A	C	A	G	C	A	C	A	G	A	A	T	A	
<i>Bos indicus</i>	G	T	T	T	C	A	T	A	C	C	A	G	T	G	G	T	T	A	T	G	A	C	A	C	A	A	G	C	C	G	T	A	G	T	A	C	A	A	A	C	A	A	T	G	A	C	A	G	C	A	C	A	G	A	A	T	A	
<i>Bos taurus</i>	G	T	T	T	C	A	T	A	C	C	A	G	T	G	G	T	T	A	T	G	A	C	A	C	A	A	G	C	C	A	T	A	G	T	A	C	A	A	A	C	A	A	T	G	A	C	A	G	C	A	C	A	G	A	A	T	A	554
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*				

Chr5.31349118

<i>Capra hircus</i>	T	C	T	T	T	C	T	T	C	T	C	A	G	T	C	T	T	T	C	G	C	C	C	A	G	C	A	C	T	A	T	A	C	T	C	G	T	-	-	-	T	C	T	A	T	T	T	C	C	T	G	G	T	C	T	
<i>Ovis aries</i>	T	C	T	T	T	C	T	T	C	T	C	A	G	T	C	T	T	T	C	G	C	C	C	A	G	C	A	C	T	A	T	A	C	T	C	G	T	T	C	T	C	T	A	T	T	T	C	C	T	G	G	T	C	T		
<i>Bubalus bubalis</i>	T	C	T	T	T	C	T	T	C	T	C	A	G	T	C	T	T	T	C	G	T	C	C	C	A	G	C	A	C	T	A	T	A	C	T	C	T	T	C	T	C	T	A	T	T	T	C	C	T	G	G	T	C	T		
<i>Bos indicus</i>	T	C	T	T	T	C	T	T	C	T	C	A	G	T	C	T	T	T	C	G	T	C	C	C	A	G	C	A	C	T	G	T	A	C	T	C	T	T	C	T	C	T	A	T	T	T	C	C	T	G	G	T	C	T		
<i>Bos taurus</i>	T	C	T	T	T	C	T	T	C	T	C	A	G	T	C	T	T	T	C	G	T	C	C	C	A	G	C	A	C	T	A	T	A	C	T	C	T	T	C	T	C	T	A	T	T	T	C	C	T	G	G	T	C	T		
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	3186				

Chr5.31349177

Capra hircus	T	T	T	A	A	G	C	T	A	G	A	A	T	G	T	C	A	T	C	T	T	A	A	A	A	C	C	A	A	A	A	G	T	C	A	T	C	A	A	G	C	A	G	A	C	T	C	A	G	G	C	T	T	C	C	A	A	T	T	T		
Ovis aries	T	T	T	A	A	G	C	T	A	G	A	A	T	G	T	C	A	T	C	T	T	A	A	A	A	T	C	A	A	A	A	A	T	C	A	T	C	A	A	G	C	A	G	A	C	T	C	A	G	G	T	T	T	C	C	A	A	T	T	C		
Bubalus bubalis	T	T	T	A	A	G	C	T	A	G	A	A	T	G	T	A	A	T	C	T	T	A	A	A	A	C	A	A	A	A	A	A	T	C	A	T	C	A	A	G	C	A	G	A	C	T	C	C	G	G	T	T	T	C	C	A	A	T	T	T		
Bos indicus	T	T	T	A	A	G	C	T	A	G	A	A	T	G	T	A	A	T	C	T	T	A	A	A	A	C	A	A	A	A	A	A	C	T	C	A	T	C	A	A	A	C	A	G	A	C	T	C	C	G	G	T	T	T	C	C	A	A	T	T	T	
Bos taurus	T	T	T	A	A	G	C	T	A	G	A	A	T	G	T	A	A	T	C	T	T	A	A	A	A	C	A	A	A	A	A	A	A	T	C	A	T	C	A	A	A	C	A	G	A	C	T	C	C	G	G	T	T	T	C	C	A	A	T	T	T	3245
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*		

Chr5.31350957

Capra hircus	T	T	T	C	C	C	A	A	G	T	C	T	A	C	A	G	G	G	T	T	T	A	G	A	A	T	A	A	T	C	G	C	T	G	T	T	C	A	C	A	G	T	G	G	T	A	T	G	G	G	A	G	G	C	A				
Ovis aries	T	T	T	C	C	C	C	A	A	G	T	C	T	A	C	A	A	G	G	G	T	T	T	A	G	A	A	T	A	A	T	C	G	C	T	G	T	T	C	A	C	A	G	T	G	G	T	A	T	G	T	G	G	A	G	G	C	A	
Bubalus bubalis	T	T	T	C	C	C	C	A	A	G	T	C	T	A	C	A	A	G	G	T	T	T	A	G	A	A	T	A	A	T	T	G	C	T	A	T	T	C	A	C	A	G	T	T	C	T	A	T	G	T	G	G	A	G	G	C	A		
Bos indicus	T	T	T	C	C	C	C	A	A	G	T	C	T	A	C	A	A	G	G	T	T	T	A	G	A	A	T	A	A	T	T	G	C	T	G	T	T	C	A	C	A	G	T	T	G	T	A	T	A	T	G	G	A	G	G	T	A		
Bos taurus	T	T	T	C	C	C	C	A	A	G	T	C	T	A	C	A	A	G	G	T	T	T	A	G	A	A	T	A	A	T	T	C	G	C	T	G	T	T	C	A	C	A	G	T	T	G	T	A	T	A	T	G	G	A	G	G	T	A	5025
	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*					

Resultado do alinhamento múltiplo de conservação evolutiva para as SNVs BTA5:31347611, BTA5:31348386, BTA5:31349118, BTA5:31349177 e BTA5:31350957 no gene *LALBA*. O alelo divergente está presente no *Bos indicus*.

