Documentos

ISSN 1980-6841
Outubro, 2011
103

PATERNITYHD versão 1.0 – um software para cálculo da probabilidade de exclusão de paternidade com dados de genotipagem em painel de alta densidade de SNPs em bovinos

Embrapa

ISSN 1980-6841
Outubro, 2011

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

Documentos 103

Paternityhd versão 1.0 – um software para cálculo da probabilidade de exclusão de paternidade com dados de genotipagem em painel de alta densidade de SNPs em bovinos

Maurício de Alvarenga Mudadu
Luciana Correia de Almeida Regitano
Polyana Cristine Tizioto

Embrapa Pecuária Sudeste
São Carlos, SP
2011

Embrapa Pecuária Sudeste
Rod. Washington Luiz, km 234
13560 970, São Carlos, SP
Caixa Postal 339
Fone: (16) 3411- 5600
Fax: (16): 3361-5754
Home page: www.cppse.embrapa.br
Endereço eletrônico: sac@cppse.embrapa.br

Normalização bibliográfica: Sônia Borges de Alencar
Editoração eletrônica: Maria Cristina Campanelli Brito
Foto da capa: Luciana Correia de Almeida Regitano

1ª edição
1ª edição on-line (2011)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Pecuária Sudeste

Mudadu, Maurício de Alvarenga
Paternityhd versão 1.0 – um software para cálculo da probabilidade de exclusão de paternidade com dados de genotipagem em painel de alta densidade de SNPs em bovinos [Recurso eletrônico] / Maurício de Alvarenga Mudadu, Luciana Correia de Almeida Regitano. — Dados eletrônicos. — São Carlos, SP: Embrapa Pecuária Sudeste, 2011.

Sistema requerido: Adobe Acrobat Reader.

Modo de acesso: Word Wide Web: <http://www.cppse.embrapa.br/sites/default/files/principal/publicacao/Documentos103.pdf>

Título da página na Web (acesso em 30 outubro de 2011).

21p. (Documentos / Embrapa Pecuária Sudeste, 103; ISSN: 1980-6841).

1. Bioinformática – Software – Paternidade – Genotipagem – SNP. I. Mudadu, Maurício de Alvarenga. II. Regitano, Luciana Correia de Almeida. III.Tizioto, Polyana Cristine. IV.Título. V. Série.

CDD: 660.6

Autores

Maurício de Alvarenga Mudadu
Biólogo, Pesquisador da Embrapa Pecuária Sudeste, São Carlos, SP,
mudadu@cppse.embrapa.br

Luciana Correia de Almeida Regitano
Médica Veterinária, Pesquisadora da Embrapa Pecuária Sudeste, São Carlos, SP,
luciana@cppse.embrapa.br

Polyana Cristine Tizioto
Bióloga da Universidade Federal de São Carlos - UFSCar. Bolsista FAPESP, São Carlos, SP,
polytizioto@yahoo.com.br

Sumário

Introdução 6

Estrutura do software 8

Metodologia de funcionamento 10

Conexão ao banco de dados 11
Filtro para controle de qualidade 11
Comparações de amostras par a par 12
Seleção de alelos homozigotos 13
Contagem de inconsistências mendelianas 13
Estatística 14
A estatística usada no cálculo do poder de exclusão de paternidade . 14
Impressão de resultados 16
Laço de repetição 17

Conclusão 20

Referências 20

Paternityhd versão 1.0 – um software para cálculo da probabilidade de exclusão de paternidade com dados de genotipagem em painel de alta densidade de SNPs em bovinos

Mauricio de Alvarenga Mudadu
Luciana Correia de Almeida Regitano
Polyana Cristine Tizioto

Introdução

O uso do DNA em programas de melhoramento animal geralmente envolve a utilização de marcadores moleculares, como microssatélites e polimorfismos de base única (SNPs; do inglês *single nucleotide polymorphisms*). As análises do genoma bovino, incluindo mapeamento de locos de caracteres quantitativos (QTLs; do inglês *quantitative trait loci*), SNPs e, mais recentemente, a genotipagem em larga escala, poderão contribuir para a seleção precoce, assim como para determinação da paternidade e da identidade dos animais. Com o uso dos marcadores moleculares pretende-se aumentar a eficiência da seleção, antecipando e aumentando a acurácia, diretamente relacionados ao ganho genético (DAVIS & DENISE, 1998). A seleção assistida por marcadores pode auxiliar a seleção tradicional, mas não substituí-la, pois o máximo benefício será obtido quando as duas estratégias forem utilizadas simultaneamente. SNPs são bons marcadores moleculares para estudos de associação e análises de paternidade, pois são abundantes no genoma bovino, são estáveis em mamíferos e podem ser usados em técnicas de genotipagem em larga escala (HEATON et al., 2002).

Painéis com milhares de SNPs, como por exemplo, o chip *Illumina Bovine HD Genotyping Beadchip* (ILLUMINA, 2010), que possui aproximadamente 770 mil marcadores, estão disponíveis comercialmente. Esses chips são utilizados em análises em larga escala e podem ser aplicados, por exemplo, na seleção genômica. A seleção genômica, segundo Resende et al.,(2008), pode ser definida como a seleção simultânea para centenas ou milhares de marcadores, os quais cobrem o genoma densamente. É necessário que todos os genes que controlam um determinado caráter quantitativo estejam em desequilíbrio de ligação com pelo menos parte dos marcadores. Dessa forma esses marcadores podem explicar quase a totalidade da variação genética de um caráter quantitativo.

Entretanto, a metodologia de alguns desses estudos exige correta identificação parental dos animais. Alguns softwares que automatizam estudos de associação usando marcadores SNPs (PURCELL et al., 2007) usam uma classificação dos animais em famílias, o que implica em conhecer corretamente a genealogia e, portanto, a paternidade dos animais estudados.

Marcadores genéticos vêm sendo usados para resolver a paternidade pela sua habilidade de excluir falsos parentais. Existem cálculos estatísticos que levam em consideração um número *n* de alelos codominantes para resolver a paternidade. No caso de marcadores SNPs, podem ser utilizados cálculos que consideram apenas dois alelos, uma vez que marcadores SNPs são geralmente bialélicos. Marcadores do tipo SNP são uma boa opção para esse tipo de teste (PIMENTA & PENA, 2010).

Há exclusão da paternidade quando ocorre incompatibilidade de genótipos entre o animal e o suposto pai. Para realizar o teste de paternidade é importante balancear a quantidade de marcadores SNPs utilizados, de tal forma a maximizar o poder de exclusão e minimizar os custos do teste. Já foram relatados testes em humanos (PIMENTA & PENA, 2010) e em bovinos (HEATON et al., 2002), nos quais foram usados 40 e 32 marcadores do tipo SNP, respectivamente, com bons resultados.

Após a genotipagem é imprescindível realizar um teste estatístico para cálculo da probabilidade de exclusão do falso pai. Esse cálculo simples baseia-se em inconsistências mendelianas entre os genótipos do suposto pai, mãe e filho e as frequências dos alelos verificados na população. A maioria dos testes estatísticos para exclusão de paternidade que envolve marcadores de DNA superestima o poder de exclusão por excluir indivíduos randômicos em uma população. Porém, muitas vezes o indivíduo a ser excluído é um parente do pai verdadeiro, como um irmão. Existem testes que levam isso em consideração (FUNG et al., 2002) de forma a reduzir a população de indivíduos a serem excluídos. Assim, o valor do poder de exclusão torna-se mais próximo de um valor real e, no caso de resolução de paternidade entre bovinos, essa metodologia torna-se uma necessidade.

O software descrito nesse documento, denominado PaternityHD versão 1.0, aproveita dados de genotipagem em larga escala, provenientes do chip *Illumina Bovine HD Genotyping Beadchip*, para realizar testes de paternidade entre os indivíduos genotipados. PaternidtyHD v1.0 seleciona marcadores dentro de uma faixa de qualidade e determina a probabilidade de exclusão de paternidade do pai biológico levando em conta a possibilidade de excluir indivíduos que são irmãos do pai verdadeiro.

Estrutura do Software

O PaternityHD versão 1.0 é composto por:

1) Um script *stand-alone* escrito na linguagem Perl versão 5.10.1. testado em ambiente GNU/Linux, distribuição *Scientific Linux 6.0* (clone da distribuição *Red Hat versão 6.0*). O script demanda como entrada:

i) Um arquivo *db_config.pl* que possui as definições (usuário, senha e nome do banco de dados) para conexão do script ao servidor MySQL.

ii) Um servidor de banco de dados MySQL versão 5.1.52 instalado que deve conter duas tabelas: **a)** tabela *final_report* que possui os dados de genotipagem gerados pelo software *GenomeStudio* da empresa *Illumina Inc.*, fabricante do chip. Essa tabela deve conter pelo menos as seguintes colunas: *snp_name* (identificador do SNP), *sample_id* (identificador da amostra), *All1_AB* (alelo 1 no formato TOP/BOT Illumina), *All2_AB* (alelo 2 no formato TOP/BOT Illumina) e *GC_score* (acurácia do dado marcador) (ILLUMINA, 2005). **b)** tabela *missing_maf* que possui quatro colunas: *SNP* (identificador de um dado SNP), *MAF* (valor de frequência do alelo menos frequente), *F_MISS* (frequência de perda do dado SNP em todas as amostras), *HWE* (p-valor que evidencia se dado SNP está em equilíbrio de Hardy-Weinberg na população). Os dados contidos nessa tabela poderão ser gerados segundo ANDERSON et al. (2010) usando, por exemplo, o software PLINK (PURCELL et al., 2007) ou o pacote R® (R DEVELOPMENT CORE TEAM).

2) Linguagem Perl versão 5.10.1instalada com os módulos DBI instalados para conexão com o banco de dados MySQL.

O software foi testado em uma máquina Dell PowerEdge T610 com 32 Gb de RAM e 2 processadores com 6 núcleos físicos cada um. Foi testado um conjunto de amostras bovinas genotipadas com o chip *Illumina Bovine HD Genotyping Beadchip* cujos dados foram gerados pelo software *Genome Studio*, versão atual.

Metodologia de funcionamento

PaternityHD versão 1.0 funciona sendo disparado em um terminal de sistema operacional Linux (testado com *bash*) com o comando *"perl paternityhd_v1.0.pl"*, sendo que o arquivo *paternityhd_v1.0.pl* contém o código do programa escrito na linguagem Perl, e o arquivo "*db_config.pl*" descrito no tópico 2, item 1 e subitem i, deve estar no mesmo diretório em que o script for disparado.

O software funciona seguindo o algoritmo definido no fluxograma da Figura 1.

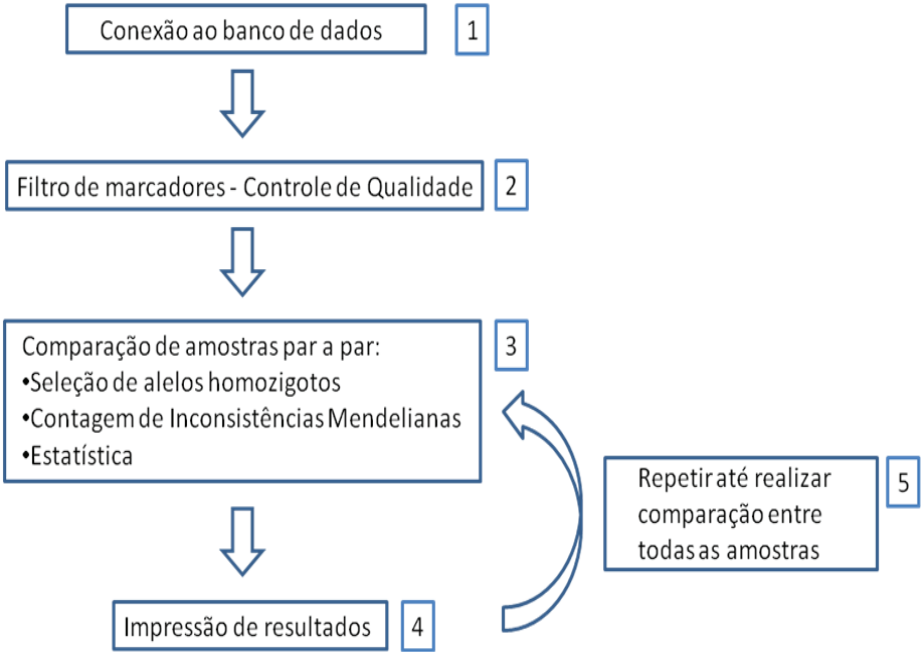


Figura 1. Fluxograma de funcionamento do PaternityHD v.1.0.

Conexão ao banco de dados

A conexão com o banco de dados MySQL (Figura 1) é feita via arquivo *db_config.pl*. Nesse arquivo as variáveis *$database*, *$server*, *$user e $pass* devem ser modificadas para conter o nome da base de dados onde estão as tabelas especificadas em 2.1, item ii (variável *$database*); o nome da máquina onde se encontra o servidor MySQL (*$server*); o nome do usuário com permissão de uso (seleção de dados e criação de tabelas temporárias) da base de dados e da tabela e sua senha (*$user* e *$pass*).

Com o arquivo configurado corretamente, a conexão do software PaternityHD com o banco de dados será efetuada.

Filtro para controle de qualidade

É utilizado um filtro (Figura 1, item 2) para selecionar apenas marcadores que passam pelos parâmetros de controle de qualidade (QC) de forma a deixar o teste mais confiável. Esse filtro é composto por três itens: i) MAF > = 0,4. Apenas marcadores que possuem a frequência do alelo menos frequente (MAF) maior ou igual a 0,4 serão selecionados. ii) F_MISS > 0. Apenas marcadores presentes em todas as amostras serão selecionados. iii) HWE > 0,05. Apenas marcadores supostamente em equilíbrio de Hardy-Weinberg (p-valor maior que 0,05) serão selecionados.

O filtro sugerido acima é realizado pela seleção de marcadores que passaram pelo QC usando a tabela *missing_maf*. Essa tabela possui os dados para MAF, F_MISS e HWE para todos os marcadores. A tabela foi gerada previamente por meio do software PLINK (PURCELL et al., 2007; ANDERSON et al., 2010).

Os marcadores que passarem pelo filtro de QC serão armazenados em um *array* associativo e, logo em seguida, em uma tabela temporária, que será usada em um passo futuro. Apenas SNPs que passarem no filtro de QC serão usados no teste de paternidade realizado pelo software.

Comparações de amostras par a par

O próximo passo realizado pelo software é a comparação de amostras par a par (Figura 1, item 3). PaternityHD seleciona e armazena em tabela temporária no banco de dados uma lista com todos os identificadores das amostras presentes na tabela *final_report*.

O software inicia, então, a comparação de todas as amostras entre si, em pares. O algoritmo de comparação de amostras funciona selecionando a primeira amostra da tabela de amostras e comparando-a com todas as outras, par a par. Findadas as comparações da primeira amostra com todas as outras, passa-se então para comparações da amostra seguinte com todas as outras, com o cuidado de não realizar comparações redundantes, por exemplo, da segunda com a primeira, e assim sucessivamente. Para evitar comparações redundantes e salvar tempo computacional, um *array* associativo é usado para guardar os identificadores das amostras que foram comparadas previamente. Esse *array* será consultado antes que um par de amostras a ser comparado for criado, para impedir que duas amostras já comparadas entrem na rodada novamente.

Iniciada a comparação entre duas amostras, passa-se para o passo seguinte que é a seleção de alelos homozigotos entre elas.

Seleção de alelos homozigotos

A seleção de alelos homozigotos (Figura 1, item 3) é feita por seleção no banco de dados, exigindo que o alelo 1 de um dado marcador seja igual ao alelo 2. As possibilidades são "AA" e "BB". Também é exigido que o marcador homozigoto tenha passado pelo filtro QC. Dessa forma são selecionados todos os alelos homozigotos do par de amostras em comparação. São criados dois *arrays* associativos, um para cada amostra em comparação, para guardar o primeiro alelo de cada marcador e o nome do marcador. Esses *arrays* serão usados no próximo passo.

Contagem de inconsistências mendelianas

Com os *arrays* associativos contendo a lista de marcadores homozigotos e seus primeiros alelos, do par de amostras em comparação, inicia-se a contagem de inconsistências mendelianas (Figura 1, item 3, subitem 2) entre os marcadores comuns às duas amostras. Para isso percorre-se o par de *arrays* associativos procurando por identificadores iguais. Se ambos os *arrays* possuírem um mesmo identificador de marcador isso significa que as duas amostras foram genotipadas para o mesmo marcador e são homozigotas. Quando isso ocorre, esse identificador será guardado em um novo *array* associativo para o cálculo da probabilidade de exclusão em passo subsequente.

A verificação de inconsistência mendeliana ocorre após a constatação de que ambas as amostras possuem um dado identificador e são homozigotas. Daí compara-se o primeiro alelo de dado marcador, verificando se são idênticos em ambas as amostras. Se não forem idênticos então está constatada uma inconsistência mendeliana, e os genótipos desses dois indivíduos para o dado marcador são incompatíveis, levando em conta a hipótese pai-filho e ou filho-pai. Uma variável é usada para contar o número de inconsistências encontradas em cada par de amostras em comparação.

Estatística

Após a contagem de inconsistências mendelianas em um dado par de amostras, passa-se para o cálculo da probabilidade de exclusão de paternidade entre as duas amostras (Figura 1, item 3). Para isso utiliza-se o *array* associativo usado para guardar todos os marcadores homozigotos comuns entre o par de amostras em comparação. Uma subrotina que recebe essa lista de marcadores é chamada e retornará os valores Q2, Q4, Q2', Q4', Q2'' e Q4''. Sendo Q2 o poder de exclusão individual para cada marcador, em que não há acesso ao genótipo da mãe. Q4 é uma modificação de Q2 em que se considera que o indivíduo testado é irmão do pai verdadeiro. Para se obter os poderes de exclusão combinados para todos os marcadores presentes no *array* associativo são calculados os valores Q2' e Q4' que combinam todos os valores individuais (Q2 e Q4, respectivamente) e ainda permitem uma inconsistência mendeliana. Os valores Q2'' e Q4'' são semelhantes a Q2' e Q4' porém permitem até duas inconsistências mendelianas.

A estatística usada no cálculo do poder de exclusão de paternidade

De acordo com Fung et al. (2002), a probabilidade de exclusão individual é a proporção de indivíduos do sexo masculino, escolhidos ao acaso, que podem ser excluídos como sendo pais de um suposto filho, baseado no genótipo do suposto filho (GF) e no genótipo da mãe (GM). Porém, nos casos em que não se possui acesso ao GM e utilizando-se marcadores bialélicos (SNPs), a fórmula para se calcular o poder de exclusão de um dado marcador (Q2) pode ser simplificada para a Equação 1:

$$Q2 = p1^2(1-p1)^2 + p2^2(1-p2)^2 + 2p1p2(1-p1-p2)^2$$

Equação 1. Poder de exclusão Q2 para marcadores individuais, onde p1 é a frequência do alelo A e p2 a frequência do alelo B encontrada na população de amostras genotipadas.

Verifica-se pela equação 1 que é necessário o cálculo da frequência (p1 e p2) dos alelos A e B. Isso é feito por meio de contagem dos alelos para um dado marcador em todas as amostras que estão sendo comparadas.

Outra possibilidade seria verificar o poder de exclusão, não usando como base todos os indivíduos do sexo masculino, mas uma subpopulação, definida como todos os indivíduos do sexo masculino, irmãos do pai verdadeiro. Para isso definem-se entre dois indivíduos I1 e I2 os coeficientes de similaridade entre eles: k0 (probabilidade de nenhum dos alelos de I1 e I2 serem idênticos por descendência) e k1 (probabilidade de um dado alelo de I1 ser idêntico por descendência a I2 e o outro não ser). No caso de irmãos, k0 = k1 = 0,25. O poder de exclusão de um dado marcador tendo em vista que os indivíduos são irmãos (Q4) é definido pela Equação 2:

$$Q4 = (k0 + k1)Q2$$

Equação 2. Poder de exclusão Q4, sem o genótipo da mãe, onde k0 = k1 = 0,25, o que implica que os indivíduos são irmãos (FUNG et al., 2002).

Para se conseguir um poder de exclusão que combine diversos marcadores, utilizam-se as equações 3, 4, 5 e 6 descritas a seguir. Dados os poderes de exclusão Q2 individuais dos diversos marcadores, definidos por Pi (i = 1,2...m), definem-se os valores Q2' (Equação 3) de modo a permitir até uma inconsistência mendeliana. Define-se Q2'' (Equação 5), de modo a permitir até duas inconsistências. Repete-se o mesmo processo utilizando os poderes de exclusão individuais Q4, para se obter os valores Q4' e Q4''(Equações 4 e 6).

$$Q2' = 1 - \prod_{i=1}^{m}(1 - Pi) - \sum_{i=1}^{m} Pi \prod_{\substack{j=1 \\ j \neq i}}^{m}(1 - Pj)$$

Equação 3. Cálculo de poder de exclusão combinado (Q2') que usa poderes de exclusão Q2 calculados para cada marcador, definidos por Pi (i = 1,2...m). Neste caso é permitida até uma inconsistência mendeliana. Adaptado de Fung et al. (2002).

$$Q4' = 1 - \prod_{i=1}^{m}(1 - Pi) - \sum_{i=1}^{m} Pi \prod_{\substack{j=1 \\ j \neq i}}^{m}(1 - Pj)$$

Equação 4. Cálculo de poder de exclusão combinado (Q4') que usa poderes de exclusão Q4 calculados para cada marcador, definidos por Pi (i = 1,2...m). Neste caso é permitida até uma inconsistência mendeliana. Adaptado de Fung et al. (2002).

$$Q2'' = Q2' - \sum_{i<j} PiPj \prod_{k=1,k\neq i,k\neq j}^{m} 1 - Pk$$

Equação 5. Cálculo de poder de exclusão combinado (Q2'') usando poderes de exclusão Q2' para cada marcador, definidos por Pi (i = 1,2...m). Permite-se até duas inconsistências mendelianas. Adaptado de Fung et al. (2002).

$$Q4'' = Q4' - \sum_{i<j} PiPj \prod_{k=1,k\neq i,k\neq j}^{m} 1 - Pk$$

Equação 6. Cálculo de poder de exclusão combinado (Q4'') usando poderes de exclusão Q2' para cada marcador, definidos por Pi (i = 1,2...m). Permite-se até duas inconsistências mendelianas. Adaptado de Fung et al. (2002).

Impressão de resultados

Finalizado o cálculo dos poderes de exclusão individuais Q2 e Q4 e dos poderes de exclusão combinados Q2', Q4', Q2'' e Q4'' é realizada a impressão dos resultados (Figura 1, item 4).

Os resultados são impressos em três arquivos separados: i) *paternityhd.mendelian.inconsistencies.txt* (Figura 2); ii) *paternityhd.inconsistencies.statistics.txt* (Figura 3) e iii) *paternityhd.power.statistic.txt* (Figura 4).

No arquivo i), são escritas quatro colunas: a) identificador do SNP, b) identificador da amostra1, c) Alelo1 da amostra1, d) identificador da amostra 2 e e) Alelo1 da amostra 2. O arquivo i) é usado apenas para conferência se há alguma inconsistência no algoritmo de PaternityHD.

No arquivo ii) são escritas oito colunas: a) identificador da amostra 1; b) identificador da amostra 2; Número de alelos homozigotos na amostra 1; c) Número de inconsistências mendelianas encontradas entre marcadores das amostras 1 e 2; d) Q2'; e) Q4'; f) Q2''; g) Q4''.

No arquivo iii) são escritas sete colunas: a) identificador da amostra 1; b) identificador da amostra 2; c) identificador do marcador; d) p1: frequência do alelo A, para o dado marcador definido em c) na população; e) p2: frequência do alelo B, para o dado marcador definido em c) na população; f) Q2 e g) Q4.

O arquivo ii) contém as informações finais necessárias que definirão se será possível excluir a paternidade de uma amostra em relação à outra, de acordo com seus genótipos, o poder de exclusão combinado dos marcadores utilizados e a contagem de inconsistências mendelianas entre as amostras.

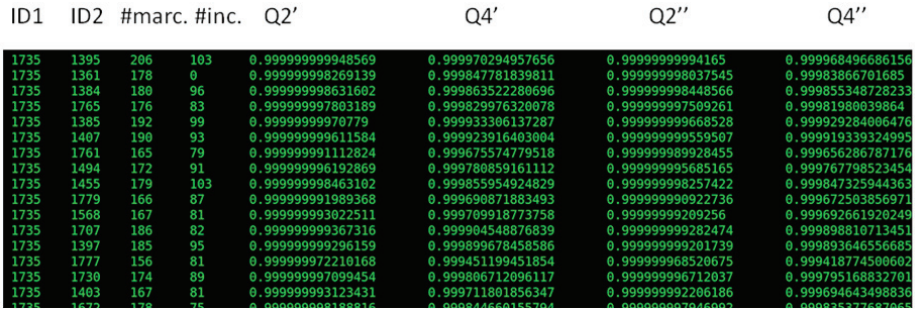
Laço de repetição

Após realizada a comparação entre um par de amostras com a seleção de alelos homozigotos comuns, a contagem de inconsistências mendelianas e o cálculo da estatística de exclusão, um laço de repetição (Figura 1, item 5) é usado para continuar formando pares de amostras ainda não comparadas, de modo a realizar a comparação entre todos os pares de amostras possíveis. Lembrando que pares de amostras redundantes são evitados pelo uso de um *array* associativo que guarda os identificadores de amostras já comparadas.

Ao final, teremos os três arquivos de saída de PaternityHD preenchidos com todos os dados relativos a todas as comparações par a par, entre todas as amostras. Esses arquivos poderão ser analisados de forma a verificar os pares de amostras cuja paternidade não pode ser excluída. Por exemplo, ao triar o arquivo ii) pode-se verificar o número de inconsistências mendelianas que estão dentro do permitido e conferir se a probabilidade de exclusão do pai verdadeiro é aceitável.

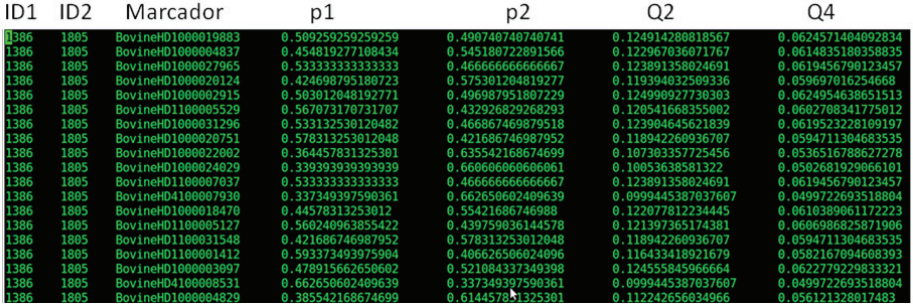
ID marc.	ID1	Al1_A1	ID2	Al1_A2
BovineHD1000025314	1735	A	1720	B
BovineHD0300018971	1735	B	1720	A
BovineHD0100024590	1735	A	1720	B
BovineHD1800016183	1735	B	1720	A
BovineHD1500023824	1735	B	1720	A
BovineHD0200030244	1735	A	1720	B
BovineHD0400022573	1735	A	1720	B
BovineHD1900018876	1735	B	1720	A
BovineHD0200007617	1735	A	1720	B
BovineHD0100017222	1735	A	1720	B
ARS-BFGL-NGS-17601	1735	B	1720	A
BovineHD1600003869	1735	B	1720	A
BovineHD1700015340	1735	B	1720	A
BovineHD1000023194	1735	A	1720	B
BovineHD1500006293	1735	B	1720	A

Figura 2. Exemplo de arquivo de saída *paternityhd.mendelian.inconsistencies.txt*. O arquivo possui cinco colunas: Identificador do marcador (ID Marc.), identificador da primeira amostra (ID1), alelo1 da primeira amostra (Al1_A1), identificador da segunda amostra (ID2), alelo 1 da segunda amostra (Al1_A2).



ID1	ID2	#marc.	#inc.	Q2'	Q4'	Q2''	Q4''
1735	1395	206	103	0.999999999948569	0.999970294957656	0.99999999994165	0.999968496686156
1735	1361	178	0	0.999999998269139	0.999847781839811	0.999999998037545	0.99983866701685
1735	1384	180	96	0.999999998631602	0.999863522280696	0.999999998448566	0.999855348728233
1735	1765	176	83	0.999999997803189	0.999829976320078	0.999999997509261	0.99981980039864
1735	1385	192	99	0.99999999970779	0.999933306137287	0.999999999668528	0.999929284006476
1735	1407	190	93	0.999999999611584	0.999923916403004	0.999999999559507	0.999919339324995
1735	1761	165	79	0.999999991112824	0.999675574779518	0.999999989928455	0.999656286787176
1735	1494	172	91	0.999999996192869	0.999780859161112	0.999999995685165	0.999767798523454
1735	1455	179	103	0.999999998463102	0.999855954924829	0.999999998257422	0.999847325944363
1735	1779	166	87	0.999999991989368	0.999690871883493	0.999999990922736	0.999672503856971
1735	1568	167	81	0.999999993022511	0.999709918773758	0.99999999209256	0.999692661920249
1735	1707	186	82	0.999999999367316	0.999904548876839	0.999999999282474	0.999898810713451
1735	1397	185	95	0.999999999296159	0.999899678458586	0.999999999201739	0.999893646556685
1735	1777	156	81	0.999999972210168	0.999451199451854	0.999999968520675	0.999418774500602
1735	1730	174	89	0.999999997099454	0.999806712096117	0.999999996712037	0.999795168832701
1735	1403	167	81	0.999999993123431	0.999711801856347	0.999999992206186	0.999694643498836

Figura 3. Exemplo de arquivo de saída *paternityhd.inconsistencies.statistics.txt*. O arquivo possui oito colunas: Identificador da primeira amostra (ID1), identificador da segunda amostra (ID2), número de marcadores homozigotos comuns utilizados (#marc.), número de inconsistências mendelianas encontradas(#inc.), poder de exclusão Q2'(Q2'), poder de exclusão Q4'(Q4'), poder de exclusão Q2''(Q2''), poder de exclusão Q4''(Q4'').



ID1	ID2	Marcador	p1	p2	Q2	Q4
1386	1805	BovineHD1000019883	0.509259259259259	0.490740740740741	0.124914280818567	0.0624571404092834
1386	1805	BovineHD1000004837	0.454819277108434	0.545180722891566	0.122967036071767	0.0614835180358835
1386	1805	BovineHD1000027965	0.533333333333333	0.466666666666667	0.123891358024691	0.0619456790123457
1386	1805	BovineHD1000020124	0.424698795180723	0.575301204819277	0.119394032509336	0.0596970162546688
1386	1805	BovineHD1000002915	0.503012048192771	0.496987951807229	0.124990927730303	0.0624954638651513
1386	1805	BovineHD1100005529	0.567073170731707	0.432926829268293	0.120541668355002	0.0602708341775012
1386	1805	BovineHD1000031296	0.533132530120482	0.466867469879518	0.123904645621839	0.0619523228109197
1386	1805	BovineHD1000020751	0.578313253012048	0.421686746987952	0.118942260936707	0.0594711304683535
1386	1805	BovineHD1000022002	0.364457831325301	0.635542168674699	0.107303357725456	0.0536516788627278
1386	1805	BovineHD1000024029	0.339393939393939	0.660606060606061	0.10053638581322	0.0502681929066101
1386	1805	BovineHD1100007037	0.533333333333333	0.466666666666667	0.123891358024691	0.0619456790123457
1386	1805	BovineHD4100007930	0.337349397590361	0.662650602409639	0.0999445387037607	0.0499722693518804
1386	1805	BovineHD1000018470	0.44578313253012	0.55421686746988	0.122077812234445	0.0610389061172223
1386	1805	BovineHD1100005127	0.560240963855422	0.439759036144578	0.121397365174381	0.0606986825871906
1386	1805	BovineHD1100031548	0.421686746987952	0.578313253012048	0.118942260936707	0.0594711304683535
1386	1805	BovineHD1100001412	0.593373493975904	0.406626506024096	0.116433418921679	0.0582167094608393
1386	1805	BovineHD1000003097	0.478915662650602	0.521084337349398	0.124555845966664	0.0622779229833321
1386	1805	BovineHD4100008531	0.662650602409639	0.337349397590361	0.0999445387037607	0.0499722693518804
1386	1805	BovineHD1000004829	0.385542168674699	0.614457831325301	0.112242656034966	0.056121328017483

Figura 4. Exemplo de arquivo de saída *paternityhd.power.statistic.txt*. O arquivo possui sete colunas: identificador da primeira amostra (ID1), identificador da segunda amostra (ID2), identificador do marcador (Marcador), frequência do aleloA (p1), frequência do alelo B (p2), poder de exclusão Q2 (Q2) e poder de exclusão Q4 (Q4).

Da forma como descrito nesse documento, PaternityHD v.1.0 utiliza todos os marcadores homozigotos comuns entre todos os pares de amostras sendo comparadas. Essa metodologia acaba por aumentar muito o custo computacional do processo, além de utilizar um número excessivo de marcadores. PaternityHD pode ser facilmente modificado para utilizar um número fixo de marcadores homozigotos comuns, que seja suficiente para verificar a paternidade entre dois indivíduos de forma mais ágil e, ainda assim, eficiente.

Conclusão

PaternityHD v1.0 é um script escrito na linguagem *Perl* que realiza o cálculo do poder de exclusão de paternidade entre indivíduos genotipados com um painel de SNPs em alta densidade, modelo *Illumina Bovine HD Genotyping Beadchip*. PaternityHD v1.0 acessa os dados da genotipagem em um banco de dados MySQL e conta o número de inconsistências mendelianas para marcadores homozigotos comuns entre todas as amostras testadas, par a par. O software realiza o cálculo do poder de exclusão para todas as amostras genotipadas, par a par, sem redundância. São gerados valores de poder de exclusão combinados para cada par de amostras, permitindo uma ou duas inconsistências mendelianas (Q2' e Q2'' respectivamente). O software também verifica o poder de exclusão combinado levando em consideração que o indivíduo testado é irmão do pai verdadeiro, com uma ou duas inconsistências mendelianas permitidas (Q4' e Q4''). Todos os dados gerados por PaternityHD v1.0 são escritos em três arquivos de saída.

Referências

ANDERSON, C. A.; PETTERSSON, F. H.; CLARKE, G. M.; CARDON, L. R.; MORRIS, A. P.; ZONDERVAN, K. T. Data quality control in genetic case-control association studies. **Nature Protocols,** v. 5, n. 9, p. 1564-1573, 2010.

DAVIS, G. P.; DANISE, S. K. The Impact of Genetic Markers on Selection. **Journal of Animal Science,** v. 76, p. 2331–2339, 1998.

FUNG, W. K.; CHUNG, Y. K.; WONG, D. M. Power of exclusion revisited: probability of excluding relatives of the true father from paternity. **International Journal of Legal Medicine**, v. 116, n. 2, p. 64-67, 2002.

HEATON, M. P.; HARHAY, G. P.; BENNETT, G. L.; STONE, R. T.; GROSSE, W. M.; CASAS, E.; KEELE, J. W.; SMITH, T. P.; CHITKO-MCKOWN, C. G.; LAEGREID, W. W. Selection and use of SNP markers for animal identification and paternity analysis in U.S. beef cattle. **Mammalain Genome,** v. 13, n. 5, p. 272-812, 2002.

ILLUMINA. **Bovine HD Genotyping BeadChip. 2010.** Disponível em: http://www.illumina.com/Documents/products/datasheets/datasheet_bovineHD.pdf. Acesso em janeiro 2011.

ILLUMINA. **Illumina GenCall Data Analysis Software. 2005.** Disponível em: http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf. Acesso em abril 2011.

PURCELL, S.; NEALE, B.; TODD-BROWN, K.; THOMAS, L.; FERREIRA, M. A.R.; BENDER, D.; MALLER, J.; SKLAR, P.; DE BAKKER, P. I.; DALY, M. J.; SHAM, P. C. PLINK: a toolset for whole-genome association and population-based Linkage analysis. **American Journal of Human Genetics**, v. 81, 2007.

PIMENTA, J. R.; PENA, S. D. Efficient human paternity testing with a panel of 40 short insertion-deletion polymorphisms. **Genetics Molecular Research,** v. 9, n. 1, p. 601-607, 2010.

R DEVELOPMENT CORE TEAM. R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing**. Vienna, 2011. ISBN 3-900051-07-0.

RESENDE, M. D. V.; LOPES, P. S.; SILVA, R. L.; PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira,** Colombo, n. 56, p. 63-77, 2008.