UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CAMPUS DE JABOTICABAL

ANÁLISE AMPLA DO GENOMA PARA DETECÇÃO DE ERROS DE MONTAGEM NO GENOMA DE REFERÊNCIA BOVINO E PARA DETECÇÃO DE LOCOS RELACIONADOS A CARACTERÍSTICAS DE PRODUÇÃO E REPRODUÇÃO DA RAÇA GIR

Adam Taiti Harth Utsunomiya

Zootecnista

2015

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CAMPUS DE JABOTICABAL

ANÁLISE AMPLA DO GENOMA PARA DETECÇÃO DE ERROS DE MONTAGEM NO GENOMA DE REFERÊNCIA BOVINO E PARA DETECÇÃO DE LOCOS RELACIONADOS A CARACTERÍSTICAS DE PRODUÇÃO E REPRODUÇÃO DA RAÇA GIR

Adam Taiti Harth Utsunomiya

Orientador: Prof. Dr. Ricardo da Fonseca

Coorientador: Dr. Marcos Vinícius Gualberto Barbosa da Silva

Coorientador: Prof. Dr. José Fernando Garcia

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Genética e Melhoramento Animal

2015

Utsunomiya, Adam Taiti Harth
U92a Análise ampla do genoma para detecção de erros de montagem no genoma de referência bovino e para detecção de locos relacionados a características de produção e reprodução da raça gir / Adam Taiti Harth Utsunomiya. – – Jaboticabal, 2015
v, 64 p. : il. ; 28 cm

Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias, 2015
Orientador: Ricardo da Fonseca
Coorientador: Marcos Vinícius Gualberto Barbosa da Silva, José Fernando Garcia
Banca examinadora: Henrique Nunes de Oliveira, Roberto Carvalheiro, Adriana Santana do Carmo, João Cláudio do Carmo Panetto
Bibliografia

1. *Bos indicus*. 2. Gir. 3. Desequilíbrio de ligação. 4. Varredura ampla do genoma. 5. Erros de montagem do genoma. I. Título. II. Jaboticabal-Faculdade de Ciências Agrárias e Veterinárias.

CDU 636.082:636.2

Ficha catalográfica elaborada pela Seção Técnica de Aquisição e Tratamento da Informação – Serviço Técnico de Biblioteca e Documentação - UNESP, Câmpus de Jaboticabal.

DADOS CURRICULARES DO AUTOR

Adam Taiti Harth Utsunomiya, filho de Takashi Utsunomiya e Josiane Cristina Harth Scanzani Camilo, nasceu em São Carlos – SP, em 22 de Maio de 1985. Em Julho de 2008 graduou-se no curso de Zootecnia pela Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP – Câmpus de Dracena, SP). Em agosto do mesmo ano ingressou no curso de Pós-graduação em Genética e Melhoramento Animal pela Faculdade de Ciências Agrárias e Veterinárias da UNESP – Câmpus de Jaboticabal, SP, sob orientação do professor Dr. Ricardo da Fonseca, obtendo seu grau de Mestre em Genética e Melhoramento Animal em Novembro de 2010. Em março de 2011 iniciou o curso de Doutorado no programa de Pós-graduação em Genética e Melhoramento Animal pela Faculdade de Ciências Agrárias e Veterinárias (UNESP – Câmpus de Jaboticabal, SP) sob orientação do professor Dr. Ricardo da Fonseca e coorientação do pesquisador da Embrapa Gado de Leite Dr. Marcos Vinícius Gualberto Barbosa da Silva e do professor Dr. José Fernando Garcia. Este também realizou o estágio de Doutorado no exterior, sendo recebido na “Universität fur Bodenkultur” - Áustria, de Janeiro à Setembro de 2014, sob orientação do professor Dr. Johann Sölkner.

AGRADECIMENTOS

Agradeço à,

Minha família, meu alicerce, que muito se dedicou para que esse objetivo fosse conquistado

Ao mestre, amigo, professor e orientador, Ricardo da Fonseca, por tantos anos de confiança no meu trabalho e na minha pessoa. Com certeza aprendi muito com seu caráter, profissionalismo e dedicação para com seus alunos

Aos coorientadores, Dr. Marcos Vinícius e Dr. José Fernando Garcia, pelas oportunidades de trabalho conjunto e pela confiança

Ao professor Dr. Johann Soelkner, pela oportunidade de poder desfrutar de um ótimo ambiente de trabalho em Vienna, no qual tive ótimas experiências e pude compartilhar do conhecimento de outros pesquisadores e construir novos laços profissionais

Aos meus amigos, de Dracena, Araçatuba, Jaboticabal, que foram indispensáveis nos momentos de descontração, aliviando os estresses desta longa jornada

Embrapa Gado de Leite e ABCGIL, por me conferirem a oportunidade de trabalhar com os dados e produzir informação de forma a contribuir para a melhoria do sistema de produção de leite nacional

Ao professor Dr. Paolo Ajmone Marsan, pela partilha de informações que complementaram este trabalho

“Cada dia sabemos mais e entendemos menos”

Albert Einstein

À meu avô, Hiroo Utsunomiya (*In memoriam*), dedico esta tese. Um exemplo de dedicação ao trabalho duro e altruísmo, permitindo que seus descendentes pudessem colher seus frutos!

SUMÁRIO

ANÁLISE AMPLA DO GENOMA PARA DETECÇÃO DE ERROS DE MONTAGEM NO GENOMA DE REFERÊNCIA BOVINO E PARA DETECÇÃO DE LOCOS RELACIONADOS A CARACTERÍSTICAS DE PRODUÇÃO E REPRODUÇÃO DA RAÇA GIR

RESUMO - A base genética que rege os processos fisiológicos para expressão dos fenótipos de produção de leite ainda não está completamente compreendida, pois poucos genes causais ou marcadores associados com a variação na expressão desses fenótipos foram relatados e espera-se que mais genes estejam envolvidos. Com o surgimento da era genômica, os esforços para identificar polimorfismos de sítio único (Single Nucleotide Polymorphisms - SNPs) foram expressivos. Os SNPs permitem estabelecer uma forte relação entre a expressão de características economicamente importantes e regiões específicas do genoma de um indivíduo. Tal relação é confirmada por estudos de associação ampla do genoma (GWAS), gerando conhecimento a cerca dos genes e fragmentos cromossômicos ligados a características importantes, os quais são posteriormente explorados na biologia dos sistemas. Qualquer inferência acerca de segmentos cromossômicos que possam estar associados a fenótipos de interesse utiliza uma montagem de um genoma de referência, onde todos os genes estão ancorados. Porém, o processo de montagem de um genoma é complexo e erros quanto ao posicionamento de sequências são esperados. Desta forma, este trabalho propõe avaliar a montagem de referência do genoma bovino produzido pelo grupo de pesquisa da universidade de Maryland e a aplicação do GWAS na raça Gir (*Bos indicus*) aos fenótipos de produção de leite, proteína e gordura, porcentagem de proteína e gordura e idade ao primeito parto, com o intuito de identificar regiões cromossômicas que possam estar relacionadas com aspectos importantes da produção de leite e fertilidade, contribuindo para a melhor compreensão dos fenômenos que regem tais aspectos.

Palavras-chave: *Bos indicus*, gir, desequilíbrio de ligação, varredura ampla do genoma, erros de montagem do genoma

GENOME WIDE ANALYSIS FOR DETECTING MIS-ASSEMBLED REGIONS IN THE BOVINE REFERENCE GENOME AND LOCI UNDERLYING PRODUCTIVE AND REPRODUCTIVE TRAITS IN GIR DAIRY CATTLE

ABSTRACT – The genetic basis of physiological processes underlying milk production traits are not completely understood, and few causal genes and markers associated with these traits have been reported to date. The emergence of the genomics era, efforts for the discovery of single nucleotide polymorphisms (SNPs) are numerous. These markers allow for establishing relationships between differences in economically important traits and specific genomic coordinates. These relationships are confirmed in genome-wide association studies (GWAS), which provide knowledge about genes and chromosomal segments affecting traits of interest that can be further explored in systems biology. Inferences about genomic localtions that are potentially implicated in phenotypic differences rely on a reference genome assembly where genes are annotated. However, genome assembly is a complex task that is prone to errors, and cases of wrong positioning of nucleotide sequences are not rare. Therefore, this thesis aimed at assessing candidate mis-assembled regions in the reference bovine genome assembly and performing a GWAS for milk traits in Gir cattle (*Bos indicus*), including milk, protein and fat yield, percentage of protein and fat, and age at first calving, targeting the identication of genomic regions that are potentially related to important aspects of fertility and milk production.

Keywords: *Bos indicus*, gir, linkage disequilibrium, genome wide mapping, misassembly genome

CAPÍTULO 1 – Considerações gerais

Introdução

A habilidade da espécie bovina em converter forragem de baixa qualidade em carne, leite e força de tração é de suma importância para o sustento de mais de 6,6 bilhões de pessoas (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al*., 2009).

Segundo projeções da FAO (2013), a população mundial deverá crescer consideravelmente nos próximos anos, embora a um ritmo mais lento do que no passado, e com diferenças consideráveis entre as regiões. Ao longo das próximas quatro décadas, a população mundial deverá aumentar em 2 bilhões e ultrapassar 9 bilhões de pessoas até 2050. As estimativas indicam que, para atender a demanda projetada, a produção agrícola global terá de aumentar em 60% a partir de seus níveis de 2005 a 2007.

Somente o Brasil foi responsável pela produção de 31 milhões de toneladas de leite em 2010, o que corresponde a aproximadamente 39% da produção de leite total da América Latina (FAO, 2013). Desta forma, o emprego de tecnologias para melhorar os sistemas de produção brasileiros, aumentando os índices de produtividade, torna-se uma necessidade, pois segundo dados da FAO (2013), o Brasil desempenha papel fundamental como agro-exportador emergente, principalmente em relação aos produtos do setor pecuário.

Na pecuária leiteira de países de clima tropical, como o Brasil, onde o sistema de produção predominante é extensivo e com forragens de baixa qualidade, raças taurinas (*Bos primigenius taurus*) não apresentam o potencial produtivo que apresentariam se estivessem produzindo em clima temperado. Desta forma, utilizaLiberation Serifr raças zebuínas (*Bos primigenius indicus*), por serem localmente adaptadas tanto ao clima quanto ao sistema de produção predominantemente extensivo, ou seus cruzamentos torna-se necessário. A raça Gir, amplamente explorada para produção de leite no Brasil, como raça pura ou em cruzamentos com taurinos leiteiros, se apresenta como uma opção a ser utilizada em sistemas de produção extensivos com forragens de baixa qualidade na maior

parte do ano.

Com o advento das tecnologias de genotipagem e o sequenciamento completo do genoma bovino (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al*., 2009), análises genômicas vêm sendo profundamente exploradas na produção animal, permitindo dissecar a porção genética que permite compreender a expressão de características complexas, tais como as produtivas e reprodutivas, e utilizar esta informação em programas de seleção. Compreender a manifestação dos fenótipos de produção e reprodução favoreceria explorar a atividade pecuária de maneira mais eficiente do que as existentes atualmente.

A abordagem que permite associar fenótipos à funções gênicas, levando ao melhor entendimento dos mecanismos subjacentes à manifestação de características complexas, é conhecida como Estudo de Associação Ampla do Genoma (do inglês “Genome-Wide Association Study” - GWAS). Outras abordagens como Mapeamento Amplo do Genoma (do inglês “Genome-Wide Mapping”) e Varredura Ampla do Genoma (do inglês “Genome-Wide Scan”) também são empregados com a finalidade de identificar regiões do genoma associadas com algum fenótipo de interesse.

Varreduras amplas do genoma bovino publicadas nos últimos anos (por exemplo, COLE *et al*., 2011; HAWKEN *et al*., 2012; MINOZZI *et al*., 2013;) tem revelado regiões cromossômicas fortemente associadas à fenótipos de produção de leite, favorecendo a melhor compreensão dos mecanismos biológicos que regulam a expressão desses fenótipos. No entanto, em revisão de literatura, somente trabalhos com raças taurinas e compostas foram encontrados.

As varreduras amplas de um genoma são baseadas em uma montagem de referência do genoma. No entanto, deve-se atentar para a qualidade de montagem do mesmo para que não hajam equívocos quanto à interpretação dos resultados de análises que dependam desta referência. De maneira superficial, um genoma de referência surge a partir de um processo no qual pequenos fragmentos de sequências de DNA de todo o genoma são alinhados uns contra os outros até que se consiga uma sequência completa de referência. Porém, sabe-se que o genoma bovino é repleto de sequências de DNA repetitivas que dificultam sua montagem,

gerando erros quanto ao posicionamento das sequências de DNA dentro do genoma. Estes erros podem ser identificados por meio de análises de desequilíbrio de ligação entre marcadores moleculares uma vez que os alelos dos marcadores que estão próximos fisicamente em um cromossomo são transmitidos de forma não independente, como colocado por O'Brien *et al*. (2014).

Desta forma, o objetivo deste trabalho foi encontrar as regiões com erros de montagem do genoma de referência por meio do desequilíbrio de ligação entre marcadores, fornecendo informação para que falsas interpretações de resultados que dependam dessa referência possam ser evitadas e realizar uma varredura ampla do genoma de bovinos da raça Gir para determinar quais regiões cromossômicas estão associadas à expressão dos fenótipos de produção de leite, proteína, gordura e sólidos totais, porcentagens de proteína, gordura e sólidos totais e idade ao primeiro parto, produzindo informação sobre os mecanismos de controle desses fenótipos.

Revisão de literatura

A raça Gir

Dentre as raças puras adaptadas às condições climáticas e ao sistema de produção extensivo brasileiro, a raça Gir apresenta os melhores índices produtivos, sendo amplamente exploradas para produção de leite e cruzamentos (LEÃO *et al*., 2013). As características morfofisiológicas da raça, como pigmentação da pele, capacidade termorregulatória, conversão alimentar, entre outras, permitem um desempenho produtivo satisfatório como raça pura ou em cruzamentos com raças leiteiras taurinas (PORCIONATO *et al*., 2010)

Embora, entre 1980 e 2010, a produtividade/vaca/ano tenha aumentado em 98,22%, ela está aquém da produtividade alcançada pelos países desenvolvidos. Para contribuir elevando a produtividade brasileira, o Programa Nacional de Melhoramento do Gir Leiteiro (PNMGL) identifica e avalia touros geneticamente superiores para produção de leite, proteína, gordura e porcentagens de proteína e gordura, além de outras características de produção, fertilidade, conformação e de

manejo (VERNEQUE *et al*., 2014).

De maneira geral, as características de produção (leite, proteína e gordura) possuem de baixa a moderada herdabilidades. As estimativas destas para as características mencionadas, obtidas a partir dos animais que integram o PNMGL, podem ser vistas na Tabela 1.

Tabela 1. Estimativas de herdabilidade (h^2) e correlações genéticas com produção de leite (r_G) para as características produtivas.

Características	**h^2**	**r_G**
Produção de leite	0,26±0,02	-
Produção de gordura	0,19±0,02	0,95
Produção de proteína	0,28±0,04	0,97

Como as herdabilidades dessas características são moderadas, apenas uma pequena porção da variação fenotípica total é explicada pela variação genética aditiva, e a seleção de reprodutores geneticamente superiores torna-se difícil e demorada. Além disso, a base genética que rege os processos fisiológicos para expressão desses fenótipos ainda não está completamente compreendida, pois poucos genes causais ou marcadores associados com a variação na expressão desses fenótipos foram relatados e espera-se que mais genes estejam envolvidos (MAI *et al*., 2010).

Além do apresentado até este ponto, vale ressaltar que a melhoria da eficiência produtiva nacional também está relacionada com o impacto das mudanças climáticas. Hayes *et al*. (2009) dizem que as mudanças climáticas, como aumento de temperatura e escassez de água, estão aumentando os preços dos grãos, também impulsionado pela maior demanda da alimentação humana e produção de biocombustíveis. Isto resultará em sistemas de produção de leite cada vez mais dependentes de pasto com animais adaptados, cenário no qual o Brasil se enquadra perfeitamente.

Contudo, a aplicação de conceitos e ferramentas da genética molecular, juntamente com a teoria da genética quantitativa, pode contribuir para uma exploração mais eficiente da atividade pecuária leiteira nacional, praticamente baseada em sistemas de produção extensivos.

Desequilíbrio de ligação

Este termo apresenta o conceito de correlação entre loci. Alelos fisicamente próximos em um cromossomo são trasmitidos conjuntamente de forma não aleatória, compondo um haplótipo.

A correlação entre loci surge principalmente devido à proximidade física entre eles mas também pode ser influenciada por diversos outros fatores, como por exemplo mutações, recombinações, deriva genética, seleção natural e artificial.

As estatísticas mais importantes e amplamente difundidas para se calcular o desequilíbrio de ligação entre loci são Lewontin's D' (LEWONTIN, 1964) e o quadrado do coeficiente de correlação genética r (HILL & ROBERTSON, 1968).

O'Brien *et al*. (2014), em revisão de literatura, mostram que estudos de desequilíbrio de ligação apresentam muitas aplicações em genética, tais como estimação de taxas de recombinação, tamanho efetivo de população e inferência de eventos históricos e demográficos de populações. A extensão do desequilíbrio de ligação também é utilizada como premissa em estudos de associação.

O conceito de desequilíbrio de ligação entre marcadores também pode ser aplicado na detecção de erros de montagem de sequências em genomas complexos, como o de bovinos.

Bohmanova *et al*. (2010), encontraram padrões não esperados de decaimento do desequilíbrio de ligação entre marcadores SNPs em 3 cromossomos bovinos. Corbin *et al*. (2012) também encontraram correlações inesperadas entre marcadores posicionados distantemente uns dos outros e até mesmo em diferentes cromossomos. Porém, nestes dois trabalhos os autores atribuem os padrões inesperados de desequilíbrio de ligação a um problema de mapeamento do posicionamento dos SNPs e não hipotetizaram problemas de montagem nos genomas de referência.

Varredura ampla do genoma

O termo *Genome Wide Association Studies* (Estudos de Associação Ampla do Genoma) foi um conceito originalmente utilizado pela comunidade científica

envolvida com a genética epidemiológica humana, e tem por finalidade a identificação de regiões do genoma relacionadas com fenótipos específicos para posterior investigação de suas funções biológicas (NATIONAL HUMAN GENOME RESEARCH INSTITUTE, 2011). Este termo possui alguns outros termos equivalentes como *Genome wide mapping* (Mapeamento amplo do genoma) ou *Genome wide scan* (Varredura ampla do genoma), que apresentam a mesma finalidade, buscar associações de regiões do genoma com a expressão de fenótipos de interesse.

No passado não distante, o mapeamento de QTL foi a estratégia criada e aplicada para identificar, a nível molecular, variações em características quantitativas. Essa estratégia é baseada na associação de regiões cromossômicas com variações fenotípicas, pois, de maneira geral, os genes são desconhecidos.

Os primeiros trabalhos para determinar se a variação a nível molecular está associada à variação na expressão de características quantitativas utilizaram marcadores tipo AFLP, RFLP e microssatélites. No entanto, a prospecção destes marcadores apresenta dificuldades técnicas inerentes à sua natureza bioquímica e os custos para obtê-los são altos. Mas, com os recentes avanços tecnológicos, surgiram metodologias de alto desempenho e acurácia e de baixo custo de mão de obra para prospecção, caracterização e genotipagem de marcadores do tipo polimorfismo de nucleotídeo único, SNPs (“Single Nucleotide Polimorphism”) (CAETANO, 2009).

Os marcadores SNPs têm tido destaque no cenário científico mundial devido ao seu alto grau de informatividade, derivada do grande volume de polimorfismos de base única encontrados nos genomas. Essa informatividade está relacionada ao conceito de desequilíbrio de ligação, pois o volume de SNPs no genoma é tão grande que espera-se que ao menos um marcador esteja em associação não aleatória com um ou mais loci que afetam a expressão de uma característica de interesse econômico.

A descoberta dos SNPs permitiu aos pesquisadores estabelecer relações entre a expressão de características economicamente importantes e regiões específicas do genoma de um indivíduo. Quando encontrados em regiões muito próximas ou até mesmo constituindo um loco que possui efeito diferenciado na

expressão de uma dada característica, os SNPs podem ser considerados marcas fiéis ao mesmo. Com isto, a capacidade de compreender a expressão de características complexas e estimar os efeitos de segmentos cromossômicos na expressão dessas características permite aumentar a eficiência de seleção pela utilização destas informações na predição de valores genéticos (FRAZER *et al*., 2009).

Considerando que existem muitos genes que exercem pequena influência na expressão de características quantitativas e poucos que oferecem grande contribuição aditiva a elas (HAYES & GODDARD, 2001), é sensato identificar os de maior contribuição para que possam ser utilizados em estratégias de seleção e cruzamentos, desde que eles expliquem uma porção da variação fenotípica que reflita em progresso genético mais acelerado do que os comumente alcançados.

Os esforços para identificar polimorfismos de nucleotídeo único foram expressivos. Durante a execução do projeto genoma bovino (THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM *et al*., 2009), foram identificados milhões de SNPs. Com tamanha densidade de cobertura do genoma, a nova ordem é identificar quais os genes que estão relacionados com a expressão de fenótipos de interesse e trazer luz de como este último é definido pela informação genética. GWAS é a abordagem amplamente difundida para tal.

Os Estudos de Associação Ampla do Genoma, aliando os conhecimentos de genética quantitativa, molecular e de populações, permitem a identificação de associações estatísticas significativas entre centenas de milhares de marcadores SNPs e fenótipos de características complexas, aumentando o entendimento sobre importantes vias biológicas (SCOTT *et al*., 2007).

Os trabalhos de GWAS realizados com animais de raças taurinas leiteiras têm encontrado resultados interessantes, fornecendo evidências confirmatórias de associação das regiões testadas com os bancos de dados gênicos e também um conjunto de novos SNPs associados com características de produção de leite (JIANG *et al*., 2010; MAI *et al*., 2010). Isso nos proporciona uma visão do sistema biológico envolvido no estabelecimento dos padrões fenotípicos dessas características.

Referências

BOHMANOVA, J.; SARGOLZAEI, M.; SCHENKEL, F. Characteristics of linkage disequilibrium in North American Holsteins. **BMC genomics**, v. 11, 2010.

CAETANO, A.R. Marcadores SNP: Conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**. n.38, p.64-71. 2009.

COLE, J. B.; WIGGANS, G. R.; MA, L.; *et al.* Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in contemporary U.S. Holstein cows. **BMC Genomics**; 12:408, 2011.

CORBIN, L. J.; BLOTT, S. C.; SWINBURNE, J. E. *et al*. The identification of SNPs with indeterminate positions using the Equine SNP50 BeadChip. **Animal Genetics**, v. 43, p. 337-339, 2012.

FAO-Food and Agriculture Organization of the United Nations. FAO Statistical Year Book 2013 – World Food and Agriculture. **FAO/Eletronic Publishing Policy and Support Branch**, p.123-199, 2013.

FRAZER, K. A.; MURRAY, S. S.; SCHORK, N. J.; TOPOL, E.J. Human genetic variation and its contribution to complex traits. **Nature Reviews**. v. 10, p. 241-251, 2009.

HAWKEN, R. J.; ZHANG, Y. D.; FORTES, M. R. S. Genome-wide association studies of female reproduction in tropically adapted beef cattle. **Journal of Animal Science**. v. 90, p. 1398-1410, 2012.

HAYES, J. B.; BOWMAN, P. J.; CHAMBERLAIN, A. J.; GODDARD, M. E. *Invited review*: Genomic selection in dairy cattle: Progress and challenges. **Journal of Dairy Science**. v. 92, p. 433–443, 2009.

HAYES, B. J.; GODDARD, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genet. Sel. Evol.**, v.33, ed.3, p.209-229. 2001.

HILL, W. G.; ROBERTSON, A. Linkage disequilibrium in finite populations. **Theoretical and Applied Genetics**, v. 38, p. 226-231, 1968.

JIANG L, LIU J, SUN D, MA P, DING X, *et al*. Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. **PloS One** 5(10), 2010.

LEÃO, G. F. M.; PIVATTO, D. R. D.; CARNIEL, H. Melhoramento genético em zebuínos leiteiros – uma revisão. **Agropecuária Científica no Semiárido,** p.9-14, 2013

LEWONTIN, R. The interaction of selection and linkage. I. General considerations; Heterotic models. **Genetics**, v. 49, p. 49-67, 1964.

MAI, M. D.; SAHANA, G.; CHRISTIANSEN, F. B.; GULDBRANDTSEN, B. A genome wide association study for milk production traits in Danish Jersey cattle using a 50K SNP chip. **J Anim Sci**., 88(11):3522-3528, 2010.

MINOZZI, G.; NICOLAZZI, E. L.; STELLA, A. Genome wide analysis of fertility and production traits in italian holstein cattle. **PloS One**, v. 8, 2013

NATIONAL HUMAN GENOME RESEARCH INSTITUTE. A Catalog of Published Genome Wide Association Studies. Disponível em: http://www.genome.gov/GWAStudies/. Acesso em: 23 de Ago de 2011.

O'BRIEN, A. M. P.; MÉSZÁROS, G.; UTSUNOMIYA, Y. T. *et al*. Linkage disequilibrium levels in *Bos indicus* and *Bos taurus* cattle using medium and high density SNP chip data and different minor allele frequency distributions. **Livestock Science**, v. 166, p. 121-132, 2014.

PORCIONATO, M. A. F.; SOARES, W. V. B.; REIS, *et al*. Milk flow, teat morphology and subclinical mastitis prevalence in Gir cows. **Pesq. Agro. Bras.**, v. 45, p. 1507-1512, 2010.

SCOTT, L. J.; MOHLKE, K. L.; BONNYCASTLE, L. L.; WILLER, C. J.; LI, Y.; DUREN, W. L. et al., A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. **Science**. v. 316, p. 1341–1345, 2007.

THE BOVINE GENOME SEQUENCING AND ANALYSIS CONSORTIUM; ELSIK, C. G.; TELLAM, R. L.; WORLEY, K. C. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science**, v.324, p.522-528. 2009.

VERNEQUE, R. S. *et al.* **Programa Nacional de Melhoramento do Gir Leiteiro: Sumário Brasileiro de Touros – Maio 2014**. Embrapa Gado de Leite, Juiz de Fora, MG, 2014.

CAPÍTULO 2 - High resolution linkage disequilibrium reveals mis-assembled regions of the bovine reference genome

Adam T H Utsunomiya[1§], Daniel J A Santos[1], Solomon A Boison[2], Yuri T Utsunomiya[1], Paolo A Marsan[3], Johann Sölkner[2], José F Garcia[1], Ricardo da Fonseca[1], Marcos V G B da Silva[4§],

[1]Univ Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias Campus de Jaboticabal, Jaboticabal, São Paulo, Brazil
[2]Universitäte fur Bodenkultur, Vienna, Austria
[3]Universidade del Sacro Cuore, Piacenza, Italy
[4]EMBRAPA dairy cattle, Juiz de Fora, Minas Gerais, Brazil

[§]Corresponding author

Email addresses:
ATHU: adamtaiti@gmail.com
DJAS: daniel_jordan2008@hotmail.com
SAB: soloboan@yahoo.com
YTU: ytutsunomiya@gmail.com
PAM: paolo.ajmone@unicatt.it
JS: johann.soelkner@boku.ac.at
JFG: jfgarcia@fmva.unesp.br
RF: ricardo@dracena.unesp.br
MVGBS: marcos.vb.silva@embrapa.br

Abstract

We analyzed the linkage disequilibrium decay using high density single nucleotide polimorphisms across all 29 bovine autosomal chromossomes in two different breeds, Holstein and Nellore, representing each subspecie, for detecting mis-assembled segments in the University of Maryland v3.1 (UMD v3.1) reference genome assembly. A total of 30 segments across 14 chromosomes and 5 misplaced single nucleotide polimorphisms (SNPs) were identified.

Keywords: Linkage disequilibrium, mis-assembly, bovine

Introduction

Since the publication of the whole genome sequence of *Bos taurus* [1] two independent initiatives have provided different assemblies of the bovine reference genome: the Baylor College of Medicine Human Genome Sequencing Center (Btau_4.6.1) [2] and the University of Maryland (UMD_3.1) assembly [3]. These two assemblies have been used in several genetic diversity, genome-wide mapping, and genomic selection studies, among others, mostly using UMD_3.1 due to its larger amount of sequences placed in the genome, (99% in UMD_3.1 against 95% in Btau_4.6.1). It is evident that all the research results obtained using these assemblies are somehow dependent of their quality.

Considering the complexity to assemble a genome, errors are expected to occur while placing sequences onto a chromosome, mainly due to the existence of repetitive sequences that tend to collapse in an assembly [2]. These repetitive sequences act as assemblers' confounders resulting in shuffling the order of multiple repeated copies, rearranging the unique sequences between them, which can produce unexpected correlations between *loci* [4]. Under the computational point of view, the repeated regions can create ambiguities during the alignments and assembling of a genome, which in turn, induces misinterpretation of the results from studies that depend on the assembly as reference [4].

Taking advantage of the correlations between *loci*, created during evolutionary

events as natural and artificial selection, linkage disequilibrium analysis can be used as a quality assessment of a genome assembly.

In this sense, we aim to explore the linkage disequilibrium decay in the 29 bovine autosomal chromosomes using the Illumina® BovineHD BeadChip (HD) assay anchored in the UMD release 3.1 assembly, capturing the unexpected correlations between single nucleotide polimorphisms (SNPs) and infer misassembly segments into the bovine genome.

Results

Our results provide information about mis-assembled genomic regions in the UMD release 3.1 bovine genome assembly using high resolution linkage disequilibrium, as several studies are dependent of its quality. Hence, we used 875 Holstein (Bos primigenius taurus) and 803 Nellore (Bos primigenius indicus) samples, genotyped with the Illumina® BovineHD BeadChip (HD) as these breeds are the most representative breeds (lowest SNP ascertainment bias) for each subspecie in this platform.

The overview of linkage disequilibrium (LD) decay per chromosome for both breeds are shown in Figure 1. We found unexpected correlations between SNPs at the same locations in both breeds on chromosomes 1-3, 5, 7, 9-16, 21, 23 and 26-27 (Table 1), which indicates errors in the placement of contigs into the assembly.

For each segment in each chromosome we plotted the pair-wise r^2 between SNPs against the physical distances to show that SNPs in a segment do not present high correlations with SNPs surrounding them but with markers positioned distantly (Figure 2). In five of the segments that we found, erratic LD was observed in a single SNP. These segments were on chromosome 3 (named CHR3SEG2 for simplicity), 5 (CHR5SEG1), 9 (CHR9SEG1), 13 (CHR13SEG1) and 14 (CHR14SEG1) (see Table 1). These SNPs were considered misplaced once markers surrounding them presented expected r2 decay (data no shown).

The average length of the mis-assembled segments is 128,819 bp with the smallest segment being located on chromosome 11 (691 bp) and largest segment being located on chromosome 1 (623,102 bp). The sizes of contigs have been used

to measure the quality of an assembly, although questionable [5]. Contig N50 is the parameter representing the size N such that 50% of the genome is contained in contigs of size N or greater [3]. In our study, the largest average contig size within a mis-assembled segment is 53,650. The current version of the UMD assembly has a N50 of 96,955, which is larger than the largest average contig size in a mis-assembled segment detected here. Small contig sizes humpers contig placement and are likely to represent repeated sequences that were collapsed by an assembler.

Bohmanova *et al*. [6], characterizing linkage disequilibrium in North American Holsteins, found SNPs producing higher than expected r^2 on chromosomes 1, 6 and 26 using the Illumina® BovineSNP50 BeadChip and applied a simplified approximation of the SNP location based on the highest pair-wise r^2 of the two markers flanking it. These SNPs are located into the limits of the segments that we found in our study. Carvalheiro *et al*. [7], detected very low imputation accuracies in different segments on chromosome 1 which also matched with the segments found here.

Instead of propose algorithms and suggest any correction for positioning the misplaced SNPs or the SNPs into segments we have detected, we compared the SNP locations in UMD_3.1 and Btau_4.6.1 for checking out discrepancies in these segments between these two assemblies. Based on SNP positions and considering they are not in the same scale to compare UMD_3.1 and Btau_4.6.1, proportionaly we could see that the sequences in Btau_4.6.1 seem to be more accurately placed into the genome compared to UMD_3.1 for the mis-assembled segments, obviously for SNPs which were mapped in Btau_4.6.1, indicating a better quality of the sequencies of these segments assembled in Btau_4.6.1. Nevertheless, linkage disequilibrium decay analysis also indicates mis-assembled sequencies in Btau_4.6.1 (data not shown).

To illustrate the previous sentence, we compared the SNP positions in the two assemblies in SEG1 on chromosome 1 (Figure 3). There are three different scenarios to be considered in this segment: 1) SNPs that were mapped in both assemblies (blue dots in Figure 3B); 2) SNPs that were mapped in UMD_3.1 but not in Btau_4.6.1 (gray dots in Figure 3B); and 3) SNPs that were mapped in different chromosomes (red dots in Figure 3B). In scenario 1 the SNPs mapped between

44.1-45.13 Mb in UMD_3.1 but between 160.8-161.4 Mb in Btau_4.6.1. Looking at the linkage disequilibrium graph of the SEG1 on chromosome 1 (Figure 2) we note that the SNPs in this segment are highly correlated with SNPs around 120 Mb, in concordance with Btau_4.6.1, saving the proportions. In scenario 2 we can see only one SNP, which could represent the difficulty to assemble the sequence where it takes place in Btau_4.6.1. In scenario 3 there is a case where SNPs in a tiny segment mapped against chromosome 1 in UMD_3.1 but against chromosome 11 in Btau_4.6.1. Plotting the LD decay of this tiny segment on chromosome 1 considering UMD_3.1 assembly there are no high correlations between SNPs within this segment with flanking markers (Figure 3C). However, plotting the r^2 between SNPs in this segment against the positions of the SNPs mapped on chromosome 11 (Figure 3D) we observed very high correlations with the region around 85 Mb. Once again, in concordance with Btau_4.6.1 assembly, where the SNPs also mapped against the region around 85 Mb.

The majority of the SNPs we prospected from the segments detected in our study are in unplaced sequences in Btau_4.6.1, evidencing that these regions present challenges to the assemblers.

It is consensus that several type of studies are dependent of the quality of an assembly. Salzberg & Yorke [5] state that not only are genes and regulatory sites anchored in the sequence, but analysis of synteny, duplications, and evolutionary relationships among species all depend on having the correct structure of the genome and our study contributes providing information about genome regions that are potentially mis-assembled.

Methods

For detecting mis-assembled genomic regions we have investigated unexpected r^2 in the linkage disequilibrium decay patterns in Holstein (*Bos primigenius taurus*) and Nellore (*Bos primigenius indicus*) cattle, since they are the most representative breeds in terms of ascertainment bias of the Single Nucletide Polimorphisms (SNPs) for each bovine subspecie in the Illumina® BovineHD Genotyping

BeadChip assay, according to the manufacturer's documentation [8]. The analyses were done based on the latest version of the UMD assembly (3.1.1) (UMD_3.1).

Genotypes and quality checks

A total of 1,009 Holstein bulls and 1,069 Nellore bulls were genotyped with the Illumina® BovineHD Genotyping BeadChip assay. Quality checks were performed using plink v1.9 [9]. Only non-coincident and autosome markers were included in further analysis. SNPs were removed from the dataset if their call rate < 95%, minor allele frequency (MAF) < 3% and Hardy-Weinberg equilibrium < 10^{-6}. Samples were removed if they exhibit call rate < 95% and mean deviation between observed and expected heterozygosity was out of ±3 standard deviations.

After this initial prunning we excluded related animals from the datasets based on the proportion of alleles identical by state (IBS) (>40%), following [10], to avoid overestimation of the correlations between markers due to family structure in our datasets.

Detecting mis-assembled regions by linkage disequilibrium decay

A SNP was considered in a mis-assembled segment based on its unexpected high correlation (> 0.5) with another SNP disposed at least 10 Mb apart.

From the position of the SNPs showing unexpected correlations we manually determined the mis-assembled segments delimitating them by the first base pair of the contig downstream and the last base pair of the contig upstream within the segment. For each segment we performed LD analysis to well delimites the regions.

Conclusions

Our findings provide useful information for helping the improvement of the bovine genome assembly UMD_3.1 and for avoiding mis-interpretation of results of studies that match with segments found here and that are dependent of SNP markers positions. Linkage disequilibrium patterns using high density SNP data showed to be

consistent for investigating mis-assembled segments in a genome.

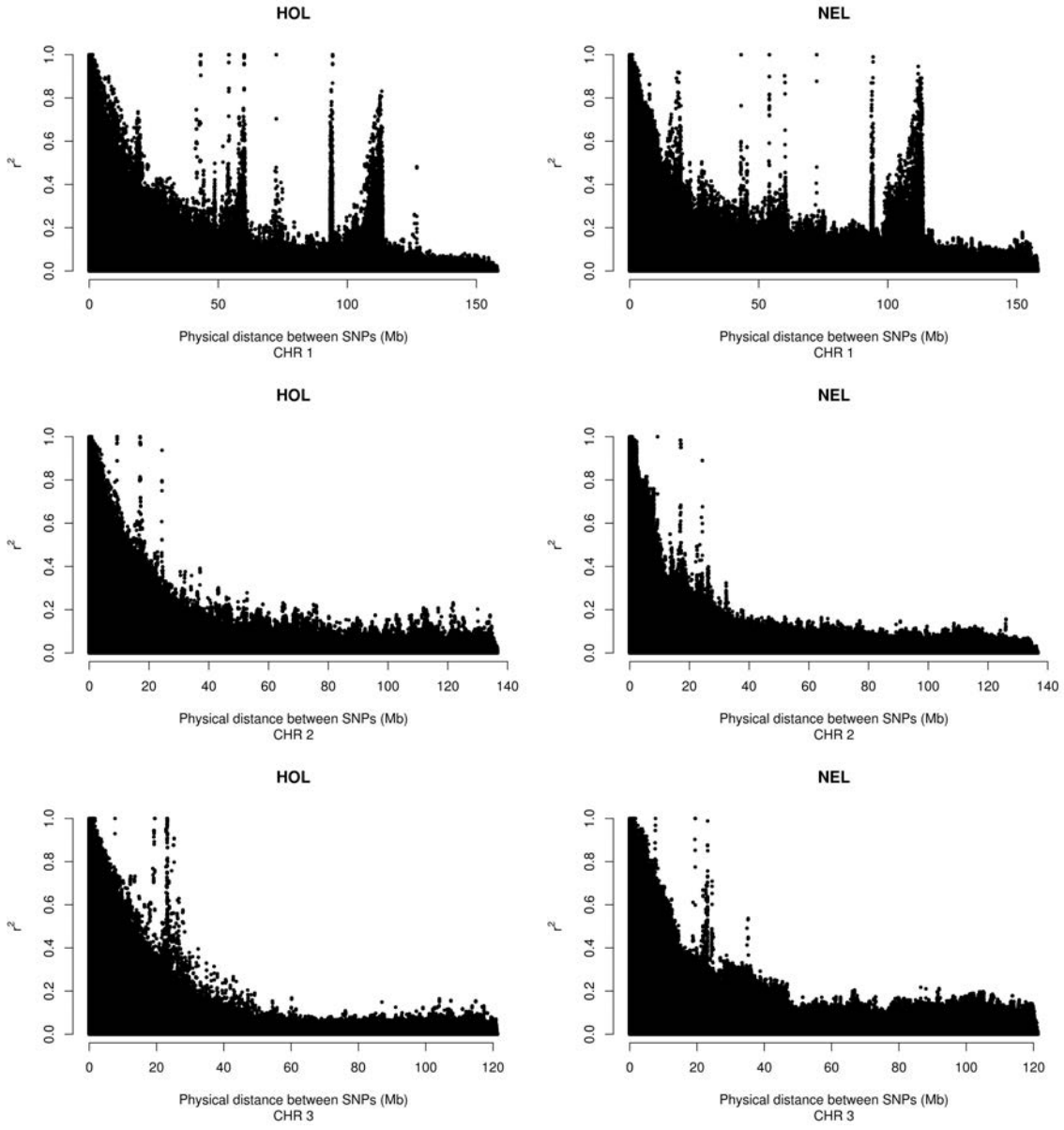
References

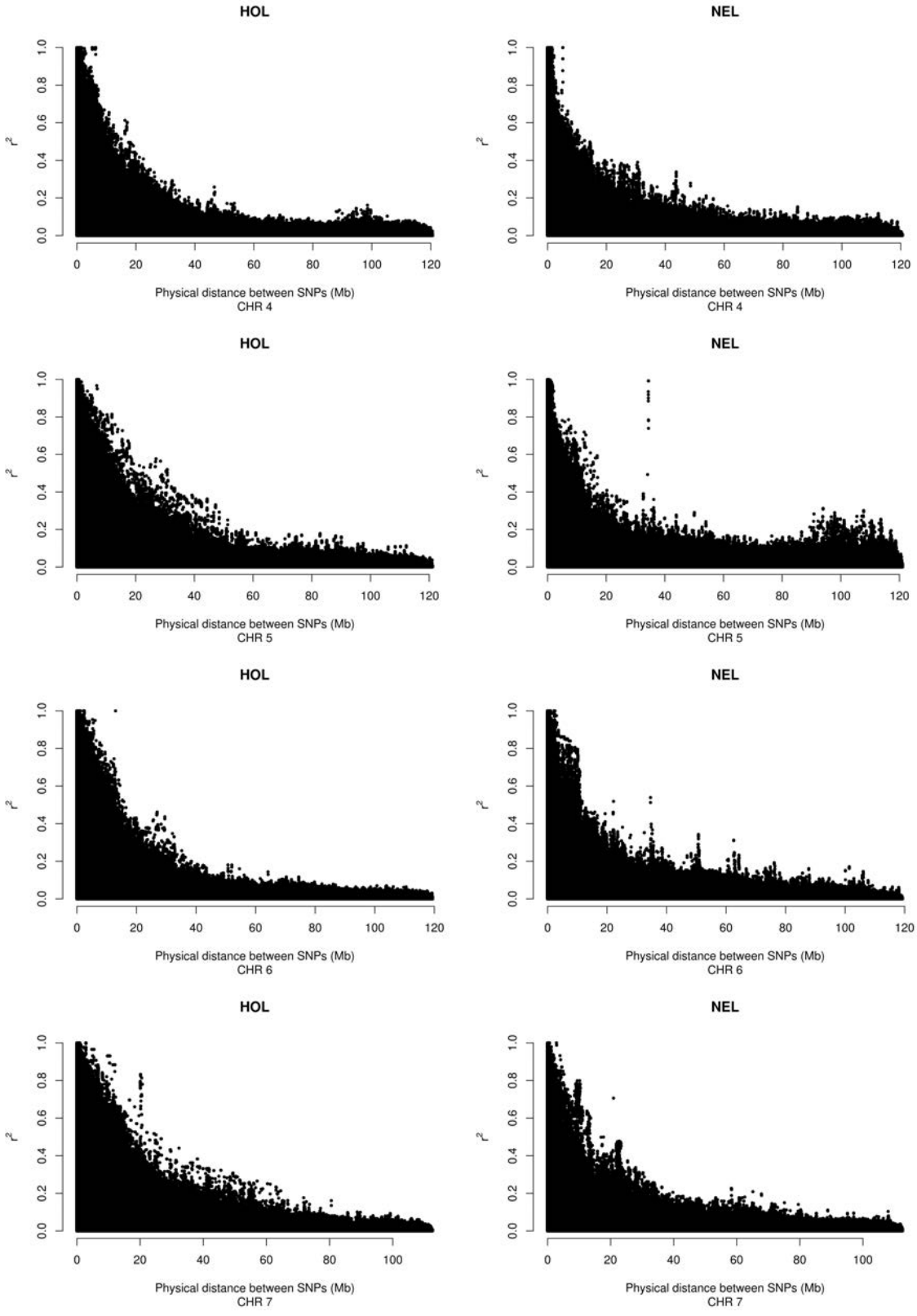
1. Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science 324: 522–528. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2943200&tool=pmcentrez&rendertype=abstract. Accessed 27 January 2014.
2. Liu Y, Qin X, Song X-ZH, Jiang H, Shen Y, et al. (2009) Bos taurus genome assembly. BMC Genomics 10: 180. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2686734&tool=pmcentrez&rendertype=abstract. Accessed 14 January 2014.
3. Zimin A V, Delcher AL, Florea L, Kelley DR, Schatz MC, et al. (2009) A whole-genome assembly of the domestic cow, Bos taurus. Genome Biol 10: R42. Available: http://genomebiology.com/2009/10/4/R42. Accessed 5 March 2014.
4. Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet 13: 36–46. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3324860&tool=pmcentrez&rendertype=abstract. Accessed 10 July 2014.
5. Salzberg SL, Yorke JA (2005) Beware of mis-assembled genomes. Bioinformatics 21: 4320–4321. Available: http://bioinformatics.oxfordjournals.org/content/21/24/4320.long. Accessed 20 January 2014.
6. Bohmanova J, Sargolzaei M, Schenkel FS (2010) Characteristics of linkage disequilibrium in North American Holsteins. BMC Genomics 11: 421. Available: http://www.biomedcentral.com/1471-2164/11/421. Accessed 14 January 2014.
7. Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, et al. (2014) Accuracy of genotype imputation in Nelore cattle. Genet Sel Evol 46: 69. Available: http://www.gsejournal.org/content/46/1/69. Accessed 10 December 2014.
8. Illumina Inc (2012) BovineHD Genotyping BeadChip. 05. Available: http://www.illumina.com/Documents/products/datasheets/datasheet_bovineHD.pdf. Accessed 20 January 2015.
9. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, et al. (2014) Second-generation PLINK: rising to the challenge of larger and richer datasets. Available: http://arxiv.org/abs/1410.4803. Accessed 7 January 2015.
10. Utsunomiya YT, Pérez O'Brien AM, Sonstegard TS, Van Tassell CP, do Carmo AS, et al. (2013) Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. PLoS One 8:

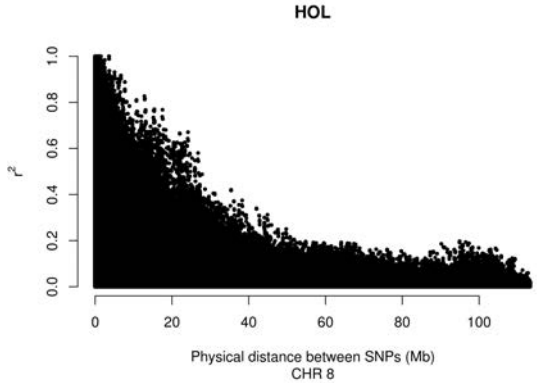
e64280. Available: http://dx.plos.org/10.1371/journal.pone.0064280. Accessed 9 January 2014.

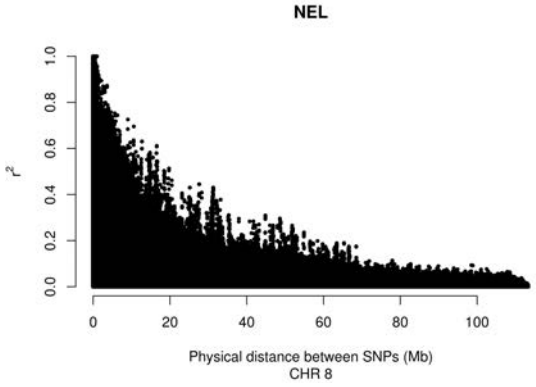
Figures

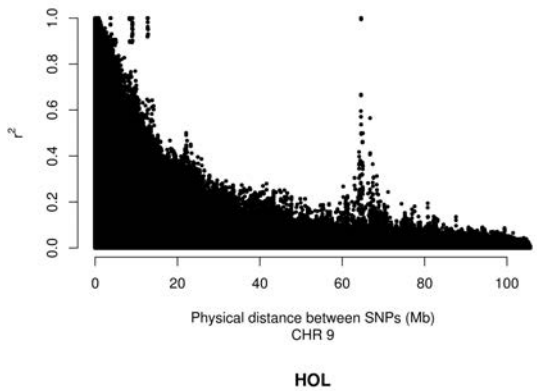
Figure 1 – Decay of linkage disequilibrium with distance in all bovine autosome chromosomes in Holstein (left) and Nellore (right) breeds.
Each black dot represents a pair-wise r2

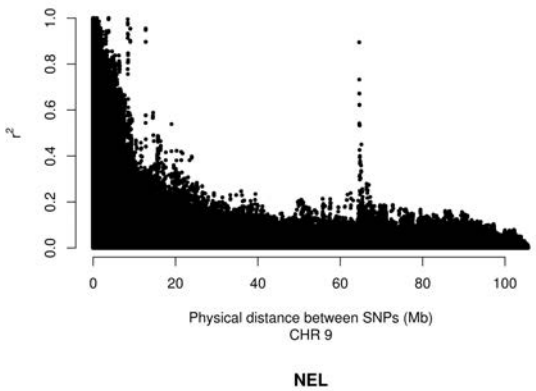


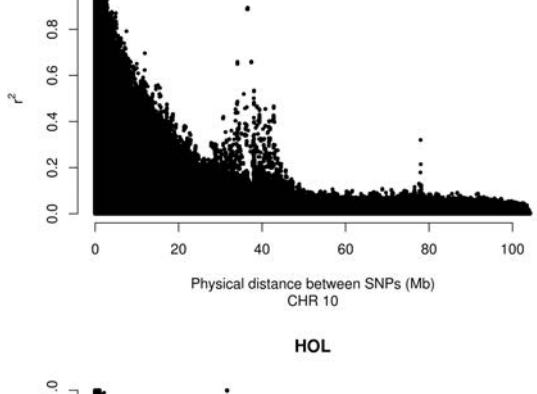

HOL
NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Physical distance between SNPs (Mb)
CHR 4
CHR 5
CHR 6
CHR 7

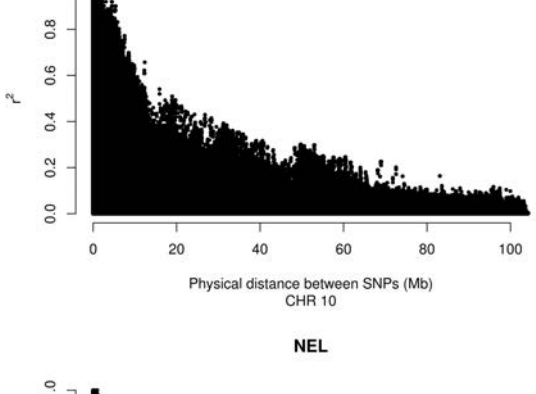

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 8

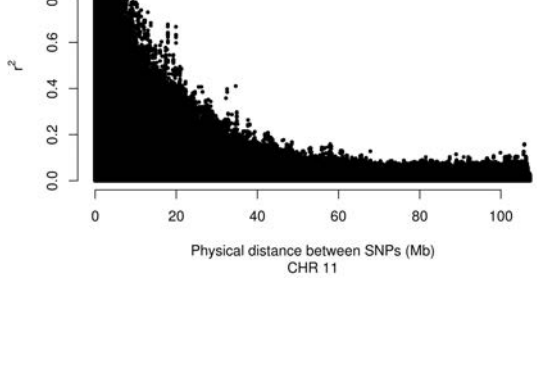

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 8

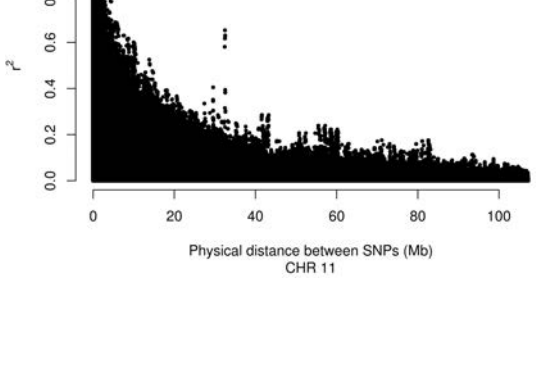

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 9

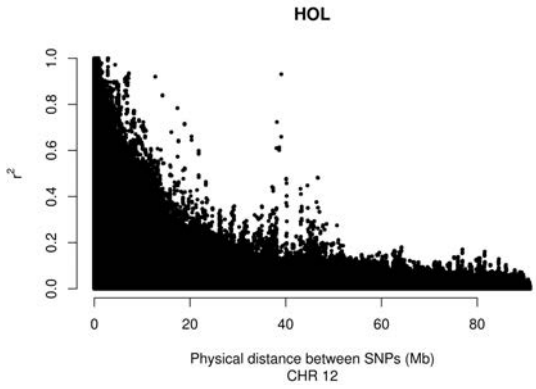

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 9

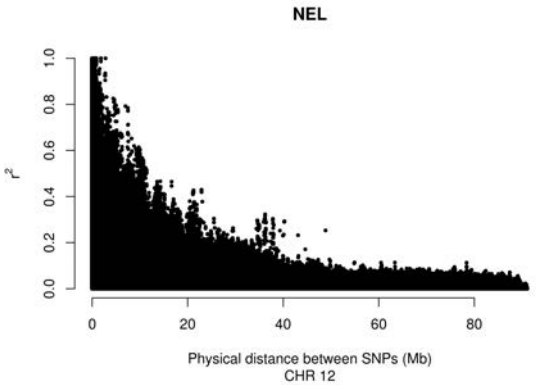

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 10

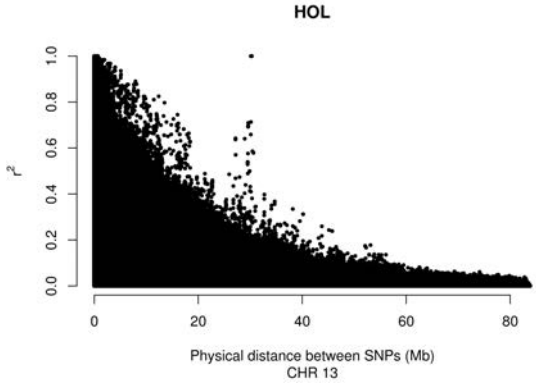

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 10

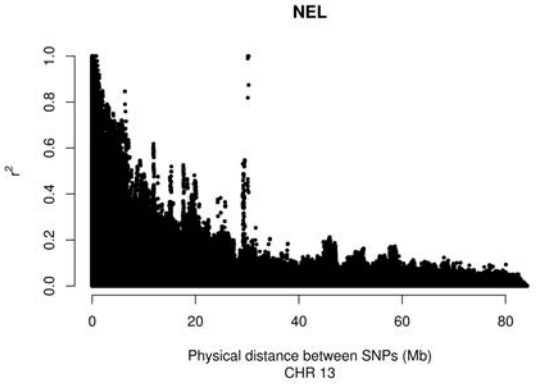

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 11

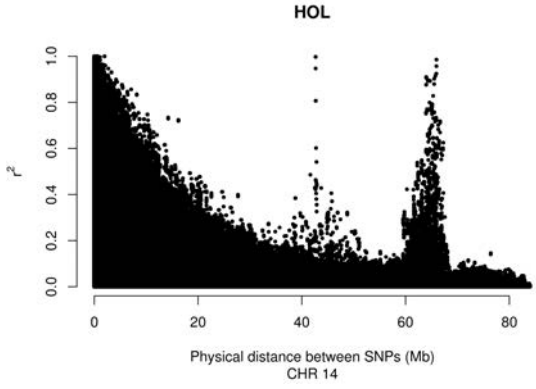

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
Physical distance between SNPs (Mb)
CHR 11

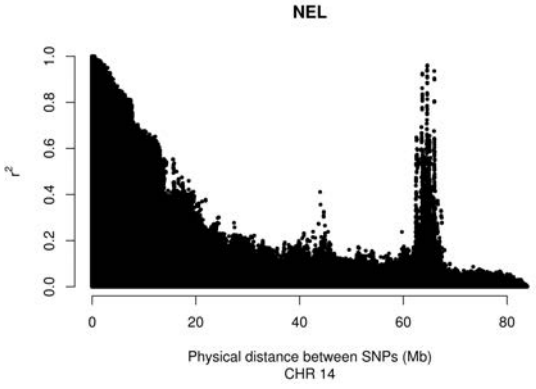

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 12

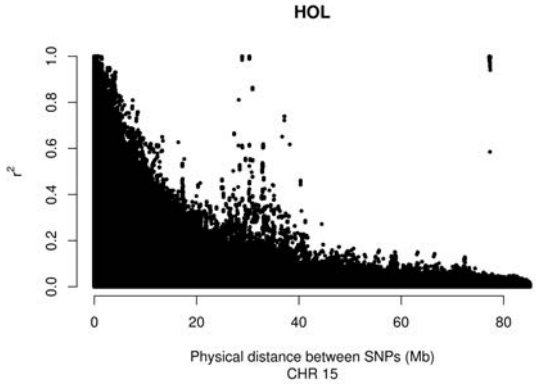

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 12

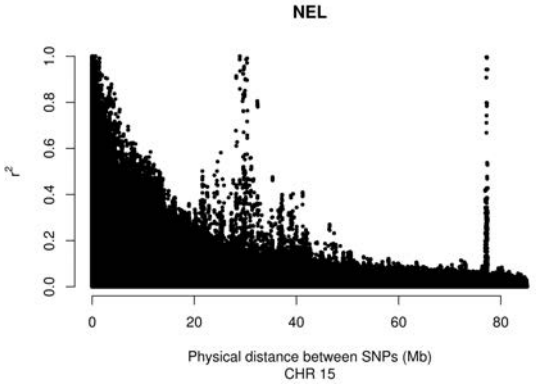

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 13

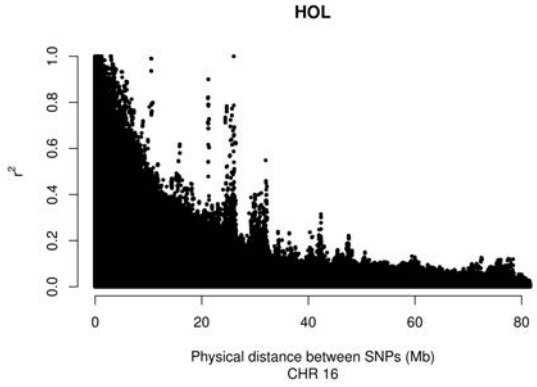

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 13

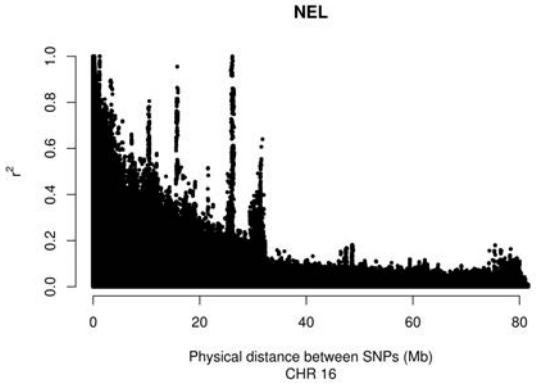

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 14

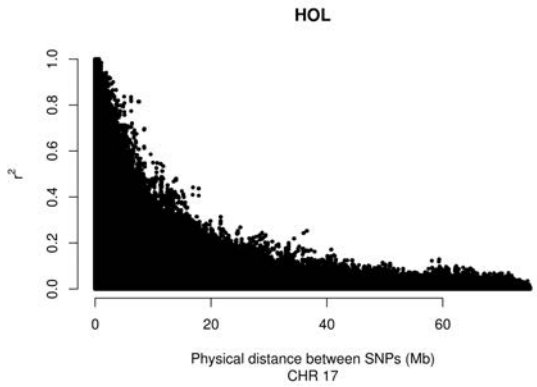

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 14

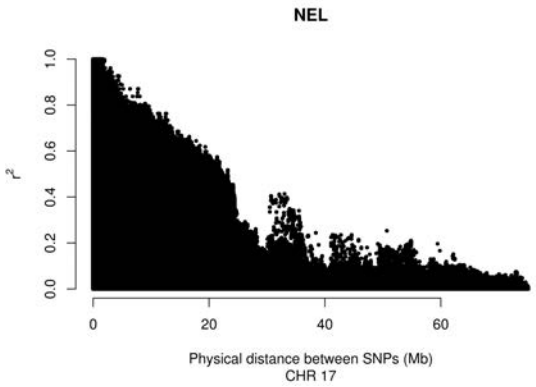

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 15

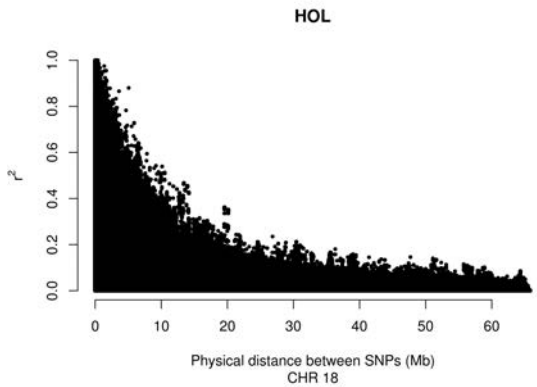

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 15

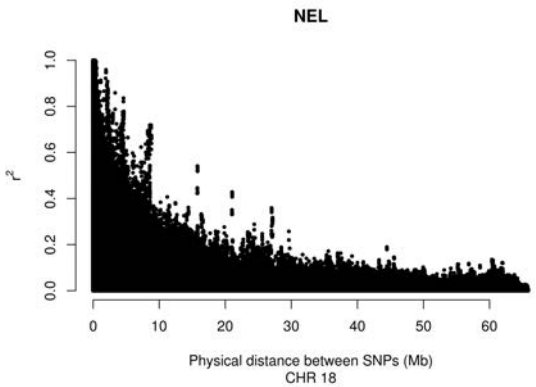

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 16

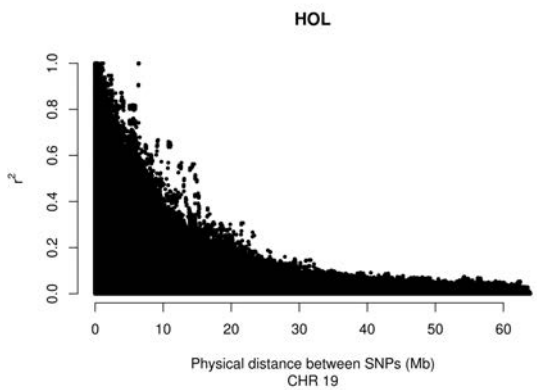

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Physical distance between SNPs (Mb)
CHR 16

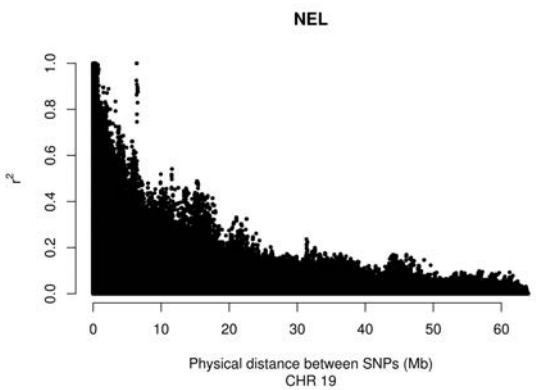

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
Physical distance between SNPs (Mb)
CHR 17

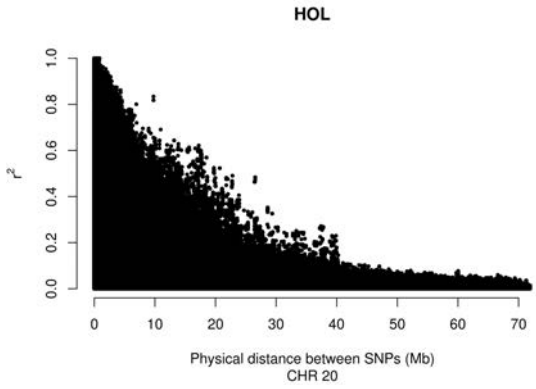

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
Physical distance between SNPs (Mb)
CHR 17

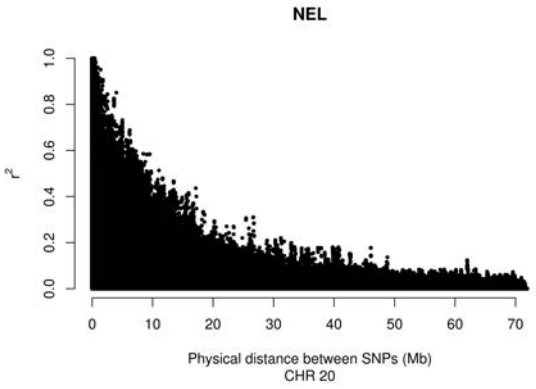

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 18

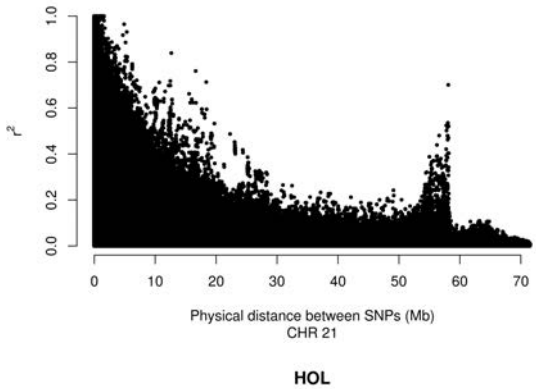

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 18

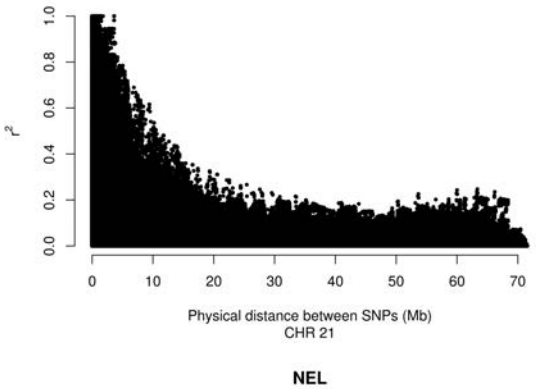

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 19

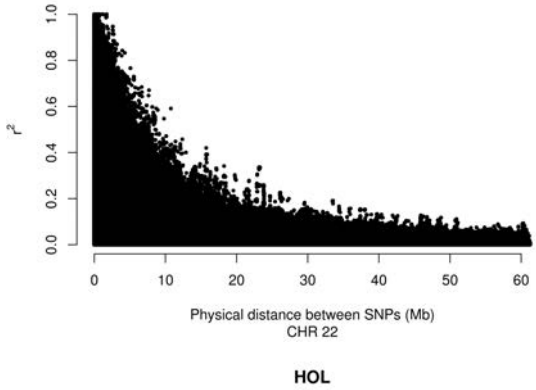

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 19

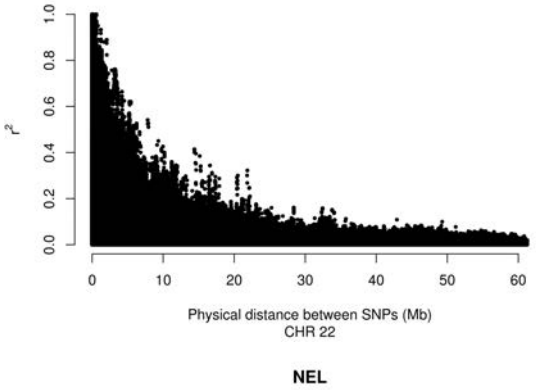

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Physical distance between SNPs (Mb)
CHR 20

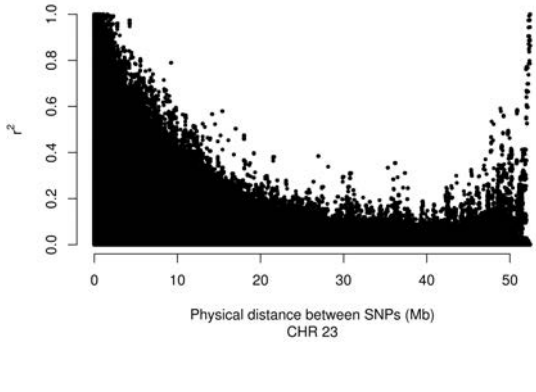

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Physical distance between SNPs (Mb)
CHR 20

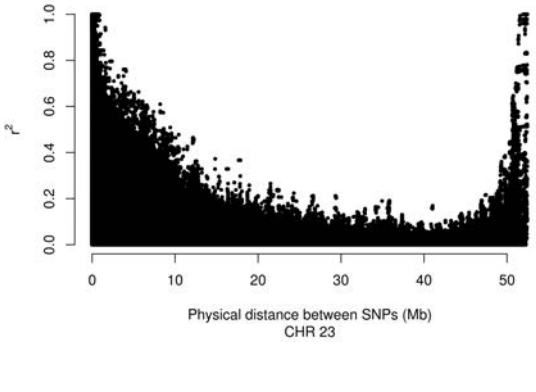

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Physical distance between SNPs (Mb)
CHR 21

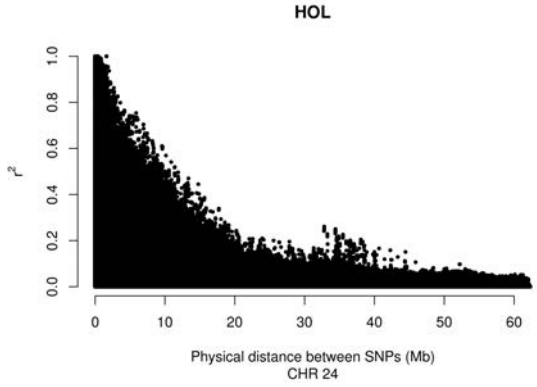

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Physical distance between SNPs (Mb)
CHR 21

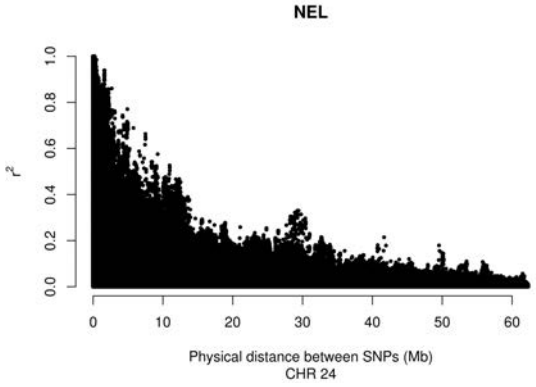

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 22

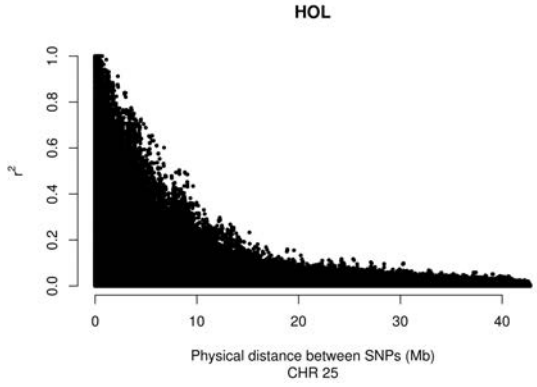

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 22

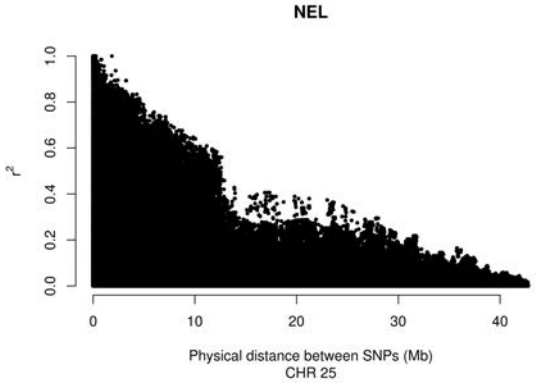

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Physical distance between SNPs (Mb)
CHR 23

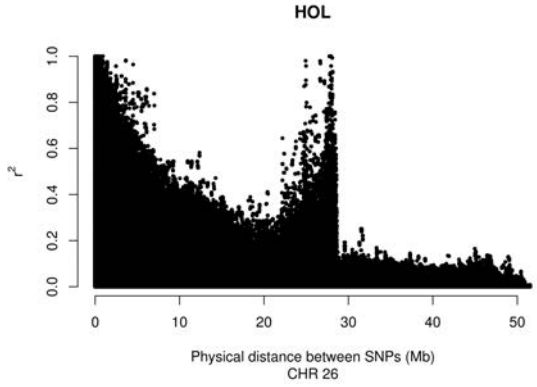

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Physical distance between SNPs (Mb)
CHR 23

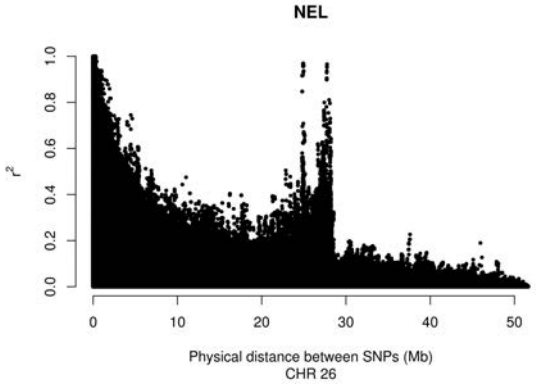

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 24

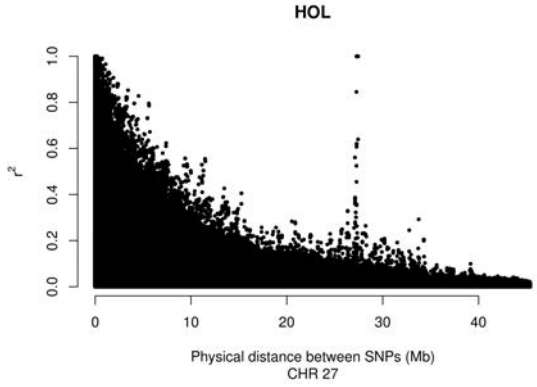

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Physical distance between SNPs (Mb)
CHR 24

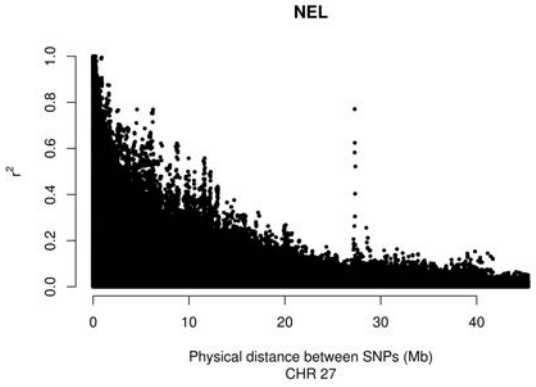

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Physical distance between SNPs (Mb)
CHR 25

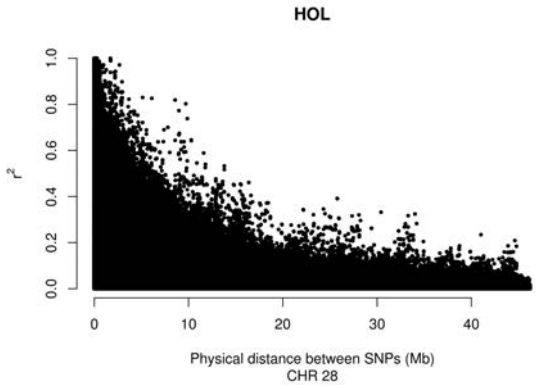

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Physical distance between SNPs (Mb)
CHR 25

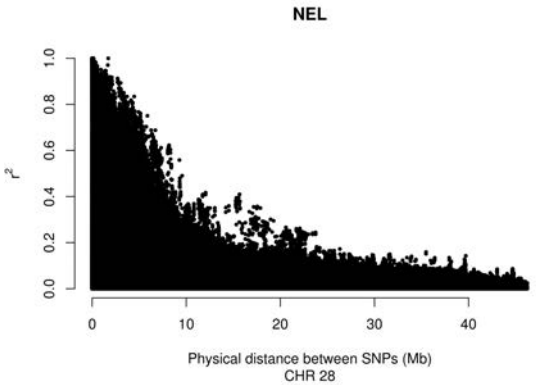

HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Physical distance between SNPs (Mb)
CHR 26

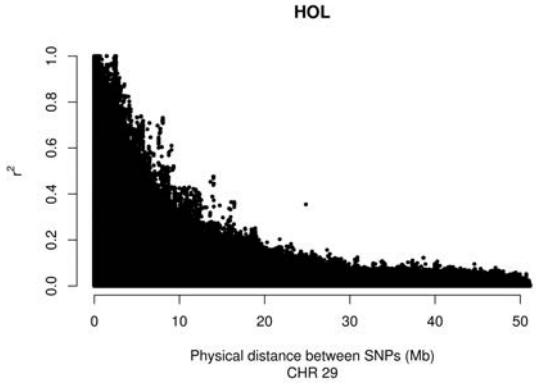

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Physical distance between SNPs (Mb)
CHR 26


HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Physical distance between SNPs (Mb)
CHR 27


NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Physical distance between SNPs (Mb)
CHR 27


HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Physical distance between SNPs (Mb)
CHR 28


NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
Physical distance between SNPs (Mb)
CHR 28


HOL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Physical distance between SNPs (Mb)
CHR 29

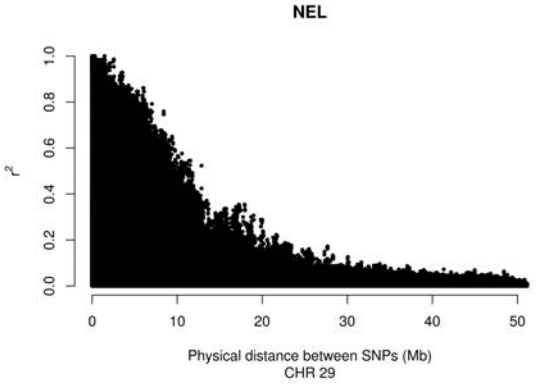
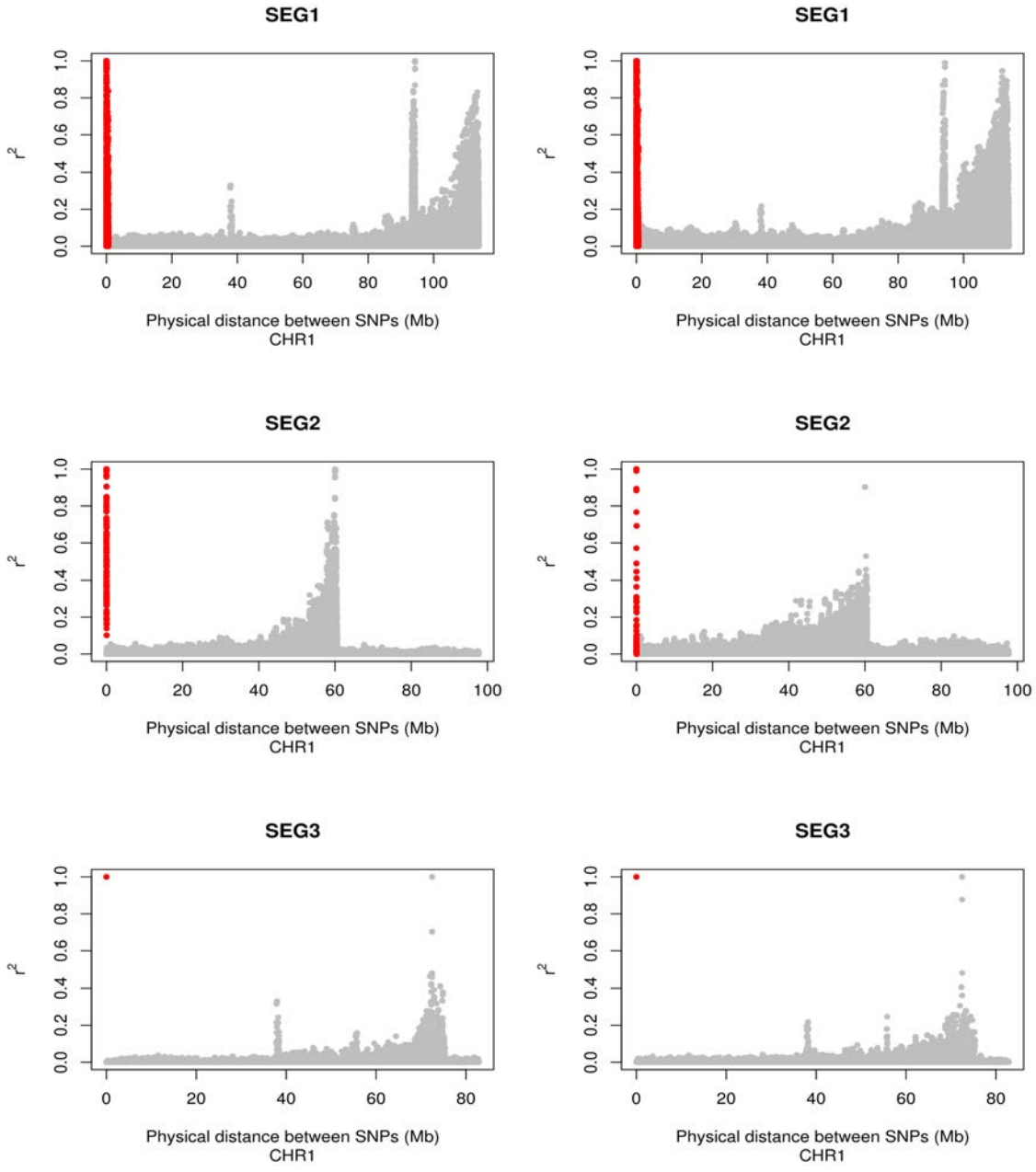
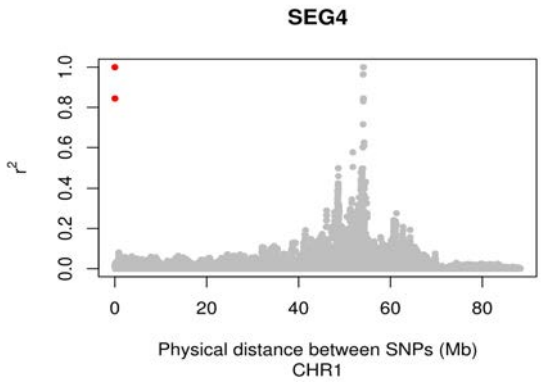
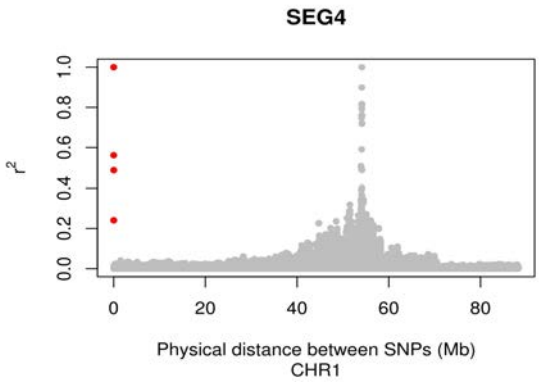
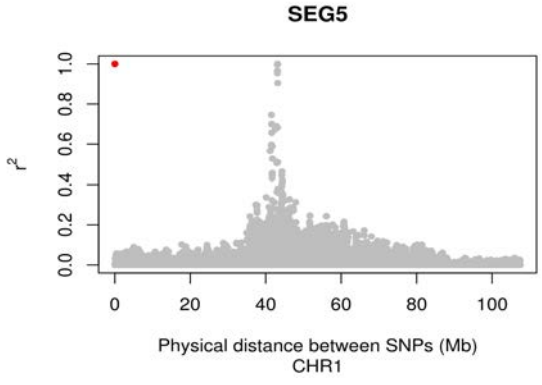

NEL
r^2
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Physical distance between SNPs (Mb)
CHR 29

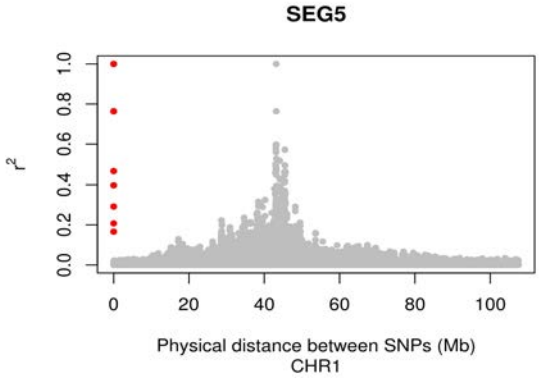
Figure 2 – Linkage disequilibrium decay of the mis-assembled segments in each chromosome in Holstein (left) and Nellore (right)
Gray points represent pair-wise r^2 between SNPs. Red points represent only pair-wise r^2 between SNPs within the segment

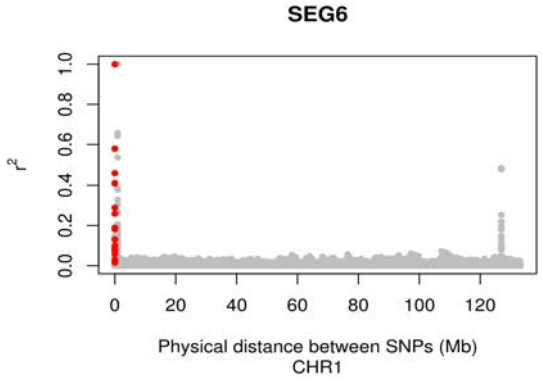


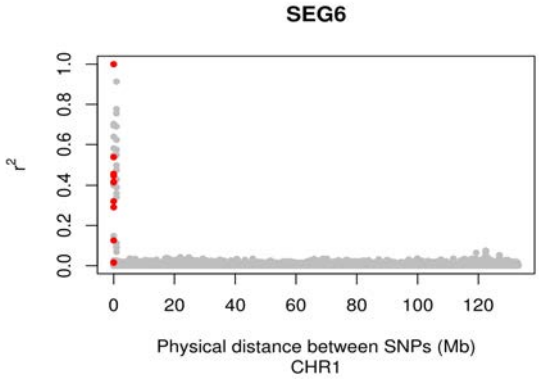

SEG4
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR1

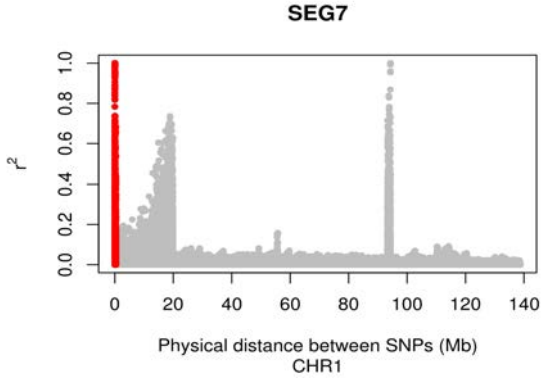

SEG4
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR1

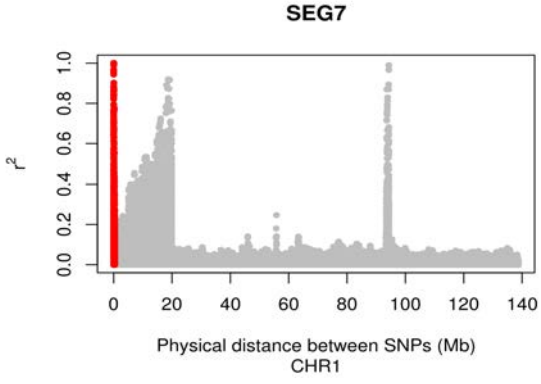

SEG5
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR1

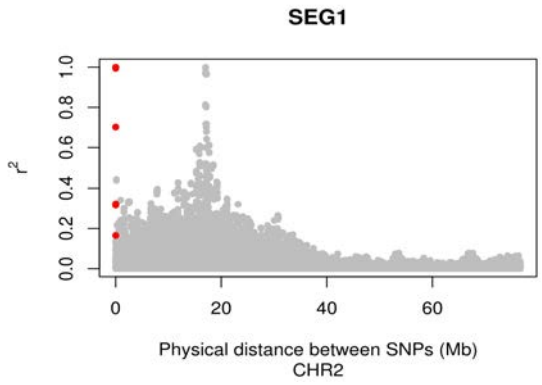

SEG5
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR1

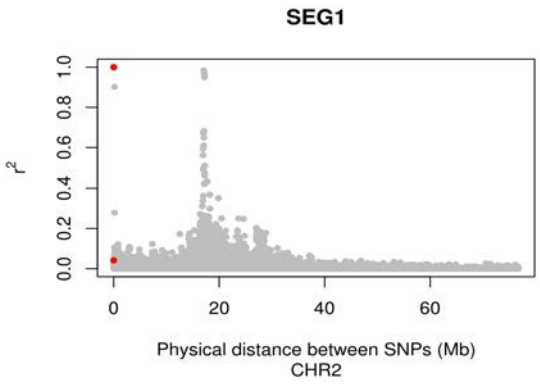

SEG6
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100 120
Physical distance between SNPs (Mb)
CHR1

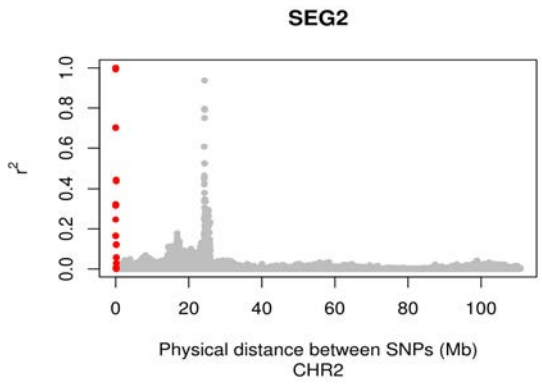

SEG6
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100 120
Physical distance between SNPs (Mb)
CHR1

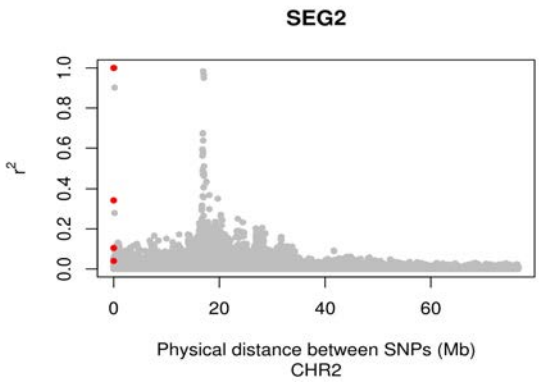

SEG7
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100 120 140
Physical distance between SNPs (Mb)
CHR1

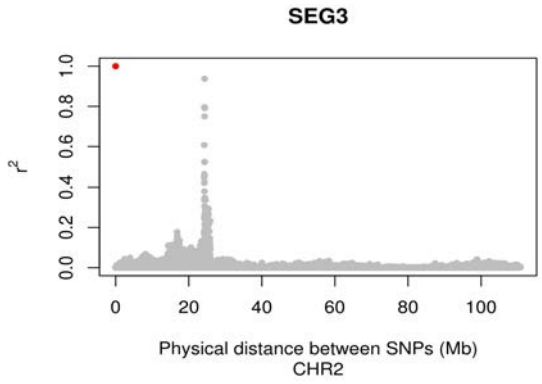

SEG7
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100 120 140
Physical distance between SNPs (Mb)
CHR1

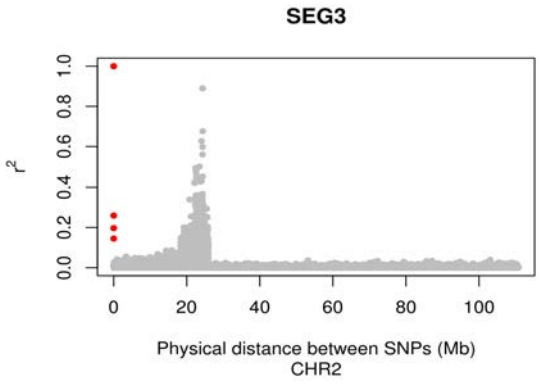

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60
Physical distance between SNPs (Mb)
CHR2

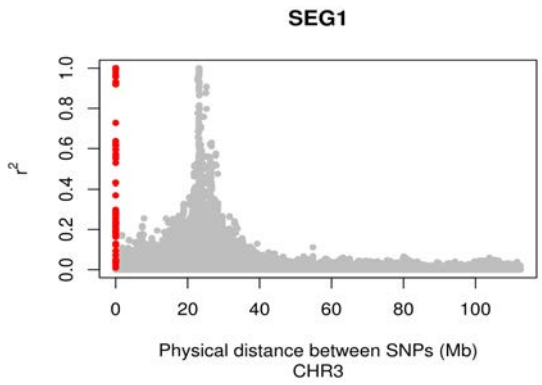

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60
Physical distance between SNPs (Mb)
CHR2

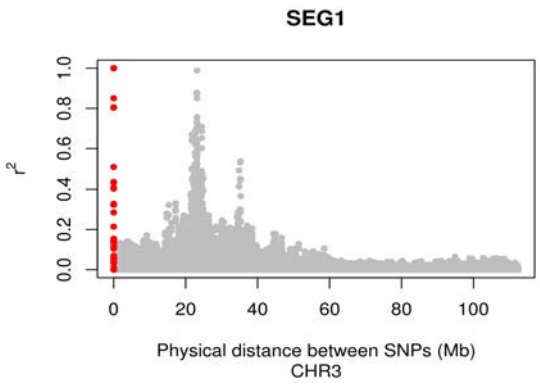

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR2

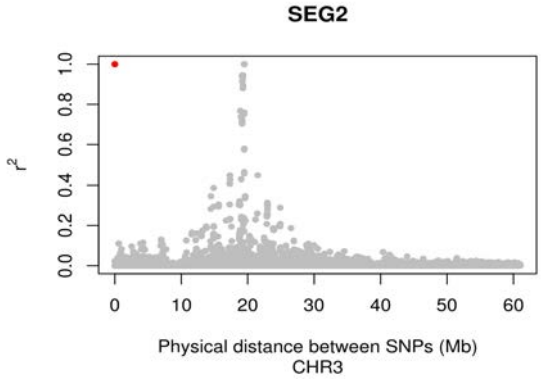

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60
Physical distance between SNPs (Mb)
CHR2

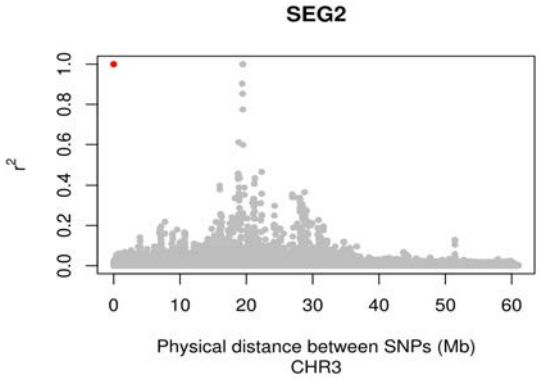

SEG3
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR2

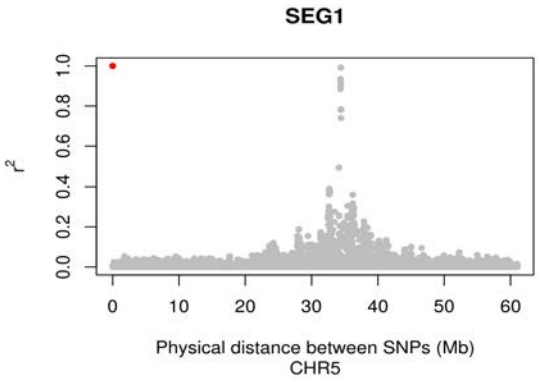

SEG3
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR2

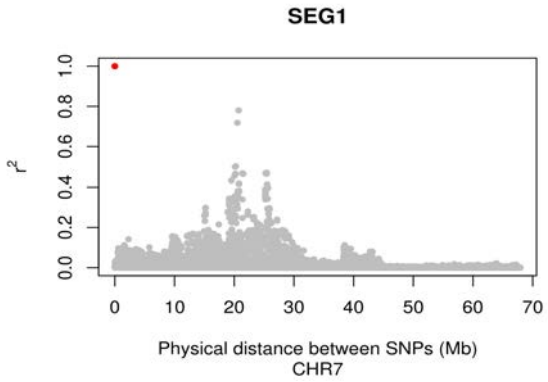

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR3

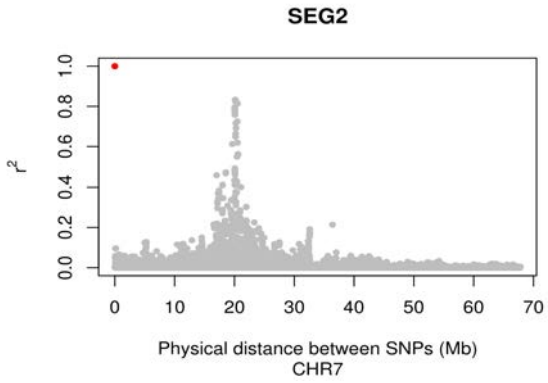

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR3

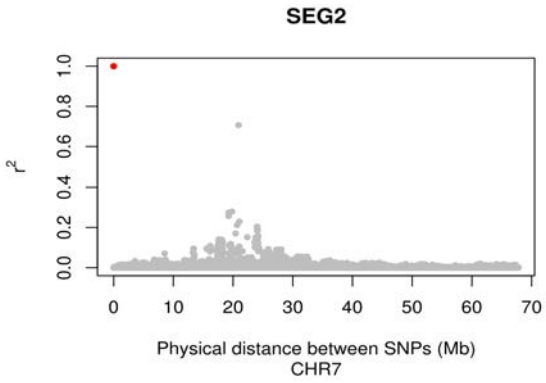

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR3

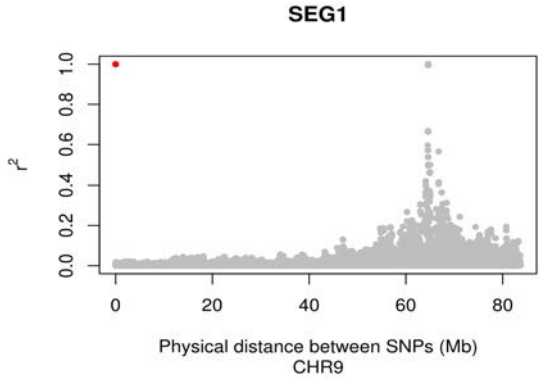

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR3

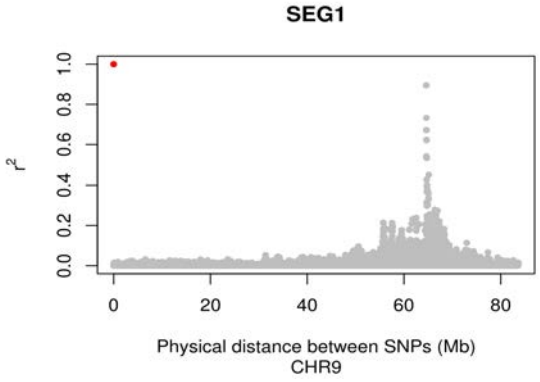

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR5

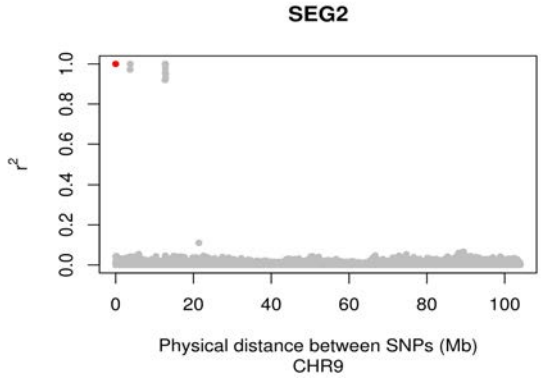

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR7

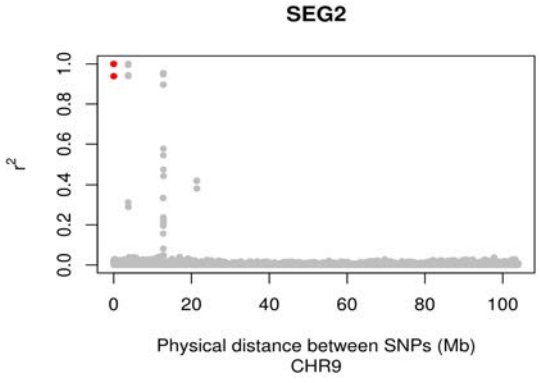

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR7

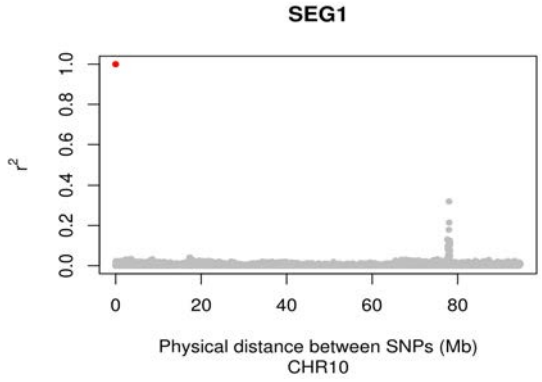

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR7

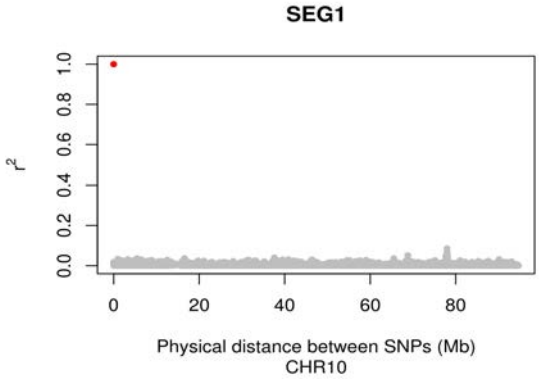

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR9

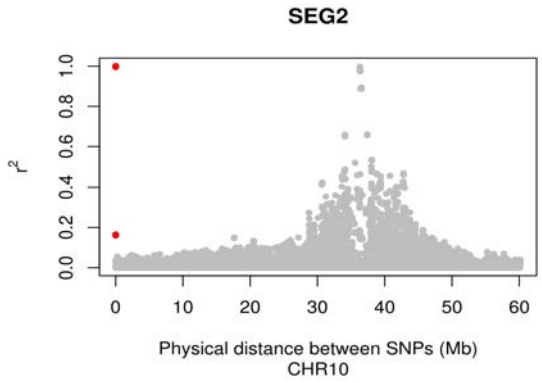

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR9

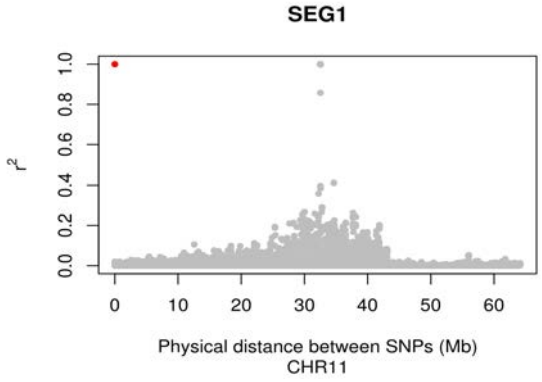

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR9

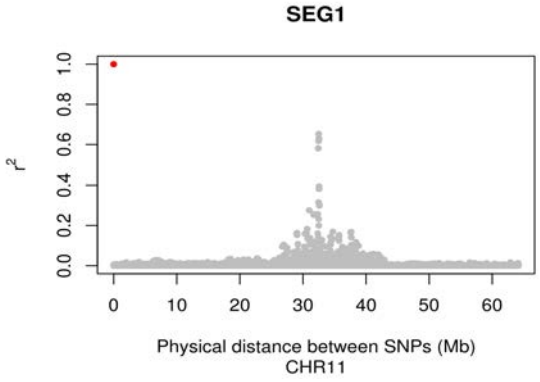

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80 100
Physical distance between SNPs (Mb)
CHR9

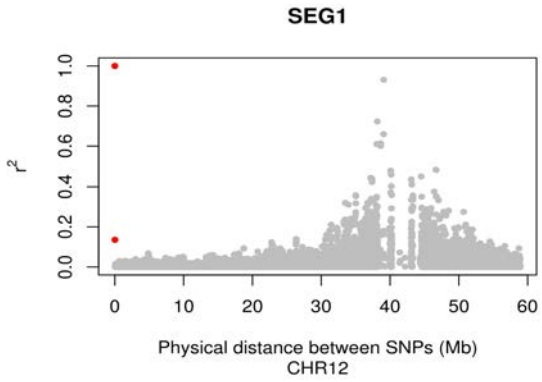

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR10

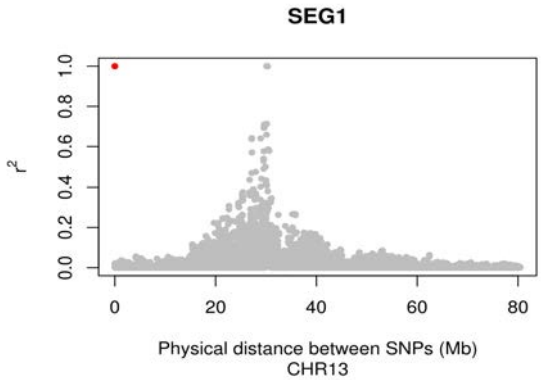

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR10

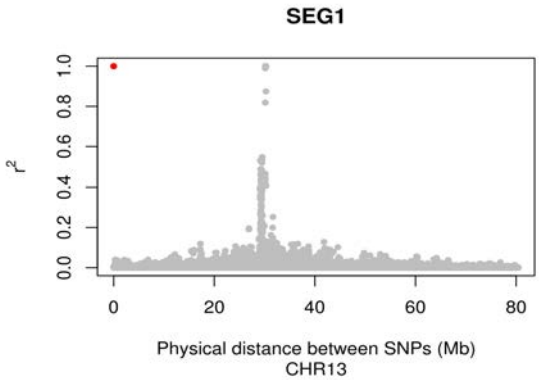

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR10

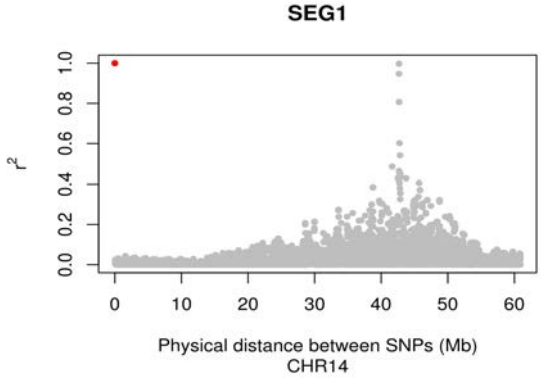

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR11

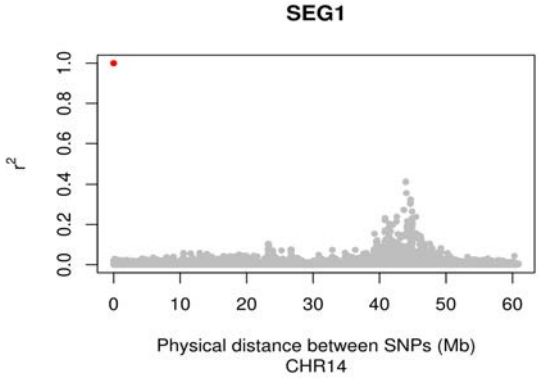

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR11

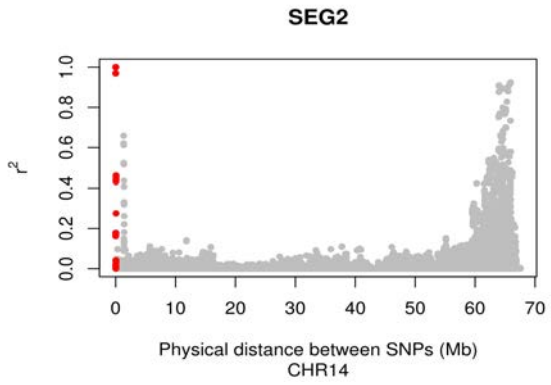

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR12

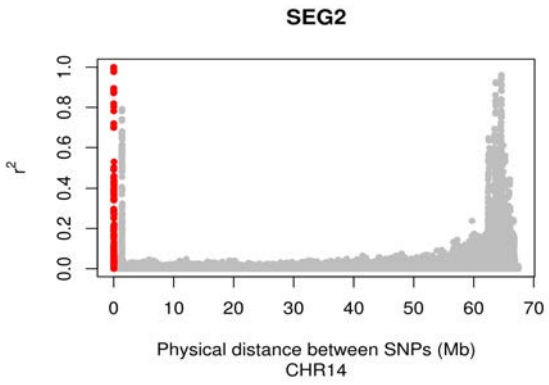

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR13

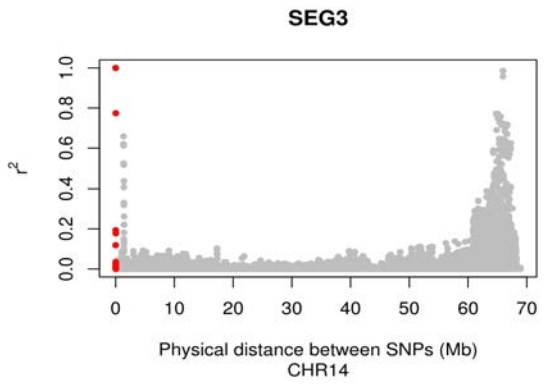

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR13

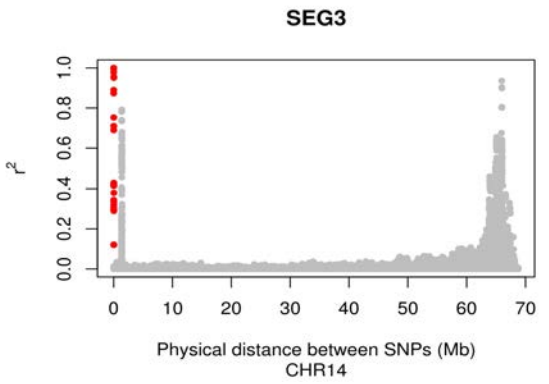

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR14

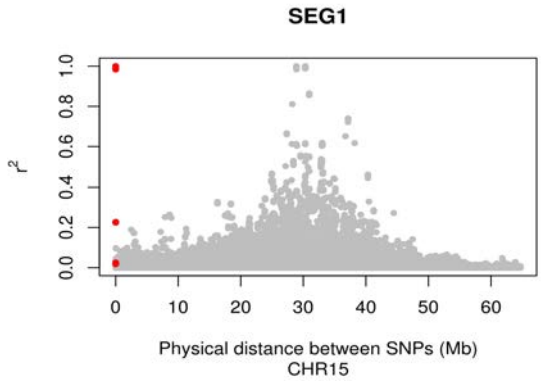

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR14

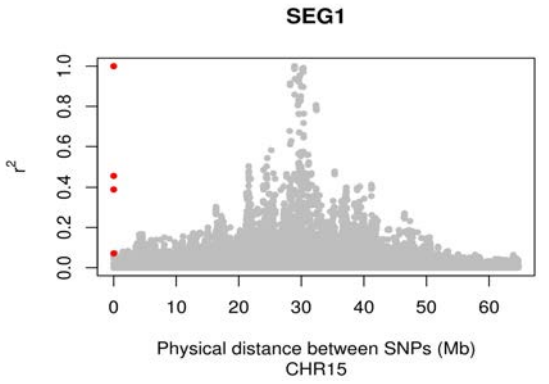

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR14

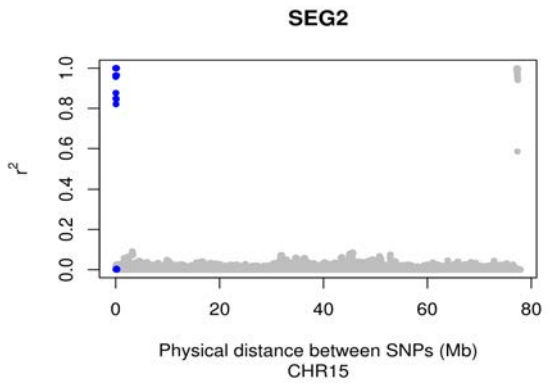

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR14

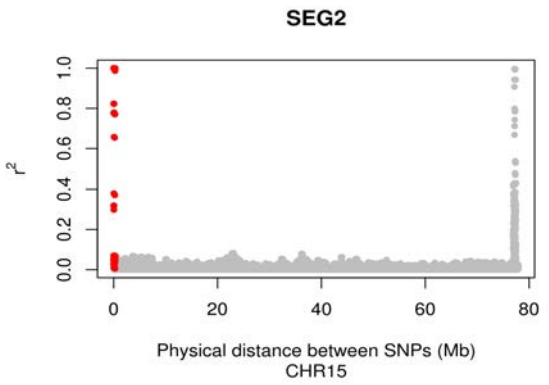

SEG3
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR14

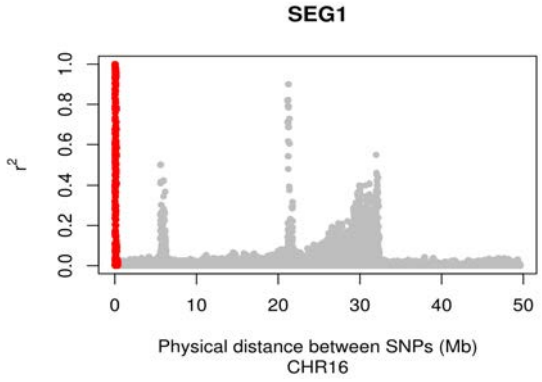

SEG3
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR14

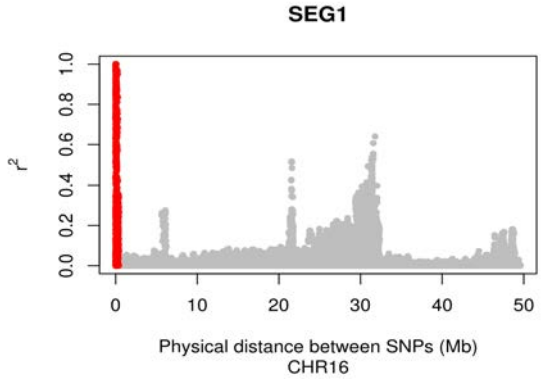

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR15

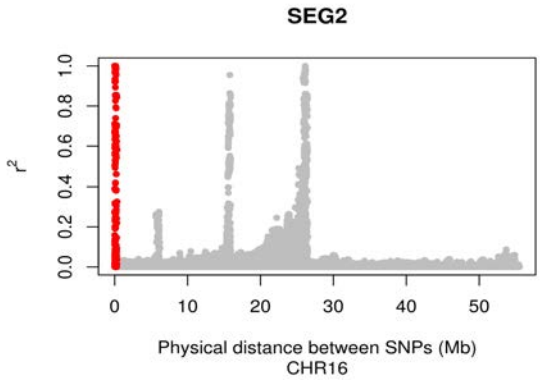

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR15

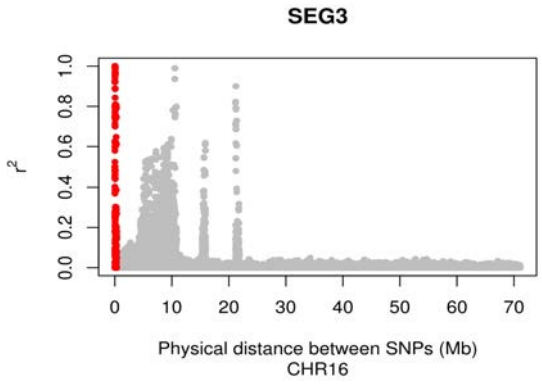

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR15

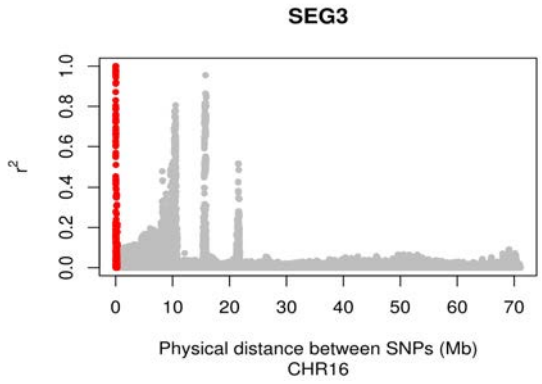

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 20 40 60 80
Physical distance between SNPs (Mb)
CHR15

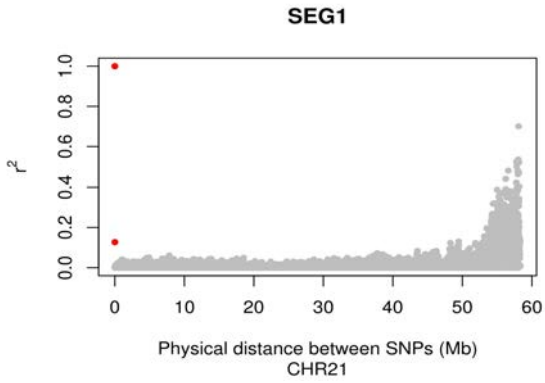

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50
Physical distance between SNPs (Mb)
CHR16

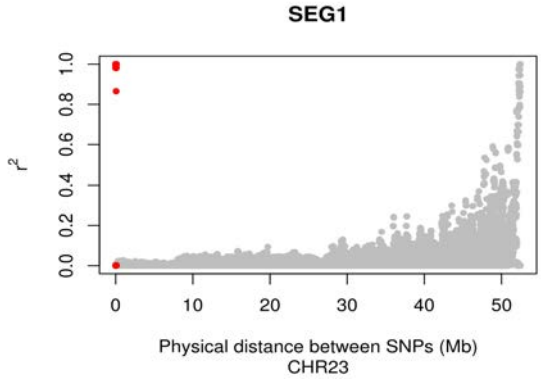

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50
Physical distance between SNPs (Mb)
CHR16

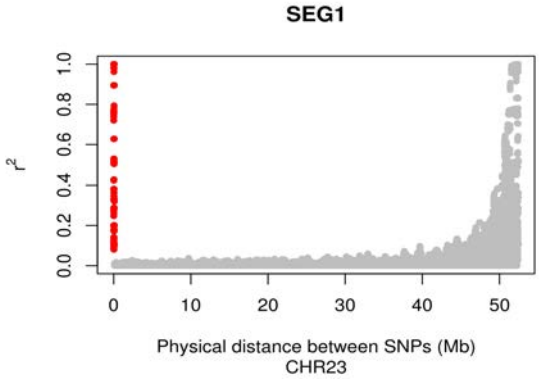

SEG2
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50
Physical distance between SNPs (Mb)
CHR16

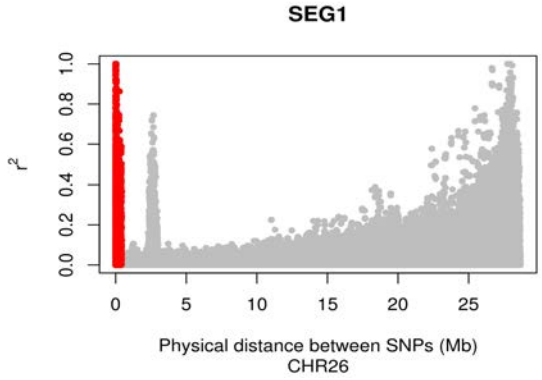

SEG3
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR16

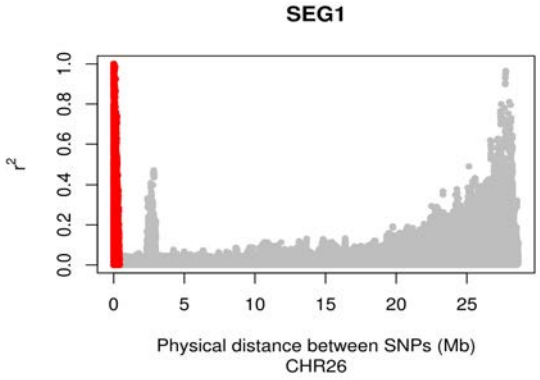

SEG3
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60 70
Physical distance between SNPs (Mb)
CHR16

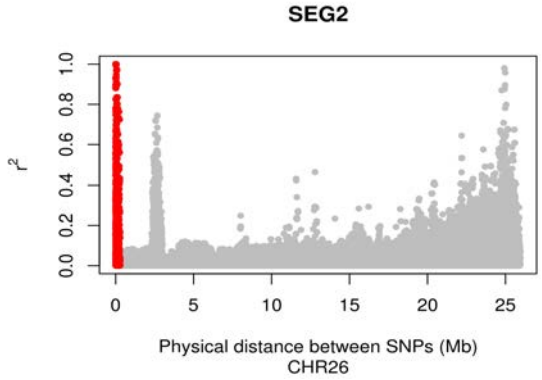

SEG1
r²
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50 60
Physical distance between SNPs (Mb)
CHR21

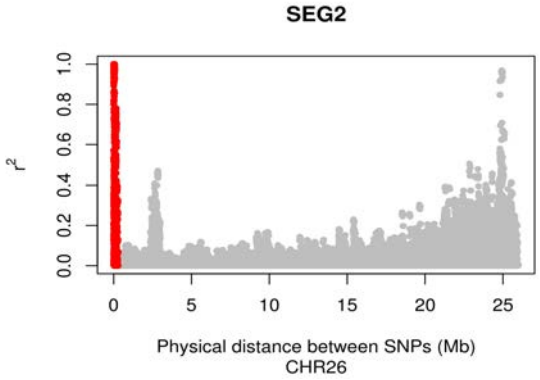

SEG1
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50
Physical distance between SNPs (Mb)
CHR23

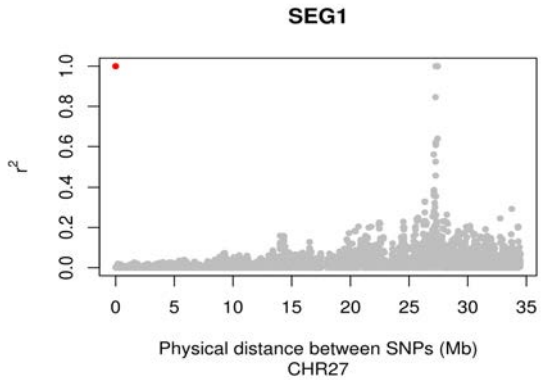

SEG1
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 10 20 30 40 50
Physical distance between SNPs (Mb)
CHR23


SEG1
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20 25
Physical distance between SNPs (Mb)
CHR26


SEG1
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20 25
Physical distance between SNPs (Mb)
CHR26


SEG2
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20 25
Physical distance between SNPs (Mb)
CHR26


SEG2
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20 25
Physical distance between SNPs (Mb)
CHR26


SEG1
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20 25 30 35
Physical distance between SNPs (Mb)
CHR27

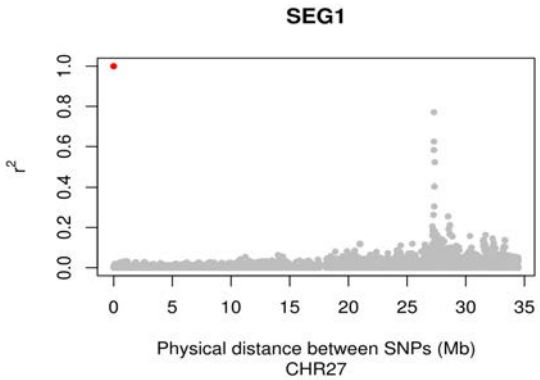
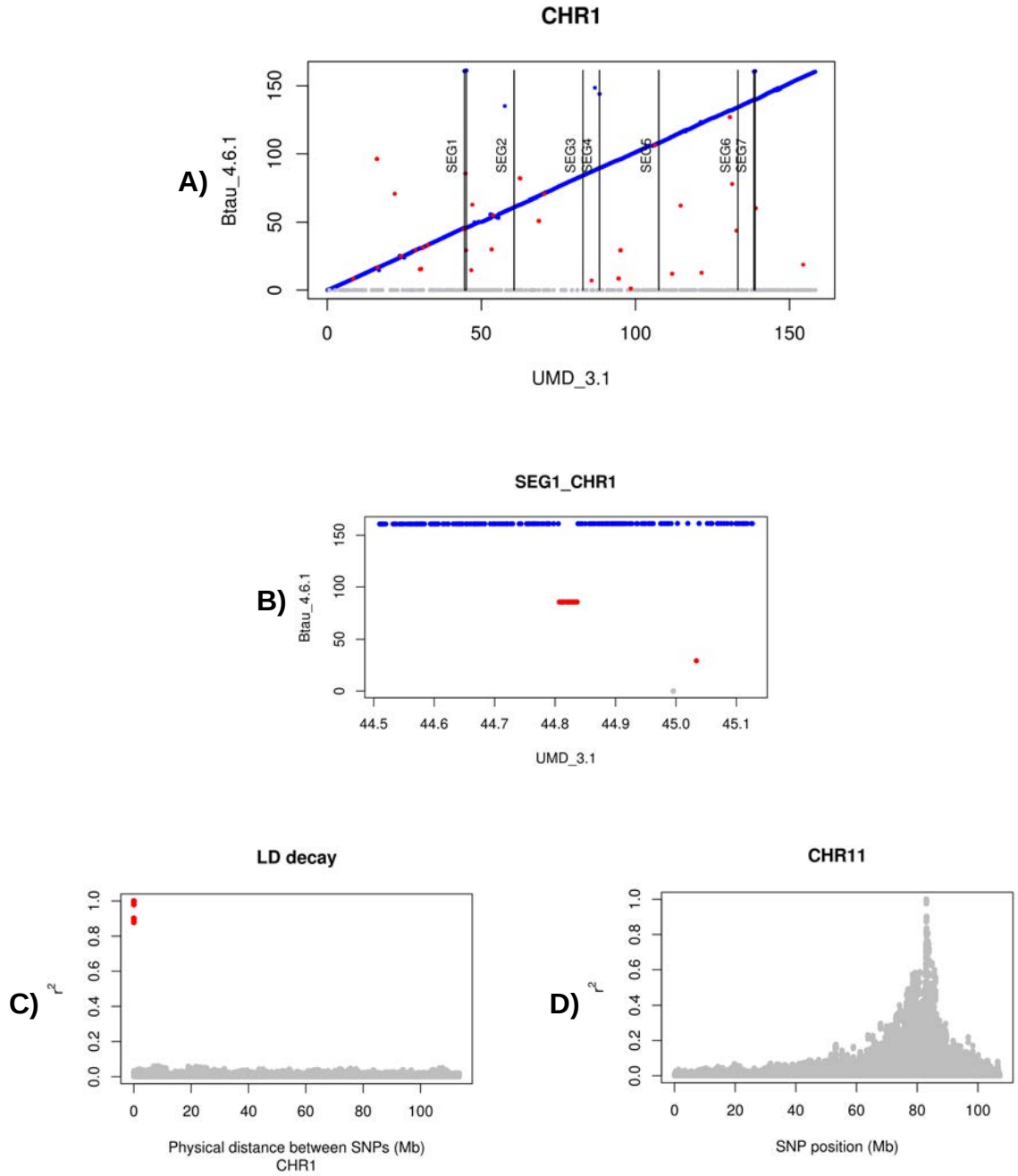

SEG1
r^2
0.0 0.2 0.4 0.6 0.8 1.0
0 5 10 15 20 25 30 35
Physical distance between SNPs (Mb)
CHR27

Figure 3 – Comparisons of SNP positions in UMD_3.1 and Btau_4.6.1 (A), positions of the SNPs in UMD_3.1 and Btau_4.6.1 in the mis-assembled segment 1 (SEG1) on chromosome 1 (B), linkage disequilibrium decay of the SNPs within SEG1 (C) and pair-wise r2 of SNPs within SEG1 with SNPs on chromosome 11 (D)

In A and B, blue dots represent SNPs mapped in both assemblies, red dots represent SNPs mapped in different chromosomes in Btau_4.6.1 and grey dots represent unmapped SNPs in Btau_4.6.1. Gray dots in C and D represent pair-wise r^2 and red dots in C represent only pair-wise r^2 of SNPs within the SEG1.



Tables

Table 1 – Mis-assembled segments in bovine UMD_3.1 assembly

CHR	Segment #	start (bp)	end(bp)	SL(bp)	# contigs/misplaced SNP*	AVCS	# of SNPs
1	SEG1	44507685	45130787	623102	22	28323	192
	SEG2	60578448	60664293	85845	6	14308	33
	SEG3	82940221	82941440	1219	1	129	1
	SEG4	88322835	88354289	31454	4	7864	5
	SEG5	107558346	107570195	11849	2	5924	7
	SEG6	133244334	133252572	8238	1	8238	7
	SEG7	138430561	138828418	397857	15	26524	121
2	SEG1	60019097	60060716	41619	3	13873	15
	SEG2	60202426	60219502	17076	1	17076	5
	SGE3	110788577	110799585	11008	1	11008	3
3	SEG1	8914006	8967656	53650	1	53650	19
	SEG2	60380428	60380428	0	BovineHD0300018117	0	1
5	SEG1	60040681	60040681	0	BovineHD0500016805	0	1
7	SEG1	44669000	44706954	37954	1	37954	1
	SEG2	44828989	44831339	2350	1	2350	1
9	SEG1	83682109	83682109	0	BovineHD0900023340		1
	SEG2	104019930	104053005	33075	1	33075	6
10	SEG1	9619857	9620972	1115	1	1115	1
	SEG2	60192840	60212611	19771	1	19771	6
11	SEG1	43128807	43129498	691	1	691	1
12	SEG1	32162421	32199628	37207	1	37207	2
13	SEG1	3775413	3775413	0	BovineHD1300001018		1
14	SEG1	66414152	66414152	0	Hapmap32295-BTA-129360		1
	SEG2	68089840	68179075	89235	4	22309	22
	SEG3	69481624	69520614	38990	4	9748	11
15	SEG1	20558399	20585624	27225	2	13612	6
	SEG2	77714908	78032207	317299	9	35255	17
16	SEG1	49258388	49671862	413474	32	19921	77
	SEG2	55233099	55573752	340653	13	26204	82
	SEG3	70813807	71152977	339170	30	11305	52
21	SEG1	13229171	13246830	17659	3	5886	2
23	SEG1	52353182	52478629	125447	4	31362	34
26	SEG1	23006493	23444622	438129	14	31295	179
	SEG2	25715286	26015674	300388	18	16688	82
27	SEG1	10912446	10914274	1828	1	1828	1

CHR – chromosome
SL – segment length
* Number of contigs within the mis-assembled segment or SNP name for the cases where the SNPs were considered misplaced
AVCS – Average contig size

CAPÍTULO 3 - Genome wide scan for milk production traits and age at first calving in Gyr dairy cattle (*Bos indicus*)

Adam Taiti Harth Utsunomiya[1§], Yuri Tani Utsunomiya[1], Daniel Jordan de Abreu Santos[1], Solomon Antwi Boison[2], Rui da Silva Verneque[4], Marco Antônio Machado[4], João Cláudio do Carmo Panetto[4], André Rabello Fernandes[4], Johann Sölkner[2], José Fernando Garcia[1], Ricardo da Fonseca[1], Marcos Vinícius Gualberto Barbosa da Silva[4§]

[1]Univ Estadual Paulista, Faculdade de Ciências Agrárias e Veterinárias Campus de Jaboticabal, Jaboticabal, São Paulo, Brazil
[2]Universität fur Bodenkultur, Vienna, Austria
[3]Universidade del Sacro Cuore, Piacenza, Italy
[4]EMBRAPA dairy cattle, Juiz de Fora, Minas Gerais, Brazil

[§]Corresponding authors

Email addresses:
ATHU: adamtaiti@gmail.com
YTU: ytutsunomiya@gmail.com
DJAS: daniel_jordan2008@hotmail.com
SAB: soloboan@yahoo.com
RSV: rui.verneque@embrapa.br
MAM: marco.machado@embrapa.br
JCP: joao.panetto@embrapa.br
ARF: andre@girleiteiro.org.br
JS: johann.soelkner@boku.ac.at
JFG: jfgarcia@fmva.unesp.br
MVGBS: marcos.vb.silva@embrapa.br
RF: ricardo@dracena.unesp.br

Abstract

Background

A genome-wide scan for milk (MY), fat (FY), protein (PY) and total solids (TSY) yield, fat (FP), protein (PP) and total solids (TSP) percentage and age at first calving (AFC) in Brazilian Gyr dairy cattle (*Bos indicus*) was performed using deregressed estimated breeding values (dEBVs), penalized for their accuracies. 471 bulls and 1,616 cows were genotyped using Illumina® BovineHD BeadChip (HD) and Illumina® Bovine50SNP BeadChip v2 (50k) assay. Cows genotypes (50k) were imputed to HD for increasing the coverage of the genome. ClueGO application of Cytoscape were used for prospecting functional terms and build up biological networks.

Results

Genome-wide scan identified a number of chromosome regions associated with 7 out of 8 traits in Brazilian Gyr dairy cattle. For MY the regions merged to exocytosis pathway which regulates milk secretion, FY merged to NADP pathway which is associated with *de novo* fatty acid synthesis, TSY presented a term related to mammary gland immune response, FP did not presented direct functional terms related to milk fat amount but the most explored gene for FP (diacylglycerol acyltransferase 1 - *DGAT1*) was detected in our study and TSP merged to immune system, that regulates the movement of lymphocytes to inflammation site in the mammary gland. Several QTL previously reported overlapped with regions detected in our study.

Conclusions

Our results overlap with QTLs previously reported in taurine dairy cattle and suggest novel candidate loci to be further investigated. More information about the mechanisms underlying dairy traits is being added. This study is believed to be the

first genome-wide scan in indicine dairy cattle

Keywords: Genome-wide scan, milk production, *Bos indicus*, Gyr dairy cattle

Background

Recently, several genome-wide scans (GWS) for milk production and reproductive traits in dairy cattle (*Bos primigenius taurus*) have been published (e.g. [1–3]). However, these complex traits express themselves differently denpending on the environmental conditions and the genetic background of the dairy breed, and additional studies including different breeds and subspecies of cattle are required. In particular, loci explaining differences in milk and fertility traits in indicine cattle (*Bos primigenius indicus*) remain undercharacterized.

It is well known that indicine breeds are well adapted to tropical and subtropical environments and low input systems, which largely affect the animal production and reproductive performance. Indicine and taurine breeds differ markedly in their milk production. Indicine breeds have prevailing daily milk yields of < 7.5 kg/day, as compared to the > 30 kg/day in taurine breeds [4]. According to [5], days in milk and nutrition largely impact milk production and composition. Therefore, it is not unexpected that metabolic processes can differentially affect milk production and reproductive performance in indicine and taurine breeds.

Here, we used genome-wide single nucleotide polymorphism (SNP) markers and deregressed estimated breeding values (dEBVs), penalized for their accuracies, to map loci affecting traits under evaluation by Embrapa dairy cattle in Brazil, including milk, fat, protein and total solids yield, fat, protein and total solids percentages, and age at first calving. We focused on the detection of previously reported quantitative trait loci (QTLs) in taurine cattle as well as on prospecting new positional candidate loci.

In face to the inability to genotype many Single Nucleotide Polimorphisms (SNPs) markers in large samples due to prohibitive costs, imputation methods have been used to increase the coverage of the genome, allowing to capture much more of the variation in a trait due to the linkage disequilibrium structure in a region of the

genome to infer the alleles of SNPs not directly genotyped in the study [6]. This strategy enhances the power of dectecting associations between genome regions and phenotypes of interest. So, we adopted this strategy to increase the coverage of the genome and the power of our study (for details see Appendice).

Results

Imputation and quality control

For the Illumina® BovineHD BeadChip (HD) data, a total of 471 bulls and 522,895 markers passed quality control (See Material and Methods for details). Additionaly, 1,616 cows and 26,333 SNPs from the Illumina® Bovine50SNP BeadChip v2 (50k) panel were imputed to HD. For the genome-wide scan, the number of animals, number of SNPs, as well as the descriptive statistics of pseudo-phenotypes (penalized dEBVs) are presented in Table 1. The distribution of the pseudo-phenotypes of all the traits were approximately normal (Figure1).

Genome-wide mapping

Considering the $6\times IQR+Q_3$ threshold for percentage of EBV variance explained by 1 Mb sliding SNP windows, an overview of windows that were declared outliers for each trait is presented in Manhattan plots in Figure 2 and a summary of the results for each trait is presented in Table 2.

Windows surpassing the adopted threshold were found in seven of the eight traits studied here, with protein percentage being the only exception. A total of 47 windows were declared outliers for MY, explaining 0.5% of the variance in EBVs, 37 windows explaining 0.57% of PY, 64 windows explaining 1.52% of FY, 69 windows explaining 1.96% of FP, 18 windows explaining 0.57% of TSY, 12 windows explaining 0.98% of TSP and 31 windows explaining 0.68% of EBV variance in AFC.

Detailed results for the seven traits are provided in the subsections below.

Milk yield

A total of 33 genes were mapped against the regions declared outliers in our study, including one on chromosome 12 and 32 on chromosome 21. These included 28 protein coding, 2 snRNA, 1 snoRNA, 1 miRNA and 1 pseudo gene. From this gene list, ClueGO was not able to build a biological network of functional terms, but two significant terms, exocyst (GO:0000145) and cellular protein metabolic process (GO:0044267), were significantly associated to 13 genes of the list (Figure 3).

A total of 22 QTLs related to milk production and milk composition traits, mined from 2 distinct publications, were found mapping against the segment on chromosome 12 for this trait. Four of them are especifically described as related to milk yield.

Fat yield

A total of 68 genes were mapped against the segments declared outliers for fat yield (Table 2), including 5 on chromosome 2, 10 on chromosome 4, 1 on chromosome 12, 13 on chromosome 16, 27 on chromosome 26 and 11 on chromosome 29. From these, 59 protein coding, 1 processed pseudo gene, 2 snRNA, 2 snoRNA , 1 rRNA, 1 miscRNA and 1 miRNA. From this gene list, ClueGO built two group of terms in a biological network involving 7 genes. The most significant terms representing the groups are NADPH oxidation (GO:0070995) and NADP binding (GO:0050661) (Figure 3).

Forty five QTLs related to milk production and milk composition traits, mined from 11 distinct publications, were also mapped against the segments described in Table 2. Nine of them were especifically described as related to fat yield.

Protein yield

A total of 82 genes were mapped against the segments (Table 2) declared outliers for protein yield, including 51 on chromosome 7 and 31 on chromosome 19. From these, 74 protein coding, 3 snRNA, 1 snoRNA and 4 miRNA. From this gene

list, ClueGO was not able to build a biological network of functional terms.

Sixteen QTLs related to milk production and milk composition traits, mined from 11 distinct publications, were also mapped against the segments on chromosome 19 for this trait. Only one is especifically described as related to protein yield.

Total Solids yield

A total of 26 genes were mapped against the significant segments (Table 2) for total solids yield, including 16 on chromosome 17 and 10 on chromosome 21. From these, 13 protein coding, 2 snRNA, 10 snoRNA and 1 processed pseudo gene. Previously described QTLs for milk traits were not found mapping against these regions.

From this gene list, ClueGO built four groups of functional terms involving 6 genes. The most significant terms representing the networks are single-organism developmental process (GO:0044767), growth factor receptor binding (GO:0070851), regulation of secretion by cell (GO:1903530) and positive regulation of interleukin-17 production (GO:0032740) (Figure 3).

Fat percentage

A total of 175 genes were mapped against the segments (Table 2) for fat percentage, including 38 on chromosome 5, 34 on chromosome 7, 28 on chromosome 12, 59 on chromosome 14 and 15 on chromosome 21. From these, 148 protein coding, 2 snRNA, 6 snoRNA, 7 rRNA, 5 pseudo gene, 2 processed pseudo gene and 5 miRNA.

A total of 130 QTLs related to milk production traits and milk composition, mined from 28 distinct publications, were found mapping against the segments on chromosomes 14 and 21. Twenty two of them were especifically described as related to fat percentage.

ClueGO was able to build a group of terms with the most significant term being intracellular part (GO:00444424). Other 5 functional terms were significant however

not grouped, including connexon complex (GO:0005922), cell part (GO:0044464), intracellular organelle (GO:0043229), cytoplasm (GO:0005737) and protein ADP-ribosylation (GO:0006471) (Figure 3).

Total solids percentage

A total of 43 genes were mapped against the segments (Table 2) declared as outliers for total solids percentage, including 2 on chromosome 4, 23 on chromosome 12 and 18 on chromosome 16. From these, 35 protein coding, 4 snRNA, 1 snoRNA, 1 rRNA, 1 processed pseudo gene and 1 miRNA. Four QTL related to milk production traits and milk composition, mined from 3 distinct publications, were found mapping against the segment, described in Table 2, on chromosome 12 for this trait.

ClueGO built a group of terms being represented by B cell chemotaxis (GO:0035754). Other 3 functional terms not grouped were significant, including negative regulation of smooth muscle cell proliferation (GO:0048662), connexon complex (GO:0005922) and kinesin binding (GO:0019894) (Figure 3).

Age at first calving

A total of 29 genes were mapped against the signifcant segments (Table 2) for age at first calving, including 16 on chromosome 13, 12 on chromosome 20 and 1 on chromosome 29. From these, 26 protein coding, 1 snRNA, 1 snoRNA and 1 pseudo gene.

A total of 9 QTLs related to reproductive traits, mined from 3 distinct publications, were found mapping against the segments on chromosomes 13, 20 and 29 for this trait. None of them is directly related to age at first calving.

ClueGO was able to build 5 groups of functional terms, being represented by nucleus (GO:0005634), nucleic acid metabolic process (GO:0005634), membrane-bounded organelle (GO:0043227), nucleotide binding (GO:0000166) and coenzyme binding (GO:0050662). Many other terms were significants, but not grouped (Figure 3).

Discussion

The objective of our study was to identify chromosomal regions associated to milk, fat, protein and total solids yield, fat, protein and total solids percentage and age at first calving recorded in Gyr, an indicine dairy cattle under genetic evaluation by Embrapa Dairy Cattle – Brazil.

In spite of the low percentage of EBV variance explained by outlier SNP windows in our study (maximum of 1.96% for fat percentage), the findings are consistent with complex and polygenic trait architecture. The low percentage of variance explained by SNP windows could also be related to the existence of rare large effects variants that were not subjected to enough selection to attain sufficient frequency to be detected.

A region of approximately 2.25 Mb length (41.4-43.65 Mb) on chromosome 12 was detected for milk and fat yield, overlapping a QTL region previously reported as associated to milk yield and milk composition traits in Holstein-Friesian [7]. There was not found studies overlapping the region declared outlier in our study on chromosome 21 for milk yield, not even QTL regions. For this reason it was considered as novel candidate locus. Although a cluster of genes were found in that region on chromosome 21 we prospected only 2 functional terms, exocyst (GO:0000145) and cellular protein metabolic process (GO:0044267). These terms did not group in a biological network and they are very general functional terms, but exocyst could be directly related to milk yield.

Exocyst is a term defined by a protein complex peripherally associated with the plasma membrane that determines where vesicles dock and fuse, regulating exocytotic secretion. The exocytotic secretion pathway in alveolar cells of mammary glands is the primary mechanism for protein secretion, which could link with cellular protein metabolic process, as well as for secretion of water, lactose, oligosaccharides, phosphate, calcium and citrate [8]. He & Guo [9] reported that exocyst mutants in a variety of species exhibit blockages of secretion and show intracellular accumulation of secretory vesicles. This mechanism is complex and not well understood yet and more studies are needed to clarify the intrinsic and extrinsic factors that regulate the secretion of milk.

Regarding the cellular protein metabolic process, it is a general term that could be associated to a wide range of pathways regulating milk yield, which preclude any other inference about this term.

The genes found associated to FY were grouped in a biological network giving two groups of functional terms, represented by NADPH oxidation (GO:0070995) and NADP binding (GO:0050661), the most significant terms associated to 7 genes anchored on chromosome 16 and chromosome 29. These terms play an important role in *de novo* synthesis of fatty acids in the mammary glands. NADPH is an essential component of the fatty acid synthesizing system. Requirements for *de novo* synthesis of fatty acids are a carbon source and reducing equivalents in the form of NADPH + H^+ [10].

Several publications reported that changes in milk fat yield and composition is dependent on complex interactions between the composition of basal diet (relation forage/concentrate) and amount and fatty acids profile of lipid supplements (e.g. [11–14]), which relate ruminal fermentation and milk fat synthesis. *B. indicus* dairy cattle breeds are known to produce more milk fat content than *B. taurus* breeds. Several extrinsic factors can induce the regulation of the difference between milk fat content produced by indicine and taurine breeds, as food supply for the cows in tropical and subtropical regions come from low quality forage, with low levels of lipid supplementation, which may be highlighting this metabolic pathway. However, more studies involving gene expression are needed to elucidate this pathway.

We found an association of a segment of 1.65 MB length on chromosome 2 (7.25-8.9 MB) with FY in our study. Despite we have found a QTL (QTLdb_ID:10127) related to FY, the reference linked to this QTL does not mention any genomic region related to FY on chromosome 2. As the two genes we found on this region (*COL3A1* and *GULP1*) could not be straightly related to any metabolic pathway that affect FY we considered this region as a novel candidate locus to be investigated. Regarding the segment of 1.95 MB length on chromosome 4 (54.9-56.85 Mb) associated to FY, no functional information was found, which can also be a novel candidate locus to be investigated. A segment of 1.3 Mb length (21.8-23.1 MB) on chromosome 26 was also detected for fat yield, overlapping a QTL region previously reported as associated with FY in Holstein [15].

Although regions were declared outliers for protein yield in our study, no genomic locations in taurine breeds were found overlapping the ones detected here. The only QTL overlapping with the region outlier on chromosome 19 in our study was reported by [16], where a single locus association showed a clear effect on protein yield, however that association is possibly due to the interactions with background genes. For chromosome 7 we found a clear peak suggesting an association between the segment and PY, but there was not found previous reports matching this segment with any QTL for milk production traits, and the cluster of genes makes it difficult to dissect about the association as ClueGO was not able to build a group of terms associated to the genes.

Regarding TSY, three most significant terms representing 3 out of 4 group of terms in a biological network built using ClueGO in our study are single-organism developmental process (GO:0044767), growth factor receptor binding (GO:0070851) and regulation of secretion by cell (GO:1903530). The GO:0044767 has direct descendant terms related to mammary gland development (GO:0061377, GO:0060592, GO:0060749), GO:0070851 has a descendant term related to fibroblast growth factor receptor binding (GO:0005104) and GO:1903530 has a descendant term related to the regulation of exocytosis (GO:0017157).

All the three terms, and child terms, mentioned before are connected by Fibroblast growth factor 2 (*FGF2*) gene. *FGF2* is expressed in the bovine mammary gland regulating its development and organization [17], and can also be associated to milk yield and composition [18], which impact in total solids yield, however it remains to be investigated by which mechanisms *FGF2* affects TSY.

The other group of terms related to TSY is represented by the positive regulation of interleukin-17 (Il-17) production (GO:0032740) and it takes place in the innate immune system. Il-17 is a cytokine that has the potential to modulate the mammary gland immune response to mastitis-causing pathogens [19]. Mastitis is a inflammatory process of the mammary gland which can reduce milk total solids production, altering the milk composition. Several studies have been published challenging mammary epithelial cells to pathogens for analysing the innate imune system response (e.g. [19–22]), however all the mechanisms involved in this pathway are still unclear and more studies are necessary, since Reis *et al*. state that

Gyr dairy herds have a high prevalence of subclinical mastitis, which negativelly affects total solids in milk [21].

For fat percentage two terms were grouped by ClueGO. The most significant term was intracellular part (GO:0044424), which cover any constituent part of the living contents of a cell, including the matter contained within the plasma membrane which is usually taken to exclude large vacuoles and masses of secretory or indigested material. Although this term is unespecific for the dissection of the mechanisms underlying fat percentage in milk, we found diacylglycerol acyltransferase 1 (*DGAT1*) gene, the most explored gene associated to milk fat percentage, associated to it. As many studies have reported [2, 3, 23–25], SNP markers in linkage disequilibrium to the *DGAT1* K232A mutation has been strongly associated to increased fat percentage and milk and fat yields. Even found in low frequencies in *Bos indicus* breeds [26], SNP markers into windows built to explore loci associated to FP in our study showed a peak on *DGAT1* region, indicating that in Gyr dairy cattle there is enough variation and this information could be used for artificial selection. However, further investigation is require to determine if the same K232A mutation or a different quantitative trait nucleotide (QTN) is responsible for variation in FP in Gyr cattle. Interestingly, if the K232A mutation is the underlying QTN, than the origin of this variant in Gyr could be traced back to reveal whether it was introgressed via matings of Gyr bulls with taurine cows during the early imports of the breed or if this QTN is a much older mutation.

Another interesting term, not grouped, that could be related not only to FP but milk production in general, is connexon complex (GO:0005922), which also appears to be related to TSP and consists in an assembly of six molecules of connexin, gap junction proteins, produced in the Golgi apparatus and subsequently transported to the plasma membrane. The establishment of functional alveoli depends on the polarization of the luminal cells and the formation of junctions between them. During pregnancy, an especific deletion of *CX26* (synonyms for *GJB2* linked to GO:0005922 – Figure 3) lead to a high level of cell death [27] precluding mammary gland cells proliferation.

Regarding total solids percentage, the most significant term was B cell chemotaxis (GO:0035754). B cells chemotaxis is the direct movement of a B cell (or

B lymphocyte) guided by a specific chemical concentration gradient. The regulation of this movement allows the migration of B cells to inflammation sites into an organism. B lymphocytes play an important role in bacterial defense mechanisms of the mammary gland for producing anti-bodies by antigen-activated B lymphocytes [28], as the first inflammatory response of the host to invading bacteria is cell-mediated immunity. As mentioned before, Gyr dairy herds in Brazil present high prevalence of subclinical mastitis [21], meaning that the immune system is frequently being challenged.

For age at first calving five groups of functional terms were built. These terms are unespecific and therefore could be involved in various molecular events and biological process. However, transcripts of the *UBE2D3* gene, shared among 3 out of 5 groups of functional terms, were found in abundance controlling cell cycle during the dominance phase of oocytes maturation of the first follicular wave of Simmental heifers [29].

Conclusions

This study is believed to be the first genome-wide scan in indicine dairy cattle. Our findings point to known QTLs affecting milk production traits and reproductive performance in taurine dairy cattle breeds and also novel genomic locations that could be used to better understand the complexity of the genetic basis of milk production and reproductive performance.

Methods

Animals and phenotypes

Estimated breeding values (EBVs) for eight traits, including 7 production and 1 reproduction trait, were studied. The EBVs were obtained from the Brazilian National Gyr Dairy Breeding Program, which is conducted by the Brazilian Agricultural Research Corporation – EMBRAPA, in partnership with the breeders association – ABCGIL. The dataset included 89,849 animals in the pedigree.

The seven production traits were milk (MY), fat (FY), protein (PY) and total solids (TSY) yields and fat (FP), protein (PP) and total solids (TSP) percentages. Age at first calving (AFC) was the only reproduction trait.

The sampling of the study population was made for having a broad representation of the Gyr breed being evaluated in Brazil. Estimated breeding values were deregressed following [30], penalized by the square root of the reliability (accuracy) of the deregressed breeding value, following [31], and treated as pseudo-phenotypes in the genome-wide scan analysis.

Genotyping, imputation and data filtering

A total of 475 Gyr bulls were genotyped using Illumina BovineHD BeadChip (HD) and 1,678 Gyr cows were genotyped using Illumina BovineSNP50 BeadChip v2 assay (50k), according to the manufacturer's protocol. As well known, 50k has high SNP ascertainment bias in indicine breeds [32] and as consequence few SNPs would remain after genotype filtering. For that reason, we decide to impute the cows using bulls as reference to proceed the genome-wide scan analysis once the bulls are highly related to the cows and influents in the Gyr population under study (For more details see Additional File 1).

The genotype and sample quality checks for the imputation were performed on the HD and 50k using PLINK v1.9 [33]. Only autosome markers were included in further analysis. SNPs were removed from the dataset if call rate < 90%, minor allele frequency (MAF) < 1% or Hardy-Weinberg equilibrium < 10^{-6} was observed. Unmapped markers and SNPs with identical physical position were also removed. Samples were removed if they presented call rate < 90%. To impute the cows we used Beagle v4 [34, 35] with default parameters.

For genome-wide scan analysis we combined bulls and imputed cows datasets and performed a second genotype quality check removing SNPs with MAF < 3% and Hardy-Weinberg equilibrium < 10^{-6}.

Genome-wide scan

Mapping was performed using the Genome-wide Rapid Association using Mixed Model and Regression (GRAMMAR) method [36], as implemented in GenABEL R (v. 1.8) package [37]. In a first step, according to [36], a polygenic model was fitted as follow:

$$y_i = \mu + G_i + e_i \quad \textbf{\{1\}}$$

where y_i is the penalized dEBV and G_i and e_i are random additive polygenic and residual effect. The random effects are assumed to follow a multivariate normal distribution with mean zero. The variance for the poligenic effects is defined as $\Phi\sigma^2_G$, where Φ is the kinship matrix and σ^2_G is the additive genetic variance due to polygenes. For the residual random effects the variance is defined as $I\sigma^2_e$ where I is an identity matrix and σ^2_e is the residual variance. In the second step the residual estimated from **{1}** (y^*_i) was used as dependent variable in a single marker linear regression:

$$y^*_i = \mu + \beta_g g_i + e_i \quad \textbf{\{2\}}$$

where y^* is the residual obteined from **{1}**, μ is the overall mean, β_g is the marker regression coefficient, g_i is the marker genotype of the animal i and e_i is the random residual. For each SNP, β_g and its respective standard error (SE_g) were estimated using ordinary least squares.

Next, aiming at mapping loci explaining differences in the traits included in this study, we investigated chromosome windows where the average phenotypic variance explained by SNPs deviated substantially from the genome background, as proposed by [38]. The percentage of phenotypic variance ($\%\hat{\pi}^2_i$) explained by each marker was calculated as:

$$\%\hat{\pi}^2_i = \frac{2p_iq_i\hat{\beta}^2_i}{\sigma^2_T} \quad \textbf{\{3\}}$$

where, relative to SNP i, $\hat{\beta}_i^2$ is the estimated allele substitution effect, p_i and q_i are allele frequencies and σ_T^2 is the total trait variance. The phenotypic variance explained by SNPs was smoothed across the genome by averaging $\%\hat{\pi}_i^2$ in sliding windows of 1 Mb, sliding 50 kb at a time. The average was used instead of the sum because they were tested as independent fixed effects. Only windows containing at least 10 SNPs were averaged, and we consider as outliers (significants) only windows with $\%\hat{\pi}_i^2 > 6\times IQR+Q_3$, where IQR is the interquartile range and Q_3 is the third quartile of the distribution.

Assessment of functional relevance

We examined the *cattle QTLdb* database [39] to find out if any genomic region identified in this study mapped against any previously described bovine quantitative trait loci (QTL). Gene coordinates anchored in the UMD v3.1 bovine genome assembly [40] were interrogated using *BioMart tool* and the *Ensembl genes 78* database [41]. Finally, we used the ClueGO Cytoscape plug-in [42] to analyze interrelations of terms and functional groups in biological networks, which uses prior biological knowledge on gene function annotations in *Bos taurus* using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and Gene Onthology (GO) database. A term was declared significant if the p-value of the right-sided hypergeometric test, corrected for Benjamini-Hochberg, was < 0.05. The terms were grouped if kappa score was > 0.7. If no functional terms were found taking into account *B. taurus* databases we went straightfoward *Homo sapiens* databases.

References

1. Kolbehdari D, Wang Z, Grant JR, Murdoch B, Prasad A, Xiu Z, Marques E, Stothard P, Moore SS: **A whole genome scan to map QTL for milk production traits and somatic cell score in Canadian Holstein bulls.** *J Anim Breed Genet* 2009, **126**:216–27.

2. Cole JB, Wiggans GR, Ma L, Sonstegard TS, Lawlor TJ, Crooker BA, Van Tassell CP, Yang J, Wang S, Matukumalli LK, Da Y: **Genome-wide association analysis of thirty one production, health, reproduction and body conformation traits in**

contemporary U.S. Holstein cows. *BMC Genomics* 2011, **12**:408.

3. Minozzi G, Nicolazzi EL, Stella A, Biffani S, Negrini R, Lazzari B, Ajmone-Marsan P, Williams JL: **Genome wide analysis of fertility and production traits in Italian Holstein cattle.** *PLoS One* 2013, **8**:e80219.

4. Collier RJ, Collier JL (Eds): *Environmental Physiology of Livestock*. John Wiley & Sons; 2012:368.

5. Lemay DG, Lynn DJ, Martin WF, Neville MC, Casey TM, Rincon G, Kriventseva E V, Barris WC, Hinrichs AS, Molenaar AJ, Pollard KS, Maqbool NJ, Singh K, Murney R, Zdobnov EM, Tellam RL, Medrano JF, German JB, Rijnkels M: **The bovine lactation genome: insights into the evolution of mammalian milk.** *Genome Biol* 2009, **10**:R43.

6. Halperin E, Stephan D a: **SNP imputation in association studies.** *Nat Biotechnol* 2009, **27**:349–51.

7. Meredith BK, Kearney FJ, Finlay EK, Bradley DG, Fahey AG, Berry DP, Lynn DJ: **Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland.** *BMC Genet* 2012, **13**:21.

8. McManaman JL, Neville MC: **Mammary physiology and milk secretion**. *Adv Drug Deliv Rev* 2003, **55**:629–641.

9. He B, Guo W: **The exocyst complex in polarized exocytosis.** *Curr Opin Cell Biol* 2009, **21**:537–42.

10. Fox PF, McSweeney PLH (Eds): *Advanced Dairy Chemistry Volume 2: Lipids*. 3rd edition. Springer Science & Business Media; 2007:801.

11. Chilliard Y, Toral PG, Shingfield KJ, Rouel J, Leroux C, Bernard L: **Effects of diet and physiological factors on milk fat synthesis, milk fat composition and lipolysis in the goat: A short review**. *Small Rumin Res* 2014, **122**:31–37.

12. Freitas Júnior JE de, Rennó FP, Gandra JR, Rennó LN, Rodrigues GH, Santos MV dos, Oliveira MDS de: **Nutrients balances and milk fatty acid profile of mid lactation dairy cows supplemented with unsaturated fatty acid**. *Rev Bras Saúde e Produção Anim* 2013, **14**:322–335.

13. Harvatine KJ, Boisclair YR, Bauman DE: **Recent advances in the regulation of milk fat synthesis.** *Animal* 2009, **3**:40–54.

14. Glasser F, Ferlay A, Chilliard Y: **Oilseed lipid supplements and fatty acid composition of cow milk: a meta-analysis.** *J Dairy Sci* 2008, **91**:4687–703.

15. Jiang Z, De S, Garcia MD, Griffin KB, Wu XL, Xiao Q, Michal JJ, Sharma BS, Jansen GB: **An independent confirmation of a quantitative trait locus for milk yield and composition traits on bovine chromosome 26.** *J Anim Breed Genet* 2005, **122**:281–4.

16. Alim MA, Wang P, Wu XP, Li C, Cui XG, Zhang SL, Zhang Q, Zhang Y, Sun DX: **Effect of FASN gene on milk yield and milk composition in the Chinese Holstein dairy population.** *Anim Genet* 2014, **45**:111–3.

17. Plath A, Einspanier R, Gabler C, Peters F, Sinowatz F, Gospodarowicz D, Schams D: **Expression and localization of members of the fibroblast growth**

factor family in the bovine mammary gland. *J Dairy Sci* 1998, **81**:2604–13.

18. Wang X, Maltecca C, Tal-Stein R, Lipkin E, Khatib H: **Association of bovine fibroblast growth factor 2 (FGF2) gene with milk fat and productive life: an example of the ability of the candidate pathway strategy to identify quantitative trait genes.** *J Dairy Sci* 2008, **91**:2475–80.

19. Bougarn S, Cunha P, Gilbert FB, Harmache A, Foucras G, Rainard P: **Staphylococcal-associated molecular patterns enhance expression of immune defense genes induced by IL-17 in mammary epithelial cells.** *Cytokine* 2011, **56**:749–59.

20. Ogorevc J, Kunej T, Razpet A, Dovc P: **Database of cattle candidate genes and genetic markers for milk production and mastitis.** *Anim Genet* 2009, **40**:832–51.

21. Malek dos Reis CB, Barreiro JR, Mestieri L, Porcionato MA de F, dos Santos MV: **Effect of somatic cell count and mastitis pathogens on milk composition in Gyr cows.** *BMC Vet Res* 2013, **9**:67.

22. Porcionato MA de F, Soares WVB, Reis CBM dos, Cortinhas CS, Mestieri L, Santos MV dos: **Milk flow, teat morphology and subclinical mastitis prevalence in Gir cows**. *Pesqui Agropecuária Bras* 2010, **45**:1507–1512.

23. Grisart B, Coppieters W, Farnir F, Karim L, Ford C, Berzi P, Cambisano N, Mni M, Reid S, Simon P, Spelman R, Georges M, Snell R: **Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition.** *Genome Res* 2002, **12**:222–31.

24. Winter A, Krämer W, Werner FAO, Kollers S, Kata S, Durstewitz G, Buitkamp J, Womack JE, Thaller G, Fries R: **Association of a lysine-232/alanine polymorphism in a bovine gene encoding acyl-CoA:diacylglycerol acyltransferase (DGAT1) with variation at a quantitative trait locus for milk fat content.** *Proc Natl Acad Sci U S A* 2002, **99**:9300–5.

25. Da Silva MVGB, Sonstegard TS, Thallman RM, Connor EE, Schnabel RD, Van Tassell CP: **Characterization of DGAT1 allelic effects in a sample of North American Holstein cattle.** *Anim Biotechnol* 2010, **21**:88–99.

26. Rosse IDC, Steinberg RDS, Coimbra RS, Peixoto MGCD, Verneque RS, Machado MA, Fonseca CG, Carvalho MRS: **Novel SNPs and INDEL polymorphisms in the 3'UTR of DGAT1 gene: in silico analyses and a possible association.** *Mol Biol Rep* 2014, **41**:4555–63.

27. Hennighausen L, Robinson GW: **Information networks in the mammary gland.** *Nat Rev Mol Cell Biol* 2005, **6**:715–25.

28. Sordillo LM, Streicher KL: **Mammary gland immunity and mastitis susceptibility.** *J Mammary Gland Biol Neoplasia* 2002, **7**:135–46.

29. Ghanem N, Hölker M, Rings F, Jennen D, Tholen E, Sirard M-A, Torner H, Kanitz W, Schellander K, Tesfaye D: **Alterations in transcript abundance of bovine oocytes recovered at growth and dominance phases of the first follicular wave.** *BMC Dev Biol* 2007, **7**:90.

30. Garrick DJ, Taylor JF, Fernando RL: **Deregressing estimated breeding values**

and weighting information for genomic regression analyses. *Genet Sel Evol* 2009, **41**:55.

31. Rodriguez-Zas SL, Southey BR, Heyen DW, Lewin HA: **Interval and composite interval mapping of somatic cell score, yield, and components of milk in dairy cattle.** *J Dairy Sci* 2002, **85**:3081–91.

32. **BovineSNP50 Genotyping BeadChip** [http://www.illumina.com/Documents/products/datasheets/datasheet_bovine_snp5O.pdf]

33. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ: **Second-generation PLINK: rising to the challenge of larger and richer datasets**. 2014.

34. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**:210–23.

35. Browning BL, Browning SR: **Improving the accuracy and efficiency of identity-by-descent detection in population data.** *Genetics* 2013, **194**:459–71.

36. Aulchenko YS, de Koning D-J, Haley C: **Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis.** *Genetics* 2007, **177**:577–85.

37. Aulchenko YS, Ripke S, Isaacs A, van Duijn CM: **GenABEL: an R library for genome-wide association analysis.** *Bioinformatics* 2007, **23**:1294–6.

38. Utsunomiya YT, Carmo AS, Neves HHR, Carvalheiro R, Matos MC, Zavarez LB, Ito PKRK, Pérez O'Brien AM, Sölkner J, Porto-Neto LR, Schenkel FS, McEwan J, Cole JB, da Silva MVGB, Van Tassell CP, Sonstegard TS, Garcia JF: **Genome-Wide Mapping of Loci Explaining Variance in Scrotal Circumference in Nellore Cattle**. *PLoS One* 2014, **9**:e88561.

39. Hu Z-L, Park CA, Wu X-L, Reecy JM: **Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era.** *Nucleic Acids Res* 2013, **41**(Database issue):D871–9.

40. Zimin A V, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS, Marçais G, Roberts M, Subramanian P, Yorke JA, Salzberg SL: **A whole-genome assembly of the domestic cow, Bos taurus.** *Genome Biol* 2009, **10**:R42.

41. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P: **Ensembl BioMarts: a hub for data retrieval across taxonomic space.** *Database (Oxford)* 2011, **2011**:bar030.

42. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman W-H, Pagès F, Trajanoski Z, Galon J: **ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.** *Bioinformatics* 2009, **25**:1091–3.

Figures

Figure 1 – Histograms of the distribution of pseudo-phenotypes. Pseudo-phenotypes were based on dEBVs penalized for the square root of the reliability of the dEBV for milk (MY), fat (FY), protein (PY) and total solids (TSY) yields, fat (FP), protein (PP) and total solids (TSP) percentages and age at first calving (AFC).

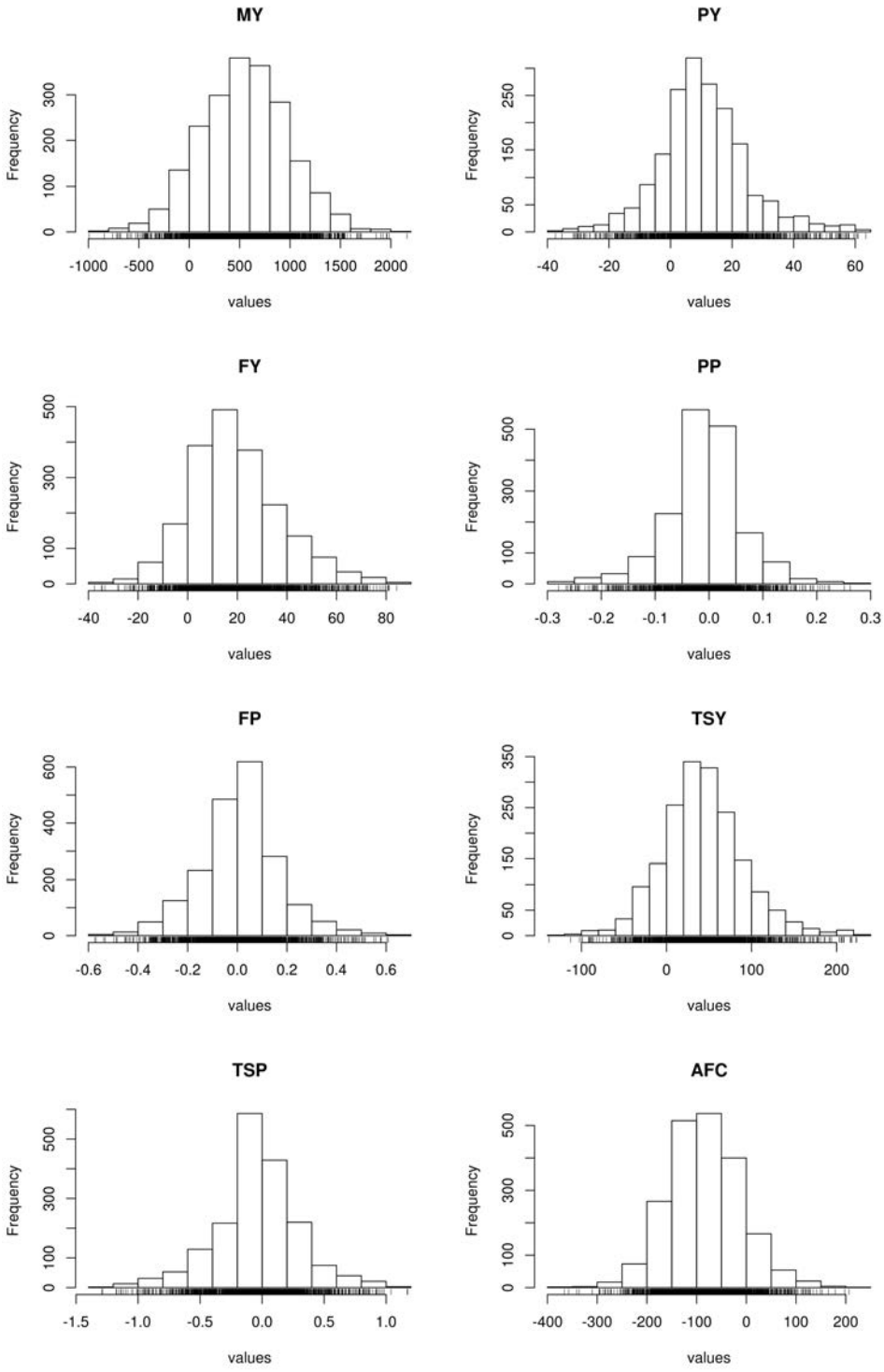
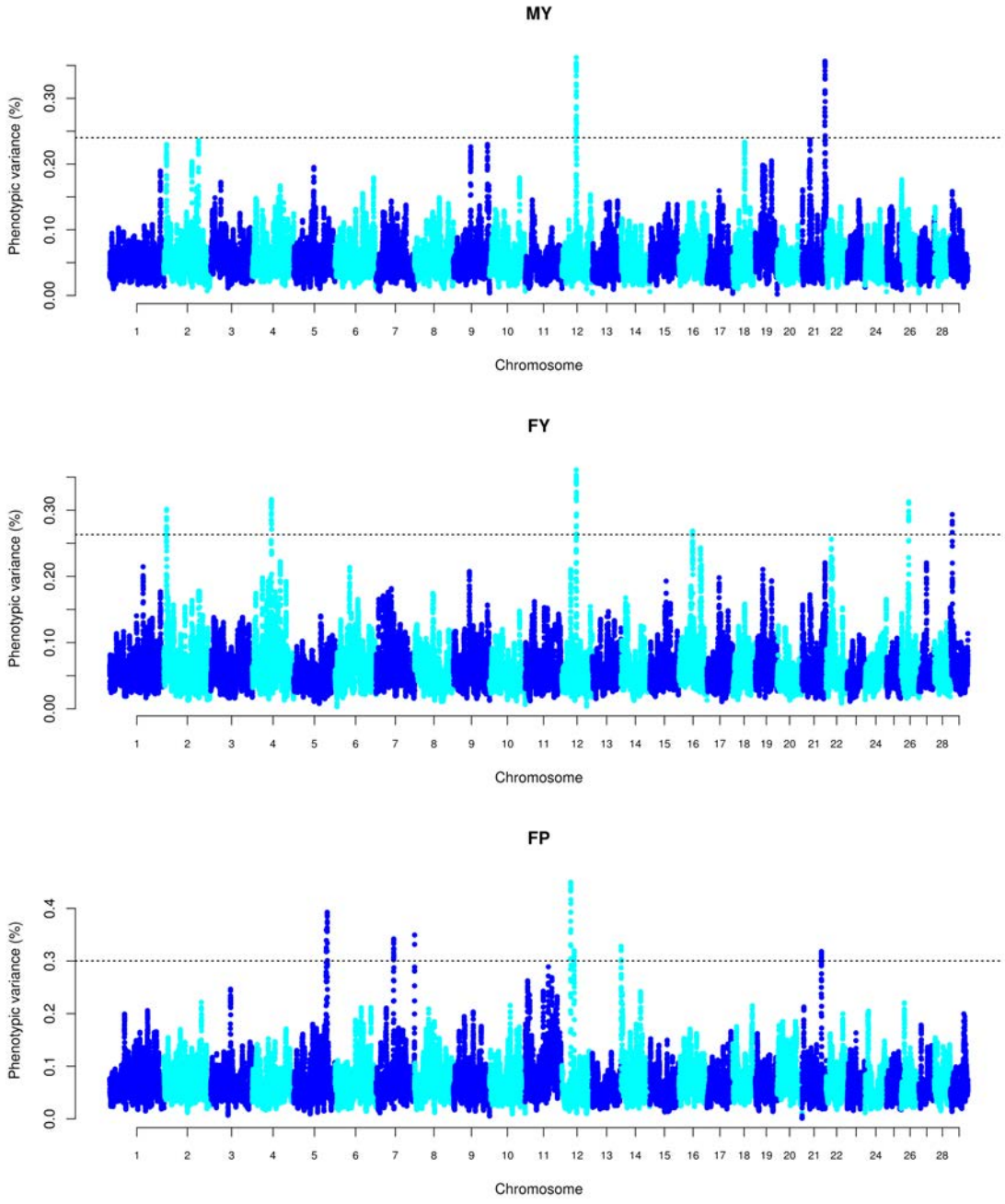
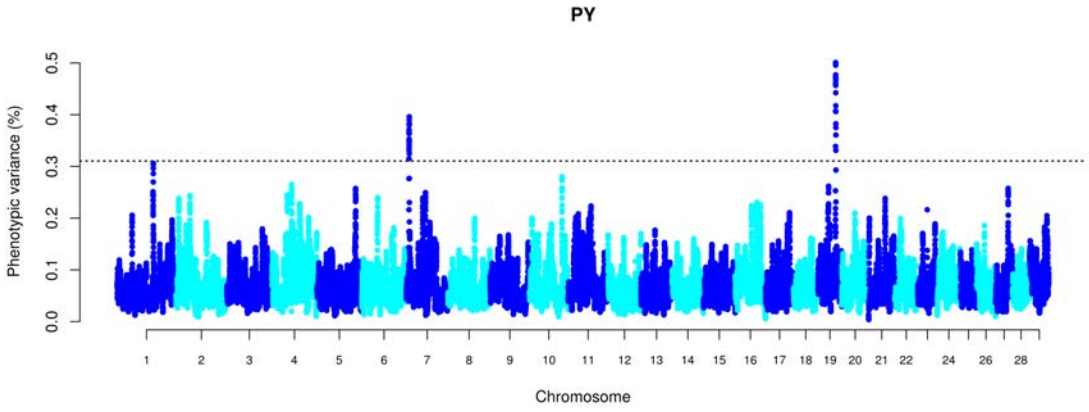
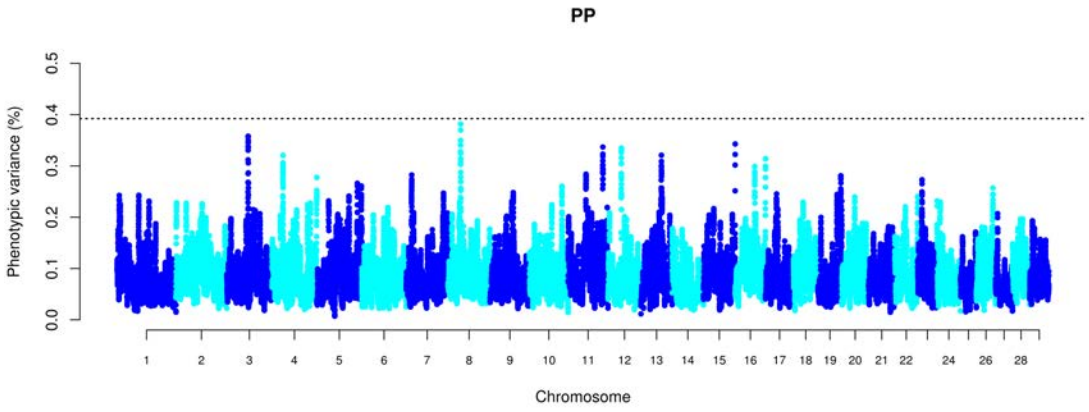
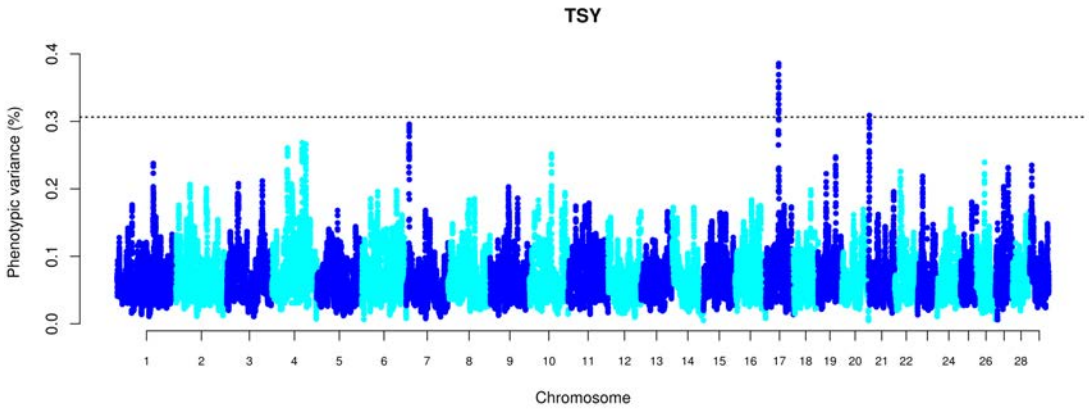


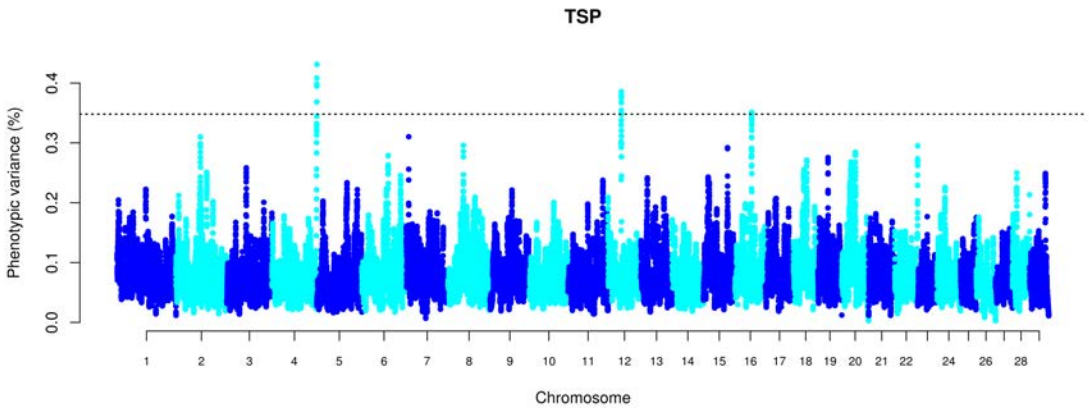
Figure 2 – Manhattan plots of phenotypic variance explained by SNP windows in Gyr dairy cattle. Pseudo-phenotypes were based on dEBVs penalized for the square root of the reliability of the dEBV for milk (MY), fat (FY), protein (PY) and total solids (TSY) yields, fat (FP), protein (PP) and total solids (TSP) percentages and age at first calving (AFC). Each dot respresents a 1 Mb SNP window. Horizontal dashed lines represent the adopted thresholds (MY = 0.24%, FY = 0.26%, PY = 0.31%, TSY = 0.31%, FP = 0.30%, PP = 0.30%, TSP = 0.35% and AFC = 0.26%)




PY
Phenotypic variance (%)
0.0
0.1
0.2
0.3
0.4
0.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 28
Chromosome


PP
Phenotypic variance (%)
0.0
0.1
0.2
0.3
0.4
0.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 28
Chromosome


TSY
Phenotypic variance (%)
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 28
Chromosome


TSP
Phenotypic variance (%)
0.0
0.1
0.2
0.3
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 24 26 28
Chromosome

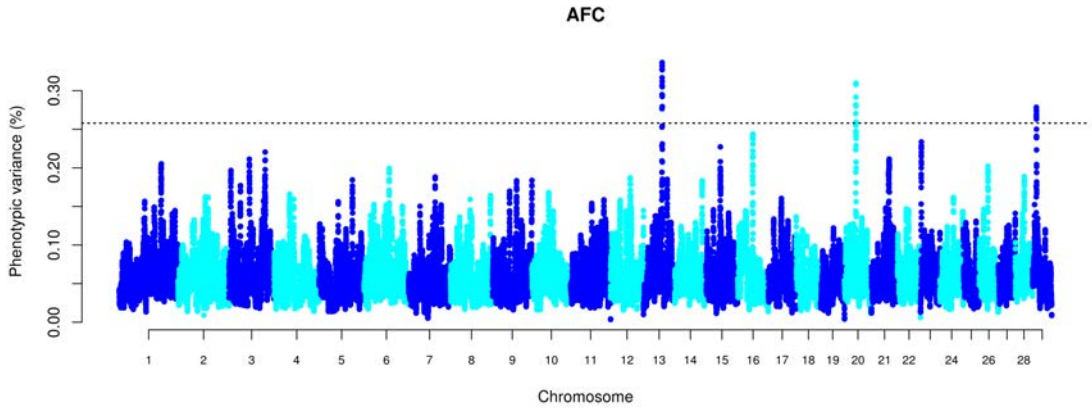
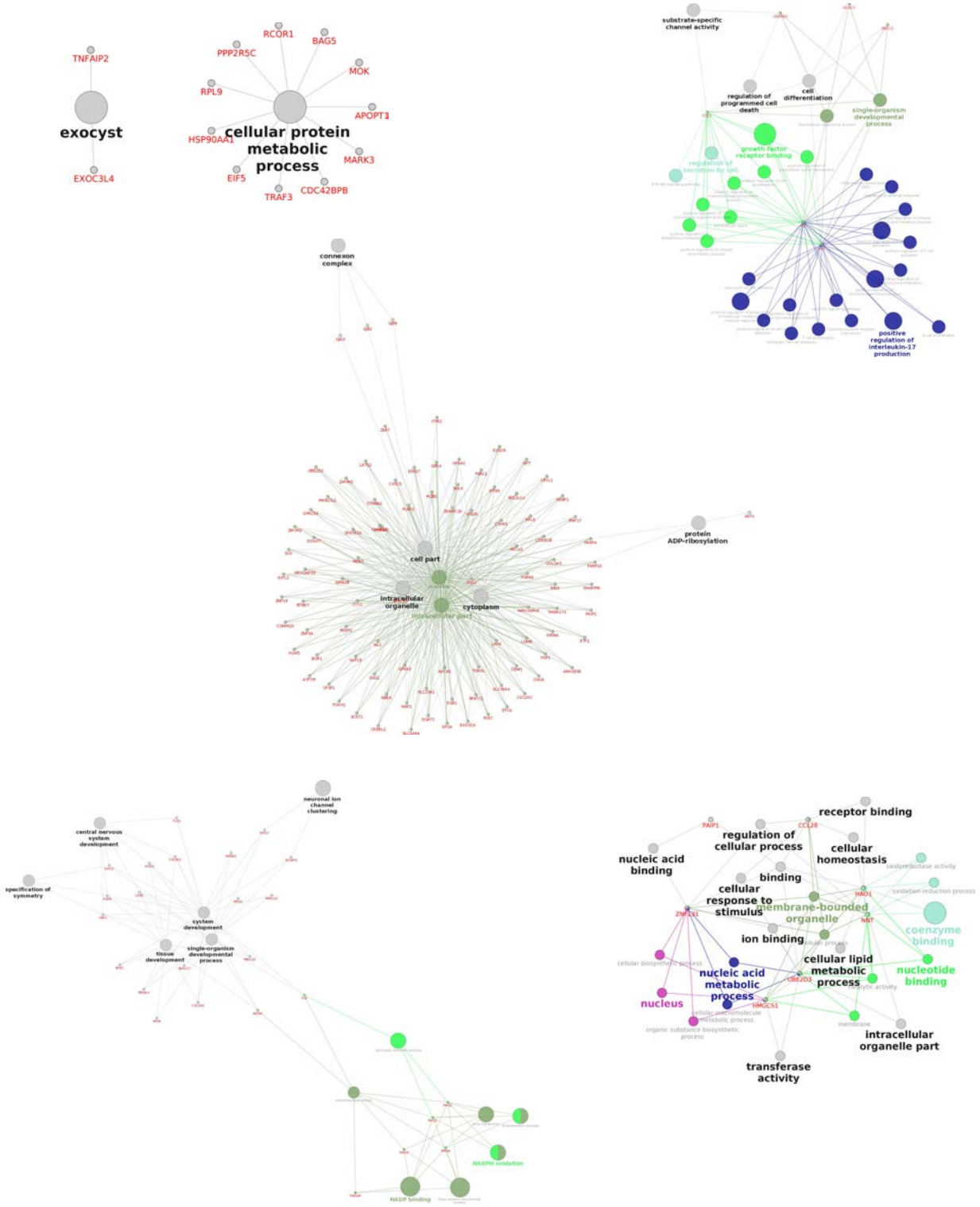

AFC
Phenotypic variance (%)
0.00
0.10
0.20
0.30
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
24
26
28
Chromosome

Figure 3 – Biological networks of functional terms for MY (Top left), FY (Bottom left), TSY (Top right), FP (center), and AFC (Bottom right) in Gyr dairy cattle



Tables

Table 1 – Description of the number of animals, number of SNPs and descriptive statistics for each trait after quality checks for genome-wide scan

Trait	# of Animals	# of SNPs	Mean (sd)	Median	Min	Max
MY	2,069	441,196	549.3	555.3	-993.2	2160.2
FY	1,995	441,478	19.232	17.203	-37.550	84.267
PY	1,799	441,457	10.735	9.702	-37.389	63.478
TSY	1,803	441,478	42.93	40.58	-138.09	223.59
FP	2,004	441,478	0.009152	0.012093	-0.5749	0.6085
PP	1710	441,457	-0.008824	-0.004695	-0.2794	0.2623
TSP	1,819	441,478	-0.4212	-0.03650	-1.29219	1.17289
AFC	2,058	441,196	-83.75	-86.63	-365.96	206.68

Table 2 – Detected major loci explaining variance for each trait in Gyr dairy cattle.

Trait (thr%[a])	Chr	Position start (Mb)	Position end (Mb)	Peak Position (Mb)	Segment length (Mb)	Number of SNPs	Average MAF[b]	Average $\%\hat{\pi}$ [c]
MY (0.24)	12	41.4	43.65	42.45	2.25	347	0.19	0.25
	21	67.7	69.9	68.7	2.2	331	0.15	0.25
FY (0.26)	2	7.25	8.9	8.2	1.65	303	0.24	0.24
	4	54.9	56.85	55.9	1.95	225	0.18	0.25
	12	41.7	43.55	42.45	1.85	294	0.18	0.26
	16	39.35	40.4	39.9	1.05	148	0.21	0.25
	26	21.8	23.1	22.55	1.3	131	0.11	0.26
	29	5.45	6.6	6	1.15	122	0.19	0.26
PY (0.31)	7	7.4	9.25	8.6	1.85	283	0.24	0.27
	19	47.1	49	47.85	1.9	312	0.18	0.3
FP (0.3)	5	95	98.9	98.15	3.9	689	0.22	0.23
	7	51.25	52.9	52.15	1.65	110	0.12	0.28
	7	112.45	112.6254	112.5627	0.175418	19	0.24	0.32
	12	25.3	27.15	26.15	1.85	215	0.16	0.3
	12	35.65	36.9	36.3	1.25	95	0.19	0.27
	14	0.9	2.1	1.4	1.2	114	0.28	0.3
	21	57.65	58.95	58.4	1.3	234	0.25	0.26

[a] thr% = threshold in percentage used for considering a window as significant for the trait
[b] MAF = Minor Allele Frequency
[c] $\%\hat{\pi}$ = Average percentage of phenotypic variance explained by the segment built from overlapping 1 Mb SNP windows.

Table 2 - Continuation

Trait (thr%[a])	Chr	Position start (Mb)	Position end (Mb)	Peak Position (Mb)	Segment length (Mb)	Number of SNPs	Average MAF[b]	Average $\%\hat{\pi}$ [c]
PP (0.39)	-	-	-	-	-	-	-	-
TSY (0.31)	17	34.9	36.5	35.8	1.6	132	0.19	0.27
	21	3.1	4.1	3.6	1	75	0.24	0.3
TSP (0.35)	4	120.35	120.6359	120.593	0.29	25	0.1	0.34
	12	35.65	36.9	36.3	1.25	95	0.2	0.31
	16	44.2	45.25	44.75	1.05	152	0.13	0.33
AFC (0.26)	13	49.65	51.35	50.4	1.7	195	0.25	0.24
	20	30.2	31.65	30.9	1.45	216	0.28	0.22
	29	10.05	11.35	10.6	1.3	159	0.23	0.22

[a] thr% = threshold in percentage used for considering a window as significant for the trait
[b] MAF = Minor Allele Frequency
[c] $\%\hat{\pi}$ = Average percentage of phenotypic variance explained by the segment built from overlapping 1 Mb SNP windows.

Appendice

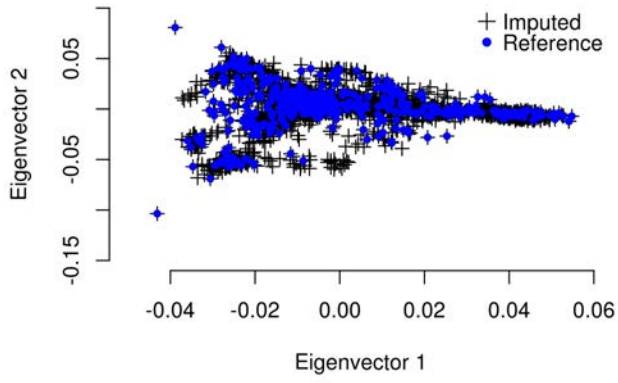
Supplementary material 1 - Imputation

As the ascertainment bias of the 50k is high for indicine breeds we decide to use an imputed dataset to perform genome wide mapping in Gyr increasing the coverage of the genome. As we could not test the accuracy of imputation because there were not cows genotyped on HD we used the relatedness between bulls and cows to support our analysis. It is known that the imputation accuracy tends to increase as the relatedness between imputed and reference animals increases [1].

Studying the accuracy of genotype imputation in Nelore cattle, [2] reported accuracies of imputation greater than 0.95 when the average top 10 genomic relationships was greater than 0.23 using software Beagle [3, 4].

In a first step we show the representativity of the bulls in Gyr population by performing a principal component analysis (Figure S1)

Second, we extracted the SNPs of the 50k after quality checks for imputation from the HD (after quality checks as well) and calculate the genomic relationships according to [5]. Then, based on the computed genomic relationships, each individual's average top 5 relatedness to the animals in the reference set was computed.



As we can see in Figure S1, through principal component analysis we can

affirm that bulls used as reference set for imputation are well distributed within the cows (used as imputation set).

The average top 5 relatedness between reference and imputed sets we found approximately 0.28. This value says that we can expect imputation accuracies greater than 0.95.

References

1. Ma P, Brøndum RF, Zhang Q, Lund MS, Su G: **Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle.** *J Dairy Sci* 2013, **96**:4666–77.

2. Carvalheiro R, Boison SA, Neves HHR, Sargolzaei M, Schenkel FS, Utsunomiya YT, O'Brien A, Sölkner J, McEwan JC, Van Tassell CP, Sonstegard TS, Garcia J: **Accuracy of genotype imputation in Nelore cattle**. *Genet Sel Evol* 2014, **46**:69.

3. Browning BL, Browning SR: **A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals.** *Am J Hum Genet* 2009, **84**:210–23.

4. Browning BL, Browning SR: **Improving the accuracy and efficiency of identity-by-descent detection in population data.** *Genetics* 2013, **194**:459–71.

5. VanRaden PM: **Efficient methods to compute genomic predictions.** *J Dairy Sci* 2008, **91**:4414–23.