



Alchemy - Analisador de qualidade de SNPs

Insira os arquivos relativos a Intensidade e Mapa de SNPs

Intensidade: Selecionar arquivo...

Mapa de SNPs: Selecionar arquivo...

[Alterar parâmetros](#)

Obs.: a geração dos resultados pode levar alguns minutos...

Alchemy Web: Uma Proposta de Interface Gráfica para o Sistema Alchemy

Marcelo Gonçalves Narciso¹
Paula Arielle Mendes Ribeiro Valdissar²
Cristyene Gonçalves Benfício³
Tereza Cristina de Oliveira Borba⁴
Cláudio Brondani⁵
Rosana Pereira Vianello⁶

Introdução

Os marcadores moleculares são ferramentas úteis na detecção de variações genéticas no DNA. Na área vegetal, são utilizados na caracterização molecular de genótipos, no mapeamento de genes e de regiões genômicas, na seleção assistida por marcadores e em estudos de diversidade genética, entre outras aplicações.

Dentre os diversos marcadores moleculares disponíveis, destacam-se os SNPs (*Single-Nucleotide Polymorphisms*), que são mutações de ponto que ocorrem no genoma em frequência alélica mínima de 1 % em dada população. Mutações de ponto consistem na substituição de um único nucleotídeo em fragmentos homólogos de DNA. As substituições mais frequentes são as denominadas transições, que são trocas entre bases nitrogenadas de mesma característica estrutural, ou seja, troca entre duas pirimidinas (T/C) ou duas purinas (A/G). Os SNPs podem ocorrer em regiões codantes ou não-codantes. Quando ocorrem em regiões codantes e resultam na substituição de aminoácidos, os SNPs são denominados não-sinônimos, porque podem provocar a alteração estrutural e funcional da proteína traduzida, dependendo das características dos aminoácidos substituídos (KWOK; GU, 1999).

Os SNPs vêm gradativamente substituindo os marcadores do tipo microssatélite (SSR) devido à sua abundância no genoma e estabilidade, aliadas à facilidade de automação e multiplexagem (YU et al.,

2011). Entretanto, para a disseminação de seu uso, faz-se ainda necessária a superação de algumas dificuldades, como o alto custo de desenvolvimento de ensaios, a dificuldade de manipulação de dados em larga escala e a necessidade de uso de softwares especializados.

A genotipagem de SNPs se baseia na produção de fragmentos alelo-específicos e posterior detecção por métodos como espectrometria de massa, fluorescência e quimiluminescência. Diferentes estratégias são adotadas para produção destes fragmentos, como extensão por primers, hibridização, ligação e clivagem por enzimas, sendo que a maioria das plataformas de genotipagem emprega a etapa de amplificação por PCR (*Polymerase Chain Reaction*) para aumentar a quantidade de produtos (YU et al., 2011). Algumas plataformas de genotipagem utilizam ainda a combinação de mais de uma das estratégias acima (KIM; MISRA, 2007).

A Embrapa Arroz e Feijão utiliza a plataforma de genotipagem BeadXpress Reader (Illumina), que utiliza a fluorescência como método de detecção. Os dados gerados pelo equipamento são analisados através do software Genome Studio (http://www.illumina.com/documents/products/datasheets/datasheet_genomestudio_software.pdf). Este software utiliza o método de clusterização (agrupamento) para chamada dos alelos, baseando-se na proporção entre os sinais de fluorescência de cy3/cy5. Para tanto, o método requer a análise de uma grande quantidade

¹ Engenheiro elétrico, Ph.D. em Bioinformática, pesquisador da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, marcelo.narciso@embrapa.br

² Farmacêutica, Especialista em Biotecnologia, analista da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, paula.valdissar@embrapa.br

³ Bióloga, Mestre em Genética e Melhoramento de Plantas, técnico administrativo superior da Universidade Federal de Goiás, Goiânia, GO, cristyene@gmail.com

⁴ Engenheira de alimentos, Doutora em Genética e Melhoramento de Plantas, pesquisadora da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, tereza.borba@embrapa.br

⁵ Engenheiro agrônomo, Ph.D. em Biologia Molecular, pesquisador da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, claudio.brondani@embrapa.br

⁶ Bióloga, Ph.D. em Genética Molecular, pesquisadora da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, rosana.vianello@embrapa.br

de indivíduos e a representatividade das três classes de genótipos: homozigotos para cada alelo (A/A ou B/B) e heterozigoto (A/B). Neste tipo de método, um cluster heterozigoto é sempre esperado para atender ao pressuposto de equilíbrio de Hardy Weinberg. No caso de amostras endogâmicas, cuja taxa de heterozigosidade é muito baixa, o Genome Studio, com o intuito de estabelecer esse equilíbrio, acaba chamando um grupo homozigoto de heterozigoto ou subdivide um grupo de homozigotos em heterozigoto e homozigoto, provocando chamadas equivocadas de alelos. Por esse motivo, Wright et al. (2010) desenvolveram um algoritmo de genotipagem denominado Alchemy, o qual utiliza o método estatístico de inferência Bayesiana, ao invés de análise de agrupamento (<http://alchemy.sourceforge.net>).

O algoritmo do Alchemy parte do princípio que o resumo das intensidades de fluorescência para cada canal (alelo) é uma distribuição composta por um componente de sinal e um componente de ruído. Quando um alelo está presente, um valor de intensidade do componente sinal é observado. Se o alelo não está presente, o valor de intensidade observado é retirado do componente ruído. Por exemplo, para um organismo diplóide com um genótipo AA será observado um sinal no canal A e um ruído no canal B, e do mesmo modo, mas no sentido inverso, para um genótipo BB. O genótipo AB produziria sinal em ambos os canais. Já um sinal de ruído, observado em ambos os canais, indica que o ensaio falhou (*no call*). Esta distribuição de sinal e ruído é razoavelmente bem aproximada pelo ajuste de uma distribuição Gaussiana. A partir disso, são calculadas as probabilidades posteriores de uma amostra, para cada marcador SNP ter o genótipo AA, BB, AB ou NC (*no call*). O genótipo considerado para a amostra (*Pcall*) é aquele que possui maior probabilidade posterior (WRIGHT et al., 2010).

Além de determinar o genótipo da amostra, o software Alchemy fornece uma avaliação da qualidade do dado, determinando qual a probabilidade da chamada alélica estar correta. A partir disto, pode ser escolhido um 'threshold', ou seja, um valor mínimo de probabilidade a ser considerado, sendo que, para valores abaixo do "threshold", as chamadas alélicas são consideradas como NC (WRIGHT et al., 2010).

A utilização do sistema Alchemy requer sua compilação e instalação local, em ambiente Linux

ou Windows, e sua execução se dá por linha de comando. Como arquivos de entrada, o Alchemy utiliza dois arquivos de saída do Genome Studio. O primeiro arquivo corresponde a um mapa que informa a substituição nucleotídica de cada SNP e o segundo informa as intensidades de fluorescência para cada alelo (A/B) por amostra e por SNP. O arquivo de saída do Alchemy é do tipo tsv (*tab-separated values*).

Para usuários não familiarizados com o Linux, o que corresponde possivelmente à realidade da maioria dos clientes do Alchemy, o uso de linhas de comando é um fator limitante, sobretudo pela necessidade para executá-lo, de conhecimento de outros comandos do ambiente Linux, além do comando *alchemy* especificamente. Assim, uma interface gráfica que permita o acesso amigável ao comando *alchemy* é bastante desejável. Neste trabalho, foi gerado o sistema Alchemy Web, que é composto por uma interface gráfica local que permite a execução remota do programa Alchemy, sem a necessidade de conhecimento e uso de linhas de comando, e que gera como resultado um arquivo em formato csv (*comma separated values*), compatível com planilhas Excel e OpenOffice Calc.

A interface gráfica foi construída utilizando-se a linguagem JavaScript. Este programa permite que o usuário, a partir do seu computador, acesse o Alchemy, através de uma página HTML (<http://200.17.55.25/alchemy>). O programa apresenta um formulário (Figura 1) onde o usuário deve inserir os dois arquivos gerados pelo Genome Studio (Figura 2) e selecionar os parâmetros para a execução do Alchemy (Figura 3). O programa também verifica se todos os parâmetros foram corretamente preenchidos e, em caso negativo, devolve uma mensagem de erro. O programa apresenta ainda uma janela suspensa de ajuda ou "Help", que mostra, passo a passo, como executar remotamente o sistema. Para desenvolver o programa que recebe os dados inseridos no formulário, executou-se o Alchemy remotamente e retornaram-se os resultados obtidos, utilizando a linguagem PHP.



Figura 1. Imagem inicial do AlchemyWeb.

Executando o Alchemy Web

Para exemplificar o acesso e o uso do Alchemy Web, foram utilizados dados obtidos em feijoeiro comum (*Phaseolus vulgaris* L.), no Laboratório de Biotecnologia da Embrapa Arroz e Feijão, através da plataforma BeadXpress (Illumina). O Alchemy Web foi acessado via web e os dois arquivos de saída do Genome Studio foram inseridos no formulário (Figura 2).



Figura 2. Inserção de arquivos oriundos do Genome Studio no Alchemy Web.

O sistema oferece um conjunto de parâmetros com valores recomendados (*default*), que podem ou não atender às necessidades e particularidades da amostra. Para personalizar os valores, deve-se selecionar "Alterar parâmetros" (Figura 3). Neste trabalho, os parâmetros *het-estimate threshold* e *inbreeding coefficient* foram configurados para 1 e 0,9, respectivamente. O primeiro parâmetro, também denominado de heterozigosidade, refere-se ao limite de amostras heterozigotas que são esperadas para a população em estudo. Este valor pode variar de 0 a 5, sendo que, quanto mais próximo de 0, menor o número de heterozigotos esperado. Já o *inbreeding coefficient* é o valor de coeficiente de endogamia das amostras, que pode variar de -0,5 a 1,0, sendo que, valores mais próximos de 1 indicam amostras mais endogâmicas. Para os demais parâmetros, utilizou-se o *default*. Informações sobre todos os parâmetros podem ser obtidas em http://sourceforge.net/apps/mediawiki/alchemy/index.php?title=Using_ALCHEMY ou no link "Parâmetros do Alchemy" do Alchemy Web. Neste caso, uma janela suspensa fornece a descrição dos parâmetros (Figura 4).

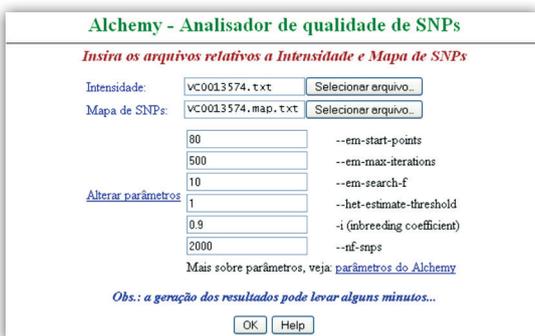


Figura 3. Seleção de valores de parâmetros no Alchemy Web.

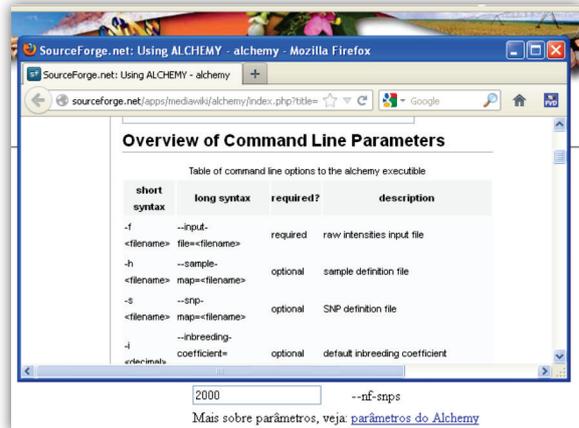


Figura 4. Janela suspensa com descrição dos parâmetros e seus valores.

Após o preenchimento do formulário, a submissão dos dados e a execução do programa Alchemy, instalado no servidor, se dá através do acionamento do botão "OK". O programa faz o carregamento do arquivo de resultados e abre uma janela, como apresentado na Figura 5. O usuário pode então escolher entre abrir ou salvar o arquivo de saída. A aparência do arquivo, em planilha Excel, é apresentada na Figura 6.



Figura 5. Aviso de conclusão da análise e arquivo de saída obtido.

#	rsnp	sample	AB call	ACGT call	P(call)	P(AA)	P(AB)	P(BB)	P(NC)	A int. (adj)	B int. (adj)	A int. (raw)	B int. (raw)
6	BARC-PV-VC001357:BB	CC	1,00E+00	1,86E-07	1,91E-07	1,00E+00	9,95E-06	6,33	9,87	6,68	9,99		
8	BARC-PV-VC001357:AA	TT	9,89E-01	9,89E-01	1,91E-03	2,66E-03	6,29E-03	8,62	8,61	8,74	8,73		
9	BARC-PV-VC001357:AA	TT	1,00E+00	1,00E+00	1,18E-06	1,48E-04	7,79E-05	9,03	6,2	9,2	6,23		
10	BARC-PV-VC001357:BB	CC	1,00E+00	1,13E-07	3,89E-07	1,00E+00	4,47E-06	6,12	10	6,17	9,9		
11	BARC-PV-VC001357:BB	CC	1,00E+00	1,98E-07	1,74E-07	1,00E+00	1,21E-05	6,22	9,77	6,21	9,89		
12	BARC-PV-VC001357:BB	CC	1,00E+00	1,18E-07	7,55E-08	1,00E+00	4,88E-06	6,16	9,98	6,28	9,96		
13	BARC-PV-VC001357:AA	TT	1,00E+00	1,00E+00	2,04E-07	3,91E-06	5,78E-06	8,96	7,58	9,05	7,65		
14	BARC-PV-VC001357:AA	TT	1,00E+00	1,00E+00	7,23E-06	3,59E-05	5,94E-05	9,01	8,31	8,91	8,22		
15	BARC-PV-VC001357:BB	CC	1,00E+00	1,36E-07	9,12E-08	1,00E+00	6,77E-06	6,09	9,85	6,03	9,77		
16	BARC-PV-VC001357:BB	CC	1,00E+00	1,30E-07	1,38E-06	1,00E+00	5,83E-06	6,22	9,96	6,23	10		
17	BARC-PV-VC001357:AA	TT	1,00E+00	1,00E+00	1,31E-04	1,28E-05	1,74E-05	9,21	8,25	9,1	8,17		
18	BARC-PV-VC001357:BB	CC	1,00E+00	1,13E-07	6,35E-08	1,00E+00	4,46E-06	6,06	9,98	6,07	10,05		
19	BARC-PV-VC001357:BB	CC	1,00E+00	1,18E-07	7,70E-08	1,00E+00	4,63E-06	6,21	10,03	6,42	10,03		
20	BARC-PV-VC001357:AA	TT	9,66E-01	9,66E-01	9,23E-03	8,64E-03	1,58E-02	8,66	8,81	8,85	8,79		
21	BARC-PV-VC001357:BB	CC	1,00E+00	1,14E-07	6,88E-08	1,00E+00	4,60E-06	6,09	9,98	6,13	9,98		
22	BARC-PV-VC001357:BB	CC	1,00E+00	1,44E-07	4,89E-07	1,00E+00	7,20E-06	6,19	9,87	6,25	9,85		
23	BARC-PV-VC001357:BB	CC	1,00E+00	1,38E-07	1,01E-07	1,00E+00	6,58E-06	6,19	9,9	6,1	9,86		
24	BARC-PV-VC001357:AA	TT	1,00E+00	1,00E+00	3,07E-05	1,06E-04	2,07E-04	8,88	8,38	8,86	8,37		
25	BARC-PV-VC001357:BB	CC	1,00E+00	1,07E-07	2,49E-07	1,00E+00	3,65E-06	6,06	10,07	6,33	9,91		
26	BARC-PV-VC001357:BB	CC	1,00E+00	1,15E-07	7,33E-08	1,00E+00	4,93E-06	6,14	9,87	6,3	10,03		

Figura 6. Arquivo de saída do Alchemy Web em Excel.

Cada coluna tem seu significado, conforme pode ser visto em http://sourceforge.net/apps/mediawiki/alchemy/index.php?title=ALCHEMY_Output_File_Format. Os valores de P(call) e P(NC) são muito relevantes e significam:

- P(call) -> probabilidade a posteriori de a chamada alélica estar correta.
- P(NC) -> probabilidade de o ensaio ter falhado (*no call*).
- P(AA) -> probabilidade de o genótipo AA ser verdadeiro.
- P(AB) -> probabilidade de o genótipo AB ser verdadeiro.

P(BB) -> probabilidade de o genótipo BB ser verdadeiro.

Os valores de P(AA), P(AB) e P(BB), somados, são iguais a 1. Assim, quando o valor de um deles aumenta, os dos outros decrescem. O índice P(call) considera a maior probabilidade posterior entre P(AA), P(AB) e P(BB), determinando qual é o genótipo da amostra (Coluna AB call da Figura 6) e o quanto esta chamada alélica pode ser verdadeira. As outras colunas, na ordem em que aparecem, se referem ao nome do SNP, nome da amostra, chamada AB (onde AA é um homozigoto para o alelo A, BB é um homozigoto para o alelo B e AB é um heterozigoto) e chamada ACGT (na qual a chamada AB é convertida para as chamadas de nucleotídeos, dependendo do mapa do SNP). Neste caso, por exemplo, se para determinado SNP a chamada A se refere à Timina e a chamada B se refere à Citosina, então AA seria TT e BB seria CC. As próximas colunas indicam as probabilidades posteriores conforme descrito acima. Para facilitar a visualização desses valores, estes podem ser convertidos em porcentagem.

Conforme exemplo mostrado na Figura 6, e ampliado na Figura 7, para o SNP BARC-PV-0006107, a amostra VC0013574-Feijao-Placa01_R001_C001 apresentou o alelo BB, que pelo mapa do SNP corresponde ao genótipo Citosina-Citosina, com 1,00E+00 (100,00%) de probabilidade P(call) de esse genótipo ser verdadeiro. Pode-se verificar que o P(call) é o valor de P(BB), ou seja, a maior probabilidade posterior neste caso.

	A	B	C	D	E	F	G	H	I
1	# ALCHEMY Std: alchemy.v.1.5 2009/12/19 16:15:12 koni Exp 5								
2	# Program copyright 2006, 2009, 2010 Mark Hamilton Wright (koni)								
3	# BUILD DATE: Qua Out 27 12:44:56 BRST 2010								
4	# RUN DATE: Thu Oct 4 12:46:48 2012								
5	[Redacted]								
6	# SNP	sample	AB call	ACGT call	P(call)	P(AA)	P(AB)	P(BB)	P(NC)
7	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C001	BB	CC	1,00E+00	1,86E-07	1,71E-07	1,00E+00	9,95E-06
8	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C002	AA	TT	9,89E-01	9,89E-01	1,91E-03	2,60E-03	6,29E-03
9	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C003	AA	TT	1,00E+00	1,00E+00	1,16E-06	1,40E-04	7,79E-05
10	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C004	BB	CC	1,00E+00	1,13E-07	3,08E-07	1,00E+00	4,47E-05
11	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C005	BB	CC	1,00E+00	1,98E-07	1,74E-07	1,00E+00	1,21E-05
12	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C006	BB	CC	1,00E+00	1,18E-07	7,55E-08	1,00E+00	4,88E-06
13	BARC-PV-0006107	VC0013574-Feijao-Placa01_R001_C007	AA	TT	1,00E+00	1,00E+00	2,04E-07	3,91E-06	3,78E-06

Figura 7. Exemplo de resultado gerado pelo Alchemy.

Conclusões

O desenvolvimento de uma interface gráfica para o Alchemy na web permitiu o acesso rápido e fácil ao programa, fornecendo resultados de genotipagem rápidos e precisos, evitando a subjetividade na análise e, simultaneamente, apresentando a avaliação da qualidade dos resultados, informando qual a probabilidade posterior das chamadas alélicas estarem corretas.

Referências

KIM, S.; MISRA, A. SNP genotyping: technologies and biomedical applications. **Annual Review of Biomedical Engineering**, Palo Alto, v. 9, p. 289-320, 2007.

KWOK, P. Y.; GU, Z. Single nucleotide polymorphism libraries: why and how are we building them? **Molecular Medicine Today**, Cambridge, v. 5, n. 12, p. 538-543, Dec. 1999.

WRIGHT, M. H.; TUNG, C. W.; ZHAO, K.; REYNOLDS, A.; MCCOUCH, S. R.; BUSTAMANTE, C. D. ALCHEMY: a reliable method for automated SNP genotype calling for small batch sizes and highly homozygous populations. **Bioinformatics**, Oxford, v. 26, n. 23, p. 2952-2960, Dec. 2010.

YU, H.; XIE, W.; WANG, J.; XING, Y.; XU, C.; LI, X.; XIAO, J.; ZHANG, Q. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. **PLoS ONE**, v. 6, n. 3, Mar. 2011. doi: 10.1371/journal.pone.0017595

Comunicado Técnico, 216



Exemplares desta edição podem ser adquiridos na:
Embrapa Arroz e Feijão
Endereço: Rod. GO 462 Km 12 Zona Rural, Caixa Postal 179 75375-000 Santo Antônio de Goiás, GO
Fone: (62) 3533 2123
Fax: (62) 3533 2100
E-mail: sac.cnpaf@embrapa.br
1ª edição
 Versão online (2014)

Comitê de publicações

Presidente: Pedro Marques da Silveira
Secretário-Executivo: Luiz Roberto R. da Silva
Membros: Camilla Souza de Oliveira, Luciene Frós Camarano de Oliveira, Flávia Rabelo Barbosa Moreira, Ana Lúcia Delalibera de Faria, Heloisa Célis Breseghello, Márcia Gonzaga de Castro Oliveira, Fábio Fernandes Nolêto

Expediente

Supervisão editorial: Luiz Roberto Rocha da Silva
Revisão de texto: Camilla Souza de Oliveira
Normalização bibliográfica: Ana Lúcia D. de Faria
Edição eletrônica: Fabiano Severino