Variants Density along DNA Sequence

Y. Pageaud

27 septembre 2016

User Guide:

In this document you will find all needed informations to use the script 'Variants_Density_Along_DNA_Sequence.R'.

This R script allow you to plot the density of variants contained in a VCF file along the DNA sequence you have annotated (Chromosome, Scaffold, Contig, Whole Genome…).

Compatibilities:

Ubuntu 16.04 LTS

R version 3.2.3 (2015-12-10) – "Wooden Christmas-Tree"

VCF File Format VCFv4.2

Prerequisite:

Install R version 3.2.3 or later version.
Optionally you can install RStudio after.
Install R packages 'VariantAnnotation' and 'plyr' from the website Bioconductor: https://www.bioconductor.org/

Package Loading:

```
library(VariantAnnotation)
library(plyr)
```

VCF File Example:

You can test the script on an example of a VCF file ('Example.VCF') in the archive .rar.

Set the Working Directory:

Put your working directory (where your VCF file is saved) between quotation marks:

```
setwd("/home/user/your_folder/")
```

Loads your VCF File:

```
vcf <- readVcf("yourfile.vcf","species")
```

DNA Sequence Variants Extraction:

```
Data_Frame_ranges<-as.data.frame(ranges(vcf))
sequence<-Data_Frame_ranges[grep("^sequence_name",Data_Frame_ranges$names),]
```

Variants Density Histograms Along DNA Sequence:

Function hist() parameters:

- Plot title: main = 'string'
- Start position: start_position = int;
- End position: end_position = int
- Limits of the Y axis: ylim = c(float/int,float/int). Useful for thresholding the minimum density of variants.
- Size of the sliding window: change the number of breakpoints of your histogram: histogram_breakpoints = int ; to increase its size choose a smaller value, to decrease its size choose a higher value. the window size depends on the length of the selected region of your sequence, and on the number of breakpoints chosen.

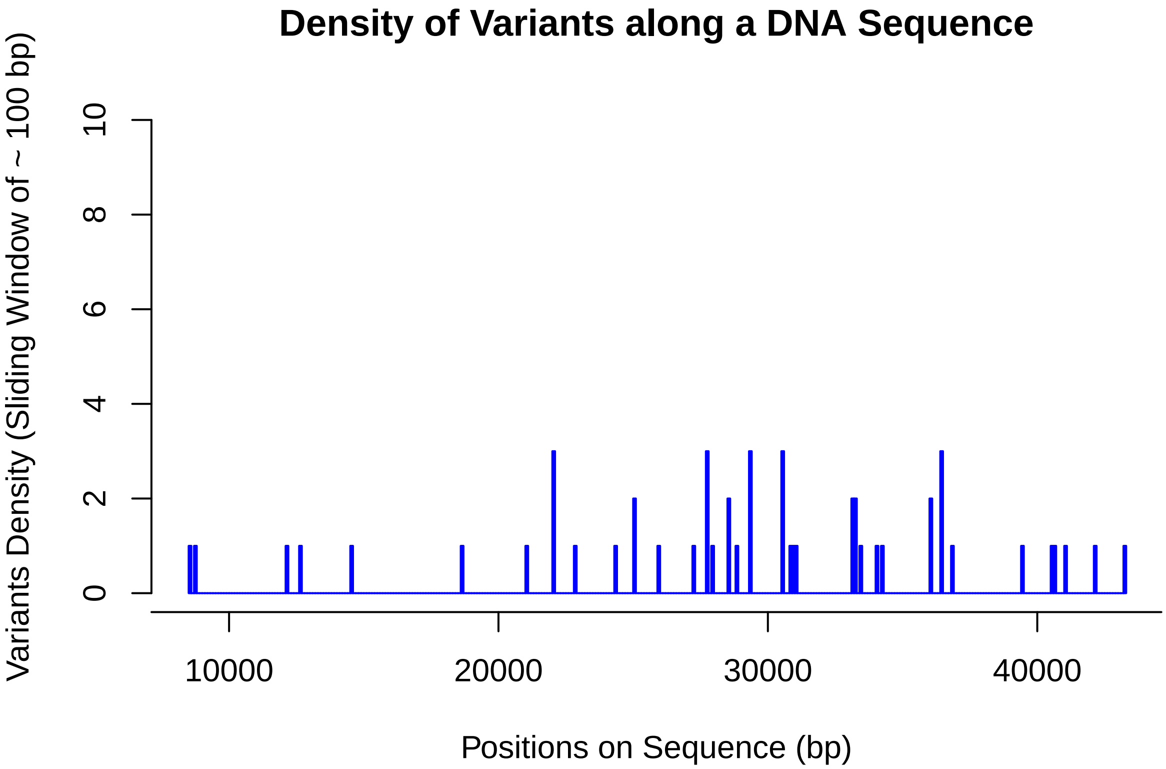
Function axis() parameters:

- Start position: start_position = int;
- End position: end_position = int
- Accuracy or step value: change the graduations of the X axis: step_accuracy = int
- Rounding parameters: round_any(. . . ,f=floor/ceiling/trunc) ; select floor if you want to round to the lower step value ; select ceiling if you want to round to the upper step value.

Default Parameters:

```
Matrix_Position_Sequence<-as.matrix(sequence[1])

start_position = head(sequence[[2]],n=1L)
end_position = tail(sequence[[2]],n=1L)

histogram_breakpoints = 347
window_size = round((end_position - start_position)/histogram_breakpoints)

hist(Matrix_Position_Sequence,
     xlim=c(start_position,end_position),
     breaks = histogram_breakpoints,
     col='blue',
     xaxt="n",
     border='blue',
     ylim=c(0,10),
     main='Density of Variants along a DNA Sequence',
     xlab='Positions on Sequence (bp)',
     ylab=paste("Variants Density (Sliding Window of ~",window_size,"bp)",sep=" "))
```

```
step_accuracy=10**4

axis(1, at = seq(from = round_any(start_position,
                                  step_accuracy,
                                  f=floor),
                 to = round_any(end_position,
                                step_accuracy,
                                f=ceiling),
                 by = round_any(end_position,
                                step_accuracy,
                                f=ceiling)/
                   ((1/step_accuracy)*
                      round_any(end_position,
                                step_accuracy,
                                f=ceiling))))
```



For any question, feel free to ask me by E-mail: yoann.pageaud@gmail.com

References:

“The Variant Call Format (VCF) Version 4.2 Specification” January 26, 2015. https://github.com/samtools/hts-specs