3

Estudo de associação genômica ampla utilizando Random Forest: estudo de caso em bovinos de corte

Roberto H. Higa
Fabiana B. Mokry
Maurício de A. Mudadu
Francisco P. Lobo
Luciana C. A. Regitano

A APLICAÇÃO de técnicas de aprendizado de máquina, que consideram o efeito de múltiplos SNPs a problemas de estudos de associação em genônica ampla (GWAS) é de grande interesse, pois essas técnicas são potencialmente capazes de identificar variantes onde o modelo causal é desconhecido e de lidar com o problema de alta dimensionalidade dos dados. Nesse cenário, Random Forest (RF) desponta como uma das técnicas mais interessantes, devido à sua simplicidade, flexibilidade, escalabilidade e capacidade de lidar com um grande número de variáveis de entrada sem incorrer em sobre-ajuste. Embora, RF já seja utilizada em GWAS em humanos, na área de ciência animal, sua utilização ainda é muito tímida. O objetivo deste trabalho é demonstrar o potencial de aplicação de

RF em GWAS na área de ciência animal, apresentando estudos de casos que utilizam dados simulados e reais.

3.1 Introdução

O sequenciamento do genoma bovino, com o subsequente mapeamento de haplótipos e descoberta de milhões de marcadores SNP (*single nucleotide polymorphisms*), distribuídos ao longo do genoma (BOVINE GENOME SEQUENCING; ANALYSIS CONSORTIUM, 2009; BOVINE HAPMAP CONSORTIUM, 2009), viabilizou a construção de painéis de genotipagem em escala genômica, contendo centenas de milhares de SNPs, e a sua utilização em estudos visando desvendar as bases genéticas de características de interesse econômico em bovinos. Esses estudos, denominados estudos de associação genômica ampla (GWAS), consistem em analisar genótipos e fenótipos de uma amostra de poucos milhares de indivíduos ou menos, procurando identificar regiões do genoma onde ocorrem variações, em uma população, associadas ao fenótipo estudado.

Em função da complexidade dos mecanismos moleculares que controlam a manifestação de fenótipos, há grande interesse em abordar GWAS por meio de técnicas multivariadas, que considerem o efeito de múltiplos SNPs. Nesse sentido, é crescente a utilização de técnicas de aprendizado de máquina e mineração de dados, principalmente em estudos envolvendo doenças em humanos (BUREAU et al., 2005; ZIEGLER; DESTEFANO; KÖNIG, 2007; SCHWARZ et al., 2007; SZYMCZAK et al., 2009; MOORE; ASSELBERGS; WILLIAMS, 2010; GOLDSTEIN et al., 2010).

Como essas técnicas não dependem do pressuposto de que o mecanismo genético subjacente assume um determinado modelo, elas são potencialmente mais apropriadas para identificar variantes, onde o mecanismo causal é desconhecido e envolve

diversos SNPs. Além disso, muitas dessas técnicas foram desenvolvidas para lidar com o problema de alta dimensionalidade dos dados, quando o número de variáveis p é muito maior que o número de observações n, $p >> n$, que é o caso de GWAS.

Random Forest (RF), uma técnica de aprendizado de máquina que constrói preditores a partir de conjuntos de árvores de classificação e regressão (CART) (BREIMAN, 2001b), figura entre as mais poderosas técnicas de aprendizado de máquina disponíveis, com estudos empíricos mostrando desempenho igual ou superior a técnicas populares como Boosting (FREUND; SHAPIRE, 1996) e máquinas de vetores de suporte (SHAWE-TAYLOR; CRISTIANINI, 2004).

Dentre suas características mais atraentes estão a sua simplicidade, seu potencial de paralelização e a capacidade de lidar com um grande número de variáveis de entrada sem incorrer em sobre-ajuste.

RF foi introduzida em GWAS em humanos por Bureau et al. (2005) e, desde então, apresenta uma lenta, mas crescente tendência de adoção e utilização (GOLDSTEIN et al., 2010). Em estudos aplicados a ciência animal, sua utilização, embora ainda pequena, também encontra-se em crescimento tanto em GWAS (MOKRY et al., 2013; YAO et al., 2011) quanto em seleção genômica (GWS) (GONZÁLEZ-RECIO; FORNI, 2011).

O objetivo deste trabalho é demonstrar o potencial de utilização de RF em GWAS na área de ciência animal. Para isso, são apresentados os resultados de dois estudos, o primeiro deles envolvendo conjuntos de dados simulados públicos, que se assemelham a populações típicas de melhoramento animal (ELSEN et al., 2012; QTL-MAS, 2012); e o segundo envolvendo um conjunto de dados reais obtidos de uma população de melhoramento de gado de corte (MOKRY et al., 2013). A apresentação desses estudos

é precedida pela exposição dos conceitos básicos envolvidos na construção de RF.

3.2 Random Forest

RF integra um conjunto de métodos de aprendizado de máquina que envolve a construção de muitos preditores (classificadores ou regressores) e cuja predição consiste na agregação das predições de todos os preditores do conjunto. O método foi inicialmente proposto por Breiman (2001b) como uma extensão de seus trabalhos anteriores com árvores CART (BREIMAN et al., 1984) e *bootstrap and aggregating* - *bagging* (BREIMAN, 2001a), também influenciado por trabalhos como de Ho (1998) e de Amit e Geman (1997), que construíram conjuntos de árvores aleatórias para problemas de classificação.

O preditor base utilizado para construção de RF é a árvore CART (BREIMAN et al., 1984). Considerando uma variável resposta Y e um conjunto de variáveis preditoras, $X_1, X_2, \ldots, X_p$, uma árvore CART é construída particionando-se, sucessivamente, o espaço de atributos, utilizando a cada passo uma única variável preditora, X_i para gerar dois sub-espaços. A raíz da árvore representa todo o espaço de atributos e cada particionamento corresponde a um nó interno da árvore. Ao particionar um nó interno da árvore, dois novos nós, e os respectivos sub-espaços associados, são criados. Ao final do processo, obtém-se uma árvore binária, que é utilizada para fazer predições por meio de um processo de busca: considerando um novo dado, $(X_1, X_2, \ldots, X_p)$, a cada nó, é realizado um teste (usando X_i) tal que, dependendo do resultado do teste, a busca prossegue pelo ramo direito ou esquerdo, até que se encontre um nó folha, onde se realiza a predição de Y. O valor da variável predita, Y, baseia-se nos valores de Y das amostras do conjunto de treinamento associadas ao nó folha. Para problemas

de regressão, a predição será igual à média dos valores de Y e, no caso de problemas de classificação, ao valor de Y mais frequente.

Bagging (BREIMAN, 2001a) é uma técnica para construção de conjuntos de preditores, construídos sucessivamente de forma independente, utilizando uma amostra *bootstrap* do conjunto de dados de treinamento. Quando comparado com o preditor base, pode-se mostrar que preditores *bagging* apresentam menor erro de predição por meio da redução do componente de variância do erro. Mas, na prática, essa redução no erro de variância é limitada pela correlação entre os preditores. Por isso, RF introduz em *bagging* um elemento adicional de aleatoriedade, visando obter um conjunto de preditores menos correlacionados: durante a construção de uma árvore CART (preditor), ao invés de analisar todas as variáveis p para determinar o melhor particionamento de um nó da árvore, um número menor de variáveis, $m < p$, selecionados de forma aleatória, é examinado.

Duas das vantagens de RF são a sua simplicidade e o número reduzido de parâmetros importantes – *ntree*, o número de árvores na floresta, e *mtry*, o número de variáveis utilizadas para particionar os nós das árvores –, além de ser robusta às variações destes valores (LIAW; WIENER, 2012). Formalmente, o algoritmo para construção de RF pode ser descrito pelo Algoritmo 1 (BREIMAN, 2001b):

Algoritmo 1: Random Forest (ntree, mtry)

```
1  para b ← 1 ... ntree faça
2      selecionar uma amostra bootstrap;
3      repita
4          selecionar aleatoriamente mtry variáveis;
5          encontrar o resultado das melhores partições;
6      até formar a árvore;
7      predizer Y para OOB;
8      predizer Y para X permutado;
9  fim para
10 calcular o erro OOB;
11 calcular a importância das variáveis;
```

Outra vantagem de RF, comparado com outros métodos de aprendizado de máquina, é que, por construção, ele possui mecanismos para estimar tanto o erro de predição quanto a importância das variáveis analisadas (Figura 3.1). RF constrói essas estimativas avaliando as amostras do conjunto de dados de treinamento que não são incluídas no conjunto de amostras *bootstrap*, dados *out-of-bag* (OOB), que em média representa 36% das amostras de treinamento.

Cada árvore construída é utilizada por RF para predizer os valores de Y para os correspondentes dados OOB; então, após finalizar a construção da floresta, essas predições são comparadas com os valores verdadeiros para obter uma estimativa de erro, denominada erro OOB. De forma similar, os dados OOB também são utilizados para identificar as variáveis importantes calculando os erros quando se permuta cada uma das variáveis utilizadas na construção de cada árvore e comparando-os com o erro OOB. A importância de uma variável é medida pelo impacto que a retirada da informação que ela traz (permutação) causa no erro de predição OOB.

Finalmente, diferentes implementações de RF estão disponíveis livremente (BREIMAN; CUTLER, 2013; LIAW; WIENER, 2012; SCIKIT, 2013; FASTRF, 2013; PARF, 2013; ZHANG; WANG; CHEN, 2009; SCHWARZ; KÖNIG; ZIEGLER, 2010), apresentando características bastante diversas. Algumas estão implementadas em Fortran (BREIMAN; CUTLER, 2013; LIAW; WIENER, 2012; PARF, 2013), enquanto outras em C/C++ (SCIKIT, 2013; ZHANG; WANG; CHEN, 2009; SCHWARZ; KÖNIG; ZIEGLER, 2010) ou mesmo em Java (FASTRF, 2013); algumas são utilizadas por linha de comando (BREIMAN; CUTLER, 2013; PARF, 2013; ZHANG; WANG; CHEN, 2009; SCHWARZ; KÖNIG; ZIEGLER, 2010) enquanto outras são utilizadas dentro de ambientes de programação como R (LIAW; WIENER, 2012) ou Python (SCIKIT, 2013); algu-

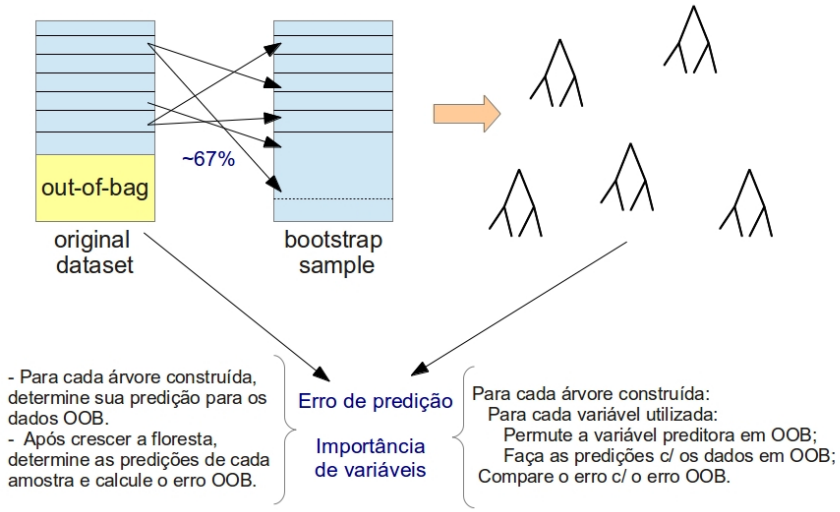

Figura 3.1 – Procedimento embutido em RF para estimar o erro OOB e da importância das variáveis.

mas exploram mecanismos de paralelismo em suas implementações (PARF, 2013; SCHWARZ; KÖNIG; ZIEGLER, 2010), enquanto algumas foram implementadas especificamente para aplicação em GWAS (ZHANG; WANG; CHEN, 2009; SCHWARZ; KÖNIG; ZIEGLER, 2010).

3.3 Estudo com dados simulados

O objetivo do estudo apresentado nesta seção é utilizar dados simulados para avaliar a capacidade de RF em identificar as regiões de QTL (*quantitative trait loci*) (regiões associadas a características quantitativas de interesse relacionadas ao melhoramento animal), utilizando seu mecanismo de mensuração de importância de variáveis, ou seja, verificando se SNPs próximos às regiões de QTL são ranqueados entre as variáveis mais importantes reportadas por RF. Para isso, foram utilizados os conjuntos de dados simulados utilizados nas edições de 2011 e 2012 do *workshop* QTL-MAS, (ELSEN et al., 2012; QTL-MAS, 2012) para avaliar novas

metodologias de GWAS e GWS, utilizando dados com as características encontradas em populações de animais utilizados em programas de melhoramento. Especificamente, os dados de 2011 representam uma população típica de melhoramento de suínos, enquanto os dados de 2012 uma população típica de melhoramento de gado de leite.

No *workshop* QTL-MAS 2011, a população utilizada consistia de 20 famílias de machos não independentes, onde cada macho foi acasalado com 10 fêmeas. Cada fêmea foi acasalada com apenas um macho e gerou dois grupos de progênies, um com 10 indivíduos e outro com 5. O primeiro grupo, contendo 2.000 indivíduos, constituiu a população experimental e continha tanto o genótipo quanto o fenótipo, enquanto o segundo grupo, com 1.000 animais, constituiu o grupo de seleção e continha apenas informação sobre genótipo. Para ambos os grupos também estava disponível o valor genético verdadeiro ou *true breeding value* (TBV). A geração parental de 20 machos e 200 fêmeas foi gerada, a partir da escolha aleatória de 2 gametas de conjuntos de 75 gerados utilizando o software LDSO (YTOURNEL et al., 2012).

A simulação consistiu de dois passos: 1.000 gerações de uma população de 1.000 gametas, seguida de uma forte redução da população (*bottleneck*) em que 150 gametas evoluíram por 30 gerações. A estrutura do genoma considerado contém 5 cromossomos (autossomos) de comprimento igual a 1 Morgan. Em cada cromossomo foram simulados 1.998 SNPs localizados a cada 0,05 cM, resultando em um painel com 9.990 SNPs. Um conjunto de 1.000 gametas foi inicialmente gerado em equilíbrio de ligação durante as 1.150 gerações simuladas, considerando-se uma taxa de mutação de 0,0002 (ELSEN et al., 2012). Foram simulados 8 QTLs, cuja segregação acrescida de ruído ambiental constituíram a variação do fenótipo. Esta variação foi ajustada para representar uma herdabilidade de 0,3 e as características dos QTLs nos diferen-

tes cromossomos escolhidas para representar situações extremas (Tabela 3.1).

Tabela 3.1 – Efeito dos QTLs nos dados do QTL-MAS 2011 (ELSEN et al., 2012). Crom: cromossomo.

QTL	Crom	Pos (cM)	Tipo				
QTL1	1	2.85	4 alelos, aditivo, grande. Alelo 1 = 0., 2 = 2., 3 = 4., 4= 6.				
					11	12	22
QTL2	2	81.9	em fase c/ QTL3	11	-4.	-2.	0.
QTL3		93.75	em fase c/ QTL2	12	-2.	0.	2.
				22	0.	2.	4.
					11	12	22
QTL4	3	5.0	em oposição c/ QTL5	11	0.	2.	4.
QTL5		15.0	em oposição c/ QTL4	12	-2.	0.	2.
				22	-4.	-2.	0.
				11	11	12	22
QTL6	4	32.2	imprinted	2.0	0.0	0.0	0.0
					11	12	22
QTL7	5	36.3	epistasia c/ QTL8	11	2.	1.	0.
QTL8		99.2	epistaisa c/ QTL7	12	0.	0.	0.
				22	0.	0.	0.

Já no *workshop* de 2012, uma população base, *G0*, com 1.020 indivíduos não relacionados, sendo 20 machos e 1.000 fêmeas, foi gerada com 5 cromossomos (autossomos), cada um com comprimento de 99,95 Mb, resultando em um genoma de comprimento 499,750 Mb. Em cada cromossomo, foram distribuídos um conjunto de 2.000 SNPs igualmente espaçados a cada 0,05 Mb (ou cM). A população base foi gerada de forma a obter um decaimento de desequilíbrio de ligação (LD) e distribuição de frequência alélica mínima (MAF) fixos. A partir de *G0* foram geradas 4 gerações, *G1* – *G4*, não sobrepostas, formadas por 20 machos e 1.000 fêmeas por acasalamento aleatório entre cada macho com 51 fêmeas. Cada fêmea gerou 1 fêmea, exceto as mães dos machos da próxima geração que geraram 1 macho e 1 fêmea. Três características correlacionadas para emular produtividade foram geradas, com efeitos distribuídos entre 50 QTLs distribuídos ao longo do genoma (Fi-

gura 3.2). Neste evento, o desafio consistiu em encontrar os QTLs e a ação em pleitropia entre eles (YTOURNEL et al., 2012). Aqui, para avaliação de RF, apenas o fenótipo 1 será utilizado.

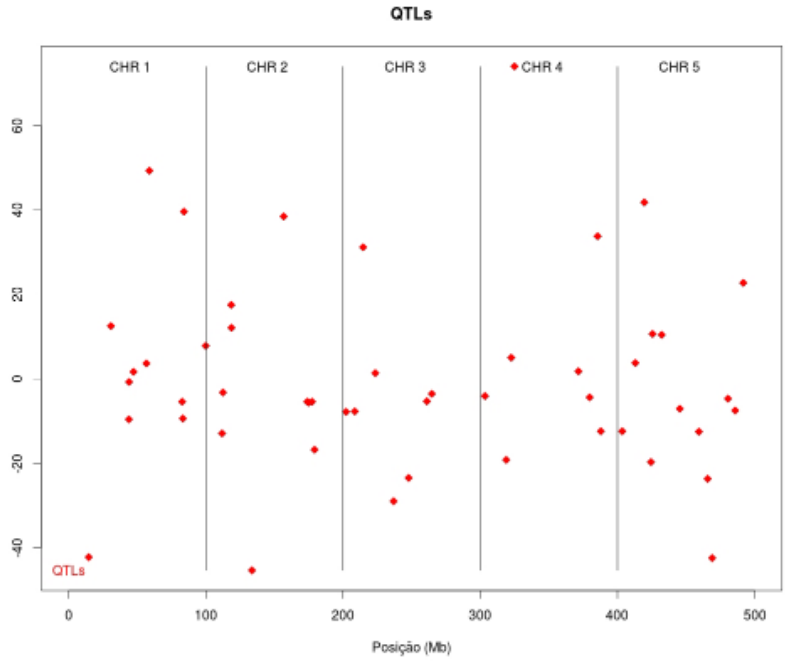

Figura 3.2 – Localização dos 50 QTLs com efeito para o fenótipo 1, ao longo do genoma, para os dados do QTL-MAS 2012. Os efeito dos QTLs são apresentados no eixo y.

Observe que, em ambos os casos, os QTLs com efeito não integram o painel utilizado para genotipar os animais (Figura 3.3).

Figura 3.3 – Ilustração do painel utilizado para genotipagem (QTL-MAS 2011). Os QTLs não integram o painel.

O pressuposto é que tanto o número de QTLs, quanto suas posições ao longo do genoma são desconhecidas, mas que um número pré-determinado, k, de SNPs no topo da lista dos reportados como variáveis importantes serão selecionados como implicados com o fenótipo analisado. Essa é a situação típica que se encontra ao analisar dados reais e, por essa razão o critério de avaliação utilizado para avaliar RF é a medida de precisão (Expressão 3.1):

$$\text{precisão} = \frac{\text{\#SNPs corretos}}{\text{\#SNPs preditos}} \tag{3.1}$$

onde um SNP predito é considerado correto se ele está a uma distância inferior a 0,25 cM (dados de 2011) ou 250 kb (dados de 2012). Os SNPs preditos são os k SNPs selecionados por RF, com os valores de k variando entre 10 e 50. A medida de precisão avalia a proporção de SNPs corretos dentre o conjunto de k preditos.

Em todos os casos, utilizou-se RF com o parâmetro mtry = 0,4p – onde p é o número de SNPs – e ntree = 2.000. Para os demais parâmetros utilizou-se os valores *default* do pacote R randomForest (LIAW; WIENER, 2012).

As Figuras 3.4 e 3.5 mostram os gráficos da precisão em função do número k de SNPs considerados como preditos por RF para os conjuntos de dados do *workshop* QTL-MAS 2011 e 2012. Em ambos os casos, a precisão apresenta um comportamento similar: quando $k = 10$, a precisão é igual a 0,5 para os dados de 2011 e 0,4 para os dados de 2012 e decaem a medida que o valor de k aumenta até 50. A curva contínua em cinza representa a precisão esperada quando a predição consiste em selecionar aleatoriamente k SNPs, enquanto a linha pontilhada cinza representa este valor acrescido de 4 desvios padrão. As Figuras 3.4 e 3.5 representam visualmente o quão estatisticamente significante é o resultado obtido.

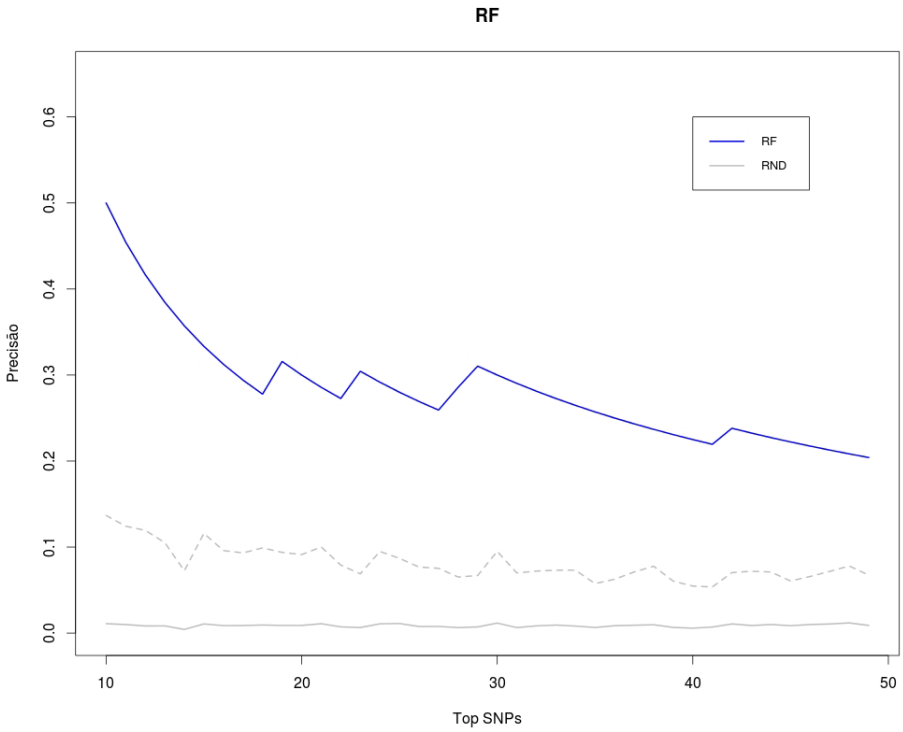

Figura 3.4 – Precisão em função do número de SNPs selecionados por RF para dados do QTL-MAS 2011.

A medida de precisão aqui utilizada considera corretos apenas os 5 SNPs mais próximos em cada lado do QTL, sendo todos os outros considerados incorretos. Para verificar o quão distante dos QTLs estão os outros SNPs preditos por RF, as Figuras 3.6 e 3.7 apresentam suas localizações no genoma, considerando sua posição na lista de ranqueamento. Na Figura 3.6 é possível notar que as predições que se encontram no topo do lista para os dados do QTL-MAS 2011 correspondem a SNPs próximos do QTL 1, que é o QTL de maior efeito. Também é possível notar que todos os 10 SNPs no topo da lista de preditos estão bem próximos de um QTL (QTL 1), apesar de apenas 5 serem considerados corretos, de acordo com a condição imposta para o cálculo de precisão. Examinando os demais SNPs na lista de ranqueamento, nota-se que grupos de SNPs próximos das posições dos QTLs 2, 3, 4 e 5

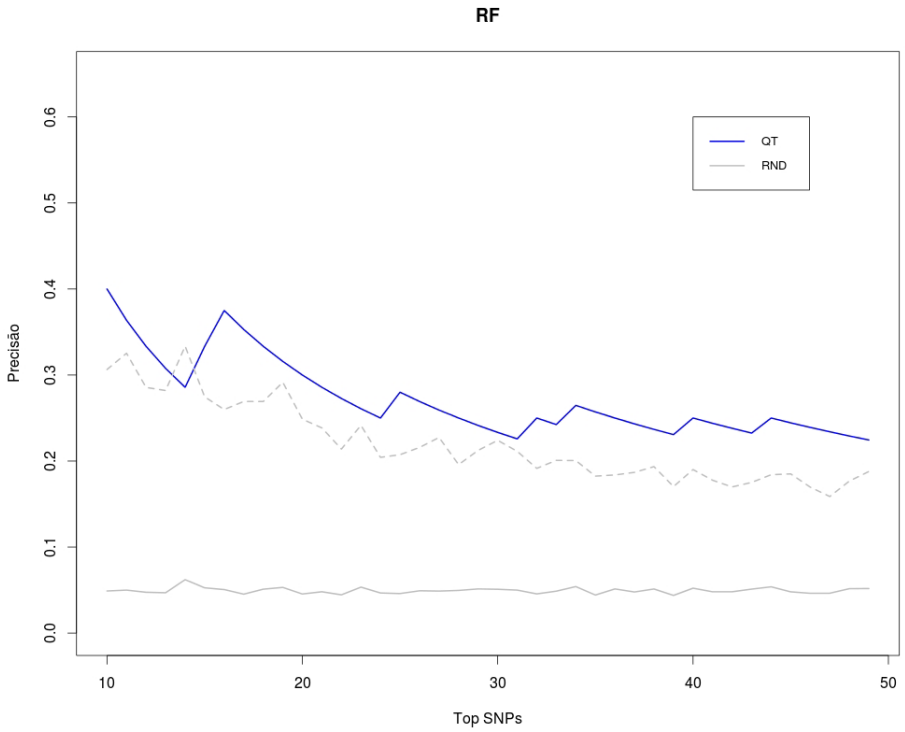

Figura 3.5 – Precisão em função do número de SNPs selecionados por RF para dados do QTL-MAS 2012.

nos cromossomos 2 e 3, mas o mesmo não ocorre para os QTLs nos cromossomos 4 e 5, que são QTLs com efeito de *imprinting* e epistasia.

A estrutura dos QTLs dos dados do QTL-MAS 2012 é bem diferente daquela exibida pelos dados do QTL-MAS 2011 (Tabela 3.1 e Figura 3.2). Enquanto os dados de 2011 apresentam apenas 8 QTLs com tipos de efeitos diversificados, os dados de 2012 contém 50 QTLs com efeitos aditivos espalhados ao longo do genoma. Contudo, da mesma forma que no caso dos dados do QTL-MAS 2011, a Figura 3.7 mostra que vários dos SNPs no topo da lista de preditos estão próximos de QTLs de grande efeito (nos cromossomos 1, 4 e 5). Com foco nos QTLs de maior efeito, isto é, $\text{efeito} \geq \text{abs}(20)$, nota-se uma concentração de SNPs preditos em torno da maioria dos QTLs (8 de 12), evidenciando a mesma

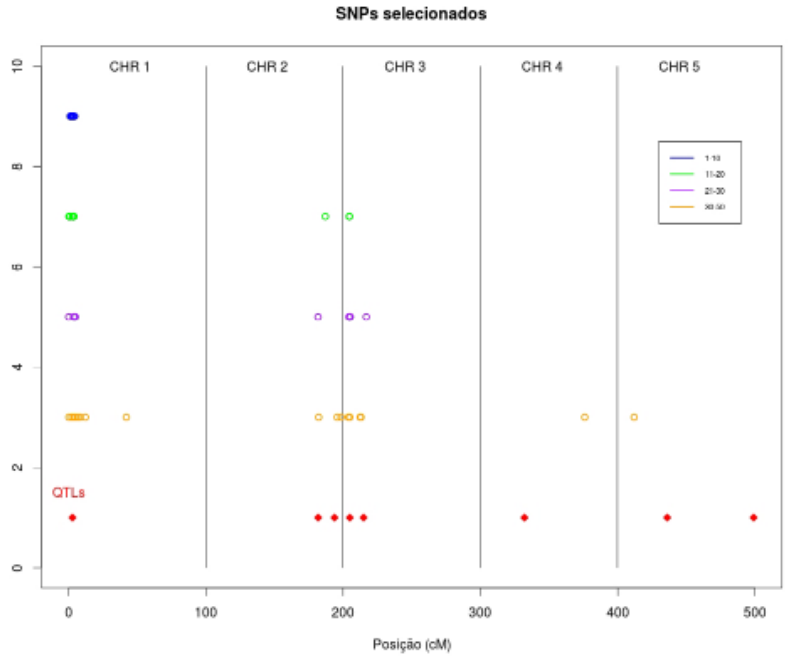

Figura 3.6 – Localização dos SNPs selecionados por RF para os dados de QTL-MAS 2011.

tendência exibida para os dados de 2011 de viés para QTLs de maior efeito.

Em função da alta dimensionalidade dos dados que, neste caso é de 10.000 SNPs, mas que para dados reais pode chegar a centenas de milhares de SNPs – por exemplo, 770.000 SNPs –, é conveniente paralelizar a tarefa de predição de SNPs associados ao fenótipo. Por construção, RF apresenta dois pontos que se prestam à paralelização da implementação: (i) o processo de crescimento das árvores de classificação/regressão e (ii) o processo de avaliação de variáveis para particionamento de um nó de uma árvore (vide Seção 3.2). Entretanto, dada a divisão natural do genoma em cromossomos, é possível pensar em executar RF em dois passos. No primeiro passo, RF é utilizado separadamente, e em paralelo, para selecionar os t SNPs mais importantes em cada cromossomo; e, no segundo passo, RF é novamente utilizado para selecionar os

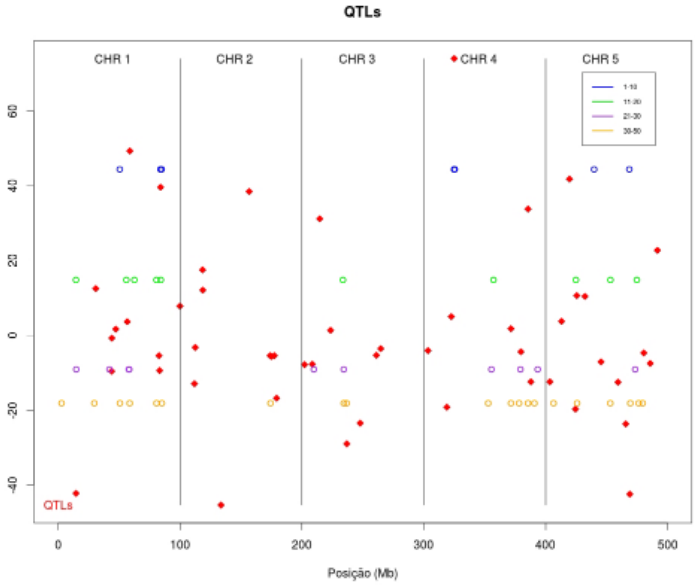

Figura 3.7 – Localização dos SNPs selecionados por RF para os dados de QTL-MAS 2012.

k SNPs mais importantes dentre a união dos conjuntos de t SNPs selecionados para cada cromossomo. Para avaliar o impacto da utilização desse processo empírico de pseudo-paralelização, as Figuras 3.8 e 3.9 apresentam os gráficos sobrepostos da precisão em função do número k de SNPs considerados como preditos por RF e por RF em dois passos para os conjuntos de dados do *workshop* QTL-MAS 2011 e 2012. As curvas de precisão utilizando RF e RF em dois passos apresentam o mesmo comportamento com níveis de precisão comparáveis e até com uma ligeira vantagem no caso dos dados do QTL-MAS 2012 (Figura 3.9).

Finalmente, mesmo não sendo o foco deste trabalho predizer os valores genéticos dos animais, em seguida comparou-se, por meio da correlação de Pearson, as estimativas obtidas utilizando RF e os 50 SNPs identificados como mais importantes com os valores de fenótipo e os TBVs fornecidos com os conjuntos de dados. No caso do fenótipo, as predições foram realizadas utilizando-se apenas as populações experimentais (2.000 animais para o ano

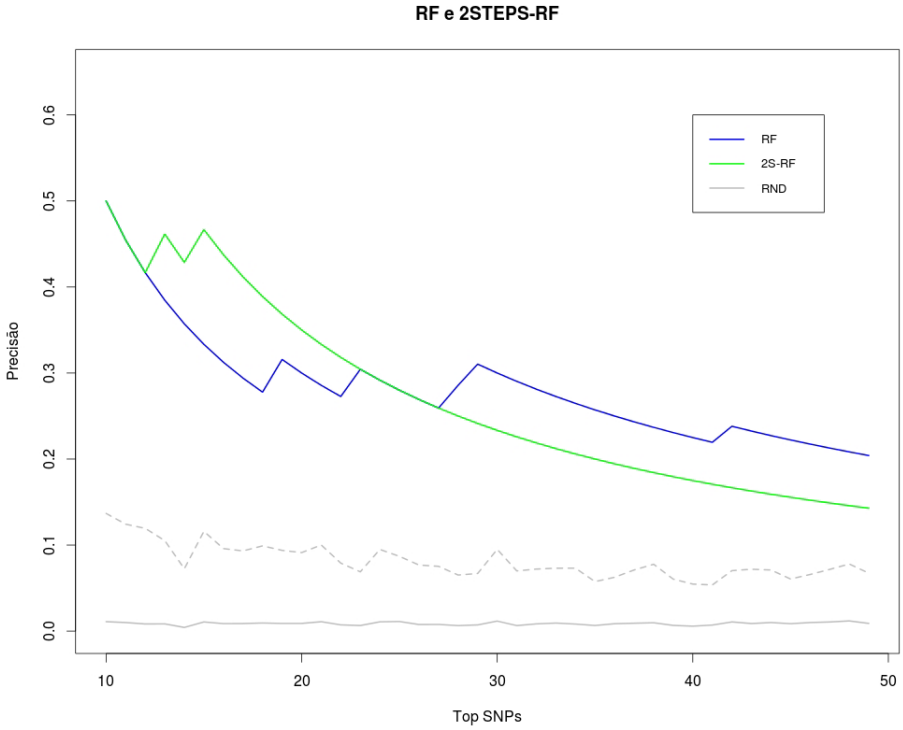

Figura 3.8 – Comparação entre a precisões em função do número de SNPs selecionados por RF e RF em dois passos para os dados do QTL-MAS 2011.

de 2011 e 3.000 animais para o ano de 2012, correspondentes às gerações, *G1 – G3*), por meio de um processo de validação cruzada com 10 partições. Já para comparação com os valores genéticos verdadeiros, foram utilizados os dados experimentais de cada ano para treinar RF e a estimativa feita para os animais do grupo de validação (1.000 animais para o ano de 2011 e 1.000 animais para o ano de 2012, correspondente aos animais da geração *G4*). Os valores de correlação obtidos em todos os casos são apresentados na Tabela 3.2 e são muito superiores aos valores de correlação obtidos utilizando-se o mesmo número de SNPs selecionados de forma aleatória ao longo do genoma. Relembrando que os SNPs foram selecionados baseados em dados de fenótipos obtidos pela combinação dos efeitos dos QTLs e ruído (aleatório) ambiental, é

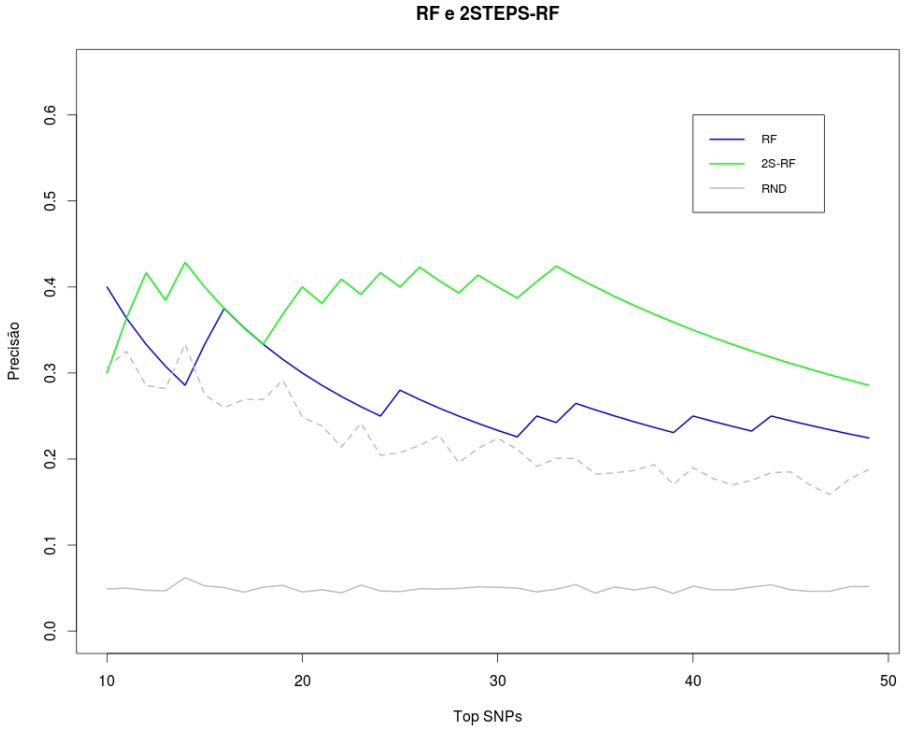

Figura 3.9 – Comparação entre a precisões em função do número de SNPs selecionados por RF e RF em dois passos para os dados do QTL-MAS 2012.

Tabela 3.2 – Correlação entre valores preditos e verdadeiros para fenótipo e valor genético verdadeiro (TBV).

	2011		2012	
	Fenótipo	TBV	Fenótipo	TBV
RF	0,49	0,73	0,44	0,74
RF em dois passos	0,51	0,77	0,45	0,70
aleatório	$0,27 \pm 0,04$	$0,5 \pm 0,05$	$0,20 \pm 0,03$	$0,31 \pm 0,05$

notável que o valor genético predito com os SNPs selecionados por RF apresentem uma correlação com o TBV acima de 0,70 em todos os casos.

3.4 Aplicação usando dados reais

A aplicação apresentada nesta seção é um estudo de associação para espessura de gordura em gado Canchim (MOKRY et al., 2013). A raça Canchim é uma raça sintética formada por $\frac{3}{8}$ Zebu e $\frac{5}{8}$ Charolês e desenvolvida no Brasil no início da década de 60 com a intenção de combinar as características de adaptação da raça Zebu e a eficiência de produção e qualidade de carne das raças taurinas (VIANNA, 1960).

Neste estudo, foram utilizados animais, registrados na Associação Brasileira de Criadores de Canchim, pertencentes a 7 diferentes rebanhos nos estados de São Paulo e Goiás. Uma amostra de 987 animais contendo machos e fêmeas, nascidos entre os anos de 2003 e 2005 e criados a pasto, foi avaliada para espessura de gordura por ultrassom *in vivo* sobre a 12ª costela. Os valores genéticos (EBVs) desses animais foram estimados por máxima verossimilhança restrita, usando o software MTDFREML (BOLDMAN et al., 1995). O modelo animal incluiu efeitos fixos de grupo de contemporâneos (sexo, ano, rebanho e grupo genético) e idade à época da medida como co-variáveis lineares, efeito genético aditivo e o erro. A partir desse conjunto de animais, foram selecionados 400, levando em consideração o EBVs, acurácia, tamanho de família e proporção entre machos e fêmeas. Esses animais foram, então, genotipados com o *chip* Illumina BovineHD, que contém 777.000 SNPs. Para o estudo de associação, considerou-se como fenótipo o valor genético derregredido (dEBV) (GARRICK; TAYLOR; FERNANDO, 2009), um pseudo fenótipo que leva em conta a matriz de *pedigree*, a herdabilidade estimada (0,16), o valor de EBV, suas correspondentes acurácias e o mesmo modelo animal descrito acima. Para a estimação de dEBV o conjunto de dados inicial foi suplementado com dados de animais nascidos entre 2005 e 2008,

totalizando 1.648 animais com fenótipos para espessura de gordura e 6.801 animais na matriz de *pedigree*.

Após um procedimento de controle de qualidade (*call rate* < 0,90 para amostras e SNPs; MAF < 0,01 e heterozigosidade < 3 desvios padrões), foram utilizados na sequência do estudo 396 animiais e 708.641 SNPs, com um *call rate* médio de 0,99. Para o estudo de associação entre SNPs e o dEBV foi utilizado o pacote randomForest (LIAW; WIENER, 2012) do software R (R CORE TEAM, 2013).

A estratégia utilizada foi a de RF em dois passos, conforme explicado na Seção 3.3. No primeiro passo, os 1% SNPs melhor ranqueados por RF em cada cromossomo foram selecionados e re-analisados em conjunto por RF em um segundo passo, novamente retendo os 1% SNPs melhor ranqueados. Foram utilizados os parâmetros ntree = 5.000 e mtry = 0,4p, enquanto que para os demais parâmetros foram adotados os valores *default* fornecidos pelo pacote randomForest. Os SNPs selecionados foram, então, utilizados no ajuste de um modelo de regressão *stepwise* usando o software SAS/STAT (SAS INSTITUTE INC., 2011) para estimar a quantidade de variância explicada pelo conjunto de SNPs selecionados. No final, 21 SNPs foram selecionados para compôr o modelo de regressão final (Tabela 3.3), explicando 53% da variância observada (R^2) no fenótipo (dEBV).

Uma segunda estratégia para utilização de RF consistiu em aplicar o procedimento acima descrito a 10 sub-amostras de 198 animais, construídas da seguinte forma: i) o primeiro animal foi escolhido de forma aleatória dentre os 396 animais genotipados; ii) o próximo animal foi escolhido considerando o menor parentesco com os animais previamente selecionados, mas representativo entre os demais animais genotipados; iii) o passo (ii) foi repetido até que 198 animais tivessem sido selecionados.

Tabela 3.3 – Os cinco primeiros SNPs selecionados ao ajustar o modelo animal. QTLs: SF – gordura subcutânea; MS – escore de marmoreio; FT12R – gordura subcutânea na 12^a costela; IF – gordura intramuscular; OAC – conteúdo de ácido oléico; PAC – conteúdo de ácido palmitoleico.

dbSNP	Crom	Pos	Genes	QTL
rs133046994	10	18129602	*THSD4, LRRC49*	SF, MS
rs137294146	1	132385787	*SOX14, CLDN18*, *DZP1L*	FT12R, IF
rs109349988	3	15814096	*KCNN3, EFNA3*, *EFNA4, DCST2*, *LOC100294774*, *PMVK, ADAR*, *CHRNB2, ADAM15*, *ZBTB7B, DCST1*, *LOC100294857*, *FLAD1, PYGO2*, *CKS1B, PBXIP1*, *SHC1*, *LOC100294894*	FT12R, MS
rs136717249	19	37969870	*B4GALNT2, GNGT2*, *ABI3, NGFR, GIP*, *PHOSPHO1, ZNF652*, *PHB, IGF2BP1*	OAC, PAC
rs134790147	13	20780821	*CCDC7, ARL5B*, *MGC152301*, *LOC100848675*, *LOC100847992*	FT12R

Os SNPs comuns entre a estratégia utilizando 398 animais e essas 10 sub-amostras foram, então, analisados para ajustar o mesmo modelo de regressão *stepwise* descrito acima. Neste caso, foram selecionados 19 SNPs que explicaram 50% da variância observada (R^2) no fenótipo (dEBV). Além disso, nas duas estratégias, os primeiros 5 SNPs do modelo de regressão foram os mesmos, com a mesma ordem de importância (rs133046994, rs137294146, rs109349988, rs136717249, rs134790147) e explicam 34% da variância observada (R^2) no fenótipo (dEBV).

Em seguida, os haplótipos para cada cromossomo foram reconstruídos utilizando o software fastPHASE, versão 1.4.0 (SCHEET; STEPHENS, 2006), e analisados por meio do software Haploview (BARRETT et al., 2005), usando seus parâmetros *default*. A estimação de blocos de haplótipos e LD baseou-se no coeficiente de correlação quadrado entre pares de SNPs (r^2). Assumiu-se o valor médio de r^2 = 0,12 como determinante da extensão de LD, tal que a distância de 250 kb em torno de cada SNP foi considerado para descobrir genes candidatos.

Esses genes e as informações contidas nas bases de dados NCBI BioSystems database (GEER et al., 2010) e Kyoto Encyclopedia of Genes and Genomes (KEGG) (KANEHISA; GOTO, 2000; KANEHISA et al., 2012) foram utilizadas para obter informações sobre processos biológicos relacionados com espessura de gordura. Dentre os 21 SNPs identificados, um SNP (cromossomo 3: rs42021729) não possui genes descritos em sua vizinhança, enquanto dentre os outros 20, apenas 4 (cromossomo 12: rs136348926; cromossomo 11: rs110833507; cromossomo 2: rs42923911; cromossomo 9: rs110025080) não estão em regiões de QTL previamente descritas na literatura (MOKRY et al., 2013). Já quando os genes candidatos identificados são analisados, observa-se que diversos deles contém anotação relacionando-os com metabolismo de lipídeos, por exemplo:

- o gene *THSD4* codifica uma proteína com domínios de disintegrina e metaloprotease, que possui uma função importante no processo de adipogênese (MCDANIEL et al., 2006);

- o gene *PMVK* (*phosphomevalonate kinase*), que cataliza a conversão de mevalonato-5-fosfato e ATP para mevalonato-5-difosfato e ADP, que é uma das reações iniciais envolvidas na via de síntese de colesterol (HERDENDORF; MIZIORKO, 2007);

- o gene *ADAR* (*adenosine deaminase, RNA-specific*), que em um estudo com humanos, foi implicado com níveis de triglicérides, adiponectina, circunferência abdominal e índice de massa corporal (OGURO et al., 2012); e
- o gene *SHC*1 (*Src homology 2 domain containing – transforming protein 1*), que foi implicado com a obesidade em humanos (EIGELSON et al., 2008).

Diferente do que ocorre quando se analisa dados simulados (Seção 3.3), ao se analisar dados reais, as respostas (posições dos QTLs) não são conhecidas. Por essa razão, não é possível avaliar o quão próximos os SNPs selecionados estão de variações com efeito no fenótipoo. Contudo, a análise funcional de genes próximos dos principais SNPs revela que eles estão envolvidos com atividades biológicas relacionadas à síntese e acúmulo de gordura, o que confere ao resultado (lista de SNPs) uma alta plausabilidade. A lista completa dos SNPs selecionados, genes candidatos e suas relações com o fenótipo analisado podem ser encontrados em Mokry et al. (2013).

3.5 Discussão e conclusão

Para demonstrar a aplicabilidade de RF a problemas de GWAS em ciência animal, nas seções anteriores utilizou-se: (i) dados simulados com a estrutura genética tipicamente encontrada em populações utilizadas em programas de melhoramento animal; e (ii) dados reais de bovinos de corte, cujos resultados foram recentemente publicados pelo nosso grupo (MOKRY et al., 2013).

Os dados simulados foram obtidos dos sites do *workshop* QTL-MAS, anos de 2011 e 2012, e se assemelham à estrutura encontrada em populações de suínos e bovinos de leite, respectivamente. Esses conjuntos de dados foram gerados por terceiros, o

que elimina a possibilidade de enviesamento em favor da metodologia avaliada. RF foi avaliado quanto à sua capacidade de identificar os QTLs presentes nesses conjuntos de dados, resultando em uma precisão entre 40% e 50%, considerando os 10 primeiros SNPs reportados e uma tolerância de 5 SNPs de distância (250 kb) para a posição da variação causal. Esses resultados demonstram a capacidade de RF de encontrar SNPs relacionados ao fenótipo de interesse em situações em que os mecanismos moleculares que controlam os fenótipos são muito diversos.

Ainda considerando os dados simulados, utilizou-se os 50 SNPs selecionados por RF (50 primeiros no ranqueamento gerado por RF) para verificar a correlação do fenótipo estimado com esses SNPs e o valor genético verdadeiro. Este variou entre 0,7 e 0,77 em ambos os casos, o que é bastante significativo quando se considera que RF baseou a seleção dos SNPs na medida de fenótipo, que inclui o valor genético acrescido de ruído ambiental. Esses resultados demonstram a robustez de RF a ruídos presentes nos dados e a diferentes estruturas genéticas relacionadas com os QTLs estudados.

Apesar de sua natureza paralelizável, os conjuntos de dados simulados também foram utilizados para avaliar uma estratégia de aplicação de RF em dois passos, o que resulta em uma pseudo-paralelização. Os níveis de desempenho obtidos utilizando essa estratégia foram comparáveis aos obtidos anteriormente, demonstrando a flexibilidade de RF, e potencial de escalabilidade para problemas envolvendo milhares de amostras e genótipos de centenas de milhares de SNPs, que é o caso encontrado em GWAS.

RF também foi aplicado a um conjunto de dados reais, a espessura de gordura em Canchim, sendo que o conjunto de 21 SNPs selecionados explicou tanto quanto 50% da variância observada no fenótipo. Uma análise funcional das regiões a que pertencem os SNPs selecionados mostrou diversos genes em QTLs

relacionados a metabolismo envolvendo gordura e genes implicados com processos biológicos como adipogênese e metabolismo de lipídeos ou associados a fenótipos fortemente relacionados a acúmulo de goudura como obesidade em humanos. Esses resultados são bastante plausíveis à luz da característica estudada, a espessura de gordura, e constitui mais uma evidência do potencial de aplicação de RF em GWAS na área de ciência animal.

Os resultados obtidos nos estudos apresentados fornecem evidências que suportam a adequabilidade da utilização de RF para GWAS, especificamente aplicado à área de ciência animal. Apesar de simples, a técnica de RF é, ao mesmo tempo, flexível, robusta e escalável para problemas da ordem encontrada ao se analisar dados de GWAS.

3.6 Referências

AMIT, Y.; GEMAN, D. Shape quantization and recognition with randomized trees. *Neural Computation*, v. 9, p. 1545–1588, 1997.

BARRETT, J. C. et al. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, v. 21, p. 263–265, 2005.

BOLDMAN, K. G. et al. *MTDFREML. A Set of Programs to Obtain Estimates of Variances and Covariances*. Washington. DC: U.S.: Department of Agriculture, Agricultural Research Service, 1995.

BOVINE GENOME SEQUENCING; ANALYSIS CONSORTIUM. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, v. 342, n. 5926, p. 522–528, 2009.

BOVINE HAPMAP CONSORTIUM. Genome-wide survey of snp variation uncovers the genetic structure of cattle breeds. *Science*, v. 324, n. 5926, p. 528–532, 2009.

BREIMAN, L. Bagging predictors. *Machine Learning*, v. 24, p. 123–140, 2001.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, p. 5–32, 2001.

BREIMAN, L.; CUTLER, A. *Random Forest*. 2013. Disponível em: <http://www.stat.berkeley.edu/b̃reiman/RandomForests/>. Acesso em: 15.9.2013.

BREIMAN, L. et al. *Classification and Regression Trees*. London: Chapman & Hall, 1984.

BUREAU, A. et al. Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, v. 28, p. 171–182, 2005.

EIGELSON, H. S. et al. Genetic variation in candidate obesity genes adrb2, adrb3, ghrl, hsd11b1, irs1, irs2, and shc1 and risk for breast cancer in the cancer prevention study ii. *Breast Cancer Research*, v. 10, p. R57, 2008.

ELSEN, J.-M. et al. XVth QTLMAS: simulated dataset. *BMC Proceedings*, v. 6(Suppl 2), n. S1, 2012. Disponível em: <https://www.biomedcentral.com/1753-6561/6/S2/S1>.

FASTRF. *fast-random-forest: An efficient implmentation of the Random forest classifier for Java*. 2013. Disponível em: <https://code.google.com/p/fast-random-forest/>. Acesso em: 15.9.2013.

FREUND, Y.; SHAPIRE, R. Experiments with a new boosting algorithm. In: SAITTA, L. (Ed.). *Proceedings of the 13th International Conference of Machine Learning*. San Francisco: Morgan Kaufmann, 1996. p. 148–156.

GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic

regression analyses. *Genetics, Selection, Evolution*, p. 41–55, 2009.

GEER, L. Y. et al. The ncbi biosystems database. *Nucleic Acid Research*, v. 38, p. D492–D496, 2010.

GOLDSTEIN, B. A. et al. An application of random forests to a genome-wide association: Methodological considerations & new findings. *BMC Genetics*, v. 11, n. 49, 2010. Disponível em: <http://www.biomedcentral.com/1471-2156/11/49>.

GONZÁLEZ-RECIO, O.; FORNI, S. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genetics Selection Evolution*, v. 43, n. 7, 2011. Disponível em: <http://www.gsejournal.org/content/43/1/7>.

HERDENDORF, T. J.; MIZIORKO, H. M. Functional evaluation of conserved basic residues in human phosphomevalonate kinase. *Biochemistry*, v. 46, p. 11780–11788, 2007.

HO, T. K. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Learning Intelligence*, v. 20, n. 8, p. 832–844, 1998.

KANEHISA, M.; GOTO, S. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acid Research*, v. 28, p. 27–30, 2000.

KANEHISA, M. et al. Kegg for integration and interpretation of large-scale molecular data sets. *Nucleic Acid Research*, v. 40, p. D109–D114, 2012.

LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2/3, p. 18–22, 2012.

MCDANIEL, A. H. et al. A locus on mouse chromosome 9 (adip5) affects the relative weight of the gonadal but not retroperitoneal adipose depot. *Mammalian Genome*, v. 17, p. 1078–1092, 2006.

MOKRY, F. B. et al. Genome-wide association study for backfat tickness in canchim beef cattle using random forest approach. *BMC Genetics*, v. 14, n. 47, 2013. Disponível em: <http://www.biomedcentral.com/1471-2156/14/47>.

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bionformatics challenges for genome-wide association studies. *Bioinformatics*, v. 26, n. 4, p. 445–455, 2010.

OGURO, R. et al. A single nucleotide polymorphism of the adenosine deaminase, rna-specific gene is associated with the serum triglyceride level, abdominal circumference, and serum adiponectin concentration. *Biochemistry*, v. 47, p. 183–187, 2012.

PARF. *parf: Parllel Random Forest Algorithm*. 2013. Disponível em: <http://code.google.com/p/parf/>. Acesso em: 15.9.2013.

QTL-MAS. *16th QTL-MAS Workshop*. 2012. Sítio do 16th QTL-MAS Workshop. Disponível em: <http://qtl-mas-2012.kassiopeagroup.com/en/index.php>. Acesso em: 15.9.2013.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

SAS INSTITUTE INC. *SAS/STAT Software. SAS Institute Inc: Version 9.3*. Cary N, 2011.

SCHEET, P.; STEPHENS, M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *American Journal of Human Genetics*, v. 78, p. 629–644, 2006.

SCHWARZ, D. F.; KÖNIG, I. R.; ZIEGLER, A. On safari to random jungle: a fast implementation of

random forests for high-dimensional data, bioinformatics. *Bioinformatics*, v. 26, n. 14, p. 1752–1758, 2010. Disponível em: <http://bioinformatics.oxfordjournals.org/content/26/14/1752.long>.

SCHWARZ, D. F. et al. Picking single-nucleotide polymorphisms in forests. *BMC Proceedings*, v. 1(suppl. 1), n. S59, 2007. Disponível em: <http://www.biomedcentral.com/1753-6561/1/S1/S59>.

SCIKIT. *scikit-learn: Machine Learning in Python*. 2013. Disponível em: <http://scikit-learn.org/stable/>. Acesso em: 15.9.2013.

SHAWE-TAYLOR, J.; CRISTIANINI, N. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press, 2004.

SZYMCZAK, S. et al. Machine learning in genome-wide association studies. *Genetics Epidemiology*, v. 33(supplement 1), p. S51–S57, 2009.

VIANNA, A. T. *Formação do gado Canchim pelo cruzamento charoles-zebu*. Rio de Janeiro: Ministério da Agricultura, 1960.

YAO, C. et al. Random forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. *Journal of Diary Science*, v. 96, n. 10, p. 6716–6729, 2011.

YTOURNEL, F. et al. LDSO: a program to simulate pedigrees and molecular information under various evolutionary forces. *Journal of Animal Breeding and Genetics*, v. 129, p. 417–421, 2012.

ZHANG, H.; WANG, M.; CHEN, X. Willows: a memory efficient tree and forest contruction package. *BMC Bioinformatics*, v. 10, n. 130, 2009. Disponível em: <http://www.biomedcentral.com/1471-2105/10/130>.

ZIEGLER, A.; DESTEFANO, A. L.; KÖNIG, I. Data mining, neural nets, trees - probems 2 and 3 of genetic analysis workshop 15. *Genetics Epidemiology*, v. 31(supplement 1), p. S51–S60, 2007.