RareVariantVis: Package for visualization of rare variants in whole genome sequencing data

Tomasz Stokowy

October 14, 2015

Introduction

This vignette was created to present how to efficiently visualize and interprete genomic variants in R. Package RareVariantVis aims to present genomic variants (especially rare ones) in a global, per chromosome way. Visualization is performed in two ways - standard that outputs png figures and interactive that uses JavaScript d3 package. Interactive visualization allows to analyze trio/family data, for example in search for causative variants in rare Mendelian diseases.

Input data

In this example we will use whole Complete Genomics genome sequencing data. Son sample from Ashkenazim trio from Stanford GIAB Personal Genome Project was used to prepare visualization input file - data frame with variants. Example of such input file is presented below:

```
library(RareVariantVis)

## Loading required package:  BiocGenerics
## Loading required package:  parallel
##
## Attaching package:  'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
##    clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##    clusterExport, clusterMap, parApply, parCapply, parLapply,
##    parLapplyLB, parRapply, parSapply, parSapplyLB
##
## The following objects are masked from 'package:stats':
##
##    IQR, mad, xtabs
##
## The following objects are masked from 'package:base':
##
##    Filter, Find, Map, Position, Reduce, anyDuplicated, append,
```

```
##    as.data.frame, as.vector, cbind, colnames, do.call,
##    duplicated, eval, evalq, get, grep, grepl, intersect,
##    is.unsorted, lapply, lengths, mapply, match, mget, order,
##    paste, pmax, pmax.int, pmin, pmin.int, rank, rbind, rownames,
##    sapply, setdiff, sort, table, tapply, union, unique, unlist,
##    unsplit
##
## Loading required package:  VariantAnnotation
## Loading required package:  GenomeInfoDb
## Loading required package:  stats4
## Loading required package:  S4Vectors
## Loading required package:  IRanges
## Loading required package:  GenomicRanges
## Loading required package:  SummarizedExperiment
## Loading required package:  Biobase
## Welcome to Bioconductor
##
##    Vignettes contain introductory material; view with
##    'browseVignettes()'.  To cite Bioconductor, see
##    'citation("Biobase")', and for packages 'citation("pkgname")'.
##
## Loading required package:  Rsamtools
## Loading required package:  XVector
## Loading required package:  Biostrings
##
## Attaching package:  'VariantAnnotation'
##
## The following object is masked from 'package:base':
##
##    tabulate
##
## Loading required package:  googleVis
## Note:  the specification for S3 class "AsIs" in package 'RJSONIO'
seems equivalent to one from package 'BiocGenerics':  not turning on
duplicate class definitions for this class.
##
## Welcome to googleVis version 0.5.10
##
## Please read the Google API Terms of Use
## before you start using the package:
## https://developers.google.com/terms/
##
## Note, the plot method of googleVis will by default use
## the standard browser to display its output.
##
## See the googleVis package vignettes for more details,
## or visit http://github.com/mages/googleVis.
##
## To suppress this message use:
```

```
## suppressPackageStartupMessages(library(googleVis))

library(AshkenazimSonChr21)
head(SonVariantsChr21)
```

Data frame consists of columns from vcf file. Mandatory colums are:

- Start.position - for location on chromosome,
- SNP.Frequency - for dbSNP frequency,
- DP - sequencing depth,
- AD - allelic depths for reference and alternative alleles.
- Gene.name - gene symbol
- Gene.component - part of gene
- Variant.type - type of variant

Gene.component field accepts such regions as `EXON_REGION`, `SA_SITE_CANONICAL`, `SD_SITE_CANONICAL`, `UTR`, `INTRON_REGION` and other. It can be also empty space for intergenic regions. Variant.type field accepts following types: Substitution - nonsynonymous, Substitution - nonsense, Complex, Deletion - frameshift, Insertion - frameshift, Substitution, Substitution - synonymous and other.

Large variant files are also accepted - computer with 16GB RAM can handle input files up to 1GB.

Visualization options

There are two main visualization functionalities of the package - static for all variants and dynamic for rare variants. Static aims for vizualization of all variants on the chromosome, whereas dynamic for interactive plotting of variants selected in the filtering procedure.

Static visualization

Static visualization is performed by a key function - chromosomeVis. ChromosomeVis provides png plot and filtered list of variants in working directory. Alternatively, it can also provide output plot to current device:

Example of chromosomeVis function that saves plot on the disk:

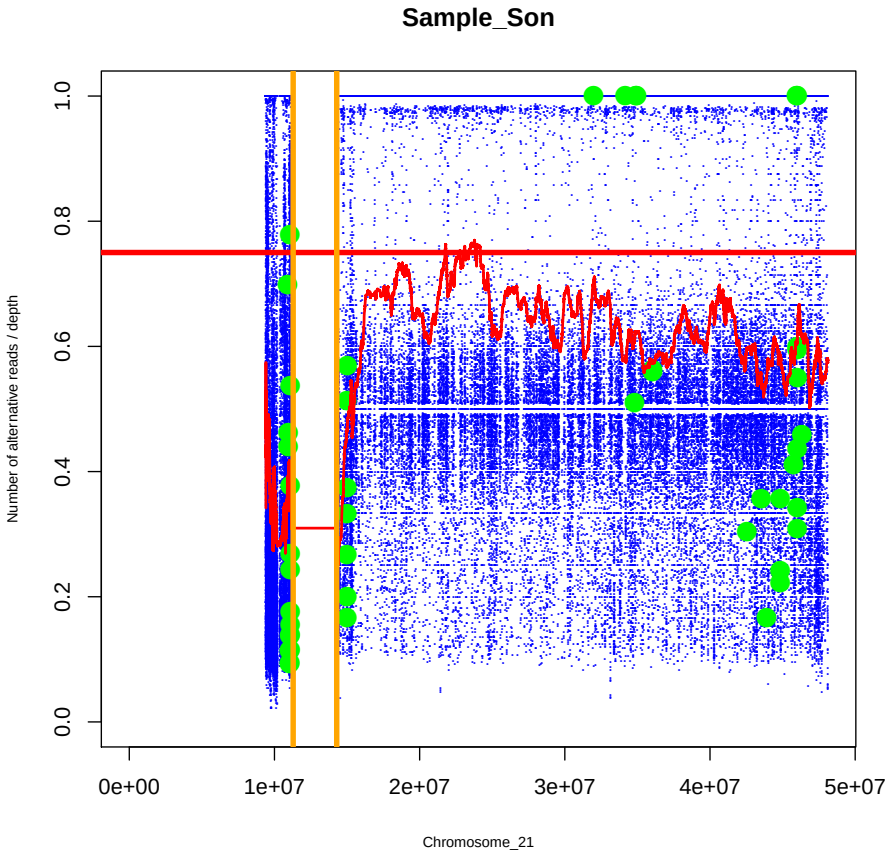
```
chromosomeVis(file = SonVariantsChr21, sample = "Son", chromosome = 21,
              centromeres = CentromeresHg19, pngWidth = 1600,
              pngHeight = 1200, plot = FALSE)

## Input file reading...
## Input file reading completed.
## Analysis finished.
## Your output files are in folder:
## /tmp/Rtmpg6pPs9/Rbuild1ad493d16d2/RareVariantVis/vignettes
```

Example of chromorsomeVis function that provides visualization to the plot device:

```
chromosomeVis(file = SonVariantsChr21, sample = "Son",
               chromosome = 21, centromeres = CentromeresHg19, plot = TRUE)

## Input file reading...
## Input file reading completed.
```




```
## Analysis finished.
## Your output files are in folder:
## /tmp/Rtmpg6pPs9/Rbuild1ad493d16d2/RareVariantVis/vignettes
```

chromosomeVis function provides also option of vcf file visualization. Example of input data and function setup are given in chromosomeVis function manual.

Functions chromosomeVis and chromosomePlot provide an output with genomic loci on x axis and ratio of alternative reads to sequencing depth on y axis. This means that heterozygous variants (for which expected ratio is 0.5) are placed below red line at 0.75. On the other hand, alternative homozygous variants are placed above 0.75, with expected ratio = 1. Some homozygous reference variants are also called, however this ones are ignored by majority of calling tools used for Illumina sequencing data.

Position of the chromosome is marked with orange vertical lines. Red continous line depicts moving average of alternative reads to sequencing depth ratio. This moving average value is based on 2000 variants and provides information about possible homozygous regions (potentially above 0.75).

Green dots depict rare, coding nonsynonymous variants, with reliable depth. Currently, only one rare variant filter setting is provided:

- dbSNP frequency lower than 0.01,
- coding,
- nonsynonymous or nonsense variant,
- sequencing depth greater than 10

This rare variants are also reported by chromosomeVis function to the output file. Output file includes the same columns as input file but for rare variants only. It is possible that input files can include more columns than in the example data (AshkenazimSonChr21). Package was tested to work efficiently with files consisting of 100 annotations (columns).

Dynamic visualization

Dynamic visualization is performed by rareVariantVis function. Function takes as an input rare variants file generated by chromosomeVis function. Example of such data frame input is data(SonRareVariantsChr21).

Example of data and function performance:

```
head(SonRareVariantsChr21)
```

```
##   Chromosome Start.position End.position Reference Variant
## 1      chr21       10910311     10910311         T       G
## 2      chr21       10942756     10942756         G       A
## 3      chr21       10943003     10943003         C       T
## 4      chr21       11049596     11049596         C       T
## 5      chr21       11049617     11049617         C       G
## 6      chr21       11058226     11058226         G       C
##   Quality.by.Depth                  Variant.type     SNP.id SNP.Frequency
## 1          3590.20 Substitution - nonsynonymous     rs9996            -1
## 2          1827.59      Substitution - nonsense  rs1810540            -1
## 3          2583.85 Substitution - nonsynonymous  rs1810856            -1
## 4           197.67 Substitution - nonsynonymous rs28571918            -1
## 5           394.81 Substitution - nonsynonymous  rs2740327            -1
## 6          1031.96 Substitution - nonsynonymous  rs4913558            -1
##   Gene.name Gene.component phyloP  DP     AD  GT
## 1      TPTE    EXON_REGION  0.077 156 47,109 0/1
## 2      TPTE    EXON_REGION  1.553 149  83,65 0/1
## 3      TPTE    EXON_REGION  2.008 176  94,81 0/1
## 4     BAGE2    EXON_REGION  0.468 322 292,30 0/1
## 5     BAGE2    EXON_REGION  0.464 342 303,39 0/1
## 6     BAGE2    EXON_REGION  4.325 356 307,49 0/1
```

```
rareVariantVis(file = SonRareVariantsChr21, sample = "Son",
        chromosome = 21, centromeres = CentromeresHg19)
```

```
## Input file reading...
## Input file reading completed.
## Analysis finished.
```

```
## Your output files are in folder:
## /tmp/Rtmpg6pPs9/Rbuild1ad493d16d2/RareVariantVis/vignettes
```

Functions rareVariantsVis provides an output html file with genomic loci on x axis and ratio of alternative reads to sequencing depth on y axis. Html file with visualized rare variants is located in current working directory. It is possible to point and zoom on the plot. Pointed variants highlight their properties. Right click on the plot cancels all changes made.

Trio analysis

RareVariantVis package is designed also for trio analysis. This approach allows to observe inheritance patterns ans and potential de novo mutations. Moreover, technical effects, regions of sequencing alignment problems and highly polymorphic genome regions are also observed in trio visualization. Function trioVis accepts chromosomeVis output from mother, index and father samples. As an output it provides interactive visualization to current working directory.

```
trioVis(fileMother = MotherRareVariantsChr21,
        fileIndex = SonRareVariantsChr21,
        fileFather = FatherRareVariantsChr21,
        sampleMother = "Mother",
        sampleIndex = "Son",
        sampleFather = "Father",
        chromosome = 21,
        centromeres = CentromeresHg19)

## Input file reading...
## Input file reading completed.
## Analysis finished.
## Your output files are in folder:
## /tmp/Rtmpg6pPs9/Rbuild1ad493d16d2/RareVariantVis/vignettes
```

Another output of the trio analysis is summary table with Inheritance and Gene Count columns. This table provides information about inheritance pattern for all variants of Index sample. It is importance to notice that some records (for example de novo) may be false positive. This fact is caused by highly polymorphic regions, alignment issues in some data or not adequate calling in mother or father samples. It is recomended to check candidate variants in raw data (bam files), for example using Integrative Genome Viewer.