ISSN 1678-9644
Abril, 2014

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

Documentos 290

Um Estudo Comparativo de Softwares para Alinhamento e Detecção de *Single Nucleotide Polymorphisms* (SNPs)

Marcelo Gonçalves Narciso
Adelmo Martins Rodrigues
Jorge Freitas Cieslak
Ricardo Diógenes Dias Silveira
Rosana Pereira Vianello
Cláudio Brondani

Embrapa Arroz e Feijão
Santo Antônio de Goiás, GO
2014

Exemplares desta publicação podem ser adquiridos na:

Embrapa Arroz e Feijão
Rod. GO 462, Km 12, Zona Rural
Caixa Postal 179
75375-000 Santo Antônio de Goiás, GO
Fone: (62) 3533 2110
Fax: (62) 3533 2100
www.cnpaf.embrapa.br
cnpaf.sac@embrapa.br

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

1ª edição
Versão online (2014)

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Arroz e Feijão

Um estudo comparativo de softwares para alinhamento e detecção de *Single Nucleotide Polymorphisms* (SNPs) / Marcelo Gonçalves Narciso ...[et al.]. - Santo Antônio de Goiás : Embrapa Arroz e Feijão, 2014.
37 p. - (Documentos / Embrapa Arroz e Feijão, ISSN 1678-9644 ; 290)

1. Alinhamento de sequências – software. 4. Marcador molecular. 3. Polimorfismo genético. 2. Detecção de SNPs. 5. Marcador genético – software. I. Narciso, Marcelo Gonçalves. II. Embrapa Arroz e Feijão. III. Série.

CDD 660.65 (21. ed.)

Autores

Marcelo Gonçalves Narciso
Engenheiro eletrônico, Ph.D. em Bioinformática, pesquisador da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, marcelo.narciso@embrapa.br

Adelmo Martins Rodrigues
Engenheiro agrônomo, Doutorando em Genética e Biologia Molecular pela Universidade Federal de Goiás, bolsista da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, adelmomartinsrodrigues@gmail.com

Jorge Freitas Cieslak
Biólogo, Mestrando em Genética e Biologia Molecular pela Universidade Federal de Goiás, bolsista da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, jorge_cieslak@hotmail.com

Ricardo Diógenes Dias Silveira
Biólogo, Doutorando em Biologia Celular e Molecular pela Universidade Federal de Goiás, bolsista da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO, ricardo_biologia@hotmail.com

Rosana Pereira Vianello
Bióloga, Doutora em Biologia Molecular Vegetal, pesquisadora da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO,
rosana.vianello@embrapa.br

Cláudio Brondani
Engenheiro agrônomo, Doutor em Biologia Molecular, pesquisador da Embrapa Arroz e Feijão, Santo Antônio de Goiás, GO,
claudio.brondani@embrapa.br

Apresentação

Single Nucleotide Polymorphisms (SNPs) representam uma fonte abundante de variação genética, já que são variações na sequência do DNA e ocorrem quando um único nucleotídeo difere no genoma entre os indivíduos de uma espécie ou entre pares de cromossomos.

Marcadores moleculares SNPs podem ser utilizados em estudos de associação e mapeamento genético, assim como em ensaios diagnósticos para confirmação de paternidade, identificação individual (rastreabilidade), detecção de doenças genéticas e polimorfismos associados a características de interesse (produtividade de grãos, cultivares com tolerância à seca e outras).

Dada a importância desses marcadores, este trabalho tem como enfoque a apresentação de softwares para alinhamento de sequências genômicas e posterior detecção de SNPs e também a comparação destes quanto ao tempo de execução e à qualidade de resultados.

Flávio Breseghello
Chefe Geral da Embrapa Arroz e Feijão

Sumário

Um Estudo Comparativo de Softwares para Alinhamento e Detecção de *Single Nucleotide Polymorphisms* (SNPs)

Marcelo Gonçalves Narciso
Adelmo Martins Rodrigues
Jorge Freitas Cieslak
Ricardo Diógenes Dias Silveira
Rosana Pereira Vianello
Cláudio Brondani

Introdução

Existem vários programas para identificar SNPs (*Single Nucleotide Polymorphisms*). Estes programas, geralmente, têm como entrada dados arquivos de ressequenciamento de uma variedade em estudo e arquivo fasta de uma variedade referência com genoma conhecido, ambos da mesma espécie, ou em caso de não existir o genoma de referência, é possível usar um genoma de referência de espécie similar.

Cada software tem sua complexidade de instalação e curva de aprendizado. Assim, vale a pena conhecer cada um para melhor usar, ou ainda escolher um que melhor se adapte às necessidades de pesquisa. No caso deste trabalho, os softwares enfocados são para alinhamento de sequências (Bowtie2, BWA, Mosaik, Panati, CLC Genomics) e posteriormente usados para a descoberta de SNPs (SAMtools, Panati, GigaBayes, GATK, VarScan e CLC Genomics). Estes softwares foram escolhidos por serem muito usados e também estarem citados em vários artigos, conforme pode ser visto em (GRATTAPAGLIA et al., 2011; KOBOLDT et al., 2009). Tais softwares podem ser instalados e executados em ambiente Linux ou Windows (CLC Genomics). Os procedimentos de instalação destes softwares estão amplamente divulgados na Internet. O maior desafio, portanto, é a curva de aprendizado, assim como ocorre com qualquer software que se deseja utilizar.

Estes softwares podem obter SNPs que existem, SNPs que não existem (falso-positivo) e também não mostrar SNPs que existem (falso-negativo). Tudo depende de como os parâmetros de cada um dos softwares são configurados. Desta forma, os parâmetros considerados quando os programas de busca de SNPs são executados são muito importantes.

Este trabalho enfoca softwares para alinhamento e identificação de SNPs e compara estes quanto ao tempo de execução e qualidade de resultados. Considerações sobre SNPs obtidos e que são falso-positivos e SNPs que não são obtidos e que existem também serão considerados neste trabalho e, para que evite ao máximo a perda de SNPs verdadeiros e minimize a quantidade de SNPs falsos, parâmetros de execução dos programas serão também discutidos.

Materiais e Métodos

Os softwares usados para identificar SNPs a partir de um genoma referência e um arquivo contendo ressequenciamento de outra variedade de mesma espécie, a serem usados neste trabalho, estão descritos brevemente a seguir.

BWA (BWA, 2013), Mosaik (MOSAIK, 2013), e Bowtie2 (BOWTIE2, 2013) são usados para alinhamento de sequências (variedade não referência em relação à referência). SAMtools (SAMTOOLS, 2013) contém um conjunto de programas para obter SNPs a partir destes alinhamentos, bem como VarScan (VARSCAN, 2013), gigaBayes (GIGABAYES, 2013) e GATK (GATK, 2013). Assim, serão usados softwares em conjunto, como por exemplo, Bowtie2 (é um ALIGNER, isto é, faz alinhamento da variedade referência e a variedade em estudo) e SAMtools (é um CALLER, isto é, obtém SNPs a partir dos resultados de um ALIGNER) e também BWA e SAMtools para obter SNPs, e outros a serem mostrados mais adiante. O software Mosaik será usado em conjunto com o software GigaBayes (GIGABAYES, 2013) para obter SNPs. O software Panati (PANATI, 2013) faz tanto o

alinhamento das sequências (referência e não referência) quanto obtém os SNPs e demais dados associados. GATK e VarScan também são tais como SAMtools, isto é, obtêm SNPs a partir de um resultado de um ALIGNER e um CALLER, o software SAMtools. Todos estes são softwares disponíveis para download (software livre). Um software proprietário, a ser usado para melhor comparar estes sistemas é o CLC Genomics (CLCBIO, 2013). Este software oferece uma interface gráfica e também mais recursos para análise de SNPs. Todos estes softwares possuem uma série de parâmetros para execução com a finalidade de melhor filtrar e avaliar os SNPs.

Estes programas são preferencialmente para ambiente Linux, com exceção do CLC, mas podem ser instalados em ambiente Windows, com algumas adaptações em seu arquivo de instalação, conhecidos como Makefile, o qual contém os comandos em sequência para instalação. É necessário que se entenda cada um destes comandos e veja os análogos que o Windows tem para então fazer as alterações no arquivo Makefile e então executar o mesmo para que seja feita a instalação. Não é um processo simples. Portanto, neste caso, a instalação é feita através de linha de comando. Para ambiente Linux, a instalação é relativamente simples com respeito aos softwares citados, desde que as bibliotecas usadas pelos programas estejam instaladas. Sobre o software CLC Genomics, este tem versão para Linux e Windows e possui instalador com interface gráfica, o que facilita bastante a instalação.

Os softwares BWA (Burrows-Wheeler Aligner), Panati, Mosaik, Bowtie2, GATK, VarScan e SAMtools são executados em mais de um comando, os quais serão vistos mais a frente no item “Resultados e Discussão”. Existem comandos destes softwares que podem usar mais de um processador e assim paralelizar o processamento. Para o CLC Genomics, será usada a versão que utiliza mais de um processador.

Estes softwares permitem ao usuário usar parâmetros diversos para a execução, os quais serão exemplificados mais a frente neste trabalho.

Destes parâmetros, alguns geram SNPs que são confiáveis (mais restritivos) e, conforme valores atribuídos a estes parâmetros, SNPs do tipo "falso-positivo" podem aparecer. Diversos parâmetros foram testados no sentido de obter SNPs verdadeiros (sem falso-positivo ou com uma probabilidade muito baixa de existir falso-positivo), os quais serão explicados na seção "Resultados e Discussão". Cada software tem parâmetros diversos para que possa ter um comportamento mais rigoroso ou não durante a detecção de SNPs.

Os dados usados para se obter SNPs são das variedades de arroz Nipponbare (referência) e Caiapó. Os dados relativos à variedade Nipponbare são conhecidos e podem ser obtidos a partir do site do NCBI (NCBI, 2013) e os dados relativos à variedade Caiapó foram obtidos a partir de ressequenciamento destas no laboratório *Plant and Breeding*, na Universidade de Cornell, USA, e ainda não foram divulgados ao público.

Resultados e Discussão

Os softwares Bowtie2, SAMtools, Panati, Mosaik, GigaBayes, VarScan, GATK e BWA foram instalados em ambiente Linux, em uma máquina com 16 processadores (2.8 MHz de *clock* por processador) e memória RAM 4 GB, sistema operacional CentOS v5.5. Cada um destes softwares tem um modo de execução que paraleliza o processamento, e assim são executados mais rapidamente. Para mais informações de como instalar estes softwares, basta consultar os sites descritos para Bowtie2 (BOWTIE2, 2013), SAMtools (SAMTOOLS, 2013), Panati (PANATI, 2013), BWA (BWA, 2013), GATK (GATK, 2013), VarScan (KOBOLDT et al., 2009) e Mosaik (MOSAIK, 2013).

O software CLC Genomics foi executado em ambiente Windows, sistema operacional Windows 7, 8 GB de memória RAM e processador de 8 GHz de *clock*. A instalação do software é tal como descrito no manual que vem com o mesmo, e que também pode ser obtido em CLC Genomics Workbench (2013).

Para se obter um conjunto de SNPs com estes softwares, considerando como variedade referência tropical Japônica Nipponbare (genoma completo em arquivo formato fasta) e como variedade que se deseja ter SNPs, a Caiapó (esta tem arquivo ressequenciado, em formato fastq), serão discutidos cada software caso a caso, isto é, como rodar estes softwares e os seus resultados.

BWA e SAMtools

A forma de executar estes softwares, em ambiente Linux (seria similar em ambiente Windows), está descrita a seguir para o caso paired-end (alinhamento paired-end), que pode ser visto em BWA (2013). Cada linha contém um comando a ser executado e cada comando deve ser executado exatamente na ordem como está a seguir.

Observe que cada linha começa com um comando, alguns parâmetros e também os arquivos de entrada e saída. A seguir, os comandos, numerados na sequência de execução.

- bwa index -a bwtsw nipponbare.fasta
- bwa aln nipponbare.fasta caiapo-f.fastq -t 16 > saida.aln
- bwa aln nipponbare.fasta caiapo-r.fastq -t 16 > saida2.aln
- bwa sampe -f saida.sam nipponbare.fa saida.aln saida2.aln caiapo-f.fastq caiapo-r.fastq
- samtools view -bS saida.sam > saida.bam
- samtools sort saida.bam saidaSorted.bam
- samtools mpileup -uf nipponbare.fasta saidaSorted.ba -B | bcftools view -bvcg - > saidaFinalSNPs.txt
- bcftools view saidaFinalSNPs.txt | vcfutils.pl varFilter –d 10 > var.flt.vcf

Um trecho do arquivo saidaFinalSNPs.txt (primeiras colunas e duas primeiras linhas) está descrito a seguir

#CHROM	POS	ID	REF	ALT	QUAL
chr01 | 13101	478	.	T	G	3.54

Observe que, do trecho descrito acima, para o cromossomo 1, na posição 478, tem-se que a variedade Nipponbare tem a base nitrogenada T e a variedade Caiapó tem a base nitrogenada G e a qualidade deste SNP vale 3,54, a qual é considerada boa (valor acima de 2). Outras informações existem neste arquivo, porém estas são as informações mais relevantes.

Os comandos iniciais do BWA são usados para alinhamento de sequências. Este programa alinha sequências de nucleotídeos relativamente curtas contra uma sequência de referência longa (no caso, trata-se da variedade referência Nipponbare). Nestes comandos, o parâmetro "-t 16" significa que o comando deverá usar os 16 processadores do hardware para paralelizar o processamento. O arquivo "saida.sam" contém dados referentes ao alinhamento de Caiapó com Nipponbare. O arquivo do tipo SAM (Sequence Alignment/Map) contém um formato genérico para armazenar alinhamentos de sequência de nucleotídeos muito grandes. A opção "sampe" é para o pareamento início fim da sequência "paired-end". Existe também a opção "samse" que é para o caso "single-end". Para esta opção, não se faz necessário o arquivo reverso (caiapo-r.fastq), apenas o forward (caiapo-f.fastq). Estes arquivos contêm sequências ressequenciadas em sentido reverso e direto, respectivamente. Para o caso da opção "samse", os comandos seriam:

- bwa index -a bwtsw nipponbare.fasta
- bwa aln nipponbare.fasta caiapo-f.fastq -t 16 > saída.aln
- bwa samse -f saida.sam nipponbare.fa saída.aln caiapo-f.fastq

Mais informações sobre estas e outras opções de comando podem ser vistas em BWA (2013).

Após o uso dos comandos do software BWA, são usadas algumas ferramentas do kit SAMtools, isto é, os comandos *samtools, bcftools e vcfutils.pl*. A partir do arquivo *saida.sam*, é gerado o arquivo *saida.bam* (Binary Alignment/Map) e posteriormente este mesmo arquivo, só que

com os dados ordenados. Em seguida, é feita a identificação de SNPs através da opção "–mpileup" e posteriormente os dados são colocados em formato legível ao usuário através do comando bcftools. A última linha de comando, na qual aparece o seguinte comando:

bcftools view saidaFinalSNPs.txt | vcfutils.pl varFilter –d 10 > var.flt. vcf

Inicialmente, o arquivo *saidaFinalSNPs.txt* é convertido para o format VCF, um tipo de arquivo texto, com o comando (*bcftools view saidaFinalSNPs.txt*). Após isto, a saída em formato VCF é usada como entrada para o comando vcfutils.pl, a qual contém a opção para filtro vrFilter e o valor de parâmetro –d 10. Isso significa uma filtragem de SNPs que estejam em *reads* com pelo menos 10x de cobertura. Assim, a opção varFilter é usada para dar o caráter de filtro de SNPs, como por exemplo a filtragem por cobertura mínima e máxima, o tamanho da janela de leitura, o p-value, entre outras. Mais detalhes podem ser vistos em SAMtools (2013).

Os comandos citados acima, ao serem inseridos em um arquivo do tipo Shell script, podem ser executados em sequência e assim o usuário não precisa ficar executando comando por comando, mas apenas o arquivo Shell. Para exemplificar, basta fazer o seguinte, em ambiente Linux:

- criar o arquivo bwa-samtools.sh ou qualquer outro nome, com a extensão .sh.
- editar o arquivo e inserir os comandos
  - #!/bin/bash
  - *bwa index -a bwtsw nipponbare.fasta*
  - *bwa aln nipponbare.fasta caiapo-f.fastq -t 16 > saida.aln*
  - *bwa aln nipponbare.fasta caiapo-r.fastq -t 16 > saida2.aln*
  - *bwa sampe -f saida.sam nipponbare.fa saida.aln saida2.aln caiapo-f.fastq caiapo-r.fastq*

- *samtools view -bS saida.sam > saida.bam*
- *samtools sort saida.bam saidaSorted.bam*
- *samtools mpileup -uf nipponbare.fasta saidaSorted.ba -B | bcftools view -bvcg - > saidaFinalSNPs.txt*
- *bcftools view saidaFinalSNPs.txt | vcfutils.pl varFilter –d 10 > var.flt.vcf*

- Salvar o arquivo
- Mudar a permissão para execução (chmod 700 bwa-samtools.sh) e executar.

O comando "#!/bin/bash", no passo 2, indica que o arquivo deverá ser executado usando a linguagem Shell chamada Bash. Os demais comandos são os que foram descritos anteriormente. A seguir, basta salvar o arquivo contendo os comandos, mudar a permissão do mesmo para que possa ser executado (executar o comando "*chmod 700 bwa-samtools.sh*") e, em seguida, executar o Shell script *bwa-samtools.sh* em linha de comando e esperar o resultado. Esta ação vale para todos os demais softwares que serão descritos (Bowtie2 + SAMtools, Panati, Mosaik, etc.), com exceção do CLC.

Bowtie2 e SAMtools

Bowtie2 tem a mesma função que BWA, porém algumas opções mudam. Para exemplificar, segue a forma de obter SNPs com Bowtie2 e SAMtools, no modo "single end". A sequência de comandos está numerada conforme a ordem de execução.

- bowtie2-build -f nipponbare.fasta saida.idx
- bowtie2 -x saida.idx -q caiapo-f.fastq -S saida.sam
- samtools view -bS saida.sam > saida.bam
- samtools sort saida.bam saidaSorted.bam
- samtools mpileup -uf nipponbare.fasta saidaSorted.bam -B | bcftools view -bvcg - > saidaFinalSNPs.txt
- bcftools view saidaFinalSNPs.txt | vcfutils.pl varFilter –d 10 > var.flt.vcf

Para executar o software Bowtie2, o primeiro comando é bowtie2-build, que faz uma indexação do genoma referência (nipponbare.fasta), e a saída é o arquivo saida.idx, o qual serve de entrada para o comando bowtie2, logo a seguir. Neste, tem-se outra entrada, o arquivo fastq de Caiapó e gera uma saída no formato SAM, tal como foi feito para o BWA. A seguir, os comandos do kit SAMtools são executados para gerar o arquivo final relativo aos SNPs obtidos, tal como explicado no item *BWA e SAMtools*.

Para o caso "paired end", o comando boltiew2 mudaria e ficaria desta forma:

bowtie2 -x saida.idx -q -1 caiapo-f.fastq -2 caiapo-r.fastq -S saida.sam

As opções -1 e -2 são para os arquivos fastq relativos a Caiapó, tal como descrito acima. Para executar estes comandos, basta fazer um script com os comandos citados, tal como foi feito para BWA e SAMtools, descrito no item BWA e SAMtools. O arquivo de saída é tal como o que foi descrito para o caso anterior, item BWA e SAMtools, visto que foi gerada pela ferramenta SAMtools.

Uma observação a ser feita para o caso do Bowtie2 é que este funciona bem para *reads* com pelo menos 50 pb, conforme descrito em (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml). Se o número for menor do que isto, vale a pena usar a versão 1 do bowtie, que está descrita também em Bowtie2 (2013). A forma de usar é praticamente a mesma do Bowtie2, isto é, nos comandos mostrados, basta trocar bowtie2 por bowtie.

Panati

Panati (PANATI, 2013) é um software feito para fazer alinhamento de sequências e também descoberta de SNPs. É executado no modo "paired-end" por default e assim tem como entrada de dados os arquivos forward e reverse. Os comandos usados para obter SNPs, na sequência correta de execução, são:

- panati-build -f nipponbare.fasta -o nipponbare.idx -l refseq -w 16 -s 1 -m 3000000000
- fastq-qc -f caiapo-f.fastq -r caiapo-r.fastq -i 300 -o qcInput.txt -t 30 --3p-trim=15 -z

- panati -r nipponbare.idx -f qcInput.txt -l 1024 -m 0.10 -g 0.013 --scan-shift = 16 --n-threads = 16 -o qcInputPanati.txt
- coverage-report -p qcInputPanati.txt -d 5 -m 0 > saidaEstatistica.txt
- echo "sample1 qcInputPanati.txt" > lista.txt
- cat lista.txt | combine-samples -d 2 -r 0 -p nipponbare.idx > saidaSNPs.txt
- impute-missing -f saidaSNPs.txt -l 100 -p 0.95 -h 0.01 --exclude = my-outgroup-sample > saidaSNP-v2s.txt

O primeiro comando (panati-build) é usado para construir um índice relativo ao genoma referência, tal como foi feito nos casos de BWA e Bowtie2. O segundo comando é para, a partir dos arquivos com formato fastq de Caiapó, construir um arquivo de saída que contém informações sobre estes dois arquivos fastq de entrada. Em seguida, as saídas dos dois comandos iniciais são usadas para fazer o alinhamento, tal como foi feito para os softwares Bowtie2 e BWA. Neste caso, a opção --n-threads = 16 indica que o processamento deverá ser feito usando os 16 processadores do hardware considerado e –g 0.013 indica que os gaps nos alinhamentos deverão ser levados em conta e eliminados se forem mais que 1,3% da sequência considerada. O quarto comando (coverage-report) mostra uma estatística dos dados obtidos, para ver se houve erro de execução. O quinto e sexto comandos são para obter os SNPs e o último comando, impute-missing, gera mais estatísticas sobre os SNPs obtidos. Mais informações sobre os parâmetros que são usados pelo programa Panati podem ser vistos em Panati (2013). Um exemplo de um trecho da saída do programa Panati (saidaSNPs.txt) está descrito a seguir para os parâmetros g = 0,01 e m = 0,01.

refseq	pos	offset	ref.allele	var.allele
chr01|13101	8541	0	C	A
chr01|13101	8544	0	C	T
chr01|13101	8547	0	A	G
chr01|13101	8584	0	A	C

Conforme observado, o programa mostra o número do cromossomo, a posição do SNP no cromossomo e os SNPs obtidos. A saída não mostra algum parâmetro ou análise estatística que permita medir a qualidade do SNP e esta é uma possível desvantagem deste software.

Para finalizar, a Tabela 1 mostra a quantidade de SNPs obtidos conforme os parâmetros m e g, descritos anteriormente, para o genoma do arroz, variedade Nipponbare.

Tabela 1. Variação da quantidade de SNPs obtidos com parâmetros de g e m.

m	g	Quantidade de SNPs obtidos
0,050	0,010	62854
0,060	0,010	68750
0,070	0,010	72784
0,080	0,010	73550
0,090	0,010	75857
0,10	0,010	77410
0,10	0,011	77410
0,10	0,015	707595
0,10	0,020	718308
0,10	0,030	715411
0,10	0,040	784119
0,10	0,050	786196
0,10	0,060	820013
0,10	0,070	822918

Na Tabela 1, observa-se que quanto menores são os valores de m e g, menos SNPs são obtidos. Porém, como Panati não faz uma estatística que mostre valores de p-value, por exemplo, de cada SNP, tem-se que não vale a pena ter valores de m e g maiores do que m = 0,10 e g = 0,010, pois embora o número de SNPs aumente, o número de SNP falso-positivo também aumenta consideravelmente, conforme testes realizados e confrontados com outros softwares. Assim, para este trabalho, vamos adotar os valores m = 0,05 e g = 0,01.

Mosaik

Este software, tal como BWA e Bowtie2, faz o alinhamento de sequências. O software Mosaik, após todo o trabalho de alinhamento, fornecerá dados que servirão de entrada para o programa GigaBayes

para então obter SNPs. Para executar o Mosaik, são executados comandos em uma sequência, a qual está descrita a seguir, para o caso "paired-end".

- MosaikBuild -fr nipponbare.fasta -oa nipponbare.dat
- MosaikBuild -q caiapo-f.fastq –q2 caiapo-r.fastq -out caiapo.dat -st illumina
- MosaikAligner -in caiapo.dat -out caiapo_aligned.dat -ia nipponbare.dat -hs 14 -act 17 -mm 2 -m unique -p 16
- MosaikSort -in caiapo_aligned.dat -out caiapo_sorted.dat
- MosaikAssembler -in caiapo_sorted.dat -out caiapo_test -ia nipponbare.dat -f gig
- MosaikCoverage -in caiapo_aligned.dat -ia nipponbare.dat -od graphs –cg > saidaCobertura.txt
- gigaBayes --gig caiapo_cromo1.gig --gff caiapo_NC_008394.gff3 --R 10000000 --ploidy diploid --anchor --indel --PSL 0.9 --CAL1 5 --CAL2 5
- gigaBayes --gig caiapo_cromo2.gig --gff caiapo_NC_008395.gff3 --R 10000000 --ploidy diploid --anchor --indel --PSL 0.9 --CAL1 5 --CAL2 5

Para cada cromossomo, é gerado um arquivo de saída em formato *gff3*. Assim, para cada cromossomo deverá ser executado o programa GigaBayes com os seus respectivos parâmetros. Abaixo, tem-se o comando para obter SNPs do cromossomo 12.

- *gigaBayes --gig caiapo_cromo12.gig --gff caiapo_NC_008405.gff3 --R 10000000 --ploidy diploid --anchor --indel --PSL 0.9 --CAL1 5 --CAL2 5*

No exemplo dado anteriormente, comando *MosaikBuild* serve para construir um índice relativo ao arquivo fasta de nipponbare e os arquivos fastq relativos ao ressequenciamento illumina de Caiapó (forward e reverse). Os arquivos nipponbare.dat e caiapo.dat são arquivos gerados neste processo. Em seguida, o comando *MosaikAligner* faz um alinhamento dos *reads* de Caiapó com a variedade referência Nipponbare. Este comando pode ser dividido entre os 16 processadores através da opção "–p 16". Em seguida,

é ordenado o resultado obtido pelo comando *MosaikAligner* através do comando *MosaikSort*. Em seguida, são criados alinhamento de sequências múltiplas através do comando *MosaikAssembler* e então é gerado um arquivo no formato .gig (ou .ace, se usar opção –ace em lugar de --gig). Este arquivo, ".gig", pode ser lido com o software gigaBayes (lê arquivos no formato .gig), que foi o software usado neste trabalho. Existe uma possibilidade de se ver graficamente os resultados através da ferramenta EagleView, que pode ser vista em Eagleview (2013). Um exemplo de saída do programa GigaBayes está descrito abaixo, com um trecho do arquivo de saída *caiapo_cromo1.gig*.

- chr01|13101 gigaBayes SNP 2500058 2500058 1.. alleles=C,T;individualGenotypes=HWI:CT|1,chr01|13101:CC|0.6666
- chr01|13101 gigaBayes INDEL 6102825 6102826 1..alleles=C,-;individualGenotypes=HWI:C-|1,chr01|13101:--|0.6666
- chr01|13101 gigaBayes SNP 7228192 7228192 1.. alleles=A,G;individualGenotypes=HWI:AG|1,chr01|13101:GG|0.6666
- chr01|13101 gigaBayes SNP 12584224 12584224 1.. alleles=G,A;individualGenotypes=HWI:GA|1,chr01|13101:GG|0.6666
- chr01|13101 gigaBayes SNP 12899395 12899395 1.. alleles=C,T;individualGenotypes=HWI:CT|1,chr01|13101:CC|0.6666
- chr01|13101 gigaBayes SNP 15386823 15386823 1.. alleles=C,T;individualGenotypes=HWI:CT|1,chr01|13101:TT|0.6666

No exemplo acima, em cada registro de saída, tem-se o número do cromossomo, se existe um SNP ou INDEL, a posição do SNP/INDEL, e a mudança de base do SNP/INDEL e demais valores, que podem ser vistos em Mosaik (2013).

VarScan

Sejam os softwares Bowtie/Bowtie2, BWA ou outro aligner, por exemplo, que forneçam arquivo do tipo sam como saída, conforme mostrado anteriormente. Este arquivo em formato sam poderá ser transformado em formato bam e depois ordenado (sorted bam). Com este arquivo, é possível usar uma ferramenta de análise conhecida por varScan, a qual fornece como resultados a obtenção de SNPs e

InDels de forma bastante confiável. Esta avaliação é feita baseada na quantidade de reads, frequência alélica e qualidade do SNP (calculado estatisticamente). Mais informações podem ser vistas em Koboldt et al. (2009) e VarScan (2013). Para exemplificar o uso, um possível script (sequência de comandos na ordem correta de execução) é o que está a seguir, para o caso de se usar o bowtie2.

- bowtie2-build -f nipponbare.fasta saida.idx
- bowtie2 -x saida.idx -q caiapo-f.fastq -S saida.sam
- samtools view -bS saida.sam > saida.bam
- samtools sort saida.bam saidaSorted.bam
- samtools mpileup -f . /nipponbare.fasta saidaSorted.bam > caiapo.mpileup
- java -jar ./VarScan.v2.3.2.jar mpileup2snp caiapo.mpileup --min-coverage 8 --min-reads2 2 --min-var-freq 0.01 --min-avg-qual 15 --p-value 0.05 --strand-filter 0 --min-freq-for-hom 0.75 > caiapo.txt

No exemplo acima, no último comando, relativo ao varScan, o arquivo VarScan.v2.3.2.jar estava no diretório corrente no qual foi executado o comando "java -jar jar ./VarScan.v2.3.2.jar... "Se o comando for executado em diretório no qual o referido arquivo não esteja, então o caminho completo do diretório no qual VarScan.v2.3.2.jar esteja, deverá ser especificado (por exemplo, java –jar /usr/local/lib/varScan/VarScan.v2.3.2.jar....).

A saída do comando varScan, redirecionada para o arquivo caiapo.txt, tem várias colunas, e as principais são: cromossomo, posição do SNP, base nitrogenada da variedade referência e base nitrogenada da variedade qualquer (para mostrar o snp), cobertura, p-value, frequência, etc. Um trecho da saída é tal como abaixo:

Chrom Position Ref Var Cons:Cov:Reads1:Reads2:Freq:P-value

chr01|13101 2812358 G A A:8:0:8:100%:7,77E-5
chr01|13101 7760777 T G K:38:32:6:15,79%:1,2628E-2

chr01|13101 7762448 G A A:63:0:63:100%:1,657E-7
chr01|13101 7916470 T C C:11:0:11:100%:1,4176E-6
chr01|13101 7916491 T C C:10:0:10:100%:5,4125E-6
chr01|13101 7916630 T C C:12:0:12:100%:3,698E-7
chr01|13101 7916713 T C C:13:2:11:84,62%:1,0096E-5
chr01|13101 9957361 A G R:9:5:4:44,44%:4,1176E-2
chr01|13101 11104595 A G R:9:5:4:44,44%:4,1176E-2

A saída acima descreve o cromossomo analisado, a posição de cada SNP, as bases nitrogenadas das variedades referência e não referência e dados importantes como p-value.

Assim, com valores de p-value, podem-se filtrar os resultados a partir de um p-value máximo (por exemplo, 0.01 ou 1.0E-2) e ter os SNPs com uma boa margem de segurança de não serem falso-positivos.

GATK

GATK (Genome Analysis Toolkit) é um pacote de software desenvolvido para analisar dados provenientes de ressequenciamento de nova geração. Este conjunto de ferramentas (Toolkit) serve para a descoberta de conhecimento em relação aos dados e também para verificar a qualidade dos mesmos. GATK é altamente genérico e pode ser aplicado em vários tipos de análise a partir de um conjunto de dados obtido por qualquer que seja a tecnologia (Illumina, SOLID, etc.).

Para o caso de se identificar SNPs, GATK é usado em conjunto com uma ferramenta tal como BWA (ou Bowtie2) e SAMtool. Para exemplificar o uso deste software para obter SNPs, pode ser feito o seguinte script, usando BWA e SAMtool, descrito a seguir.

- bwa index -a bwtsw nipponbare.fasta -p nipponbare
- bwa aln -t 16 -f caiapo.sai -l nipponbare ../caiapo-f.fastq
- bwa aln -t 16 -f caiapo2.sai -l nipponbare ../caiapo-r.fastq
- bwa sampe -f saida.sam -r "@RG\tID:id\tLB:foo\tSM:id\

tPL:ILLUMINA" nipponbare caiapo.sai caiapo2.sai ../caiapo-f.fastq ../caiapo-r.fastq

- samtools faidx nipponbare.fasta

Após a geração destes arquivos, é usado então o pacote GATK, que vem com uma série de arquivos do tipo .jar (FixMateInformation.jar, SortSam.jar, MarkDuplicates.jar, GenomeAnalysisTK.jar, que é o mais usado, e vários outros). Para se obter SNPs, a partir dos dados já gerados, tem-se uma sequência de execuções de comandos GATK, os quais estão descritos a seguir.

- java -Xmx5g -jar FixMateInformation.jar I=saida.sam O=fixed.sam SO=coordinate VALIDATION_STRINGENCY=LENIENT
- java -jar AddOrReplaceReadGroups.jar I=saidaSorted.bam O=saida-v2.bam LB=SNP PL=illumina PU=ILLUMINA SM=idSM VALIDATION_STRINGENCY=LENIENT CREATE_INDEX=true SORT_ORDER=coordinate
- Pode ser usado o parâmetro LENIENT, IGNORE ou SILENT para VALIDATION_STRINGENCY
- java -jar SortSam.jar I=saida-v2.bam O=saida-v3.bam SORT_ORDER=coordinate VALIDATION_STRINGENCY=LENIENT
- java –Xmx5g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R nipponbare.fasta -o saida2.bam.list -I saida-v3.bam –U

Este passo foi feito para inserir a tabela gerada no arquivo saida2.bam.list. Quando terminado este passo, então é executado o comando abaixo:

- java –Xmx5g -Djava.io.tmpdir=/tmp -jar GenomeAnalysisTK.jar -I saida-v3.bam -R nipponbare.fasta -T IndelRealigner -targetIntervals saida2.bam.list -o saida.realigned.bam –U

A identificação de SNP é feita através de várias opções (UnifiedGenotyper, VariantFiltration, BaseRecalibrator, PrintReads). Através destas opções, são detectados SNPs e INDELs (sequência

curta) ao mesmo tempo em que gera a saída em formato VCF (Variant Call Format). Este formato é um texto padrão para representar SNP, Indel, outros. Mais informações sobre VCF podem ser vistas em http://www.broadinstitute.org/gatk/guide/topic?name=intro. Em seguida, é executado o seguinte conjunto de comandos:

- java -Xmx5g -jar GenomeAnalysisTK.jar -glm BOTH -R nipponbare.fasta -T UnifiedGenotyper -I saida.realigned.bam -o snps.vcf -metrics snps.metrics -stand_call_conf 50.0 -stand_emit_conf 10.0 -dcov 1000 -A DepthOfCoverage -A AlleleBalance
- java -Xmx5g -jar GenomeAnalysisTK.jar -R nipponbare.fasta -T VariantFiltration -V snps.vcf -o snp.filtered.vcf --clusterWindowSize 10 --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) > 0.1)" --filterName "HARD_TO_VALIDATE" --filterExpression "DP < 5 " --filterName "LowCoverage" --filterExpression "QUAL < 30.0 " --filterName "VeryLowQual" --filterExpression "QUAL > 30.0 && QUAL < 50.0 " --filterName "LowQual" --filterExpression "QD < 1.5 " --filterName "LowQD" --filterExpression "SB > -10.0 " --filterName "StrandBias"
- java -Xmx5g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R nipponbare.fasta -I saida.realigned.bam -knownSites snps.vcf -o recal_data.grp
- java -Xms5g –jar GenomeAnalysisTK.jar -T PrintReads -R nipponbare.fasta -I saida.realigned.bam --BQSR recal_data.grp --baq CALCULATE_AS_NECESSARY -o output.bam
- java -Xmx5g -jar GenomeAnalysisTK.jar -nt 16 -T UnifiedGenotyper -R nipponbare.fasta -I output.bam -o resultSNPs.vcf -metrics UniGenMetrics -stand_call_conf 50.0 -stand_emit_conf 10.0 -dcov 1000 -A DepthOfCoverage -A AlleleBalance

Após esta sequência de comandos, tem-se então o arquivo *resultSNPs.vcf,* o qual contém todos os SNPs identificados para Caiapó em relação à variedade Nipponbare. Um trecho do arquivo (cabeçalho e registros) é tal como abaixo:

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT idSM

chr01|13101 45471 . A G 36.97 LowQual ABHom=0.750;AC=2;AF=1.00;AN=2;DP=4;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ=25.11;MQ0=1;OND=0.250;QD=9.24;SB=-6.519e-03 GT:AD:DP:GQ:PL 1/1:0,3:4:9:69,9,0

chr01|13101 48412 . G T 36.97 LowQual ABHom=0.889;AC=2;AF=1.00;AN=2;DP=9;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ=21.00;MQ0=1;OND=0.111;QD=4.11;SB=-2.221e+01 GT:AD:DP:GQ:PL 1/1:0,8:9:9:69,9,0

chr01|13101 48413 . A T 15.88 LowQual ABHom=1.00;AC=2;AF=1.00;AN=2;DP=9;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ=21.00;MQ0=1;QD=1.76;SB=-6.519e-03 GT:AD:DP:GQ:PL 1/1:0,9:9:6:47,6,0

chr01|13101 94445 . G A 59.51 . ABHom=0.909;AC=2;AF=1.00;AN=2;DP=11;Dels=0.00;FS=0.000;HaplotypeScore=0.0000;MLEAC=2;MLEAF=1.00;MQ=23.97;MQ0=1;OND=0.091;QD=5.41;SB=-4.375e+01 GT:AD:DP:GQ:PL 1/1:0,10:11:12:92,12,0

CLC Genomics

Este software, para este estudo, foi instalado em ambiente Windows, sistema operacional Windows 7, o qual é executado em hardware que tem uma CPU de 8 GHz de clock e possui também 8 GB de memória RAM. O sistema CLC Genomics, ao iniciar, oferece as opções “File”, “Edit”, “Download”, “View”, “Toolbox”, “Workspace” e “Help”. Para obter SNPs com este software, o primeiro passo é importar o arquivo do tipo fasta da variedade referência (por exemplo, Nipponbare) e arquivos ressequenciados das variedades, a partir dos quais são obtidos SNPs (arquivos com formato fastq, por exemplo). A Figura 1 ilustra o que foi mencionado.

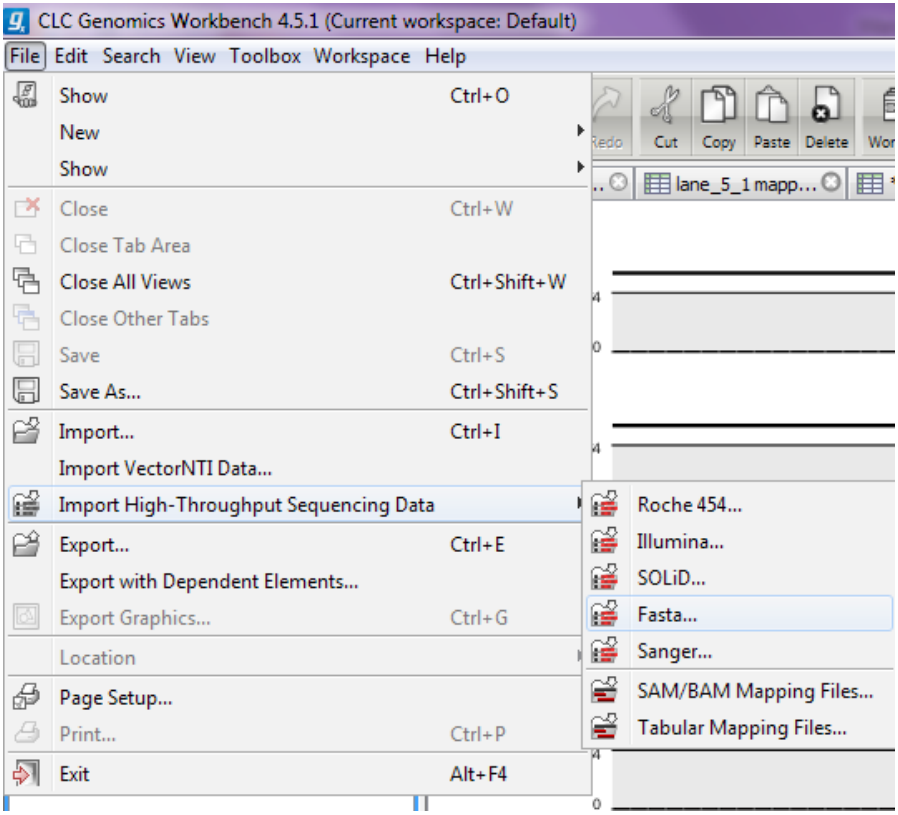

Figura 1. Importação de arquivos – fasta, fastq Illumina, etc.

Após importar estes arquivos, é usada a opção ToolBox para obter um mapa de cada variedade (escolher a opção “Map Reads to Reference” e depois carregar o arquivo fastq ou proveniente do ressequenciamento), o que leva algumas horas ou até mesmo um dia. A Figura 2 ilustra como construir este mapa.

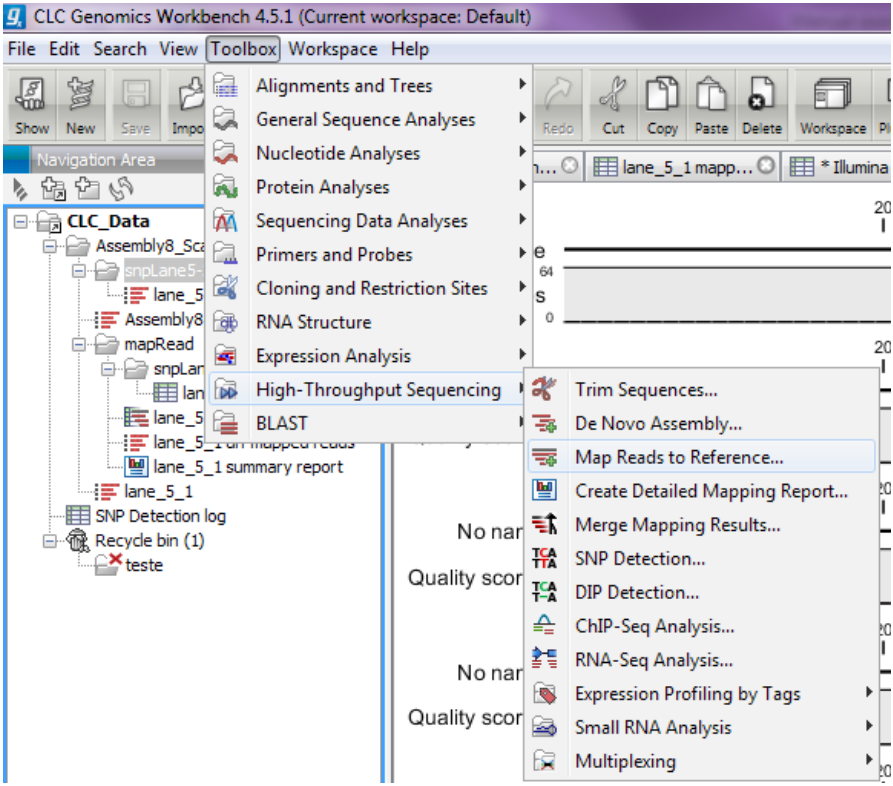

Figura 2. Construção de mapas (variedades não referência).

Após o carregamento das variedades referência e não referência, o próximo passo é a identificação de SNPs. Para isto, basta acessar a opção "ToolBox" na barra de ferramentas, depois a opção "High-Throughput Sequencing" e em seguida a opção "SNP Detection", descrito na Figura 3, a seguir.

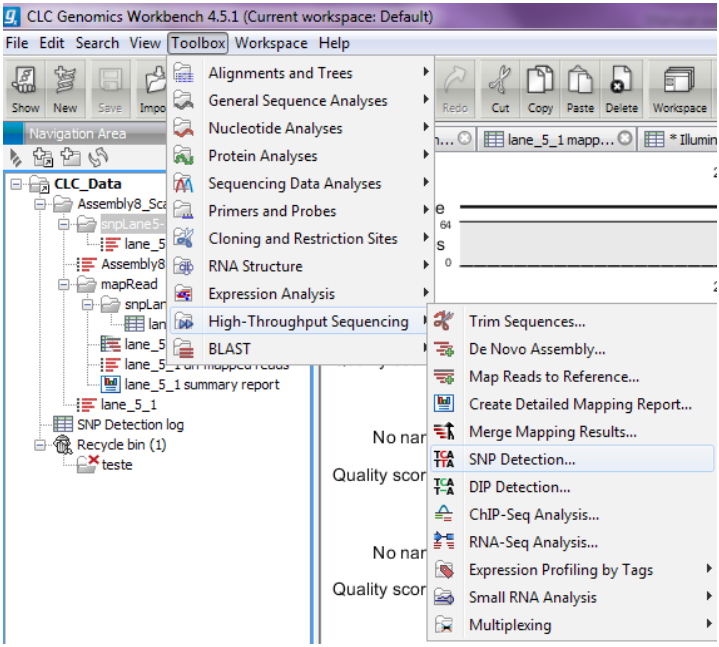

Figura 3. Como rodar o programa CLC para obter SNPs.

Para ver um resumo do que foi obtido, basta acessar a opção "ToolBox" na barra de ferramentas, depois a opção "Resequence Analysis" e em seguida a opção "Quality-based Variant Detection", e a saída é tal como descrito na Figura 4, a seguir.

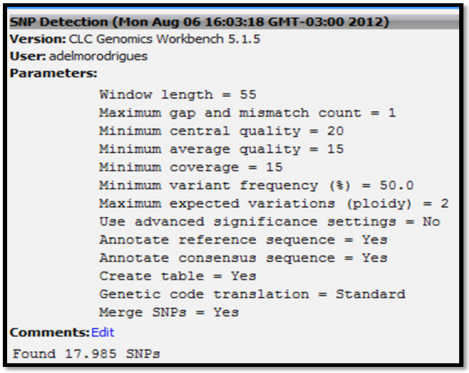

Figura 4. Resumo dos resultados obtidos sobre SNPs.

Mais informações sobre os parâmetros a serem usados quando da execução do CLC Genomics, visto que são muitos, podem ser vistas em CLC Genomics Workbench (2013). Neste documento estão citados os parâmetros e também os exemplos de como configurar os mesmos antes de executar o programa, o que facilita o entendimento.

As saídas dos SNPs podem ser salvas em arquivo do tipo texto (.txt) e também podem ser vistas graficamente. A Figura 5 ilustra a saída do sistema em sua interface gráfica.

Rows: 4.405 SNP Detection Table Filter:

Mapping	Reference ...	Consensus...	Variation T...	Length	Reference	Variants	Allele Varia...	Frequencies	Counts	Coverage
chr01|1310...	247043	245861	SNP	1	A	1	C	91.7	55	60
chr01|1310...	247047	245865	SNP	1	T	1	C	77.8	21	27
chr01|1310...	247066	245885	SNP	1	T	1	C	90.7	39	43
chr01|1310...	247071	245892	SNP	1	T	1	C	95.3	82	86
chr01|1310...	247104	245932	SNP	1	A	1	G	96.5	110	114
chr01|1310...	247115	245944	SNP	1	T	1	G	95.5	84	88
chr01|1310...	842604	782924	SNP	1	C	1	A	96.4	27	28
chr01|1310...	922242	857588	SNP	1	G	1	C	52.0	13	25
chr01|1310...	922304	857650	SNP	1	G	2	C/G	50.0/50.0	15/15	30
chr01|1310...	922822	858225	SNP	1	T	1	G	61.3	19	31
chr01|1310...	923730	859150	SNP	1	T	1	A	56.7	17	30
chr01|1310...	957134	891584	SNP	1	A	1	G	64.0	16	25
chr01|1310...	1476328	1384050	SNP	1	T	1	C	55.6	20	36
chr01|1310...	1961827	1851211	SNP	1	C	1	T	80.0	24	30
chr01|1310...	1961854	1851238	SNP	1	T	1	C	63.0	17	27
chr01|1310...	1961874	1851260	SNP	1	A	1	G	57.1	16	28
chr01|1310...	2057791	1918612	SNP	1	A	1	C	100.0	25	25
chr01|1310...	2803061	2597356	SNP	1	C	1	T	82.6	57	69
chr01|1310...	2803067	2597362	SNP	1	A	1	T	87.0	60	69
chr01|1310...	2803092	2597387	SNP	1	A	1	G	90.4	66	73
chr01|1310...	2803170	2597471	SNP	1	A	1	C	88.2	45	51
chr01|1310...	2898393	2687886	SNP	1	G	1	A	52.0	13	25
chr01|1310...	3075932	2840050	SNP	1	C	1	T	100.0	80	80
chr01|1310...	3075934	2840052	SNP	1	T	1	A	98.8	82	83
chr01|1310...	3075970	2840088	SNP	1	T	1	C	100.0	120	120
chr01|1310...	3076025	2840148	SNP	1	A	1	G	100.0	26	26
chr01|1310...	4164510	3830190	SNP	1	A	1	G	80.5	33	41
chr01|1310...	4164531	3830211	SNP	1	C	1	T	75.0	27	36
chr01|1310...	4416745	4077596	SNP	1	A	1	G	100.0	31	31
chr01|1310...	5573845	5203606	SNP	1	G	1	C	52.0	13	25
chr01|1310...	5922995	5547622	SNP	1	A	1	C	75.0	24	32
chr01|1310...	5923043	5547675	SNP	1	A	1	G	55.6	15	27
chr01|1310...	5923044	5547676	SNP	1	A	1	G	55.6	15	27
chr01|1310...	6315111	5926637	SNP	1	A	1	G	96.0	24	25
chr01|1310...	6355262	5963533	SNP	1	T	1	C	100.0	25	25
chr01|1310...	6869319	6466834	SNP	1	T	1	G	88.0	22	25
chr01|1310...	6869320	6466836	SNP	1	T	1	C	88.0	22	25
chr01|1310...	6869378	6466905	SNP	1	A	1	G	78.7	37	47
chr01|1310...	6869481	6467019	SNP	1	T	1	C	82.1	32	39

Figura 5. Saída relativa aos SNPs obtidos para a variedade Caiapó.

Para a obtenção dos SNPs, portanto, é contrastado o genoma da cultivar Caiapó contra o genoma de referência, Nipponbare, utilizando o alinhamento global com uma similaridade mínima de 90%. Os arquivos

do tipo fastq gerados pelo ressequenciamento da variedade Caiapó tiveram 501.700 sequências reverse e 3.179.445 de sequências forward que não foram lidas pelo CLC, quando da execução do programa. Utilizando um tamanho de janela de 55 (window length), cobertura mínima de 20x e um número máximo de erros e gaps igual a 2, foi encontrado um total de 8.521 SNPs verdadeiros. Quando esta cobertura é aumentada para 25x, o número de SNPs encontrados cai para 4.405. Deste mesmo modo, quando o tamanho da janela é diminuído para 25, são encontrados 16.428 e 8.375 SNPs para as coberturas de 20x e 25x, respectivamente.

A Tabela 2 ilustra a quantidade de SNPs encontrada, conforme a cobertura e o tamanho da janela.

Tabela 2. Quantidade de SNPs de Caiapó obtida conforme parâmetros.

Tamanho da janela	Cobertura	Quantidade de SNPs
25	20	16.341
25	25	6.032
55	15	17.986
55	20	8.524
55	25	2.727

Para a Tabela 4, que será mostrada mais adiante, a quantidade de SNPs considerada será a com janela = 55 e cobertura = 25x.

Muitas análises podem ser feitas com os resultados textuais ou gráficos nesta ferramenta. Na verdade, ela engloba uma série de funcionalidades e vai muito além de apenas detectar SNPs. Mais informações sobre CLC Genomics podem ser vistas na página http://www.clcbio.com/desktop-applications/all-features.

Tempo de execução do software

Os tempos de execução dos programas citados acima, com seus comandos executados por linha de comando ou interface gráfica, estão resumidos na Tabela 3.

Tabela 3. Tempo de execução dos softwares para alinhamento e obtenção de SNPs.

Software	Tempo (s)	Tempo(HH:MM:SS)
Bowtie2 + SAMtools	10.302	02h51min42s
BWA + SAMtools	6.867	01h54min27s
CLC	82.939	23h2min19s
Mosaik + GigaBayes	64.843	18h0min43s
Panati	6.040	1h40min40s
Bowtie2 + VarScan	27.183	7h33min3s
BWA + SAMtools + GATK	29.952	8h19min12s

De acordo com a Tabela 3, observa-se que o Mosaik é o que mais demora para ser executado, e o software Panati é o mais rápido. Porém, mais vale a qualidade dos SNPs obtidos do que o tempo de execução. No caso, CLC demorou praticamente um dia, mas seus resultados contêm uma série de dados para avaliar se um SNP é falso-positivo ou não. Serve até de referência para ver se os demais softwares estão com parâmetros corretos ou se é necessário restringir mais a parte de "descoberta de SNPs". O mesmo pode ser dito do GATK, que aproveita os resultados obtidos a partir do kit SAMtools e faz uma nova análise dos resultados para evitar falso-positivos. A quantidade de SNPs obtida para a variedade Caiapó, em cada um dos seus 12 cromossomos, foi a seguinte:

Tabela 4. Resultados obtidos com os softwares usados para obter SNPs de Nipponbare.

Cromos-somo	BWA + SAMtools	Bowtie2 + SAMtools	Mosaik + GigaBayes	Panati	VarScan	BWA + SAMtools + GATK	CLC
1	101.314	103.878	68	7.234	237	295	263
2	75.340	78.718	37	5.852	71	155	210
3	73.378	79.025	97	5.656	59	149	310
4	88.677	81.794	97	5.870	229	140	316
5	77.475	73.820	32	5.000	65	83	119
6	81.584	77.913	75	5.339	260	100	183
7	62.423	62.146	78	4.636	62	72	161
8	83.482	75.925	25	5.017	42	68	140
9	50.765	50.283	29	3.739	432	23	202
10	79.897	70.469	54	4.677	146	50	144
11	69.786	64.558	79	5.093	32	77	181
12	66.312	60.396	95	4.741	259	63	414

De acordo com a Tabela 4, foi observado que VarScan, Mosaik, GATK e CLC fizeram uma filtragem de SNPs e retiraram aqueles que eram falso-positivos, após uma análise estatística feita por estes programas. Basta comparar a coluna com rótulo "BWA + SAMtools" com a coluna com rótulo "BWA + SAMtools + GATK", por exemplo, para ver a grande quantidade de SNPs falso-positivos. Assim, ao se usar algum Alligner com algum Caller, para o caso de softwares freeware, deve-se usar também GATK ou VarScan para melhorar a qualidade dos resultados retirando-se SNPs que são falso-positivos. Para o caso do CLC, software proprietário, deve-se configurar o software com os parâmetros ilustrados na Figura 4.

Vale a pena mencionar que os softwares citados no parágrafo anterior não são suficientes para definir se os SNPs são 100% confiáveis, porém dão uma margem grande de segurança quanto ao resultado obtido, isto é, SNPs válidos. O trabalho de (Grattapaglia et al. (2011) sugere o uso de BWA, GATK (2013) e SAMtools (2013), porém sugerem mais uma série de medidas para se avaliar a qualidade dos SNPs usando filtros, que podem ser configurados em VarScan, GigaBayes, GATK e CLC. Esses filtros, resumidamente, são:

- SNPs bi-alélicos com mais de cinco *reads* na posição
- Minor Allele Frequency (MAF) ≥ 0.2;
- 100 pares de base, pelo menos, ao lado do SNP (flanqueadores à esquerda e à direita) e nestes não deve haver tandem repeats (cadeias repetitivas) e nestes 100 pares de base, se houver outro SNP, este deverá estar a pelo menos 20 pb de distância de um outro SNP.

Existem mais filtros, que podem ser vistos em Grattapaglia et al. (2011). Estes podem ser configurados para serem usados com GigaBayes, VarScan e GATK (ver o comando que usa a opção *-T VariantFiltration*, como exemplo, na secção GATK), através de parâmetros, ou ainda podem ser feitos mais programas para

implementarem os filtros faltantes. No software CLC, é possível também configurar todos ou pelo menos a maioria dos filtros. Caso não seja viável ou possível configurar algum filtro, então deverá ser feito algum programa para criar o filtro e rodar o mesmo a partir dos resultados obtidos pelo CLC, tal como seria feito para o GATK e VarScan ou ainda o Mosaik. Em suma, se o filtro existe, basta usar, se não existe, então é necessário um programa para criar o filtro e este usa algum arquivo gerado como saída de um dos softwares citados neste trabalho (CLC, Mosaik, etc.).

Resumo sobre cada software

GATK, GigaBayes e VarScan fazem análises estatísticas diversas, tanto quanto a qualidade do alinhamento obtido quanto dos valores de p-value obtidos para cada SNP. Por um lado, garantem que os SNPs obtidos são confiáveis, porém, por outro lado, existe a possibilidade de SNPs verdadeiros serem rejeitados. Em primeira análise, é melhor perder um SNP verdadeiro do que contabilizar um SNP falso. Os valores configurados para se obter SNPs com qualidade e confiabilidade podem ser relaxados para se obter mais SNPs, caso o número obtido seja muito baixo e se saiba que as variedades contrastantes (referência e não referência) tenham muitas diferenças fenotípicas, para as mesmas condições de estudo.

O software CLC Genomics demora um pouco mais para ser executado, porém oferece resultados de qualidade e também resultados intermediários diversos como, por exemplo, a cobertura de *reads* considerada (5x, 10x, 25x, etc.), mostra os SNP e posições graficamente e assim facilita ao usuário ver todos os passos intermediários para que se possa ter segurança de como todo o processamento foi feito.

O software Panati ainda necessita de um amadurecimento e uma nova versão deverá estar disponível na rede futuramente, conforme descrito em Panati (2013). Porém, os resultados devem ser considerados, desde que os parâmetros usados sejam aqueles

referenciados na seção referente ao software Panati, discutido anteriormente.

O software Mosaik teve um tempo de execução longo e como resultado teve uma quantidade muito baixa de SNPs. Isto é devido aos parâmetros usados nos programas do kit do Mosaik e também aos parâmetros de GigaBayes. O usuário poderá relaxar as restrições para obter uma quantidade maior de SNPs. Por exemplo, alterar os parâmetros PSL, CAL1 e CAL5 de GigaBayes para outros valores e testar se os valores de p-value são aceitáveis (menor que 0,01, por exemplo). Vale a pena mencionar que o tempo de execução de GigaBayes é bem menor do que todos os programas relativos ao Mosaik (em torno de 5% do tempo total).

Conclusões

Este trabalho mostrou alguns dos softwares usados para alinhamento de sequências (Bowtie2, BWA, Mosaik, Panati, CLC) e callers de SNPs (SAMtools, Mosaik + GigaBayes, Panati, VarScan, GATK e CLC). Os softwares Panati e CLC têm programas que fazem o alinhamento e a descoberta de SNPs. Os softwares BWA e Bowtie2 podem rodar em conjunto com SAMtools para uma descoberta inicial de SNPs e os softwares VarScan e GATK são usados para fazer uma filtragem mais rigorosa dos SNPs e impedir que algum falso-positivo seja obtido e venha a atrapalhar as análises.

Para cada software foi mostrado como fazer a execução do mesmo, seus resultados em termos de tempo e quantidade de SNPs obtida e considerações sobre cada execução. Além disso, estes softwares foram comparados com o CLC Genomics, reconhecido no mercado como software de qualidade, para ver o tempo de execução de cada um comparativamente e a quantidade de SNPs obtida, para o caso da variedade de arroz Caiapó.

Referências

BOWTIE2. Disponível em: <http://bowtie-bio.sourceforge.net/index.shtml>. Acesso em: 10 dez. 2013.

BWA. **Burrows-Wheeler Aligner.** Disponível em: <http://bio-bwa.sourceforge.net/>. Acesso em: 20 dez. 2013.

CLCBIO. Disponível em: <http://clcbio.com>. Acesso em: 20 dez. 2013.

CLC Genomics Workbench: user manual. Disponível em: <http://www.clcbio.com/files/usermanuals/CLC_Genomics_Workbench_User_Manual.pdf>. Acesso em: 25 set. 2013.

EAGLEVIEW. Disponível em: < http://bioinformatics.bc.edu/marthlab/wiki/index.php/EagleView>. Acesso em: 18 dez. 2013.

GATK. Disponível em: <http://www.broadinstitute.org/gatk/guide>. Acesso em: 25 set. 2013.

GIGABAYES. Disponível em: <http://bioinformatics.bc.edu/marthlab/wiki/index.php/GigaBayes>. Acesso em: 25 set. 2013.

GRATTAPAGLIA, D.; SILVA JUNIOR, O. B. da; KIRST, M.; LIMA, B. M. de; FARIA, D. A.; PAPPAS, G. J. High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. **BMC Plant Biology**, Cambridge, v. 11, p. 65, Apr. 2011. doi: 10.1186/1471-2229-11-65.

KOBOLDT, D. C.; CHEN, K.; WYLIE, T.; LARSON, D. E.; MCLELLAN, M. D.; MARDIS, E. R.; WEINSTOCK, G. M.; WILSON, R. K.; DING, L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. **Bioinformatics**, Oxford, v. 25, n. 17, p. 2283-2285, Sept. 2009.

MOSAIK. Disponível em: <https://github.com/wanpinglee/MOSAIK/wiki/QuickStart>. Acesso em: 20 dez. 2013.

NCBI. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acesso em: 25 set. 2013.

PANATI. Disponível em: <http://panati.sourceforge.net/>. Acesso em: 25 set. 2013.

SAMTOOLS. Disponível em: <http://samtools.sourceforge.net>. Acesso em: 25 set. 2013.

TRAPNELL, C.; ROBERTS, A.; GOFF, L.; PERTEA, G.; KIM, D.; KELLEY, D. R.; PIMENTEL, H.; SALZBERG, S. L.; RINN, J. L.; PACHTER, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**, London, v. 7, n. 3, p. 562-578, Mar. 2012.

VARSCAN. Disponível em: <http://varscan.sourceforge.net/>. Acesso em: 20 dez. 2013.