UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA GERAL
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

DISSERTAÇÃO DE MESTRADO

DESENVOLVIMENTO DE UMA INTERFACE WEB PARA A PLATAFORMA DIVERGENOME:

UMA FERRAMENTA BIOINFORMÁTICA PARA ESTUDOS DE GENÉTICA DE POPULAÇÕES E EPIDEMIOLOGIA GENÉTICA

AUTORA: MÁRCIA LOBÃO IANNINI
ORIENTADOR: EDUARDO MARTÍN TARAZONA SANTOS

Belo Horizonte
2010

MÁRCIA LOBÃO IANNINI

DESENVOLVIMENTO DE UMA INTERFACE WEB PARA A PLATAFORMA DIVERGENOME - UMA FERRAMENTA BIOINFORMÁTICA PARA ESTUDOS DE GENÉTICA DE POPULAÇÕES E EPIDEMIOLOGIA GENÉTICA

Dissertação de Mestrado apresentada à Universidade Federal de Minas Gerais – UFMG, Programa de Pós Graduação em Genética, para obtenção do título de Mestre em Genética.

Orientador: Eduardo Martín Tarazona Santos

Co-orientadora: Alessandra Faria-Campos

Belo Horizonte
2010

À minha querida mãe
e ao meu desejado filho,
Dedico

AGRADECIMENTOS

Agradeço ao Professor Eduardo Martin Tarazona Santos pela orientação e a co-orientadora Dra. Alessandra Faria-Campos, pelo apoio e incentivo. Aos colegas do Laboratório de Diversidade Genética Humana (LDGH) e aos colegas do Laboratório de Universalização do Acesso (LUAR) do Departamento de Ciência da Computação pela colaboração.

Agradeço aos amigos e familiares pela presença constante, em especial a minha mãe e irmãos por tudo que representam pra mim.

Agradecimento especial ao colega de trabalho Daniel que sempre me prestou auxílio e esclarecimento das dúvidas todas as vezes que precisei.

Agradeço ao meu marido pelo companheirismo e pela alegria de ter me feito mãe.

RESUMO

A plataforma bioinformática DIVERGENOMEdb consiste de um banco de dados relacional que permite organizar e reunir uma série de informações genéticas e fenotípicas provenientes de vários projetos de genética de populações e epidemiologia genética. DIVERGENOMEdb facilita o compartilhamento dos dados obtidos das pesquisas científicas, através de um sistema que permita inserir novas informações e navegar pelo banco de dados de maneira ágil e simples. Com esse objetivo foi construída a interface gráfica para o banco de dados relacional DIVERGENOMEdb. Foram utilizadas as linguagens de programação HTML, PHP e JavaScript embutidos. A plataforma utilizada foi o Linux/Intel e a aplicação hospedada em um servidor Apache. Com a interface desenvolvida é possível ao usuário executar diversos tipos de buscas de dados genéticos através da filtragem de pesquisas, além de possibilitar aos usuários cadastrados a inserção de dados científicos no DIVERGENOMEdb, sem que os mesmos tenham conhecimento das linhas de comando SQL, iniciando um ciclo de realimentação positiva que envolve a utilidade científica e as contribuições da comunidade.

Palavras-chave: DIVERGENOME, genética de populações, epidemiologia genética, banco de dados biológico, bioinformática.

ABSTRACT

The DIVERGENOMEdb bioinformatics platform consists of a relational database which allows the user to organize and gather a series of genetic and phenotypic information from multiple population genetics and genetic epidemiology projects. DIVERGENOMEdb facilitates to share data obtained from scientific research through a system that allows to enter new information and access the database quickly and simply. With this purpose, we built a graphical interface for the relational database DIVERGENOMEdb. We used the programming language HTML, PHP and JavaScript embedded. The platform used was the Linux / Intel and the application was hosted on an Apache server. With the developed interface the users can perform many kinds of data queries using filters, and enable registered users to insert scientific data in DIVERGENOMEdb without the need of previous knowledge of SQL command lines.

Keywords: DIVERGENOME, population genetics, genetic epidemiology, biological database, bioinformatics.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CEPH – *Centre d'Etude du Polymorphisme Humain* (Centro de Estudo do Polimorfismo Humano)

CGI - Common Gateway Interface

CNV - Copy Number Variation

DBA - Administrador do Banco de Dados (DBA

dbSNP – The Single Nucleotide Polymorphism Database

DDBJ – DNA Data Bank of Japan

DER - Diagrama de Entidade-Relacionamento

DNA - Ácido Desoxirribonucléico

DNAsp – DNA Sequence Polymorphism

EMBL – Laboratório de Biologia Molecular Europeu

ER - Entidade-Relacionamento

GenBank - Genetic Sequence Data Bank

GNU General Public License (Licença Pública Geral)

GWAS: Genome-Wide Association Studies

HGDP - Human Genome Diversity Project (Projeto da Diversidade do Genoma Humano)

HGP - Human Genome Project (Projeto Genoma Humano)

HTML - HyperText Markup Language (Linguagem de Marcação de Hipertexto)

Perl – Practical Extraction and Report Language

PHP - Hypertext Preprocessor

SGBD - Sistema de Gerenciamento de Banco de Dados

SNP – Single Nucleotide Polymophisms (Polimorfismo de Base Única)

SQL - *Structured Query Language*

WWW - World Wide Web

SUMÁRIO

1. INTRODUÇÃO

Com a invenção da técnica da Reação em Cadeia da Polimerase nos anos 80 e com o surgimento dos seqüenciadores automáticos de DNA na década de 90, aumentou consideravelmente a quantidade de informações genéticas a serem armazenadas (CAMARGO FILHO, 2002).

Os avanços das tecnologias de análise do genoma devem ser acompanhados do aumento da capacidade de computação e do desenvolvimento de *softwares* que possibilitem armazenar em bancos de dados informações genéticas em grandes quantidades. As soluções prévias, constituídas principalmente de bancos de dados centrais, estão sendo reformuladas com o desenvolvimento de novos sistemas com gestão automatizada de fluxos de dados, juntamente com tecnologias emergentes que realçam a conectividade e a recuperação de dados (GUDMUNDUR et al., 2009).

Um sistema de banco de dados é basicamente um sistema computadorizado de manutenção de registros, cuja finalidade é armazenar informações e permitir que os usuários as busquem e as atualizem quando necessário (DATE, 2003).

Sistemas de banco de dados são de grande relevância para as análises genômicas, pois além de manterem todo o volume de dados organizado, também executam tarefas e comandos que podem ser previamente programados por uma pessoa que define os serviços a serem realizados pelo sistema baseado nas rotinas e de acordo com a necessidade de cada usuário (ELSMARI, 2005).

Os bancos de dados integrados podem ser considerados como uma unificação de vários arquivos distintos, com a eliminação de redundância parcial ou total entre esses arquivos (DATE, 2003). A organização das informações pode ser feita com o uso de mais de uma tabela (RESENDE e SILVA, 2008). Já no caso de bancos de dados compartilhados, este pode ser compartilhado entre diferentes usuários que podem ter acesso aos mesmos dados de forma parcial ou total simultaneamente (DATE, 2003).

O elevado número de informações geradas todos os dias pelo mapeamento de genes necessitam ser armazenadas de forma sistemática em bancos de dados computacionais, servindo de base para estudos médicos e biológicos através da Bioinformática (LEAL, 2003).

Os bancos de dados genéticos assim desenvolvidos possibilitam gerenciar o grande volume de informação genética e viabilizam análises destes em pesquisas científicas. Os primeiros bancos de destaque na genética foram projetados para armazenar dados de seqüências de DNA. No início de 1980, logo que o uso de tecnologias de seqüenciamento de DNA tornou-se mais comum, tais depósitos foram necessários para facilitar a troca e a comparação de seqüências de DNA. Três grandes bancos de dados centrais foram construídos com esta finalidade: o banco de dados de DNA do Japão (DDBJ), GenBank (baseado nos Estados Unidos) e o Laboratório de Biologia Molecular Europeu (EMBL) (GUDMUNDUR et al., 2009).

De acordo com Suarez et al. (2008), alguns bancos de dados de seqüências especializaram-se ao longo dos anos no armazenamento de tipos específicos de informação. Um exemplo deste tipo de banco especializado são os bancos de Polimorfismos de Base Única (Single Nucleotide Polymorphisms, SNPs). SNP é uma variante genética entre indivíduos, limitada a um único par de bases, o qual pode ser substituído, inserido ou removido. A correlação entre uma doença e um SNP específico é uma vantagem para a prática clínica, pois torna relativamente fácil a identificação de pessoas afetadas ou portadoras. (LESK, 2008). Os SNPs são os marcadores mais amplamente utilizados em estudos para avaliar as associações entre variantes genéticas e fenótipos complexos como as doenças multifatoriais. Eles também são cada vez mais importantes em estudos de evolução humana e de outras espécies (SUAREZ et al., 2008).

Sherry (2001) descreve que entre os bancos de SNPs pode-se citar o dbSNP (www.ncbi.nlm.nih.gov/dbSNPs), que armazena informação de domínio público para uma ampla coleção de SNPs em diferentes espécies. Outro banco de dados particularmente relevante é o do projeto Internacional HapMap (www.hapmap.org), que é utilizado para selecionar SNPs para estudos em larga escala de estudos de associação genótipo-fenótipo e mapeamento de doenças complexas baseado no desequilíbrio de ligação no genoma humano. O objetivo do projeto HapMap é catalogar os SNPs comuns (nas fases I e II do projeto) do genoma humano e determinar o padrão de desequilíbrio de ligação entre eles em quatro populações humanas (europeus, africanos, chineses e japoneses). Na fase III do projeto, incluíram-se SNPs raros e outras cinco populações humanas. A informação

disponível no projeto HapMap possibilitou o desenho eficiente de arranjos que permitindo a genotipagem de 1-2.5 milhões de SNPs no genoma de um indivíduo em estudos de associação de varredura genômia (GWAS: genome-wide association studies), tem levado a descoberta de SNPs associados a doenças comuns, e ao desenvolvimento de ferramentas de diagnóstico, assim como uma melhora da capacidade de escolher alvos para intervenção terapêutica) (INTERNATIONAL HAPMAP CONSORTIUM, 2003).

Outro exemplo é a base de dados SNP500Cancer (http://snp500cancer.nci.nih.gov), um banco de dados que fornece a seqüência e a informação sobre protocolos de genotipagem de SNPs em genes relevantes na carcinogênese, imunidade e farmacogenética, sendo muito útil no mapeamento de doenças complexas. SNP500Cancer fornece a informação de seqüenciamento bidirecional sobre um conjunto de amostras de DNA provenientes de indivíduos anônimos representativos de quatro grupos étnicos: afro-americanos, caucasianos, hispânicos e asiáticos (PACKER et al., 2004).

1.1 Estudo da Genética Epidemiológica e Populacional

A epidemiologia genética evoluiu de um enfoque em estudos sobre doenças mendelianas raras para a análise genética de características complexas. Com o advento de informações sobre seqüência completa do genoma humano e de outros organismos, o interesse da epidemiologia genética em desvendar a natureza dos fatores que influenciaram estas características se tornou fundamental (FEITOSA e KRIEGER, 2002).

Em estudos de populações, o conhecimento dos fatores genéticos e não-genéticos, bem como, os fenótipos relacionados às doenças e as adaptações aos fármacos é de grande relevância em estudos epidemiológicos (ROSENBERG, et. al, 2002).

O potencial dos dados genéticos para prover informações sobre a história e geografia das populações humanas já era conhecido a partir do estudo de proteínas, ainda no começo do século XX (CAVALLI-SFORZA, 2005). Entretanto, apenas quando o Projeto Genoma Humano (*Human Genome Project* – HGP) estava em

plena atividade, surgiu a idéia de um estudo sistemático e em larga escala das variações do genoma humano (CAVALLI-SFORZA, 1990). Mais especificamente, notou-se que amostras renováveis de populações bem escolhidas ao longo de todo mundo, para as quais qualquer parte do genoma poderia ser examinada, poderiam facilitar enormemente os estudos genéticos de geografia e história da espécie humana. Nascia, assim, a idéia do *Human Genome Diversity Project* (HGDP).

Estudos de genética de populações têm sido utilizados tanto para explicar os padrões de diversidade genética humana em termos de história populacional, quanto para entender as bases genéticas das adaptações fenotípicas. Por trás dessas diferenças há eventos evolutivos moldando a variabilidade genética, seja a deriva genética, seja a seleção natural. Entretanto, um dos principais obstáculos referente às inferências evolutivas repousa justamente na identificação de quais variantes evoluem por deriva genética e quais, devido às pressões seletivas (BALARESQUE; BALLEREAU; JOBLING, 2007).

A história da América Latina implicou um processo complexo de miscigenação de populações nativas e imigrantes recentes em toda uma vasta região geográfica. Poucos detalhes são conhecidos sobre este processo e sobre como isso afetou a composição genética de populações latino-americanas contemporâneas (WANG, et. al., 2008).

Os dados sobre a história demográfica da população latino-americana são escassos e os estudos no nível genômico nesta região são até agora bastante restritos em termos de número de populações e/ou marcadores que foram analisados. Uma pesquisa genômica sobre a miscigenação das populações em toda a América Latina é, portanto, de interesse considerável sendo também importante para avaliar o contexto no qual o mapeamento por miscigenação pode ser aplicado em populações desta região (WANG, et. al., 2008).

No contexto de grandes quantidades de informações sobre a diversidade genômica humana que estão emergindo, torna-se de fundamental importância que os grupos de pesquisa contêm com ferramentas computacionais para o desenvolvimento de novos algoritmos e de bancos de dados e softwares que possibilitem armazenar informações genéticas em grandes quantidades, além de auxiliar nas análises e comparações entre os dados disponíveis e o cruzamento

entre as informações biológicas inferidas em projetos independentes e entre bases de dados diferentes. A integração de dados poderá ajudar a responder questões específicas experimentais e a descobrir novos relacionamentos e detectar padrões gerais de dados (SEARLS, 2005).

Com o objetivo de contribuir no desenvolvimento de um sistema adequado para um grupo de pesquisa de médio porte, Magalhães e colaboradores (2010, manuscrito em preparação) e outros integrantes do Laboratório de Diversidade Genética Humana, vêm desenvolvendo a plataforma DIVERGENOME que apresenta dois componentes: um banco de dados relacional, DIVERGENOMEdb; e um conjunto de ferramentas bioinformáticas para facilitar a manipulação e análise dos dados: DIVERGENOME tools.

DIVERGENOMEdb é um banco de dados relacional que reúne dados genotípicos e fenotípicos provenientes de diferentes estudos nas áreas de Genética de Populações, Genética Clínica e Epidemiológica realizados em populações indígenas ou miscigenadas da América Latina e outras regiões.

1.2 O Projeto DIVERGENOME

O DIVERGENOME tem como objetivo principal o armazenamento e manipulação eficiente de dados de projetos de Genética de Populações e Epidemiologia Genética do Laboratório de Diversidade Genética Humana.

A plataforma DIVERGENOME possui dois componentes: um banco de dados relacional – DIVERGENOMEdb, que possibilitará o armazenamento seguro e eficiente de dados de projetos de genética de populações e epidemiologia genética, e um conjunto de ferramentas para facilitar a análise dos dados - DIVERGENOMEtools.

O modelo ER (entidade-relacionamento) descreve os dados como entidades, relacionamento e atributos (ELMASRI, 2005). Um banco de dados relacional é um banco de dados percebido por seus usuários como uma coleção de tabelas (DATE, 2003). A tabela, também chamada de entidade, tem um nome e várias colunas que representam propriedades particulares que descrevem cada tabela (ELMASRI, 2005). Cada coluna na tabela tem um nome único e contém dados diferentes. As

colunas são chamadas como campos ou atributos (WELLING, 2005). Os valores dos atributos que descrevem cada entidade serão os dados armazenados no banco de dados, cada linha é chamada de tupla. (ELMASRI, 2005). Os bancos de dados relacionais são compostos de relações e utilizam uma chave como uma referência de uma tabela para outra. As chaves primárias são atributos ou combinação de atributos que possuem a propriedade de identificar de forma única uma linha da tabela. Uma chave estrangeira (foreign key) é um atributo de uma tabela que contém dados de uma chave primária de outra tabela. Uma chave estrangeira é um campo, que aponta para a chave primária de outra tabela. Ou seja, passa a existir uma relação entre essas duas tabelas. A finalidade da chave estrangeira é garantir a integridade dos dados referenciais. O DIVERGENOMEdb é um banco de dados relacional pois as tabelas são interligadas por chaves estrangeiras, o que permite a filtragem de pesquisa interligando campos de diferentes tabelas.

O DIVERGENOMEdb foi desenvolvido em MySQL. É constituído de vinte tabelas, cada uma contendo um número variável de campos como pode ser visto no Diagrama de Entidade-Relacionamento (DER) do mesmo (Figura 1).

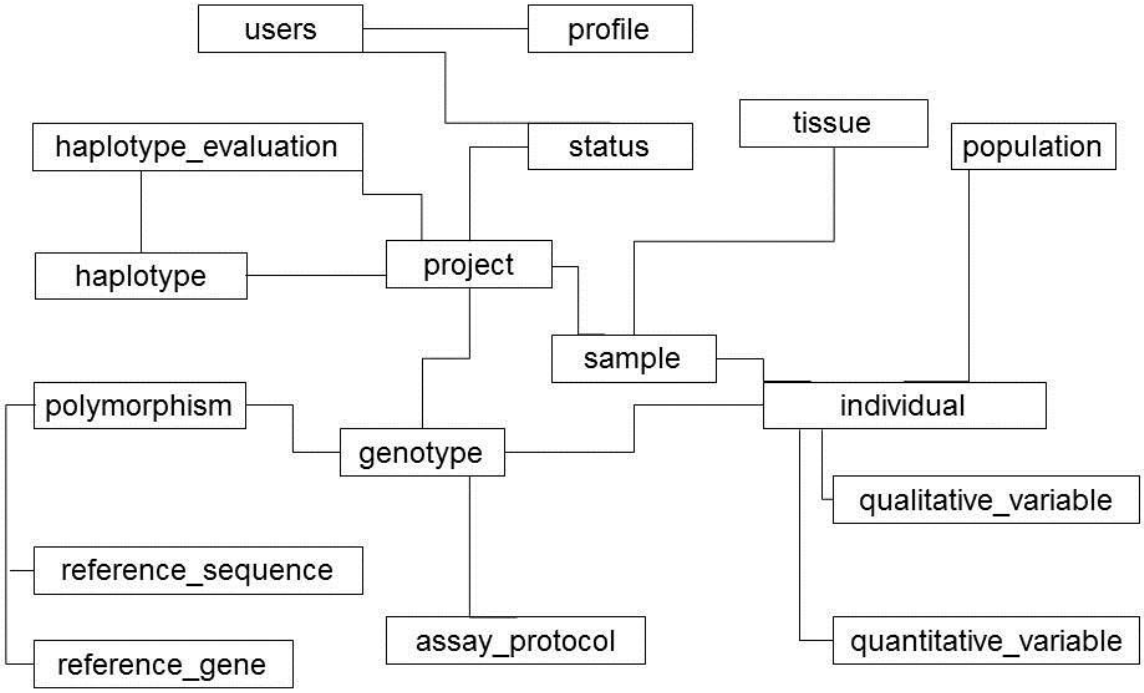


FIGURA 1: DER DIVERGENOMEdb Magalhães e cols. (2010, manuscrito em preparação). Diagrama que apresenta as relações entre as tabelas que compõe o DIVERGENOMEdb. Na terminologia do modelo relacional formal, a entidade se refere às tabelas e os atributos aos campos que descrevem cada entidade em particular.

No modelo de banco de dados relacional, a organização das informações pode ser feita de maneira descomplicada com o uso de mais de uma tabela. Outro benefício é o compartilhamento de dados, o que de certa maneira possibilita um controle no que se refere à redundância de informações (LESK, 2008). A criação de bases de dados permite o armazenamento, a administração, a extração e a difusão generalizada da informação biológica, disponibilizando, ferramentas computacionais desenvolvidas para atualizar, pesquisar e recolher dados armazenados no sistema (RESENDE e SILVA, 2008). A informação genética com ênfase especial na análise da seqüência do DNA precisa ser ampliada para um escopo mais abrangente para controlar todos os tipos de informação biológica – sua modelagem, armazenamento, recuperação e gerenciamento. Além disso, as aplicações de Bioinformática se estendem por projetos de alvos para drogas, estudos de mutações e doenças relacionadas, investigações antropológicas sobre padrões de migração de tribos, e

tratamento terapêutico, enfim, servindo de base para estudos médicos e biológicos (LEAL, 2003). Hoje, a Bioinformática é uma ciência aplicada. Utiliza-se de programas de computador para fazer inferências a partir de dados obtidos da biologia molecular moderna, para fazer conexões entre eles e para derivar predições importantes e relevantes (LESK, 2008). Buscando integrar essa complexa variedades de dados biológicos, o DIVERGENOMEdb aceita dados genotípicos individuais de quatro tipos: contigs, SNPs, INDELs e microssatélites, e ainda pode ser facilmente modificado para incorporar polimorfismos como *Copy Number Variation* (CNV). Os genótipos são ligados a uma descrição dos protocolos de laboratório usados para gerar dados. Os indivíduos podem estar ligados a dados fenotípicos coletados em estudos epidemiológicos, que podem incluir o status da doença como atributo binário, variáveis quantitativas como idade, ou variáveis fisiológicas. O acesso a estes dados pode ser variável conforme os níveis de hierarquia que serão obedecidas para a manipulação dos dados.

Em associação ao banco de dados vêm sendo desenvolvido também o DIVERGENOME Tools, um conjunto de scripts desenvolvidos em Perl que disponibilizará aos usuários o acesso a ferramentas de manipulação de saídas (*outputs*) do DIVERGENOMEdb, para a geração de entradas (*inputs*) para programas de genética de população (PHASE, DNAsp, Structure), genética médica (Haplopainter) e pacotes de análises estatísticas (MAGALHÂES 2010, manuscrito em preparação). A maioria dos programas usa um determinado formato de arquivo de dados. Os usuários normalmente precisam analisar o mesmo conjunto de dados com vários programas, que descrevem os seus respectivos formatos de entrada, por isso torna-se necessária especial atenção à sua interoperabilidade, como um importante fator que limita o uso de um determinado programa, em que muitas vezes existe a necessidade de reformatar os dados brutos para utilização em um programa específico. Ferramentas de conversão dos dados para formatos específicos disponibilizados em um único sistema são essenciais para as análises, evitando-se assim a re-formatação manual dos dados. Daí a importância do DIVERGENOME Tools, que disponibilizará aos usuários o acesso a ferramentas de conversão de base de dados para diversos formatos de arquivos utilizados em diferentes programas.

A Figura 2 ilustra a comunicação entre os *scripts* já desenvolvidos e a base de dados DIVERGENOME (MAGALHÂES 2010, manuscrito em preparação).

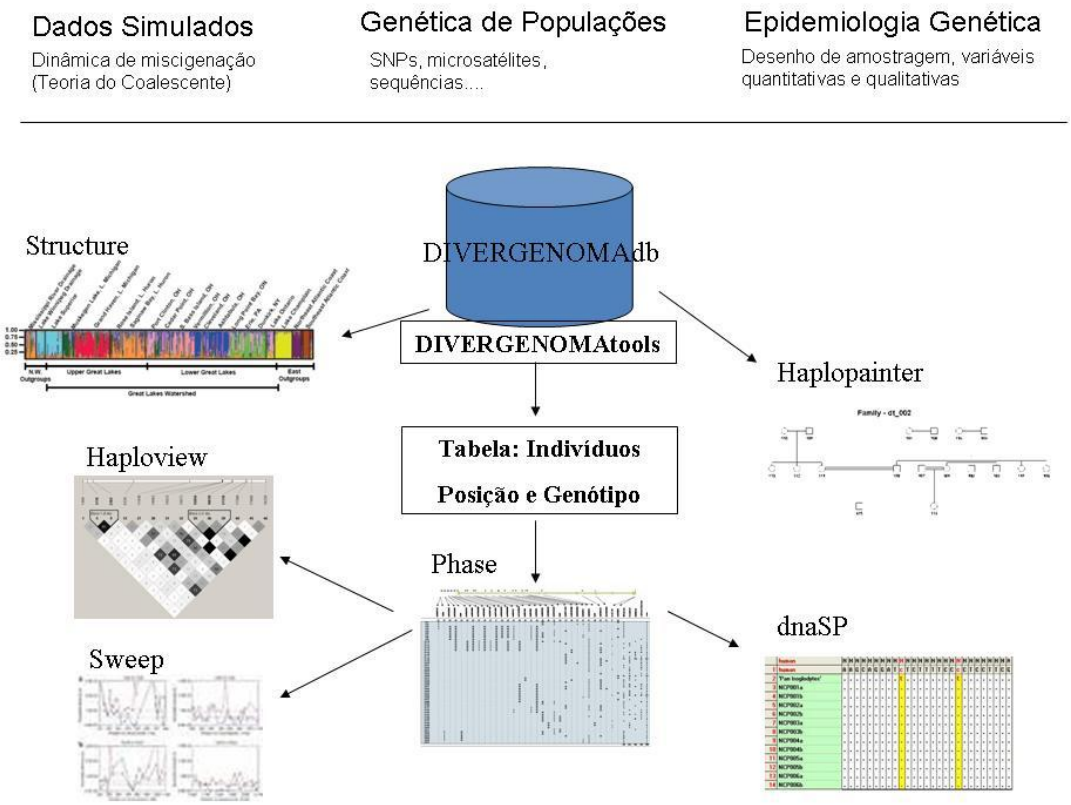


FIGURA 2: DIVERGENOMEtools
Fonte: Magalhães, 2010 – manuscrito em preparação. Comunicação entre os *scripts* já desenvolvidos e base de dados DIVERGENOME.

Contudo, grande parte dos pesquisadores não tem o conhecimento das linhas de comando necessárias para a manipulação dos dados, sendo assim, para que o DIVERGENOME seja acessível a todo o público alvo a que é direcionado, torna-se necessário à criação da interface gráfica do usuário, através de elementos gráficos e outros indicadores visuais.

Em contraste à interface de linha de comando, o ambiente gráfico, facilita e torna prática a navegação pelo banco de dados. Ela vem auxiliar os usuários as escolhas dos tipos de análises a serem realizadas de uma maneira mais dinâmica e de fácil compreensão (RESENDE e SILVA, 2008).

Numa base de dados a qualidade, quantidade e originalidade de dados, bem como a qualidade da interface *web* são importantes. Bons dados com uma interface

pobre, ou vice-versa, nunca são suficientemente apreciados (BATEMEN, 2007). Através do desenvolvimento de uma interface *web* será possível aumentar o número de usuários do sistema DIVERGENOME.

Importa considerar que a necessidade da troca de dados advindos de diferentes fontes vem crescendo relativamente, devido principalmente ao grande tráfego de informação que é gerado tanto em sistemas convencionais como em sistemas *web*. Tais dados são escritos de várias maneiras, sendo que, em alguns casos, um mesmo objeto do mundo real pode ser representado de várias formas. Neste cenário, tem-se a necessidade de identificar similaridades entre tais objetos (SUDER e DORNELES, 2008).

Entretanto, vale ressaltar que não é uma tarefa fácil criar e manter um banco de dados, pois existem dificuldades para o desenvolvimento de plataformas que consigam representar fielmente ou aproximadamente as relações que podem ser feitas entre os componentes de um ou de vários sistemas biológicos.

Segundo Silva (2004) essa elevada quantidade de informações precisa ser cuidadosamente decifrada e com a organização pode-se encontrar informações muito valiosas sobre os mecanismos biológicos.

Leal (2003) descreve que para resolver este problema, deve ser feita uma padronização de tipos de dados e métodos que os bancos de dados devem suportar, relacionando-os e descrevendo suas funcionalidades. Surge assim necessidade de se possuir diferenciadas formas de armazenamento, tratamento, acesso e pesquisa dos dados, para que se consiga trazer a informação da melhor maneira desejada possível, desse modo, destaca-se a necessidade de propor sistemas que permitam a integração dos dados. Estes, em linhas gerais possuem o objetivo principal de tratar a comparação dos dados gerados, com vários níveis de alinhamento (DORNELES et al. 2003).

Desta forma, a integração eficaz dos dados e do conhecimento de muitas fontes diferentes passa a ser crucial para a descoberta de fârmacos e é um elemento-chave da realização de investigações científicas. Além disso, o desenvolvimento de uma interface *web* eficiente como a proposta neste trabalho para o sistema DIVERGENOME constitui uma importante adição para aumentar a eficiência e usabilidade do mesmo.

2. OBJETIVOS

2.1. Objetivo geral

Desenvolver a interface gráfica para acesso ao banco de dados DIVERGENOMEdb.

2.2. Objetivos Específicos

- Construir uma interface web de acesso ao DIVERGENOMEdb
- Construir nesta interface mecanismos que possibilitem ao usuário executar buscas de dados genéticos através da filtragem de pesquisas;
- Construir nesta interface mecanismos para possibilitar aos usuários cadastrados a inserção de dados científicos no DIVERGENOMEdb.
- Desenvolver sistemas de *login* e senha, bem como o sistema de privilégios dos usuários para controle de acesso ao DIVERGENOMEdb (módulos básico de segurança).

3. METODOLOGIA

3.1 Descrição do banco de dados DIVERGENOME

DIVERGENOME DATABASE foi desenvolvido pelo aluno de doutorado em Bioinformática Wagner Magalhães e outros integrantes do Laboratório de Diversidade Genética Humana (MAGALHÃES et al. 2010, manuscrito em preparação). Este banco tem como objetivo contribuir no desenvolvimento de sistemas de Análise Genética, possibilitando o armazenamento seguro e eficiente de dados de projetos de Genética de Populações e Epidemiologia Genética.

Abaixo segue uma breve descrição das entidades, bem como dos atributos disponibilizados no database. A descrição foi dividida em dois grupos:

- Entidades relacionadas ao controle de usuários e projetos;
- Entidades relacionadas a dados genéticos.

3.1.1 Entidades relacionadas ao controle de usuários e projetos

As entidades referentes ao controle de usuário visam definir os acessos e permissões dos mesmos. Os usuários são previamente cadastrados e inseridos na entidade *Users*. Abaixo segue uma breve descrição de seus atributos.

- User_name: nome completo do usuário;
- Login: login de acesso do usuário;
- Pwd: password de acesso do usuário;
- Citation: escrita do nome do usuário em citações;
- Instituition: instituição do coordenador ou usuário;
- Email: email de contato com o usuário.

A entidade “*Profile*” define a hierarquia dos potencias usuários. Estes podem ser administradores, coordenadores ou simples usuários (membros de projetos). O acesso às informações irá variar conforme os níveis pré-estabelecidos para a manipulação dos dados, sendo os administradores com acesso ilimitado a base de dados, os coordenadores que poderão inserir/editar dados de seus projetos e os

simples usuários que podem fazer buscas e utilizar-se das ferramentas disponíveis no DIVERGENOME Tools. Os coordenadores terão acesso ao banco de dados através do cadastramento prévio à Plataforma DIVERGENOME e posterior controle de privilégios pelo administrador.

A entidade "*statu*" realiza o controle de privilégios desses usuários, bem como define o status do projeto. Possui dois atributos:

- Dsc_status: define o tipo do status do usuário ou do projeto;
- Name_table: define se o controle se refere ao usuário ou ao projeto. Se a tabela se referir a usuários, os possíveis valores de status (dsc_status) serão: aguardando, concluído e recusado. No que se refere à tabela projetos, os possíveis valores de status (desc_status) serão: público e privado.

  A Tabela 1 apresenta os possíveis valores de status para usuários e projetos. A coluna "Dsc_name" se refere ao nome que define os possíveis valores de status, a coluna "Name_table" se refere a tabela relacionada, ou seja, se o status descrito em Dsc_name se refere a usuário ou a um projeto, e a coluna "Utilização" se refere ao emprego dessas funcionalidades.

TABELA 1

Possíveis valores de status para usuários e projetos

Dsc_name	Name_table	Utilização
Aguardando	Users	Usuário realizou o cadastro e aguarda aprovação do administrador
Concluído	Users	O cadastro realizado pelo usuário foi aprovado pelo administrador
Recusado	Users	O cadastro realizado pelo usuário foi recusado pelo administrador
Público	Project	Todos os dados inseridos para o projeto público serão visualizados por qualquer usuário do DIVERGENOMEdb
Privado	Project	Os dados inseridos para o projeto privado serão visualizados apenas pelo coordenador responsável pelo projeto e futuramente pelos membros cadastrados para este projeto (ainda em desenvolvimento)

Os usuários podem ter seu acesso aprovado, ou recusado pelo administrador. Quando do *login* pelo usuário, a mensagem do status será apresentada na tela, caso o mesmo esteja aguardando ou tenha seu cadastro recusado pelo administrador, como apresentado nas Figuras 10B e 10C contidas na parte de Resultados deste material. Os projetos podem ser públicos ou privados. Se públicos, todas as informações inseridas serão visualizadas pelos usuários, se privados, apenas os coordenadores (em projetos de autoria do próprio coordenador) e os membros de seus projetos poderão visualizar os dados inseridos no banco (controle em desenvolvimento).

A entidade "*Project*" disponibiliza informações referentes aos projetos inseridos no database DIVERGENOME. Possui os seguintes atributos:

- Project_name: nome do projeto;
- Id_author: identificação do nome do coordenador de autoria deste projeto;
- Data_first_upload: início da inserção de dados - data da primeira inserção;
- Data_last_upload : data da última inserção de dados;

- Investigators: membros do projeto;
- Publication: publicações;
- Publication_date: data da publicação;
- Description: campo de texto que permite a adição de informações pertinentes ao projeto;
- Dsc_project: informações descritivas sobre o projeto.

3.1.2 Entidades relacionadas a dados genéticos.

O banco DIVERGENOME está direcionado para as necessidades de grupos de pesquisas de pequeno porte. Pode-se extrair no banco subconjuntos de dados, ou combinar dados de diferentes fontes e oferecer formas avançadas de acesso. É possível procurar informações de polimorfismos por sequência de referência ou pelo gene de referência. Registros adicionais são fornecidos para representar dados fenotípicos. Trata-se de uma base para a integração de uma grande massa de dados diversos. São diversas as aplicações do banco de dados.

3.1.2.1 Descrição das entidades

A entidade "*Polymorphism*" armazena informações sobre os polimorfismos. Possui chaves estrangeiras com diferentes entidades, entre elas: "reference_sequence", "reference_gene" e "haplotype". O que permite relacionar o polimorfismo identificado a um gene ou sequência ou até mesmo a um haplótipo. A entidade *Polymorphism* é constituída pelos seguintes atributos:

- Polymorphism_code: representa o código de identificação do polimorfismo. Esse identificador corresponde ao código de identificação interno do banco de dados SNP500Cancer ou código de identificação para polimorfismos descobertos ou submetidos ao projeto *The Single Nucleotide Polymorphism* database (dbSNP).
- Kind: podem ser SNPs, contig (produtos de sequenciamento), INDELs ou microssatélites;

- Sub_kind: indica se o polymorphism encontrado está em região de íntron ou éxon;
- Reference_value: posição do polimorfismo no contig de seqüenciamento;
- Coord_relative_to_gene: indica a coordenada do polimorfismo relativa ao gene do polimorfismo;
- Chromossome: indica em qual cromossomo está localizado o polimorfismo;
- Position_reference_sequence: posição do polimorfismo na sequência de referência;
- Assm_build_version: indica a versão de montagem do genoma correspondente às informações disponibilizadas pelo banco;
- Assm_coord_start: indica a coordenada de início;
- Assm_coor_end: indica a coordenada de fim;
- Description: campo de texto que permite a adição de informações pertinentes ao polymorphism, tais como, nomes anteriores, posições diferentes em relação ao anotado no dbSNP e quaisquer outras anotações de importância para projetos desenvolvidos a partir do banco.

A entidade "*Genotype*" armazena os dados genotípicos provenientes da amostragem dos indivíduos. É representada pelos atributos genotype, definindo assim os alelos encontrados no indivíduo genotipado. É referenciada por algumas chaves estrangeiras; uma proveniente da entidade *Polymorphism* que permite a identificação do polimorfismo e uma proveniente da entidade *Individual* que identifica o sujeito, permitindo assim o relacionamento do genótipo ao polimorfismo e ao indivíduo. Existem também as chaves estrangeiras com a entidade *Project*, *Assay Protocol* e *Haplotype Evaluation.*

A entidade "*Sample*" armazena informações relativas às amostras dos sujeitos para os quais dispomos de dados. É constituída pelos seguintes atributos relacionados as outras entidades, como as chaves estrangeiras para a entidade "*Tissue*", "*Project*" e "*Individual*". Este último recebe o código de identificação do sujeito genotipado. Atualmente contém três identificadores correspondentes aos dados armazenados: HGDP, SNP500Cancer e Nativos-Americanos genotipados no National Cancer Institute (NIH). Além disso apresenta os seguintes atributos:

- Sample_code: código de identificação da amostra;
- Collection_data: referente a informações sobre a data de coleta;
- Source: whole genome amplification ou genômico;
- Sample_strategy: define a estratégia de amostragem utilizada, sendo possível saber quais os tipos de análises são mais adequados a esse conjunto de dados.

A entidade "*Population*" armazena informações relativas às populações de origem dos indivíduos genotipados. Constituída pelos seguintes atributos:

- Population_name: nome da população à qual pertence o indivíduo. O código de identificação da população é a chave estrangeira que relaciona a tabela Population à tabela Individual;
- Geographic_origin: indica em qual continente a população se encontra.
- Country: localização geográfica da população , podendo ser definida por região ou país;
- Coordinates_of_population_sample: indica as coordenadas geográficas dos locais onde os indivíduos foram amostrados.

A entidade "*Individual*" armazena informações dos indivíduos genotipados. Constituída pelos seguintes atributos:

- Individal_code: código de indentificação dos indíviduos;
- Family_id: identificação da família a qual o indivíduo pertence;
- Father_id: identificação do pai;
- Mother_id: identificação da mãe;
- Sex: indica o sexo do sujeito amostrado;
- Aff_status: indica se o indivíduo é ou não afetado;
- Live_status: status de vida do indivíduo (se vivo ou falecido);

- Info1: permite o armazenamento de quaisquer informações relevantes referentes àquela indivíduo que não tenham sido anteriormente anotados em algum dos demais campos.

A entidade "*assay protocol*" armazena informações dos protocolos de laboratório utilizados para a geração de dados. Possui os seguintes atributos:

- Assay_protocol_name: o nome do protocolo utilizado para genotipagem ou sequenciamento;
- Kind: o tipo de protocolo utilizado;
- Description: permite o armazenamento de quaisquer informações relevantes referentes ao protocolo.

A entidade "*reference gene*" contém informações do gene de referência. É constituída pelos seguintes atributos:

- Reference_gene_name – refere-se ao nome do gene de referência;
- Symbol: símbolo de identificação do gene;
- Chromosome: o cromossomo em que o gene está localizado;
- Sequence: a sequência do gene;
- Ass_build_version: indica a versão de montagem do genoma correspondente às informações disponibilizadas pelo banco;
- Assm_coord_start: coordenada de início do gene;
- Assm_coord_end: coordenada de fim do gene;
- Description: permite o armazenamento de quaisquer informações relevantes referentes ao gene.

A entidade "*reference sequence*" contêm registros provenientes de uma região definida de uma sequência genômica. Possui como atributos:

- Reference_seq_name: nome da sequência de referência;
- Chromosome: o cromossomo em que a sequência está localizada;

- Cytogenetic_band: indica os padrões de bandeamento. (número do cromossomo, indicação do braço p ou q do cromossomo e os dígitos subseqüentes indicam as subdivisões das bandas);
- Contig: código de identificação dos produtos de seqüenciamento;
- Ass_build_version: indica a versão de montagem do genoma correspondente às informações disponibilizadas pelo banco;
- Assm_coord_start: coordenada de início da sequência;
- Assm_coord_end: coordenada de fim da sequência;

O DIVERGENOMEdb permite a inserção de variáveis que podem ser medidas na escala quantitativa como idade, peso, altura ou variáveis fisiológicas como pressão arterial e a inserção de variáveis qualitativas, que irão representar uma classificação do indivíduo como cor do olhos, fumante ou não fumante, etc. Sendo assim é possível ligar os indivíduos a dados fenotípicos coletados em estudos epidemiológicos, que podem incluir neste caso o status da doença. As entidades criadas por Wagner Magalhães que se referem a essas duas possíveis varáveis - quantitativas e qualitativas - possuem como atributos o nome da variável (quant_var_name e qual_var_name), valor (value) e permitem o armazenamento de quaisquer informações relevantes referentes às variáveis (description).

3.2 Ferramentas utilizadas para implementação do sistema

A plataforma utilizada foi o Linux, e o sistema é executado em um servidor *web* Apache 2.0. A representação gráfica do diagrama de Entidades-Relacionamento foi construída utilizando a biblioteca gráfica DBDESIGNER 4.5.6.

O DIVERGENOMEdb foi construído seguindo um modelo de aplicação de três camadas. A camada de dados foi implementada usando o MySQL versão 5.1.31. A camada de lógica de programação da interface foi construída usando a linguagem PHP versão 5.2.6 e JavaScript, para inserção, controle de validação e recuperação de dados. A camada de apresentação foi construída usando o HTML e o sistema de gerenciamento de layout foi desenvolvido em CSS.

3.2.1 Sobre o sistema de pesquisa ao banco de dados relacional

O DIVERGENOMEdb foi construído usando o servidor MySQL. A criação do banco e suas tabelas, bem como a administração, foram efetuadas através da plataforma de gerenciamento de banco de dados *on-line PHPMyAdmin* v. 4.3.1.2 capaz de executar, importar e exportar códigos SQL.

3.2.2 Sobre o PHP, JavaScript e HTML para criação de páginas Web interativos

A finalidade do DIVERGENOMEdb é permitir que nosso grupo de pesquisa e outros colaboradores compartilhem dados com simplicidade e eficiência, através de uma interface *web*. Várias tecnologias permitem a comunicação entre páginas Web e banco de dados. A mais antiga é denominada programação CGI (Common Gateway Interface), que foi aperfeiçoada por outras tecnologias como XML e PHP (GIBAS, 2001).

A interface gráfica com o usuário foi criada através da programação em HTML e JavaScript com PHP embutido. Optou-se pelo HTML (HyperText Markup Language) por se tratar de uma linguagem padrão para divulgação de documentos na rede (um documento escrito em HTML é, em geral, chamado de página *web*), que pode ser transferida de uma plataforma computacional para outra. Isto significa que se pode escrever código-fonte HTML sem se preocupar em qual computador e por qual sistema operacional este documento será visualizado portanto, qualquer computador deve ser capaz de interpretá-lo (VENETIANER, 1996). Conhecida também como uma extensão da linguagem HTML, os comandos JavaScript são embutidos nas páginas HTML e interpretados pelo navegador, ou seja, o JavaScript não possui nenhum procedimento de compilação. O JavaScript é usado normalmente pelos programadores que fazem uso da linguagem HTML para controlar dinamicamente o comportamento de objetos nas páginas.

Outro fator importante na escolha do HTML e também do JavaScript é que os códigos são executados diretamente no navegador do próprio cliente, o que é chamado de *script Client-Side* (lado cliente). Um cliente é em geral uma máquina de usuário que tem as funcionalidades de interface com o usuário e processamento local. Quando um cliente precisa de uma funcionalidade adicional, como acesso ao banco de dados, inexistente naquela máquina, ela se conecta a um servidor que

disponibiliza a funcionalidade. Um servidor é uma máquina que pode fornecer serviços para as máquinas clientes, como acesso a arquivos ou acesso a um banco de dados, podendo encontrar o *script* responsável pela busca e processá-lo (ELSMARI, 2005). O resultado é colocado em “pacotes” e enviado de volta ao computador que requisitou a pesquisa. Um servidor pode ter muitos clientes realizando requisições desse tipo, portanto, é recomendado que pequenas verificações e tarefas menores sejam executadas por aplicações *Client-Side*, deixando as requisições mais importantes (como conferir senhas por exemplo) para o Servidor. O código HTML, em outras palavras, tira a responsabilidade do processamento do servidor *web*, que fica livre para executar os *scripts Server-Side* (lado servidor), que são, por exemplo, os scripts PHP.

PHP (Hypertext Preprocessor) é uma linguagem de programação de computadores interpretada, livre e muito utilizada para gerar conteúdos dinâmicos na *World Wide Web*. PHP é multiplataforma: Inicialmente foi desenvolvido para ser usado em servidores Unix/Linux, ganhando uma versão para Windows e para Macintosh posteriormente. Isso faz do PHP uma linguagem capaz de ser executada independente da plataforma utilizada. Além desses fatores, a escolha pelo PHP se deu pelo fato de que o PHP também tem suporte incorporado para interação com banco de dados MySQL (GIBAS, 2001).

Com o uso dessa linguagem foi possível construir a busca de dados genéticos através da filtragem de pesquisas. Quando um dos filtros estiver ativo, toda a operação realizada com o DIVERGENOMEdb afetará apenas as amostras selecionadas no filtro. As linhas de comando em SQL foram montadas conforme as escolhas das variáveis selecionadas pelo usuário para filtragem da pesquisa, e o comando SQL da pesquisa foi disponibilizado para os usuários.

3.3 Módulos de Visualização *Web*

Grande parte dos usuários não possui, em geral, conhecimento da estrutura interna do banco de dados. As interfaces dos bancos de dados biológicos devem, portanto, exibir informação para o usuário de maneira que ela seja aplicável para o problema que está sendo tratado e que reflita a estrutura subjacente dos dados. Os usuários normalmente sabem quais dados eles necessitam, mas não têm

conhecimento técnico da estrutura de dados ou de como um sistema de gerenciamento de banco de dados representa estes dados (ELMASRI, 2005).

Foi discutida nas seções anteriores sobre as entidades de relacionamento, bem como os atributos referentes a cada entidade para a construção do sistema DIVERGENOMEdb. Para ampliar o uso do DIVERGENOMEdb foi preciso desenvolver interfaces amigáveis Web. A disponibilização dos recursos é apresentada conforme o modelo criado para o projeto DIVERGENOME, que contém duas plataformas com acessos distintos: DIVERGENOMEdb e DIVERGENOMEtools.

3.4 Recursos disponíveis no DIVERGENOMEdb através da interface *web*

3.4.1 Proteção da base de dados

Um problema comum a todos os sistemas de computação é prevenir que pessoas não autorizadas acessem o sistema, seja para obter informação, seja para realizar alterações mal-intencionadas em uma parte do banco de dados. O mecanismo de segurança de um banco de dados deve incluir providências para a restrição de acesso ao sistema de banco de dados como um todo. Essa função é chamada de controle de acesso e é tratada por meio da criação de contas de usuários e senhas para controlar o processo de *login* pelo sistema de gerenciamento de banco de dados (SGBD) (ELMASRI, 2005). O DIVERGENOMEdb possui esse controle de acesso. Neste caso o acesso aos dados através da interface web varia conforme os níveis de hierarquia estabelecidos quando o usuário é cadastrado. Os usuários podem ser administradores, coordenadores ou membros de um projeto.

O administrador do banco de dados (DBA) possui acesso irrestrito. Ele é a autoridade principal para o gerenciamento de um sistema de banco de dados. As responsabilidades do DBA incluem a concessão de privilégios a usuários que precisam utilizar o sistema e a classificação de usuários e dados de acordo com a política da organização. O DBA possui uma conta de DBA no SGBD, às vezes chamada conta de sistema ou de superusuário, que habilita as capacidades que não estão disponíveis para as contas e usuários comuns do banco de dados. Os comandos de privilégios do DBA incluem comandos para conceder e revogar

privilégios para contas individuais de usuários, e comandos para a realização dos seguintes tipos de ações:

- Criação de contas: Cria uma nova conta e senha para um usuário ou um grupo de usuários para habilitar o acesso ao SGBD;
- Concessão e Revogação de privilégios: Permite que o DBA conceda ou revogue (cancele) certos privilégios as contas de usuários previamente cadastrados;
- Atribuição de nível de segurança: Consiste em atribuir as contas de usuários ao nível de classificação de segurança adequado (ELMASRI, 2005).

Ao acessar o DIVERGENOMEdb como administrador é possivel utilizar este sistema de controle de privilégios: - a criação de contas, criação de projetos, revogação e concessão de privilégios,, através dos links “Register New User”, “Register New Project” e “Manager User Privileges” (Figura 3).

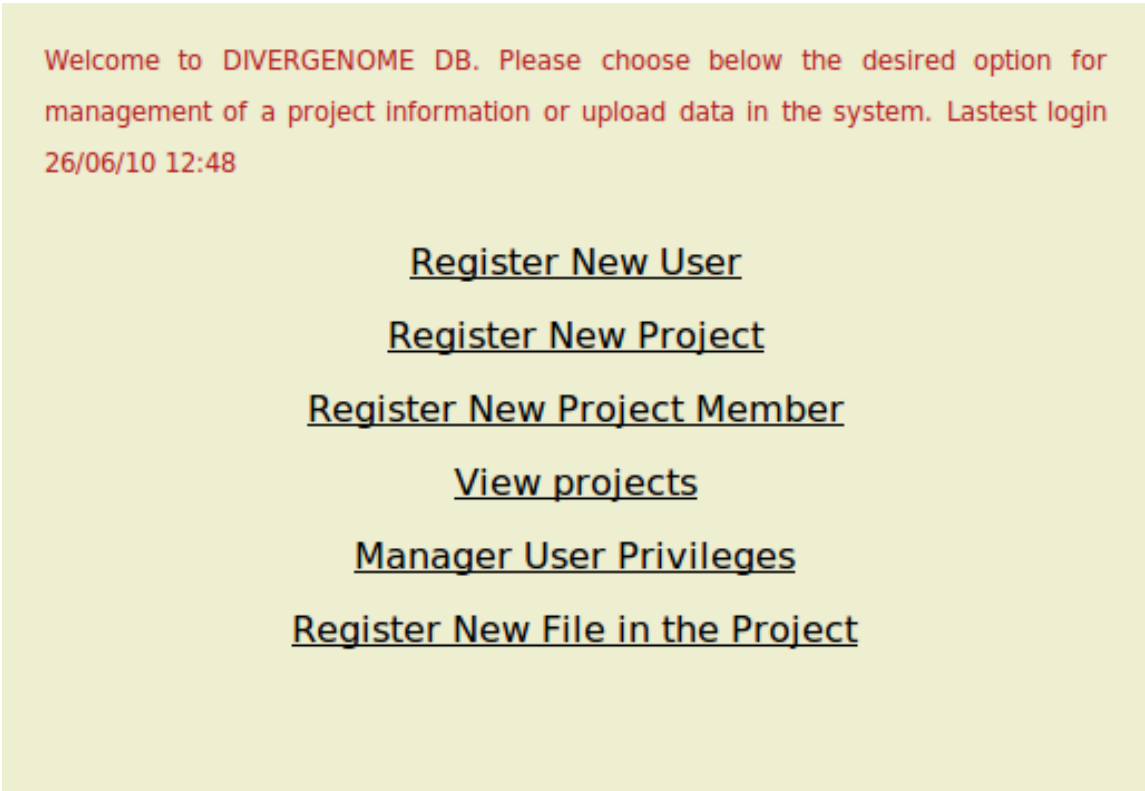


FIGURA 3: Tela principal de acesso ao sistema de administração disponibilizada ao administrador do DIVERGENOMEdb. Para os coordenadores são disponibilizados os ícones: *register new project, register new project member, view project e register new file in project.* Para

membros de projetos cadastrados é permitida apenas a visualização de projetos. A opção de cadastrar usuário é usada para coordenadores que desejam adicionar membros ao projeto. Posteriormente será realizado o controle de dados privados, e esses usuários terão acesso aos dados inseridos para um projeto específico em que estão cadastrados.

Através do controle de privilégios, as operações executadas pelo usuário ao acessar o banco podem ser controladas; assim, possuir uma conta não necessariamente habilita o usuário a utilizar todas as funcionalidades oferecidas pelo sistema. Para controlar o acesso foi desenvolvido um *script* que verifica esses níveis de permissão do usuário e formata a tela de apresentação conforme as atividades permitidas pelo nível atribuído ao mesmo (Tabela 2).

TABELA 2
Níveis de acesso em relação aos privilégios dos usuários cadastrados

	Administrador	Coordenador	Membro de projeto
Register New User	X		
Register New Project	X	X	
Register New Project Member	X	X	
View Projects	X	X	X
Manager User Privileges	X		
Register New File in the Project	X	X	X

3.4.2 Envio das informações

As informações de um formulário podem ser enviadas através dos métodos GET ou POST. O método GET é um método rápido de envio de informações, porém menos seguro, pois se utiliza da barra de endereços do navegador para enviar as informações para o servidor, sendo assim as informações ficam explicitamente visíveis na barra de endereços. Além disso, o método GET pode ser utilizado para enviar pequenas quantidades de informação, visto que possui limite de capacidade de envio de 1024 caracteres, sendo que o excedente pode ser perdido durante a transmissão.

O sistema de envio de dados desenvolvido para o DIVERGENOMEdb usando php utiliza o método POST que representa uma alternativa ao método GET para as questões de capacidade e segurança. Nesse método uma conexão paralela é aberta e os dados são passados por ela. As informações são enviadas de forma mais segura, possibilitando o uso de criptografia e outros recursos de segurança. Não há um limite para a quantidade de informações que pode ser enviada através deste método. Este método é usado quando queremos enviar dados a serem gravados em um banco de dados ou uma pesquisa cujos dados sejam grandes o suficiente para não caber na URL da página.

3.4.3 Tipos de buscas

A definição e a representação de consultas complexas são extremamente importantes para o biólogo. Por isso, os sistemas de informações de dados biológicos precisam dar suporte a consultas complexas. Sem nenhum conhecimento da estrutura de dados, os usuários comuns não podem construir por conta própria uma consulta complexa através dos conjuntos de consultas. Conforme mencionado previamente, muitos sistemas fornecem *templates* (modelos) predefinidos de consultas (ELMASRI, 2005).

Foram desenvolvidos *scripts* php com a finalidade de realizar consultas no DIVEGENOMEdb. Podem-se realizar dois tipos de consultas: (a) buscas simples em tabelas únicas, (b) buscas avançadas relacionando duas, três ou quatro tabelas. Em cada tabela de busca há campos de filtro e campos de seleção.

As consultas simples nas tabelas quantitative_variable, qualitative_variable e tissue não permitem filtragem, por isso suas buscas são direcionadas diretamente para os *scripts* de resultado e todos os campos presentes no banco de dados são apresentados.

Para as buscas avançadas são possíveis vários tipos de combinações entre as tabelas do DIVERGENOMEdb:

(a) Buscas relacionando campos de duas tabelas:

- Population e individual;

- Individual e Quantitative variable;
- Individual e Qualitative variable;
- Individual e Sample;
- Sample e Tissue;
- Polymorphism e Genotype;
- Sample e Population;
- Polymorphism e Reference Gene;
- Polymorphism e Reference Sequence.

(b) Buscas relacionando campos de três tabelas:

- Individual, polymorphism e genotype;
- Population, individual e variable;
- Population, individual e sample.

(c). Buscas relacionando campos de quatro tabelas:

- Population, individual, polymorphism e genotype

3.4.3.1 Especificações gerais dos scripts de busca

Para buscas simples ou avançadas há o direcionamento para um *script* específico de interface em HTML. Neles a programação PHP foi desenvolvida em um formulário que armazena todos os campos de filtro e busca selecionados e então a página é encaminhada para outro arquivo em que será feita a montagem do SELECT conforme as variáveis recebidas pelo método POST. O PHP abre uma conexão com o servidor do MySQL e envia a consulta apropriada. O servidor do MySQL recebe a consulta de banco de dados, processa e envia os resultados de volta para o mecanismo de PHP que então termina de executar o *script* envolvendo a formatação da consulta que retorna os dados ao usuário em uma tabela.

3.4.3.2 Especificação da construção do SELECT

A SQL possui um comando básico para a recuperação de informações de um banco de dados: o comando SELECT. A consulta seleciona as tuplas da entidade que satisfazem a condição da cláusula WHERE, então projeta o resultado dos atributos solicitados na cláusula SELECT. Em geral, pode ser especificado qualquer número de condições de seleção e junção em uma única consulta SQL (ELMASRI, 2005).

Em SQL, um mesmo nome pode ser usado para dois (ou mais) atributos (campos) desde que esses atributos estejam em relações diferentes (ou tabelas diferentes). Se esse é o caso e se uma consulta se refere a dois ou mais atributos com o mesmo nome, é preciso qualificar o nome do atributo com o nome da relação, de modo a prevenir ambigüidade. Isso é feito por meio da prefixação do nome da relação ao nome do atributo, separados por um ponto (ELMASRI, 2005).

Como exemplo citamos os atributos “Kind” presentes nas entidades *Polymorphism* e *Assay Protocol* (descritos no item 3.1.2.1). Apesar de possuírem o mesmo nome, possuem significado diferentes e estão em tabelas diferentes. Desta forma é necessário qualificar o nome do atributo com o nome da entidade ou relação. Para a entidade *Polymorphism* definiu-se o termo “pl” como prefixação e para a entidade *Assay Protocol* definiu-se o termo “a” como prefixação. Então nas construções das cláusulas quando se usa o termo a.kind, este se refere ao atributo kind da entidade *Assay Protocol*, e quando se usa o termo pl.kind, este se refere ao atributo kind da entidade *Polymorphism*. O mesmo princípio foi utilizado para a construção das buscas avançadas que relacionam duas ou mais entidades.

3.4.3.3 Validação dos dados em buscas por tabelas únicas

Nas buscas em tabela única foi desenvolvida uma função em Java Script que valida e averigua se pelo menos um campo e um filtro foram selecionados para montagem posterior da cláusula de instrução SELECT. Caso a condição pré-estabelecida não tenha sido atendida, a mensagem de alerta é acionada.

Os *scripts* para retorno dos resultados da busca montam a instrução SELECT conforme os campos selecionados na página anterior de consulta disponibilizada na interface.

3.5 Testes de verificação

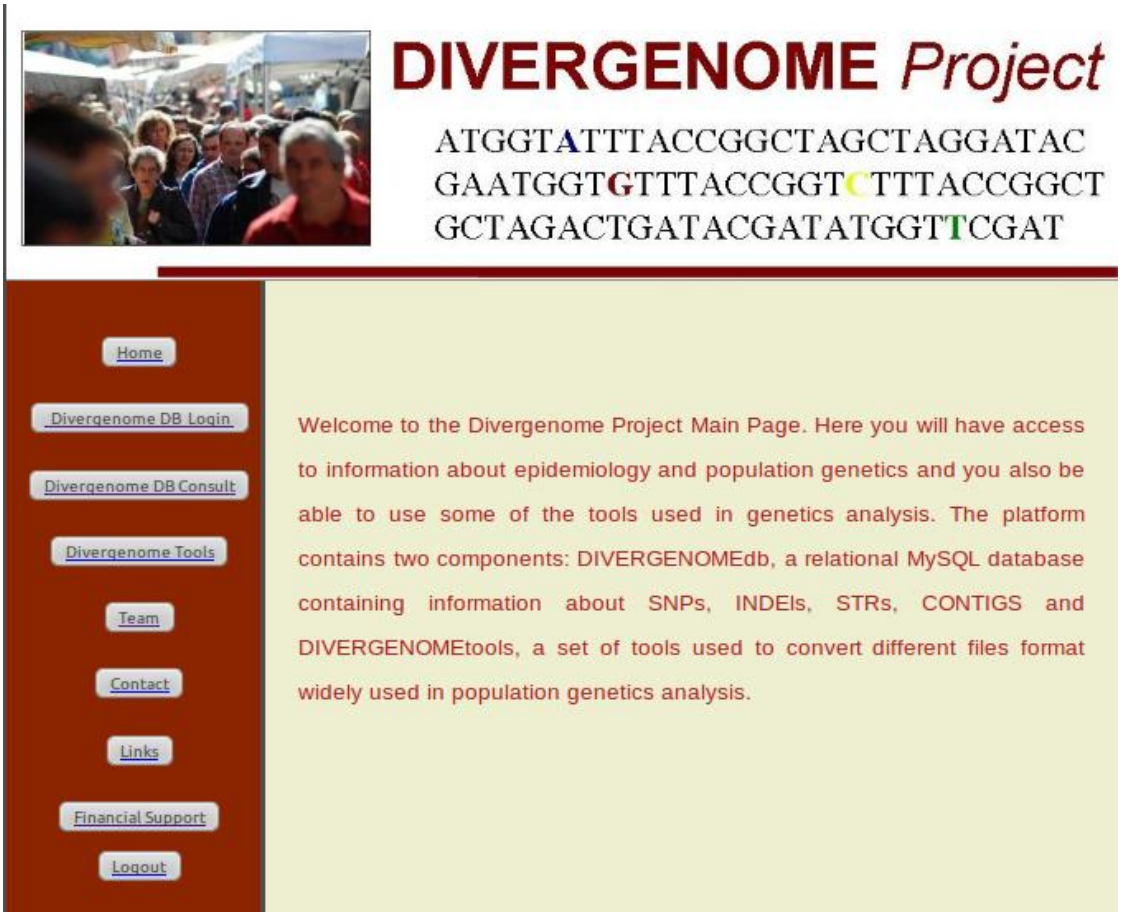
A qualidade é um aspecto que deve ser tratado simultaneamente ao processo de desenvolvimento do *software*. Abordagens complementares, como, introduzir em cada fase ou etapa do processo de desenvolvimento de software, uma atividade de verificação, podem nortear a estruturação das atividades de garantia de qualidade de software. Dentre as atividades de garantia de qualidade de software estão as de validação e teste, com o objetivo de minimizar a ocorrência de erros e riscos associados. O objetivo da validação é assegurar que o software que está sendo desenvolvido é o *software* correto de acordo como os requisitos do usuário. A atividade de teste é considerada um elemento crítico para a garantia de qualidade de *software*. Consiste na análise dinâmica do mesmo, ou seja, na execução do produto de *software* com o objetivo de verificar a presença de defeitos no produto e aumentar a confiança de que o produto esteja correto (ROCHA, 2001).

Buscou-se durante o desenvolvimento da interface DIVERGENOMEdb, verificar os possíveis usos do banco de dados e as possíveis buscas que pudessem ser realizadas (validação). O sistema foi submetido a testes durante o decorrer da construção da interface com dados fictícios em algumas tabelas, para verificação de sua correta funcionalidade. Ao final foi realizado o teste com dados reais trabalhados por G. Souza em sua tese de mestrado (Identificação de genes com alta diferenciação entre populações humanas: inferências evolutivas e aplicações biomédicas, 2010).

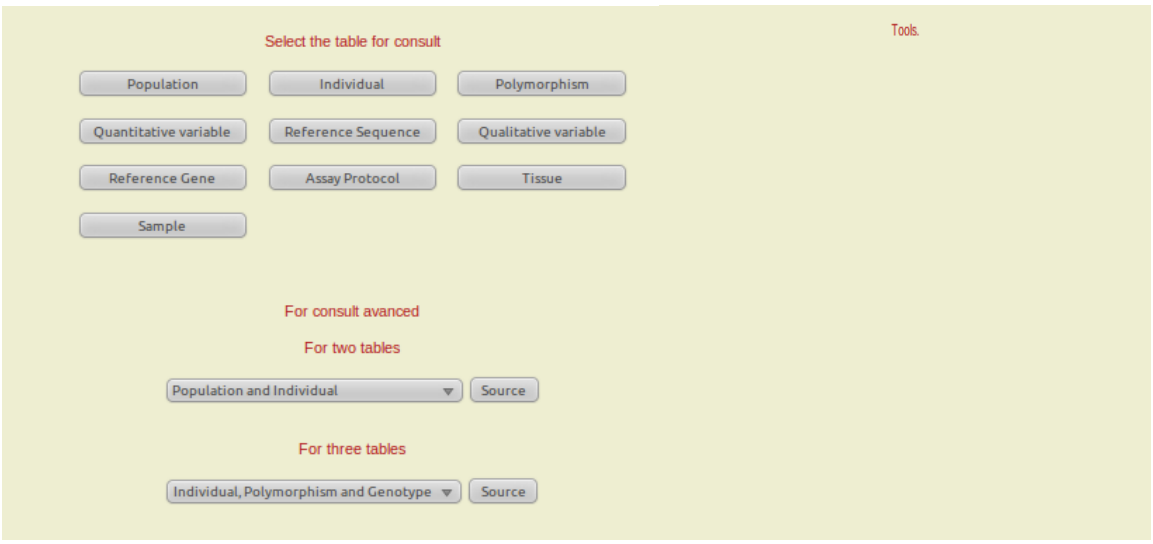
Para auxiliar na manutenção da interface *web* para acesso ao DIVERGENOMEdb e sua utilização foi desenvolvido além da documentação técnica do sistema o manual de usuário ou documentação do usuário.

4. RESULTADO E DISCUSSÃO

A interface *Web* para o DIVERGENOME Project foi desenvolvida seguindo a disponibilização dos recursos para dois ambientes: DIVERGENOMEdb (para acesso e administração do banco de dados) e DIVERGENOMEtools (acesso a programas de conversão de formatos de dados, em implementação). A Figura 4 apresenta as telas de interface para a disponibilização destes recursos.



A – Tela Inicial



B - Tela de consulta DIVERGENOMEdb **C - Link para DIVERGENOME Tools**

FIGURA 4: Interface *web* de acesso ao sistema DIVERGENOME. A. Tela inicial de apresentação do DIVERENOME Project. O menu contém o link para acesso a informações gerais, bem como um link para o DIVERGENOMEdb (B) e outro para o DIVERGENOMEtools (C).

A partir da tela inicial (Figura 4A) é possível ter acesso ao DIVERGENOMEdb (Figura 4B) e realizar consultas simples, por tabela única, ou consultas avançadas por duas, três ou quatro tabelas (ícone DIVERGENOME DB Consult). Essa tela permite também a inserção de dados pelos coordenadores, após prévia aprovação pelo administrador do banco de dados (ícone DIVERGENOME DB Login). Além disso, nessa tela visualiza-se o link para o DIVERGENOME Tools (Figura 4C), que permite a conversão dos dados para o formato específico exigido por diversos programas usados em pesquisas genéticas. O DIVERGENOME Tools está em implementação no Laboratório de Diversidade Genética Humana pelo aluno de doutorado Wagner Magalhães, a pós-doutora Maira Rodrigues e os alunos de Iniciação Científica Bruno Araujo e Allan Sene.

O foco do presente trabalho foi o desenvolvimento da interface para consulta e inserção de informações no banco de dados (DIVERGENOMEdb). Informações gerais podem ser visualizadas sem cadastro prévio, assim como a realização de consultas a dados públicos cadastrados no mesmo. O controle para dados privados, que serão visualizados apenas pelos coordenadores (pesquisadores) cujo projeto seja de sua autoria e pelos membros cadastrados pelo pesquisador para um projeto específico está em desenvolvimento.

4.1 Visualização dos recursos gerais

Os recursos gerais disponíveis podem ser visualizados na página de navegação do DIVERGENOMEdb apresentado na Figura 5.

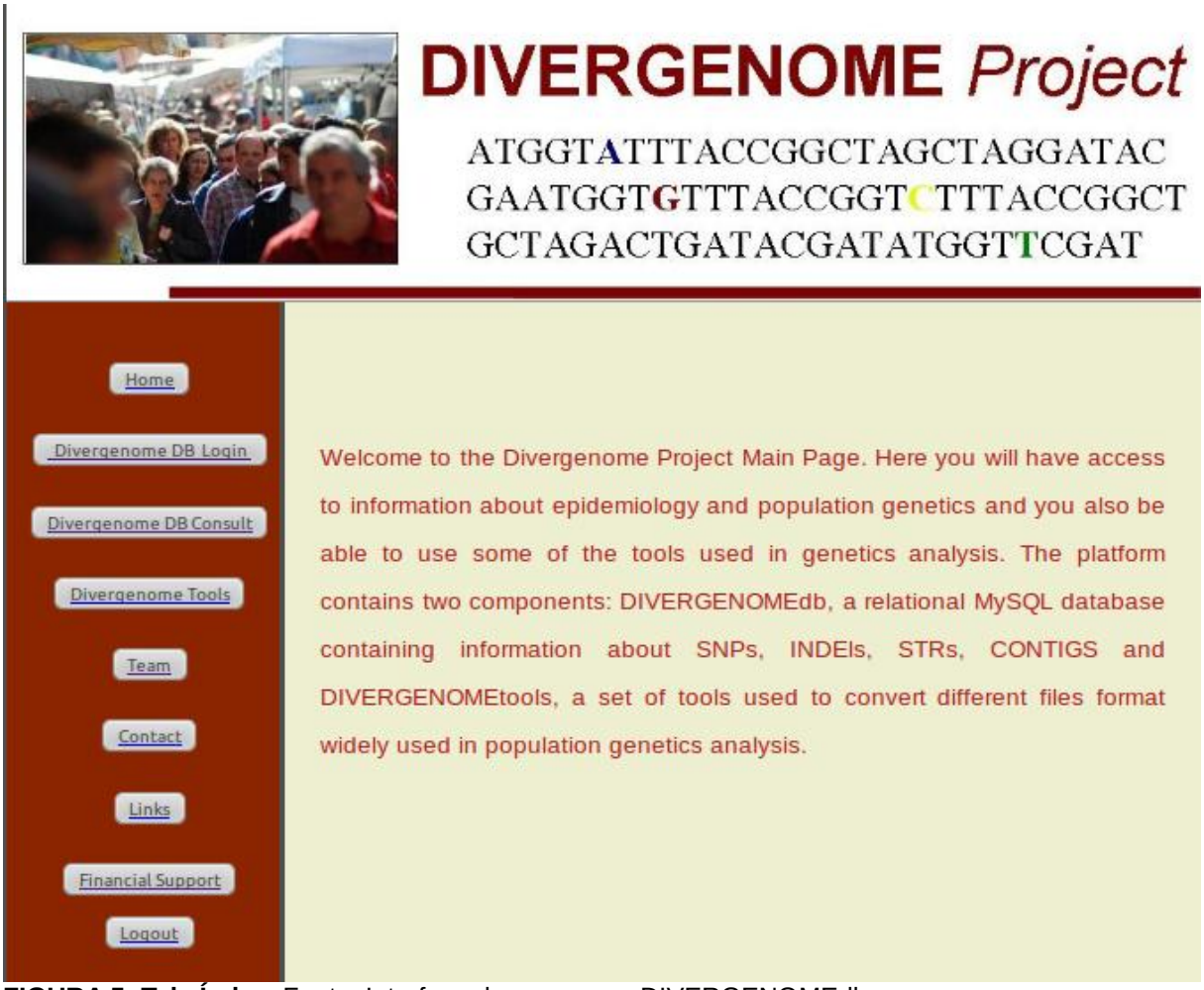


FIGURA 5: Tela Índex. Fonte: Interface de acesso ao DIVERGENOMEdb

Na Tabela 3 é apresentada uma breve descrição dos ícones disponíveis na tela inicial.

TABELA 3

Botões contidos no menu e descrição dos campos

Botão	Descrição
Home	▪Página inicial do DIVERGENOME. Contém uma breve descrição do projeto.
DIVERGENOMEdb Login	▪Acesso ao banco de dados, e cadastro de usuário.
DIVERGENOMEdb Consult	▪Consulta a dados públicos presentes no DIVERGENOMEdb.
DIVERGENOME Tools	▪Acesso as ferramentas de conversão de formato de dados biológicos.
Team	▪Lista de desenvolvedores do projeto, contendo breve descrição de suas formações.
Contact	▪Disponibiliza o email de contato com o Laboratório de Diversidade Genética Humana (LDGH) no Instituto de Ciências Biológicas – UFMG
Links	▪Links voltados ao estudo de genética de populações e epidemiologia genética.
Financial Support	▪Listas de financiadores.
Logout	▪Saída da conexão.

Na Figura 6 é apresentado um diagrama dos ícones gerais, bem como o fluxo destes ícones dentro do programa.

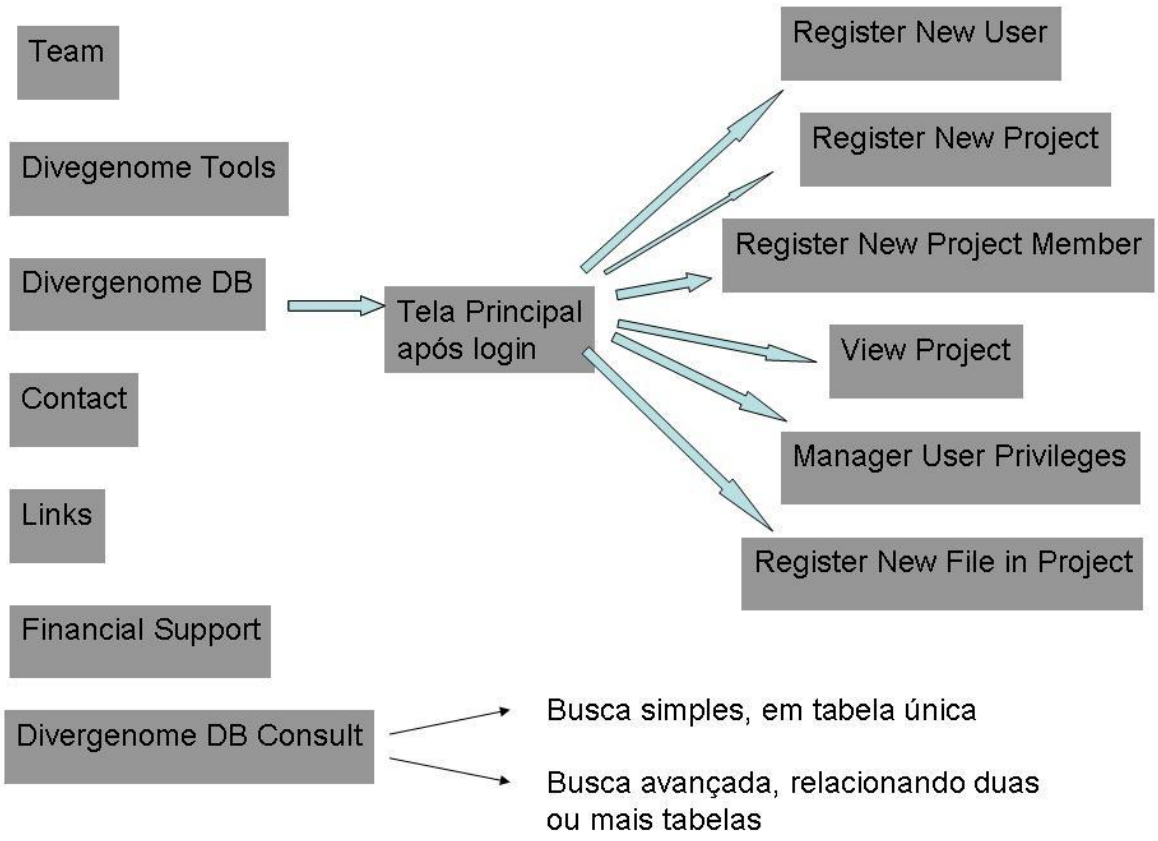


FIGURA 6: Diagrama dos ícones dentro do programa. Através do login obtido pelo acesso ao ícone DIVERGENOME DB se tem disponibilizado os recursos aos usuários cadastrados que é variável conforme os níveis de permissão de cada um (administradores, coordenadores ou membros de projeto). O ícone DIVERENOME DB Consult, de acesso livre a dados públicos, permite a busca por tabela única ou por mais de uma tabela (busca avançada), usando assim o relacionamento entre as tabelas.

Os ícones da esquerda são disponibilizados na tela principal do DIVERGENOME PROJECT, sem nenhuma restrição de acesso. O ícone Divergenome DB Consult permite a realização de buscas ao banco de dados. Atualmente todos os dados inseridos são considerados como públicos. Posteriormente será realizado o controle de dados privados. O ícone Divergenome DB permite, após o *login* do usuário, o acesso aos ícones disponibilizados à direita e que irão variar conforme o privilégio do usuário logado. Como já apresentado anteriormente, para os administradores são disponibilizados todos os recursos apresentados na Figura 7 (ícones da direita da Figura 6). Para os demais (coordenadores e membros de projeto), a disponibilização dos recursos é variável conforme os níveis de acesso (vide Tabela 2).

FIGURA 7: Tela de disponibilização dos recursos para os administradores
Fonte: Interface desenvolvida para DIVERGENOMEdb

4.1.1 Descrição breve dos recursos de cada ícone

4.1.1.1 A - Register New User

O cadastro do usuário (Figura 8B) pode ser feito diretamente pelo administrador através do *link* “Register New User” (Figura 7), ou pelo novo usuário através da tela de *login* de acesso (Figura 8A), em que há um *link* para cadastro de usuário (Register Now). Neste caso o cadastro necessita de uma posterior aprovação pelo administrador.

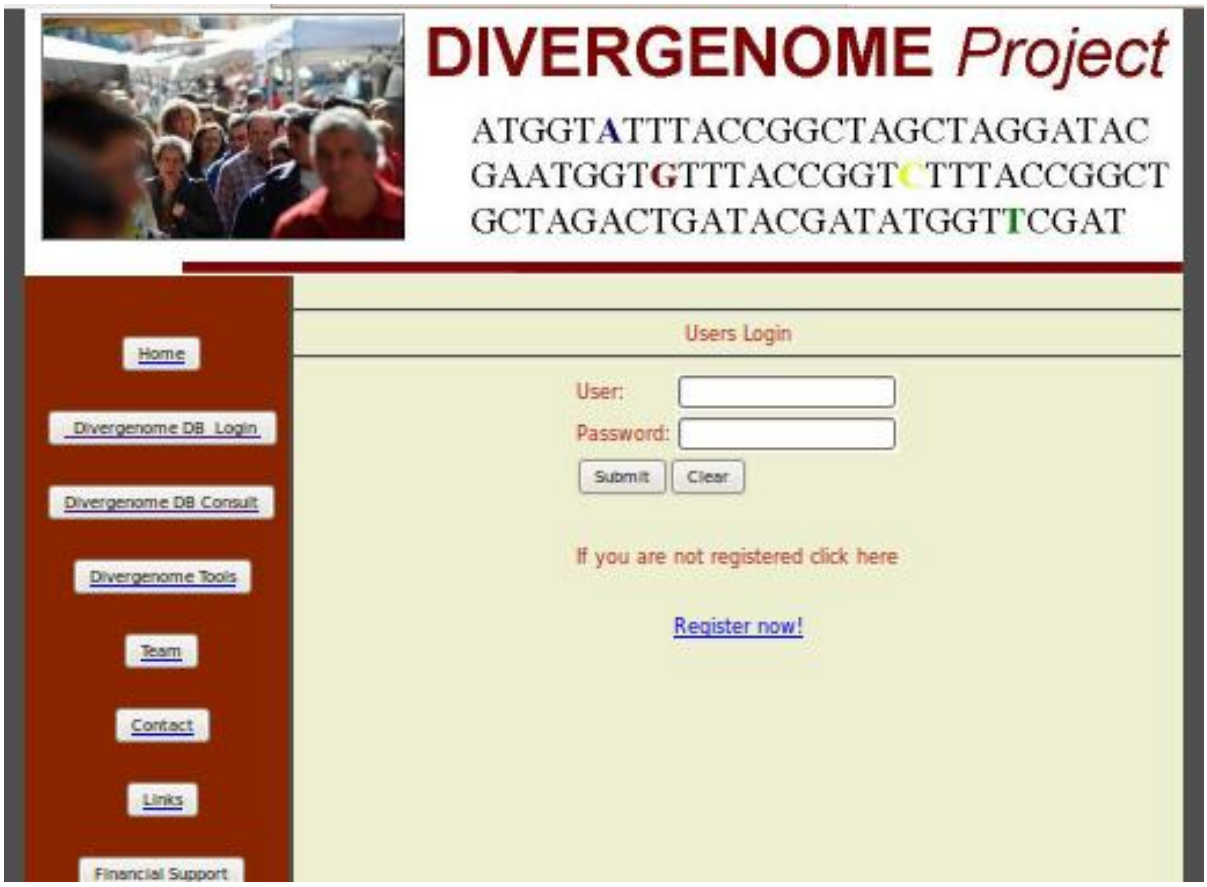


Figura 8A – Tela de *login*

Caso o usuário ainda não seja cadastrado, pode-se realizar seu registro prévio através do *link* "Register now" disponível na tela de *login*.

FIGURA 8B: Tela de cadastro do usuário

4.1.1.2 B – Register New Project

Através desta opção é possível cadastrar um novo projeto ao DIVERGENOMEdb. É apresentando um formulário para preenchimento dos campos referente ao projeto. Um destes campos se refere ao status do projeto como público ou privado. Ao se cadastrar um projeto como público todos os dados inseridos para esse projeto serão disponibilizados abertamente ao público. Já para projetos privados, posteriormente será realizado o controle de acesso aos dados inseridos para esse projeto, em que apenas o coordenador que registrou o projeto e os membros cadastrados por esse coordenador terão acesso às informações pertinentes ao projeto especificado. Atualmente, todos os dados inseridos no banco de dados são tratados como públicos.

A Figura 9 se refere à tela para cadastro de novo projeto.

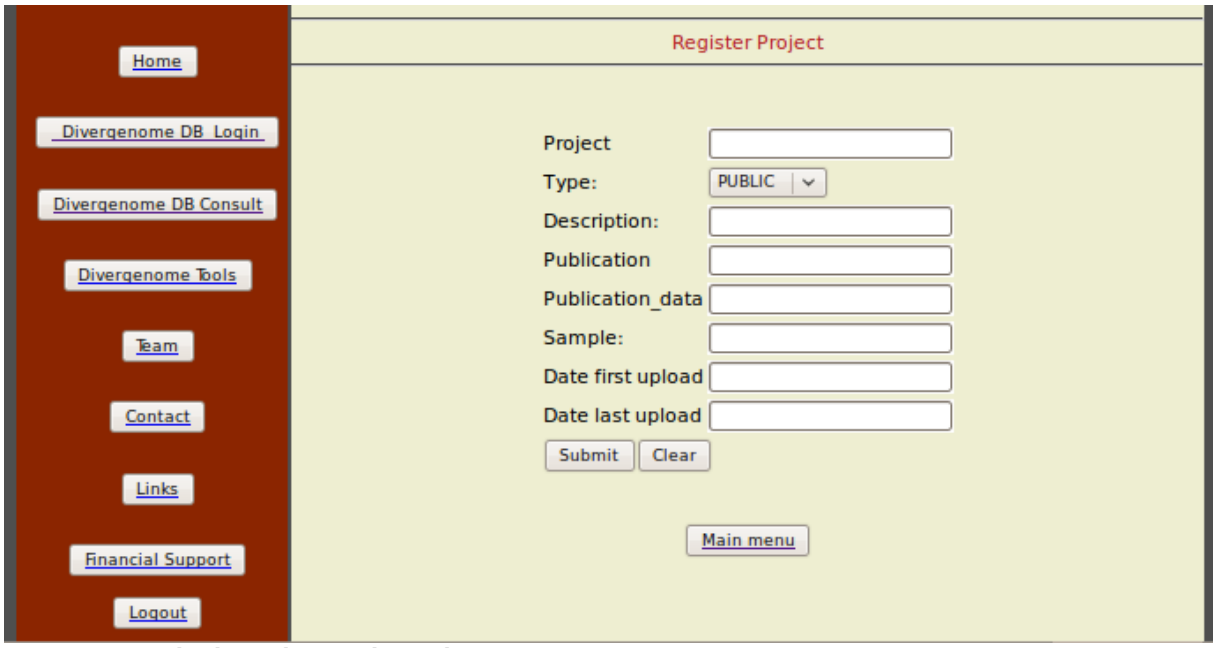


FIGURA 9: Tela de cadastro de projeto

4.1.1.3 C – Manager User Privileges

O ícone “Manager User Privileges” permite ao administrador controlar os acessos dos usuários previamente cadastrados e que estão aguardando aprovação. É possível alterar o status do usuário, bem como o privilégio do mesmo. A Figura 10A se refere à tela de controle do usuário pelo administrador e as Figuras 10B e

10C se referem às mensagens disponibilizadas aos usuários quando da tentativa de logar ao sistema caso o acesso ainda não tenha sido aprovado pelo administrador (Figura 10B) ou o usuário tenha tido seu acesso negado pelo administrador (Figura 10C).

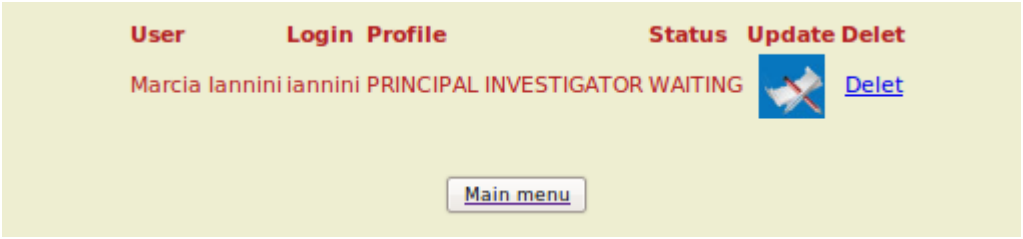


FIGURA 10A: Tela de controle do usuário pelo administrador

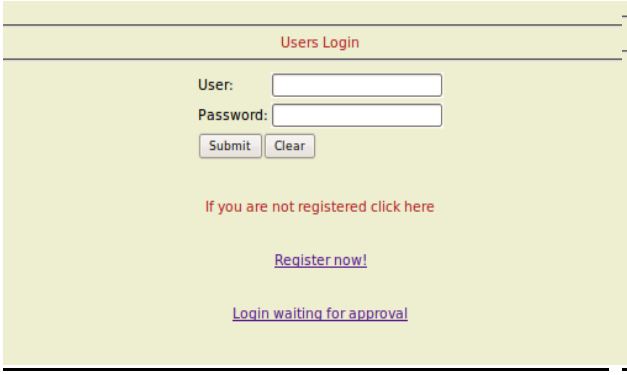


Figura 10B: Usuário aguardando aprovação pelo administrador

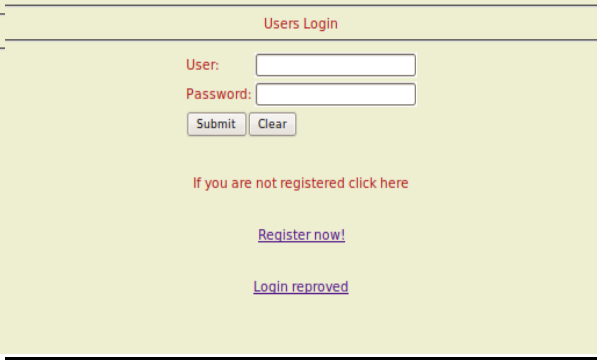


Figura 10C: Usuário reprovado pelo administrador

4.1.1.4 D – Register New Project Member

Essa opção permite aos administradores e coordenadores adicionarem um novo membro ao seu projeto anteriormente cadastrado. A função já é realizada, porém ainda sem eficácia, uma vez que o controle de dados privados não está operante. A Figura 11 apresenta a tela para inserção de novo membro no projeto.

FIGURA 11: Tela de cadastro do novo membro do projeto

4.1.1.5 E – View Project

Obtém-se visualização dos dados gerais cadastrados no projeto. É permitido acesso aos dados de projetos apenas de autoria do coordenador. A visualização de projetos por administradores e coordenadores já está em funcionamento. A disponibilização desse ícone para usuários ainda não está em operação, mas justifica-se pelo funcionamento posterior de controle de dados privados, em que após o cadastro prévio do usuário como membro de um projeto será possível a visualização dos campos de especificação desse projeto, bem como de seus dados inseridos.

4.1.1.6 F – Register New File in the Project

Através desse ícone é possível a inserção de dados no database pelos coordenadores de projetos e administradores do banco de dados. Para as entidades *genotype*, *haplotype*, *haplotype_evaluation* e *sample* que possuem o atributo project_id_project como chave estrangeira que relaciona ao atributo id_project da entidade *Project*, será possível posteriormente o controle dos dados como públicos ou privados. Sendo assim, para projetos privados, as informações inseridas pelo pesquisador poderão ser visualizadas apenas pelo próprio pesquisador ou pelos

membros do projeto cadastrados pelo pesquisador. Anteriormente no tópico 3.1.2 (Entidades relacionadas a dados genéticos) foram detalhadas as chaves estrangeiras específicas de cada tabela. No tópico 4.3 (Inserção de dados no DIVERGENOMEdb) há a descrição detalhada do recurso de recuperação de chaves estrangeiras na inserção de arquivo de dados.

4.2 Buscas e Disponibilização dos resultados

Em todas as buscas (simples ou avançadas), a tela disponibilizada ao usuário é disposta em sua parte superior por campos de filtro, e a parte inferior por *checkbox* (Figura 12). O *checkbox* é responsável pelos itens (colunas da tabela) a serem listados dentro da instrução SELECT. Os campos de filtro são os responsáveis pela recuperação de dados com critérios específicos, ou seja, acessam um subconjunto de linhas em uma tabela, especificando os critérios de seleção através da cláusula WHERE.

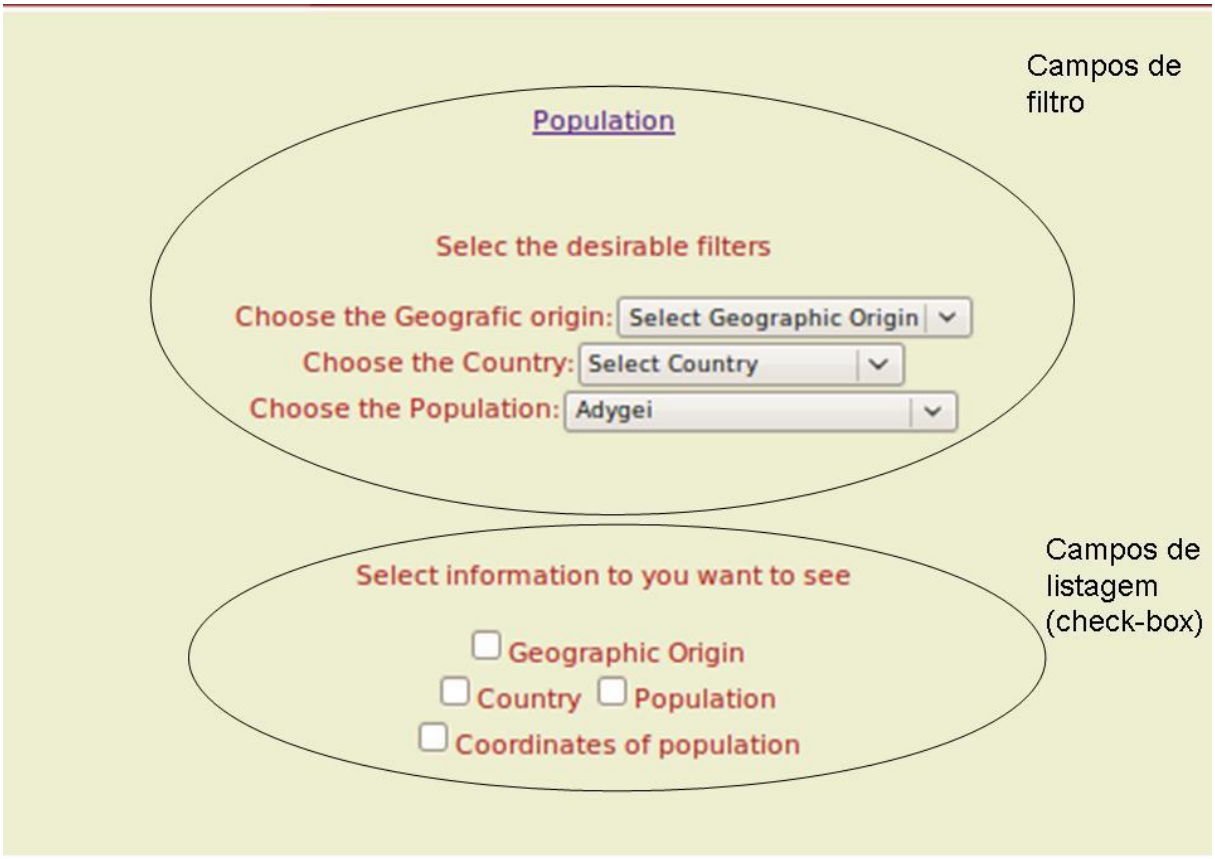


Figura 12: Campos de filtro e check-box. Tela visualizando os campos de filtros e os campos a serem listados, para montagem do select.

As buscas permitem gerar um arquivo de saída no formato texto. Tanto os arquivos de entrada de dados para DIVERGENOMEdb quanto os arquivos de saída obedecem o formato padrão de banco de dados – onde os mesmos estão dispostos linearmente e os campos separados um do outro por tabulação, como por exemplo a disposição de dados apresentada na Figura 13. Primeiramente o resultado da busca é disponibilizado na tela como uma tabela (veja Figura 18). Posteriormente é possível gerar um arquivo (registro.txt) com os valores retornados da busca e também é possível gerar a linha de comando do select usado na busca em um arquivo de texto (select.txt).

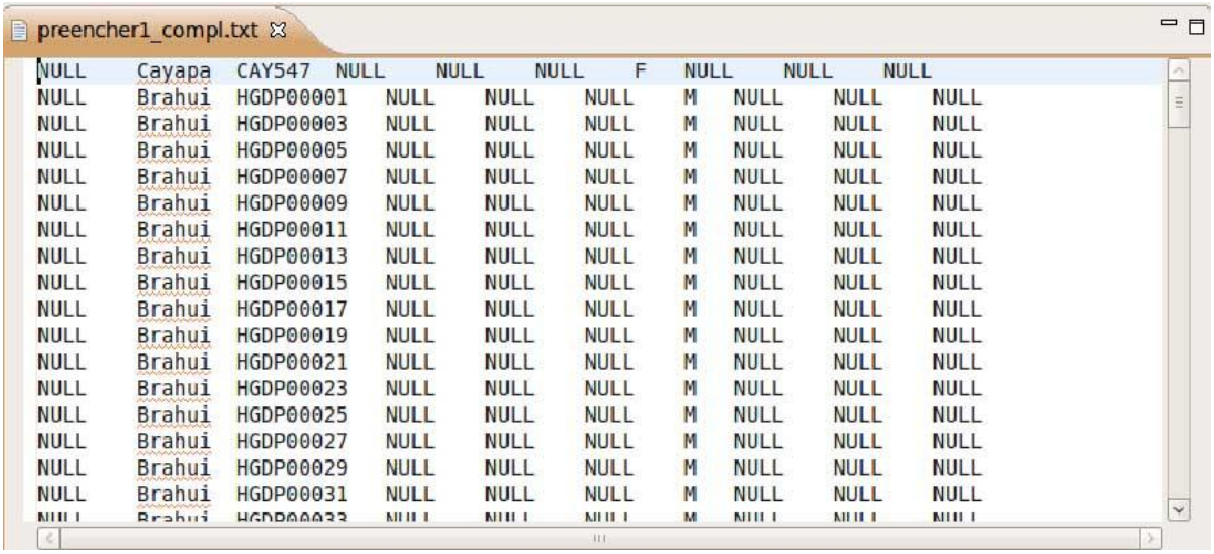


FIGURA 13: Arquivo de inserção da entidade *individual*
Fonte: Dados obtidos do projeto de mestrado de Souza, 2010

A partir do resultado obtido no arquivo registro.txt é possível utilizar-se de ferramentas do DIVERGENOME Tools, que permite a conversão dos dados para formatos específicos requeridos por diferentes softwares de genética de populções (PHASE, DNAsp, Structure), genética médica (Haplopainter) e pacotes estatísticos de uso comum (MAGALHÂES 2010, manuscrito em preparação).

4.3 Inserção de dados no DIVERGENOMEdb

Muitos recursos de Bioinformática não apenas oferecem a recuperação de informações, mas também facilitam o processamento subsequente das entradas selecionadas. Essas compreendem resultados experimentais brutos e informações suplementares, ou anotações (LESK, 2008).

Os pedidos de informações às fontes de Bioinformática devem iniciar um ciclo de realimentação positiva que envolve três componentes: a utilidade científica, o uso da comunidade, e contribuições da comunidade. Na ausência deste ciclo de realimentação positiva, estas aplicações não irão atingir a massa crítica de utilizadores e nem atingirão os objetivos para os quais se destinam. A idéia primordial é de que uma grande comunidade de usuários possa coletivamente e colaborativamente sintetizar o conhecimento. É a grande população de usuários que faz individualmente pequenas (mas coletivamente grandes) contribuições de

conteúdo. Esses esforços têm sido aplicados para objetivos científicos, e especificamente com o objetivo de anotação de genes de genomas completos. Exemplos recentes incluem o WikiProteins e WikiGenes (WU, 2009).

A acessibilidade via *web* de fonte de dados de Bioinformática é crítica para o sucesso das pesquisas científicas. Uma das metodologias adotadas é a realização da integração dos dados em um repositório integrado. Outro método de integração seria o modelo da federação, em que a integração dos dados é feita quando necessário através do acesso as fontes de dados originais (GAASTERLAND, 2005).

No DIVERGENOME Project buscou-se criar um repositório integrado. Através da interface do DIVERGENOMEdb é possível a inserção de arquivos txt no formato padrão de banco de dados. Essas inserções podem ser realizadas por coordenadores previamente cadastrados ou por administradores do banco de dados.

Para as entidades que possuem atributos relacionados em várias tabelas (chaves estrangeiras) é possível recuperar automaticamente os identificadores primários nas tabelas de origem. A Figura 13 mostra um exemplo de arquivo a ser inserido no banco de dados na entidade *individual* e na Figura 14 é apresentada a estrutura desta entidade no banco de dados.

Visualizar | Estrutura | SQL | Procurar | Inserir | Exportar | Importar | Operações | Limpar | Eliminar

	Campo	Tipo	Collation	Atributos	Nulo	Padrão	Extra	Ação
☐	id_individual	int(10)		UNSIGNED	Não	*Nenhum*	auto_increment	
☐	population_id_population	int(10)		UNSIGNED	Não	*Nenhum*		
☐	individual_code	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	family_id	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	father_id	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	mother_id	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	sex	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	aff_status	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	live_status	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	info_1	varchar(45)	latin1_swedish_ci		Sim	*NULL*		

FIGURA 14: Estrutura da entidade *Individual* no DIVERGENOMEdb

No arquivo da Figura 13, a primeira coluna que recebe valores nulos se refere ao código de identificação dos indivíduos (id-individual) que é definido como auto-increment no database (receberá valores sequenciais quando da inserção) e a segunda coluna se refere ao atributo population-id-population (chave estrangeira

com a tabela *population*). O script desenvolvido compara os valores dos campos da segunda coluna do arquivo (Cayapa e Brahui) com as populações cadastradas no DIVERGENOMEdb, e retorna, caso encontre a população já cadastrada no banco de dados, o respectivo valor de identificação da população, ou seja, o id_population. A inserção no banco de dados para este campo na entidade *individual* será do id_population (id de identificação da população na tabela *population*). Caso existissem nesta entidade outros atributos que fossem chaves estrangeiras de outras tabelas, o procedimento de recuperação do id seria o mesmo. O resultado da inserção da tabela *individual* com a devida recuperação dos ids das populações está apresentado na Figura 15.

id_individual	population_id_population	individual_code	family_id	father_id	mother_id	sex	aff_status	live_status	info_1
1	18	CAY547	NULL	NULL	NULL	F	NULL	NULL	NULL
2	15	HGDP00001	NULL	NULL	NULL	M	NULL	NULL	NULL
3	15	HGDP00003	NULL	NULL	NULL	M	NULL	NULL	NULL
4	15	HGDP00005	NULL	NULL	NULL	M	NULL	NULL	NULL
5	15	HGDP00007	NULL	NULL	NULL	M	NULL	NULL	NULL
6	15	HGDP00009	NULL	NULL	NULL	M	NULL	NULL	NULL
7	15	HGDP00011	NULL	NULL	NULL	M	NULL	NULL	NULL
8	15	HGDP00013	NULL	NULL	NULL	M	NULL	NULL	NULL
9	15	HGDP00015	NULL	NULL	NULL	M	NULL	NULL	NULL
10	15	HGDP00017	NULL	NULL	NULL	M	NULL	NULL	NULL
11	15	HGDP00019	NULL	NULL	NULL	M	NULL	NULL	NULL
12	15	HGDP00021	NULL	NULL	NULL	M	NULL	NULL	NULL
13	15	HGDP00023	NULL	NULL	NULL	M	NULL	NULL	NULL
14	15	HGDP00025	NULL	NULL	NULL	M	NULL	NULL	NULL
15	15	HGDP00027	NULL	NULL	NULL	M	NULL	NULL	NULL
16	15	HGDP00029	NULL	NULL	NULL	M	NULL	NULL	NULL

FIGURA 15: Tela de visualização da entidade *Individual* inserida no database

Repare que para o indivíduo 1 (primeira linha do arquivo apresentado na Figura 13), o campo population_id_population (segunda coluna Figura 15) recebeu o valor 18 que se refere ao código de identificação da população Cayapa no banco de dados, e para os demais indivíduos (segunda linha em diante dos dados apresentados na Figura 13), o campo population_id_population recebeu o valor 15 que se refere ao código de identificação da população Brahui cadastrada no banco de dados.

É possível também para cada entidade que possui atributos relacionados a outras tabelas (chaves estrangeiras) a recuperação de ids primários nas entidades de origem, através da seleção do valor da variável disponibilizada em um menu drop-down (valores retornados conforme os dados disponíveis no banco e que foram anteriormente inseridos), como apresentado na Figura 16, em que permite a inserção de um arquivo de texto contendo atributos referente à entidade *individual*. Neste caso, o arquivo de texto para inserção apresentará no atributo correspondente a chave estrangeira (neste exemplo o atributo population_id_population) o valor NULL. O usuário deve selecionar a população (valores retornados do banco de dados) para a qual se deseja inserir os indivíduos (atributos presentes no arquivo texto). Veja tela para esse tipo de inserção na Figura 16.

FIGURA 16: Tela de inserção do arquivo texto no formato padrão de banco de dados para a tabela individual. Deve-se selecionar a população para qual se deseja inserir os indivíduos e escolher o arquivo texto contendo os dados dos indivíduos.

Neste caso, para todos os indivíduos do arquivo txt será inserido no banco de dados o código de identificação correspondente a população selecionada no menu drop-down (no nosso exemplo o id da população “Adygei”). Ou seja, uma população apenas para todos os indivíduos do arquivo inserido. Esse recurso difere do anteriormente citado, pois no exemplo anterior, o arquivo txt de inserção na entidade *individual*, contém o nome da população no atributo referente a chave estrangeira (population_id_population), sendo possível através de inserção única de arquivo, a recuperação automática de ids da chave estrangeira, ou seja, é possível inserir um único arquivo contendo indivíduos de diferentes populações.

Os recursos para recuperação automática de identificadores (ids), bem como para recuperação de identificadores de chaves estrangeiras através de seleção de valores retornados do banco de dados em menu drop-down, estão disponíveis para as entidades *genotype*, *polymorphism*, *individual* e *sample* (entidades que possuem atributos relacionados a outras entidades – chave estrangeira).

Outra utilidade da criação da interface se refere as facilidades de pesquisa no banco de dados. A construção da interface para procura em DIVERGENOMEdb requer a compreensão da estrutura do banco de dados, de forma que possa traduzir as pesquisas via *web* em comandos em SQL. Para atingir este objetivo é importante conhecer os tipos de procuras mais comuns dos usuários. A seguir são apresentadas algumas das aplicações práticas do DIVERGENOMEdb utilizando as facilidades da interface gráfica.

4.4 Aplicação em estudo de caso

Para exemplificar a utilização do sistema DIVERGENOME optou-se por apresentar alguns casos que possam demonstrar o funcionamento do sistema desenvolvido.

Para a aplicação e testes foram utilizados os dados trabalhados por Juliana Chevitarese e Giordano Bruno Soares Souza do Laboratório de Diversidade Genética Humana. O trabalho de caracterização genética das populações do CEPH-HGDP, SNP500Cancer e Nativo-Americanos foi realizado por Juliana Chevitarese (2009) em sua dissertação de mestrado. As mesmas populações e loci foram utilizados no estudo realizado por Giordano Souza e por isso, a descrição da estrutura genética desse conjunto de dados é equivalente para os dois estudos. Enquanto no primeiro trabalho o enfoque foi a descrição da estrutura genética populacional, no segundo o objetivo principal foi a caracterização da estrutura genética para cada loci. Os dados utilizados foram dados disponibilizados de genotipagem de 1442 SNPs em 411 genes (1421 SNPs do SNPCancerPanel da Plataforma de Genotipagem Illumina Golden Gate® e 21 SNPs adicionais) para 52 populações do CEPH – HGDP (Centre d'Etude du Polymorphisme Humain - Human Genome Diversity Cell Line Panel) (CANN, 1998; 2002; CAVALLI-SFORZA, 2005), 4 populações nativo-americanas do Peru e Equador (Quechua, San Martin, Cayapa e

Matsiguenga – dados não publicados) e 4 populações do painel de 102 indivíduos do SNP500Cancer (Afro-americanos, Euro descendentes residentes em Utah, Hispânicos e Asiáticos) (PACKER et al., 2004; 2006), perfazendo um total de 1198 indivíduos (SOARES-SOUZA, 2010). Estes dados foram obtidos em colaboração com o Dr. Stephen Chanock (NIH – NCI).

4.4.1 Diversidade genética na antropologia

As informações obtidas com os SNPs são de grande utilidade na antropologia, fornecendo dados sobre variações históricas no tamanho da população e padrões de migração (LESK, 2008).

Os graus de diversidade genética são interpretáveis em termos do tamanho da população fundadora. Fundadores são o conjunto original de indivíduos dos quais uma população inteira descende. Esses fundadores podem ser os colonizadores originais, ou simplesmente os sobreviventes de um evento de quase extinção (LESK, 2008).

O trabalho de caracterização genética das populações do CEPH-HGDP, SNP500Cancer e Nativo-Americanos foi realizado por Chevitarese (2009) em sua dissertação de mestrado. Para obtenção dos valores de F_{ST} e F_{IT} (refletem a estruturação entre as populações estudadas) é necessário inicialmente obter as tabelas contendo todos os indivíduos com os determinados polimorfismos de interesse, bem como os valores do genótipo para esses indivíduos, sendo fundamental a filtragem dessas buscas por população. Estas filtragens iniciais também seriam necessárias para a descrição da estrutura genética dos loci estudados através de índices clássicos de variabilidade genética: Equilíbrio de Hardy-Weinberg, Heterozigosidade Esperada e F_{IS} (Coeficiente de Endocruzamento).

Um dos passos iniciais para a realização dos cálculos desses índices, seria realizar as buscas iniciais que poderia ser feita através da interface gráfica de DIVERGENOMEdb. Um exemplo seria buscarmos todos os dados de genotipagem dos indivíduos para todos os SNPs cadastrados no banco de dados para a população Adygei. Para realizar tal busca seria necessário utilizar a linha de comando como a apresentada abaixo:

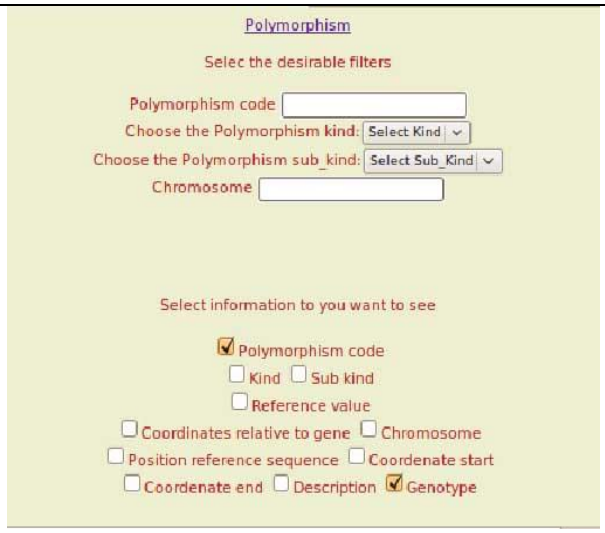
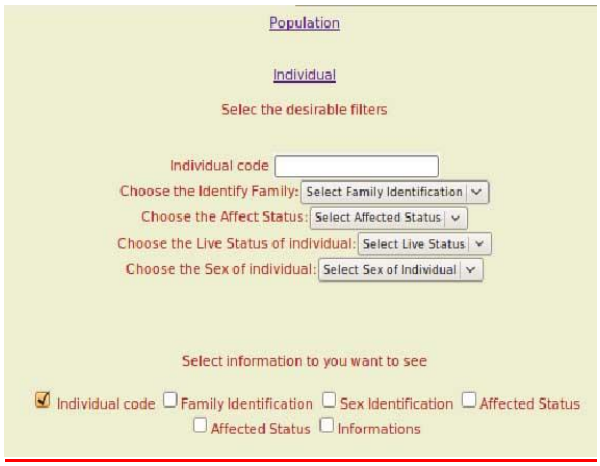
```
SELECT  ind.individual_code, pl.polymorphism_code, ge.value
FROM  population.pop,  individual.ind,  polymorphism.pl,  genotype.ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and pop.population_name  = 'Adygei'
```

A cláusula SELECT do SQL especifica os atributos para a projeção, e a cláusula WHERE especifica a condição de seleção e somente aqueles valores que satisfazem a condição serão selecionados. (ELMASRI, 2005). Na busca acima especificada, os atributos para projeção, ou seja, o que se deseja visualizar são os polimorfismos, o valor do genótipo e os indivíduos (*SELECT ind.individual_code, pl.polymorphism_code, ge.value*), que sejam da população Adygei (*WHERE pop.population_name = Adygei)*.

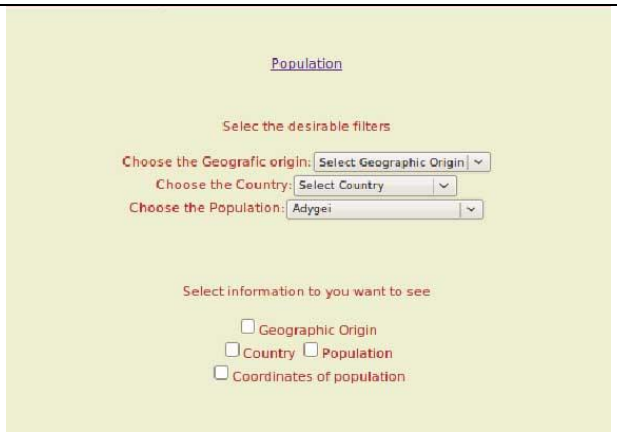
Os bancos de dados consistem em várias tabelas e utilizam uma chave como uma referência de uma tabela para outra. Então em se tratando do banco de dados relacional torna-se necessário na construção da sintaxe do select definir o relacionamento entre a chave estrangeira (polymorphism_id_polymorphism) da tabela *genotype* com a sua respectiva chave primária (id_polymorphism) na tabela relacional (*polymorphism*), além do relacionamento entre a chave estrangeira (individual_id_individual) da tabela *genotype* com a respectiva chave primária (id_individual) na tabela *individual*. Por isso na construção do SELECT acrescentamos as respectivas igualdades abaixo listadas:

```
ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
```

A construção de linhas complexas de comando é facilitada pela interface gráfica disponibilizada para o DIVERGENOMEdb (Figura 17).

	Tela contendo os campos para seleção (polymorphism code e genotype) e filtro da tabela *polymorphism*
	Tela contendo os campos para seleção (individual code) e filtro da tabela *individual*

SELECT ***ind.individual_code, pl.polymorphism_code, ge.value***
FROM population pop, individual ind, polymorphism pl, genotype ge

	Tela evidenciando os campos de busca e filtragem (população Adygei) da tabela *population*

WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and
ge.individual_id_individual = ind.id_individual
and
ind.population_id_population = pop.id_population
and
pop.population_name = Adygei.

FIGURA 17: Tela de busca e construção do select correspondente para consulta
Na esquerda se observa as telas disponibilizando os campos de busca e filtragem das respectivas tabelas descritas a direita. Abaixo em vermelho e preto a construção do respectivo select.

Em preto foi listado a cláusula obtida conforme os campos selecionados pelo usuário. Em vermelho destaque da cláusula gerada em programação para a construção completa do SELECT. O resultado da busca é visualizado como uma tabela, sendo cada coluna referente aos campos solicitados pelo usuário, como mostrado na Figura 18.

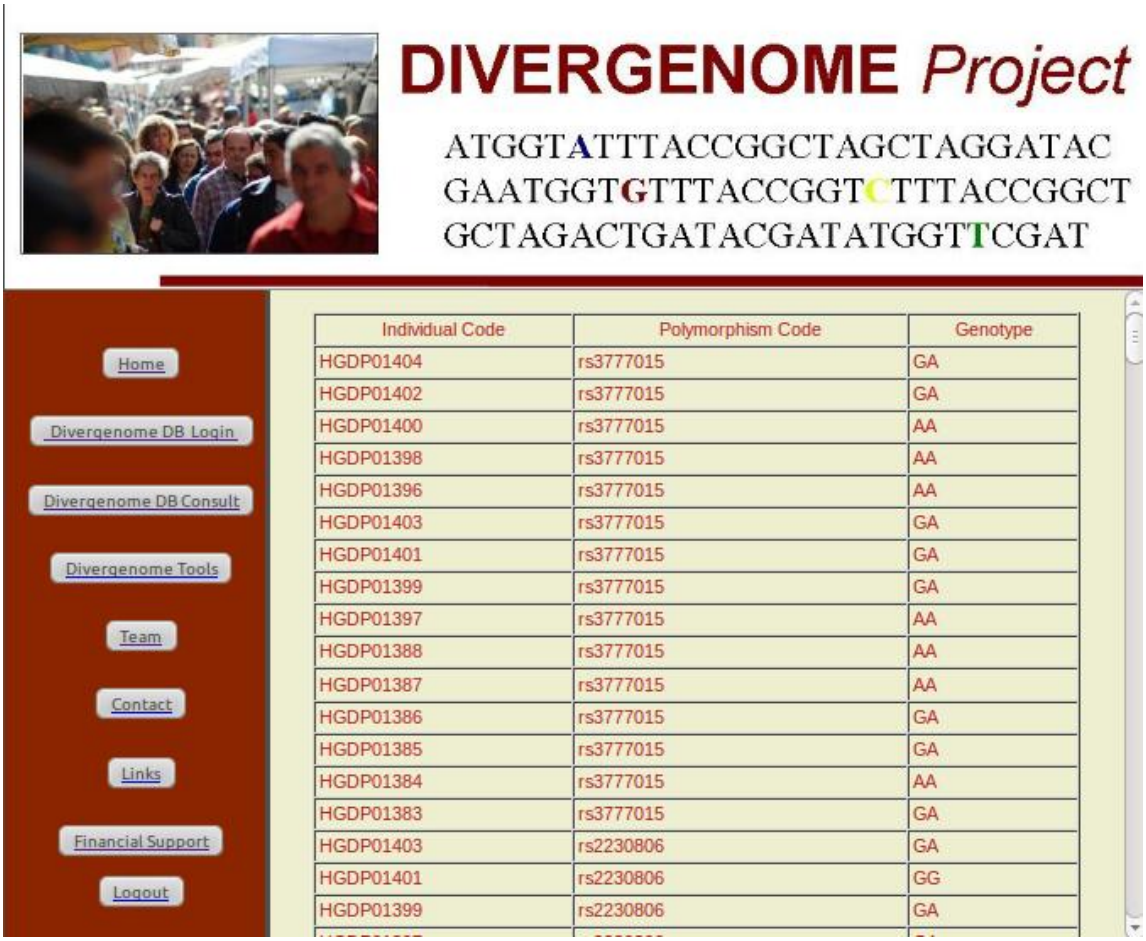


FIGURA 18: Resultado da busca, conforme os campos e filtros selecionados pelo usuário.
Todo o resultado é visualizado através do uso da barra de rolagem disponibilizada a direita da tela.

Após a disposição dos resultados na tabela existem dois botões distintos (Figura 19), através dos quais é possível:

(1) obter um arquivo contendo o select gerado na busca (*Do you want to print the select*); e

(2) fazer um download de arquivo texto no formato padrão de banco de dados com os valores retornados da busca (*Download*).

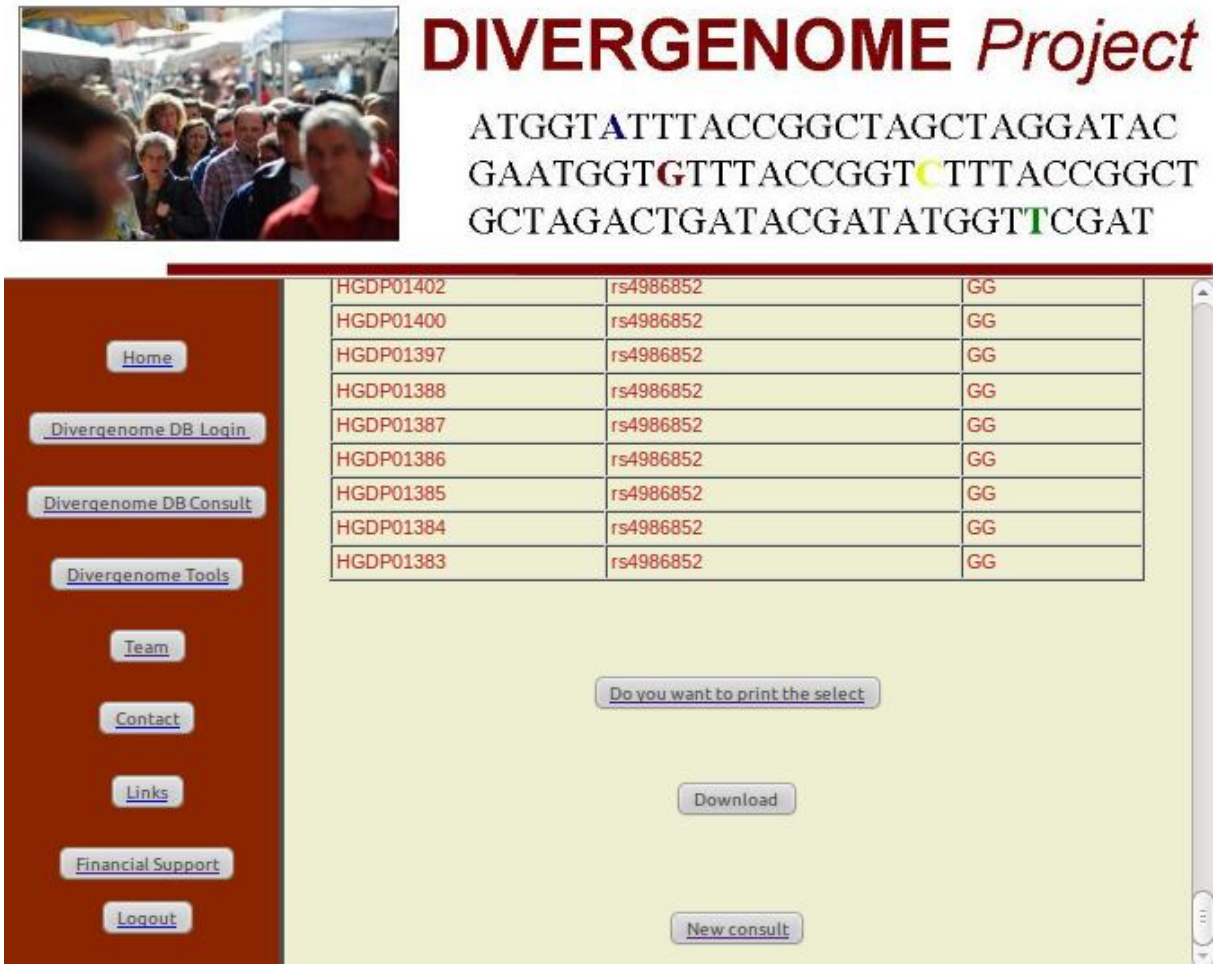


FIGURA 19: Tela evidenciando o final da visualização dos resultados. Disponibilizando as teclas de impressão do SELECT gerado na busca, bem como download deste mesmo resultado visualizado inicialmente na tela

A interface gráfica permite ao usuário realizar essas pesquisas de maneira mais prática, sem a necessidade do conhecimento de sintaxe de busca em SQL.

Com os dados de polimorfismos e genótipos dos indivíduos em diferentes populações de interesse, posteriormente poderíamos utilizar ferramentas do DIVERGENOME Tools para gerar os inputs no formato específico requerido por diferentes programas utilizados em genética de população. No trabalho de Soares-

Souza, as buscas pelo banco de dados foram utilizadas para gerar a matriz no formato SDAT (código do indivíduo x código do polimorfismo) no programa Poly_out, para posteriormente a matriz ser convertida e utilizada em diferentes programas de Genética de Populações que retornou cálculos estatísticos de heterozigosidade, divergência, EHW, etc. No trabalho de Soares-Souza (2010) o filtro de busca inicialmente apresentado foi utilizado para elucidar a existência de SNPs com alta diferenciação entre populações, que podem estar sob seleção natural positiva ou devem sua diferenciação ao allele surfing (SOARES-SOUZA, 2010).

Para as buscas avançadas também é possível fazermos pesquisas em que o filtro de busca não é apenas um valor único, mas um arquivo contendo um conjunto de valores deste campo de filtro. Esse tipo de busca é disponibilizada para alguns atributos específicos, implementada conforme as necessidades atuais de buscas realizadas no LDGH, mas que poderão ser ampliadas conforme o surgimento de novas necessidades. Um exemplo é dado para pesquisas em que se busca a filtragem por diversos códigos de polimorfismos, ou seja, desejo saber valores específicos, como os genótipos, indivíduos, as populações e origens geográficas disponíveis no banco de dados para uma lista de polimorfismos presentes em um arquivo de busca do usuário. Neste caso, ao invés de fazermos uma busca para cada código de polimorfismo, há a possibilidade de realizarmos uma única busca, em que o arquivo disponibilizado pelo usuário conteria todos os códigos de polimorfismos que se deseja obter como filtro de busca. Para esse tipo de consulta há a necessidade de utilizarmos uma sintaxe que permita juntar valores de duas (*ou mais*) consultas em um único resultado. Isso é obtido pelo comando *SQL* **UNION.**

Um exemplo de consulta recuperando diversos valores de *polymorphism code* disponíveis em um arquivo fornecido pelo usuário pode ser visto na Figura 20.

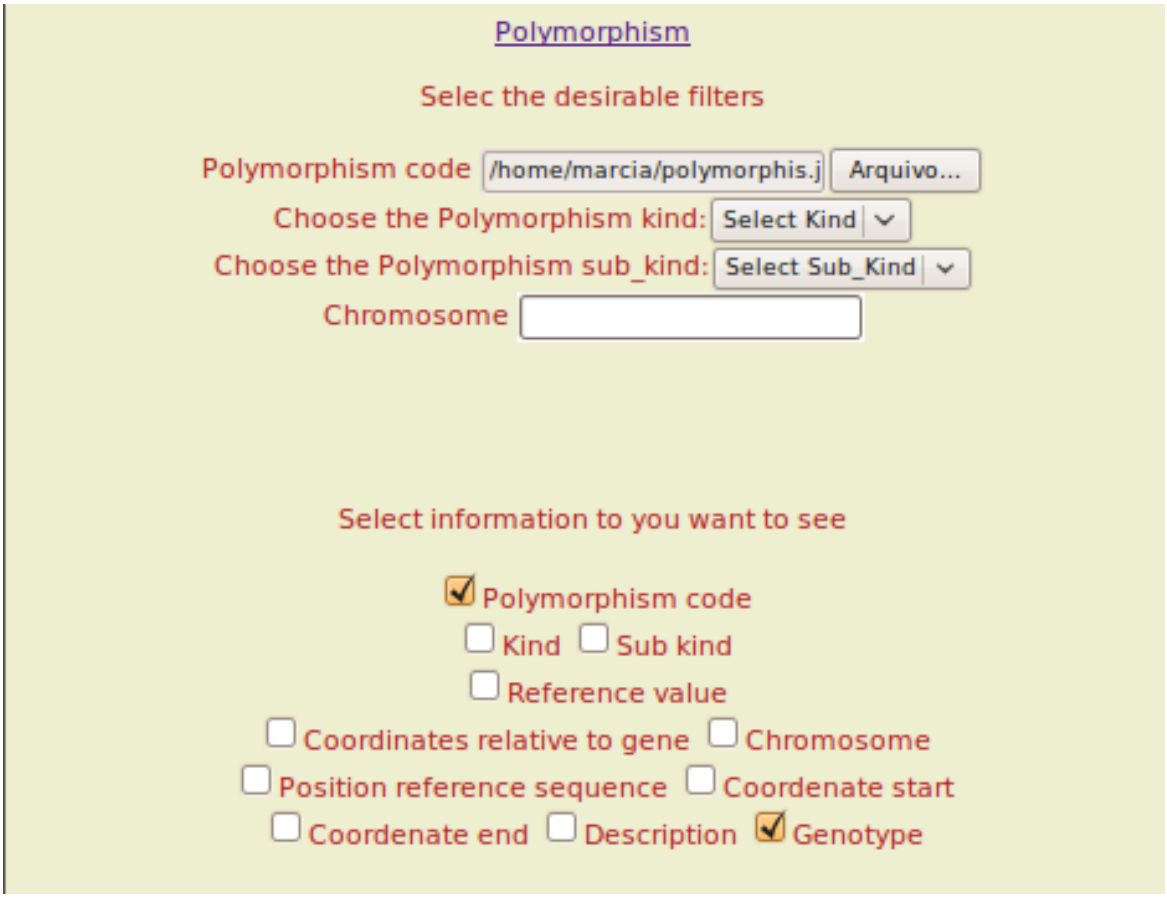


FIGURA 20: Tela apresentando os dados disponíveis para filtragem e seleção da tabela *polymorphism*[1]

Essa busca se refere ao relacionamento de quatro tabelas, polymorphism, genotype, individual e population. Foi solicitada ao DIVERGENOMEdb a seleção de todos os indivíduos, bem como a população e origem geográfica as quais tais indivíduos pertencem, seus genótipos para os polimorfismos contidos no arquivo de entrada pelo usuário no campo de filtro Polymorphism code (campos superiores). Esse arquivo em específico contém nove códigos de polimorfismos (Figura 21).

[1] No campo de filtragem de busca para o atributo polymorphism code o usuário fornece o nome do arquivo que contém vários códigos de polimorfismos

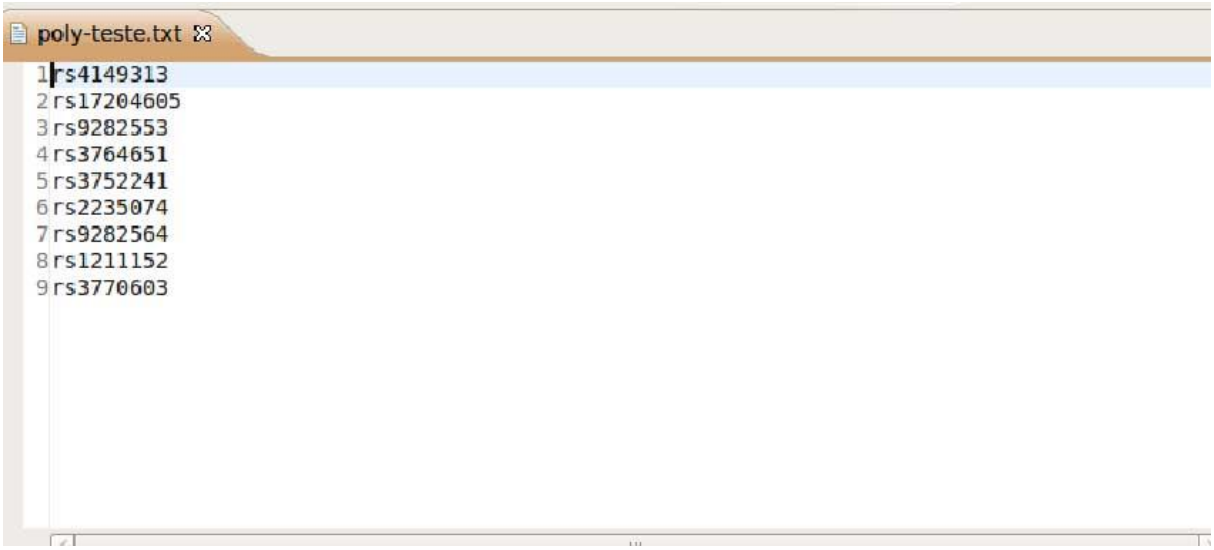


FIGURA 21: Visualização do arquivo contendo os nove códigos de polimorfismos para pesquisa no banco de dados

A consulta no DIVERGENOMEdb retorna uma tabela contendo os resultados (Figura 22).

HGDP00672	EUROPE	Sardinian	rs4149313	AG
HGDP00671	EUROPE	Sardinian	rs4149313	AG
HGDP00670	EUROPE	Sardinian	rs4149313	AG
HGDP00669	EUROPE	Sardinian	rs4149313	AA
HGDP00668	EUROPE	Sardinian	rs4149313	AA
HGDP00667	EUROPE	Sardinian	rs4149313	AG
HGDP00666	EUROPE	Sardinian	rs4149313	AA
HGDP00665	EUROPE	Sardinian	rs4149313	AA
CAY561	AMERICA	Cayapa	rs17204605	CC
QT073	AMERICA	Quetchua	rs17204605	CC
QT059	AMERICA	Quetchua	rs17204605	CC
QT137	AMERICA	Quetchua	rs17204605	CC
CAY559	AMERICA	Cayapa	rs17204605	CC
CAY547	AMERICA	Cayapa	rs17204605	CC
QT134a	AMERICA	Quetchua	rs17204605	CC
QT134b	AMERICA	Quetchua	rs17204605	CC
QT143	AMERICA	Quetchua	rs17204605	CC
QT145	AMERICA	Quetchua	rs17204605	CC
QT110	AMERICA	Quetchua	rs17204605	CC

FIGURA 22: Resultados retornados da consulta ao DIVERGENOMEdb. Colunas da esquerda para a direita: indivíduo, origem geográfica da população, população, código do polimorfismo e genótipo do indivíduo.

Este tipo de busca tem a seguinte linha de comando embutida para o retorno do resultado:

```
SELECT  pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,    polymorphism pl,    genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and  pl.id_polymorphism = 3
UNION
SELECT  pop.geographic_origin, pop.population_name, ind.individual_code,
         pl.polymorphism_code, ge.value
FROM population pop, individual ind,    polymorphism pl,    genotype ge
WHERE ge.polymorphism_id_polymorphism   = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and  pl.id_polymorphism = 604
UNION
SELECT pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,    polymorphism pl,    genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and pl.id_polymorphism =  10
UNION
SELECT pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,    polymorphism pl,    genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and pl.id_polymorphism =  11
UNION
SELECT pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,    polymorphism pl, genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and pl.id_polymorphism =  12
UNION
SELECT pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,    polymorphism pl, genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
```

```
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and pl.id_polymorphism =  13
UNION
SELECT  pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,     polymorphism pl, genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and  pl.id_polymorphism = 14
UNION
SELECT  pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,     polymorphism pl, genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and ind.population_id_population = pop.id_population
and  pl.id_polymorphism = 15
UNION
SELECT  pop.geographic_origin, pop.population_name, ind.individual_code,
        pl.polymorphism_code, ge.value
FROM population pop, individual ind,     polymorphism pl, genotype ge
WHERE ge.polymorphism_id_polymorphism = pl.id_polymorphism
and ge.individual_id_individual = ind.id_individual
and  ind.population_id_population = pop.id_population
and  pl.id_polymorphism = 16
```

A interface gráfica permite realizar as buscas sem que o usuário tenha que conhecer linhas de comandos avançadas e extensas como essa apresentada. Vale ressaltar que no *select* em questão trabalhamos com a busca filtrando por um arquivo que continha apenas nove códigos de polimorfismos (um SELECT para cada código de polimorfismo do arquivo – todos valores unidos pelo comando UNION). Uma única consulta pôde retornar todos os resultados para os nove códigos de polimorfismos contidos no arquivo apresentado no campo filtro. Sem esse recurso teríamos que realizar uma busca para cada polimorfismo. A mesma pesquisa pode ser realizada com centenas de códigos de polimorfismos como filtro, o que para o usuário seria um tempo dispendioso para a construção da linha de comando sem a disponibilização desse recurso. Além disso, o usuário pode gerar um arquivo de saída em txt (select.txt) contendo o select acima apresentado, bem como pode gerar um arquivo txt (registros.txt) contendo o mesmo resultado apresentado na tela

(Figura 22) no formato padrão de saída de dados, para uso posterior nas ferramentas disponíveis no DIVERGENOME TOOLS. Não é o foco deste trabalho, mas no caso em questão, o arquivo registros.txt (contendo o resultado da busca realizada anteriormente), foi utilizado como arquivo de entrada para o algoritmo Poly_out de DIVERGENOMEtools, que manipula o arquivo de saída disponibilizado na busca gerando uma tabela onde na primeira coluna (eixo vertical) estão os indivíduos dispostos em linhas, na primeira linha e demais colunas (eixo horizontal) estão as posições dos SNPs na seqüência referência e nas intercessões entre colunas linhas os respectivos genótipos (Tabela 4). Com essa tabela é possível ter uma visão geral de quais indivíduos possuem polimorfismos em determinada posição, além disso, essa tabela é o arquivo de entrada para as próximas etapas de manipulação de arquivos de análise.

Tabela 4
Matriz SDAT

Sdat	ABCA1_04	ABCA1_12	ABCA1_15	ABCA1_17	ABCA1_26
CDP0847	AG	AG	TT	GG	AC
CDP0848	AG	GG	TT	AG	AA
CDP0849	AG	AG	TT	AG	AC
CDP0850	AG	AG	TT	AG	AC
CDP0851	AA	AG	TT	GG	AC
CDP0852	AA	GG	TT	GG	AA
CDP0853	GG	GG	TT	GG	AA
CDP0854	AG	AG	TT	GG	AC
CDP0855	AA	GG	TT	AA	AA
CDP0856	AA	AA	TT	AG	CC

4.4.2 Outros tipos de aplicação

Em estudos de Epidemiologia Genética, desenhos de estudo de caso-controle são agora amplamente utilizados para estudar o papel da suscetibilidade genética na etiologia de doenças complexas. Normalmente, um estudo caso-controle envolve a totalidade ou uma grande parte dos indivíduos doentes (casos) que surgem num estudo de base subjacente e, em seguida, um número comparável de amostragem de indivíduos saudáveis (controles), sendo de preferência exatamente a mesma base do estudo, e eventualmente combinados com os casos de algumas

características sócio-demográficas, como idade e sexo (NILANJAN, 2009). Estudos caso-controle de pessoas não aparentadas e projetos de base familiar são os mais amplamente utilizados. Apresentamos uma das aplicações do DIVERGENOMEdb nesses tipos de estudos caso-controle.

Umas das entidades do DIVERGENOMEdb é a tabela *Individual*. Nela são encontradas informações referentes ao indivíduo, bem como à família do indivíduo (Figura 23).

Visualizar | Estrutura | SQL | Procurar | Inserir | Exportar | Importar | Operações | Limpar | Eliminar

	Campo	Tipo	Collation	Atributos	Nulo	Padrão	Extra	Ação
☐	id_individual	int(10)		UNSIGNED	Não	*Nenhum*	auto_increment	
☐	population_id_population	int(10)		UNSIGNED	Não	*Nenhum*		
☐	individual_code	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	family_id	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	father_id	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	mother_id	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	sex	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	aff_status	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	live_status	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	info_1	varchar(45)	latin1_swedish_ci		Sim	*NULL*		

FIGURA 23: Tela do phpMyAdmin, visualizando a estrutura da tabela *Individual*

Uma das chaves estrangeiras desta tabela é population_id_population que relaciona o indivíduo a uma população inserida na entidade *population*. Além disso, o código de identificação do indivíduo esta relacionado ao fenótipo (variáveis quantitativas ou qualitativas), bem como ao genótipo, sendo possível estudo da penetrância e a expressividade da doença em uma família específica.

Uma das possíveis procuras (*selects*) que poderia ser utilizada nos estudos utilizando-se das variáveis quantitativas esta apresentada na Figura 24, que tem por objetivo listar as idades dos indivíduos cadastrados no banco de dados.

select-4.txt

```
SELECT  ind.individual_code,quant.value FROM  individual ind, quantitative_variable quant,
individual_has_quantitative_variable quanthas WHERE ind.id_individual =
quanthas.individual_id_individual     AND quanthas.quantitative_variable_id_quantitative_variable
= quant.id_quantitative_variable     and   quant.qual_var_name = IDADE
```

FIGURA 24: Possível select a ser utilizado para busca no DIVERGENOMEdb.

A relação como as variáveis quantitativas e qualitativas podem ser usadas para controle de amostragem em busca da mesma base de estudo, contendo características sócio-demográficas, como idade e sexo.

Utilizando a interface podemos fazer buscas pelas tabelas indivíduos e variáveis quantitativas ou qualitativas. O mesmo *select* seria obtido de forma simples através da interface gráfica apresentada na Figura 25 e na Figura 26.

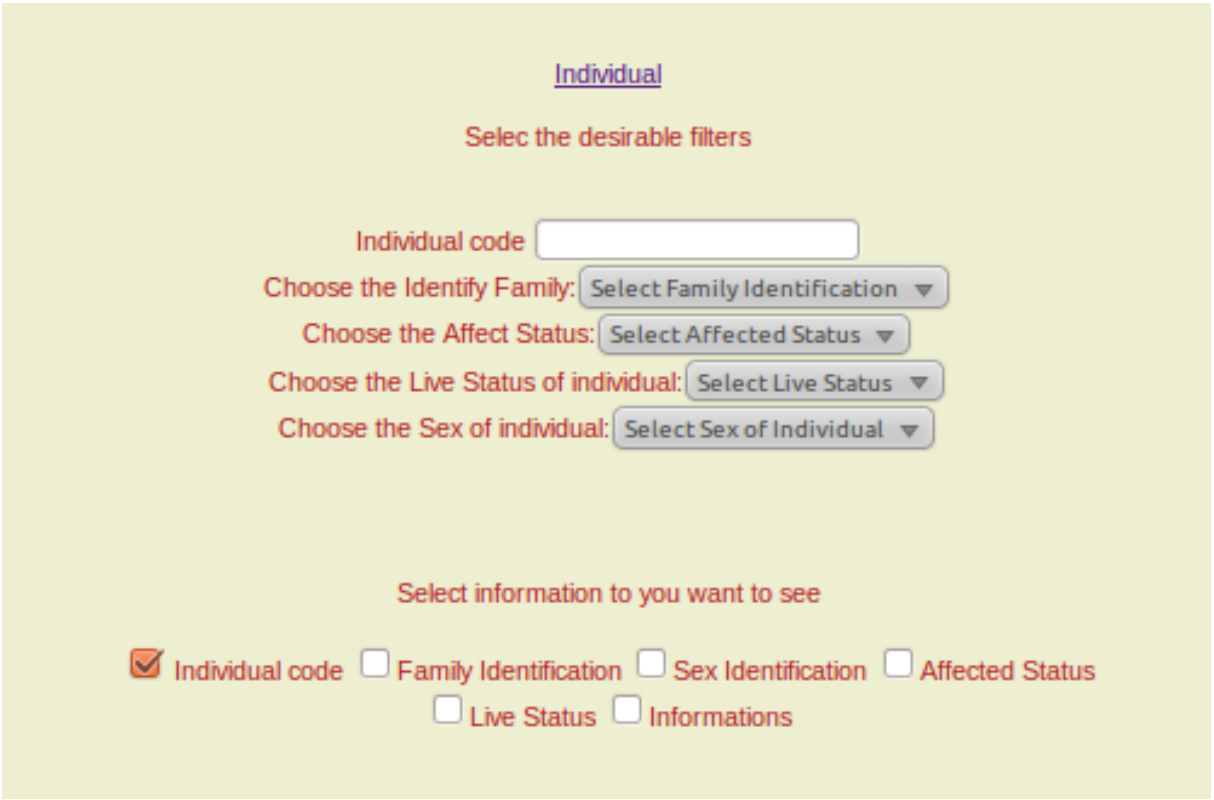


FIGURA 25: Tela de interface para busca na tabela *individual*.

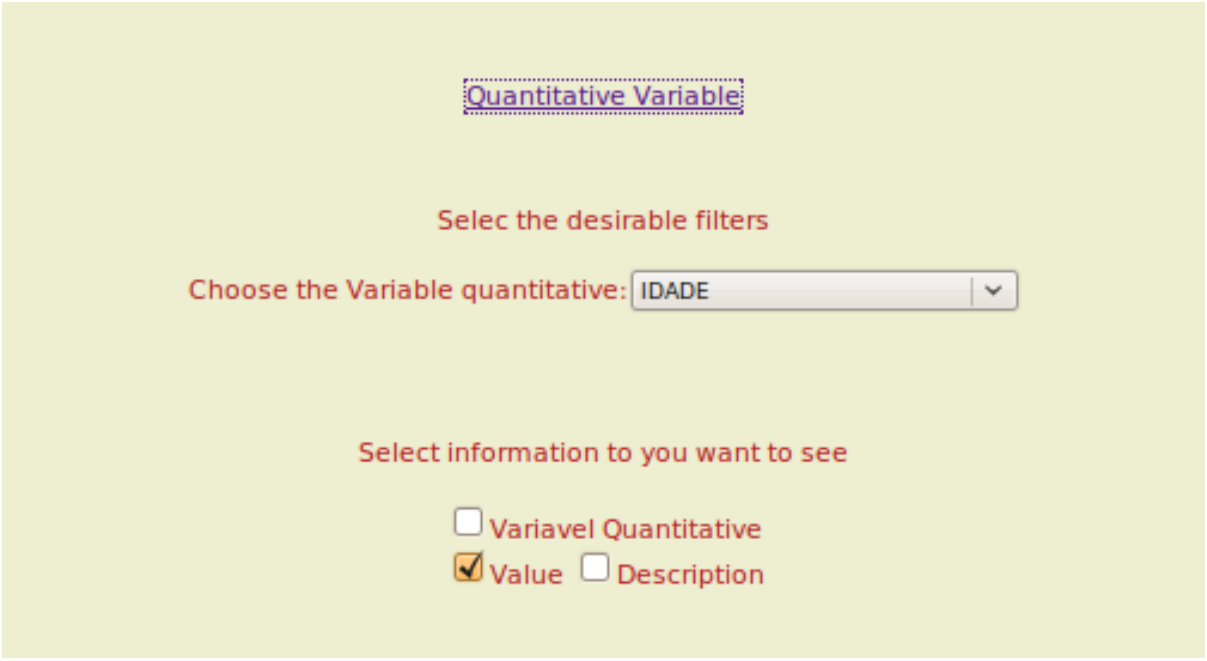


FIGURA 26: Tela de interface para a entidade *variable_quantitative*

Através dos exemplos apresentados, é possível perceber a aplicação prática e a viabilidade da interface como facilitadora na utilização do DIVERGENOMEdb.

5. CONCLUSÃO

Foi criada a interface gráfica da plataforma DIVERGENOMEdb, e com isso espera-se facilitar o compartilhamento dos dados obtidos de pesquisas científicas, através de um sistema que seja capaz de inserção de novas informações, bem como a navegação pelo banco de dados de maneira ágil e simples. Os pesquisadores que não possuem a habilidade para compor uma consulta através da linha de comando podem ver a interface do DIVERGENOMEdb como uma facilidade para trabalhar com os seus dados.

Com a possibilidade de inserção de dados por usuários, esperamos que o número de utilizadores aumente e estes mesmos pesquisadores encorajados e capacitados para se tornar contribuidores, aumentem o nível e volume de informação científica na base de dados do DIVERGENOME.

5.1. Perspectivas

Atualmente, DIVERGENOMEdb permite a inserção e consulta de dados, não sendo possível a alteração de dados inseridos no database. É necessário definir o mecanismo para o controle de dados privados, bem como será realizado o controle das possíveis alterações, para posterior implementação desse recurso.

Não foi desenvolvida a interface gráfica para consulta e inserção de dados das entidades *haplotype* e *haplotype_evaluation* (vide diagrama que apresenta as relações entre as tabelas – Figura 1). Tais entidades estão ainda em debate sobre seus atributos, bem como será realizada a disponibilização de seus dados.

Esperamos que com o aumento do uso dos recursos disponibilizados para o DIVERGENOME Project, novas necessidades surjam e possam ser implementadas ao projeto em uma nova versão.

O compartilhamento dos dados só será atingido após a disponibilização da plataforma via internet, que será feita posteriormente após maiores testes e implementação de novos recursos que julgarmos necessários para novas aplicações. Será publicado um artigo científico com a descrição de DIVERGENOME. Na nossa experiência, vários pesquisadores do ICB estão utilizando as ferramentas de DIVERGENOMEtools. Por este motivo, podemos esperar que o artigo com a descrição de DIVERGENOME terá um elevado número de citações. Para aumentar o impacto científico de DIVERGENOME, os scripts originais desenvolvidos em DIVERGENOMEdb serão disponibilizados sob *GNU General Public License*.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, F. N. **Implementação de um Banco de Dados de Proteomas de Bactérias associadas a plantas**: PROBACTER. Modelagem Computacional com Ênfase em Bioinformática e Biologia Computacional, LNCC. Tese – Laboratório Nacional de Computação Científica, LNCC. Petrópolis, RJ - 2007.

ARBEX, W.; COSTA, V. S.; SILVA, M. V. G. **Bioinformática como ferramenta nas Pesquisas atuais**. 2008. Disponível em: <http://www.genmelhor.ufv.br/materiais/III%20egm/textobioinformatica.pdf>. Acesso em: 11 de junh de 2010.

ASHBURNER, M.; [et al.] Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nat Genet**; v. 25, n.1, p. 25-9, 2000.

BALARESQUE, P. L.; BALLEREAU, S. J.; JOBLING, M. A. Challenges in human genetic diversity: demographic history and adaptation. **Hum Mol Genet**; v.6 Spec No. 2, p. 34-9, 2007.

BATEMEN, A. Editorial: What makes a good database? **Nucleic Acids Research, Database issue**; v. 35, n. 10, p. 1.093-105, 2007.

CAMARGO FILHO, F. [et. al.]. **Bioinformática: Manual do Usuário**. 2008. Disponível em: < http://www.biotecnologia.com.br/revista/bio29/bioinf.pdf>. Acesso em: 11 de junho de 2010.

CAVALLI-SFORZA L. L. How can one study individual variation for three billion nucleotides of the human genome? **The American Journal of Human Genetics;** v. 46, p. 649-651, 1990.

CAVALLI-SFORZA, L. L. The Human Genome Diversity Project: past, present and future. **Nat Rev Genet**; v.6, n. 4, p. 333-40, 2005.

CHEVITARESE, J. **Determinação da estrutura genética das populações humanas e inferência dos fatores evolutivos que contribuíram para sua formação**. 2009. 99f. Dissertação (Mestrado em Genética) – Universidade Federal de Minas Gerais, Belo Horizonte.

CUNHA, D. A. S., [et al]. Site Piatam: **Pesquisa Científica e Tecnologia da Comunicação em Simbiose**. Trabalho apresentado ao Altercom – Jornada de Inovações Midiáticas e Alternativas Experimentais . Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação - XXIX Congresso Brasileiro de Ciências da Comunicação – UnB – setembro. 2006.

DATE, C. J. **Introdução a sistemas de bancos de dados.** 8. ed. Rio de Janeiro: Eselvier, 2003.

DORNELES, C. F. [et. al.]. **Pesquisa por similaridade em dados XML**. Relatório de Projeto de Pesquisa – RP – 327. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2003.

ELMASRI, R. **Sistemas de banco de dados**. 4. ed. São Paulo: Pearson Addison Wesley, 2005.

EVANGELOS, E., [et al]. Family-Based versus Unrelated Case-Control Designs for Genetic Associations. **PLoS Genet.** August. 2006.

FEITOSA, M. F.; KRIEGER, H. O futuro da epidemiologia genética de características complexas. **Ciência & Saúde Coletiva.** v. 7, n. 1, 2002.

GAASTERLAND, T. [et al]. Special issue on data management, analysis, and mining for the life sciences. **The VLDB Journal**. v. 14, n. 3, p. 279-280, 2005.

GIBAS, C.; JAMBECK, P. **Desenvolvendo bioinformática**: ferramentas de software para aplicações em biologia. Rio de Janeiro: Editora Campos, 2001.

GONCALVES, A.; GONCALVES, N. N. S. Epidemiologia genética: epidemiologia, genética ou nenhuma das anteriores? **Cad. Saúde Pública**, Rio de Janeiro, v. 6, n. 4, Dec. 1990.

GUDMUNDUR, A. T. [et al.]. Genotype-phenotype databases: challenges and solutions for the post-genomic era. **Nature Reviews in Genetics**, v.10, nº.1, p.9-18, January, 2009.

INTERNATIONAL HAPMAP CONSORTIUM. The International HapMap Project. Nature, v. 8,n. 426, p. 789-96, Dec, 2003.

JAKOBSSON, M. [et al.]. Genotype, haplotype and copy-number variation in worldwide human populations. **Nature Reviews in Genetics,** n. 451, p.998-1003, Feb, 2008.

LEAL, E.; WIECZOREK, E. M. **Caminhos e Tendências do uso de Banco de Dados em Bioinformática** 2003. Disponível em: <http://www.wieczorek.com.br/publicacoes/artigo_IVencoinfo.pdf>>. Acesso em: 11 de julho de 2010.

LESK, A. M. **Introdução à Bioinformática**. 2. ed. Porto Alegre: Artmed, 2008.

LIFSCHITZ, S. Algumas Pesquisas em Banco de Dados e Bioinformática. Anais do XXVI Congresso da SBC. Workhop de Biologia Computacional. Campo Grande, MS, Julho. 2006.

NIEDERAUER, J. **Desenvolvendo Websites com PHP**. São Paulo: Editora Novatec, 2004.

NILANJAN, C. [et, al.]. Analysis of Case-Control Association Studies: SNPs, Imputation and Haplotypes. **Stat Sci.**; v.24, n. 4, p. 489–502, November, 2009.

PACKER, B. R. [et al.]. SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes**. Nucleic Acids Research**, v.32, p. D528-D532, 2004.

PROSDOCIMI, F; [et. al]. Bioinformática. Manual do usuário. Um guia básico e amplo sobre os diversos aspectos dessa nova ciência. **Biotecnologia Ciência e Desenvolvimento**, v. 29, n. 4, p. 12-21, 2008.

RESENDE, B. F. ; SILVA, D. S. **Bioinformática.** 2008. Disponível em : http://www.merit.unu.edu/MEIDE/papers/2009/1228690340_DS.pdf. Acesso em 11 de julho de 2010.

RYAN, E.M., [et, al.]. An initial map of insertion and deletion (INDEL) variation in the human genome. **Genome Res**. September; v. 16, n. 9, p. 1182–1190. 2006.

ROCHA, A. R. C. [et al.]. **Qualidade de software:** Teoria e Prática. São Paulo: Prentice Hall, 2001.

ROSENBERG, N. A. [et. al.]. Genetic structure of human populations. **Science**; v. 298, n. 5602, p. 2381-5, 2002.

SEARLS, D. B. Data integration: challenges for drug Discovery. Nature reviews. **Drug discovery**, v.4, n. 1, p.45-58, 2005.

SHERRY, S.T. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. v.1, n. 29, p. 308-11, jan. 2001.

SILVA, W. L. S. **Análise in silico de uma matriz DRE na seqüência promotora de genes da Levedura Saccharomyces cerevisiae** 2004. Disponível em < http://biolab.cin.ufpe.br/team/dissertacao_wlss.pdf> Acesso em 05 de junho de 2010.

SOUZA, G. B. S. **Identificação de genes com alta diferenciação entre populações humanas:** inferências evolutivas e aplicações biomédicas. 2010. 104f. Dissertação (Mestrado em Genética) – Universidade Federal de Minas Gerais, Belo Horizonte.

SUAREZ, C. V. [et al.]. SNP analysis to results (SNPator): a web-based environment oriented to statistical genomics. **Analyses upon SNP data**. v.24. p. 1643-1644, May, 2008.

SUDER, R. L; DORNELES, C. F. **Integração de dados em múltiplos níveis**. 2008. Disponível em: <http://www.upf.br/erbd/download/16194.pdf>. Acesso em 15 de julho de 2010.

VENETIANER, T. **HTML:** Desmitificando a Linguagem da Internet. São Paulo: Makron Boocks, 1996.

WANG, R. N.[et al.]. Geographic patterns of genome admixture in Latin American Mestizos. **PLoS Genet**; v. 4 n. 3p. 1000-37, 2008.

WELLING, L.; THOMSON, L. **PHP e MySQL desenvolvimento Web**. Rio de Janeiro: Elsevier, 2005.

WU, C. [et al.]. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. **Genome Biology**, november 2009.

Universidade Federal de Minas Gerais

MANUAL DO USUÁRIO

Sistema DIVERGENOMEdb

SUMÁRIO

1 - INTRODUÇÃO

A plataforma bioinformática DIVERGENOME tem como objetivo auxiliar no armazenamento de dados genéticos em estudos epidemiológicos e de genética de populações. Possui dois componentes: um banco de dados relacional (DIVERGENOMEdb) que possibilita o armazenamento seguro e eficiente de dados de projetos de Genética de Populações e Epidemiologia Genética, e um conjunto de ferramentas para facilitar a análise dos dados (DIVERGENOMEtools).

O DIVERGENOMEdb reúne dados genotípicos e fenotípicos proveniente de diferentes estudos nas áreas de Genética de Populações, Genética Clínica e Epidemiologia realizados em populações indígenas ou miscigenadas da América Latina e outras regiões.

O DIVERGENOMEtools corresponde a um conjunto de scripts que disponibiliza aos usuários o acesso a ferramentas de manipulação de *outputs* do DIVERGENOMEdb, para a geração de *inputs* para *softwares* de Genética de População (PHASE, DNAsp, Structure), Genética Médica (Haplopainter) e pacotes de análises estatísticas, ou seja disponibiliza aos usuários o acesso a ferramentas de conversão de base de dados para diversos formatos de arquivos utilizados em diferentes programas. A Figura 1 representa a comunicação entre os *scripts* desenvolvidos e a base de dados DIVERGENOME.

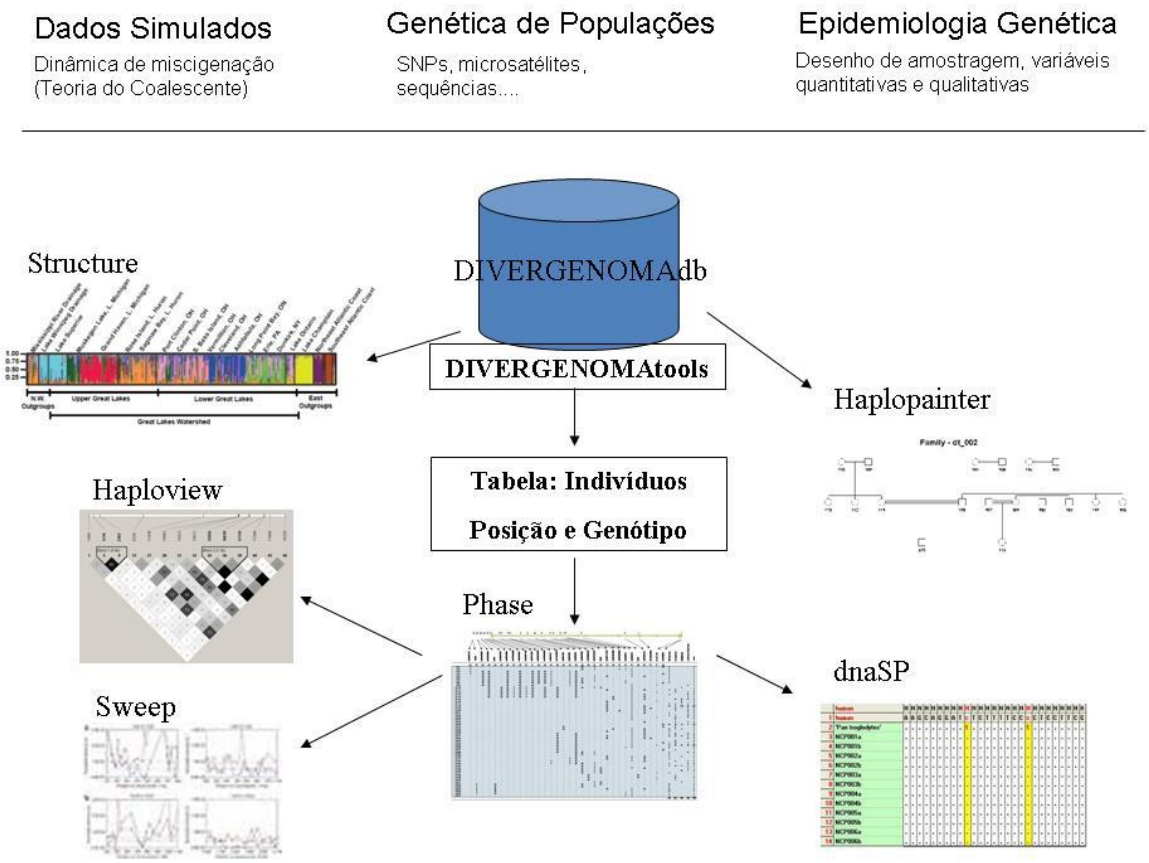


FIGURA 1: DIVERGENOMEtools
Fonte: Magalhães, 2010 – manuscrito em preparação. Comunicação entre os *scripts* já desenvolvidos e base de dados DIVERGENOME.

Esta documentação se refere à interface de consulta do DIVERGENOMEdb e tem por finalidade servir de auxílio aos usuários, apresentando as funcionalidades e recursos do DIVERGENOME database. O banco de dados é constituído de vinte tabelas, cada uma contendo um número variável de campos (Diagrama de Entidade-Relacionamento (DER) – Anexo 1). As tabelas são interligadas por chaves estrangeiras (campos que conectam duas tabelas), o que permite a filtragem de pesquisa relacionando campos de diferentes tabelas. A plataforma aceita dados genotípicos individuais de quatro tipos: contigs, SNPs, INDELs e microssatélites, e ainda pode ser facilmente modificada para incorporar polimorfismos como *Copy Number Variation* (CNV). Os genótipos são ligados a uma descrição dos protocolos de laboratório usados para gerar dados. Os indivíduos podem estar ligados a dados

fenotípicos coletados em estudos epidemiológicos, que podem incluir o *status* da doença como atributo binário, variáveis quantitativas como idade, ou variáveis fisiológicas.

2 – DESCRIÇÕES GERAIS DAS FUNCIONALIDADES

2.1 – Interface

A interface *web* foi desenvolvida para o acesso e administração do DIVERGENOME Database. Informações gerais sobre o DIVERGENOMEdb podem ser visualizadas sem cadastro prévio. A interface é composta de uma página inicial ou home (Figura 2) que informa ao visitante o objetivo geral da plataforma, bem como informações sobre cadastro de usuário, membros desenvolvedores, financiadores e disposição de links. Além disso, sem cadastro prévio é permitido ao usuário a realização de consultas a dados públicos inseridos no DIVERGENOMEdb.

2.2 – Layout

As páginas de navegação do DIVERGENOMEdb possuem o layout apresentado na Figura 2.

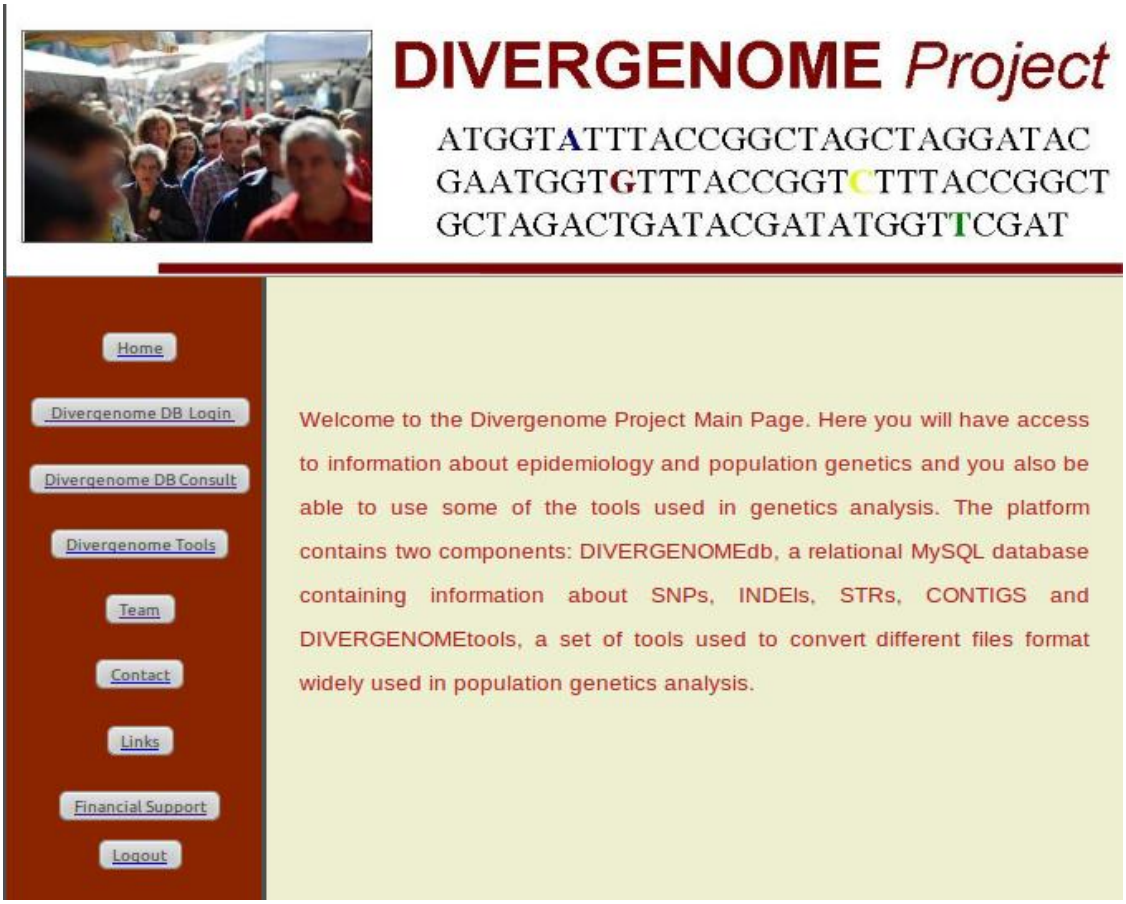


Figura 2: Tela Inicial de apresentação do DIVEREGENOME Project. O menu contém *links* para acesso a informações gerais, bem como *links* para o DIVERGENOMEdb e o DIVERGENOMEtools.

Na Tabela 1 é apresentada uma breve descrição dos ícones disponíveis na tela inicial, bem como o respectivo direcionamento feito a partir dos links desta tela.

Botão	Descrição
Home	Página inicial do DIVERGENOME. Contém uma breve descrição dos benefícios do projeto.
Divergenome DB Login	Disponibiliza o acesso ao banco de dados, bem como permite o cadastro de usuário.
Divergenome DB Consult	Permite a consulta a dados públicos presentes na database do Divergenome.

Divergenome Tools	Acesso às ferramentas de conversão de dados biológicos.
Team	Listas de desenvolvedores do projeto, contendo breve descrição de suas formações.
Contact	Disponibiliza o email de contato com o Laboratório de Diversidade Genética Humana (LDGH) no Instituto de Ciências Biológicas – UFMG
Links	Links voltados ao estudo de Genética de Populações e Genética Médica.
Financial Support	Listas de financiadores.
Logout	Saída da conexão.

Tabela 1. Botões contidos no menu inicial da interface de acesso ao DIVERGENOMEdb , bem como descrição breve dos campos.

2.3 – Acesso do Usuário

O acesso ao sistema varia de acordo com os níveis de hierarquia estabelecidos para a manipulação dos dados. Os administradores possuem acesso ilimitado a base de dados, os coordenadores poderão inserir dados de seus projetos e os simples usuários ou membros de projetos podem visualizar dados gerais de projetos ao quais são membros. Usuários não cadastrados podem fazer consultas a dados públicos e utilizar as ferramentas disponíveis no DIVERGENOME Tools.

2.4 – Formato de entrada e saída de dados do Divergenome

Tanto os arquivos de entrada para DIVERGENOMEdb, quanto os arquivos de saída, obedecem o formato padrão de banco de dados – onde os mesmos estão dispostos linearmente e os campos separados um do outro por tabulação . Primeiramente o resultado da busca é disponibilizado na tela como uma tabela. Posteriormente é possível gerar um arquivo (registro.txt) com os valores retornados

da consulta e também é possível gerar a linha de comando do select usado na busca em um arquivo de texto (select.txt).

A partir do resultado obtido no arquivo registro.txt é possível utilizarmos de ferramentas do DIVERGENOME Tools, que permite a conversão dos dados para formatos específicos requeridos por diferentes *softwares* de Genética de Populções (PHASE, DNAsp, Structure), Genética Médica (Haplopainter) e pacotes estatísticos de uso comum.

3 – DIVERGENOMEdb LOGIN

Para acesso as funcionalidades disponibilizadas para o Sistema DIVERGENOMEdb basta clicar no botão “DIVERGENOMEdb Login” – Figura 2. A tela apresentada na Figura 3 será apresentada. O código de usuário e a senha deverão ser informados e em seguida a tecla *Submit* deverá ser acionada. Os testes de verificação de usuário e senha serão realizados. Se o usuário ainda estiver aguardando a aprovação do administrador, ou o seu cadastro tiver sido recusado, o mesmo não conseguirá logar ao sistema (mensagem de alerta será apresentada).

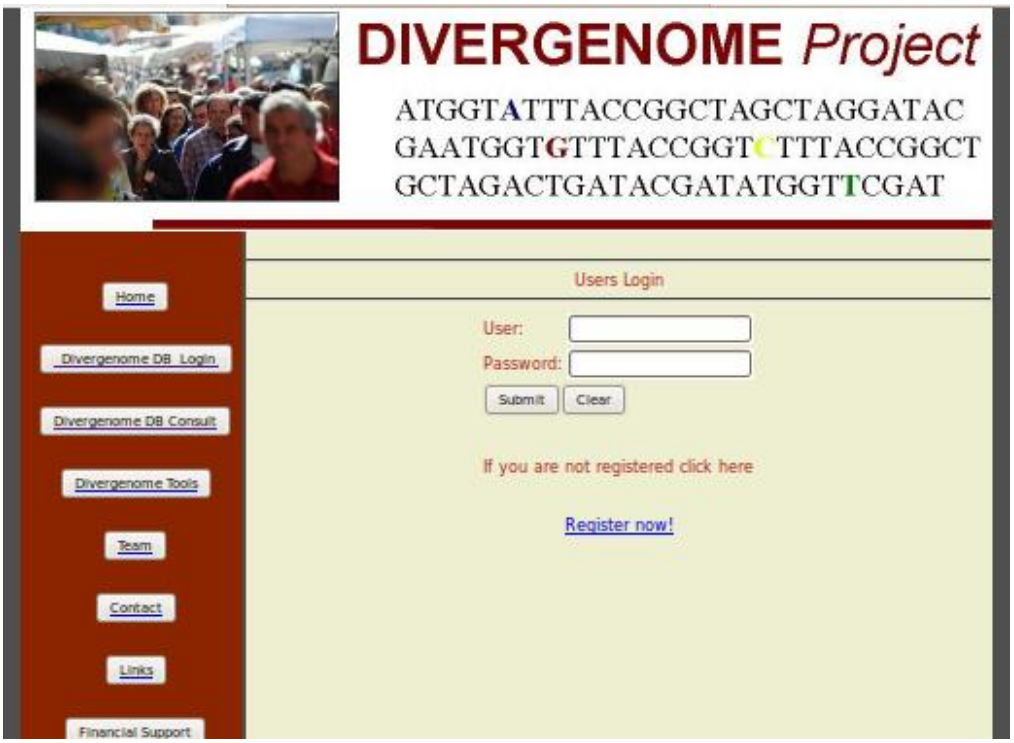


Figura 3: Tela para login do usuário no sistema DIVERGENOME

Após o *login* são verificados os níveis de permissão do usuário e a tela de apresentação é disponibilizada de acordo com o nível de permissão do usuário. Ao administrador é dado o maior nível de controle sobre o banco de dados e sobre os usuários. Os ícones disponíveis são:

- Register New User
- Register New Project
- Register New Project Member
- View Projects
- Manager User Privileges- Register New File in the Project

Os ícones para administradores serão disponibilizados conforme apresentado na Figura 4.

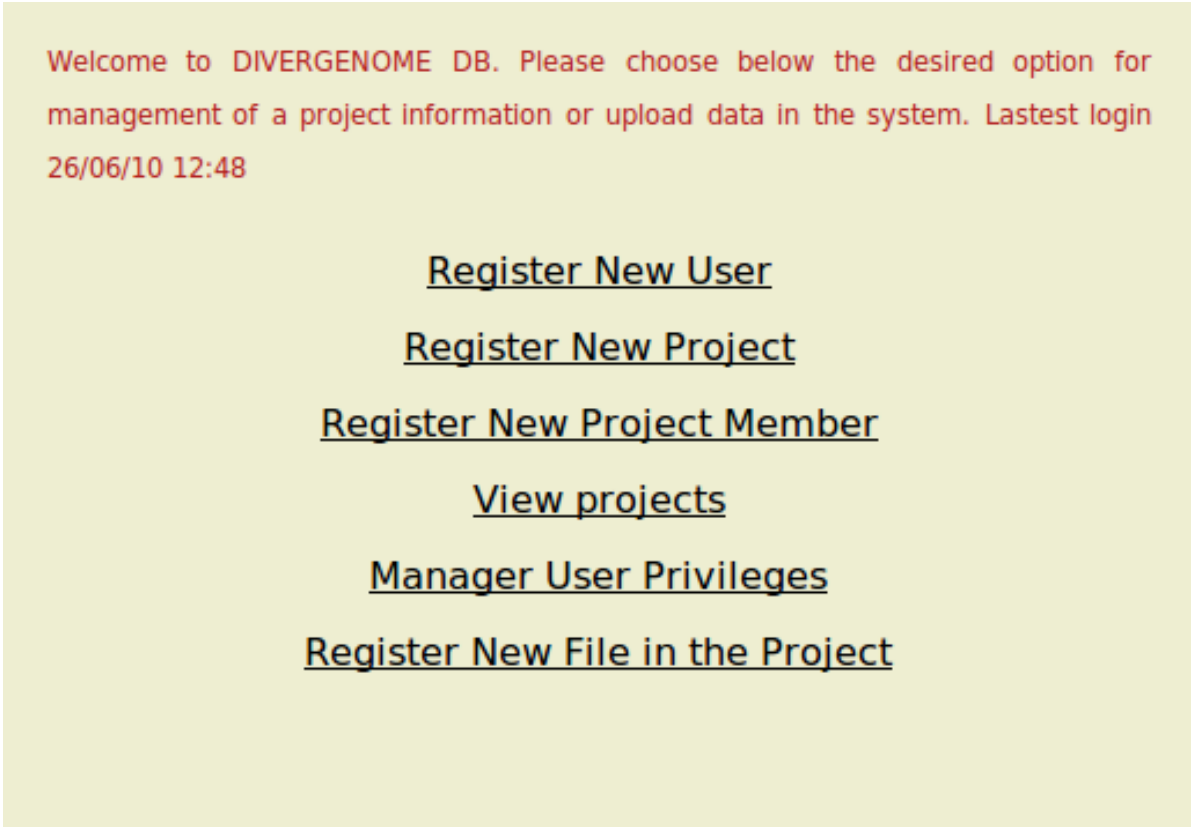


FIGURA 4: Tela de disponibilização dos recursos para os administradores

Para os coordenadores (pesquisadores) são disponibilizados os ícones: register new project, register new project member, view project e register new file in Project – Figura 5.

Figura 5: Tela de disponibilização dos recursos para coordenadores (pesquisadores)

A opção de cadastrar membro ao projeto é usada para pesquisadores que desejam adicionar membros ao projeto e posteriormente quando for realizado o controle de dados privados, apenas esses coordenadores e os membros inseridos para um projeto em específico terão acesso a esses dados. Atualmente para os membros cadastrados em projetos é possível apenas a visualização de dados gerais de projetos.

A Tabela 2 fornece detalhes sobre os níveis de acesso conforme os privilégios.

	Administrador	Coordenador (Pesquisador)	Membros de projetos (Simples Usuários)
Register New User	X		
Register New Project	X	X	
Register New Project Member	X	X	
View Projects	X	X	X
Manager User Privileges	X		
Register New File in the Project	X	X	

Tabela 2. Níveis de acesso em relação aos privilégios dos usuários cadastrados.

3.1 - Detalhamento de cada ícone disponibilizados conforme o nível de permissão do usuário

3.1.2 – Register New User

Opção disponibilizada apenas para administradores.

O cadastro do usuário (Figura 6) pode ser feito diretamente pelo administrador através do *link* “Register New User” (Figura 4), ou pelo novo usuário através da tela de *login* de acesso (Figura 3), em que há um *link* para cadastro de usuário (Register Now). A diferença entre as duas opções reside no fato que o cadastro de usuário realizado por um simples usuário necessita de uma posterior aprovação pelo administrador (Manager User Privileges) – Figura 4. Já o cadastro realizado pelo administrador receberá aprovação direta.

Figura 6: Tela de cadastro do usuário no sistema DIVERGENOME

3.1.2 – Register New Project

Opção disponibilizada para administradores e pesquisadores.

Através desta opção é possível cadastrar um novo projeto no DIVERGENOMEdb. É apresentando um formulário para preenchimento dos campos referente ao projeto (Figura 7). Um destes campos se refere ao *status* do projeto como público ou privado. Se público todos os dados inseridos para esse projeto poderão ser visualizados por qualquer usuário. Já os privados, posteriormente será realizado o controle de acesso aos dados inseridos para esse projeto, em que apenas o pesquisador que o registrou e os membros cadastrados por esse pesquisador para um projeto em específico terão acesso às informações inseridas no mesmo.

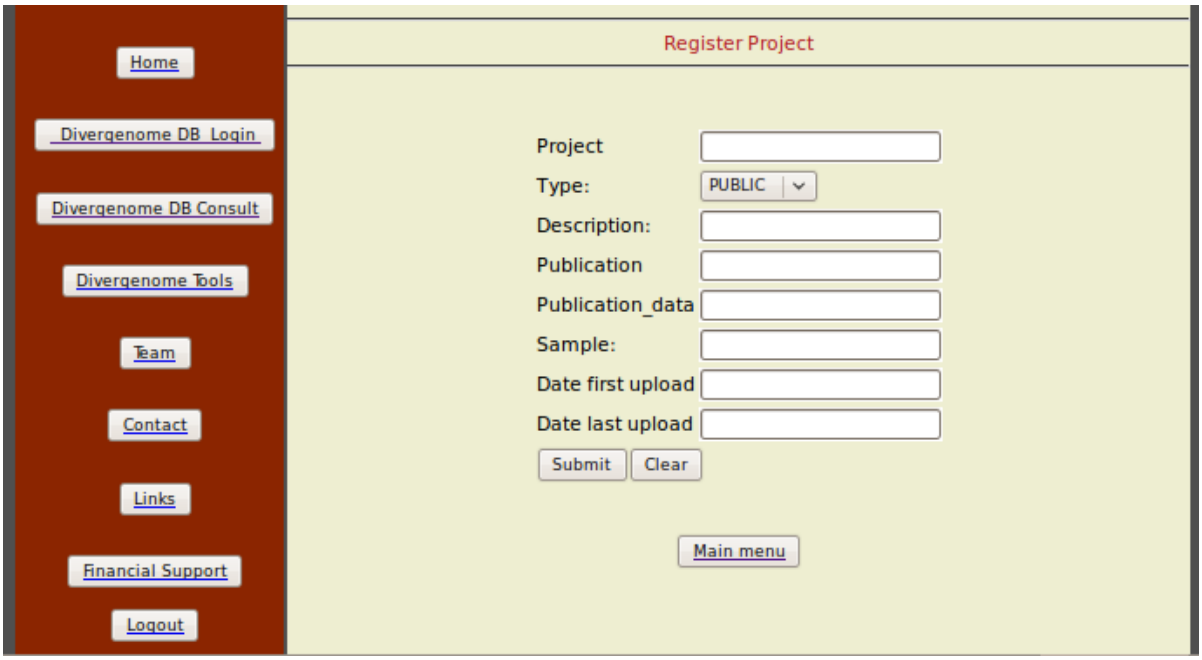


Figura 7: Tela de cadastro de projeto no sistema DIVERGENOME

3.1.3 – Manager User Privileges

Opção disponibilizada apenas para administradores.

O ícone *Manager User Privileges* (Figura 4) permite ao administrador controlar os acessos dos usuários previamente cadastrados e que estão aguardando aprovação (Figura 8). É possível excluir usuário, alterar o *status* do usuário, bem como o privilégio do mesmo. Para cadastro de um novo administrador, o atual administrador pode inseri-lo diretamente no sistema através do ícone *Register new use*r, ou alterar o privilégio de um usuário ou pesquisador que tenha se cadastrado recentemente e ainda aguarda a aprovação do administrador. Os tipos de *status* e privilégio para os usuários no sistema são:

1. Status:
- aguardando;
- concluído;
- recusado.

2. Privilégios:

- administrador;
- coordenador (pesquisador);
- membro de projeto (simples usuário).

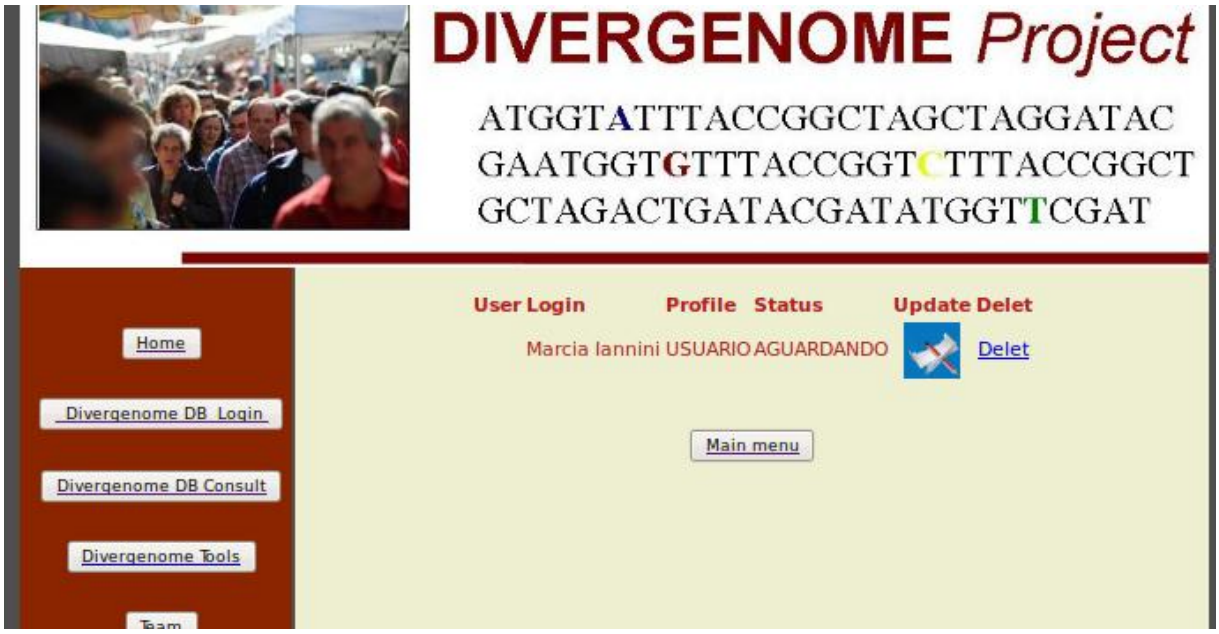


Figura 8: Tela de controle do usuário pelo administrador no sistema DIVERGENOME

3.1.4 – Regiser New Project Member

Opção disponibilizada para administradores e pesquisadores.

Essa opção permite que administradores e pesquisadores adicionem um novo membro a um projeto anteriormente cadastrado (Figura 9). No menu drop-down são apresentados os usuários cadastrados que já tiveram o acesso concluído e permitido pelo administrador. Nesse menu também são disponibilizados os projetos inseridos pelo administrador no sistema.

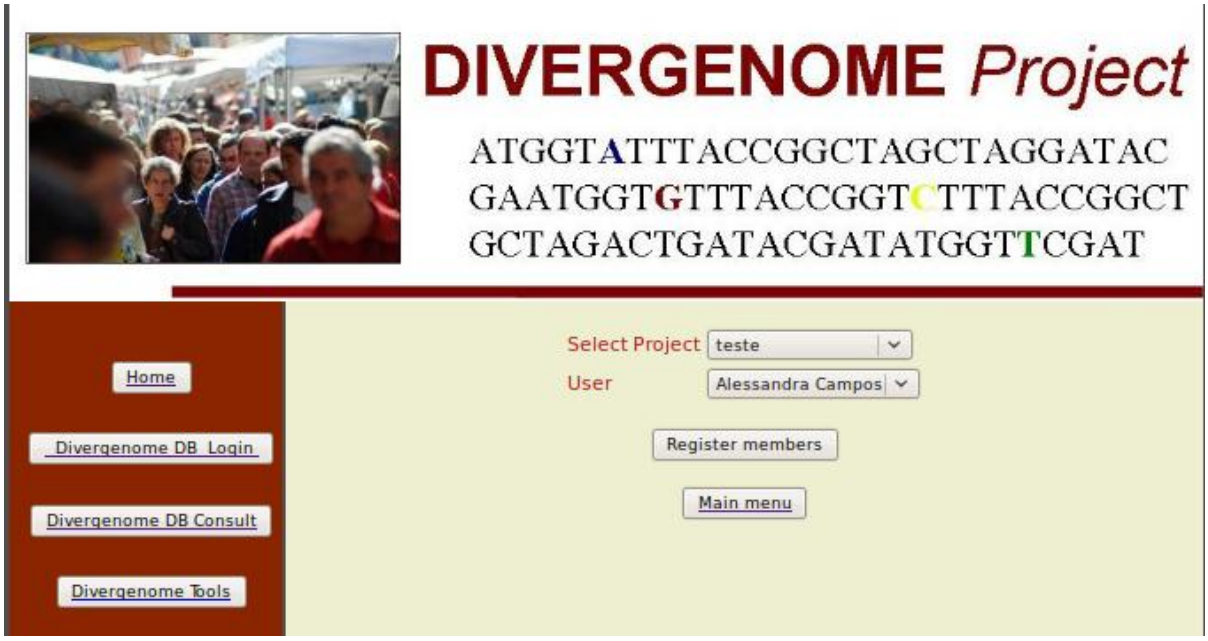


Figura 9: Tela de cadastro de novo membro no projeto no sistema DIVERGENOME

3.1.5 – View Project

Opção disponibilizada para administradores, pesquisadores e usuários.

Este ícone permite a visualização das informações cadastradas sobre um projeto. Nesta tela é permitido o acesso apenas aos dados de projetos ao qual o usuário foi inserido anteriormente como membro. A visualização de projetos por administradores e pesquisadores já está em funcionamento. A disponibilização desse ícone para usuários justifica-se pelo funcionamento posterior de controle de dados privados, em que após o cadastro prévio do usuário como membro de um projeto será possível a visualização dos campos de especificação desse projeto, bem como de seus dados inseridos pelo usuário-membro.

3.1.6 – Register New File in the Project

Opção disponibilizada para administradores e pesquisadores.

Através desse ícone é possível o cadastro de novo arquivo no banco de dados por pesquisadores e administradores. É possível o cadastro de dados apenas para projetos de autoria do pesquisador ou administrador logado. Veja detalhes sobre a inserção de dados no item 5 desse material.

4 – DIVERGENOMEdb CONSULT

Este ícone é utilizado para consultas no DIVERGENOMEdb. É possível realizar dois tipos de consultas no banco de dados:

- Buscas simples em tabelas únicas;
- Buscas avançadas relacionando duas, três ou quatro tabelas.

Em todas as buscas (simples ou avançadas) a tela disponibilizada ao usuário é disposta em sua parte superior por campos de filtro, e a parte inferior por *checkbox* (Figura 10). O *checkbox* é responsável pelos itens (colunas da tabela) a serem listados dentro da instrução SELECT. Os campos de filtro são os responsáveis pela recuperação de dados com critérios específicos, ou seja, acessam um subconjunto de linhas em uma tabela, especificando os critérios de seleção através da cláusula WHERE.

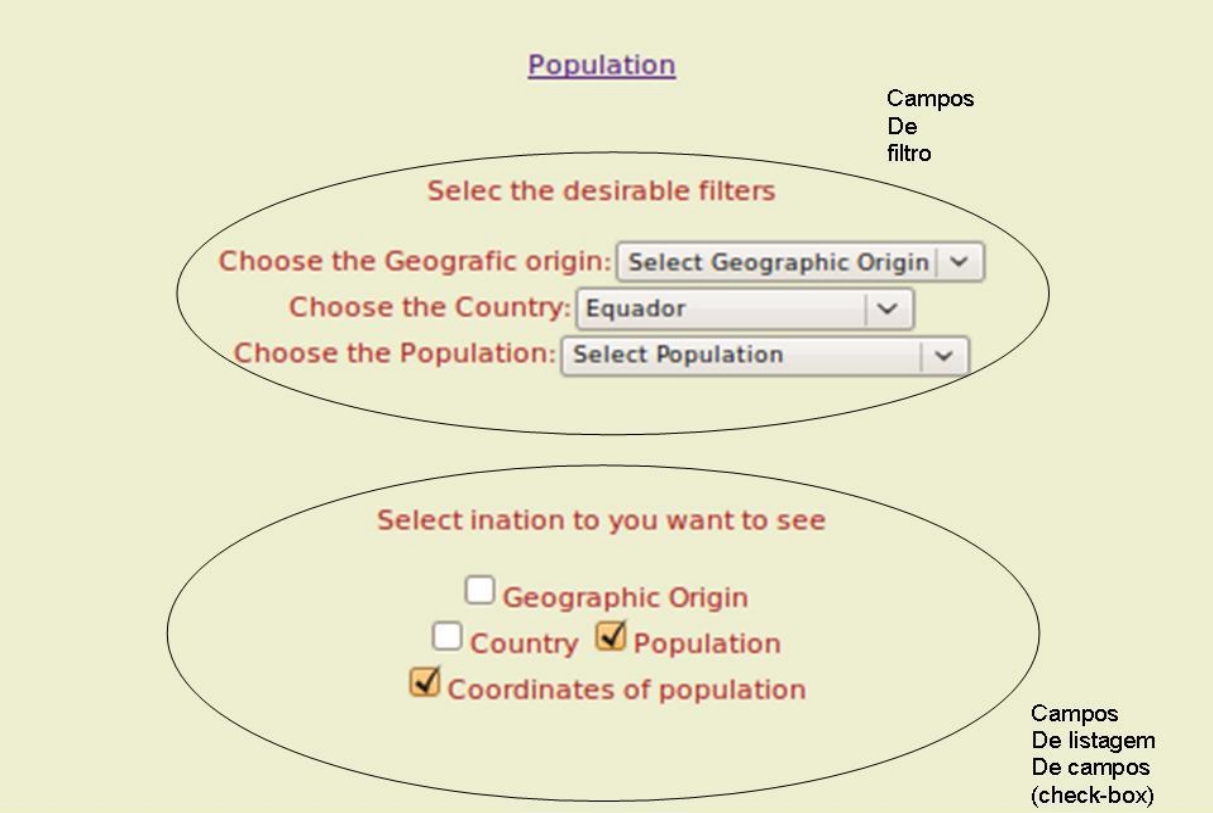
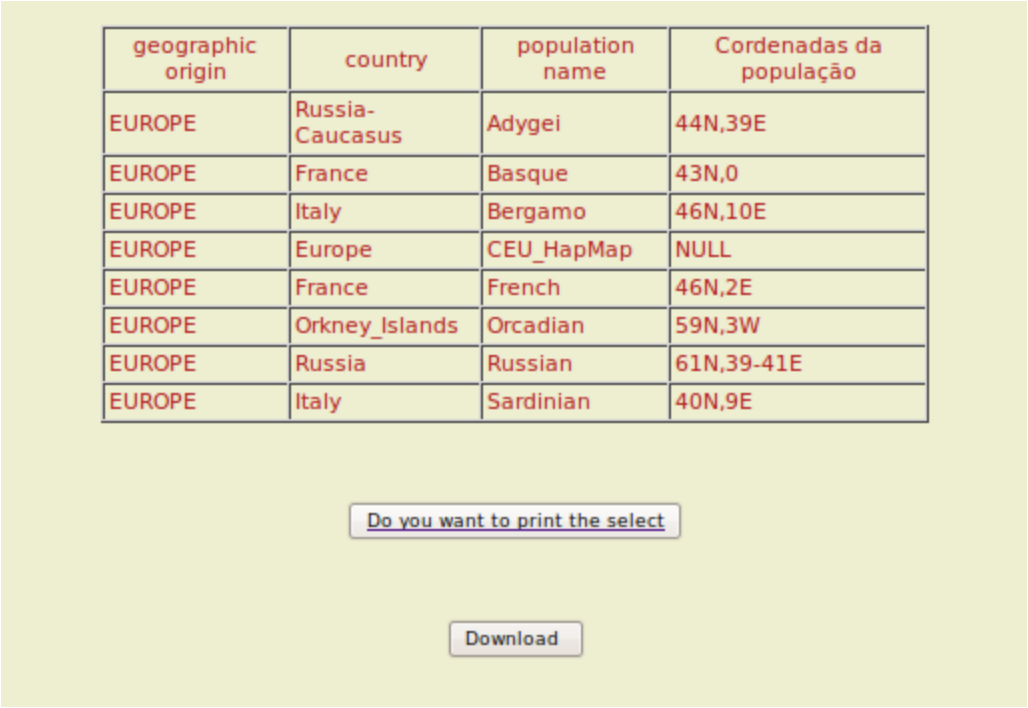


Figura 10: Tela visualizando os campos de filtros e os campos a serem listados, para montagem do select

Ambas as buscas são possíveis gerar um arquivo (registro.txt) com os valores retornados da consulta. Também é possível gerar a linha de comando do select em um arquivo em formato de texto (select.txt). Estes links são disponibilizados após a apresentação total dos resultados (Figura 11).



geographic origin	country	population name	Cordenadas da população
EUROPE	Russia-Caucasus	Adygei	44N,39E
EUROPE	France	Basque	43N,0
EUROPE	Italy	Bergamo	46N,10E
EUROPE	Europe	CEU_HapMap	NULL
EUROPE	France	French	46N,2E
EUROPE	Orkney_Islands	Orcadian	59N,3W
EUROPE	Russia	Russian	61N,39-41E
EUROPE	Italy	Sardinian	40N,9E

Figura 11: Tela de apresentação dos resultados com botões disponíveis para impressão da linha de comando do select gerado e do botão para *download* do resultado apresentado inicialmente na tela.

4.1 – Para buscas em tabela única

Os possíveis tipos de buscas realizadas em uma única tabela são apresentados na FIGURA 12.

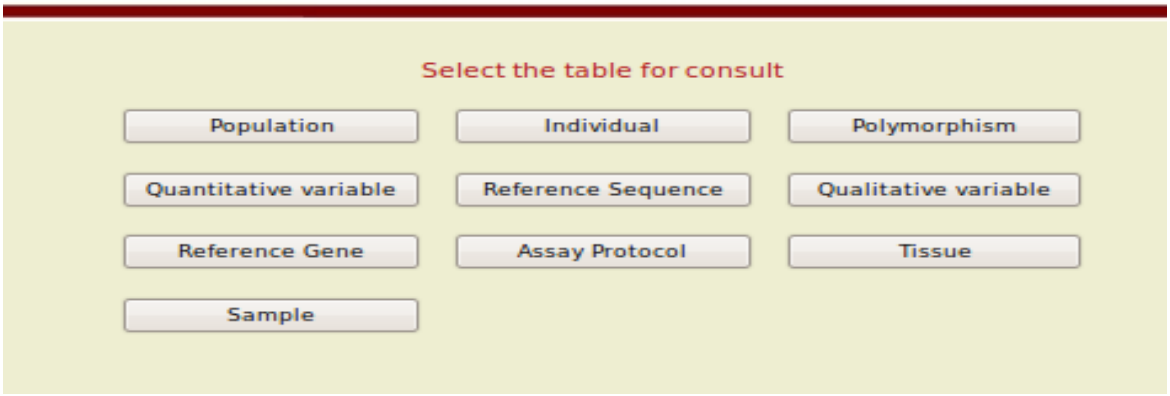


Figura 12: Tela exibindo as possíveis consultas por tabela única no sistema DIVERGENOME

Nas buscas em tabela única é feita a validação de campos. Verifica-se se pelo menos um campo e um filtro foram selecionados para montagem posterior da cláusula de instrução SELECT. Caso a condição pré-estabelecida não tenha sido atendida, a mensagem de alerta será acionada (FIGURA 13 – FIGURA 14).

Figura 13: Tela apresentando a mensagem de alerta quando um filtro não é selecionado

O mesmo ocorrerá se um dos campos do *checkbox* não for selecionado (Figura 14).

Figura 14: Tela evidenciando a mensagem de tela apresentada quando um item do *checkbox* não foi selecionado

Os *scripts* de apresentação dos resultados da busca montam a instrução SELECT conforme os campos selecionados, devendo haver pelo menos um campo do *checkbox* e um filtro selecionados.

As consultas nas tabelas quantitative_variable, qualitative_variable e tissue não permitem filtragem, por isso suas buscas são direcionadas diretamente para a tela de resultado e todos os campos inseridos no banco de dados são apresentados.

4.2 – Para buscas relacionando tabelas

É possível realizar buscas avançadas utilizando os campos disponíveis em duas, três ou quatro tabelas (FIGURA 15).

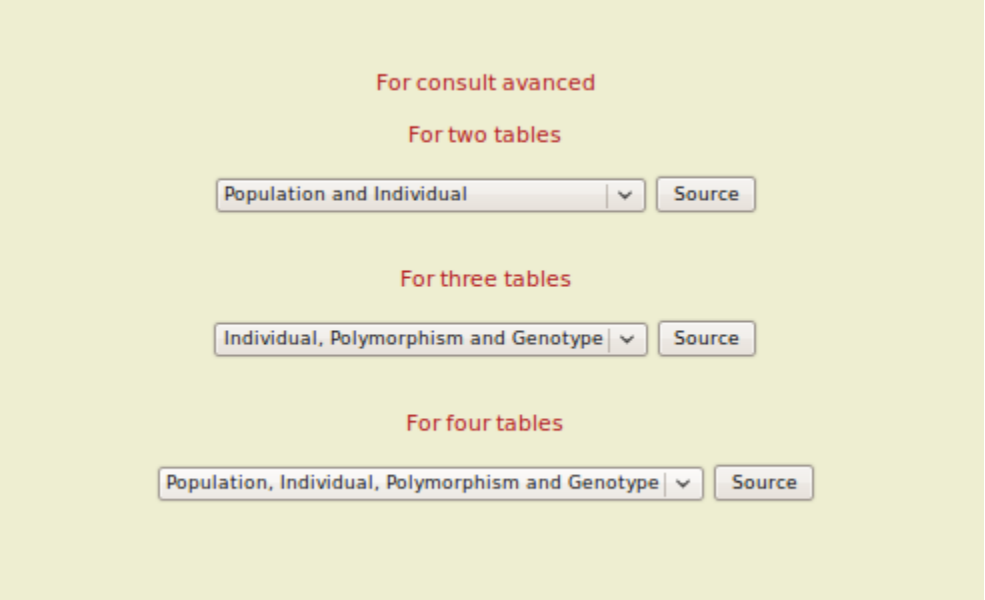


Figura 15: Tela evidenciando os tipos de consultas avançadas. Acesse o menu drop-down para visualizar todas as possíveis buscas entre duas, três ou quatro tabelas.

É possível executar as seguintes combinações:

Busca relacionando duas tabelas:
- Population e individual;
- Individual e Quantitative variable;
- Individual e Qualitative variable;
- Individual e Sample;
- Sample e Tissue;
- Polymorphism e Genotype;
- Sample e Population;
- Polymorphism e Reference Gene;
- Polymorphism e Reference Sequence.

Busca relacionando três tabelas:
- Individual, polymorphism e genotype;
- Population, individual e variable;
- Population, individual e sample.

Obs.: A procura avançada relacionando as tabelas *individual*, *polymorphism* e *genotype* foi especificada como valor *default*, por se tratar da busca mais freqüente que outras opções de buscas relacionando três tabelas.

Busca relacionando quatro tabelas:
Population, individual, polymorphism e genotype.

Através do menu drop-down é possível a escolha das buscas relacionando duas, três ou quatro tabelas. Após escolha, aparecerá uma tela (Figura 16) contendo um *link* para as duas, três ou quatro tabelas conforme o relacionamento escolhido na tela apresentada na Figura 15. Tomemos como exemplo a busca relacionando as tabelas *population* e *individual*.

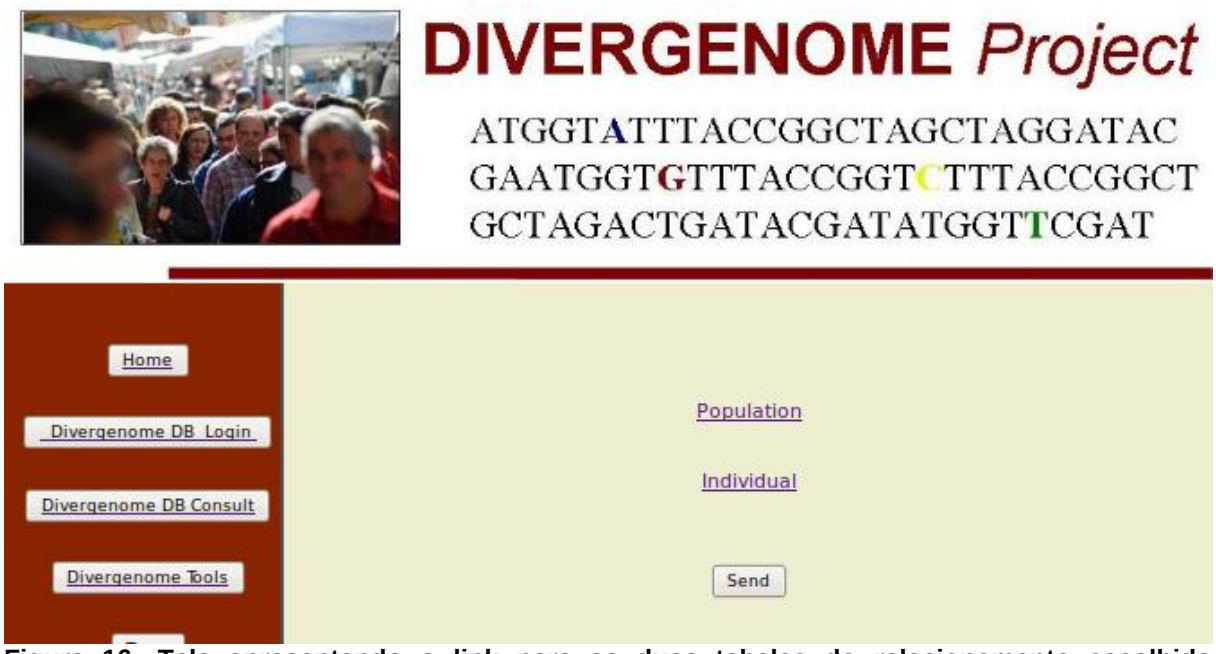


Figura 16: Tela apresentando o link para as duas tabelas do relacionamento escolhido anteriormente (Figura 15).

Foi disponibilizado um recurso que permite exibir ou ocultar os campos de filtragem e de listagem das tabelas. Quando acionada, os dados disponíveis para a montagem do SELECT são apresentados, ou seja, quando selecionado uma das tabelas, haverá a expansão da mesma, apresentando assim os campos de filtros e *checkbox* para seleção (Figura 17).

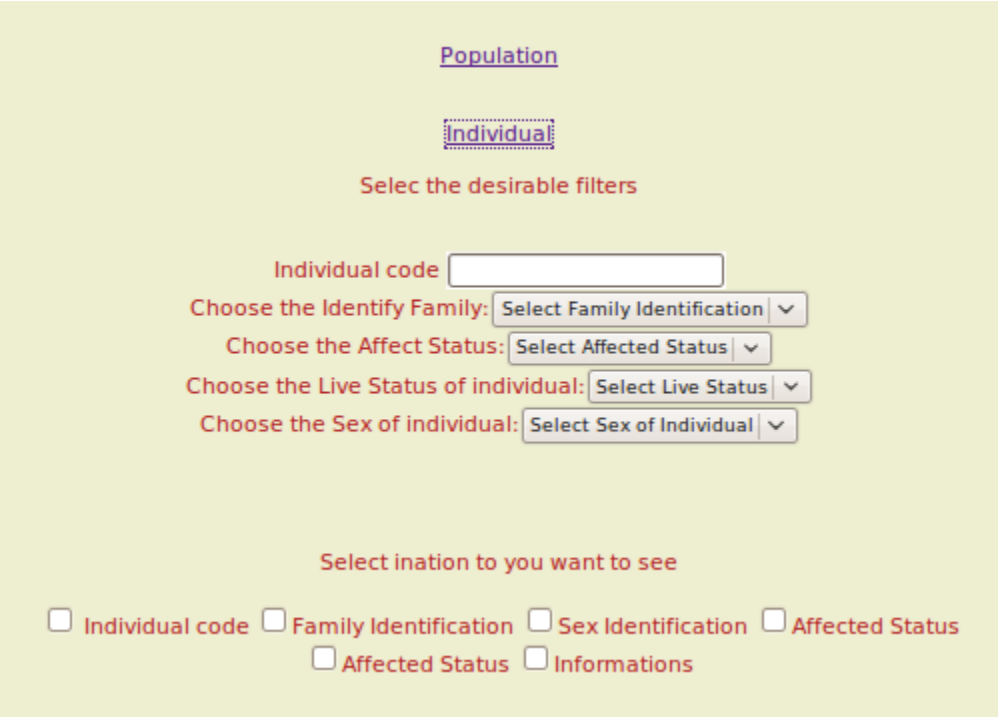
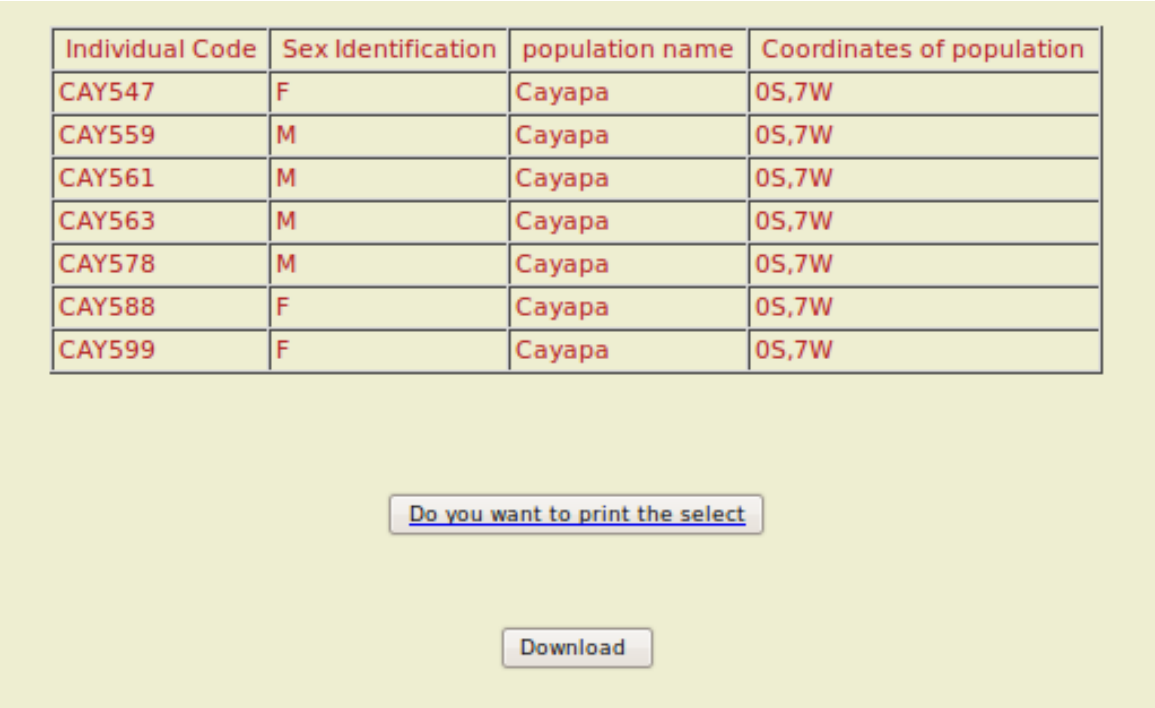


Figura 17: Tela visualizando os campos de filtro e de busca da tabela *individual*. Repare que o link para a tabela *population* continua ativo na parte superior, porém com os seus campos ocultos. Se clicarmos no link para *population*, haverá a expansão dos dados disponíveis de filtro e seleção para a tabela *population* e os campos da tabela *individual* se tornarão ocultos.

A tecla “Send” deve ser acionada após a escolha dos respectivos campos de seleção e filtro das tabelas disponibilizadas, conforme o relacionamento escolhido.

Os *scripts* de apresentação dos resultados de busca montam a instrução SELECT conforme os campos selecionados. A busca é feita diretamente nas tabelas correspondentes. O resultado para a busca realizada pode ser visualizada na Figura 18.

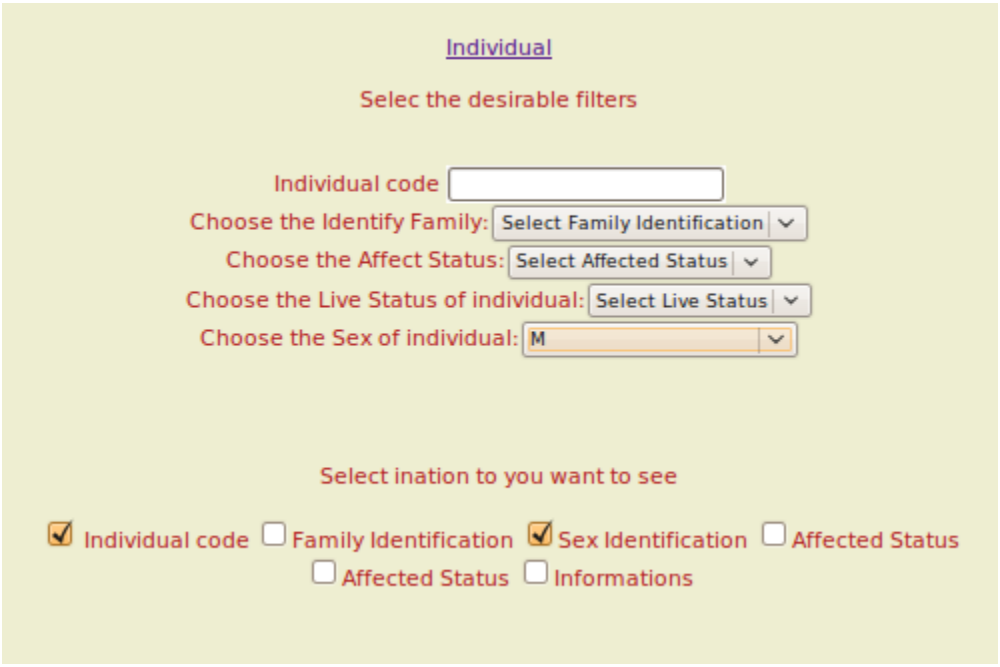


Individual Code	Sex Identification	population name	Coordinates of population
CAY547	F	Cayapa	0S,7W
CAY559	M	Cayapa	0S,7W
CAY561	M	Cayapa	0S,7W
CAY563	M	Cayapa	0S,7W
CAY578	M	Cayapa	0S,7W
CAY588	F	Cayapa	0S,7W
CAY599	F	Cayapa	0S,7W

Figura 18: Tela visualizando o resultado da busca solicitada utilizando o relacionamento entre as tabelas *population* e *individual*. Os campos de filtro e de busca foram: na tabela *population* o filtro de busca foi pelo país Equador e os campos de visualização foram a população e coordenada geográfica. Na tabela *individual*, não houve campo de filtro e os campos de visualização foram sexo e o código dos indivíduos. Resumindo: foi solicitado que me retornasse todos os indivíduos, bem como o sexo e a população a que pertencem, apresentando também a coordenada geográfica da população correspondente, para todos aqueles que estejam cadastrados no banco de dados e que sejam do país Equador. Se houvesse mais de uma população do Equador cadastrada, estas seriam também apresentadas.

A mesma busca pode ser feita usando mais de um filtro – Figura 19.

Tela A



Tela B

Figura 19: Telas visualizando os campos disponíveis para a tabela *population* (B) e para a tabela *individual* (A)

O select construído nessa busca:

SELECT pop.coordinates_of_population_sample, pop.population_name (B1), ind.individual_code, ind.sex (A1) FROM population pop, individual ind WHERE pop.id_population = ind.population_id_population and pop.country = Equador (B2) and ind.sex = M (A2)

A – dados extraídos da tabela *individual*
B – dados extraídos da tabela *population*
1 – dados obtidos do checkbox
2 – dados obtidos do campo de filtro

O resultado dessa busca é apresentado na Figura 20.

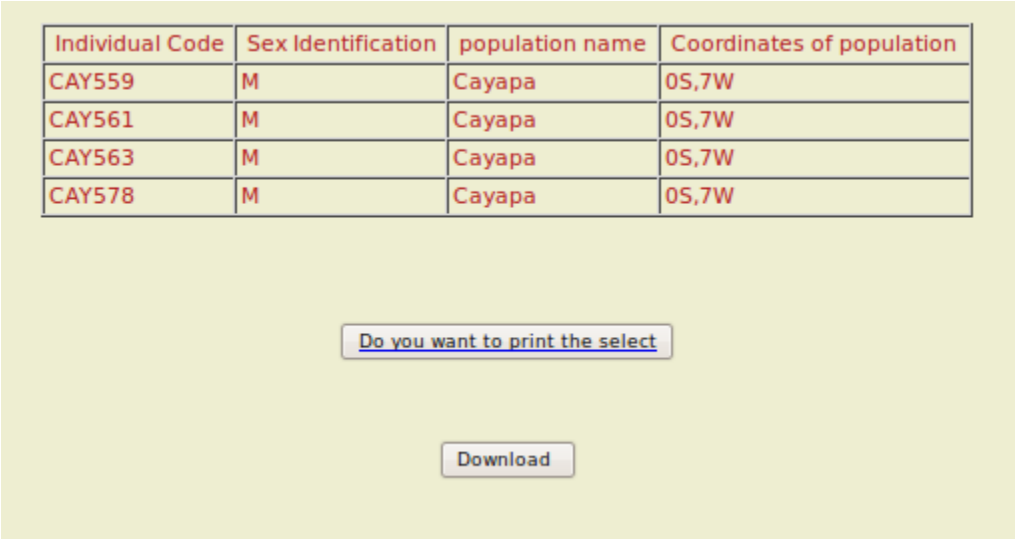


Figura 20: Tela de apresentação do resultado da busca. Observe que agora obtivemos a lista apenas dos indivíduos do sexo masculino, conforme solicitado como filtro da busca.

Como explicado anteriormente é possível gerar um arquivo 'txt' contendo o select gerado na consulta (Figura 21), bem como *download* do resultado em arquivo 'txt' no formato padrão de banco de dados (Figura 22).

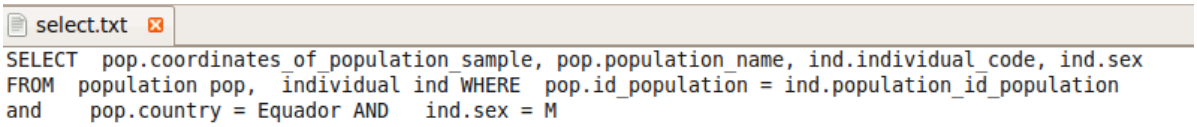


Figura 21: Tela apresentando o arquivo (select.txt) contendo a linha de comando select gerado na busca.

Figura 22: Tela apresentando o arquivo 'registro.txt' contendo o resultado da busca anteriormente apresentado na tela como uma tabela.

4.3 – Para buscas relacionando tabelas filtrando por arquivo

Para as buscas avançadas também é possível fazer pesquisas em que o filtro não é apenas um valor único, mas um arquivo contendo um conjunto de valores deste campo de filtro. Esse tipo de busca é disponibilizado para alguns relacionamentos de tabelas em específico. Um exemplo é dado para buscas em que há a necessidade de filtragem por diversos códigos de polimorfismos. Para esse tipo de busca há a necessidade de utilizar uma sintaxe que permita juntar valores de duas (*ou mais*) colunas em um único resultado. Isso é obtido pelo comando não muito comum, mas muito útil do *SQL* que é o **UNION.**

Na Figura 24 é apresentada a interface para a consulta recuperando diversos valores de *polymorphism code* disponíveis em um arquivo fornecido pelo usuário.

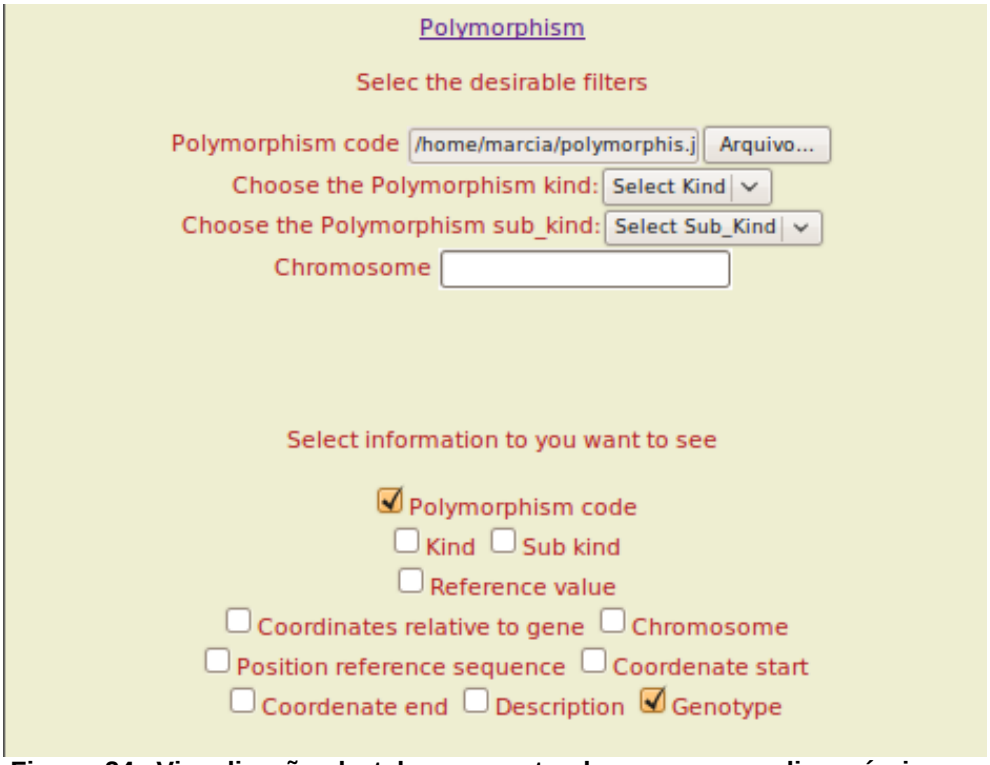


Figura 24: Visualização da tela apresentando os campos disponíveis para filtragem e seleção na tabela *polymorphism*. No campo de filtragem de busca *polymorphism code*, o usuário fornece o nome do arquivo que contém vários códigos de polimorfismos.

Essa busca se refere ao relacionamento de quatro tabelas, *polymorphism*, *genotype*, *individual* e *population*. Foi solicitada ao Divergenome database a seleção de todos os indivíduos, bem como a população e origem geográfica as quais tais indivíduos pertencem, seus genótipos para os polimorfismos contidos no arquivo de entrada pelo usuário. Esse arquivo em específico contém nove códigos de polimorfismos – Figura 25.

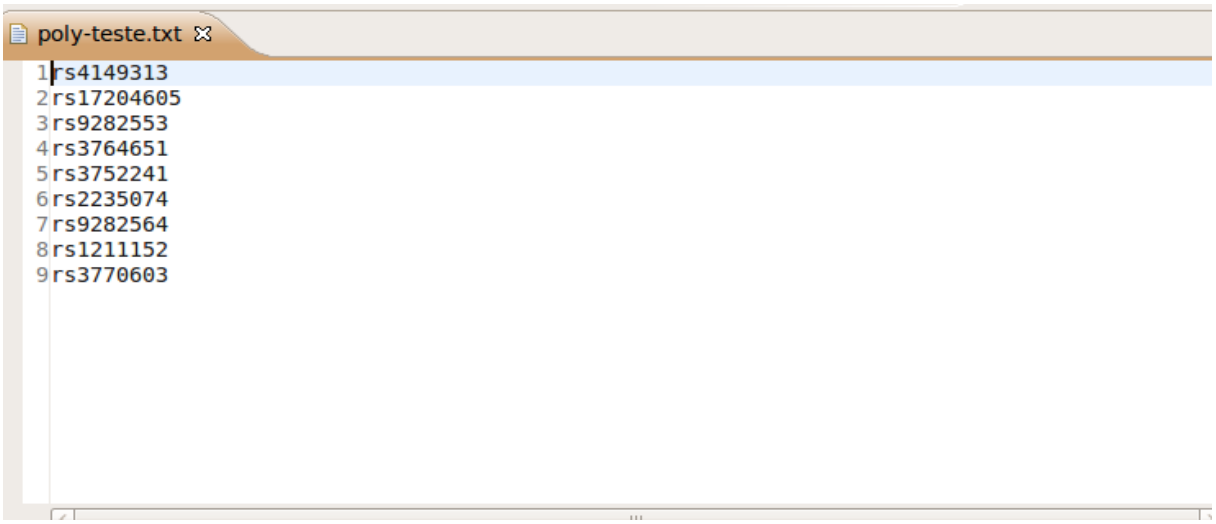


Figura 25: Visualização do arquivo contendo os nove códigos de polimorfismos para pesquisa no banco de dados.

A consulta no DIVERGENOME database retornará a tela contendo a tabela com os resultados como visualizado na Figura 26.

HGDP00672	EUROPE	Sardinian	rs4149313	AG
HGDP00671	EUROPE	Sardinian	rs4149313	AG
HGDP00670	EUROPE	Sardinian	rs4149313	AG
HGDP00669	EUROPE	Sardinian	rs4149313	AA
HGDP00668	EUROPE	Sardinian	rs4149313	AA
HGDP00667	EUROPE	Sardinian	rs4149313	AG
HGDP00666	EUROPE	Sardinian	rs4149313	AA
HGDP00665	EUROPE	Sardinian	rs4149313	AA
CAY561	AMERICA	Cayapa	rs17204605	CC
QT073	AMERICA	Quetchua	rs17204605	CC
QT059	AMERICA	Quetchua	rs17204605	CC
QT137	AMERICA	Quetchua	rs17204605	CC
CAY559	AMERICA	Cayapa	rs17204605	CC
CAY547	AMERICA	Cayapa	rs17204605	CC
QT134a	AMERICA	Quetchua	rs17204605	CC
QT134b	AMERICA	Quetchua	rs17204605	CC
QT143	AMERICA	Quetchua	rs17204605	CC
QT145	AMERICA	Quetchua	rs17204605	CC
QT110	AMERICA	Quetchua	rs17204605	CC

Figura 26: Tela apresentando parte dos resultados retornados da consulta.

No select em questão trabalhamos com a busca filtrando por um arquivo que continha apenas nove códigos de polimorfismos. A mesma pesquisa pode ser realizada com centenas de códigos de polimorfismos como filtro. A vantagem desse recurso é que podemos em única busca realizar várias consultas. Da mesma forma que as demais buscas, o usuário pode gerar um arquivo no formato texto (select.txt) contendo a linha de comando gerada na consulta, bem como pode gerar um arquivo txt (registros.txt) contendo o mesmo resultado apresentado na tela, para uso posterior nas ferramentas disponíveis no DIVERGENOME TOOLS.

5 – INSERÇÕES DE DADOS

O cadastro de dados na database pode ser realizado por pesquisadores ou administradores do banco de dados.

Para a inserção de dados, as tabelas podem ser classificadas como:
- necessariamente públicas;
- públicas ou privadas;
- sem recuperação de chaves estrangeiras;
- com recuperação de chaves estrangeiras.

A Tabela 3 apresenta um quadro resumido com as possíveis classificações das tabelas.

Classificação / Tabelas	Pública	Pública ou privada	Sem recuperação de chaves estrangeira	Com recuperação de chave estrangeira (em destaque as chaves recuperadas)
Population	X		x	
Individual	X			X (population)
Tissue	X		x	
Assay_protocol	X		x	
Polymorphism	X			X (haplotype, reference_gene, reference_sequence)
Reference_gene	X		x	
Reference_sequence	X		x	
Qualitative_variable	X		X	X (individual)
Quantitative_variable	X		x	X (individual)
Genotype		x		X (individual, polymorphism, assay_protocol, haplotype_evaluation, project)
Haplotype		x		X (haplotype_evaluation, project)
Haplotype_evaluation		x		X (project)
Sample		x		X (individual, tissue, project)

Tabela 3: Classificação das tabelas quanto ao status (exclusivamente pública ou não) e quanto a recuperação ou não de chaves relacionadas a outras tabelas.

Além disso, a inserção pode ser feita de duas formas:

- Recuperação de id's (chave que relacionam tabelas) por seleção de dados recuperados do banco de dados e disponibilizados em menu drop-down;

- recuperação automática de id's.

5.1 – Tabelas exclusivamente públicas

Para as tabelas *individual*, *population*, *tissue*, *polymorphism*, *reference-gene*, *reference-sequence*, *assay protocol*, *quantitative-variable* e *qualitative-variable*, os dados inseridos serão necessariamente públicos. Já para as tabelas *genotype*, *haplotype*, *haplotype-evaluation* e *sample*, os dados podem ser públicos ou privados. Neste caso, os dados estarão necessariamente ligados a um projeto específico de um pesquisador.

5.1.1 – Tabelas sem recuperação de chaves estrangeiras

Para tabelas em que não há recuperação de chaves estrangeiras (*population*, *tissue*, *assay_protocol*, *qualitative_variable*, *quantitative_variable*, *reference_sequence* e *reference_gene*), (Figura 27) há o direcionamento para tela em que o usuário escolhe a tabela, bem como o arquivo texto que deseja inserir (Figura 28). O arquivo deverá estar no formato padrão de banco de dados (dados estão dispostos em linhas e o campos em colunas separados por tabulação).

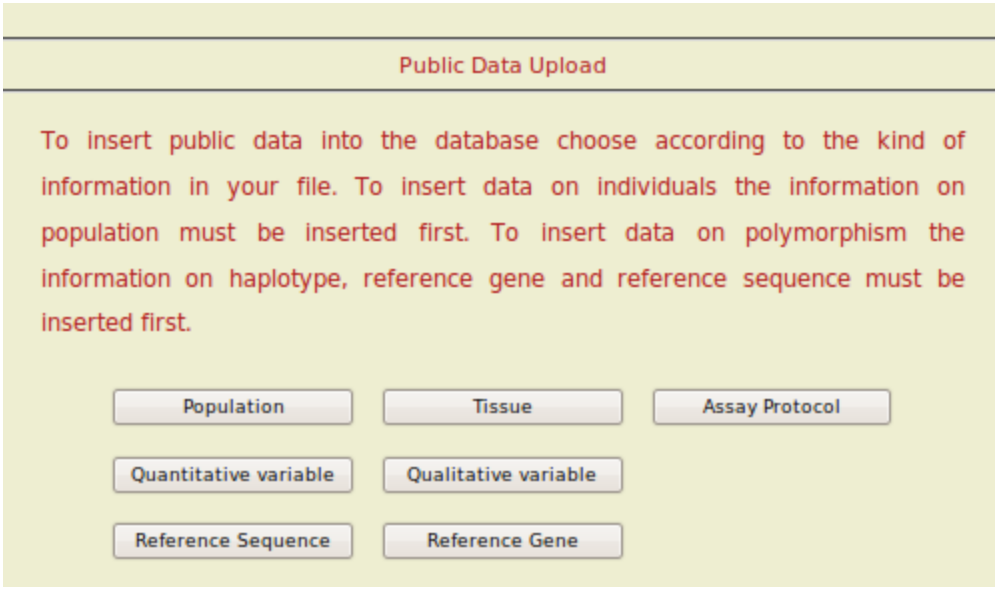


Figura 27: Botões para inserção em tabelas de dados necessariamente públicos e sem recuperação de chaves estrangeiras.

Figura 28: Tela de inserção de tabelas públicas sem recuperação de chaves estrangeiras.

Para os dados das tabelas *quantitave_variable* e *qualitative_variable* há a possibilidade de inserção apenas das variáveis, sem associação a um indivíduo específico, o que será feito na tela definida na Figura 27 ou há a possibilidade de após inserção dessas variáveis haver a junção ou associação da variável a um indivíduo (Figura 29B) ou vários indivíduos contidos em um arquivo (Figura 29C).
(Observação: Nestes dois últimos casos define-se a inserção de arquivo com recuperação de chave estrangeira – item 5.1.2 deste material. A Figura 29A apresenta a tela contendo os botões para esse tipo de inserção e a Figura 29D apresenta o esquema com o detalhamento dos dois tipos de inserção de variáveis quantitativas e qualitativas.)

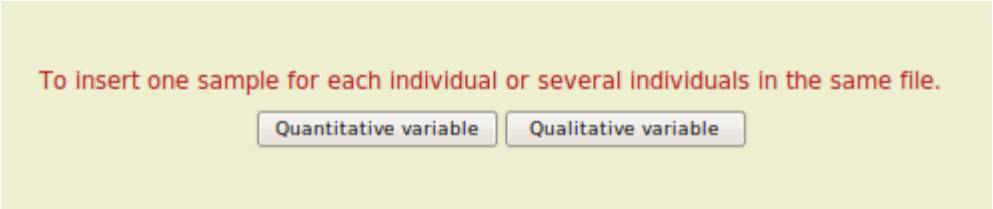


Figura 29A: Botões para inserção de uma variável qualitativa ou quantitativa para um ou vários indivíduos.

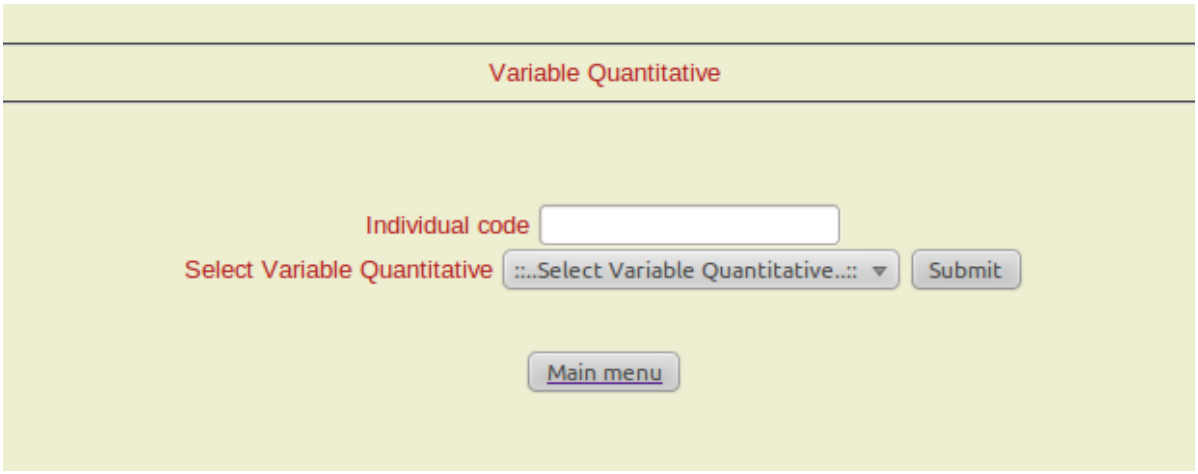


Figura 29B: Tela para associação de uma variável a um indivíduo.

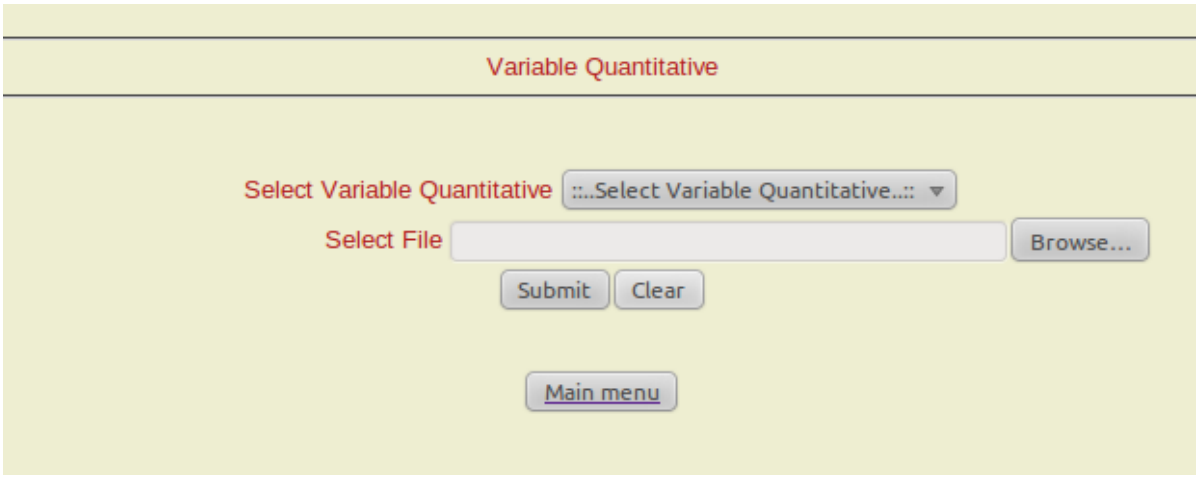


Figura 29C: Tela para associação de uma variável a vários indivíduos contidos em um arquivo.

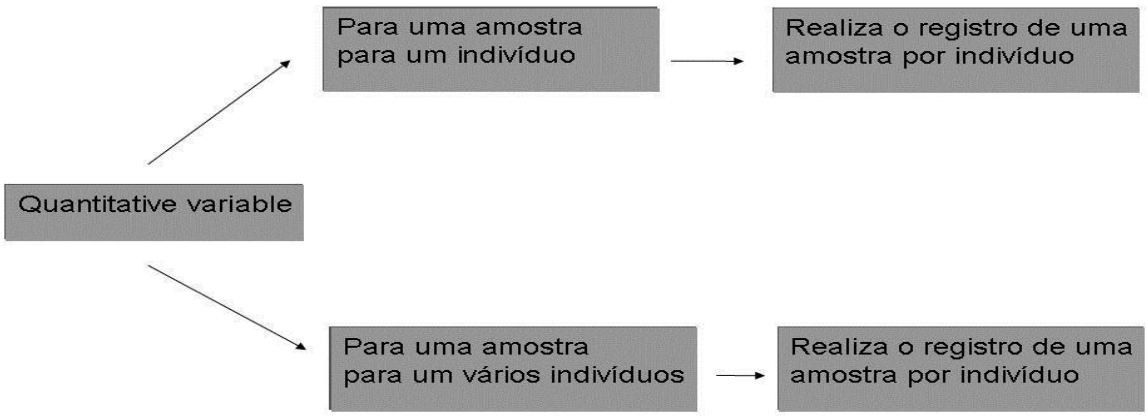


Figura 29D: Esquema de inserção de variáveis qualitativas e quantitativas para um ou vários indivíduos.

5.1.2 – Tabelas com recuperação de chaves estrangeiras

As tabelas *individual* e *polymorphism* são tabelas de dados necessariamente públicas em que há recuperação de chaves estrangeiras de relacionamentos com outras tabelas.

Para a tabela *individual* há a recuperação do campo population_id_population, chave estrangeira com a tabela *population*. O usuário deve selecionar a população para a qual se deseja inserir os dados dos indivíduos contidos no arquivo texto. Veja tela de inserção da tabela *individual* na Figura 30.

Figura 30: Tela de inserção do arquivo texto no formato padrão de banco de dados para a tabela *individual*. Selecionar a população para qual se deseja inserir os indivíduos e escolha o arquivo texto contendo os dados dos indivíduos.

Para a tabela de *polymorphism* há recuperação do campo haplotype_id_haplotype (chave estrangeira da tabela *haplotype*), reference_sequence_id_reference_sequence (chave estrangeira da tabela *reference_sequence*), reference_gene_id_reference_gene (chave estrangeira da tabela *reference_gene*). Na Figura 31 é apresentada a tela de inserção da tabela *polymorphism*.

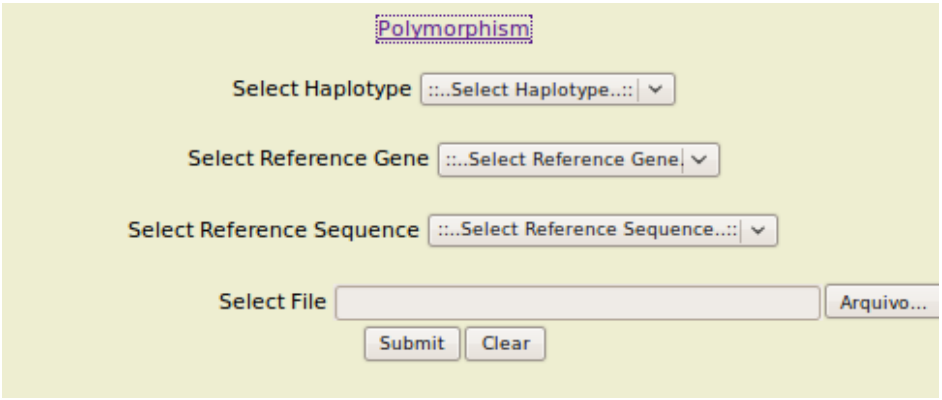


Figura 31: Detalhes para inserção de dados na tabela *polymorphism*.

5.2 – Tabelas públicas ou privadas

Os dados das tabelas *sample* e *genotype* podem ser públicos (visualizados por qualquer usuário) ou podem ser privados (visualizados por pesquisador de autoria do projeto ou membro cadastrado para o projeto). Essa é a proposta do

controle de dados, porém ainda em desenvolvimento. Atualmente todos os dados inseridos no banco de dados são considerados como públicos. Para ambas há recuperação de chaves estrangeiras de relacionamento com outras tabelas, sendo que o campo project_id_project é recuperado da tabela *project*. Na Figura 32 é apresentada a tela para inserção das tabelas que podem ser privadas.

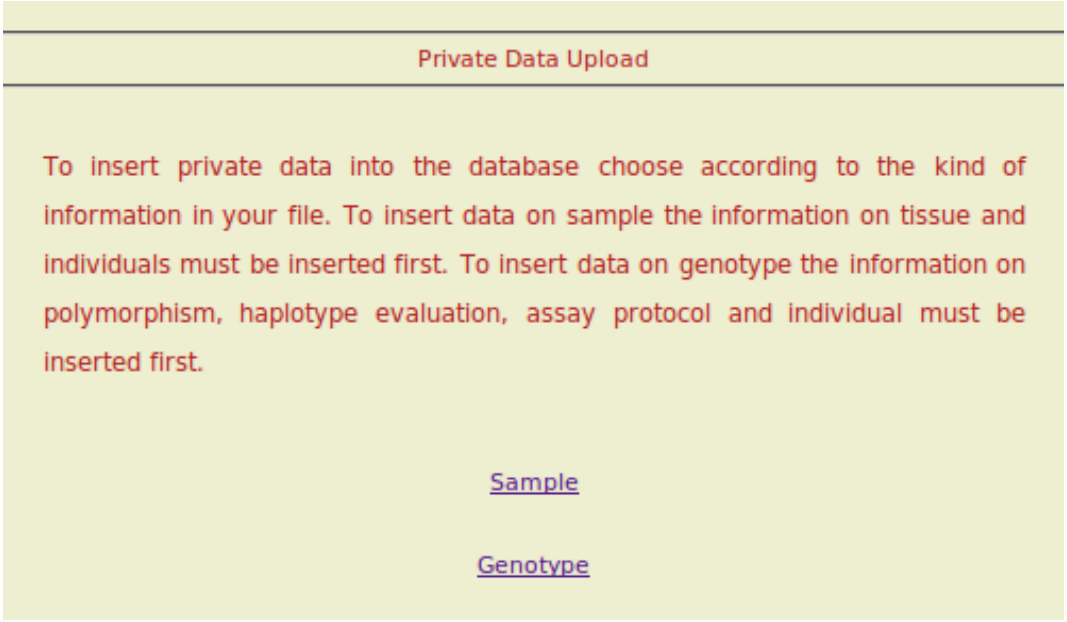


Figura 32: Tela para inserção de tabelas privadas na database.

Para a tabela *sample* há recuperação dos campos tissue_id_tissue (chave estrangeira da tabela *tissue*), individual_id_individual (chave estrangeira da tabela *individual*), além da recuperação do campo project_id_project (chave estrangeira da tabela *project*) para a qual se deseja inserir os dados. Na Figura 33 é apresentada a tela para inserção da tabela *sample*.

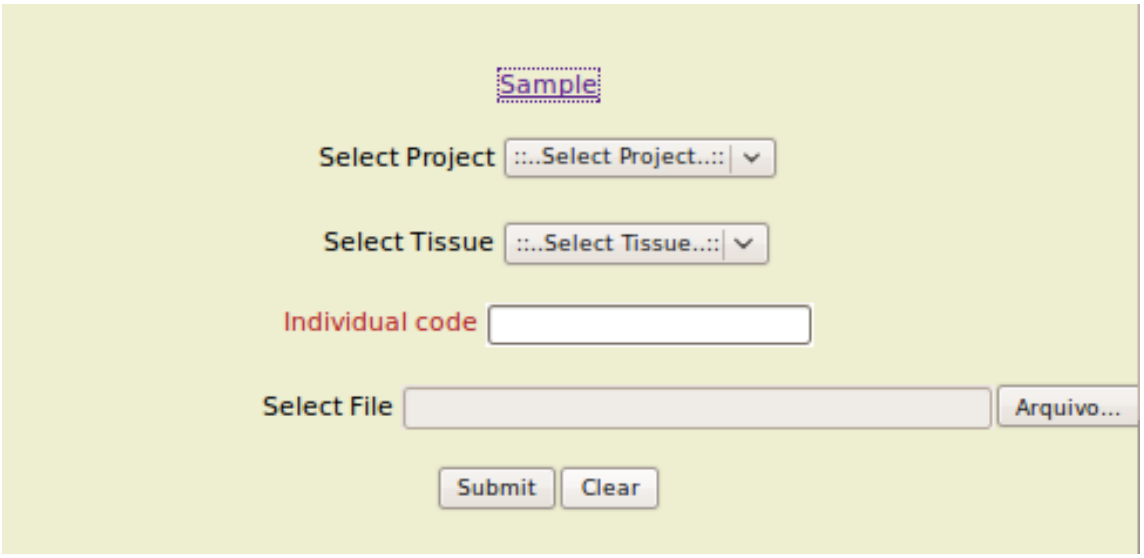


Figura 33: Tela de inserção da tabela *sample*.

Para a tabela *genotype* há recuperação dos campos polymorphism_id_polymorphism (chave estrangeira da tabela *polymorphism*), haplotype_evaluation_id_haplotype_evaluation (chave estrangeira da tabela *haplotype_evaluation*), assay_protocol_id_assay_protocol (chave estrangeira da tabela *assay_protocol*), individual_id_individual (chave estrangeira da tabela *individual*), além da recuperação do campo project_id_project (chave estrangeira da tabela *project*) para a qual se deseja inserir os dados. Na Figura 34 é apresentada a tela para inserção da tabela *genotype*.

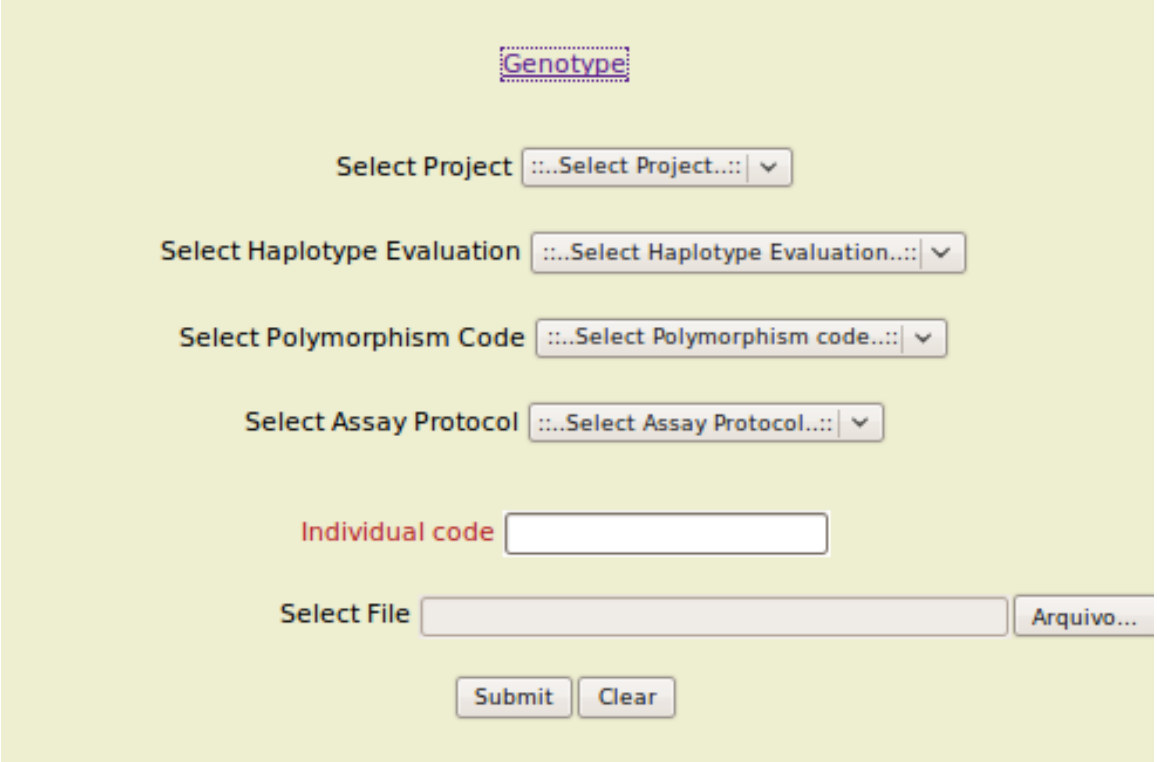


Figura 34: Tela de inserção da tabela *genotype*.

5.3 - Recuperação de id's (chave que relacionam tabelas) por seleção de campos recuperados do banco de dados e disponibilizados através de menu drop-down

É possível para cada entidade (tabela) que possui campos relacionados a outras tabelas (chaves estrangeiras) a recuperação de ids primários nas entidades de origem, através da seleção do valor da variável disponibilizada em um menu drop-down (valores retornados conforme os dados disponíveis no banco e que foram anteriormente inseridos). Neste caso, o arquivo de texto para inserção apresentará no atributo correspondente à chave estrangeira o valor NULL. O usuário deve selecionar um dos valores retornados do banco de dados para a qual se deseja inserir os campos presentes no arquivo texto. Neste caso, para todos os campos chave estrangeira do arquivo txt será inserido no banco de dados o código de identificação correspondente ao campo selecionado no menu drop-down. Tome

como exemplo a inserção de um arquivo de texto contendo atributos referente à tabela *individual*. Neste caso, o arquivo de texto para inserção apresentará no campo correspondente a chave estrangeira (neste exemplo o campo population_id_population) o valor NULL. O usuário deve selecionar a população (valores retornados do banco de dados) para a qual se deseja inserir os indivíduos (atributos presentes no arquivo texto). Veja tela para esse tipo de inserção na Figura 30. Neste caso, para todos os indivíduos do arquivo txt será inserido no banco de dados o código de identificação correspondente a população selecionada no menu drop-down (no nosso exemplo o id da população “Adygei – Figura 30”). Ou seja, uma população apenas para todos os indivíduos do arquivo inserido.

5.4 – Inserção de tabelas com recuperação automárica de id’s (chaves estrangeiras)

Para algumas tabelas que possuem campos relacionados a outras tabelas (chaves estrangeiras) é possível recuperar automaticamente os ids primários nas tabelas de origem (Figura 35). Na Figura 36 é apresentado um arquivo de exemplo para esse tipo de inserção.

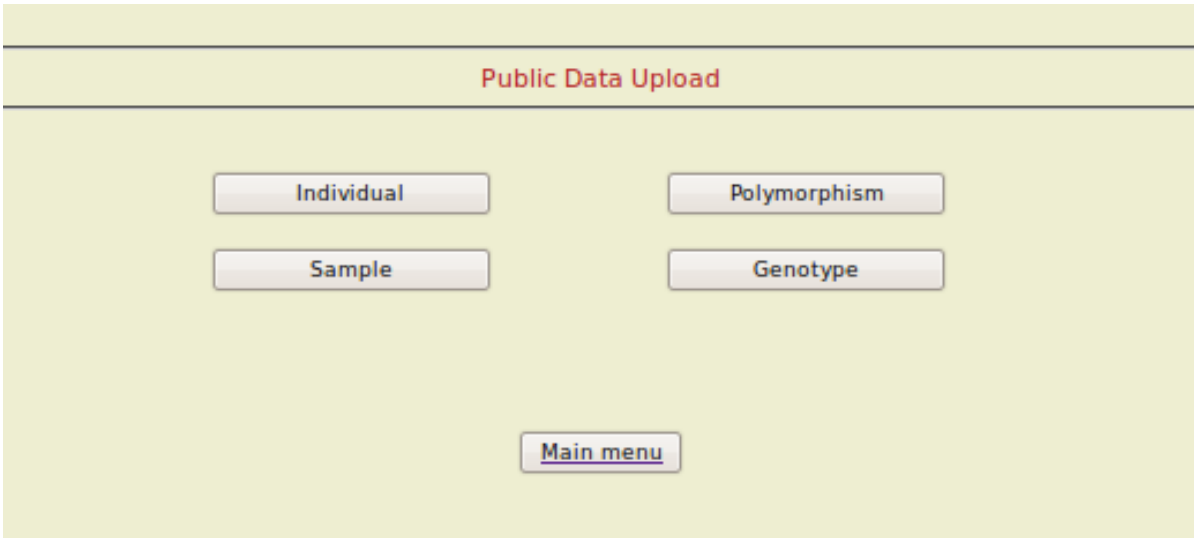


Figura 35: Tela evidenciando os ícones para recuperação automática de campos referentes a outras tabelas.

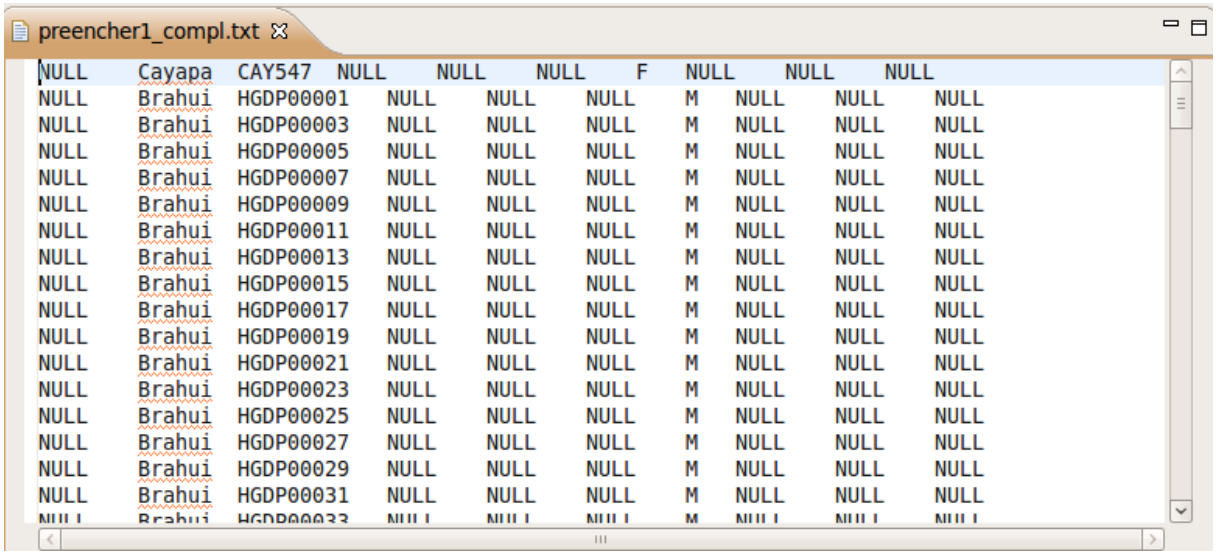
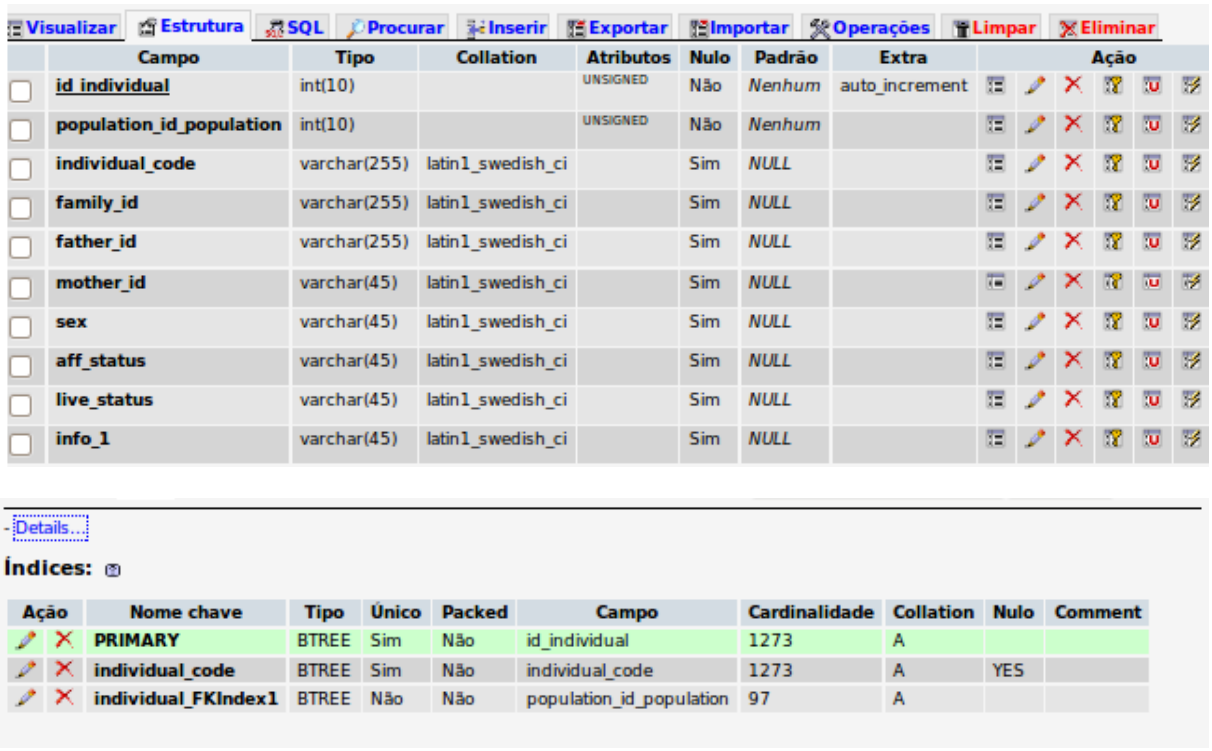


Figura 36: Tela apresentado um arquivo de inserção da tabela *individual*.

A estrutura da entidade *Individual* no DIVERGENOME database está representada na Figura 37.



Visualizar | Estrutura | SQL | Procurar | Inserir | Exportar | Importar | Operações | Limpar | Eliminar

	Campo	Tipo	Collation	Atributos	Nulo	Padrão	Extra	Ação
☐	id_individual	int(10)		UNSIGNED	Não	*Nenhum*	auto_increment	
☐	population_id_population	int(10)		UNSIGNED	Não	*Nenhum*		
☐	individual_code	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	family_id	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	father_id	varchar(255)	latin1_swedish_ci		Sim	*NULL*		
☐	mother_id	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	sex	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	aff_status	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	live_status	varchar(45)	latin1_swedish_ci		Sim	*NULL*		
☐	info_1	varchar(45)	latin1_swedish_ci		Sim	*NULL*		

Details...

Índices:

Ação	Nome chave	Tipo	Único	Packed	Campo	Cardinalidade	Collation	Nulo	Comment
	PRIMARY	BTREE	Sim	Não	id_individual	1273	A		
	individual_code	BTREE	Sim	Não	individual_code	1273	A	YES	
	individual_FKIndex1	BTREE	Não	Não	population_id_population	97	A		

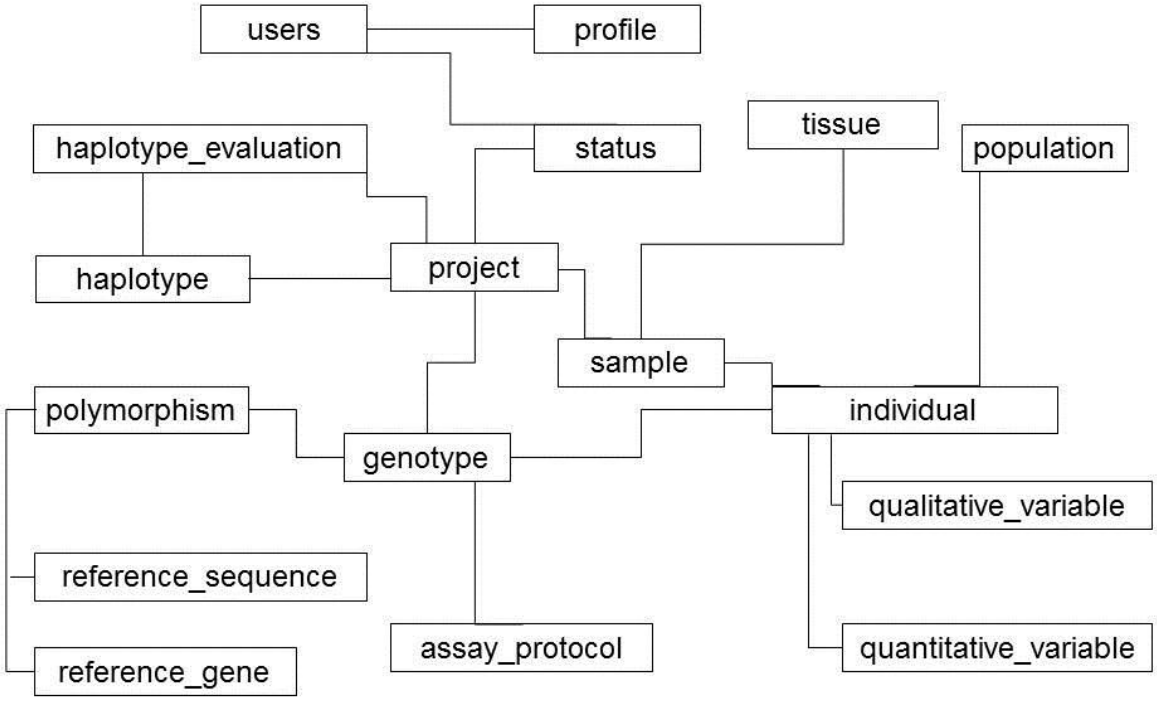
Figura 37: Tela apresentando a estrutura da entidade *Individual*.

No arquivo apresentado na Figura 36, a primeira coluna que recebe valores nulos se refere ao código de identificação do indivíduo (id-individual), que ao ser inserido no banco de dados receberá valor sequencial após o último valor contido na database. A segunda coluna se refere ao campo population-id-population (chave estrangeira com a tabela *population*). Para esta coluna, o script desenvolvido compara os valores do campo da segunda coluna do arquivo texto com as populações cadastradas no Divergenome Database, e retorna, caso encontre a população no banco de dados, o valor do id_population (código de identificação da população - valor sequencial). A inserção no banco de dados para este campo na tabela *individual* será do id_population (valor númerico de identificação da população na tabela *population*). Caso existissem nesta tabela outros campos que fossem chaves estrangeiras de outras tabelas, o procedimento de recuperação dos id's seria o mesmo. O resultado da inserção da tabela *individual* com a devida recuperação dos id's das populações está apresentado na Figura 38.

id_individual	population_id_population	individual_code	family_id	father_id	mother_id	sex	aff_status	live_status	info_1
1	18	CAY547	NULL	NULL	NULL	F	NULL	NULL	NULL
2	15	HGDP00001	NULL	NULL	NULL	M	NULL	NULL	NULL
3	15	HGDP00003	NULL	NULL	NULL	M	NULL	NULL	NULL
4	15	HGDP00005	NULL	NULL	NULL	M	NULL	NULL	NULL
5	15	HGDP00007	NULL	NULL	NULL	M	NULL	NULL	NULL
6	15	HGDP00009	NULL	NULL	NULL	M	NULL	NULL	NULL
7	15	HGDP00011	NULL	NULL	NULL	M	NULL	NULL	NULL
8	15	HGDP00013	NULL	NULL	NULL	M	NULL	NULL	NULL
9	15	HGDP00015	NULL	NULL	NULL	M	NULL	NULL	NULL
10	15	HGDP00017	NULL	NULL	NULL	M	NULL	NULL	NULL
11	15	HGDP00019	NULL	NULL	NULL	M	NULL	NULL	NULL
12	15	HGDP00021	NULL	NULL	NULL	M	NULL	NULL	NULL
13	15	HGDP00023	NULL	NULL	NULL	M	NULL	NULL	NULL
14	15	HGDP00025	NULL	NULL	NULL	M	NULL	NULL	NULL
15	15	HGDP00027	NULL	NULL	NULL	M	NULL	NULL	NULL

Figura 38: Tela de visualização da tabela *Individual* inserida no database. Observe que o campo population_id_population (segunda coluna) apresentado contém o id recuperado da tabela population correspondente à população Cayapa (id-population=18) para o indivíduo CAY547 (id-individual=1) e para os demais indivíduos apresentados na tela (id_individual de 2 a 15) foi recuperado o id-population=15 que corresponde à população Brahui. Vide arquivo apresentado na Figura 36.

Anexo 1



DER DIVERGENOMEdb Magalhães e cols. (2010, manuscrito em preparação). Diagrama que apresenta as relações entre as tabelas que compõe o DIVERGENOMEdb. Na terminologia do modelo relacional formal, a entidade se refere às tabelas e os atributos aos campos que descrevem cada entidade em particular.