UNIVERSIDADE FEDERAL DE JUIZ DE FORA
PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

SELEÇÃO E AVALIAÇÃO DE MARCADORES MOLECULARES COM GRANDE INFORMATIVIDADE PARA A PREDIÇÃO DO VALOR GENÔMICO

Bruno Zonovelli da Silva

Juiz de Fora
Março de 2018

Bruno Zonovelli da Silva

Seleção e avaliação de marcadores moleculares com grande informatividade para a predição do valor genômico

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Orientador: Prof. D.Sc. Carlos Cristiano Hasenclever Borges

Coorientador: Prof. D.Sc. Wagner Antonio Arbex

Juiz de Fora

2018

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Silva, Bruno Zonovelli da.
Seleção e avaliação de marcadores moleculares com grande informatividade para a predição do valor genômico / Bruno Zonovelli da Silva. -- 2018.
178 f.

Orientador: Carlos Cristiano Hasenclever Borges
Coorientador: Wagner Antonio Arbex
Tese (doutorado) - Universidade Federal de Juiz de Fora, ICE/Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2018.

1. Aprendizado de Máquina. 2. Bioinformática. 3. Inteligência Computacional. 4. Seleção Genômica. I. Borges, Carlos Cristiano Hasenclever, orient. II. Arbex, Wagner Antonio, coorient. III. Título.

Bruno Zonovelli da Silva

Seleção e avaliação de marcadores moleculares com grande informatividade para a predição do valor genômico

Tese apresentada ao Programa de Pós-graduação em Modelagem Computacional, da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do grau de Doutor em Modelagem Computacional.

Aprovada em 7 de Março de 2018.

BANCA EXAMINADORA

Prof. D.Sc. Carlos Cristiano Hasenclever Borges - Orientador
Universidade Federal de Juiz de Fora

Prof. D.Sc. Wagner Antonio Arbex - Coorientador
Empresa Brasileira de Pesquisa Agropecuária

Prof. D.Sc. Moyses Nascimento
Universidade Federal de Viçosa

Prof. D.Sc. Claudio Napolis Costa
Empresa Brasileira de Pesquisa Agropecuária

Prof. D.Sc. Raul Fonseca Neto
Universidade Federal de Juiz de Fora

Prof. D.Sc. Saulo Moraes Villela
Universidade Federal de Juiz de Fora

Dedico este trabalho a minha esposa Débora e a minha filha Bruna Karla. Obrigado por tornarem esse sonho possível!

AGRADECIMENTOS

A trilha percorrida para que essas poucas palavras pudessem ser escritas foi longa, cansativa e extenuante. Mas, durante todo esse percurso conheci pessoas, fiz amigos e principalmente recebi apoio de pessoas próximas. Por esse motivo vou tentar expressar nessas palavras toda a gratidão por cada apoio, carinho, amizade e força recebida.

Certamente a citação nominal de algumas pessoas não é a forma mais justa de se agradecer a todos que de alguma forma apoiaram minha formação, mas as pessoas que estão nominadas são demasiadamente importantes em minha, e espero que continuem sendo por muitos anos.

Muitas são as pessoas a quem desejo agradecer. Primeiramente agradeço a meu Deus por guiar em todos os momentos, matérias e escolhas, sendo sempre o refugio certo em momentos de dúvidas e angustias.

A minha esposa Débora Cristina, mulher, amiga, parceira de vida, rainha do meu coração, minha companheira e inspiração para prosseguir a cada passo dado. Obrigado por estar sempre ao meu lado me incentivando a prosseguir. Débora te amo mais do que a mim. E obrigado pelo meu presentinho lindo, minha filhinha Bruna Karla, que ocupa um lugar imenso no meu coração. Minha filha obrigado por entender quando as vezes o papai precisava trabalhar no computador e não podia brincar com você.

Os amigos sim, eles, pessoas especiais, que te acompanham, ajudam, e escutam. Quero agradecer a todos. Em especial aos doutores Fabrizzio, Aldemon e Rafael Veiga, pessoas que aprendi a respeitar, não somente pela inteligência impar de cada um, mais pelo caráter e disposição de ambos, obrigado pelos conselhos. Aos amigos da cobertura foi um prazer dividir esse caloroso espaço com todos vocês.

A todos os outros que não citei, saibam que não é por esquecer, pois guardo todos em minhas melhores lembranças. E que Deus possa retribuir a cada um todas as ajudas que me deram.

Aos orientadores Carlos Cristiano e Wagner Arbex pela confiança a mim depositada, por me ouvirem, orientarem, mostrando o caminho a ser seguido, mas deixando-me livre para traçar meu próprio caminho, sempre se posicionando mais como conselheiros do que autoridades. Obrigado por confiar em mim.

Aos professores pela paciência e dedicação oferecidas aos alunos, pela disposição em

atenderem e instruírem nos mais variados assuntos, e por muitas vezes repetidamente. Aos coordenadores Rafael e Luís Paulo pela compreensão e paciência cada qual em seu momento, mas sempre com a mesmo energia exigida pelo nosso PGMC.

A UFJF, pelo financiamento da minha pesquisa, ajuda essa sem a qual não seria possível nem começar esse trabalho. Ao PGMC pela oportunidade oferecida. Aos funcionários e técnicos administrativos pela atenção e apoio, o grande Reginaldo sempre solícito e atencioso, as secretárias sempre carinhosas e pacientes.

A TODOS OBRIGADO !!!

NOVAMENTE OBRIGADO !!!

"Bem-aventurado o homem que acha sabedoria, e o homem que adquire conhecimento; Porque é melhor a sua mercadoria do que artigos de prata, e maior o seu lucro que o ouro mais fino. Mais preciosa é do que os rubis, e tudo o que mais possas desejar não se pode comparar a ela."

Provérbios 3:13-15

RESUMO

A seleção dos melhores indivíduos busca aprimorar uma característica ao longo do tempo. O uso de dados genômicos deram origem ao que é conhecido como seleção genômica. A construção de modelos genéticos eficientes para a avaliação do mérito de um indivíduo é complexa e no geral se baseia no pressuposto da herança aditiva. Entretanto, na presença de variabilidade genética não-aditiva os modelos podem não comportar toda a complexidade de possíveis interações entre os genes, a epistasia. O S4GS é um simulador de dados genômicos que busca mimetizar características importantes para o estudo em seleção genômica como, desequilíbrio de ligação, inseminação artificial e cruzamento geracional. Outro fator relevante é a capacidade de simular diferentes ações gênicas e interações em múltiplos níveis. Sendo utilizado na simulação de 8 cenários de estudo, com destaque para o cenário 8 que procurou simular o cruzamento do Girolando opção B. O método proposto consiste em duas etapas: seleção e a avaliação, gerando uma combinação ótima para o aumento de acurácia. Os algoritmos escolhidos para a etapa de seleção de atributos foram: o FFS; O SMS; e a CART como uma alternativa rápida. A etapa de avaliação utilizou duas técnicas clássicas o RR-BLUP e o BLASSO como referência, e o SVR. A associação das técnicas utilizadas na etapa de seleção e avaliação levam a três modelos: SVR + FFS; SVR + SMS; e SVR + CART. Nos resultados obtidos a seleção de atributos se mostrou um importante recurso no aumento da acurácia, em todos os 8 cenários. O processo de simulação possibilitou a obtenção de dados até a 15^a geração permitindo treinar os modelos na 1^a ou 4^a e aplicá-los nas subsequentes. A seleção de atributos aumentou de forma significativa a acurácia dos modelos utilizando dados genômicos, com exceção para o conjunto com amostra pequena e em dados totalmente lineares. O método proposto conseguiu para as bases com as características descristas serem eficientes, gerando um aumento significativo na correlação final.

Palavras-chave: Aprendizado de Máquina. Bioinformática. Inteligência Computacional. Seleção Genômica.

ABSTRACT

The animal breeding seeks to maximize of a characteristic over time. The use of genomic data gave rise that we know as genomic selection. The made of efficient models for evaluate the merit of an animal is complex and generally is based on the assumption of additive genetic effects. However, in the presence of non-additive genetic variability, the models may not contain all the complexity of possible interactions between the genes, the epistasis. The S4GS is a genomic data simulator developed in this thesis, which seeks to mimic important features for the study in genomic selection such as linkage disequilibrium, artificial insemination and crossing over. Another relevant factor is the ability to simulate different gene actions and interactions at multiple levels. It was used in the creation of 8 study scenarios, highlighting the scenario 8 that sought to simulate the Girolando option B. The proposed method consists of a two-step selection and evaluation, generating an optimal combination for the increase of accuracy. The algorithms chosen for the feature selection step were: the FFS that was developed in this thesis; The SMS; and CART as a quick alternative. The evaluation stage used two classical techniques, the RR-BLUP and the BLASSO as a reference, and the SVR. The association of the techniques used in the selection and evaluation stage leads us to three models: SVR + FFS; SVR + SMS; and SVR + CART. In the results obtained, the selection of attributes proved to be an important resource in increasing accuracy in all 8 scenarios. The simulation process allowed data to be obtained up to 15th generation allowing the models generated in 1th or 4th to in subsequent ones to be applied. The application of feature selection significantly increased accuracy in genomic data, except for the small sample set and in completely linear data. The proposed method was able to the bases with the descriptive characteristics to be efficient, generating a significant increase in the final correlation.

Keywords: Bioinformatics. Computational Intelligence. Genomic Selection. Machine Learning.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 Introdução

A seleção dos melhores indivíduos, visando o aprimoramento de uma característica em um rebanho, é uma prática comum e recorrente desde os tempos mais remotos da humanidade. Os bons resultados observados após uma primeira etapa de seleção incentivam a manutenção e aperfeiçoamento dessa prática. Melhorias são aplicadas de forma continuada gerando um importante aumento na qualidade final dos próximos rebanhos.

A definição correta dos melhores indivíduos é complexa, pois a expressão de uma característica pode envolver estruturas biológicas nem sempre conhecidas. Atualmente o processo de escolha utiliza métodos estatísticos próprios, pois a previsão da expressão futura de um fenótipos é claramente um problema estatístico (GEISSER, 1993; GIANOLA, 2013). Entretanto, mesmo que grande parte da expressões possua caraterísticas quantitativas, é possível observar características inferenciais, logo a decisão pelo uso do método estatístico, bem como sua definição, pode ser tornar mais uma decisão no já complexo processo de melhoramento genético.

As características quantitativas podem apresentar variações contínuas ou descontínuas, tendo também parte de sua origem não-genética. As bases para os estudos de genética quantitativa são os trabalhos de Haldane (1927), Fisher (1930), Wright (1931) e Falconer (1975), onde definem que a característica observada em um indivíduo é influenciada em parte pelo genótipo e outra parte pelo ambiente. Os valores genotípicos (G) podem ser decompostos em valor genético aditivo (A), de dominância (D) e de epistasia (I) que somados ao efeito ambiental (E) produzem o valor expressão fenotípica (F), bem como a interação genótipo ambiente, conforme Equação 1.1.

$$\begin{cases} G &= A + D + I \\ F &= G + E + G * E \end{cases} \tag{1.1}$$

A expressão do fenótipo observado em uma determinada população não é constante, como visto ela sofre múltiplas influência. A variação encontrada em uma determinada população (variação fenotípica - V_F) pode ser de duas origens: variação devido ao ambiente (V_E) e variação devido a diferenças genéticas (V_G). A caracterização do fenótipo é importante, pois grande parte da identidade de um genótipo é determinada por múl-

tiplos genes que possuem grande influência do ambiente. Esta influência é consequência da interação do genótipo mais o ambiente na manifestação de um determinado fenótipo. Portanto, procura-se determinar qual a proporção da variação fenotípica que se refere ao genótipo e ao ambiente dado pela Equação 1.2 (GRIFFITHS, 2008).

$$V_F = V_G + V_E \tag{1.2}$$

A variância fenotípica de uma população segregante, pode ser desdobrada para se estimar a proporção da variação que corresponde aos fatores genéticos da população selecionada e a proporção da variação devido ao fatores ambientais. Quanto esta população é testada em vários ambientes, pode ser quantificado a proporção da variação que corresponde a interação genótipo x ambiente (V_{GE}). Portanto o fenótipo pode ser expresso pela Equação 1.3:

$$V_F = V_G + V_E + V_{GE} \tag{1.3}$$

O genótipo refere-se à constituição genética de um organismo, representada por todos os genes que possui um indivíduo de uma espécie. O fenótipo é uma característica observada, identificada e individualizada, de difícil repetição, que se expressa através de um genótipo em um determinado ambiente (GRIFFITHS, 2008).

A herdabilidade é a proporção de variância genética sobre a fenotípica total, ou seja, a proporção herdável da bilinearidade total. Este proporção herdável é alterada pelo efeito do ambiente. Portanto, com o aumento da variabilidade proporcionado pelo efeito do ambiente, a seleção de novos genótipos torna-se mais difícil (GRIFFITHS, 2008).

A herdabilidade, Equação 1.4, é definida como o coeficiente de determinação entre a variação valor genotípico (V_G) e o valor fenotípico (V_F), ou a regressão do valor genotípico sobre o valor fenotípico, e é representada H^2.

$$H^2 = \frac{V_G}{V_F} \tag{1.4}$$

A seleção genômica possui um relevante papel, pois permite prever a expressão do valor futuro, ou do fenótipo, com base em dados genômicos ou genótipo. Nesse contexto, o correto cálculo do **F** de um indivíduo permite selecioná-lo no início da vida, trazendo economia e produtividade para os criadores. Esses dados são interessantes por possuírem

valores constantes ao longo da vida do animal. Dessa forma a extração do material genético e a qualidade do modelo de previsão são os pontos de importância no cálculo da expressão do fenótipo futuro ou previsto.

O principal objetivo da seleção dos melhores indivíduos é maximizar ou reduzir a expressão de uma determinada característica ao longo do tempo. A estimativa da expressão do fenótipo futuro é feita por meio de uma função que busca estimar o valor de **F** Figura 1.1. O uso de dados genômicos para a construção dessas funções deram origem ao processo conhecido como seleção genômica Figura 1.2. A característica em análise pode ser expressa por meio de um mapeamento linear ou não-linear, e sofrerá impacto do meio ambiente. Contudo, a construção de modelos eficientes para a avaliação do mérito genético de um indivíduo é complexa. Dessa forma na ausência de conhecimento genético detalhado a função deverá ser construída utilizando todo o conjunto de dados genéticos e fenotípicos disponíveis, o que também pode ser ineficiente.

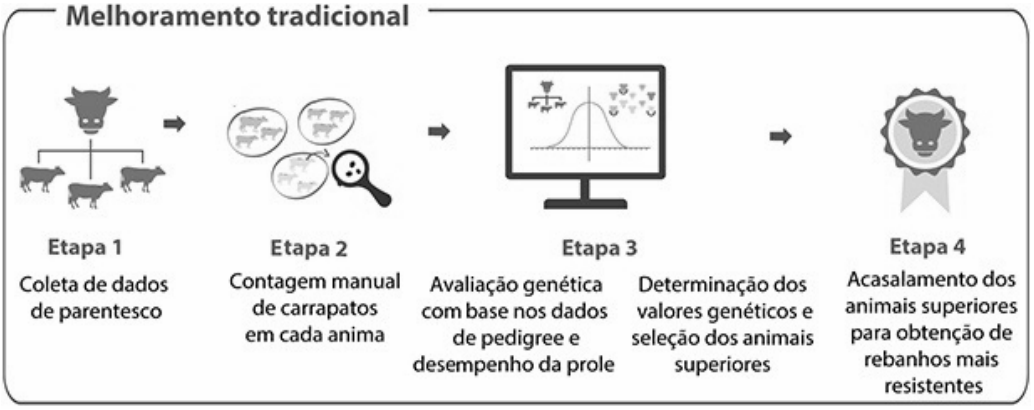


Figura 1.1: Método Tradicional para a Seleção Genômica - adaptada de Agropecuária (2017).

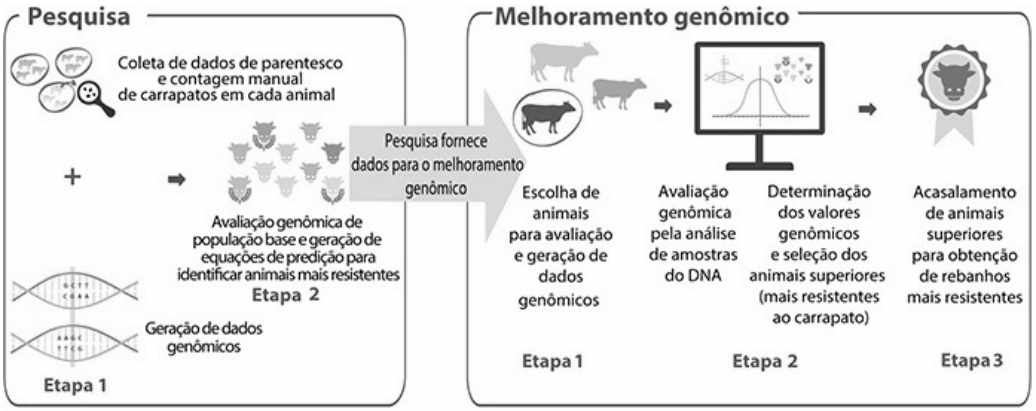


Figura 1.2: Melhoramento Genético por meio da Seleção Genômica - adaptada de Agropecuária (2017).

O valor genômico previsto (do inglês *genomic estimated breeding values* - GEBV) dos indivíduos da próxima geração é calculado tendo como base as equações de predição obtidas de uma população de referência, dessa forma é possível efetuar a seleção via dados genômicos (MEUWISSEN; HAYES; GODDARD, 2001; HAYES; GODDARD, 2010). A Figura 1.3 mostra, de forma simplificada, o processo de seleção genômica. O GEBV consiste no somatório dos efeitos de cada marcador, permitindo assim a predição do valor genômico da população futura ($GEBV = w_1x_1 + w_2x_2 + w_3x_3 \ldots$), sendo w o efeito do marcador e x o valor genômico dele (MEUWISSEN; HAYES; GODDARD, 2001). A redução do custo operacional para a obtenção dos marcadores aliada a precisão dos GEBV obtidos levaram a uma rápida adoção da seleção genômica (SCHAEFFER, 2006).

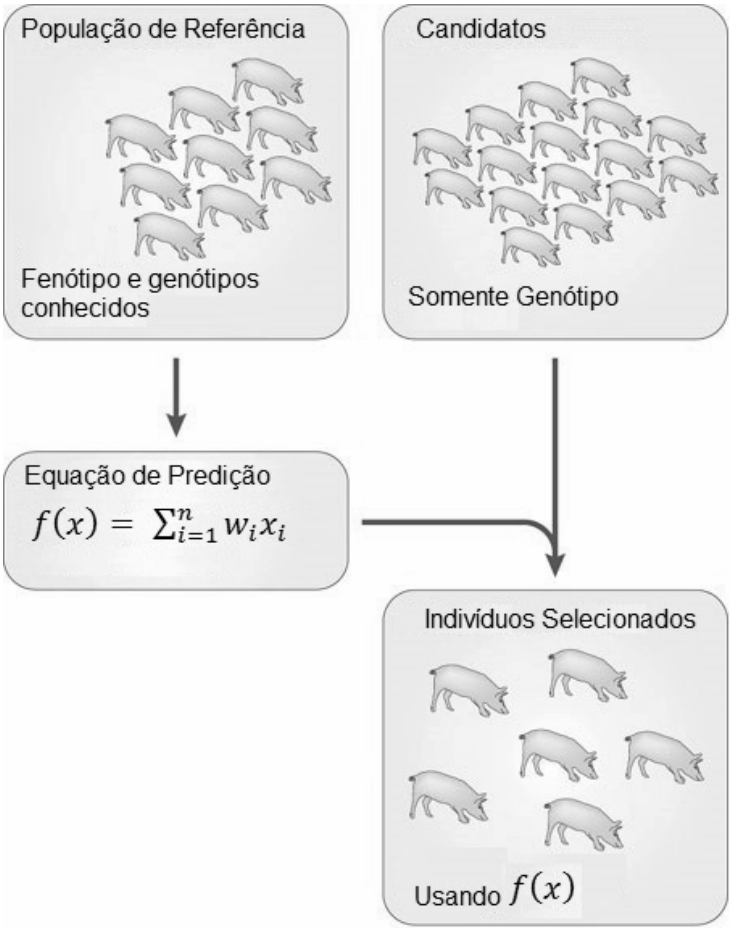


Figura 1.3: Seleção Genômica - adaptada:Goddard e Hayes (2009).

O sucesso na seleção genômica depende, em geral, de três itens: o tamanho da população de referência; a herdabilidade; e o tamanho do desiquilíbrio de ligação (do inglês *linkage disequilibrium* - LD[1]) entre os marcadores e o locus de características quantitativas

[1]O desequilíbrio de ligação, em genética populacional, é a associação não-aleatória de alelos em dois ou mais loci, não necessariamente no mesmo cromossomo (BROWN, 2006).

(do inglês *Quantitative Trait Locus* - QTL[2]) (GODDARD; HAYES, 2009). A seleção genômica para estimar com precisão o GEBV necessita, em geral, de um grande conjunto de dados para o treinamento conforme sugerem Meuwissen, Hayes e Goddard (2001), Hayes e Goddard (2010). A precisão do GEBV, segundo a fórmula de Daetwyler, Villanueva e Woolliams (2008), é diretamente proporcional à herdabilidade, onde traços com maior herdabilidade gera GEBVs mais precisos do que aqueles com herdabilidades menores. O aumento do desiquilíbrio de ligação entre os marcadores e o QTL, segundo Goddard e Hayes (2009), também geram uma maior precisão no cálculo do GEBV.

O conhecimento genético ainda é pequeno em face a crescente dimensão de informações biológicas que é descoberta. Dessa forma, os modelos estatísticos se tornam uma regra no processo de seleção genômica. As métricas utilizadas na confecção desses modelos se baseiam, em grande parte, no pressuposto da herança aditiva de Crow, Kimura et al. (1970) e Ewens (1977), bem como no teorema fundamental da seleção natural de Fisher (1919), Fisher (1930) e Robertson (1960). Entretanto, na presença de variabilidade genética não-aditiva, os modelos podem não comportar toda a complexidade da dominância ou de possíveis interações entre os genes, a epistasia (GALLAIS, 1974). Outro ponto de destaque é a presença de LD que pode dificultar a divisão da variabilidade genética, e as técnicas estatísticas existentes para efetuar corretamente essa divisão exigem conhecimento prévio da estrutura genética, sendo necessário para isso a identificação e associação dos genes ao fenótipo.

A identificação dos genes, ou marcadores e sua associação a uma determinada característica é também uma área de interesse da bioinformática. O seu objetivo é associar um limitado conjunto de marcadores a uma característica específico, aumentando assim o conhecimento e detalhamento biológico sobre os mecanismos envolvidos na confecção do fenótipo em estudo.

A correta associação de um limitado conjunto de marcadores a um fenótipo, contínuo ou discreto, é um linha de estudo completa e amplamente utilizada em trabalhos que visam identificar os genes associados a doenças em humanos, geralmente em problemas de caso controle. Entretanto, esse tipo de trabalho vem ganhando força no estudo de características em animais e plantas, onde em geral os fenótipos são contínuos.

O detalhamento dos mecanismos envolvidos na produção do fenótipo permitem a me-

[2]QTL são regiões do genoma relacionadas à variação das características quantitativas.

lhoria dos modelos utilizados em sua previsão. A melhoria associada ao aumento no conhecimento sobre determinada característica resulta em maiores acurácias no processo de previsão do fenótipo dos indivíduos. Contudo, o tempo despendido no estudo e detalhamento pode vir a ser elevado, principalmente se comparado com a expectativa do mercado em relação a melhora do modelo de previsão.

O aumento na qualidade e dimensão dos chips genômicos, bem como a exigência do mercado por melhorias no processo de seleção genômica impulsiona o desenvolvimento e aplicação de técnicas capazes de trabalhar com uma grande dimensão de dados, e também inferir de forma correta fenótipos expressos por meio de mapeamentos lineares e não-lineares do genótipo, sendo sua expressão discreta ou contínua. Cada conjunto de dados genotípicos e fenotípicos podem exigir um conjunto de técnicas dependendo da característica de interesse, fazendo-se necessário o conhecimento prévio da relação genótipo x fenótipo. Nesse contexto, as técnicas de inteligência computacional podem ser capazes de melhorar a identificação do mapeamento linear ou não-linear, mesmo desconhecendo o detalhamento biológico e tendo por base somente os dados amostrais do estudo.

A quantidade de indivíduos genotipados necessários no processo de seleção genômica é alto, segundo Meuwissen, Hayes e Goddard (2001), Hayes e Goddard (2010) para se obter uma acurácia de $0,8$ o valor mínimo é de 2.000 (duas mil) amostras, sendo maior para herdabilidades menores que 0,2. O número de amostra exigido pode ser difícil de se conseguir, seja devido a restrições orçamentárias, pequeno número de indivíduos, ou mesmo uma característica do problema como o pseudo-fenótipo. O valor do pseudo-fenótipo atribuído ao um animal consiste no valor médio do fenótipo de suas proles, aumentando assim a quantidade de indivíduos a serem avaliadas.

Os atuais chips utilizados no processo de seleção genômica possuem milhares de marcadores e um pequeno número de entradas. As técnicas mais comuns para se trabalhar com a atual dimensão de dados consiste no RR-BLUP eno BLASSO, ambas ferramentas de regularização e encolhimento. Os dois processos são similares e consistem em anular ou reduzir o efeito de um conjunto de marcadores, permitindo assim a criação de um modelo viável com os atributos restantes. Esse método pode ser visto como uma seleção de atributos, pois, apesar de não retirá-los do conjunto de dados, eles são ignorados por premissas da modelagem.

A seleção de atributos é uma técnica, muito utilizada em inteligência computacional,

que procurar reduzir a dimensionalidade do problema por meio da diminuição do número de variáveis e a manutenção do conhecimento. Cada modelo de seleção de atributos possui métodos próprios, de forma que os mecanismos envolvidos na expressão do fenótipo não são previamente conhecidos. A escolha das ferramentas de seleção e avaliação dos subconjuntos de marcadores, bem como a escolha dos critérios de parada de cada modelo são decisivos no resultado final. A ferramenta necessita ser capaz de trabalhar com uma grande dimensão de dados, ser útil em características lineares e não-lineares, na presença de interação entre as variáveis e principalmente ser robusta em relação ao número de indivíduos, mantendo a qualidade dos resultados mesmo em cenários com população considerada pequena.

1.1 Motivação

Diversos são os desafios encontrados no processo de seleção genômica, mas nesse trabalho objetiva-se abordar, inicialmente, dois tópicos: o impacto da redução do tamanho da amostra ou população de referência e a dificuldade da obtenção de um modelo para predição quando existe epistasia entre os alelos (GODDARD; HAYES, 2009; HAYES et al., 2009). A dificuldade em se obter um conjunto de dados, onde as associações génicas são conhecidas, incentiva o desenvolvimento de uma ferramenta de simulação, produzindo dados com comportamento próximo a dados reais e controlando, principalmente, a expressão gênica na produção do fenótipo.

Algumas populações podem não possuir o tamanho necessário para satisfazer os requisitos mínimos para a obtenção de resultados precisos, como os presentes em raças de grande porte. Deste modo, torna-se necessário o estudo do impacto das populações pequenas na seleção genômica. Porém, em estudo prévio, Mészáros et al. (2015) utilizou três bases de dados pequenas e discutiu o efeito da junção das mesmas, obtendo uma melhora de aproximadamente 10% em uma delas. Assim, nesse trabalho é investigada se a identificação e seleção dos atributos largamente informativos podem melhorar a precisão do GEBV. O estudo no impacto do uso de uma pequena amostra de dados é computacional, sendo que bons resultados podem ser obtidos utilizando amostra menores que 2000 indivíduos, sendo esse valor considerado mínimo para o trabalho com seleção genômica (GODDARD; HAYES, 2009).

O estudo com epistasia é definido por Hayes et al. (2009) como um dos desafios para a seleção genômica e tem sido utilizado para justificar uma série de fenômenos, tais como, a interação funcional entre os genes, o resultado de mutações genéticas que atuam dentro da mesma via metabólica e o desvio estatístico da ação aditiva (PHILLIPS, 2008). Desta forma, estudos recentes buscam adaptar métodos clássicos para serem utilizados em bases de dados com efeitos epistáticos (XU; JIA, 2007; WANG et al., 2012; HOWARD; CARRIQUIRY; BEAVIS, 2014). Nesse trabalho são aplicados métodos de inteligência computacional com o objetivo de verificar uma possível melhora na capacidade de generalização frente as bases de dados com efeitos epistáticos comparado aos métodos clássicos em seleção genômica.

A combinação desses diferentes desafios em um mesmo conjunto de dados tendem a dificultar bastante a obtenção de um GEBV mais preciso. Neste sentido, esse trabalho visa avaliar se a identificação e seleção de atributos largamente informativos, associado ao uso de técnicas de inteligência computacional, podem influenciar de forma positiva na predição do valor genômico, por meio da redução do número de marcadores e o aumento da acurácia do GEBV.

1.2 Objetivos

O objetivo primário desse trabalho é estabelecer e desenvolver um procedimento metodológico para predição de valor genômico com seleção de um conjunto informativo de marcadores moleculares do tipo SNP. Para que se possa alcançar esse objetivo, algumas etapas complementares são necessárias, a saber:

- Desenvolver uma ferramenta de simulação de dados genômicos que possuam interações epistática entre mais de dois genes, herdabilidade, LD e o cruzamento entre populações distintas.

- Avaliar, em populações com um número reduzido de amostra, o impacto da seleção de marcadores moleculares do tipo SNP, informativos e relacionados com a predição do valor genômico.

- Analisar o potencial, a adequação aos dados e, principalmente, os limites da aplicação de métodos estatísticos e computacionais modernos de inteligência computacional para seleção de atributos e predição do valor genômico;

1.3 Estrutura do Texto

O caráter multidisciplinar do trabalho gera a necessidade do suporte de conhecimentos de áreas diversas. Esse trabalho envolve conceitos das áreas de: biologia, zootecnia, genética, estatística, matemática e computação. Apresenta-se, a seguir, a estruturação do texto no que tange a configuração adotada para os capítulos.

O **Capítulo 1**, Introdução, apresenta uma breve introdução, resgata o conteúdo biológico do problema, apresentando os conceitos necessários para o seu entendimento, assim como objetivo do trabalho.

O **Capítulo 2**, Referencial Teórico, descreve as técnicas estatísticas e computacionais utilizadas para o desenvolvimento do método em estudo. Bem como uma breve revisão bibliográfica.

O **Capítulo 3**, Simulador de dados genômicos, descreve o seu desenvolvimento, a saber S4GS, como ferramenta de apoio aos estudos desenvolvidos nessa tese.

O **Capítulo 4**, Dados Simulados, é o local onde são apresentados os conjuntos de dados utilizados e a forma como foram simulados.

O **Capítulo 5**, Modelo Proposto, apresenta o método utilizado de forma a cumprir os objetivo estabelecidos, bem como as técnicas utilizadas na solução.

O **Capítulo 6**, Experimentos Computacionais, apresenta e discute os resultados dos experimentos realizados.

O **Capítulo 7**, Conclusão, são apresentados os comentários finais que levam à conclusão do trabalho e o direcionamento para a continuidade em trabalhos futuros.

2 Referencial Teórico

Neste capítulo é apresentada a revisão bibliográfica da aplicação de seleção de atributos em seleção genômica, bem como seus métodos mais comuns. Alem disso, apresentam-se os métodos utilizados no decorrer do trabalho e os conceitos biológicos necessários para o melhor entendimento dos problemas.

2.1 Conceitos Biológicos

2.1.1 Ação Gênica

A ação gênica consiste basicamente na forma com os genes atuam para a síntese de proteínas, regulação das ações das células entre outros, e podem possuir várias formas diferentes, influenciando diretamente na expressão dos fenótipos associados. Nessa seção são mostradas as ações relevantes para o entendimento do problema de interesse.

2.1.1.1 Ação Gênica Aditiva

A ação gênica aditiva é o tipo de ação onde cada um dos inúmeros genes que constituem o genótipo do indivíduo provocam um efeito na manifestação do fenótipo, independente da ação dos outros genes que constituem o genótipo. O efeito provocado pela ação independente de cada um dos genes adiciona-se aos efeitos dos demais na expressão do fenótipo (ELER, 2014). A ação aditiva é também chamada de ausência de dominância ou codominância.

Se uma característica é determinada por um par de alelos (loco A, por exemplo, com os alelos A_1 e A_2), na ação aditiva o valor expresso pelo genótipo heterozigoto (A_1A_2) é exatamente o valor médio dos genótipos homozigotos (A_1A_1 e A_2A_2), Figura 2.1.

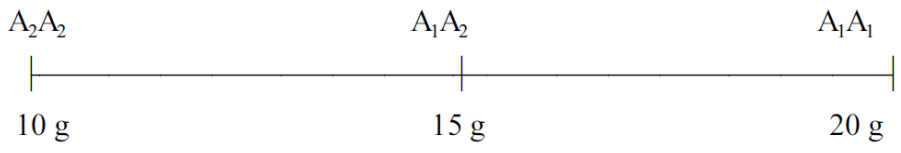


Figura 2.1: Exemplo de ação gênica aditiva com um alelo

A herança aditiva é o pressuposto básico do melhoramento genético animal. Se a expressão fenotípica dependesse apenas da ação aditiva, bastaria selecionar os indivíduos superiores, para mudar no sentido desejado. A modificação que ocorre no fenótipo corresponde à unidade gênica "introduzida" ou "retirada" do genótipo dos indivíduos. O problema se resumiria, assim, na avaliação do potencial genético ou do valor genotípico dos indivíduos e, então, se escolheriam os melhores para a reprodução (ELER, 2014).

2.1.1.2 Dominância

Os resultados dos experimentos de melhoramento de ervilhas feitos por Mendel (1866) mostraram que cada planta possui dois alelos por gene, exibindo, contudo, apenas um fenótipo. Isto é fácil de compreender se a planta é pura, ou homozigoto, para uma característica em particular, uma vez que, em seguida, possui dois alelos idênticos e exibe o fenótipo adequado (Figura 2.2 A). No entanto, Mendel mostrou que, se duas plantas puras com diferentes fenótipos são cruzadas, e toda a descendência apresentará os mesmos fenótipos observados na geração anterior, sendo conhecida como F_1. Logo as plantas F_1 são heterozigotos, o que significa que elas possuem dois alelos diferentes para cada fenótipo, um alelo herdado da mãe e outro do pai. Mendel postulou que nesta condição um alelo sobrepõe os efeitos do outro alelo de forma que o fenótipo expresso nas plantas F_1 são descritos como sendo dominante em relação ao segundo, que é recessivo (Figura 2.2 B), definindo assim o conceito de dominância e recessividade.

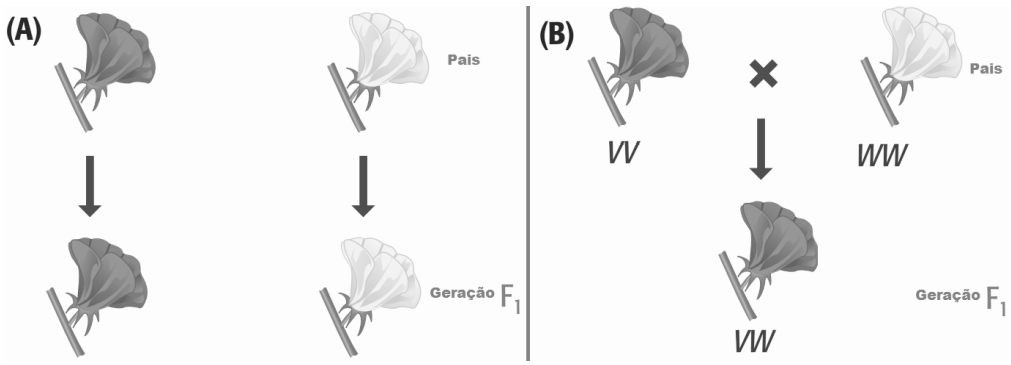


Figura 2.2: Mendel estudou sete pares de características em suas plantas de ervilha, uma das quais era violeta(cinza) e branco cor de flor, como mostrado aqui. (A) plantas pais que são puras sempre darão origem a flores com a cor dos pais. (B) Quando duas plantas reprodutores puras são cruzadas, apenas um dos fenótipos é visto na geração F_1. Mendel deduziu que o genótipo das plantas F_1 foi VW, então V é o alelo dominante e W é o alelo recessivo.

Mendel realizou experimentos adicionais que lhe permitiram estabelecer suas duas leis da genética. A primeira lei declara que os alelos segregam aleatoriamente, em outras palavras, se os alelos dos pais são A e a, em seguida, um membro da geração F_1 tem a mesma chance de herdar A como de herdar a. A segunda lei determina que pares de alelos segregam de forma independente, de modo que a herança dos alelos do gene A é independente da herança dos alelos do gene B. Devido a estas leis, os resultados de cruzamentos genéticos são previsíveis (BROWN, 2006).

2.1.1.3 Epistasia

A epistasia é caracterizada como a interação entre alelos de diferentes locus, com alteração conjunta na expressão fenotípica, podendo ou não estar no mesmo cromossomo. A ação epistática pode ser observada tanto em características qualitativas, quanto nas quantitativas. Um exemplo clássico deste tipo de ação gênica, para características qualitativas, é a cor da pelagem em cavalos. As cores preta e castanha são básicas e controladas por dois alelos, o B, dominante, que condiciona a cor preta e o b, recessivo, que condiciona a cor castanha. Portanto, os indivíduos BB e Bb são pretos enquanto que os bb são castanhos. Existe, no entanto um outro locus, W, que apresenta ação dominante e assim "mascara" a ação do locus B. O genótipo WW é inviável, morrendo na fase embrionária. O genótipo Ww impede a expressão do locus B e o genótipo ww não interfere na sua expressão, ver Tabela 2.1 (ELER, 2014).

Tabela 2.1: Genótipos e fenótipos para a pelagem dos cavalos sem o locus a (ELER, 2014).

Genótipo	Fenótipo
$wwBB$	Preto
$wwBb$	Preto
$wwbb$	Castanho
$WwBB$	Branco
$WwBb$	Branco
$Wwbb$	Branco

Na verdade, neste caso específico, ocorre ainda mais um locus de ação epistática, o A, que completa a determinação da cor da pelagem dos cavalos, condicionando ainda as cores baia e alazão, conforme Tabela 2.2 (ELER, 2014).

Em características quantitativas as combinações epistáticas podem ser numerosas. Na formação das raças, por exemplo, a seleção natural tende a fixar combinações epistáticas

Tabela 2.2: Genótipos e fenótipos para a pelagem dos cavalos com o locus A (ELER, 2014).

Genótipo	Fenótipo
$wwaaB-$	Preto
$wwA-B-$	baio
$wwA-bb$	alazão
$wwaabb$	castanho
$Ww----$	branco

favoráveis, melhorando o desempenho do animal, por isto, admite-se que nas diversas raças existem mais combinações favoráveis do que desfavoráveis.

2.1.2 Marcadores Moleculares do tipo SNP

Os marcadores moleculares para serem úteis devem existir em pelo menos duas formas ou alelos, especificando fenótipos diferentes. Existem três tipos de sequência que satisfazem este requisito: polimorfismos de comprimento de fragmentos de restrição (do inglês *restriction fragment length polymorphisms* - RFLPs), polimorfismos de comprimento de sequência simples (do inglês *simple sequence length polymorphisms* - SSLPs), e polimorfismos de base única (do inglês *single nucleotide polymorphisms* SNPs) (BROWN, 2006). Nesse trabalho os marcadores utilizados são do tipo SNP.

Os projetos de sequenciamento de genomas trouxeram muitas revelações para a ciência, sendo uma delas a descoberta, por meio do Projeto Genoma Humano, de que o código genético mostrou-se mais variado e complexo do que propriamente maior, quando comparado ao de outras espécies (CONSORTIUM et al., 2001).

Em geral, as "regras" que regem o estudo do genoma podem ser aplicadas a qualquer espécie viva, com diferenças apenas entre organismos procariotos e eucariotos. Uma das muitas variações e particularidades do genoma, humano ou de qualquer espécie, são os SNPs, modificações de um único nucleotídeo, em uma dada sequência, quando comparada a outra (Figura 2.3). Ou seja, SNPs são pares de bases em uma única posição no DNA genômico, que se apresentam com diferentes alternativas nas sequências, em uma fração significativa da população, sendo $\geq 1\%$, e podem ser encontrados no genoma de indivíduos normais em algumas populações ou grupos (HAPMAP, 2003). Geralmente, SNPs são encontrados em um vasto número e em qualquer genoma (BROWN, 2006).

Figura 2.3: Exemplo de um Polimorfismos de Base Única - SNP

A maior parte do genoma entre os indivíduos de uma mesma espécie é idêntica, porém existe a variabilidade genética que são as diferenças encontradas em algumas regiões do genoma (BRONDANI; BRONDANI, 2004). A variabilidade consiste na alteração em sequências de bases ao longo do DNA que ocorrem por substituição, ausência ou duplicação de bases e, os SNPs, essas diferenças pontuais entre pares de bases de diferentes sequências alinhadas, são o tipo mais comum de variabilidade genética (HAPMAP, 2003).

Assim, tais diferenças são importantes no estudo da variabilidade das espécies, pois podem provocar alterações funcionais ou fenotípicas, que, por sua vez, podem implicar em consequências evolutivas ou bioquímicas nos indivíduos em que os SNPs se manifestam (ARBEX, 2009).

2.2 Conceitos Computacionais Aplicados em Seleção Genômica

2.2.1 Seleção de Atributos

Em muitas tarefas de classificação ou regressão, o número total de possíveis atributos associados as instâncias que definem a base de dados é relativamente alta (STAŃCZYK; JAIN, 2015). Esta alta dimensionalidade tende a dificultar o processamento, ou até mesmo torná-lo impraticável. A presença de muitas variáveis é um desafio para muito dos indutores nas tarefas de classificação/regressão mesmo quando são atributos relevantes. Em caso de variáveis irrelevantes ou redundantes o processo de indução poderá ser comprometido (JOHN et al., 1994).

De maneira geral, os métodos para seleção de atributos podem ser agrupados em três grandes classes:

- Métodos baseados em filtro;
- Métodos encapsulados (do inglês, *wrapper*);
- Métodos embutidos (do inglês, *embedded*).

Os métodos com base em filtros trabalham de forma independente do classificador ou regressor envolvido no reconhecimento de padrões, independentemente das suas especificidades e de seus parâmetros (GUYON; ELISSEEFF, 2003). Essas abordagens podem ser tratadas como um procedimentos de pré-processamento. Elas exploram informações contidas no conjunto de dados de entrada buscando atributos que produzam ganho de informação, entropia ou consistência (DASH; LIU, 2003). Além disso, são particularmente efetivas em tempo computacional e robustos em relação ao *overfitting* (HAMON, 2013). A natureza geral dos filtros os tornam aplicáveis em todos os casos, no entanto, o fato de eles não considerarem um sistema de classificação que empregará o conjunto de variáveis selecionadas tendem a gerar resultados, geralmente, de qualidade inferior em relação às outras abordagens, o que pode ser uma desvantagem (STAŃCZYK; JAIN, 2015).

Métodos encapsulados avaliam subconjuntos de variáveis que permitem, diferentemente das abordagens de filtro, detectar possíveis interações entre elas em relação ao indutor adotado através de um procedimento de otimização. Apresentam custo computacional elevado porém, com grande efetividade nos resultados da seleção (PHUONG; LIN; ALTMAN, 2005). Espera-se uma melhor avaliação de alguns subconjuntos de variáveis devido a busca do subconjunto ótimo estar associado ao classificador de interesse (KOHAVI; JOHN, 1997). Desta forma, nessa abordagem um algoritmo de aprendizagem é usado para medir a qualidade de subconjuntos de variáveis na classificação ou regressão, e, portanto, otimizando a seleção em relação à ferramenta a ser utilizada para indução (LAL et al., 2006). Os métodos embutidos, por sua vez, usam características intrínsecas do próprio algoritmo de aprendizado para selecionar os atributos.

O ajustamento do processo de busca e seleção à característica específica do indutor pode gerar um viés, resultando em um aumento do desempenho do classificador escolhido, mas piores resultados para outro, especialmente quando eles variam significativamente em propriedades. Em outras palavras modelos encapsulados tendem a construir conjuntos de atributos que são ajustados, feitos sob medida para alguma tarefa especial e/ou algum sistema particular. Outra desvantagem desta abordagem é o custo computacional necessá-

rio. A execução do algoritmo de aprendizado para muitos subconjuntos de recursos pode se tornar inviável, não só quando há um número muito elevado de atributos a considerar, mas também nos casos em que o processo de aprendizado é complexo e consome tempo, mesmo para um número pequeno de variáveis (STAŃCZYK; JAIN, 2015).

Métodos embutidos diferem em relação ao modo como a seleção de variáveis e o processo de aprendizagem interagem (LAL et al., 2006). Como exemplo, Weston et al. (2000) mede a importância do atributo usando um limite que é válido somente para SVM, portanto, não é possível usar esse método com uma árvores de decisão, por exemplo.

Há também combinações de abordagens, onde, por exemplo, em uma primeira fase um filtro é utilizado, em seguida, um método encapsulado complementa o processo de seleção. Também é possível algoritmos específicos visando obter uma ordenação de atributos em relação à alguma medida de interesse, a qual será base para seleção ou redução de atributos que poderá ser executada posteriormente (STAŃCZYK; JAIN, 2015).

2.2.2 Valor-p

O valor-p pode ser considerada a probabilidade de se encontrar o valor observado ou mais extrema, sob uma hipótese nula - a definição de "extrema" depende de como a hipótese está sendo testada. A hipótese nula é geralmente a de "nenhuma diferença" sendo necessário sua definição antes do início do estudo (AGRESTI, 2007).

A hipótese alternativa (H_1) é o oposto da hipótese nula sendo, que esta é geralmente a hipótese a ser investigada. Por exemplo, para a questão: "há uma diferença significativa (não devido ao acaso) na pressão arterial entre os grupos A e B, sendo administrado ao grupo A uma droga de teste e ao grupo B uma pílula de placebo?" e a hipótese alternativa é: "existe uma diferença na pressão arterial entre os grupos A e B quando apresentado a um fármaco de ensaio e o outro a um comprimido de placebo".

Se o valor-P é menor que o nível de significância definido, então a hipótese nula é rejeitada ou seja, há amostras de provas razoáveis apoiando a hipótese alternativa. A escolha do nível de significância para rejeitar H_0 é geralmente empírica. Usualmente os valores 5%, 1% e 0,1% (P $<$ 0,05, 0,01 e 0,001) são níveis que podem ser utilizados. Contudo a adoção destes valores podem induzir a resultados imprecisos e uma falsa sensação de segurança (AGRESTI, 2007; AGRESTI, 2010).

A maioria dos autores refere-se que estatisticamente significativa é $P < 0{,}05$ e alta-

mente significativa como $P < 0{,}001$ contudo Welter et al. (2014) refere-se a $P< 10^{-5}$ com um valor significante para estudo em GWAS.

2.2.3 BLASSO

O método de regressão LASSO (*Least Absolute Shrinkage and Selection Operator*, (TIBSHIRANI, 1996)) combina a seleção de variáveis com a regularização via redução dos coeficientes de regressão. A implementação Bayesiana da regressão LASSO (PARK; CASELLA, 2008) foi adaptada para seleção genômica por Campos et al. (2009). Nesta adaptação, informações de parentesco e outras covariáveis que não sofrem o efeito da regularização são consideradas no modelo.

A metodologia LASSO consiste na obtenção de estimadores de coeficientes de regressão que resolvam o problema de otimização, $min\{\Sigma_i(y_i - x_i\beta)^2 + \lambda(t)\Sigma_j|\beta_j|\}$, onde y_i é o numero de observações do indivíduo i; β é o vetor de coeficientes de regressão; e x_i é um vetor das covariáveis; $\sum_j |\beta_j|$ é a soma dos valores absolutos dos coeficientes de regressão contidos no vetor β, de modo que soluções nas quais os coeficientes de regressão se afastam de 0 sofrem penalização. Adicionalmente, λ é um parâmetro de suavização que controla a força da regularização. Quando este último parâmetro é igual a zero, não há regularização. No LASSO Bayesiano, esse parâmetro controla a precisão da distribuição *a priori* atribuída aos coeficientes de regressão.

A implementação desse tipo de regularização envolve uma redução mais forte no sentido de que alguns coeficientes da regressão tenham valores iguais a zero, o que pode ser demonstrado de diversas maneiras. Uma alternativa é pela própria implementação bayesiana do LASSO (CAMPOS et al., 2009). Ela impõe como distribuição *a priori* dos p coeficientes da regressão um produto de densidades exponenciais duplas, dada pela Equação 2.1.

$$p(\beta|\lambda) = \Pi_{j=1}^{p} \frac{\lambda}{2} exp(-\lambda|\beta_j|) \tag{2.1}$$

Por sua vez, a regressão bayesiana por padrão utiliza uma distribuição normal multivariada, dada pela Equação 2.2

$$p(\beta|\sigma_\beta^2) = \Pi_{j=1}^{p} \frac{1}{\sqrt{2\pi\sigma_\beta^2}} exp(-\frac{\beta_j}{2\sigma_\beta^2}) \tag{2.2}$$

As duas distribuições podem ser comparadas na Figura 2.4, na qual se observa que a densidade *a priori* utilizada no LASSO Bayesiano (curva sólida) apresenta maior massa de densidade no valor zero e caudas mais robustas, exercendo maior redução nos coeficientes de regressão próximos a zero e menor nas pontas, Figura 2.4.

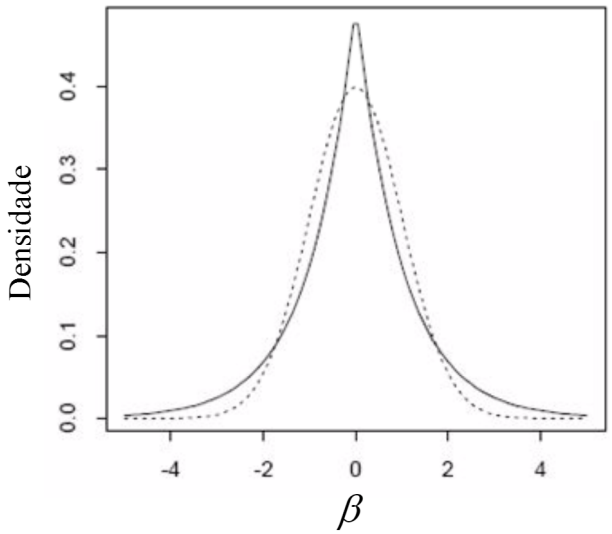


Figura 2.4: Densidades das distribuições normal (curva pontilhada) e exponencial dupla (curva sólida), ambas com médias iguais a zero e variâncias iguais à unidade.

O modelo proposto por Campos et al. (2009) é apresentado na Equação 2.3,

$$y_i = \mu + x'_{ri}\beta_r + x'_{li}\beta_l + e_i \tag{2.3}$$

onde y_i é o fenótipo mensurado no indivíduo i; μ é a média da característica estudada; x'_{ri} e x'_{li} são as covariáveis atribuídas ao individuo i a serem tratadas pela regressão bayesiana padrão e o LASSO bayesiana, respectivamente; β_r e β_l são os coeficientes da regressão bayesiana-padrão e do LASSO bayesiana, respectivamente; e e_i é o resíduo aleatório do modelo. Adicionalmente, será considerado que $e \sim N(0, \sigma_e^2)$.

Para construção da distribuição conjunta *a priori* dos parâmetros, os autores exploram o fato de a distribuição exponencial dupla poder ser representada como uma mistura de densidades normais com parâmetro de escala, com o processo de mistura de variâncias controlado por distribuição exponencial. Na distribuição *a priori* conjunta construída, a densidade atribuída aos coeficientes de regressão regularizados por LASSO será: $\Pi_{j=1}^{p} N(\beta_{j1}|0, \sigma_e^2\tau_e^2)$ resultando em variâncias específicas para cada coeficiente de regressão. Por sua vez, a distribuição *a priori* para o parâmetro de escala τ_j^2 é representado por: $\Pi_{j=1}^{p} = exp(\tau_e^2|\lambda)$, pela qual o parâmetro de suavização λ influência o ajuste dos coeficien-

tes de regressão. A informação *a priori* para esse parâmetro é dada por uma distribuição com hiperparâmetros conhecidos. Caso a distribuição escolhida seja conjugada, é possível obter amostras da distribuição *a posteriori* conjunta por meio de um amostrador de Gibbs (CAMPOS et al., 2009).

2.2.4 RR-BLUP

O método utiliza os preditores do tipo BLUP (*Best Linear Unbiased Prediction*) (GOLDBERGER, 1962), onde os efeitos dos marcadores não são ajustados como variáveis classificatórias, mas como explicativas ou explanatórias. Essas são variáveis de regressão ajustadas como covariáveis de efeitos aleatórios, ou seja, os fenótipos são regredidos com base nessas covariáveis. O nome mais apropriado é Regressão de Cumeeira (*Ridge Regression*) do tipo BLUP (RR-BLUP) aplicada à Seleção Genômica Ampla (RESENDE et al., 2010).

O ajuste para estimar os efeitos dos marcadores utiliza uma regressão linear mista conforme a Equação 2.4

$$y = Xb + Zm + e \tag{2.4}$$

onde y é um vetor de observações fenotípicas; b é um vetor de efeitos fixos; m é o vetor de efeitos dos marcadores assumidos como aleatórios; e se refere ao vetor de erros aleatórios; e X e Z são as matrizes de incidência para b e m, respectivamente. A matriz de incidência Z contém os valores dos alelos.

As equações de modelo misto para predição de m por meio do método RR-BLUP é dada pela Equação 2.5

$$\begin{bmatrix} X'X & X'Z \\ X'Z & Z'Z + I\dfrac{\sigma_e^2}{\sigma_g^2/n} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \tag{2.5}$$

onde σ_g^2 se refere à variância genética da característica, σ_e^2 à variância residual e n é função do número total de marcadores ponderados por suas frequências alélica, sendo dado por $n = 2\sum_i p_i(1 - P_i)$, em que p_i é a frequência do alelo i. Dessa forma considera-se que cada loco explica $(1/n)\sigma_g^2$, ou seja, partes iguais da variância genética são atribuídas a todos os locus.

O GEBV do indivíduo j é dado por:

$$GEBV = \hat{y} = \hat{u} + \sum_i Z_i \hat{m}_i \tag{2.6}$$

onde Z_i equivale a 0, 1 ou 2 para os genótipos dos tipos aa, Aa e AA, respectivamente, para marcadores bialélicos e codominantes como os SNPs

2.2.5 Regressão com Máquina de Vetores Suporte (Support Vector Regression - SVR)

A Máquina de Vetores Suporte (*Support Vector Machine* - SVM) é uma técnica de aprendizado supervisionado que analisa padrões entre os dados de entrada, caraterizados por variáveis numéricas contínuas ou discretas, com os dados de saída designados por um atributo dicotômico (problema de classificação). Esse modelo foi desenvolvido por Cortes e Vapnik (1995) e é baseado na ideia de encontrar o hiperplano ótimo que separa as duas classes por meio da maximização da margem.

A primeira versão do SVM com regressão foi proposta em 1997 por Drucker et al. (1997), e foi denominada como SVR. Dentre as vantagens do SVR, vale citar que este método não pressupõe linearidade do modelo, desde que se adote função *kernel* [1] não-linear, não necessita de normalidade dos resíduos e adapta-se facilmente a dados de alta dimensionalidade (número de instâncias menor que o número de atributos).

Smola e Schölkopf (2004) demonstram que uma premissa no modelo expresso pela função objetivo 2.7 e pelas restrições 2.8 é indicativo de que existe a função f que aproxima todos os pares (x_i, y_i) com uma precisão ε, ou seja, o problema de otimização convexo é viável. Entretanto, algumas vezes, tal problema pode ser inviável, ou até mesmo, pode-se permitir alguns erros superiores à margem definida pela precisão ε. Deste modo, para flexibilizar o modelo anterior a aceitar erros superiores à ε, são introduzidas as variáveis de folga ξ_i e ξ_i^*. Com isso, obtém-se a formulação proposta por Vapnik (1995) denotada pelas Expressões 2.9 e 2.10.

$$\text{Minimizar } Z(w, b) = \frac{1}{2}\|w\|^2 \tag{2.7}$$

[1] Um *kernel* é uma função K tal que para todo $\mathbf{x}, \mathbf{z} \in X$ satisfaz $K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$ onde ϕ é uma função de X para um espaço de características com produto interno F, onde $\phi : \mathbf{x} \mapsto \phi(\mathbf{x}) \in F$.

$$\begin{cases} y_i - \langle w, x_i \rangle - b & \leq \quad \varepsilon \\ \langle w, x_i \rangle + b - y_i & \leq \quad \varepsilon \end{cases} \tag{2.8}$$

$$\text{Minimizar } Z(w, b, \xi_i, \xi_i^*) = \left[\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \right] \tag{2.9}$$

sujeita às restrições:

$$\begin{cases} y_i - \langle w, x_i \rangle - b & \leq \quad \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq \quad \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq \quad 0 \end{cases} \tag{2.10}$$

Outra forma de escrever a Equação 2.9 é mostrada em 2.11.

$$\text{Minimizar} \left[\frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} L_\varepsilon(f(x_i), y_i) \right] \tag{2.11}$$

Assim, a função de perda ε-insensível é definida como indicado em 2.12.

$$L_\varepsilon(f(x_i), y_i) = \begin{cases} 0 & \text{se } |y_i - f(x_i)| \leq \varepsilon; \\ |y_i - f(x_i)| - \varepsilon & \text{se } |y_i - f(x_i)| > \varepsilon. \end{cases} \tag{2.12}$$

Por outro lado, de acordo com a Expressão 2.11, o termo $\frac{1}{2}\|w\|^2$ indica a complexidade do modelo e o termo $L_\varepsilon(f(x_i), y_i)$ traduz a função de perda ε-insensível que penaliza somente os valores fora do tubo, ou seja, com erros maiores que ε. Já o parâmetro C é chamado de constante de regularização e traduz o equilíbrio entre a complexidade de f e a quantidade de desvios maiores do que ε que podem ser tolerados (ÜNSTü; MELSSEN; BUYDENS, 2006). Assim, quanto menor o tubo (menor ε), mais complexa é a função f e, de forma contrária, quanto maior o tubo (maior ε), menos complexidade é necessária para f. A função de perda ε-insensível com SVR linear é mostrada na Figura 2.5.

Segundo Ünstü, Melssen e Buydens (2006), com a introdução de variáveis de folga ξ_i e ξ_i^* e devidas manipulações algébricas, as Expressões 2.7 e 2.8 se transformam na função objetivo 2.9 e nas restrições 2.10. Tal formulação é chamada de primal, pois a regressão é baseada no espaço original dos dados. Já as variáveis de folga têm por objetivo possibilitar a ocorrência de vetores fora do tubo, sendo os mesmos chamados vetores suporte, pois são somente eles que contribuem para a regressão. Desta forma, todos os outros vetores dentro do tubo podem ser removidos após a construção do modelo. Essa propriedade permite

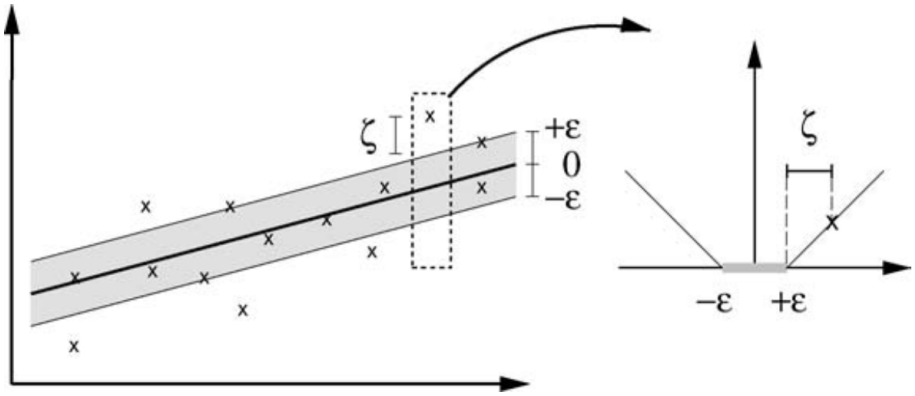


Figura 2.5: Função de perda com margem flexível no SVR linear (adaptado de Smola e Schölkopf (2004)).

que o SVR modele relações em que o número de variáveis dependentes seja superior ao número de instâncias na amostra de treinamento.

No caso dos padrões de entrada x_i não possuírem relação linear com a variável dependente designada pelos valores y_i, a função f do modelo primal é reformulada para o modelo dual como mostra a Equação 2.13. Com isso, o espaço original é mapeado para um novo, denominado espaço de características, por meio da função ϕ e do produto interno $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$, sendo K chamada de função *kernel*. Esta função traduz a relação subjacente entre os dados de entrada e os dados de saída.

$$f(x) = \left[\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \langle \phi(x_i), \phi(x_j) \rangle \right] + b \tag{2.13}$$

As variáveis duais α_i e α_i^* representam os multiplicadores de Lagrange que satisfazem as desigualdades e que podem ser obtidos pela e pelas Equações 2.14 e 2.15.

$$\begin{aligned} \text{Maximizar } Q(\alpha_i, \alpha_i^*) = \Bigg\{ & -\frac{1}{2} \left[\sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(x_i, x_j) \right] \\ & - \varepsilon \left[\sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \right] + \left[\sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*) \right] \Bigg\} \end{aligned} \tag{2.14}$$

sujeita às restrições:

$$\begin{cases} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) & = & 0 \\ \alpha_i, \alpha_i^* & \in & [0, C] \end{cases} \tag{2.15}$$

Este procedimento é chamado de expansão de vetores suporte, isto é, w pode ser determinado por uma combinação linear dos padrões de treinamento x_i. Com essa observação, conclui-se que a representação da complexidade de uma função por vetores suporte é independente da dimensionalidade do espaço de entrada $\mathcal{X}$, dependendo somente do número de vetores suporte (SMOLA; SCHÖLKOPF, 2004).

Uma primeira forma para tratar do SVR não linear seria realizar um mapeamento dos dados de entrada para o espaço de características a partir da função $\phi : \mathcal{X} \mapsto F$, e, em seguida, aplicar o SVR linear padrão nos dados transformados. Com isso, a linearidade da regressão é obtida no espaço de características e não no espaço original como pode ser notado na Figura 2.6. Entretanto, esses dois passos em problemas com uma grande quantidade de dados de treinamento torna o SVR computacionalmente inviável. Para superar essa dificuldade são escolhidas funções *kernel* que podem ser escritas em função dos dados de treinamento. Com isso, o cálculo do produto interno entre os vetores transformados do espaço de característica não é mais necessário, sendo feito implicitamente pela função *kernel*, que utiliza como base os dados de treinamento em um único passo.

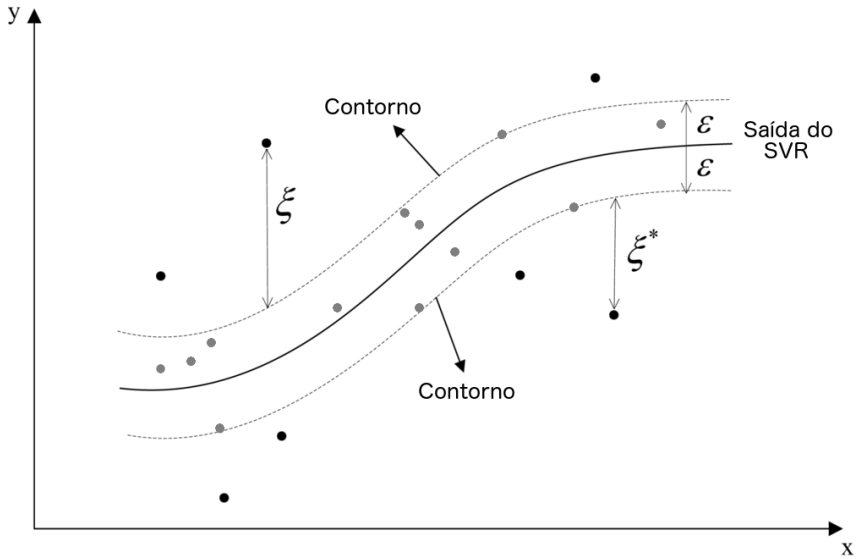


Figura 2.6: Regressão com *kernel* não linear com função de perda ε-insensível - os pontos em preto são os vetores suportes (adaptado de Ma, Song e Xiao (2012)).

Caso os parâmetros do SVR/SVM sejam otimizados, tem-se uma efetividade no ajuste do modelo aos dados e na generalização das predições, como é discutido em Ünstü, Melssen e Buydens (2006).

2.2.6 SMS

O método *SNP Markers Selector* (SMS), cuja tradução livre é Seletor de Marcadores SNP, foi desenvolvido pelo grupo de pesquisa onde está inserido este trabalho. Sua primeira versão foi publicada por Oliveira et al. (2014b), Oliveira et al. (2014a). A versão utilizada nesse trabalho corresponde ao trabalho final de Oliveira (2015).

O método busca combinar, de forma otimizada, técnicas da inteligência computacional, e utiliza: Floresta Aleatória (do inglês, *Random Forests - RF*), Máquinas de Vetores Suporte (do inglês, *Support Vector Machines - SVM*) e Algoritmos Genéticos (do inglês, *Genetic Algorithms - GA*). Cada técnica foi combinada de forma a se obter o máximo de eficiência em cada etapa. O SMS combina a seleção de atributos por filtro e por encapsulamento em etapas distintas e complementares.

A primeira etapa utilizada é o filtro, por ser menos custoso, e consiste em utilizar a RF para ordenar os marcadores por sua relevância em relação ao fenótipo. Em seguida é utilizado o SVM/SVR que incrementa o conjunto de teste de n em n elementos (em geral $n = 10$), selecionando o subconjunto com menor erro. Essa primeira seleção só é possível pois, em geral, o número de marcadores causais é bem menor que o sequenciado. A segunda seleção é um encapsulamento combinando o AG com o SVM/SVR, onde o AG seleciona o melhor subconjunto seguindo a avaliação feita pelo SVM/SVR. O resultado final do SMS é o menor subconjunto de marcadores com a maior informação relativa ao fenótipo estudado.

O método desenvolvido possui as seguintes etapas, conforme Figura 2.7.

Detalhadamente, os SMS é composto dos seguintes passos:

1. Executa-se o modelo da RF sobre todos os marcadores de cada cromossomo em conjunto com o fenótipo avaliado para se obter o ordenamento decrescente de importância dos marcadores. Essa importância é mensurada pelo acréscimo percentual do MSE, para problemas de regressão onde o fenótipo é contínuo, ou pelo decréscimo percentual da acurácia para problemas de classificação com fenótipo binário.

2. Constrói-se uma sequência crescente de subconjuntos de marcadores usando o ranque gerado pela RF com passo igual a 10 marcadores. A cada interação são adiciona mais 10 ao subconjunto atual gerando assim um novo. Esse procedimento é repetido para cada cromossomo.

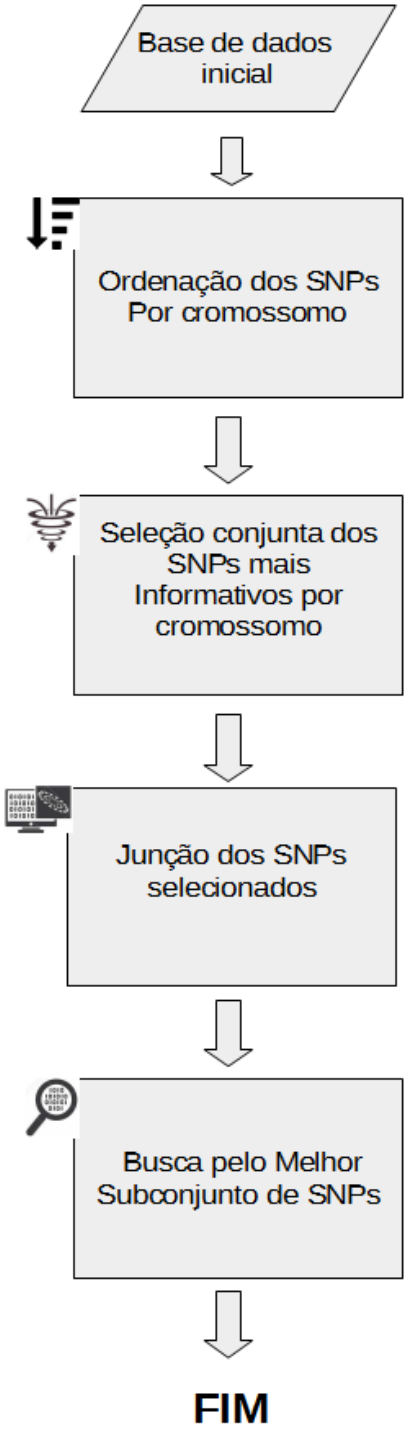


Figura 2.7: Fluxograma do SMS.

3. Implementa-se o SVR (regressão) ou o SVM (classificação) com validação cruzada com 10-partes sobre cada um dos subconjuntos gerados no passo 2 e avalia-se a correlação para regressão ou a média da AUC-ROC para classificação de cada um dos modelos construídos. Esse passo é também realizado para cada cromossomo.

4. Calcula-se a maior média da correlação ou a maior média da AUC-ROC, dependendo do tipo de problema (regressão ou classificação), obtendo-se o grupo de marcadores que maximiza a correlação média ou a AUC-ROC média. Como existe a possibilidade de imprecisão na correlação média do grupo avaliado, acrescenta-se uma margem de erro no índice deste subconjunto, permitindo a entrada de mais marcadores. Com isso, aumenta-se a probabilidade da entrada de verdadeiros positivos que não estão no grupo com correlação máxima ou com AUC-ROC máxima, mas em algum outro grupo que contém este. É importante ressaltar que esse artifício permite também a entrada de marcadores falso positivos, entretanto, o objetivo principal dessa etapa é maximizar o número de verdadeiros positivos na amostra. Ao final constrói-se uma base de dados intermediária formada pela união de todos os SNPs selecionados por cromossomo juntamente com o fenótipo. Esse passo representa o primeiro filtro do método SMS.

5. Na base de dados intermediária, aplica-se um GA para a selecionar o subconjunto final de marcadores. A última seleção com uso do GA objetiva maximizar a média da correlação de Pearson para o SVR ou a média da AUC-ROC do SVM, sendo ambas as médias geradas pela validação cruzada com 10-partes.

2.2.7 Árvores de Classificação e Regressão

As árvores de regressão e classificação (do inglês - *Classification And Regression Tree - CART*) fazem uso de dados históricos para construir as chamadas árvores de decisão. A metodologia CART foi desenvolvida por Breiman et al. (1984). A construção da CART é feita por meio da amostra de aprendizagem.

As CARTs são métodos de inteligência computacional para a construção de modelos de predição a partir dos dados. Os modelos são obtidos dividindo recursivamente o espaço de dados e ajustando um modelo de predição simples dentro de cada partição. Como resultado, o particionamento pode ser representado graficamente como uma árvore de decisão. As árvores de classificação são projetadas para variáveis dependentes que tomam um número finito de valores não ordenados, com previsão de erro medido em termos de custo de erros de classificação. As árvores de regressão são para variáveis dependentes que tomam valores discretos contínuos ou ordenados, com erro de previsão tipicamente medido pela diferença quadrática entre os valores observados e preditos.

Os algoritmos C4.5 e CART são utilizados na construção de árvores de classificação. O C4.5 utiliza a entropia para a sua função de impureza, enquanto a CART utiliza uma generalização da variância binomial denominada índice de Gini. A CART utiliza validação cruzada com 10-partes, e a C4.5 uma fórmula heurística para estimar taxas de erro.

A árvore de regressão utilizada foi implementada em R utilizando o pacote RPART (*Recursive Partitioning and Regression Trees* - Árvores de regressão e Particionamento recursivo) (THERNEAU; ATKINSON; RIPLEY, 2015).

2.3 Revisão Bibliográfica Relativa à Seleção de Atributos em Seleção Genômica

A seleção de atributos visando o aumento da acurácia em seleção genômica já foi utilizada por diferentes autores, seja em um contexto de classificação ou redução na dimensão dos dados. Long et al. (2007) apresentam uma técnica de seleção dos SNPs mais relevantes por meio de um processo de adição à frente. O trabalho apresenta um problema de classificação e busca identificar os SNPs relacionados à morte prematura de frangos. Os autores ressaltam a necessidade de técnicas que busquem interação entre os marcadores, pois a ferramenta de avaliação utilizada é linear e não contempla esse cenário.

O trabalho apresentado por Verbyla et al. (2009) consiste na aplicação de técnicas de seleção de atributos por meio da anulação do efeito dos marcadores irrelevantes, sendo elas: Bayes, Bayes B, BLUP e a seleção de variáveis por busca estocástica (do inglês *stochastic search variable selection* - SSVS). Os métodos foram aplicados em um problema característico de seleção genômica, os dados utilizados continham 1498 touros Holstein-Friesian australianos nascidos entre 1940 e 2000 e genotipados utilizando o chip *Illumina BovineSNP50K* com 39.048 marcadores após o controle de qualidade, contendo também 5 fenótipos, sendo eles: proteína em quilos, quilos de gordura, porcentagem de proteína, porcentagem de gordura e fertilidade das filhas. Os métodos foram validados utilizando um segundo conjunto com 400 touros nascidos entre 2005, 2006 e 2007. Os autores utilizaram técnicas lineares para a escolha dos atributos, sendo que a ferramenta de seleção estocástica SSVS obteve excelentes resultados, melhorando a acurácia por meio da seleção de atributos.

O uso da seleção de atributos nem sempre resulta em aumento de acurácia. O trabalho

de Moser et al. (2009) faz o comparativo entre cinco métodos distintos,a saber: regressão fixa utilizando mínimo quadrados (do inglês *fixed regression using least squares* - FR-LS), RR-BLUP, regressão Bayseana (Bayes-R), regressão de mínimos quadrados parciais (do inglês *partial least squares regression* - PLSR) e máquina de vetor e suporte com regressão (do inglês *support vector regression* - SVR). Os métodos foram apresentados a dados reais contendo 1945 touros e dois fenótipos, porcentagem de proteína e índice de rendimento (*Australian Selection Index*, ASI). O método FR-LS executa uma seleção por meio da adição pra frente onde o critério de escolha é o valor-p e a avaliação é feita por meio de uma regressão linear. O método mais acurado foi o SVR, que foi apresentado ao conjunto completo dos dados. A ferramenta de seleção de atributos pode ter sido impactada pelo uso de uma técnica linear para a escolha dos melhores marcadores.

O trabalho de Solberg et al. (2009) apresenta um comparativo entre as técnicas de redução de dimensionalidade PLSR, regressão via componentes principais (do inglês, *principal component regression* - PCR) e o método de predição BayesB. Os autores concluem que a seleção de atributos não foi eficiente para o aumento da acurácia nos conjuntos de dados utilizados. A única vantagem apresentada pelos autores foi o tempo computacional baixo necessário para executar as técnicas de PCR e PLSR, principalmente se comparada com o BayesB. Entretanto, o método BayesB anula o efeito de alguns marcadores podendo ser entendido também como uma ferramenta de seleção de atributos.

O uso de técnicas de seleção se mostrou vantajosa no trabalho de Long et al. (2011), que utilizou o PLS e o PCR. O uso da seleção de atributos aumentou a acurácia no valor genômico médio obtido por meio de regressão. Segundo os autores o uso combinado de técnicas redução da dimensionalidade por meio da escolha dos melhores atributos, associadas a métodos precisos de regressão podem aumentar a precisão na previsão genômica.

Qiu et al. (2016) comparam cinco diferentes técnicas para a classificação de milho sendo eles: RR-BLU; BayesA; BayesB; BayesC; e Random Forest (RF). Os autores selecionaram os marcadores utilizando a ordenação da RF escolhendo os melhores 50, 100, 1000, assim sucessivamente ate o total de marcadores. Apesar dos autores não discutirem a qualidade da seleção, o trabalho demonstra uma melhora na regressão quando aplicado a seleção de atributos.

A seleção de atributos pode ser lenta ou computacionalmente custosa, sendo esse um outro enfoque do trabalho. Liu et al. (2017) apresentam um novo índice baseado em SVM

visando minimizar o custo computacional da seleção de atributos. A alteração do índice utilizado pelo SVM para que o mesmo possa efetuar a seleção de atributos durante o processo de regressão também foi alvo da pesquisa de Yao et al. (2017).

3 S4GS: Simulador para Seleção Genômica

O objetivo desse capítulo é apresentar o simulador para seleção genômica (***S**imulator for **Ge**nomic **S**election* - S4GS), suas principais características, processo de geração das populações e do fenótipo, bem como alguns cenários para análise dos parâmetros de entrada. A construção do S4GS permitiu um melhor desenvolvimento do trabalho, pois atende a todas as necessidades iniciais de forma satisfatória. Outros simuladores da literatura possuem somente parte dos requisitos esperados.

O uso de dados simulados permite a avaliação e escolha de novos modelos, com destaque à aplicação em seleção genômica que é objeto de estudo dessa tese. A simulação de dados permite o controle da forma como o fenótipo será expresso, bem como a atuação das diferentes ações gênicas e forças evolutivas. As ações gênicas mais comuns são: aditiva, dominância e epistasia. As forças evolutivas são várias, mas nesse trabalho foram utilizados: mutação, cruzamento aleatório, seleção dos mais aptos e recombinação. O uso em conjunto das forças evolutivas e ações gênicas busca aproximar a simulação do cenário real, um fator importante para a qualidade estudo.

A simulação de dados genéticos de animais difere dos aspectos necessários para a simulação em seres humanos. O trabalho de Andersson (2001) apresenta algumas características presentes em populações de gado como: genealogias multigeracionais, famílias muito grandes, uso de inseminação artificial bem como o cruzamento controlado e planejado. O LD em gado se estende por distâncias maiores que nos seres humanos, uma característica que pode ser explicada pelo aumento do tamanho efetivo da população em humanos e sua diminuição nos animais (FARNIR et al., 2000). Essas características combinadas geram uma forte estrutura familiar que é utilizada no mapeamento de LD e QTL (MEUWISSEN; GODDARD, 2000).

O LD é uma característica importante no estudo de seleção genômica devido a sua relevância em bases de dados reais. O LD é um pressuposto em seleção genômica, sendo objeto de estudo para vários autores (HILL; ROBERTSON, 1968; LEWONTIN, 1988; MEUWISSEN; GODDARD, 2000; REICH et al., 2001; FARNIR et al., 2002; MCKAY et

al., 2007; ROOS et al., 2008).

Os fenótipos avaliados devem ser contínuos ou divididos em duas ou mais classes. O NIH (2016) lista alguns simuladores para dados genômicos, contemplando dos mais simples aos mais complexos, porém grande parte geram somente fenótipos dicotômicos, ou seja, com duas classes. O foco do estudo desse trabalho é com uso de variáveis contínuas com interação entre marcadores. Entre os simuladores disponíveis quatro foram selecionados para estudo de suas características principais, seja devido a facilidade de uso ou sua ocorrência em outros estudos da mesma área, sendo eles: o SCRIME (SCHWENDER, 2007), o QTLCart - (WANG; BASTEN; ZENG, 2007), o QMSIM (SARGOLZAEI; SCHENKEL, 2009) e o LDSO (YTOURNEL et al., 2012). Cada qual com suas particularidades construtivas com especificidades e variabilidades que tornam complexa a tarefa de simulação que abranja diversos aspectos gênicos.

O SCRIME trabalha com interação entre os marcadores, e definição dos SNPs causais, porém não simula LD, nem o cruzamento entre pares e o processo geracional. O QTLCart possui interface gráfica simples de utilizar, entretanto o cruzamento simula o comportamento observado em plantas e para simular o cenário de cruzamento de animais exige a configuração por meio de linhas de comando em um processo mais complexo. O QMSIM possui exemplos fáceis de serem reproduzidos e alterados, contudo simula somente o efeito aditivo. O LDSO simula uma variedade de ações gênicas e forças evolutivas, contudo simula somente interação entre 2 marcadores e não permite a definição dos SNPs causais, sendo também de uso de configurações mais complexas que os anteriores.

Desta forma, o S4GS visa associar, de forma simplificada, as melhores características observadas em cada um desses simuladores, tais como: facilidade de uso, exemplos fáceis de serem replicados, simulação de múltiplas ações gênicas, definição dos SNPs causais, controle personalizado da forma de cruzamento e inseminação artificial. As ações gênicas modeladas foram a aditiva, dominância e a epistasia em diferentes níveis. As forças evolutivas a serem implementadas para a simulação prevista são: a mutação, a seleção dos pais, a recombinação, o cruzamento entre populações e a opção de uso de um banco de sêmen. Todas essas funções devem estar associadas a um uso simplificado e integrado ao software R (R Core Team, 2015). A seguir são descritas as principais características e a dinâmica do S4GS.

3.1 Método

O Simulador para Seleção Genômica S4GS foi desenvolvido seguindo o modelo de Wirght e Fish (WRIGHT, 1931), para um genoma diploide com genes bi-alélicos e sem sobreposição de indivíduos entre as gerações. Permite a simulação de múltiplas populações, incluindo o uso da inseminação artificial via banco de sêmen. As ações gênicas podem ser declaradas de acordo com o perfil desejado, sendo possível definir a interação entre múltiplos marcadores, bem como a força da expressão dos mesmos. A combinação das expressões é utilizada para calcular o fenótipo de cada indivíduo.

3.2 Simulação das populações

A simulação populacional é subdividida em duas fases: histórica e recente. A população histórica visa estabelecer o LD após um determinado número de gerações. A última geração da população histórica é utilizada como a parental para a construção da população recente. Em ambas as populações o cruzamento é aleatório e sem sobreposição entre as gerações. Na população recente é possível selecionar os melhores pais. A simulação de duas ou mais populações pode utilizar ou não o mesmo conjunto de parâmetros, assim cada população gerada pode possuir um grupo de parâmetros específicos. A junção das populações pode ser feita de forma automática pelo S4GS ou pelo próprio usuário.

O LD é estabelecido na população histórica por meio da aplicação das forças evolutivas e da recombinação. A recombinação visa representar o processo que ocorre durante a meiose, a combinação dos alelos é codificada em 1, 2, 3 e 4 podendo ser descrita em (1,1), (1,2), (2,1) e (2,2) respectivamente, onde 1 é o A e 2 o a. A recombinação é aplicada durante o cruzamento dos pais onde, inicialmente, os alelos são separados em lados distintos, em seguida são sorteados pontos de quebra no cromossomo sendo gerado na sequência o lado que foi utilizado como "cromossomo sexual", e com a posse de cada lado o simulador gera então o genótipo do novo indivíduo. A Figura 3.1 representa o processo de recombinação executado pelo simulador.

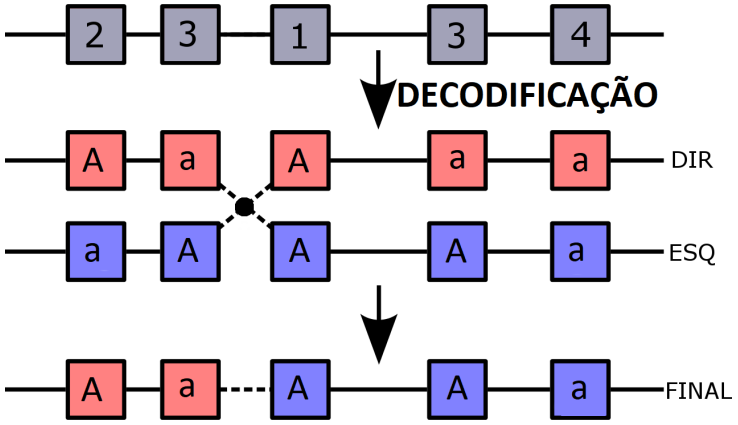


Figura 3.1: Processo de recombinação implementado

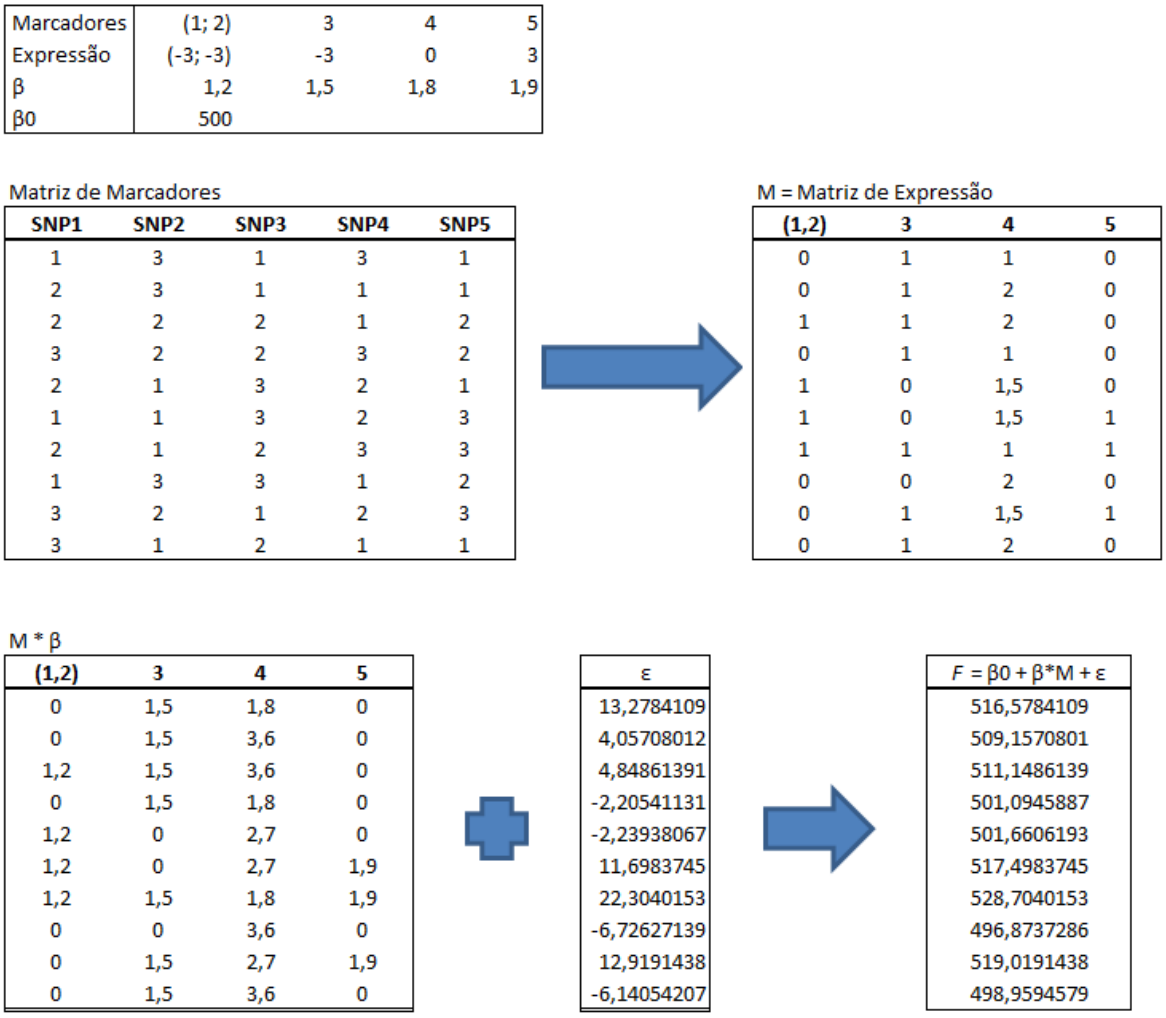
3.3 Simulação do Fenótipo

As ações gênicas clássicas de dominância, codominância e epistasia foram implementadas. As ações gênicas de dominância e recessividade são mapeadas de acordo com sua expressão de forma que na dominância os valores AA e Aa expressam 1 e aa expressam 0 e o contrário a recessividade. A codominância é dada pela expressão de AA, Aa e aa em 2, 1,5 e 1 respectivamente. A epistasia segue o modelo proposto por Cordell (2002), e é dada pela forma como a interação é mapeada. A matriz de mapeamento (M) é calculada de acordo com as interações definidas pelo usuário. O fenótipo (F) é simulado pelo somatório dos efeitos (β) de cada marcador somado a um efeito residual (ϵ) dado por uma distribuição normal com média nula e desvio padrão (σ) definido por parâmetro. O fenótipo consiste na multiplicação da matriz de expressão pelo vetor de β´s somados ao $\beta 0$ e a ϵ, ou seja: $F = \beta 0 + \beta \cdot M + \epsilon$. A Figura 3.2 mostra um exemplo de cálculo do fenótipo.

3.4 Parâmetros de Entrada

As variáveis de entrada do simulador permitem que o usuário personalize a simulação, onde cada parâmetro atua de forma efetiva e diferente no resultado final da simulação. Cada conjunto de dados visa representar uma amostra real, assim sendo o usuário pode refinar a simulação de forma a aproximá-la do cenário de interesse. Os parâmetros utilizados pelo S4GS são:

- **Número de gerações**: Valor numérico positivo e diferente de 0, define a quantidade de gerações utilizadas na construção da população histórica.



Marcadores	(1; 2)	3	4	5
Expressão	(-3; -3)	-3	0	3
β	1,2	1,5	1,8	1,9
β0	500			

Matriz de Marcadores

SNP1	SNP2	SNP3	SNP4	SNP5
1	3	1	3	1
2	3	1	1	1
2	2	2	1	2
3	2	2	3	2
2	1	3	2	1
1	1	3	2	3
2	1	2	3	3
1	3	3	1	2
3	2	1	2	3
3	1	2	1	1

M = Matriz de Expressão

(1,2)	3	4	5
0	1	1	0
0	1	2	0
1	1	2	0
0	1	1	0
1	0	1,5	0
1	0	1,5	1
1	1	1	1
0	0	2	0
0	1	1,5	1
0	1	2	0

M * β

(1,2)	3	4	5
0	1,5	1,8	0
0	1,5	3,6	0
1,2	1,5	3,6	0
0	1,5	1,8	0
1,2	0	2,7	0
1,2	0	2,7	1,9
1,2	1,5	1,8	1,9
0	0	3,6	0
0	1,5	2,7	1,9
0	1,5	3,6	0

ε
13,2784109
4,05708012
4,84861391
-2,20541131
-2,23938067
11,6983745
22,3040153
-6,72627139
12,9191438
-6,14054207

F = β0 + β*M + ε
516,5784109
509,1570801
511,1486139
501,0945887
501,6606193
517,4983745
528,7040153
496,8737286
519,0191438
498,9594579

Figura 3.2: Exemplo de cálculo do fenótipo.

- **Taxa de recombinação**: Valor utilizado como parâmetro em uma função de probabilidade com o objetivo de sortear a quantidade os locais de ocorrência da recombinação. A cada ponto sorteado ocorre uma mudança de lado. A quantidade de pontos é proporcional ao valor da taxa .

- **Taxa de mutação**: Probabilidade de ocorrer uma mudança pontual, SNP, em uma determinada posição do vetor de genótipo.

- **Mais aptos**: Quantidade dos melhores indivíduos, machos e/ou fêmeas, que são considerados no sorteio para cruzamento.

- **Tamanho da população**: Número de indivíduos presentes na população. Todas as gerações possuem a mesma quantidade de indivíduos.

- **Número de marcadores**: Define a quantidade de genótipos cada indivíduo da

população possuirá.

- **Indivíduos no banco de sêmen**: Quantidade de indivíduos, machos, estão presentes no banco de sêmen. A cada geração novos indivíduos são enviados ao banco, em seguida eles são ordenados e somente a quantidade definida é mantida.

- **Envio para o banco**: É o número máximo de machos enviados ao banco de sêmen após cada geração de evolução .

- **Inseminação Artificial**: Define o comportamento a ser adotado pelo simulador quanto ao uso da inseminação artificial, com **0** não utiliza banco de sêmen, **1** utiliza machos do banco e do rebanho a uma proporção de 20/80 e **2** utiliza somente machos do banco de sêmen.

- **Desvio padrão**: Utilizado no cálculo do resíduo, ou erro, ou efeito ambiental.

- $\beta 0$ Valor inicial utilizado como referência para o incremento ou decremento de acordo com somatório dos valores de expressão de cada marcadores.

- β**´s**: Efeito da expressão de cada marcador causal. O valor de β é multiplicado pela mapeamento da ação gênica, que podendo envolver múltiplos marcadores.

- **Marcadores causais**: É lista que contem os marcadores com maior força de expressão dentro do genótipo. Essa lista contêm todos os marcadores considerados importantes no processo de simulação.

- **Lista de interações**: Contêm a forma como os marcadores causais interagem para expressar o fenótipo. As interações dependem da forma como cada ação gênica foi mapeada. O primeiro elemento dessa lista possui como efeito o primeiro item da da lista de β´s e assim sucessivamente.

Exemplos da lista de interação de marcadores causais e β: Lista de interação *((-3, -3), 0 ,c(-3, -3), c(-3, -3), 0, c(-3, -3));* lista de marcadores = *((10, 20), 50, c(30, 40), c(60, 70), 78, c(85, 95));* e β *´s = (1.8, 1.9, 1.7, 1.8, 1.9, 1.7)*

3.5 Implementação

Nas seções anteriores foi apresentada a dinâmica de funcionamento do S4GS. Nessa seção é explicado a estrutura computacional utilizada para produzir cada um dos elementos anteriormente apresentados. A Figura 3.3 apresenta o fluxograma do S4GS, mostrando a separação entre população histórica e recente, onde é possível ver em destaque as funções *setPairs, GetProgeny, GetPhenotype*. Além dessas, existem as funções *InitialPopulation* que gera a população histórica algoritmo 1, e a *ExecSimulation* que cruza uma população dada por N gerações algoritmo 2. O pseudocódigo das funções *setPairs, GetProgeny* também são apresentados em algoritmo 3 e algoritmo 4. A função *GetPhenotype* já foi explicada na seção 3.3 com um exemplo apresentado pela Figura 3.2.

Visando facilitar a troca de informação entre as funções e organizar a disposição dos dados duas estruturas distintas foram criadas, a primeira é *generation* e a segunda *individual*. A estrutura *generation* armazena os valores: lista com a estrutura *individual* de cada geração, número de Gerações simuladas, número de Indivíduos em cada geração, último índice utilizado e a herdabilidade em cada geração. A estrutura *individual* registra as seguintes informações sobre cada indivíduo: identificador único, identificador do pai, identificador da mãe, sexo, matriz de genótipo, fenótipo, O valor fenotípico excluindo-se o erro ou efeito do ambiente (*True Breeding Value* - TBV), erro, GEBV. Softwares como haploview (BARRETT et al., 2005) utilizam arquivos do tipo *PED* e *INFO* para gerar os gráficos de desequilíbrio de ligação entre outros, por esse motivo o S4GS permite exportar as saídas de cada geração em arquivos desse formato, sendo necessário informar somente o nome de base dos arquivo de saída, que é formado pela combinação do nome base e a geração separados pelo simbolo de sublinha.

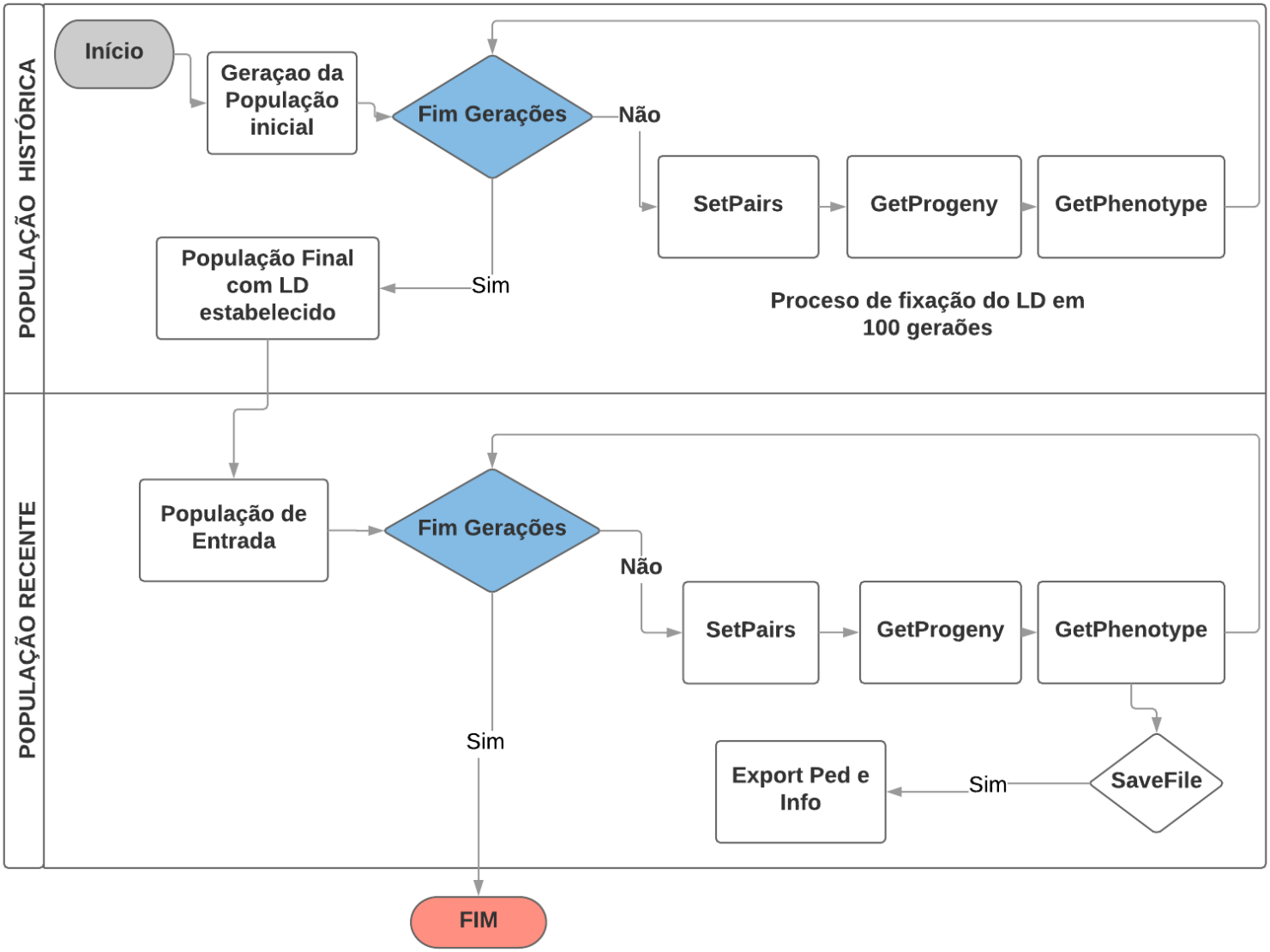


Figura 3.3: Fluxograma detalhada do simulador S4GS.

Algoritmo 1: Pseudocódigo da função *InitialPopulation*

Entrada: Gerações, número de individuos, número de snp, desvio padrão, SNPs Causais,β0, lista de β´s, Interações, Geração Arquivo PED, nome arquivo PED

Resultado: lista com a Estrutura indivíduo + resumo das gerações com GEBV, Herdabilidade e TBV, bem como os arquivos PED e INFO caso seja escolhido

```
Geração do Primeiro Indivíduo totalmente heterozigoto;
Atualiza HERDABILIDADE;
Atualiza lista do GEBV;
Atualiza lista do TBV;
para (i=2; i<gerações; i++) faça
    Chama a função SetPairs;
    Chama a função getProgenyGenotype;
    Cálculo do fenótipo por meio da função getPhenotype();
    Atualiza ind;
    Atualiza HERDABILIDADE;
    Atualiza lista do GEBV;
    Atualiza lista do TBV;
    se Geração Arquivo PED for VERDADEIRA então
        Gera os Arquivos PED e INFO;
    fim se
fim para
```

Algoritmo 2: Pseudocódigo da função *ExecSimulation*

Entrada: População de Entrada, Nº de indivíduos, gerações, quantidade de machos e fêmeas para seleção, uso de inseminação artificial, Geração e nome do Arquivo PED

Resultado: Estrutura GENERATION

```
A geração 1 recebe a população inicial;
para (i=2; i<gerações; i++) faça
    Geração de um novo INDIVIDUAL;
    SetPairs(...);
    getProgenyGenotype(...);
    getPhenotype(...);
    se uso de inseminação artificial == 2 então
        Gera somente fêmeas;
    senão
        Gera fêmeas e machos em proporções aleatórias;
    fim se
    Atualiza a estrutura GENERATION;
    Armazena a estrutura INDIVIDUAL na GENERATION;
    se Geração Arquivo PED for VERDADEIRA então
        Gera os Arquivos PED e INFO;
    fim se
fim para
```

O algoritmo 3 representa a função que define os pares para o cruzamento. A função *SetPairs* apresenta mais três novas funções, sendo elas: *paretoPairs, justBank* e *simplePairs*. Na *paretoPairs* define os pares escolhendo 80% do total de machos dos melhores do rebanho e os outros 20% no banco de sêmen e, entre as fêmeas 80% do número é escolhida dentro das melhores e o restante em todo o rebanho. A *justBank* por sua vez escolhe os machos exclusivamente no banco de sêmen e as fêmeas entre as melhores do rebanho. Por fim, a função *simplePairs* seleciona os pares entre os melhores do rebanho.

Algoritmo 3: Pseudocódigo da função *SetPairs*

Entrada: Número de Indivíduos, quantidade de Macho, quantidade de Fêmea, Estrutura INDIVIDUAL, geração atual, uso de Inseminação Artificial

Resultado: Vetor de Pares

```
Geração dos vetores com os índices dos machos e das fêmeas extraídas de
 INDIVIDUAL;
se quantidade_macho == 0 então
 | quantidade_macho = Total de Machos em INDIVIDUAL
fim se
se quantidade_fêmea == 0 então
 | quantidade_fêmea = Total de Fêmeas em INDIVIDUAL
fim se
se useArtificialInsemination==1 então
 | paretoPairs(...)
senão
 | se useArtificialInsemination==2 então
 |  | justBank(...)
 | senão
 |  | simplePairs(...)
 | fim se
fim se
```

O algoritmo 4 apresenta a função *GetProgeny* que é responsável por gerar o genótipo do filho com base nos pares definidos pela função *SetPairs*. A função é responsável por aplicar a recombinação e a mutação.

Algoritmo 4: Pseudocódigo da função *GetProgeny*

Entrada: genótipo do Macho, genótipo da Fêmea

Resultado: Genotipo do Filho

Sorteio dos pontos de recombinação;

Quebra do genótipo dos pais em dois lados;

Geração do "Cromossomo Sexual" após recombinação;

Obtenção do genótipo do filho por meio da combinação do dos pais;

Inserção de possível mutação;

3.6 Simulações de Teste e Verificação

Essa seção apresenta uma breve análise de sensibilidade das variáveis visando demonstrar o potencial do simulador e possíveis aplicações. As possibilidades de combinações e cenários são inúmeras, contudo foi avaliado o impacto dos parâmetros Taxa de recombinação e de mutação na geração do LD, e o comportamento do simulador com e sem o uso da inseminação artificial. A Tabela 3.1 exibe os parâmetros utilizados em cada um dos seis blocos iniciais de teste. Conforme objetivos traçados, os experimentos computacionais conduzidos nessa tese utilizaram o S4GS como referência sendo possível ver outros cenários na Capítulo 4.

Tabela 3.1: Parâmetros utilizados nos seis blocos de teste do simulador S4GS.

Teste	Taxa Recombinação	Taxa Mutação	Uso Inseminação Artificial
1	0,500	0,001	Não
2	0,090	0,001	Não
3	0,001	0,001	Não
4	0,090	1,000	Não
5	0,090	0,001	Parcial
6	0,090	0,001	Completo

3.6.1 1^o *Teste de sensibilidade de parâmetros*

Os três primeiros blocos de teste têm por objetivo avaliar a sensibilidade ao parâmetros taxa de recombinação, utilizando no 1^o bloco o maior valor entre os três.

A Figura 3.4 mostra a evolução da MAF ao longo das gerações. É possível observar uma uniformidade no início do processo evolutivo sendo seguido por uma maior distribui-

ção da MAF no decorrer das gerações. As gerações da população recente estão destacadas em cinzas, onde o tom mais claro exibe as gerações com seleção dos mais aptos as mais escura indicando a população recente já melhorada. A MAF nas primeiras gerações é muito próxima entre os marcadores, principalmente devido ao fato dos alelos possuírem valor similar. Contudo, com o processo de evolução, ocorre o surgimento da diversidade e uma maior diferença entre as MAFs dos marcadores.

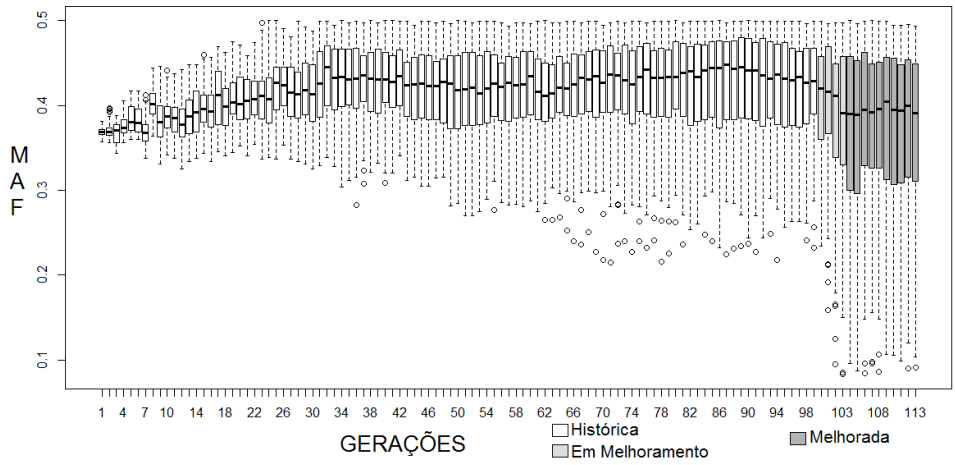


Figura 3.4: Evolução da MAF ao longo das populações histórica e recente. A população recente é divida em duas partes: em melhoramento e melhorada.

A Figura 3.5 mostra a evolução do GEBV de cada geração. Como é possível observar, o GEBV médio é similar em todas as 100 gerações da população histórica, ocorrendo um aumento com o processo de melhoramento genético. A seleção dos melhores animais para o cruzamento melhora o valor médio do fenótipo dos indivíduos. A variação existente na primeira geração é decorrente do erro imputado no cálculo do fenótipo.

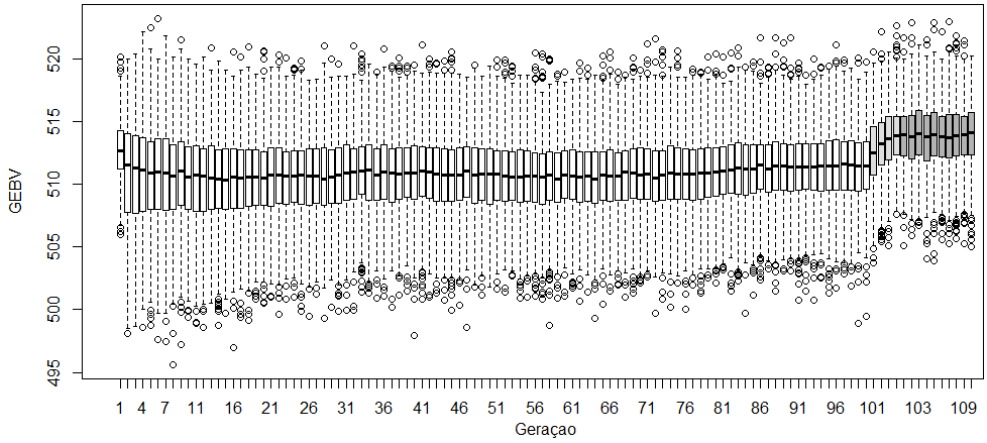


Figura 3.5: Evolução do GEBV ao longo das populações histórica e recente.

A Figura 3.6 mostra a variação da herdabilidade ao longo do processo de simulação, indicando que durante a geração da população histórica ficou entre 40% e 50% e cai com o processo de seleção genômica. A seleção genômica diminui a variação genética dos animais aumentando assim, a influência do ambiente na variação fenotípica da população, por isso, tem-se a queda na herdabilidade.

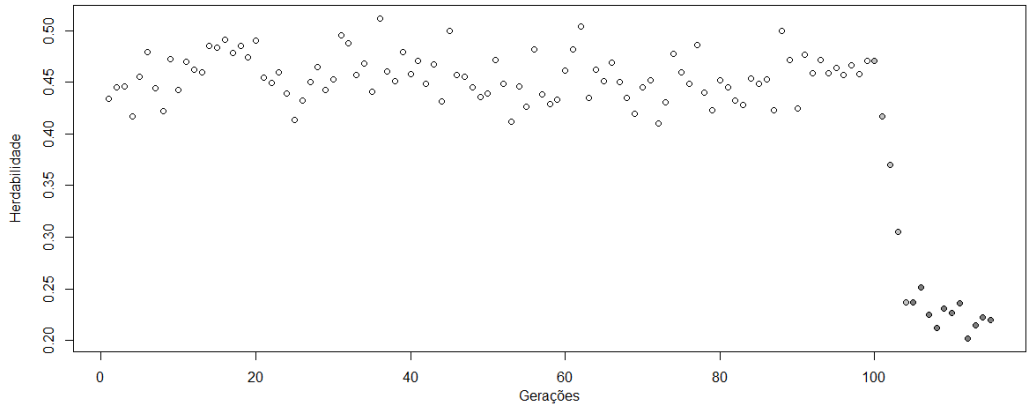
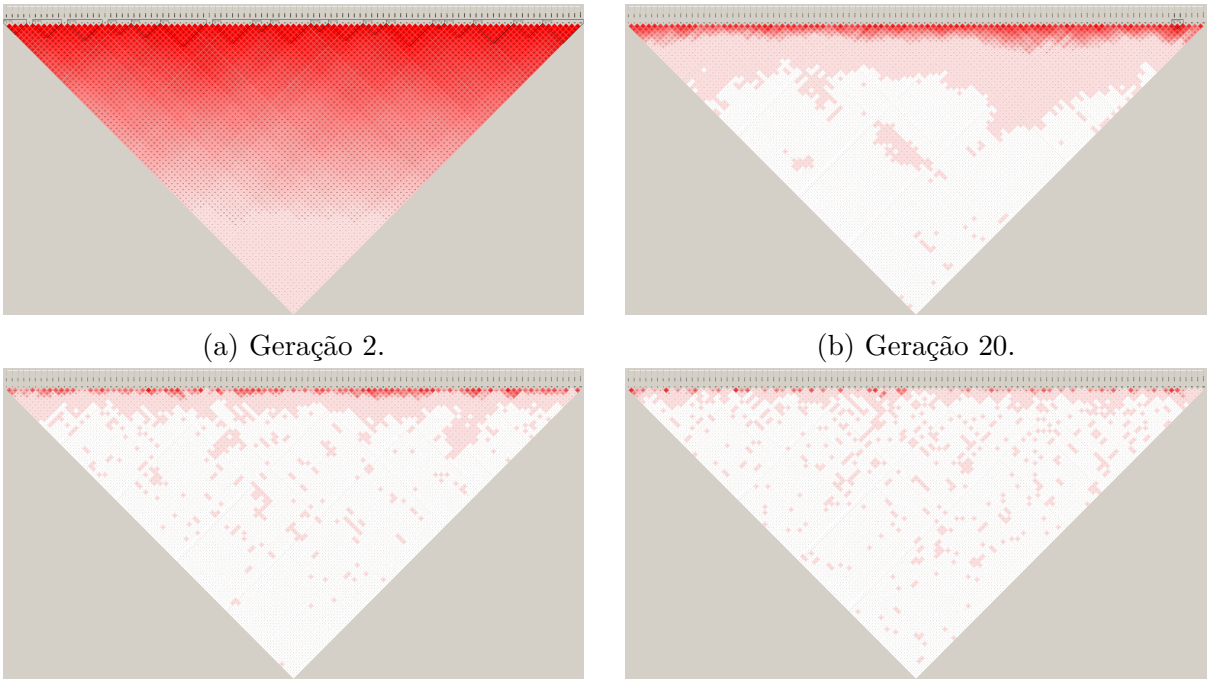


Figura 3.6: Variação da herdabilidade longo das populações histórica e recente.

A Figura 3.7 mostra o LD em quatro momentos diferentes durante a geração da população histórica. A Figura 3.7a exibe o mapa de LD na geração 2, o nível de LD é alto inclusive com a presença de alguns blocos haplótipos. As Figura 3.7b, Figura 3.7c e Figura 3.7d demonstram o LD nas gerações 20, 50 e 100, respectivamente, sendo que a última é utilizada como a parental para a construção da população recente. Como visto, uma taxa de recombinação alta gera uma diminuição do LD muito brusca na constituição da população histórica.

A Figura 3.8 exibe o LD durante as quatro gerações de melhoramento da população recente, sendo possível observar que não ocorrem mudanças significativas no LD com a seleção dos melhores animais.

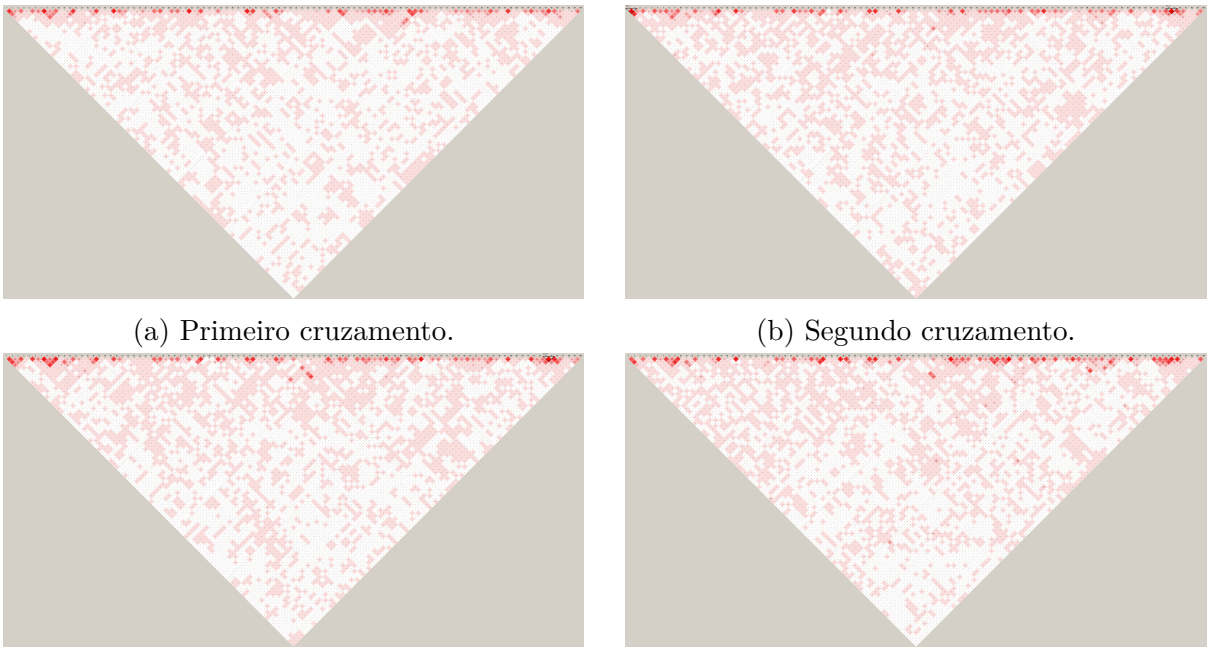
A alta taxa de recombinação gerou um menor nível de LD na população histórica, tendo por objetivo fixar o LD, que é uma característica essencial no estudo em seleção genômica. O nível de LD observado pode variar de acordo com o objetivo do estudo, logo esse parâmetro tem um importante papel no processo de simulação das populações.



(a) Geração 2. (b) Geração 20.

(c) Geração 50 (d) Geração 100.

Figura 3.7: Mapa de LD de algumas gerações da população histórica.



(a) Primeiro cruzamento. (b) Segundo cruzamento.

(c) Terceiro cruzamento. (d) Quarto cruzamento.

Figura 3.8: Mapa de LD de cada geração da população recente durante a etapa de melhoramento.

3.6.2 2º Teste de sensibilidade de parâmetros

O segundo bloco de teste busca igualmente avaliar o impacto da taxa de recombinação. O valor utilizado é $0,09$, menor que o utilizado no primeiro bloco.

A Figura 3.9 exibe os gráficos da variação da MAF em cada geração. É possível observar um comportamento similar ao primeiro bloco durante a evolução da população histórica, porém na população recente ocorre uma variação mais brusca e com uma queda maior após o processo de seleção genômica.

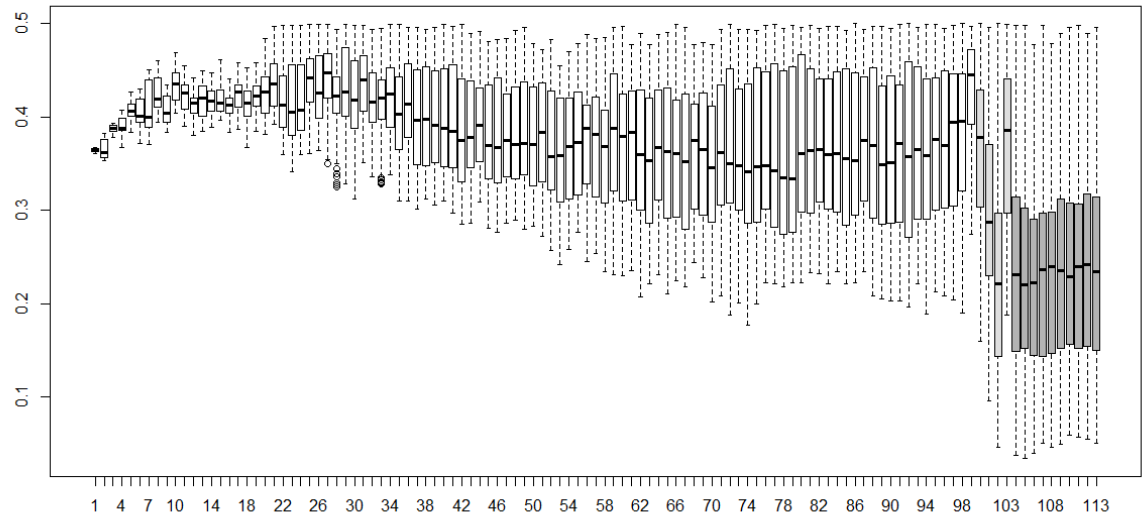


Figura 3.9: Evolução da MAF no segundo bloco ao longo das populações histórica e recente. A população recente é divida em duas partes: em melhoramento e melhorada.

A Figura 3.10 mostra a evolução do GEBV em cada geração. É possível verificar um comportamento similar ao observado no primeiro bloco onde o GEBV médio é similar em todas as 100 gerações da população histórica e com um aumento pelo processo de melhoramento genético.

A Figura 3.11 apresenta o gráfico da herdabilidade ao longo das gerações e, diferente do primeiro bloco de teste, o valor inicial é maior que o anterior, variando entre 80% e 55% na população histórica.

A Figura 3.12 mostra os mapas de LD em quatro gerações distintas da população histórica. A taxa de recombinação menor permitiu o surgimento de blocos haplótipos e marcadores desequilibrados entre si. O mapa da geração 2 contém grandes blocos haplótipos que nas gerações seguintes vão se segmentando. Porém, mesmo na geração 100, os blocos estão presentes, em maior quantidade, mas menores em distância.

O segundo bloco de teste utilizou uma taxa de recombinação menor. Ocorre o surgimento e manutenção de blocos haplótipos e de LD na população recente. A Figura 3.13

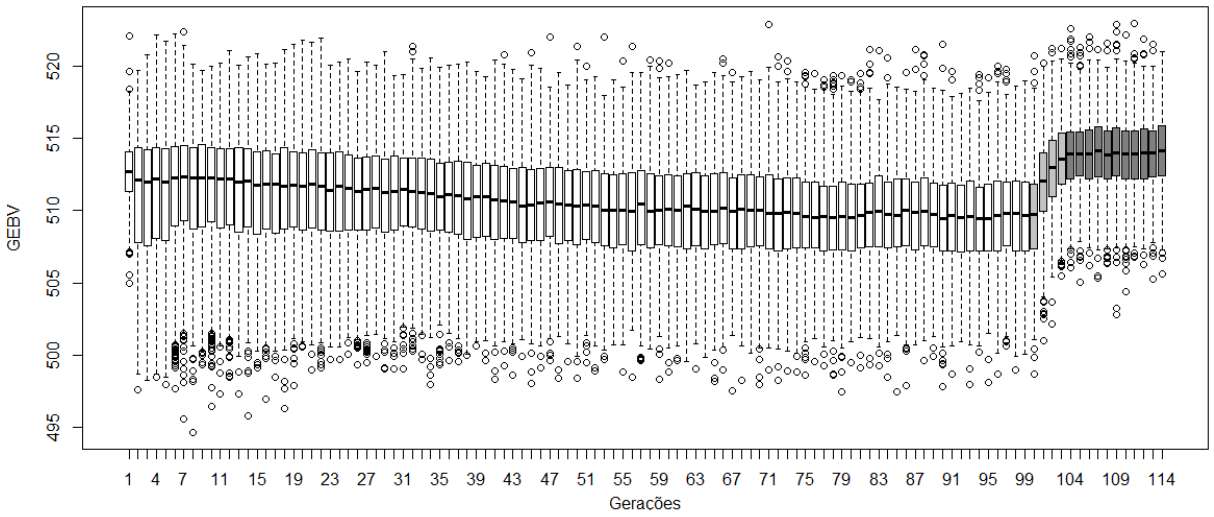


Figura 3.10: Evolução do GEBV no segundo bloco de teste ao longo das populações histórica e recente.

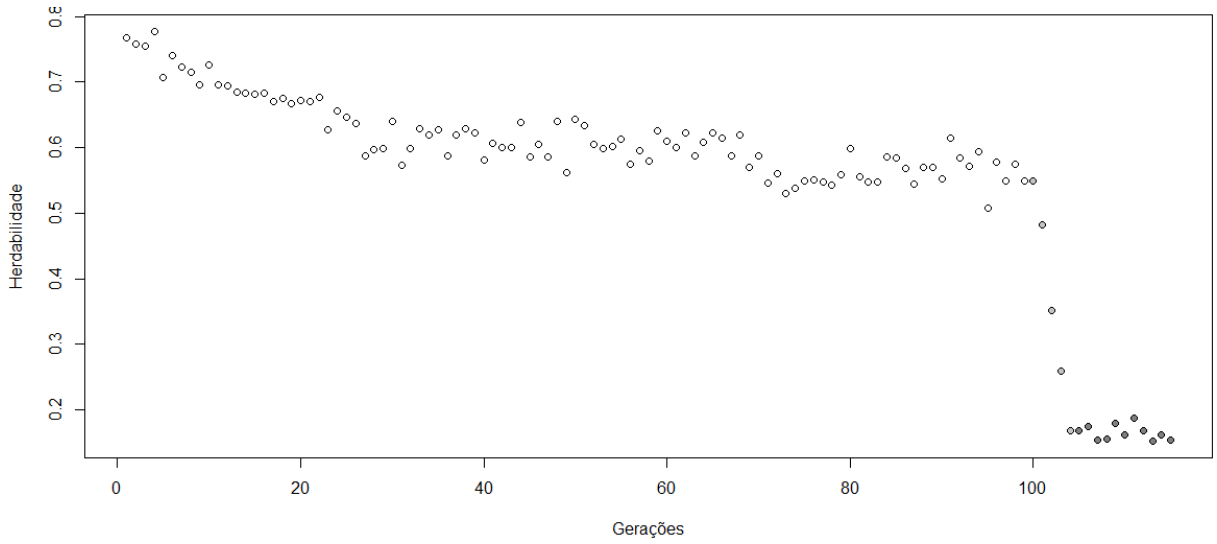
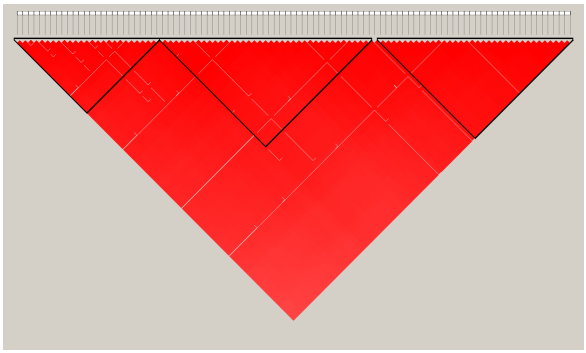


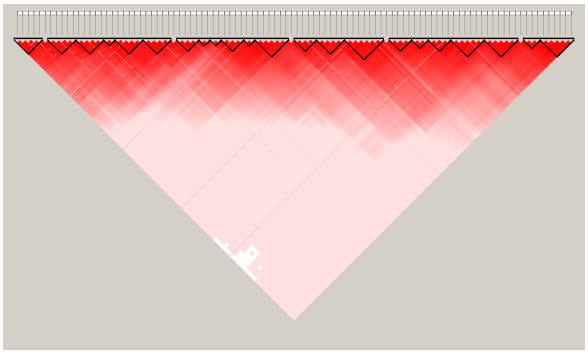
Figura 3.11: Variação da herdabilidade no segundo bloco ao longo das populações Histórica e recente.

mostra o resultado dessa mudança na população recente onde é possível ver a manutenção do LD gerado pela população histórica, (Figura 3.12d).

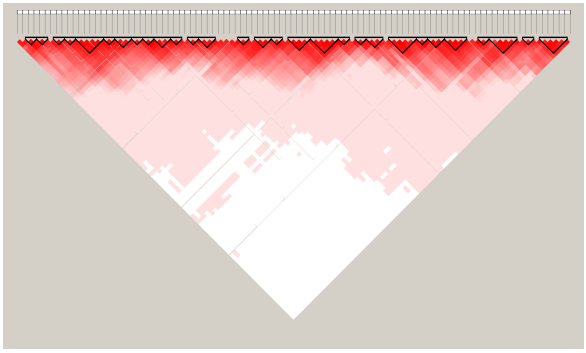
A variação na taxa de recombinação de $0,5$ para $0,09$ permitiu o surgimento de blocos haplótipos e LD nos marcadores. O nível de LD varia de acordo com a combinação de todos os parâmetros do simulador contudo, pode ser ajustado com uma correção na taxa de recombinação.



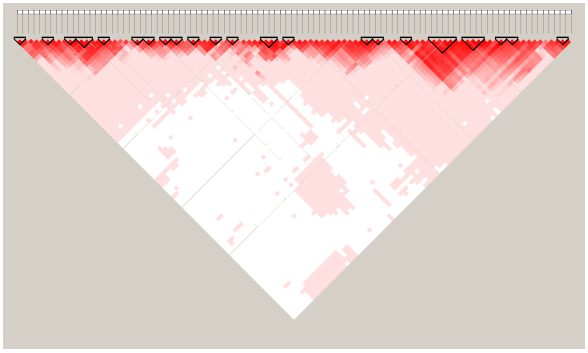
(a) Geração 2.



(b) Geração 20.

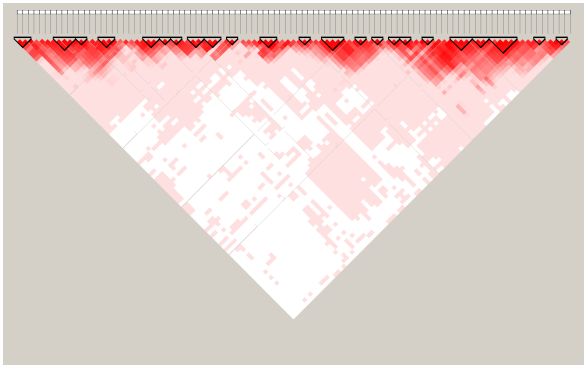


(c) Geração 50

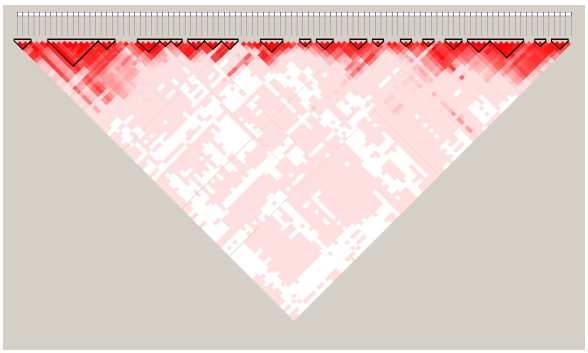


(d) Geração 100.

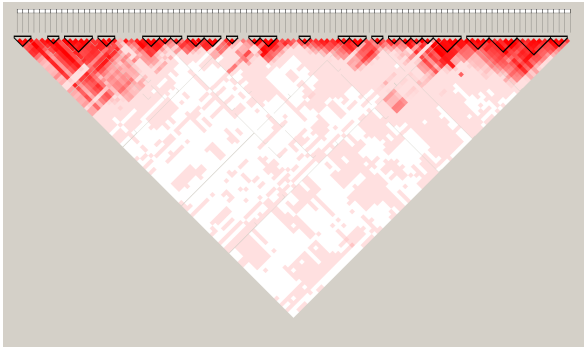
Figura 3.12: Mapa de LD de algumas gerações da população histórica do segundo bloco.



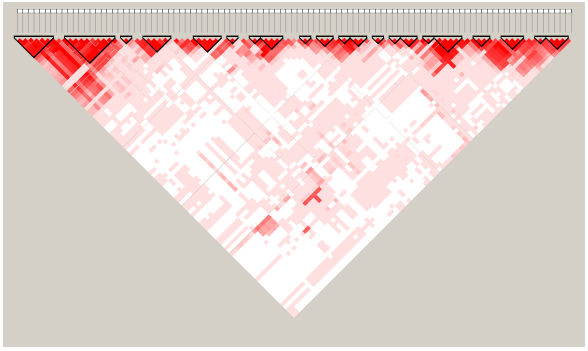
(a) Primeiro Cruzamento.



(b) Segundo.



(c) Terceiro.



(d) Quarto.

Figura 3.13: Mapa de LD de cada geração da população recente do bloco 2, durante a etapa de melhoramento.

3.6.3 3^0 *Teste de sensibilidade de parâmetros*

O terceiro bloco de teste possui uma taxa menor de recombinação como parâmetro variante. Os dois primeiros conjuntos avaliaram o impacto da variação da taxa de recombinação, assim como este bloco. A seguir são mostradas as análises e os gráficos resultantes do uso de uma taxa de recombinação igual a $0,001$.

A Figura 3.14 mostra a evolução da MAF ao longo das gerações. No início da geração histórica não é notada variação, mas após a geração 22 ocorre uma queda chegando a zero na geração 40 e seguindo com poucas variações. Devido a uma taxa de recombinação menor praticamente não ocorrem quebras, dificultando o surgimento de variação no genótipo da população. Vale ressaltar que o LD e as variações surgem por causa da recombinação, sendo esse um importante parâmetro do S4GS.

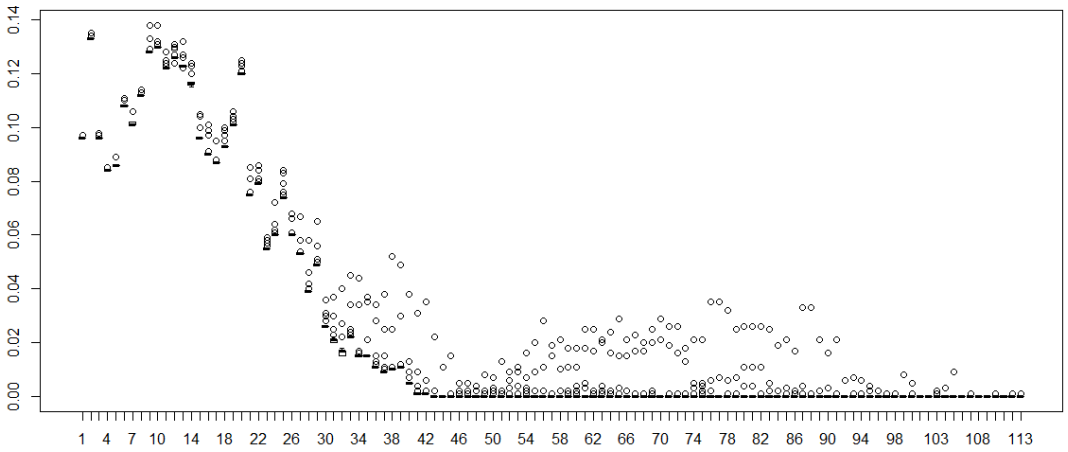


Figura 3.14: Evolução da MAF no terceiro bloco ao longo das populações histórica e recente.

A Figura 3.15 exibe a evolução do GEBV em cada geração. É possível verificar um comportamento diferente do observado nos dois primeiros blocos. Como não ocorre variação no genótipo, o fenótipo calculado se mantém próximo, sendo a variação oriunda somente do erro. Como os indivíduos presentes possuem valores de genótipos próximos e fenótipos similares, a escolha dos melhores para a seleção genômica não leva a um aumento do valor médio devido a ausência de variação genotípica.

A Figura 3.16 mostra o gráfico da herdabilidade, onde verifica-se que, como toda a variação do fenótipo é devida ao erro (que no simulador faz o papel do ambiente), a

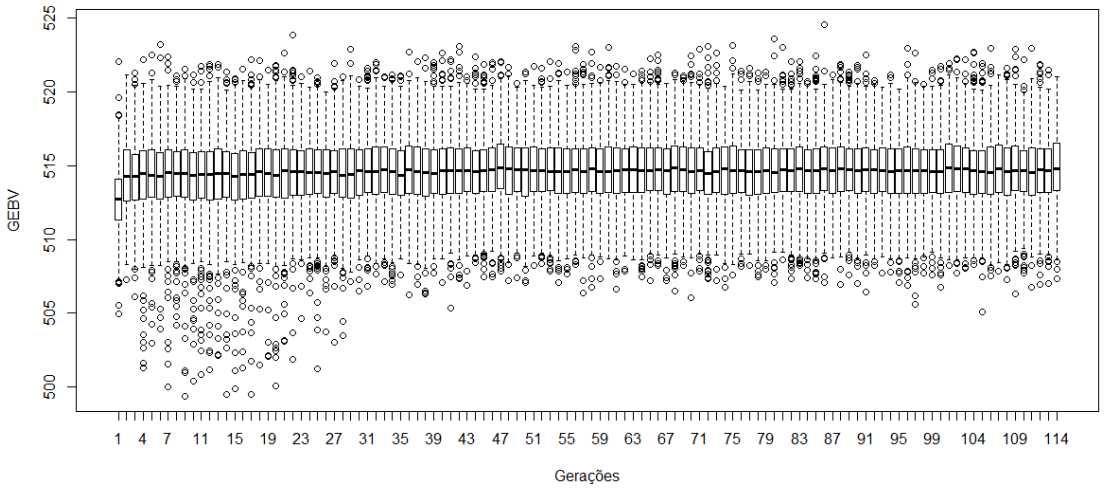


Figura 3.15: Evolução do GEBV no terceiro bloco ao longo das populações histórica e recente.

herdabilidade se anula com a ausência de variação genotípica.

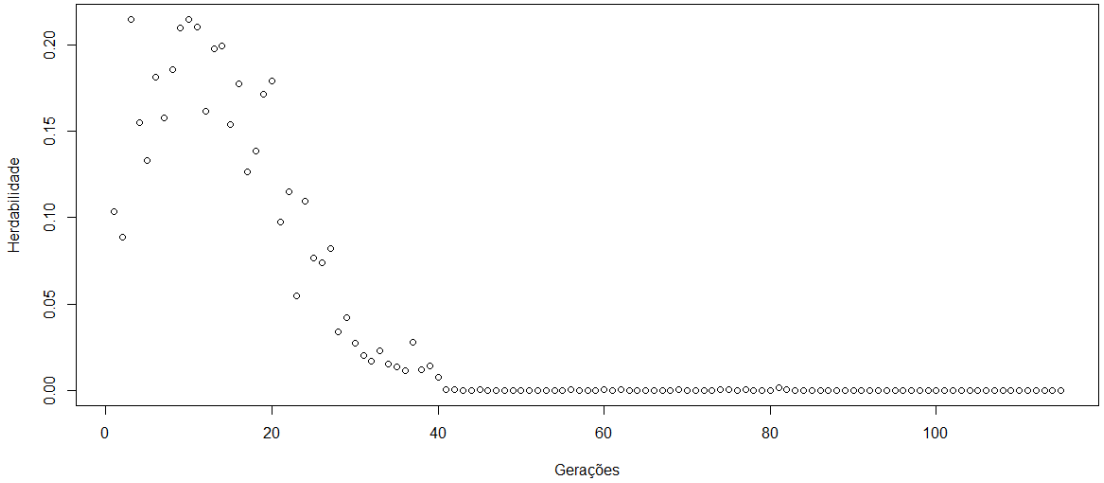
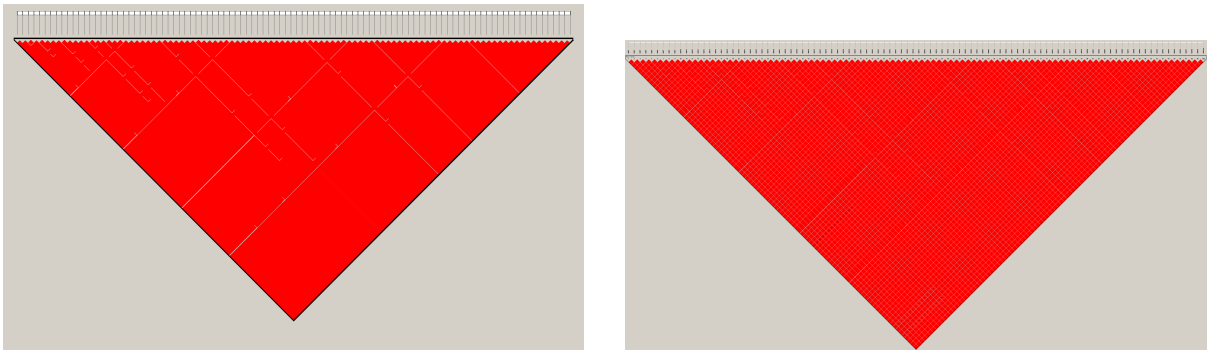


Figura 3.16: Variação da herdabilidade no terceiro bloco de teste ao longo das populações histórica e recente.

A Figura 3.17 contém os mapas de LD das gerações 2 e 20. Devido a ausência de variabilidade genética não há o surgimento de LD nem formação de blocos, onde, mesmo com o processo de seleção, nenhuma variação ocorreu.

Como visto, a taxa de recombinação possui um papel importante no S4GS, sendo uma das principais responsáveis pelo surgimento e manutenção de LD e blocos haplótipos. A variação do nível de LD permite ao usuário do S4GS aproximar o cenário simulado do



(a) Geração 2. (b) Geração 20.

Figura 3.17: Mapa de LD de cada geração da população histórica do terceiro bloco.

real. A variação dos parâmetros do simulador pode trazer um melhor entendimento do cenário de interesse inferindo seu possível comportamento. Nos próximos blocos de teste o valor da taxa de recombinação foi fixado em $0,09$

3.6.4 4^o Teste de sensibilidade de parâmetros

Os três primeiros blocos avaliaram a taxa recombinação, agora nesse conjunto foi avaliado o impacto de uma taxa de mutação alta, com o valor alterado de $0,001$ para 1. Os parâmetros são os mesmo do teste 2, com exceção da taxa de mutação, possibilitando comparar o seu impacto no processo de simulação das populações.

A Figura 3.18 exibe a evolução da MAF, que se inicia com pouca variação obtendo valores estáveis após a geração 50, aumentando em amplitude com o processo de seleção genômica. O comportamento ficou similar ao observado no bloco 2, ou seja a variação na taxa de mutação não gerou um impacto significativo no comportamento da MAF.

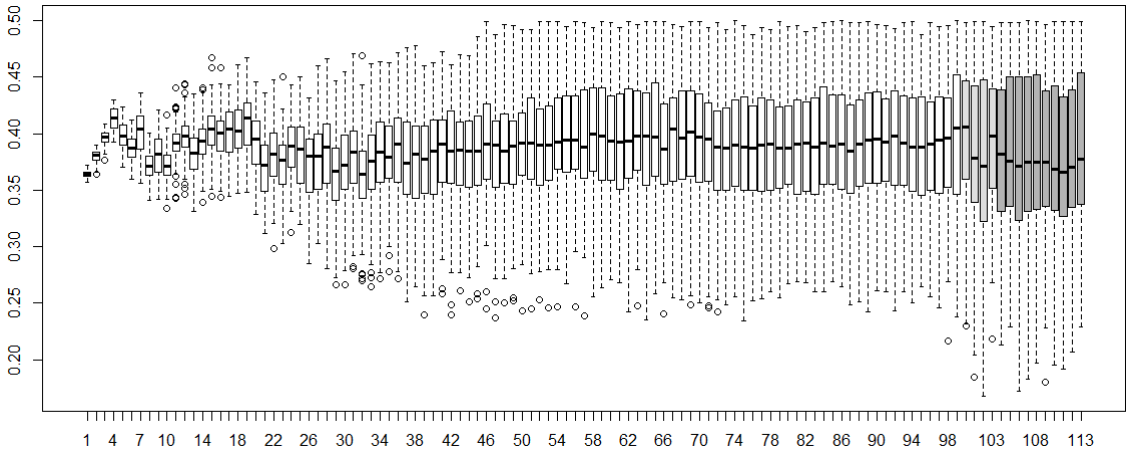


Figura 3.18: Evolução da MAF no quarto bloco ao longo das populações histórica e recente.

A evolução do GEBV observado na Figura 3.19 segue de forma similar à Figura 3.10 do bloco 2.

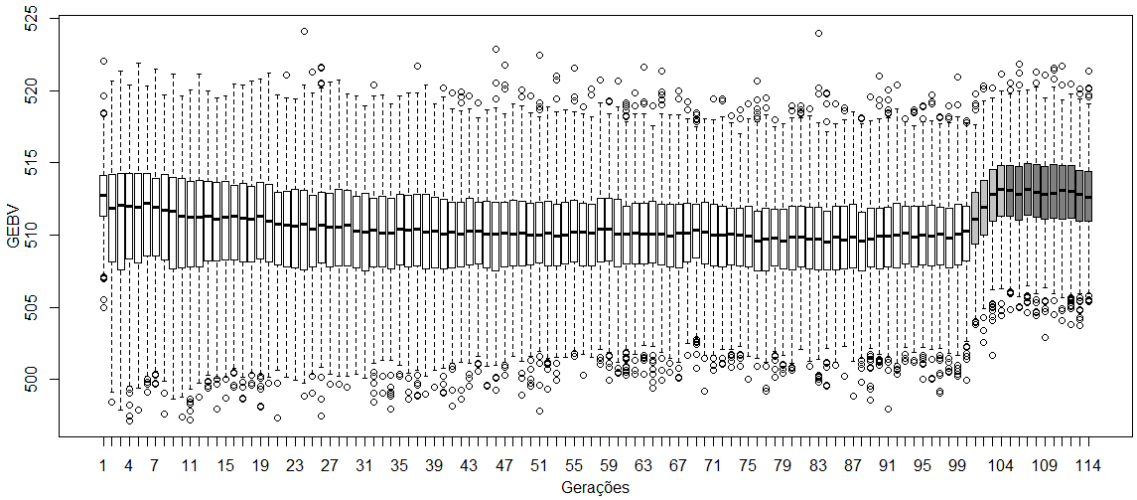


Figura 3.19: Evolução do GEBV no quarto bloco ao longo das populações histórica e recente.

No caso da herdabilidade, é possível observar uma alteração em variação, pois no bloco 4 a Figura 3.20 exibe uma queda contínua durante a simulação da população histórica enquanto no bloco 2, Figura 3.11, a herdabilidade é praticamente constante durante a mesma fase.

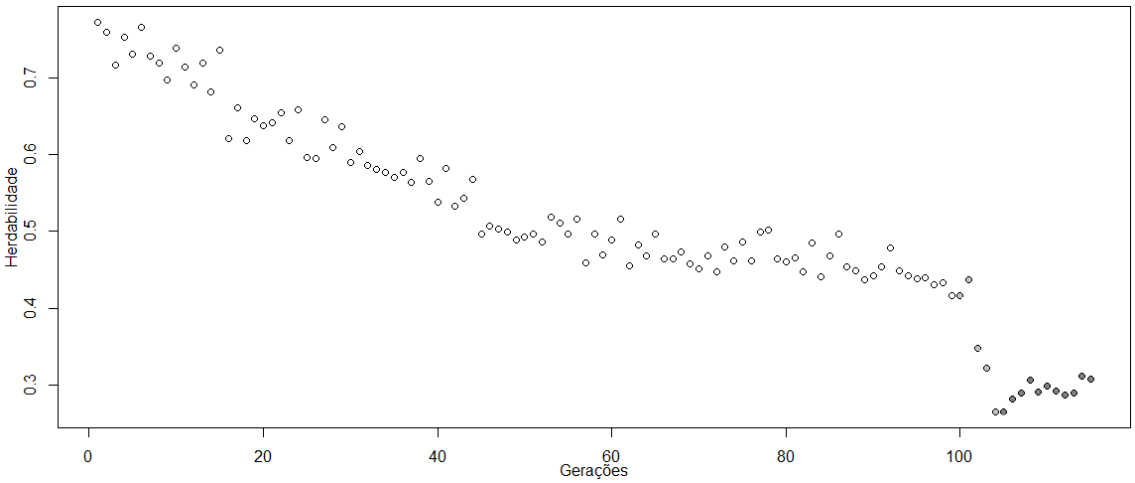
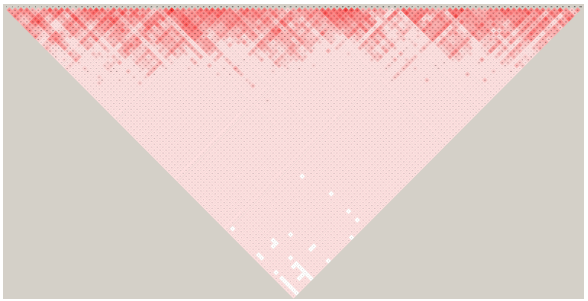


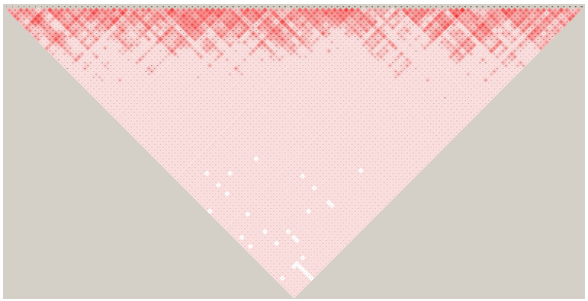
Figura 3.20: Variação da herdabilidade no quarto bloco ao longo das populações histórica e recente.

O grande impacto da mudança na taxa de mutação pode ser visto nos gráficos de LD, onde devido a uma taxa de mutação muito alta, não ocorre a fixação do LD e nem o surgimento de blocos haplótipos. O comportamento observado nas Figuras 3.21 e 3.22 é diferente se comparado com os testes anteriores, pois nas primeiras gerações da

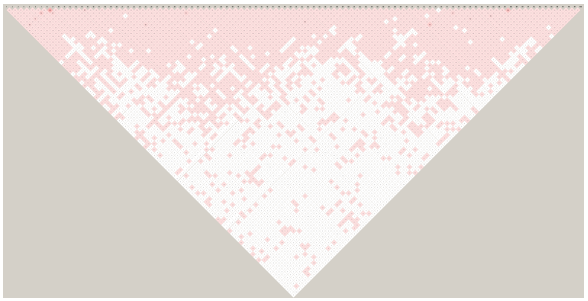
população histórica há a incidência de LD que se perde nas gerações seguintes. Os gráficos de LD, da população recente mostram uma distribuição maior dispersão das combinações se comparada com os 1^o e o 2^o teste.



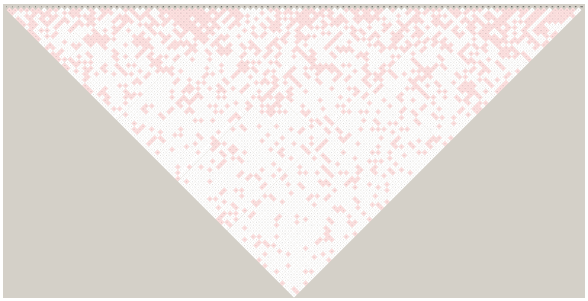
(a) Geração 2.



(b) Geração 20.

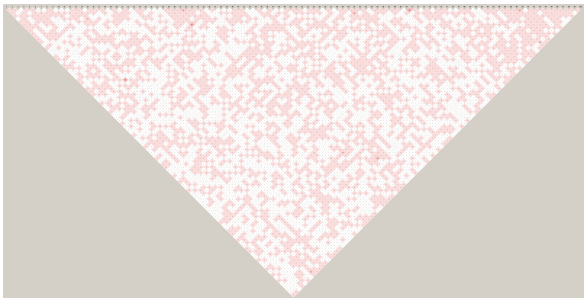


(c) Geração 50.

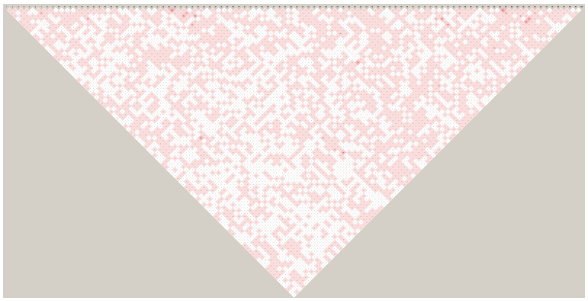


(d) Geração 100.

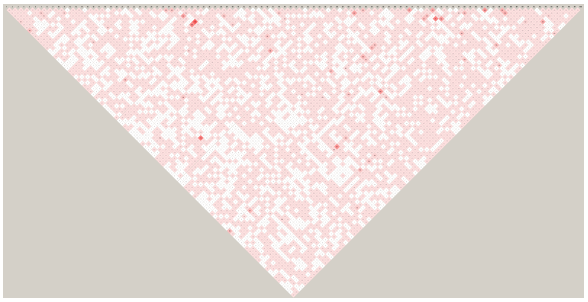
Figura 3.21: Mapa de LD de algumas gerações da população histórica do quarto bloco.



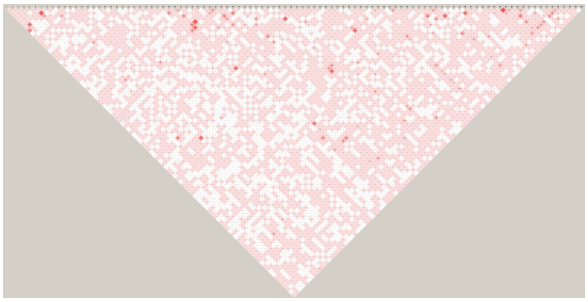
(a) Primeiro Cruzamento.



(b) Segundo.



(c) Terceiro.



(d) Quarto.

Figura 3.22: Mapa de LD de cada geração da população recente do quarto bloco, durante a etapa de melhoramento.

3.6.5 5^o *Teste de sensibilidade de parâmetros*

O último parâmetro avaliado é o uso de inseminação artificial. Os blocos 5 e 6 são variações do 2^o com o uso combinado de machos do rebanho e do banco de sêmen no bloco 5 e uso exclusivo da inseminação artificial no bloco 6.

As Figuras 3.23, 3.24 e 3.25 mostram a varição da MAF, GEBV e herdabilidade, respectivamente. O comportamento é similar ao observado no segundo bloco, com a manutenção durante a simulação da população histórica e uma variação devido ao processo de seleção genômica. O destaque fica por conta da variação da herdabilidade com um leve aumento nas últimas gerações, que pode ter ocorrido devido a variabilidade inserida pela inseminação artificial.

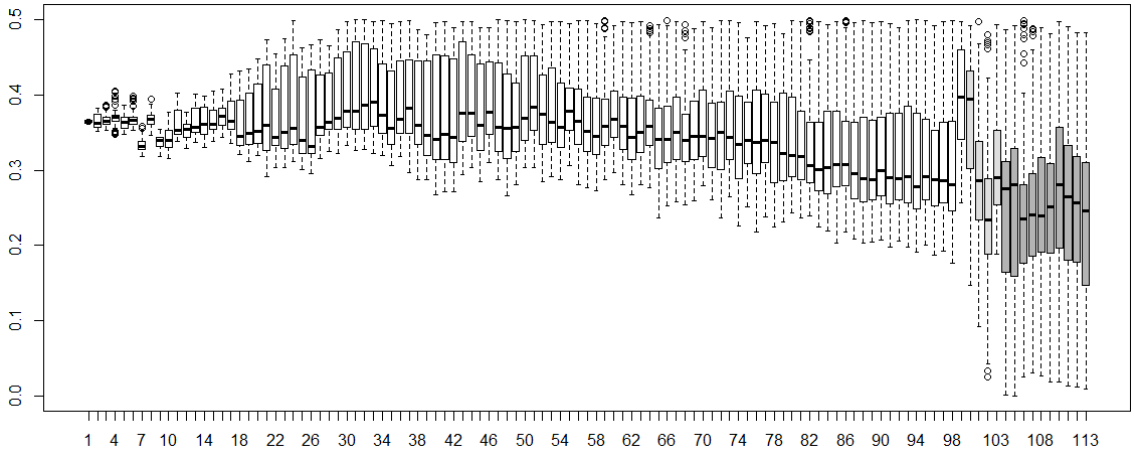


Figura 3.23: Evolução da MAF no quinto bloco ao longo das populações histórica e recente.

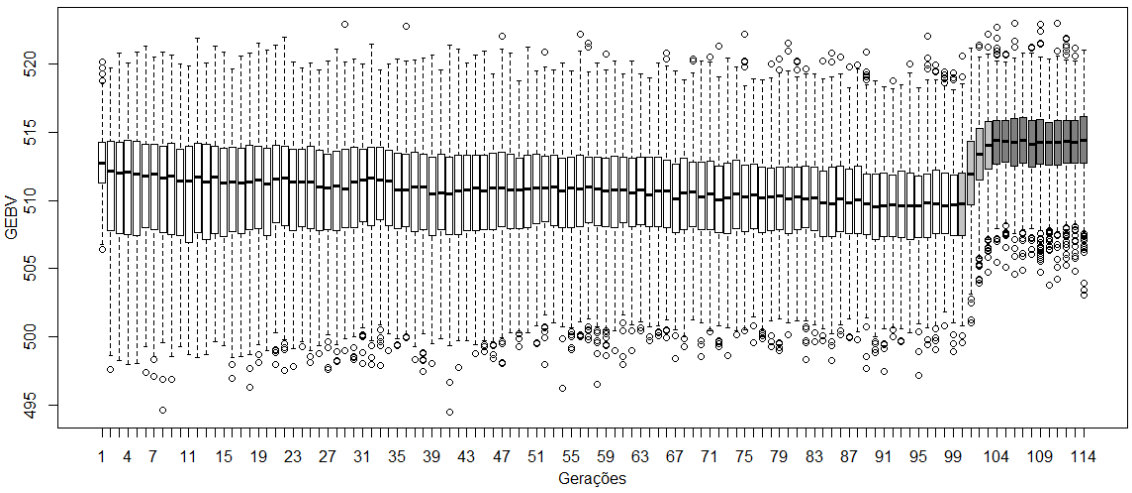


Figura 3.24: Evolução do GEBV no quinto bloco ao longo das populações histórica e recente.

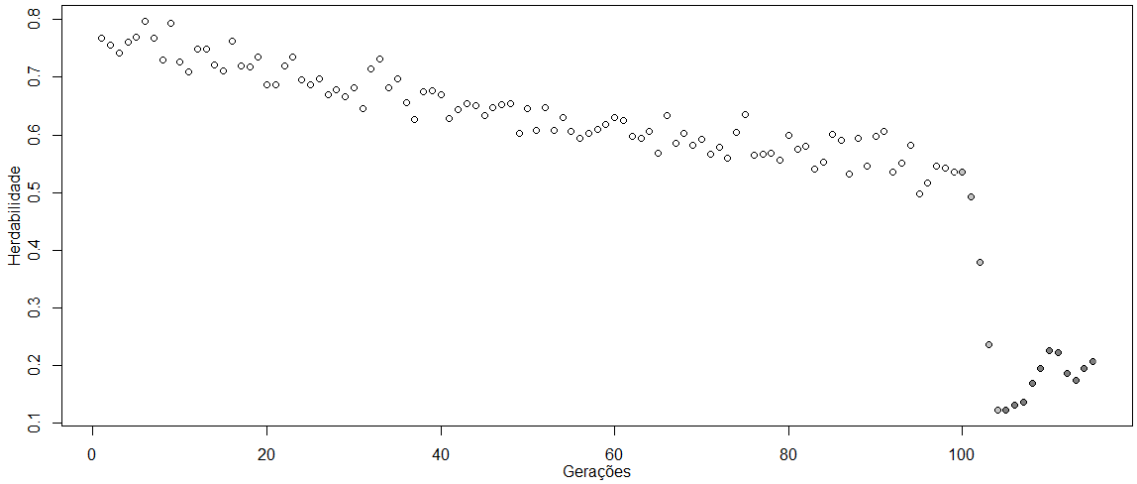
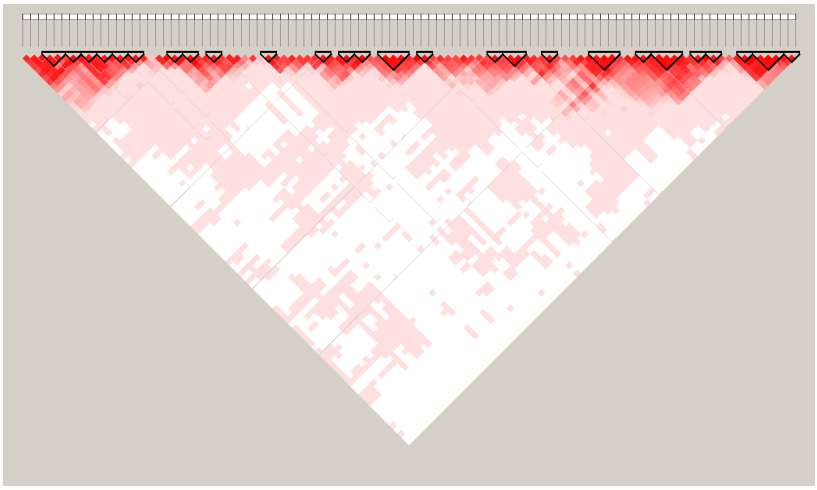


Figura 3.25: Variação da herdabilidade no quinto bloco ao longo das populações histórica e recente.

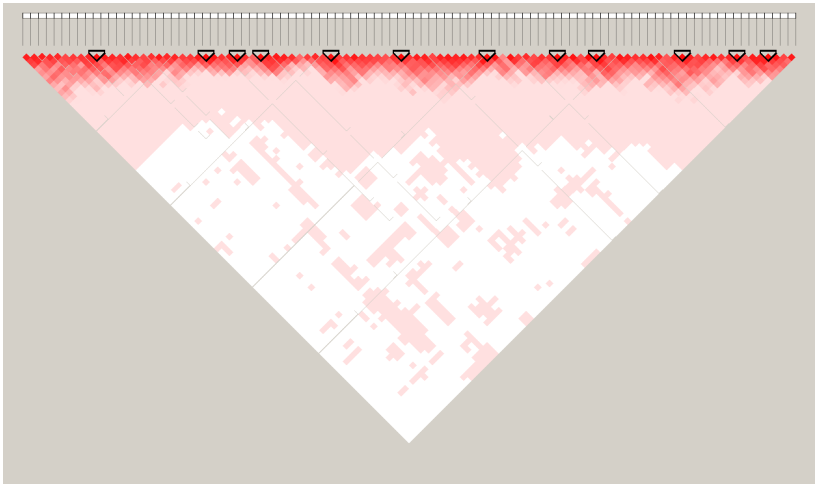
A Figura 3.26 exibe os gráficos de LD das primeiras populações envolvidas na geração da população recente. A Figura 3.26a exibe a última geração da população histórica, que é utilizada como entrada para cruzamento e geração da população recente. A Figura 3.26b apresenta o gráfico da população histórica utilizada para evoluir os machos do banco de sêmen. E, por último, a Figura 3.26c indica o resultado do cruzamento da população da Figura 3.26a com os machos da Figura 3.26b, onde é possível notar alguns padrões de LD que podem ter sido herdados de cada uma das populações utilizadas.

O bloco 5 é o primeiro dos blocos a avaliar o impacto do uso da inseminação artificial, sendo possível observar que a MAF, o GEBV e a herdabilidade não sofreram variações sensíveis. O mapa de LD não demonstrou variações nítidas da geração parental para a F1. Em partes isso pode ser explicado pelo uso combinado de machos do banco de sêmen e do rebanho, e também pelo uso dos mesmos valores de β e $\beta 0$ para a geração das duas populações. Contudo, essa característica pode impactar os resultados na etapa de seleção dos marcadores como é mostrada em uma análise feita em outro conjunto de dados no Capítulo 6.

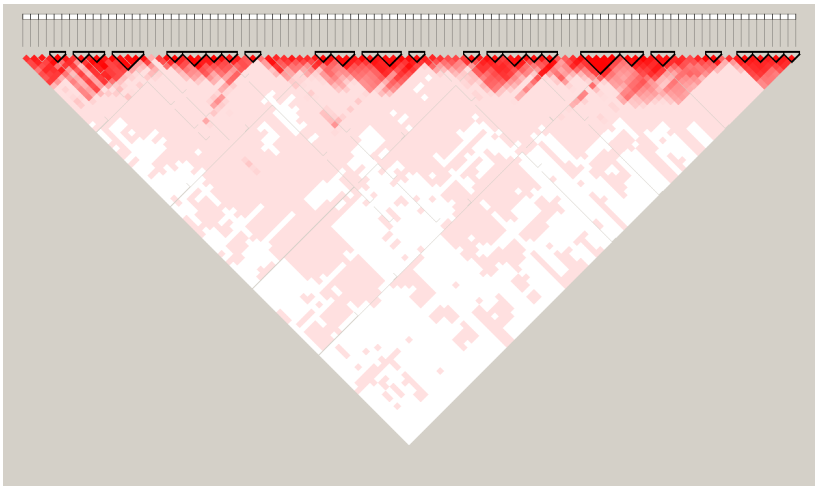
O uso da inseminação artificial é um importante recurso do S4GS, sendo possível simular a população do banco de sêmen com parâmetros diferentes da população de referência. Esse tipo de simulação exige do usuário o controle refinado do processo de cruzamento entre as diferentes populações, e da variação dos valores de β e $\beta 0$ em cada etapa de cruzamento. Esse processo é melhor explicado na geração do Girolando PS no Capítulo 4.



(a) Parental gerada pela população histórica.



(b) População de origem dos machos do banco de sêmen.



(c) F1 originada do cruzamento da parental com parte dos machos do banco de sêmen.

Figura 3.26: Mapa de LD das diferentes populações utilizadas para a geração da F1 no bloco 5

3.6.6 6^o *Teste de sensibilidade de parâmetros*

O sexto bloco de teste é o último a ser avaliado, assim como no quinto, estuda o impacto do uso da inseminação artificial, contudo com o uso exclusivo, ou seja, a totalidade dos machos utilizados no cruzamento são provenientes do banco de sêmen.

As variações de MAF, GEBV e herdabilidade vista nas Figuras 3.27, 3.28 e 3.29 não sofreram grandes diferenças se comparado com os blocos 5 e 2. O comparativo é devido ao uso do bloco 2 como referência para a montagem do bloco 5 e do bloco 6, como as taxas de recombinação e mutação utilizadas no bloco 2 obtiveram resultados de LD próximos ao esperado.

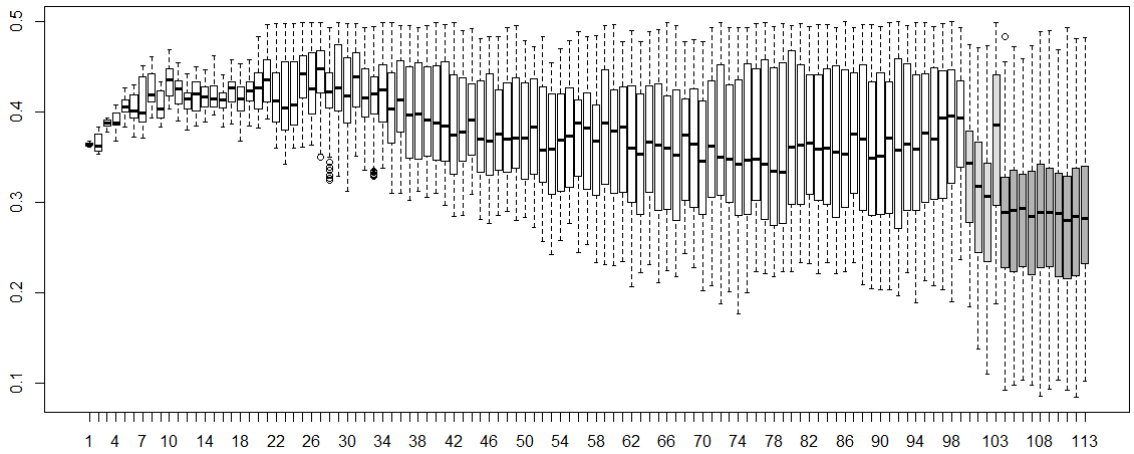


Figura 3.27: Evolução da MAF no sexto bloco ao longo das populações histórica e recente.

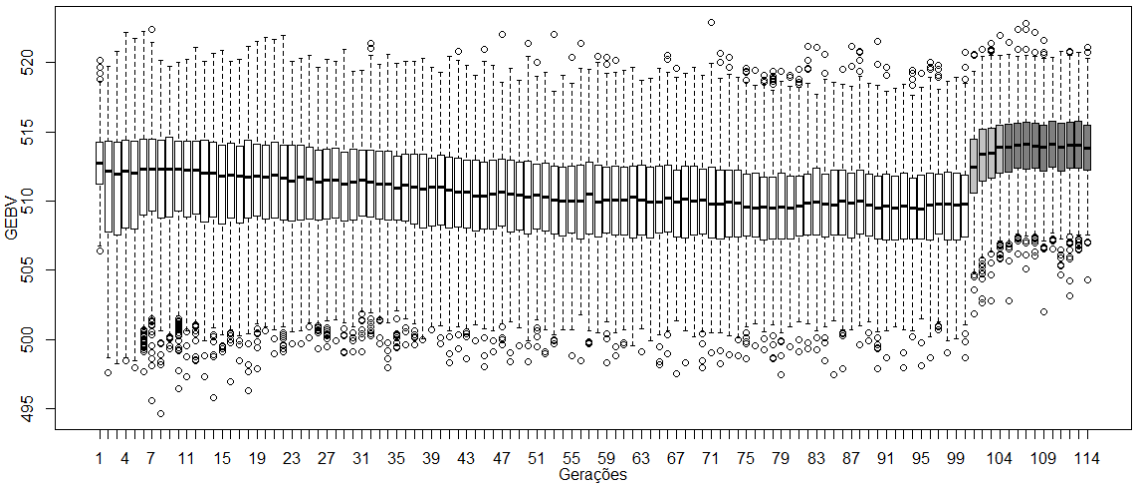


Figura 3.28: Evolução do GEBV no sexto bloco ao longo das populações histórica e recente.

A Figura 3.30 exibe os mapas de LD das populações utilizadas na obtenção da geração F1, Figura 3.30c. O cruzamento da população histórica, Figura 3.30a, e os machos da

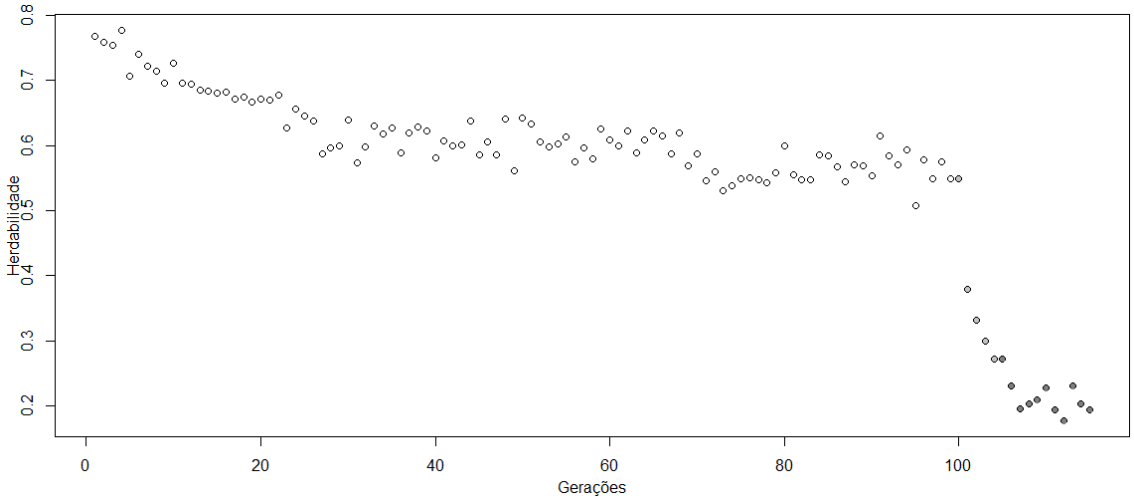


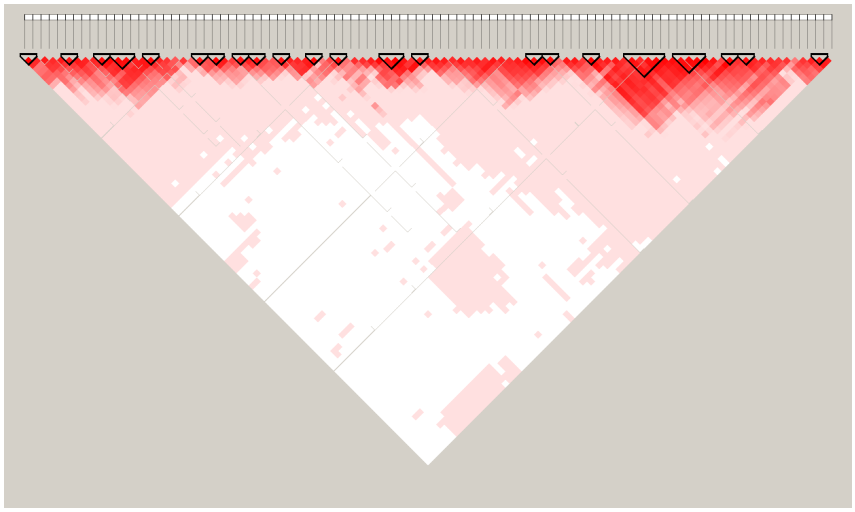
Figura 3.29: Variação da herdabilidade no sexto bloco ao longo das populações histórica e recente.

do banco de sêmen extraído de outra população histórica, Figura 3.30b, geram a F1, Figura 3.30c, e, como é possível observar, alguns blocos haplótipos e regiões de LD foram extraídas de cada uma das populações de origem. Como bloco 6 utiliza exclusivamente os machos do banco de sêmen, uma parte mais consistente de regiões desequilibradas foi importada dos pais.

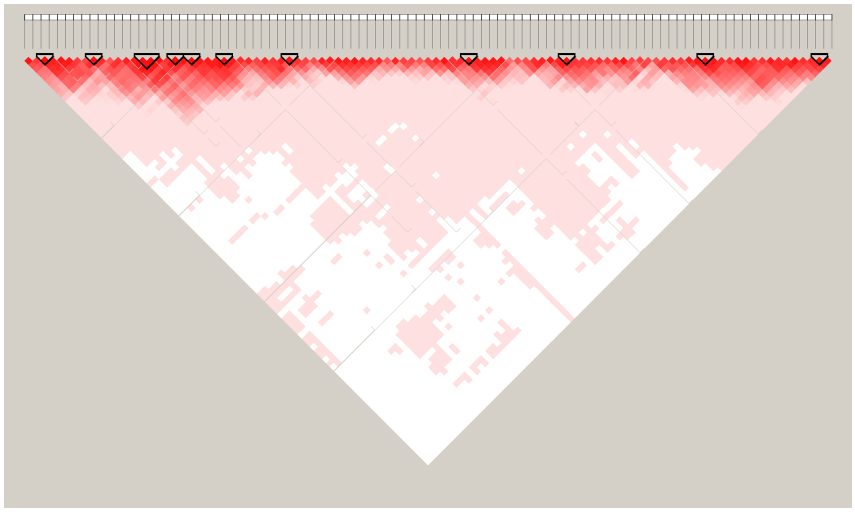
3.7 Considerações

O S4GS conseguiu mimetizar características importantes para o estudo em seleção genômica, como LD, inseminação artificial e cruzamento geracional. Outro fator relevante é a capacidade de simular diferentes ações gênicas e interações em múltiplos níveis. Como visto, o parâmetro taxa de recombinação permite controlar o nível de LD em uma determinada amostra. A implementação utilizando o R permite ao usuário a geração de gráficos e a manipulação dos dados para uso em outras rotinas e bibliotecas, inclusive para controle de qualidade, análise estatísticas ou inferência via inteligência computacional.

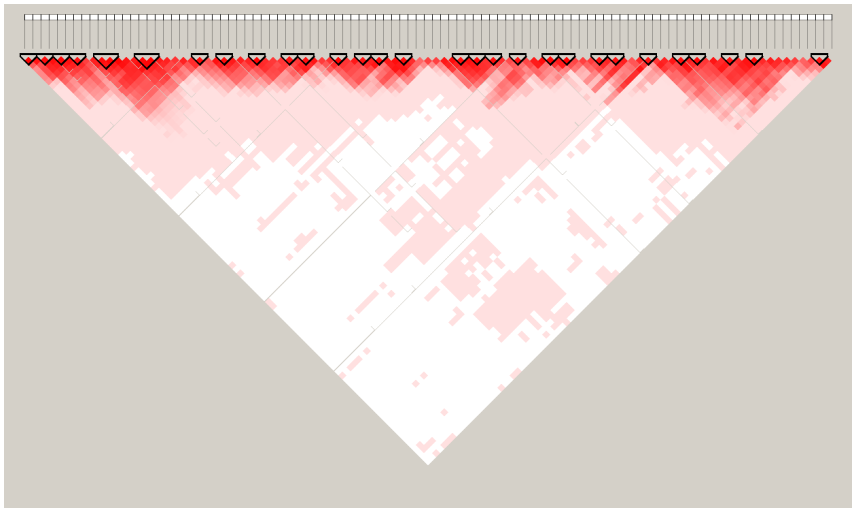
Porém, o mais relevante na construção do S4GS está na sua completude em relação aos principais simuladores usualmente utilizados, o que traz uma grande flexibilidade para o usuário, permitindo uma ampla gama de experimentos sem a necessidade do uso de múltiplas ferramentas.



(a) Parental gerada pela população histórica.



(b) População de origem dos machos do banco de sêmen.



(c) F1 originada do cruzamento da parental com parte dos machos do banco de sêmen.

Figura 3.30: Mapa de LD das diferentes populações utilizadas para a geração da F1 no sexto bloco

4 Geração e Construção dos Dados Simulados

O propósito desse capítulo é apresentar a forma como os dados utilizados na etapa de experimentos computacionais foram simulados. Como apresentado no Capítulo 3, o uso do S4GS permite definir a forma como o fenótipo será expresso, utilizando para isso as ações gênicas modeladas e a aplicação de efeito por marcador. A seguir, é explicada a organização e simulação de cada cenário. Visando uma avaliação ampla do modelo em estudo foram criados oito cenários distintos, a saber:

- Cenário 1: Completo, contendo todas as ações gênicas permitidas pelo simulador e sem o uso da inseminação artificial;
- Cenário 2: Completo, com todas as ações gênica e utilizando inseminação Artificial Exclusiva;
- Cenário 3: Exclusivamente aditiva e utilizando inseminação artificial Mista;
- Cenário 4: Ênfase na interação epistática e utilizando inseminação artificial Mista;
- Cenário 5: Completo, com todas as ações gênicas e utilizando inseminação artificial mista e um número reduzido de indivíduos;
- Cenário 6: Ênfase na interação epistática utilizando inseminação artificial mista, foram simuladas interações com até 3 marcadores;
- Cenário 7: Ênfase na interação epistática utilizando inseminação artificial mista, foram simuladas interações com até 4 marcadores;
- Cenário 8: Simulação do Girolando PS utilizando a opção de cruzamento B conforme Girolando (2017).

O processo de simulação, parâmetros e código são detalhados a seguir. Visando facilitar a visualização do LD, somente os SNPs causais em uma janela de 30 marcadores com 15 de cada lado, foram utilizados na geração do mapa de LD. As populações exibidas são as utilizadas nos experimentos do Capítulo 6.

4.1 Cenário 1 Completo, contendo todas as ações gênicas permitidas pelo simulador e sem o uso da inseminação artificial

O cenário 1 é composto de duas interações epistáticas dominantes entre os pares (100 e 200) e (1700 e 1900), os marcadores 400, 600 e 900 atuando como aditivos, os marcadores 1200 e 1400 como dominante e o marcador 1500 como recessivo. O código de geração do cenário 1 pode ser visto no Apêndice A. A seguir tem-se uma prévia de como os SNPs interagem bem como os valores dos efeitos de cada ação gênica ou *beta*.

```
list.ia<-list(c(-3,-3),0,0,0,3,3,-3,c(-3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))

list.snp<-list(c(100,200),400,600,900,1200,1400,1500,c(1700,1900))
list.snp<-c(list.snp, seq(1,num_snp)

beta=c(1.8,1.2,1.4,1.5,1.9,1.7,1.8,1.9,runif(1990,0,0.001))
```

O objetivo do primeiro cenário é avaliar a capacidade das ferramentas em selecionar corretamente os diversos marcadores envolvidos na geração do fenótipo, mesmo com ações gênicas diferentes.

A Figura 4.1 exibe a evolução do GEBV, as primeiras 100 gerações compreendem a parte histórica do processo de evolução. Como visto, existe uma manutenção no comportamento devido a ausência de pressão de seleção. A partir da geração 100 ocorre a escolha dos melhores machos, gerando um aumento no valor médio de GEBV de cada geração.

A Figura 4.2 mostra a evolução da herdabilidade e, como visto nas primeiras 100 gerações, o valor médio é próximo de 0,60, diminuindo com a seleção dos melhores machos para valores próximos a 0,35. A redução na herdabilidade pode gerar uma maior dificuldade para a seleção dos melhores marcadores.

A Figura 4.3 mostra o mapa de LD da população parental utilizada no cenário 1 da população recente. A Figura 4.4 exibe o LD da geração 4 em seleção. Após as 100 primeiras gerações ocorre a fixação do LD e o surgimento de blocos haplótipos, com

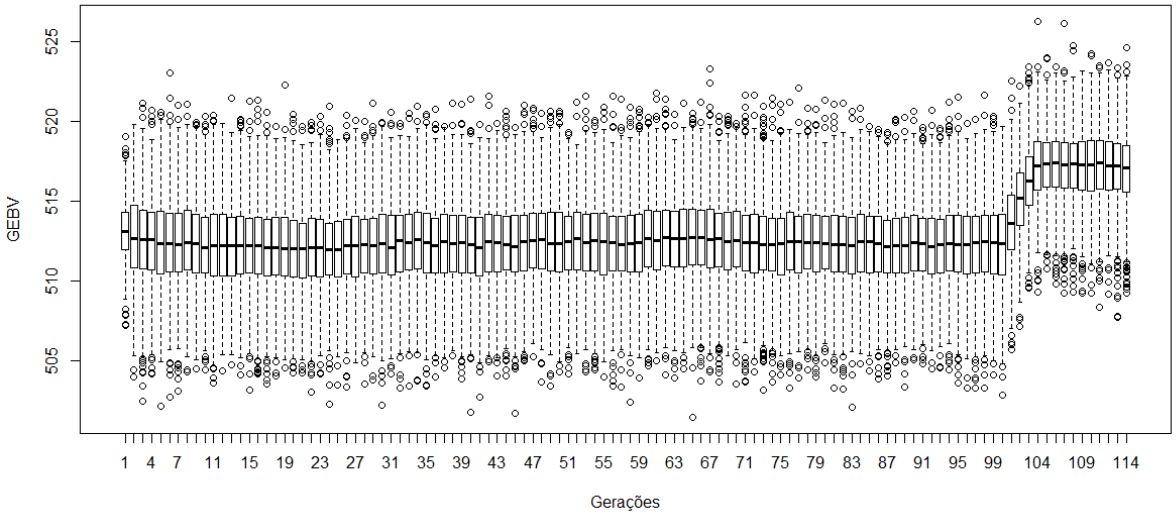


Figura 4.1: Evolução do GEBV do cenário 1.

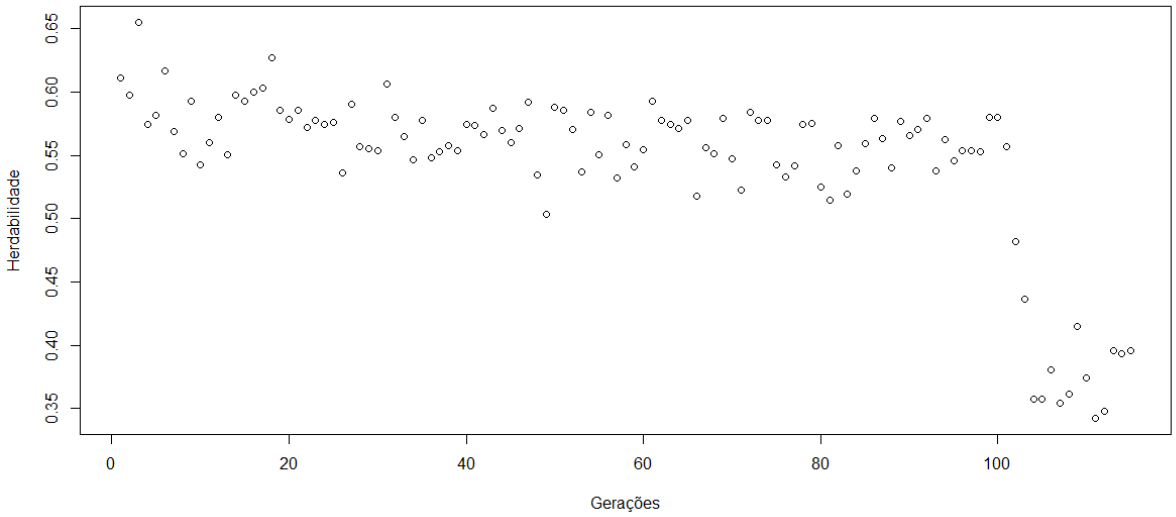


Figura 4.2: Evolução da herdabilidade no cenário 1.

o processo de seleção alterando a conformação dos blocos e o comportamento do LD, contudo ambos são similares aos encontrados na literatura.

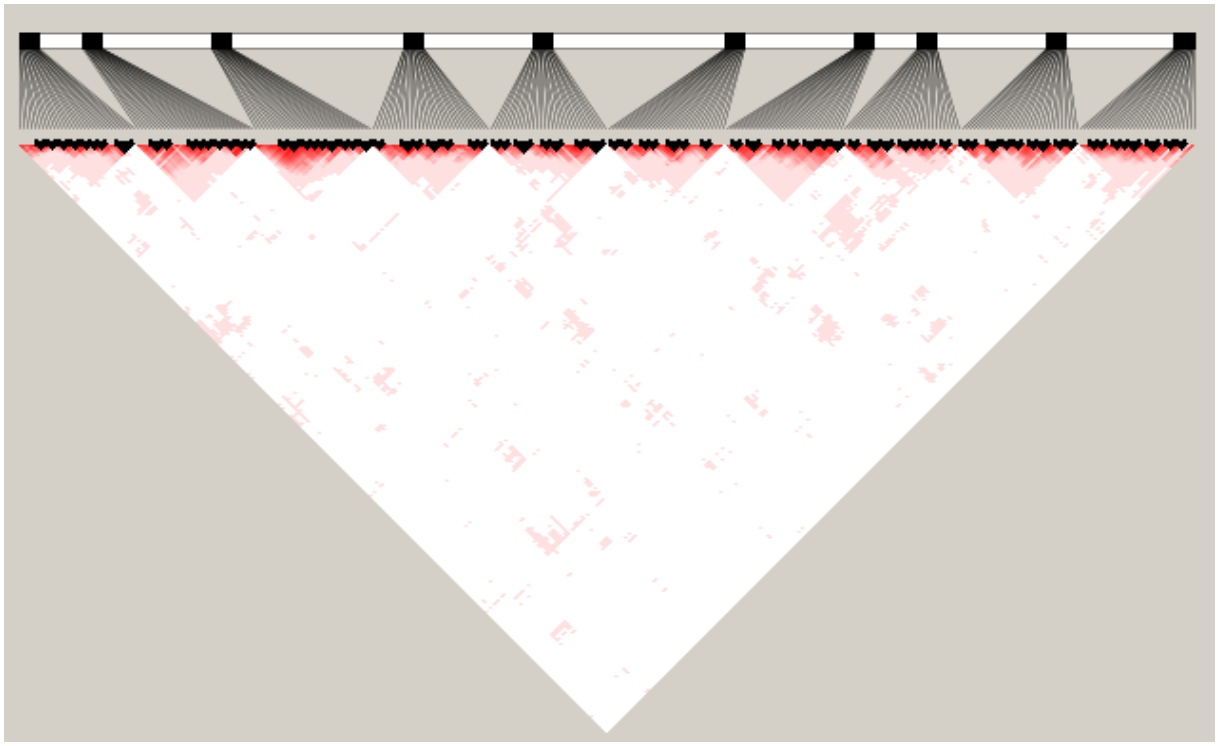


Figura 4.3: Mapa de LD da população parental utilizada no cenário 1.

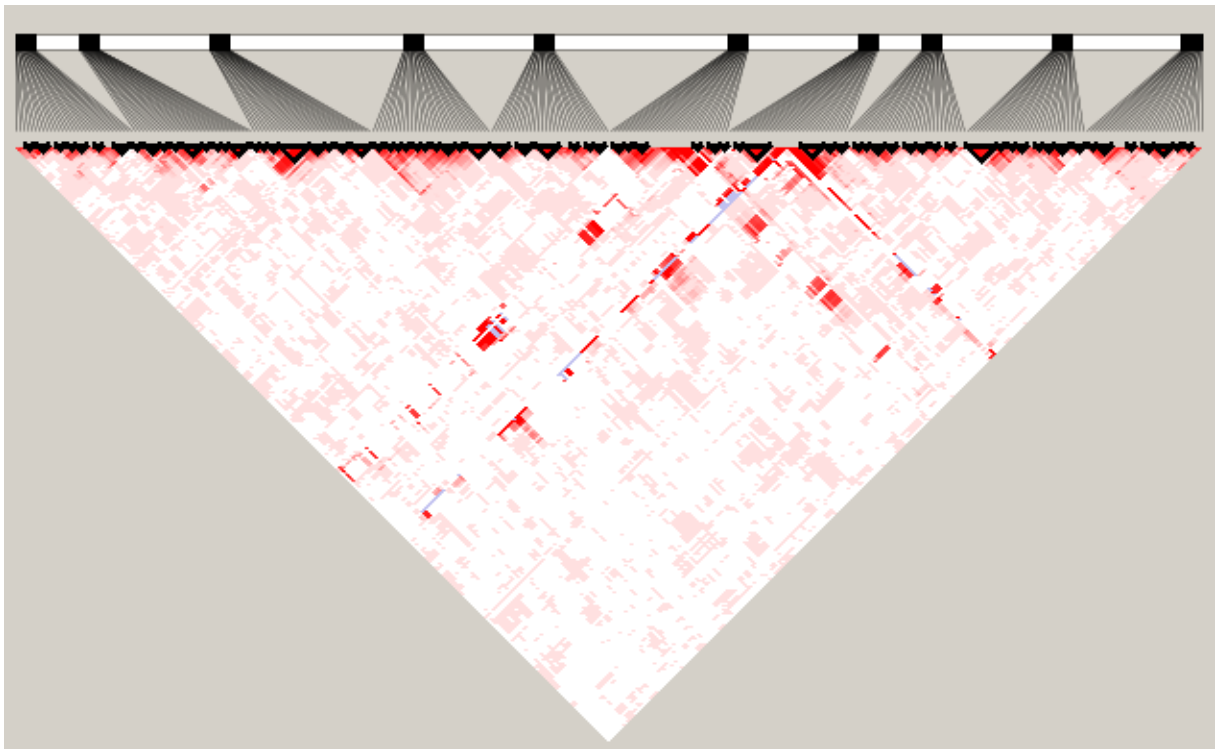


Figura 4.4: Mapa de LD da geração 4 em seleção no cenário 1.

4.2 Cenário 2 completo, com todas as ações gênicas e utilizando inseminação artificial exclusiva

O cenário 2 é composto de duas interações epistáticas dominantes duplas entre os pares (100 e 200) e (1700 e 1900), com os marcadores 400, 600 e 900 atuando como aditivos, os marcadores 1200 e 1400 como dominante e o marcador 1500 como recessivo. Os marcadores e os βs são iguais aos do cenário 1, porém com o uso do banco de sêmen. O código de geração do cenário 2 pode ser visto no Apêndice B.

A evolução do GEBV observada no cenário 2 é similar a do primeiro, conforme pode ser visto na Figura 4.5, porém a evolução do GEBV durante as gerações no processo de melhoramento é maior que no cenário 1.

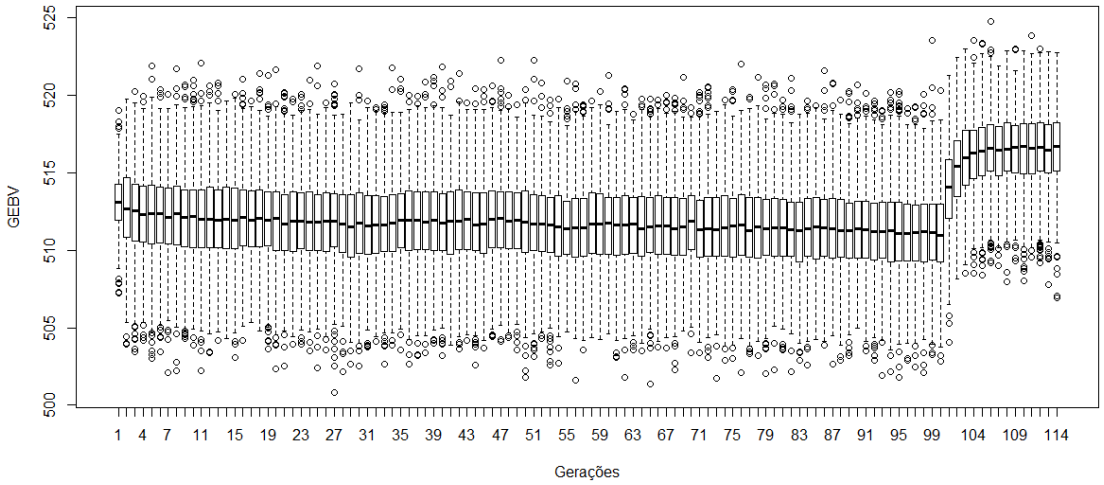


Figura 4.5: Evolução do GEBV no cenário 2.

O comportamento observado na Figura 4.6 indica o impacto da aplicação do banco de sêmen, pois a queda na herdabilidade é menor que a observada no cenário 1. A redução na queda da herdabilidade pode ser devido a variabilidade adicional oriunda do banco de sêmen.

O LD e os blocos observados na Figura 4.7 são similares aos do primeiro cenário mas, após o processo de seleção, Figura 4.8, é nítida a diferença entre os mapas de LD dos dois cenários. É possível notar no mapa de LD da geração 4 do cenário 1 a presença de alguns marcadores com MAF muito baixa. O uso do banco de sêmen trouxe variabilidade aos indivíduos, dessa forma não foi observado marcadores com MAF baixa no cenário 2.

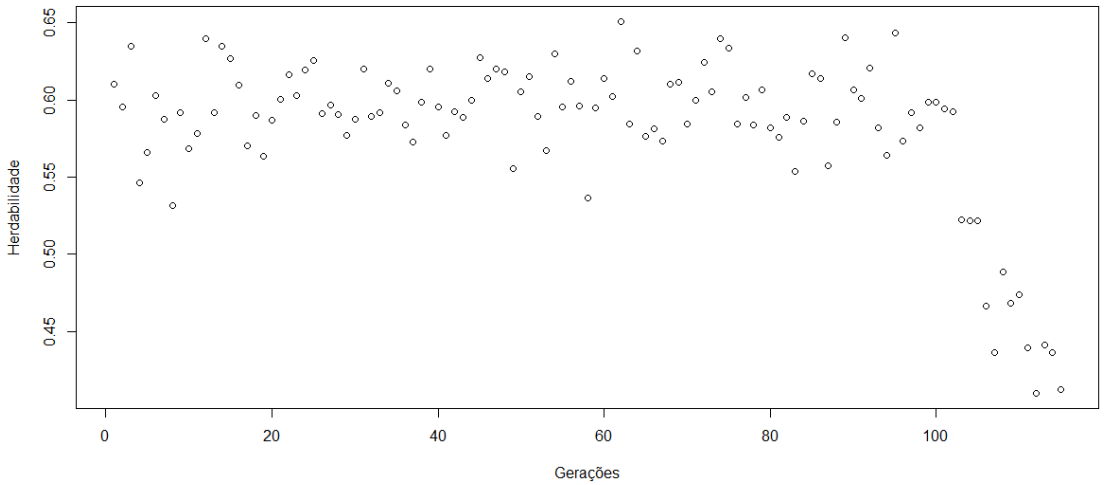


Figura 4.6: Evolução da herdabilidade no cenário 2.

A diferença mais significativa entre os dois cenários está no mapa de LD da última geração em seleção, bem como na herdabilidade, que teve uma queda menor.

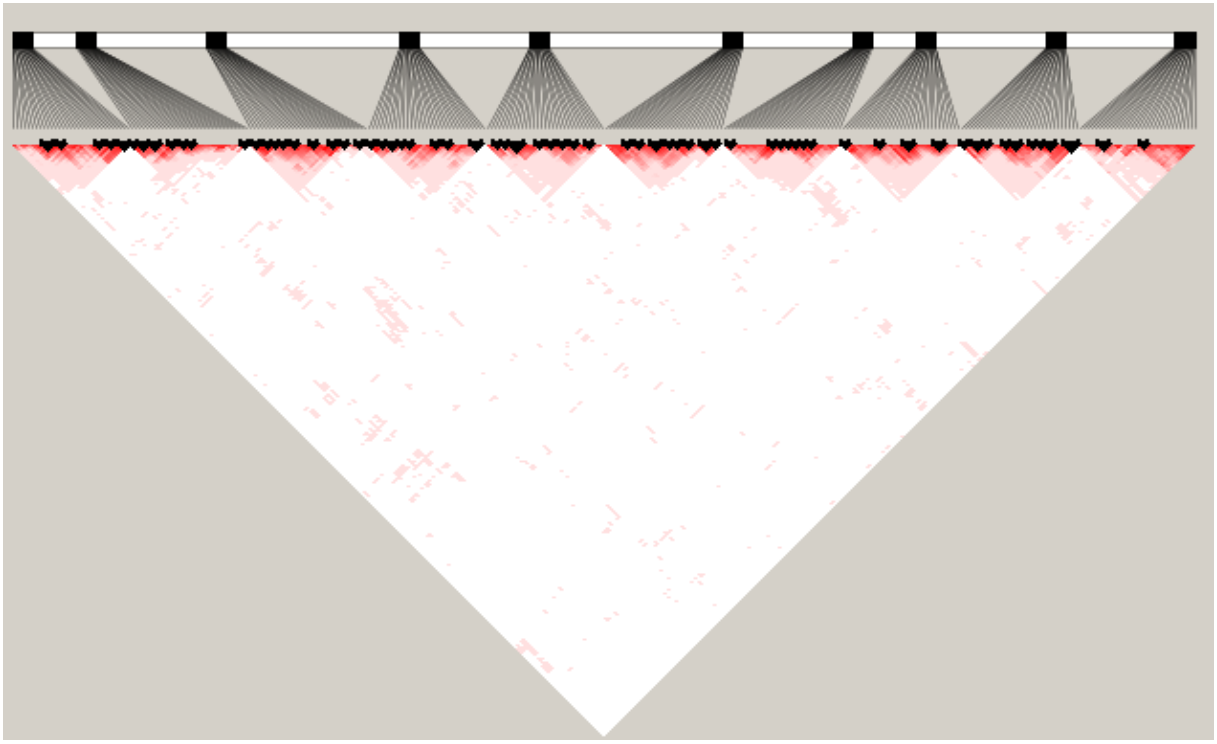


Figura 4.7: Mapa de LD da população parental utilizada no cenário 2.

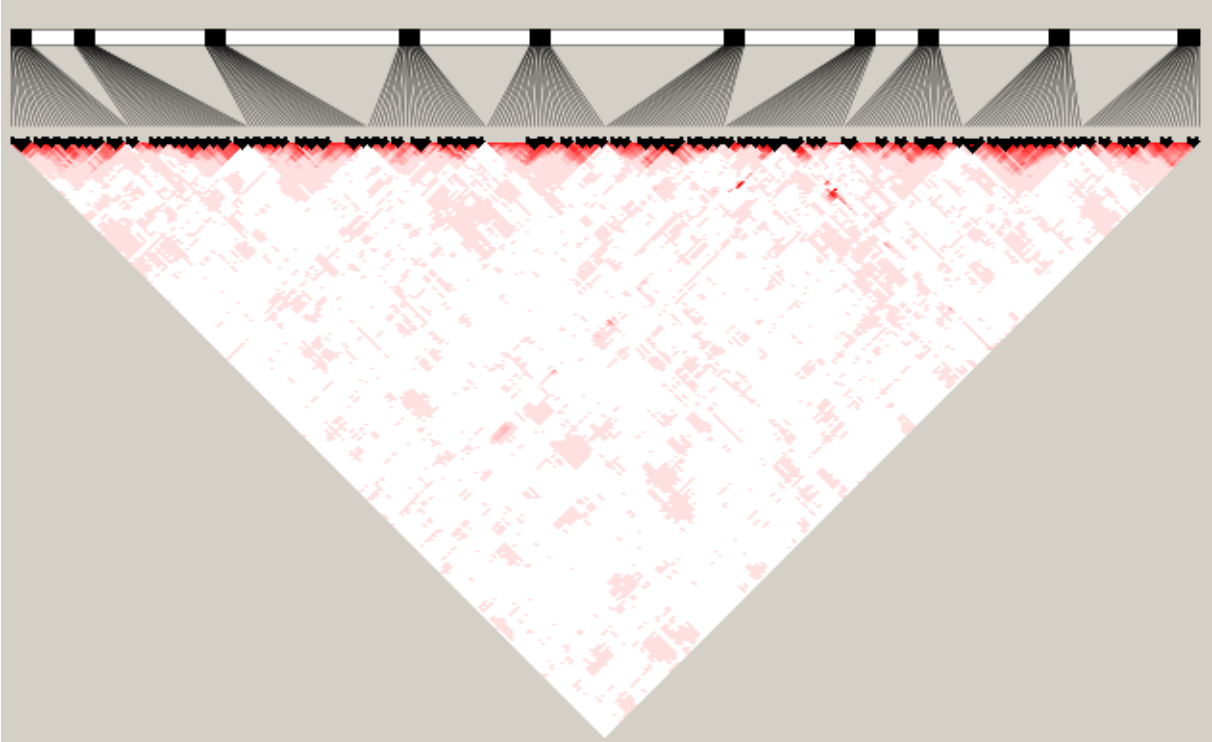


Figura 4.8: Mapa de LD da geração 4 em seleção no cenário 2.

4.3 Cenário 3 - Exclusivamente aditiva e utilizando inseminação artificial mista

O cenário 3 foi montado exclusivamente com ação gênica aditiva e uso de inseminação artificial mista, ou seja, com machos oriundos do banco de sêmen e do rebanho. O código de geração do cenário 3 pode ser visto no Apêndice C. Este cenário servirá de comparativo, pois muitos trabalhos de seleção genômica utilizam modelos lineares por acreditar que a informação aditiva é abundante e suficiente para explicar a expressão gênica.

A evolução do GEBV é maior que a observada nos cenários anteriores, conforme mostra a Figura 4.9. O comportamento pode ser explicado, pois a seleção dos marcadores aditivos se mostra mais simples que de outras ações gênicas.

A variação observada na herdabilidade é maior que nos outros dois cenários, conforme pode ser visto na Figura 4.10. A queda na herdabilidade ocorre mesmo sem a seleção de animais, o que pode ser explicado pelo uso do machos do banco de sêmen, pois o uso de uma quantidade menor de pais diminuirá a variabilidade genética, destacando o impacto do ambiente na variação fenotípica.

O mapa de LD observado na Figura 4.11 é similar ao observado no cenários anteriores,

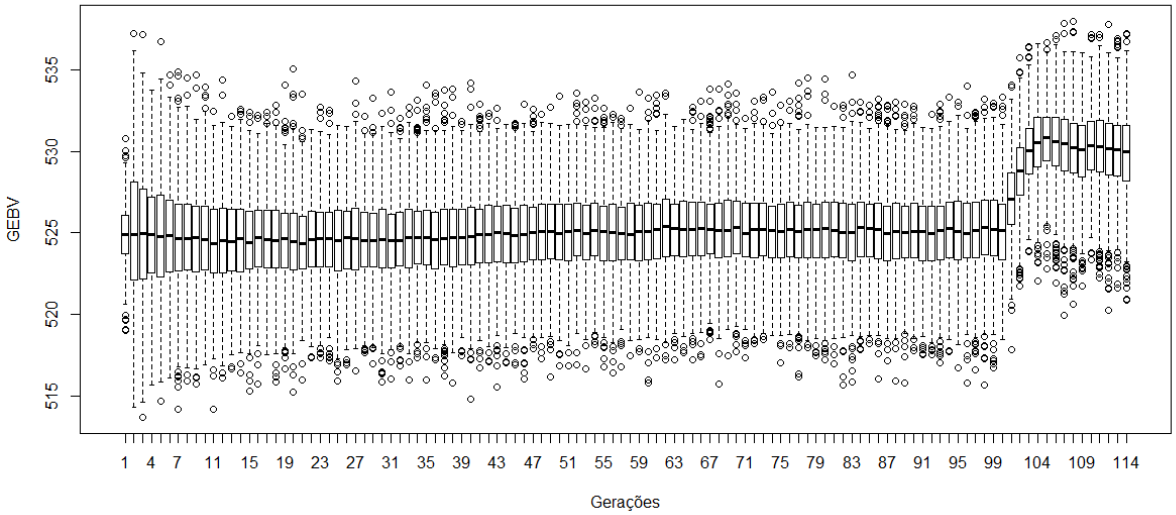


Figura 4.9: Evolução do GEBV do cenário 3.

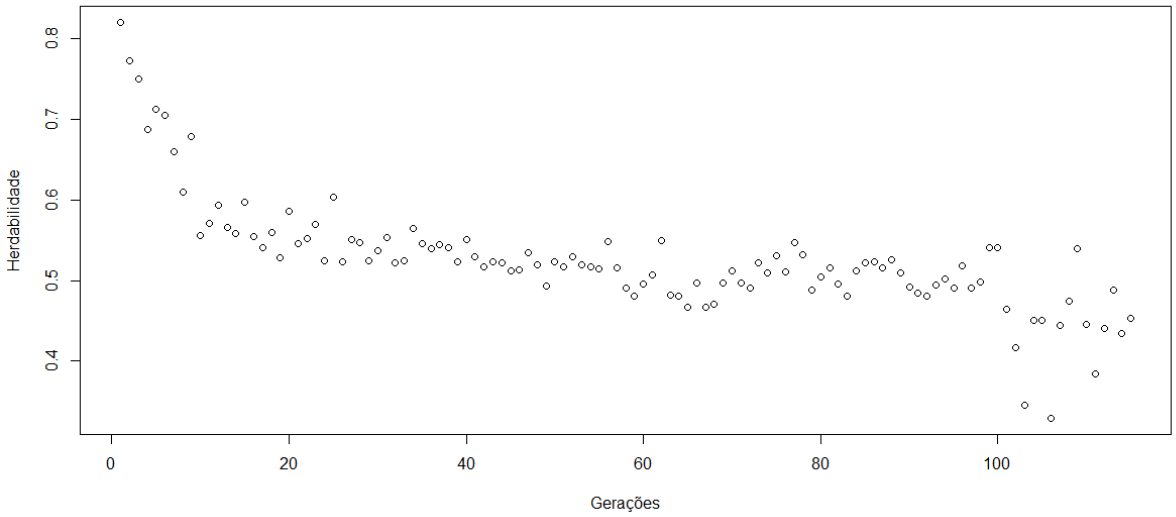


Figura 4.10: Evolução da herdabilidade no cenário 3.

mas a expectativa é que com a seleção ocorra uma diminuição brusca da MAF. Contudo, conforme pode ser observar na Figura 4.12, um comportamento similar ao visto no cenário 2, sendo nítido o impacto do uso da inseminação artificial na manutenção da MAF. Vale ressaltar que pequenas diferenças podem gerar resultados diferentes nos processos de seleção de atributos e regressão. Espera-se que o modelo proposto seja capaz de capturar essas sutis diferenças.

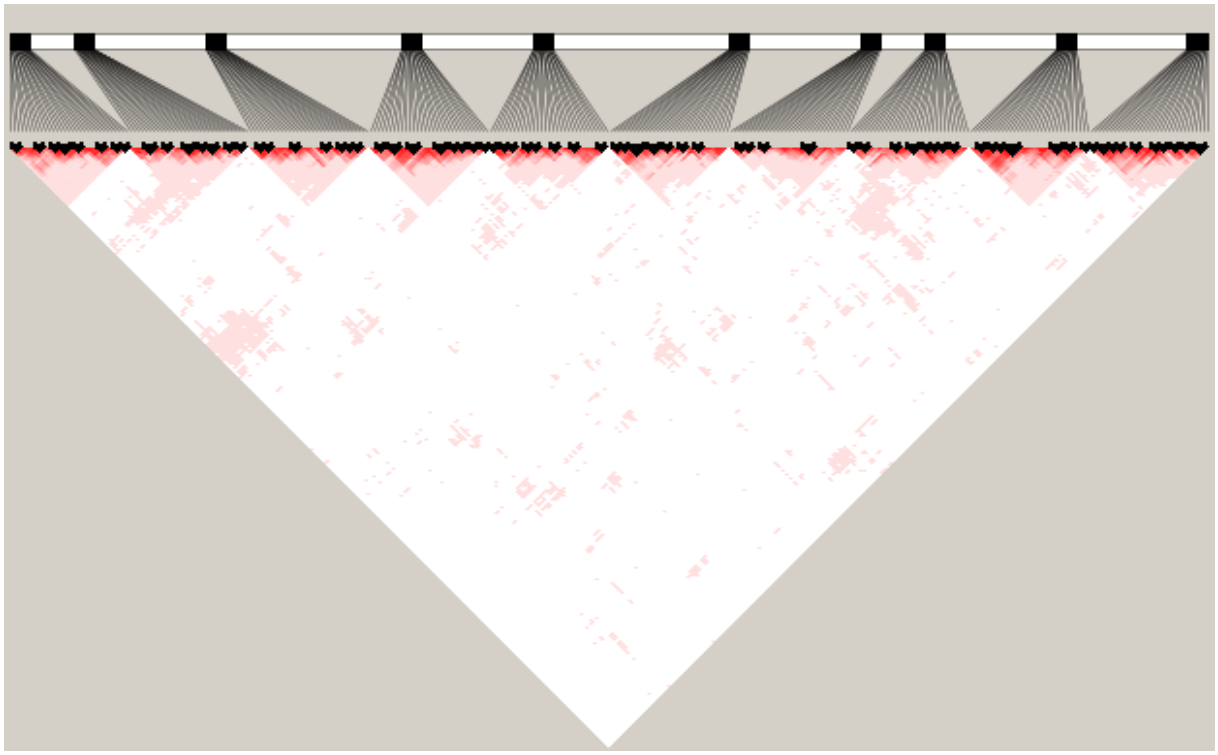


Figura 4.11: Mapa de LD da população parental utilizada no cenário 3.

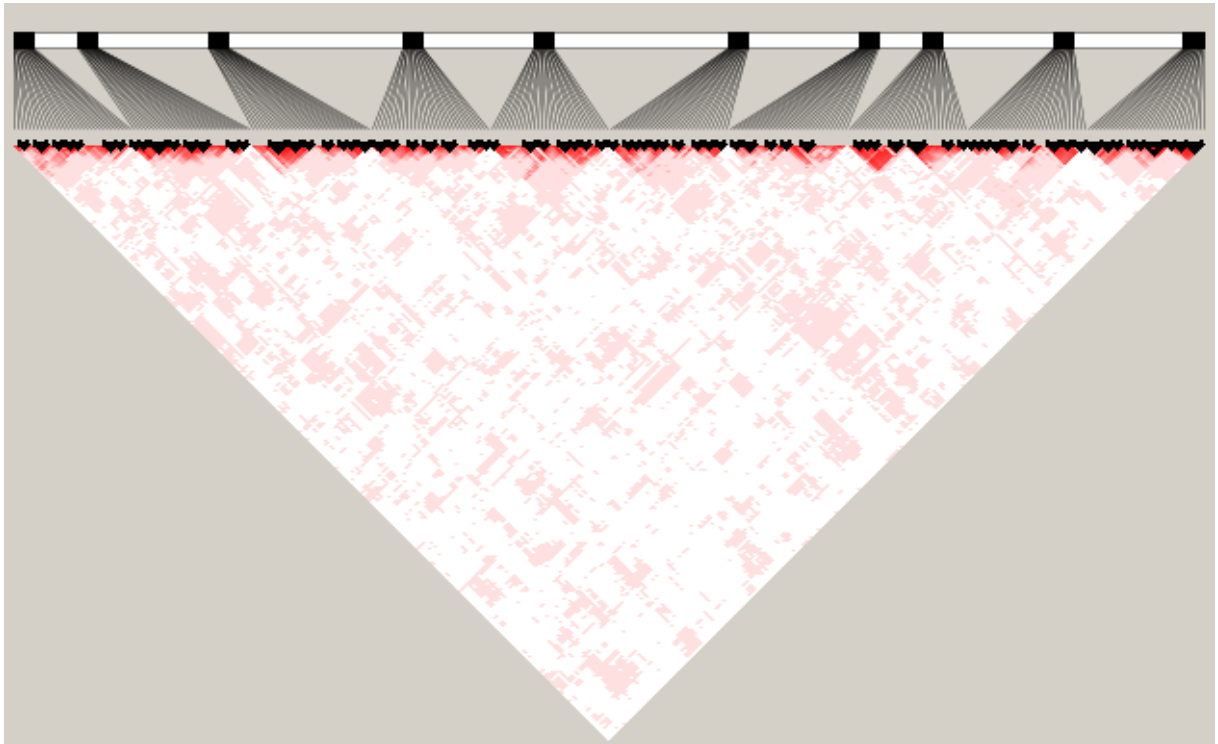


Figura 4.12: Mapa de LD da geração 4 em seleção no cenário 3.

4.4 Cenário 4 - Ênfase na interação epistática utilizando inseminação artificial mista

O cenário 4 é o primeiro a focar nas interações epistáticas, sendo nesse cenário de ordem 2. As interações ocorrem entre as seguintes duplas de marcadores: 100 e 200; 400 e 600; 900 e 1200; 1400 e 1500; e 1700 e 1900. O Apêndice D exibe o código de geração do conjunto de dados simulados.

A evolução do GEBV exibida na Figura 4.13 é menor que a observada nos cenários anteriores, chegando no máximo a 510 de média, com os valores de média também menores que o observado nos cenários anteriores. As interações impactaram diretamente na dinâmica de geração do fenótipo.

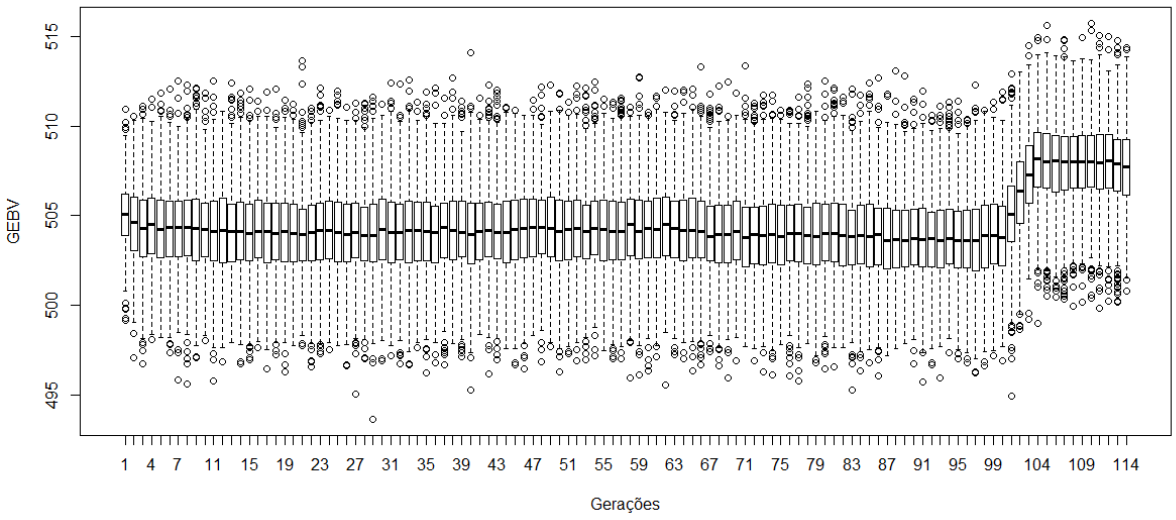


Figura 4.13: Evolução do GEBV do cenário 4.

O comportamento observado pela herdabilidade, Figura 4.14, é bem diferente se comparado com outros cenários, sendo nítida a variação ao longo de todo o processo de simulação. A seleção dos melhores animais estabilizou a variação da herdabilidade.

O mapa de LD exibido na Figura 4.15 é similar ao dos cenários anteriores. Contudo, na Figura 4.16 é possível observar a diminuição da MAF, que são as linhas em azul, dos marcadores em desequilíbrio com o marcador 1400. A redução da MAF gera uma diminuição na capacidade de identificação desses marcadores por ferramentas de seleção de atributos, e também impacta na geração de modelos para a regressão.

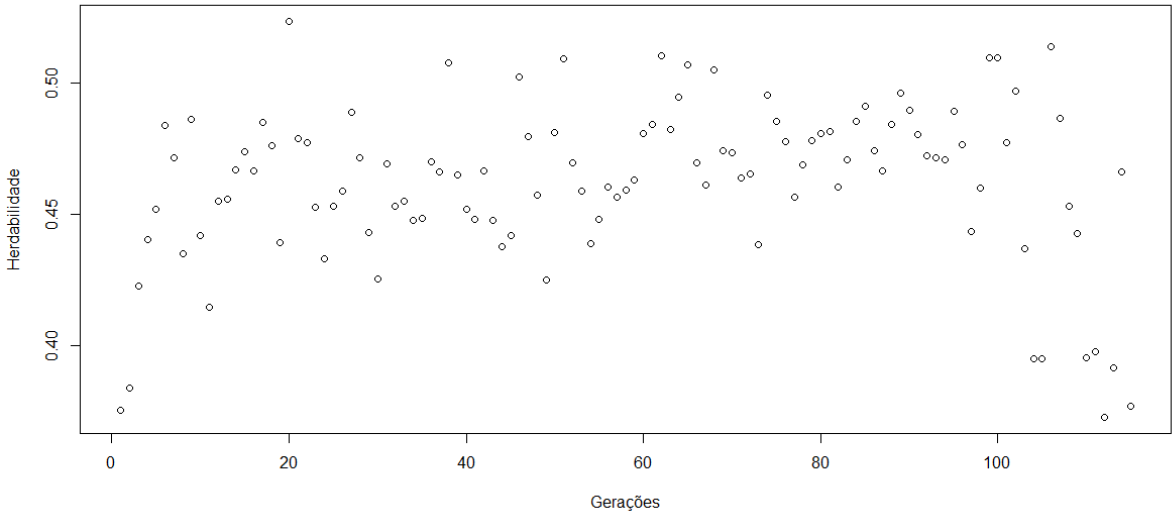


Figura 4.14: Evolução da herdabilidade no cenário 4.

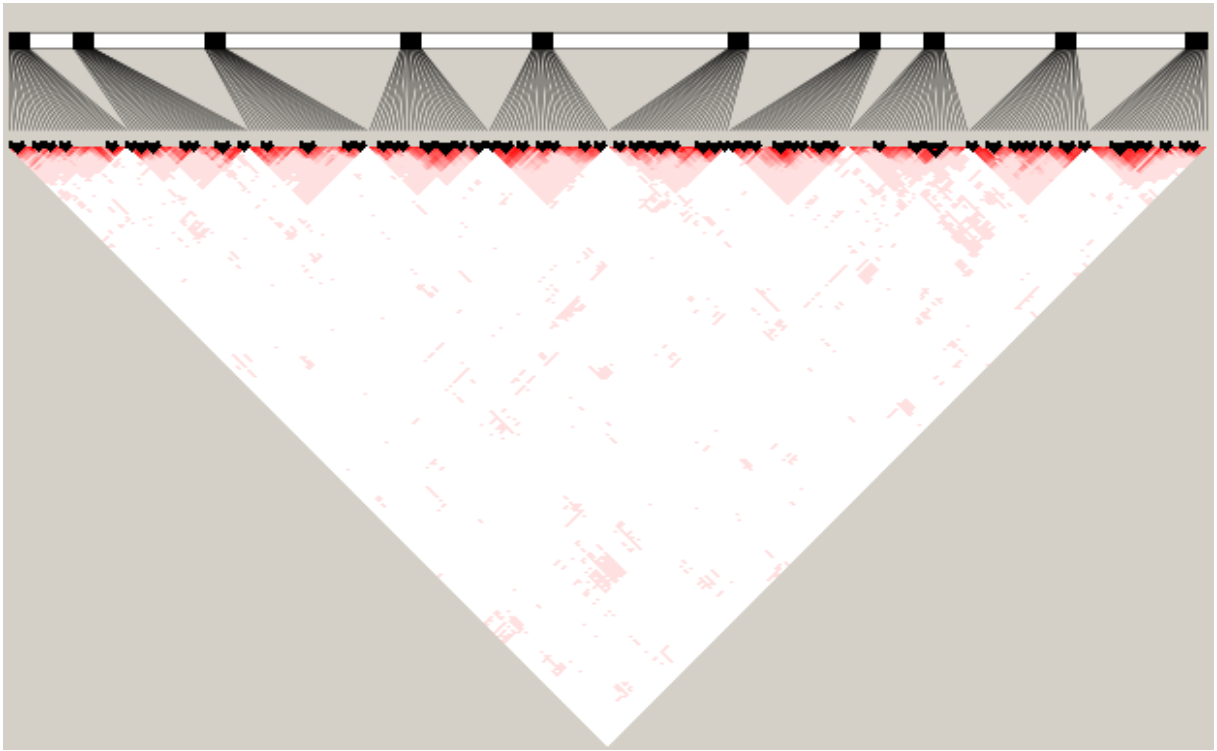


Figura 4.15: Mapa de LD da população parental utilizada no cenário 4.

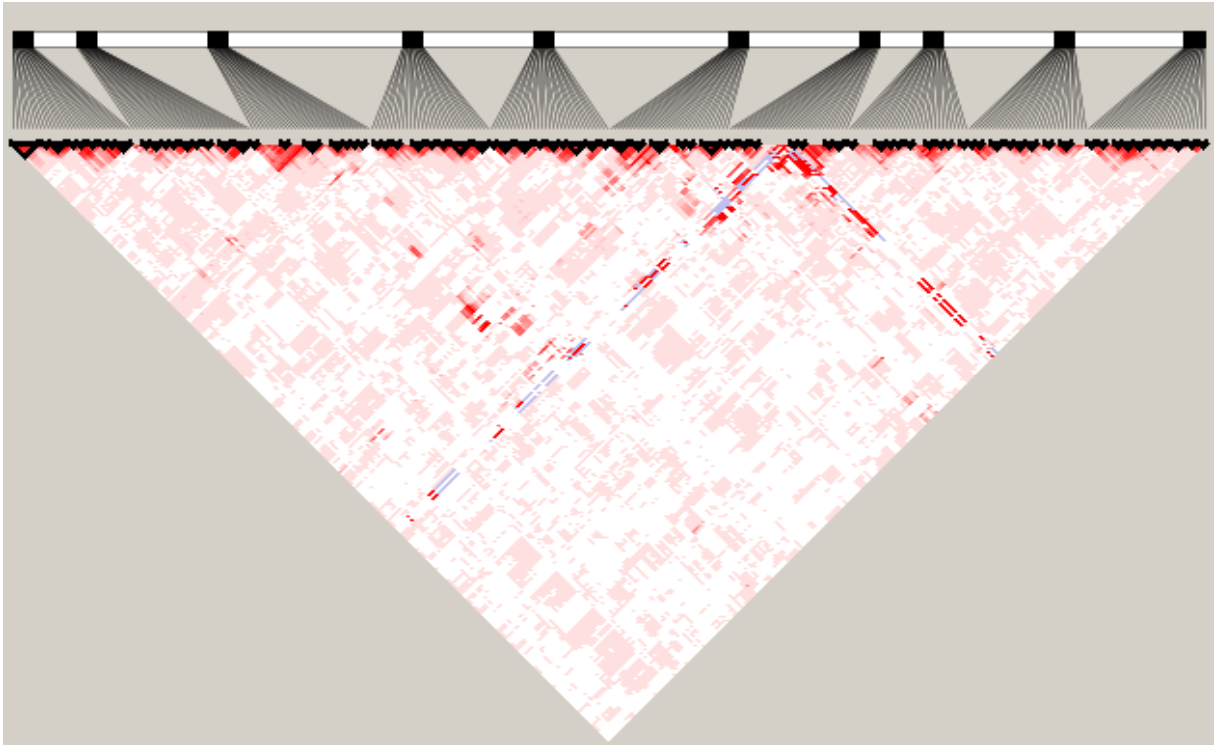


Figura 4.16: Mapa de LD da 4 geração em seleção no cenário 4.

4.5 Cenário 5 Completo, com todas as ações gênicas e utilizando inseminação artificial e com um número reduzido de indivíduos

O cenário 5 é uma variação do cenário 2, ou seja, possui todas as ações gênicas disponíveis no S4GS, contudo com um número reduzido de indivíduos, sendo esse um dos pontos de interesse na pesquisa dessa tese. O código de geração do cenário 5 pode ser visto no Apêndice E.

A redução no número de indivíduos não gerou impacto significativo ns dinâmica do GEBV, mas é possível observar uma variação nos valores de média e máximo ao longo das gerações da população histórica, conforme mostra a Figura 4.17. O processo de melhoramento se mostrou estável, porém com um crescimento menor se comparado com outros cenários com populações maiores.

A evolução da herdabilidade pode ser vista na Figura 4.18, onde é possível perceber um padrão a cada 20 gerações, onde a herdabilidade aumenta, decrescendo em seguida. Novamente, o processo de melhoramento estabiliza a herdabilidade.

O mapa de LD da geração parental, visto na Figura 4.19, e o da geração 4, exibido

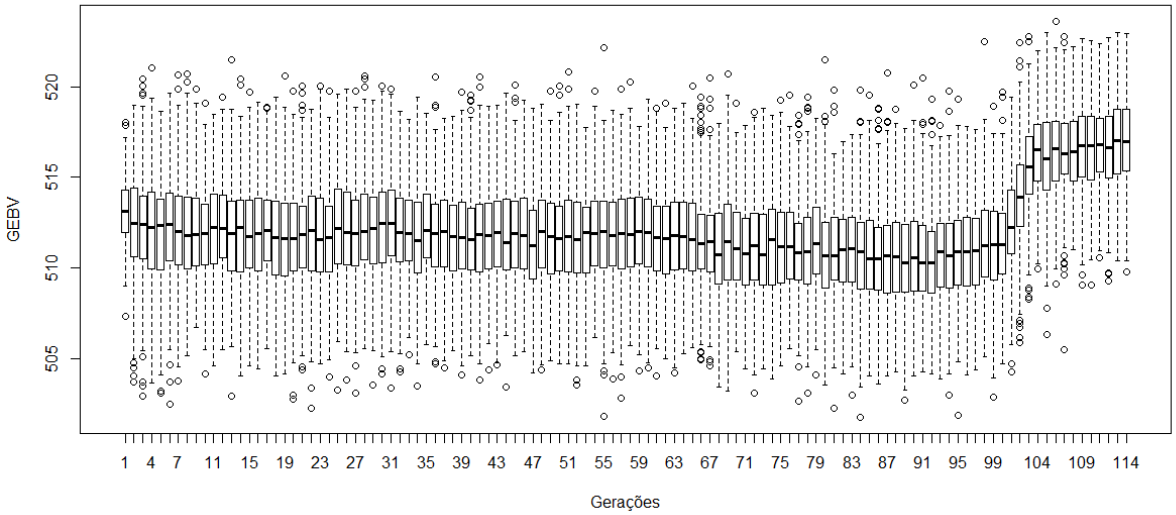


Figura 4.17: Evolução do GEBV do cenário 5.

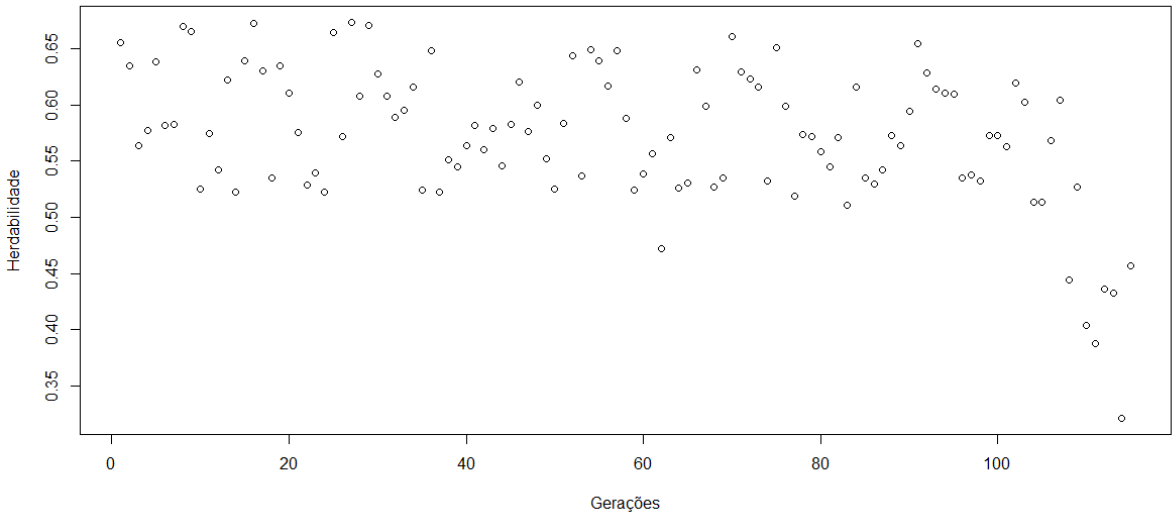


Figura 4.18: Evolução da herdabilidade no cenário 5.

na Figura 4.20, não demonstram diferenças significativas se comparado com o cenário 2. O destaque fica por conta de uma pequeno grupo de marcadores desequilibrados com o marcador 1400, onde é possível notar uma redução da MAF, comportamento observado no cenário 1.

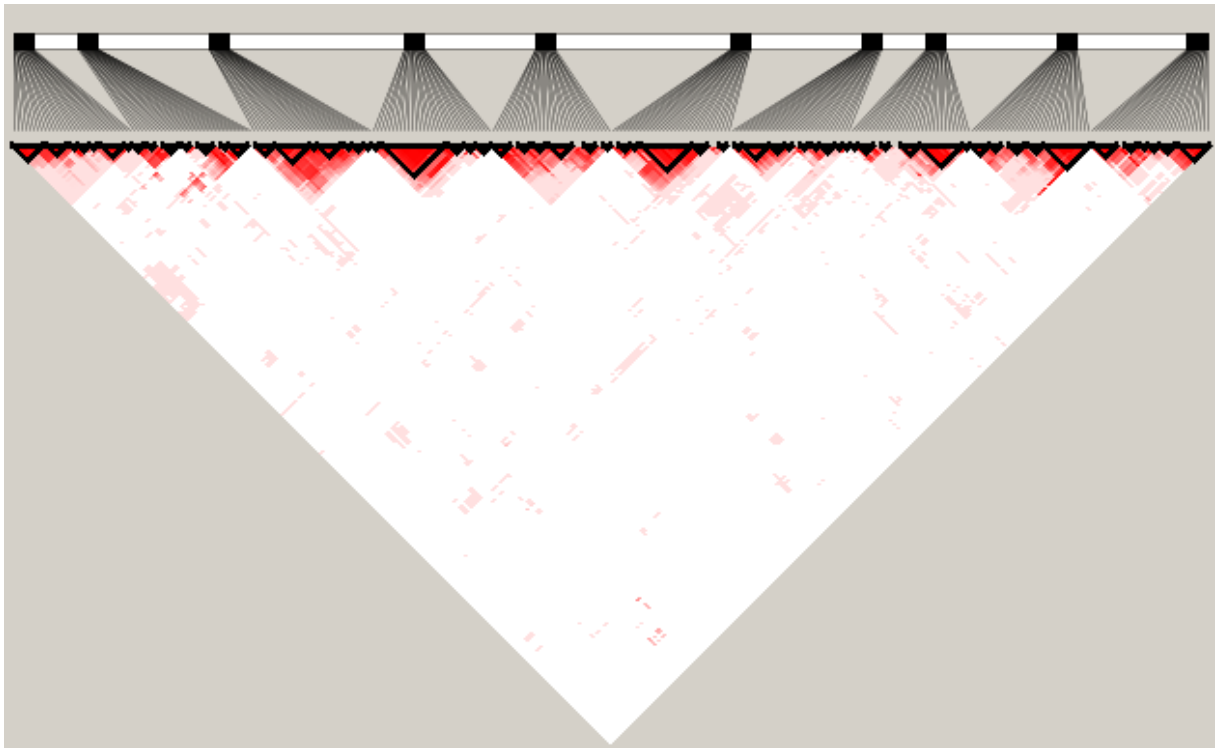


Figura 4.19: Mapa de LD da população parental utilizada no cenário 5.

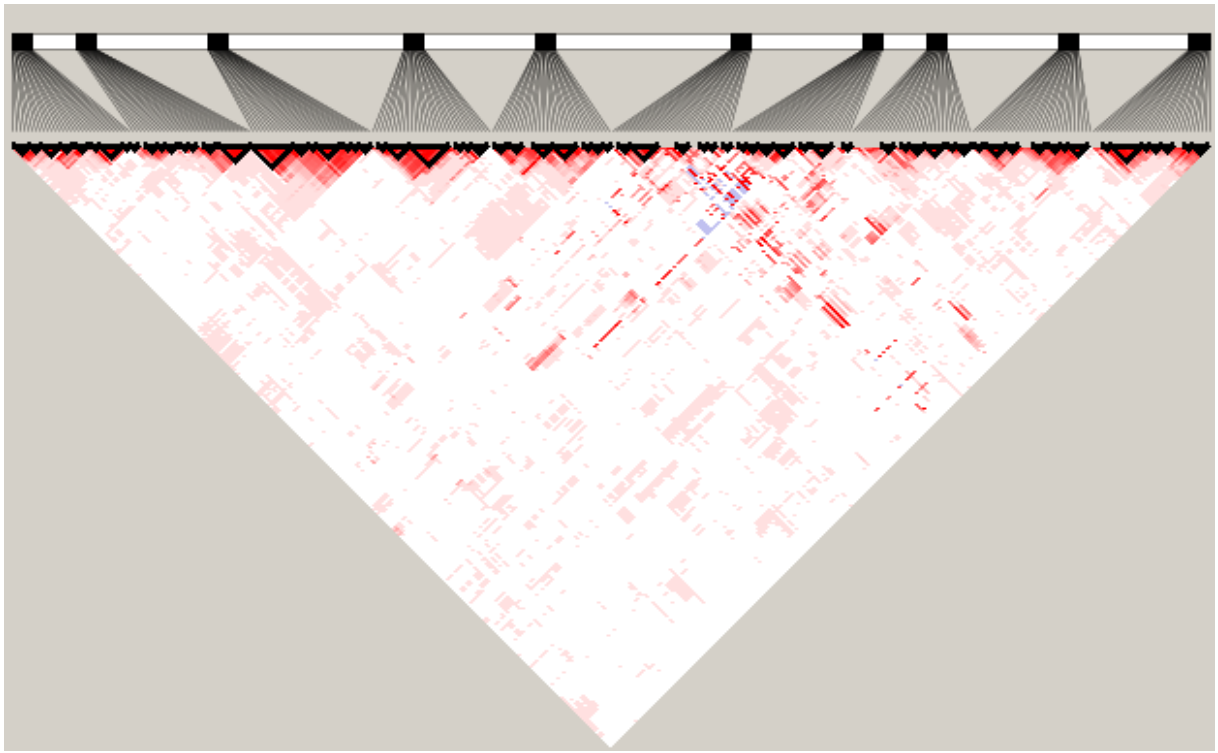


Figura 4.20: Mapa de LD da geração 4 em seleção no cenário 5.

4.6 Cenário 6 - Interação epistática de ordem 3 com o uso de inseminação artificial mista

As interações epistáticas simuladas no cenário 6 combinam 3 marcadores, sendo eles: 100, 200 e 400; 900, 1200 e 1400; 1500, 1700 e 1900. Além disso, o marcador 600 possui efeito aditivo. O Apêndice F exibe o código utilizado para a geração do conjunto de dados simulados.

A evolução do GEBV, vista na Figura 4.21, é diferente da observada nos cenários anteriores, possuindo um crescimento menor e com uma distância maior entre os valores de mínimo e máximos.

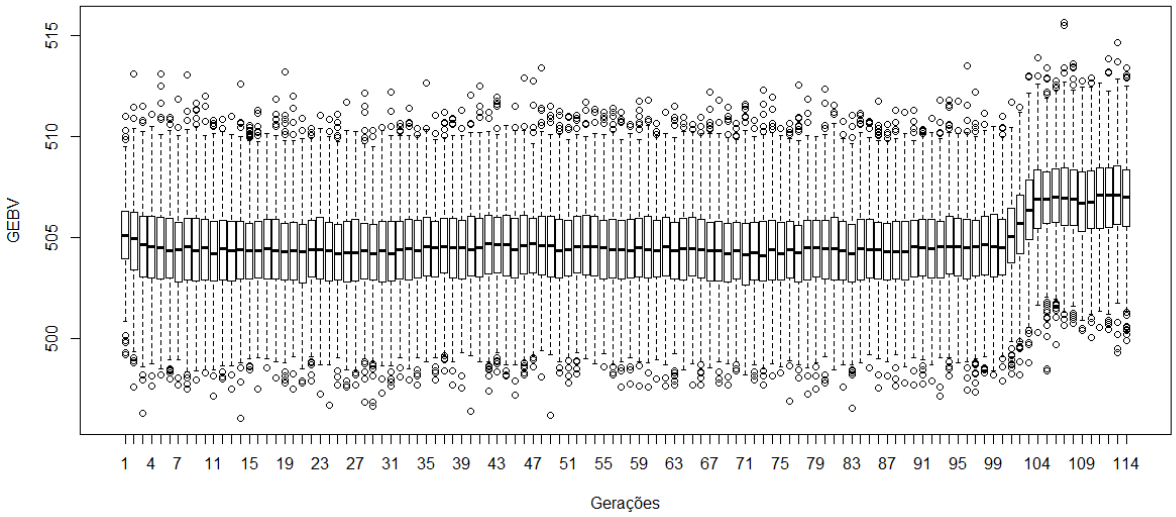


Figura 4.21: Evolução do GEBV do cenário 6.

O comportamento observado pela herdabilidade é diferente do visto nos cenários anteriores, Figura 4.22, apresentando uma variação maior. Os valores de herdabilidade obtidos são menores que nos outros cenários, provavelmente como reflexo da epistasia. É possível observar um comportamento cíclico como visto no cenário 5, porém ocorrendo a cada 10 gerações.

O mapa de LD da população parental, visto na Figura 4.23, não demonstra diferenças significativas se comparado com os cenários anteriores. Os gráficos de LD de todas as gerações parentais são similares, pois o processo de simulação da população histórica é idêntico variando o erro, os pontos de quebra da recombinação e as ações gênicas.

A dinâmica da 4^a geração do processo de melhoramento, exibido na Figura 4.24,

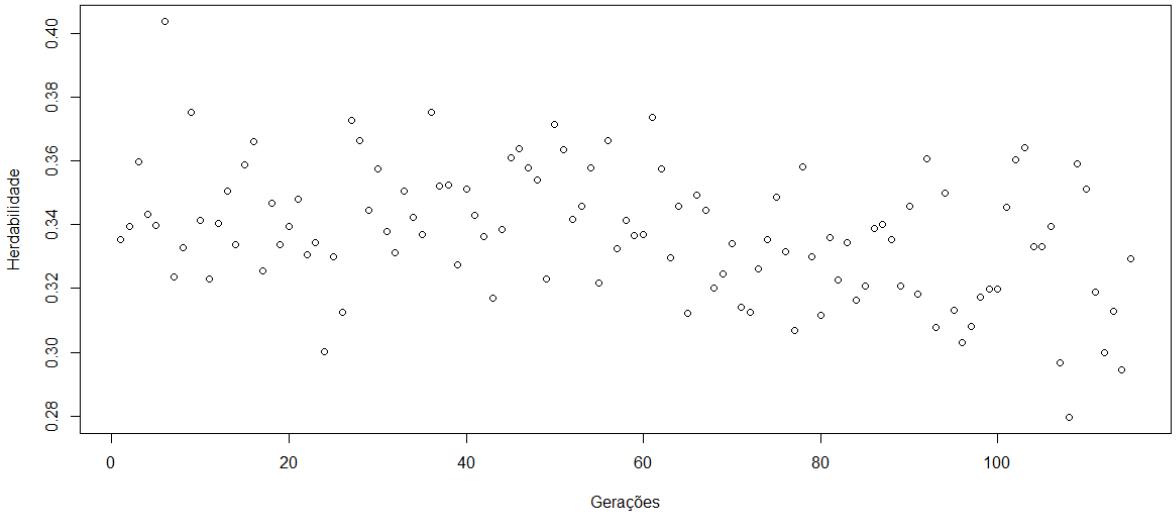


Figura 4.22: Evolução da herdabilidade no cenário 6.

é diferente da observada nos outros cenários, demonstrando uma dispersão menor e o surgimento de novos blocos. A interação entre marcadores, certamente, trouxe impactos também nos mapas de LD.

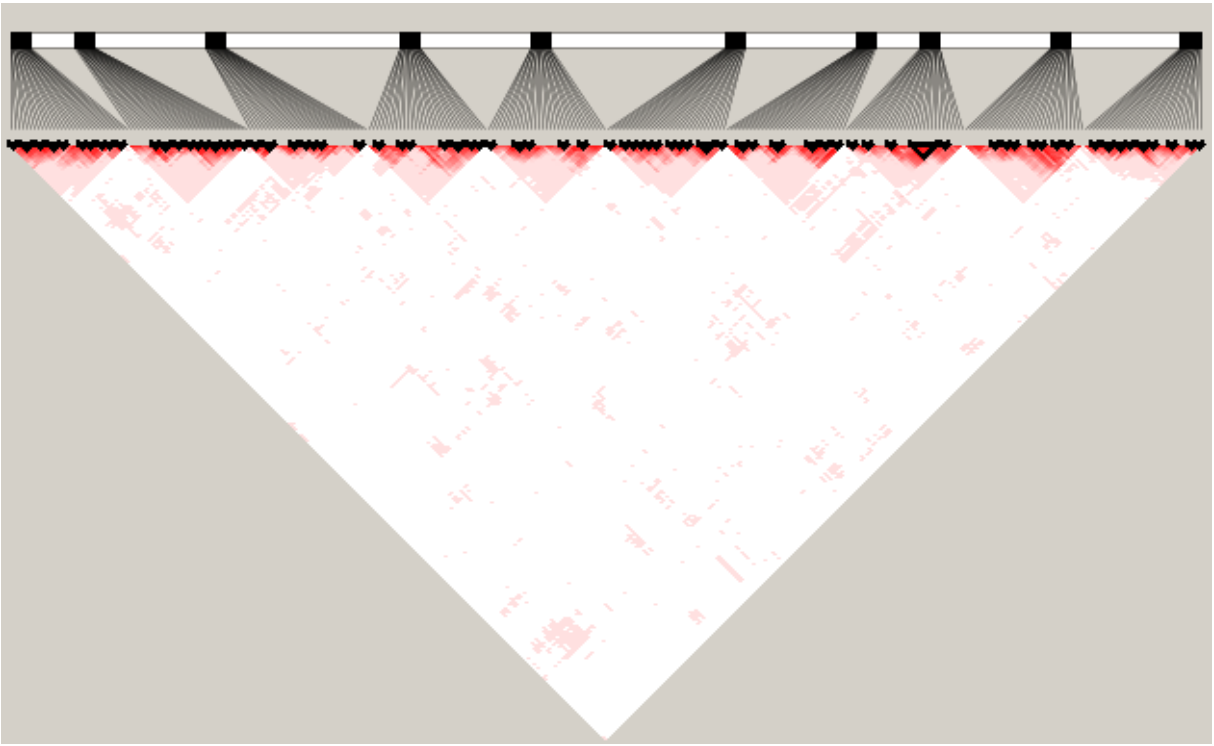


Figura 4.23: Mapa de LD da população parental utilizada no cenário 6.

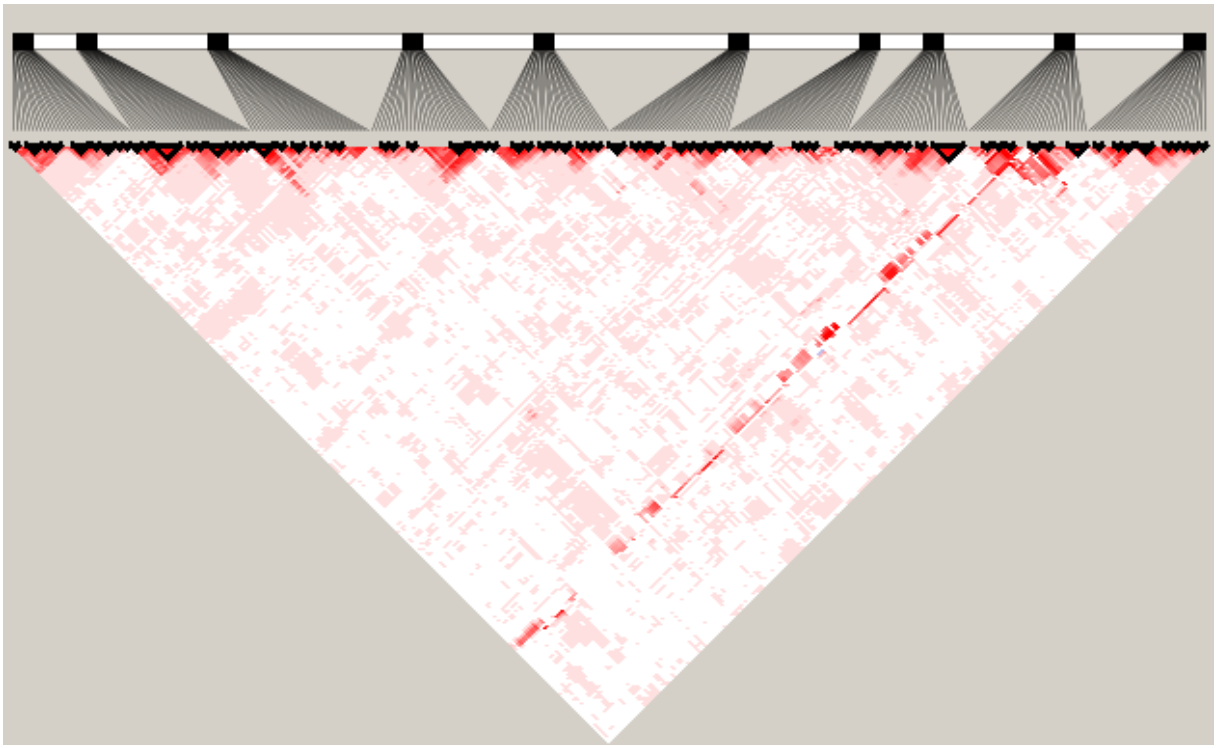


Figura 4.24: Mapa de LD da geração 4 em seleção no cenário 6.

4.7 Cenário 7 - Interação epistática de ordem 4 utilizando inseminação artificial mista

O cenário 7 é o último desse estudo com foco nas interações epistáticas, que foram mapeadas da seguinte forma: 100, 200, 400 e 600; e 1400, 1500, 1700 e 1900 gerando interações de ordem 4, e uma interação de ordem 2 entre os marcadores 900 e 1200. O código de geração do cenário 7 pode ser visto no Apêndice G.

A evolução do GEBV, observada na Figura 4.25, é similar a vista no cenário 6, sendo possível notar um incremento menor durante o processo de melhoramento. Os valores médios são os menores entre todos os cenários.

A variação da herdabilidade apresentada na Figura 4.26 é também a menor entre todos os cenários, com os valores de herdabilidade são menores de que nos cenários anteriores, sendo possível perceber o impacto das ações gênicas na herdabilidade.

A Figura 4.27 mostra o mapa de LD da população parental utilizada no cenário 7, e assim como nos cenários anteriores, o mapa exibe o LD e os blocos gerados pela população histórica. A posição dos blocos é diferente em cada cenário e geração, onde, mesmo com os mesmos parâmetros, cada simulação gerou mapas de LD e blocos distintos.

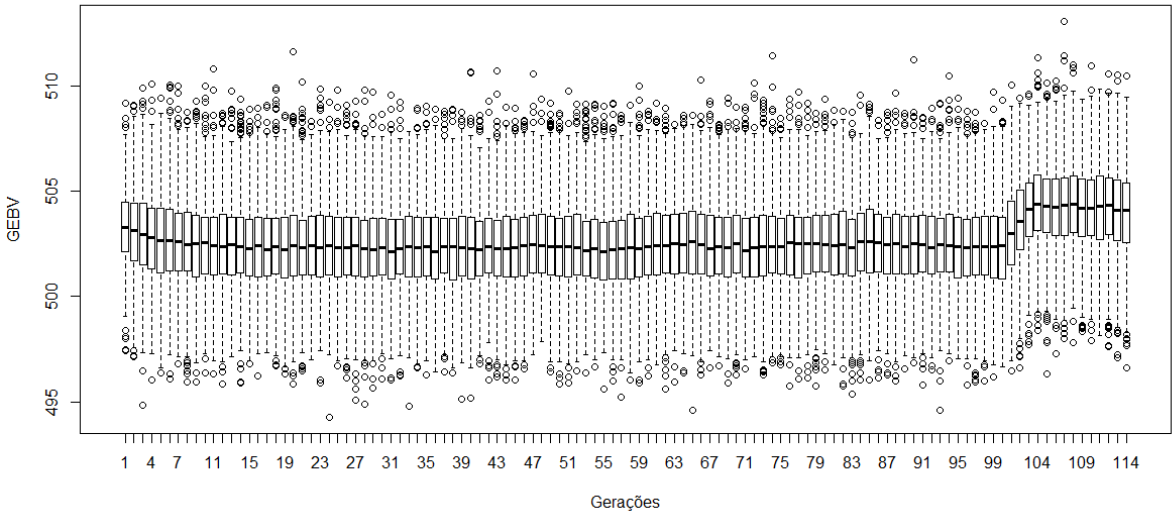


Figura 4.25: Evolução do GEBV do cenário 7.

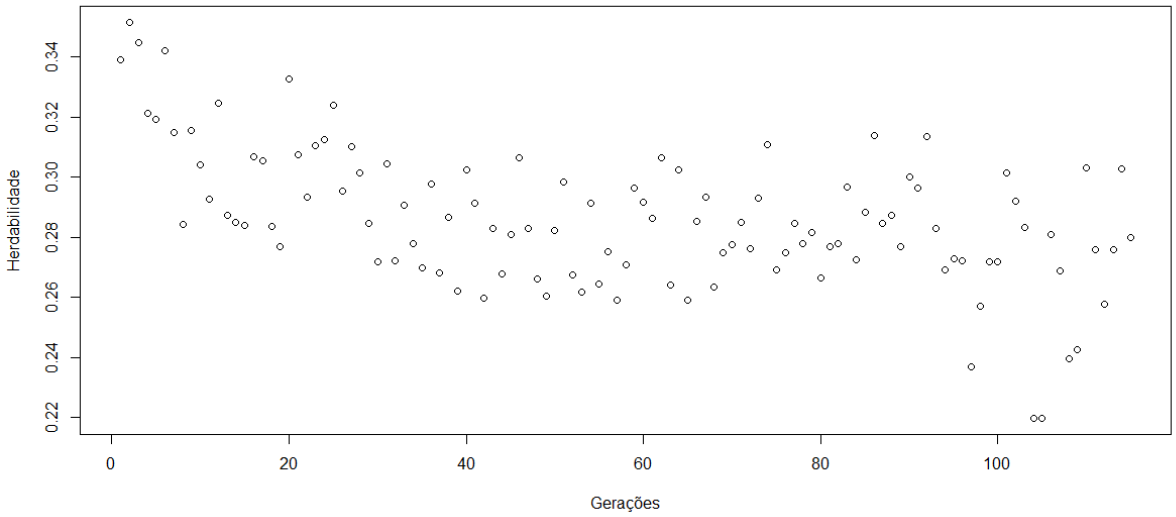


Figura 4.26: Evolução da herdabilidade no cenário 7.

A Figura 4.28 exibe o LD da 4^a geração em seleção. O mapa exibido na Figura 4.28 mostra uma menor dispersão dos marcados em LD se comparado aos exibidos na geração parental, bem como uma manutenção de boa parte dos blocos haplótipos. O marcadores desequilibrados com o marcador 1200 exibiram baixa frequência alélica após o melhoramento genético.

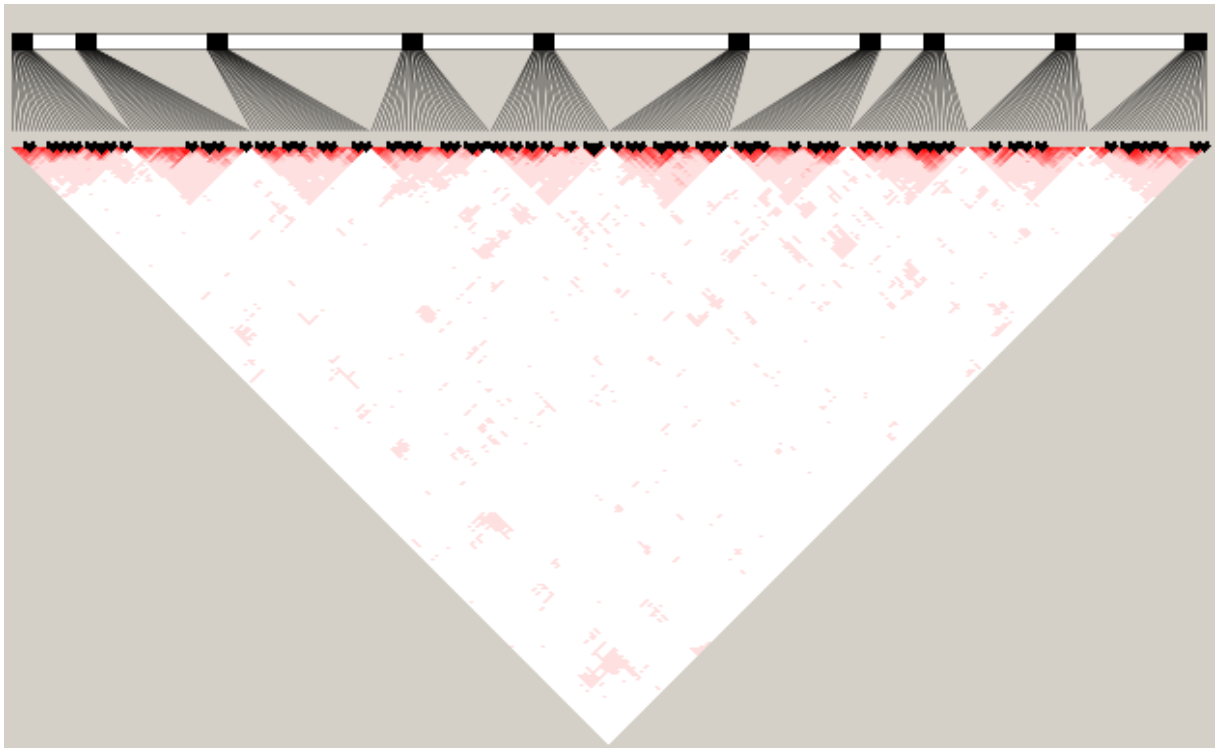


Figura 4.27: Mapa de LD da população parental utilizada no cenário 7.

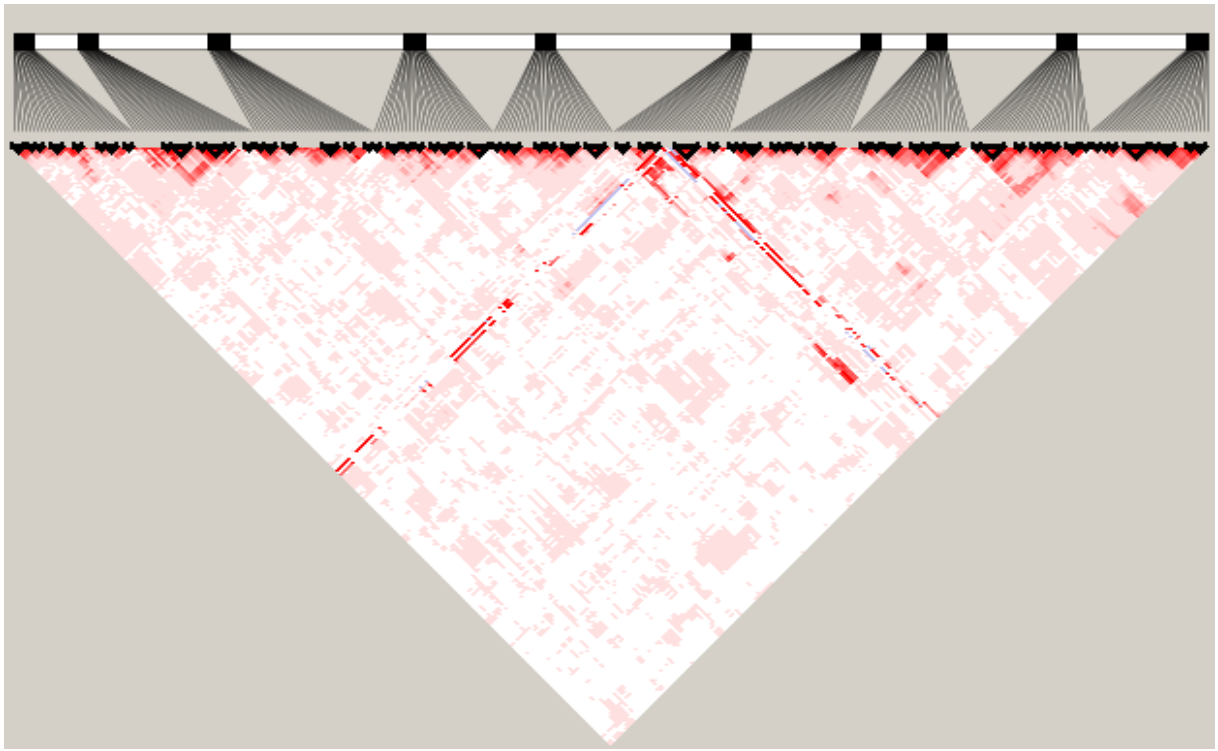


Figura 4.28: Mapa de LD da geração 4 em seleção no cenário 7.

4.8 Cenário 8 - Girolando PS

Os primeiros animais oriundos do cruzamento de gados Gir e Holandês surgiram de forma não planejada no Vale do Paraíba nas decadas de 1940 e 1950. As características exibidas por esses animais se mostraram atraentes para os produtores de leite brasileiro, levando a um difusão da prática desse cruzamento para outras regiões leiteiras do Brasil. O cruzamento padronizado foi elaborado em 1989 visado aumentar a probabilidade de acerto.

A raça é fundamentalmente produto do cruzamento do Holandês com o Gir, passando por variados graus de sangue. O direcionamento dos acasalamentos busca a fixação do padrão racial, no grau de 5/8 Holandês + 3/8 Gir, objetivando um gado produtivo e padronizado, de forma a consolidar a raça Puro Sintético do Girolando (PS). Quaisquer combinações entre a raça Holandesa e a raça Gir, e seus mestiços, podem ser utilizados para obtenção do PS.

O cenário 8 visa simular o cruzamento do Girolando opção B, conforme Figura 4.29. O código de geração do cenário 8 pode ser visto no Apêndice H.

Figura 4.29: Estratégia de cruzamento do Girolando - opção B.

A variância da herdabilidade das raças envolvidas no cruzamento do Girolando PS pode ser vista na Tabela 4.1, com os valores obtidos na população recente. Os valores de média e mediana do Girolando são maiores do que a do Holandês devido aos valores de 0,9993 e 0,9996 das duas primeiras gerações de cruzamento, onde a variação genética é muito alta devido ao cruzamento entre raças distintas.

Tabela 4.1: Variância da Herdabilidade

x	**Mínimo**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
Gir	0,4035	0,4415	0,4510	0,4549	0,4658	0,5210
Holandês	0,6544	0,7001	0,7115	0,7132	0,7267	0,7690
Girolando	0,6154	0,6574	0,8353	0,8214	0,9993	0,9996

As Tabelas 4.2 e 4.3 exibem as variações do GEBV das duas gerações utilizadas na obtenção do Girolando PS, sendo possível notar um valor maior para o Holandês se comparado com o Gir, assim como ocorre na natureza.

Tabela 4.2: Variação do GEBV das duas gerações de Gir utilizadas no cruzamento do Girolando

x	**Mínimo**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
Parental	1202	1209	1210	1210	1212	1219
F1	1203	1209	1210	1210	1212	1217

Tabela 4.3: Variação do GEBV das duas gerações do Holandês utilizadas no cruzamento do Girolando

x	**Mínimo**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
Parental	1507	1516	1518	1518	1520	1527
F1	1508	1516	1518	1518	1520	1527

A variação do GEBV do Girolando durante o processo de cruzamento pode ser vista na Tabela 4.4, onde é possível observar que nas primeiras duas gerações os valores oscilam entre o mínimo do Gir e o máximo do Holandês. A distância entre os valores de mínimo e máximo é menor nas duas últimas gerações, porém a média é maior.

Tabela 4.4: Variação do GEBV nas 4 gerações de cruzamento do Girolando P.S

x	**Mínimo**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
1ª G. - 1/2	1202	1210	1217	1363	1518	1527
2ª G. - 3/4	1202	1210	1217	1363	1518	1527
3ª G. - 5/8	1394	1401	1403	1402	1404	1411
4ª G. - PS	1393	1401	1403	1403	1405	1411

O objetivo do melhoramento genético é aumentar o valor médio da próxima geração e, como visto, foi alcançado com o Girolando PS.

Os mapas de LD dos animais Holandês e Gir utilizados na primeira geração podem ser vistas nas Figuras 4.30 e 4.31. Os mapas são similares entre si, bem como com dos outros cenários.

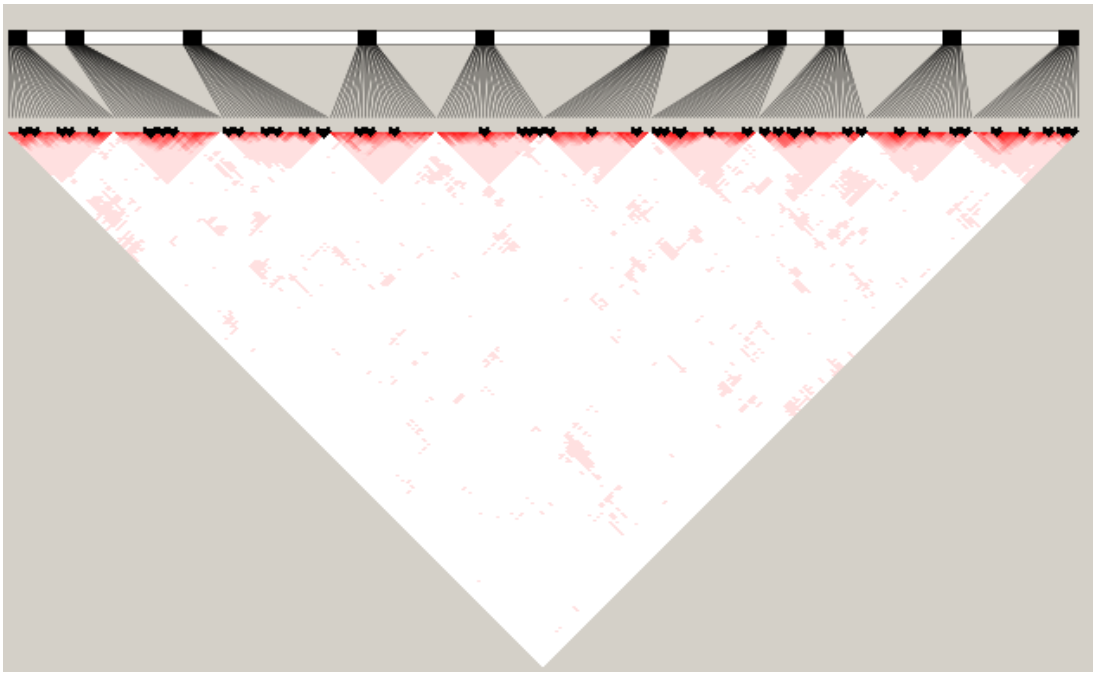


Figura 4.30: Mapa de LD da geração parental do HOLANDÊS.

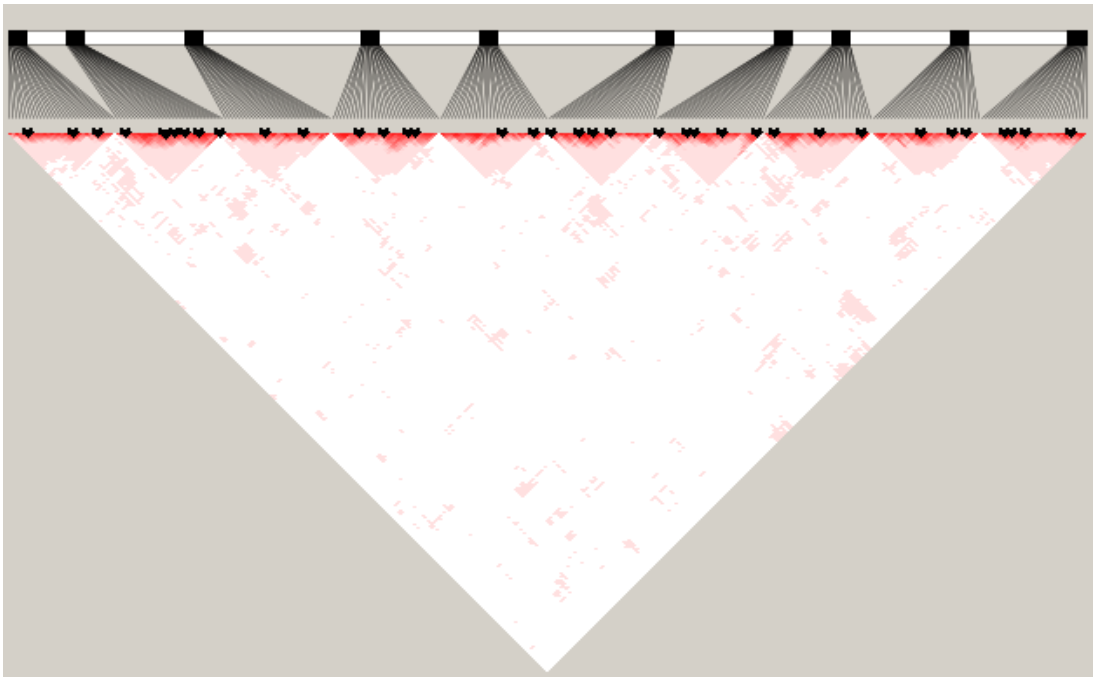


Figura 4.31: Mapa de LD da geração parental do Gir.

O Gir e o Holandêss são utilizados também na segunda geração do Girolando, e os mapas são exibidos nas Figuras 4.32 e 4.33. O salto de somente uma geração não foi suficiente para gerar nenhuma mudança significativa nos mapas de LD. É possível observar o surgimento de alguns blocos haplótipos e a manutenção dos existentes.

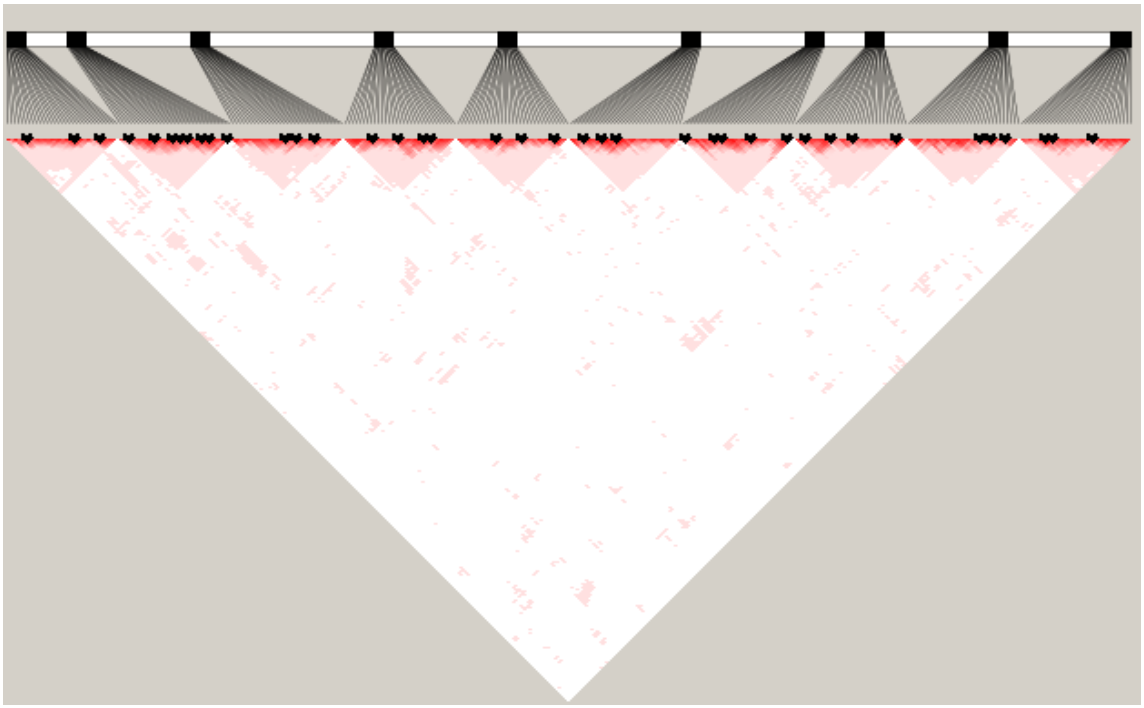


Figura 4.32: Mapa de LD da geração 1 do Gir.

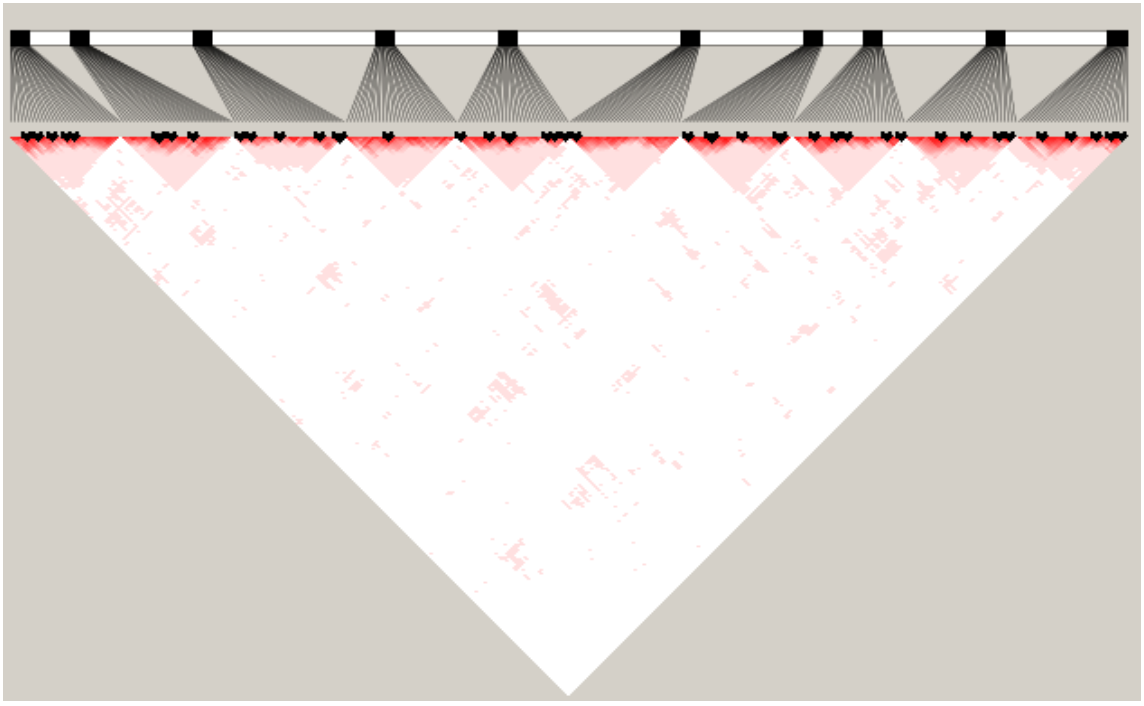


Figura 4.33: Mapa de LD da geração 1 do HOLANDÊS.

Os mapas de LD das gerações 1, 2, 3 e 4 da formação do Girolando PS são mostrados nas Figuras 4.34, 4.35, 4.36 e 4.37. É possível observar blocos oriundos do Gir ou do Holandês no Girolando, porém o mapa é mais comportado que o observado nos cenários de 1 a 7, onde existe o processo de melhoramento animal por meio do cruzamento seletivo.

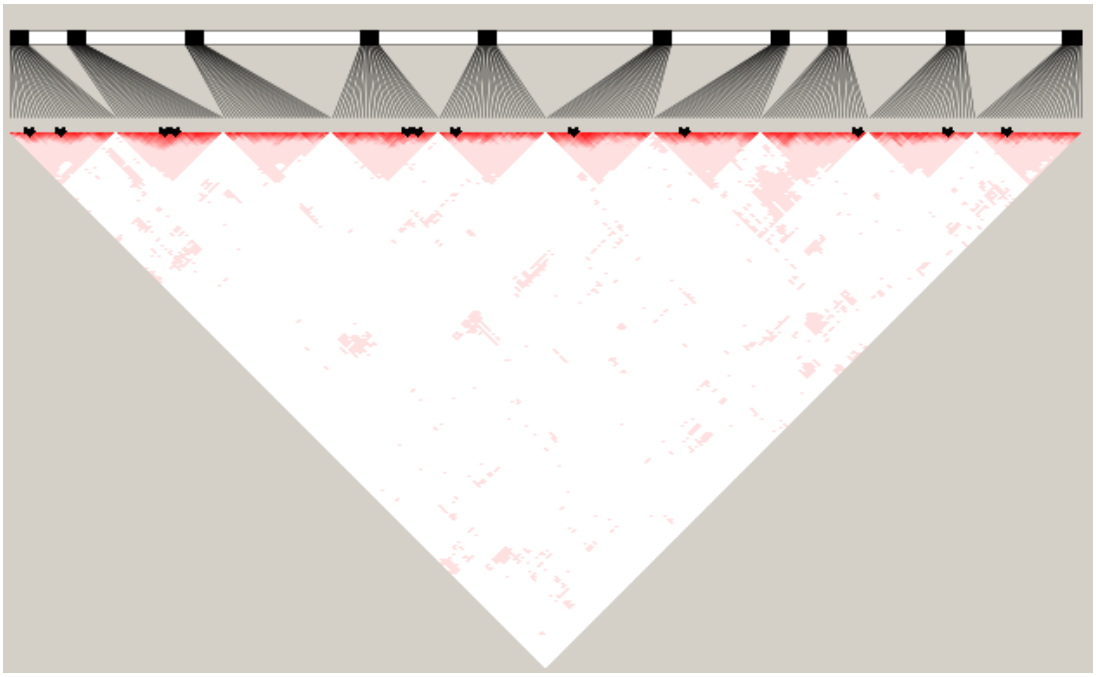


Figura 4.34: Mapa de LD da geração 1 do GIROLANDO.

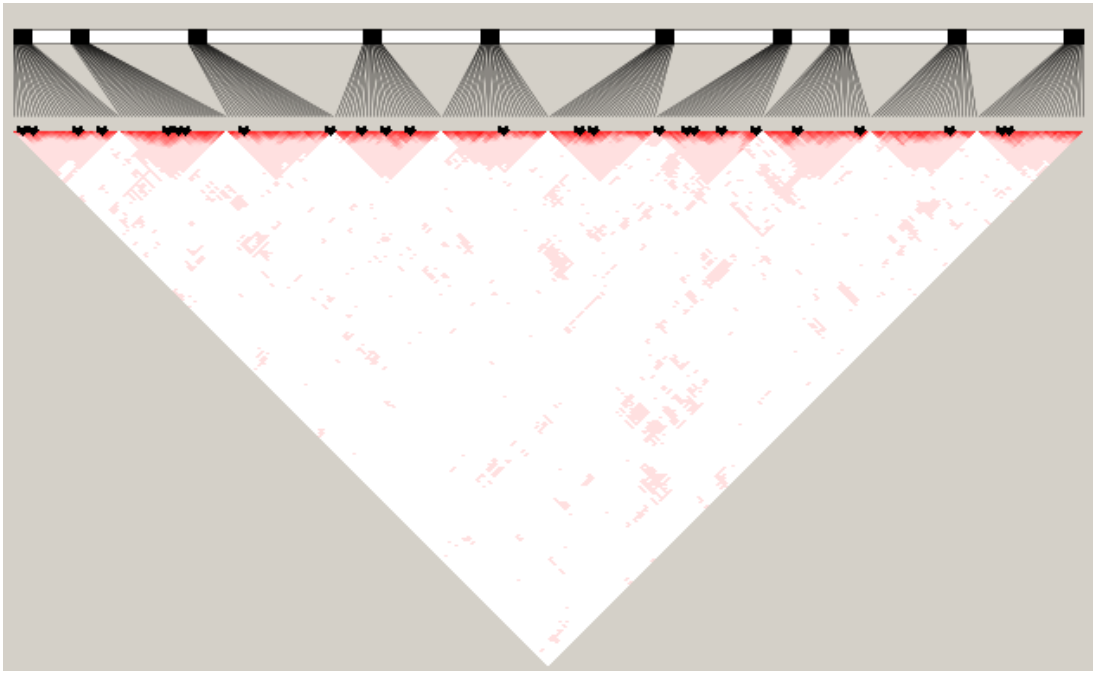


Figura 4.35: Mapa de LD da geração 2 do GIROLANDO.

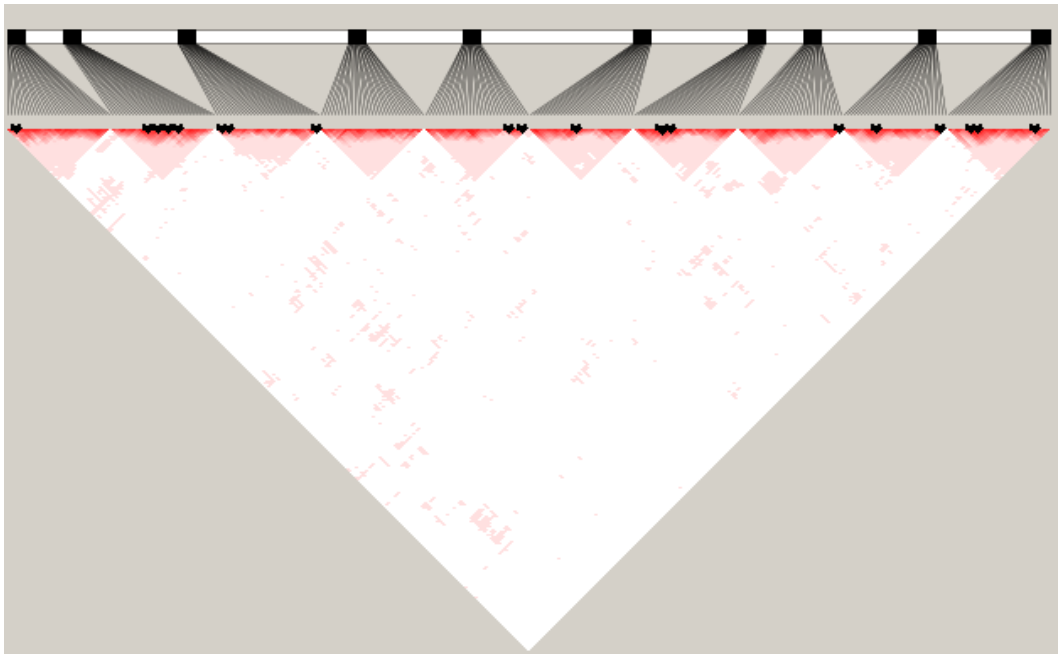


Figura 4.36: Mapa de LD da geração 3 do GIROLANDO.

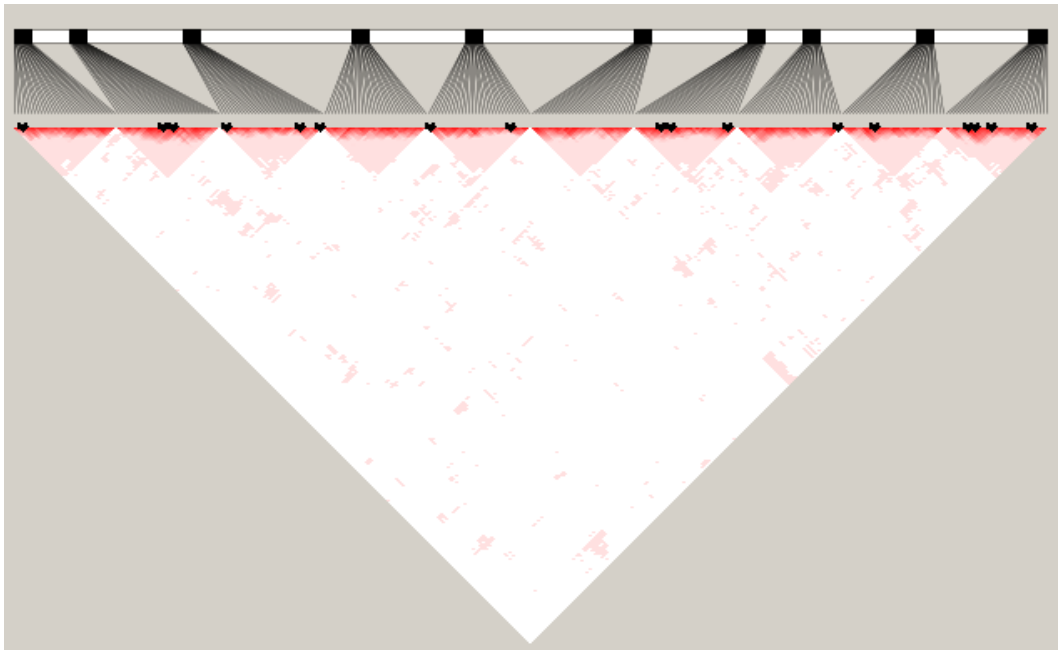


Figura 4.37: Mapa de LD do GIROLANDO PS.

4.9 Considerações

O uso do S4GS permitiu simular cenários com características variadas, possibilitando múltiplas análises fundamentais para seleção genômica. A expectativa é que nos experimentos computacionais, essas distinções produzam conclusões diferentes sobre o impacto da herdabilidade, tamanho da população do estudo, ações gênica e, em especial da epistasia, no aumento da acurácia por meio da seleção de atributos.

5 Modelo Proposto

A seleção de atributos é um técnica que busca encontrar, entre todas as variáveis de um determinado problema, as que sejam mais representativas e explicativas do mesmo, proporcionando, assim, uma redução em sua dimensionalidade. O seu uso em problemas de bioinformática é recorrente, seja buscando diminuir a dimensionalidade do problema ou encontrar os atributos mais relevantes. Muitos são os usos para técnicas de seleção de atributos em bioinformática, em especial na predição de proteínas e/ou suas sequências (SAEYS; INZA; LARRAÑAGA, 2007). Os chips atuais possuem milhares de marcadores de SNPs, em contrapartida a um número não tão alto de indivíduos, assim o uso de técnicas de seleção de atributos se mostra uma alternativa para o aumento da acurácia em seleção genômica.

O uso da seleção de atributos utilizando dados genômicos já foi feito por outros autores, porém com limitações nos algoritmos de avaliação que, em geral, são lineares (LONG et al., 2007; VERBYLA et al., 2009; LONG et al., 2011). Contudo, a seleção de atributos é alvo de críticas, pois para alguns autores ela é ineficaz e não gera aumento significativo na acurácia (MOSER et al., 2009; SOLBERG et al., 2009; HAWS et al., 2015).

As características aditivas ou lineares são uma importante fonte de informação para os programas de seleção genômica, contudo vale ressaltar a crescente pesquisa por interações entre genes ou epistasia. Esse tipo de interação produz uma relação não-linear entre o genótipo e o fenótipo. Pérez-Rodríguez et al. (2012), em seu artigo, demonstra a importância da pesquisa de métodos capazes de tratar esse tipo de interação. O autor utilizou diferentes conjuntos de dados sujeitos a vários métodos distintos, lineares e não-lineares, entre eles uma técnica de redução de dimensionalidade via BLASSO para seleção das entradas para uma rede neural. Os métodos não-lineares se mostraram mais acurados nos conjuntos de dados apresentados, com destaque para a rede neural com entrada penalizada via BLASSO. A seleção de atributos em cenários não-lineares se mostrou promissora, porém ainda aberta para novas pesquisas.

A interação entre marcadores é um problema complexo de ser tratado e recebe destaque neste trabalho. Os cenários simulados buscaram mimetizar essa interação, trazendo assim um desafio maior para o modelo apresentado.

O alto número de marcadores em contraste a quantidade de indivíduos fenotipados torna os problemas com chips de alta densidade um campo relevante para a aplicação da seleção de atributos. Wimmer et al. (2013), em seu artigo, analisaram duas técnicas de seleção de variáveis por meio do anulamento do efeito das que não são de interesse, para isso os autores utilizaram a técnica de BLASSO e BayesB comparando com o RR-BLUP. Os autores demonstram o potencial da seleção de atributos concluindo que o processo de seleção é dependente do número de indivíduos fenotipados ser maior que a quantidade de mutações causais que contribuem para a característica de interesse. O tamanho da amostra também é inversamente proporcional a herdabilidade e a extensão do desequilíbrio de ligação. Os autores destacam a importância da pesquisa de comportamento em cenários não-lineares mediante o processo de seleção de atributos, que é o foco dessa tese.

O trabalho de He, Wang e Parada (2015) apresenta um método para detecção de epistasia utilizando o RR-BLUP como técnica de referência. O algoritmo consiste em criar modelos com interações entre 2 marcadores. O método visa aumentar a acurácia buscando um modelo ótimo ao extrapolar os modelos aditivos do RR-BLUP.

O LD é uma importante característica no estudo de seleção genômica. As bases de dados de animais ou planta possuem, em geral, um alto LD entre os marcadores. Atualmente tem-se o indicativo de que a intensidade do LD pode influenciar na acurácia dos modelos utilizados em seleção genômica (LIU et al., 2015).

A seleção de atributos vem sendo utilizada como uma das alternativas para o aumento da acurácia em dados genômicos. O BLASSO e o RR-BLUP podem ser considerados técnicas de seleção de atributos embutidos, uma vez que o impacto de alguns marcadores é anulado durante o processo de construção dos modelos. As técnicas apresentadas, em geral, são lineares, privilegiando interações aditivas. De uma forma distinta, o método consiste em separar a etapa de seleção dos marcadores mais relevantes da etapa de avaliação, permitindo a aplicação de métodos de seleção de atributos mais robustos e independentes, em uma etapa anterior a construção do modelo de seleção.

A separação do processo em duas etapas permite também identificar os marcadores mais relevantes para a característica de interesse. Essa identificação permite estender o estudo buscando um entendimento do comportamento biológico da característica. As técnicas de seleção de atributos utilizadas entregam explicitamente um conjunto menor de dados, enfatizando os marcadores mais relevantes. Enquanto os métodos clássicos buscam

reduzir ou anular o efeito dos marcadores e para identificá-los é necessário uma avaliação dos efeitos calculados, o que em alguns cenários é custoso e por vezes ineficiente.

A seguir, serão descritas as técnicas utilizadas nas etapas de seleção de atributos bem como na regressão. Essas técnicas compõem o modelo proposto.

5.1 *Forward Features Selection* - FFS

O modelo *Forward Features Selection* é considerado um método clássico para a seleção de atributos, selecionando as variáveis através de adições sucessivas ao modelo, que é mantido caso a adição seja benéfica (DRAPER; SMITH; POWNELL, 1966; HOCKING, 1976). O primeiro modelo de FFS testando o ganho de informação por meio do incremento do *Mean Square Error* (MSE) ao se adicionar variáveis foi desenvolvido por Draper, Smith e Pownell (1966). A escolha do melhor modelo pode ser feita por meio de medidas estatísticas conhecidas, como teste-f, teste-t, valor-p, R-quadrado ajustado, critério de informação de Akaike, Critério de Informação Bayesiano, MSE, entre outras.

O processo de FFS consiste em adicionar variáveis a cada passo do método, que se inicia avaliando cada uma separadamente, em seguida a mais significativa gera o primeiro modelo, que é o modelo referência M_f. Então ela é combinada com todas as variáveis restantes gerando $n-1$ modelos e se o novo modelo M_n for melhor que o M_f ele sera o novo modelo M_f, caso contrario o algoritmo termina por já ter encontrado o modelo ótimo. Os passos são executados até que o M_f possua a melhor qualificação (DRAPER; SMITH; POWNELL, 1966; GUYON; ELISSEEFF, 2003).

O algoritmo 5 exibe o pseudo-código do FFS implementado nessa tese. Conforme a proposta inicial do FFS o código avalia o impacto individual de cada marcador sobre o fenótipo, contudo nesse trabalho a medida utilizada foi a correlação. Os marcadores são inicialmente ordenados da maior para a menor correlação. O marcador mais influente é utilizado para montar o primeiro modelo. O processo então segue adicionando marcadores enquanto a correlação atual for maior que a anterior.

O uso de FFS combinado com alguma técnica específica para o problema de seleção genômica é de aplicação recorrente, entretanto os métodos mais comuns de serem utilizados são o BLASSO e/ou o RR-BLUP que, no geral, são lineares. Desta forma, procura-se substituir as chamadas medidas estatísticas por uma mais relacionada e associada ao pro-

cesso de seleção genômica. Nesse trabalho o método de FFS foi implementado utilizando o SVR como modelo de regressão e a medida de correlação como decisão de escolha entre os modelos. O SVR permite avaliar cenários lineares e não-lineares, logo a expectativa é que o uso do SVR permita identificar relações não lineares entre os marcadores.

Buscando reduzir o tempo gasto no processo e seleção de atributos, o FFS foi implementado utilizando processamento paralelo. O processamento em paralelo permitiu uma redução de 2,5 a 4 vezes o tempo computacional gasto.

Algoritmo 5: Pseudocódigo do FFS

Entrada: Matriz de dados com os Marcadores do tipo SNP + Fenótipo X número de indivíduos.
Resultado: Conjunto reduzido de dados contendo somente os SNPs relevantes + Fenótipo e a totalidade dos indivíduos.

```
inicialização;
// Avaliação de cada variável
sizeDados = tamanho da matriz de dados;
para (i=1; i< sizeDados; i++) faça
    matrizUnitaria[i] <- SVM(dados[, i] ∼ ·fenotipo);
fim para
matrizUnitaria é adicionada a matrizDados;
matrDados é ordenada;
ModeloFoiAtualizado = TRUE;
o vetor de correlações é atualizado utilizando o SVR;
enquanto ModeloFoiAtualizado faça
    ModeloFoiAtualizado = FALSE;
    A variável modelos é reiniciada;
    matriz modeloDados declarada;
    para (i=2; i< sizeDados; i++) faça
        se (o marcador já está presente no modelo) então
            modelos[i] = 0;
        fim se
        else modelos[i] = Avaliação via SVR
    fim para
    modeloDados é ordenada;
    se (se a correlação obtida for maior que a anterior) então
        correlação e atualizada;
        vetor indices é atualizado;
        ModeloFoiAtualizado = TRUE;
    fim se
fim enquanto
SAIDA = indices;
```

A correlação entre os marcadores e o fenótipo foi escolhida como critério de avaliação dos modelos, pois permite selecionar os que mantêm a ordenação mais próxima entre os

indivíduos, e não somente o quão próximo os valores preditos estão dos reais. O uso do MSE como medida de qualidade foi avaliado, porém o mesmo selecionou os modelos onde os valores preditos possuía o menor erro em detrimento a ordem entre os indivíduos. Quando os valores de herdabilidade eram baixos, ou o efeito dos marcadores eram pequenos, a seleção utilizando a correlação se mostrou mais eficiente que com o MSE.

5.2 Método Para Aumento de Acurácia em Seleção Genômica

O processo de seleção genômica já pressupõe a linearidade ao utilizar como referência para o calculo do GEBV o somatório de cada marcador multiplicado pelo seu efeito já conhecido *a priori*. Nessa tese propõe-se a separação do cálculo do GEBV em duas etapas, sendo elas: **Seleção** e **Avaliação**. A etapa de seleção consiste na aplicação de uma técnica de redução de dimensionalidade, com um maior nível de conservadorismo, ou seja, aceitando a presença de falsos positivos, desde que os marcadores causais continuem presentes. A seguir, em uma etapa de avaliação, aplica-se um modelo de regressão capaz de trabalhar com um conjunto de dados com ações gênicas que expressarão mapeamentos lineares e não lineares.

Desta forma, o método proposto consiste em uma combinação eficiente entre um processo de seleção de atributos com um maior nível de falsos positivos, e procedimentos para a determinação do GEBV com o máximo de acurácia. Como visto no materiais e métodos o cálculo do GEBV consiste no somatório dos marcadores e seus pesos ($GEBV = \sum_{j=1}^{n} w_j x_j$). O uso de outros método no cálculo do GEBV exige uma alteração nessa formulação para a $GEBV = f(x)$.

A codominância é a única ação gênica conhecidamente linear, as outras expressão mapeamento não linear. Logo é de interesse que as técnicas utilizadas na etapa de seleção e avaliação sejam capazes de capturar tanto mapeamentos lineares quanto não lineares, incluindo, inclusive combinações mistas. A expectativa com um método que consiga detectar tais níveis de interação consiste em poder avaliar se esses atributos, ou suas associações, podem impactar de forma positiva ou negativa no ganho de acurácia.

O método é dividido em duas etapas: seleção e a avaliação. A primeira fase consiste em selecionar as variáveis mais relevantes por meio da seleção de atributos, e avaliar se

ela será capaz de melhorar o modelo, ou $f(x)$, obtido. A segunda etapa é a avaliação de cada subconjunto, comparando-os com o grupo completo, o uso de um conjunto de dados simulados permite conhecer os marcadores causais, dessa forma o subconjunto escolhido foi comparado também com um contendo somente os SNPs causais. A seleção genômica busca o modelo com maior correlação ou acurácia, contudo a dimensionalidade dos dados pode impactar diretamente na capacidade de obtenção de um modelo eficiente. Os chips atuais possuem um número elevado de marcadores, muitas vezes maior que a quantidade de amostras. A seleção dos atributos mais relevantes, mesmo com a presença de falsos positivos, contribui diretamente na redução da dimensionalidade, mitigando, assim, a questão associada a imprecisão na correlação ou acurácia.

A divisão em duas etapas visa obter um procedimento que seja complementar em relação às técnicas que as compõem, gerando uma combinação ótima para o aumento de acurácia, sendo que a seleção dos marcadores mais relevantes traz o benefício do destaque das possíveis regiões genômicas responsáveis pela característica em estudo.

Os algoritmos escolhidos para a etapa de seleção de atributos utilizam aprendizado supervisionado como base para a construção dos modelos de seleção de atributos. As técnicas escolhidas para a etapa de seleção foram: o FFS-SVR desenvolvido nessa tese, apresentado na seção 5.1; o SMS desenvolvido por Oliveira (2015); e a CART como uma alternativa rápida, porém menos robusta que as outras duas (BREIMAN et al., 1984).

A fase de avaliação utilizou duas técnicas clássicas o RR-BLUP e o BLASSO como referência, e, de acordo com as expectativas descritas, adotou-se para esta etapa o SVR, devido a sua robustez e capacidade de trabalhar com dados lineares e não-lineares.

A associação das técnicas utilizadas na etapa de seleção (**S**) e avaliação (**A**) leva a três modelos: **S**→FFS + **A**→SVR; **S**→SMS + **A**→SVR; e **S**→CART + **A**→SVR. Os três modelos são comparados com os clássicos BLASSO e RR-BLUP e também com o SVR aplicados diretamente em todo o conjunto de dados.

Com a aplicação dos modelos desenvolvidos nos conjuntos de dados simulados no Capítulo 4, objetiva-se responder as seguintes questões:

- É possível aumentar a acurácia em dados genômicos aplicando técnicas de seleção de atributos?

- A qualidade da seleção é proporcional ao aumento da acurácia?

- A seleção é eficiente em qualquer cenário?
- As interações epistáticas geram impacto no modelo proposto?
- A ordem das interações gera algum impacto no processo de obtenção dos modelos?

Em resumo, o objetivo é avaliar se o modelo proposto é capaz de aumentar a acurácia em dados genômicos oriundos de diferentes contextos e interações gênicas como a epistática. Outro ponto de destaque é o impacto do tamanho populacional na capacidade dos modelos, pois em cenários onde a população possui um número reduzido de amostra, as ferramentas clássicas podem não obter valores de acurácia satisfatórios. Nesse contexto, espera-se que o método **Seleção** e **Avaliação**, utilizando técnicas atuais de inteligência computacional, apresentado a seguir, seja capaz de melhorar a acurácia desse conjunto de dados.

5.3 Parâmetros e Configurações

A seguir são explicados os parâmetros e configurações utilizado pelos métodos, bem como as bibliotecas e as referências para as implementações.

5.3.1 FFS

O FFS foi implementado utilizando o software R (R Core Team, 2015) com algoritmo próprio, aplicando a seleção à frente e utilizando o SVR como modelo de avaliação. A variável é mantida no modelo se gerar um aumento superior ao erro quando adicionada. O SVR utilizou kernel radial com validação cruzada de 10-partes. Os parâmetros utilizados foram: erro=0,0001, γ=0,01, C=1 e ϵ=0,1. Bem como os pacotes doParallel e foreach para o processamento paralelo (ANALYTICS; WESTON, 2014a; ANALYTICS; WESTON, 2014b).

5.3.2 SMS

O SMS foi implementado utilizando o software R (R Core Team, 2015) com algoritmo disponibilizado pelo autor. O SMS é um combinação de diferentes técnicas, demandando um maior número de parâmetros, a saber: Para a floresta randômica ($ntree$ = 4000 e $mtry$ = número de atributos); Para o AG (população = 100, e como critério de parada máximo

interações = 10000 e gerações iguais = 30); e o SVR (validação cruzada 10-partes, γ = 0,01, $C = 1$, ϵ = 0,1). Bem como os pacotes doParallel e foreach para o processamento paralelo (ANALYTICS; WESTON, 2014a; ANALYTICS; WESTON, 2014b).

5.3.3 CART

A CART foi implementada utilizando o software R (R Core Team, 2015) e o pacote RPART com os parâmetros padrão do pacote, com destaque para a medida de erro utilizada, o modelo *ANOVA* (THERNEAU; ATKINSON; RIPLEY, 2015).

5.3.4 RR-BLUP

O RR-BLUP foi implementado utilizando o software R (R Core Team, 2015), e os pacotes RR-BLUP de Endelman (2011) para a regressão e o pacote HapEstXXR de Knueppel e Rohde (2015) para os cálculos das medidas genômicas, utilizando os parâmetros padrões dos pacotes.

5.3.5 BLASSO

O BLASSO foi implementado utilizando o software R (R Core Team, 2015), e os pacotes BLR de Campos et al. (2013) para a regressão e o pacote HapEstXXR de Knueppel e Rohde (2015) para os cálculos das medidas genômicas. Os parâmetros utilizados no pacote BLR foram: número de interação =200; $burnIn = 1$. Por sua a *priori* do BLR utiliza os seguintes elementos em sua configuração:$varE = (S$=4,5, df=3), $varBR = (S$=0,s009,df=3) e $lambda$ = (tipo='random',$value$=30,$shape$=0,52,$rate$=2e-5).

5.3.6 SVR

O SVR foi implementado utilizando o software R (R Core Team, 2015), e os pacotes e1071 de Meyer et al. (2014). Bem como os pacotes doParallel e foreach para o processamento paralelo (ANALYTICS; WESTON, 2014a; ANALYTICS; WESTON, 2014b), devido a grande demanda computacional necessária por este modelo.

O SVR pode se adaptar às mais variadas relações entre genótipo/fenótipo desde que se escolha parametrização correta. Dessa forma o *kernel* escolhido foi o radial com as seguintes configurações: validação cruzada 10-partes, $\gamma = 0{,}01$, $C = 1$, $\epsilon = 0{,}1$.

5.4 Considerações

Finalizando, é importante ressaltar que os modelos propostos foram construídos baseando-se em padrões intrínsecos de dados genômicos relacionados a indivíduos fenotipados. Todo o entendimento das características destas bases, bem como das possíveis ações gênicas, direcionaram o desenvolvimento dos métodos complementares nas etapas de seleção e avaliação.

Em uma primeira análise, pode-se ter a impressão que o modelo Seleção-Avaliação proposto seria ajustado para qualquer tipo de base de dados, não necessariamente para dados de seleção gênica. Porém, algumas sutilezas construtivas, a saber: a herdabilidade que distribui o efeito entre ambiente e genético, as ações gênicas que promovem interações complexas entre os marcadores, a proporção entre o número de indivíduos e de marcadores e o LD, tornam, realmente, o modelo bastante específico para a busca de marcadores causais em diversos padrões de interação gênica.

Tem-se a expectativa que os experimentos computacionais identifiquem o modelo mais adequado e robusto para os diversos cenários gerados pelo simulador S4GS descrito no Capítulo 3.

6 Experimentos Computacionais

Esse capítulo exibe os resultados obtidos com a execução dos experimentos computacionais. O cenário 1 foi utilizado como referência para explicar o processo de análise dos resultados obtidos durante os testes. A construção do S4GS permitiu a geração de um amplo conjunto de dados, enriquecendo os resultados e a avaliação do uso de técnicas de seleção de atributos, que utilizem inteligência computacional, aplicadas ao melhoramento genético animal por meio da seleção genômica.

6.1 Cenário 1

Os conjuntos de dados utilizados nos experimentos foram descritos em detalhes no Capítulo 4. Visando facilitar o entendimento são explicados os resultados obtidos com o cenário 1, permitindo dessa forma uma discussão mais aprimorada do processo de seleção de atributos aplicada a dados genômicos.

O primeiro cenário contém todas as ações gênicas permitidas pelo simulador. O processo de simulação dos dados permite controlar a forma como os marcadores interagem e geram o fenótipo de interesse. Nesse ponto é possível avaliar a dificuldade encontrada na construção de modelos para as diferentes configurações genéticas. Assume-se que os dados simulados são simplificações dos cenários reais, dessa forma as análises estão restritas aos dados disponibilizados pela simulação.

O cenário 1 é composto de duas interações epistáticas dominantes entre os pares (100 e 200) e (1700 e 1900), os marcadores 400, 600 e 900 atuando como aditivos e o 1200 e 1400 como dominante e o 1500 como recessivo. O objetivo da etapa de seleção e escolher os SNPs causais ou algum desequilibrado com eles. Ressalta-se que o enfoque do modelo é aumentar a acurácia por meio da seleção de atributos, com a qualidade da seleção sendo também foco de avaliação.

Os métodos RR-BLUP e BLASSO são de uso corrente no processo de melhoramento genético por meio da seleção genômica. Os modelos gerados por essas técnicas são, em geral, lineares, pois consideram somente o efeito isolado de cada marcador, por esse motivo existe a expectativa de que uma ferramenta de regressão mais geral, como o SVR, consiga

obter melhores resultados em cenários não lineares.

6.1.1 1^a *Geração*

O processo de melhoramento genético foi simulado escolhendo os melhores animais durante 4 gerações consecutivas. Por esse motivo a primeira análise foi na 1^a geração, ou a parental.

O resultado apresentado na tabela 6.1 é diferente do esperado. As ferramentas RR-BLUP e o BLASSO obtiveram resultados próximos entre si, porém a correlação encontrada pelo SVR foi muito baixa. Vale destacar que o MSE encontrado é baixo, se comparado às outras duas ferramentas. O baixo MSE é importante pois é utilizado no calculo da precisão da ferramenta.

Tabela 6.1: Resultado da aplicação de cada uma das ferramentas no cenário 1

	MSE	**DP**	**% Var.**	**COR**	**DP**	**% Var**
RR-BLUP	191,15	27,29	14,28	0,50	0,10	19,18
BLASSO	10,67	2,70	25,35	0,49	0,09	18,81
SVR	7,14	1,18	16,47	0,08	0,08	101,30

A seleção de atributos é aplicada visando o aumento da acurácia. A forma mais simples de se fazer a seleção de atributos é o método de *Forward Features Selection - FFS* que consiste em um método que adiciona atributos ao modelo visando atingir um objetivo. O SVR foi utilizado como técnica de avaliação, sendo o objetivo obter a maior correlação no menor subconjunto de marcadores. Além do FFS, foram aplicados as técnicas de SMS e CART.

A redução da dimensionalidade dos dados por meio da seleção de atributos gerou um considerável aumento na correlação, conforme é possível observar na Tabela 6.2. O uso do FFS associado ao SVR aumentou em 8,87 vezes a correlação obtida usando somente o SVR, sendo também 42% maior que a obtida pelo RR-BLUP. A menor correlação obtida com o uso da seleção de atributos foi com o uso da CART, com um valor 22% maior que o RR-BLUP. O MSE é uma importante característica a ser avaliada, sendo menor nas técnicas associadas com a redução da dimensionalidade dos dados. A covariância (CV) entre a correlação e o desvio padrão mede o quanto o último impacta no primeiro. Nesse ponto é possível observar uma proporção menor no uso do SVR associado a qualquer técnica de seleção de atributos.

Tabela 6.2: Resultado da aplicação de cada uma das ferramentas no cenário 1 com o uso da seleção de atributos.

	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
RR-BLUP	191,15	27,29	14,28	0,50	0,10	19,18
BLASSO	10,67	2,70	25,35	0,49	0,09	18,81
SVR	7,14	1,18	16,47	0,08	0,08	101,30
SVR + SMS	3,73	0,46	12,25	0,68	0,07	10,70
SVR + CART	4,42	0,53	11,94	0,61	0,06	10,34
SVR + FFS	3,50	0,58	16,68	0,71	0,05	7,15

O valor-p é uma medida estatística de uso comum quando se deseja diminuir o número de marcadores em um dado conjunto de dados, logo é possível avaliar a qualidade da seleção das ferramentas apresentadas em contraste com a efetuada pelo valor-p e suas variações. Para isso foram utilizados os filtros: estatístico básico com valor-p $< 0,05$ ou 95%; o critério utilizado pelo GWAS *catalog* de valor-p $< 10e^{-8}$ (MACARTHUR et al., 2017); e os critério estatístico, valor-p $< 0,05$, no conjunto corrigido por Bonferoni. Durante os testes foi avaliada a seleção utilizando a variância explicada calculada pelo BLASSO com o corte $> 1\%$. Porém, só foram encontrados marcadores no cenário 8 e na 1^a geração, devido a altíssima herdabilidade resultante da grande variação genética nesse caso. Por esse motivo esse critério de seleção não foi avaliado.

A Tabela 6.3 exibe os marcadores selecionados que estão desequilibrados com os causais, de forma a facilitar o entendimento eles foram separados por marcador. Os marcadores selecionados podem ser considerados verdadeiros positivos. O valor-p utilizando o corte de 95% seleciona uma grande quantidade de marcadores, com o valor-p $< 10e^{-8}$ o número diminui de forma acentuada, mas cobrindo somente 7 dos 10 marcadores de interesse. As técnicas de IC selecionaram 8 dos 10, mas considerando os desequilibrados tem-se todos os de interesse.

Tabela 6.3: Resultados da seleção de atributos distribuídos por marcadores causais

	100	**200**	**400**	**600**	**900**	**1200**	**1400**	**1500**	**1700**	**1900**
VP 0,05	13	18	21	18	16	21	21	20	17	19
VP e-08	0	0	13	6	1	7	9	6	3	5
Vp-BonFerroni	0	3	0	1	1	0	1	0	0	0
SMS	4	5	1	4	1	5	1	4	5	7
CART	7	6	0	7	0	6	6	6	9	8
FFS	3	4	1	3	1	5	1	4	4	6

A Tabela 6.4 permite avaliar a qualidade da seleção de cada ferramenta, para isso é importante avaliar 3 das medidas exibidas, o percentual filtrado (Filtro %), a sensibilidade da seleção e a sensibilidade da seleção levando em conta o LD. A razão entre a quantidade selecionada e o número de marcadores desequilibrados com os causais, e a razão entre quais dos 10 marcadores causais foram selecionados permite avaliar a cobertura de cada seleção e sua eficácia. O percentual filtrado demonstra o quanto a ferramenta foi capaz de reduzir o conjunto de dados, nesse ponto quanto menor o valor melhor. Logo, a melhor ferramenta é aquela com maior sensibilidade(s) e menor filtro. O SMS foi a técnica com a melhor capacidade de filtragem, selecionando somente 2% da população inicial e com a maior sensibilidade média, seguido do valor-p $< 10e^{-8}$, FFS, CART, valor-p<0,05 e por último o valor-p corrigido por Bonferroni.

Tabela 6.4: Análise da seleção de atributos

	SELEÇÃO				**10 SNPS CAUSAIS**		
	Total	**em LD**	**Sens.**	**Filtro %**	**Causais**	**em LD**	**Sens. LD**
VP 0,05	439	184	41,91	21,95	10	10	100
VP e-08	50	50	100,00	2,50	7	8	80
VP-Bonferroni	98	6	6,12	4,90	0	4	40
SMS	40	37	92,50	2,00	8	10	100
CART	73	55	75,34	3,65	8	8	80
FFS	66	32	48,48	3,30	8	10	100

A seleção de atributos melhorou a acurácia em dados de seleção genômica, contudo a qualidade da seleção não foi o principal fator no aumento da acurácia, pois a técnica com melhor qualidade não gerou o resultado com a maior correlação média. Entretanto, ao levar em conta o desvio padrão, as correlações podem ser consideradas equivalentes.

6.1.2 4^a *geração*

O uso de uma ferramenta para simulação dos dados utilizados permite um maior controle sobre os procedimentos. O processo de melhoramento genético animal pode ser feito de muitas formas, por exemplo, Falconer et al. (1960) orientam a selecionar os melhores animais por 4 gerações consecutivas, logo após seguindo com o cruzamento normal. Essa orientação foi aplicada, permitindo, assim, uma análise da aplicação da metodologia na 1^a e 4^a gerações, ou seja, antes e após o processo de seleção genômica animal. A expectativa é a ocorrência de uma menor variabilidade genética, em consequência uma menor

herdabilidade, logo dificultando a seleção eficaz dos marcadores causais e também uma menor correlação.

A Tabela 6.5 exibe os resultados obtidos com a aplicação das técnicas escolhidas no conjunto de dados da 4^a geração, ou seja a última geração após a seleção genômica. Como esperado a correlação média cai onde, nas técnicas que utilizaram o conjunto completo dos dados a redução foi de 30%, contudo o uso da seleção de atributos minimizou esse impacto quase pela metade, em torno de 16%.

Tabela 6.5: Erro MSE e Correlação obtidas utilizando a 4^a geração como referência no treinamento dos modelos.

	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
RR-BLUP	7,33	4,23	57,63	0,35	0,06	16,93
BLASSO	8,01	2,62	32,76	0,34	0,06	17,56
SVR	4,81	0,93	19,22	0,04	0,11	280,01
SVR + SMS	3,24	0,48	14,90	0,57	0,06	9,63
SVR + CART	3,81	0,67	17,49	0,47	0,09	18,29
SVR + FFS	3,11	0,53	17,17	0,60	0,05	7,95

A redução da herdabilidade e variância genética impactou também na qualidade da escolha dos marcadores causais, conforme é possível observar na Tabela 6.6. As ferramentas de seleção que utilizam IC selecionaram um número maior de marcadores comparando com a 1^a geração, contudo o valor-p selecionou um número menor. Porém, independente da quantidade de marcadores selecionados a sensibilidade, ou seja, a razão entre o total selecionado e o total desequilibrados com os causais, diminuiu consideravelmente. Somente no subconjunto selecionado pelo valor-p $< 10e^{-8}$ que se manteve igual, entretanto quando se avalia exclusivamente os causais encontrados o percentual diminui de 80% para 20%, ou seja a qualidade do subconjunto é menor, conforme mostra a Tabela 6.7.

Tabela 6.6: Seleção por marcadores na geração 4

	100	**200**	**400**	**600**	**900**	**1200**	**1400**	**1500**	**1700**	**1900**
SMS	4	4	1	3	1	4	1	4	3	1
CART	1	6	0	6	6	6	6	5	6	2
FFS	3	3	1	3	1	1	1	5	2	2
VP 0,05	7	16	19	18	7	13	18	11	10	13
VP e-08	0	0	0	2	0	8	4	0	0	0
VP Bonf	1	0	0	0	1	0	0	0	0	0

Tabela 6.7: Análise da seleção de atributos - geração 4

	Seleção				**Causais (10 Marcadores)**		
	Total	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
VP 0,05	443	132	29,80	22,15	10	10	100
VP e-08	14	14	100,00	0,70	2	3	30
VP Bonferroni	26	2	7,69	1,30	0	2	20
SMS	52	26	50,00	2,60	8	10	100
CART	79	44	55,70	3,95	8	9	90
STEPWISE	38	22	57,89	1,90	6	10	100

6.1.3 Considerações

A seleção de marcadores se mostrou relevante na melhoria da acurácia, proporcionando, inclusive, sua manutenção nos diferentes conjuntos de dados. Enquanto a redução da acurácia foi de aproximadamente 30% nas ferramentas clássicas após o processo de melhoramento, ou seja o teste na 4^a geração, ela foi menor no SVR associado ao SMS e ao FFS.

A Tabela 6.8a exibe um resultado comparativo amplo, mostrando as três técnicas utilizando os dados selecionados de 6 formas diferentes. Como é possível observar a seleção dos marcadores melhorou o valor médio da correlação, contudo os valores são similares entre si devido ao desvio padrão. As técnicas de SMS e FFS aumentaram de forma significativa a acurácia, obtendo no mínimo 20% de aumento. O SVR foi certamente a ferramenta mais beneficiada com a redução na dimensionalidade dos dados, chegando a quase 900% da acurácia inicial.

O comparativo entre as técnicas de seleção e regressão pode ser visto na Tabela 6.8b. O aumento na acurácia proveniente da seleção de atributos é mais nítido na 4^a geração, sendo em muitos casos superior a 40% nas ferramentas clássicas. O aumento da acurácia também é maior que o desvio padrão superior, mostrando que o valor é realmente maior.

A seleção de atributos em muitos cenários pode não ser uma alternativa funcional para o aumento da acurácia. Observa-se que na 1^a geração o aumento percebido é pequeno e muitas vezes dentro do desvio padrão do valor total contudo, na 4^a geração, a seleção de atributos se mostrou vantajosa, aumentando a acurácia de forma significativa. A seleção utilizando valor-p corrigido por Bonferroni gerou uma correlação menor em todos os dados, ou seja a redução dos dados se não aplicada corretamente pode diminuir a acurácia dos dados.

Tabela 6.8: Comparativo entre as técnicas de seleção no cenário 1

(a) 1^a **GERAÇÃO**

RR-BLUP						
	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
Total	191,15	27,29	14,28	0,50	0,10	19,18
VP 0,05	285,45	25,71	9,01	0,57	0,08	13,96
VP e-8	147,15	9,09	6,18	0,57	0,07	12,90
Bonf.	25,52	9,03	35,40	0,26	0,14	51,14
SMS	265,09	11,32	4,27	0,60	0,08	12,60
CART	88,42	7,40	8,37	0,55	0,08	14,59
FFS	241,44	16,93	7,01	0,62	0,06	10,06
BLASSO						
	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
Total	10,67	2,70	25,35	0,49	0,09	18,81
VP 0,05	31,73	11,97	37,73	0,55	0,08	14,32
VP e-8	91,64	15,59	17,02	0,57	0,07	11,97
Bonf.	8,94	3,23	36,18	0,27	0,13	49,82
SMS	98,02	12,04	12,28	0,59	0,07	12,62
CART	35,05	12,58	35,90	0,54	0,07	12,95
FFS	52,89	20,24	38,27	0,60	0,06	10,18
SVR						
	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
Total	7,14	1,18	16,47	0,08	0,08	101,30
VP 0,05	6,85	1,11	16,25	0,49	0,08	16,57
VP e-8	4,38	0,43	9,90	0,62	0,06	9,79
Bonf.	6,85	0,77	11,20	0,23	0,13	55,39
SMS	3,73	0,46	12,25	0,68	0,07	10,70
CART	4,42	0,53	11,94	0,61	0,06	10,34
FFS	3,50	0,58	16,68	0,71	0,05	7,15

(b) 4^a **GERAÇÃO**

RR-BLUP						
	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
TOTAL	7,33	4,23	57,63	0,35	0,06	16,93
VP	4,41	1,09	24,63	0,43	0,06	13,34
VP e-8	364,66	19,28	5,29	0,41	0,07	16,23
Bonf.	7,25	2,14	29,46	0,12	0,10	84,25
SMS	103,67	26,09	25,17	0,49	0,06	11,99
CART	94,97	16,31	17,17	0,45	0,06	14,30
FFS	51,25	13,31	25,96	0,53	0,05	9,80
BLASSO						
	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
TOTAL	8,01	2,62	32,76	0,34	0,06	17,56
VP 0,05	6,41	3,16	49,26	0,40	0,06	15,73
VP e-8	71,35	13,49	18,91	0,37	0,07	20,35
Bonf.	6,31	1,96	30,97	0,11	0,10	93,98
SMS	4,89	1,53	31,28	0,46	0,05	9,89
CART	10,68	3,76	35,17	0,43	0,05	12,20
FFS	5,38	2,16	40,16	0,50	0,06	11,51
SVR						
	MSE	**FP**	**CV**	**COR**	**DP**	**CV**
TOTAL	4,81	0,93	19,22	0,04	0,11	280,01
VP 0,05	4,60	0,89	19,39	0,40	0,05	13,49
VP e-8	4,03	0,67	16,73	0,41	0,06	14,66
Bonf.	4,82	1,04	21,50	0,14	0,10	70,26
SMS	3,24	0,48	14,90	0,57	0,06	9,63
CART	3,81	0,67	17,49	0,47	0,09	18,29
FFS	3,11	0,53	17,17	0,60	0,05	7,95

A associação do SVR com o FFS ou SMS aumentou a acurácia mesmo se comparado com os melhores resultados do BLASSO e do RR-BLUP. O MSE obtido pelo RR-BLUP e o BLASSO foi menor nos dados completos. Após a redução é possível observar o aumento significativo na acurácia, exceto para o SVR onde os valores diminuem.

6.2 Cenários de 2 a 7

Os cenários de 2 a 7 são variações do cenário 1 como: tamanho da população, ações gênicas e o uso, ou não, da inseminação artificial. Essas varições permitem avaliar a qualidade da seleção de atributos em diferentes contextos. A variação dos cenários permite avaliar a eficiência, ou não, da seleção de atributos como estratégia para o aumento da acurácia em seleção genômica. Um dos objetivos de grande interesse do trabalho é avaliar o comportamento da acurácia quando o número de indivíduos genotipados for muito baixo. O valor de 1000 foi utilizado como base para a montagem da maioria dos cenários é considerado baixo (GODDARD; HAYES, 2009).

6.2.1 1^a *Geração*

A Tabela 6.9 exibe os resultados consolidados do comparativo entre diferentes técnicas de seleção na 1^a geração, de cada um dos cenários de 2 a 7.

A seleção utilizando valor-p $< 0,05$ possui a menor capacidade de filtro entre as técnicas comparadas, selecionando em média 19% da base, sendo que as outras ferramentas essa seleção não foi maior que $4,95\%$, com vários casos próximos de 2%. Apesar da grande quantidade selecionada as sensibilidades obtidas foram próximas a da ferramenta SMS. O corte no valor-p utilizando a regra do GWAS diminui a quantidade selecionada, obtendo um conjunto de marcadores onde praticamente a totalidade estava desequilibrada com os causais, contudo não cobrindo todos os 10 marcadores. A correção de Bonferroni se mostrou muito restritiva e com baixa cobertura dos marcadores causais em todos os cenários.

As técnicas que utilizam IC tiveram um comportamento mais estável se comparado com o valor-p. Os cenários 5, 6 e 7 também se mostraram mais complexos, pois é possível observar uma sensibilidade menor em relação aos outros porém, os valores encontrados ainda são superiores aos obtidos pelo valor-p. A CART é uma técnica rápida, contudo se

comparada com as outras duas é menos eficiente, mas obteve melhores resultados que o valor-p $< 5x10^{-8}$ e o corrigido por Bonferroni.

Tabela 6.9: Resultados consolidados da etapa de seleção nos cenários de 2 a 7 no conjunto de dados da 1^a geração.

Valor-p <0,05							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	467	158	33,83	23,35	9	9	90
3	480	199	41,46	24,00	10	10	100
4	363	137	37,74	18,15	10	10	100
5	404	128	31,68	20,20	9	10	100
6	214	85	39,72	10,70	5	7	70
7	225	62	27,56	11,25	4	7	70
Valor-p <10e-8							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	66	66	100,00	3,30	6	6	60
3	68	66	97,06	3,40	6	10	100
4	31	31	100,00	1,55	3	3	30
5	0	0	0,00	0,00	0	0	0
6	4	4	100,00	0,20	2	2	20
7	7	7	100,00	0,35	1	1	10
Valor -p Corrigido por Bonferroni							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	99	7	7,07	4,95	0	6	60
3	147	18	12,24	7,35	1	9	90
4	55	4	7,27	2,75	0	3	30
5	11	0	0,00	0,55	0	0	0
6	17	5	29,41	0,85	1	5	50
7	15	2	13,33	0,75	0	2	20
SMS							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	56	39	69,64	2,80	6	10	100
3	32	27	84,38	1,60	7	10	100
4	48	39	81,25	2,40	8	10	100
5	53	33	62,26	2,65	8	9	90
6	66	24	36,36	3,30	5	8	80
7	52	23	44,23	2,60	5	9	90
CART							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	68	50	73,53	3,40	6	7	70
3	98	60	61,22	4,90	8	8	80
4	68	33	48,53	3,40	5	6	60
5	112	40	35,71	5,60	4	7	70
6	75	24	32,00	3,75	4	4	40
7	78	14	17,95	3,90	2	3	30
FFS							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	73	32	43,84	3,65	8	10	100
3	65	19	29,23	3,25	9	10	100
4	59	20	33,90	2,95	7	7	70
5	46	13	28,26	2,30	4	8	80
6	45	16	35,56	2,25	4	8	80
7	57	24	42,11	2,85	5	8	80

O uso da inseminação artificial exclusiva (cenário 2), ou mista (cenários 3 e 4) nessa etapa do trabalho não gerou impacto significativo nas análises. O objetivo do uso da inseminação artificial consiste em manter a variabilidade genética dentro do rebanho, bem como adicionar melhores características provenientes de indivíduos avaliados e selecionados.

A Figura 6.1 exibe um comparativo entre os cenários de 1 a 7, distribuídos por ferramenta. A principal diferença entre o cenário 1 e o 2 é o uso da inseminação artificial, sendo possível verificar que as ferramentas conseguiram obter uma maior correlação no cenário 2, o que pode ser justificado pela presença de uma maior variabilidade genética. O uso da seleção de atributos trouxe benefícios nítidos para o aumento da acurácia em dados genômicos.

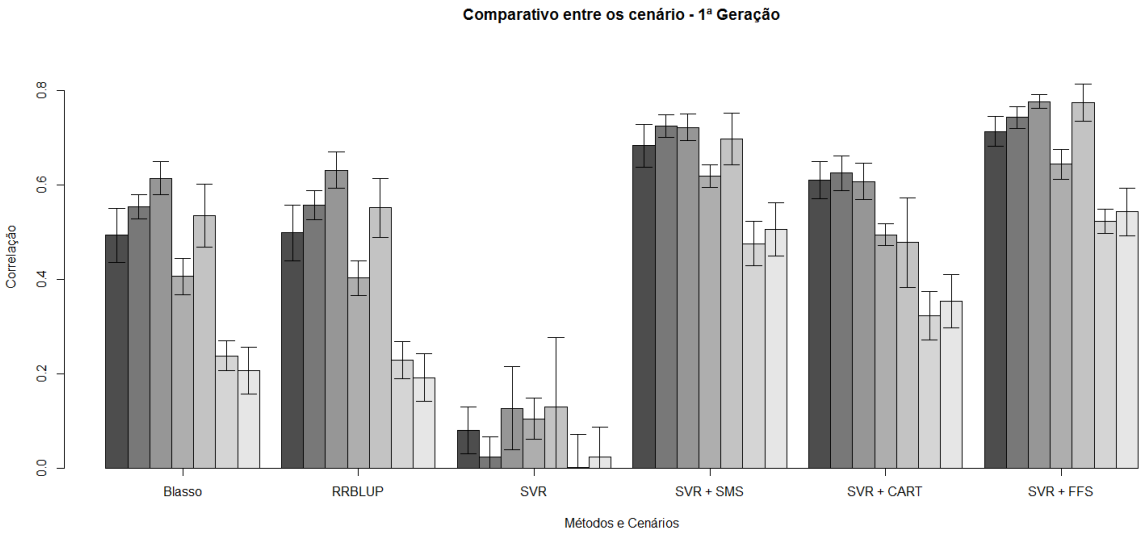


Figura 6.1: Comparativo completo agrupado por ferramentas - cenários de 1 a 7.

A Tabela 6.10 exibe o resultado consolidado de cada ferramenta utilizada na etapa de avaliação, onde observa-se que os cenários mais complexos foram os cenários 6 e 7 que possuem interações de ordem 3 e 4 respectivamente.

O cenário 3 por ser totalmente aditivo favoreceu as ferramentas lineares BLASSO e RR-BLUP, que obtiveram correlações maiores. O uso da seleção de atributos gerou um significativo aumento na correlação, em muitos casos chegando a dobrar a correlação encontrada pelo BLASSO e o RR-BLUP, principalmente nos cenários 6 e 7. O uso de dados após a seleção de atributos permite a obtenção de um MSE menor.

Tabela 6.10: Resultados consolidados da etapa de avaliação utilizando os dados da 1^a geração de cada cenário

RR-BLUP						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	253,13	31,23	12,34	0,56	0,05	8,82
3	989,76	91,34	9,23	0,63	0,06	9,84
4	40,58	9,61	23,69	0,4	0,06	14,71
5	206,4	49,39	23,93	0,55	0,1	18,32
6	26,86	8,37	31,18	0,23	0,06	27,35
7	5,81	0,64	10,98	0,19	0,08	41,84
BLASSO						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	28,58	6,16	21,54	0,55	0,04	7,53
3	28,63	8,6	30,03	0,61	0,06	9,27
4	13,93	4,72	33,85	0,41	0,06	15,15
5	20,27	12,66	62,46	0,54	0,11	20,07
6	4,87	1,09	22,29	0,24	0,05	21,43
7	5,48	1,29	23,63	0,21	0,08	39,03
SVR						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	7,5	0,86	11,4	0,02	0,07	301,13
3	6,4	0,44	6,94	0,13	0,14	111,83
4	5,76	0,77	13,36	0,1	0,07	66,68
5	6,13	1,8	29,44	0,13	0,24	182,26
6	4,15	0,48	11,6	0	0,11	7455,14
7	4,15	0,53	12,74	0,02	0,1	438,11
SVR + SMS						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	3,54	0,51	14,3	0,73	0,04	5,35
3	3,05	0,4	13,25	0,72	0,05	6,31
4	3,6	0,54	14,86	0,62	0,04	6,33
5	3,23	0,76	23,54	0,7	0,09	12,64
6	3,23	0,55	16,97	0,48	0,08	15,97
7	3,12	0,42	13,62	0,51	0,09	18,08
SVR + CART						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	4,54	0,59	13,09	0,63	0,06	9,39
3	4,03	0,5	12,4	0,61	0,06	10,29
4	4,41	0,64	14,43	0,5	0,04	7,54
5	4,72	1,08	22,86	0,48	0,15	31,99
6	3,78	0,46	12,23	0,32	0,08	25,85
7	3,67	0,43	11,65	0,35	0,09	25,81
SVR + FFS						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	3,38	0,61	17,95	0,74	0,04	5,01
3	2,56	0,27	10,4	0,78	0,02	3,12
4	3,46	0,58	16,85	0,64	0,05	7,99
5	3,03	0,85	28,04	0,77	0,06	8,32
6	3,03	0,36	11,75	0,52	0,04	7,8
7	2,97	0,45	15,12	0,54	0,08	14,96

6.2.2 4^a Geração

A análise dos dados dos indivíduos da 4^a geração, ou seja, após a etapa de seleção dos mais aptos permite avaliar o impacto que o processo tem em relação ao modelo. A seleção dos melhores animais aumenta o valor médio do fenótipo, contudo ocorre uma redução na variabilidade genética. A seleção dos melhores animais pode reduzir a MAF dos marcadores causais, dificultando o processo de seleção de atributos.

O comparativo entre as técnicas ao longo dos cenários de 1 a 7 pode ser visto na Figura 6.2. O uso da inseminação artificial gerou uma queda na acurácia, pois o número de machos utilizados é menor, logo a aplicação contínua desse método pode diminuir a variabilidade genética. O cenário 2 faz uso exclusivo de inseminação artificial, obtendo no modelo treinado na 1^a geração acurácia maior que do cenário 1, que não utiliza a inseminação, porém utilizando a 4^a geração ocorre uma queda na acurácia em todas as ferramentas. Os modelos treinados nessa etapa obtiveram acurácias menores que na 1^a geração, com uma maior influência da seleção de atributos, pois a associação do SVR com uma técnica de redução de dimensionalidade manteve a acurácia similar nos dois conjuntos de dados.

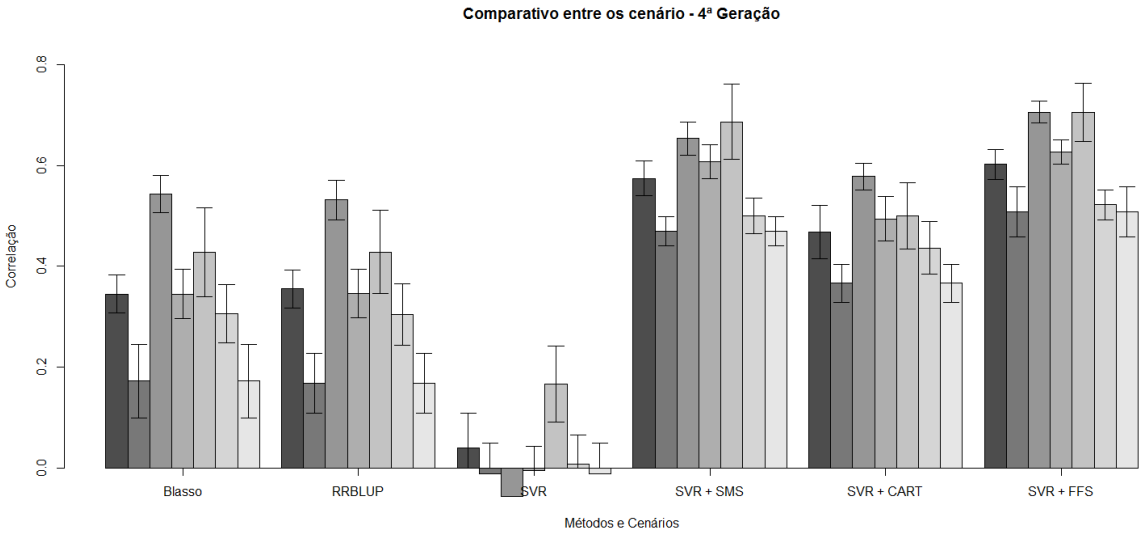


Figura 6.2: Comparativo completo agrupado por ferramentas na geração 4 - cenários de 1 a 7.

A Tabela 6.11 exibe o resultado da etapa de seleção nos cenários de 2 a 7 na 4^a geração. A sensibilidade em LD do valor-p $< 0,05$ foi maior que nos dados da 1^a geração, entretanto a sensibilidade geral reduziu de 35% para 24%, e o percentual filtrado aumentou de 19

para 24. Dessa forma, pode-se observar uma maior cobertura em relação aos causais, porém o valor-p se mostrou menos restritivo. As variações do valor-p não demonstraram diferenças significativas em relação a restrição e cobertura na seleção dos marcadores.

Tabela 6.11: Resultados consolidados da etapa de seleção nos cenários de 2 a 7 no conjunto de dados da 4^a geração.

Valor-p <0.05							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	598	140	23,41	29,90	10	10	100
3	687	173	25,18	34,35	10	10	100
4	696	125	17,96	34,80	10	10	100
5	325	72	22,15	16,25	7	9	90
6	378	100	26,46	18,90	9	10	100
7	314	81	25,80	15,70	7	8	80
Valor-p<10e-8							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	12	12	100,00	0,60	3	3	30
3	61	59	96,72	3,05	10	10	100
4	21	18	85,71	1,05	5	5	50
5	3	0	0,00	0,15	1	1	10
6	7	7	100,00	0,35	2	2	20
7	0	0	0,00	0,00	0	0	0
Valor-p Corrigido por Bonferroni							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	98	2	2,04	4,90	0	2	20
3	117	9	7,69	5,85	1	6	60
4	84	9	10,71	4,20	6	6	60
5	13	2	15,38	0,65	0	1	10
6	15	1	6,67	0,75	0	1	10
7	7	0	0,00	0,35	0	0	0
SMS							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	36	28	77,78	1,80	8	10	100
3	37	22	59,46	1,85	7	9	90
4	43	26	60,47	2,15	10	10	100
5	49	19	38,78	2,45	5	7	70
6	35	22	62,86	1,75	6	8	80
7	34	20	58,82	1,70	5	7	70
CART							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	73	55	75,34	3,65	7	8	80
3	60	48	80,00	3,00	6	9	90
4	82	49	59,76	4,10	10	10	100
5	114	30	26,32	5,70	5	6	60
6	88	53	60,23	4,40	9	10	100
7	51	28	54,90	2,55	6	6	60
FFS							
Cenário	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
2	46	28	60,87	2,30	9	10	100
3	53	18	33,96	2,65	8	10	100
4	47	18	38,30	2,35	8	10	100
5	45	9	20,00	2,25	2	4	40
6	45	16	35,56	2,25	4	6	60
7	38	18	47,37	1,90	4	7	70

Na Tabela 6.12, que exibe os resultados da etapa de avaliação, é possível observar uma redução no MSE de cada ferramenta, bem como na correlação média.

Tabela 6.12: Resultados consolidados da etapa de avaliação utilizando os dados da 4^a geração de cada cenário

RR-BLUP						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	191,15	27,29	14,28	0,50	0,10	19,18
3	283,89	47,74	16,82	0,53	0,06	11,87
4	15,43	5,07	32,86	0,35	0,08	22,46
5	99,15	26,37	26,59	0,43	0,13	31,16
6	4,35	0,75	17,12	0,30	0,10	32,49
7	5,21	1,45	27,78	0,17	0,10	57,54
BLASSO						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	10,67	2,70	25,35	0,49	0,09	18,81
3	4,81	1,16	24,16	0,54	0,06	11,08
4	5,46	1,91	35,01	0,34	0,08	22,79
5	29,60	15,52	52,42	0,43	0,14	33,24
6	4,67	1,14	24,50	0,31	0,09	30,70
7	4,80	1,16	24,19	0,17	0,12	68,14
SVR						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	7,14	1,18	16,47	0,08	0,08	101,30
3	5,27	0,60	11,47	-0,06	0,05	-86,07
4	5,10	0,50	9,78	-0,00	0,08	-1565,38
5	5,06	1,28	25,33	0,17	0,12	73,12
6	4,41	0,60	13,65	0,01	0,09	1242,61
7	3,94	0,36	9,16	-0,01	0,10	-877,12
SVR + SMS						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	3,73	0,46	12,25	0,68	0,07	10,70
3	3,05	0,36	11,78	0,65	0,05	8,06
4	3,20	0,38	11,77	0,61	0,05	9,00
5	2,71	0,52	19,17	0,69	0,12	17,59
6	3,31	0,31	9,27	0,50	0,06	11,46
7	3,05	0,38	12,56	0,47	0,05	10,13
SVR + CART						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	4,42	0,53	11,94	0,61	0,06	10,34
3	3,53	0,48	13,70	0,58	0,04	7,56
4	3,84	0,34	8,90	0,49	0,07	14,53
5	3,77	1,06	28,16	0,50	0,11	21,32
6	3,58	0,40	11,08	0,44	0,08	19,26
7	3,41	0,27	8,00	0,37	0,06	16,81
SVR + FFS						
Cenário	**MSE**	**DP**	**CV**	**COR**	**DP**	**CV**
2	4,29	0,47	10,97	0,62	0,08	13,37
3	2,69	0,40	15,04	0,71	0,04	4,99
4	3,12	0,41	13,15	0,63	0,04	6,09
5	2,68	0,64	23,87	0,71	0,09	13,23
6	3,21	0,48	15,10	0,52	0,05	9,13
7	2,93	0,50	17,07	0,51	0,08	15,90

A redução da correlação foi maior nas ferramentas BLASSO e RR-BLUP do que na associação do SVR com alguma técnica de seleção de atributos. A correlação média do BLASSO e do RR-BLUP reduziu de 47 na primeira geração para 38 no conjunto de dados da 4^a geração, enquanto no SVR + SMS ou SVR + FFS a redução foi de aproximadamente 66 para 63, ou seja um redução pouco significativa. O subconjunto de dados selecionado pela CART não gerou variação na correlação média, se comparado os dados das duas gerações distintas.

6.2.3 Considerações

A seleção de atributos se mostrou um importante recurso no aumento da acurácia, em todos os 7 cenários, pois a redução no número de marcadores gerou um ganho na acurácia média. A etapa de seleção dos mais aptos diminui o número de pares, produzindo uma menor variabilidade genética, dificultando assim a seleção dos marcadores causais. As ferramentas clássicas obtiveram menores correlação, provavelmente pela dificuldade em se mensurar o efeito dos marcadores causais, bem como anular aqueles que não produzem impacto no valor final do fenótipo. A interação entre marcadores aumentou a complexidade no aumento da acurácia e na capacidade de seleção dos marcadores de interesse.

6.3 Análise das próximas gerações

O processo de simulação permitiu a obtenção de dados até a 15^a geração, dessa forma possibilitando avaliar os modelos gerados na 1^a ou 4^a geração no conjunto de dados futuros. O modelo foi treinado em uma geração e somente aplicados nas gerações seguintes.

6.3.1 1^a Geração

A seguir, são apresentados os resultados da aplicação do modelo treinado na 1^a geração e aplicado nas próximas 14 gerações. Em todas as figuras é possível observar a correlação baixa obtida pelo SVR com todos os dados. As Figuras 6.3 e 6.4 exibem os resultados dos cenários 1 e 2, respectivamente, onde observa-se resultados próximos entre as ferramentas, porém a seleção de atributos melhorou a acurácia de forma nítida, com destaque para a associação SVR + SMS. Na etapa de seleção, os dados selecionados pelo FFS obtiveram

acurácias maiores que o SMS, contudo na análise das gerações os dados selecionados pelo SMS obtiveram maiores acurácias.

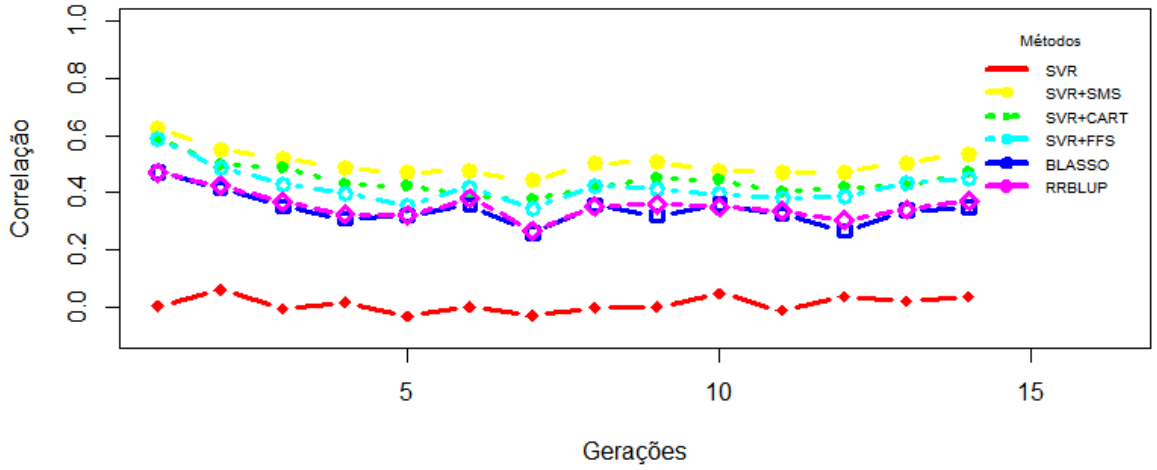


Figura 6.3: Aplicação do modelo treinado na 1^a geração nas gerações subsequentes do cenário 1.

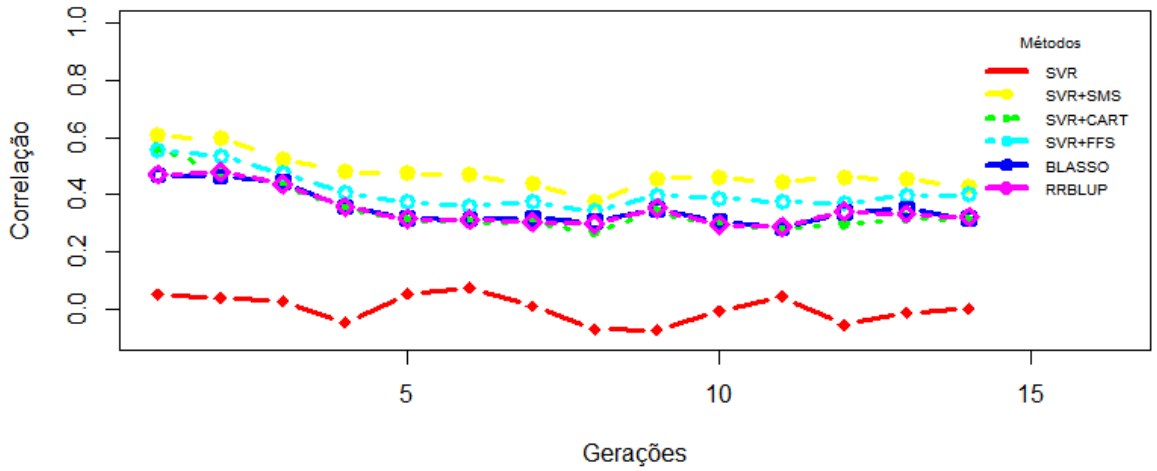


Figura 6.4: Aplicação do modelo treinado na 1^a geração nas gerações subsequentes do cenário 2.

A Figura 6.5 mostra o resultado do cenário 3, onde todas as ações gênicas são aditivas. O comportamento apresentado no gráfico é de acordo com o esperado, ou seja ele é similar entre as ferramentas. O BLASSO e o RR-BLUP são técnicas, em geral, lineares, dessa forma o desempenho delas em cenários aditivos é ótimo. Como visto, todas as ferramentas

possuem comportamento similar entre si, mesmo com o destaque para a associação SVR + SMS. Nesse conjunto de dados o uso da seleção de atributos pode não ser uma solução eficiente, pois o ganho é pequeno. Vale destacar que os modelos associados de SVR + Seleção de Atributos obtiveram correlações, em média, acima de 50%.

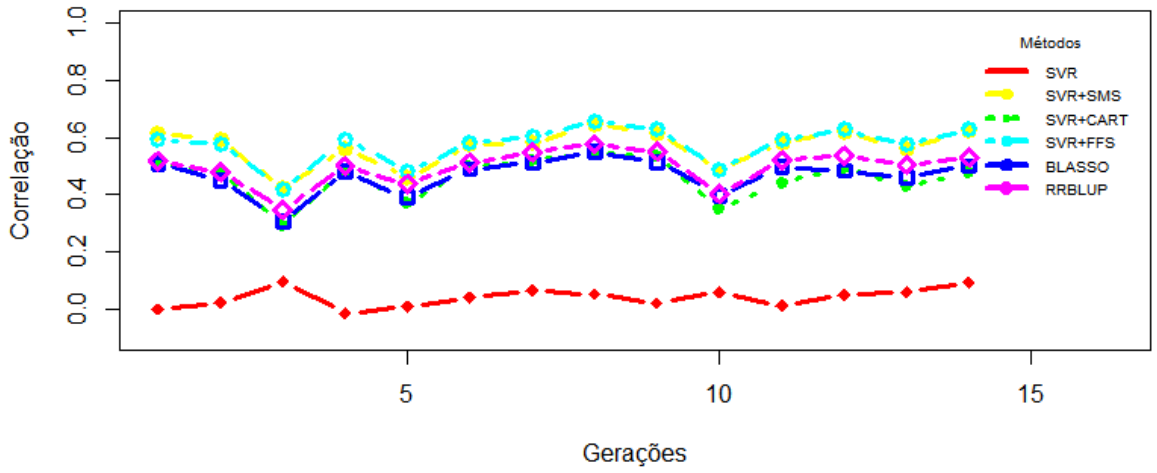


Figura 6.5: Aplicação do modelo treinado na 1^a geração nas gerações subsequentes do cenário 3.

O cenário 4, mostrado na Figura 6.6, é composto por interações epistáticas de ordem 2 e alguns marcadores aditivos. Como é possível observar, a seleção melhorou a acurácia, com destaque para o SMS que obteve melhores resultados comparado com as outras técnicas.

O tamanho da população de indivíduos é um importante recurso para as ferramentas de seleção genômica, com o cenário 5 apresentando um desafio devido ao baixo número de indivíduos. A Figura 6.7 mostra os resultados da análise ao longo das gerações do cenário 5, onde o baixo número de indivíduos impactou todas as ferramentas. A seleção de atributo melhorou a acurácia do modelo, contudo após 7 gerações os valores se aproximam. Esse comportamento não foi observado em outros cenários, onde o ganho obtido pela seleção de atributos se manteve ao longo das gerações.

Os cenários 6 e 7 são apresentados na Figura 6.8 e Figura 6.9, sendo que os dois possuem interações de ordem 3 e 4, gerando assim uma dificuldade para todas as ferramentas. Observa-se que a seleção de atributos não aumentou a acurácia de foram tão significativa como nos cenários anteriores. A seleção efetuada pelo FFS se mostrou o melhor no con-

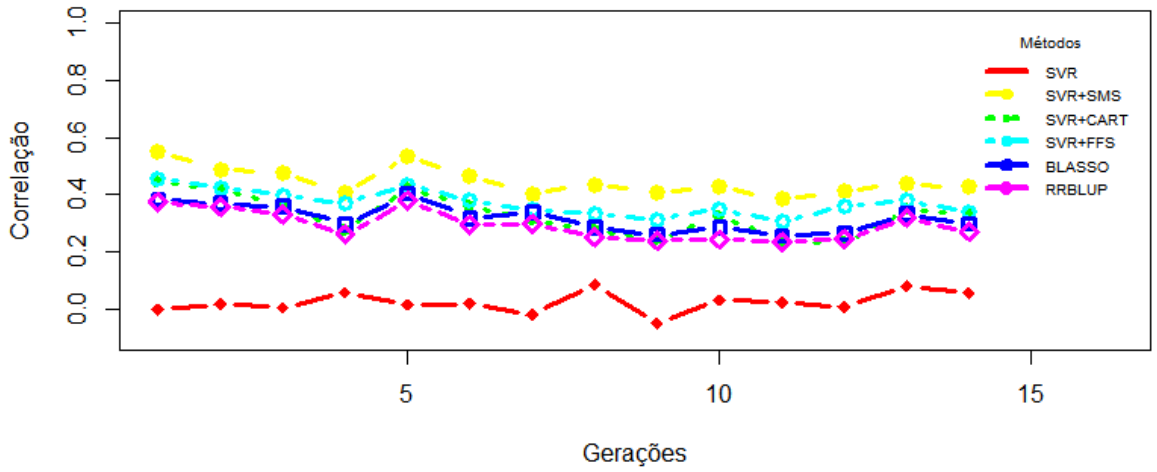


Figura 6.6: Aplicação do modelo treinado na 1^a geração nas subsequentes do cenário 4.

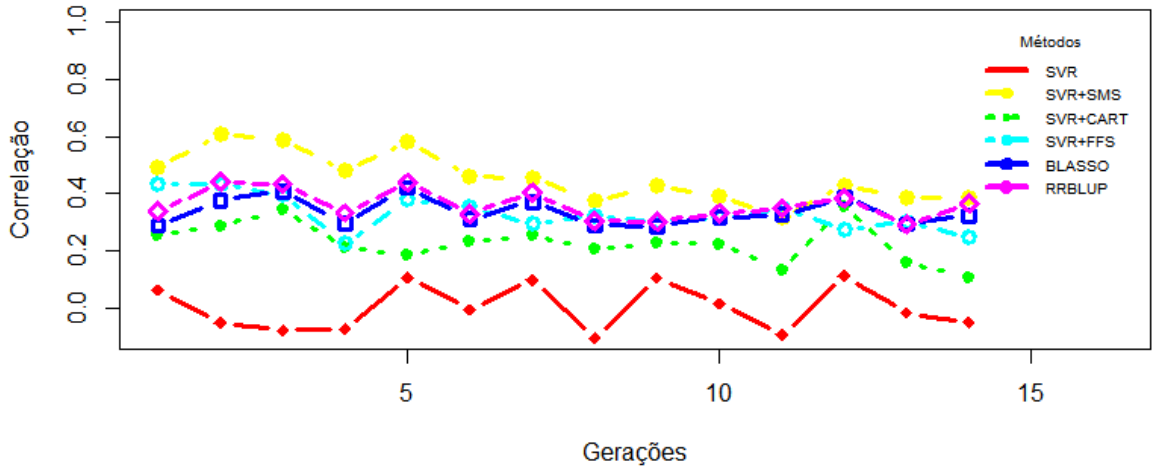


Figura 6.7: Aplicação do modelo treinado na 1^a geração nas gerações subsequentes do cenário 5.

junto do cenário 7. O cenário 6 se mostrou complexo, pois os valores de acurácias ficaram baixos e o comportamento das técnicas foram similares. A correlação obtida pelas ferramentas no cenário 7 mostraram comportamentos diferentes, onde em alguns conjuntos de dados o valor aumenta para algumas ferramentas e diminui em outras.

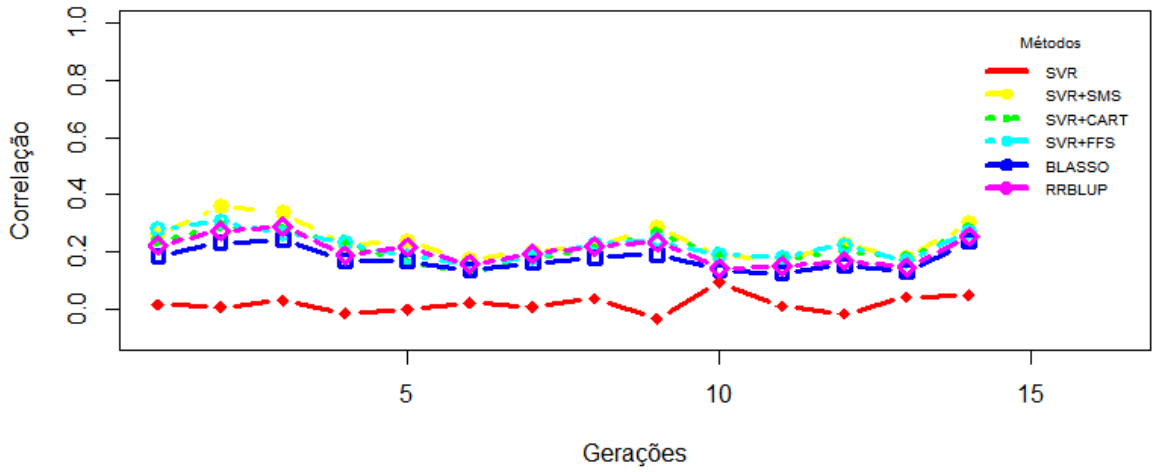


Figura 6.8: Aplicação do modelo treinado na 1^a geração nas gerações subsequentes do cenário 6.

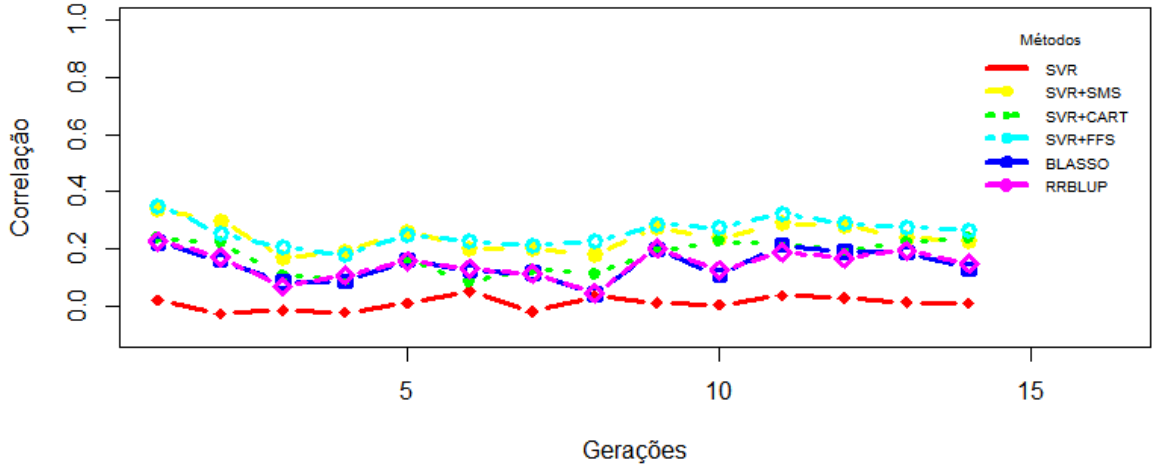


Figura 6.9: Aplicação do modelo treinado na 1^a geração nas subsequentes do cenário 7.

6.3.2 4^a *Geração*

O modelo treinado na 4^a geração foi aplicado nas 10 gerações seguintes. É possível observar nos gráficos um comportamento diferente do modelo treinado na 1^a geração, com um destaque para o uso das técnicas de seleção de atributos. As Figuras 6.10 e 6.11 exibem os resultados dos cenários 1 e 2 onde o aumento da acurácia obtido com a aplicação da seleção de atributos é nítido, com as três técnicas demonstrando resultados similares nesses dois cenários.

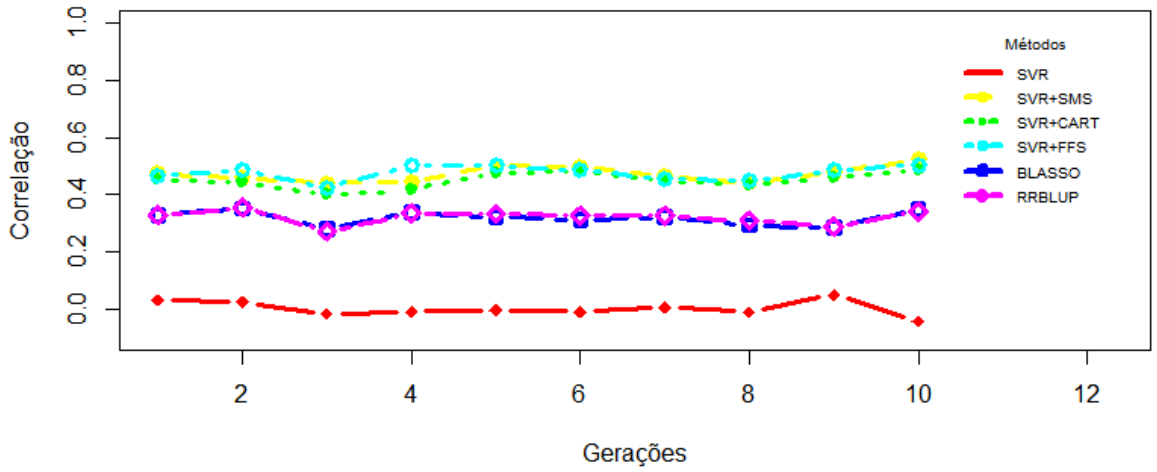


Figura 6.10: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 1.

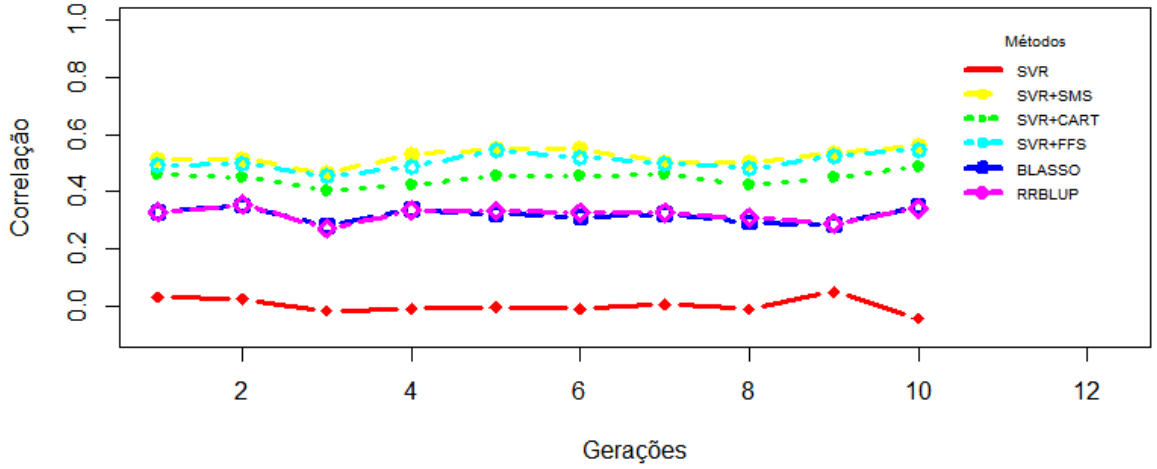


Figura 6.11: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 2.

A Figura 6.12 exibe o resultado no cenário 3, onde, conforme esperado, todas as técnicas obtêm resultados similares. Nesse ponto é possível destacar que o uso de seleção de atributos não é vantajoso em cenários totalmente lineares ou aditivos, contudo é improvável saber a dinâmica de interação dos marcadores sem antes um longo processo de pesquisa.

O cenário 4, mostrado na Figura 6.13, exibe novamente um destaque para as ferra-

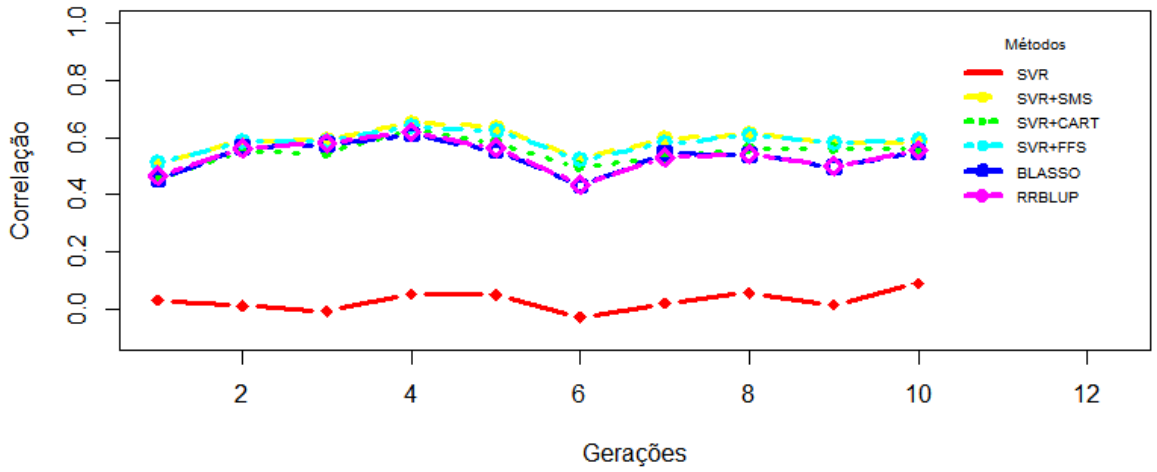


Figura 6.12: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 3.

mentas de seleção de atributos. O aumento na acurácia obtido é próximo dos 40%, sendo um ganho considerável para o cenário em questão. O mesmo comportamento não é observado na Figura 6.14 devido ao baixo número de indivíduos. O enfoque da pesquisa desse trabalho é o uso da seleção de atributos no aumento da acurácia em dados genômicos e, como visto, esse incremento pode não ocorrer na presença de uma baixo número de amostras.

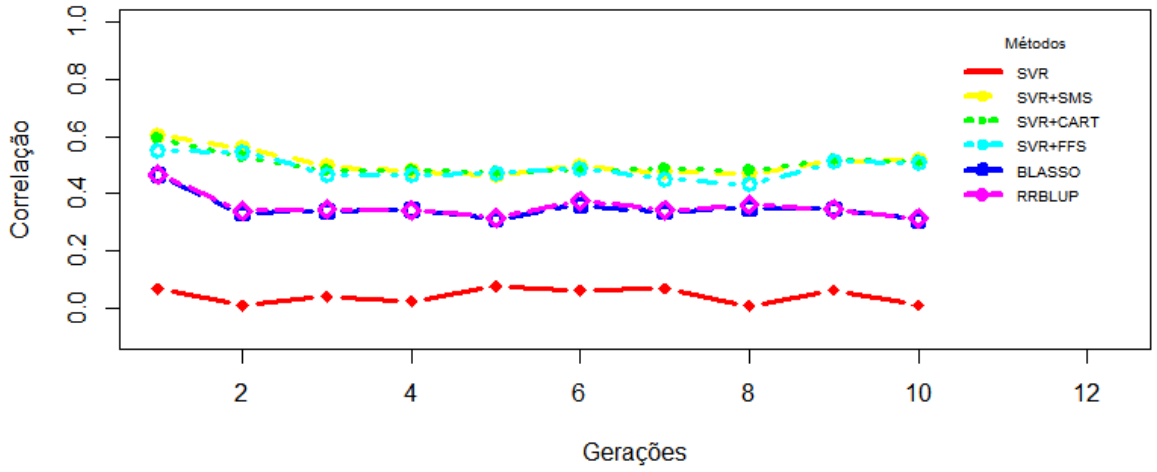


Figura 6.13: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 4.

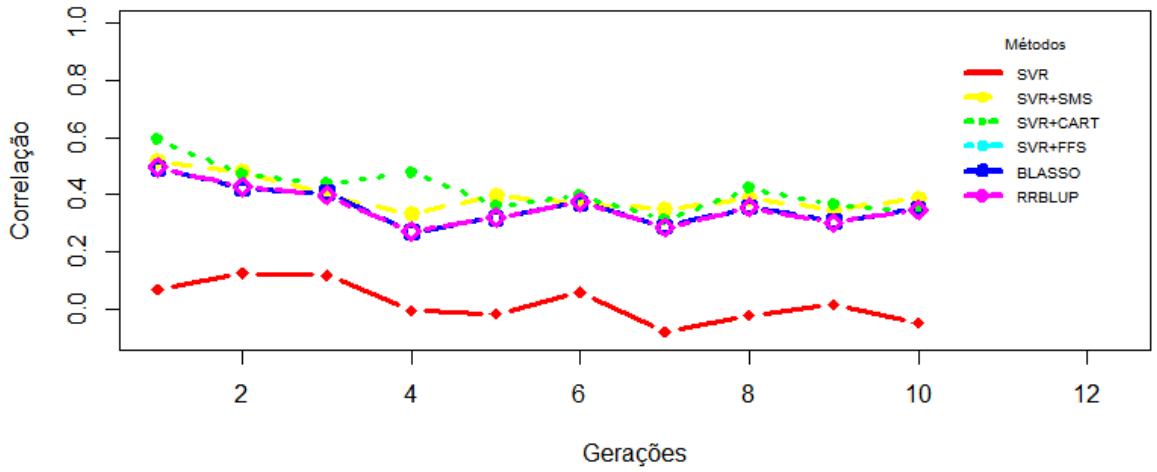


Figura 6.14: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 5.

A complexidade presente nos cenários 6 e 7 dificulta técnicas lineares, como pode ser visto na Figura 6.15 e Figura 6.15. Nessas condições, o uso da seleção de atributos se mostrou eficiente para o aumento da acurácia, chegando a quase dobrar o valor no cenário 7.

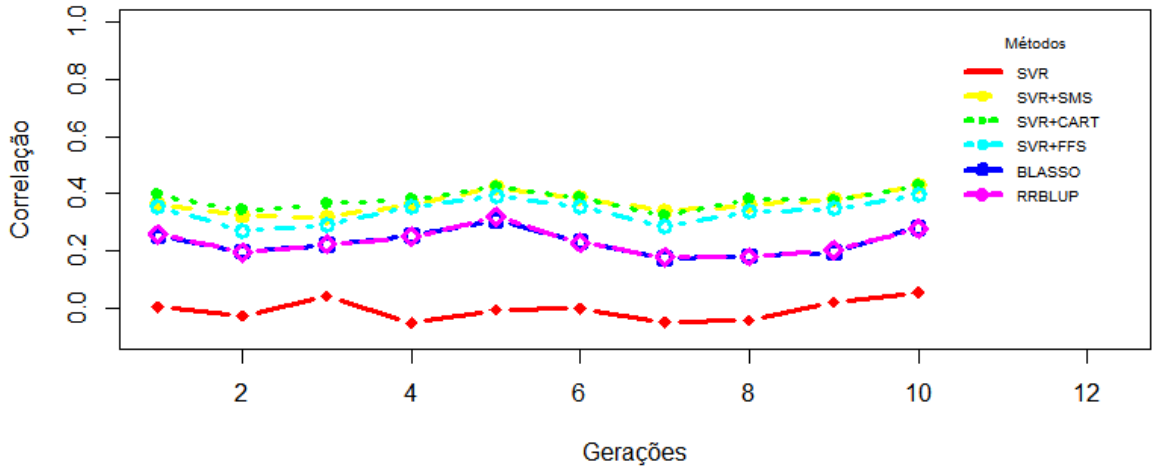


Figura 6.15: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 6.

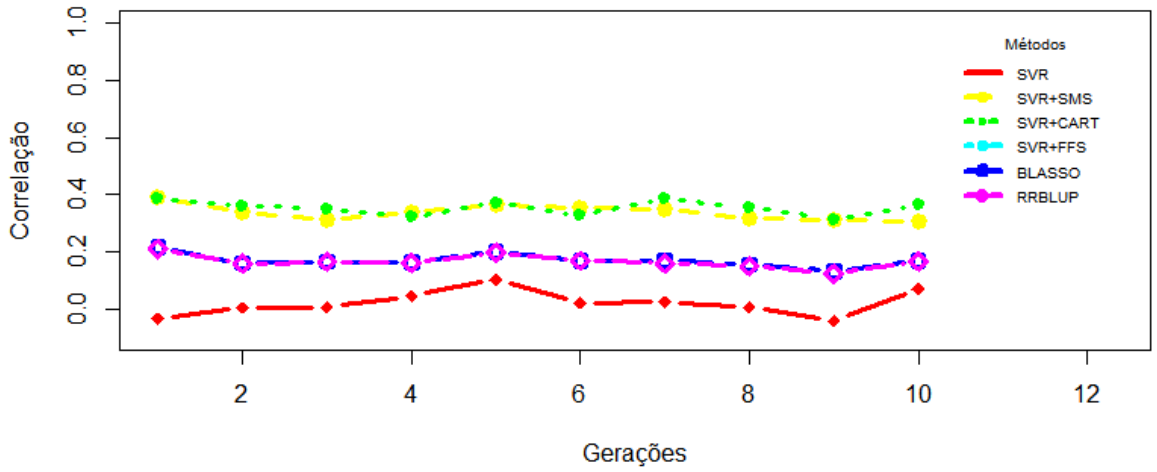


Figura 6.16: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 7.

6.3.3 Considerações

A seleção de atributos em alguns contextos é apresentada como sendo ineficiente no aumento da acurácia em seleção genômica, contudo como mostrado nos experimentos sua aplicação aumentou de forma significativa a acurácia em dados genômicos, com exceção para o conjunto com amostra pequena e em dados totalmente lineares. A seleção se mostrou mais vantajosa quando aplicada na 4^a geração da população gerada. O modelo construído na 1^a geração é importante pois poderá ser o modelo utilizado durante o processo de melhoramento genético animal, logo o uso de técnicas que aumentem a acurácia trazem o benefício de melhoria para o modelo e o resultado por ele apresentado.

6.4 Conjunto de dados baseado no Girolando - cruzamento B

A complexidade e dimensão dos estudos com o Girolando, certamente, não serão esgotados com uma única simulação. O principal objetivo desse conjunto de dados é identificar uma possível dificuldade existente na aplicação da seleção de atributos em raças híbridas. Uma das muitas particularidades reside na dificuldade em se montar um modelo eficiente nas primeiras gerações, devido ao fato de serem animais mistos, logo com características

biológicas distintas. A seguir, será mostrado a tabela com a correlação média dos modelos treinados na geração 1, com touros GIR e vacas HOLANDESAS.

Os modelos treinados na 1^a geração tem o seu resultado expostos na Tabela 6.13. Como é possível verificar, com exceção do SVR na base toda, todos as técnicas obtiveram acurácias altas, com valores próximos de 1. Os valores de desvios padrão baixos, entretanto o MSE é alto para todas as ferramentas. Nos cenários anteriores, o SVR e suas associações obtinham MSE baixos, porém nesse cenário o comportamento foi similar às outras ferramentas. Nesse contexto, o uso da seleção de atributos não melhorou o resultado, pelo contrário, diminuiu a correlação média se comparada com o BLASSO ou RR-BLUP.

Tabela 6.13: Resultado da etapa de avaliação na 1^a geração do cenário 8

	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
RR-BLUP	2428,72	1403,97	57,81	0,97	0,00	0,34
BLASSO	6421,19	3679,62	57,30	0,96	0,00	0,42
SVR	23743,07	83,98	0,35	0,18	0,06	34,30
SVR + SMS	3517,31	278,15	7,91	0,92	0,01	0,67
SVR + CART	6891,70	950,33	13,79	0,84	0,02	2,90
SVR + FFS	3021,30	277,90	9,20	0,93	0,01	0,67

A Tabela 6.14 mostra o resultado da 4^a geração e, comparando-se os modelos, nota-se uma queda na acurácia em todas as ferramentas. O MSE diminui de forma considerável, com destaque para a associação SVR + FFS e SVR + SMS. O desvio padrão das ferramentas ficou baixo mesmo com os dados da 4^a geração, o que pode ser explicado pelo número de indivíduos do estudo, que no cenário 8 é de 2000 amostras, que é o mínimo recomendado para estudos em seleção genômica ampla. Vale destacar o aumento da acurácia obtido pelo uso da seleção de atributos.

Tabela 6.14: Resultado da etapa de avaliação na 4^a geração do cenário 8

	MSE	**DP**	**CV**	**COR**	**DP**	**CV**
RR-BLUP	493,58	30,88	6,26	0,57	0,04	7,44
BLASSO	21,10	8,33	39,47	0,57	0,05	8,59
SVR	8,04	0,81	10,02	0,08	0,06	68,33
SVR + SMS	3,68	0,50	13,53	0,74	0,04	4,98
SVR + CART	4,78	0,53	11,08	0,64	0,03	5,31
SVR + FFS	3,67	0,46	12,65	0,74	0,03	4,24

A diferença entre os resultados da 1^a e a 4^a é nítida também na capacidade de seleção

de atributos, como pode ser visto na Tabela 6.15 e Tabela 6.16. A seleção no primeiro conjunto de dados foi prejudicada, com o valor-p selecionando uma grande quantidade de marcadores, e as técnicas de IC sendo muito restritiva, cobrindo menos da metade dos marcadores causais. O resultado do segundo conjunto de dados é similar ao encontrado nos cenários anteriores, com destaque para a técnicas de IC que foram restritivas e com alta cobertura.

Tabela 6.15: Resultado da etapa de seleção na 1^a geração do cenário 8

	Seleção				**Causais (10 Marcadores)**		
X	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
VP 0,05	1531	168	10,97	76,55	8	10	100
VP e-08	831	101	12,15	41,55	4	9	90
Vp-BonFerroni	1052	111	10,55	52,60	4	10	100
SMS	96	7	7,29	4,80	0	5	50
CART	83	9	10,84	4,15	0	2	20
STEPWISE	74	8	10,81	3,70	0	6	60

Tabela 6.16: Resultado da etapa de seleção na 4^a geração do cenário 8

	Seleção				**Causais (10 Marcadores)**		
X	**Total**	**em LD**	**Sens.**	**Filtro**	**Causais**	**em LD**	**Sens. LD**
VP 0,05	538	194	36,06	26,90	10	10	100
VP e-08	98	87	88,78	4,90	8	8	80
Vp-BonFerroni	163	18	11,04	8,15	0	9	90
SMS	34	31	91,18	1,70	10	10	100
CART	46	46	100,00	2,30	7	7	70
STEPWISE	33	25	75,76	1,65	10	10	100

A acurácia média dos modelos treinados na 1^a geração é alta, contudo esse valor não é mantido nas próximas gerações conforme é mostrado na Figura 6.17. Os valores se mantêm altos durante o cruzamento híbrido, entretanto com a formação do girolando PS a acurácia fica próximo de 10%.

A Figura 6.18 exibe o resultado do treinamento dos modelos na 4^a geração. Como é possível avaliar os modelos mantêm o comportamento ao longo das gerações. O aumento na acurácia obtido com a aplicação da seleção de atributos ficou em torno de 22%, com destaque para a associação entre o SVR + SMS ou o SVR + FFS.

O cenário 8 é complexo em muitos fatores, principalmente devido a junção de dois conjuntos de dados simulados para a geração de um novo. Como visto, não é eficiente

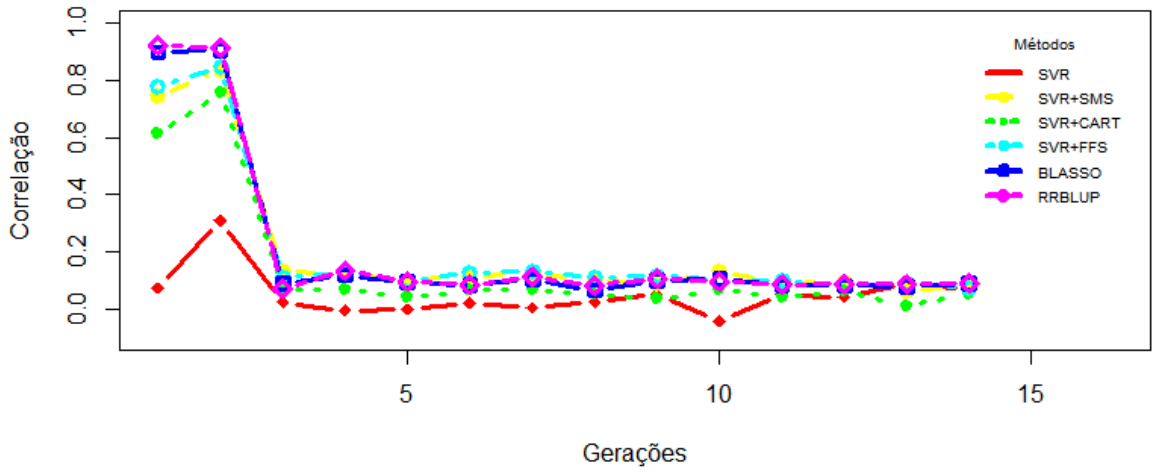


Figura 6.17: Aplicação do modelo treinado na 1^a geração nas subsequentes do cenário 8.

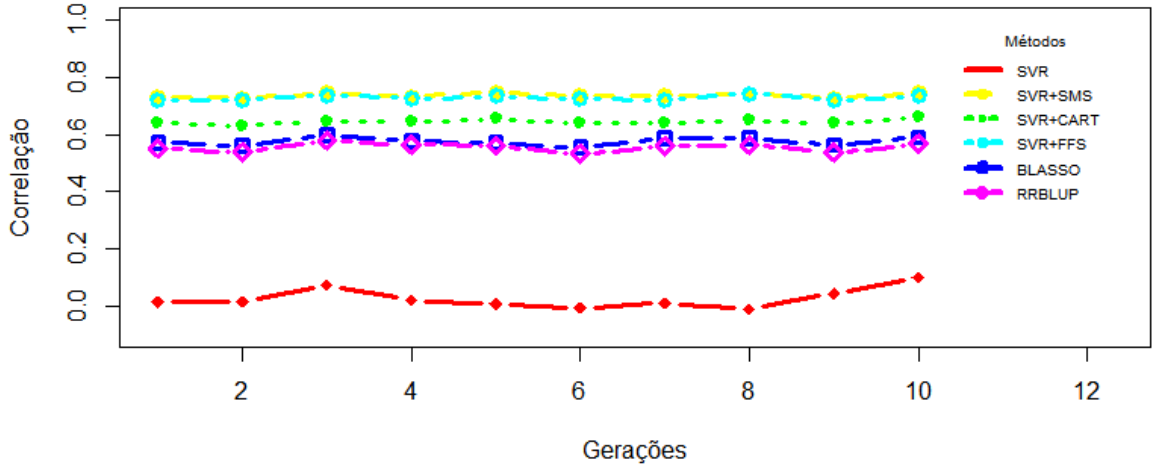


Figura 6.18: Aplicação do modelo treinado na 4^a geração nas gerações subsequentes do cenário 8.

treinar os modelos com os conjuntos de dados da 1^a geração. Os cenários de 1 a 7 possuem população simulada com 1000 indivíduos, e 300 no cenário 5, sendo nesse cenário a população de 2000, o que contribuiu para a manutenção da acurácia do modelo ao longo das gerações, apesar da maior complexidade deste caso.

6.5 Considerações Finais

A seleção de atributos se mostrou uma alternativa para o aumento da acurácia em seleção genômica. A Figura 6.19 exibe um panorama completo com todos os cenários e ferramentas aplicados, nos conjuntos de dados de duas gerações distintas. Como é possível observar, a seleção de atributos aumentou a acurácia média, se comparado com o BLASSO ou o RR-BLUP. Outra questão que pode ser destacada é a diferença entre as duas gerações que foi menor com uso da seleção de atributos.

As ferramentas que utilizam IC, com destaque para o SMS e o FFS, possuem um custo computacional alto. O processo é custoso, pois trabalha com múltiplas soluções e necessita avaliar cada uma delas até encontrar a melhor dentro das disponíveis. Apesar da busca não ser exaustiva é ainda longa. Os métodos clássicos, BLASSO e RR-BLUP, são computacionalmente mais eficientes, entretanto estão restritos a um conjunto de dados específico.

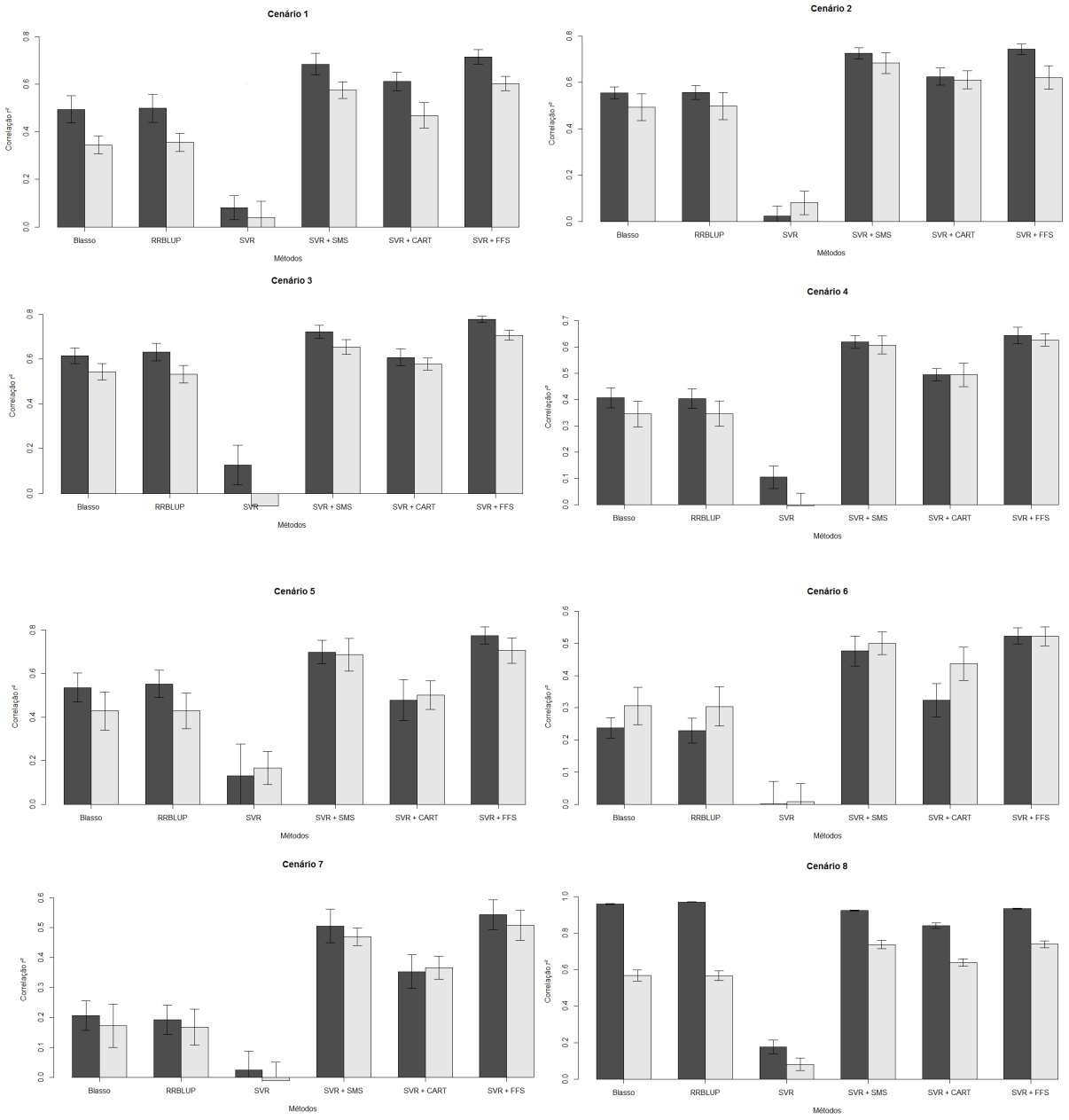


Figura 6.19: Comparativo geral dos cenários.

7 Conclusão

A seleção genômica, certamente, é um campo de estudo amplo e com muitos desafios. Nesse trabalho foi avaliado o impacto da diminuição populacional e da presença de epistasia entre os SNPs causais. O método proposto conseguiu para as bases com as características descritas serem eficientes, gerando um aumento significativo na correlação final.

O modelo apresentado pode ser definido como um fluxo de trabalho onde aplicam-se diversos métodos em uma sequência que visa otimizar a seleção gênica, até mesmo em bases complexa. É importante ressaltar que melhorias pontuais neste fluxo de trabalho podem trazer ganhos de forma geral na seleção.

7.1 Conclusão

O simulador S4GS foi proposto, implementado e utilizado na simulação dos mais variados cenários permitindo uma vasta gama de análises. A simulação de dados é uma importante etapa no processo de avaliação de novas técnicas. A simulação de dados genômicos é complexa e envolve uma diversidade de características e padrões construtivos distintos, e, como visto, o S4GS desenvolvido nesse trabalho se mostrou uma ferramenta eficiente, confiável e bastante útil como suporte ao estudo de seleção genômica.

O uso do S4GS permitiu a simulação computacional de uma raça híbrida de interesse, a saber, o Girolando PS. A simulação dessa raça não foi encontrada em outro trabalho, dessa forma, abrindo possibilidade para pesquisas futuras com o S4GS para analisar outras características do Girolando, que é uma raça de importância econômica para produtores brasileiros.

Os conjuntos de dados com pequenas populações se mostraram um problema complexo, tanto para a seleção quando para a obtenção do GEBV. A seleção de atributos conseguiu aumentar os valores de acurácia se comparando com as ferramentas clássicas, entretanto esse aumento só foi mantido por 7 gerações. Assim, mostra-se necessário um estudo mais amplo nessa linha, buscando novas soluções como o aumento da população via métodos estatísticos ou de inteligência computacional.

A estruturação de um procedimento para o aumento da acurácia por meio da seleção de atributos foi alcançado. A associação do SVR com o SMS ou FFS trouxe nítidas melhorias nos conjuntos analisados, com destaque para os cenários com a presença de interações epistáticas.

A aplicação do método em um cenário com população pequena trouxe benefício, contudo o aumento ficou abaixo do obtido em cenários anteriores. Logo é necessário uma ampliação no estudo de populações com número de amostras considerado reduzido, com a investigação de outras técnicas de regressão ou seleção de atributos, ou mesmo procedimentos para o aumento da população.

O SVR quando aplicado no conjunto completo não se mostrou uma ferramenta eficiente para o uso em seleção genômica, porém quando associado com métodos de seleção de atributos melhorou a acurácia da técnica. Dessa forma, pode-se concluir que o uso do SVR combinado com outro método é viável em problemas de seleção genômica, obtendo, em geral, valores de acurácia maiores que os métodos clássicos analisados nesse trabalho.

É possível aumentar a acurácia em dados genômicos aplicando técnicas de seleção de atributos. Como pode-se avaliar, mostrou-se uma solução eficiente para o aumento da acurácia. Como visto, a qualidade da seleção influencia diretamente na melhoria dos modelos de regressão. Logo, a seleção utilizando o valor-p corrigido por Bonferroni reduziu a acurácia, demonstrando que o processo de redução dos dados necessita ser efetuado de forma orientada e organizada.

A seleção de atributos não se mostrou eficiente em todos os cenários, pois no conjunto com população pequena o aumento foi pouco expressivo, e no cenário composto somente por ações aditivas a acurácia foi similar a obtida pelas técnicas tradicionais.

As interações epistáticas se mostraram um desafio para as ferramentas analisadas. Nos cenários onde essa ação foi mapeada as acurácias foram menores. O aumento na ordem da interação de 2 para 3 ou 4 marcadores interagindo gerou uma nítida diminuição na acurácia dos modelos, conforme era de se esperar. Nesse contexto, a seleção de atributos se mostrou mais relevante, chegando em alguns experimentos a quase dobrar a correlação encontradas pelas ferramentas clássicas.

7.2 Contribuições

Conclui-se como principais contribuições desse trabalho os seguintes pontos:

- A adaptação do FFS utilizando o SVR como avaliador;
- O Simulador S4GS ;
- O cenário simulado do Girolando PS;
- A proposta de separar a etapa de seleção de atributos e avaliação, permitindo uma melhor análise de cada uma delas;
- A análise da aplicação de técnicas de IC em problemas de seleção genômica.

7.3 Trabalhos Futuros

Trabalhos utilizando seleção de atributos em seleção genômica são recorrente, pois o problema de dimensionalidade é real. Logo como sugestão de trabalhos futuros, tem-se:

- Aprimorar o modelo proposto de seleção e avaliação;
- Incrementa de forma eficaz a interação entre as fases;
- Ajustar métodos de seleção de atributos que sejam mais adequados ao problema de seleção genômica sendo computacionalmente eficientes;
- Aumentar o entendimento de seleção em populações com amostras reduzidas, visando buscar um modelo que seja mais robusto e confiável em tais casos;
- Aprimorar, no modelo, a captura dos mapeamentos não lineares;
- Avaliar possíveis melhorias para o simulador, aprimorando o mapeamento das ações gênicas e as formas de sorteio dos pares;
- Novos estudos com simulação de populações híbridas, como a do Girolando utilizada nesse trabalho.

REFERÊNCIAS

AGRESTI, A. An introduction to categorical data analysis. Hoboken, NJ: Wiley-Interscience, 2007.

AGRESTI, A. *Analysis of ordinal categorical data*. [S.l.]: John Wiley & Sons, 2010. v. 656.

AGROPECUáRIA, E. E. B. de P. *Novas ferramentas genômicas mudam a cara do melhoramento genético*. 2017. Disponível em: <https://www.embrapa.br/xxi-ciencia-para-a-vida/busca-de-noticias/-/noticia/21255873/novas-ferramentas-genomicas-mudam-a-cara-do-melhoramento-genetico>.

ANALYTICS, R.; WESTON, S. *doParallel: Foreach parallel adaptor for the parallel package*. [S.l.], 2014. R package version 1.0.8. Disponível em: <http://CRAN.R-project.org/package=doParallel>.

ANALYTICS, R.; WESTON, S. *foreach: Foreach looping construct for R*. [S.l.], 2014. R package version 1.4.2. Disponível em: <http://CRAN.R-project.org/package=foreach>.

ANDERSSON, L. Genetic dissection of phenotypic diversity in farm animals. *Nature Reviews Genetics*, Nature Publishing Group, v. 2, n. 2, p. 130–138, 2001.

ARBEX, W. A. *Modelos Computacionais para Identificação de Informação Genômica Associada à Resistência ao Carrapato Bovino*. Tese (Doutorado) — UFRJ/COPPE/Programa de Engenharia de Sistemas e Computação, 2009.

BARRETT, J. C. et al. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, Oxford Univ Press, v. 21, n. 2, p. 263–265, 2005.

BREIMAN, L. et al. Classification and regression trees. Wadsworth, 1984.

BRONDANI, R. P. V.; BRONDANI, C. Germoplasma: base para a nova agricultura. *Ciência Hoje*, v. 35, n. 207, p. 70–73, 2004.

BROWN, T. A. *Genomes*. [S.l.]: Garland science, 2006.

CAMPOS, G. D. L. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, Genetics Soc America, v. 182, n. 1, p. 375–385, 2009.

CAMPOS, G. de L. et al. Genome-enabled prediction using the BLR (Bayesian Linear Regression) R-package. *Methods in molecular biology (Clifton, N.J.)*, v. 1019, p. 299–320, 2013. ISSN 1940-6029. Disponível em: <http://view.ncbi.nlm.nih.gov/pubmed/23756896>.

CONSORTIUM, I. H. G. S. et al. Initial sequencing and analysis of the human genome. *Nature*, Nature Publishing Group, v. 409, n. 6822, p. 860, 2001.

CORDELL, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics*, Oxford Univ Press, v. 11, n. 20, p. 2463–2468, 2002.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine learning*, v. 20, n. 3, p. 273–297, 1995.

CROW, J. F.; KIMURA, M. et al. An introduction to population genetics theory. *An introduction to population genetics theory.*, New York, Evanston and London: Harper & Row, Publishers, 1970.

DAETWYLER, H. D.; VILLANUEVA, B.; WOOLLIAMS, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*, v. 3, n. 10, p. e3395, 2008.

DASH, M.; LIU, H. Consistency-based search in feature selection. *Artificial intelligence*, Elsevier, v. 151, n. 1, p. 155–176, 2003.

DRAPER, N. R.; SMITH, H.; POWNELL, E. *Applied regression analysis*. [S.l.]: Wiley New York, 1966. v. 3.

DRUCKER, H. et al. Support vector regression machines. *Advances in neural information processing systems*, Morgan Kaufmann Publishers, v. 9, p. 155–161, 1997.

ELER, J. P. *TEORIAS E MÉTODOS EM MELHORAMENTO GENÉTICO ANIMAL*. [S.l.], 2014. Disponível em: <www.usp.br/gmab/discip/apos1.pdf>.

ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with r package rrblup. *Plant Genome*, v. 4, p. 250–255, 2011.

EWENS, W. Foundations of mathematical genetics. *American journal of human genetics*, Elsevier, v. 29, n. 5, p. 545, 1977.

FALCONER, D. et al. Introduction to quantitative genetics. *Introduction to quantitative genetics.*, Edinburgh and London: Oliver & Boyd Ltd., 1960.

FALCONER, D. S. *Introduction to quantitative genetics*. [S.l.]: Pearson Education India, 1975.

FARNIR, F. et al. Extensive genome-wide linkage disequilibrium in cattle. *Genome research*, Cold Spring Harbor Lab, v. 10, n. 2, p. 220–227, 2000.

FARNIR, F. et al. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics*, Genetics Soc America, v. 161, n. 1, p. 275–287, 2002.

FISHER, R. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the royal society of Edinburgh*, Cambridge Univ Press, v. 52, n. 02, p. 399–433, 1919.

FISHER, R. The genetical theory of natural selection. Clarendon Press, 1930.

GALLAIS, A. Covariances between arbitrary relatives with linkage and epistasis in the case of linkage disequilibrium. *Biometrics*, JSTOR, p. 429–446, 1974.

GEISSER, S. *Predictive Inference: An Introduction*. [S.l.]: Chapman & Hall, 1993.

GIANOLA, D. Priors in whole-genome regression: the bayesian alphabet returns. *Genetics*, v. 194, p. 573–596, 2013.

GIROLANDO, A. B. dos Criadores de. *Generalidade: FATOS E DADOS HISTÓRICOS*. 2017. Disponível em: <http://www.girolando.com.br/index.php?paginasSite/girolando,2,pt>.

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

GOLDBERGER, A. S. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 57, n. 298, p. 369–375, 1962.

GRIFFITHS, A. J. *Introdução à genética*. [S.l.]: Guanabara Koogan, 2008.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, JMLR. org, v. 3, p. 1157–1182, 2003.

HALDANE, J. B. S. A mathematical theory of natural and artificial selection, part v: selection and mutation. In: CAMBRIDGE UNIVERSITY PRESS. *Mathematical Proceedings of the Cambridge Philosophical Society*. [S.l.], 1927. v. 23, n. 7, p. 838–844.

HAMON, J. *Optimisation combinatoire pour la sélection de variables en régression en grande dimension: Application en génétique animale*. Tese (Doutorado) — Université des Sciences et Technologie de Lille-Lille I, 2013.

HAPMAP, C. I. The international hapmap project. *Nature*, v. 426, n. 6968, p. 789 – 96, 2003. Disponível em: <http://dx.doi.org/10.1038/nature02168>.

HAWS, D. C. et al. Variable-selection emerges on top in empirical comparison of whole-genome complex-trait prediction methods. *PloS one*, Public Library of Science, v. 10, n. 10, p. e0138903, 2015.

HAYES, B. et al. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, Elsevier, v. 92, n. 2, p. 433–443, 2009.

HAYES, B.; GODDARD, M. Genome-wide association and genomic selection in animal breeding. *Genome*, NRC Research Press, v. 53, n. 11, p. 876–883, 2010.

HE, D.; WANG, Z.; PARADA, L. Mined: An efficient mutual information based epistasis detection method to improve quantitative genetic trait prediction. In: SPRINGER. *International Symposium on Bioinformatics Research and Applications*. [S.l.], 2015. p. 108–124.

HILL, W.; ROBERTSON, A. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, Springer, v. 38, n. 6, p. 226–231, 1968.

HOCKING, R. R. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, International Biometric Society, v. 32, n. 1, p. pp. 1–49, 1976. ISSN 0006341X. Disponível em: <http://www.jstor.org/stable/2529336>.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes| Genomes| Genetics*, Genetics Society of America, v. 4, n. 6, p. 1027–1046, 2014.

JOHN, G. H. et al. Irrelevant features and the subset selection problem. In: *Machine Learning: Proceedings of the Eleventh International Conference*. [S.l.: s.n.], 1994. p. 121–129.

KNUEPPEL, S.; ROHDE, K. Cran - package hapestxxr. *CRAN*, 2015. Disponível em: <http://cran.r-project.org/web/packages/HapEstXXR/index.html>.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. *Artificial Intelligence*, v. 97, p. 273–324, 1997.

LAL, T. N. et al. Embedded methods. In: *Feature extraction*. [S.l.]: Springer, 2006. p. 137–165.

LEWONTIN, R. On measures of gametic disequilibrium. *Genetics*, Genetics Soc America, v. 120, n. 3, p. 849–852, 1988.

LIU, C. et al. A new validity index of feature subset for evaluating the dimensionality reduction algorithms. *Knowledge-Based Systems*, Elsevier, 2017.

LIU, H. et al. The impact of genetic relationship and linkage disequilibrium on genomic selection. *PloS one*, Public Library of Science, v. 10, n. 7, p. e0132379, 2015.

LONG, N. et al. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in holsteins. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 128, n. 4, p. 247–257, 2011.

LONG, N. et al. Machine learning classification procedure for selecting snps in genomic selection: application to early mortality in broilers. *Journal of animal breeding and genetics*, Wiley Online Library, v. 124, n. 6, p. 377–389, 2007.

MA, J.; SONG, A.; XIAO, J. A robust static decoupling algorithm for 3-axis force sensors based on coupling error model and ε-svr. *Sensors*, Molecular Diversity Preservation International, v. 12, n. 11, p. 14537–14555, 2012.

MACARTHUR, J. et al. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, Oxford University Press, v. 45, n. D1, p. D896–D901, 2017.

MCKAY, S. D. et al. Whole genome linkage disequilibrium maps in cattle. *BMC genetics*, BioMed Central Ltd, v. 8, n. 1, p. 74, 2007.

MENDEL, G. Versuche über pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn 4: 3*, v. 44, 1866.

MÉSZÁROS, G. et al. Genomic analysis for managing small and endangered populations: A case study in tyrol grey cattle. *Frontiers in Genetics*, Frontiers, v. 6, p. 173, 2015.

MEUWISSEN, T. H. E.; GODDARD, M. E. Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, v. 155, n. 4, p. 421–430, 2000.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, v. 157, n. 4, p. 1819–1829, 2001. Disponível em: <http://www.genetics.org/content/157/4/1819.abstract>.

MEYER, D. et al. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*. [S.l.], 2014. R package version 1.6-3. Disponível em: <http://CRAN.R-project.org/package=e1071>.

MOSER, G. et al. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. *Genet Sel Evol*, v. 41, n. 1, p. 56, 2009.

NIH, N. C. I. *Genetic Simulation Resources (GSR): Software Resource List*. 2016. Acessado em: 17-10-2016. Disponível em: <https://popmodels.cancercontrol.cancer.gov/gsr/packages/>.

ÜNSTü, B.; MELSSEN, W.; BUYDENS, L. Facilitating the application of support vector regression by using a universal pearson vii function based kernel. *Chemometrics and Intelligent Laboratory Systems*, v. 81, p. 29–40, 2006.

OLIVEIRA, F. C. de. *Um método para seleção de atributos em dados genômicos*. Tese (Doutorado) — UFJF/PGMC/Programa de Pós Graduação em Modelagem Computacional, 2015.

OLIVEIRA, F. C. de et al. Metodologia para seleção de marcadores com máquina de vetores suporte com regressão. In: _____. [S.l.]: Embrapa, 2014. p. 101–126. ISBN 978-85-7035-382-5.

OLIVEIRA, F. C. de et al. Snps selection using support vector regression and genetic algorithms in gwas. *BMC genomics*, BioMed Central Ltd, v. 15, n. Suppl 7, p. S4, 2014.

PARK, T.; CASELLA, G. The bayesian lasso. *Journal of the American Statistical Association*, Taylor & Francis, v. 103, n. 482, p. 681–686, 2008.

PÉREZ-RODRÍGUEZ, P. et al. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes| Genomes| Genetics*, Genetics Society of America, v. 2, n. 12, p. 1595–1605, 2012.

PHILLIPS, P. C. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, Nature Publishing Group, v. 9, n. 11, p. 855–867, 2008.

PHUONG, T. M.; LIN, Z.; ALTMAN, R. B. Choosing snps using feature selection. In: IEEE. *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*. [S.l.], 2005. p. 301–309.

QIU, Z. et al. Application of machine learning-based classification to genomic selection and performance improvement. In: SPRINGER. *International Conference on Intelligent Computing*. [S.l.], 2016. p. 412–421.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2015. Disponível em: <http://www.R-project.org/>.

REICH, D. E. et al. Linkage disequilibrium in the human genome. *Nature*, Nature Publishing Group, v. 411, n. 6834, p. 199–204, 2001.

RESENDE, M. D. V. d. et al. Computação da seleção genômica ampla (gws). Embrapa Florestas, 2010.

ROBERTSON, A. A theory of limits in artificial selection. *Proceedings of the Royal Society of London B: Biological Sciences*, The Royal Society, v. 153, n. 951, p. 234–249, 1960.

ROOS, A. D. et al. Linkage disequilibrium and persistence of phase in holstein–friesian, jersey and angus cattle. *Genetics*, Genetics Soc America, v. 179, n. 3, p. 1503–1512, 2008.

SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. *bioinformatics*, Oxford Univ Press, v. 23, n. 19, p. 2507–2517, 2007.

SARGOLZAEI, M.; SCHENKEL, F. S. Qmsim: a large-scale genome simulator for livestock. *Bioinformatics*, Oxford Univ Press, v. 25, n. 5, p. 680–681, 2009.

SCHAEFFER, L. Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 123, n. 4, p. 218–223, 2006.

SCHWENDER, H. *Statistical Analysis of Genotype and Gene Expression Data*. Tese (Doutorado) — the Department of Statistics of the University of Dortmund, 2 2007.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, n. 3, p. 199–222, 2004.

SOLBERG, T. R. et al. Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution*, BioMed Central, v. 41, n. 1, p. 29, 2009.

STAŃCZYK, U.; JAIN, L. C. *Feature Selection for Data and Pattern Recognition*. [S.l.]: Springer, 2015. v. 584.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. *rpart: Recursive Partitioning and Regression Trees*. [S.l.], 2015. R package version 4.1-10. Disponível em: <http://CRAN.R-project.org/package=rpart>.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, JSTOR, p. 267–288, 1996.

VAPNIK, V. N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995. v. 1.

VERBYLA, K. L. et al. Accuracy of genomic selection using stochastic search variable selection in australian holstein friesian dairy cattle. *Genetics research*, Cambridge University Press, v. 91, n. 5, p. 307, 2009.

WANG, D. et al. Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. *Heredity*, Nature Publishing Group, v. 109, n. 5, p. 313–319, 2012.

WANG, S.; BASTEN, C.; ZENG, Z. Windows qtl cartographer 2.5. *Department of statistics, North Carolina state university, Raleigh, NC*, 2007.

WELTER, D. et al. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic acids research*, Oxford Univ Press, v. 42, n. D1, p. D1001–D1006, 2014.

WESTON, J. et al. Consistency-based search in feature selection. *Advances in Neural Information Processing Systems*, MIT Press, v. 12, p. 526?–532, 2000.

WIMMER, V. et al. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*, Genetics Soc America, v. 195, n. 2, p. 573–587, 2013.

WRIGHT, S. Evolution in mendelian populations. *Genetics*, Genetics Soc America, v. 16, n. 2, p. 97–159, 1931.

XU, S.; JIA, Z. Genomewide analysis of epistatic effects for quantitative traits in barley. *Genetics*, Genetics Soc America, v. 175, n. 4, p. 1955–1963, 2007.

YAO, L. et al. Sparse support vector machine with lp penalty for feature selection. *Journal of Computer Science and Technology*, Springer US, v. 32, n. 1, p. 68–77, 2017.

YTOURNEL, F. et al. Ldso: A program to simulate pedigrees and molecular information under various evolutionary forces. *Journal of Animal Breeding and Genetics*, Wiley Online Library, v. 129, n. 5, p. 417–421, 2012.

APÊNDICE A - Código para geração do cenário 1

```
library('methods')
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 1000



num_individuos<-pop; #Número de indivíduos na amostra
num_snp<-2000; #Total de marcadores simulados
sd=sqrt(3) # Variavel Importante na simulação

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 0

gen = 100

list.ia<-list(c(-3,-3),0,0,0,3,3,-3,c(-3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))
```

```
list.snp<-list(c(100,200),400,600,900,1200,1400,1500,c(1700,1900))
list.snp<-c(list.snp, seq(1,num_snp)
[-c(100,200,400,600,900,1200,1400,1500,1700,1900)])

beta=c(1.8,1.2,1.4,1.5,1.9,1.7,1.8,1.9,runif(1990,0,0.001))

beta0=500

source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')

PopEmMelhoramento<-execSimulation(PopInicial$ind,
num_individuos,
4,
topMacho = topMacho,
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
```

```
nameFile='PopEmMelhoramento')

PosSelecao<-execSimulation(PopEmMelhoramento@generation[
[PopEmMelhoramento@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')

save.image('Cenario1.RData')
```

APÊNDICE B - Código para geração do cenário 2

```
library('methods')
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 1000



num_individuos<-pop;
num_snp<-2000;
sd=sqrt(3)

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 2
gen = 100

list.ia<-list(c(-3,-3),0,0,0,3,3,-3,c(-3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))
```

```
list.snp<-list(c(100,200),400,600,900,1200,1400,1500,c(1700,1900))
list.snp<-c(list.snp, seq(1,num_snp)[-c(100,200,400,600,
900,1200,1400,1500,1700,1900)])

beta=c(1.8,1.2,1.4,1.5,1.9,1.7,1.8,1.9,runif(1990,0,0.001))

beta0=500

source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')




PopEmMelhoramento<-execSimulation(PopInicial$ind,
num_individuos,
4,
topMacho = topMacho,
```

```
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PopEmMelhoramento')



PosSelecao<-execSimulation(PopEmMelhoramento@generation[[temp@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')


save.image('Cenario2.RData')
```

APÊNDICE C - Código para geração do cenário 3

```
library('methods')
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 1000

num_individuos<-pop; #Número de indivíduos na amostra
num_snp<-2000; #Total de marcadores simulados
sd=sqrt(3) # Variavel Importante na simulação

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 1
gen = 100

list.ia<-list(0,0,0,0,0,0,0,0,0,0)
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))

list.snp<-list(100,200,400,600,900,1200,1400,1500,1700,1900)
```

```
list.snp<-c(list.snp, seq(1,num_snp)[-c(100,200,400,600,900,
1200,1400,1500,1700,1900)])

beta=c(1.8,1.3,1.1,1.2,1.4,1.5,1.9,1.7,1.8,1.9,runif(1990,0,0.001))
beta0=500

source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')

PopEmSelecao<-execSimulation(PopInicial$ind,
num_individuos,
4,
topMacho = topMacho,
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PopEmSelecao')
```

```
PosSelecao<-execSimulation(PopEmSelecao@generation[[temp@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')

save.image('Cenario3.RData')
```

APÊNDICE D - Código para geração do cenário 4

```
library('methods')
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 1000



num_individuos<-pop;
num_snp<-2000;
sd=sqrt(3)

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 1
gen = 100

list.ia<-list(c(-3,-3),c(3,3),c(-3,3),c(3,-3),c(-3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))
```

```
list.snp<-list(c(100,200),c(400,600),c(900,1200),c(1400,1500),c(1700,1900))
list.snp<-c(list.snp, seq(1,num_snp)[-c
    (100,200,400,600,900,1200,1400,1500,1700,1900)])

beta=c(1.7,1.4,1.6,1.8,1.9,runif(1990,0,0.001))

beta0=500

source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')


PopEmMelhoramento<-execSimulation(PopInicial$ind,
num_individuos,
4,
topMacho = topMacho,
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
```

```
pedFile=T,
nameFile='PopEmMelhoramento')



PosSelecao<-execSimulation(PopEmMelhoramento@generation[[
    PopEmMelhoramento@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')

save.image('Cenario4.RData')
```

APÊNDICE E - Código para geração do cenário 5

```
library('methods')
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 30
topFemea = 60
pop = 300



num_individuos<-pop;
num_snp<-2000;
sd=sqrt(3)

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 1
gen = 100

list.ia<-list(c(-3,-3),0,0,0,3,3,-3,c(-3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))
```

```
list.snp<-list(c(100,200),400,600,900,1200,1400,1500,c(1700,1900))
list.snp<-c(list.snp, seq(1,num_snp)[-c
    (100,200,400,600,900,1200,1400,1500,1700,1900)])

beta=c(1.8,1.2,1.4,1.5,1.9,1.7,1.8,1.9,runif(1990,0,0.001))

beta0=500
beta0=500


source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')




PopEmMelhoramento<-execSimulation(PopInicial$ind,
num_individuos,
```

```
4,
topMacho = topMacho,
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PopEmMelhoramento')



PosSelecao<-execSimulation(PopEmMelhoramento@generation[[
    PopEmMelhoramento@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')


save.image('Cenario5.RData')
```

APÊNDICE F - Código para geração do cenário 6

```
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 1000

num_individuos<-pop;
num_snp<-2000;
sd=sqrt(3)

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 1
gen = 100

list.ia<-list(c(-3,-3,-3),0,c(3,3,3),c(-3,3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))

list.snp<-list(c(100,200,400),600,c(900,1200,1400),c(1500,1700,1900)) #Indica
    quais são os marcadores causais
```

```
list.snp<-c(list.snp, seq(1,num_snp)[-c
    (100,200,400,600,900,1200,1400,1500,1700,1900)])

beta=c(1.8,1.2,1.5,1.9,runif(1990,0,0.001))

beta0=500

source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')


PopEmMelhoramento<-execSimulation(PopInicial$ind,
num_individuos,
4,
topMacho = topMacho,
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PopEmMelhoramento')
```

```
PosSelecao<-execSimulation(PopEmMelhoramento@generation[[
    PopEmMelhoramento@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')


save.image('Cenario6.RData')
```

APÊNDICE G - Código para geração do cenário 7

```
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 1000

num_individuos<-pop
num_snp<-2000
sd=sqrt(3)

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 1
gen = 100

list.ia<-list(c(-3,-3,-3,-3),c(-3,3),c(3,3,3,3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))

list.snp<-list(c(100,200,400,600),c(900,1200),c(1400,1500,1700,1900)) #Indica
    quais são os marcadores causais
```

```
list.snp<-c(list.snp, seq(1,num_snp)[-c
    (100,200,400,600,900,1200,1400,1500,1700,1900)])

beta=c(1.8,1.7,1.9,runif(1990,0,0.001))

beta0=500

source('SimuladorNOVAVERSAO.R')
source('GerarBancoSemen.R')

updateSpermBank()
updateSpermBank()
updateSpermBank()
updateSpermBank()

PopInicial<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0,
beta,
list.ia,
pedFile=T,
nameFile = 'PopInicial')


PopEmMelhoramento<-execSimulation(PopInicial$ind,
num_individuos,
4,
topMacho = topMacho,
topFemea = topFemea,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PopEmMelhoramento')
```

```
PosSelecao<-execSimulation(PopEmMelhoramento@generation[[
    PopEmMelhoramento@TotalGen]],
num_individuos,
10,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = useArtificialInsemination,
pedFile=T,
nameFile='PosSelecao')


save.image('Cenario7.RData')
```

APÊNDICE H - Código para geração do cenário 8

```
library('methods')
library("HapEstXXR")

breaks<-vector()
pointBreak<-vector()
spermBank<-data.frame()

cross=.09
mutation=.001

topMacho = 50
topFemea = 200
pop = 2000

num_individuos<-pop;
num_snp<-2000;
sd=sqrt(3)

sendSpermBank = 150
numberSpermBank = 150

useArtificialInsemination = 0
gen = 100

source('SimuladorNOVAVERSAO.R')

list.ia<-list(c(-3,-3),0,0,0,3,3,-3,c(-3,-3))
list.ia<-c(list.ia,sample(0, 1990,replace=TRUE))
```

```
list.snp<-list(c(100,200),400,600,900,1200,1400,1500,c(1700,1900)) #Indica
    quais são os marcadores causais
list.snp<-c(list.snp, seq(1,num_snp)[-c
    (100,200,400,600,900,1200,1400,1500,1700,1900)])

betaGIR=c(1.1,1.15,1.2,1.25,1.29,1.37,1.48,1.69,runif(1990,0,0.001))
beta0GIR=1200

GIR<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0GIR,
betaGIR,
list.ia,
pedFile=T,
nameFile = 'PopInicialGIR')


betaHOL=c(2.1,2.15,2.2,2.25,2.29,2.37,2.48,2.69,runif(1990,0,0.001))
beta0HOL=1500

HOL<-initialPopulation(gen,
num_individuos,
num_snp,sd,
list.snp,
beta0HOL,
betaHOL,
list.ia,
pedFile=T,
nameFile = 'PopInicialHOL')
```

```
beta<-betaGIR
beta0<-beta0GIR
F1Gir<-execSimulation(GIR$ind,
num_individuos,
1,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = 0,
pedFile=T,
nameFile='F1Gir')



beta<-betaHOL
beta0<-beta0HOL
F1Hol<-execSimulation(HOL$ind,
num_individuos,
1,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = 0,
pedFile=T,
nameFile='F1Hol')



#População Parental, formada por machos GIR e Fêmeas Holandes.
ind<-SimSNPglmIND()

#Touro GIR
ind@father<-which(GIR$ind@sex==1)
#Vaca HOLANDESA
```

```
ind@mother<-which(HOL$ind@sex==2)

num_individuos<-length(ind@father) + length(ind@mother)
ind@id<-seq(1,num_individuos)

Girolando<-SimSNPglmGEN()
Girolando@TotalGen = 1
Girolando@totalInd = num_individuos
Girolando@lastId = num_individuos + 1

ind@sex   <- c(rep(1,length(ind@father)) , rep(2,length(ind@mother)) )
ind@genotype<-rbind(GIR$ind@genotype[ind@father,],HOL$ind@genotype[ind@mother
    ,])
ind@phenotype<-c(GIR$ind@phenotype[ind@father],HOL$ind@phenotype[ind@mother])
ind@err<-c(GIR$ind@err[ind@father],HOL$ind@err[ind@mother])
ind@GEBV <- ind@phenotype
ind@TBV <- ind@phenotype - ind@err
ind@SVR <- rep(0,num_individuos)#predict(svm.model,ind@genotype)

Girolando@generation[[1]]<-ind
Girolando@h2[1]<-var(ind@TBV)/var(ind@GEBV)

beta<-betaHOL/2 + betaGIR/2
beta0<-beta0HOL/2 + beta0GIR/2
F1Girolando<-execSimulation(ind,
num_individuos,
1,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = 0,
pedFile=T,
nameFile='F1Girolando')
```

```
#
# Segunda Geração do girolando Macho Gir F1 e Femea Meio Sangue Girolando.
#


ind<-SimSNPglmIND()

#Touro GIR
ind@father<-which(F1Gir@generation[[2]]@sex==1)
#Vaca F1 Girolando
ind@mother<-which(F1Girolando@generation[[2]]@sex==2)

num_individuos<-length(ind@father) + length(ind@mother)
ind@id<-seq(Girolando@lastId,(Girolando@lastId + num_individuos-1))

Girolando@TotalGen = Girolando@TotalGen + 1
Girolando@totalInd = Girolando@totalInd + num_individuos
Girolando@lastId = Girolando@lastId + num_individuos


ind@sex   <- c(rep(1,length(which(F1Gir@generation[[2]]@sex==1))) , rep(2,
    length(which(F1Girolando@generation[[2]]@sex==2))))
ind@genotype<-rbind(F1Gir@generation[[2]]@genotype[which(F1Gir@generation[[2]]
    @sex==1),],F1Girolando@generation[[2]]@genotype[which(
    F1Girolando@generation[[2]]@sex==2),])

ind@phenotype<-c(F1Gir@generation[[2]]@phenotype[which(F1Gir@generation[[2]]
    @sex==1)],F1Girolando@generation[[2]]@phenotype[which(
    F1Girolando@generation[[2]]@sex==2)])


ind@err<-c(F1Gir@generation[[2]]@err[which(F1Gir@generation[[2]]@sex==1)],
    F1Girolando@generation[[2]]@err[which(F1Girolando@generation[[2]]@sex==2)])
```

```
ind@GEBV <- ind@phenotype
ind@TBV <- ind@phenotype - ind@err
ind@SVR <- rep(0,num_individuos)#predict(svm.model,ind@genotype)

Girolando@generation[[2]]<-ind
Girolando@h2[2]<-var(ind@TBV)/var(ind@GEBV)



beta<-(betaHOL/4) + (3*betaGIR/4)
beta0<-(beta0HOL/4) + (3*beta0GIR/4)
F2Girolando<-execSimulation(ind,
num_individuos,
1,
topMacho = 0,
topFemea = 0,
useArtificialInsemination = 0,
pedFile=T,
nameFile='F2Girolando')



#
# Segunda Geração do girolando Macho Gir F1 e Femea Meio Sangue Girolando.
#



ind<-SimSNPglmIND()

#Touro Holandes
ind@father<-which(F1Hol@generation[[2]]@sex==1)
#Vaca F2 Girolando
ind@mother<-which(F2Girolando@generation[[2]]@sex==2)

num_individuos<-length(ind@father) + length(ind@mother)
ind@id<-seq(Girolando@lastId,(Girolando@lastId + num_individuos-1))
```

```
Girolando@TotalGen = Girolando@TotalGen + 1
Girolando@totalInd = Girolando@totalInd + num_individuos
Girolando@lastId = Girolando@lastId + num_individuos

ind@sex   <- c(rep(1,length(which(F1Hol@generation[[2]]@sex==1))) ,
rep(2,length(which(F2Girolando@generation[[2]]@sex==2))) )
ind@genotype<-rbind(F1Hol@generation[[2]]@genotype[which(F1Hol@generation[[2]]
    @sex==1),],
F2Girolando@generation[[2]]@genotype[which(F2Girolando@generation[[2]]@sex==2)
    ,])
ind@phenotype<-c(F1Hol@generation[[2]]@phenotype[which(F1Hol@generation[[2]]
    @sex==1)],
F2Girolando@generation[[2]]@phenotype[which(F2Girolando@generation[[2]]@sex==2)
    ])

ind@err<-c(F1Hol@generation[[2]]@err[which(F1Hol@generation[[2]]@sex==1)],
F2Girolando@generation[[2]]@err[which(F2Girolando@generation[[2]]@sex==2)])

ind@GEBV <- ind@phenotype
ind@TBV <- ind@phenotype - ind@err
ind@SVR <- rep(0,num_individuos)#predict(svm.model,ind@genotype)

Girolando@generation[[2]]<-ind
Girolando@h2[2]<-var(ind@TBV)/var(ind@GEBV)

beta<-(5*betaHOL/8) + (3*betaGIR/8)
beta0<-(5*beta0HOL/8) + (3*beta0GIR/8)

# A generation [[2]] é a F3 e
# A generation [[3]] é a PS
F3_PSGirolando<-execSimulation(ind,
num_individuos,
2,
```

```
topMacho = 0,
topFemea = 0,
useArtificialInsemination = 0,
pedFile=T,
nameFile='F3_PSGirolando')


save.image('Cenario8.RData')
```