Caderno 3/6: Identificação de Variantes Genéticas (Variant Call) de SARS-CoV-2 com arquivos FASTA

Gepoliano Chaves, Ph. D.

Outubro, 2020

Contents

Para plotagem PPT ou PDF, incluir os seguintes comandos:

output: powerpoint_presentation

ou

output: pdf_document: toc: yes toc_depth: '5' html_document: df_print: paged toc: yes number_sections: no toc_depth: 5 toc_float: yes, above

1) Introdução

No Caderno 2, ilustramos uma das aplicações do conhecimento ensinado neste curso: o estabelecimento da Associação Biológica entre genótipo (os polimorfismos de DNA) e fenótipo usando-se um software estatístico (PLINK). Antes disto, no Caderno 1, tivemos uma introdução à Programação, com instalação de livrarias em R e utilização de programas para visualização de arquivos de texto. Aqui, vamos expandir a noção de arquivo de texto, para a noção de um arquivo de texto em que podemos também armazenar uma sequências biológicas. O primeiro arquivo de armazenamento de sequêcia biológica que estudaremos será o arquivo FASTA. O arquivo FASTA pode armazenar uma sequência de DNA em formato de texto.

No presente caderno, aprofundamos a análise de sequências com o início da Identificação de Variantes genéticas. Aqui, devemos tratar as variantes genéticas como polimorfismos de DNA. Um polimorfismo de DNA nada mais é que uma mutação ou variante, que é diferente de uma sequência FASTA, usada como referência. Queremos definir o protocolo computacional para Associação Biológica ou Genética como uma pipeline de identificação de variantes genéticas. Em inglês, a lingua da literatura científica, este protocolo é chamado de *Variant Call*. No presente Caderno, faremos uma *Variant Call* usando arquivos FASTA de contendo o genoma de SARS-CoV-2, ou a sequência biológica do genoma deste vírus, em um arquivo de texto, especificamente chamado FASTA (Figura 1). Identificamos variantes de SARS-CoV-2 utilizando as ferramentas samtools e bcftools para extrair polimorfismos genéticos dos arquivos FASTA. Neste Caderno, os arquivos FASTA devem ser armazenados localmente, o que significa que devem ser bbaixados no computador do pesquisador. As variantes genéticas, também chamdas SNPs (*Single Nucleotide Polymorphisms*), são determinadas pela comparação de um arquivo FASTA usado como referência, à sequência isolada em outras regiões do planeta.

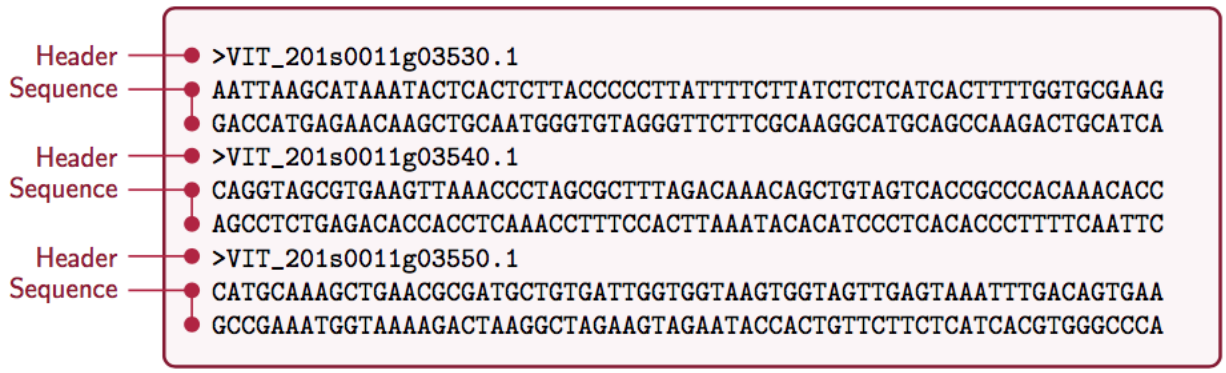

Figure 1: Formatação do arquivo FASTA: um arquivo de texto contendo uma sequência biológica.

Ao comparar a sequência FASTA de cada região do planeta a uma sequência FASTA referência, por exemplo a primeira sequência FASTA correspondente ao primeiro paciente que teve SARS-CoV-2 isolado e sequenciado na China, podemos anotar cada posição em que haja um nucleotídeo diferente da sequência FASTA de referência, obtida do primeiro paciente chinês. Dizemos então, que nas demais regiões, há mutações do vírus, ou, de forma mais técnica, polimorfismos do material genético viral. Arquivos FASTA provenientes de virtualmente todas as regiões do planeta, podem ser obtidos da base de dados alemã GISAID (Global initiative on sharing all influenza data). Informações sobre a base de dados GISAID podem ser obtidas no site da mesma:

https://www.gisaid.org

No Caderno de Biologia Computacional Número 6, aprenderemos como efetuar a Identificação de Variantes genéticas utilizando arquivos FASTQ, ao invés de arquivos FASTA. Arquivos FASTQ podem ser analisados diretamente de uma base de dados chamada *Gene Expression Omnibus* (GEO), e se usada a pipeline apropriada de identificação de variantes genéticas, não é necessário o armazenamento local de arquivos de sequenciamento.

2) Submissão de *script* a servidor computacional.

Minhas análises envolvendo os genomas humano e viral, foram em sua imensa maior parte feitas utilizando-se o servidor computacional da Universidade da California em Santa Cruz (UCSC), onde primeiro sequenciou-se o Genoma Humano em 2001. A parte abaixo, codificada em script bash, é feita usando o Sistema Operacinal Linux. O codigo representa um tipo de "cabecalho" que deve ser incluído para submissão de scripts usando-se o sistema Linux a servidores computacionais.

```
#!/bin/bash
#SBATCH --partition=128x24
##SBATCH --job-name=Variants_BWA # Job name
##SBATCH --mail-type=ALL              # Mail events (NONE, BEGIN, END, FAIL, ALL)
##SBATCH --mail-user=gchaves@ucsc.edu  # Where to send mail
##SBATCH --nodes=1                    # Use one node
##SBATCH --ntasks=1                   # Run a single task
##SBATCH --cpus-per-task=4            # Number of CPU cores per task
##SBATCH --output=Variants_BWA     # Standard output and error log
#
# module load gcc/5.4.0
source ~/.bashrc
```

3) Definição das pastas de trabalho e do Genoma de Referência

Nesta etapa, devemos definir uma pasta onde serao armazenados os arquivos VCF resultantes da pipeline. Como visto no Caderno 2, uma pipeline é o conjunto de todos os passos, em sequência, utilizados em uma analise computacional. Esta pasta, chamada ProjectDirectory abaixo, pode ser criada utilizando-se o comando *mkdir*, o qual cria a pasta dentro do diretorio onde o usuario encontra-se no presente momento (conferir o comando *cd*, *change directory*).

Tambem deve ser criada, uma pasta onde localiza-se o Genoma de Referencia, chamada Reference. Nesta pasta, deve estar armazanada a sequencia genomica a ser analisada, no caso, a sequencia FASTA de SARS-CoV-2. Esta sequencia fasta pode ser baixada a partir do Genome Browser da Universidade da California em Santa Cruz.

Os demais arquivos FASTA, contendo as mutacoes ou polimorfismos de RNA/DNA a serem identificados, devem ser baixados e salvos na pasta FastaDirectory. A abordagem desta pipeline, facilita a organizacao computacional ao criar uma pasta para o projeto, onde sao salvos todos os arquivos resultantes da pipeline.

Finalmente, o comando *vim* cria uma lista, salva em formato *txt*, a qual contem todos os arquivos FASTA a serem analisados pela pipeline. Como afirmado acima, os arquivos FASTA devem ser salvos na pasta FastaDirectory. Esta abordagem utiliza uma *for loop*, a qual facilita a analise de varias sequencias FASTA menor intervalo de tempo possivel.

```
ProjectDirectory=~/Desktop/Gepoliano/UFSB/COVID_BWA_Variant_Call
mkdir $ProjectDirectory
#vim ~/$ProjectDirectory/COVID_List_Region.txt
```

4) Alinhamento e *Variant Call*

A abordagem deste algoritmo é a criação das pastas para armazenar os resultados da pipeline (VCFs e demais arquivos) para cada arquivo FASTA usando uma "for loop". Desta forma, usamos a *for loop* para efetuar a pipeline para quantidades grandes de arquivos de sequenciamento (FASTA ou FASTQ).

```
source ~/.bash_profile
conda install -c bioconda bwa
bwa --help
```

Este post ajudou a resolver o problema de instalacao de samtools:

https://github.com/samtools/samtools/issues/974

Isto resolveu:

```
conda uninstall samtools
conda update --all
conda install samtools
```

Isto resolveu no caso de bcftools:

```
conda install -c bioconda/label/cf201901 bcftools
head ~/Desktop/Gepoliano/Corona\ Virus/genome_assemblies/ncbi-genomes-2020-03-21/GCF_009858895.2_ASM985
```

4.1) Alinhamento

Este alinhamento deve ser executado em uma "for loop", para que o máximo número de amostras seja processado ao mesmo tempo.

```
## Tentei executar bwa através da definição da variável abaixo, mas isso não funcionou:
ComandoBwa=~/anaconda3/bin/bwa

REFERENCE=~/Desktop/Gepoliano/CoronaVirus/genome_assemblies/ncbi_genomes_2020_03_21/GCF_009858895_2_ASM
ProjectDirectory=~/Desktop/Gepoliano/UFSB/COVID_BWA_Variant_Call
Regiao=Australia_GISAID

## Criar pasta para salvar outputs da pipeline, por região
mkdir $ProjectDirectory/$Regiao

## Para o alinhamento, se a linha começando com @RG, abaixo, não for incluída,
## o erro "the read group line is not started with @RG" é produzido:
for fasta_file in $(cat $ProjectDirectory/COVID_List_Region.txt); do
  mkdir $ProjectDirectory/$Regiao/$fasta_file

  ~/anaconda3/bin/bwa mem -M -R \
  '@RG\tID:SampleCorona\tLB:sample_1\tPL:ILLUMINA\tPM:HISEQ\tSM:SampleCorona' \
  $REFERENCE \
  ~/COVID_BWA_Variant_Call/$Regiao/Australia_EPI_ISL_416412.fasta  > \
  $ProjectDirectory/$Regiao/$fasta_file/$fasta_file".sam"

  cd $ProjectDirectory/$Regiao/$fasta_file/

  ## Samtools conversão de formatos SAM para BAM
  samtools view -S -b $fasta_file".sam" > $fasta_file".bam"

  ## Samtools usa arquivo FASTA referência para detectar "empacotamento" das sequências
  samtools mpileup -g -f $REFERENCE $fasta_file".bam" > $fasta_file".bcf"

  ## Bcftools extrai colunas especificas
  ~/anaconda3/bin/bcftools query -f '%CHROM %POS %REF %ALT\n' $fasta_file".bcf" | head -50000

  ## Bcftools extrai SNPs
  ~/anaconda3/bin/bcftools view -v snps $fasta_file".bcf"  > $fasta_file"_snps.vcf"

  ## Bcftools extrai indels
  ~/anaconda3/bin/bcftools view -v indels $fasta_file".bcf" > $fasta_file"_indels.vcf"

done
```