

UNIVERSIDADE FEDERAL DE PELOTAS  
Programa de Pós-Graduação em Zootecnia



Tese

**Prospecção de marcadores microssatélites potenciais  
para o melhoramento genético de bovinos.**

**Marco André Paldês da Costa**

Pelotas, 2012

MARCO ANDRÉ PALDÊS DA COSTA

PROSPECÇÃO DE MARCADORES MICROSSATÉLITES POTENCIAIS PARA O  
MELHORAMENTO GENÉTICO DE BOVINOS

Tese apresentada ao Programa de Pós-Graduação em Zootecnia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciências (área do conhecimento: Melhoramento Genético Animal).

Orientador: Prof. Dr. Heden Luiz Marques Moreira  
Co-orientadores: Prof. Dr. Vitor Hugo Borba Manzke  
Prof. PhD. Frank Siewerdt

Pelotas, 2012

Dados de catalogação na fonte:  
Ubirajara Buddin Cruz – CRB-10/901  
Biblioteca de Ciência & Tecnologia - UFPel

C837p

Costa, Marco André Paldês da

Prospecção de marcadores microssatélites potenciais para o melhoramento genético de bovinos / Marco André Paldês da Costa. – 57f. ; gráf. – Tese (Doutorado). Programa de Pós-Graduação em Zootecnia. Área de concentração: Melhoramento genético animal. Universidade Federal de Pelotas. Faculdade de Agronomia Eliseu Maciel. Pelotas, 2012. – Orientador Heden Luiz Marques Moreira ; co-orientador Vitor Hugo Borba Manzke; Frank Siewerdt.

1.Zootecnia. 2.SSR. 3.Predição in silico. 4.Anotação genômica. 5.Seleção assistida. I.Moreira, Heden Luiz Marques.. II.Manzke, Vitor Hugo Borba. III.Siewerdt, Frank. IV.Título.

CDD: 636.0821

Banca examinadora:

Prof. Dr. Heden Luiz Marques Moreira – UFPel (Presidente)

Profa. Dra. Beatriz Helena Gomes Rocha - UFPel

Prof. Dr. Luciano Carlos da Maia - UFPel

Prof. Dr. Nelson José Laurino Dionello - UFPel

Prof. Dr. Valmor Antonio Konflanz - UNOESC

Profa. Dra. Cecilia Irene Pérez Calabuig – UFRGS (Suplente)

## Agradecimentos

Em breves linhas aqui reconheço aqueles que grato me ponho, não somente por apoio ou compreensão, mas sim pelas “pedras” que alicerçam a base da minha formação:

- Aos amigos docentes e discentes dos Programas de Pós-Graduação da Universidade Federal de Pelotas, em especial àqueles ligados ao Programa de Pós-Graduação em Zootecnia, por todo auxílio e conhecimentos transmitidos;

- Ao meu orientador, Prof. Dr. Heden Luiz Marques Moreira, e co-orientadores, Prof. Dr. Vitor Hugo Borba Manzke e Prof. PhD. Frank Siewerdt, pelas importantes colaborações no planejamento e condução deste trabalho de pesquisa;

- Aos componentes da banca examinadora, Profa. Dra. Beatriz Helena Gomes Rocha, Profa. Dra. Cecilia Irene Pérez Calabuig, Prof. Dr. Nelson José Laurino Dionello e Prof. Dr. Valmor Antonio Konflanz, pela contribuição para o desfecho deste manuscrito;

- Aos professores Dr. Luciano Carlos da Maia e PhD. Antônio Costa de Oliveira, pela inestimável colaboração para a execução e aprimoramento do trabalho desenvolvido;

- A equipe do Laboratório de Engenharia Genética Animal/UFPel, representados pela figura do amigo Dr. Diones Bender Almeida, valorosos colegas com quem compartilhei anos de convivência e trabalho;

- Aos meus familiares, esteios que jamais esmorecem;

- A minha namorada, Darliane da Rosa Luz, fiel parceira e meu incentivo à perseverança;

- A Deus, por me conceder o luxo de matear sempre cercado de amigos!...

Reconheço também o suporte da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão de minha bolsa de estudos e auxílio financeiro.

A todos, por tudo, em fim: *mil gracias!*

*Dedico esse trabalho a minha namorada, meus familiares, amigos e professores,  
que juntos proporcionaram esse momento de realização.*

*“Bendito aquele que estuda  
porque estudar é importante,  
embora o ignorante  
tem sempre um santo que ajuda,  
às vezes a sorte muda,  
quando existe um santo forte,  
cada qual procura um norte,  
por isso não encabulo  
- que a tava que bota culo  
é a mesma que bota sorte!...”*

*Sem Diploma (Jayme Caetano Braun)*

## Resumo

COSTA, Marco André Paldês da. **Prospecção de marcadores microssatélites potenciais para o melhoramento genético de bovinos**. 2012. 57f. Tese (Doutorado) - Programa de Pós-Graduação em Zootecnia. Universidade Federal de Pelotas, Pelotas.

Bovinos (*Bos taurus*) representam um importante recurso filogenético para compreensão de aspectos evolutivos e características complexas. Sobre seu genoma, repetições microssatélites constituíram a base para genotipagem em diferentes estudos, como o mapeamento genético. Assim, esse trabalho teve por objetivo descrever a distribuição destas sequências ao longo do genoma bovino taurino, atribuindo a elas anotação estrutural e funcional, além de disponibilizar um recurso para que tal informação possa auxiliar investigações futuras sobre tais regiões. Com uma cobertura genômica estimada em 0,8%, cerca de 1.3 milhões de *loci* foram identificados, dos quais 90,5% classificados como perfeitos. Monômeros ocorreram em maior número e frequência, seguidos por tetrâmeros, dímeros, trímeros, pentâmeros e hexâmeros. Contudo dímeros e pentâmeros contribuíram em densidade mais que tetrâmeros. Independente do tipo de repetição, sempre a maior proporção dos trechos repetitivos distribuíram-se sobre regiões IG, Introns e GPG. Apesar disso, densidade e frequência em 5'UTR e PR superaram IG. Através desses descritores semelhanças puderam ser evidenciadas entre IG, Introns e 3'UTR, como também para PR e 5'UTR. Anotações para processo biológico, função molecular e componente celular estavam super-representadas em genes abrigando sequências repetitivas. Utilizando referências de mapeamento, como anotação gênica, STS, SNPs e outros elementos repetitivos já descritos, a plataforma *on-line Pampeano Animal Genetic Resources* fornece um importante recurso para que a informação reunida possa auxiliar o planejamento de novas abordagens, em especial para obtenção de marcadores moleculares destinados a seleção de características de interesse ao melhoramento genético da espécie.

Palavras-chave: ssr, predição *in silico*, anotação genômica, seleção assistida.



## Abstract

COSTA, Marco André Paldês da. **Prospecção de marcadores microssatélites potenciais para o melhoramento genético de bovinos**. 2012. 57f. Tese (Doutorado) - Programa de Pós-Graduação em Zootecnia. Universidade Federal de Pelotas, Pelotas.

Cattle (*Bos taurus*) represent an important phylogenetic resource for understanding of evolutionary aspects and complex traits. On its genome, microsatellites repeats constituted the basis for genotyping in different studies as the genetic mapping. So, this study has by aim to describe the distribution of these sequences along of the taurine bovine genome, ascribing to them structural and functional annotation, besides supply a resource for that such information may assist future investigations on these regions. With a genomic coverage estimated at 0.8%, around 1.3 million loci were identified, of which 90.5% classified as perfect. Monomers occurred in greater number and frequency, followed by tetramers, dimers, trimers, pentamers and hexamers. However dimers and pentamers contributed in density more than tetramers. Regardless of the type of repetition, always the largest proportion of repetitive stretches distributed on IG, Introns and GPG regions. Nevertheless, density and frequency on PR and 5'UTR overcame IG. Through of these descriptors similarities could be observed between IG, Introns and 3'UTR, as too for PR and 5'UTR. Annotation for biological process, molecular function and cellular component were super-represented in genes harboring repetitive sequences. Using mapping references, as gene annotation, STS, SNPs and other repetitive elements already described, the Pampeano Animal Genetic Resources online platform provides an important resource for that the information gathered can to assist the planning of new approaches, in especial to obtaining molecular markers destined to selection of traits of interest to the genetic improvement of the species.

Keywords: *ssr*, *in silico* prediction, genomic annotation, assisted selection.

## Lista de Figuras

Figura 1: Ocorrências de <i>loci</i> microssatélites ao longo do genoma bovino taurino. ...	24
Figura 2: Comparações das ordenações médias de ocorrência, densidade e frequência de repetições microssatélites sobre o genoma bovino taurino pelo Teste da Soma de Ordens de Friedman. ....	26
Figura 3: Ocorrência de repetições microssatélites entre regiões do genoma bovino taurino. ....	27
Figura 4: Frequência de repetições microssatélites entre regiões do genoma bovino taurino. ....	28
Figura 5: Densidade de repetições microssatélites entre regiões do genoma bovino taurino. ....	29
Figura 6: Ocorrência dos motivos de repetições microssatélites ao longo do genoma bovino taurino. ....	30
Figura 7: Distribuição da anotação funcional para função molecular, processo biológico e componente celular em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites. ....	41
Figura 8: Distribuição ao segundo nível da anotação funcional para função molecular em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites. ....	42
Figura 9: Distribuição ao segundo nível da anotação funcional para componente celular em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites. ....	43
Figura 10: Distribuição ao segundo nível da anotação funcional para processo biológico em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites. ....	44

## Lista de Tabelas

Tabela 1: Ocorrência de <i>loci</i> microssatélites sobre o genoma bovino taurino.....	23
Tabela 2: Distribuição de repetições microssatélites sobre o genoma bovino taurino. .....	25
Tabela 3: Ontologia Gênica recuperada para genes transcritos mapeados sobre o genoma bovino taurino, utilizando dados dos repositórios B2G-FAR, Gene/NCBI e GOC. ....	40

## Sumário

Resumo .....	6
Abstract .....	7
Lista de Figuras .....	8
Lista de Tabelas .....	9
Sumário .....	10
1 Introdução .....	11
2 Distribuição de repetições microssatélite sobre o genoma bovino taurino. ....	16
2.1 Material e Métodos .....	16
2.2 Resultados .....	18
2.3 Discussão .....	31
2.4 Conclusão .....	35
3 Anotação funcional de repetições microssatélites identificadas sobre o genoma bovino taurino .....	36
3.1 Material e Métodos .....	36
3.2 Resultados .....	37
3.3 Discussão .....	45
3.4 Conclusão .....	47
4 Conclusões .....	48
Referências .....	49

## 1 Introdução

Para a economia mundial, os bovinos representam uma espécie de animal doméstico com grande importância para exploração pecuária (LEWIN, 2009; TELLAM et al., 2009). Desde a última década, o Brasil se destaca no mercado internacional pelo maior rebanho comercial, somando 205,3 milhões de cabeças em 2009 (POLL et al., 2011a). Ao mesmo tempo, o país mantém a posição de principal exportador, seguido pela Austrália e Estados Unidos (JOSÉ, 2012). Aliado a uma série de vantagens geográficas e climáticas, a produção exclusivamente a pasto e resíduos agrícolas proporciona um produto extremamente valorizado (POLL et al., 2011a). Após anos de baixa rentabilidade, a cadeia produtiva nacional volta a se recuperar fortalecida pelo aumento da demanda dos mercados interno e externo (POLL et al., 2011b), e pela expectativa de um significativo aumento das exportações para 2012 (JOSÉ, 2012).

No entanto, a importância econômica não foi o principal motivo que levou a espécie bovina a integrar um grupo seletivo de animais com o genoma completamente sequenciado. Dentro da classe *Mammalia*, eles pertencem a ordem *Cetartiodactyla*, uma ordem filogeneticamente distante dos primatas, tendo divergido de um ancestral comum a aproximadamente 85 milhões de anos (KUMAR; HEDGES, 1998). Análises comparativas entre espécies relacionadas permitem elucidar funções e aspectos evolutivos dos genomas (LARKIN et al., 2003). Assim, os bovinos representam um importante recurso para compreensão de mecanismos envolvidos na regulação e controle de características complexas (ELSIK et al., 2009; GIBBS et al., 2009; LEWIN, 2009; TELLAM et al., 2009), em especial para outras espécies do gênero *Bos* com menor volume de informação.

O seqüenciamento completo do genoma humano proporcionou um importante avanço nas tecnologias de seqüenciamento e a redução de seus custos, viabilizando muitos projetos em andamento (MCCARTHY et al., 2006). Após os taurinos (ELSIK et al., 2009), uma nova perspectiva surge pelo anúncio de término do seqüenciamento do primeiro genoma zebuino (*Bos indicus*), provenientes de um animal da raça Nelore, e de outros quatro (Gir Leiteiro, Guzará, Red Sindi e Girolando) que estão em andamento (POLL et al., 2011c). Esses recursos possibilitarão diversas investigações comparativas com extrema acurácia.

Por décadas, o melhoramento genético para espécies de interesse zootécnico proporcionou ganhos representativos a produção pecuária utilizando

informações sobre fenótipo e *pedigree* para predizer o valor genético dos animais (GODDARD; HAYES, 2007) . Surgem então os marcadores moleculares como uma nova ferramenta para maximizar seus avanços, em especial para seleção de características de difícil mensuração (GODDARD; HAYES, 2009). Essa metodologia desencadeou inúmeras investigações, culminando recentemente com desenvolvimento de painéis de alta cobertura genômica (MATUKUMALLI et al., 2009). No entanto, novas gerações dessa tecnologia trazem a expectativa da possibilidade de sequenciar genomas inteiros a custos e tempo extremamente reduzidos (EID et al., 2009).

Ao longo do curso histórico da pesquisa genômica, várias sequências de nucleotídeos serviram de referência para descrever variações no código genético. Um grupo particular permaneceu em foco pela ampla distribuição tanto em regiões codificantes como não codificantes do genoma de eucariotos e procariotos, perfil multialélico e herança codominante (LI et al., 2002; SUBRAMANIAN; MISHRA; SINGH, 2003; AISHWARYA; GROVER; SHARMA, 2007; FAN; CHU, 2007; MAIA et al., 2009). Com diferentes nomenclaturas, microssatélites (ou repetições de sequência simples/SSR ou repetições curtas em tandem/STR) são combinações de um a seis nucleotídeos repetidos sequencialmente (TAUTZ, 1993) onde eventos como escorregões da DNA Polimerase, formação de estruturas secundárias e recombinação geram taxas de mutação superiores a observadas sobre sítios pontuais (LI et al., 2002; KASHI; KING, 2006; FAN; CHU, 2007; CRUZ; ROUX; ROBINSON-RECHAVI, 2009). Essas sequências constituíram a base principal para genotipagem em diferentes abordagens (KRUGLYAK et al., 1998; TEMNYKH et al., 2001; JU et al., 2005; AISHWARYA; GROVER; SHARMA, 2007; MAIA et al., 2008; YAN et al., 2008; ZHANG et al., 2010), fornecendo uma ferramenta eficiente para ligar variação fenotípica e genotípica (TEMNYKH et al., 2001; KASHI; KING, 2006; MAIA et al., 2009; VICTORIA; MAIA; OLIVEIRA, 2011).

Muito se discute sobre importância estrutural e funcional de regiões repetitivas sobre o genoma de vários organismos (GLENN et al., 1996; SUBRAMANIAN; MISHRA; SINGH, 2003; ALBA; GUIGO, 2004; AISHWARYA; GROVER; SHARMA, 2007; BRANDSTROM; ELLEGREN, 2008; MAIA et al., 2008; MAIA et al., 2009; ADELSON et al., 2010) inclusive para bovinos (YAN et al., 2008; ADELSON; RAISON; EDGAR, 2009). Conhecer a distribuição de repetições microssatélite sobre as diferentes regiões do genoma pode fornecer indícios de sua

influencia na organização e regulação gênica (SUBRAMANIAN; MISHRA; SINGH, 2003; AISHWARYA; GROVER; SHARMA, 2007). Para cada região existe ao menos uma possível contribuição dessas sequências exaustivamente discutidas (LI et al., 2002; KASHI; KING, 2006; FAN; CHU, 2007). No entanto, é esperado que a pressão de seleção sobre regiões expressas e principalmente traduzidas ocorra em maior magnitude que em outras. Isso confere com o nível de conservação já descrito nessas regiões (GLENN et al., 1996; SUBRAMANIAN; MISHRA; SINGH, 2003; ALBA; GUIGO, 2004; SERAPION et al., 2004; JU et al., 2005; ZHAN et al., 2009) ou para a presença de pseudo-genes (SVENSSON; ARVESTAD; LAGERGREN, 2006).

Atualmente, a gama de bancos de dados genômicos existentes torna público e acessível todo o conhecimento gerado sobre qualquer espécie. No entanto, cada repositório possui particularidades que dificilmente atende todas as demandas. Com as iniciativas de sequenciamento completo de genomas, um grande volume de dados tem sido acumulado (SUBRAMANIAN; MISHRA; SINGH, 2003; JU et al., 2005; KIM et al., 2006; AISHWARYA; GROVER; SHARMA, 2007; HUGHES et al., 2008; OZDEMIR OZGENTURK et al., 2010; ZHANG et al., 2010), permitindo análises amplas (SUBRAMANIAN; MISHRA; SINGH, 2003; KIM et al., 2006; BRANDSTROM; ELLEGREN, 2008), e fornecendo um excelente recurso para o desenvolvimento de marcadores moleculares (TEMNYKH et al., 2001; SERAPION et al., 2004; JU et al., 2005; AISHWARYA; GROVER; SHARMA, 2007; MAIA et al., 2008; YAN et al., 2008; MAIA et al., 2009; ZHAN et al., 2009; OZDEMIR OZGENTURK et al., 2010; ZHANG et al., 2010). Nesse momento, garimpar e combinar a informação disponível em busca de um determinado propósito é um modo rápido e eficiente de maximizar resultados e atingir os objetivos de pesquisa (CONESA et al., 2005; CONESA; GOTZ, 2008; GOTZ et al., 2008; GOTZ et al., 2011; LECERF et al., 2011).

O *Gene Ontology Consortium* (GOC) é um importante exemplo de esforço cooperativo entre grupos de pesquisa que surgiu como resultado do acúmulo de informação em bancos de dados genômicos. Lacunas terminológicas, sintáticas e semânticas criaram a necessidade do uso de ontologias para que diferentes pesquisadores pudessem ter acesso a recursos inter-relacionados de modo consistente e não ambíguo (ZHOU; SU, 2007; HUGHES et al., 2008). Assim, a ontologia gênica apóia-se na utilização de um vocabulário com termos controlados para padronizar a representação de genes e seus produtos, independente de

espécie ou repositório (BLUTHGEN et al., 2005). Os termos são manualmente anotados por curadores treinados, extraindo informação de dados experimentais publicados ou inferindo anotação por homologia com produtos gênicos relacionados para os quais dados experimentais estão disponíveis (GGR-GOC, 2009). Métodos automatizados baseados em similaridade ou na composição de domínios também são utilizados para maximizar a velocidade de anotação (MCCARTHY et al., 2006; CONESA; GOTZ, 2008). Porém, para cada fonte um código de evidência fornece o nível de confiança para determinada anotação (CONESA; GOTZ, 2008; GOTZ et al., 2011). Atualmente existem mais de 26 mil termos arranjados em três ramos: função molecular (atividade elementar do produto gênico a nível molecular), processo biológico (conjunto de eventos reunindo uma ou mais funções moleculares) e componente celular (parte de uma célula ou seu ambiente extracelular). Neles identificadores são interligados por relações bem definidas, particularmente de subsunção, participativas e de regulação biológica (GGR-GOC, 2009). Desse modo, conjuntos de organismos modelo e banco de proteínas são utilizados para transferir informação funcional para outras espécies com menor volume de investigações (CONESA et al., 2005; MCCARTHY et al., 2006; CONESA; GOTZ, 2008; GOTZ et al., 2008; GGR-GOC, 2009; GOTZ et al., 2011).

A diversidade racial é outra peculiaridade do rebanho bovino brasileiro. Nas zonas tropicais predominam os zebuínos (*Bos indicus*), representando cerca de 80% do contingente nacional. À raça Nelore (e cruzas) é atribuída 90% desta parcela. Já nos estados ao sul, a prevalência é dos taurinos (*Bos taurus*), principalmente Hereford e Angus (POLL et al., 2011a). Embora tenham divergido de um ancestral comum por eventos independentes de domesticação (ACHILLI et al., 2009), suas características particulares (HANSEN, 2004; CHAN; NAGARAJ; REVERTER, 2010) e capacidade de produção de híbridos férteis (HIENDLEDER; LEWALSKI; JANKE, 2008), fornecem um importante recurso para programas de cruzamento e seleção, servindo de base para formação de várias raças modernas (BOLORMAA et al., 2011). Essa valiosa fonte de variabilidade genética reflete uma inestimável ferramenta a disposição para o melhoramento genético do rebanho brasileiro.

Assim sendo, o objetivo deste trabalho foi identificar e descrever a distribuição de repetições microssatélites ao longo do genoma bovino taurino, atribuindo anotação estrutural e funcional, bem como disponibilizar um recurso para



que a informação reunida possa auxiliar estudos futuros sobre tais regiões ou a seleção de características de interesse ao melhoramento genético da espécie.

## 2 Distribuição de repetições microssatélite sobre o genoma bovino taurino.

Nessa primeira etapa, o objetivo foi identificar e descrever a distribuição de repetições microssatélites ao longo do genoma bovino taurino.

### 2.1 Material e Métodos

#### *Informação genômica*

A construção do genoma bovino taurino Btau5.2 (*Bos taurus*) consta de uma montagem de referência (Btau4.2), disponibilizada pelo *Bovine Genome Sequencing Consortium/BGSC*, e outra alternativa (UMD3.1), fornecida pelo *Center for Bioinformatics and Computational Biology* (Universidade de Maryland, EUA). Estão inclusas as sequências para os vinte e nove cromossomos autossômicos e um sexual (BtaX). Somente uma delas contempla o DNA mitocondrial (mtBta, em Btau4.2), enquanto nada está disponível sobre o cromossomo sexual Y. Existe também um grande volume de segmentos ainda não alocados (*contigs* não mapeados). Todos esses dados, além das informações de mapeamento sobre a anotação gênica, polimorfismos de nucleotídeo único (SNP), sítios de sequência marcada (STS) e outros elementos repetitivos já descritos, foram obtidos através do repositório para projetos de anotação e montagem de genomas do *National Center for Biotechnology Information* (<ftp://ftp.ncbi.nih.gov/genomes/>), em agosto de 2011.

#### *Identificação das repetições microssatélites*

Arquivos individuais contendo os cromossomos fracionados foram triados pelo programa SSRLocator (MAIA et al., 2008) para identificar microssatélites com mínimo de doze repetições para monômeros, seis para dímeros, quatro para trímeros, e três para tetrâmeros, pentâmeros e hexâmeros. Ao término de cada entrada, o aplicativo gera arquivos descritivos em formato de texto puro. Um *pipeline* desenvolvido com duas linguagens de programação (PASCAL e PERL) e integrado a um banco de dados (MySQL) permitiu importar, classificar, anotar e descrever todos os trechos repetitivos.

#### *Classificações e agrupamentos*

Dois critérios determinaram a classificação dos *loci*: intervalo e integridade das repetições. *Loci* imperfeitos apresentavam segmentos repetitivos interrompidos

por até seis pares de base. Dois microssatélites adjacentes caracterizavam *loci* ligados. Com intervalos maiores ou iguais a cem pares de base os *loci* eram intitulados perfeitos, do contrário como compostos e identificados conforme o número de repetições que abrigavam. No intuito de evitar fragmentos demasiadamente longos por aglomerações de microssatélites, *loci* compostos com mais de 500pb foram isolados ao encontrar uma extensão mínima de 40pb a partir da última sequência repetitiva, possibilitando assim um intervalo adequado para desenho de *primers* (JU et al., 2005).

Sete regiões foram consideradas ao descrever a localização dos microssatélites, a saber: intergênica (IG), promotor (PR; 500 pares de base antes do sítio para início de transcrição), 5' não traduzida (5'UTR), codificadora (CDS), intragênica (Intron), 3' não traduzida (3'UTR) e genes sem regiões descritas ou pseudo-genes (GPG). Devido a presença de limites adjacentes, por exemplo, entre CDS e Introns, para que um microssatélite fosse atribuído a determinado região mais que 50% de sua sequência deveria estar presente. Repetições anotadas em trechos intergênicos receberam as referências de mapeamento mais próximas *up* e *dowstream*.

Como a nível de *locus* marcador um motivo de repetição independe de fita ou orientação, as ocorrências ao longo do genoma referem-se ao somatório de todas as suas possíveis combinações, por exemplo, AGC inclui AGC/GCT/CGA/TCG. No entanto, quando regiões estruturais passam a ser foco da discussão, fita e orientação tornam-se pertinentes. Desse modo, AGC passa a representar apenas si próprio.

#### *Medidas descritivas e análise estatística*

Três medidas descreveram a distribuição das repetições. A Ocorrência informa o somatório dos microssatélites identificados. A Densidade demonstra a contribuição das sequências repetitivas para um intervalo de um milhão de pares de base, expressa em pb/Mb, enquanto a Frequência estima a ocorrência sobre a mesma distância, expressa em SSR/Mb. Utilizando o pacote estatístico R (2011), testes não-paramétricos complementaram as inferências, evidenciando associações entre os descritores e o número de genes ou tamanho dos cromossomos (Teste para Correlações de Ordem de Spearman), e descrevendo diferenças entre os tipos de *loci* ou repetições (Teste da Soma de Ordens de Friedman).

## 2.2 Resultados

Todos os descritores alcançaram estimativas extremamente similares entre as duas montagens do genoma bovino. Por esse motivo, serão discutidos apenas os resultados obtidos a partir da sequência de referência Btau4.2, sem que haja perda de informação sobre a construção Btau5.2. Quando necessário existem referências individuais. Detalhes a respeito de ambas as montagens podem ser encontrados no material suplementar (S1) disponível através da plataforma *Pampeano Animal Genetic Resources* (<http://www.agrpampeano.org/>).

### *Distribuição e cobertura*

Cerca de 1,3 milhões de *loci* microssatélites foram identificados sobre o genoma bovino (Tab. 1), sendo 90,5% perfeitos, 0,3% imperfeitos, 1,2% ligados e 8,0% compostos. A grande maioria deste último (81,2%) abrigava duas sequências repetitivas. Tais proporções permaneceram constantes entre os cromossomos (Fig. 1), tendo as diferenças evidenciadas pelas ordenações médias (Tab. 1). Além disso, uma forte associação entre a extensão dos cromossomos e o somatório das observações pode ser descrita ( $r=0,99$ ;  $p<0,001$ ). Assim, cromossomos maiores (por exemplo, Bta1 e Bta2) também continham maior número de *loci* que os menores (Bta25 e Bta28). É importante ressaltar a discrepância entre as duas montagens para o número de *loci* identificados sobre segmentos não alocados (Btau4.2: 130.301; UMD3.1: 2.982). Como apenas três repetições estavam presentes sobre a única sequência do DNA mitocondrial –  $(C)_{12}$ ,  $(CCTC)_3$  e  $(TAC)_4$ , todos *loci* perfeitos – essas não contribuíram para a discussão.

Sequências microssatélites representaram 0,8% do genoma bovino, onde monômeros ocorreram em maior número e frequência (Tab. 2), seguidos por tetrâmeros, dímeros, trímeros, pentâmeros e hexâmeros. Contudo dímeros e pentâmeros contribuíram em densidade mais que tetrâmeros. A não ser pelas densidades de tetrâmeros e pentâmeros, as demais ordenações médias distinguiram os tipos de repetição (Fig. 2). Identificou-se também uma importante correlação entre a ocorrência de microssatélites e o número de genes por cromossomo ( $r=0,54$ ;  $p<0,01$ ). Do mesmo modo, a extensão dos cromossomos foi associada a ocorrência das repetições ( $r=0,99$ ,  $p<0,001$ ). No entanto, densidade e frequência demonstraram correlações baixas ( $r=0,30$  e  $r=0,23$ , respectivamente) e não significativas ( $p>0,05$ ).

### *Importância regional*

Independente do tipo de repetição (Tab. 2), sempre a maior proporção dos trechos repetitivos (98,7%) distribui-se sobre IG (68,8%), Introns (28,2%) e GPG (1,7%). O restante é atribuído a PR, CDS, 3'UTR, e 5'UTR. No entanto, densidade e frequência em 5'UTR e PR superaram IG (Tab. 2). Apenas a frequência em 3'UTR ultrapassou este último, mas superior a Introns também em densidade. Exceto por 3'UTR e CDS, todas as regiões diferiram quanto as ordenações médias para ocorrência de microssatélites (Fig. 2). Densidade em 5'UTR e PR foram equiparáveis (Fig. 2), ocorrendo o mesmo para frequência entre 3'UTR e IG.

Monômeros, tetrâmeros e dímeros atingiram ocorrência e frequência maiores em Introns, 3'UTR e IG (Fig. 3 e 4), alternando apenas a importância de trímeros por pentâmeros para IG. No entanto, dímeros superaram tetrâmeros em densidade (Fig. 5), sendo semelhantes as ordenações médias em 3'UTR, assim como entre monômeros e dímeros em IG (Fig. 2). Para GPG, ao contrário de Introns e 3'UTR, trímeros e pentâmeros tiveram maior ocorrência e frequência que dímeros (Fig. 3 e 4), e ao contrário de IG, trímeros contribuíram mais em densidade que tetrâmeros (Fig. 5). Pelas ordenações médias de ocorrência e frequência, monômeros se equipararam aos tetrâmeros, como também trímeros para pentâmeros e dímeros (Fig. 2). Já dímeros e pentâmeros ou trímeros e tetrâmeros foram semelhantes em densidade.

Somente duas alternâncias diferenciaram PR e 5'UTR: monômeros contribuíram mais que tetrâmeros na primeira, enquanto hexâmeros mais que pentâmeros na segunda (Fig. 5). Pelas ordenações médias trímeros e monômeros ou tetrâmeros e dímeros foram comparáveis por densidade em PR (Fig. 2). Além disso, dímeros e pentâmeros possuíam ocorrência e frequência similares, assim como tetrâmeros e monômeros com os tipos com índices maiores aos seus. O mesmo ocorreu em 5'UTR para dímeros, pentâmeros e hexâmeros, a não ser pela densidade semelhante entre eles. Já em CDS, trímeros, tetrâmeros e hexâmeros prevaleceram em ocorrência, frequência e densidade, seguidos por pentâmeros, dímeros e monômeros (Fig. 3-5). No entanto, as ordenações médias equipararam ocorrência e frequência de pentâmeros e dímeros (Fig. 2).

### *Repetições de mononucleotídeos*

Salvo os extremos em IG, Introns e GPG, maior proporção de monômeros concentraram-se em 3'UTR e PR, com menor ocorrência sobre 5'UTR e CDS (Fig.3). No entanto, densidade e frequência em 3'UTR e PR superaram todas as demais (Fig.4-5). As ordenações médias corroboraram com as divergências, equiparando apenas as densidades para Introns e IG ou GPG e 5'UTR, e as frequências entre os dois últimos ou PR e Introns (Fig.2). A discrepância entre o somatório das combinações de adenina e citosina permaneceu constante entre os cromossomos (Fig.6a), sendo T e A prevalente em todas as regiões (S1). Além disso, trechos poliG não foram vistos em 5'UTR.

### *Repetições de dinucleotídeos*

CDS e 5'UTR concentraram o menor contingente de dímeros (Fig.3), sendo similares pelas ordenações de média (Fig.2). Somas maiores distribuíram-se sobre 3'UTR e PR (Fig.3), fora os extremos em GPG, Introns e IG. A não ser por IG, PR apresentou as mais altas densidade e frequência, seguido por 3'UTR apenas em frequência (Fig.4-5). Pelas ordenações médias, densidade e frequência não diferiram entre GPG e 5'UTR, assim como a densidade de Introns e PR ou 3'UTR (Fig.2). No entanto, esses três foram semelhantes em frequência, do mesmo modo que IG e PR. Combinações de AC e AT, juntas, somaram aproximadamente 87,3% das observações (Fig.6a). Exceto em CDS (onde AG/GA prevaleceram, sendo TA raro e AT ausente) e 5'UTR (com pouquíssimos AT/TA e CA), suas combinações também contribuíram em maior número sobre as diferentes regiões (S1).

### *Repetições de trinucleotídeos*

Ambas as UTRs registraram menor ocorrência de trímeros (Fig.3). Apesar dos somatórios em IG, Introns e GPG, esse último esteve muito próximo a CDS, a ponto de se assemelhar pelas ordenações médias (Fig.2). Já suas contribuições e ocorrências em intervalos regulares foram maiores em 5'UTR, PR e CDS, sendo 3'UTR também superior a GPG (Fig.4-5). No entanto, todas as regiões diferiram pelas ordenações médias de densidade e frequência (Fig.2). Quanto aos motivos de repetição, combinações AGC destacaram-se pelo volume ao longo do genoma (Fig.6a), mais que dobro da segunda maior ocorrência, ACG. Todavia, AGC/GCT acumularam-se principalmente em IG, Introns, 3'UTR e GPG, enquanto GGA e GAG

prevaleceram em CDS (S1). Embora as combinações CGC tenham ocorrido em menor número, CGC/GCG concentraram-se em PR, 5'UTR e CDS. O conteúdo GC médio entre os motivos raros variou de 22% em CDS a 66% em Introns.

#### *Repetições de tetranucleotídeos*

Ao contrário de dímeros, 5'UTR somou menor número de tetrâmeros que CDS (Fig.3), com divergência pelas ordenações médias entre todos os contrastes (Fig.2). Além disso, densidade e frequência em 5'UTR e PR superaram todas as demais regiões, com maiores valores em 3'UTR que Introns e IG (Fig.4-5). No entanto, pelas ordenações médias dessas medidas os três últimos foram comparáveis entre si (Fig.2). Combinações ricas em adenina e timina destacaram-se ao longo do genoma (Fig.6b), dispersando-se principalmente sobre IG, Introns, 3'UTR e GPG (S1), embora repetições AAAT/TTTA tenham sido raras em CDS e ausentes em 5'UTR. Sobre PR, 5'UTR e CDS, prevaleceram CTCC, GGCG e CCAG, respectivamente. Já o conteúdo GC médio entre os motivos menos frequentes variou de 40% em PR a 76,8% em GPG. Combinações de CTAG (em menor número sobre o genoma) não existiram em ambas as UTR e CDS, sendo raras em Introns e IG.

#### *Repetições de pentanucleotídeos*

Mesma distribuição das ocorrências de tetrâmeros foi observada entre os pentâmeros (Fig.2-3). No entanto, maiores densidade e frequência foram descritas sobre IG, enquanto ambas as UTRs superaram apenas CDS (Fig.4-5). Pelas ordenações médias, as UTRs foram similares em densidade e frequência, como também GPG e 5'UTR por esta última medida (Fig.2). Dentre os motivos mais frequentes (Fig.6b), combinações ACTGA somaram 29,6% das ocorrências, ou mais que o dobro do segundo, AAGTC. No entanto, AGTCA/TGACT distribuíram-se apenas em IG e Introns, e em poucas ocasiões (S1). A não ser em CDS e 3'UTR, onde predominaram combinações CCCCCG e AAAAC (respectivamente), ACTGA/TCAGT dispersaram-se em grande número sobre todas as regiões. Alto conteúdo GC médio caracterizou as repetições raras, variando de 47,1% em GPG a 63,2% em IG.

### *Repetições de hexanucleotídeos*

A não ser pela maior ocorrência de hexâmeros em 3'UTR que 5'UTR (Fig.3), e densidade e frequência que IG (Fig.4-5), a dispersão destas repetições seguiu igual importância àquela dos trimeros. No entanto, PR e 5'UTR foram equiparados pelas ordenações médias de densidade e frequência, assim como 3'UTR, IG e Introns (Fig.2). Apesar a ocorrência extremamente baixa dos hexâmeros, combinações ACTTTC prevaleceram ao longo do genoma (Fig.6b), exceto em 5'UTR, CDS e 3'UTR em que predominaram ACGCCG, TCTTCA e GTTTTT, respectivamente (S1). Motivos raros também possuíam o mais alto conteúdo GC médio variando de 46,6% em GPG a 76,4% em 5'UTR.



Tabela 1: Ocorrência de *loci* microssatélites sobre o genoma bovino taurino.

	Total
Número de <i>loci</i>	1.330.781
Perfeitos	1.204.500 <sup>a</sup>
Imperfeitos	4.431 <sup>e</sup>
Ligados	16.025 <sup>c</sup>
Compostos	-
por 2	86.002 <sup>b</sup>
por 3	15.008 <sup>d</sup>
por 4	2.877 <sup>f</sup>
por 5	1.938 <sup>g</sup>

Montagem de referência Btau4.2 para a espécie *Bos taurus*, identificador taxológico 9913; *loci* perfeitos: intervalo entre repetições maior ou igual a cem pares de base; *loci* imperfeitos: sequência repetitiva interrompida por até seis pares de base; *loci* ligados: dois microssatélites adjacentes; *loci* compostos por 2, 3, 4 e 5 ou mais repetições, respectivamente; <sup>2</sup>Letras distintas na coluna informam diferenças significativas para ordenação média de ocorrências pelo Teste da Soma de Ordens de Friedman ( $p < 0,001$ ).

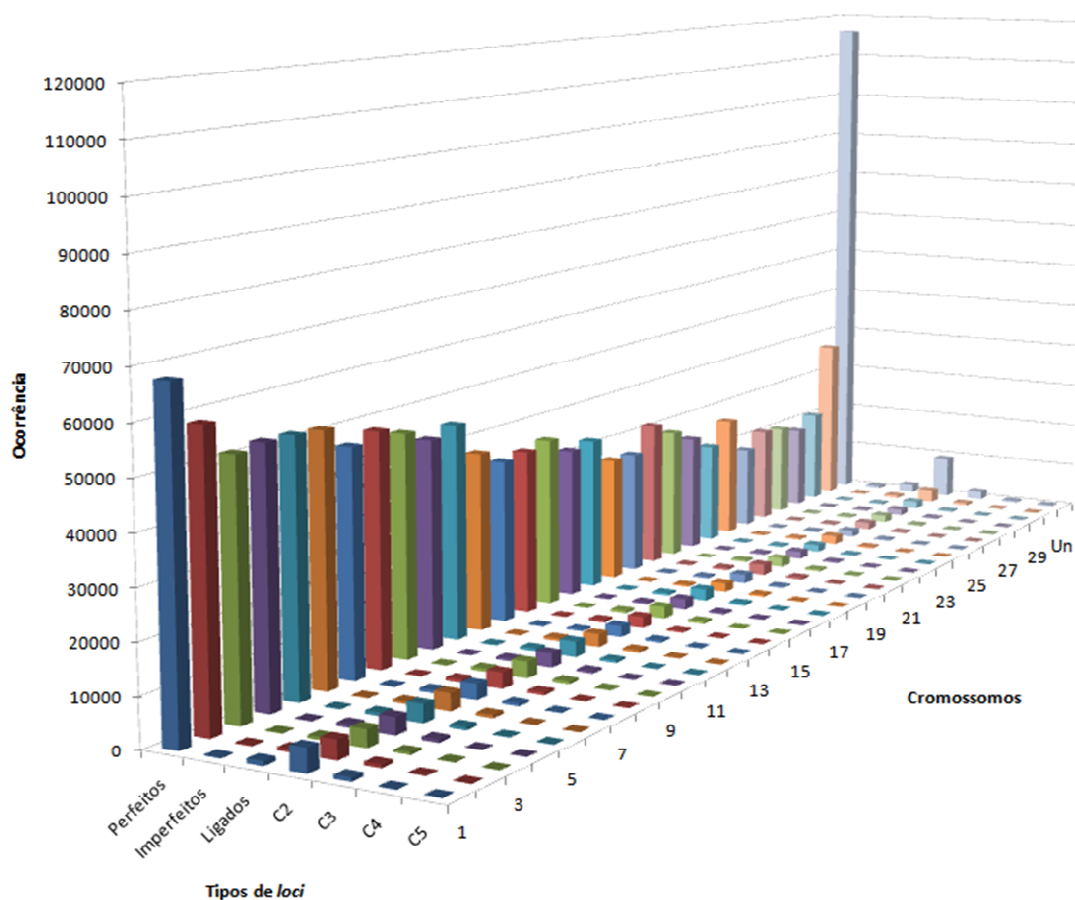


Figura 1: Ocorrências de *loci* microssatélites ao longo do genoma bovino taurino.

Montagem de referência Btau4.2 para a espécie *Bos taurus*, identificador taxológico 9913; *loci* perfeitos: intervalo entre repetições maior ou igual a cem pares de base; *loci* imperfeitos: sequência repetitiva interrompida por até seis pares de base; *loci* ligados: dois microssatélites adjacentes; *loci* compostos (C2, C3, C4 e C5): abrigam 2, 3, 4 e 5 ou mais repetições, respectivamente.

Tabela 2: Distribuição de repetições microssatélites sobre o genoma bovino taurino.

	Ocorrência	Densidade	Frequência
Monômeros	353.742	1.807,80	121,2
Dímeros	271.446	1.730,30	93,0
Trímeros	263.936	1.330,10	90,5
Tetrâmeros	323.646	1.428,30	110,9
Pentâmeros	254.166	1.464,90	87,1
Hexâmeros	20.903	134,2	7,2
PR	7.412	9.584,85	627,36
5'UTR	1.482	11.072,90	756,89
CDS	5.817	2.485,68	177,74
Intron	425.789	7.562,66	497,59
3'UTR	5.602	7.771,54	529,46
IG	1.038.314	8.039,75	516,92
GPG	25.149	4.125,47	265,67

Montagem de referência Btau4.2 para a espécie *Bos taurus*, identificador taxológico 9913; PR: promotor, 500 pares de base antes do sítio para início de transcrição; 5'UTR: região 5' não traduzida; CDS: região codificadora; 3'UTR: região 3' não traduzida; INTRON: região intragênica; IG: região intergênica; GPG: genes sem regiões descritas ou pseudo-genes; Ocorrência: somatório das observações; Densidade: contribuição das repetições a cada milhão de pares de base, expressa em pb/Mb; Frequência: somatório das observações a cada milhão de pares de base, expressa em SSR/Mb.



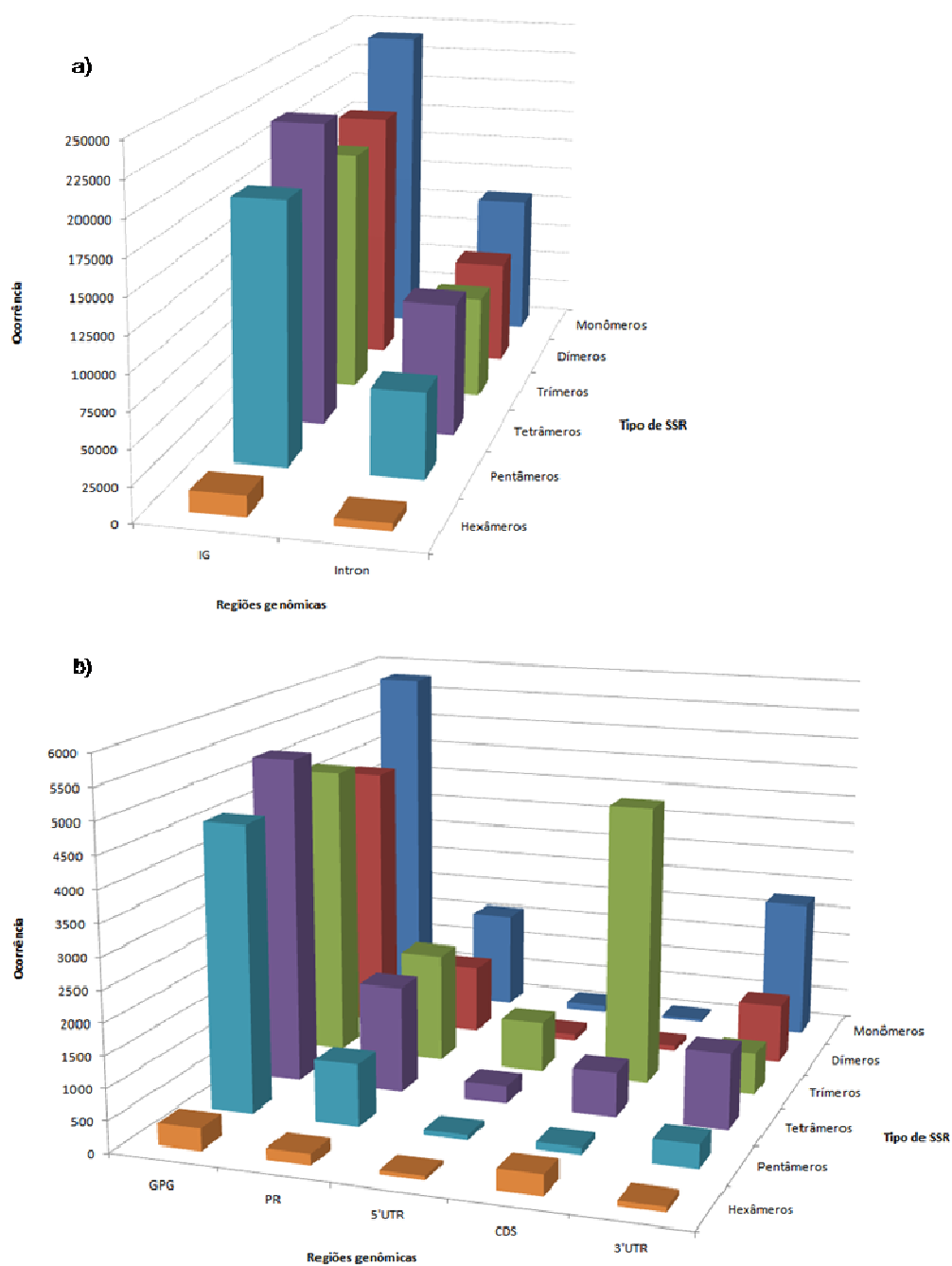


Figura 3: Ocorrência de repetições microssatélites entre regiões do genoma bovino taurino.

Montagem de referência Btau4.2 para a espécie *Bos taurus*, identificador taxológico 9913; PR: promotor, 500 pares de base antes do sítio para início de transcrição; 5'UTR: região 5' não traduzida; CDS: região codificadora; 3'UTR: região 3' não traduzida; INTRON: região intragênica; IG: região intergênica; GPG: genes sem regiões descritas ou pseudo-genes; Ocorrência: somatório das observações. IG e Intron em (a) foram ser agrupadas dado o volume de ocorrências.

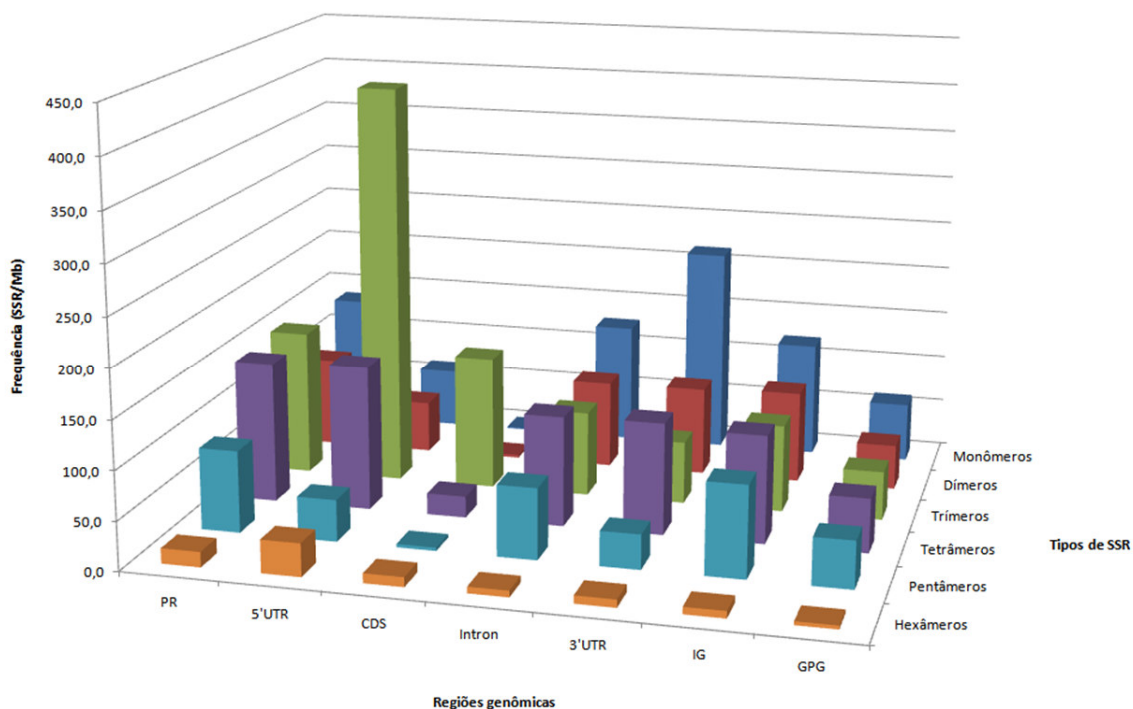


Figura 4: Frequência de repetições microsatélites entre regiões do genoma bovino taurino.

Montagem de referência Btau4.2 para a espécie *Bos taurus*, identificador taxológico 9913; PR: promotor, 500 pares de base antes do sítio para início de transcrição; 5'UTR: região 5' não traduzida; CDS: região codificadora; 3'UTR: região 3' não traduzida; INTRON: região intragênica; IG: região intergênica; GPG: genes sem regiões descritas ou pseudo-genes; Frequência: somatório das observações a cada milhão de pares de base, expressa em SSR/Mb.

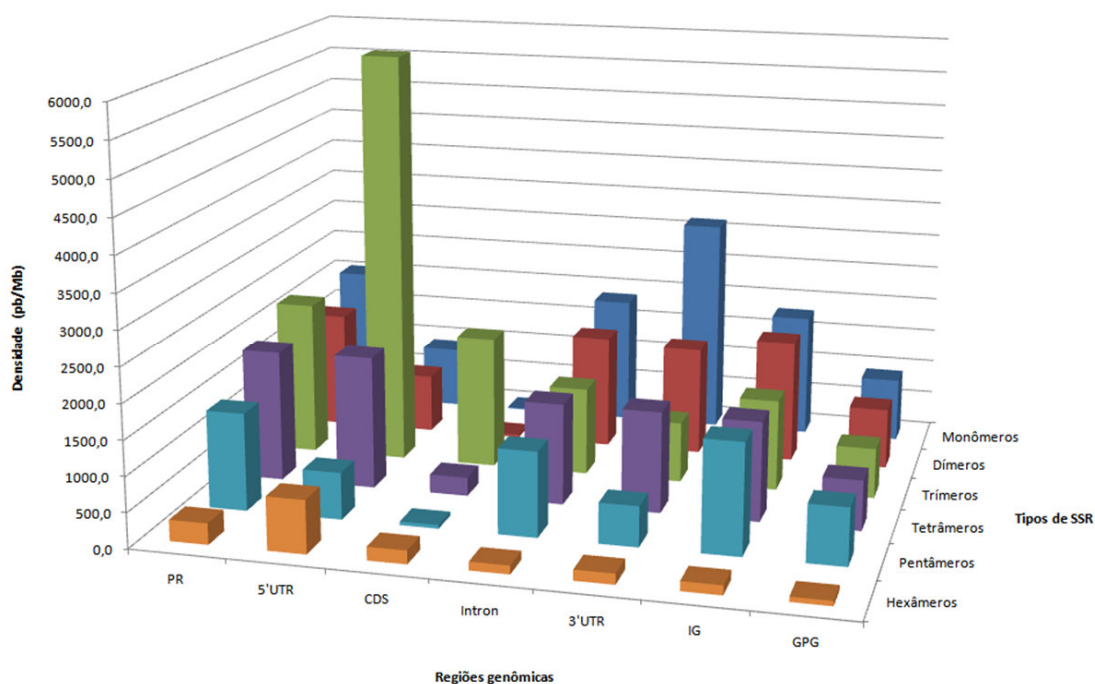


Figura 5: Densidade de repetições microssatélites entre regiões do genoma bovino taurino.

Montagem de referência Btau4.2 para a espécie *Bos taurus*, identificador taxológico 9913; PR: promotor, 500 pares de base antes do sítio para início de transcrição; 5'UTR: região 5' não traduzida; CDS: região codificadora; 3'UTR: região 3' não traduzida; INTRON: região intragênica; IG: região intergênica; GPG: genes sem regiões descritas ou pseudo-genes; Densidade: contribuição das repetições a cada milhão de pares de base, expressa em pb/Mb.





### 2.3 Discussão

Embora a montagem alternativa UMD3.1 tenha sido gerada a partir de material cedido pelo BGSC, existem consideráveis diferenças em relação a montagem de referência Btau4.2, como extensão dos cromossomos, anotação gênica e número de segmentos não alocados. Desse modo, uma não exclui a importância da outra. No entanto, devido à similaridade dos descritores, discuti-las simultaneamente produziria redundância das informações. No entanto, detalhes de ambas as montagens podem ser encontrados como material suplementar (S1).

Comumente, inexistente a preocupação em caracterizar *loci* microssatélite (SUBRAMANIAN; MISHRA; SINGH, 2003; SERAPION et al., 2004; JU et al., 2005; YAN et al., 2008; ADELSON; RAISON; EDGAR, 2009; ADELSON et al., 2010; ZHANG et al., 2010), nem mesmo em descrever aqueles não perfeitos (TEMNYKH et al., 2001; BRANDSTROM; ELLEGREN, 2008; ZHAN et al., 2009). Isso talvez pela parcela que esses últimos representam (9,5% neste estudo). No entanto, é importante lembrar que em se tratando de um *locus* microssatélite, os alelos são atribuídos conforme o número de repetições de seu motivo, que conseqüentemente geram variações no tamanho de fragmentos amplificados pela Reação em Cadeia da Polimerase - PCR (GAO et al., 2009). Pequenos intervalos entre as repetições permitem que um mesmo segmento abrigue mais de uma sequência repetitiva. Embora tais *loci* possam ser mais instáveis, a medida que incluem mais repetições (LI et al., 2002; AISHWARYA; GROVER; SHARMA, 2007; FAN; CHU, 2007), seu uso pode levar a erros. Caso uma repetição seja estendida enquanto outra for reduzida, dois alelos terão o mesmo tamanho, mas com conteúdos diferentes, sendo possível diferenciá-los apenas por sequenciamento. Além disso, um intervalo mínimo é necessário para que *primers* específicos sejam desenhados a fim de capturar uma região de interesse. No presente trabalho, ao atingir 500pb um *locus* composto era isolado após encontrar uma extensão mínima de 40pb a partir da última sequência repetitiva. Essa característica permite que os segmentos flanqueantes as repetições alcancem uma eficiência adequada para desenho de *primers* (JU et al., 2005), além de evitar *loci* extremamente grandes.

Embora a nomenclatura para microssatélites inclua motivos de um a seis nucleotídeos (TAUTZ, 1993), monômeros comumente são ignorados ao estabelecer os parâmetros de busca (KRUGLYAK et al., 1998; SERAPION et al., 2004; JU et al., 2005; BRANDSTROM; ELLEGREN, 2008; YAN et al., 2008; ZHANG et al., 2010),

seja pela difícil utilização como marcadores - uma vez que diferenças entre os alelos são de apenas um nucleotídeo - ou pela maior chance de erros na genotipagem. Contudo, eles representaram a maior parcela das observações no presente estudo. Ignorá-los poderia subestimar a presença de *loci* compostos ou ligados. Consequentemente, a seleção de marcadores para posterior utilização poderia ser prejudicada, uma vez que muitos podem ser erroneamente classificados como perfeitos, apenas por tal critério. Além disso, um número expressivo de monômeros foi também identificado sobre segmentos promotores, e em volume maior que dímeros, pentâmeros ou hexâmeros.

Estimativas da cobertura genômica por sequências microssatélites oscilam entre menos de um por cento entre plantas (MORGANTE; HANAFEY; POWELL, 2002) a 3% para humanos (SUBRAMANIAN; MISHRA; SINGH, 2003). No entanto essa proporção é fortemente influenciada pelos parâmetros de busca. Quanto maior o número de unidades de repetição menor a sua ocorrência (YAN et al., 2008). Embora não tenha sido alvo deste trabalho descrever quanto cada estimativa é inflacionada ao reduzir o limiar em uma unidade de repetição, a partir de certo ponto (principalmente para motivos maiores, como hexâmeros) muitas repetições não haviam sido incluídas em busca anterior por conter mutações de ponto sobre os motivos de repetição. Como as taxas de mutação por escorregões da DNA Polimerase são superiores a aquelas para alterações pontuais (KRUGLYAK et al., 1998; LI et al., 2002; KASHI; KING, 2006; FAN; CHU, 2007), variações de expansão ou contração de trechos repetitivos são mais recentes no histórico evolutivo, mas comumente são estabilizadas através de mutações de ponto (FAN; CHU, 2007; BRANDSTROM; ELLEGREN, 2008; CRUZ; ROUX; ROBINSON-RECHAVI, 2009) fornecendo um mecanismo para controlar o crescimento desordenado dos microssatélites (KRUGLYAK et al., 1998). Assim, ao reduzir o limiar dos parâmetros de busca poderia apenas inflacionar uma estimativa com um maior número de trechos monomórficos e não informativos. Quando uma triagem foi conduzida para identificar hexâmeros com apenas duas repetições, a cobertura genômica aumentou em mais de 1%, chegando próximo a aquela atribuída para o genoma humano (SUBRAMANIAN; MISHRA; SINGH, 2003).

Existem evidências que a taxa de mutação em microssatélites aumenta com o número de cópias dos motivos de repetição (GAO et al., 2009) ou com o tipo de repetição (KRUGLYAK et al., 1998). Por esse motivo, constantemente,

microssatélites são classificados em classe I - hipervariáveis - ou classe II - potencialmente variáveis (TEMNYKH et al., 2001). No entanto, uma maior proporção das repetições curtas - classe II - tende a ser variável, embora com menor número de alelos (menos polimórficas) que aquelas longas – classe I (BRANDSTROM; ELLEGREN, 2008). Além disso, tipos de repetições menores ocorrem em maior proporção que os demais (SERAPION et al., 2004; JU et al., 2005; YAN et al., 2008; ZHANG et al., 2010). No entanto é importante notar um vies que surge dos algoritmos empregados. A busca por repetições inicia sempre por um dos extremos, monômeros (na grande maioria das vezes) ou hexâmeros. Ao identificar uma sequência repetitiva ele impede que outra seja alocada sobre o mesmo local. Em alguns casos surgem dúvidas, por exemplo, ao discriminar dímeros (TG)<sub>9</sub> ou hexâmeros (TGTGTG)<sub>3</sub>. É oportuno lembrar que este dímero ocorreu com extrema abundância no presente estudo, sem nenhum caso do hexâmero. Visto a dificuldade de determinar qual a repetição apropriada, ao iniciar por motivos menores acaba-se por subestimar aqueles maiores, e vice versa.

Anteriormente, a ocorrência de repetições microssatélites foi também associada ao tamanho do genoma (SUBRAMANIAN; MISHRA; SINGH, 2003), assim como ao conteúdo CG das sequências vizinhas (GLENN et al., 1996; TEMNYKH et al., 2001; ALBA; GUIGO, 2004; BRANDSTROM; ELLEGREN, 2008; ADELSON; RAISON; EDGAR, 2009; CRUZ; ROUX; ROBINSON-RECHAVI, 2009). No entanto, tão importante quanto isso é a capacidade de discriminar a distribuição sobre as regiões estruturais (AISHWARYA; GROVER; SHARMA, 2007), visto que certas repetições são frequentemente predominantes em determinados locais do genoma (SUBRAMANIAN; MISHRA; SINGH, 2003) ou grupos de genes (ALBA; GUIGO, 2004; CRUZ; ROUX; ROBINSON-RECHAVI, 2009). Embora seja esperado que regiões não codificantes incluam maior proporção de repetições microssatélite, as medidas de ocorrências, densidades e frequências sobre o genoma bovino permitiram que algumas semelhanças fossem descritas entre IG, Introns e 3'UTR, como também entre PR e 5'UTR. A não ser para monômeros, os componentes das combinações dos motivos de repetição prevalentes no genoma distribuíam-se principalmente em IG, Introns, GPG, 3'UTR e PR, com menor número sobre 5'UTR. Já muitas das repetições raras concentravam-se sobre regiões codificantes de proteínas (CDS). Atenção especial também deve ser atribuída ao somatório de microssatélites descritos muito próximos a sítios de início de transcrição (prováveis

regiões promotoras), os quais podem estar influenciando a expressão de seus genes ou o reconhecimento de fatores de transcrição.

Para a gama de resultados, alguns achados mereceram maior discussão. Dímeros AC/TG preditos como dominantes entre regiões expressas (SUBRAMANIAN; MISHRA; SINGH, 2003; SERAPION et al., 2004; JU et al., 2005; ZHANG et al., 2010) inclusive de bovinos (YAN et al., 2008) não demonstraram igual importância em CDS no presente estudo. Já motivos de trímeros sugeridos como menos passíveis de escorregões da DNA Polimerase e a formação de estruturas secundárias (BRANDSTROM; ELLEGREN, 2008) predominaram entre aqueles aqui descritos sobre CDS. Apesar da abundância de repetições ATT no genoma humano (SUBRAMANIAN; MISHRA; SINGH, 2003), motivos AGC já haviam sido descritos como prevalentes em ovinos, suínos, como também em bovinos (ELSIK et al., 2009), porém sua super-representação sobre sequências codificadoras (YAN et al., 2008) não concordaram com resultados encontrados. Essas discrepâncias surgem não por erros de predição, mas devido a diferenças entre as sequências utilizadas, algoritmos, parâmetros de busca (MAIA et al., 2008) e possíveis agrupamentos dos motivos de repetição.

Como a seleção de um conjunto de *loci* para estudo depende dos objetivos da pesquisa ou do grupo de genes sob investigação, não houve interesse em disponibilizar *primers* para cada *locus*, uma vez que é aconselhável a criação de arranjos *multiplex* para maximizar a eficiência das genotipagens. Se apenas *loci* perfeitos de dímeros ou trímeros fossem selecionados dentre todas as repetições microssatélites identificadas no presente estudo para segmentos mapeados haveriam 371.015 trechos candidatos (um *locus* a cada 7100,9pb), com aproximadamente 32,7% destes alocados sobre regiões gênicas. Identificar e descrever a distribuição dessas sequências ao longo genoma bovino, além de apontar sua orientação com relação a anotação gênica, marcadores moleculares (STS e SNP) e outros elementos repetitivos, fornece um recurso capaz de ajudar a elucidar a influencia de tais regiões sobre características de interesse, bem como o desenvolvimento de estratégias de seleção assistida (DEKKERS, 2004). Dada a importância filogenética dos bovinos (ELSIK et al., 2009; GIBBS et al., 2009; LEWIN, 2009; TELLAM et al., 2009), a transferência dessa informação pode também beneficiar aquelas espécies com menor nível de conhecimento (SERAPION et al.,

2004; JU et al., 2005; AISHWARYA; GROVER; SHARMA, 2007; LEWIN, 2009), como os zebuínos e outros do gêneros *Bos*.

#### 2.4 Conclusão

Grande número de repetições microssatélite foi identificado e descrito ao longo do genoma bovino taurino, com ampla distribuição sobre as diferentes porções estruturais.

### 3 Anotação funcional de repetições microssatélites identificadas sobre o genoma bovino taurino.

Nessa segunda etapa o objetivo foi atribuir anotação funcional baseada em ontologia gênica para as repetições microssatélites identificadas ao longo do genoma bovino taurino.

#### 3.1 Material e Métodos

##### *Informação genômica*

A anotação funcional baseada em ontologia gênica para a espécie bovina foi obtida a partir do repositório B2G-FAR (GOTZ et al., 2011), em outubro de 2011. Tal escolha reflete a abrangência e acurácia da informação ofertada, aliada a uma metodologia de anotação conceituada (CONESA et al., 2005; CONESA; GOTZ, 2008; GOTZ et al., 2008). Como os identificadores de sequências diferem daqueles empregados na anotação estrutural de ambas as montagens genômicas, recursos (gene2accession, gene2ensembl, gene2go, gene\_info, gene\_refseq\_uniprotkb\_collab) do repositório para informações baseada em genes de organismos completamente seqüenciados do *National Center for Biotechnology Information* (<ftp://ftp.ncbi.nlm.nih.gov/gene/>) foram necessários para interligar os dados. Além disso, registros (gene\_ontology\_ext.obo, grath\_path, term, term\_synonym) disponibilizados pelo *Gene Ontology Database* (<http://www.geneontology.org/>) permitiram reconstruir a estrutura hierárquica dos termos. Ao final, a anotação funcional completa para cada montagem (Btau4.2 e UMD3.1) serviu como referencia em análises comparativas posteriores.

##### *Repetições microssatélites*

Como a ontologia é atribuída a produtos gênicos, somente aquelas repetições microssatélites mapeadas sobre genes transcritos foram selecionadas. Conseqüentemente, a partir de tais genes, um subconjunto da anotação funcional completa de cada montagem (SSR4.2 e SSR3.1) representou estas sequências em análises comparativas posteriores. Dado a gama de informações reunidas, houve a necessidade de construir um novo *pipeline* a fim de transferir, mesclar e descrever os dados apurados. Para isso duas linguagens de programação (PASCAL e PERL) estruturam as informações através de um banco de dados (MySQL).

### *Medidas descritivas e análise estatística*

Como cada termo é interligado a outros por uma relação hierárquica (em níveis) e a regra de anotação retém a somente aqueles mais específicos atribuídos a determinada sequência (CONESA; GOTZ, 2008; GOTZ et al., 2008; GOTZ et al., 2011), cada identificador contribui para o montante de seus relativos de modo ascendente até o termo raiz (ZHOU; SU, 2007). Após apurar os somatórios de termos dentro de cada um dos domínios principais (função molecular, processo biológico e componente celular) as descrições prosseguiram até a sua máxima profundidade (nível mais baixo).

A fim de identificar a existência de termos em abundância sobre o conjunto de genes transcritos abrigando repetições microssatélites, uma função implementada junto ao aplicativo estatístico R (2011) executou sucessivas análises de enriquecimento (BLUTHGEN et al., 2005) através do Teste Exato de Fisher, preservando o nível da ontologia. Ajustes de Bonferroni corrigiram as probabilidades para erro tipo I entre os múltiplos contrastes.

### 3.2 Resultados

Assim como no estudo anterior, os descritores foram muito similares entre as duas montagens gênicas. No entanto, Btau4.2 apresentou maior número de genes transcritos com anotação GO. Desse modo, para facilitar a interpretação dos resultados, novamente a discussão segue referindo-se sobre a montagem de referência, sem perda de informação sobre construção Btau5.2. Informações detalhadas sobre ambas as montagem consta em material suplementar (S2) disponível pela plataforma *Pampeano Animal Genetic Resources*.

O subconjunto genes transcritos abrigando repetições microssatélites (Tab. 3) conservou grande proporção tanto da anotação funcional do genoma bovino (82,6%) como do total de termos GO atribuídos a ela (89,8%). Ao mesmo tempo, as médias para o número de identificadores por sequência anotada chegaram a estimativas muito próximas. Para os termos raiz (domínios GO), considerados o primeiro nível da ontologia, a semelhança entre as distribuições foi evidente entre Btau4.2 e SSR4.2 (Fig. 7). Sempre um maior volume destes referiu-se a processo biológico que função molecular ou componente celular, independente do conjunto em questão. No entanto, pela análise de enriquecimento, houve um acúmulo

significativo de termos em SSR4.2 para processo biológico ( $p < 0,01$ ) e componente celular ( $p < 0,05$ ).

Dentre os vinte possíveis termos ao segundo nível da ontologia gênica para função molecular, dezoito caracterizaram os genes mapeados em Btau4.2, sendo a maior parcela atribuídos a ligação, atividade catalítica, atividade transdutor molecular e atividade de receptor. Contudo, em SSR4.2, atividade transportadora recebeu maior importância que os dois últimos (Fig. 8), assim como em SSR3.1 e UMD3.1. Além destes cinco termos, atividade fator de transcrição ligado a proteína e atividade regulador de enzima apresentaram enriquecimento significativo em SSR4.2 ( $p < 0,05$ ). Dentre 3.887 termos anotados até o décimo segundo nível GO, apenas 75 estavam significativamente enriquecidos (S2).

Treze dos quatorze termos atribuídos ao segundo nível da ontologia para componente celular possuíam genes transcritos relacionados (Fig. 9). Identificadores para célula, parte de célula, organela e parte de organela estavam em maior número tanto em Btau4.2 como SSR4.2. Apesar disso, pelo teste exato de Fisher ( $p < 0,05$ ), somente região extracelular, complexo macromolecular, sinapse e parte da sinapse estavam enriquecidos em SSR4.2 além de organela. Nesse domínio, a profundidade máxima alcançou apenas dez níveis (S2). Neles, 2.247 termos GO tinham genes transcritos atribuídos, enquanto que 75 estavam significativamente enriquecidos.

Já processo biológico alcançou a profundidade máxima da anotação gênica, ou seja, doze níveis. Ao segundo destes (Fig. 10), para os 28 termos descritos, 26 foram relacionados para ao menos um gene transcrito. Termos para processo celular, regulação biológica, processo de organismo multicelular e resposta a estímulo prevaleceram tanto em Btau4.2 como SSR4.2. Junto a outros onze termos, processo de organismo multicelular apresentou enriquecimento significativo em SSR4.2 ( $p < 0,05$ ), sendo eles: morte celular, processo metabólico, proliferação celular, adesão biológica, sinalização, processo desenvolvimental, locomoção, localização, estabelecimento de localização, processo de multi-organismos, e biogênese e organização de componente celular. Dentre 17.774 termos anotados até o décimo segundo nível GO, 169 estavam significativamente enriquecidos (S2).

Através da plataforma *Pampeano Animal Genetic Resources*, material suplementar (S2) fornece listas completas com todos os termos GO, estando ordenados por nível da ontologia e probabilidade para o erro tipo I. As informações foram também individualizadas para processo biológico, componente celular e



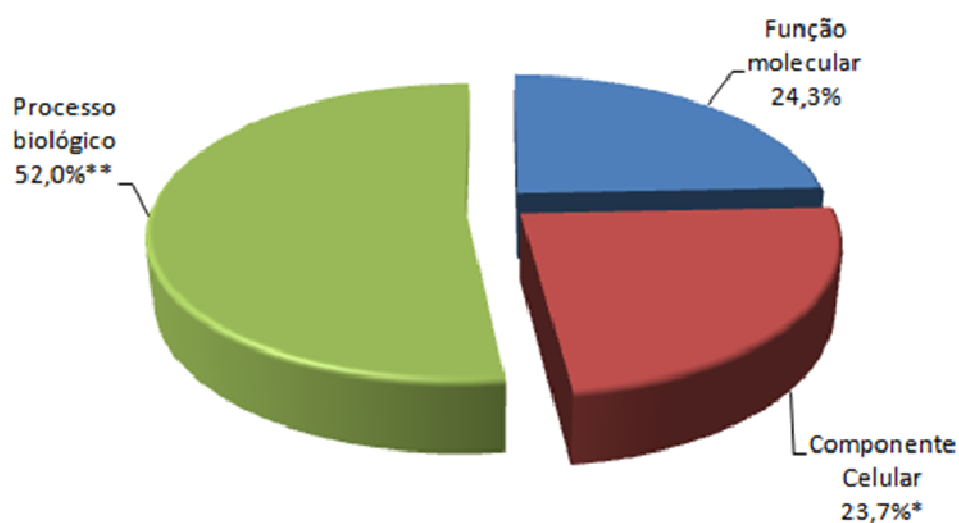
função molecular. Ao mesmo tempo, a ferramenta de buscas por *loci* microssatélites foi otimizada para selecionar repetições através da anotação GO de genes transcritos abrigando trechos repetitivos.

Tabela 3: Ontologia Gênica recuperada para genes transcritos mapeados sobre o genoma bovino taurino, utilizando dados dos repositórios B2G-FAR, Gene/NCBI e GOC.

	SSR4.2	Btau4.2
Transcritos	19.536	23.644
Transcritos com GO	11.793	13.666
Total de GO	100.009	111.410
GO por Transcrito( $\bar{x} \pm EP$ )	8,48 $\pm$ 0,09	8,15 $\pm$ 0,08

SSR4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2 abrangendo repetições microssatélites; Btau4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2. B2G-FAR: <http://bioinfo.cipf.es/b2gfar/home>; Gene/NCBI: <ftp://ftp.ncbi.nlm.nih.gov/gene/>; GOC: <http://www.geneontology.org/>.

## a) SSR4.2



## b) Btau4.2

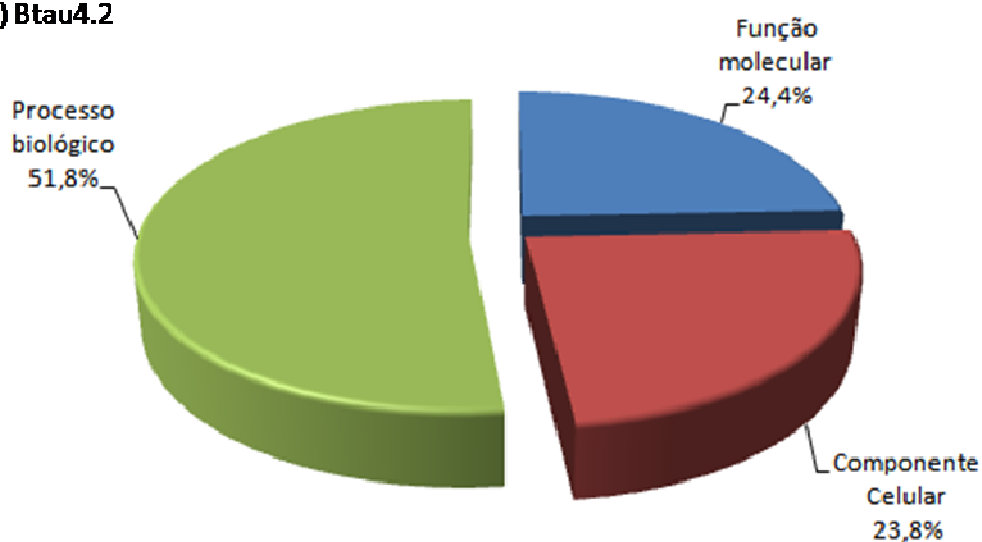


Figura 7: Distribuição da anotação funcional para função molecular, processo biológico e componente celular em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites.

SSR4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2 abrigando repetições microssatélites; Btau4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2; \*, \*\*, \*\*\*: diferenças significativas pela análise do enriquecimento de termos GO, com probabilidade para o erro tipo I de 5%, 1% e 0,1%, respectivamente.

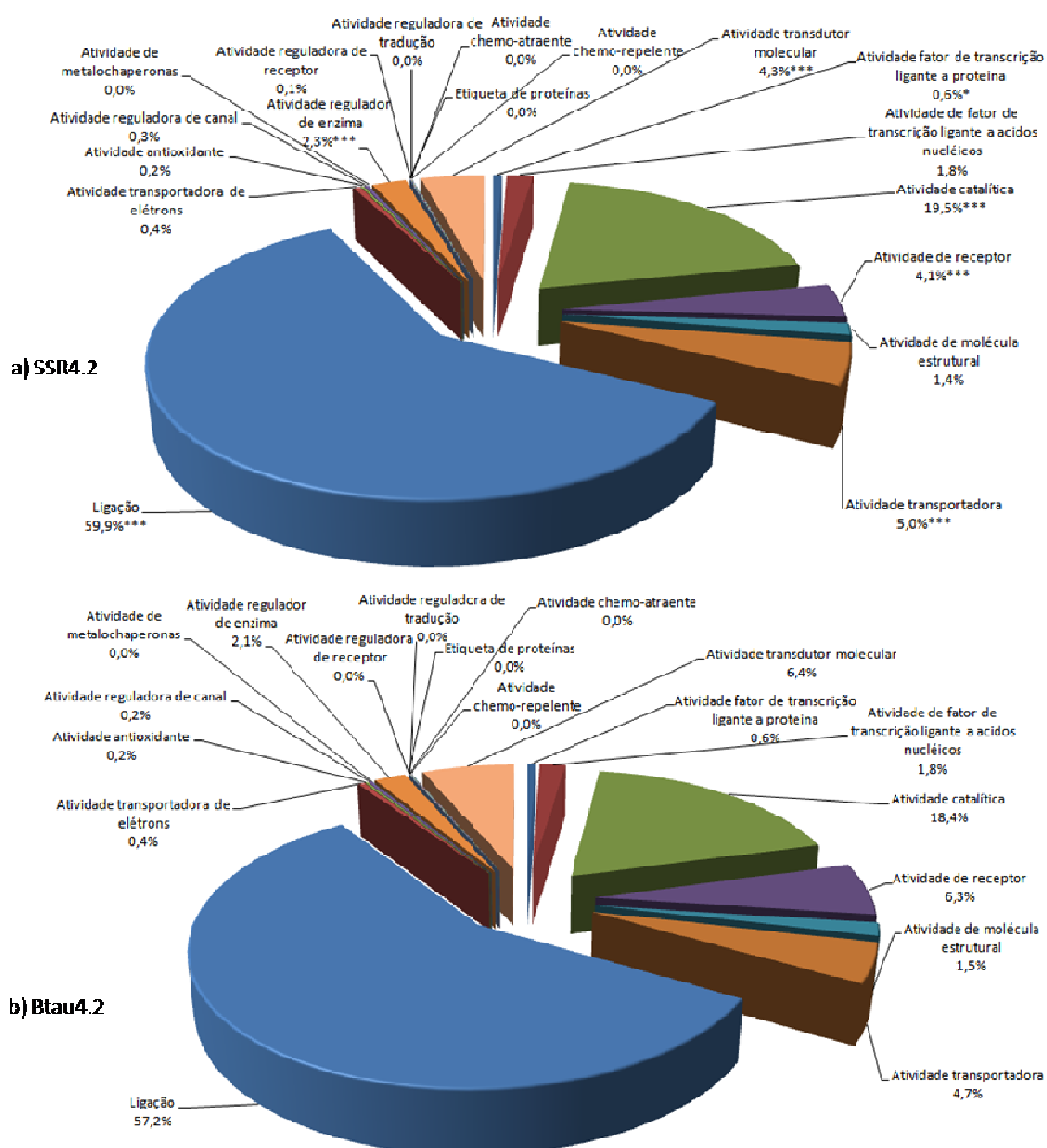


Figura 8: Distribuição ao segundo nível da anotação funcional para função molecular em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites.

SSR4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2 abrigando repetições microssatélites; Btau4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2; \*, \*\*, \*\*\*: diferenças significativas pela análise do enriquecimento de termos GO, com probabilidade para o erro tipo I de 5%, 1% e 0,1%, respectivamente.

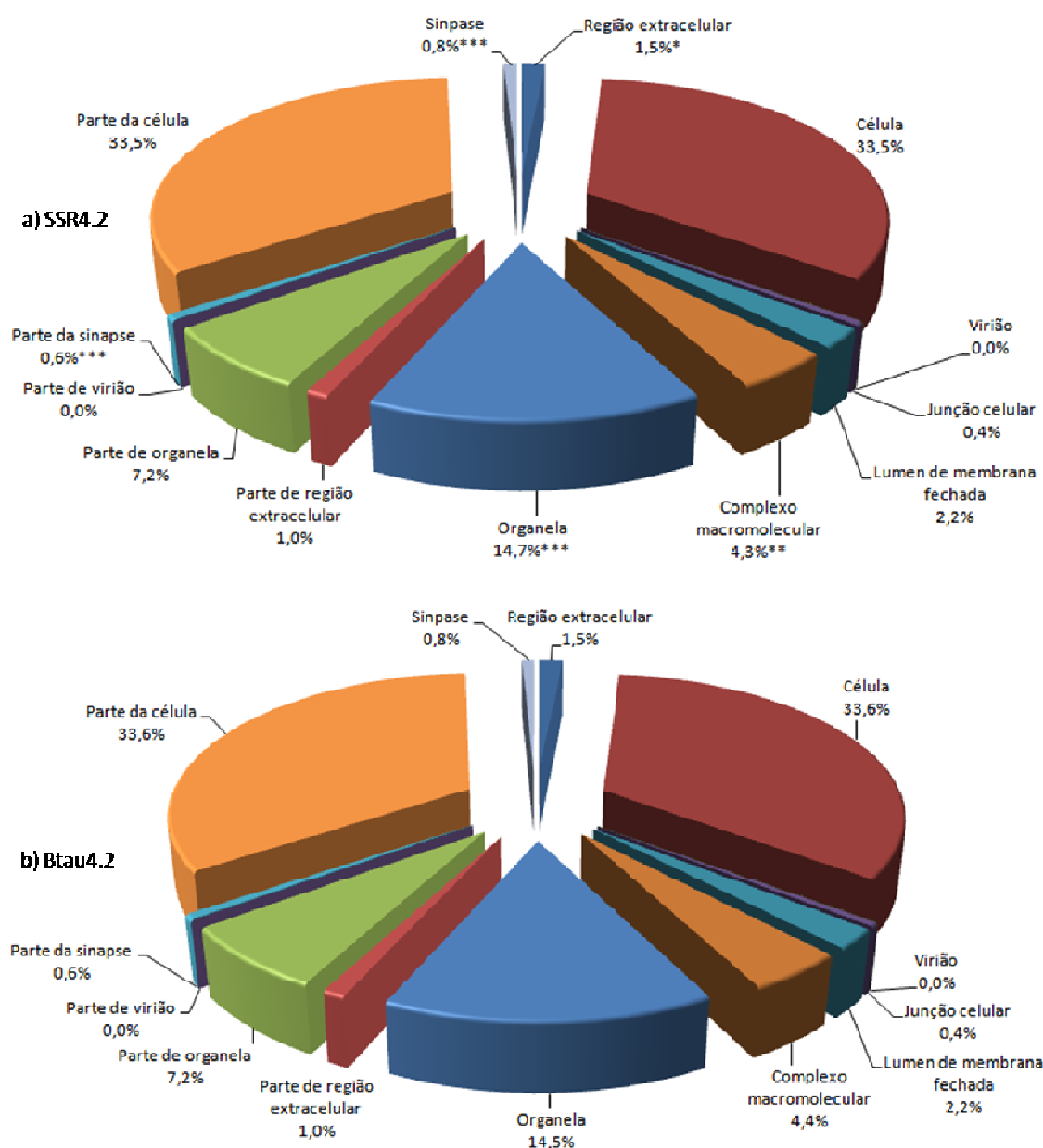


Figura 9: Distribuição ao segundo nível da anotação funcional para componente celular em genes transcritos mapeados sobre o genoma bovino taurino abrigando repetições microssatélites.

SSR4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2 abrigando repetições microssatélites; Btau4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2; \*, \*\*, \*\*\*: diferenças significativas pela análise do enriquecimento de termos GO, com probabilidade para o erro tipo I de 5%, 1% e 0,1%, respectivamente.

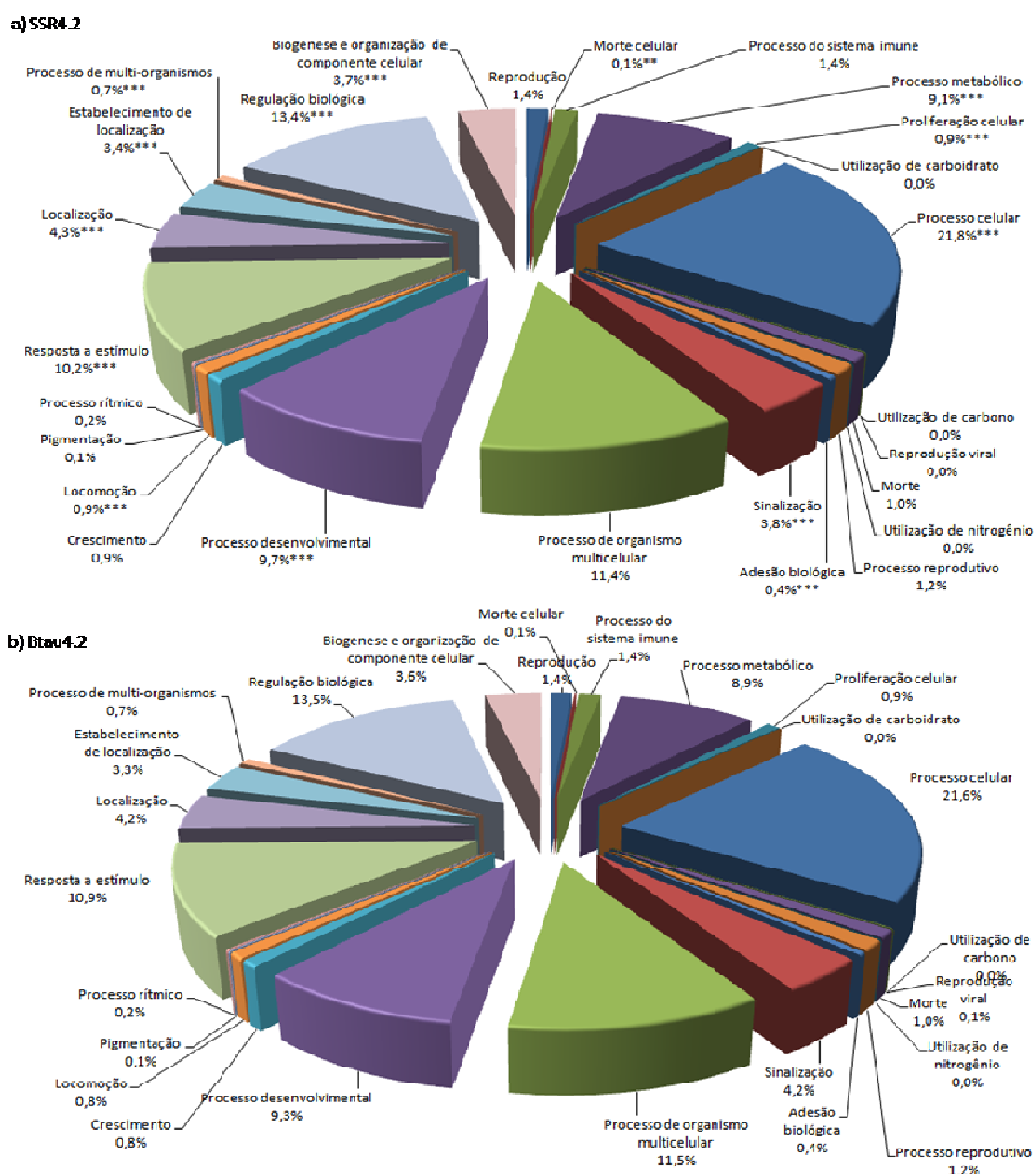


Figura 10: Distribuição ao segundo nível da anotação funcional para processo biológico em genes transcritos mapeados sobre o genoma bovino taurino abrangendo repetições microssatélites.

SSR4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2 abrangendo repetições microssatélites; Btau4.2: conjunto de genes transcritos mapeados sobre a montagem de referência do genoma bovino Btau4.2; \*, \*\*, \*\*\*: diferenças significativas pela análise do enriquecimento de termos GO, com probabilidade para o erro tipo I de 5%, 1% e 0,1%, respectivamente.

### 3.3 Discussão

Em construções anteriores do genoma bovino, comumente duas montagens estavam disponíveis. Uma permitia a análises de genes específicos (anotação estrutural), oferecendo também uma sequência consenso alternativa. Já a outra servia como referencia para construção dos cromossômicos e demais análises globais (ELSIK et al., 2009). Assim, seus objetivos explicavam suas divergências. No entanto, no presente estudo, Btau4.2 apresentou maior proporção de genes transcritos com anotação funcional, além de um maior número de termos GO atribuídos. É oportuno lembrar que durante a estruturação do repositório B2G-FAR, as regras de anotação (CONESA; GOTZ, 2008) inibiram a presença de termos anotados por meio eletrônico (GOTZ et al., 2011). Desse modo, é possível que variações nas predições e curadoria do conjunto de genes atribuído a cada uma das montagens levem as diferenças observadas. Contudo, para esclarecer tal questão seria necessário um estudo aprofundado sobre o tema.

Investigação funcional, ao contrário da anotação estrutural, busca descrever como diferentes porções do genoma interagem frente a variações ambientais que promovem mudanças quantitativas e qualitativas em função (MCCARTHY et al., 2006). No entanto, para comparar resultados de estudos experimentais complexos é fundamental a padronização de um vocabulário. Isso elimina a chance de termos redundantes e permite que os agentes subjacentes façam inferências ou descobertas sobre os dados (HUGHES et al., 2008; GGR-GOC, 2009). Por essa razão, iniciativas como *Gene Ontology* impulsionaram grandes avanços na pesquisa genômica, tornaram essencial a caracterização funcional de regiões estruturais dos genomas (CONESA; GOTZ, 2008; GOTZ et al., 2008; GGR-GOC, 2009; LECERF et al., 2011). É difícil encontrar um domínio biológico sem projetos de mesmo âmbito, tanto que mais de 2000 espécies já possuem anotação baseada em ontologia gênica (GOTZ et al., 2011). Apesar disso, somente em uma espécie de planta (*Physcomitrella patens*) tal abordagem é empregada de modo semelhante a esse manuscrito (VICTORIA; MAIA; OLIVEIRA, 2011). Embora com semelhanças entre alguns termos com maior frequência, as divergências são passíveis, visto as fontes de dados utilizadas e a distância filogenética para os bovinos.

Até o momento, a caracterização de sequências genômicas abrigando repetições microssatélites não considerava anotação baseada em ontologia (TEMNYKH et al., 2001; SUBRAMANIAN; MISHRA; SINGH, 2003; SERAPION et al.,

2004; JU et al., 2005; AISHWARYA; GROVER; SHARMA, 2007; YAN et al., 2008; ADELSON; RAISON; EDGAR, 2009; ZHAN et al., 2009; ZHANG et al., 2010). Possivelmente, a finalização de projetos de seqüenciamento genômico para outras espécies (OZDEMIR OZGENTURK et al., 2010; GOTZ et al., 2011), a expansão gradual de repositórios públicos (KIM et al., 2006; OZDEMIR OZGENTURK et al., 2010), ou mesmo a reavaliação de seus conteúdos (VICTORIA; MAIA; OLIVEIRA, 2011) possibilitem novos estudos comparativos.

Sobre cada termo da ontologia gênica, ao realizar uma análise do enriquecimento o objetivo é saber se um termo particular está enriquecido em um grupo teste quando comparado a um grupo de referencia (CONESA et al., 2005; ZHOU; SU, 2007; CONESA; GOTZ, 2008; GOTZ et al., 2008), ou seja, a hipótese de nulidade considera que o grupo teste representa uma amostra aleatória do grupo de referência (BLUTHGEN et al., 2005; ZHOU; SU, 2007). Em nosso estudo, o conjunto de todos os genes transcritos de uma determinada montagem constitui o grupo de referencia, sendo o grupo teste o subconjunto destes que continham repetições microssatélite mapeadas. Esse tipo de análise é constante utilizada para diferenciar anotação funcional entre grupos de genes com distintos níveis de expressão em ensaios com microarranjos (CONESA et al., 2005; ZHOU; SU, 2007; CONESA; GOTZ, 2008; ZHANG; ZAN; WANG, 2011; ZHAO et al., 2011) ou de expressão diferencial (CONESA et al., 2005; GALLA et al., 2009; ZHANG; ZAN; WANG, 2011; ZHAO et al., 2011). Infelizmente ainda não existem estudos aplicando esta metodologia para os mesmo fins que o presente estudo. No entanto, ao apontar termos enriquecidos são identificados processos biológicos, funções moleculares ou componentes celulares sendo monitorados por um número maior de genes abrigando repetições microssatélites que seria esperado por um processo aleatório. Devido a diferenças em especificidade, todas as análises realizadas preservaram a estrutura hierárquica entre os níveis da ontologia, comparando apenas termos sobre o mesmo nível.

Ao transferir anotação funcional para uma parcela das repetições microssatélites identificadas sobre o genoma bovino taurino, ampliou-se o horizonte de aplicações para os dados gerados no presente estudo. Grupos de pesquisa interessados em estudar, por exemplo, microssatélites sobre genes transcritos que contribuem para processos metabólicos (GO:0008152) teria inicialmente 5.791 genes alvo. Obviamente que garimpando para termos mais específicos seria



possível criar conjuntos *multiplex* para avaliar grupos menores de genes. Por outro lado, pesquisadores munidos de informações de outros repositórios (BARRETT et al., 2005) ou evidências experimentais de genes diferencialmente expressos atuando sobre características de interesse (TAN et al., 2006; FINUCANE et al., 2008; ZHANG; ZAN; WANG, 2011; ZHAO et al., 2011) poderiam selecionar *loci* pela anotação GO no intuito de identificar marcadores microssatélites ligados aos interesses de estudos. Desse modo, a informação reunida, e conseqüentemente a plataforma *Pampeano Animal Genetic Resources*, fornecem um recurso em potencial para estudar a influencia de repetições microssatélite sobre características de interesse, além de proporcionar embasamento para fundamentar aplicação de marcadores moleculares em abordagens de seleção assistida (DEKKERS, 2004), em especial focando sobre genes candidatos (LECERF et al., 2011; ZHANG; ZAN; WANG, 2011; ZHAO et al., 2011). Ao mesmo tempo, a transferência dessa informação pode beneficiar espécies onde o nível de conhecimento é limitado (SERAPION et al., 2004; JU et al., 2005; AISHWARYA; GROVER; SHARMA, 2007; GGR-GOC, 2009; LEWIN, 2009; GOTZ et al., 2011), como zebuínos ou outros do gêneros *Bos*.

### 3.4 Conclusão

Grande proporção das repetições microssatélite mapeadas sobre genes transcritos ao longo do genoma bovino receberam anotação funcional baseada em ontologia gênica.

#### 4 Conclusões

A execução do presente trabalho permitiu identificar e descrever grande número de repetições microssatélites ao longo do genoma bovino taurino, caracterizando sua distribuição sobre regiões codificantes e não codificantes, além de preservar referências de mapeamento (anotação gênica, STS, SNPs e outros elementos repetitivos). Aliado a isso, uma importante parcela destas sequências recebeu anotação funcional baseada em ontologia gênica. Ao final, toda a informação reunida foi disponibilizada através da plataforma *Pampeano Animal Genetic Resources*, fornecendo a comunidade científica uma ferramenta em potencial para estudos sobre a contribuição desses trechos repetitivos ao genoma bovino taurino, bem como para a obtenção de marcadores moleculares destinados a seleção de características de interesse ao melhoramento genético da espécie.

## Referências

- ACHILLI, A.; BONFIGLIO, S.; OLIVIERI, A.; MALUSA, A.; PALA, M.; HOOSHIAR KASHANI, B.; PEREGO, U.A.; AJMONE-MARSAN, P.; LIOTTA, L.; SEMINO, O.; BANDELT, H.J.; FERRETTI, L.; TORRONI, A. The multifaceted origin of taurine cattle reflected by the mitochondrial genome. **PLoS One**, v. 4, n. 6, p.e5753, 2009.
- ADELSON, D.L.; RAISON, J.M.; EDGAR, R.C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. **Proceedings of the National Academy of Sciences USA**, v. 106, n. 31, p.12855-12860, 2009.
- ADELSON, D.L.; RAISON, J.M.; GARBER, M.; EDGAR, R.C. Interspersed repeats in the horse (*Equus caballus*); spatial correlations highlight conserved chromosomal domains. **Animal Genetics**, v. 41 Suppl 2, n. p.91-99, 2010.
- AISHWARYA, V.; GROVER, A.; SHARMA, P.C. EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. **BMC Genomics**, v. 8, n. p.225, 2007.
- ALBA, M.M.; GUIGO, R. Comparative analysis of amino acid repeats in rodents and humans. **Genome Research**, v. 14, n. 4, p.549-554, 2004.
- BARRETT, T.; SUZEK, T.O.; TROUP, D.B.; WILHITE, S.E.; NGAU, W.C.; LEDOUX, P.; RUDNEV, D.; LASH, A.E.; FUJIBUCHI, W.; EDGAR, R. NCBI GEO: mining millions of expression profiles--database and tools. **Nucleic Acids Research**, v. 33, n. Database issue, p.D562-566, 2005.
- BLUTHGEN, N.; BRAND, K.; CAJAVEC, B.; SWAT, M.; HERZEL, H.; BEULE, D. Biological profiling of gene groups utilizing Gene Ontology. **Genome Informatics**, v. 16, n. 1, p.106-115, 2005.
- BOLORMAA, S.; HAYES, B.J.; HAWKEN, R.J.; ZHANG, Y.; REVERTER, A.; GODDARD, M.E. Detection of chromosome segments of zebu and taurine origin and their effect on beef production and growth. **Journal of Animal Science**, v. 89, n. 7, p.2050-2060, 2011.
- BRANDSTROM, M.; ELLEGREN, H. Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. **Genome Research**, v. 18, n. 6, p.881-887, 2008.
- CHAN, E.K.; NAGARAJ, S.H.; REVERTER, A. The evolution of tropical adaptation: comparing taurine and zebu cattle. **Animal Genetics**, v. 41, n. 5, p.467-477, 2010.

CONESA, A.; GOTZ, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. **International Journal of Plant Genomics**, v. 2008, n. p.619832, 2008.

CONESA, A.; GOTZ, S.; GARCIA-GOMEZ, J.M.; TEROL, J.; TALON, M.; ROBLES, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p.3674-3676, 2005.

CRUZ, F.; ROUX, J.; ROBINSON-RECHAVI, M. The expansion of amino-acid repeats is not associated to adaptive evolution in mammalian genes. **BMC Genomics**, v. 10, n. p.619, 2009.

DEKKERS, J.C. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, v. 82 E-Suppl, n. p.E313-328, 2004.

EID, J.; FEHR, A.; GRAY, J.; LUONG, K.; LYLE, J.; OTTO, G.; PELUSO, P.; RANK, D.; BAYBAYAN, P.; BETTMAN, B.; BIBILLO, A.; BJORNSON, K.; CHAUDHURI, B.; CHRISTIANS, F.; CICERO, R.; CLARK, S.; DALAL, R.; DEWINTER, A.; DIXON, J.; FOQUET, M.; GAERTNER, A.; HARDENBOL, P.; HEINER, C.; HESTER, K.; HOLDEN, D.; KEARNS, G.; KONG, X.; KUSE, R.; LACROIX, Y.; LIN, S.; LUNDQUIST, P.; MA, C.; MARKS, P.; MAXHAM, M.; MURPHY, D.; PARK, I.; PHAM, T.; PHILLIPS, M.; ROY, J.; SEBRA, R.; SHEN, G.; SORENSON, J.; TOMANEY, A.; TRAVERS, K.; TRULSON, M.; VIECELI, J.; WEGENER, J.; WU, D.; YANG, A.; ZACCARIN, D.; ZHAO, P.; ZHONG, F.; KORLACH, J.; TURNER, S. Real-time DNA sequencing from single polymerase molecules. **Science**, v. 323, n. 5910, p.133-138, 2009.

ELSIK, C.G.; TELLAM, R.L.; WORLEY, K.C.; GIBBS, R.A.; MUZNY, D.M.; WEINSTOCK, G.M.; ADELSON, D.L.; EICHLER, E.E.; ELNITSKI, L.; GUIGO, R.; HAMERNIK, D.L.; KAPPES, S.M.; LEWIN, H.A.; LYNN, D.J.; NICHOLAS, F.W.; REYMOND, A.; RIJNKELS, M.; SKOW, L.C.; ZDOBNOV, E.M.; SCHOOK, L.; WOMACK, J.; ALIOTO, T.; ANTONARAKIS, S.E.; ASTASHYN, A.; CHAPPLE, C.E.; CHEN, H.C.; CHRAST, J.; CAMARA, F.; ERMOLAEVA, O.; HENRICHSEN, C.N.; HLAVINA, W.; KAPUSTIN, Y.; KIRYUTIN, B.; KITTS, P.; KOKOCINSKI, F.; LANDRUM, M.; MAGLOTT, D.; PRUITT, K.; SAPOJNIKOV, V.; SEARLE, S.M.; SOLOVYEV, V.; SOUVOROV, A.; UCLA, C.; WYSS, C.; ANZOLA, J.M.; GERLACH, D.; ELHAIK, E.; GRAUR, D.; REESE, J.T.; EDGAR, R.C.; MCEWAN, J.C.; PAYNE, G.M.; RAISON, J.M.; JUNIER, T.; KRIVENTSEVA, E.V.; EYRAS, E.; PLASS, M.;

DONTHU, R.; LARKIN, D.M.; REECY, J.; YANG, M.Q.; CHEN, L.; CHENG, Z.;  
CHITKO-MCKOWN, C.G.; LIU, G.E.; MATUKUMALLI, L.K.; SONG, J.; ZHU, B.;  
BRADLEY, D.G.; BRINKMAN, F.S.; LAU, L.P.; WHITESIDE, M.D.; WALKER, A.;  
WHEELER, T.T.; CASEY, T.; GERMAN, J.B.; LEMAY, D.G.; MAQBOOL, N.J.;  
MOLENAAR, A.J.; SEO, S.; STOTHARD, P.; BALDWIN, C.L.; BAXTER, R.;  
BRINKMEYER-LANGFORD, C.L.; BROWN, W.C.; CHILDERS, C.P.; CONNELLEY,  
T.; ELLIS, S.A.; FRITZ, K.; GLASS, E.J.; HERZIG, C.T.; IIVANAINEN, A.;  
LAHMERS, K.K.; BENNETT, A.K.; DICKENS, C.M.; GILBERT, J.G.; HAGEN, D.E.;  
SALIH, H.; AERTS, J.; CAETANO, A.R.; DALRYMPLE, B.; GARCIA, J.F.; GILL, C.A.;  
HIENDLEDER, S.G.; MEMILI, E.; SPURLOCK, D.; WILLIAMS, J.L.; ALEXANDER,  
L.; BROWNSTEIN, M.J.; GUAN, L.; HOLT, R.A.; JONES, S.J.; MARRA, M.A.;  
MOORE, R.; MOORE, S.S.; ROBERTS, A.; TANIGUCHI, M.; WATERMAN, R.C.;  
CHACKO, J.; CHANDRABOSE, M.M.; CREE, A.; DAO, M.D.; DINH, H.H.; GABISI,  
R.A.; HINES, S.; HUME, J.; JHANGIANI, S.N.; JOSHI, V.; KOVAR, C.L.; LEWIS,  
L.R.; LIU, Y.S.; LOPEZ, J.; MORGAN, M.B.; NGUYEN, N.B.; OKWUONU, G.O.;  
RUIZ, S.J.; SANTIBANEZ, J.; WRIGHT, R.A.; BUHAY, C.; DING, Y.; DUGAN-  
ROCHA, S.; HERDANDEZ, J.; HOLDER, M.; SABO, A.; EGAN, A.; GOODELL, J.;  
WILCZEK-BONEY, K.; FOWLER, G.R.; HITCHENS, M.E.; LOZADO, R.J.; MOEN,  
C.; STEFFEN, D.; WARREN, J.T.; ZHANG, J.; CHIU, R.; SCHEIN, J.E.; DURBIN,  
K.J.; HAVLAK, P.; JIANG, H.; LIU, Y.; QIN, X.; REN, Y.; SHEN, Y.; SONG, H.; BELL,  
S.N.; DAVIS, C.; JOHNSON, A.J.; LEE, S.; NAZARETH, L.V.; PATEL, B.M.; PU, L.L.;  
VATTATHIL, S.; WILLIAMS, R.L., JR.; CURRY, S.; HAMILTON, C.; SODERGREN,  
E.; WHEELER, D.A.; BARRIS, W.; BENNETT, G.L.; EGGEN, A.; GREEN, R.D.;  
HARHAY, G.P.; HOBBS, M.; JANN, O.; KEELE, J.W.; KENT, M.P.; LIEN, S.;  
MCKAY, S.D.; MCWILLIAM, S.; RATNAKUMAR, A.; SCHNABEL, R.D.; SMITH, T.;  
SNELLING, W.M.; SONSTEGARD, T.S.; STONE, R.T.; SUGIMOTO, Y.;  
TAKASUGA, A.; TAYLOR, J.F.; VAN TASSELL, C.P.; MACNEIL, M.D.;  
ABATEPAULO, A.R.; ABBEY, C.A.; AHOLA, V.; ALMEIDA, I.G.; AMADIO, A.F.;  
ANATRIELLO, E.; BAHADUE, S.M.; BIASE, F.H.; BOLDT, C.R.; CARROLL, J.A.;  
CARVALHO, W.A.; CERVELATTI, E.P.; CHACKO, E.; CHAPIN, J.E.; CHENG, Y.;  
CHOI, J.; COLLEY, A.J.; DE CAMPOS, T.A.; DE DONATO, M.; SANTOS, I.K.; DE  
OLIVEIRA, C.J.; DEOBALD, H.; DEVINOY, E.; DONOHUE, K.E.; DOVC, P.;  
EBERLEIN, A.; FITZSIMMONS, C.J.; FRANZIN, A.M.; GARCIA, G.R.; GENINI, S.;  
GLADNEY, C.J.; GRANT, J.R.; GREASER, M.L.; GREEN, J.A.; HADSELL, D.L.;

HAKIMOV, H.A.; HALGREN, R.; HARROW, J.L.; HART, E.A.; HASTINGS, N.; HERNANDEZ, M.; HU, Z.L.; INGHAM, A.; ISO-TOURU, T.; JAMIS, C.; JENSEN, K.; KAPETIS, D.; KERR, T.; KHALIL, S.S.; KHATIB, H.; KOLBEHDARI, D.; KUMAR, C.G.; KUMAR, D.; LEACH, R.; LEE, J.C.; LI, C.; LOGAN, K.M.; MALINVERNI, R.; MARQUES, E.; MARTIN, W.F.; MARTINS, N.F.; MARUYAMA, S.R.; MAZZA, R.; MCLEAN, K.L.; MEDRANO, J.F.; MORENO, B.T.; MORE, D.D.; MUNTEAN, C.T.; NANDAKUMAR, H.P.; NOGUEIRA, M.F.; OLSAKER, I.; PANT, S.D.; PANZITTA, F.; PASTOR, R.C.; POLI, M.A.; POSLUSNY, N.; RACHAGANI, S.; RANGANATHAN, S.; RAZPET, A.; RIGGS, P.K.; RINCON, G.; RODRIGUEZ-OSORIO, N.; RODRIGUEZ-ZAS, S.L.; ROMERO, N.E.; ROSENWALD, A.; SANDO, L.; SCHMUTZ, S.M.; SHEN, L.; SHERMAN, L.; SOUTHEY, B.R.; LUTZOW, Y.S.; SWEEDLER, J.V.; TAMMEN, I.; TELUGU, B.P.; URBANSKI, J.M.; UTSUNOMIYA, Y.T.; VERSCHOOR, C.P.; WAARDENBERG, A.J.; WANG, Z.; WARD, R.; WEIKARD, R.; WELSH, T.H., JR.; WHITE, S.N.; WILMING, L.G.; WUNDERLICH, K.R.; YANG, J.; ZHAO, F.Q. The genome sequence of taurine cattle: a window to ruminant biology and evolution.

**Science**, v. 324, n. 5926, p.522-528, 2009.

FAN, H.; CHU, J.Y. A brief review of short tandem repeat mutation. **Genomics Proteomics Bioinformatics**, v. 5, n. 1, p.7-14, 2007.

FINUCANE, K.A.; MCFADDEN, T.B.; BOND, J.P.; KENNELLY, J.J.; ZHAO, F.Q. Onset of lactation in the bovine mammary gland: gene expression profiling indicates a strong inhibition of gene expression in cell proliferation. **Functional & Integrative Genomics**, v. 8, n. 3, p.251-264, 2008.

GALLA, G.; BARCACCIA, G.; RAMINA, A.; COLLANI, S.; ALAGNA, F.; BALDONI, L.; CULTRERA, N.G.; MARTINELLI, F.; SEBASTIANI, L.; TONUTTI, P. Computational annotation of genes differentially expressed along olive fruit development. **BMC Plant Biology**, v. 9, n. p.128, 2009.

GAO, H.; CAI, S.; YAN, B.; CHEN, B.; YU, F. Discrepancy variation of dinucleotide microsatellite repeats in eukaryotic genomes. **Biological Research**, v. 42, n. 3, p.365-375, 2009.

GGR-GOC / Reference Genome Group of the Gene Ontology Consortium. The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. **PLoS Computational Biology**, v. 5, n. 7, p.e1000431, 2009.

GIBBS, R.A.; TAYLOR, J.F.; VAN TASSELL, C.P.; BARENDSE, W.; EVERSOLE, K.A.; GILL, C.A.; GREEN, R.D.; HAMERNIK, D.L.; KAPPES, S.M.; LIEN, S.;

MATUKUMALLI, L.K.; MCEWAN, J.C.; NAZARETH, L.V.; SCHNABEL, R.D.; WEINSTOCK, G.M.; WHEELER, D.A.; AJMONE-MARSAN, P.; BOETTCHER, P.J.; CAETANO, A.R.; GARCIA, J.F.; HANOTTE, O.; MARIANI, P.; SKOW, L.C.; SONSTEGARD, T.S.; WILLIAMS, J.L.; DIALLO, B.; HAILEMARIAM, L.; MARTINEZ, M.L.; MORRIS, C.A.; SILVA, L.O.; SPELMAN, R.J.; MULATU, W.; ZHAO, K.; ABBEY, C.A.; AGABA, M.; ARAUJO, F.R.; BUNCH, R.J.; BURTON, J.; GORNI, C.; OLIVIER, H.; HARRISON, B.E.; LUFF, B.; MACHADO, M.A.; MWAKAYA, J.; PLASTOW, G.; SIM, W.; SMITH, T.; THOMAS, M.B.; VALENTINI, A.; WILLIAMS, P.; WOMACK, J.; WOOLLIAMS, J.A.; LIU, Y.; QIN, X.; WORLEY, K.C.; GAO, C.; JIANG, H.; MOORE, S.S.; REN, Y.; SONG, X.Z.; BUSTAMANTE, C.D.; HERNANDEZ, R.D.; MUZNY, D.M.; PATIL, S.; SAN LUCAS, A.; FU, Q.; KENT, M.P.; VEGA, R.; MATUKUMALLI, A.; MCWILLIAM, S.; SCLEP, G.; BRYC, K.; CHOI, J.; GAO, H.; GREFENSTETTE, J.J.; MURDOCH, B.; STELLA, A.; VILLA-ANGULO, R.; WRIGHT, M.; AERTS, J.; JANN, O.; NEGRINI, R.; GODDARD, M.E.; HAYES, B.J.; BRADLEY, D.G.; BARBOSA DA SILVA, M.; LAU, L.P.; LIU, G.E.; LYNN, D.J.; PANZITTA, F.; DODDS, K.G. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. **Science**, v. 324, n. 5926, p.528-532, 2009.

GLENN, T.C.; STEPHAN, W.; DESSAUER, H.C.; BRAUN, M.J. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. **Molecular Biology and Evolution**, v. 13, n. 8, p.1151-1154, 1996.

GODDARD, M.E.; HAYES, B.J. Genomic selection. **Journal of Animal Breeding and Genetics**, v. 124, n. 6, p.323-330, 2007.

GODDARD, M.E.; HAYES, B.J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Review Genetics**, v. 10, n. 6, p.381-391, 2009.

GOTZ, S.; ARNOLD, R.; SEBASTIAN-LEON, P.; MARTIN-RODRIGUEZ, S.; TISCHLER, P.; JEHL, M.A.; DOPAZO, J.; RATTEI, T.; CONESA, A. B2G-FAR, a species-centered GO annotation repository. **Bioinformatics**, v. 27, n. 7, p.919-924, 2011.

GOTZ, S.; GARCIA-GOMEZ, J.M.; TEROL, J.; WILLIAMS, T.D.; NAGARAJ, S.H.; NUEDA, M.J.; ROBLES, M.; TALON, M.; DOPAZO, J.; CONESA, A. High-throughput functional annotation and data mining with the Blast2GO suite. **Nucleic Acids Research**, v. 36, n. 10, p.3420-3435, 2008.

- HANSEN, P.J. Physiological and cellular adaptations of zebu cattle to thermal stress. **Animal Reproduction Science**, v. 82-83, n. p.349-360, 2004.
- HIENDLEDER, S.; LEWALSKI, H.; JANKE, A. Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication. **Cytogenetic and Genome Research**, v. 120, n. 1-2, p.150-156, 2008.
- HUGHES, L.M.; BAO, J.; HU, Z.L.; HONAVAR, V.; REECY, J.M. Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. **Journal of Animal Science**, v. 86, n. 6, p.1485-1491, 2008.
- JOSÉ, M. Liderança assediada. **Anuário DBO**, São Paulo, v. 30, n. 375, p.6-7, 2012.
- JU, Z.; WELLS, M.C.; MARTINEZ, A.; HAZLEWOOD, L.; WALTER, R.B. An in silico mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, *Fundulus*, and *Xiphophorus*. **In Silico Biology**, v. 5, n. 5-6, p.439-463, 2005.
- KASHI, Y.; KING, D.G. Simple sequence repeats as advantageous mutators in evolution. **Trends in Genetics**, v. 22, n. 5, p.253-259, 2006.
- KIM, T.H.; KIM, N.S.; LIM, D.; LEE, K.T.; OH, J.H.; PARK, H.S.; JANG, G.W.; KIM, H.Y.; JEON, M.; CHOI, B.H.; LEE, H.Y.; CHUNG, H.Y.; KIM, H. Generation and analysis of large-scale expressed sequence tags (ESTs) from a full-length enriched cDNA library of porcine backfat tissue. **BMC Genomics**, v. 7, n. p.36, 2006.
- KRUGLYAK, S.; DURRETT, R.T.; SCHUG, M.D.; AQUADRO, C.F. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. **Proceedings of the National Academy of Sciences USA**, v. 95, n. 18, p.10774-10778, 1998.
- KUMAR, S.; HEDGES, S.B. Amolecular timescale for vertebrate evolution. **Nature**, v. 392, n. p.917-920, 1998.
- LARKIN, D.M.; EVERTS-VAN DER WIND, A.; REBEIZ, M.; SCHWEITZER, P.A.; BACHMAN, S.; GREEN, C.; WRIGHT, C.L.; CAMPOS, E.J.; BENSON, L.D.; EDWARDS, J.; LIU, L.; OSOEGAWA, K.; WOMACK, J.E.; DE JONG, P.J.; LEWIN, H.A. A cattle-human comparative map built with cattle BAC-ends and human genome sequence. **Genome Research**, v. 13, n. 8, p.1966-1972, 2003.
- LECERF, F.; BRETAUDEAU, A.; SALLOU, O.; DESERT, C.; BLUM, Y.; LAGARRIGUE, S.; DEMEURE, O. AnnotQTL: a new tool to gather functional and



comparative information on a genomic region. **Nucleic Acids Research**, v. 39, n. Web Server issue, p.W328-333, 2011.

LEWIN, H.A. Genetics. It's a bull's market. **Science**, v. 324, n. 5926, p.478-479, 2009.

LI, Y.C.; KOROL, A.B.; FAHIMA, T.; BEILES, A.; NEVO, E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. **Molecular Ecology**, v. 11, n. 12, p.2453-2465, 2002.

MAIA, L.C.; DE SOUZA, V.Q.; KOPP, M.M.; DE CARVALHO, F.I.; DE OLIVEIRA, A.C. Tandem repeat distribution of gene transcripts in three plant families. **Genetics and Molecular Biology**, v. 32, n. 4, p.822-833, 2009.

MAIA, L.C.; PALMIERI, D.A.; DE SOUZA, V.Q.; KOPP, M.M.; DE CARVALHO, F.I.; COSTA DE OLIVEIRA, A. SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. **International Journal of Plant Genomics**, v. 2008, n. p.412696, 2008.

MATUKUMALLI, L.K.; LAWLEY, C.T.; SCHNABEL, R.D.; TAYLOR, J.F.; ALLAN, M.F.; HEATON, M.P.; O'CONNELL, J.; MOORE, S.S.; SMITH, T.P.; SONSTEGARD, T.S.; VAN TASSELL, C.P. Development and characterization of a high density SNP genotyping assay for cattle. **PLoS One**, v. 4, n. 4, p.e5350, 2009.

MCCARTHY, F.M.; WANG, N.; MAGEE, G.B.; NANDURI, B.; LAWRENCE, M.L.; CAMON, E.B.; BARRELL, D.G.; HILL, D.P.; DOLAN, M.E.; WILLIAMS, W.P.; LUTHE, D.S.; BRIDGES, S.M.; BURGESS, S.C. AgBase: a functional genomics resource for agriculture. **BMC Genomics**, v. 7, n. p.229, 2006.

MORGANTE, M.; HANAFEY, M.; POWELL, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. **Nature Genetics**, v. 30, n. 2, p.194-200, 2002.

OZDEMIR OZGENTURK, N.; ORUC, F.; SEZERMAN, U.; KUCUKURAL, A.; VURAL KORKUT, S.; TOKSOZ, F.; UN, C. Generation and Analysis of Expressed Sequence Tags from *Olea europaea* L. **Comparative and Functional Genomics**, v. 2010, n. p.757512, 2010.

POLL, H.; VENCATO, A.Z.; KIST, B.B.; SANTOS, C.; REETZ, E.R.; BELING, R.R. Grande como o Brasil. **Anuário Brasileiro da Pecuária**, Tubarão, v. n. p.19, 2011a. POLL, H.; VENCATO, A.Z.; KIST, B.B.; SANTOS, C.; REETZ, E.R.; BELING, R.R. A passo lento. **Anuário Brasileiro da Pecuária**, Tubarão, v. n. 25, p.35, 2011b.

- POLL, H.; VENCATO, A.Z.; KIST, B.B.; SANTOS, C.; REETZ, E.R.; BELING, R.R. Segredos a desvendar. **Anuário Brasileiro da Pecuária**, Tubarão, v. n. p.97, 2011c.
- RDCT / **R Development Core Team R: A language and environment for statistical computing**. Vienna, Austria: R Foundation for Statistical Computing, 2011.
- SERAPION, J.; KUCUKTAS, H.; FENG, J.; LIU, Z. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). **Marine Biotechnology (NY)**, v. 6, n. 4, p.364-377, 2004.
- SUBRAMANIAN, S.; MISHRA, R.K.; SINGH, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. **Genome Biology**, v. 4, n. 2, p.R13, 2003.
- SVENSSON, O.; ARVESTAD, L.; LAGERGREN, J. Genome-wide survey for biologically functional pseudogenes. **PLoS Computational Biology**, v. 2, n. 5, p.e46, 2006.
- TAN, S.H.; REVERTER, A.; WANG, Y.; BYRNE, K.A.; MCWILLIAM, S.M.; LEHNERT, S.A. Gene expression profiling of bovine in vitro adipogenesis using a cDNA microarray. **Functional & Integrative Genomics**, v. 6, n. 3, p.235-249, 2006.
- TAUTZ, D. Notes on the definition and nomenclature of tandemly repetitive DNA sequences. **EXS**, v. 67, n. p.21-28, 1993.
- TELLAM, R.L.; LEMAY, D.G.; VAN TASSELL, C.P.; LEWIN, H.A.; WORLEY, K.C.; ELSIK, C.G. Unlocking the bovine genome. **BMC Genomics**, v. 10, n. p.193, 2009.
- TEMNYKH, S.; DECLERCK, G.; LUKASHOVA, A.; LIPOVICH, L.; CARTINHO, S.; MCCOUCH, S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. **Genome Research**, v. 11, n. 8, p.1441-1452, 2001.
- VICTORIA, F.C.; MAIA, L.C.; OLIVEIRA, A.C. In silico comparative analysis of SSR markers in plants. **BMC Plant Biology**, v. 11, n. p.15, 2011.
- YAN, Q.; ZHANG, Y.; LI, H.; WEI, C.; NIU, L.; GUAN, S.; LI, S.; DU, L. Identification of microsatellites in cattle unigenes. **Journal of Genetics and Genomics**, v. 35, n. 5, p.261-266, 2008.
- ZHAN, A.; WANG, Y.; BROWN, B.; WANG, H.P. Isolation and characterization of novel microsatellite markers for yellow perch (*Perca flavescens*). **International Journal of Molecular Sciences**, v. 10, n. 1, p.18-27, 2009.

ZHANG, W.; WANG, Z.; ZHAO, Z.; ZENG, X.; WU, H.; YU, P. Using bioinformatics methods to develop EST-SSR markers from Sheep's ESTs. **Journal of Animal and Veterinary Advances**, v. 9, n. 21, p.2759-2762, 2010.

ZHANG, Y.; ZAN, L.; WANG, H. Screening candidate genes related to tenderness trait in Qinchuan cattle by genome array. **Molecular Biology Reports**, v. 38, n. 3, p.2007-2014, 2011.

ZHAO, C.; TIAN, F.; YU, Y.; LUO, J.; HU, Q.; BEQUETTE, B.J.; BALDWIN VI, R.L.; LIU, G.; ZAN, L.; SCOTT UPDIKE, M.; SONG, J. Muscle transcriptomic analyses in Angus cattle with divergent tenderness. **Molecular Biology Reports**, v. n. 2011.

ZHOU, X.; SU, Z. EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. **BMC Genomics**, v. 8, n. p.246, 2007.