Using WhopGenome

Ulrich Wittelsbürger

August 26, 2014

1 Handling VCF files

1.1 Opening a VCF file

Suppose we want to have a look at some of the 1000 Genome Project data for chromosome Y, available on the EBI ftp at ftp:ftp.1000genomes.ebi.ac.ukvol1ftpphase1analysis_resultsintegrat
.

```
> library(WhopGenome) # load WhopGenome first
>
> # Make things a bit easier by having a variable instead of a lengthy filename
> yvcffile <- "ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated
>
> yvcfh <- vcf_open(yvcffile)
[get_local_version] downloading the index file...
(II) VCF version is 4.1
vcff::open : file opened, contains 526 samples
>
```

As a demonstration for vcf_reopen, let's quit R and restart it:

```
> q()
Save workspace image? [y/n/c]: y
user@computer:~\$ R

R version 3.0.1 (2013-05-16) -- "Good Sport"
Copyright (C) 2013 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

[...]
[Previously saved workspace restored.]

> library(WhopGenome)
> yvcfh
<pointer: (nil)>
attr(,"VCF.filename")
[1] "ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets
>
```

As you can see, the filehandle variable *yvcfh* is present but the data it refers to isn't anymore, because it was managed by WhopGenome and not R. This is the typical use-case for vcf_reopen (and tabix_reopen, fai_reopen, see below), shown next:

```
> vcf_reopen( yvcfh )
(II) VCF version is 4.1
vcff::open : file opened, contains 526 samples
[1] TRUE
>
```

The VCF file is now usable again. Before we can begin reading data from it, we need to take care of three points:

- select the samples we are interested in (REQUIRED)
- set which region we are interested in (OPTIONAL)
- set up filtering step (OPTIONAL)

1.2 Samples

In order to know which samples we can select, we query their names:

```
> GET SAMPLE NAMES
```

```
> SELECT SAMPLES
```

```
> GET SELECTED SAMPLES
```

Now that we have taken care of that, we can specify a region to get data from. Regions are specified by three components: chromosome, start position, end position. A VCF file can contain any number of chromosomes or contigs, provided that they have unique identifiers. NOTE: the naming scheme of chromosomes can vary: sometimes they are named Chr1, Chr2, ChrX, ChrY and sometimes it is just 1,2,X,Y. WhopGenome expects the correct identifier and will not try to guess.