BDGF: um sistema web para recuperação de informação de genótipos e fenótipos

Fábio Danilo Viiera[1], Danilo Gomes de Moura[1], Diego Félix da Silva[2], Roberto Hiroshi Higa[1], Adhemar Zerlotini[1]

[1] Embrapa Informática Agropecuária, Campinas, São Paulo, Brasil

[2] Programa Geneplus-Embrapa, Campo Grande, Mato Grosso do Sul, Brasil

fabio.vieira@embrapa.br, danilo.moura@colaborador.embrapa.br, dfelixsilva@hotmail.com, roberto.higa@embrapa.br, adhemar.zerlotini@embrapa.br

RESUMO

Nos últimos anos, o uso de genotipagem em grande escala de dezenas ou centenas de milhares de polimorfismos de nucleotídeo único (SNP) para estimar o perfil genômico de animais permitiu o desenvolvimento de estudos de associação genótipo-fenótipo em escala genômica (GWAS) e a introdução da tecnologia de seleção genômica em programas de melhoramento genético. No entanto, esta situação implica na necessidade de armazenamento de grande volume de dados de genotipagem, fenotipagem e pedigree de um elevado número de animais. Para integrar esse volume de dados distintos, é primordial utilizar uma estrutura de armazenamento robusta, como um SGBD. Assim, uma questão importante a considerar é o equilíbrio entre normalização e desempenho durante o estágio de modelagem de banco de dados, pois isso terá um impacto direto na usabilidade e na experiência do usuário. Buscando resolver esse problema de armazenamento eficiente e consultas rápidas num grande volume de dados, este trabalho apresenta o sistema BDGF (Banco de Dados de Genótipos e Fenótipos). Seu modelo de dados permite a implementação do tipo JSON em campos de tabelas relacionadas a fenótipos e do tipo texto em campos da tabela genótipos. BDGF foi projetado para dar suporte a projetos de criação de animais da Embrapa, mas pode ser facilmente ajustado para armazenar dados de diversas fontes, tais como dados de plantas. Além disso, o sistema implementa políticas de acesso e segurança para fenótipos, genótipos e pedigree dos animais.

PALAVRAS-CHAVE: Tecnologias Java EE, Sistema Gerenciador de Banco de Dados, PostgreSQL, Polimorfismo de Nucleotídeo Unico, JSON.

ABSTRACT

In recent years, the use of large scale genotyping of tens or hundreds of thousands of Single Nucleotide Polymorphisms (SNP) to estimate the genomic profile allowed the development of both genotype-phenotype association studies in genomic scale (genome-wide association studies-GWAS) and the introduction of genomic selection technology in breeding programs. However, this situation implies the need of storing large volumes of genotyping, phenotyping and pedigree data from large numbers of animals. In order to effectively integrate such amount of distinctive datasets, it's advisable to use a robust storage structure, such as a DBMS. Therefore, a major issue to consider is the trade off between normalization and performance during the database modeling stage, as this will have a direct impact on the usability and user experience. In order to get efficient storage and fast queries in this high volume of data, in this work we present the BDGF system (Genotypes and Phenotypes Database). Its data model allows the implementation of JSON type in table fields related to phenotypes and character varying in table fields related to genotypes. BDGF is designed to support the animal breeding projects of Embrapa, but can be easily adjusted to store data from diverse sources, such as plant data. Furthermore, the system implements access and security policies to phenotypes, genotypes and pedigree of the animals.

KEYWORDS: Java EE Technologies, Database Management System, PostgreSQL, Single Nucleotide Polymorphism, JSON.

INTRODUÇÃO

Nos últimos anos, a utilização da tecnologia de genotipagem em larga escala 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. As tecnologias atuais para geração de dados moleculares são capazes de genotipar de dezenas até centenas de milhares de marcadores SNP em um único ensaio para cada indivíduo, com enorme velocidade e automação (CAETANO, 2009).

Contudo, essa situação implica na necessidade de armazenamento de um grande volume de dados, não somente de genótipos, mas também de fenótipos e pedigree de um

número cada vez maior de animais. Dessa forma, realizar o armazenamento adequado e a extração de conhecimento útil a partir dessa quantidade de dados torna-se um grande desafio. 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 do banco de dados à aplicação desejada, pois esta terá impacto direto nos tempos de consulta e escrita em sistemas gerenciadores de bancos de dados relacionais (RDBMS – do inglês *Relational Database Management System*) onde essa informação estará armazenada.

Diante disso, com o objetivo de fornecer uma solução que fosse eficiente tanto no armazenamento quanto na consulta desse alto volume de dados, o sistema Web Banco de Dados de Genótipos e Fenótipos (BDGF) foi desenvolvido utilizando um diagrama de dados inicialmente proposto por (HIGA e OLIVEIRA, 2015). Esse diagrama foi redesenhado de forma que possibilitasse a implementação do tipo *JavaScript Object Notation* (JSON) em campos de tabelas relacionadas a fenótipos e do tipo texto (*character varying*) em campos de tabelas relacionadas a genótipos do diagrama. Com a implementação do tipo JSON e do tipo texto nessas tabelas, foi possível o uso da abordagem *Not Only SQL* (NoSQL - http://nosql-database.org) para armazenar parte dos dados sem que seja necessário se importar com a normalização dos mesmos, agilizando consultas que necessitariam realizar junções (*joins*) com outras tabelas.

MATERIAL E MÉTODOS

Material

O sistema está hospedado em computadores dedicados ao desenvolvimento e servidores (virtuais) para testes e homologação. Para o desenvolvimento do sistema Web, foram escolhidos componentes de tecnologia de informação (TI) disponível no mercado, ou seja, dentro da filosofia do uso de *software* livre. Primeiramente, como sistema operacional, optou-se pelo Linux Ubuntu (https://www.ubuntu.com) em todos os equipamentos.

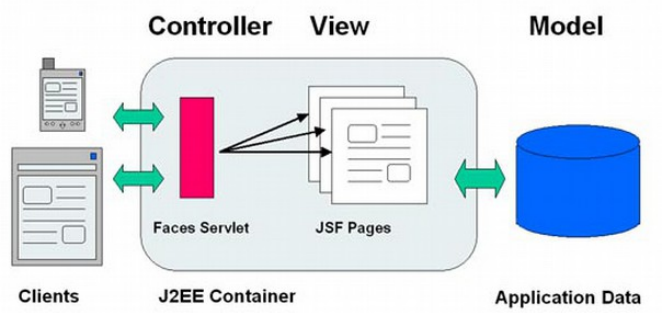
O sistema gerenciador de banco de dados (SGBD) escolhido foi o PostgreSQL (https://www.postgresql.org), versão 9.5. O PostgreSQL foi selecionado por ser um SGBD confiável, amplamente utilizado no mercado e pelo qual a equipe de desenvolvimento tem conhecimentos mais avançados. Além disso, oferece o formato JSON para tipificar os campos de tabelas. O tipo JSON é oferecido pelo PostgreSQL em suas últimas versões, a partir da 9.2. A tecnologia JSON consiste num formato de padrão aberto que consiste de conjuntos de pares

na forma "chave:valor". Ela foi utilizada para que fosse possível viabilizar o uso da abordagem NoSQL para armazenar parte dos dados sem que seja necessário se importar com a normalização dos mesmos.

O *software* utilizado para controle de versão foi o Subversion (SVN), versão 1.8.8. Já a linguagem de programação escolhida foi Java (https://www.oracle.com/br/java/), versão 8, e seus componentes da tecnologia *Java Enterprise Edition* (Java EE), que consiste de um conjunto de serviços e de interfaces de programa de aplicação (API) para o desenvolvimento de sistemas corporativos.

Dentre as tecnologias Java EE disponíveis e utilizadas pelo BDGF destaca-se, entre outras, a estrutura *Java Server Faces* (JSF). A arquitetura do *framework* JSF emprega o modelo MVC (*Model, View, Controller*), que faz a separação entre as camadas de apresentação e de aplicação. Na sua implementação como modelo MVC, o JSF possui uma camada de visualização bem distinta do conjunto de classes de modelo (Fig. 1). O JSF ainda se destaca por ser uma especificação do Java EE, isto é, todo servidor de aplicações Java tem que vir com uma implementação do *framework*.

Figura 1: Ilustração do modelo MVC (Model View Controller) adotado pelo JSF (Fonte: http://luissoares.com/jsf-parte-3-backing-beans/).



Todo sistema Web desenvolvido em Java que segue a especificação Java EE deve ser instalado e executado dentro de um servidor de aplicação. O servidor escolhido para abrigar o sistema BDGF foi o WildFly (http://wildfly.org/downloads/), versão 10, que é um dos servidores de aplicação mais seguros do mundo.

A aplicação foi desenvolvida na IDE (*Integrated Development Environment*) NetBeans (versão 8.1), que pode ser executada em diversas plataformas, como Windows, Linux, Solaris e MacOS. Essa versão do NetBeans já oferece o pacote JDK, versão 8, como linguagem Java padrão, além de variados recursos para se criar aplicativos profissionais para Web.

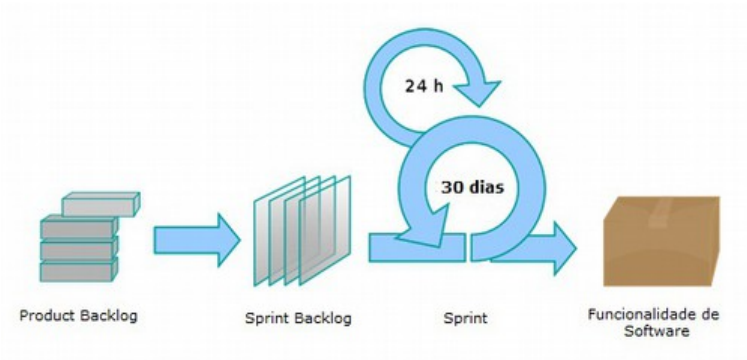
Métodos

Com base na lista de requisitos obtida para desenvolvimento do sistema de banco de dados, elaborada por meio da identificação do problema e das necessidades dos usuários, casos de uso e diagramas de sequência foram desenhados. A partir desse ponto, o projeto de desenvolvimento do sistema utilizou alguns conceitos do Scrum, que é um *framework* ágil para a realização de projetos complexos. O Scrum destaca-se dos demais métodos ágeis pela maior ênfase dada ao gerenciamento do projeto. Reúne atividades de monitoramento e *feedback*, em geral, por meio de reuniões rápidas e diárias com toda a equipe, procurando identificar e corrigir quaisquer deficiências no processo de desenvolvimento (SCHWABER, 2004).

Além disso, o método Scrum baseia-se em fundamentos como: equipes pequenas (no máximo, sete pessoas); requisitos desconhecidos e iterações curtas, dividindo o desenvolvimento em intervalos de tempos pequenos (no máximo, trinta dias), também chamados de *Sprints*. Ademais, existem três personagens importantes nesse processo: *Product Owner* (descreve os interesses de todos no projeto), *Time* (desenvolve as funcionalidades do produto) e *ScrumMaster* (responsável por garantir que todos sigam as práticas do Scrum).

O desenvovimento de um projeto Scrum inicia-se com uma visão do produto, a qual contém todas as características e restrições do produto determinadas pelo cliente (SCHWABER, 2004). Em seguida, cria-se o *Product Backlog* contendo a lista de todos os requisitos conhecidos, sendo então priorizados e divididos em tarefas, ou *sprints* (Fig. 2).

Figura 2: Representação geral do funcionamento Scrum (Fonte: http://www.devmedia.com.br/desenvolvimento-agil-com-scrum-uma-visao-geral/26343)



Dessa forma, a fase de construção compreendeu varias iterações (*sprints*), nas quais, em cada iteração, procurava-se desenvolver e corrigir pequenas partes do sistema, sendo esses testados e integrados no final, procurando satisfazer um subconjunto de requisitos do projeto. Tendo completado todas as tarefas do *backlog*, o desenvolvimento encontra-se agora em fase

de homologação pelos usuários em que, por meio de interações diretas com os usuários, o código fonte e a estrutura do sistema estão sendo ajustados e estabilizados.

Como ponto inicial do desenvolvimento do sistema, partiu-se pelo diagrama de dados inicialmente proposto por (HIGA e OLIVEIRA, 2015). Depois de diversas análises e levando-se em consideração que o SGBD utilizado seria o PosgreSQL e que a quantidade de dados a ser armazenada seria sempre crescente, optou-se por utilizar o formato JSON (*jsonb*) em campos de tabelas relacionadas a fenótipos e indivíduos (animais), e formato texto (*character varying*) em campos da tabela de genótipos do diagrama.

RESULTADOS E DISCUSSÃO

O sistema BDGF já possui muitos recursos implementados e está em processo de homologação pelos usuários. Por meio de sua interface Web é possível realizar consultas e importações de dados fenotípicos, genotípicos e de pedigree de diversas espécies de animais e, futuramente, de plantas também.

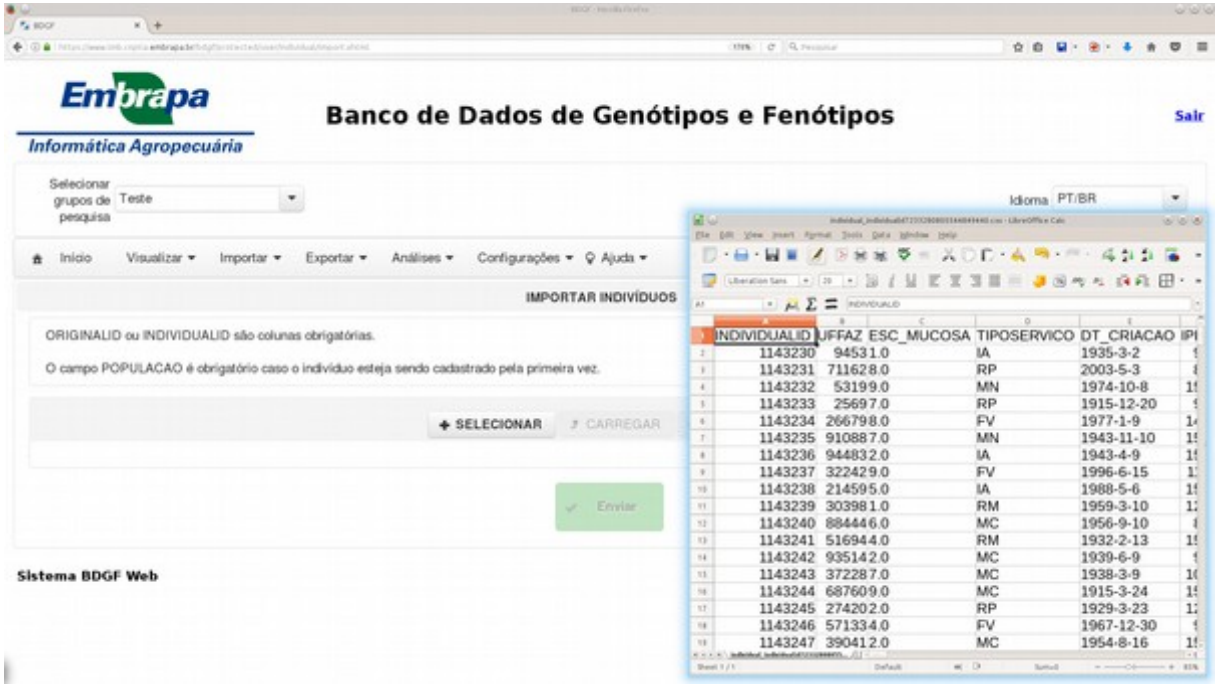
O usuário deve acessar o sistema BDGF por meio de um navegador Web tradicional. Ao acessar o endereço do sistema, a página de login será exibida (Fig. 3):

Figura 3: Tela de login do sistema BDGF.

Entre suas funcionalidades, destaca-se a função de importação de dados de animais. É possível importar dados de arquivos com colunas separadas por tabulações (TSV), que podem ser abertos e manipulados por *softwares* como o LibreOffice (Fig. 4). Esses arquivos precisam

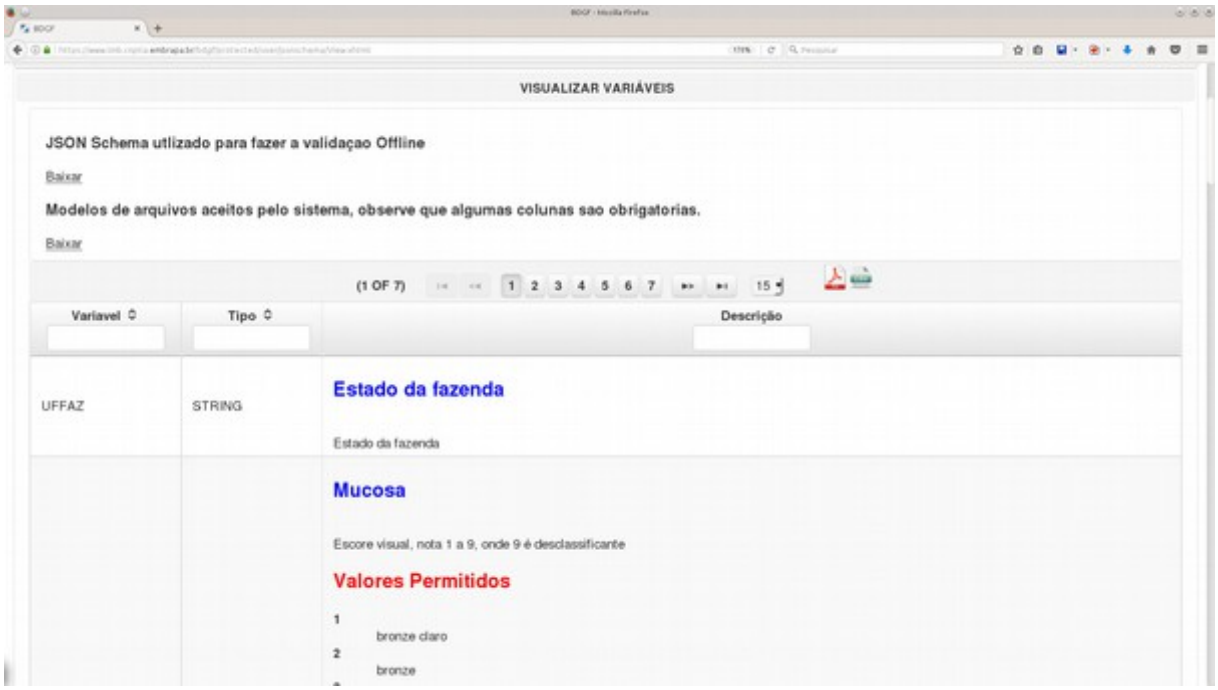
seguir um formato padronizado. O sistema fornece arquivos de exemplo para guiar o usuário a usar o formato apropriado.

Figura 4: Tela para importação de arquivo CSV com dados de animais



Contudo, cabe ressaltar que as variáveis dos fenótipos relacionados às espécies consideradas pelo sistema devem ser previamente registradas (Fig. 5). Essas variáveis devem ser importadas do Sistema de Experimentos da Embrapa - SIEXP (APOLINÁRIO *et al.*, 2016), onde foram definidas para a espécie com a qual o usuário trabalhará no seu grupo de usuários (ex: bovinos, suínos, etc.).

Figura 5: Tela para visualização das variáveis fenotípicas aceitas para um organismo importadas do SIExp.



Depois de importar os dados, é possível visualizar o pedigree de um animal listado na página de visualização de animais clicando em um ícone ao lado de cada registro (Fig. 6). A janela de pedigree pode ser expandida para facilitar a visualização dos animais e dos seus antepassados.

Figura 6: Janela de visualização de pedigree do animal selecionado.

Outra característica importante do BDGF é a capacidade de se exportar os genótipos dos animais para arquivos nos formatos MAP e PED (Fig. 7).

Figura 7: Tela para exportação de genótipos nos formatos MAP e PED.

Avaliando outros softwares com interface web desenvolvidos pela Embrapa Informática Agropecuária (VIEIRA, 2010; 2012a; 2012b) com funcionalidade de armazenamento de genótipos e fenótipos, e que contemplam algumas consultas básicas ao conjunto de dados moleculares (SNP), uma consulta simples em cerca de 800 animais e 700 mil marcadores SNP cada um demora pelo menos uma hora para ser processada. Uma consulta semelhante (utilizando a mesma máquina) feita no banco BDGF leva pouco menos de um minuto, pois a utilização de campos dos tipos JSON e texto nas tabelas de fenótipos e genótipos, respectivamente, retira parte da normalização necessária do desenho tradicional de bancos de dados, agilizando as consultas.

O sistema BDGF está sendo documentado e testado e espera-se que, em breve, esteja totalmente operacional.

CONCLUSÕES

O sistema BDGF foi projetado para apoiar programas de melhoramento animal da Embrapa, mas pode ser facilmente ajustado para armazenar dados de diversas fontes, como dados de plantas. Ele permite o armazenamento e acesso rápido a dados de pedigree, fenótipos e genótipos, utilizados para realização de avaliações genéticas. Além disso, o sistema

implementa políticas de acesso e segurança para fenótipos, genótipos e pedigree dos animais. Como trabalhos futuros, pretende-se integrar o BDGF com sistemas de avaliação genética, bem como construir Web Services para que outros sistemas possam obter informações da base de dados BDGF.

REFERÊNCIAS

APOLINÁRIO, D. R. de F.; QUEIROS, L. R.; VACARI, I.; CRUZ, S. A. B. da SIExp - Sistema de Informação de Experimentos da Embrapa. Versão v. 1.7.6. Campinas: Embrapa Informática Agropecuária, 2016.

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. R. Bras. Zootec., Viçosa , v. 38, n. spe, p. 64-71, 2009.

HIGA, R. H.; OLIVEIRA, G. B. Banco de Dados de Genótipos e Fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal. Campinas: Embrapa, 2015 (Série Documentos). Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/138127/1/Doc133.pdf>. Acesso em: 19 de Jul. 2017.

SCHWABER, Ken. Agile project management with Scrum, Microsoft Press. 2004.

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.