UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA
DISSERTAÇÃO DE MESTRADO

Montagem, Anotação e Análises Comparativas dos Genomas Mitocondriais de Animais Representantes das Raças Bovinas: Gir e Guzerá e o Desafio da Montagem *De novo* do Genoma Nuclear dessas duas Raças Usando Sequenciamento de Nova Geração

Por:

Juliana Assis Geraldo

Orientador: Dr. Guilherme Corrêa de Oliveira

Belo Horizonte, Março de 2015

Juliana Assis Geraldo

Montagem, Anotação e Análises Comparativas dos Genomas Mitocondriais de Animais Representantes das Raças Bovinas: Gir e Guzerá e o Desafio da Montagem *De novo* do Genoma Nuclear dessas duas Raças Usando Sequenciamento de Nova Geração

Dissertação apresentada ao Programa de Pós Graduação em

Bioinformática da Universidade Federal de Minas

Gerais como requisito parcial a obtenção do título de

Mestre em Bioinformática.

ÁREA DE CONCENTRAÇÃO: BIOINFORMÁTICA GENÔMICA

Orientador: Dr. Guilherme Corrêa de Oliveira

Belo Horizonte, Março de 2015

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE DISSERTAÇÃO

5/2015
entrada
1º/2013
CPF:
093.400.836-10

Juliana Assis Geraldo

Às nove horas do dia **12 de março de 2015**, 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: **"Montagem, Anotação e Análises comparativas dos Genomas Mito condriais de animais representantes das Raças bovinas: Gir e Guzerá e o Desafio da Montagem De novo do Genoma Nuclear dessas duas Raças Usando Sequenciamento de Nova Geração"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Guilherme Correa de Oliveira**, 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
Dr. Guilherme Correa de Oliveira	FIOCRUZ/CPqRR	686551186 72	APROVADA
Dr. Lucas Bleicher	UFMG	813528 593-00	APROVADO
Dr. Henrique Cesar Pereira Figueiredo	UFMG	952 711.716-04	APROVADO

Pelas indicações, a candidata foi considerada: APROVADA
O resultado final foi comunicado publicamente à candidata pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 12 de março de 2015.

Dr. Guilherme Correa de Oliveira - Orientador _____

Dr. Lucas Bleicher _____

Dr. Henrique Cesar Pereira Figueiredo _____

Esse trabalho foi desenvolvido no Centro de Pesquisas René Rachou – CPqRR - FIOCRUZ em colaboração com a Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) - Gado de Leite. Parte do trabalho foi desenvolvido na *University of Georgia*, durante estágio realizado no período de fevereiro de 2014 a junho de 2014, em colaboração com o grupo da Dra. Jéssica Kissinger (*Center for Tropical and Emerging Global Diseases* da *University of Georgia* (CTEGD-UGA-EUA). Posteriormente, foi concluído no CPqRR, sob a orientação do Dr. Guilherme Oliveira. O projeto conta com o suporte financeiro da Fundação de Amparo do Estado de Minas Gerais (FAPEMIG).

AGRADECIMENTOS

Agradeço ao meu orientador Dr. Guilherme Oliveira. Boss! Obrigada pela orientação, pelos ensinamentos, incentivos, cobranças, questionamentos, puxões de orelha, pela confiança, mas, principalmente por todas as oportunidades de crescimento que você me proporcionou. Sou muito feliz por isso, e espero poder retribuir contribuindo para o crescimento do grupo.

Agradeço a Dra. Jéssica Kissinger por me acolher tão bem em seu laboratório, sua casa, seu país. Agradeço mais ainda por todas as inquietações, questionamentos durante o trabalho. Sem seus ensinamentos esse trabalho não seria possível!

A Embrapa Gado de Leite pela autorização de desenvolvimento desse trabalho, sob a coordenação do Dr. Marcos Vinícos Gualberto.

Aos doutores Magdi e Raji da Universidade da Geórgia, por todos os ensinamentos sobre montagem de grandes genomas.

A Dra. Ângela Volpini, Eliane Moura, deixo meu muito obrigada! Vocês são inenarráveis.

Betsy Winter, não tenho palavras para agradecer tudo o que você fez por mim, antes, durante e após minha estadia nos EUA. Foi um imenso prazer te conhecer e trabalhar com você. Obrigada pela amizade e momentos de descontração!

A Dr. Maria Raquel pelo suporte durante o desenvolvimento desse trabalho.

Agradeço ao Programa de Pós Graduação em Bioinformática da Universidade Federal de Minas Gerais. Meu muito obrigada a todos os professores do curso e também á todos da secretaria e coordenação

A Luiza Andrade e Marina Mourão, minhas eternas chefas!

Izinara Rosse, por todos os ensinamentos bovólogos, biológicos etc. Agradeço por toda a ajuda, você foi essencial nesse trabalho. É uma honra poder trabalhar com você! Agradeço e muito pela amizade e por todos os momentos de descontração no Cebio e nas nossas aventuras por aí. =p

Anderson Dominiti e Fausto, por todo suporte em TI. Mariana Maga por toda parte burocrática, Mariana Oliveira por toda criação e talento como webdesigner. E agradeço a todos pela amizade e carinho!

Luis Martinez, a mente mais brilhante do grupo. Agradeço pelas ideias e por não me deixar chutar o balde em algumas horas! haha

Francislon o futuro da bioinformática no mundo! Muito obrigada pelos ensinamentos de programação pra grandes dados! Obrigada por corrigir meus programas, rs, e por desenvolver várias vezes em que estava desesperada com os erros! Você é o cara!

Agradeço a Laura Leite por todas as conversas e discussões sobre avanços em bioinformática e por todo suporte na utilização de ferramentas. Yesid pelo suporte nas análises filogenéticas.

Larissa por todas as conversas!

Deixo meu muito obrigado á todos do Cebio por toda companhia, amizade, ajuda e principalmente por me aguentarem todos os dias! Haha. Mesmo vocês roubando meus biscoitos eu gosto muito de vocês!

Gabriel, pelas sugestões e revisão.

Agradeço a todos do laboratório da Jéssica Kissinger, em especial Ousman, Ranjani, Jeremy, por toda paciência, amizade e ensinamentos (principalmente no inglês).

Patchara e Juntiwana, minhas roomates! Agradeço pelos ensinamentos de como preparar uma belíssima comida tailandesa! Agradeço por terem tornado minha estadia nos EUA mais feliz!

A banca examinadora por ter aceitado o convite para avaliar e ajudar a finalizar esse trabalho

Às Agências Financiadoras que contribuíram para a realização deste trabalho: National Institutes of Health - NIH/Fogarty International Center, FAPEMIG e a CAPES pela minha bolsa de mestrado.

Agradeço aos meus pais e minhas irmãs, mesmo que vocês não compreendam nada do que eu faço e acham que eu passo o dia inteiro jogando joguinhos na tela preta do meu computador rosa, eu amo vocês!

A ciência nunca resolve um problema sem criar pelo menos outros dez".

Shaw, Bernard

Sumário

LISTA DE TABELAS

LISTA DE FIGURAS

LISTA DE SIGLAS

Ala - Alanina
Arg - Arginina
Asn - Asparagina
Asp - Aspartato
CDS - Conding DNA sequencing (Sequência codificadora do DNA)
COI - Citocromo Oxidase subunidade I
COII - Citocromo Oxidase II
COIII- Citocromo oxidase subunidade 3
CYB - Citocromo B
Cys - Cisteína
Gln - Glutamina
Glu - Glutamato
Gly - Glicina
His - Histidina
Ile - Isoleucina
Kb - kilobases (1000 pares de bases de DNA ou RNA)
Leu1, Leu2 – Leucina 1 e Leucina 2
Lys - Lisina
m.a.a. - milhões de anos atrás
Mb - megabases (1.000.000 de pares de bases de DNA ou RNA)
Met - Metionina
mtDNA - Genoma Mitocondrial
NCBI - *National Center for Biotechnology Information*
ND1- NADH Desidrogenase Subunidade 1
ND3- NADH Desidrogenase Subunidade 3
ND4- NADH Desidrogenase Subunidade 4
ND5- NADH Desidrogenase Subunidade 5
ND7- NADH Desidrogenase Subunidade 7
ND8 - NADH Desidrogenase Subunidade 8
ND9 - NADH Desidrogenase Subunidade 9
nDNA - Genoma Nuclear
Pb - pares de bases
Phe - Fenilalanina
Pro - Prolina
rRNAs: 12S e 16S – Ácido ribonucleico ribossomal, subunidades 12 e 16.
Ser1, Ser2 – Serina 1 e Serina 2
Thr - Treonina
Trp - Triptofano
tRNAs - RNA transportador
Tyr - Tirosina
Val - Valina

RESUMO

Com o objetivo de melhorarmos o entendimento molecular do genoma mitocondrial (mtDNA) das atuais raças bovinas, nesse trabalho foram montados e anotados quatro mtDNA de duas raças zebuínas de grande importância na produção de leite no Brasil: Guzerá e Gir. Por meio das análises comparativas com outros genomas bovinos foi possível identificar em dois dos genomas montados o mtDNA de origem taurina e dois de origem zebuína. Alterações das funções celulares provenientes das variações entre os mitogenomas não foram encontradas. Com a reconstrução filogenética foi possível classificar os animais Gir 1 e Guzerá 3 no haplogrupo I2, tendo o subcontinente indiano como provável ponto de origem de domesticação desses animais. O indivíduo Gir 2 foi classificado no haplogrupo T3 tendo sua origem de domesticação no Oriente Próximo e o indivíduo Guzerá 4 foi classificado no haplogrupo T1c "*Africa-derived-American*". Os valores de apoio estatístico das árvores geradas com a sequência mitocondrial completa se mostraram mais robustos em relação às demais árvores geradas (região hipervariável D-loop e todos os genes codificadores de proteínas concatenados). Esse fato nos indicou a necessidade da utilização de uma sequência mitocondrial ampla, cobrindo regiões codificadoras e não codificadoras, de modo a obter valores de apoio estatísticos maiores, a fim de garantir a confiabilidade das conclusões sobre a história evolutiva das raças que tais análises nos proporcionam. As sequências mitocondriais completas serão depositadas em bancos de dados públicos para contribuir com trabalhos futuros de genética bovina. Além disso, no presente estudo foi realizado um projeto piloto de montagem *de novo* do genoma nuclear de seis animais zebuínos das raças Gir e Guzerá. Nesta estratégia foram utilizados dados de Sequenciamento de nova geração de quatro diferentes plataformas (SOLiD, Miseq, PacBio e Hiseq). Mesmo não tendo concluido um *draft* do genoma, conseguimos trazer boas contribuições para trabalhar com grandes genomas bovinos. Através dos resultados foi possível concluir que as *reads* trimadas com alto valor de qualidade aumentam a confiabilidade dos dados e ajuda na redução da fragmentação da montagem. A utilização de diferentes plataformas nos permitiu concluir que as *reads* oriundas do Hiseq são as mais indicadas para trabalhar com genomas complexos como o de bovinos. Desta forma, o presente trabalho, fornece um modelo de delineamento experimental e idealização dos dados que poderão ser utilizadas em projetos futuros, além de apontarem a direção para se concluir a montagem *de novo* do genoma das raças Gir e Guzerá.

ABSTRACT

This study aimed at assembling the mitochondrial genome (mtDNA) of two Asian cattle breeds (*Bos indicus*), Guzerá and Gir. Both of these breeds are the main milk production in Brazil. We assembled the mtDNA of both breeds to improve the understanding of the mtDNA molecular diversity within the *Bos* genus. Comparative analysis with other cattle genomes was performed and we found two of the genomes assembled, carry the mtDNA of taurine origin and two the zebu origin. Changes in cellular functions from the variations between mtDNAs were not found. Through phylogenetic reconstruction was possible to classify the animals Gir 1 and 3 in Guzerá haplogroup I2, and the Indian subcontinent as probable point of origin of domestication of animals. The Gir 2 individual was classified as haplogroup T3 having its domestication originated in the Near East and the individual Guzerá 4 was classified as haplogroup T1c "Africa-derived-American." The statistical support values of the trees generated with the complete mitochondrial sequence were more robust compared to other trees generated (hypervariable region D-loop and all the genes encoding proteins concatenated). The complete mitochondrial sequences will be deposited in public databases to contribute to future work of bovine genetics. Furthermore, in this study we also work with the pilot project for assembly a nuclear genome from six animals (three Gir and three Guzerá). For this strategy was used a Next Generation Sequencing of four different platforms (SOLiD, Miseq, PacBio and Hiseq). The HiSeq platform was the most suitable for working with complex genomes such as cattle. This study provides a model of experimental design and idealization of data that could be used in future projects, and also point out the direction to complete the new genome assembly to Gir and Guzerá breeds.

I- INTRODUÇÃO GERAL

Esse trabalho apresenta a montagem e anotação dos genomas mitocondriais de quatro animais representantes de duas raças de bovinos leiteiros pertencentes ao rebanho brasileiro. Através de uma abordagem filogenética o presente estudo conta uma história evolutiva sobre o relacionamento desses animais com os demais bovinos.

Essa dissertação ainda traz o enfoque da montagem *de novo* do genoma nuclear bovino por meio do sequenciamento de nova geração.

O presente trabalho é apresentado em duas partes divididas da seguinte forma:

Capítulo 1: Montagem, anotação análises comparativas e estudos evolutivos do genoma mitocondrial bovino. Nessa primeira parte é apresentado o trabalho no qual foi realizada a montagem, anotação e análises comparativas dos genomas mitocondriais de quatro indivíduos das raças bovinas Gir e Guzerá. O estudo será submetido à revista Mitochondrial DNA.

Capítulo 2: Montagem *de novo* do genoma nuclear. Nessa parte serão expostas as tentativas de montagens *de novo* dos genomas bovinos de seis animais pertencentes as raças Gir e Guzerá, descrevendo tudo o que se pode alcançar com os dados presentes. Esse estudo é um projeto piloto da montagem dos genomas dessas duas raças bovinas em que foi possível avaliar e estabelecer diferentes estratégias para a melhor montagem desses genomas.

Na Fundamentação Teórica será apresentada uma breve introdução das raças bovinas estudadas e ao final do trabalho será mostrada uma breve consideração de ambos os capítulos.

Todos os trabalhos desenvolvidos em paralelo a este, bem como participações em congressos, cursos e estágios são apresentados na seção anexos.

1.1 FUNDAMENTAÇÃO TEÓRICA

Acredita-se que a maior parte dos bovinos do mundo seja constituída por duas subespécies: *Bos taurus* (taurinos) *e Bos indicus* (zebuínos) [LOFTUS *et al*., 1994]. Levando-se em consideração o conceito biológico de espécie proposto por Mayr em 1963, ambos podem ser considerados subespécies, visto que os descendentes entre taurinos e zebuínos apresentam completa fertilidade [LOFTUS *et al*., 1994, HIENDLEDER, *et al*., 2008].

As principais características morfológicas as quais permitem a separação dos taurinos de zebuínos são a presença de cupim e barbela grande (zebu) ou ausência (taurus).

A complexidade de classificação bovina não para na subdivisão de espécies/subespécies, ela é fortemente agravada por se tratar de animais domesticados pelo homem e deve-se acrescentar devido a esse fator a complexidade de formação das raças.

Estima-se que hoje existam aproximadamente 800 diferentes raças bovinas em todo o mundo [ELSIK *et al*., 2009]. As diferenças entre as raças não são tão evidentes como entre espécies, podendo ser bem mais sutis. As raças foram originadas há relativamente pouco tempo a partir de um conjunto de genes comuns e o isolamento genético raramente é existente [FELIUS *et al*., 2011]. Várias raças são mais diferentes atualmente do que era há apenas 20 anos. Na verdade, a criação seletiva tem acelerado a evolução dos bovinos até o ponto que os dois últimos séculos viram-se mais mudanças na aparência e na produção do que nos milênios anteriores [GARCIA *et al*., 2010].

As múltiplas origens das raças só tornam mais complexo os estudos filogenéticos entre bovinos, pois estas são formadas entre os cruzamentos entre raças já existentes (sejam elas taurinas ou zebuínas), sempre no intuito do melhoramento das características de produção, seja ela de leite, carne ou ambos. [GARCIA *et al*., 2010].

1.1.1 As raças estudadas: Gir e Guzerá

Esse presente trabalho teve como alvo de estudo duas raças zebuínas: Gir e Guzerá. Ambas são de grande importância para a formação do rebanho brasileiro tanto para a produção de leite quanto para a produção de carne.

Em uma breve comparação aos bovinos europeus (*Bos taurus*), os zebuínos têm mais glândulas sudoríparas e são capazes de lidar com climas bem quentes e úmidos. Além disso, o zebu apresenta uma maior resistência a pragas (tais como carrapatos) do que taurinos [BAIG et al., 2005]. Os zebuínos são os principais componentes do rebanho brasileiro, onde cerca de 75% do rebanho é composto por animais com sangue zebuíno, tanto puros quanto mestiços.

Uma pequena descrição das raças é apresentada a seguir, sendo as informações provenientes da página da Associação Brasileira dos Criadores de Zebu [http://www.abcz.org.br/].

A raça Gir é originária da Índia, das regiões de Gir na Península de Kathiawar e foi trazida para o Brasil no início do século XIX. É uma raça muito bem adaptada ao clima tropical, ocupando o segundo lugar no controle leiteiro oficial no Brasil. Neste trabalho os animais sequenciados são para a produção de leite, também denominados Gir-leiteiro.

A raça guzerá é uma raça originária do subcontinente indiano, tendo sido introduzido no Brasil no século XVIII. É uma raça com aspectos de rusticidade, habilidade materna, grande fertilidade e de dupla aptidão, ou seja, pode ser utilizada para a produção de leite e a maioria selecionada para corte. Entretanto, neste trabalho os animais são touros selecionados para as características de produção de leite.

Visto que a maior parte do plantel brasileiro é composta de zebuínos e as raças leiteiras Gir e Guzerá são de grande importância para a formação do rebanho bovino brasileiro, a EMBRAPA em parceria com outras instituições vêm desenvolvendo o programa de melhoramento genético dessas raças. Um ganho genético de 1% ao ano vem sendo alcançado, porém a maior parte dos marcadores utilizados nesses programas são identificados em raça taurinas. Esse fato é devido ao genoma das raças zebuínas leiteiras ainda estarem disponíveis em pequenas quantidades e as raças Gir e Guzerá ainda não estarem com os genomas montados e disponíveis para inclusão nesses chips. Nesse contexto, o objetivo do capítulo dois do presente trabalho é iniciar o projeto de montagem *de* novo do genoma nuclear dessas duas raças bovinas de grande importância no Brasil.

No capítulo 1 apresentamos a sequência completa do genoma mitocondrial de dois animais da raça Gir e dois animais da raça Guzerá. Para este efeito, utilizamos a tecnologia de sequenciamento de Nova Geração - NGS. Esse estudo é o primeiro a descrever a sequência completa do genoma mitocondrial destas raças. O capítulo 1 teve por objetivo utilizar os genomas montados para melhorar a compreensão da diversidade molecular de mtDNA entre as raças bovinas.

II - OBJETIVOS

2.1 Objetivos Gerais:

Capítulo 1:

- Montar o genoma mitocondrial de animais pertencentes às raças bovinas Gir e Guzerá e analisar as sequências em relação a outros organismos utilizando métodos comparativos e estudos evolutivos, visando contribuir para uma melhor compreensão das características dessas raças.

Capítulo 2:

- Estabelecer um delineamento experimental da montagem *de novo* do genoma nuclear dos seis animais das raças bovinas Gir e Guzerá, visando contribuir para a conclusão da montagem desses genomas bem como de outros futuros projetos.

2.2 Objetivos Específicos:

Capítulo 1:

- Montar e anotar os genomas mitocondriais das duas raças bovinas: Gir e Guzerá, identificando a estrutura desses genomas, bem como sua composição;
- Comparar os genomas montados com as demais raças com genomas disponíveis publicamente, a fim de compreender as variações entre esses animais;
- Classificar os genomas em haplogrupos com o objetivo de identificar o possível ponto de origem de domesticação desses animais;

Capítulo 2:

- Testar diferentes estratégias de montagem *de novo* com diferentes programas e parâmetros, a fim de estabelecer qual se adequa melhor aos dados desses seis animais;
- Contribuir para a conclusão da montagem *De* novo do genoma nuclear desses animais.

III - CAPÍTULO 1: GENOMA MITOCONDRIAL

3.1 Genoma mitocondrial bovino

As mitocôndrias possuem o seu próprio DNA que é distinto do DNA nuclear. Em bovinos este DNA mitocondrial é circular e de fita dupla com aproximadamente 16,338 pb. Não possui íntrons e contém 37 genes. Treze codificam 13 proteínas da cadeia respiratória: ND1, ND2, ND3, ND4, ND4L, ND5 e ND6 (componentes do completo NADH desidrogenase), ATP6 e ATP8 (polipeptídios do complexo atpase), COI, COII, COIII (subunidades do complexo citocromo oxidase) e Citocromo B. Vinte e dois codificam os tRNAs: Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu1, Leu2, Lys, Met, Phe, Pro, Ser1, Ser2, Thr, Trp, Tyr, Val . Dois codificam os rRNAs: 12S e 16S.

O genoma mitocondrial apresenta algumas características peculiares como, por exemplo, a herança uniparental (geralmente materna) a ausência de recombinação e as altas taxas evolutivas (quando comparado ao genoma nuclear). Devido a estas características o mtDNA tornou-se uma ferramenta importante no estudo das relações evolutivas entre indivíduos, espécies e populações [LI E GRAUR, 1991].

As taxas de substituições sinônimas no genoma mitocondrial de vertebrados foram estimadas em $5{,}7 \times 10^{-8}$/sítio/ano [BROWN *et al*., 1982]. Esse valor é cerca de 10x maior que o encontrado no genoma nuclear (nDNA). Para os genes codificadores de proteínas os sítios sinônimos apresentam taxas de substituição 22x superiores no mtDNA. Nas regiões não codificadoras as taxas evolutivas de maneira geral são mais elevadas [BROWN *et al*., 1982]. O aumento na taxa de mutação é uma das explicações sugeridas para a taxa de substituição de nucleotídeos ser bem mais alta que no nDNA. Algumas das causas da alta taxa de mutação são os excessos de resíduos metabólicos, pela baixa fidelidade na replicação das mitocôndrias e pela ausência de mecanismos de reparo [LI E GRAUR, 1991].

O genoma mitocondrial taurino foi completamente sequenciado pela primeira vez em 1982, por Anderson e colaboradores. Até a presente data (23/12/2014) existem depositadas no GenBank [http://www.ncbi.nlm.nih.gov/genbank/] 270 sequências completas de genomas bovinos, totalizando 58 diferentes raças, sendo uma híbrida (*Bos taurus* e *Bison bison*), 50 taurinas e apenas sete delas são pertencentes aos zebuínos.

Estudos do DNA mitocondrial bovino indicam uma divergência de raças zebuínas e taurinas a partir de um ancestral comum de aproximadamente 1,7-2,0 milhões de anos [HIENDLEDER *et al*., 2008].

Marcadores mitocondriais propostos por Meirelles (1999) conseguem separar os taurinos de zebuínos, entretanto é válido ressaltar que um animal considerado taurino devido as suas características morfológicas, cromossomo Y e genoma nuclear, podem conter o genoma mitocondrial zebuíno (e vice-versa). Essa mistura ocorre devido à introgressão de uma raça sobre a outra.

A população do zebu brasileiro subdivide-se em descendentes de animais de origem importada (POI, puro sangue de origens importados) e aqueles possivelmente derivados localmente por acasalamentos e retro cruzamentos para zebus machos. A segregação mendeliana purificou o genótipo nuclear de origem pura (PO) até um ponto em que não é mais possível diferencia-los do POI. Entretanto, os genes citoplasmáticos podem ser uma fonte significativa de polimorfismos entre essa subespécie [MEIRELLES *et al*., 1999].

Meirelles e colaboradores (1999), propõem uma hipótese em que o zebu americano é composto por dois grupos levando-se em consideração o mtDNA e que a cada três animais analisados apenas um possui o mtDNA zebuíno. Estes resultados sugerem que uma grande parte da matrilinhagem do zebu americano foi obtida através do cruzamento de fêmeas "nativas" de origem taurina com touros importados do subcontinente indiano (zebuínos).

O mtDNA vem sendo frequentemente aplicado em estudos evolutivos devido a facilidade de obtenção da sequência, quando comparado ao genoma nuclear e como já descrito anteriormente, o mtDNA também é empregado nesses estudos devido a sua alta taxa evolutiva e sua origem na maioria das vezes ser uniparental (materna).

3.1.1 Origem dos taurinos e zebuínos

Embora haja inúmeras controvérsias sobre a origem das subespécies *Bos taurus* e *Bos indicus*, existe hoje um consenso em aceitar que o extinto auroque (*Bos primigenius*), em pelo menos dois eventos distintos de domesticação no final do Pleistoceno e Holoceno, foi o progenitor de ambas subspécies [ZEUNER *et al*., 1963, HIENDLEDER, *et al*., 2008]. Devido a este fator, existe atualmente uma nova proposta de denominação para as duas subespécies *Bos primigenius taurus* (originado do *Bos primigenius primigenius*) para os taurinos e *Bos primigenius indicus* (originado do *Bos primigenius namadicus*) para os zebuínos [HIENDLEDER *et al*., 2008]. Entretanto, no presente estudo, apesar de considera-los subespécies, os denominaremos a todo o tempo apenas como *Bos taurus* e *Bos indicus*.

Sobre os eventos que deram origem aos bovinos atuais, um destes teria originado os zebuínos, tendo ocorrido no Baluquistão (hoje a região do Paquistão, subcontinente indiano [BAIG *et al*., 2005]. Embora as evidências arqueológicas apontem que a domesticação ocorreu

no subcontinente indiano, a origem geográfica exata e história filogenética dos zebuínos permanecem incertas [CHEN *et al*., 2010].

O outro evento teria ocorrido no Sudoeste da Ásia (região asiática da Turquia), Líbano, Israel, Iraque e Palestina, dando origem aos taurinos [LOFTUS *et al*., 1994].

Alguns estudos [TROY *et al*., 2001 e LOFTUS *et al*., 1994, BRADLEY *et al*., 1996] tiveram como alvos os bovinos europeus, africanos e indianos. Estes estudos apontaram as raças europeias e africanas pertencentes a uma linhagem e as indianas a outra diferente linhagem. Entretanto, estes estudos foram baseados apenas em genomas mitocondriais parciais e mesmo estudando mais de 300 animais, esses se limitaram a apenas 14 diferentes raças.

A origem dos bovinos Africanos é citada em diversos trabalhos como sendo talvez a mais complexa de todas principalmente pela falta de evidências arqueológicas. As raças africanas geralmente apresentam um mosaico entre *Bos taurus* e *Bos indicus* [GRIGSON, 1991]. FRISCH e colaboradores (1997) apontam evidências para classificação dos bovinos do leste africano ser denominado de "taurindicus" e os bovinos do sul africano de sanga.

Os estudos de Bruford e colaboradores (2003) revelaram a grande complexidade de marcadores de DNA em animais domesticados, tendo como grande surpresa o elevado número de eventos de domesticação e os diversos locais onde estes ocorreram. Diversos trabalhos reportaram a presença de haplogrupos indicando o possível local onde ocorreu a domesticação dos animais pertencentes a estes haplogrupos.

Uma domesticação independente pode ter ocorrido para os zebuínos no norte da África e no leste da Ásia com possibilidade de introgressão local com o auroque silvestre. Chen e colaboradores (2010) propuseram dois haplogrupos, I1 e I2, no Subcontinente indiano consistente com a hipótese de que todos os zebuínos tiveram origem nesta região. Para o haplogrupo I1 foi sugerido que o Vale do rio Indo teria sido o ponto de origem da domesticação, mas a origem do haplogrupo I2 permanece incerta, sabendo-se apenas que foi no subcontinente indiano.

Assim como os zebuínos, os taurinos são divididos em haplogrupos. De acordo com o ponto origem de domesticação, taurinos podem ser divididos em seis haplogrupos (T-T5) sendo o haplogrupo T3 o mais dominante do continente europeu [TROY *et al*., 2001], alguns estudos como o de Beja-Pereira e colaboradores (2006) sugerem a origem múltipla do bovino europeu, baseados no estudo de animais de raças italianas. Um trabalho detalhado sobre o haplogrupo T1 [BONFLIGIO *et al*., 2012] revelou seis subdivisões deste haplogrupo (T1a-f) tendo sido possível inferir que o haplogrupo T1d surgiu no norte da África, pouco depois de sua chegada do Oriente Médio.

A origem das raças bovinas vem sendo estudada e a cada dia mais novos pontos de origens são descobertos. Com as novas tecnologias de sequenciamento, diversos estudos estão sendo publicados [LIU *et al*., 2014; KAI-XING *et al*., 2006; LE *et al*., 2001; SLOMOVIC *et al*., 2005, ACHILLI *et al*., 2008, ACHILLI *et al*., 2009 BONFIGLIO *et al*.,2012], principalmente envolvendo os genomas mitocondriais completos e cada vez mais estes dados contribuem para o conhecimento das raças bovinas em todo o mundo. Com esses novos trabalhos, é provável que importantes fatores surjam e certamente estes poderão contribuir para mudar a nossa abordagem para a conservação dos recursos da biodiversidade dos bovinos no futuro.

As variações estruturais no mtDNA têm sido associada com fenótipos específicos em várias espécies, incluindo a humana [http://www.mitomap.org] camundongo [ROUBERTOUX *et al*., 2003]. Em bovinos, variantes no mtDNA podem afetar a produção de gordura do leite, a composição de carcaça, e traços de fertilidade [HIENDLEDER, 1998; HIENDLEDER *et al*., 2005]. Diante do exposto, o estudo da sequência completa do mtDNA pode ser considerado de grande importância. As diferenças entre taurinos e zebuínos poderiam, portanto, contribuir para as diferenças de fenótipo entre essas subspécies.

O presente estudo envolve o mapeamento e anotação de quatro genomas mitocondriais zebuínos, sendo dois animais pertencentes a raça Gir e dois animais da raça Guzerá e faz a reconstrução filogenética desses animais sequenciados comparando com todas as raças com os genomas mitocondriais completos disponíveis no banco de dados do GenBank [http://www.ncbi.nlm.nih.gov/ genbank/].

Os objetivos deste trabalho consistiram em mapear e anotar os genomas mitocondriais de quatro animais zebuínos para melhor entendimento de suas características genéticas, compreensão de suas histórias evolutivas e relações filogenéticas.

As perguntas principais desse trabalho consistiram em saber se os animais Gir e Guzerá, raças zebuínas do rebanho brasileiro, são realmente *Bos indicus* e se essas raças estariam mais próximas filogeneticamente dos zebuínos das Américas ou de outras localizações geográficas.

Com base nessas perguntas foi possível elaborar a seguinte hipótese:

Os zebuínos do Brasil estariam mais próximos filogeneticamente dos demais zebuínos quando comparado aos taurinos. O Gir e o Guzerá ainda teriam o seu ponto de origem próximo um dos outros, visto que ambas as raças são consideradas como as mais antigas do mundo tendo se originado possivelmente no subcontinente Indiano.

Esse estudo será submetido a revista Mitochondrial DNA para publicação.

3.2 MATERIAIS E MÉTODOS

3.2.1 Amostras coletadas e animais

O DNA total foi extraído do sangue e/ou sêmen, tendo sido coletado a partir de quatro animais de duas raças bovinas: dois touros da raça Gir e dois touros da raça Guzerá, de acordo com protocolos padrão [MILLER *et al*., 1988]. Para proteger a identidade dos animais, estes foram numerados de 1 a 4, como mostrado na Tabela 1. Os critérios para a escolha dos indivíduos 1 e 3 (um touro Gir e um Guzerá) foram: uma menor endogamia; o maior número de registros de filhas; participação em programas oficiais de criação desenvolvido pela EMBRAPA Gado de Leite, e serem de propriedade de empresas públicas (EMBRAPA e EPAMIG – Empresa de Pesquisa Agropecuária de Minas Gerais). Os critérios para a escolha dos demais indivíduos foram: a diversidade, de modo que estes touros são representantes de diferentes rebanhos do país; e possuírem relações genéticas relativamente baixas. Todos estes procedimentos foram realizados na EMBRAPA gado de leite e/ou na Universidade Federal de Minas Gerais - UFMG.

Tabela 1: Animais sequenciados – Plataformas, bibliotecas

Raça	Indivíduo	Plataforma	Tipo-Biblioteca	Tamanho *read*	Inserto
Gir	1	SOLiD	*Mate-pair*	50	1-2kb e 3-4kb
Gir	2	HiSeq	*Paired-end*	100	300-500pb
Guzerá	3	SOLiD	*Mate-pair*	50	1-2kb e 3-4kb
Guzerá	4	HiSeq	*Paired-end*	100	300-500pb

3.2.2 Sequenciamento dos genomas

Os sequenciadores Applied Biosystems SOLiD v4 e o Ilumina HiSeq 1000 foram as plataformas utilizadas para o sequenciamento do tipo *Shotgun* do DNA total dos quatro animais. O DNA dos animais Gir indivíduo 1 e Guzerá indivíduo 3 foram sequenciados no SOLiD v4. Duas bibliotecas do tipo *Mate-pair* (2x50), com tamanhos de insertos de 1-2kb e 3-4kb, foram construídas para cada animal. O DNA dos animais Gir indivíduo 2 e Guzerá indivíduo 4 foram submetidos ao sequenciamento utilizando a plataforma Illumina HiSeq 1000. Uma única biblioteca do tipo *Paired-end* (2x100) foi construída com um tamanho médio de inserto 300-500 bp.

3.2.3 Análises dos dados, mapeamento e anotação

As *reads* brutas oriundas de ambas as plataformas de sequenciamento foram pré-processadas e avaliadas estatisticamente por valores de qualidade dos dados, seguidos por filtragem por qualidade, tamanho e presença de bases ambíguas. As sequências SOLiD foram filtradas pelo programa csfasta_quality_filter e corrigidas pelo *SOLiD Accuracy Enhancement Tool-SAET* [ambos desenvolvidos pela *Applied Biosystems Technologies* http://bcc.bx.psu.edu/ download/ saet. 2.2/]. O tamanho mínimo estabelecido da sequência foi de 50pb e o valor de qualidade PHRED 30. As *reads* oriundas do Hiseq foram checadas por qualidade através do programa FastQC [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/], filtradas usando o programa trimmomatic [http://www.usadellab.org/cms/?page=trimmomatic] e corrigidas pelo programa RACER - *Rapid Accurate correction of erros in reads* [RACER - http://www.csd.uwo.ca/~ilie/RACER/]. O tamanho mínimo estabelecido para as sequências foi de 75pb e o valor de qualidade PHRED 30.

Após o pré-processamento dos dados, estes foram mapeadas contra a sequência pública do genoma mitocondrial completo de *Bos taurus* (V00654) por meio do programa *LifeScope* para as *reads* SOLiD [http://www.lifetechnologies.com/br/~lifescope-genomic-analysis-software.html] e BWA-MEM [http://bio-BWA.sourceforge.net/] para as *reads* Hiseq.

É importante resssaltar, que a estratégia selecionada foi o mapeamento dos genomas e não uma abordagem *De novo*. O mapeamento foi definido, visto que o viés dessa estratégia para genomas mitocondriais é muito menor do que para genoma nuclear, devido as características desses genomas mitocondriais (citadas anteriormente). Contudo, selecionamos uma referência bem curada e próxima dos organismos de estudo.

Para a anotação do genoma, as sequências de cada animal foram submetidas separadamente ao programa MitosWebServer, versão 5.83 (2014-11-04) [http://mitos.bioinf.uni-leipzig.de/index.py], que é uma plataforma online para anotação de genomas mitocondriais. A anotação automática foi seguida por uma minuciosa curadoria manual por meio do programa Artemis versão 13.2.0 [https://www.sanger.ac.uk/ resources/software/artemis/]. As sequências dos tRNAs passaram por uma inspeção manual adicional da estrutura secundária e anticódons.

3.2.4 Composição de nucleotídeos e uso dos códons

A composição de nucleotídeos e os valores de *Relative Synonymous Codon Usage* (RSU) e análises de conteúdo GC foram calculados pelo programa MEGA versão 6. As análises

de enviesamento de uso de códons foram realizadas por meio do programa CodonO webserver [http ://sysbio.cvm.msstate.edu/CodonO/index.php].

3.2.5 Reconstrução Filogenética

3.2.5.1 Conjunto de dados

Para comparações com as quatro sequências geradas nesse estudo, foram obtidas sequências completas do genoma mitocondrial de 58 diferentes raças bovinas, incluindo as subespécies *Bos indicus* e *Bos taurus*, como podem ser observadas na Tabela 2 (arquivo com informações completas no material suplementar online). O conjunto de dados foi constituído por um indivíduo de cada raça cuja sequência completa do mtDNA estivesse disponível publicamente no banco de dados do GenBank. A sequência completa do genoma do Bisão (*Bison bison*) foi incluído nesse conjunto de dados como o grupo externo.

A fim de avaliar a diversidade molecular três conjuntos de dados foram formados, ambos utilizando apenas as sequências de nucleotídeos:

Estratégia 1: Contendo o genoma completo dos 63 animais (58 raças, + 4 montadas nesse estudo + grupo externo),

Estratégia 2: Contendo somente a região hipervariável (D-loop) dos mesmos 63 animais.

Estratégia 3: Contendo as 13 sequências codificadoras de proteínas concatenadas dos 63 animais.

Essas estratégias foram escolhidas em virtude de trabalhos anteriores [LIU *et al*., 2014; KAI-XING *et al*., 2006; LE *et al*., 2001; LOGUE *et al*., 2013] que abordaram cada um, uma dessas estratégias em seus estudos (com bovinos e outros organismos), portanto escolhemos as três para fazer uma comparação de qual conjunto de dados seria a melhor para esse estudo.

3.2.5.2 Alinhamento e curadoria

Para as estratégias 1 e 2 (genoma completo e região D-loop) as sequências foram alinhadas pelo programa MAFFT, versão 7, selecionando um único parâmetro alterado do *default* – G-INS-i, que apesar de muito lento é o mais recomendado para conjuntos de dados menores que 200 sequências com homologia global [http://align.bmr.kyushu-u.ac.jp/mafft/online/server/].

Para a estratégia 3 (genes codificadores de proteínas), os alinhamentos múltiplos foram construídos, gene-por-gene, com a sequência de nucleotídeos dos 13 genes codificadores de proteínas. Para cada gene codificador de proteína, a sequência de DNA foi traduzida para aminoácido selecionando o código de mitocôndria para vertebrados, alinhadas uns contra os

outros, em seguida foram inversamente traduzidas para sequências de nucleotídeos. O alinhador ClustalW, incorporado pelo MEGA, foi utilizado para construir os alinhamentos múltiplos. Os alinhamentos foram corrigidos manualmente quando necessário. Esses genes foram concatenados (um seguido do outro) através do programa FASconCAT [http://software.zfmk.de].

3.2.5.3 Análises Filogenéticas

Para todas as estratégias o programa JModelTest 2.0 implementado na plataforma online Phylemon2 Web server [http://phylemon.bioinfo.cipf.es/index.html] foi usado para predizer qual o melhor modelo para a reconstrução filogenética, baseado nas sequências de nucleotídeos alinhadas.

Árvores filogenéticas foram construídas pelo método modificado de máxima verossimilhança GTR (melhor modelo) - PhyML (version PhyML 3.0.0) implementado na plataforma online Phylemon2.

O programa FigTree [http://tree.bio.ed.ac.uk/] foi escolhido para edição gráfica das árvores geradas.

3.2.6 Identificação de variações do tipo SNV (Variação de uma troca de base - nucleotídeo)

Para avaliar o número total de sítios polimórficos no mtDNA de todas as raças, os programas SNP-Sites [https://github.com/sanger-pathogens/snp_sites] e DnaSP (v.5) [http://www.ub.edu/dnasp/] foram escolhidos. Os números de mudanças sinônimas e não sinônimas foram registrados. As alterações não-sinônimas foram avaliadas basicamente em quatro diferentes classes de amino ácidos determinadas pela diferença em suas cadeias: (1) não-polar e neutro, (2) polar e neutra, (3) ácido e polar, (4) básico e polar. Alterações nos tRNAs, rRNAs, região da D-loop também foram avaliadas.

Tabela 2: Conjunto de dados

Subspecie	Raça	País-Região	Genoma-pb	N.acesso	PubMed	Referência
Bos indicus	Zwergzebu	Germany	16339	AF492350	18467841	HIENDLEDER 2008
Bos indicus	Nelore	Brazil	16341	AY126697	NA	NA
Bos taurus	Fleckvieh	Germany	16338	AF492351	18467841	HIENDLEDER 2008
Bos taurus	Korean native	Korean	16338	AY526085	NA	NA
Bos taurus	Beef cattle	Korean	16340	DQ124389	NA	NA
Bos taurus	Holstein-Friesian	Korean	16340	DQ124403	NA	NA
Bos taurus	Hereford? 1º boi	?	16338	V00654	7120390	ANDERSON
Bos taurus	Iraqi	Iraq	16339	EU177868	18302915	ACHILLI 2008
Bos indicus	deqin	China	16338	GU256940	NA	NA
Bos indicus	Boran	Ethiopia	16339	JN817299	22685589	BONFLIGIO 2012
Bos indicus	Abigar	Ethiopia	16339	JN817298	22685589	BONFLIGIO 2012
Bos indicus	Horro	Ethiopia	16339	JN817330	22685589	BONFLIGIO 2012
Bos indicus	Arsi	Ethiopia	16340	JN817302	22685589	BONFLIGIO 2012
Bos taurus	Mong	Mongolia	16339	FJ971088	19484124	ACHILLI 2009
Bos taurus	Iranian	Iran	16339	EU177870	18302915	ACHILLI 2008
Bos taurus	Ukrainian grey	Ukraine	16340	GQ129208	NA	NA
Bos taurus	Romagnola	Italy	16339	HQ184033	21209945	BONFLIGIO 2010
Bos taurus	Chianina	Italy	16339	FJ971081	19484124	ACHILLI 2009
Bos taurus	Heck cattle	Poland	16340	HM045018	NA	NA
Bos taurus	Hungarian Grey	Hungary	16341	GQ129207	NA	NA
Bos taurus	Angus-X	USA	16341	AY676872	NA	NA
Bos taurus	Calvana	Italy	16340	JN817306	22685589	BONFLIGIO 2012
Bos taurus	Charolais	USA	16341	AY676861	NA	NA
Bos taurus	White Park	Germany	16339	KC153977	23350719	LUDWIG 2013
Bos taurus	Pettiazza	Italy	16338	EU177832	18302915	ACHILLI 2008
Bos taurus	Cinisara	Italy	16340	JN817319	22685589	BONFLIGIO 2012
Bos taurus	Angus	USA	16340	AY676859	NA	NA
Bos taurus	Italian Red Pied	Italy	16339	FJ971082	19484124	ACHILLI 2009
Bos taurus	Friesian	Italy	16339	EU177821	18302915	ACHILLI 2008
Bos taurus	Agerolese	Italy	16341	JN817341	22685589	BONFLIGIO 2012
Bos taurus	Maremmana	Italy	16340	JN817332	22685589	BONFLIGIO 2012
Bos taurus	hybrid bison	USA	16340	GU947009	20870040	KORY 2012
Bos taurus	Valdostana	Italy	16341	EU177817	18302915	ACHILLI 2008
Bos taurus	Piedmontese	Italy	16341	EU177815	18302915	ACHILLI 2008
Bos taurus	Menofi	Egypt	16339	JN817327	22685589	BONFLIGIO 2012
Bos taurus	Angus mix	USA	16339	GU947019	20870040	KORY 2012
Bos taurus	Betizuak	Spain	16339	EU177833	18302915	ACHILLI 2008
Bos taurus	Podolica	Italy	16338	EU177830	18302915	ACHILLI 2008
Bos taurus	Simmental-X	USA	16339	AY676855	NA	NA
Bos taurus	Japanese Black	Japan	16337	AB074964	NA	NA
Bos taurus	Marchigiana	Italy	16340	JN817335	22685589	BONFLIGIO 2012
Bos taurus	Italian Brown	Italy	16338	JN817312	22685589	BONFLIGIO 2012
Bos taurus	Italian Podolian	Italy	16338	EU177843	18302915	ACHILLI 2008
Bos taurus	Pampa Chaquen Creole	Italy	16338	JN817309	22685589	BONFLIGIO 2012
Bos taurus	Sheko	Italy	16338	JN817348	22685589	BONFLIGIO 2012
Bos taurus	Alentejana	Portugal	16339	JN817300	22685589	BONFLIGIO 2012
Bos taurus	Cabannina	Italy	16339	EU177840	22685589	BONFLIGIO 2012
Bos taurus	Chihuahua Creole	Mexico	16341	JN817308	22685589	BONFLIGIO 2012
Bos taurus	Creole	Mexico	16339	JN817307	22685589	BONFLIGIO 2012
Bos taurus	Domiaty	Egypt	16338	JN817323	22685589	BONFLIGIO 2012
Bos taurus	Galbvieh	USA	16340	AY676860	NA	NA
Bos taurus	Greek	Greece	16340	EU177852	18302915	ACHILLI 2008
Bos taurus	Grey Alpine	Italy	16339	HQ184036	NA	NA
Bos taurus	Limousin	France	16340	JN817331	22685589	BONFLIGIO 2012
Bos taurus	Longhorn	USA	16339	GU947021	20870040	KORY 2012
Bos taurus	Modicana	Italy	16338	EU177831	18302915	ACHILLI 2008
Bos taurus	Rendena	Italy	16339	EU177861	18302915	ACHILLI 2008
Bos taurus	Red Mountain Cattle	Germany	16340	KJ709681	NA	NA
Bison bison	Bison	American	16319	NC012346	18302915	ACHILLI 2008

NA: Not Available

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Organização e estrutura dos mitogenomas

A sequência completa dos mtDNAs contém 16,339pb para os animais Gir 1 e Guzerá 3 e 16,338pb para os animais Gir 2 e Guzerá 4. A anotação automática seguida pela minuciosa curadoria manual permitiram identificar a região não codificadora (D-loop), 13 genes codificadores de proteínas, 22 tRNAs e dois rRNAs genes (12s e 16s) em cada um dos genomas analisados. Esses resultados correspondem aos 37 típicos genes encontrados nos animais vertebrados. A Tabela 3 apresenta a estrutura do genoma das quatro sequências geradas nesse estudo que serão disponibilizadas no GenBank (números de acesso ainda não disponíveis). A ordem dos genes mitocondriais é conservada entre todos os animais (Figura1). Para três dos quatro genomas aqui sequenciados (Gir 1, Guzerá 3, Guzerá 4) foi possível identificar regiões não cobertas ou cobertas por apenas uma *read* de profundidade. No total 120 *gaps* foram encontrados nos indivíduos Gir 1, Guzerá 3 e Guzerá 4. Exceto para o gene ND6, todos os outros genes codificadores de proteínas são codificados na fita pesada (*heavy strand*).

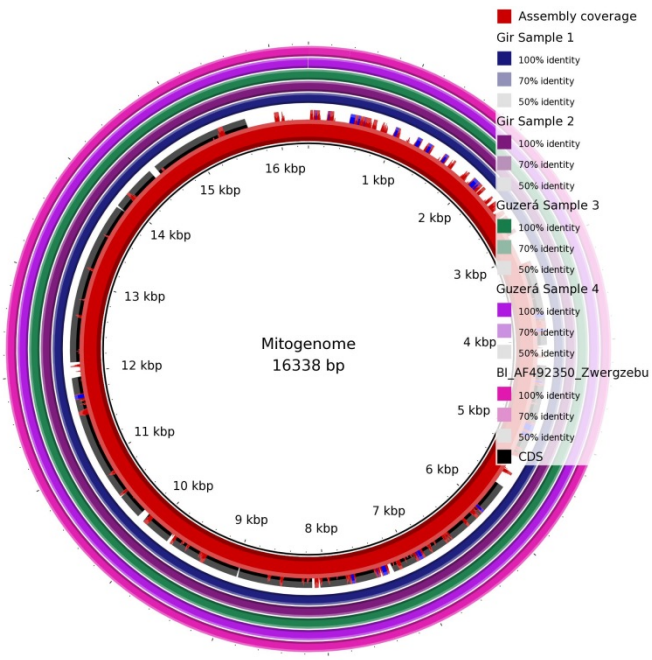


Figura 1: Genomas mitocondriais anotados: A Figura apresenta a comparação entre os quatro genomas desse estudo, o genoma de referência de *Bos taurus* (em vermelho ao centro) e o genoma de outro zebuíno (Zwegzebu, em rosa no último círculo). A sequência dos quatro genomas do presente estudo está entre os dois. As CDS estão plotadas em preto evidenciando que não houve inversões entre esses genomas.

Tabela 3: Estrutura do genoma dos quatro animais sequenciados

		Gir 1			Gir 2			Guzerá 3			Guzerá 4			
Order	Gene/Feature	Start	Stop	Size-bp	Start	Stop	Size-bp	Start	Stop	Size-bp	Start	Stop	Size-bp	*Strand*
	D-Loop	1	362	363	1	363	363	1	363	363	1	363	363	
1	tRNA-Phe (GAA)	363	429	67	364	430	67	364	430	67	364	430	67	*Heavy*
2	12s-rRNA	432	1387	956	431	1385	955	432	1387	956	431	1385	955	*Heavy*
3	tRNA-Val (TAC)	1387	1453	67	1386	1452	67	1387	1453	67	1386	1452	67	*Heavy*
4	16S-rRNA	1450	3022	1573	1450	3023	1574	1450	3022	1573	1450	3023	1574	*Heavy*
5	tRNA-Leu (TAA)	3024	3098	75	3024	3098	75	3024	3098	75	3024	3098	75	*Heavy*
6	ND1	3101	4052	952	3101	4051	951	3101	4052	952	3101	4057	957	*Heavy*
7	tRNA-Ile (GAT)	4057	4125	69	4057	4125	69	4057	4125	69	4057	4125	69	*Heavy*
8	tRNA-Gln (TTG)	4123	4194	72	4123	4194	72	4123	4194	72	4123	4194	72	*Light*
9	tRNA-Met (CAT)	4197	4265	69	4197	4265	69	4197	4265	69	4197	4265	69	*Heavy*
10	ND2	4251	5304	1054	4251	5304	1054	4251	5304	1054	4251	5304	1054	*Heavy*
11	tRNA-Trp (TCA)	5308	5374	67	5308	5374	67	5308	5374	67	5308	5374	67	*Heavy*
12	tRNA-Ala (TGC)	5376	5444	69	5376	5444	69	5376	5444	69	5376	5444	69	*Light*
13	tRNA-Asn (GTT)	5446	5518	73	5446	5518	73	5446	5518	73	5446	5518	73	*Light*
14	tRNA-Cys (GCA)	5551	5618	68	5551	5617	67	5551	5618	68	5551	5617	67	*Light*
15	tRNA-Tyr (GTA)	5618	5685	68	5618	5685	68	5618	5685	68	5618	5685	68	*Light*
16	COI	5688	7232	1545	5687	7231	1545	5688	7232	1545	5687	7231	1545	*Heavy*
17	tRNA-Ser (TGA)	7229	7297	69	7229	7297	69	7229	7297	69	7229	7297	69	*Light*
18	tRNA-Asp (GTC)	7305	7372	68	7305	7372	68	7305	7372	68	7305	7372	68	*Heavy*
19	COII	7374	8056	683	7374	8056	683	7374	8056	683	7374	8056	683	*Heavy*
20	tRNA-Lys (TTT)	8061	8127	67	8061	8127	67	8061	8127	67	8061	8127	67	*Heavy*
21	ATP8	8129	8328	200	8129	8328	200	8129	8328	200	8129	8328	200	*Heavy*
22	ATP6	8290	8970	681	8290	8970	681	8290	8970	681	8290	8970	681	*Heavy*
23	COIII	8967	9751	785	8967	9751	785	8967	9751	785	8967	9751	785	*Heavy*
24	tRNA-Gly (TCC)	9754	9822	69	9754	9822	69	9754	9822	69	9754	9822	69	*Heavy*
25	ND3	9824	10168	345	9823	10167	345	9824	10168	345	9823	10167	345	*Heavy*
26	tRNA-Arg (TCG)	10170	10238	69	10170	10238	69	10170	10238	69	10170	10238	69	*Heavy*
27	ND4 (L)	10224	10533	310	10224	10533	310	10224	10533	310	10225	10534	310	*Heavy*
28	ND4	10530	11898	1369	10529	11897	1369	10530	11898	1369	10529	11897	1369	*Heavy*
29	tRNA-His (GTG)	11907	11976	70	11907	11976	70	11907	11976	70	11907	11976	70	*Heavy*
30	tRNA-Ser2 (GCT)	11977	12036	60	11977	12036	60	11977	12036	60	11977	12036	60	*Heavy*
31	tRNA-Leu2 (TAG)	12038	12108	71	12038	12108	71	12038	12108	71	12038	12108	71	*Heavy*
32	ND5	12110	13930	1821	12109	13929	1821	12110	13930	1821	12109	13929	1821	*Heavy*
33	ND6	13915	14441	527	13914	14440	527	13915	14441	527	13914	14440	527	*Light*
34	tRNA-Glu (TTC)	14441	14509	69	14441	14509	69	14441	14509	69	14441	14509	69	*Light*
35	CYTB	14514	15653	1140	14514	15653	1140	14514	15653	1140	14514	15653	1140	*Heavy*
36	tRNA-Thr (TGT)	15657	15726	70	15657	15726	70	15657	15726	70	15657	15726	70	*Heavy*
37	tRNA-Pro (TGG)	15726	15791	66	15726	15791	66	15726	15791	66	15726	15791	66	*Light*
	D-Loop	15792	16339	548	15792	16338	547	15792	16339	548	15792	16338	547	

3.3.2 Composição de nucleotídeo

A composição de nucleotídeos parece seguir um padrão entre os quatro animais sequenciados, com uma pequena variação na região das CDS. Na Figura 2 é possível observar um possível padrão da composição de nucleotídeos entre todos os quatro indivíduos quando comparamos todos os genes codificadores de proteínas, genes de tRNA e rRNA, região D-loop e o genoma completo. A imagem foi plotada com base nos resultados da composição de nucleotídeos.

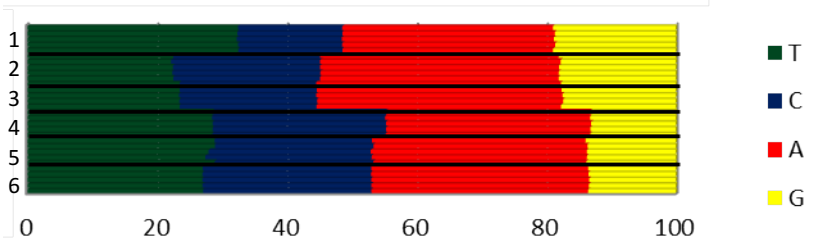


Figura 2: Composição de bases dos genomas: Os números representam as sequências de todos os quatro genomas: 1: tRNAs, 2:12s, 3:16s, 4:D-loop, 5: CDS, 6: Genoma Completo. Cada linha representa um animal. Resultados plotados com base nos resultados obtidos pela MEGAv6.

3.3.3 Proteínas, códons e variações

Treze genes codificadores de proteínas foram identificados em cada um dos genomas mtDNA dos animais sequenciados e estes são semelhantes aos de outros bovinos. Exceto para os genes ND2, ND5 e COIII que possuem TAA como códon inicial, para todos os outros genes o ATG é o códon inicial. Três genes, ND3, COIII e ND4 possuem o códon de parada incompletos que podem ser concluídos através de poliadenilação do mRNA [SLOMOVIC *et al*., 2005], esses resultados são mostrados no material suplementar online.

O uso de códons apresenta um pequeno padrão entre os animais Gir 1 e Guzerá 3 diferente do padrão Gir 2 e Guzerá 4, como pode ser visualizado na Figura 3. O cálculo completo do uso dos códons pode ser visualizado na material suplementar online.

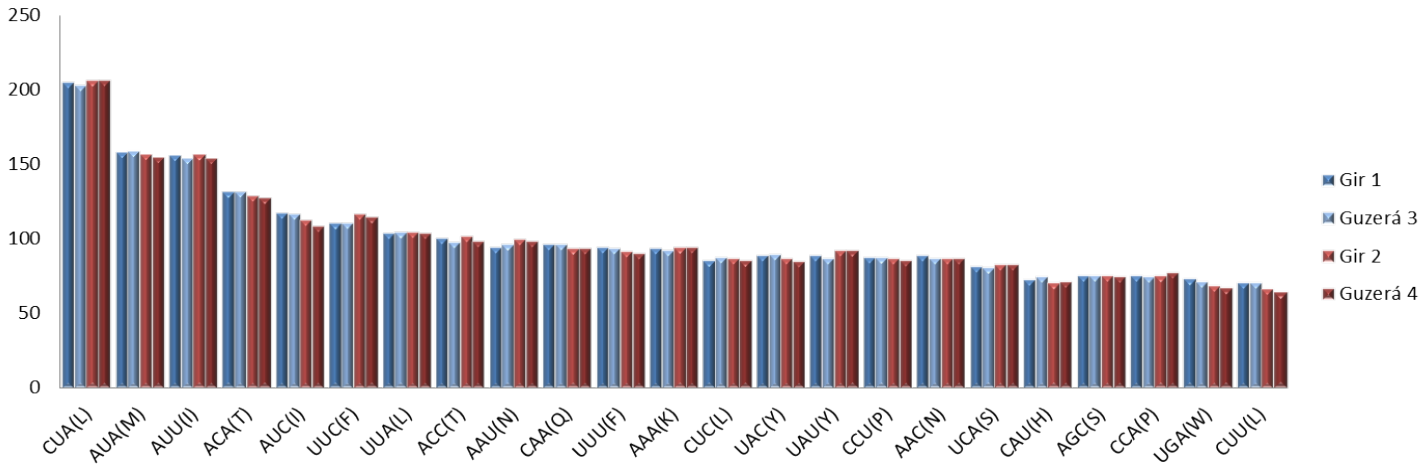


Figura 3: Uso dos códons: A figura apresenta os 23 códons mais usados para cada um dos animais em todos os genes codificadores de proteínas. O eixo Y identifica os códons, entre parênteses os aminoácidos. O eixo X mostra a quantidade de códons. As barras em azul escuro representa o animal Gir 1, em azul claro o animal Guzerá 3, em vermelho claro o animal Gir 2 e vermelho escuro o animal Guzerá 4.

Apesar de encontrarmos uma pequena diferença na utilização dos códons entre os Animais Gir 1e Guzerá 3; Gir 2 e Guzerá 4, quando avaliamos a contribuição de cada códon no total de aminoácidos estes apresentam uma distribuição quase idêntica para todos os animais (Figura 4).

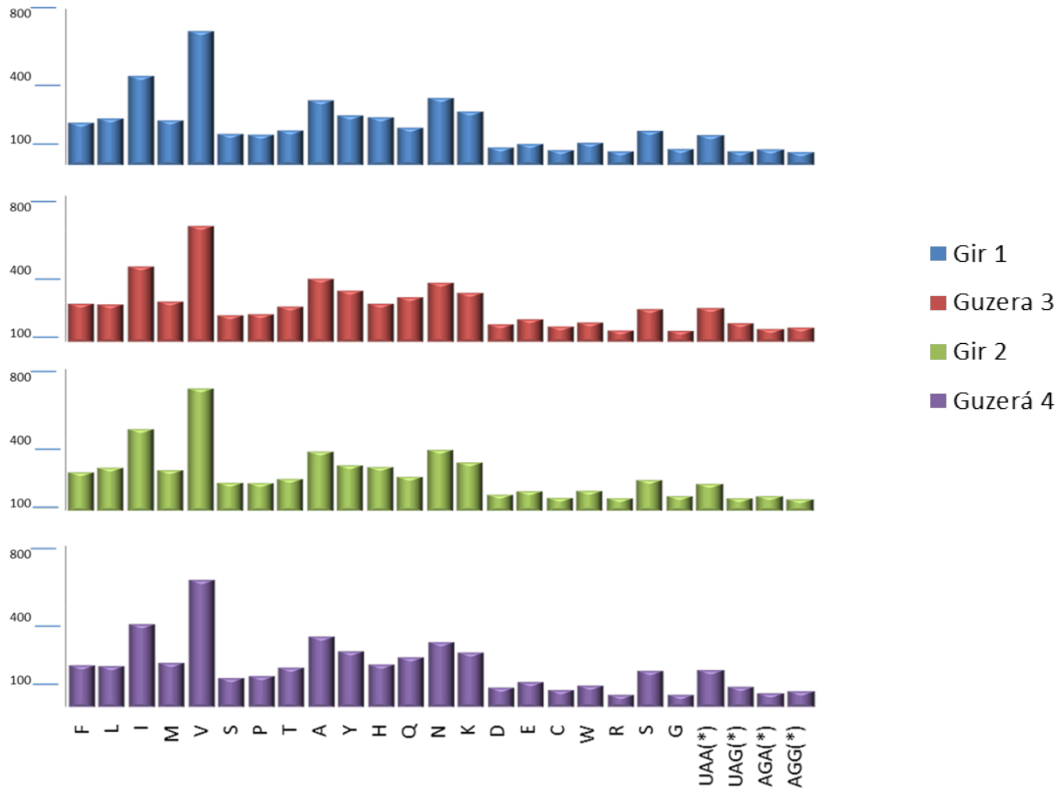


Figura 4: Distribuição dos códons: O eixo Y representa a contribuição de cada um dos códons para o total do aminoácido. O nome dos aminoácidos é indicado no eixo X. O animal Gir 1 está representado pela cor azul, Guzerá 3 em vermelho, Gir 2 verde e Guzerá 4 em lilás. Os (*) indicam códons de parada.

Estes resultados sugerem trocas sinônimas na utilização dos códons. Acredita-se que o enviesamento na utilização dos códons possa resultar de um desvio nas taxas de substituição e/ou da ação da seleção atuando sobre as trocas "silenciosas" no DNA, ou seja, substituições de nucleotídeos que não acarretam a substituição de aminoácidos na sequência de proteínas. A utilização dos códons sinônimos pode refletir a variação na composição dos nucleotídeos, observada nos genomas distintos. Nesse estudo conseguimos ver uma pequena variação na composição de nucleotídeos (Figura 2) principalmente nas regiões codificadores de proteínas.

3.3.4 Genes de tRNA, rRNA e região não codificadora D-loop

Vinte e duas sequências de nucleotídeos (variando de 67pb a 73bp) foram identificados em ambas as raças e as estruturas secundária foram preditas.

O tamanho das sequências dos genes que transcrevem os RNA ribossomais 12s são de 956pb para os animais Gir 1 e Guzerá 3 e 955pb para os animais Gir 2 e Guzerá 4. Para o rRNA

16s os tamanhos são de 1573pb para os animais Gir 1 e Guzerá 3 e 1574pb para os animais Gir 2 e Guzerá 4.

O tamanho total da sequência D-loop foi novamente igual para os dois animais: Gir 1 e Guzerá 3 com um total de 911pb e para os animais Gir 2 e Guzerá 4 com um tamanho de 910pb.

3.3.5 Variações nos genomas (SNVs) em relação à sequência de *Bos taurus* (V00654)

Os SNVs em todos os 63 animais (incluindo os quatro mapeados nesse estudo + o *Bison bison*) selecionadas para este estudo e a diversidade de pares de nucleotídeos resultante ($\boldsymbol{\pi}$), tendo como referência o genoma mitocondrial de *Bos taurus* (V00654) foram analisados.

Em resumo o resultado apresentado é apenas para os quatro genomas mapeados neste estudo contra o genoma referência de *Bos taurus* (V00654) (os dados completos podem ser acessados no material suplementar online). Foi possível encontrar maior similaridade de sequência entre todos os genes de tRNA. Quando analisamos todo o genoma, foi possível encontrar regiões altamente variáveis e a maioria das substituições estão presentes na terceira posição do códon (arquivo suplementar online).

A maior diferença observada em todos os mitogenomas analisados foi observada na região D-loop, seguida pelo gene ND5, mas se analisarmos todos os SNVs em todos os genes codificadores de proteínas (13 genes), este número é maior do que a região D-loop sozinha (Figura 5).

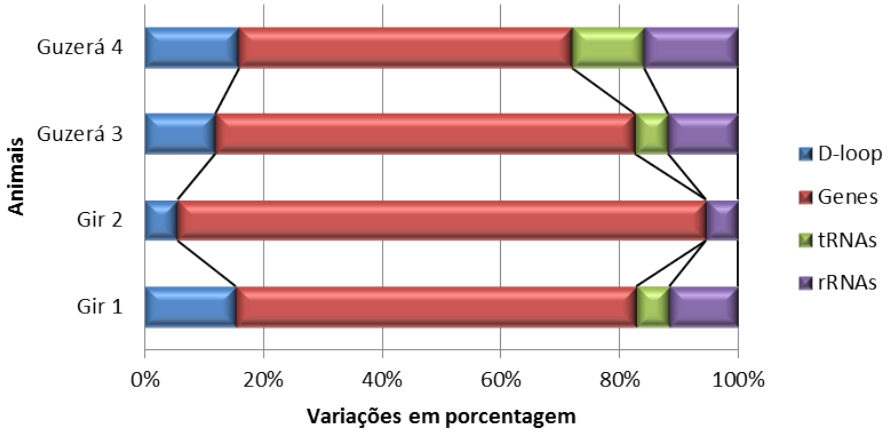


Figura 5: Variações tipo SNV por região nos genomas: A figura apresenta as variações encontradas nos quatro genomas por região quando comparados ao genoma referência de *Bos taurus* (V00654). O eixo Y representa os animais e X as variações em porcentagens. As variações na região D-loop estão representadas em azul, todos SNVs dos 13 genes codificadores de proteína estão em vermelho, SNVs dos tRNAs em verde e os SNVs dos dois rRNAs em roxo.

No total foram identificados 237 SNVs para o Gir indivíduo 1, 18 SNVs para o Gir indivíduo 2, 285 SNVs para o Guzerá indivíduo 3 e 81 SNVs para o Guzerá indivíduo 4 (Figura 6).

A Tabela 4 mostra a quantidade de variações por gene e/ou região encontrada em todos os quatro animais quando comparados a sequência de referência (V00654).

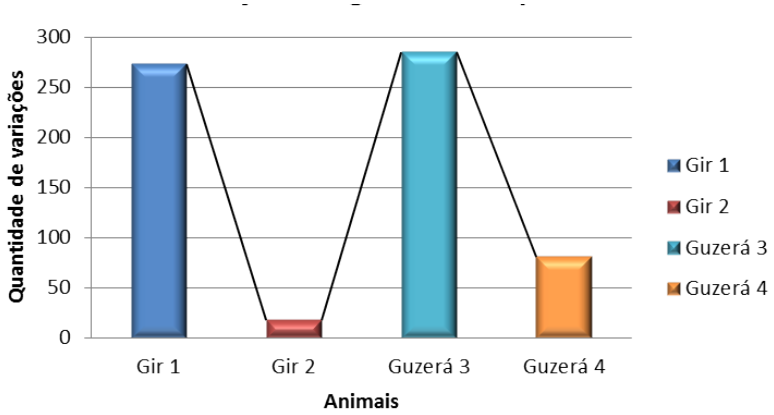


Figura 6: Variações tipo SNV no genoma completo: A figura apresenta as variações encontradas nas sequências completadas dos quatro animais quando comparados ao genoma referência de *Bos taurus* (V00654). O eixo Y representa a quantidade de variações encontradas e o X os animais. A barra em azul escuro mostra as variações para o animal Gir 1, em vermelho para o animal Gir 2, azul claro o animal Guzerá 3 e laranja o animal Guzerá 4.

Em resumo, as variações encontradas nos genes codificadores de proteínas foram: 179 (correspondendo a 65% dos SNVs) para o Gir 1, 16 (correspondendo a 88% dos SNVs) para o Gir 2, 190 (correspondendo a 56% dos SNVs) para o Guzerá 3, 46 (correspondendo a 56% dos SNVs) para o Guzerá 4 .

Dos SNVs capazes de alterar o aminoácido (não-sinônimos) apenas 21 no Gir 1 (correspondendo a 11,7% dos SNVs), três no Gir 2 (correspondendo a 18,7% dos SNVs), 20 no Guzerá 3 (correspondendo a 10% dos SNVs), quatro no Guzerá 4 (correspondendo a 8,6% dos SNVs) foram encontrados (Figura 7). Destes, apenas nove são capazes de alterar as classes de aminoácidos nos animais Gir 1 e Guzerá 3, e somente um é capaz de alterar a classe de aminoácido nos animais Gir 2 e Guzerá 4.

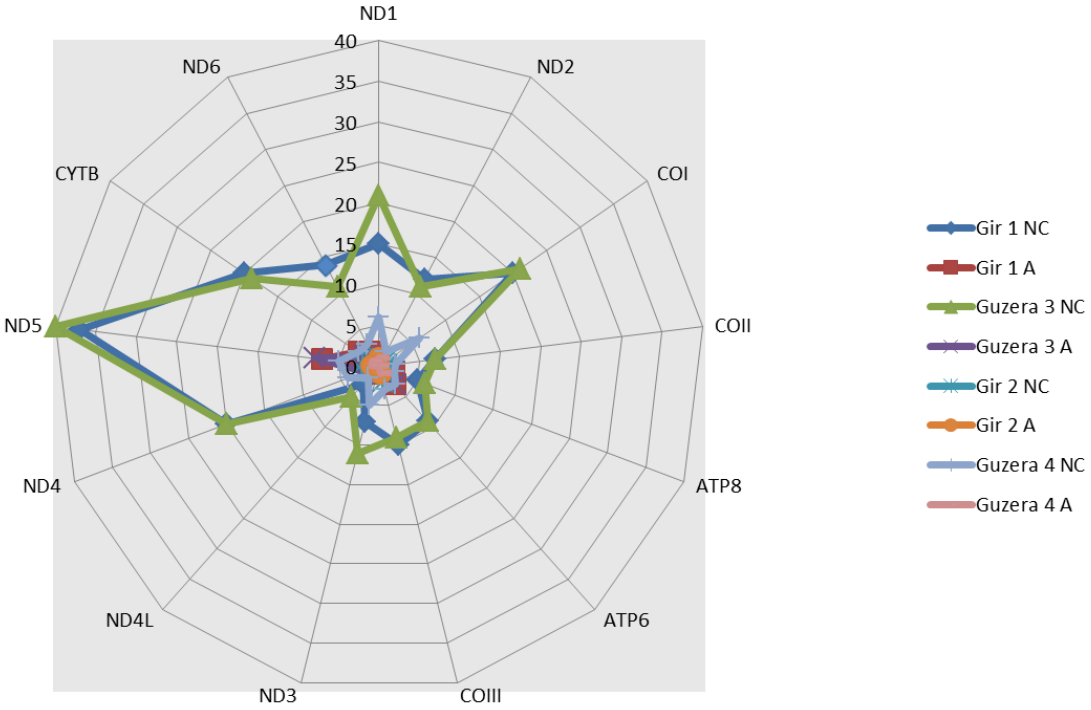


Figura 7: Variações tipo SNV por gene: A figura apresenta as variações encontradas nas sequências dos treze genes codificadores de proteínas dos quatro animais quando comparados ao genoma referência de *Bos taurus* (V00654). Os animais estão representados por cores, NC indica as variações em nucleotídeos, A indica as variações em aminoácidos. Os genes estão representados nos pontos do radar.

Para ambos os rRNAs, foi possível encontrar 41 SNVs no Gir 1, um para o Gir 2, 32 no Guzerá 3 e 13 para o Guzerá 4. Estes resultados podem ser visualizados na Tabela 4. A Tabela 4 também mostra os SNVs encontrados na região D-loop para ambos os animais, sendo possível identificar 40 SNVs para o Gir indivíduo 1, um SNV para o Gir 2, 44 SNVs para Guzerá 3 e nove SNVs para o Guzerá 4.

Além disso, foram identificados SNVs em 15 dos 22 tRNAs para os animais Gir 1 e Guzerá 3, dez para o Guzerá 4 e nenhuma alteração foi encontrada no Gir indivíduo 2. Nenhuma dessas variações levou a mudanças nas estruturas secundárias quando comparamos à s outras raças de bovinos.

Tabela 4: Distribuição dos SNVs no genoma completo dos quatro animais sequenciados comparados ao genoma referências de *Bos taurus*

Genome Order	Gene/Feature	Gir 1	Gir 2	Guzerá 3	Guzerá 4
1	tRNA-Phe (GAA)	0	0	0	0
2	12s-rRNA	11	0	9	6
3	tRNA-Val (TAC)	2	0	0	1
4	16S-rRNA	30	1	23	7
5	tRNA-Leu (TAA)	1	0	1	1
6	ND1	15	1	21	6
7	tRNA-Ile (GAT)	0	0	0	0
8	tRNA-Gln (TTG)	1	0	1	1
9	tRNA-Met (CAT)	0	0	0	0
10	ND2	12	1	11	2
11	tRNA-Trp (TCA)	1	0	1	0
12	tRNA-Ala (TGC)	0	0	1	1
13	tRNA-Asn (GTT)	0	0	1	1
14	tRNA-Cys (GCA)	0	0	1	1
15	tRNA-Tyr (GTA)	0	0	1	1
16	COI	20	1	21	6
17	tRNA-Ser (TGA)	0	0	0	0
18	tRNA-Asp (GTC)	3	0	5	0
19	COII	7	1	7	2
20	tRNA-Lys (TTT)	0	0	0	0
21	ATP8	5	1	6	2
22	ATP6	9	1	9	3
23	COIII	10	2	9	3
24	tRNA-Gly (TCC)	1	0	1	0
25	ND3	7	1	11	5
26	tRNA-Arg (TCG)	0	0	0	0
27	ND4 (L)	3	1	5	2
28	ND4	20	2	20	4
29	tRNA-His (GTG)	0	0	0	0
30	tRNA-Ser2 (GCT)	1	0	2	2
31	tRNA-Leu2 (TAG)	0	0	0	0
32	ND5	37	2	40	5
33	ND6	14	1	19	3
34	tRNA-Glu (TTC)	1	0	1	0
35	CYTB	20	1	11	3
36	tRNA-Thr (TGT)	0	0	1	2
37	tRNA-Pro (TGG)	2	0	2	2
	D-Loop	40	1	44	9
	Total	273	18	285	81

Esses resultados da detecção de variações destacaram uma separação entre os animais Gir indivíduo 1 e Guzerá indivíduo 3 dos outros dois animais (Gir 2 e Guzerá 4).

A fim de confirmar e compreender a separação entre estes animais, foi utilizado um mapa de restrição para a identificação de polimorfismos capazes de identificar e separar os genomas mitocondriais de *B. taurus* e *B. indicus,* usando três enzimas de restrição propostas por Meirelles *et al*., (1999). Todas as análises foram realizadas in *silico* através do *programa* Webcutter [http://rna.lundberg.gu.se/cutter2/] (resultados em arquivo suplementar online).

Foi possível identificar a presença de genoma mitocondrial de taurinos nos animais Gir 2 e Guzerá 4 e a presença de mtDNA de zebuínos para os outros dois animais: Gir 1 e Guzerá

3. Nenhum destes animais apresentou caracterísicas para ambos os genótipos mitocondriais. Tratando-se de animais domesticados isso é muito comum de ocorrer devido a introgressão de uma raça sobre a outra [BRUFORD *et al*., 2003].

3.3.5.1 Das alterações de aminoácidos

O gene ND5 apresentou maior número de SNVs quando comparado aos outros genes (Gir 1= 37, Gir 2 = 2, Guzerá 3 = 40, Guzerá 4 = 5). Estas variações estão localizados na região mais variável da sequência de aminoácidos do ND5. Segundo Meirelles e colaboradores (1999), essas alterações não influenciam o fornecimento de energia para os tecidos, devido a mudanças na eficiência da fosforilação oxidativa. Os estudos de Meirelles foram realizados em animais das raças Gir e Nellore contra a mesma referência que utilizamos de taurus (V00654).

No total, nove alterações de aminoácidos que levam a alterações na classe foram identificados para os animais Gir 1 e Guzerá 3, e um para os animais Gir 2 e Guzerá 4. Quatro dos nove aminoácidos que mudam a classe estavam presentes no gene ND5 para os animais Gir 1 e Guzerá 3. A única alteração de aminoácido que leva a alteração na classe que foi identificada nos animais Gir 2 e Guzerá 4 também estava presente no gene ND5.

As demais alterações de classes de aminoácidos vistas aqui estão presentes nos genes ND1 (1), ATP8 (1), ATP6 (2) ND3 (1) para ambos os animais (Gir1 e Guzerá 3). Essas trocas de aminoácidos podem afetar a afinidade de ligação peptídica [BETTS *et al.,* 2003]. Entretanto análises funcionais precisam ser realizadas para que se possa conhecer o fenótipo completo.

3.3.6 Análises filogenéticas

Para todos os três conjuntos de dados (genoma completo, todos os genes codificadores de proteínas e região D-loop) o modelo GTR foi indicado como o melhor para ser utilizado. Nesse modelo a frequência das bases nitrogenadas se mostra desigual, sendo as taxas de substituição AC, AG, AT, CG, CT, GT.

3.3.6.1 Reconstrução filogenética utilizando as sequências completas dos genomas

Foi realizada a reconstrução filogenética pelo método de máxima verossimilhança - GTR, do conjunto de dados de sequências de nucleotídeos do genoma completo (~ 16339 nucleotídeos) para os 63 animais, incluindo as quatro sequências geradas nesse trabalho mais o grupo externo.

A árvore filogenética (Figura 8) apresenta todos os animais denominados taurinos, levando-se em consideração características morfológicas e /ou cromossomo Y em preto, os zebuínos em rosa e o grupo externo em laranja.

É reportado na árvore os haplogrupos aos quais os animais pertencem. Vários animais tiveram seus haplogrupos informados em trabalhos anteriores [ANDERSON *et al*., 1982, BONFIGLIO *et al*., 2012, ACHILLI *et al*., 2008, ACHILLI *et al*., 2009, BRUFORD *et al*., 2003, BONFIGLIO *et al*., 2010, BAIG *et al*., 2005, CHEN *et al*., 2010]. Para aqueles animais cujo a informação era ausente (11 no total) análises de detecção dos haplogrupos foram realizadas (descrição completa no material suplementar online). Em taurinos são reportadas a existência dos haplogrupos do tipo T (1-5) e nos zebuínos os haplogrupos do tipo I (I1 e I2), ambos podendo ter subdivisões (T1a,b,c,d,e,f,e,g). Os haplogrupos do tipo T e I são identificados levando-se em consideração transições encontradas na região hipervariável. Os haplogrupos do tipo P,E,Q,R são utilizados para reportar uma ancestralidade mitocondrial não taurina e zebuína, indicando uma introgressão das espécies *Bos grunniens*, *Bos javanicus, Bos pirimigenius*. A introgressão de haplogrupo P provavelmente ocorreu tanto no Norte ou na Europa Central, enquanto o haplogrupo Q possivelmente foi adquirido a partir de uma população diferente de auroques que poderia ter variado apenas ao sul dos Alpes. A região utilizada para a classificação é a hipervariável não codificadora mais uma pequena região codificadora (tRNA) [BONFLIGIO *et al*., 2012, ACHILLI *et al*., 2008,. BEJA-PEREIRA *et al*., 2006].

É válido ressaltar, que em vários pontos a árvore não tem uma boa resolução sendo o valor de apoio muito baixo. Esse fato é devido aos genomas apresentarem uma grande similaridade, nenhum deles é idêntico, mas baseado nas análises realizadas nesse trabalho [*Decrease redundancy* - http://web.expasy.org/decreaseredundancy/] as sequências apresentavam similaridade acima de 90% (90-98%). Essa baixa divergência foi retratada no trabalho de Achilli e colaboradores (2009) que reportou a divergência do mtDNA bovino (entre raças) sendo 8x menor que a divergência no mtDNA de humanos.

Na árvore foram reportados apenas os valores de apoio maiores que 75% e por isso, só será discutido os ramos que contém um valor estatístico dentro desse critério. Essa decisão foi tomada por considerarmos que valores abaixo de 75% poderiam ser resultados aleatórios [SILVA *et al*., 2012].

Quatro grandes subdivisões (além do grupo externo) podem ser visualizadas na árvore denominadas de acordo com a origem do genoma mitocondrial: Clado I (zebuínos), Clado II (taurinos), Clado III (Q) e Clado IV (R). Todas essas quatro divisões apresentaram um alto valor de apoio como pode ser visualizado na Figura 8.

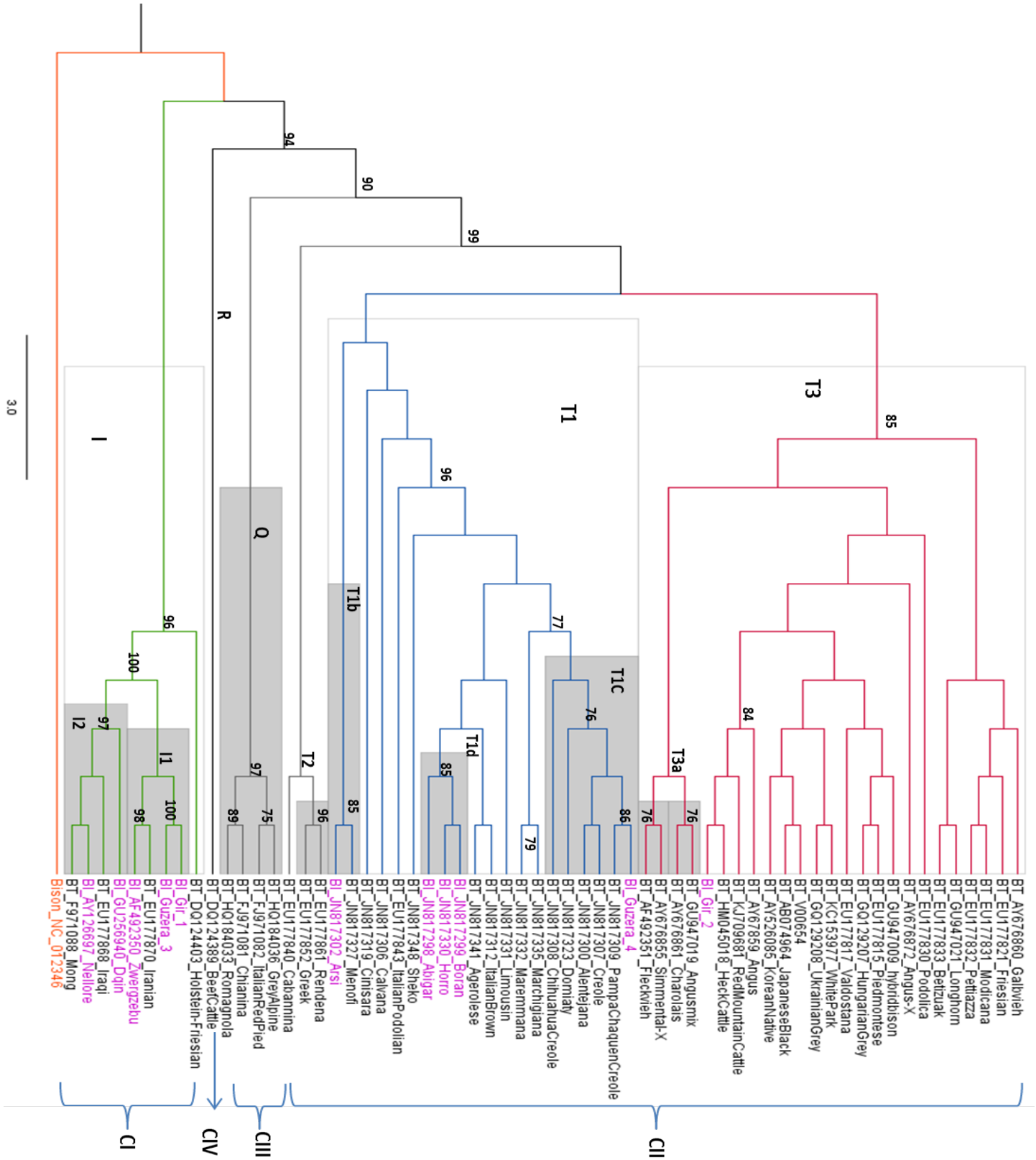


Figura 8: A árvore filogenética utilizando a sequência completa dos genomas: Árvore construída pelo método de máxima verossimilhança – GTR. BT (preto): representa as raças taurinas definidas por características físicas, Chry. BI (rosa): representa as raças zebuínas definidas por características físicas. *Bison*: grupo externo (laranja). CI: clado contendo as raças zebuínas definidas pelo mtDNA. CII: clado contendo as raças taurinas definidas pelo mtDNA. CII: clado contendo os mtDNAs do haplogrupo Q. CIV: clado contendoo haplogrupo R. Os sub-haplogrupos são representados pelas letras seguidas por números: I1, I2, T1a, T1b, T1c, T1d. Os valores de apoio estão representados em porcentagem, ramos contendo valores inferior a 75% foram ocultados da árvore.

A primeira grande divisão a ser discutida é a do clado I (zebuínos - verde). Na árvore, todos os animais cujo genoma mitocondrial é pertencente aos zebuínos estão presentes no clado I (CI). Trabalhos anteriores (ACHILLI *et al*., 2008; ACHILLI *et al*., 2009) reportaram que os animais das raças Iraqi, Iranian, Mong, apesar de serem taurinos por características morfológicas e analise do cromossomo Y, possuem o mtDNA de zebuínos.

Dentro do CI ocorrem mais duas divisões: I1 e I2 com um alto valor de apoio nos ramos. O sub-haplogrupo do tipo I1 é composto pelos animais Gir 1 e Guzerá 3 montadas nesse trabalho e os animais das raças Zwergzebu (Alemanha), Iranian (Irã), Holstein (Coreia). Já o sub-haplogrupo I2 é composto pelas raças Mong (Mongólia), Nellore (Brasil), Iraq (Iraqui) e Dequin (China). Como já citado anteriormente a divisão de haplogrupos sugere o ponto de origem de domesticação do animal, sendo reportado na literatura que o haplogrupo I1 teve forte evidência apontada para o Vale do Rio Indo e o haplogrupo I2 tem seu ponto de origem de domesticação ainda incerto no subcontinente indiano [CHEN *et al*., 2010]. Com esse resultado conseguimos confirmar as hipóteses de que dois dos animais desse estudo são realmente zebuínos e que estes estão mais próximas filogeneticamente dos demais animais das raças com mtDNA zebuínos e ainda que as raças Gir e Guzerá provavelmente tiveram a origem de domesticação ocorridas em lugares próximos e/ou no mesmo local.

O clado II contendo todos os animais cujo mtDNA é pertencente aos taurinos apresenta nesse trabalho três grandes divisões: T1, T2, T3. É possível observar que seis zebuínos estão mais próximos de taurinos (Gir 2 e Guzerá 4 montados nesse trabalho e as raças Africanas: Arsi, Abigar, Horro, Boran) o mesmo caso anterior se aplica aqui, mas desta vez apesar do animal apresentar características morfológicas zebuínas o genoma mitocondrial pertence ao grupo de taurinos. O trabalho de Bonfiglio e colaboradores (2012) descreve os zebuínos Africanos contendo realmente o mtDNA taurino, o que justifica a posição destes animais fora do grupo dos zebuínos no clado II (CII). Como já abordado anteriormente, tratando-se de animais domesticados isso é muito comum de ocorrer devido a introgressão de uma raça sobre a outra.

A raça Egípicia Menofi e a raça Arsi da Etiópia foram agrupadas no sub-haplogrupo T1b. O sub-haplogrupo T1c é representado pelos animais: Guzerá 4 montado nesse trabalho e as raças Italianas: Pampa-Chaquen-Creole, Creole, Alentejana, Domiaty, Chihuaua-Creole. Esse sub-haplogrupo denomina os animais de "*Africa-derived-American*" hipotetiza que estes animais tiveram uma origem de domesticação próxima ao Norte da África, atingindo a Península Ibérica e navegou para a América, com os primeiros colonizadores europeus.

O sub-haplogrupo T1d apresentou um bom valor de apoio agrupando as raças Africanas: Boran, Horro e Abigar. Segundo Bonfligio e colaboradores (2010) esse sub-haplogrupo

mostrou um processo diferente de domesticação das demais raças podendo ter surgido no norte da África, pouco depois de sua chegada do Oriente Médio.

As raças Italianas Rendena e Cabanina e a raça Grega Greek foram agrupadas no haplogrupo T2.

O animal Gir 2 (animal desse trabalho) está presente no haplogrupo T3, que é o haplogrupo com maior número de representantes na Europa. O ponto de origem da domesticação desses animais ocorreu no Oriente Próximo.

Os animais das raças Angus-Mix, Charolais, Simmental-X, Fleckvieh foram agrupadas no sub-haplogrupo T3a.

O haplogrupo T4 e T5 não foram encontrados em nossos estudos.

Achilli e colaboradores (2009) sugerem com seus achados que o haplogrupo T teve uma domesticação de origem neolítica. Bongflio e colaboradores (2012) acreditam que apesar da identificação de numerosos novos polimorfismos ter revelado a existência de seis prováveis tipos de sub-haplogroupos (T), 7-8 fêmeas do mesmo ancestral poderiam ter sofrido domesticação no mesmo local, sendo este local original de todos os T haplogrupos.

O clado III é representado pelos animais com o haplogrupo do tipo Q. Os animais das raças Italianas Grey-Alpine, Italian-Red-Pied, Chianina, Romagnola apresentaram esse tipo de haplogrupo. Bonfiglio e colaboradores (2010) descreveram essas raças Italianas tendo uma origem enigmática e sugeriram que estes haplogrupos representam excelentes ferramentas para avaliar os cruzamentos ou eventos esporádicos da domesticação dos bovinos atuais.

A raça Coreana Beef-Cattle foi a única a apresentar o haplogrupo do tipo P, representada pelo clado IV. É característico das raças coreanas apresentarem esse tipo de haplogrupo devido a introgressão de *Bos grunniens* e *Bos javanicus* provavelmente tendo ocorrido tanto no Norte ou na Europa Central [KIKKAWA *et al*., 2003; NIJMAN *et al*., 2003].

Todos os haplogrupos descritos já haviam sido reportados na literatura, o que foi feito nesse estudo foi a classificação daqueles que ainda não haviam sido classificados. Entretanto o nosso estudo foi o único que utilizou pelo menos um representante de cada raça com genoma disponível publicamente para a reconstrução filogenética

Em conclusão dos resultados obtidos dessa análise filogenética, foi possível separar os quatro grandes clados I (zebuínos) T (taurinos) Q e R. Acredita-se que os zebuínos foram originados na mesma região (subcontinente indiano) e todos os taurinos de outra região (Oriente Próximo). O haplogrupo do tipo Q teria surgido de populações de auroques presentes nos Alpes, enquanto o haplogrupo P pode ter ocorrido no Norte ou na Europa Central. Mesmo com um baixo valor estatístico os ramos que evidenciam as grandes separações se mostraram confiáveis estatisticamente, ajudando assim a suportar as nossas hipóteses.

3.3.6.2 Reconstrução Filogenética usando a região hipervariável D-loop

Foi realizada a reconstrução filogenética pelo método de máxima verossimilhança - GTR, do conjunto de dados de sequências de nucleotídeos da região hipervariável D-loop (~ 948 nucleotídeos) para os 63 animais, incluindo as quatro sequências geradas nesse trabalho mais o grupo externo.

A árvore filogenética (Figura 9) mostra todos os animais taurinos (características morfológicas e/ou cromossomo Y), em preto, os zebuínos em rosa e o grupo externo em laranja. Os haplogrupos dos animais são reportados nos ramos.

Quando comparado aos resultados da estratégia anterior (genoma completo), é possível ressaltar diferenças na topologia da árvore. Como é possível observar na Figura 9, a raça Holstein não está mais presente no clado I (zebuínos), tendo passado para o clado II (taurinos). Árvore representada na Figura 9.

A fim de entender o porquê da separação do animal da raça Holstein, esse genoma foi analisado mais profundamente tendo sido possível inferir que esse animal apresenta características de ambas subspécies em sua composição. Nossas análises foram baseadas no mapa de restrição proposto por Meirelles e colaboradores (1999). Vale ressaltar que a sequência do Holstein foi depositada em 2005 no GenBank (DQ12440), entretanto não houve a publicação de um artigo que fizesse uma análise comparativa do genoma. No trabalho de Kai-Xing e colaboradores (2006) em que houve a utilização do genoma desse animal, os autores só utilizaram a região D-loop para a construção da árvore filogenética sendo possível encontrar o Holstein presente no clado de taurinos. Hiendleder e colaboradores (2008) reportaram em seus achados uma surpresa ao encontrar esse animal no clado I (zebuínos) ao utilizarem apenas as sequências dos genes codificadores de proteínas para reconstrução filogenética e quando utilizaram a região D-loop esse animal estava associado ao clado dos taurinos. Entretanto, parece que o fato do genoma apresentar características de ambas subspécies passou despercebido por estes autores. Estes resultados evidenciam a necessidade de se trabalhar com o genoma completo em estudos filogenéticos e comparativos e sugerem a presença de heteroplasmia na composição desse genoma.

Ainda comparando com a árvore gerada anteriormente, conseguimos ver uma subdivisão do haplogrupo T3. O haplogrupo do tipo Q apesar de permanecer unindo todas as raças Italianas que apresentou esse tipo de haplogrupo, não conseguiu irradiar antes do haplogrupo T. Também podemos observar que mais uma vez o animal da raça coreana Beef-Cattle permaneceu em um ramo sozinho (haplogrupo P), mas dessa vez sem um apoio estatístico.

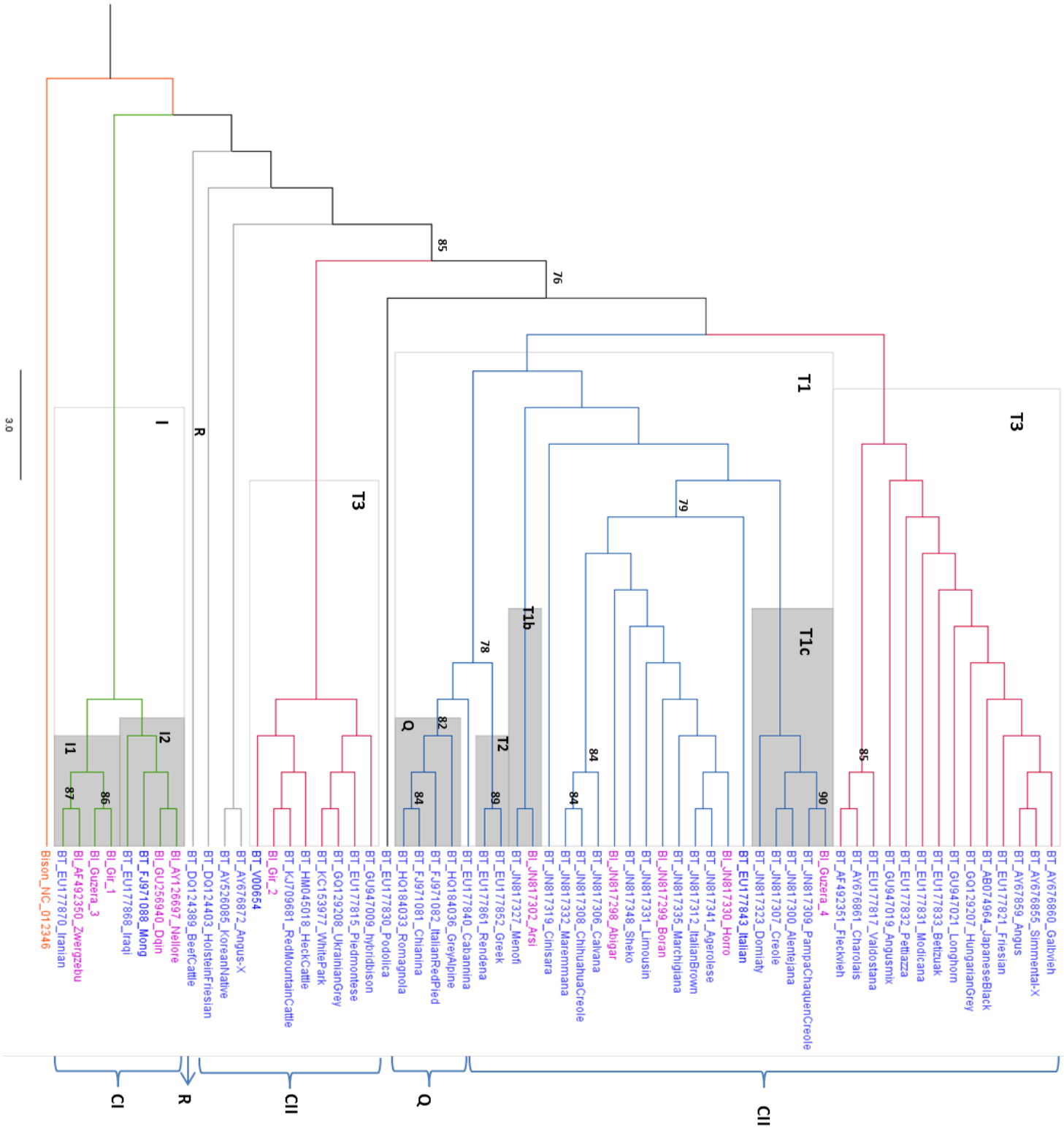


Figura 9: A árvore filogenética utilizando a região D-loop: Árvore construída pelo método de máxima verossimilhança – GTR. BT (azul): representa as raças taurinas definidas por características físicas, Chry. BI (rosa): representa as raças zebuínas definidas por características físicas. *Bison*: grupo externo (laranja). CI: clado contendo as raças zebuínas definidas pelo mtDNA. CII: clado contendo as raças taurinas definidas pelo mtDNA. CII: clado contendo os mtDNAs do haplogrupo Q. CIV: clado contendoo haplogrupo R. Os sub-haplogrupos são representados pelas letras seguidas por números: I1, I2, T1a, T1b, T1c, T1d. Os valores de apoio estão representados em porcentagem, ramos contendo valores inferior a 75% foram ocultados da árvore.

É provável que os haplogrupos Q e R não foram separados (com valor de apoio dos taurinos, visto que a classificação desses haplogrupos utiliza não somente a região hipervariável D-loop, mas também as regiões codificadoras.

Um fator muito relevante de ser observado é a diminuição do apoio estatístico na presente estratégia. Alguns ramos permaneceram com um apoio alto como pode ser observado no clado I e em todos os clados destacados em cinza.

3.3.6.3 Reconstrução Filogenética baseadas em clusters de genes

Foi realizada a reconstrução filogenética pelo método de máxima verossimilhança - GTR, do conjunto de dados de sequências de nucleotídeos dos genes codificadores de proteína concatenados para os sessenta e três animais, incluindo as quatro sequências geradas nesse trabalho mais o grupo externo.

O valor de apoio estatístico para os ramos se mostrou muito inferior ao das estratégias anteriores (Figura 10). Visto que os valores de apoio são baixos resolvemos não utilizar essa estratégia para resolução da filogenia.

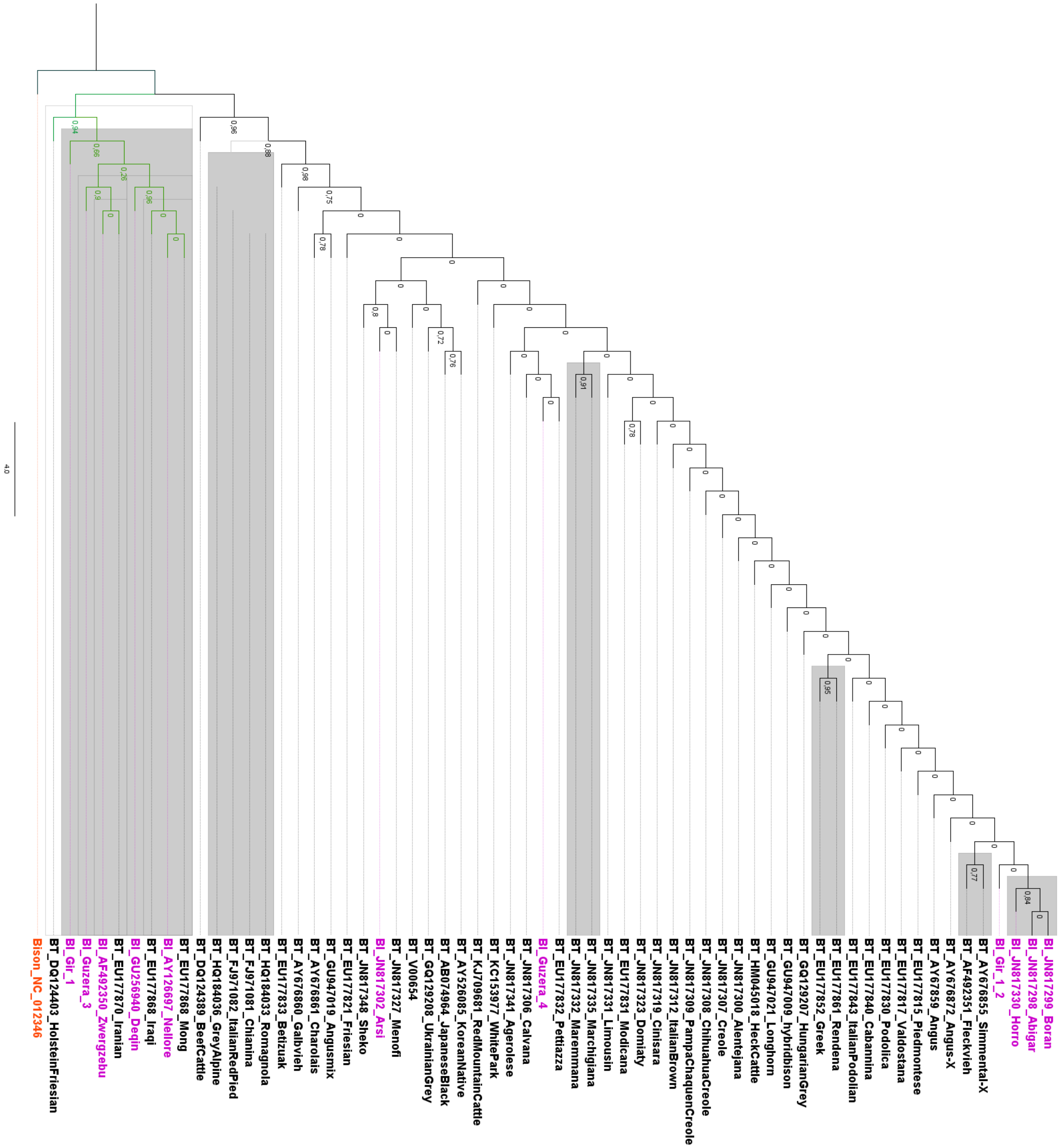


Figura 10: A árvore filogenética utilizando os genes codificadores de proteínas concatenados: Árvore construída pelo método de máxima verossimilhança – GTR. BT (preto): representa as raças taurinas definidas por características físicas, Chry. BI (rosa): representa as raças zebuínas definidas por características físicas. *Bison*: grupo externo (laranja). CI: clado contendo as raças zebuínas definidas pelo mtDNA. CII: clado contendo as raças taurinas definidas pelo mtDNA. CII: clado contendo os mtDNAs do haplogrupo Q. CIV: clado contendoo haplogrupo R. Os sub-haplogrupos são representados pelas letras seguidas por números: I1, I2, T1a, T1b, T1c, T1d. Os valores de apoio estão representados em porcentagem.

Os valores de apoio estatístico das árvores geradas com a sequência mitocondrial completa se mostraram mais robustos em relação às demais árvores geradas. Esse fato nos indicou a necessidade de uma sequência mitocondrial ampla, cobrindo regiões codificadoras e não codificadoras, de modo a obter valores de apoio estatísticos maiores, a fim de garantir a confiabilidade das conclusões sobre a história evolutiva das raças que tais análises nos proporcionam.

Por longos anos, as relações filogenéticas inferidas a partir do mtDNA, baseando-se apenas para a região D-loop, forneceram um quadro alternativo para a classificação dos bovinos. No entanto, os estudos em seres humanos utilizando a sequência dos genomas mitocondriais completos, mostraram que, quando utilizada a sequência completa dos genomas a resolução filogenética foi grandemente melhorada quando comparada a análise de uma pequena porção do genoma [TORRONI *et al.,* 2006]. Os nossos achados corroboram com esses resultados em humanos, visto que o ponto de apoio das árvores aumentou quando utilizamos o genoma completo havendo uma melhor separação das raças.

3.4 CONCLUSÕES

No presente trabalho o genoma mitocondrial de quatro representantes das duas raças zebuínas de maior contribuição para o rebanho leiteiro do país foram sequenciadas pela primeira vez com o objetivo de utilizar as sequências geradas para melhorar a compreensão da diversidade molecular de mtDNA entre as raças bovinas.

Através das análises comparativas foi possível identificar que dois dos genomas aqui montados transportam o mtDNA de origem taurina (Gir 2 e Guzerá 4) e dois origem zebuína (Gir 1 e Guzerá 3).

A quantidade de SNVs encontradas corrobora com a quantidade descrita em trabalhos anteriores [HIENDLEDER *et al*., 2008]. Alterações das funções celulares provenientes destes SNVs não foram encontradas.

A reconstrução filogenética nos permitiu classificar os genomas nos haplogrupos já propostos anteriormente, nos fazendo classificar os animais Gir 1 e Guzerá 3 no haplogrupo I2 tendo o subcontinente Indiano como provável ponto de origem da domesticação desses animais. Para o animal Gir 2 foi possível classifica-lo no haplogrupo T3 tendo sua provável origem de domesticação no Oriente Próximo e o Guzerá indivíduo 4 classificado no haplogrupo T1C "*Africa-derived-American*".

A montagem do genoma mitocondrial dessas raças certamente irá fornecer bases genéticas para várias outras pesquisas

3.4.1 Limitações das análises

O estudo filogenético apresentou algumas limitações como: A alta similaridade entre as raças dificultando a resolução das árvores geradas. Pouca quantidade de genomas zebuínos sequenciados dificultando a comparação entre esses genomas. Utilização apenas do genoma mitocondrial para análises filogenéticas (origem geralmente uniparental). Dificuldade de obtenção dos meta-dados. País de origem dos animais sequenciados, tipo de material coletado (sangue, sêmen), sexo do animal.

IV - CAPÍTULO 2: GENOMA NUCLEAR

Este trabalho faz parte de um projeto maior tendo como objetivo geral identificar SNPs nos genomas zebuínos para inclusão em Chips de genotipagem, mas para isso é necessário que o genoma nuclear e mitocondrial dos zebuínos das raças Gir e Guzerá sejam montados. Contudo, esse é um trabalho piloto de tentativas de montagem dos genomas e teve como objetivo estabelecer as melhores estratégias de montagem *de* novo, bem como, direcionar o melhor caminho para futura conclusão do genoma nuclear dessas raças e de outros projetos de montagem de grandes genomas desenvolvidos pelo nosso grupo de trabalho.

Os animais e as sequências utilizadas nesse capítulo são os mesmos do capítulo 1, acrescidos de mais dois animais, sendo um de cada raça.

4.1 Genoma nuclear bovino

Um dos objetivos mais visados ao sequenciar o genoma nuclear bovino consiste em identificar genes que possam estar associados às características mais apreciadas em produção, como por exemplo, os genes de metabolismo do lipídeo, de grande interesse na produção de leite, bem como os genes relacionados à reprodução, dentre outros fatores. Atualmente as informações de alterações na sequência do genoma que possam estar relacionadas a características de produção vêm sendo integradas aos chips de genotipagem.

O primeiro sequenciamento completo do genoma nuclear bovino foi concluído e publicado em 2009 [ELSIK, TELLAM, & WORLEY, with The Bovine Genome Sequencing and Analysis Consortium, 2009]. A raça sequenciada foi a Hereford, pertencente à subespécie *Bos taurus*. Esse estudo revelou o tamanho do genoma nuclear de aproximadamente 2670.14Mb e 22 mil genes com grande complexidade do genoma, apresentando alta densidade de segmentos duplicados e inúmeros elementos repetitivos.

Outros estudos completos de genomas nuclear bovinos envolvendo outras raças de *Bos taurus* e *Bos indicus* [BT flekvieh: ECK *et al*., 2009, BI Gir: LIAO *et al*., 2013, BI: Nellore: CANAVEZ *et al.*, 2011] foram publicados, mas nenhum dos estudos adoratam a abordagem *de novo*. Todos os estudos abordaram a estratégia de mapeamento tendo a sequência do Hereford como referência. Nenhum zebuíno teve o genoma nucelar montado com a abordagem *de novo.* Uma das vantagens da montagem *de novo* sobre o mapeamento contra a referência, é que através da abordagem é possível identificar inversões entre os genomas que poderiam ter passado despercebido na estratégia de mapeamento. Já a desvantagem é a dificuldade de utilização desse método.

Visto que as raças leiteiras Gir e Guzerá são de grande importância para a formação do rebanho bovino brasileiro e que estas raças ainda não estão com os genomas montados e

disponíveis para inclusão nos chips de genotipagem para o melhoramento genético, esse trabalho teve como objetivo iniciar o projeto de montagem *de* novo do genoma nuclear dessas duas raças bovinas.

Neste estudo, seis genomas zebuínos de duas diferentes raças foram sequenciados utilizando-se sequenciadores de nova geração e submetidos ao processo de montagem *de novo*.

4.1.2 Desafios da montagem *de novo* para grandes genomas eucariotos

Diversos autores relatam a dificuldade em montar grandes genomas sequenciados com a tecnologia de sequenciamento de nova geração - NGS [SCHATZ *et al*., 2010, BRADNAM *et al*., 2013, SIMPSON *et al*., 2014, CHU *et al*., 2013, YOUNG *et al*., 2010]. Um grande gargalo para os projetos de montagem consiste em converter os dados brutos dos sequenciadores em dados de alta qualidade. Outro fator é a complexidade do processo de montagem do genoma devido aos diferentes comprimentos e quantidade das sequências e as taxas de erros produzidas por diferentes tecnologias de NGS. Um desafio adicional é a mistura de sequências produzidas por diferentes tecnologias.

Na montagem dos grandes genomas o desafio é aumentado, sendo as maiores dificuldades relacionadas à complexidade desses genomas, à grande quantidade de elementos repetitivos (muitas vezes estes são responsáveis pela confusão nos algoritmos dos programas de montagem) e a imensa quantidade de dados (dificuldade de processamento e armazenamento). Torna-se um desafio maior trabalhar com mamíferos, pois estes genomas costumam ser maiores e mais complexos comparados a outros animais, bactérias, fungos e até mesmo algumas plantas.

Não existe ainda uma definição oficial para se denominar um genoma de grande ou pequeno, mas baseado em alguns trabalhos [BRADNAM *et al*., 2013, Li *et al*.,2010], chamaremos aqui, de grandes genomas todos aqueles que sejam maiores ou iguais a 1gb.

Alguns grandes genomas eucariotos foram sequenciados, montados (*de novo*) publicados e disponibilizados publicamente utilizando somente sequenciamento de nova geração [Cabra: DONG *et al*., 2013, Conífera: NYSTEDT *et al*., 2013, Humano: LI *et al*., 2010, Panda: LI *et al*., 2009, Tartaruga: SHAFFER *et al*., 2013, Peru: DALLOUL *et al*., 2010; : Peixe, Cobra, Pássaro: BRADNAM *et al*., 2013].

4.1.3 As plataformas de sequenciamento de nova geração

Diversas plataformas de sequenciamento estão disponíveis para o sequenciamento dos genomas. Definem-se como sequenciamento de nova geração todas as tecnologias de sequenciamento desenvolvidas após o do método de sequenciamento por Sanger [Nature.com http://www.nature.com/subjects/next-generation-sequencing, SANGER *et al*., 1977] e são

baseadas em alta geração de dados. Milhões ou bilhões de sequências de DNA podem ser sequenciadas em paralelo diminuindo o tempo de sequenciamento e o custo por base [Nature.com http://www.nature.com/subjects/next-generation-sequencing].

As primeiras novas plataformas de sequenciamento comercializadas foram denominadas de Segunda Geração, a partir da evolução das técnicas de sequenciamento denominou-se Terceira Geração e alguns autores denominam até Quarta Geração [GUT *et al*., 2013]. Esses termos ainda são controversos, não sendo objetivo desse trabalho discutir nomenclaturas. Aqui iremos denominar o sequenciamento apenas como Sequenciamento de Nova Geração ou pela sigla NGS.

Nesse trabalho quatro diferentes plataformas de NGS foram selecionadas: SOLiD V4 (Applied Biosystems® SOLiD™ 4 System), Illumina HiSeq 1000, Illumina MiSeq e PacBio V2 (Pacific Biosciences).

As plataformas SOLiD e HiSeq têm como característica vantajosa gerar uma alta quantidade de dados (500gb SOLiD e 10-300gb HiSeq 1000). A principal desvantagem é o pequeno comprimento da sequência (2x50pb SOLiD V4 e 2x100pb HiSeq 1000).

Vale ressaltar que os valores informados de todas as plataformas são referentes aos *kits* usados para gerar os dados neste trabalho, as informações completas da quantidade de sequências geradas por cada plataforma podem ser encontradas no site dos fabricantes [Illumina:http://www.illumina.com/systems/sequencing.html,SOLiD:https://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms091372.pdf].

Já o MiSeq tem a vantagem do maior comprimento das sequências (200-400pb), mas como desvantagem a geração de dados é pequena quando comparado ao SOLiD V4 e HiSeq 1000, podendo variar de 8gb á 15gb de dados, dependendo do *kit* de preparação da amostra. Neste trabalho utilizamos a versão de 8gb.

As sequências oriundas do PacBio apresentam a vantagem do tamanho da sequência ser longa, média de 6kb na versão 2 e 10kb na versão 3 (dados brutos), entretanto a quantidade de dados gerados é baixa se comparados as outras plataformas disponíveis, aproximadamente 50.000bp. Outra desvantagem é a taxa de erro gerada pela plataforma. No presente trabalho utilizamos a versão 2 do PacBio.

Diversos autores fizeram comparações entre as diferentes plataformas [LIU *et al*., 2012, JÜNEMANN *et al*., 2013, LOMAN *et al*., 2012, QUAIL *et al*., 2012]. Nenhuma das comparações utilizaram o SOLiD V4. As conclusões são bem variadas, dependendo do organismo trabalhado, do *kit* de sequenciamento usado para cada plataforma, das métricas de avaliação da qualidade, entre outros fatores. Mesmo não sendo possível gerar um consenso

sobre essas avaliações, o que fica claro entre todos os resultados é que não existe uma plataforma melhor do que a outra, sempre vai depender da pergunta do projeto, do tipo de dado analisado (organismo de estudo) etc. Todas as plataformas apresentam vantagens e desvantagens quando comparadas uma com a outra. Entretanto, para montagem de grandes genomas, também podemos ver outro consenso onde a plataforma mais usada é a HiSeq, provavelmente devido a característica de alta geração de dados e tamanho das sequências [BRADNAM *et al* ., 2013].

4.1.4 Algoritmos dos atuais programas de montagem de genomas

Com o avanço das tecnologias de sequenciamento de nova geração, os desenvolvedores de softwares se viram obrigados a acompanhar esse crescimento. Entretanto os programas computacionais para montagem não acompanharam o crescimento da tecnologia de sequenciamento, tornando-se este, talvez, um dos maiores gargalos hoje da montagem dos genomas.

Algumas plataformas de sequenciamento desenvolveram seus próprios programas para correção das *reads* e para montagem das mesmas. Como exemplo temos os programas Newbler [http://www.454.com/products/analysis-software/] desenvolvido para trabalhar com *reads* 454, SMRTAnalysis [http://www.pacb.com/devnet/] desenvolvido para filtrar, corrigir e montar as *reads* de PacBio e o programa fornecido pela *Life Technologies SOLiD Accuracy Enhancement Tool* [SAET - http://bcc.bx.psu.edu/download/saet.2.2/] desenvolvido para filtrar e corrigir as *reads* SOLiD. Esses programas são os mais indicados para trabalhar nas etapas de filtragem e correção das sequências, de acordo com a plataforma escolhida, sendo possível obter geralmente os melhores resultados. Especificamente nesse trabalho os melhores e mais confiáveis resultados sempre foram obtidos utilizando o software desenvolvido pela equipe da plataforma.

Diversos programas foram desenvolvidos para montagem de genomas através de dados de NGS: HyDA [SHARIAT *et al*., 2014], ABySS [SIMPSON *et al*., 2009], Newbler [http://www.454.com/products/analysis-software/], SOAPdenovo [LI *et al*., 2008], Ray [BOISVERT et al., 2012], SGA [SIMPSON *et al*., 2012], Velvet [ZERBINO *et al*., 2008], ALLPHATs [BUTLER *et al*., 2008] entre outros, e até mesmo programas que foram desenvolvidos na época do método de Sanger podem ser aplicados na montagem de pequenos genomas: MIRA [CHEVREUX *et al*., 1999], Celera Assembler [MYERS *et al*., 2000], CAP3 [HUANG *et al*., 1999].

A maioria dos programas aplicados para montagem de grandes genomas são baseados na utilização de grafos de *Brujin*, a grande vantagem da utilização desse tipo de grafo é a velocidade e capacidade de processamento de grandes volumes de dados.

Resumidamente este é um grafo de representação de uma sequência (ou conjunto de sequências) através de sua decomposição de subsequências de tamanho K (k-mer). O tamanho de K não pode ser muito grande ou muito pequeno. A utilização deste grafo pode ser vantajosa por ter sido desenvolvido para lidar com problemas complexos, grandes volumes de dados gerados pelo NGS e a rápida detecção de K-mers compartilhados reduzindo assim o custo computacional em relação a busca de sobreposições em alinhamentos pareados, ou seja, não é necessário comparar par a par. Porém, existem também alguns pontos que não podem passar despercebidos, tais como o alto uso de memória, são mais sensíveis a repetições e erros de sequenciamento, e ainda podem perder algumas sobreposições verdadeiras dependendo do tamanho do K, do tamanho da sobreposição e a taxa de erro nas *reads*.

Apesar de grande parte dos programas integrarem o grafo de *Bruijn*, cada um tem sua característica específica, podendo ser vantajosa ou não para o dado de interesse. O montador de leituras curtas Velvet, por exemplo, utiliza o grafo de *Bruijn* para o processo de comparação e montagem das sequências. Zerbino e colaboradores (2008) afirmam que em montadores tradicionais que não utilizam o grafo De *Bruijn*, cada leitura é tratada como um nó em um grafo de sobreposição, o que, considerando a quantidade de informação gerada pelos sequenciadores de *reads* curtas, torna o processamento do grafo extremamente custoso computacionalmente. Já o grafo de *Bruijn* compõe uma representação das leituras em pequenas palavras com tamanho pré-definido K, como já foi falado anteriormente. Outra importante característica do Velvet é a possibilidade de entrada de diferentes conjuntos de dados para a realização da montagem, porém sem o uso efetivo destes dados. Por exemplo, leituras longas são usadas apenas no tratamento de repetições e erros de montagem. Além disto, o montador Velvet não faz uso da qualidade de bases [ZERBINO *et al*., 2008]. Neste trabalho o programa Velvet não conseguiu concluir nenhuma das montagens, provavelmente devido à grande quantidade de dados (dados não mostrados).

Outro programa que integra o grafo de *Bruijn* é o SOAPdenovo, este foi projetado para leituras curtas de sequências geradas a partir do Illumina GA, mas pode ser aplicado por várias outras plataformas, como SOLiD, 454 entre outros, criando novas oportunidades para construção de sequências de referência e realização de análises precisas de genomas inexplorados. O SOAPdenovo aceita o formato de entrada FASTA para referência que também é o formado de saída, este programa emprega um único modelo de linha de comando. No modelo de computação paralela, as sequências dos índices da Tabela *hash* irão se manter na

memória e o alinhamento dos procedimentos serão realizados por vários conjuntos de dados consultados em uma ordem. Este modelo evita entrada/saída (E/S), tempo gasto no carregamento de referências e criação de Tabelas *hash* várias vezes sendo adequado também para serviços web em tempo real [LI *et al*., 2009].

4.1.5 Estratégias de montagens

Diante das diferentes plataformas de sequenciamento e dos distintos programas existentes, diferentes estratégias de montagens de grandes genomas podem e vêm sendo aplicadas. No trabalho de Li e colaboradores (2010) o genoma do Panda foi montado, para isso os autores escolheram as sequências oriundas do sequenciador Genome Analyser (GA) da Illumina. Trinta e duas bibliotecas de diferentes tamanhos de insertos, variando de 150pb a 10kb foram sequenciadas, o que correspondeu a uma alta cobertura de 92x após a filtragem por qualidade. O genoma não foi concluído, mas estimou-se que 94% foram montados. Já a montagem da Cabra [DONG *et al*., 2013] foi uma montagem mais elegante e complexa, porém não mais eficiente que a do Panda, tendo sido adotada a plataforma GA Illumina com 14 bibliotecas com insertos variando de 180pb até 40kb, correspondendo a uma cobertura de 65x. Os autores empregaram a montagem híbrida integrando o mapeamento óptico aos dados GA. Assim como no genoma do Panda, o genoma da Cabra não foi completamente montado, neste caso 92% do genoma foi coberto.

O que pode ser observado de estratégia de montagem de grandes genomas é que em todas foram utilizadas diferentes tipos de bibliotecas (*Paired-end* e *Mate-pair*) com diferentes tamanhos de insertos, e uma alta cobertura das *reads* sobre o genoma. Algumas também adotaram a montagem híbrida integrando dados de diferentes plataformas [Tartaruga: SHAFFER *et al*., 2013, Peru: DALLOUL *et al*., 2010; Cobra, Pássaro: BRADNAM *et al*., 2013].

Em relação aos programas de montagem escolhidos, para o genoma do Panda apenas um foi utilizado: SOAPdenovo e todas as funções (como filtragem e correção das *reads*, pré-grafos, *contigs*, *scaffolds*, *gapcloser*), o mesmo autor que montou o genoma do Panda foi quem desenvolveu o SOAPdenovo (LI *et al*., 2008). O que fica comprovado com a utilização deste programa é que o mesmo consegue lidar com uma imensa quantidade de dados. O genoma da Cabra também escolheu o programa SOAPdenovo (DBG) e para montagem híbrida foram desenvolvidos programas "*in-house*".

Assim, diante do êxito obtido com a utilização do SOAPdenovo para montagem de genomas complexos, o mesmo será utilizado no presente trabalho para a montagem *de novo* do genoma de animais representantes da raça Gir e Guzerá.

4.1.6 Principais parâmetros considerados na avaliação da qualidade da montagem *de novo*

Após a montagem do genoma é preciso avaliar a qualidade da montagem gerada. Para poder determinar qual a melhor montagem, seja ela oriunda de diferentes programas ou de um mesmo programa com diferentes parâmetros, é necessário adotar a mesma métrica de avaliação de montagem para todas, pois só assim será possível determinar qual foi a melhor, de acordo com as métricas utilizadas.

A outra grande questão é como definir quais são as melhores métricas de avaliação. Em um recente grande projeto, denominado Assemblathon [BRADNAM *et al* ., 2013], os autores após testarem 100 diferentes métricas propuseram 10 como chave na avaliação das montagens *de novo*, essas métricas avaliaram parâmetros estatísticos como valor de N50, NG50 como também parâmetros biológicos do tipo a presença de genes eucariotos centrais.

O valor de N50 corresponde ao N, tal que 50% do total de pares de bases do genoma esteja contida em *contigs* >=N pb. A medida NG50 é a utilização do N50 *versus* o tamanho do *contig/scaffold*. Define-se como cobertura do genoma o total de pares de bases em *reads* dividido pelo tamanho do genoma, o que na prática corresponde a quantas vezes em média, cada base do genoma foi sequenciada

Neste trabalho não foram abordadas todas as 10 métricas chaves devido à falta de dados para avaliar todos os parâmetros, como por exemplo, não temos dados de mapeamento óptico e biblioteca de fosmídeos.

Os parâmetros selecionados foram os estatísticos como: valor do N50, quantidade de *contigs/scaffolds*, maior *contig/scaffold*, média mediana e quantidade das bases presentes nos *contigs/scaffolds*.

Diante do exposto, o presente capítulo teve como objetivos principais fazer a montagem *de novo* do genoma nuclear de seis genomas bovinos através do sequenciamento de nova geração de diferentes plataformas. Esse trabalho foi embasado em algumas questões centrais: A utilização de diferentes plataformas de sequenciamento poderia contribuir para uma melhor montagem *de novo*? Haveria uma melhor plataforma de sequenciamento para esses genomas bovinos e uma melhor estratégia?

Diante dessas questões foi possível formular a hipótese de que a utilização de diferentes plataformas de sequenciamento poderia contribuir para uma melhor montagem *de novo*, pois as mesmas seriam capazes de fechar *gaps* oriundos de uma primeira plataforma abordada. E a utilização de diferentes programas poderiam mostrar melhorias gradativas da qualidade da montagem.

Os animais e as sequências trabalhadas nesse capítulo são os mesmos do capítulo 1, acrescido de mais um animal da raça Gir e um da raça Guzerá.

A seguir serão descritos todos os dados que foram disponibilizados para o desenvolvimento dessa dissertação, bem como, todas as tentativas de montagem *de novo* dos dados recebidos.

4.2 MATERIAIS E MÉTODOS

4.2.1 Dados disponíveis – Sequenciamento

Como parte desse trabalho foram recebidos dados de seis indivíduos sendo três pertencentes a raça Gir e três pertencentes a raça Guzerá. A descrição dos indivíduos é mostrada na Tabela 5. Os dados foram recebidos em diferentes momentos: dados *Mate-pair* do SOLiD (1/2013), dados de PacBio (1/2013), MiSeq (2/2013) e por último os dados do HiSeq (2/2014).

Tabela 5: Animais Sequenciados

Raça	Indivíduo	Plataforma	Tipo biblioteca	Tamanho *Read*	Tamanho do Inserto	Cobertura Esperada
Gir	1	SOLiD	*Mate-pair*	50	1a2kb	17,857
Gir	1	SOLiD	*Mate-pair*	50	3a4kb	17,849
Gir	1	MiSeq	*Paired-end*	250	700pb	2,02
Gir	1	PacBio	*Standard sequencing*	~6000	*Standard sequencing*	0,169
Gir	2	HiSeq	*Paired-end*	100	300-500pb	14,268
Gir	5	HiSeq	*Paired-end*	100	300-500pb	13,33
Guzerá	3	SOLiD	*Mate-pair*	50	1a2kb	19,384
Guzerá	3	SOLiD	*Mate-pair*	50	3a4kb	21,064
Guzerá	3	MiSeq	*Paired-end*	250	700pb	3,073
Guzerá	3	PacBio	*Standard sequencing*	~6000	*Standard sequencing*	0,169
Guzerá	4	HiSeq	*Paired-end*	100	300-500pb	11,795
Guzerá	6	HiSeq	*Paired-end*	100	300-500pb	15,402

Os dados SOLiD e MiSeq foram gerados na plataforma de Sequenciamento da FIOCRUZ-Minas. Os dados da plataforma PacBio RS foram gerados pela GATC Biotech AG, Konstanz, Alemanha. As sequências HiSeq foram geradas na plataforma de sequenciamento da Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo – (ESALQ USP Piracicaba). A construção das bibliotecas e descrição da obtenção dos dados é descrita a seguir.

4.2.1.1 Dados oriundos das plataformas SOLiD e HiSeq

A metodologia de geração dos dados SOLiD e HiSeq já foram descritos anteriormente no capítulo do genoma mitocondrial (página 29).

Os animais sequenciados com a plataforma SOLiD foram o Gir 1 e o Guzerá 3, tendo sido geradas duas bibliotecas (1-2kb e 3-4kb) para cada indivíduo.

Já para a plataforma HiSeq os animais sequenciados foram o Gir 2 e 5, Guzerá 3 e 6 (uma biblioteca para cada animal).

4.2.1.2 Dados oriundos da plataforma de sequenciamento PacBio

Para o sequenciamento dos dados da plataforma PacBio foi utilizado DNA dos mesmos animais sequenciados com as plataformas SOLiD e MiSeq (Gir indivíduo 1 e Guzerá indivíduo 3). A metodologia do PacBio é baseada no sequenciamento de DNA em tempo real a partir da observação, em ordem temporal, da incorporação de nucleotídeos marcados com fluorescência durante a síntese de DNA por uma molécula única de polimerase [EID *et al*., 2009]. O sequenciamento das amostras foi terceirizado com a empresa GATC Biotech AG [Konstanz, Alemanha]. Uma biblioteca do tipo *Standard sequencing* (ideal para leituras aleatórias, longas e contínuas) foi construída para cada animal e sequenciada. Um total de 75.000 leituras foram geradas para cada animal com tamanho médio de 6.000 kb. A cobertura esperada foi de 0,16x para cada um.

4.2.1.3 Dados oriundos da plataforma de sequenciamento MiSeq

Para os animais Gir indivíduo 1 e Guzerá indivíduo 3 foram geradas sequências com a plataforma MiSeq. A preparação da biblioteca do tipo *Paired-end* de DNA genômico foi construída a partir de 50 ng de DNA. Em seguida, a amostra foi submetida a uma reação de fragmentação aleatória na qual o DNA foi simultaneamente fragmentado e ligado a adaptadores específicos utilizando *o kit Nextera® XT DNA Sample Preparation* (Illumina) conforme instrução do fabricante. Em seguida, o DNA genômico foi purificado e submetido a uma reação de amplificação utilizando iniciadores complementares aos adaptadores. Os produtos foram quantificados através do qPCR utilizando o Kit Sybr Fast qPCR kit (Kapa). As bibliotecas foram diluídas em uma solução de Tris-HCl e Tween 0,1%, depositadas em uma flowchip e submetidas a 500 ciclos (2x250bp) de sequenciamento utilizando o MiSeq Reagent Kit v2 (Illumina). As imagens obtidas foram processadas e analisadas pelo programa fornecido pelo fabricante. A cobertura esperada é de ~3x para cada indivíduo sequenciado.

4.2.2 Avaliação da qualidade e pré-processamento dos dados

Todas as sequências geradas, independente da plataforma, foram submetidas a análises de qualidade. Com exceção dos dados PacBio, para as sequências de todas as outras plataformas o programa FastQc [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/] foi selecionado. Esse programa permite extrair relatórios dos conjuntos de sequências a partir dos dados brutos, gerando gráficos da distribuição das sequências levando em consideração a qualidade média total por base, a qualidade média de todas as *reads* e o conteúdo G + C dos

dados. Para os dados de PacBio o programa SMRTanalysis fornecido pelos fabricantes para a extração do relatório de qualidade das sequências foi selecionado.

Após avaliação da qualidade, as *reads* foram submetidas à etapa de pré-processamento por qualidade e tamanho, garantindo assim que apenas as sequências de boa qualidade sejam aplicadas para a montagem do genoma.

4.2.2.1 Pré-processamento das reads SOLiD

O arquivo de saída do sequenciador SOLiD tem o formato "ColorSpace" ou também chamado de csfasta (característico da química do sequenciador). A filtragem e correção dos dados, portanto, ocorrem nesse formato.

Para a filtragem dos dados foi utilizado o *script* csfasta_quality_filter (desenvolvido pelo fabricante). Foram realizados dois testes, sendo o primeiro foi estabelecido um tamanho mínimo de 50pb e qualidade PHRED 20, e no segundo permaneceu o tamanho da *read*, mas alterou-se a qualidade PHRED para 30. Em seguida essas sequências foram submetidas ao processo de correção de possíveis erros gerados pela química do sequenciamento pelo programa fornecido pela *Life Technologies*: SAET.

Após essas etapas os arquivos csfasta foram convertidos para o formato FASTQ, através do programa fornecido pela própria *Life Technologies* (SOLiD2std.pl) e submetidos ao programa FASTQc para análise de qualidade e conteúdo das sequências.

4.2.2.2 Pré-processamento das reads HiSeq e MiSeq

Para os dados do HiSeq e MiSeq o programa Trimmomatic [http://www.usadellab.org/cms/?page=trimmomatic] foi escolhido para filtragem dos dados. Para ambos foi estabelecido um alto valor de qualidade PHRED (mínimo 30) com tamanho mínimo da sequência de 75pb. Para correção de possíveis erros nas *reads* foi utilizado o programa *Rapid Accurate correction of erros in reads* [RACER - http://www.csd.uwo.ca/~ilie/RACER/].

4.2.2.3 Pré-processamento das reads PacBio

A análise da qualidade dos dados e o pré-processamento das sequências oriundas do sequenciador PacBio foram realizadas através do programa SMRTAnalysis. Assim como para os dados SOLiD, submetemos as *reads* PacBio a dois tipos de filtro, sendo um mais rigoroso que o outro. Neste caso, os valores foram de 75% e 80% da qualidade PacBio.

4.2.3 Montagens

Após o pré-processamento dos dados, estes foram submetidos a diferentes tentativas de montagem *de novo* dos genomas. Para essas montagens dois diferentes programas foram testados: SOAPdenovo [LI *et al*., 2008], ABySS [SIMPSON *et al*., 2009] e PacBioToCA [KOREN *et al*., 2012] e diferentes estratégias, como será descrito a seguir.

4.2.3.1 Decisão dos melhores valores de k (k-mers)

Visto que os programas SOAPdenovo e ABySS são baseados em estruturas de grafos de *Brujin*, o programa Kmergenie [http://Kmergenie.bx.psu.edu/] foi selecionado para tomada de decisão do melhor valor de k para cada um dos conjuntos de dados. A partir do valor de k indicado pelo Kmergenie dois valores abaixo e dois acima do melhor valor estimado foram testados nas montagens. Em caso do programa retornar mais de um melhor valor estes foram acrescentados, totalizando sete diferentes valores de K para cada tentativa de montagem.

4.2.3.2 Separando os pares das Reads

Com exceção do PacBio, todas as outras plataformas englobaram o sequenciamento do tipo pares: *Paired-end* (Illumina: MiSeq e HiSeq) e *Mate-pair* (SOLiD). No pré-processamento das *reads* muitas vezes os pares são perdidos (por baixa qualidade e entre outros fatores), mas ocorre com grande frequência a perda de apenas uma *read* do par. Nesse trabalho denominaremos as sequências sem pares como fragmentos.

Uma abordagem capaz de facilitar a montagem dos genomas é a separação dos pares das sequências dos fragmentos, para isso o *script* mergeshuffed.pl, é capaz de separar os pares e os fragmentos em arquivos diferentes, através de expressões regulares, foi adotado nesse estudo.

Os dados oriundos das quatro plataformas de sequenciamento (SOLiD, MiSeq, HiSeq e PacBio) foram montados separadamente. Para a montagem híbrida, em uma primeira estratégia reunimos os dados das plataformas SOLiD e MiSeq (mesmos animais sequenciados) e depois unimos os dados das plataformas SOLiD + MiSeq + HiSeq (diferentes animais, mas mesma raça). A Tabela 6 apresenta um resumo comparativo de todas as estratégias de montagem realizadas nesse trabalho. Para todas essas estratégias representadas na Tabela 6, ainda houve um acréscimo de estratégia que foi inserir os fragmentos (*reads* sem par) em todas essas estratégias (fazendo a montagem com os fragmentos e sem os fragmentos).

Tabela 6: Estratégias das Montagens *De novo*

Plataformas	Raça	Indivíduos	Estratégias			Programas	
SOLiD	Gir	1	1 Biblioteca 1a2kb	1 Biblioteca 3a4kb	2 Bibliotecas	SoapDenovo	-
SOLiD	Guzerá	3	1 Biblioteca 1a2kb	1 Biblioteca 3a4kb	2 Bibliotecas	SoapDenovo	-
MiSeq	Gir	1	1 Biblioteca	-	-	SoapDenovo	ABySS
MiSeq	Guzerá	3	1 Biblioteca	-	-	SoapDenovo	ABySS
SOLiD + MiSeq	Gir	1	Todas Bibliotecas			SoapDenovo	-
SOLiD + MiSeq	Guzerá	3	Todas Bibliotecas			SoapDenovo	-
HiSeq	Gir	2	1 Biblioteca	-	-	SoapDenovo	ABySS
HiSeq	Guzerá	4	1 Biblioteca	-	-	SoapDenovo	ABySS
HiSeq	Gir	5	1 Biblioteca	-	-	SoapDenovo	ABySS
HiSeq	Guzerá	6	1 Biblioteca	-	-	SoapDenovo	ABySS
HiSeq	Gir	2 e 5	Todas Bibliotecas	-	-	SoapDenovo	ABySS
HiSeq	Guzerá	4 e 6	Todas Bibliotecas	-	-	SoapDenovo	ABySS
SOLiD + MiSeq + HiSeq	Gir	1, 2 e 5	Todas Bibliotecas			SoapDenovo	-
SOLiD + MiSeq + HiSeq	Guzerá	3, 4 e 6	Todas Bibliotecas			SoapDenovo	-
PacBio	Gir	1	1 Biblioteca	-	-	PacBioToca	-
PacBio	Guzerá	3	1 Biblioteca	-	-	PacBioToca	-

4.2.3.3 Montagens SOLiD

Em uma primeira estratégia, os dados oriundos das quatro plataformas de sequenciamento foram montados separadamente, assim para a montagem *de novo* das sequências oriundas do SOLiD (animais Gir indivíduo 1e Guzerá indivíduo 3) seguimos três diferentes estratégias:

- Primeira estratégia: montagem com o conjunto de dados da biblioteca *Mate-pair* de 1-2kb (para cada raça separada).

- Segunda estratégia: montagem com o conjunto de dados da biblioteca *Mate-pair* de 3-4kb (para cada raça separada).

- Terceira estratégia: montagem com ambos os conjuntos de dados para cada raça (bibliotecas de 1-2kb e 3-4kb).

Seguimos essas estratégias para saber o quanto cada biblioteca poderia contribuir separadamente para a montagem final.

Todas as três estratégias foram submetidas à apenas o programa SOAPdenovo. O programa ABySS foi testado, mas ele não é capaz de incorporar sequências *Mate-pair*s no processo de montagem de *contigs*, apenas no de *scaffolds*.

- SOAPdenovo

Diferentes parâmetros foram testados, gerando duas possiblidades para cada valor de k-mer. Como por exemplo: Para um valor de k (23) duas diferentes montagens foram testadas alterando-se as restrições dos parâmetros (-D e -M). O parâmetro –D controla as bordas das sequências, fazendo a exclusão das mesmas, caso estas estejam abaixo do valor indicado. O parâmetro –M controla a força da fusão das sequências similares durante a formação dos *contigs*.

Testamos ainda mais um tipo de montagem que foi acrescentando as *reads* sem pares (chamadas de fragmentos).

As Figuras 11 e 12 representam a montagem do animal Gir 1, para a biblioteca de 1-2kb com o valor de k=23. Na Figura 11 é mostrado o arquivo contendo as duas parametrizações do arquivo *bos.cfg*. Os parâmetros contidos na figura 11 significam: max_rd_len: o tamanho máximo da *read*, LIB: sempre que uma biblioteca for acrescentada, name: nome da biblioteca, min-avg-max: tamanhos do inserto, reverse_seq1: *Mate-pair,* asm_flag3*:* utiliza a biblioteca na construção de *contigs* e *scaffolds*, rank: ordem de processamento das bibliotecas, fragmentos: *reads* sem par.

```
> Biblioteca de 1a2

bos.cfg Montagem com fragmentos:
max_rd_len=50
[LIB]
name=1a2kb
min_ins=1275
avg_ins=1500
max_ins=1875
reverse_seq=1
asm_flags=3
rank=1
q1=/sto4data-2/zebu4/result/UGA_Juliana/result/Gir/SAET/1a2/MergeFR/mergedSequencesGirSOLiD1a2.1.fastq
q2=/sto4data-2/zebu4/result/UGA_Juliana/result/Gir/SAET/1a2/MergeFR/mergedSequencesGirSOLiD1a2.2.fastq
[LIB]
name=fragmentos
reverse_seq=1
asm_flags=1
q=/sto4data-2/zebu4/result/UGA_Juliana/result/Gir/SAET/1a2/MergeFR/mergedSequencesGirSOLiD1a2.nomatch1.fastq
q=/sto4data-2/zebu4/result/UGA_Juliana/result/Gir/SAET/1a2/MergeFR/mergedSequencesGirSOLiD1a2.nomatch2.fastq


bos.cfg Montagem sem Fragmentos
max_rd_len=50
[LIB]
name=1a2kb
min_ins=1275
avg_ins=1500
max_ins=1875
reverse_seq=1
asm_flags=3
rank=1
q1=/sto4data-2/zebu4/result/UGA_Juliana/result/Gir/SAET/1a2/MergeFR/mergedSequencesGirSOLiD1a2.1.fastq
q2=/sto4data-2/zebu4/result/UGA_Juliana/result/Gir/SAET/1a2/MergeFR/mergedSequencesGirSOLiD1a2.2.fastq
```

Figura 11: Arquivo de configuração do SOAPdenovo

A Figura 12 mostra o arquivo de configuração criado em *shell script* contendo a biblioteca com e sem os fragmentos. Os parâmetros significam: Pregraph: formação dos pré grafos, -s: arquivo de entrada, -d: controla a frequência do k-mer, fazendo a exclusão caso os valores estejam abaixo do indicado, -a: quantidade de memória Ram em GB, -p: número de processadores, -k: valor k-mer, -o arquivo de saída, contig: construção dos *contigs*, -g: arquivo de entrada, -D e –M (parâmetros alterados para as diferentes estratégias), map: mapear as *reads* nos *contigs*, scaff: construção dos *scaffolds*, -F: preencher com N as lacunas nos *scaffolds*, -L: tamanho mínimo do *contig* para formar os *scaffolds*.

Pipele.sh

```
#!/bin/sh

echo "SOAPdenovo-31mer pregraph -s bos.cfg -d 4 -a 800G -p 32 -K 23 -o montagem1" >> montagem1.out
time SOAPdenovo-31mer pregraph -s bos.cfg -d 4 -a 800G -p 32 -K 23 -o montagem1 >> montagem1.out
echo "SOAPdenovo-31mer contig -g montagem1 -D 15 -M 3" >> montagem1.out
time SOAPdenovo-31mer contig -g montagem1 -D 15 -M 3 >> montagem1.out
echo "SOAPdenovo-31mer map -s bos.cfg -p 32 -g montagem1" >> montagem1.out
time SOAPdenovo-31mer map -s bos.cfg -p 32 -g montagem1 >> montagem1.out
echo "SOAPdenovo-31mer scaff -F -L 200 -g montagem1 -p 32" >> montagem1.out
time SOAPdenovo-31mer scaff -F -L 200 -g montagem1 -p 32 >> montagem1.out

echo "SOAPdenovo-31mer pregraph -s bos.cfg -a 800G -p 32 -K 23 -o montagem2" >> montagem2.out
time SOAPdenovo-31mer pregraph -s bos.cfg -a 800G -p 32 -K 23 -o montagem2 >> montagem2.out
echo "SOAPdenovo-31mer contig -g montagem2 -M 3" >> montagem2.out
time SOAPdenovo-31mer contig -g montagem2 -M 3 >> montagem2.out
echo "SOAPdenovo-31mer map -s bos.cfg -p 32 -g montagem2" >> montagem2.out
time SOAPdenovo-31mer map -s bos.cfg -p 32 -g montagem2 >> montagem2.out
echo "SOAPdenovo-31mer scaff -F -g montagem2 -p 32" >> montagem2.out
time SOAPdenovo-31mer scaff -F -g montagem2 -p 32 >> montagem2.out
```

Figura 12: Pipeline de montagem do SOAPdenovo

4.2.3.4 Montagens MiSeq

Para as montagens das sequências MiSeq (Gir 1, Guzerá 3) testamos os dois diferentes programas, neste caso, diferentemente das *reads* SOLiD, só tínhamos uma biblioteca (700pb) para o MiSeq.

- SOAPdenovo

  Os mesmos parâmetros do SOAPdenovo foram utilizados nessa montagem.

- ABySS

Os parâmetros utilizados para o ABySS estão descritos na Figura 13, onde –pe indica sequências paired-end, –j: número de processadores, k: valor de k-mer, n: número mínimo de pares necessários para considerar unir dois *contigs*, lib: bibliotecas, se: fragmentos (*reads* sem pares).

```
abyss-pe -j100 k=31 n=10 name=Guzera lib='lib1 lib2' \
    lib1='lib1_1.fa lib1_2.fa' lib2='lib2_1.fa lib2_2.fa' \
    se='se1.fa se2.fa'
```

Figura 13: Arquivo de configuração do ABySS

4.2.3.5 Montagem PacBio

Primeiramente foi realizada a correção das *reads* PacBio pelas sequências da Plataforma MiSeq. Em seguida as sequências corrigidas foram montadas *de novo*. O programa PacBiotoCA [http://wgs-assembler.Sourceforge.net/wiki/index.php/PacBioToCA] foi adotado para essas duas etapas.

4.2.3.6 Montagens híbridas – SOLid + MiSeq

Para essa estratégia de montagem híbrida as duas bibliotecas do SOLiD mais a única biblioteca do MiSeq foram integradas. Nesse caso apenas o programa SOAPdenovo foi testado, isso porque este apresentou o melhor resultado para as plataformas sozinhas, como será descrito na sessão resultados e também porque o ABySS não acrescentaria as *reads* do SOLiD na formação dos *contigs*.

A construção da estratégia de montagem híbrida contou com diferentes parametrizações, como alternância na ordem de montagem do programa (*rank* das bibliotecas), indicação da construção de *contigs*, *scaffolds*, fechamento de *gaps* ou ambas as opções, utilização ou não dos fragmentos, mais a alternância entre os diferentes valores de k.

4.2.3.7 Montagens HiSeq

Assim como as sequências oriundas do MiSeq, para o HiSeq só tínhamos uma biblioteca para cada animal. Lembrando que neste caso, apesar de serem as mesmas raças (Gir e Guzerá), os animais sequenciados não foram os mesmos das estratégias anteriores.

Aqui também testamos os dois diferentes programas, seguindo os mesmos parâmetros já descritos.

4.2.3.8 Montagens híbridas – HiSeq + MiSeq + SOLiD

Mesmo se tratando de indivíduos diferentes, submetemos todos os três animais de cada raça à montagem híbrida, unindo os dados das três diferentes plataformas. Os dados da correção

das sequências do PacBio não foram incluídas nessa etapa do trabalho devido a baixa cobertura das *reads* sobre o genoma.

Mais uma vez, diferentes parâmetros foram testados (como descritos na montagem híbrida do SOLiD + MiSeq).

4.2.4 Avaliação das Montagens

As montagens foram validadas através de *scripts* desenvolvidos pela equipe do Grupo de Genômica e Biologia Computacional da FIOCRUZ-Minas. O *script* calcN50.pl foi desenvolvido para calcular o valor de N50, N90 dos *contigs* e *scaffolds*, bem como a quantidade, média, mediana, tamanho quantidade de bases dos mesmos (*contigs/scaffolds*). A outra métrica de avaliação de qualidade de montagem escolhida foi a avaliação da cobertura das bases em extensão no genoma, essa etapa foi sugerida pelo projeto de competição de montagem de genomas Assemblathon2 [BRADNAM *et al*., 2013] e será descrita a seguir.

4.2.4.1 Mapeamento dos contigs contra o genoma referência de *Bos taurus* (raça Hereford)

Uma das métricas de avaliação de montagem é a análise da cobertura do genoma em relação a uma referência, sendo que esta referência pode ser de outro organismo (mesma espécie) ou em caso de sequenciamento de fosmídeos a utilização destes.

Neste trabalho, como não temos dados de fosmídeos e nem o genoma da mesma raça sequenciado, o genoma de *Bos taurus* (Hereford) foi escolhido como referência. A versão abordada deste genoma foi a UMD3.1, pois devido ao artigo publicado em 2012 [ZIMIN *et al*., 2012] os autores analisaram todas as versões disponíveis desse genoma e chegaram a conclusão que esta versão seria a melhor, baseado em observações do tipo duplicações, fragmentação entre outros.

O mapeamento foi executado para as três melhores montagens dos resultados estatísticos para cada animal. Para isso selecionamos dois diferentes programas para avaliar a cobertura em extensão e profundidade dos genomas montados. Os dois diferentes programas de mapeamento foram: BWA [http://bio-BWA.sourceforge.net/] e SOAPAligner [http://soap.genomics.org.cn/soapaligner.html] selecionados para mapear os *contigs* contra a referência.

O pacote BedTools [https://bedtools.readthedocs.org/] foi integrado para as comparações dos genomas. Esse pacote foi utilizado a fim de responder a questão da sobreposição entre os conjuntos de dados, neste caso o conjunto de dados foi o resultado do

mapeamento dos *contigs* contra a referência do Hereford (UMD3.1) no formato de arquivo Bam, característico de mapeamento [detalhes do formato: http://samtools.github.io/hts-specs/SAMv1.pdf]. Contudo, foi possível analisar o quanto cada biblioteca do sequenciamento contribui em cobertura de extensão e/ou só profundidade para o fechamento do genoma e também para contabilizar o quanto cada biblioteca poderia sobrepor a outra.

4.2.5 Infraestrutura de informática

Para o desenvolvimento desse trabalho foram utilizadas as plataformas de alto desempenho em bioinformática da FIOCRUZ – Centro de Excelência em Bioinformática (CEbio) e da Universidade da Geórgia - *Georgia Advanced Computing Resource Center* (GACRC).

Da plataforma do CEbio utilizamos um único servidor onde todos os programas estão instalados:

- SGI Autix UV 100 128 Cores / 4 sockets Intel Xeon Octo-Core E78837 de 2.66-GHz, com 24MB cachê, 2(dois) Discos de 600-GB SAS 10K RPM; 2-TB de Memória DDR3 1066 MHz. SO Red Hat Enterprise Linux 6.

Da plataforma da GACRC utilizamos o Z-cluster (Linux Cluster):

- O Linux cluster é composto por nós de computação com 4, 6, 8, e 12 núcleos com processadores da Intel e AMD. Subconjuntos de nós com "grande memória" (por exemplo, 128, 256 ou 512 GB de RAM), enquanto outros têm capacidades de conectividade ou GPU InfiniBand. Potência total de computação CPU é de 25,9 Tflops.

Computadores de uso pessoal foram utilizados para as análises filogenéticas e todas as análises pós-montagem do genoma mitocondrial:

- Linux: SO Ubuntu. Processador: Intel® Core™2 Duo CPU E7400 @ 2.80GHz x 2. 4GB Ram, 250 HD.
- MacBook Pro Retina: SO: OS X Yosemite. 2,5GHz dual core. Intel Core i5 (3,1GHz) com 3MB de cache L3. 8GB Ram, 250 HD.

4.3 RESULTADOS

Para facilitar a compreensão, os resultados serão apresentados seguindo uma ordem lógica e não cronológica do que foi realmente foi realizado.

4.3.1 Pré-processamento dos dados

Após avaliação da qualidade dos dados pelos programas FastQC e SMRTanalysis (PacBio) todas as sequências foram submetidas a etapa de pré-processamento, visto que a qualidade dos dados estava abaixo do ideal para se iniciar uma montagem. Dados completos dos relatórios de qualidade podem ser acessados no material suplementar online.

4.3.1.1 Reads SOLiD

Dois conjuntos de dados foram gerados para os dados de SOLiD. O primeiro conjunto apresentado é aquele cujo valor de PHRED foi estabelecido como 30 (Figura 14) e o segundo conjunto estabeleceu-se o valor de PHRED 20 (Figura 15). Essa decisão foi tomada visto que com o valor 30 grande quantidade de dados foi perdida. Os resultados de ambos os valores de PHRED foram comparados a fim de estabelecer qual o melhor conjunto de dados para aplicar a montagem *de novo*.

A Figura 16, mostra a quantidade de dados gerada pelo sequenciador SOLiD para os indivíduos de cada raça e a as duas bibliotecas (1-2kb e 3-4kb), os resultados da filtragem dos dados com um parâmetro de PHRED 20 e outro sendo um pouco mais rigorosos usando PHRED 30.

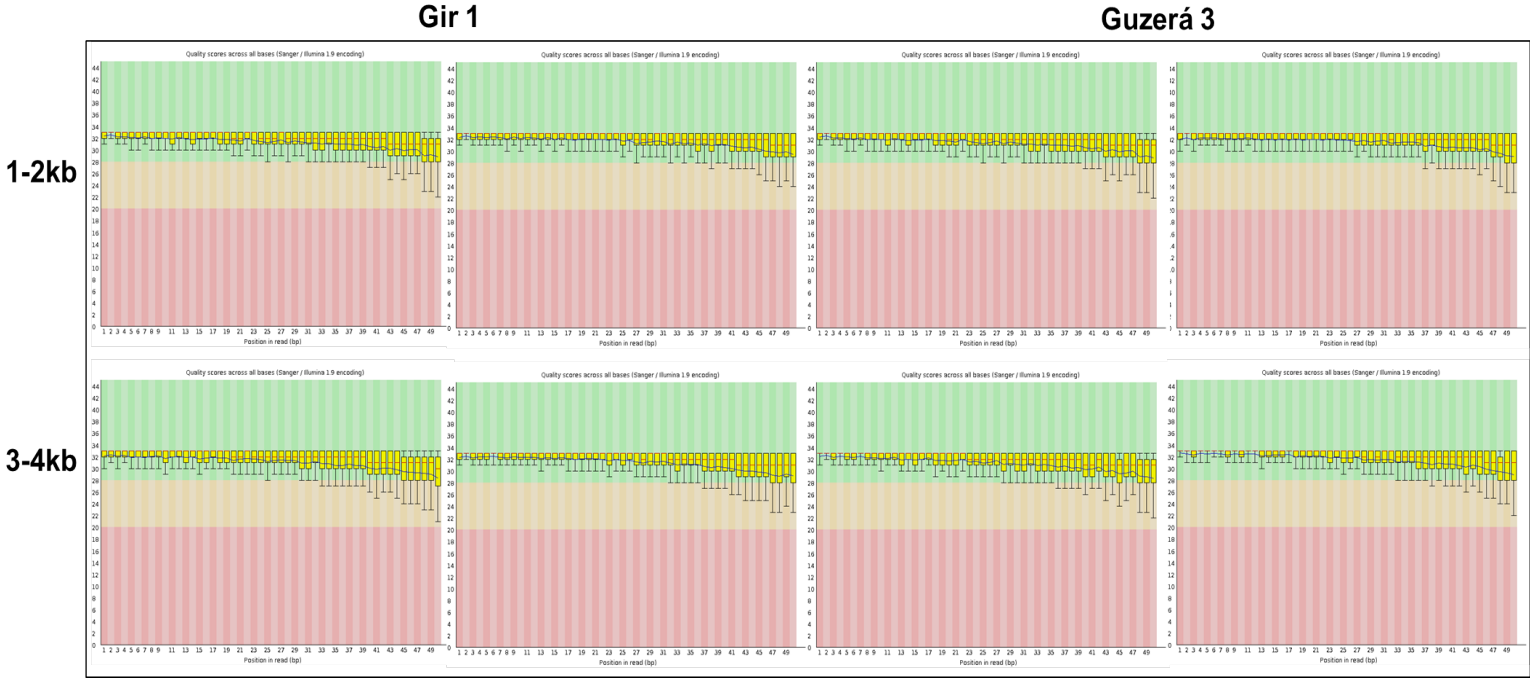


Figura 14: Qualidade por base das *Reads* SOLiD PHRED30: Qualidade por base obtida através da ferramenta FASTQC da sequência, após a filtragem dos dados por valor de PHRED30. Aqui são mostrados os dados senso e anti senso para as duas bibliotecas e as duas raças. As duas primeiras figuras superiores e inferiores representam a raça Gir, o primeiro gráfico superior e inferior representam o senso e os segundos gráficos são anti senso, o mesmo se aplica para o Guzerá. O eixo X de cada gráfico representa o score de qualidade, dividido em alta qualidade (verde, 28 a40), média (laranja, 20 a 28) e baixa (rosa, 0 a20). O eixo Y representa a posição da base nas *reads* (0 a 50 pb). A linha central vermelha é o valor mediano, a linha azul é a qualidade média, a caixa amarela representa o intervalo interquartil (25 75%), os segmentos verticais representam o maior e menor valor observado.

Nota-se nas Figuras 14 e 15, que as bases no final da sequência tiveram uma qualidade menor, pois na maioria das plataformas a qualidade vai diminuindo ao longo da corrida.

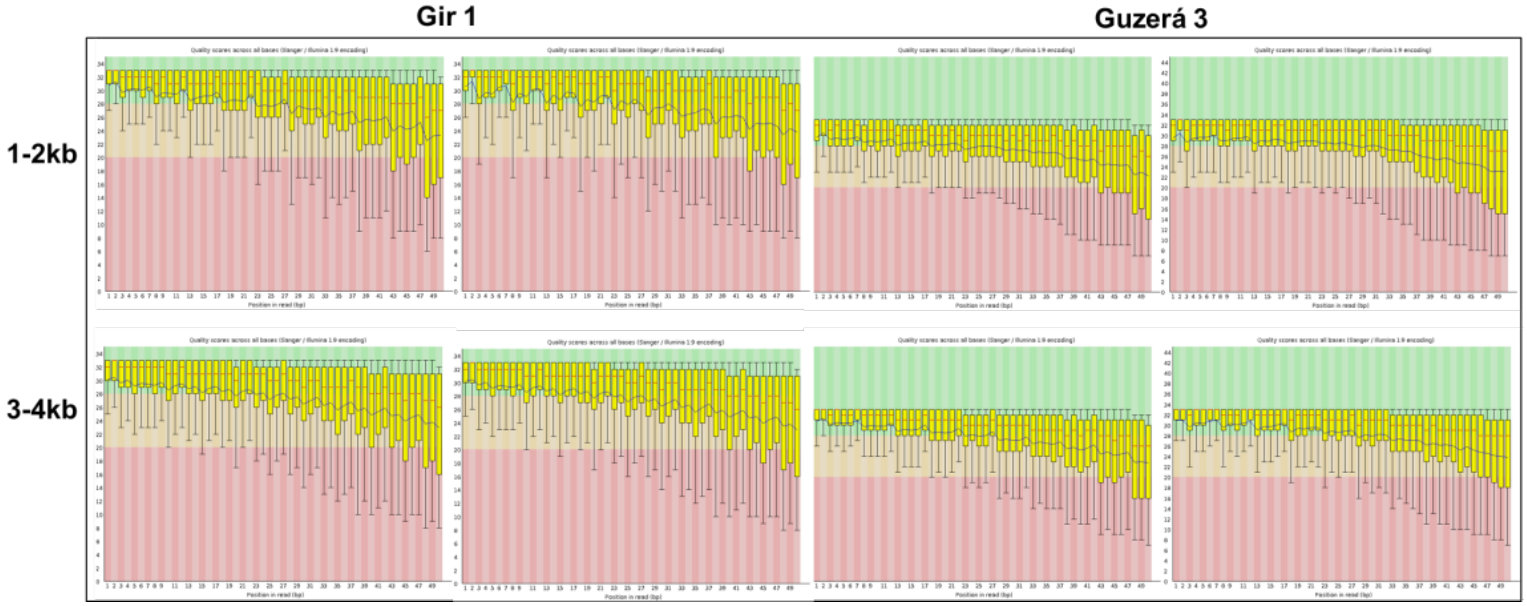


Figura 15: Qualidade por base das *Reads* SOLiD PHRED20: Qualidade por base obtida através da ferramenta FASTQC da sequência, após a filtragem dos dados por valor de PHRED20. Aqui são mostrados os dados senso e anti senso para as duas bibliotecas e as duas raças. As duas primeiras figuras superiores e inferiores representam a raça Gir, o primeiro gráfico superior e inferior representam o senso e os segundos gráficos são anti senso, o mesmo se aplica para o Guzerá. O eixo X do gráfico representa o score de qualidade, dividido em alta qualidade (verde, 28 a40), média (laranja, 20 a 28) e baixa (rosa, 0 a20). O eixo Y representa a posição da base nas *reads* (0 a 50 pb). A linha central vermelha é o valor mediano, a linha azul é a qualidade média, a caixa amarela representa o intervalo interquartil (25 75%), os segmentos verticais representam o maior e menor valor observado.

Como pode ser observado na Figura 16, mais da metade das sequências são descartadas quando filtradas por um valor de qualidade PHRED20 e mais de 80% dos dados são eliminados com o valor de qualidade mais alto. É válido ressaltar que além do valor de PHRED também utilizamos como parâmetro o tamanho total da *read*, isso porque as sequências geradas pelo SOLiD apresentam um tamanho pequeno de 50pb e valores menores do que estes poderiam dificultar ainda mais a montagem desses genomas.

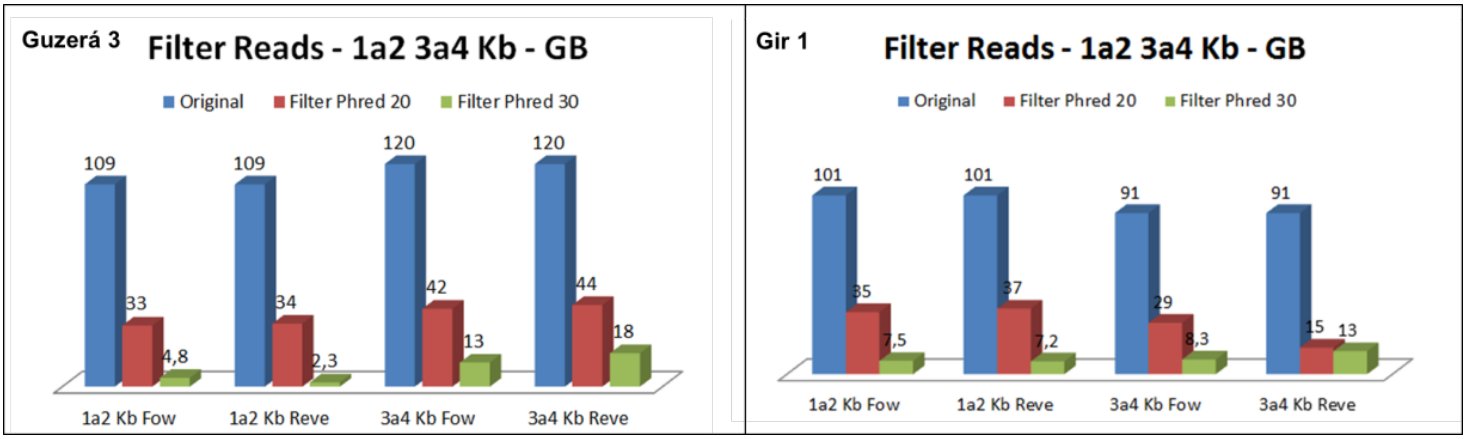


Figura 16: Dados reads SOLiD antes e após filtragem e correção (PHRED20 e 30): Representação da quantidade de dados geradas pelo sequenciador SOLiD para as raças Guzerá (esquerda) e Gir (direita). As barras em azul mostram a quantidade de dados originais geradas pelo sequenciador em GB. As barras vermelhas mostram as reads após o filtro de qualidade pelo valor de PHRED20 e as barras em verdes mostram os resultados após filtragem das reads pelo valor de PHRED30.

Além da qualidade por base, outros parâmetros foram avaliados e estão disponíveis no arquivo suplementar online: nível de duplicação, perfil de k-mers, conteúdo GC por base, conteúdo de n por bases, qualidade por base, distribuição por tamanho.

Após o pré-processamento dos dados a cobertura alcançada foi de 5,8x para o Gir 1 e 6,4x para o Guzerá 3.

4.3.1.2 Reads MiSeq e HiSeq

Para as sequências oriundas dos sequenciadores da Illumina: MiSeq e HiSeq, o valor de qualidade de PHRED foi igual a 30, com o tamanho mínimo da *read* de 75pb. A penalidade para essas sequências foram ser cortadas em uma janela deslizante de até cinco nucleotídeos. Contudo é possível observar nas Figuras 17 e 18(A,B), que o novo conjunto de dados é formado apenas por sequências de alta qualidade. Diferente do SOLiD as *reads* após filtragem para o valor de qualidade de PHRED igual a 30 correspondem a mais de 88% do valor inicial, tanto para os dados MiSeq quanto para o HiSeq, o que pode caracterizar uma alta qualidade da química utilizada pelo sequenciador.

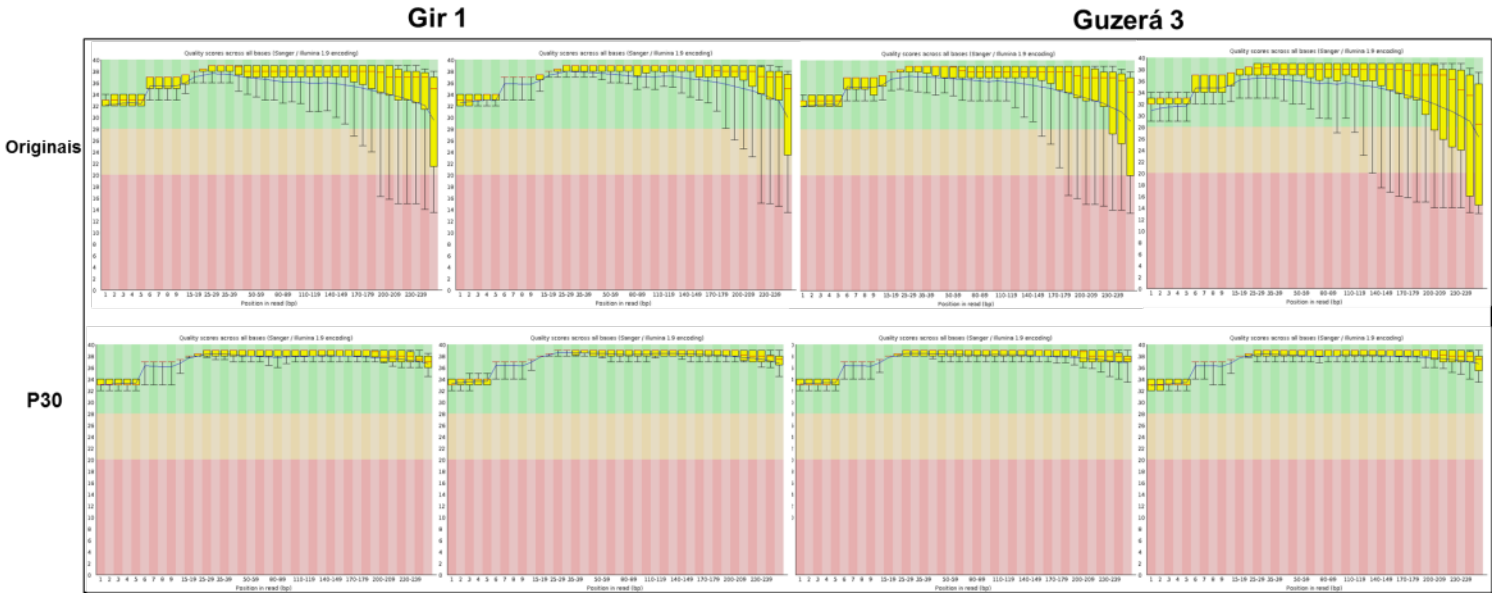


Figura 17: Qualidade por base das *Reads* MiSeq: Qualidade por base obtida através da ferramenta FASTQC da sequência, antes (Figura superior) e após a filtragem dos dados por valor de PHRED30 (Figura inferior). Aqui são mostrados os dados senso e anti senso para as duas bibliotecas e as duas raças. As duas primeiras figuras superiores e inferiores representam a raça Gir, o primeiro gráfico superior e inferior representam o senso e os segundos gráficos são anti senso, o mesmo se aplica para o Guzerá. O eixo X de cada gráfico representa o score de qualidade, dividido em alta qualidade (verde, 28 a40), média (laranja, 20 a 28) e baixa (rosa, 0 a 20). O eixo Y representa a posição da base nas *reads* (0 a 250 pb). A linha central vermelha é o valor mediano, a linha azul é a qualidade média, a caixa amarela representa o intervalo interquartil (25-75%), os segmentos verticais representam o maior e menor valor observado.

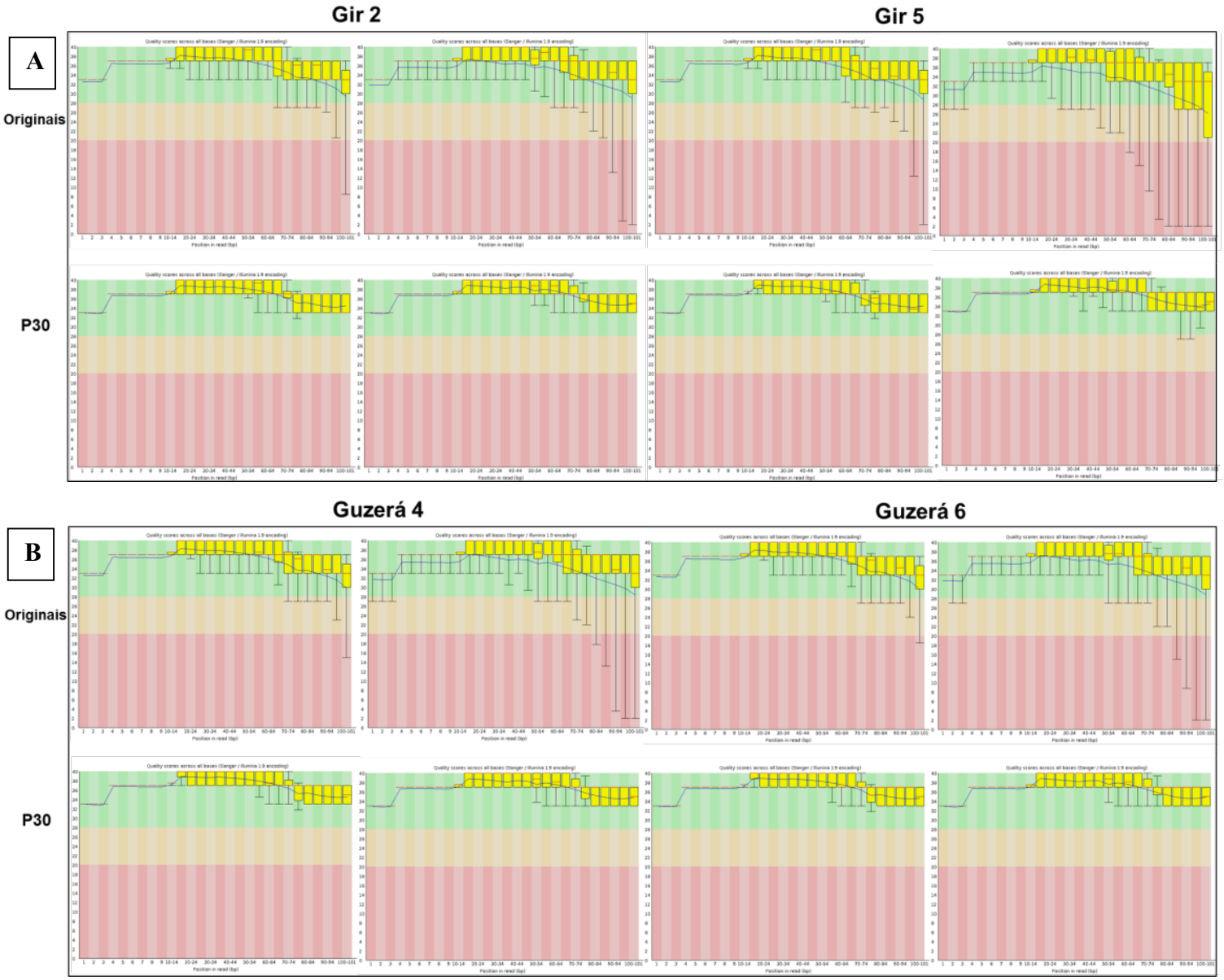


Figura 18 A e B: Qualidade por base das *Reads* HiSeq: Qualidade por base obtida através da ferramenta FASTQC da sequência, antes (Figura A e B superior) e após a filtragem dos dados por valor de PHRED30 (Figura A e B inferior). Aqui são mostrados os dados senso e anti senso para os 4 animais (2 de cada raça). As duas primeiras figuras superiores e inferiores (A) representam o animal Gir 2, o primeiro gráfico superior e inferior representam o senso e os segundos gráficos são o anti senso, o mesmo se aplica para os demais animais.
O eixo X de cada gráfico representa o score de qualidade, dividido em alta qualidade (verde, 28 a40), média (laranja, 20 a 28) e baixa (rosa, 0 a 20). O eixo Y representa a posição da base nas *reads* (0 a 100 pb). A linha central vermelha é o valor mediano, a linha azul é a qualidade média, a caixa amarela representa o intervalo interquartil (25-75%), os segmentos verticais representam o maior e menor valor observado.

Após o pré-processamento dos dados a cobertura alcançada foi de 1,5x para o Gir 1 e 1,9x para o Guzerá 3 (MiSeq). Para os dados HiSeq: 13,9x Gir (2), 12,1x Gir (5), 11,7x Guzerá (3), 14,9x Guzerá (6).

4.3.1.3 Reads PacBio

Diferente de todas as outras plataformas, o PacBio apresenta uma outra maneira de avaliar e filtar as sequências geradas por qualidade. Para essa avaliação o programa SMRTAnalysis da PacBio foi selecionado para gerar as '*filtered* sub-*reads*" do instrumento. O termo "*sub-reads*" refere-se à parcela de uma *read*. Filtragem refere-se a um processo no programa para identificar a qualidade da leitura. "*Filtered* sub-*reads*" são gerados seguindo uma análise primária em que os adaptadores do sequenciador são separados das longas *reads*, e as bases de baixa qualidade relatados pelo instrumento são removidos, dando origem a sub-*reads* de 1000 bases de comprimento, em média. A segunda etapa do programa não foi utilizada para correção, para esta etapa selecionamos o programa PacBioToCa que será descrito na seção montagem das *reads* PacBio.

As sequências foram submetidas a dois diferentes testes de qualidade: 75% e 80%, representados nas Figuras 19 e 20. Diferentemente da abordagem para as *reads* SOLiD, neste caso a montagem só foi realizada com os valores 75% devido a baixa quantidade de dados gerados com o sequenciamento PacBio. Com o valor de qualidade 75%, 30% das sequências foram descartadas, enquanto que para o valor mais alto da qualidade apenas 28% não foram descartadas. Após a filtragem dos dados a cobertura passou para 0,13X para cada animal.

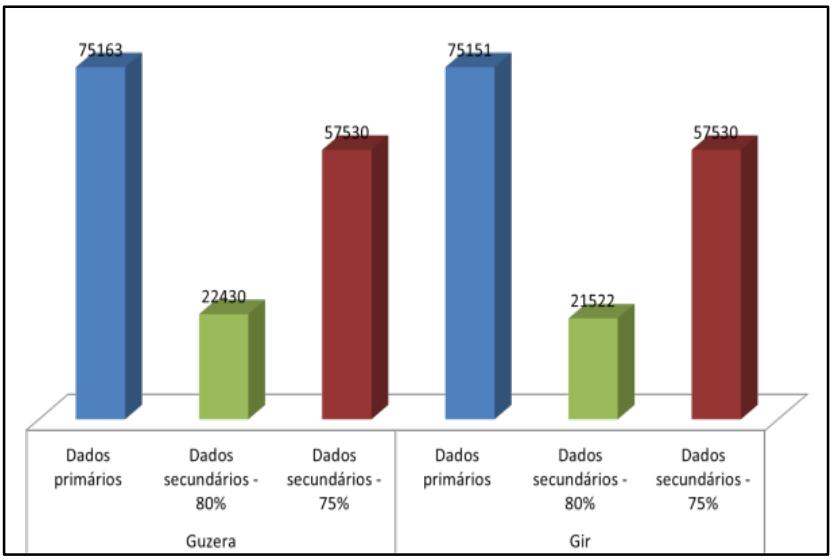


Figura 19: Dados PacBio antes e após filtragem (V=75% e 80%): 1 - Representação da quantidade de dados geradas pelo sequenciador PacBio para as raças Guzerá (esquerda) e Gir (direita). As barras em azul mostram a quantidade de dados originais geradas pelo sequenciador. As barras vermelhas mostram as *reads* após o filtro de qualidade pelo valor de 75% e as barras em verdes mostram os resultados após filtragem das *reads* pelo valor 80%.

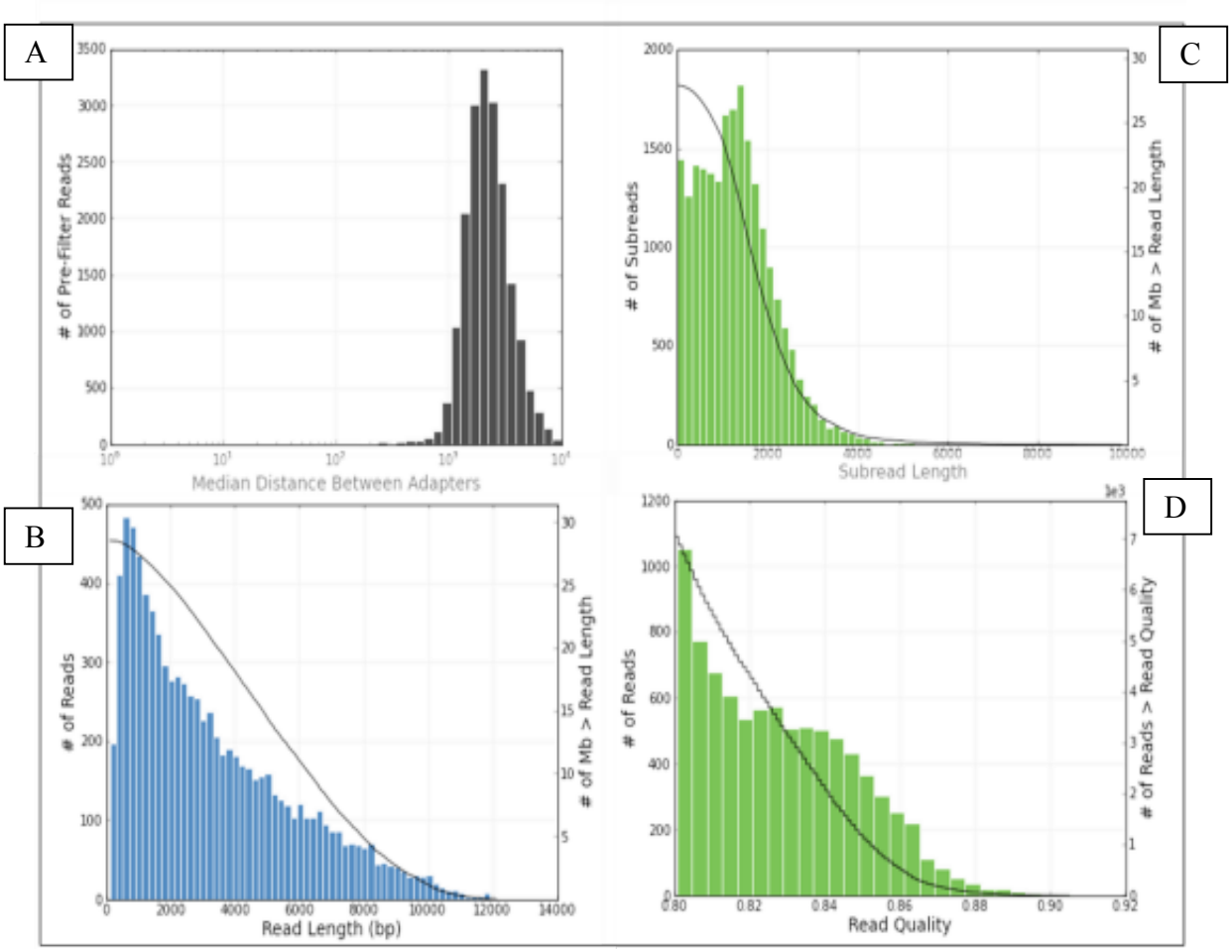


Figura 20: Dados qualidade PacBio: A= Distância média dos adaptadores, B= tamanho das *reads* (antes filtragem), C= Tamanho das sub*reads* D= Qualidade das sub*reads* (após filtragem).

4.3.2 Montagem *de novo*

Ao total 100 diferentes montagens para cada animal utilizando diferentes parâmetros e diferentes programas foram realizadas.

4.3.2.1 Melhores Valores de K-mer

Os resultados dos valores de K foram computados pelo programa Kmergenie. A Figura 21 ilustra um dos resultados obtidos para o Guzerá indivíduo 6 (HiSeq, 100pb). Nesse caso o programa gerou dois melhores valores de k: 31 e 37. Para o processo de montagem foram selecionados os valores sugeridos pelo Kmergenie bem como os valores k=29, k=33, k=35 e k= 39, podendo trabalhar assim com os intervalos destes valores.

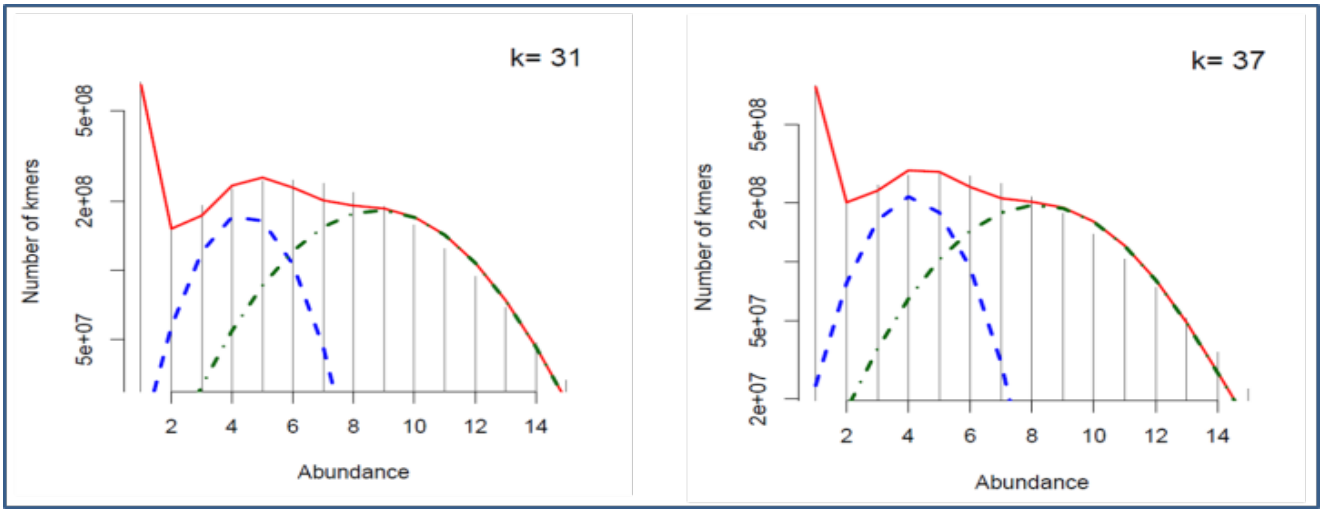


Figura 21: Melhores valores de k-mer: A linha vermelha é o ajuste do modelo estatístico completo do histograma (k-mers errados + k-mers genômicos). Para o modelo diplóide, verde representa apenas k-mers heterozigotos, azul são apenas os homozigotos.

Os resultados para os dados do SOLiD (50pb) foram: 23 para o Gir, 31 para o Guzerá. MiSeq (250pb): 27, 31, 67 para o Gir e 31 e 67 para o Guzerá. HiSeq (100pb) Gir indivíduo 2: 31 e 33, Gir indivíduo 5: 35 e Guzerá indivíduo 4: 35.

Os resultados completos dos gráficos podem ser visualizados no material suplementar online.

4.3.2.2 Melhores Montagens

A primeira decisão das melhores montagens levou em consideração apenas resultados estatísticos, como maior valor de N50, menor quantidade de *contigs* e maior cobertura em bases do genoma, selecionando-se as três melhores de cada programa por animal sequenciado para serem representados aqui e para dar procedência as próximas métricas de avaliação.

- Resultados das análises estatísticas

Nessa sessão é apresentada a média das três melhores montagens para cada estratégia (diferentes plataformas, combinações entre bibliotecas e plataformas). Os resultados completos para cada estratégia podem ser visualizados no material suplementar online.

4.3.2.3 Resultados das Montagens das *reads* SOLiD

A Figura 22 apresenta o resultado da comparação entre os valores médios de N50 das três melhores montagens da estratégia de montagem das sequências SOLiD (resultados plotados apenas para os dados de *contigs* gerados). Para essa montagem dois conjuntos de dados com valores de PHRED20 e PHRED30 para as bibliotecas de 1-2kb sozinha, 3-4kb sozinha e a união das duas bibliotecas, foram selecionados.

Os valores de N50 variaram de 250pb á 400pb, lembrando que o tamanho das sequências é de 50pb. A primeira comparação a ser evidenciada é a de diferentes valores de PHRED. Como podemos perceber com um valor mais rigoroso a montagem apresenta melhores resultados, o que significa uma montagem com mais bases dentro dos *contigs*, sendo, portanto menos fragmentada.

A segunda comparação em relação aos resultados da Figura 22 é sobre a adição das duas bibliotecas juntas contribuem para uma montagem menos fragmentada. No caso do Gir 1 com valor de PHRED20 esse padrão não foi visualizado o que pode ter sido devido à inserção de bases errôneas pelo não tão acurado filtro de qualidade das bases.

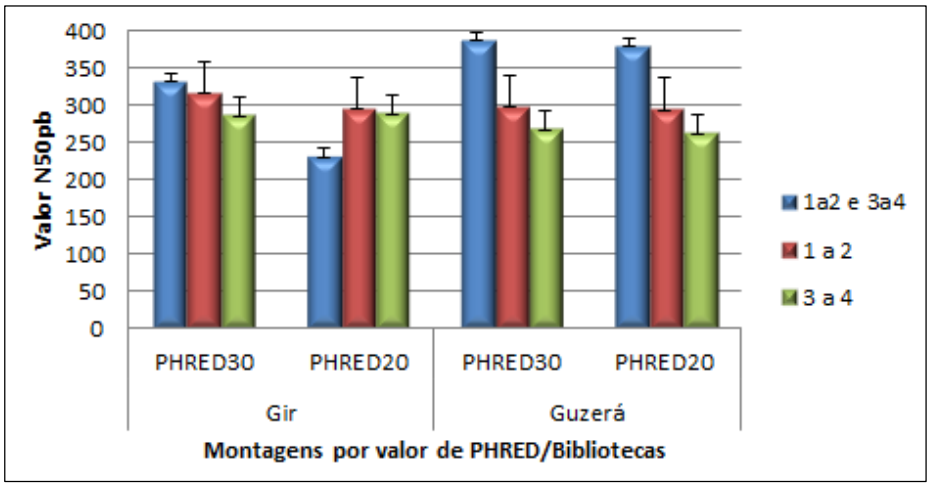


Figura 22: Média dos valores de N50 das 3 melhores montagens SOLiD: A Figura apresenta as montagens para as duas raças, Gir (esquerda) Guzerá (direita). Dois valores de PHREDs foram testados e estão indicados na Figura. As barras azuis apresentam os resultados das montagens das duas bibliotecas juntas, as barras vermelhas a biblioteca de 1-2kb e as barras verdes representam as montagens da biblioteca de 3-4kb. O eixo X representa os valores de N50 indo de 0 à 400pb e o eixo Y os animais por valor de qualidade. O desvio padrão é referente ao desvio encontrado entre as três melhores montagens para cada estratégia.

Em relação às outras métricas avaliadas como quantidade de *contigs* e cobertura das bases sobre os genomas, estes, assim como os valores de N50 não se mostraram muito satisfatórios. A média da cobertura das bases sobre o genoma encontradas nas montagens com o valor de PHRED30 para as duas bibliotecas foi de 40% para o Guzerá 3 e apenas 20% para o Gir indivíduo 1. Em relação a quantidade de *contigs* para ambas as bibliotecas, a média da quantidade foi de 2.603.451 para o Guzerá 3 e 1.191.745 para o Gir 1.

Os resultados da avaliação dessas métricas (N50, quantidade de *contigs* e cobertura das bases) indicam que as nossas melhores montagens obtidas não conseguem cobrir nem 50% do tamanho do genoma e que as bases que o cobrem ainda estão muito fragmentadas.

Podemos atribuir estes resultados a vários fatores: a baixa cobertura inicial das bases para começar o processo de montagem (~6x), a utilização de apenas duas bibliotecas de tamanhos próximos e mesmo tipo (*Mate-pair* 1-2 e 3-4kb) e o pequeno comprimento das sequências 50pb.

Vale ressaltar que esses *contigs* são os resultados brutos pós montagem, ou seja, eles podem conter 200pb. Para fechamento de genomas, certamente os pequenos *contigs* são ignorados, o que aumenta consideravelmente o valor de N50.

4.3.2.4 Resultados das montagens das *reads* MiSeq

A Figura 23 apresenta o resultado da comparação entre os valores de N50 da estratégia de montagem com as sequências do MiSeq (dados de *contigs* gerados). Para essa montagem dois diferentes programas foram testados: SOAPdenovo e ABySS.

Quando comparado aos resultados do SOLiD foi possível obter uma melhora nos valores de N50 dos *contigs*, o que já era esperado visto que as *reads* MiSeq tem o tamanho de 250pb. O programa SOAPdenovo mostrou melhores resultados que o ABySS para ambos os animais. Os valores de N50 variaram de 350 a 500pb em ambos os programas.

Em relação a quantidade de *contigs* gerados e quantidade de bases totais, novamente o SOAPdenovo apresentou melhores resultados que o ABySS, sendo os resultados para o SOAPdenovo de: 30% de cobertura das bases no genoma do Gir em 2.472.854 *contigs* e 42% de cobertura do Guzerá em 2.740.334 *contigs*. Já para o ABySS: 15% de cobertura do Gir em 1.200.483 *contigs* e 25% do Guzerá 3em 1.744.682 *contigs*.

Apesar de apresentar uma cobertura maior sobre o genoma do que quando usamos o SOLiD, a montagem com o MiSeq também apresenta limitações em relação à pouca quantidade de dados iniciais para as montagens, o que resultou em uma montagem incompleta e fragmentada. Nesse caso podemos atribuir os resultados também a alguns fatores, além da baixa cobertura inicinal(~2x), como por exemplo, a utilização de apenas uma biblioteca para cada raça (*Paired-end* com pequeno tamanho de inserto ~700pb).

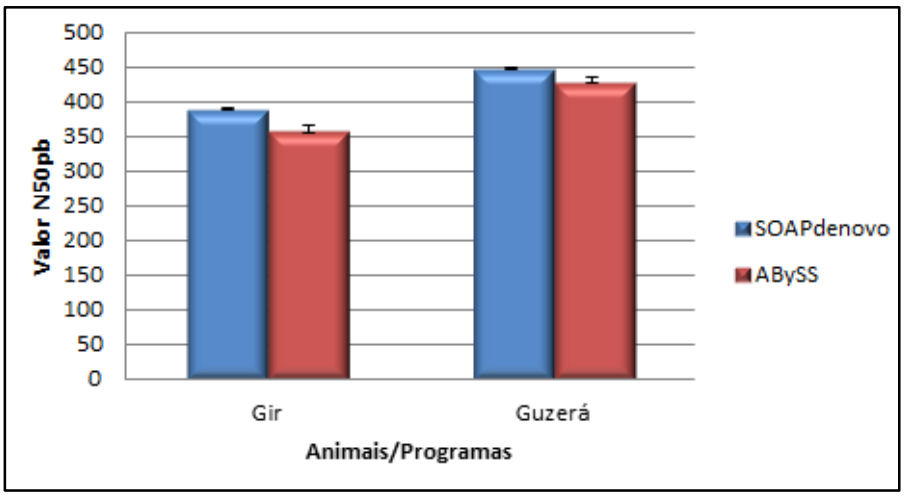


Figura 23: Média dos valores de N50 das três melhores montagens MiSeq: A Figura apresenta as montagens para as duas raças, Gir (esquerda) Guzerá (direita). Dois porgoramas foram testados e estão representados pela cor das barras: azul (SOAPdenovo) vermelhor (ABySS). O programa SOAPdenovo foi superior ao ABySS em relação ao N50 para as duas raças analisadas. O eixo X representa os valores de N50 indo de 0 á 500pb e o eixo Y os animais por programa. O desvio padrão é referente ao desvio encontrado entre as três melhores montagens para cada estratégia.

4.3.2.5 Resultados das Montagens Híbridas: SOLiD + MiSeq

Visto que os mesmos animais foram sequenciados com as plataformas SOLiD e MiSeq, a estratégia de montagem híbrida entre essas plataformas foi realizada. A hipótese para realização dessa estratégia consistiu em que as sequências obtidas em cada uma das plataformas poderiam ser complementares e assim melhorar a montagem final.

A Figura 24 apresenta os valores médios de N50 obtidos das três melhores montagens para cada animal. Uma vez que na etapa anterior foi possível observar que o SOAPdenovo era um programa mais adequado para lidar com os dados do presente trabalho, apenas esse programa foi utilizado.

Os valores de N50 variaram de 500 a 700pb para ambos os animais. A cobertura média das bases sobre o genoma foi de 40% para o Gir em 3.000.000 *contigs* , 54% para o Guzerá em 2.556.641 *contigs*.

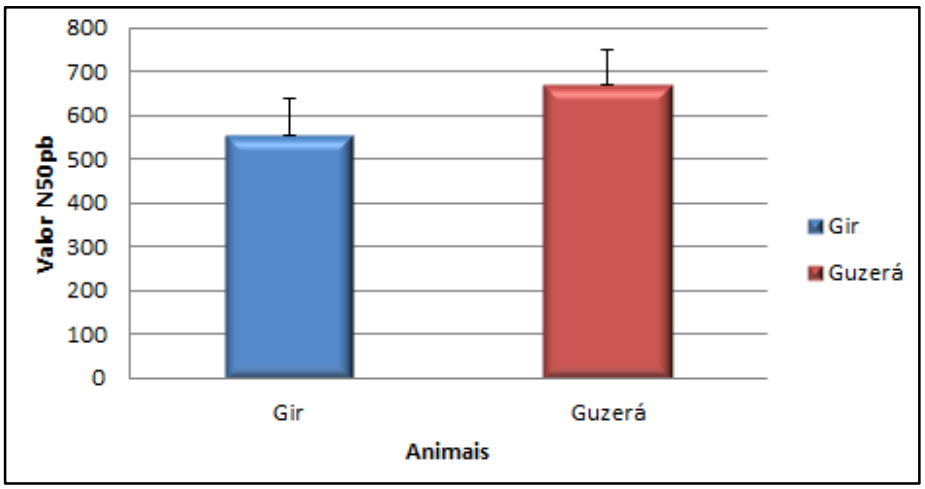


Figura 24: Média dos valores de N50 das três melhores montagens Híbridas: SOLiD + MiSeq: A Figura apresenta as montagens para as duas raças, Gir (esquerda, azul) Guzerá (direita, vermelho). O eixo X representa os valores de N50 indo de 0 á 800pb e o eixo Y os animais. O desvio padrão é referente ao desvio encontrado entre as três melhores montagens para cada estratégia.

Com a união das sequências das duas plataformas a cobertura inicial dos dados aumentou, passando de aproximadamente 5x para ~7x. É provável que a maior contribuição seja devido ao diferente tamanho dos insertos. Ainda que a montagem permaneça muito fragmentada e incompleta, esses resultados são melhores do que quando utilizamos uma única plataforma, evidenciando que a montagem híbrida é uma boa estratégia e deve ser utilizada em projetos de montagens de genomas grandes.

4.3.2.6 Resultados das Montagens PacBio

Dos programas escolhidos para fazer as montagens de genomas nesse trabalho, o PacBioToCA (Celera assembler) é o único que utiliza a estratégia OLC (*overlap, layout e consensus*). O grande fator limitante desse tipo de programa é a ineficiência em processamento de grandes dados, a vantagem é a maior acurácia.

A Tabela 7 apresenta o resultado da montagem das sequências PacBio corrigidas pelas sequências MiSeq. É possível observar o quão melhor fica o valor de N50 quando comparado as estratégias anteriores, isso porque as *reads* PacBio são maiores quando comparado as demais plataformas. Entretanto, em termos de cobertura do genoma, esses dados não tem valor significativo, uma vez que a cobertura esperada foi menor do que 1% do genoma. Devido a este fato, resolvemos não utilizar as sequências PacBio para as demais etapas de montagem híbrida.

Tabela 7: Montagem PacBio

N. *contigs*	1.094
Total bases	493.645
Longest contig	24.071
N50	5.927
N90	2.406
N95	1.830
Cobertura	0,00018

4.3.2.7 Resultados das montagens HiSeq

A Figura 25 apresenta o resultado da comparação entre os valores médios de N50 da estratégia de montagem das sequências HiSeq por meio do programa SOAPdenovo. Os animais sequenciados nessa plataforma não são os mesmos que os utilizados pelo MiSeq e SOLiD. Os animais sequenciados no Hiseq foram o Gir 2 e 5 e Guzerá 4 e 6.

Para a montagem das sequências dessa plataforma foram realizadas duas estratégias: na primeira estratégia um único animal de cada raça foi montado e na segunda estratégia as sequências dos dois animais foram unidas (não misturando as raças, apenas os indivíduos).

Os valores médios de N50 variaram de 900pb a 1,2kb para ambos os animais. Quando unimos as bibliotecas dos diferentes animais não observamos melhoras no valor de N50 quando comparamos ao melhor valor individual. Percebemos melhoras quando analisamos o Gir indivíduo 5 que quando unido ao Gir indivíduo 2 aumentou seu valor de N50.

A cobertura média das bases sobre o genoma e a quantidade de *contigs* foram as seguintes: Gir (2) cobertura de 69% em 2.286.380 *contigs*, Gir (5) cobertura de 51% em 2.075.433 *contigs*, Gir (2e5) cobertura de 69% em 2.312.153 *contigs*, Guzerá (4) cobertura de 68% em 2.407.142 *contigs*, Guzerá (6) cobertura 69% em 2.246.759 *contigs*, Guzerá (4e6) cobertura de 69% em 2.262.155 *contigs*. Assim como nos valores de N50 não foi possível perceber uma melhora ao unirmos os animais.

Montagens com o ABySS também foram realizadas para os dados HiSeq, entretanto os resultados se mostraram inferiores ao SOAPdenovo. A média do valor de N50 foi de 780pb para ambos os animais com a cobertura de 58% do genoma. Os resultados completos podem ser visualizados no material suplementar online.

No caso das sequências Illumina Hiseq a cobertura inicial obtida das *reads* sobre o genoma para todos os animais foi melhor que das outras plataformas e considerada aceitável para se iniciar um processo de montagem, uma vez que foi observada uma cobertura de ~10x.

O fato da cobertura (após montagem) não ter sido melhorada quando unimos os animais pode ter sido devido as bibliotecas conterem o mesmo tamanho de inserto. Outro fator limitante é o pequeno tamanho do inserto (~300pb), este tipo de inserto é o ideal para formação de *contigs* [BRADNAMAN *et al.,* 2013], mas para isso a cobertura sobre o genoma deveria ser aumentada.

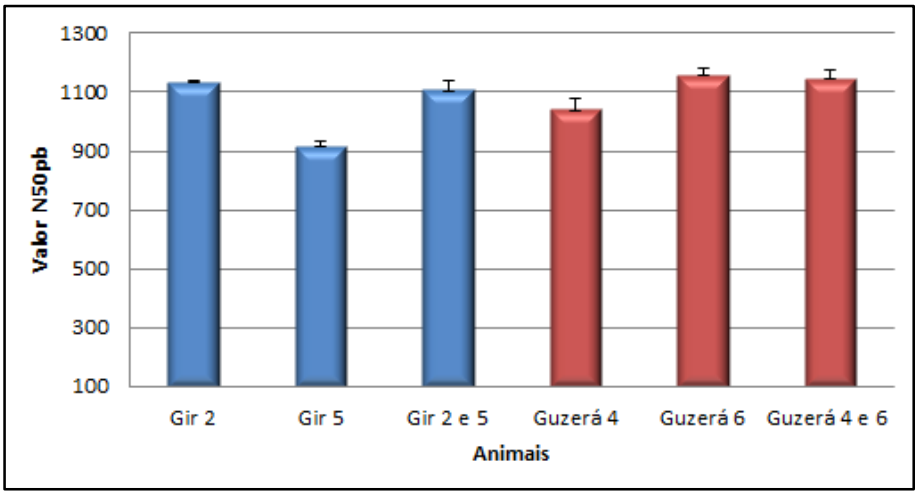


Figura 25: Média dos valores de N50 das 3 melhores montagens HiSeq: A Figura apresenta as montagens para as duas raças, Gir (esquerda, azul) Guzerá (direita, vermelho). O eixo X representa os valores de N50 indo de 100 a 1300pb e o eixo Y os animais. O desvio padrão é referente ao desvio encontrado entre as três melhores montagens para cada estratégia.

4.3.2.8 Resultados das montagens híbridas: SOLiD + MiSeq + HiSeq

Mesmo se tratando de diferentes indivíduos os dados de todas as plataformas (exceto PacBio) foram unidos. Essa estratégia foi realizada para sabermos o quanto essa montagem poderia ser melhorada (ou não) diante dessa estratégia.

A Figura 26 apresenta o resultado dessa estratégia por raça. A média dos valores de N50 foi de 1,2kb para o Gir e 1,0kb para o Guzerá. A cobertura média das bases sobre o genoma foi de 70% para o Gir e 68% para o Guzerá em 2.369.271 *contigs*.

Os resultados obtidos foram muito próximos de quando só os dados do HiSeq foram montados. No caso do Gir houve uma pequena melhora, para o Guzerá, os dados HiSeq sozinhos apresentaram um melhor resultado. O que podemos inferir desse resultado é que mesmo unindo todos os dados a cobertura inicial ficou muito similar ao do HiSeq (SOLiD + MiSeq = ~7x). Para discutir melhor o quanto cada biblioteca contribuiu para a montagem, análises mais específicas foram realizadas e serão apresentadas a seguir, no tópico de mapeamento e saturação das bibliotecas.

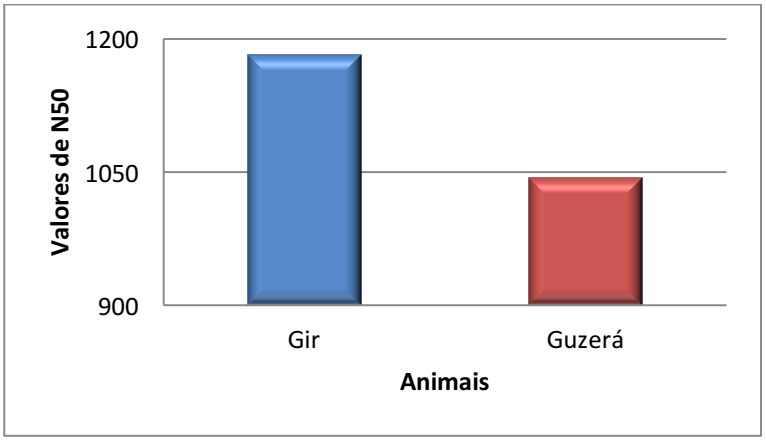


Figura 26: Média dos valores de N50 das montagens de todas as plataformas: A Figura apresenta as montagens para as duas raças, Gir (esquerda, azul) Guzerá (direita, vermelho). O eixo X representa os valores de N50 indo de 900 á 1200pb e o eixo Y os animais.

4.3.3 Resultados da análise de cobertura sobre o genoma de *Bos taurus*

Os *contigs* oriundos das melhores montagens (seguindo os critérios já descritos) foram mapeados contra o genoma de referência do *Bos taurus* (UMD3.1).

Para isso dois diferentes programas com os parâmetros iguais de penalidade de mapeamento foram utilizados: BWA e SOAPAligner (SOAP2).

O BWA mostrou melhores resultados que o SOAP2, sendo possível mapear maior quantidade de sequências. O algoritmo do BWA é capaz de lidar com sequências maiores que o SOAP2, o que pode ter contribuído nessa estratégia, em que os *contigs* (e não as *reads*) foram mapeados. A Tabela 8 apresenta o resultado da comparação dos resultados dos dois programas, tendo sido mapeado os *contigs* oriundos da montagem SOLiD gerados no programa SOAPdenovo. Os resultados completos também podem ser acessados no material suplementar online.

Tabela 8: Comparação *Contigs* mapeados – SOAP2 x BWA

Animal	*Contigs* Mapeados BWA	*Contigs* Mapeados SOAP2
Gir 1	94,15%	78,52%
Guzerá 3	93,58%	76%

Para continuidade da avaliação, o resultado do mapeamento com o BWA (arquivo Bam) foi utilizado para compararmos onde cada *contig* consegue mapear na referência de *Bos taurus*. É válido ressaltar que o genoma utilizado como referência é taurino. Apesar do genoma taurino ser considerado próximo aos nossos genomas de estudo, sabemos que certamente existem diferenças entre eles, entretanto este é o dado que temos disponível no momento para trabalho.

O resultado do mapeamento, arquivo no formato Bam foi convertido em formato Bed e analisado com o pacote do BedTools. As posições de cada montagem referentes ao genoma taurino foram obtidas e as sobreposições entre os *contigs* das mesmas montagens (quando contabilizado na montagem de uma só estratégia) e das diferentes plataformas (quando unimos os arquivos Bam a procura das sobreposições) foram computadas.

O que podemos observar com estes resultados foi que com o acréscimo de diferentes bibliotecas a cobertura em extensão foi aumentada. Quando analisamos a montagem híbrida do SOLiD + MiSeq, as bibliotecas de ambos mapearam em diferentes posições da referência taurina, o que fortalece nossa hipótese das bibliotecas de diferentes tamanho de insertos mapeiam em diferentes posições do genoma. Já quando avaliamos os resultados do HiSeq os dois indivíduos sequenciados com o mesmo tipo e tamanho de biblioteca mapearam praticamente nas mesmas posições.

Apesar do valor de N50 não ter sido muito alterado ao unirmos todos os dados (resultados do tópico montagem), essa união ajudou na cobertura em extensão (pois estamos trabalhando com posições sem sobreposições dos dados). Estes resultados sugerem que bibliotecas oriundas do HiSeq mapeiam em posições diferentes das *reads Mate-pair* do SOLiD e *Paired-end* do MiSeq. Apesar da dificuldade de parametrização dos dados a montagem híbrida entre diferentes plataformas se mostrou interessante por cobrir diferentes regiões do genoma. A Tabela 9 mostra o resultado do mapeamento contra o genoma taurino, mostrando a porcentagem de mapeamento em posições únicas das montagens (cobertura em extensão).

Tabela 9: Mapeamento dos *Contigs* X *Bos taurus*

	SOLiD			MiSeq	SOLiD+MiSeq	HiSeq			Todas Juntas
	1a2kb	3a4kb	1a2e3a4	-		1	2	Juntos	
Guzerá	15%	13%	19%	48%	53%	55%	54%	57%	74%
Gir	14%	12%	17%	23%	37%	55%	42%	57%	63%

4.3.3.1 Saturação das Bibliotecas

Com o objetivo de saber o quanto a biblioteca de mesmo tamanho de inserto pode contribuir em uma mesma montagem, alguns testes foram realizados. Para isso os dados HiSeq desse trabalho (2 bibliotecas iguais para dois indivíduos de cada raça) e os dados de HiSeq do genoma de um projeto desenvolvido em paralelo foram utilizados.

Os dados HiSeq desse projeto consistiram em duas bibliotecas com tamanho de inserto de 300-500pb e os dados de Hiseq do genoma da planta (diploide) consistiram em três bibliotecas iguais de 700pb.

A Tabela 10 apresenta os resultados das três bibliotecas de mesmo tamanho do genoma da planta (diploide).

Tabela 10: Saturação das Bibliotecas de Mesmo Tamanho de Inserto Genoma Planta

	Biblioteca 1	**Bilioteca 2**	**Bilioteca 3**	**2 Biliotecas**	**3 bibliotecas**
Cobertura em extensão:	85,80%	88,66%	84,03%	88,84%	89,03%
Cobertura em profundidade	9,2x	52x	35x	62x	87x

A cobertura em extensão não aumenta significativamente com o acréscimo das bibliotecas (variando de 85% á 89%), enquanto a cobertura em profundidade teve grande aumento (variando de 9x á 87x). Esses resultados sugerem que as mesmas regiões do genoma foram sequenciadas repetidas vezes.

4.3.3.2 Resumo das montagens

A Figura 27 e a Tabela 11 resumem todos os resultados apresentados no tópico decisão da melhor montagem. Os resultados obtidos foram da média dos valores de N50 por plataforma/estratégia variaram de 250pb a 1,3kb. Os valores de N50 e cobertura, em ordem crescente do menor para o maior valor, foram encontrados na seguinte ordem: SOLiD, MiSeq, SOLiD + MiSeq, HiSeq e todas juntas.

Com os resultados de mapeamento dos contigs contra a referência de *Bos taurus* conseguimos inferir que os nossos dados apresentam uma grande redundância, ou seja, podemos ter sequenciado várias vezes a mesma região do genoma. Isso porque a cobertura em profundidade aumenta, mas o mesmo não pode ser notado para a cobertura em extensão.

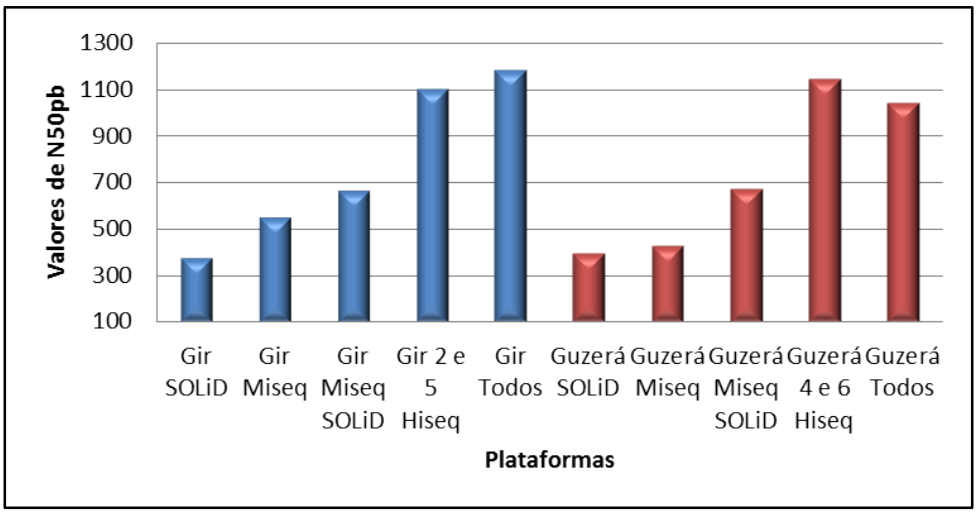


Figura 27: Média dos valores de N50 por plataforma: A Figura apresenta as montagens para as duas raças, Gir (esquerda, azul) Guzerá (direita, vermelho). O eixo X representa os valores de N50 indo de 100 á 1300pb e o eixo Y os animais por plataforma/estratégia.

Tabela 11: Resumo das montagens dos genomas por plataforma/estratégia

	N. contigs	N50	Cobertura
Gir SOLiD	1.191.745	381	20%
Gir Miseq	2.472.854	554	30%
Gir Miseq + SOLiD	3.000.000	670	40%
Gir 2 Hiseq	2.286.380	1131	69%
Gir 5 Hiseq	2.075.433	914	51%
Gir 2 e 5 Hiseq	2.312.153	1106	69%
Gir todos	2.369.271	1182	70%
Guzerá SOLiD	2.603.451	393	40%
Guzerá Miseq	2.740.334	427	42%
Guzerá Miseq + SOLiD	2.556.641	670	54%
Guzerá 4 Hiseq	2.407.142	1040	68%
Guzerá 6 Hiseq	2.246.759	1159	69%
Guzerá 4 e 6Hiseq	2.262.155	1145	69%
Guzerá Todos	2.369.271	1043	68%

Resultados que evidenciam nossas conclusões são as análises dos gráficos gerados das *reads* utilizando o programa FastQC. A Figura 28 apresenta o nível de duplicação das *reads* para os dados de HiSeq, SOLiD e MiSeq. Esses resultados foram de 38,9% 29% e 18% para o HiSeq, SOLiD e MiSeq, respectivamente.

O objetivo dessa abordagem do FastQC é informar até que ponto estamos perdendo a capacidade de sequenciamento e passando simplesmente a ressequenciar as mesmas regiões.

Em um “dado ideal” para uma biblioteca diversificada, os valores que estão acima do nível 1 (duplicados) devem decair rapidamente e permanecer no zero. No entanto, como pode ser visto nos nossos dados (Figura 28) não ocorre o decaimento com os valores de duplicação ultrapassando 30%, como por exemplo, nos dados do HiSeq.

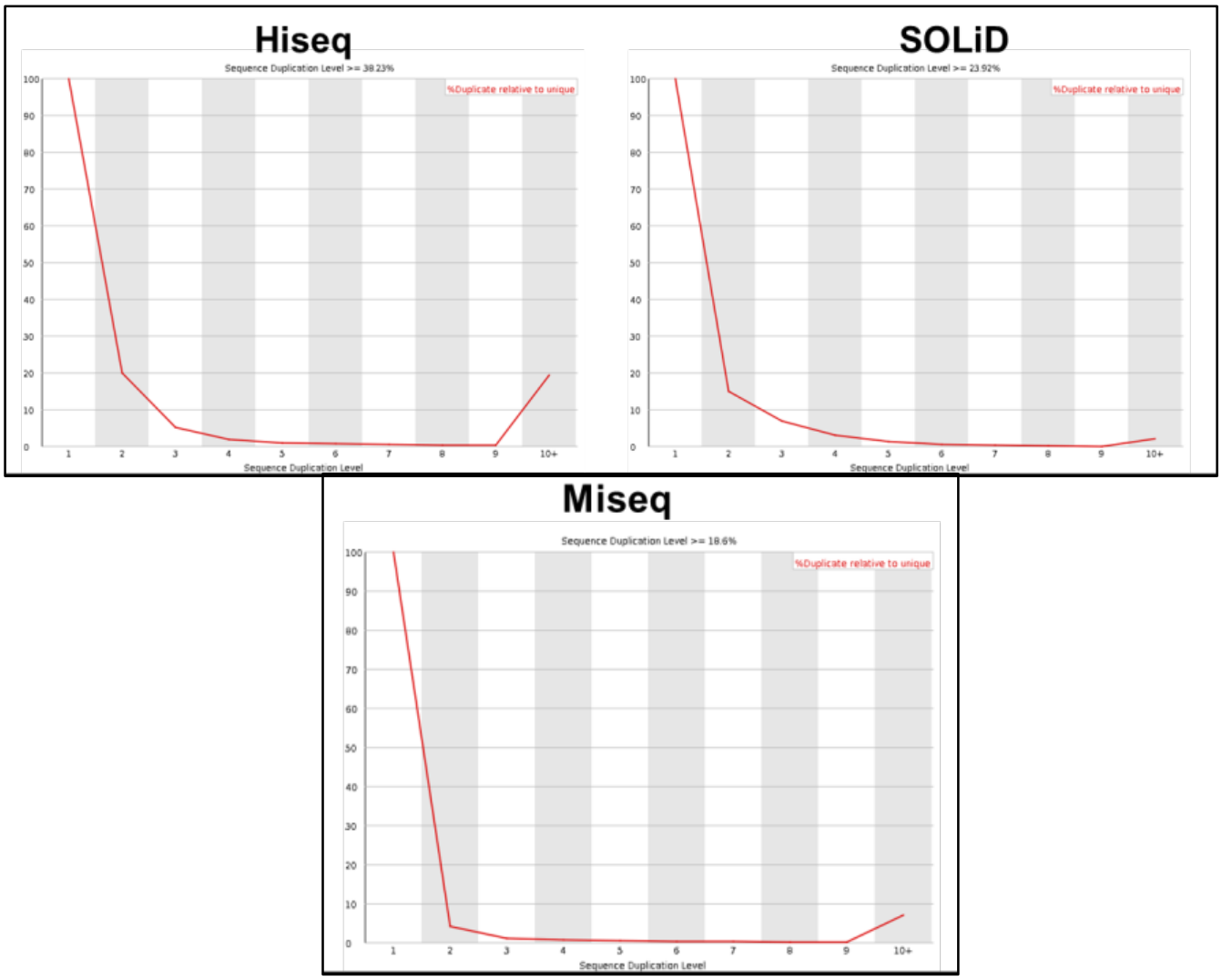


Figura 28: Duplicação das *Reads*

As Figuras 29 e 30 apresentam o resultado da distribuição média das bases dos *contigs* da montagem de todas as plataformas juntas (exceto PacBio) por cromossomo (referência UMD3.1). O cálculo do Z-score é capaz de determinar quantos desvios padrão acima ou abaixo da média a distribuição está. Para isso o desvio padrão e a média da cobertura dos *contigs* sobre o genoma (por cromossomo) foi calculado, em seguida a diferença entre a amostra e a média foram divididas pelo desvio padrão (resultando no valor de distribuição normal).

Nas figuras 29 e 30 os valores plotados para cima indicam uma forte probabilidade de "super cobertos", enquanto um resultado inferior (negativo) indica uma baixa probabilidade cobertura.

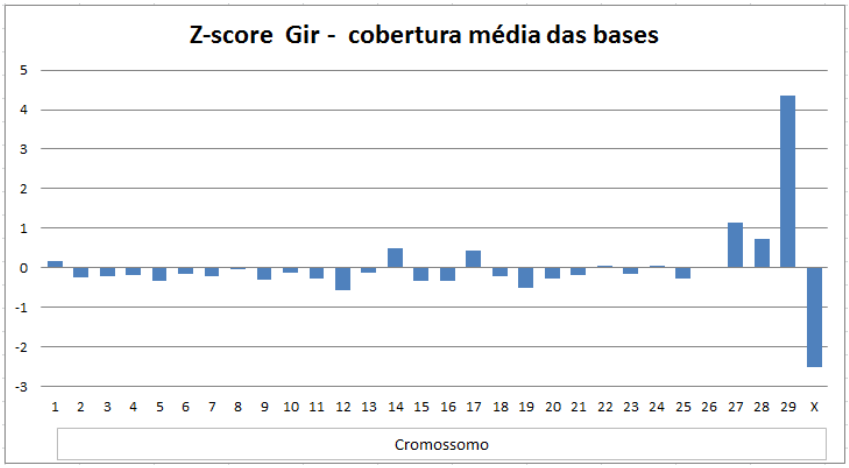


Figura 29: Cobertura média das bases Gir

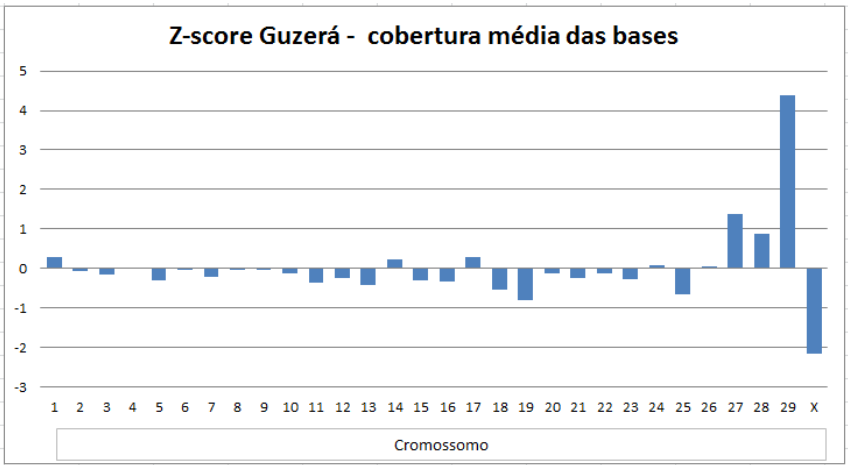


Figura 30: Cobertura média das bases Guzerá

A partir dos resultados do z-score também podemos inferir outra hipótese para a redundância dos nossos dados. A super-representação do cromossomo 29, por exemplo, pode estar relacionada aos elementos repetitivos presentes nesse cromossomo, como já relatado por Adelson e colaboradores (2009). Ou seja, essa poderia ser uma evidência de que as pequenas *reads* fornecidas pela tecnologia de NGS quando presentes em baixa cobertura podem não ajudar a resolver os problemas de grandes genomas com regiões repetitivas, como os bovinos que apresentam aproximadamente 40% de regiões repetitivas no genoma [ELSIK *et al*., 2009].

4.4 Resumo dos Resultados

O pré-processamento das *reads* antes de iniciar a montagem se mostrou de grande importância, mesmo sabendo que os programas de montagem fazem um pré-processamento por qualidade. De acordo com os nossos resultados os programas de montagem não são eficientes nessa etapa de pré-processamento da qualidade das *reads*, esse resultado corrobora com MOLNAR (2014) onde foi relatado a necessidade de correção de sequências (Illumina) antes de iniciarem o processo de montagem. Em um futuro próximo, é bem provável que os programas montadores de genomas consigam incorporar uma melhor estratégia de pré-processamento das *reads*, assim como muitos programas de mapeamento já o fazem (como por exemplo, o BWA).

Para trabalhar com montagem dos genomas com dados provenientes da plataforma SOLiD é preferível trabalhar com alto valor de qualidade. Com um valor de qualidade maior (PHRED 30) foi possível reduzir mais de 1.000.000 de *contigs* gerados, quando comparado ao valor de PHRED 20. Apesar da redução dos *contigs* e consequentemente dos pares de bases totais, a quantidade de bases dentro dos *contigs* aumenta, com isso, podemos concluir que neste caso, trabalhar com um valor de qualidade maior além de aumentar a confiabilidade dos dados ajuda a montagem a ficar menos fragmentada. No site do fabricante do SOLiD (https://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf), os autores fazem uma escala da utilização do valor de PHRED, onde foi determinado que a utilização do valor de PHRED 20 tem a acurácia de 99%, e PHRED 30 tem a acurácia seria de 99.9%. Contudo seria possível inferir que a diferença das montagens *de novo* possam estar relacionadas sim a qualidade dos dados, mas também com a possível saturação das bibliotecas sequenciadas (tendo coberto regiões iguais ou muito similares do genoma) e a baixíssima cobertura dos dados.

Não foi possível trabalhar com os dados de PacBio devido a baixa cobertura sobre o genoma (0,16%). Em um cálculo realizado juntamente à equipe da Pacbio na Universidade da Geórgia, chegamos a conclusão que seriam necessários 134 *flow cells* para se concluir o genoma, lembrando que nesse trabalho utilizamos apenas uma para cada indivíduo (Gir 1, Guzerá 3).

Dados MiSeq se mostraram de boa importância para a montagem do genoma, provavelmente devido ao tamanho das *reads*, porém não eficientes para esse tamanho de genoma. Com apenas 1x de cobertura das sequências sobre o genoma não foi possível fornecer dados suficientes para os grafos encontrarem o melhor caminho para construção dos pré-grafos,

contigs etc. Se todo o genoma fosse montado com MiSeq precisaríamos de no mínimo mais oito bibliotecas para cada animal trabalhado.

Os dados de HiSeq superaram as demais plataformas. Quando comparada ao MiSeq, as *reads* HiSeq são consideradas pequenas, entretanto devido a característica de alta geração de dados e consequentemente maior cobertura inicial (aproximadamente 10x para cada animal) o resultado da montagem mesmo não sendo tão satisfatório foi o melhor dentre todas as plataformas usadas nesse trabalho. A maior cobertura das sequências facilita o processo de montagem pelos programas montadores de genoma, pois, a maior cobertura ajuda a solucionar possíveis erros na geração dos pré-grafos. A Ilumina sugere que aproximadamente 50x de cobertura seja gerada para se obter uma montagem com 99,47% de cobertura e uma geração de 21x para uma cobertura de 99,38% (essa última é mais viável financeiramente para um projeto de montagem de grandes genomas) [http://res.illumina.com/documents/products/technotes/technotedenovoassemblyecoli.pdf].

Sobre a estratégia de unir os diferentes animais obtivemos dois diferentes resultados de acordo com a raça trabalhada. Para os animais da raça Gir, apesar da cobertura ter sido aumentada quando unimos os três animais, a montagem não apresentou uma melhora expressiva nos resultados, isso porque, como demonstrado, houve uma saturação da biblioteca de mesmo tamanho. Com isso o que pode ser observado foi um grande aumento na cobertura de profundidade, mas não na cobertura de extensão, evidenciando assim que bibliotecas de mesmo tamanho não contribuem para fechamento da montagem do genoma, mas sim para aumento em profundidade. Já para os animais da raça Guzerá a montagem com apenas a plataforma HiSeq apresentou melhores resultados do que unindo todos os dados oriundos de todas as plataformas. Mesmo quando unimos as duas bibliotecas HiSeq (dois diferentes animais) essa não apresentou uma melhora da montagem (lembrando que inúmeros parâmetros foram utilizados). Outra hipótese sobre esse resultado é que esses animais podem ser um pouco mais divergentes podendo ter complicado as montagens utilizando esse "*pool* de genomas".

Para ambos os animais, mesmo unindo todos os dados a montagem continua muito fragmentada, isso porque o maior tamanho de inserto adotado foi de 3-4kb, considerado pequeno para a construção de *scaffolds*, e os dados de PacBio que poderiam contribuir para essa estratégia de fechamento de *scaffolds* não puderam ser aproveitados devido a baixa cobertura dos mesmos sobre o genoma.

4.5 CONCLUSÕES

O presente capítulo teve como objetivo estabelecer as melhores estratégias de montagem *de novo* e direcionar o melhor caminho para futura conclusão do genoma nuclear das raças Gir e Guzerá. Para alcançar esse objetivo foram abordadas diferentes estratégias de montagem e tratamento dos dados das diferentes plataformas de NGS. Diante do exposto foi possível obter inúmeras conclusões:

As sequências oriundas do HiSeq seriam as mais indicadas para trabalhar com esses genomas bovinos. A montagem híbrida entre as três diferentes plataformas (HiSeq, MiSeq e SOLiD) é possível e se presente em uma quantidade alta de cobertura inicial do processo de montagem pode servir como uma estratégia híbrida de fechamento de um genoma complexo.

Em relação ao melhor programa para montagem, o SOAPdenovo apresentou maior versatilidade para trabalhar com todos os tipos de dados presentes nesse projeto, tendo sido possível trabalhar com bibliotecas *Mate-pair*, *Paired-end* com diferentes tamanhos de insertos e diferente tipos de plataformas de sequenciamento. Não foi possível trabalhar com os dados SOLiD no programa ABySS por isso não tentamos uma abordagem híbrida nesse programa. Outros programas como o ALLPATHS-LG podem ser testados para montagem de mamíferos, entretanto não adotamos esse programa nesse estudo por ele exigir uma alta cobertura inicial dos dados.

É válido ressaltar, que todas essas conclusões são baseadas nos resultados que obtivemos utilizando os genomas bovinos (melhor programa, plataforma entre outros), como já discutidos anteriormente, para cada organismo trabalhado pode ser possível obter resultados diferentes. Entretanto, para as análises de saturação da biblioteca de mesmo tamanho, testes em dois diferentes organismos (bovinos desse estudo e uma planta) foram realizados e foi possível obter os mesmos resultados e conclusões, o que só reforça a necessidade de sequenciamento de bibliotecas de diferentes tamanhos de inserto.

Mesmo unindo todos os nossos dados, ainda não foi possível concluir um *draft* do genoma desses animais.

4.5.1 Tempo Computacional, Processamento e Armazenamento dos dados

Para a montagem de um grande genoma utilizando NGS é obrigatória uma boa infraestrutura em bioinformática. Deve-se ter o espaço de armazenamento dos dados e capacidade de processamento.

A Tabela 12 descreve resumidamente os principais programas abordados nesse trabalho e o quanto foi utilizado de processamento dos dados. *Scripts* não foram acrescentados.

Tabela 12: Programas Utilizados x Tempo Computacional

Principais programas usados	CPU/RAM requeridos (por indivíduo)
FASTQc	2 cores, 20-60 minutos (dependendo dos dados)
SMRTanalaysis	40 cores, 3 horas
RACER	40 cores, 5 horas
SAET	12 cores, 4 horas
Kmergenie	15 cores, 200GB RAM, 3 horas
SOAPdenovo	850GB RAM, 20–60 cores. Tempo 8–24h (dependendo da estratégia).
PacBioToCA (PBcR)	450GB RAM, 48 cores 3 semanas
ABySS	650GB RAM, 60 cores 24-120h (dependendo da estratégia).
BWA	10 cores, 1-6 horas (dependendo da estratégia).
SOAP2	10 cores, 20-60 minutos (dependendo da estratégia).

Em termos de armazenamento foram gerados para esse trabalho 18TB de dados.

4.5.2 Ganhos e Limitações do trabalho

Esse foi o primeiro projeto de montagem de grandes genomas desenvolvido totalmente pelo nosso grupo e um dos primeiros do Brasil. A participação de diferentes instituições como EMBRAPA, FIOCRUZ e UFMG contribuíram muito para o trabalho, visto que é de extrema importância a presença de uma equipe multidisciplinar na montagem de um grande genoma e mais importante ainda a comunicação entre todas as partes. Os resultados desse estudo certamente irão contribuir para as montagens futuras de grandes genomas pelo nosso grupo.

Das limitações, esse trabalho apresentou inúmeras, como por exemplo, apesar da grande quantidade de dados inicial, estes se tornaram pouco informativos ao desenrolar do trabalho, principalmente após o pré-processamento por qualidade dos dados gerados.

Acredito que o grande gargalo do trabalho foi a baixa cobertura das *reads* sobre o genoma em si. Certamente é necessária a geração de mais bibliotecas com diferentes tamanhos de inserto, de preferência com grandes insertos para que a montagem fique menos fragmentada e quase completa.

Não considero que a utilização de diferentes plataformas foi uma limitação do trabalho, acredito que com mais dados e maior cobertura de cada uma, essa possa ser uma interessante estratégia.

A infraestrutura computacional não foi um fator limitante deste trabalho.

4.5.3 Dados reais x Dados ideais

Contudo, diante do exposto é possível se fazer sugestões para concluir um trabalho de montagem de genomas complexos:

Delineamento experimental:

- Conhecimento prévio do genoma a ser estudado

É de extrema importância um bom conhecimento prévio do organismo a ser trabalhado, conhecer as características gerais desse genoma, saber fazer a predição do tamanho, elementos repetitivos entre outros.

- Infraestrutura de TI, bioinformática e sequenciamento

Servidores de alto desempenho, capacidade de armazenamento e processamento de dados são obrigatórios nesse tipo de trabalho, bem como quem saiba manipulá-los. É indispensável a presença de algum profissional que saiba calcular o espaço de dados que será gasto e o tipo de máquina a ser utilizado, conhecimento dos programas necessários e de como usá-los ou até mesmo desenvolver esses programas quando necessário.

- Plataformas de sequenciamento

A partir do conhecimento prévio e estimativa do tamanho do genoma, é possível delinear qual a plataforma mais apropriada para esse tipo de dado. Para genomas grandes fica evidente aqui e em outros trabalhos [LI *et al*., 2009, BRADNAM *et al*., 2013] que é necessário utilizar plataformas com características de gerar uma grande quantidade de dados, como HiSeq. Em caso de se utilizar plataformas compactas como MiSeq e Ion deve-se levar em consideração a construção de muito mais bibliotecas e sequenciamentos para se obter uma alta cobertura. A utilização de mais de uma plataforma é válida, desde que se busquem várias alternativas, teste de diferentes programas e/ou desenvolvimento de novos métodos para essa integração.

- Construção das bibliotecas

Diferentes construções de biblioteca são vitais para uma cobertura completa do genoma. Tamanhos menores facilitam a formação dos *contigs* e os maiores contribuem para unir esses contis formando os *scaffolds*. Com a utilização de diferentes bibliotecas certamente o custo do projeto vai cair, visto que podemos obter efetivamente a mesma quantidade de informações fazendo menos sequenciamento.

- Programas

Diferentes programas podem ser testados para se concluir uma montagem. Deve-se levar em consideração qual plataforma e tipo de biblioteca foram utilizados, alguns programas não conseguem trabalhar com sequências *Mate-pair*s na formação de *contigs*.

- Validação da montagem

Deve-se avaliar a montagem com parâmetros estatísticos, como valor de N50, NG50, quantidade de *contigs* no primeiro momento, mas a montagem também deve ser validada biologicamente. Os dados de montagem de genomas não podem ser tratados apenas como valores estatísticos e matemáticos, deve-se lembrar que trata-se de um ser vivo, composto por DNA e torna-se de extrema importância a avaliação do sentido biológico. Como proposto por BRADNAM *et al*., 2013, é válido utilizar as 10 métricas propostas pelo grupo.

4.5.4 Perspectivas do método de NGS

Os desafios da utilização do sequenciamento de nova geração, seja para grandes ou pequenos genomas, certamente consiste na alta geração de dados e no pequeno tamanho das *reads* geradas. Não há dúvidas que o método NGS vem revolucionando a área da genômica, proteômica, transcriptômica, assim como não há dúvidas da necessidade de aprimoramente deste método. Esses desafios e falhas vêm sendo aos poucos resolvidos com o desenvolvimento e aperfeiçoamento das técnicas de sequenciamento.

Recentemente a Illumina apresentou sua nova plataforma de sequenciamento *Moleculo Long Read Sequencing* capaz de gerar *reads* de até 10kb.

A PacBio também anunciou o melhoramento da técnica de sequenciamento apresentando menor quantidade de erros e geração de *reads* maiores com maior geração de dados.

A *Life Technologies* também já apresentou uma nova versão do SOLiD V5, mas seus maiores investimentos vem sendo nos sequenciadores compactos do Tipo Ion.

Das plataformas ainda não disponíveis comercialmente, também existem novos anúncios, como a *Oxford Nanopore Technologies* (MinIONT).

Certamente o aumento do tamanho das *reads* e da acurácia irão facilitar o processo de montagem, desafiando mais uma vez os bioinformatas a aperfeiçoarem os programas de montagem para trabalhar com esse tipo de dado.

V - CONSIDERAÇÕES FINAIS DE AMBOS OS CAPÍTULOS

O objetivo geral da presente dissertação foi inicar o projeto piloto da montagem *de novo* dos genomas de animais representantes das raças Gir e Guzerá. Várias estratégias foram desenvolvidas para alcançar esse objetivo. Até o momento temos 69 % do genoma dessas duas raças montados. As estratégias utilizadas nos permitiram desenvolver um *pipeline* de montagem que poderá ser utilizado em todos os demais projetos de montagem de genomas grandes. Além disso, o genoma mitocondrial dessas raças foi montado pela primeira vez e nos permitiu fazer uma reconstrução filogenética desses animais e contar uma história evolutiva dos bovinos dos rebanhos.

Mesmo diante de todas as dificuldades e desafios da montagem de grandes genomas eucariotos utilizando dados de NGS, esse trabalho mostrou o que é possível fazer com poucos dados desses genomas.

REFERÊNCIAS

1. Achilli, A. *et al.* The multifaceted origin of taurine cattle reflected by the mitochondrial genome. *PLoS One* **4,** (2009).
2. Achilli, M., Pellecchia, M., Uboldi, C. & Uboldi, C. Mitochondrial genomes of extinct aurochs survive in domestic cattle. *Curr. Biol.* 157–158. (2008).
3. Adelson, D. L., Raison, J. M. & Edgar, R. C. Characterization and distribution of retrotransposons and simple sequence repeats in the bovine genome. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 12855–12860 (2009).
4. Anderson, S., Bruijn, M. H. L. D. E., Coulson, A. R., Sanger, F. & Medical, T. Complete Sequence of Bovine Mitochondrial. 683–717 (1982).
5. Baig *et al*., Intro-, I. & Proceedings, I. Phylogeography and origin of Indian domestic cattle. **89,** 9–11 (2005).
6. Beja-Pereira, A. *et al.* The origin of European cattle: evidence from modern and ancient DNA. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 8113–8 (2006).
7. Bonfiglio, S. *et al.* Origin and spread of *Bos taurus*: New clues from mitochondrial genomes belonging to haplogroup T1. *PLoS One* **7,** 1–10 (2012).
8. Bonfiglio, S. *et al.* The enigmatic origin of bovine mtDNA haplogroup R: Sporadic interbreeding or an independent event of Bos primigenius domestication in Italy? *PLoS One* **5,** (2010).
9. Bradley, D. G., MacHugh, D. E., Cunningham, P. & Loftus, R. T. Mitochondrial diversity and the origins of African and European cattle. *Proc. Natl. Acad. Sci.* **93,** 5131–5135 (1996).
10. Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *J. Mol. Evol.* **18,** 225–239 (1982).
11. Bruford, M. W., Bradley, D. G. & Luikart, G. DNA markers reveal the complexity of livestock domestication. *Nat. Rev. Genet.* **4,** 900–10 (2003).
12. Butler, J. *et al.* ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Res.* **18,** 810–820 (2008).
13. Canavez, F. C. *et al.* Genome sequence and assembly of Bos indicus. *J. Hered.* **103,** 342–8 (2012).
14. Chen, S. *et al.* Zebu cattle are an exclusive legacy of the South Asia neolithic. *Mol. Biol. Evol.* **27,** 1–6 (2010).
15. Chinnery, P. F., Elliott, H. R., Hudson, G., Samuels, D. C. & Relton, C. L. Epigenetics, epidemiology and mitochondrial DNA diseases. *Int. J. Epidemiol.* **41,** 177–187 (2012).
16. Chu, T.-C. *et al.* Assembler for de novo assembly of large genomes. *Proc. Natl. Acad. Sci. U. S. A.* **110,** E3417–24 (2013).
17. Dalloul, R. a *et al.* Multi-platform next-generation sequencing of the domestic turkey (Meleagris gallopavo): genome assembly and analysis. *PLoS Biol.* **8,** (2010).
18. Denisov, G. *et al.* Consensus generation and variant detection by Celera Assembler. *Bioinformatics* **24,** 1035–1040 (2008).
19. Dong, Y. *et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (Capra hircus). *Nat. Biotechnol.* **31,** 135–41 (2013).
20. Eck, S. H. *et al.* Whole genome sequencing of a single *Bos taurus* animal for single nucleotide polymorphism discovery. *Genome Biol.* **10,** R82 (2009).
21. Elsik CG, Tellam RL, Worley KC with The Bovine Genome Sequencing and Analysis Consortium. The Genome Sequence of Taurine Cattle: A window to ruminant biology and evolution. *Science (New York, N.Y.)* (2009);.
22. Felius, M., Koolmees, P. a., Theunissen, B. & Lenstra, J. a. On the Breeds of Cattle—Historic and Current Classifications. *Diversity* **3,** 660–692 (2011).
23. Frisch *et al*. Classification of the southern African sanga and east African shorthorned zebu. PMID:9172304. (1997)
24. Garcia, F., Lenstra, J. A. & Ajmone-marsan, P. On the Origin of Cattle : How Aurochs Became Cattle and Colonized the World. **157,** 148–157 (2010).
25. Grigson, C. An African origin for African cattle? ? some archaeological evidence. *African Archaeol. Rev.* **9,** 119–144 (1991).
26. Hiendleder S, Zakhartchenko V, Wolf E: Mitochondria and the success of somatic cell nuclear transfer cloning: from nuclear-mitochondrial interactions to mitochondrial complementation and mitochondrial DNA recombination. Reprod Fertil Dev 17: 69–82 (2005).
27. 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. *Cytogenet. Genome Res.* **120,** 150–6 (2008).

28. Iacobazzi, V., Castegna, A., Infantino, V. & Andria, G. Mitochondrial DNA methylation as a next-generation biomarker and diagnostic tool. *Mol. Genet. Metab.* **110,** 25–34 (2013).

29. Jünemann, S. *et al.* Updating benchtop sequencing performance comparison. *Nat. Biotechnol.* **31,** 294–6 (2013).

30. Kikkawa Y, Takada T, Sutopo, Nomura K, Namikawa T, et al: Phylogenies using mtDNA and *SRY* provide evidence for male-mediated introgression in Asian domestic cattle. Anim Genet 34: 96–101 (2003).

31. Le, T. H. *et al.* Mitochondrial gene content, arrangement and composition compared in African and Asian schistosomes. *Mol. Biochem. Parasitol.* **117,** 61–71 (2001).

32. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20,** 265–72 (2010).

33. Li, R. *et al.* The sequence and de novo assembly of the giant panda genome. *Nature* **463,** 311–318 (2009).

34. Li, W-H. e Graur, D. Fundamentals of Molecular Evolution. Sinaur Associates, Sunderland, Massachustes, 284pp (1991).

35. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** 251364 (2012).

36. Liu, W. *et al.* African origin of the malaria parasite Plasmodium vivax. *Nat. Commun.* **5,** 3346 (2014).

37. Loftus, R. T., Machugh, D. E., Bradley, D. G. & Sharp, P. M. Evidence for two independent domestications of cattle. **91,** 2757–2761 (1994).

38. Logue, K. *et al.* Mitochondrial genome sequences reveal deep divergences among Anopheles punctulatus sibling species in Papua New Guinea. *Malar. J.* **12,** 64 (2013).

39. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30,** 434–9 (2012).

40. Meirelles, F. V, Rosa, A. J. M., Lôbo, R. B. & Garcia, J. M. IS THE AMERICAN ZEBU REALLY Bos indicus ? **546,** 543–546 (1999).

41. Miller, S. a., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16,** 1215 (1988).

42. Molnar, M. & Ilie, L. Correcting Illumina data. *Brief. Bioinform.* (2014). doi:10.1093/bib/bbu029

43. Nijman IJ, Otsen M, Verkaar EL, de Ruijter C, Hanekamp E, et al: Hybridization of banteng (*Bos javanicus*) and zebu (*Bos indicus*) revealed by mitochondrial DNA, satellite DNA, AFLP and microsatellites. Heredity 90: 10–16 (2003).

44. Nystedt, B. *et al.* The Norway spruce genome sequence and conifer genome evolution. *Nature* **497,** 579–84 (2013).

45. Quail, M. a *et al.* A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13,** 341 (2012).

46. Roubertoux PL, Sluyter F, Carlier M, Marcet B, Maarouf-Veray F, et al: Mitochondrial DNA modifies cognition in interaction with the nuclear genome and age in mice. Nat Genet 35: 65–69 (2003).

47. Sanger, F., S. Nicklen, and A. R. Coulson. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977)

48. Schatten, H., Prather, R. S. & Sun, Q. Y. The significance of mitochondria for embryo development in cloned farm animals. *Mitochondrion* **5,** 303–321 (2005).

49. Schatz, M. C., Delcher, A. L. & Salzberg, S. L. Assembly of large genomes using second-generation sequencing. 1165–1173 (2010). doi:10.1101/gr.101360.109.20

50. Shaffer, H. B. *et al.* The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol.* **14,** R28 (2013).

51. Shariat, B., Movahedi, N. S., Chitsaz, H. & Boucher, C. HyDA-Vista : towards optimal guided selection of k -mer size for sequence assembly. *BMC Genomics* **15,** S9 (2014).

52. Silva, L. L. *et al.* The Schistosoma mansoni phylome: using evolutionary genomics to gain insight into a parasite's biology. *BMC Genomics* **13,** 617 (2012).

53. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–23 (2009).

54. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19,** 1117–23 (2009).

55. Simpson, J. T. *et al.* Efficient de novo assembly of large genomes using compressed data structures sequence data. 549–556 (2012). doi:10.1101/gr.126953.111

56. Simpson, J. T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* 1–8 (2014). doi:10.1093/bioinformatics/btu023

57. Slomovic, S., Laufer, D., Geiger, D. & Schuster, G. Polyadenylation and Degradation of Human Mitochondrial RNA : the Prokaryotic Past Leaves Its Mark Polyadenylation and Degradation of Human Mitochondrial RNA : the Prokaryotic Past Leaves Its Mark †. *Society* **25,** 6427–6435 (2005).

58. Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. In: Some mathematical questions in biology - DNA sequence analysis. Providence, RI: Amer. Math. Soc., 1986. p. **57-86** (1986)

59. Torroni, A., Achilli, A., Macaulay, V., Richards, M. & Bandelt, H. J. Harvesting the fruit of the human mtDNA tree. *Trends Genet.* **22,** 339–345 (2006).

60. Troy, C. S., Machugh, D. E. & Bailey, J. F. Genetic evidence for Near-Eastern origins of European cattle. **410,** 1088–1091 (2001).

61. Young, A., Abaan, H. & Zerbino, D. A new strategy for genome assembly using short sequence reads and reduced representation libraries. *Genome ...* 249–256 (2010). doi:10.1101/gr.097956.109.20

62. Zerbino, D. R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).

63. Zeuner D.Sc .*A History of Domesticated Animals*. Hutchinson, London, 1963. 84*s*.

64. Zimin, A. V *et al.* Mis-assembled "segmental duplications" in two versions of the *Bos taurus* genome. *PLoS One* **7,** e42680 (2012).

65. Liao, X. *et al.* and loci under selection 1. **7,** 1–7 (2013).

66. EID, D. *et al.* Single Polymerase Molecules. 133–138 (2009).

67. Bradnam, K. R. *et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2,** 10 (2013).

68. Huang, X. & Madan, a. CAP 3: A DNA sequence assembly program. *Genome Res.* **9,** 868–877 (1999).

69. Gut, I. G. New sequencing technologies. *Clin. Transl. Oncol.* **15,** 879–881 (2013).

70. Kai-Xing *et al*., 2006: Mitochondrial DNA D-Loop Variation and Genetic Background of Brahman Cattle. Zoological Research. **0254**-5853. (2006).

ANEXOS

Material Suplementar Online:

https://www.dropbox.com/sh/pdtnnsmt3bas96d/AADR5It0Jj8mwT6aYA0miqrWa?dl=0

PRODUÇÃO CIENTÍFICA, PARTICIPAÇÕES EM CONGRESSOS, CURSOS, ESTÁGIOS

Além dos trabalhos já mencionados, desde o ingresso no Mestrado atuo também como colaboradora em outros projetos e uma pequena síntese de alguns estudos é apresentada abaixo.

Durante o mestrado tive seis resumos publicados em congressos científicos, um artigo publicado. Participei da organização de um evento internacional.

Fiz dois cursos avançados. Lecionei algumas disciplinas e realizei estágio fora do país.

Descrição das atividades a seguir:

Whole-genome sequencing of Guzerá cattle: SNPs and INDELs in genes associated with production traits, disease resistance and heat tolerance

Izinara C. Rosse, **Juliana A. Geraldo**, Francislon S. Oliveira, Laura R. Leite, Flávio Araujo, Adhemar Zerlotini, Angela Volpini, Anderson J. Dominitini, Beatriz C. Lopes, Wagner A. Arbex, Marco A. Machado, Maria G.C.D. Peixoto, Rui S. Verneque, Marta F.Martins, Roney S. Coimbra, Marcos V.G.B. Silva, Guilherme Oliveira, Maria Raquel S. Carvalho

Abstract

Background: The Guzerá is an indicine dual-purpose breed, well adapted to the tropical climate, resistant to parasites, that has low susceptibility to mastitis. However, current SNP arrays include relatively few Guzerá variations. In this context, the objective of this work was to sequence and assemble the genome of one Guzerá to identify breed-specific variations that might be useful in breeding programs. Mate-pair libraries, with inserts of 1-2 and 3-4 kb, were generated with the ABI SOLiD system. Sequences were mapped to *Bos taurus* reference genome (UMD 3.1) using LifeScope. A list of putative SNPs and INDELs was generated using LifeScope and SAMtools, respectively, and their functional repercussion was investigated with NGS-SNP package.

Results: An average depth of coverage of 26X was achieved and 87% of the reference genome was covered. After quality filtering, 4,200,936 SNPs and 664,704 INDELs were identified. Sixty-five percent of the SNPs and 89% of the INDELs were previously unknown. Additionally, 2,676,067 (64%) of the SNPS and 466,005 (70%) of the INDELs were homozygous and not found in any database searched and may represent true differences between Guzerá and *Bos taurus*. From all the 3,142,072 genetic differences in Guzerá, 1,069 variations were classified as new non-synonymous SNPs, splice-site variants and coding INDELs (NS/SS/I) which have larger potential to cause functional repercussion. These variations were detected in 935 genes, which 105 were assigned as QTL for milk, meat and carcass, production, reproduction and health traits based on QTLdb and literature search. Additionally, the enrichment analysis showed that cell communication, environmental adaptation, signal transduction, sensory and immune systems were the KEGG categories with the highest number of genes containing homozygous NS/SS/I variants. These categories includes pathways involved in characteristics such as health, adaptation to the environment and behavior, disease resistance, and heat tolerance.

Conclusions: Substancial genetic differences were found between Guzerá and the taurine reference sequence, and some this variation is predict to affect the hardiness of the Guzerá. Thus, our findings provide groundwork for unraveling key genes and mutations behind disease resistance and heat tolerance that characterize the zebu breeds and may be used for customization of more effective arrays.

Key words: Guzerá cattle, whole-genome sequencing, SNP, INDEL, *Bos indicus*

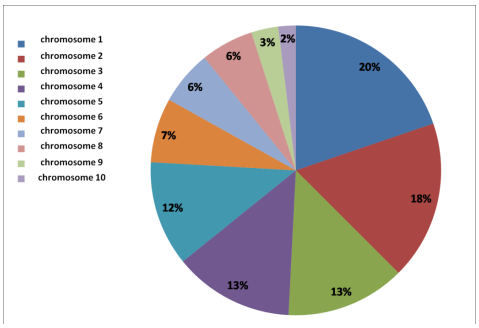
Esse trabalho está sendo finalizado e será submetido ainda nesse semestre a BMC Genomics

Comparative genomics in human parasite flatworms: Ehinococuccus granulosus s.s. (G1 genotype) and Echinococcus canadensis (G7 genotype)

Lucas L Maldonado, **Juliana Assis**, Flávio Gomes Araújo, Natalia Macchiaroli, Marcela Cucher, Mara Rosenzvit, Guilherme Oliveira and Laura Kamenetzky
1-IMPaM, CONICET, Fac. de Medicina - Univ. de Buenos Aires, Argentina 2- Genomics and Computational BiologyGroup, CPqRR - Oswaldo Cruz Foundation, Belo Horizonte, MG, Brazil.

Background. Echinococcus canadensis is a platyhelminth parasite which keeps close phylogenetic relationship with *Echinococcus granulosus* and *Echinococcus multilocularis*, members of the class Cestoda that are involved in hydatid infections of humans and animals. In South America three species of Echinococcus sensu lato have been reported *E. granulosus* sensu stricto (G1 and G2 genotypes), E. canadensis (G6 and G7 genotypes) and *E. ortleppi* (G5 genotype) (Kamenetzky and Cucher, 2014). Only limited genetic information of E. canadensis G7 was reported so far. In this work we have sequenced the genome of this species.
Methods. High quality genomic DNA has been extracted and two paired-end libraries have been sequenced by Illumina technology. Several pipelines of assembly have been evaluated. The genome has been *de novo* assembled with Velvet using different parameters until the best assembly was obtained. Also, *reads* have been mapped over E. multilocularis reference genome with BWA . Genes have been annotated by CEGMA and MAKER softwares with flatworm data for gene model training.

Results. Comparative studies have revealed high levels of nucleotidic identity of E. canadensis G7 with *E. multilocularis* as well as with *E. granulosus* s. s. G1. Almost all *contigs* have a correlation in *E. multilocularis* genome (Figure 1). Interestingly, the procedure for in silico annotation employed in this work allowed to identify 86% (387/450) of highly conserved genes (Table 1).



Conclusions. This is the first report of E. canadensis G7 genome. It was obtained by high throughput sequencing, allowing a broad genome view of this particular species that shows important biological and epidemiological features. The knowledge of this new genome would provide information for comparative genomics allowing adapting prevention and diagnosis tools to each epidemiological situation.

Esse trabalho está sendo finalizado e será submetido esse ano.

PLoS Negl Trop Dis. 2014 Jun 19;8(6):e2949. doi: 10.1371/journal.pntd.0002949. eCollection 2014.

Regulation of Schistosoma mansoni development and reproduction by the mitogen-activated protein kinase signaling pathway.

Andrade LF, Mourão Mde M, **Geraldo JA**, Coelho FS, Silva LL, Neves RH, Volpini A, Machado-Silva JR, Araujo N, Nacif-Pimenta R, Caffrey CR,Oliveira G.

Abstract

BACKGROUND:

Protein kinases are proven targets for drug development with an increasing number of eukaryotic Protein Kinase (ePK) inhibitors now approved as drugs. Mitogen-activated protein kinase (MAPK) family members connect cell-surface receptors to regulatory targets within cells and influence a number of tissue-specific biological activities such as cell proliferation, differentiation and survival. However, the contributions of members of the MAPK pathway to schistosome development and survival are unclear.

METHODOLOGY/PRINCIPAL FINDINGS:

We employed RNA interference (RNAi) to elucidate the functional roles of five S. mansoni genes (SmCaMK2, SmJNK, SmERK1, SmERK2 and SmRas) involved in MAPK signaling pathway. Mice were injected with post-infective larvae (schistosomula) subsequent to RNAi and the development of adult worms observed. The data demonstrate that SmJNK participates in parasite maturation and survival of the parasites, whereas SmERK are involved in egg production as infected mice had significantly lower egg burdens with female worms presenting underdeveloped ovaries. Furthermore, it was shown that the c-fos transcription factor was overexpressed in parasites submitted to RNAi of SmERK1, SmJNK and SmCaMK2 indicating its putative involvement in gene regulation in this parasite's MAPK signaling cascade.

CONCLUSIONS:

We conclude that MAPKs proteins play important roles in the parasite in vivo survival, being essential for normal development and successful survival and reproduction of the schistosome parasite. Moreover SmERK and SmJNK are potential targets for drug development.

Trabalho publicado na PLoS Negl Trop Dis.

Resumos:

Assis,J.G ; Rosse, I. C. ; Oliveira FS ; ARAUJO, F. ; SILVA, M. V. G. ; CARVALHO, M. R. S. ; OLIVEIRA, G. . Mitochondrial Genome Assembly of the Guzerá Breed. In: 10 th ISCB Student Council Symposium, 2014, Boston. Student Council Symposium, 2014. Referências adicionais: Classificação do evento: Internacional; Estados Unidos/ Inglês; Meio de divulgação: Vários; Homepage:http://scs2014.iscbsc.org/files/scs2014/SCS2014booklet.pdf.

Assis JG ; ROSSE, I.C. ; OLIVEIRA, F. S. ; ARAUJO, F. ; SILVA, M. V. G. B. ; CARVALHO, M.R.S. ; OLIVEIRA, G. . Mitochondrial genome assembly of the Guzerá breed. In: ISCB-Latin American x-Meeting on Bioinformatics with BSB & SoiBio, 2014, Belo Horizonte. ISCB-Latin American x-Meeting on Bioinformatics with BSB & SoiBio, 2014. Referências adicionais: Classificação do evento: Internacional; Brasil/ Inglês; Meio de divulgação: Digital.

ROSSE, I.C**.** ; **Assis JG** ; OLIVEIRA, F. S. ; LEITE, L. R. ; ARAUJO, F. ; Zerlotini, A. ; LOPES, B. C. ; ARBEX, W. A. ; MACHADO, MA ; PEIXOTO, MGCD ; Verneque, RS ; GUIMARAES, M. F. M. ; SILVA, M. V. G. B. ; COIMBRA, R. S. ; CARVALHO, M.R.S. ; OLIVEIRA, G. . Whole-Genome sequencing of Guzerá breed revealed SNPs with potential implication for milk production. In: Plant & Animal Genome XXII, 2014, San Diego. Plant & Animal Genome XXII, 2014. Referências adicionais: Classificação do evento: Internacional; Estados Unidos/ Inglês.

ROSSE, I.C**.** ; **Assis JG** ; OLIVEIRA, F. S. ; LEITE, L. R. ; ARAUJO, F. ; Zerlotini, A. ; LOPES, B. C. ; ARBEX, W. A. ; MACHADO, MA ; PEIXOTO, MGCD ; Verneque, RS ; GUIMARAES, M. F. M. ; SILVA, M. V. G. B. ; COIMBRA, R. S. ; OLIVEIRA, G. ; CARVALHO, M.R.S. . Novel Polymorphisms in genes associated with milk and meat production and disease resistance in the Guzerá breed identified by whole-genome sequencing. In: V Encontro de Genética de Minas Gerais, 2014, Belo Horizonte. V Encontro de Genética de Minas Gerais, 2014. Referências adicionais: Classificação do evento: Nacional; Brasil/ Inglês; Meio de divulgação: Vários.

ROSSE, I.C. ; **Assis JG** ; OLIVEIRA, F. S. ; LEITE, L. R. ; ARAUJO, F. ; Zerlotini, A. ; LOPES, B. C. ; ARBEX, W. A. ; MACHADO, MA ; PEIXOTO, MGCD ; Verneque, RS ; GUIMARAES, M. F. M. ; SILVA, M. V. G. B. ; COIMBRA, R. S. ; CARVALHO, M.R.S. ; OLIVEIRA, G. . New single nucleotide polymorphisms in Guzerá breed revealed by whole-genome re-sequencing. In: International Conference of the AB3C and Brazilian Symposium on Bioinformatics(X-meeting), 2013, Recife. International Conference of the AB3C and Brazilian Sympsium on Bioinformatics(X-meeting), 2013. Referências adicionais: Classificação do evento: Internacional; Brasil/ Inglês; Meio de divulgação: Digital.

ROSSE, I.C**.** ; **Assis JG** ; FONSECA, P. A. S. ; SANTOS, F. C. ; Pedro Lamounier Faria ; Steinberg, RS ; MIRANDA, M. ; OLIVEIRA, G. ; PIRES, M. F. A. ; PEIXOTO, MGCD ; CARVALHO, M.R.S. . Functional analysis in intronics SNPs. In: International Conference of the AB3C and Brazilian Symposium on Bioinformatics(X-meeting), 2013, Recife. International Conference of the AB3C and Brazilian Sympsium on Bioinformatics(X-meeting), 2013. Referências adicionais: Classificação do evento: Internacional; Brasil/ Português; Meio de divulgação: Digital.

Apresentações Orais:

ROSSE, I.C. ; Assis JG ; OLIVEIRA, F. S. ; LEITE, L. R. ; ARAUJO, F. ; Zerlotini, A. ; LOPES, B. C. ; ARBEX, W. A. ; MACHADO, MA ; PEIXOTO, MGCD ; Verneque, RS ; GUIMARAES, M. F. M. ; COIMBRA, R. S. ; OLIVEIRA, G. ; CARVALHO, M.R.S. . Novel Polymorphisms in genes associated with milk and meat production and disease resistance in the Guzerá breed identified by whole-genome sequencing (participação em mesa redonda). 2014. (Apresentação de Trabalho/Comunicação). Referências adicionais: Brasil/Português; Cidade: Belo Horizonte; Evento: V Encontro de Genética de Minas Gerais; Inst. promotora/financiadora: SBG-MG e Pós-Graduação em Genética da UFMG.

ROSSE, I.C. ; Assis JG ; OLIVEIRA, F. S. ; LEITE, L. R. ; ARAUJO, F. ; Zerlotini, A. ; LOPES, B. C. ; ARBEX, W. A. ; MACHADO, MA ; PEIXOTO, MGCD ; Verneque, RS ; GUIMARAES, M. F. M. ; SILVA, M. V. G. B. ; COIMBRA, R. S. ; CARVALHO, M.R.S. ; OLIVEIRA, G. . New Single Nucleotide Polymorphisms in Guzerá Breed Revealed by Whole-Genome Re-sequencing. Referências adicionais: Brasil/Inglês; Local: Mar Hotel; Cidade: Recife; Evento: International Conference of the AB3C and Brazilian Symposium on Bioinformatics - X-meeting; Inst. promotora/financiadora: Associação Brasileira de Bioinformática e Biologia Computacional (AB3C).

Participação em eventos internacionais:

10 th ISCB Student Council Symposium. Boston, USA. 2014.

22nd Intelligent Systems for Molecular Biology (ISMB). Boston, USA. 2014.

International Conference of the AB3C and Brazilian Symposium on Bioinformatics - X-meeting. Recife, Brasil, 2013

ISCB LA/X-Meeting/BSB/SolBio – Belo Horizonte, Brasil, 2014.

Latin American Student Council Symposium, Belo Horizonte, 2014.

Organização de Eventos:

Latin American Student Council Symposium, Belo Horizonte, 2014.

Cursos:

Exploring variation in animal genomes. EBI, 2014.

Aulas Ministradas:

Aula ministrada na disciplina de Biologia do Desenvolvimento para o curso de Ciências Biológicas manhã-UFMG, com carga horária de 5 horas

Aulas ministradas na disciplina de Bioinformática para o curso de Ciências Biológicas manhã-UFMG, durante o segundo semestre de 2013. Temas principais: Evolução e Análise de Sequências.

Estágio:

Mestrado "Sanduiche" no Instituto de Bioinformática da Unversity of Georgia, Athens - Georgia,com bolsa provida pelo "U.S. National Institute of health - Infectious Disease Genomics and Bioinformatics Training Grant in Brazil"