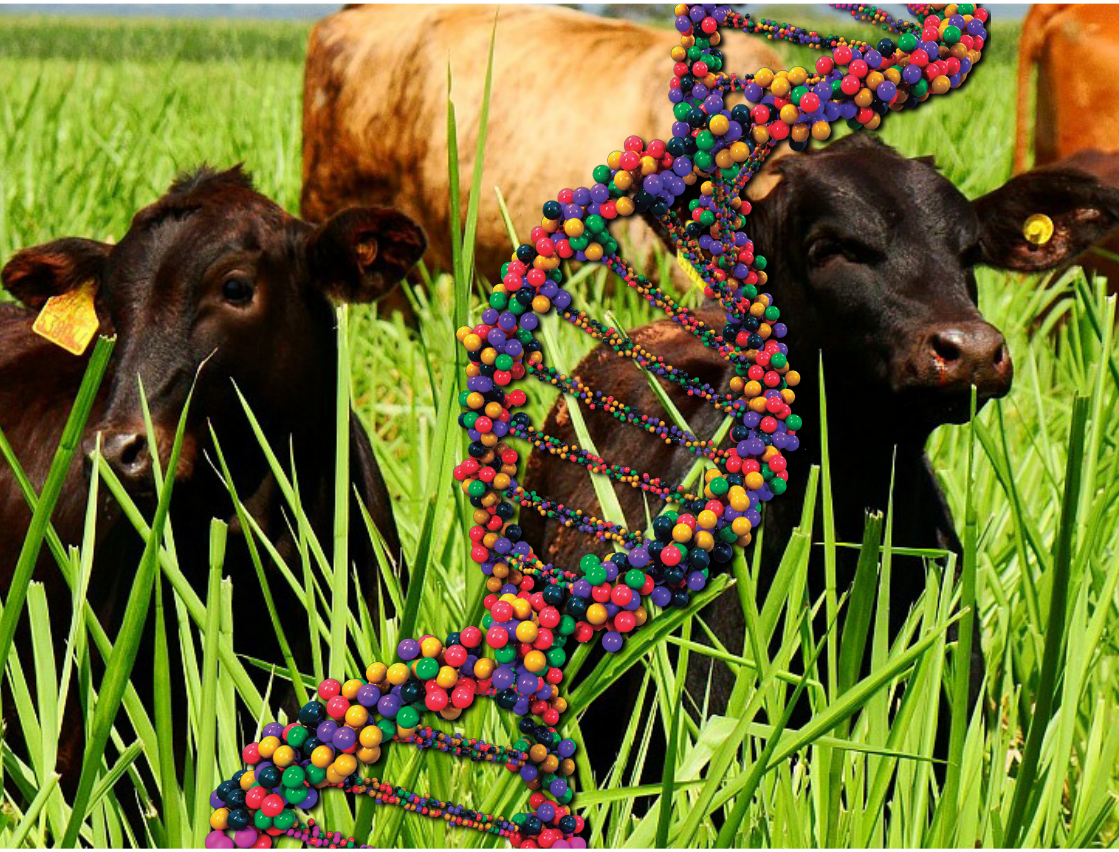
Documentos

ISSN 1677-9274
Agosto, 2015
133

Banco de Dados de Genótipos e Fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal



Embrapa

ISSN 1677-9274
Agosto, 2015

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

Documentos 133

Banco de Dados de Genótipos e Fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal

Roberto Hiroshi Higa
Gabriel Bueno de Oliveira

Embrapa Informática Agropecuária
Campinas, SP
2015

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700
www.embrapa.br/informatica-agropecuaria
SAC: www.embrapa.br/fale-conosco/sac/

1ª edição
publicação digital 2015

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

Higa, Roberto Hiroshi.
Banco de dados de genótipos e fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal / Roberto Hiroshi Higa, Gabriel Bueno de Oliveira. - Campinas : Embrapa Informática Agropecuária, 2015.

30 p. : il. - (Documentos / Embrapa Informática Agropecuária, ISSN; 1677-9274; 133).

1. Genótipo. 2. Fenótipo. 3. Associação genômica. I. Oliveira, Gabriel Bueno. II.Título. III. Embrapa informática Agropecuária. IV. Série.

CDD (21. ed.) 576.53

Autores

Roberto Hiroshi Higa
Engenheiro eletricista, doutor em Engenharia Elétrica
Pesquisador da Embrapa Informática Agropecuária, Campinas, SP

Gabriel Bueno de Oliveira
Estudante de Engenharia da Computação
Bolsista CNPq/PIBIC, Campinas, SP

Apresentação

A tecnologia de genotipagem em larga escala de dezenas ou centenas de milhares de marcadores moleculares do tipo *Single Nucleotide Polymorphisms* (SNP) para estimar o perfil genômico de animais permitiu o desenvolvimento tanto de estudos de associação genótipo-fenótipo em escala genômica (do inglês *genome-wide association studies* - GWAS) quanto a introdução da tecnologia de seleção genômica em programas de melhoramento genético.

Contudo, isto implica na necessidade de armazenamento de grande volume de dados de genotipagem, fenotipagem e pedigree de um grande número de animais, uma tendência que possivelmente aumentará ao longo dos próximos anos, dado a diminuição dos custos para produção dos dados experimentais. Visando suplantar essa dificuldade atual, o projeto componente 1 da Rede Genômica Animal II, liderado pela Embrapa Informática Agropecuária, desenvolveu um modelo de banco de dados para acondicionamento e manipulação de grandes volumes de dados de genotipagem de animais, o Bando de Dados de Genótipos (BDG) (HIGA et al., 2013; DIAS & HIGA, 2013; BUENO et al., 2014).

Este documento complementa os trabalhos previamente desenvolvidos, apresentando um modelo de dados para armazenamento de dados de genótipos, fenótipos e pedigree de animais de interesse agropecuário para suporte tanto a experimentos de GWAS quanto a programas de melhoramento genético animal, Banco de Dados de Genótipos e Fenótipos(BDGF). Neste cenário, é suposto que os conjuntos de dados a serem manipulados são compostos por centenas de milhares de amostras de animais genotipados em plataformas com dezenas ou centenas de milhares de marcadores.

Silvia Maria Fonseca Silveira Massruhá

Chefe-geral
Embrapa Informática Agropecuária

Sumário

Banco de Dados de Genótipos e Fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal

Roberto Hiroshi Higa
Gabriel Bueno de Oliveira

1 Introdução

A tecnologia de genotipagem em larga escala de dezenas ou centenas de milhares de marcadores moleculares do tipo SNP para estimar o perfil genômico de animais permitiu o desenvolvimento tanto de estudos de associação genômica ampla (do inglês Genome-Wide Association Studies - GWAS) quanto a introdução da tecnologia de seleção genômica em programas de melhoramento genético.

A Empresa Brasileira de Pesquisa Agropecuária (Embrapa), em conjunto com seus parceiros, já realizou alguns estudos de associação em escala genômica com animais das raças Canchim (MOKRY et al., 2013) e Nelore (TIZIOTO et al., 2013). Também já foi publicada, por dois anos seguidos, a avaliação genômica para resistência a carrapato de touros Herefore e Braford (CARDOSO et al., 2013). Além disso, outros estudos de GWAS e ações no sentido de incorporação de seleção genômica em programas de melhoramento genético animal, coordenados pela Embrapa, encontram-se atualmente em desenvolvimento, como por exemplo as ações desenvolvidas nos projetos PC1 (SILVA et al., 2013) e PC 4 (ROSA et al., 2013) da

Rede do Macroprograma 1 “Rede Genômica Animal(RGA)” (CAETANO et al., 2013).

Uma necessidade básica e comum a todas essas ações é a utilização de uma estrutura para armazenamento dos conjuntos de dados, incluindo genótipos, fenótipos e pedigree. Dado o volume de dados considerado (centenas de milhares de animais fenotipados e, possivelmente, genotipados para centenas de milhares de marcadores do tipo SNPs), uma questão importante a se considerar no desenvolvimento de uma solução computacional é a adequabilidade da modelagem à aplicação desejada, pois esta terá impacto direto nos tempos das consultas nos bancos de dados onde essa informação estará armazenada. Por exemplo, foram desenvolvidas pela Embrapa Informática Agropecuária (VIEIRA, 2010; 2012a; 2012b) diferentes softwares com a funcionalidade de armazenamento de genótipos e fenótipos. 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: SNPs monomórficos, ou seja, sem variação entre os animais incluídos no conjunto de dados). Apesar de os softwares terem apresentado comportamento adequado para um cenário envolvendo poucos milhares de animais (2 a 4 mil) genotipados utilizando painéis com 50 mil ou 60 mil marcadores alterando este cenário 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 um dos motivos apontados para este comportamento foi uma modelagem de dados com normalização “excessiva”.

Visando suplantar essa limitação, desenvolveu-se, no escopo do projeto componente 1 da RGA (HIGA et al., 2013a), liderado pela Embrapa Informática Agropecuária, um novo modelo de dados para acomodar esse tipo de dados, utilizando uma modelagem mais simples (BUENO et al., 2014; DIAS; HIGA, 2013; HIGA et al., 2013b) e o conceito de campo Binary Large OBject (BLOB) em bancos de dados relacionais. O trabalho aqui apresentado complementa esses trabalhos, apresentando um modelo de dados para armazenamento de dados de genótipos, fenótipos e pedigree de animais de interesse agropecuário para suporte tanto a experimentos

de GWAS quanto a programas de melhoramento genético animal. Neste cenário, é suposto que os conjuntos de dados a serem manipulados são compostos por centenas de milhares de amostras de animais genotipados em plataformas com dezenas ou centenas de milhares de marcadores e, por essa razão, também é apresentado um estudo relacionado ao tempo de recuperação dos dados em função do tamanho do conjunto de dados, visando analisar a escalabilidade do modelo de dados para o cenário considerado.

O documento está organizado da seguinte forma: na seção 2 são apresentadas as principais diferenças entre o BDG (BUENO et al., 2014; DIAS; HIGA, 2013; HIGA et al., 2013b) e o BDGF, apresentado neste documento; na seção 3 é apresentado o modelo de dados do BDGF; já a seção 4 apresenta um estudo do desempenho do BDGF na inserção e recuperação de grandes volumes de dados de genotipagem. Finalmente, a seção 6 encerra o documento apresentando as discussões e principais conclusões.

2 Principais diferenças entre BDG e BDGF

As principais diferenças entre o Banco de Dados de Genótipos (BDG) (HIGA et al., 2013b) e o Banco de Dados de Genótipos e Fenótipos (BDGF) compreendem:

- A extensibilidade do conjunto de atributos associados aos animais a partir de um conjunto mínimo considerado comum aos diferentes estudos de GWAS ou programas de melhoramento genético animal. Esta característica visa permitir que o mesmo modelo de dados possa ser utilizado por diferentes estudos de GWAS e programas de melhoramento genético, independente da espécie animal considerada.

- A inclusão de dados de fenótipos no banco de dados, o que o torna diretamente utilizável para armazenamento de dados tanto em aplicações de GWAS quanto na implantação de seleção genômica em programas de melhoramento animal, o que justifica sua renomeação com o acréscimo de um “F” no final do nome original. A modelagem utilizada não

restringe quantos e quais fenótipos podem ser associados a um animal, e importa as descrições de fenótipos cadastrados na base de dados do sistema SIEXP (QUEIROS et al., 2012), definindo um nível básico de integração dos dados. Além disso, assim como já acontece com os dados de genótipos, o conceito de conjunto de dados também é preservado para os fenótipos, o que permite não só rastrear a origem do dado no banco de dados, mas também o aproveitamento de dados originalmente coletados para estudos de GWAS em avaliações genéticas de programas de melhoramento genético animal.

- Separação de dados e metadados em bases de dados diferentes. Da mesma forma que o modelo de dados do BDG, o modelo de dados do BDGF foca na descrição dos dados (genótipos e pedigree no caso do BDG e mais fenótipos no caso do BDGF) e seus relacionamentos. Entretanto, ao invés de utilizar o conceito de campo BLOB, no caso do BDGF os dados de genótipos e fenótipos, são armazenados diretamente em tabelas de uma base de dados diferente (Figura 1).

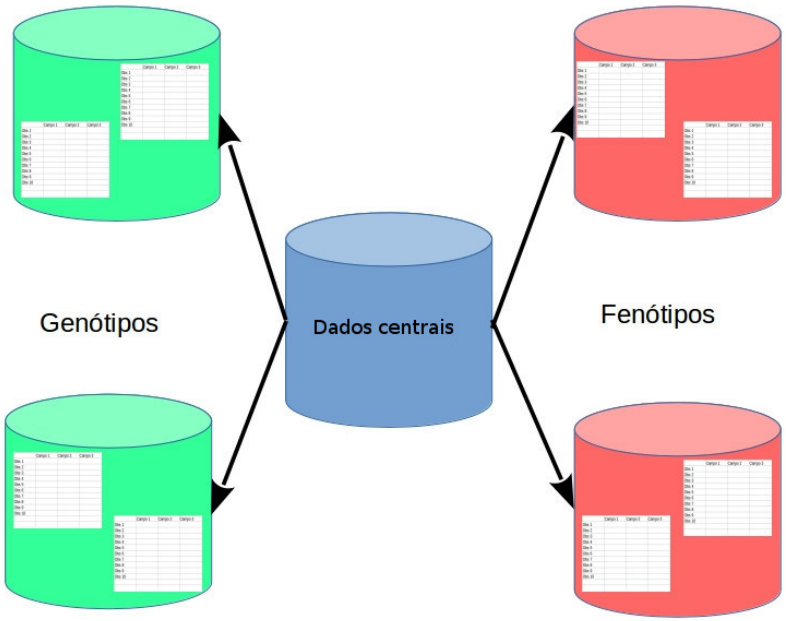


Figura 1. Organização dos dados no BDGF.

3 Modelo de dados

De forma similar ao BDG, o BDGF funciona como um repositório de dados de genótipos, fenótipos e pedigree, sendo esperado a inserção periódica de dados de poucos milhares de animais e que as consultas sejam realizadas sobre os conjuntos de dados acumulados, podendo chegar a centenas de milhares de animais. A granularidade requerida para se manipular dados de genótipos 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 tabelas de um banco de dados à parte para armazenamento dos dados genotipagem de cada animal. Com este tipo de estratégia 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. De forma similar, os conjuntos de dados de fenótipos também são armazenados como tabelas em uma base de dados à parte.

Consultas ao BDGF inicialmente identificam os conjuntos de dados (genótipos e/ou fenótipos) no banco de dados central, acessando-os em seguida diretamente nos bancos de dados que armazenam os dados, conforme ilustrado pela Figura 1. O modelo de dados da base de central é formado por cinco tipos de informações, conforme cores das tabelas na Figura 2:

- Informações sobre os animais, seu relacionamento de parentesco e com uma população (tabelas *individual*, *population*, *member*, *attribute* e *indiv_atrib*).

- Informações sobre os marcadores moleculares utilizados para genotipagem (tabelas *species*, *genome*, *map*, *snp* e *snpset*).

- Informações sobre as pessoas responsáveis pelos dados (tabelas *researcher* e *institution*).

- Informações sobre dados de fenotipagem, sua localização e pessoas com permissão de acesso (tabelas *fenotype*, *dts_pheno*, *dts_pheno_cols*, *pheno_cols* e *permission_pheno*).

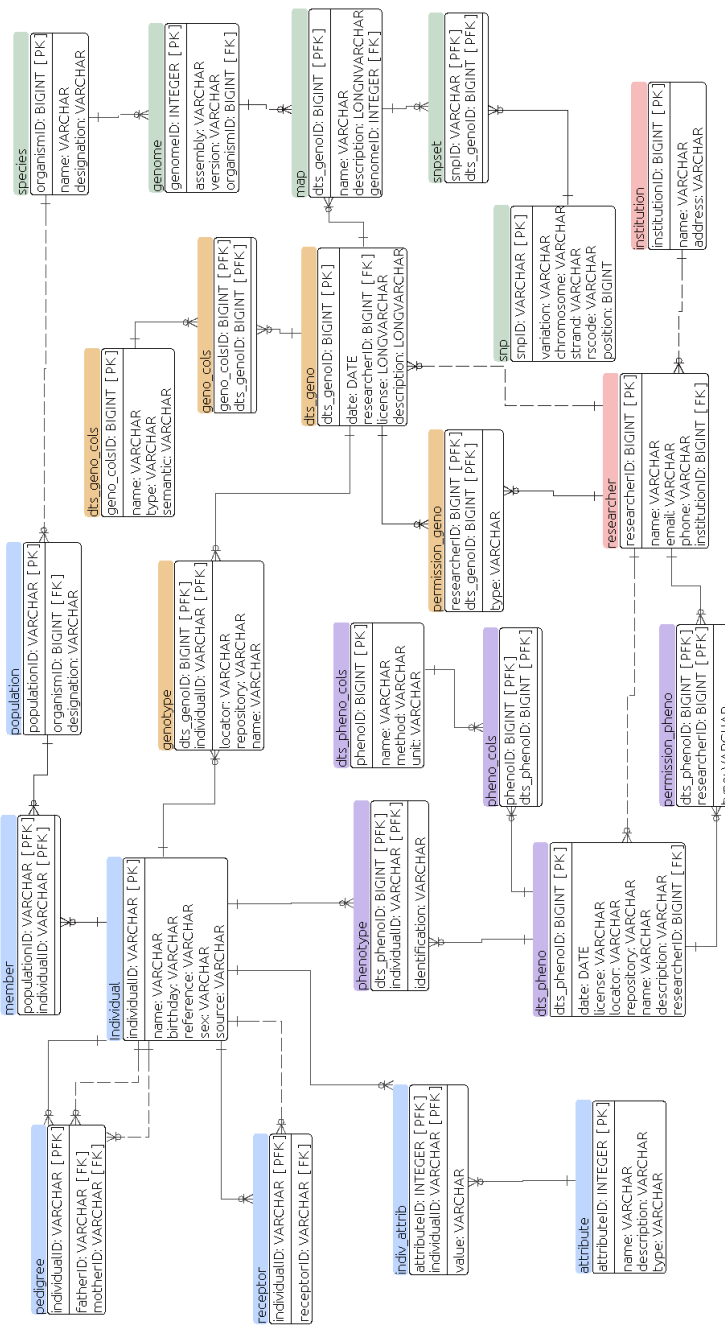


Figura 2. Modelo de dados do BDGF.

- Informações sobre dados de genotipagem, sua localização e pessoas com permissão de acesso (tabelas *genotype*, *dts_geno*, *dts_geno_cols*, *geno_cols* e *permission_geno*).

3.1 Descrição das tabelas do banco de dados central

3.1.1 Tabela *individual*

Esta tabela representa um animal (indivíduo) pertencente a uma população (ex: vinculado 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 na base de dados.
- *birthday*: data de nascimento do animal.
- *sex*: sexo do animal.
- *source*: rebanho, granja ou fazenda de origem do animal.
- *receptorID*: identificador da fêmea que gestou o animal, quanto utilizado processo de transferência de embriões.
- *fatherID*: identificador do pai do animal.
- *motherID*: identificador da mãe do animal.

Observe que os dados de pedigree são definidos pelos campos *individualID*, *fatherID* e *motherID* da tabela *individual*.

3.1.2 Tabela *pedigree*

Esta tabela representa a relação de filiação dos animais, ou sjea, quem é o seu pai e sua mãe. Seus campos incluem:

- *individualID*: identificador único do animal.
- *fatherID*: identificador do pai do animal.
- *motherID*: identificador da mãe do animal.

3.1.3 Tabela *receptor*

Esta tabela representa a relação entre um animal e a doadora de material genético materno, no caso utilização de tecnologia de transferência de embriões. Seus campos incluem:

- *individualID*: identificador único do animal.
- *receptorID*: identificador da fêmea que gestou o animal, quando utilizado processo de transferência de embriões.

3.1.4 Tabela *attribute*

Esta tabela contém a definição dos atributos específicos, não definidos na tabela *individual*. Esses atributos são utilizados quando, para alguma aplicação (ex: programa de melhoramento genético), há interesse em se armazenar um conjunto de atributos diferente daquele definido pela tabela *individual*. Seus campos incluem:

- *name*: nome do atributo.
- *description*: descrição textual da semântica do atributo e sua manipulação.
- *type*: indicação do tipo de atributo (int, float, character, etc.).

3.1.5 Tabela *indiv_attrib*

Esta tabela contém os valores dos atributos específicos para os indivíduos que as possuem. Seus campos incluem:

- *individualID*: identificador do animal.
- *atributeID*: identificador do atributo.
- *value*: valor do atributo.

3.1.6 Tabela *population*

Esta tabela representa um conjunto de animais que definem uma população (ex: nelore, canchim, braford, etc.), podendo estar vinculada a um programa de melhoramento genético específico. Seus campos incluem:

- *populationID*: identificador único da população.
- *designation*: designação utilizada para referenciar a população (ex: raça Y, programa X).
- *organismID*: identificada o organismo a que a população se refere (bovinos, suínos, etc.), de acordo com a conteúdo da tabela *species*.

3.1.7 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.

3.1.8 Tabela *species*

Esta tabela registra os organismos incluídas no BDGF. Seus campos incluem:

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

3.1.9 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.
- *assembly*: designação da montagem (ex: Btau ou UMD).
- *version*: versão da montagem (ex: 4.1 ou 3.1).
- *organismID*: identificador do organismo ao qual o genoma pertence.

3.1.10 Tabela *map*

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

- *dts_genoID*: identificador único para o map associado a um conjunto de dados de genotipagem.
- *name*: nome que identifica o mapa.

- *description*: descrição textual do mapa, contendo informações sobre fabricante (ou se é customizado), versão, número de SNPs, etc.

- *genomeID*: identificador do genoma referente ao mapa.

3.1.11 Tabela *snp*

Esta tabela representa SNPs que compõem os painéis utilizados na genotipagem. 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.

- *strand*: indica o sentido da leitura do SNP.

- *rscode*: indica o rscode (se houver).

- *position*: indica a posição do SNP no cromossomo, por exemplo 4323456.

3.1.12 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.

3.1.13 Tabela *researcher*

Esta tabela representa o usuário com acesso ao BDGF via sistema de informação e é utilizado para modelar permissões de acesso. 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*).

3.1.14 Tabela *institution*

Esta tabela representa o cadastro da instituição a que todo usuário deve estar vinculado. 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.).

3.1.15 Tabela *dts_geno*

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 conjunto 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. Seus campos incluem:

- *dts_genoID*: designação única do conjunto de dados.
- *date*: data em que o conjunto de dados foi inserido no banco de dados.
- *researcherID*: identificação do pesquisador proprietário (*owner*) dos dados.
- *license*: referência ao documento que disciplina o uso dos dados.
- *description*: descrição do conjunto de dados, como procedência, projeto, etc.

3.1.16 Tabela *dts_geno_cols*

Esta tabela indica quais colunas estão presentes em uma determinada tabela contendo informações de genótipos. Seus campos incluem:

- *geno_colsID*: identificador único para a coluna.
- *name*: nome da coluna na tabela.
- *type*: tipo de dado da coluna.
- *semantic*: tipo de terminologia utilizada. (Ex.:alleles AB, alleles forward (AC, AG, AT, CG, CT, GT), haplótipos, etc.)

3.1.17 Tabela *geno_cols*

Esta tabela relaciona as tabelas *dts_geno_cols* e *dts_geno*, ou seja, para um determinado conjunto de dados de genótipos, ela descreve as tabelas que contém esses dados no banco de dados de genótipos. Seus campos incluem:

- *dts_genoID*: designação única do conjunto de dados.
- *geno_colsID*: identificador único para a coluna.

3.1.18 Tabela *genotype*

Esta tabela contém a indicação da localização do genótipo de um animal. O genótipo está vinculado a um conjunto de dados. Seus campos incluem:

- *dts_genoID*: designação única do conjunto de dados.
- *individualID*: identificador único para o indivíduo.
- *locator*: local onde se encontra o banco de dados.
- *repository*: nome do banco de dados.
- *name*: nome da tabela onde estão os genótipos.

3.1.19 Tabela *permission_geno*

Esta tabela representa as permissões de acesso ao conjunto de dados de genótipos. Seus campos incluem:

- *dts_genoID*: designação única do conjunto de dados.
- *researcherID*: identificador de pesquisador.
- *type*: tipo de permissão concedida.

3.1.20 Tabela *dts_pheno*

Esta tabela representa os conjuntos de dados de fenótipos armazenados no BDGF, compostos por dados observados (fenótipos) em um conjunto de animais, previamente especificados, e que fazem parte de um projeto/experimento conduzido por um pesquisador responsável, que também desempenha função de *owner* dos dados. Seus campos incluem:

- *dts_phenoID*: identificador único para o conjunto de dados.
- *name*: nome do conjunto de fenótipos.

- *date*: data em que o conjunto de dados foi inserido no banco de dados.
- *license*: referência ao documento que disciplina o uso dos dados.
- *locator*: banco de dados onde se encontram as informações de fenótipos.
- *repository*: tabela contendo dados de fenótipos.
- *description*: descrição do conjunto de dados, como procedência, projeto, etc.
- *researcherID*: identificação do pesquisador proprietário (*owner*) dos dados.

3.1.21 Tabela *dts_pheno_cols*

Esta tabela representa as colunas presentes no arquivo que contém as informações de fenótipos.

- *phenoID*: identificador único da coluna.
- *name*: nome da coluna na tabela.
- *method*: descrição do método de medida.
- *unit*: unidade padrão para o fenótipo.

3.1.22 Tabela *pheno_cols*

Esta tabela relaciona o conjunto de dados presentes na tabela *dts_pheno_cols* e *dts_pheno*, ou seja, indica os conjuntos de medidas de fenótipos associados a cada conjunto de dados de fenótipos. Seus campos contêm:

- *phenoID*: identificador único da coluna.
- *dts_phenoID*: identificador único para o conjunto de dados.

3.1.23 Tabela *phenotype*

Esta tabela relaciona os indivíduos com os conjuntos de dados que contêm informações de fenótipos relacionadas a ele. Seus campos incluem:

- *dts_phenoID*: designação única do conjunto de dados.
- *individualID*: identificador do pesquisador.
- *identification*: identifica a observação correspondente ao indivíduo no conjunto de dados de fenótipos.

3.1.24 Tabela *permission_pheno*

Esta tabela representa as permissões de acesso ao conjunto de dados de fenótipos. Seus campos incluem:

- *dts_phenoID*: designação única do conjunto de dados.
- *researcherID*: identificador de pesquisador.
- *type*: tipo de permissão concedida.

3.2 Banco de dados de genótipos

O banco de dados de genótipos é composto por tabelas que representam o genótipo de animais, sendo que cada linha representa um SNP do chip utilizado na genotipagem e cada coluna representa um tipo de informação obtido no processo de genotipagem (ex: genótipo segundo a denominação A/B, genótipo segundo os nucleotídeos variantes, *GC-Score – Genotype Calling Score*, *BAF – B Allele Frequency*, *LLR – Log-Likelihood Ratio*, etc.).

3.3 Banco de dados de fenótipos

O banco de dados de fenótipos é composto por tabelas que representam os conjuntos de fenótipos medidos em conjuntos de animais. Os fenótipos avaliados em cada conjunto de dados encontram-se especificado na tabela *pheno_cols*. Além dessas colunas, a tabela correspondente a um conjunto de dados de fenótipos que também possui um campo específico para designação de cada observação. Este campo está em correspondência com o campo *identification* da tabela *phenotype* e faz o mapeamento entre o animal na tabela *individual* e sua correspondente observação no conjunto de dados.

4 Manipulação dos dados de genotipagem

O BDG (BUENO et al., 2014; DIAS; HIGA, 2013; HIGA et al., 2013b) tinha como foco um modelo para armazenamento de dados de genotipagem de animais num contexto de sua utilização na avaliação dos valores genéticos genômicos em programas de melhoramento genético animal. Dado o volume de dados necessários (milhares de animais com dados de genotipagem, obtidos experimentalmente ou por imputação, para centenas de milhares de marcadores do tipo SNP), foi analisado o desempenho do banco de dados ao inserir e recuperar conjuntos de dados dessa magnitude. Um comportamento linear para a operação de inserção e quasi-linear para a recuperação foram observados (em torno de 5h33min para inserir 10.000 amostras de animais genotipados com 780.000 SNPs e 52 minutos para recuperar 10.000 amostras).

Uma vez que o BDGF altera a forma como os dados de genótipos são armazenados e considerando que este tipo de dado representa a maior parte do volume dos dados manipulados pelo BDGF, foram realizados novos experimentos para avaliar o tempo de inserção e recuperação de dados de genotipagem, agora considerando a estrutura de acondicionamento desses dados proposta pelo BDGF.

Utilizando o mesmo tipo de plataforma de genotipagem (*chip* de 780.000 SNPs) e a mesma configuração de equipamento *desktop* (processador Intel Core i5-3470 – 4 núcleos, 64 bits, 3,20 GHz, 6 MB de *cache* e 16 Gb de RAM), para instalação do banco de dados (SGBD *postgresql* 9.2 (POSTGRESQL, 2014)), observou-se o seguinte comportamento:

- No caso do processo de inserção (Figura 3), obteve-se um tempo de espera de 2h38min para inserir 10.000 amostras, o que resulta em aproximadamente 16 minutos para cada mil amostras inseridas.

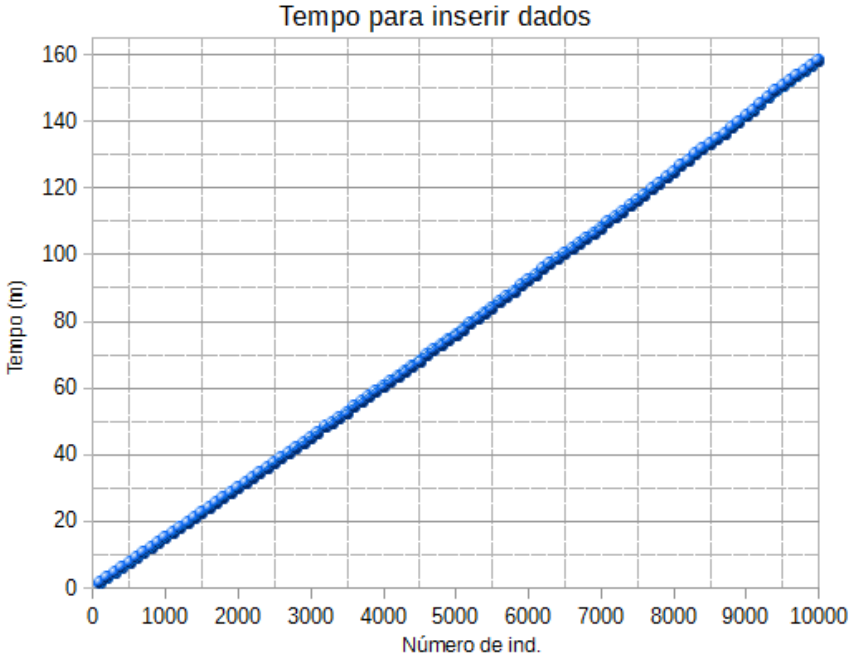


Figura 3. Tempo para inserção de dados em função do número de genótipos.

- No caso do processo de recuperação (Figura 4), o tempo para se consultar dados de genotipagem para 100.000 animais foi avaliado recuperando-se cada um dos 10.000 genótipos 10 vezes. Assim, para um total de 100.000 genótipos, obteve-se um tempo de 14h41min, equivalente a 1h28min para recuperar 10.000 amostras.

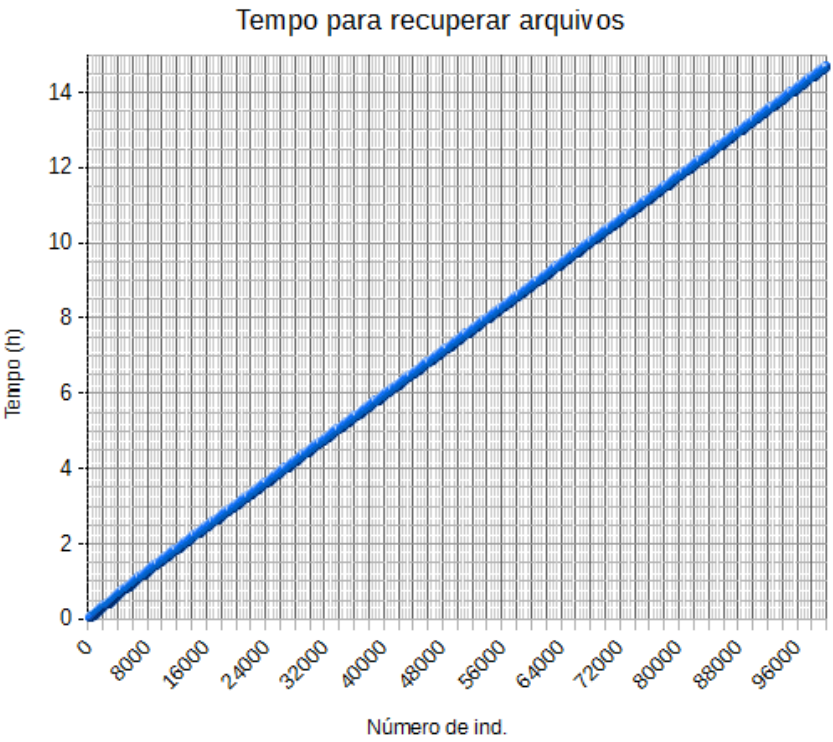


Figura 4. Tempo para recuperação de dados em função do número de genótipos.

5 Discussão

Neste documento foi apresentado um modelo de dados para acondicionamento de dados de genotipagem, fenotipagem e pedigree no contexto de estudos de associação genômica ampla e estimação de valores genéticos em programas de melhoramento animal - BDGF, considerando tanto os modelos tradicionais, que se baseiam em dados de pedigree, quanto os modelos genômicos, que se baseiam na utilização de dados de genotipagem.

O modelo proposto estende o modelo para acondicionamento de dados de genotipagem proposto anteriormente (BDG - BUENO et al., 2014; DIAS;

HIGA, 2013 HIGA et al., 2013b) por integrar a parte de dados de fenótipos, de forma integrada com outras propostas institucionais em andamento (QUEIROS et al., 2012), é aplicável ao mesmo tempo aos contextos de estudos de associação genômica ampla e inclusão de seleção genômica em programas de melhoramento genético animal.

Em termos desempenho ao se manipular grandes volumes de dados de genótipos, o BGDF apresentou uma redução no tempo de inserção de 52,55%, enquanto que para o caso de consulta de 100.000 animais genotipados com 700.000 SNPs houve um aumento de 17,78% no tempo de latência. Considerando que o tempo de consulta no caso do BDG é não linear (BUENO et al., 2014; DIAS; HIGA, 2013; HIGA et al., 2013b) e linear no caso do BGDF (Figura 4), espera-se um desempenho superior do BGDF em relação ao BGF aumentando-se o número de animais consultados na base de dados.

Conclui-se, portanto, que o modelo de dados proposto, BDGF, é adequado para implementação de um banco de dados para acondicionamento de dados de genótipo, fenótipo e pedigree tanto no contexto de estudos de associação genômica ampla quanto de estimação de parâmetros genéticos em programas de melhoramento animal.

6 Referências

CAETANO, A. R. et al. **Rede nacional para o desenvolvimento e adaptação de estratégias genômicas inovadoras aplicadas ao melhoramento, conservação e produção animal**. Brasília, DF: Embrapa Recursos Genéticos e Biotecnologia, 2013. (Embrapa Macroprograma 1 – Genômica animal - SEG: 01.11.07.002.00.00). Projeto em execução.

CARDOSO, F. F.; YOKOO, M. J. I.; GULIAS-GOMES, C. C.; SOLLERO, B. P.; COSTA, R. F. da; ROSO, V. M; BRITO, F. V.; CAETANO, A. R.; AGUILAR, I. **Avaliação genômica para resistência ao carrapto de touros Hereford e Braford**. Bagé: Embrapa Pecuária Sul, 2013. 40 p. (Embrapa Pecuária Sul. Documentos, 133).

DIAS, V. F.; HIGA, R. H. Banco de dados de genótipos da Rede Genômica Animal. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 9., 2013, Campinas. **Resumos**… Brasília, DF: Embrapa, 2013. p. 55-57.

HIGA, R. H. et al. **PC1 - Desenvolvimento e aplicação de ferramentas de bioinformática em suporte a projetos de melhoramento e sistemas de produção animal**. Campinas: Embrapa Informática Agropecuária, 2013a. (Embrapa. Macroprograma 1 – Genômica animal. SEG: 01.11.07.002.06.00). Projeto em execução.

HIGA, R. H.; DIAS, V. F.; CORRÊA, J. L.; OLIVEIRA, G. B. **Banco de dados de genótipos para suporte à seleção genômica em programas de melhoramento animal**. Campinas: Embrapa Informática Agropecuária, 2013b. (Embrapa Informática Agropecuária. Documentos, 128).

MOKRY, F. B.; HIGA, R. H.; MUDADU, M. M; LIMA, A. O.; MIRELLES, S. L.C.; SILVA, M. V. G. B.; CARDOSO, F. F.; OLIVEIRA, M. M.; URBINATI, I.; NICIURA, S. C. M.; TULLIO, R. R.; ALENCAR, M. M.; REGITANO, L. C. A. Genome-wide association study for backfat thickness in Canchim beef cattle using Random Forest approach. **BMC Genetics**, v. 14, n. 47, 2013. DOI: 10.1186/1471-2156-14-47.

OLIVEIRA, G. B.; DIAS, V. F.; PODESTÁ, E. V.; CORRÊA, J. L.; HIGA, R. H. Banco de dados de genótipos para melhoramento genético animal. In: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 8., 2014, Campinas. **Anais**... Campinas: IAC, 2014. p. 1-2.

POSTGRESQL. 2013. Disponível em: <http://www.postgresql.org/.> Acesso em: 30 dez. 2014.

QUEIROS, L. R. et al. **Gestão dos dados experimentais da Embrapa**. Campinas: Embrapa Informática Agropecuária, 2012. (Embrapa. Macroprograma 5 – Gestão da informação e do conhecimento. SEG: 05.11.11.007.00.00). Projeto em execução.

ROSA, A. J. de M. et al. **PC4 – Identificação e uso de genes de interesse em sistemas de produção**. Planaltia, DF: Embrapa Cerrados, 2013. (Embrapa. Macroprograma 1 – Genômica animal). Código SEG: 01.11.07.002l.04.00. Projeto em execução.

SILVA, M. V. G. B. et al. **PC3 – Estratégias de seleção genômica nos programas de melhoramento animal**. Juiz de Fora: Embrapa Gado de Leite, 2013. ((Embrapa. Macroprograma 1 – Genômica animal. SEG: 01.11.07.002.01.00). Projeto em execução.

TIZIOTO, P. C.; DECKER, J. E.; TAYLOR, J. F. N.; SCHNABEL, R. D.; MUDADU, M. A.; SILVA, F. L.; MOURÃO, G. B.; COUTINHO, L. L.; THOLON, P.; SONSTERGARD,T. S.; ROSA, A. I. F.; ALENCAR, M. F. L.; TULLIO, R. R.; MEDEIROS, S. R.; NASSU, R.T.; FEIJÓ, G. L. D.; SILVA, L. O. C.; TORRES, R.; SIQUEIRA, F.; HIGA, R. H.; REGITANO, L. C .A. Genome scan for meat quality traits in Nelore beef cattle. **Physiological genomics**. v. 25, n. 21, p. 1012-1020, 2013. DOI: 10.1152/physiolgenomics.00066.2013.

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.

Embrapa
Informática Agropecuária

Embrapa
Ministério da
Agricultura, Pecuária
e Abastecimento
GOVERNO FEDERAL
BRASIL
PÁTRIA EDUCADORA

CGPE 12150