# RBcf: An VCF API for R.

Pierre Lindenbaum / yokofakun/ Institut du Thorax . Nantes.

April 23, 2020

## 1 Abstract

RBcf uses the Htslib C API for parsing VCF and BCF files. This API was written by a regular user of the htsjdk library who doesn't like R.

## 2 Examples

### 2.1 Open and close a VCF file

Code:

```
# load rbcf
library(rbcf)
# we don't need the index for this file
fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
# dispose the vcf reader
bcf.close(fp)
```

Output:

```
> # load rbcf
> library(rbcf)
> # we don't need the index for this file
> fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
>
```

## 2.2 Print the INFOs in the VCF header

Code:

```
# load rbcf
library(rbcf)
# we don't need the index for this file
fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
info <- bcf.infos(fp)
# dispose the vcf reader
bcf.close(fp)
# print the table
info
```

Output:

```
> # load rbcf
> library(rbcf)
> # we don't need the index for this file
> fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
> info <- bcf.infos(fp)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
> # print the table
> info
         ID Number    Type
INDEL INDEL      0    Flag
IDV     IDV      1 Integer
IMF     IMF      1   Float
DP       DP      1 Integer
VDB     VDB      1   Float
RPB     RPB      1   Float
MQB     MQB      1   Float
BQB     BQB      1   Float
MQSB   MQSB      1   Float
SGB     SGB      1   Float
MQOF   MQOF      1   Float
ICB     ICB      1   Float
HOB     HOB      1   Float
AC       AC      A Integer
AN       AN      1 Integer
```

```
DP4     DP4      4 Integer
MQ       MQ      1 Integer

INDEL                                                          "Indicates that the v
IDV                                                          "Maximum number of reads
IMF                                                        "Maximum fraction of reads
DP
VDB   "Variant Distance Bias for filtering splice-site artefacts in RNA-seq data
RPB                                       "Mann-Whitney U test of Read Position Bias
MQB                                     "Mann-Whitney U test of Mapping Quality Bias
BQB                                       "Mann-Whitney U test of Base Quality Bias
MQSB                         "Mann-Whitney U test of Mapping Quality vs Strand Bias
SGB                                                                          "Segreg
MQ0F                                                       "Fraction of MQ0 reads
ICB                                             "Inbreeding Coefficient Binomial test
HOB                                               "Bias in the number of HOMs number
AC                        "Allele count in genotypes for each ALT allele, in the s
AN                                                         "Total number of alleles
DP4            "Number of high-quality ref-forward , ref-reverse, alt-forward an
MQ                                                                           "Aver
>
```

## 2.3 Print the FORMATs in the VCF header

Code:

```
# load rbcf
library(rbcf)
# we don't need the index for this file
fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
fmts <- bcf.formats(fp)
# dispose the vcf reader
bcf.close(fp)
# print the table
fmts
```

Output:

```
> # load rbcf
> library(rbcf)
> # we don't need the index for this file
> fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
```

```
> fmts <- bcf.formats(fp)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
> # print the table
> fmts
   ID Number    Type                                    Description
PL PL      G Integer "List of Phred-scaled genotype likelihoods"
GT GT      1  String                                     "Genotype"
>
```

## 2.4 Print the FILTERs in the VCF header

Code:

```
# load rbcf
library(rbcf)
# we don't need the index for this file
fp <- bcf.open("../tests/data/gnomad.exomes.r2.0.1.sites.vcf",FALSE)
flt <- bcf.filters(fp)
# dispose the vcf reader
bcf.close(fp)
# print the table
flt
```

Output:

```
> # load rbcf
> library(rbcf)
> # we don't need the index for this file
> fp <- bcf.open("../tests/data/gnomad.exomes.r2.0.1.sites.vcf",FALSE)
> flt <- bcf.filters(fp)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
> # print the table
> flt
                             ID
PASS                       PASS
AC0                         AC0
InbreedingCoeff InbreedingCoeff
LCR                         LCR
```

```
RF                              RF
SEGDUP                      SEGDUP

PASS
AC0             "Allele Count is zero (i.e. no high-confidence genotype (GQ >= 2
InbreedingCoeff
LCR
RF                                                  "Failed random forests filte
SEGDUP
>
```

## 2.5 Print the Samples in the VCF header

Code:

```
# load rbcf
library(rbcf)
# we don't need the index for this file
fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
# print the number of samples
cat(paste("Num. Samples=",bcf.nsamples(fp),".\n"))
# get the name for the 1st sample
cat(paste("First sample is ",bcf.sample1(fp,1),".\n"))

# get the samples
samples <- bcf.samples(fp)
# dispose the vcf reader
bcf.close(fp)
# print the list
samples
```

Output:

```
> # load rbcf
> library(rbcf)
> # we don't need the index for this file
> fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
> # print the number of samples
> cat(paste("Num. Samples=",bcf.nsamples(fp),".\n"))
Num. Samples= 5 .
> # get the name for the 1st sample
> cat(paste("First sample is ",bcf.sample1(fp,1),".\n"))
```

```
First sample is  S1 .
>
> # get the samples
> samples <- bcf.samples(fp)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
> # print the list
> samples
[1] "S1" "S2" "S3" "S4" "S5"
>
```

## 2.6 Print the Dictionary in the VCF header

Code:

```
# load rbcf
library(rbcf)
# we don't need the index for this file
fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
dict <- bcf.dictionary(fp)
# dispose the vcf reader
bcf.close(fp)
# print the table
dict
```

Output:

```
> # load rbcf
> library(rbcf)
> # we don't need the index for this file
> fp <- bcf.open("../tests/data/rotavirus_rf.01.vcf",FALSE)
> dict <- bcf.dictionary(fp)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
> # print the table
> dict
     chrom size
RF01  RF01 3302
RF02  RF02 2687
RF03  RF03 2592
```

```
RF04  RF04 2362
RF05  RF05 1579
RF06  RF06 1356
RF07  RF07 1074
RF08  RF08 1059
RF09  RF09 1062
RF10  RF10  751
RF11  RF11  666
>
```

## 2.7 Print the Indexed Chromosomes

Code:

```
# load rbcf
library(rbcf)
# Open the indexed VCF
fp <- bcf.open("../tests/data/rotavirus_rf.02.vcf.gz")
# get the indexed contigs
contigs <- bcf.contigs(fp)
# dispose the vcf reader
bcf.close(fp)
# print the table
contigs
```

Output:

```
> # load rbcf
> library(rbcf)
> # Open the indexed VCF
> fp <- bcf.open("../tests/data/rotavirus_rf.02.vcf.gz")
> # get the indexed contigs
> contigs <- bcf.contigs(fp)
> # dispose the vcf reader
> bcf.close(fp)
[1] TRUE
> # print the table
> contigs
 [1] "RF01" "RF02" "RF03" "RF04" "RF05" "RF06" "RF07" "RF08" "RF09" "RF10"
[11] "RF11"
>
```