Documentos

ISSN 1677-9274
Dezembro, 2013
128

Banco de Dados de Genótipos para Suporte à Seleção Genômica em Programas de Melhoramento Animal

Embrapa 40 anos

ISSN 1677-9274
Dezembro, 2013

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Documentos 128

Banco de Dados de Genótipos para Suporte à Seleção Genômica em Programas de Melhoramento Animal

Roberto Hiroshi Higa
Vinícius Fernandes Dias
Jorge Luiz Corrêa
Gabriel Bueno de Oliveira

Embrapa Informática Agropecuária
Campinas, SP
2013

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700 - Fax: (19) 3211-5754
www.cnptia.embrapa.br
sac@cnptia.embrapa.br

1ª edição
on-line 2013

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Banco de Dados de Genótipos para suporte à seleção genômica em programas de melhoramento animal / Roberto Hiroshi Higa... [et al.]. - Campinas : Embrapa Informática Agropecuária, 2013.
30 p. : il. - (Documentos / Embrapa Informática Agropecuária , ISSN 1677-9274 ; 128).

1. Melhoramento genético. 2. Genótipos. 3. Banco de dados. I. Higa, Roberto Hiroshi. II. Embrapa Informática Agropecuária. III. Série.

572.8 CDD (21. ed.)

Autores

Roberto Hiroshi Higa
Engenheiro eletricista, doutor em engenharia elétrica
Pesquisador da Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
roberto.higa@embrapa.br

Vinícius Fernandes Dias
Graduando em Engenharia da Computação
Estagiário da Embrapa Informática Agropecuária
v.fernandesdias@gmail.com

Jorge Luiz Corrêa
Cientista da computação, mestre em ciência da computação
Analista da Embrapa Informática Agropecuária
jorge.l.correa@embrapa.br

Gabriel Bueno de Oliveira
Graduando em Engenharia da Computação
Bolsista PIBIC/CNPq na Embrapa Informática Agropecuária
xgabriel.bueno@gmail.com

Apresentação

A seleção genômica utiliza dezenas ou centenas de milhares de marcadores moleculares do tipo *Single Nucleotide Polymorphisms* (SNP) para estimar o perfil genômico dos animais em avaliação e predizer características fenotípicas de interesse econômico. Ela permite que tanto o intervalo entre gerações quanto a quantidade de animais fenotipados sejam reduzidos, com ganhos no custo de manutenção de programas de melhoramento. O momento atualmente enfrentado pelos programas de melhoramento genético animal, coordenados pela Embrapa, é o de incorporação dessas tecnologias em suas rotinas.

Contudo, isto implica na necessidade de armazenamento de grande volume de dados de genotipagem, que deverão se acumular ao longo dos próximos anos. Visando suplantar essa dificuldade, o projeto componente 1 da Rede Genômica Animal II, liderado pela Embrapa Informática Agropecuária, vem realizando ações no sentido de desenvolver um Banco de Dados de Genótipos (BDG) para acondicionamento de dados de genotipagem de animais utilizados em avaliações vinculadas a programas de melhoramento genético de animais.

Este documento representa o primeiro passo nessa direção, onde aproveitando-se de experiências anteriores, um novo modelo de dados é proposto tendo em vista o cenário em que centenas de milhares de amostras de animais genotipados em plataformas com dezenas ou centenas de milhares de marcadores terão que ser manipulados.

Kleber Xavier Sampaio de Souza

Chefe-geral

Embrapa Informática Agropecuária

Sumário

1 Introdução 9

2 Escopo 11

3 Modelo de dados 13

3.1 Descrição de genótipos 17

3.2 Descrição das tabelas 18

3.2.1 Tabela *individual* 18
3.2.2 Tabela *population* 19
3.2.3 Tabela *member* 19
3.2.4 Tabela *pedigree* 19
3.2.5 Tabela *species* 20
3.2.6 Tabela *genome* 20
3.2.7 Tabela *panel* 20
3.2.8 Tabela snp 21
3.2.9 Tabela *snpset* 21
3.2.10 Tabela *dataset* 21
3.2.11 Tabela *genotype* 22
3.2.12 Tabela *sampleset* 23
3.2.13 Tabela *researcher* 23
3.2.14 Tabela *institution* 23
3.2.15 Tabela *permission* 24

4 Entrada e saída de dados 24

5 Análise de desempenho 25

6 Discussão 28

7 Referências 29

Banco de Dados de Genótipos para Suporte à Seleção Genômica em Programas de Melhoramento Animal

Roberto Hiroshi Higa
Vinícius Fernandes Dias
Jorge Luiz Corrêa
Gabriel Bueno de Oliveira

1 Introdução

Atualmente, os programas de melhoramento genético animal coordenados pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) encontram-se em um momento de grandes mudanças, com a perspectiva de incorporação em suas rotinas de metodologias baseadas em seleção genômica ampla. Uma das consequências esperadas para os próximos anos é um acúmulo de dados de genotipagem, devido às atividades de avaliação genética. Em resposta a esse desafio, a rede de pesquisa vinculada ao Macroprograma 1 “Rede nacional para o desenvolvimento e adaptação de estratégias genômicas inovadoras aplicadas ao melhoramento, conservação e produção animal”, conhecido como Rede Genômica Animal II e seus projetos associados, vem desenvolvendo ações específicas em seus projetos componentes para incluir estratégias de seleção genômica em programas de melhoramento de bovinos e ovinos. Dentre esses esforços, o projeto componente 1, liderado pela Embrapa Informática Agropecuária, inclui ações no sentido de desenvolver um banco de dados de genótipos

para acondicionamento de dados de genotipagem de animais utilizados em avaliações genéticas para produção de sumários de touros.

Anteriormente, diferentes softwares com a funcionalidade de armazenamento de genótipos e fenótipos foram desenvolvidos pela Embrapa Informática Agropecuária (VIEIRA, 2011; 2012a; 2012b). Esses softwares possuem interface web para interação *online* com o usuário e, além de armazenar dados de genótipos e fenótipos, também contemplam algumas consultas básicas ao conjunto de dados (ex: identificação de *Single Nucleotide Polymorphisms* (SNPs) monomórficos, ou seja, sem variação entre os animais incluídos no conjunto de dados). Os softwares, entretanto, revelaram-se adequados para um cenário envolvendo poucos milhares de animais (2 a 4 mil) genotipados utilizando painéis com 50 mil ou 60 mil marcadores. Quando este cenário foi alterado para poucos animais genotipados na plataforma bovine HD (800 animais e 770 mil marcadores), sua utilização *online* mostrou-se insatisfatória, visto que uma consulta simples demora pelo menos 1 hora para ser processada. Os autores do software não realizaram estudos para avaliar o tempo de recuperação dos dados em função do tamanho do conjunto de dados, mas indicaram que um dos motivos desse comportamento era a normalização "excessiva" do modelo de dados utilizado. Por essa razão, o modelo implementado por esses softwares não é considerado neste trabalho.

O modelo de dados apresentado neste documento considera o cenário esperado para os programas de melhoramento genético animal em que será necessário acondicionar dados de centenas de milhares de animais genotipados em plataformas de 50 mil, 60 mil ou 770 mil marcadores, e que o escopo da aplicação a que está vinculado, diferente das aplicações acima mencionadas, é de utilização *offline*, tendo como funcionalidade básica apenas o acondicionamento de dados de genotipagem. Em adição ao modelo de dados, também é apresentado um estudo relacionado o tempo de recuperação dos dados em função do tamanho do conjunto de dados, visando analisar a escalabilidade do modelo de dados para o cenário pretendido.

O documento está organizado da seguinte forma: na seção 2 é apresentado o escopo do Banco de Dados de Genótipos (BDG), suas principais funções e os processos com que interage; na seção 3 é apresentado o

modelo lógico de dados e os principais conceitos envolvidos; a seção 4 aborda a questão de entrada e saída dos dados armazenados no BDG; a seção 5 apresenta um estudo do desempenho do BDG, quanto a inserção e consulta de grandes volumes de dados de genótipos. As discussões, conclusões e trabalhos futuros são apresentados na seção 6.

2 Escopo

O escopo do BDG compreende o armazenamento organizado e padronizado de conjuntos de dados resultantes de procedimentos de genotipagem em larga escala e alguns resultados de análises, consideradas básicas, que podem ser utilizados por diversas análises subsequentes. Por exemplo, a partir da análise de Controle de Qualidade (*Quality Control* - QC), pode-se decidir por excluir das análises subsequentes algumas das amostras e/ou alguns marcadores. Neste caso, o BDG deve permitir a utilização dos dados pós QC sem que seja necessário executar a análise de QC novamente. Da mesma forma, após uma análise de haplótipos em que as fases dos marcadores são inferidas, estes resultados devem ser armazenados de forma que esses resultados possam ser recuperados sem que a análise tenha que ser realizada novamente.

Genericamente, o objetivo do BDG é prover suporte para sistemas computacionais que implementem a (semi-)automação de processos de análises vinculados a programas de melhoramento animal que utilizem dados de genotipagem, no que se refere à segurança, privacidade e disponibilidade desses dados. A Figura 1 ilustra o BDG e os processos que o circundam. A interação com o BDG se dá de forma padronizada, implementada por dois processos: a) “Padronizar e armazenar”, que converte os dados de seu formato de chegada para um formato padronizado de armazenamento; este processo também deve garantir que sejam especificados os metadados que permitirão uma posterior recuperação destes conjuntos de dados; e b) “Recuperar e padronizar”, que retira conjuntos de dados do BDG no formato padrão para permitir sua conversão para o formato de utilização, que é dependente da análise; nesse processo, utilizam-se os metadados que foram associados durante o processo de inclusão dos

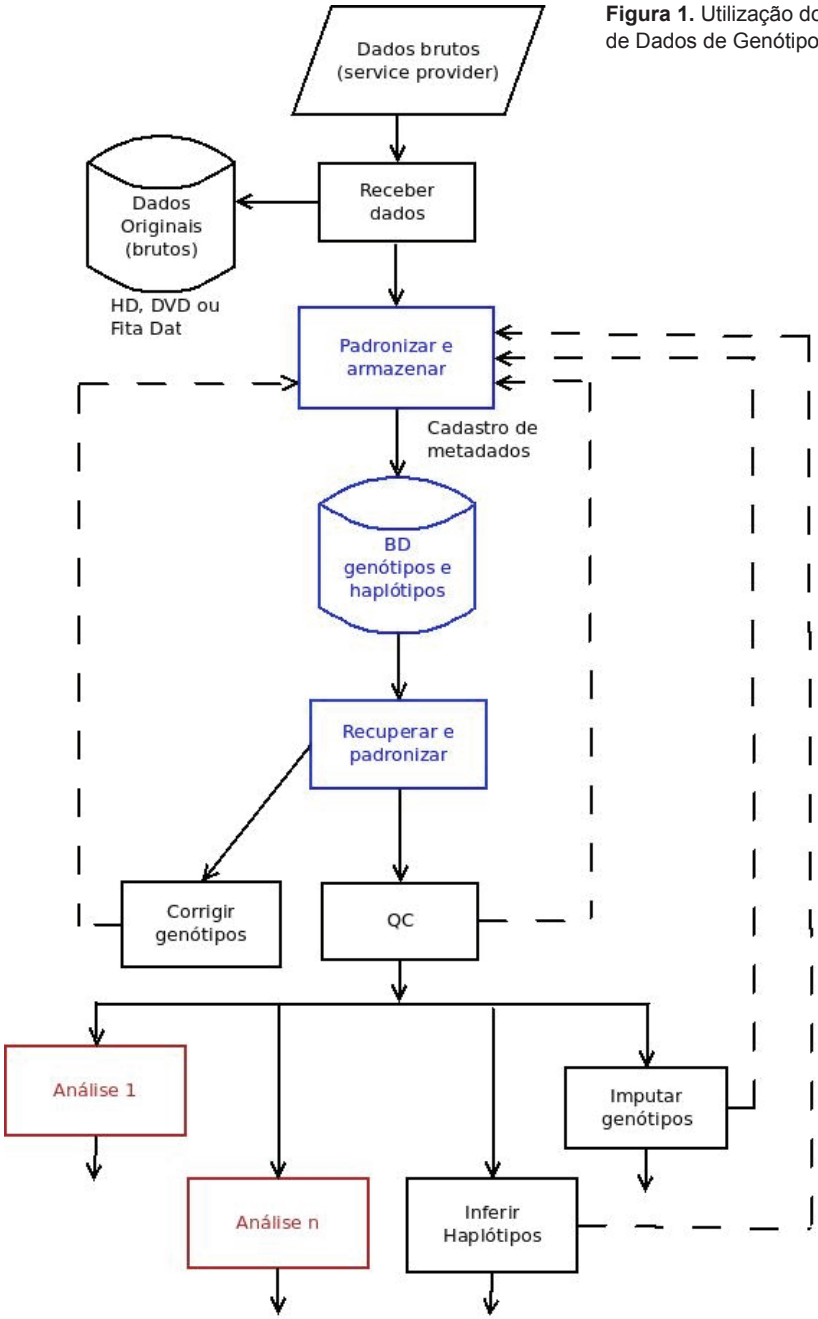

Figura 1. Utilização do Banco de Dados de Genótipos.

dados, o que pode resultar inclusive na recuperação de um conjunto de dados composto por diversos outros.

No escopo do BDG entende-se como conjunto de dados:

- dados de genotipagem de amostras de uma população, obtidos por procedimentos experimentais.
- dados de genotipagem derivados de processos computacionais, como imputação ou procedimentos de correção de genótipos e controle de qualidade.
- dados de genótipos com fase conhecida, derivados de processos computacionais para inferência de fases.

Embora o objetivo final do BDG seja permitir a utilização de dados de genótipos em conjunto com dados de medidas fenotípicas, o armazenamento de dados de fenotípicos está fora do escopo do BDG. Assume-se que estes estão armazenados em outros banco de dados, mas que o mapeamento dos indivíduos no BDG e no banco de dados de fenótipos é obtido por meio de um identificador comum.

3 Modelo de dados

No contexto em que está inserido (seleção genômica), o BDG funciona como um repositório de dados de genotipagem, baseado na utilização de painéis de SNPs e/ou obtidos por processamento computacional (ex: QC, imputação, inferência de fase). É esperado que conjuntos de dados de genotipagem contendo poucos milhares de animais sejam periodicamente inseridos e que as consultas sejam realizadas sobre os conjuntos de dados acumulados, podendo chegar a centenas de milhares de animais.

A granularidade requerida ao se manipular os conjuntos de dados armazenados no BDG é o conjunto de genótipos de um indivíduo e o respectivo painel de marcadores utilizado. Por esse motivo, e considerando o volume de dados a ser manipulado, utiliza-se o conceito de campo *Binary Large*

Object (BLOB) para armazenar a informação de genótipo de um indivíduo. Com este tipo de modelagem evita-se uma granularidade excessiva dos dados e uma consequente superpopulação de registros em algumas tabelas, o que dificulta a realização de consultas sobre esses dados. A desvantagem é a possibilidade de armazenamento redundante de dados, pois parte da responsabilidade por manter a consistência dos dados acessados passa a ser das aplicações. Além disso, o BDG também deve manter informações adicionais sobre os conjuntos de dados, visando qualificar as consultas, controlar o acesso e realizar o cruzamento com os dados de fenótipos.

O modelo dados do BDG considera que a aplicação a que está vinculado acessa o banco de dados para manipular (cadastrar e consultar) conjuntos de dados de genótipos baseados em painéis com 50 mil, 60 mil ou 770 mil SNPs (tabelas *dataset*, *sampleset*, *genotype*, *panel* e snp na Figura 2), ou seja, os dados são manipulados com uma granularidade em que não é necessário acessar o valor de cada SNP individualmente, como acontece com os sistemas bifequali, suínos e ovinos. Neste contexto, um conjunto de dados é composto por um conjunto de amostras (tabela *sampleset*) correspondente a genótipos de animais (tabela *genotype*), cujo valor é codificado na forma de um campo BLOB que representa um arquivo compactado. Associado a cada conjunto de dados está um painel contendo a descrição de um conjunto de SNPs, que pode ser oriundo de um chip comercial, de um painel customizado ou de subconjuntos determinados por processamentos como a análise de QC. Cada painel, por sua vez, está associado a uma montagem específica de um genoma de referência de um organismo (tabelas *genome* e *species*). No caso da não existência de um genoma de referência associado ao painel, o campo correspondente na tabela *panel* deve conter o valor NULL.

Os genótipos cadastrados no BDG referem-se a indivíduos pertencentes a uma população conhecida (programa de melhoramento) de uma espécie animal (tabelas *individual*, *population* e *species*). Esta modelagem de dados prevê a possibilidade de genotipagem de um mesmo animal por diferentes tecnologias ou como resultado de procedimentos computacionais como a imputação de genótipos. O resultado da estimação de fase também pode ser registrada como um “genótipo com fase conhecida”.

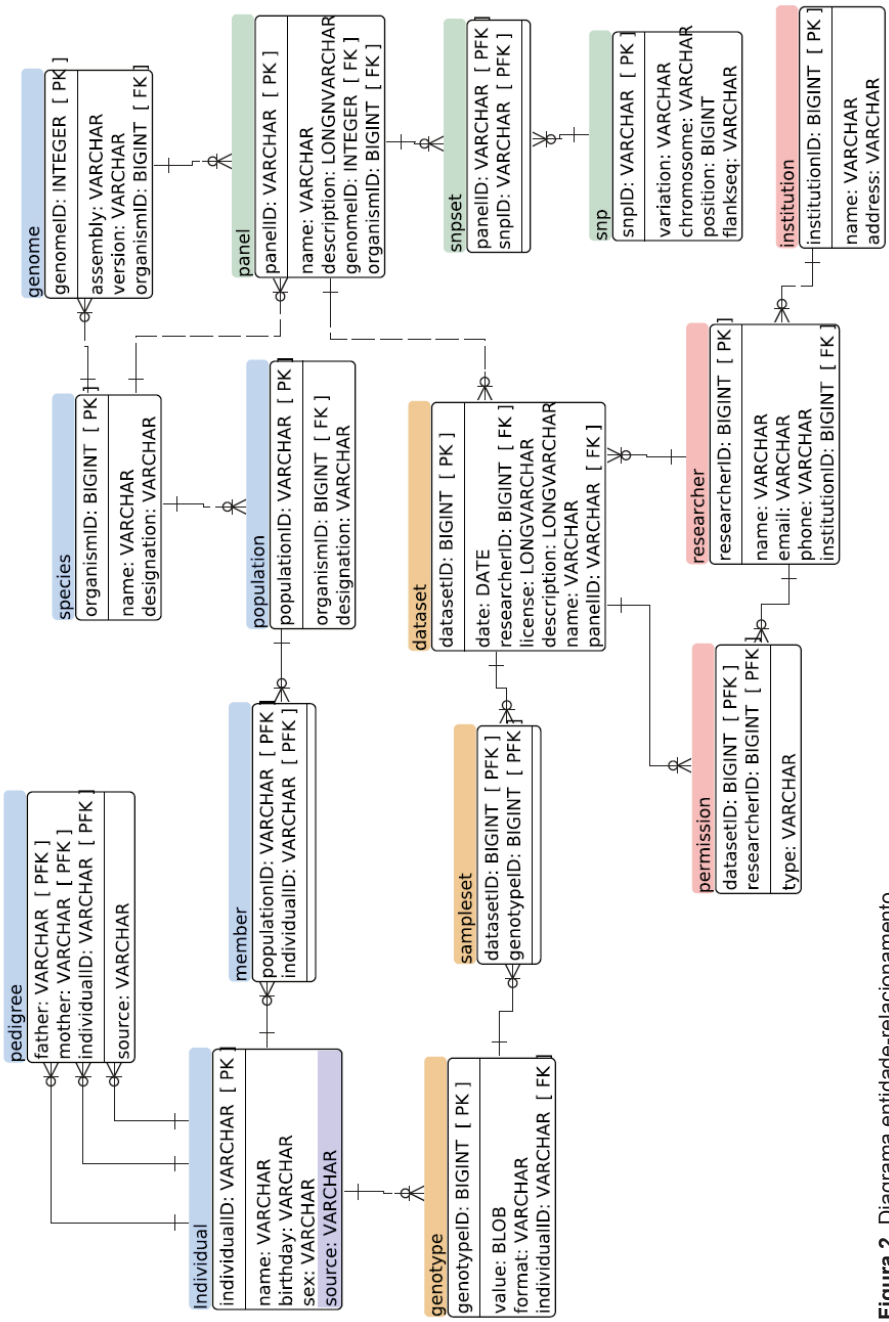

Figura 2. Diagrama entidade-relacionamento.

Conforme mencionado, cabe à aplicação que utiliza o BDG zelar pela consistência dos dados de genótipos, em termos dos marcadores utilizados, uma vez que esses são armazenados em um único campo BLOB. Existe o risco de incorrer em inconsistência quando novos conjuntos de dados são criados a partir de conjuntos de dados já existentes. Por isso, esse procedimento é padronizado da seguinte forma:

- quando são executadas operações de criação de genótipos para marcadores até então não genotipados (imputação) ou correção de erros, um novo registro de genótipo será criado exatamente como se o mesmo animal tivesse sido genotipado experimentalmente; o resultado da operação de estimação de haplótipos (ou determinação de fases) é tratada da mesma forma que um experimento de genotipagem.
- quando são executadas operações de redução do número de marcadores, como na análise de QC, é criado um novo painel (tabela *panel*) que mapeia (tabela *snpset*) apenas os marcadores remanescentes (tabela snp).
- quando operações de redução do número de amostras são realizadas a partir de um mesmo conjunto de dados (ex: análise de QC), o novo conjunto de dados (tabela *dataset*) mapeia (tabela *sampleset*) apenas um subconjunto dos genótipos do conjunto de dados original (tabela *genotype*).
- operações que produzem conjuntos de dados combinando outros conjuntos de dados devem se basear no painel de SNPs correspondentes à interseção dos conjuntos de SNPs associados com os conjuntos de dados originais.

Além disso, visando atender a requisitos de programas de melhoramento animal, o modelo lógico do BDG também inclui a tabela *pedigree*, que relaciona cada indivíduo à seu pai e à sua mãe, ambos considerados pertencentes à mesma população. Também assume-se que a chave primária de cada indivíduo (campos *individualID* e *populationID*) será utilizada para cruzamento com dados de fenótipos, assumidos como armazenados em outro repositório de dados. A tabela *pedigree* mantém os registro de paternidade e maternidade dos indivíduos cadastrados.

Por fim, para permitir que um sistema de informação implemente uma política de acesso aos conjuntos de dados, cada um deles está associado

a um *owner*, representado na tabela *researcher*, e permissões de acesso (leitura dos dados), de acordo com a tabela *permission*. Cada pesquisador, por sua vez, deve estar associado a uma instituição. Estas informações devem ser consultadas por sistemas de informação antes de prover acesso a conjuntos de dados armazenados no BDG.

3.1 Descrição de genótipos

O genótipo (campo BLOB) é descrito por um arquivo em formato CSV (valores dos campos separados por vírgula), onde cada arquivo armazenado na tabela *genotype* refere-se ao genótipo de um indivíduo. Como existe a possibilidade dos parâmetros BAF (*B allele frequency*) e LRR (*Log R ratio*), para inferência de CNVs, estarem ou não disponíveis, são adotados dois formatos de arquivo CSV sem a especificação dos nomes das colunas na primeira linha. Além disso, uma terceira possibilidade de formato é utilizada quando as fases são conhecidas. É importante observar que as designações utilizadas para descrever os alelos sejam compatíveis com a descrição do painel, armazenado na tabela *panel*.

- CSV1 (sem parâmetros de CNV): SNP ID, *genotype* (ex: AB, AA), *score* (*GenomeStudio GC-Score*). Por exemplo:

```
SNPID00001, AA, 0.9875
SNPID00002, AB, 0.9209
SNPID00003, AA, 0.9875
SNPID00004, BB, 0.9709
SNPID00005, AA, 0.9875
SNPID00006, BB, 0.9394
...............
```

- CSV2 (com parâmetros de CNV): SNP ID, *genotype* (ex: AB, AA), *score* (*GenomeStudio GC-Score*), BAF, LLR. Por exemplo:

```
SNPID00001, AA, 0.9875, 0.67458, 0.89076
SNPID00002, AB, 0.9209, 0.39756, 0.90776
SNPID00003, AA, 0.9875, 0.93747, 0.12987
```

```
SNPID00004, BB, 0.9709, 0.13947, 0.39489
SNPID00005, AA, 0.9875, 0.56989, 0.87568
SNPID00006, BB, 0.9394, 0.48957, 0.87569
..............
```

- FASES: arquivo CSV contendo com três campos, sendo SNP ID, alelo paterno, alelo materno. Por exemplo:

```
SNPID00001, A, A
SNPID00002, A, B
SNPID00003, A, A
SNPID00004, B, B
SNPID00005, A, A
SNPID00006, B, B
..............
```

Todos os genótipos em um mesmo conjunto de dados possuem o mesmo formato de descrição de genótipos. Valores faltantes são indicados por "-".

3.2 Descrição das tabelas

3.2.1 Tabela *individual*

Esta tabela representa um animal (indivíduo) pertencente a uma população vinculada a um programa de melhoramento genético. Para ser cadastrado nessa tabela, o animal não precisa ter sido genotipado. Seus campos incluem:

- *individualID*: identificador único do animal; esta informação deve ser a mesma utilizada para armazenar valores fenotípicos para este animal; portanto, é de extrema importância, pois permite o cruzamento de informações de genótipos e fenótipos.
- *name*: nome de registro do animal.
- *birthday*: data de nascimento do animal.

- *sex*: sexo do animal.
- *source*: rebanho, granja ou fazenda de origem do indivíduo.

3.2.2 Tabela *population*

Esta tabela representa populações de animais vinculadas a programas de melhoramento genético. Seus campos incluem:

- *populationID*: identificador único da população.
- *organismID*: identificação da espécie específica que está vinculada à população (ex: *bos indicus*).
- *designation*: designação utilizada para referenciar a população (ex: raça Y, programa X).

3.2.3 Tabela *member*

Esta tabela representa o relacionamento entre indivíduos e população, caracterizando indivíduos como pertencentes a uma ou mais populações. Seus campos incluem:

- *populationID*: identificador único da população;
- *individualID*: identificador único do animal.

3.2.4 Tabela *pedigree*

Esta tabela representa a relação de parentesco entre animais de uma população. Seus campos incluem:

- *father*: identificador único de indivíduo, referenciado na tabela *individual* e que representa o pai;
- *mother*: identificador único de indivíduo, referenciado na tabela *individual* e que representa a mãe.
- *individualID*: identificador único de indivíduo, referenciado na tabela *individual* e que representa o filho.

- *source*: indicação textual da fonte de onde a relação foi obtida (ex: associação de criadores da raça X).

3.2.5 Tabela *species*

Esta tabela registra as espécies de organismos incluídas no BDG. Seus campos incluem:

- *organismID*: identificador único para organismo (utilizar *TaxonID* fornecido pelo NCBI).
- *name*: designação popular (ex: bovino).
- *designation*: designação científica (ex: *bos taurus*).

3.2.6 Tabela *genome*

Esta tabela especifica o genoma de referência no qual o painel utilizado para genotipagem se baseia. Ele está associado a um organismo específico (tabela *species*) e seus campos incluem:

- *genomeID*: identificador único do genoma.
- *organismID*: identificador do organismo ao qual o genoma pertence.
- *assembly*: designação da montagem (ex: Btau ou UMD).
- *version*: versão da montagem (ex: 4.1 ou 3.1).

3.2.7 Tabela *panel*

Esta tabela representa o painel de marcadores utilizado para genotipagem de indivíduos que compõem os conjuntos de dados. Seus atributos incluem:

- *panelID*: identificador único para o painel.
- *name*: nome que identifica o painel (ex: swine60k, bovine50k, bovineHD, etc.).
- *description*: descrição textual do painel, contendo informações sobre fabricante (ou se é customizado), versão, número de SNPs, etc.

3.2.8 Tabela snp

Esta tabela representa SNPs que compõem os painéis utilizados na genotipagem. Quando existe um genoma de referência associado, ele contém informações para seu mapeamento; em qualquer caso, também é possível associar a região flanqueadora do SNP. Seus campos incluem:

- *snpID*: identificador único para o SNP.
- *variation*: indica variação (ex: A/T ou A/B).
- *chromosome*: indica o cromossomo no genoma de referência, por exemplo 1, 2, 23, etc.
- *position*: indica a posição do SNP no cromossomo, por exemplo 4323456.
- *flankseq*: indica a região flanqueadora do SNP, incluindo a variação, por exemplo ATCGGTTAAC[A/T]GGACTCATA.

Note que o os campos chromosome e position tem valor NULL quando não há um genoma de referência, enquanto o campo flankseq pode ou não ter valor NULL quando há um genoma de referência.

3.2.9 Tabela *snpset*

Esta tabela representa o relacionamento entre SNPs e painéis, caracterizando SNPs como pertencentes a um ou mais painéis. Seus campos incluem:

- *panelID*: identificador único para o painel.
- *snpID*: identificador único para o SNP.

3.2.10 Tabela *dataset*

Esta tabela representa os conjuntos de dados de genótipos armazenados no BDG. Estes são compostos pelo resultado da genotipagem de um conjunto de animais utilizando um painel de marcadores SNPs, previamente especificados, e fazem parte de um projeto/experimento conduzido por

um pesquisador responsável que também desempenha função de *owner* dos dados. Além disso, esta tabela também faz referência a campos que indicam o pesquisador *owner* do conjunto de dados e o painel utilizado na genotipagem dos animais cadastrados no conjunto de dados. Seus campos incluem:

- *datasetID*: identificador único para o conjunto de dados.
- *researcherID*: identificação do pesquisador *owner* dos dados.
- *panelID*: identificação do painel utilizado para construção do conjunto de dados (genotipagem).
- *date*: data em que o conjunto de dados foi inserido no banco de dados.
- *license*: referência ao documento que disciplina o uso dos dados.
- *description*: descrição do conjunto de dados, como procedência, projeto, etc.
- *name*: nome de designação do conjunto de dados.

3.2.11 Tabela *genotype*

Esta tabela representa o genótipo de um animal (indivíduo). O genótipo está vinculado a um indivíduo e inclui os seguintes campos:

- *genotypeID*: identificador único para o genótipo.
- *value*: é um valor BLOB correspondente ao arquivo que descreve o genótipo do animal e contém os valores de genótipos para cada marcador SNP do painel utilizado.
- *format*: indica o formato do arquivo de genótipo armazenado no campo *value* (vide seção 3.1).

Note que, apesar do genótipo de um animal ser composto por um conjunto de informações (um valor para cada marcador que integra o painel de genotipagem), ele é modelado como uma entidade única, armazenada em um campo do tipo BLOB. Considera-se que esse conjunto de marcadores deve ser armazenado e processado em conjunto.

3.2.12 Tabela *sampleset*

Esta tabela representa o relacionamento entre genótipos e conjuntos de dados, caracterizando genótipos como pertencentes a um ou mais conjuntos de dados. Seus campos incluem:

- *genotypeID*: identificador único para o genótipo.
- *datasetID*: identificador único para o dataset.

3.2.13 Tabela *researcher*

Esta tabela representa o usuário que acessa o BDG (*owner* de datasets ou não) via sistema de informação. Seus campos incluem:

- *researcherID*: Identificador único do usuário.
- *name*: nome do usuário.
- *e-mail*: endereço eletrônico para contato.
- *phone*: número para contato telefônico.

Note que todo usuário deve estar vinculado a uma instituição (tabela *institution*).

3.2.14 Tabela *institution*

Esta tabela representa o cadastro da instituição a que todo pesquisador deve estar vinculado e seus campos incluem:

- *institutionID*: identificador único para cada instituição.
- *name*: nome pelo qual a instituição é referenciada (Ex: Embrapa).
- *address* seu endereço institucional (e-mail, telefone, etc.).

3.2.15 Tabela *permission*

Esta tabela representa as permissões concedidas a pesquisadores para acesso a conjuntos de dados armazenados no BDG, caso eles não sejam os *owners*. Seus campos incluem:

- *datasetID*: identificador único para o dataset.
- *researcherID*: identificador único do usuário.
- *type*: pode ser R (apenas leitura) ou RW (leitura e escrita); Note que a leitura se refere ao acesso ao conjunto de dados já cadastrado no BDG e W à inserção de um novo conjunto de dados derivado de conjuntos de outros dados já cadastrados no BDG. Neste caso, o pesquisador torna-se automaticamente o *owner* do novo conjunto de dados. Conjuntos de dados brutos (resultado de experimento de genotipagem) são inseridos pelo administrador do BDG.

4 Entrada e saída de dados

Conjuntos de dados de genótipos, painéis de marcadores e *pedigrees* são inseridos e consultados/extraídos no BDG como arquivos de formato conhecido. No caso dos genótipos, os formatos dos arquivos de entrada são os especificados na seção 3.1. Estes, contudo, não são os formatos de arquivos de saída esperados de programas de determinação de genótipos. Assim, espera-se que um sistema de informação para acesso ao BDG implemente conversores de arquivos, a partir dos formatos mais comumente encontrados. Por exemplo:

- *Long* (PLINK, 2013): Linhas com colunas separadas por espaços, cada linha tendo como colunas os rótulos SNP ID, *Sample ID*, *A allele*, *B allele*, *GC-Score*, *BAF*, *LLR*.
- *PED* (PLINK, 2013): Linhas com colunas separadas por espaços em branco, cada linha tendo como colunas os rótulos *family ID*, *individual ID*, *father ID*, *mother ID*, *Sex* (1=macho, 2=fêmea, outro=desconhecido), fenótipo, genótipos, neste caso codificados como 1, 2, 3 e 0 (*missing*).

- No caso de fase conhecida, um formato comum utiliza uma linha para cada indivíduo, a primeira coluna com o identificador do indivíduo e as demais contendo os pares de alelos paterno e materno separados por espaço.

No caso de arquivos de descrição de painéis de marcadores, são utilizados os seguintes formatos:

- Arquivo SNP_Map da *Illumina* (*GenomeStudio*).
- Arquivo CSV com 3 colunas, *SNP ID*, região flanqueadora, *score*.

Para a descrição de *pedigree*, o formato mais comumente utilizado consiste em um arquivo CSV, separado por vírgula, onde cada linha contém o identificador do indivíduo, identificador do pai e identificador da mãe.

Da mesma forma, ao serem extraídos do banco de dados, os conjuntos de dados devem ser formatados para utilização por programas específicos. No caso de dados de genótipos, espera-se que os sistemas de informação para acesso ao BDG implementem conversores de arquivos, de acordo com o uso dos dados, ou seja, formatos requeridos por programas de análise específicos, o que inclui os formatos *Long*, PED, PHASE (STEPHENS et. al., 2001), etc. No caso de painéis de marcadores e *pedigrees*, os formatos são os mesmos tratados no processo de entrada de dados.

5 Análise de desempenho

O modelo de dados aqui proposto para armazenamento de dados de genotipagem de animais tem como objetivo o armazenamento seguro dessas informações, visando sua utilização em análises de estimação de valores genéticos genômicos em programas de melhoramento genético animal. Esta tarefa é realizada com periodicidade anual e de forma *offline*, não exigindo tempos de resposta compatíveis com interações *online* com o usuário. Contudo, dado que o volume de dados a serem processados

é enorme, ou seja, da ordem de dezenas até centenas de milhares de animais genotipados em uma plataforma de 50k ou 780k SNPs, é preciso assegurar que a modelagem utilizando campos do tipo *BLOB* permita uma velocidade de acesso aos dados compatível com a aplicação pretendida, pois a primeira operação a ser realizada é a consulta e retirada dos dados de genotipagem inseridos de forma cumulativa no BDG.

O objetivo dessa seção é estudar o comportamento de um sistema que utilize o campo do tipo BLOB na modelagem de dados do BDG, em termos de tempo de latência ao se inserir e extrair conjuntos de dados contendo dezenas de milhares de animais genotipados em uma plataforma com 50k ou 780k SNPs. O primeiro experimento foi realizado em um *desktop* funcionando como servidor com a seguinte configuração básica: processador AMD – 4 núcleos, 64 bits, 800 MHz, 512k de *cache* e 6 Gb de RAM, e o SGBD *postgresql* 9.2 (POSTGRESQL, 2013), instalado na mesma máquina, com sua configuração básica. Para simular os dados de genótipos, foram criados arquivos contendo 780k linhas de dados, gerados aleatoriamente, de acordo com o formato CSV1. Os resultados obtidos (Figura 3) indicam que tanto o processo de inserção de amostras quanto o de seleção e extração variam de forma quase linear com relação ao número

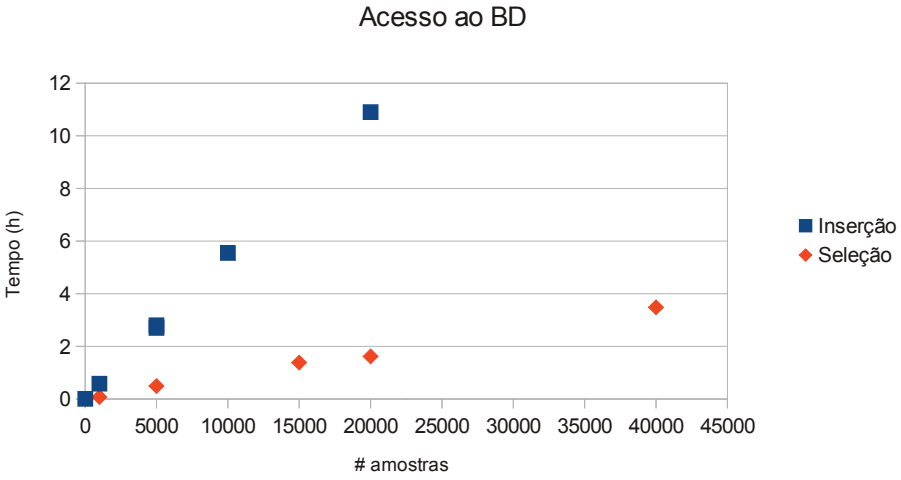

Figura 3. Análise do tempo de latência para inserção e recuperação de genótipos no BDG, utilizando um desktop como servidor.

de amostras. No caso do processo de inserção, supõe-se um tempo de espera de 5h33min para inserir 10.000 amostras e 10h54min para inserir 20.000 amostras, o que resulta em pouco mais de 30min para cada mil amostras inseridas. O caso de uso mais crítico, contudo, é a seleção de amostras, uma vez que se prevê o acúmulo de genótipos e sua utilização conjunta em procedimentos de predição do valor genético de animais. Para este caso, é possível recuperar 40.000 amostras em 3h29 min, ou seja, aproximadamente 52 min para recuperar 10.000 amostras.

O segundo experimento foi realizado em um servidor com a seguinte configuração: 2 processadores Intel Xeon – 8 núcleos, 64 bits, 2,6 GHz e 128 Gb de RAM, 1,1 TB de espaço disponível para o banco de dados, com configuração RAID 10 em hardware. Neste caso, também utilizou-se o compactador Pigz (PIGZ, 2013), que consiste em uma implementação paralela do compactador padrão gzip, para reduzir o tempo utilizado na compactação do arquivo de genótipos armazenado no campo BLOB. O SGBD postgresql 9.2 foi instalado na mesma máquina, com configuração otimizada para inserção de dados. Para simular os dados de genótipos, foram criados arquivos contendo 50k linhas de dados, gerados aleatoriamente, de acordo com o formato CSV1. O resultados obtidos (Figura 4) indicam

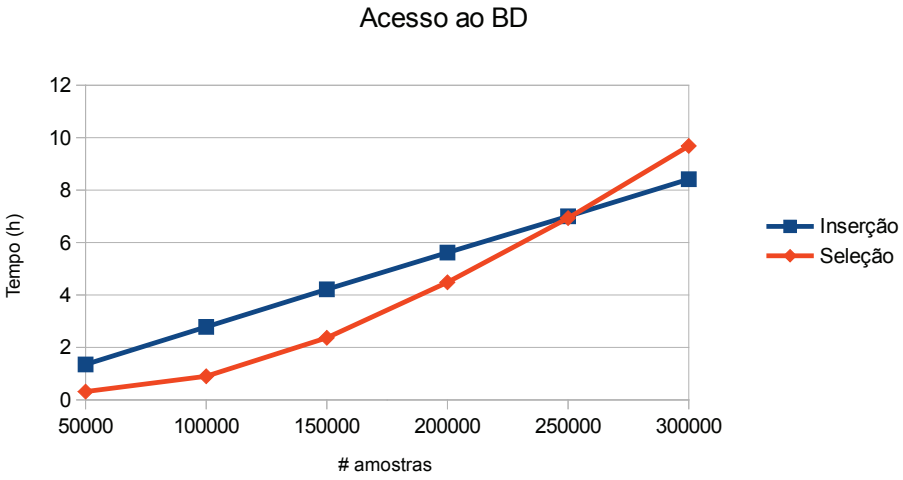

Figura 4. Análise do tempo de latência para inserção e recuperação de genótipos no BDG, utilizando um servidor.

que a tarefa de inserção varia de forma linear com relação ao número de amostras, enquanto a tarefa de recuperação apresenta uma variação de forma polinomial. O tempo esperado para inserir 300.000 amostras é da ordem de 8h25min enquanto o tempo para recuperar esse mesmo número de amostras é de 9h41min.

Ambos os experimentos reportados acima foram realizados utilizando a linguagem Python (PYTHON, 2013) para acessar o banco de dados. Foi percebido que ela consumia muitos recursos do servidor e, por isso, também foi avaliada implementação alternativa em linguagem C (KERNIGHAN; RITCHIE, 1988), que é compilada. Neste caso, foi avaliado apenas um caso extremo: recuperação de 100.000 amostras genotipadas com 700.000 SNPs. Esse processo utilizando a linguagem Python apresentou tempo de latência de 12h28min, enquanto que utilizando a implementação em linguagem C esse tempo foi de 7h54min, demonstrando que é possível um ganho de desempenho utilizando uma aplicação compilada para acesso ao banco de dados.

Finalmente, considerando a possibilidade de utilização de equipamentos *storage*, ao invés de se utilizar os próprios discos do servidor para armazenamento dos dados, mediu-se a velocidade de acesso ao realizar o segundo experimento, descrito acima. Foi observado que a taxa máxima de escrita ficou em torno de 40 MB/s, que é inferior ao que é possível atingir utilizando um equipamento *storage* acessado via protocolo *Internet Small Computer System Interface* (iSCSI)(WIKIPÉDIA, 2013). Desta forma, é provável que a utilização de um equipamento *storage* não tenha impacto significativo no desempenho da aplicação.

6 Discussão

Neste documento foi apresentado um modelo de dados para acondicionamento de grande volume de dados de genotipagem, visando sua utilização na estimação de parâmetros genéticos em programas de melhoramento animal coordenados pela Embrapa.

O modelo proposto tem como principal característica a de ser genérico para diferentes espécies de animais e programas de melhoramento. Além disso, visando um cenário futuro de acúmulo de grande volume de dados de genotipagem que precisará ser manipulado, o modelo proposto utiliza o conceito de campo BLOB para obter uma implementação mais enxuta e eficiente.

Foram realizados testes para avaliar o tempo de latência ao acessar o banco de dados de genótipos – BDG. Os resultados mostram que em um cenário com animais genotipados com chip de 50k é possível acessar 300.000 amostras em pouco mais de 9h, mesmo realizando o acesso por meio da linguagem Python, que é menos eficiente. Considerando um cenário com animais genotipados com chip HD (700k SNPs), utilizando uma implementação compilada (linguagem C), é possível acessar em torno de 100.000 amostras em aproximadamente 8h. Esses tempos de latência são bastante razoáveis quando se considera que a aplicação pretendida (estimação de parâmetros genéticos) é realizada anualmente e que espera-se que a maior parte do tempo de processamento seja utilizada na execução dos modelos de estimação dos parâmetros genéticos, o que pode consumir um ou mais semanas.

Conclui-se, portanto, que o modelo de dados proposto é adequado para implementação de um banco de dados de genótipos, utilizado por processos de estimação de parâmetros genéticos em programas de melhoramento animal coordenados pela Embrapa.

7 Referências

KERNIGHAN, B. W.; RITCHIE, D. M. **The C programming language**. 2nd ed. Englewood Cliffs: Prentice-Hall, 1988. 272 p.

PIGZ. **A parallel implementation of the gzip for modern multi-processor, multi-core machines**. 2013. Disponível em: <http://zlib.net/pigz/>. Acesso em: 27 dez. 2013.

PLINK. **Whole genome association analysis toolset**. 2013. Disponível em: <http://pngu.mgh.harvard.edu/~purcell/plink/>. Acesso em: 26 dez. 2013.

POSTGRESQL. 2013. Disponível em: <http://www.postgresql.org/.> Acesso em: 18 out. 2013.

PYTHON. **Programming Language – official Website**. 2013. Disponível em: <http://www.python.org/.> Acesso em: 27 dez. 2013.

STEPHENS, M.; SMITH, N.; DONNELY, P. A new statistical method for haplotype reconstruction from population data. **American Journal of Human Genetics**, v. 68, p. 978-989, 2001.

VIEIRA, F. D. **Sistema Consulta Dados de Ovinos**. Versão 1.0. Campinas: Embrapa Informática Agropecuária, 2010. 1 CD-ROM.

VIEIRA, F. D. **Sistema Bife de Qualidade**. Versão 1.6. Campinas: Embrapa Informática Agropecuária, 2012a. 1 CD-ROM.

VIEIRA, F. D. **Sistema Suínos**. Versão 1.1. Campinas: Embrapa Informática Agropecuária, 2012b. 1 CD-ROM.

WANG, K.; BUCAN, M. Copy number variation detection via high-density SNP genotyping. **Cold Spring Harbor Protocols**, Woodbury, v. 6, June, 2008. Doi: 10.1101/pdb.top46.

WIKIPÉDIA. **iSCSI**. 2013. Disponível em: http://pt.wikipedia.org/wiki/ISCSI. Acesso em: 27 dez. 2013.

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

CGPE 11124