snvmut: an R package for SNV extraction

Chiara Albertini

```
library(snvmut)
```

Package introduction

The package snvmut is a Bioconductor compliant R package that allows the user to search for SNV (Single Nucleotide Variant) inside of a .vcf file, to store the mutations in a vector and to graphically plot the mutation counts.

The package has three main functions.

snv_extraction This function returns a vector of SNVs and requires three parameters:

- a VCF object
- the reference genome as a a BSgenome object
- the “context_lenght” peramter, which is a numerical value that corresponds to the final window of nucleotides (mutation included) that will be stored inside the vector. The mutation is in the format [REF>ALT] and in the window are included also the upstream and downstream nucleotides as defined by the “context_lenght” parameter.

The function is optimized to take into account and solve the redundancy caused by the fact that, for example, C[G>A]A is the same as “T[C>T]G” on the reverse strand. This is solved by having all mutations report C or T as the REF base.

snv_count This function takes as input the vector obtained with the snv_extraction function and returns a data frame with three columns:

- “SNV” reports the mutations found in the vector
- “Count” reports how many time the mutation was found in the vector
- “Percentage” report the percentage of every mutation on the total number of mutations of the vector

snv_graphics This function takes as input the data frame obtained with the snv_count function and returns two ggplot2 plots:

- a bar plot
- a pie chart

The plots allow for a quick and easy visualization of the mutation types and proportion over the total number of mutations.

Function usage

Loading VCF files is extremely quick when using the VariantAnnotation package, as shown here. The reference genome can also be loaded using the BSgenome packaging and by specifying the preferred genome.

```
library(BSgenome.Hsapiens.UCSC.hg38)
library(VariantAnnotation)
#> Warning: package 'VariantAnnotation' was built under R version 4.3.2
#> Warning: package 'SummarizedExperiment' was built under R version 4.3.2
library(Biostrings)
library(GenomicRanges)
library(snvmut)

genome_ref <- BSgenome.Hsapiens.UCSC.hg38
path_to_vcf <- system.file("extdata", "chr22.vcf.gz", package="VariantAnnotation")
vcf <- readVcf(path_to_vcf, "hg38")[1:55]

vector_snv <- snv_extraction(vcf, 5, genome_ref)
vector_snv
#>  [1] "GG[T>C]GA" "CC[C>T]CT" "CG[C>T]GG" "CT[C>T]CC" "CC[C>T]TC" "CC[C>T]CG" "GG[T>C]CG" "CC[C>T]GC"
#> [10] "CA[T>C]AT" "TT[C>T]GG" "GT[C>T]CG" "CG[T>C]CC" "CA[C>T]TC" "AG[C>T]GA" "AC[C>A]TC" "GA[C>T]GC"
#> [19] "GG[C>T]GT" "GG[C>T]TC" "CC[C>T]GG" "TC[C>T]GG" "GT[T>G]TA" "CA[C>T]AG" "GC[T>C]AC" "GG[C>T]CA"
#> [28] "CT[C>T]GC" "CT[C>T]CA" "CT[T>C]TG" "CA[C>T]GC" "GC[C>T]AA" "GC[C>T]CT" "GA[C>A]AG" "GG[C>T]GG"
#> [37] "GT[C>G]AG" "TT[C>T]TG" "GT[C>G]AG" "GT[T>C]TG" "AC[T>C]CA" "AT[C>T]GG" "GG[C>T]GG" "GC[C>T]CA"
#> [46] "GG[T>C]CC" "TG[C>T]GG" "TG[C>T]TG" "CC[T>A]AG" "CA[C>T]TG" "AA[C>T]GT" "TT[T>C]GC" "CT[C>T]GG"
```

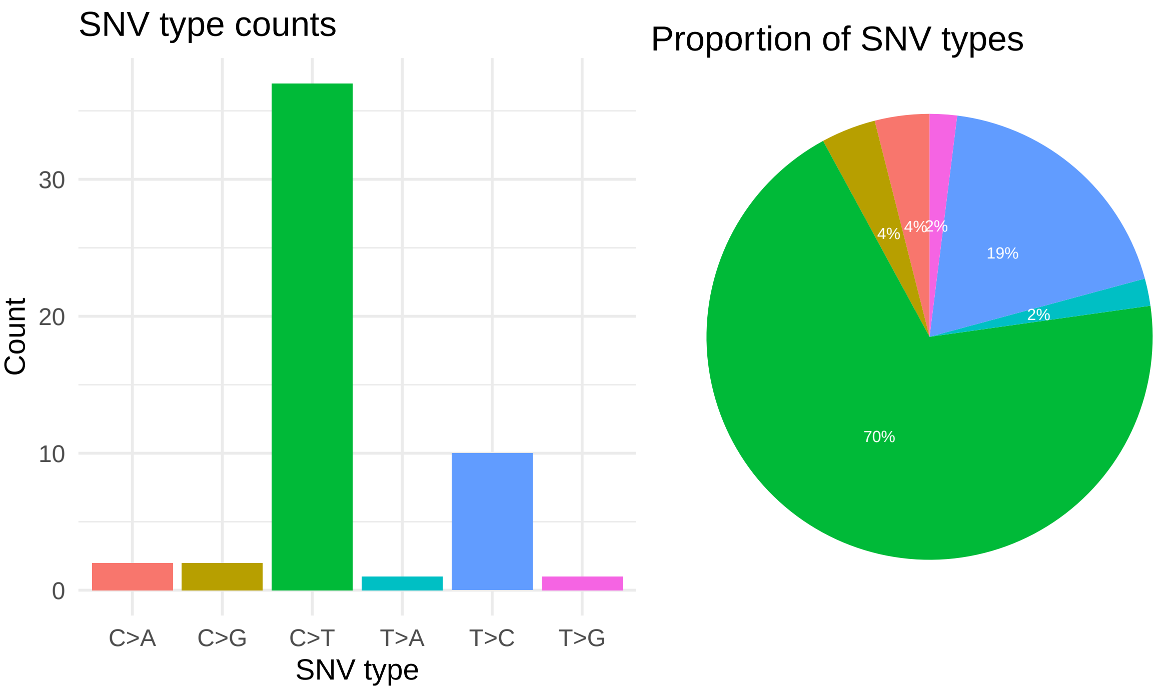
The obtained vector can then be processed using the snv_count function to obtain a data frame with the mutation counts and percentages.

```
snv_table <- snv_count(vector_snv)
snv_table
#>   SNV Count Percentage
#> 1 C>A     2          4
#> 2 C>G     2          4
#> 3 C>T    37         70
#> 4 T>A     1          2
#> 5 T>C    10         19
#> 6 T>G     1          2
```

The obtained data frame can then be processed using the snv_graphics function that assigns a color to each mutation type and plots a bar plot (displaying the mutation counts) and a pie chart (useful for a quick proportion analysis).

```
library(ggplot2)
library(patchwork)

snv_plots <- snv_graphics(snv_table)
snv_plots
```



```
sessionInfo()
#> R version 4.3.1 (2023-06-16)
#> Platform: aarch64-apple-darwin20 (64-bit)
#> Running under: macOS Sonoma 14.1.2
#>
#> Matrix products: default
#> BLAS:   /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versi
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
#>
#> locale:
#> [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> time zone: Europe/Rome
#> tzcode source: internal
#>
#> attached base packages:
#> [1] stats4    stats     graphics  grDevices utils     datasets  methods   base
#>
#> other attached packages:
#>  [1] patchwork_1.1.3                   ggplot2_3.4.4                     VariantAnnotation_1.48.1
#>  [4] Rsamtools_2.18.0                  SummarizedExperiment_1.32.0       Biobase_2.62.0
#>  [7] MatrixGenerics_1.14.0             matrixStats_1.2.0                 BSgenome.Hsapiens.UCSC.hg38
#> [10] BSgenome_1.70.1                   rtracklayer_1.62.0                BiocIO_1.12.0
#> [13] GenomicRanges_1.54.1              snvmut_0.99.1                     Biostrings_2.70.1
#> [16] GenomeInfoDb_1.38.5               XVector_0.42.0                    IRanges_2.36.0
#> [19] S4Vectors_0.40.2                  BiocGenerics_0.48.1
#> 
#> loaded via a namespace (and not attached):
#>  [1] tidyselect_1.2.0         farver_2.1.1             dplyr_1.1.4              blob_1.2.4
#>  [5] filelock_1.0.3           bitops_1.0-7             fastmap_1.1.1            RCurl_1.98-1.13
#>  [9] BiocFileCache_2.10.1     GenomicAlignments_1.38.0 XML_3.99-0.16            digest_0.6.33
#> [13] lifecycle_1.0.4          KEGGREST_1.42.0          RSQLite_2.3.4            magrittr_2.0.3
#> [17] compiler_4.3.1           rlang_1.1.2              progress_1.2.3           tools_4.3.1
#> [21] utf8_1.2.4               yaml_2.3.8               knitr_1.45               labeling_0.4.3
#> [25] prettyunits_1.2.0        S4Arrays_1.2.0           bit_4.0.5                curl_5.2.0
#> [29] DelayedArray_0.28.0      xml2_1.3.6               abind_1.4-5              BiocParallel_1.36.0
#> [33] withr_2.5.2              grid_4.3.1               fansi_1.0.6              colorspace_2.1-0
#> [37] scales_1.3.0             biomaRt_2.58.0           cli_3.6.2                rmarkdown_2.25
#> [41] crayon_1.5.2             generics_0.1.3           rstudioapi_0.15.0        httr_1.4.7
#> [45] rjson_0.2.21             DBI_1.2.0                cachem_1.0.8             stringr_1.5.1
#> [49] zlibbioc_1.48.0          parallel_4.3.1           AnnotationDbi_1.64.1     restfulr_0.0.15
#> [53] vctrs_0.6.5              Matrix_1.6-4             hms_1.1.3                bit64_4.0.5
#> [57] GenomicFeatures_1.54.1   glue_1.6.2               codetools_0.2-19         stringi_1.8.3
#> [61] gtable_0.3.4             munsell_0.5.0            tibble_3.2.1             pillar_1.9.0
#> [65] rappdirs_0.3.3           htmltools_0.5.7          GenomeInfoDbData_1.2.11  R6_2.5.1
#> [69] dbplyr_2.4.0             evaluate_0.23            lattice_0.22-5           highr_0.10
#> [73] png_0.1-8                memoise_2.0.1            SparseArray_1.2.3        xfun_0.41
#> [77] pkgconfig_2.0.3
```

sessionInfo