RareVariantVis 2: R suite for analysis of rare variants in whole genome sequencing data.

Adam Gudyś and Tomasz Stokowy

October 30, 2017

Introduction

The search for causative genetic variants in rare diseases of presumed monogenic inheritance has been boosted by the implementation of whole genome sequencing(WGS). Analysis and visualisation of WGS data is demanding due to its size and complexity. To aid this challenge, we have developed a WGS data analysis suite—RareVariantVis 2. This new, significantly extended implementation of RareVariantVis (Stokowy et al., Bioinformatics 2016) annotates, filters and visualises whole human genome in less than 30 minutes. Method accepts and integrates vcf files for single nucleotide, structural and copy number variants. Proposed method was successfully used to disclose causes of three rare monogenic disorders, including one non-coding variant.

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.

This vignette presents example of Genome in a Bottle—NA12878 sample (chromosome 19) which is analyzed in about one minute on laptop/desktop computer. Examples include reading of all chr19 variants, filtering, annotation, visualization and homozygous region calling.

Example

The analysis is preceeded by loading neccessary packages.

```
library(RareVariantVis)
library(DataRareVariantVis)
```

After that, single nucleotide variants (SNV) and structural variants (SV) for NA12878 have to be loaded from `DataRareVariantVis` package. Directories of these files are stored in `sample` and `sv_sample` variables.

```
sample = system.file("extdata", "CoriellIndex_S1.vcf.gz",
    package = "DataRareVariantVis")
sv_sample = system.file("extdata", "CoriellIndex_S1.sv.vcf.gz",
    package = "DataRareVariantVis")
```

The `RareVariantVis` package supports vcf files generated by speedseq and GATK variant callers. The discovery of rare variants in chromosome 19 is performed by executing the following command.

```
chromosomeVis(sample=sample, sv_sample=sv_sample,
    chromosomes=c("19"))
```

As a result, the package creates a set of files describing discovered rare variants:

- *RareVariants_CoriellIndex_S1.txt*—table with single nucleotide variants,
- *ComplexVariants_CoriellIndex_S1.txt*—table with complex variants,
- *StructuralVariants_CoriellIndex_S1.txt*—table with structural variants,
- *CoriellIndex_S1_chr19.png*—per-chromosome variants visualisation.

If more than one chromosome is to be processed, separate visualisation files are generated. All tables are tab-separated text files with rows corresponding to variants and columns describing their properties.

Single nucleotide variants

This table contain variants with a single alternative allele. They are described by the following columns:

- *chromosome*—number of chromosome,
- *final_positions*—position in chromosome,
- *variant_type*—type of mutation: synonymous/nonsynonymous/nonsense,
- *ref_allele*—reference allele,
- *alt_allele*—alternative allele,
- *final_variations*—homozygosity (ratio of alternative allele to depth),
- *conservation*—conservation index according to UCSC phastCons.

Additional columns are filled only for variants localized in coding regions:

- *gene_name*—primary name of gene according to UCSC Human Genome annotation,
- *in_exon*—flag indicating whether variant is in exon,
- *Entry*—corresponding UniProt entry,
- *Status*—protein status (UniProt),
- *Protein.names*—names of the corresponsing proteins (UniProt),
- *Gene.names*—all gene aliases (UniProt),
- *Annotation*—
- *Tissue.specificity*—tissue specificity (UniProt),
- *Gene.ontology..biological.process*—description of biological process the gene is involved in and ontology identifier (UniProt),
- *Involvement.in.disease*—known involvement in diseases (UniProt)
- *Cross.reference..Orphanet*—
- *PubMed.ID*—colon-separated list of related PubMed identifiers (UniProt).

Complex variants

Variants with more than one alternative allele are referred to as complex. They are described by:

- *chromosome*—number of chromosome,
- *start_position*—starting position in the chromosome,

If variant spans over coding regions, additional additional UniProt columns are filled analogously as for SNVs.

Structural variants

The table with structural variants have following columns:

- *chromosome*—number of chromosome,
- *start_position*—starting position in the chromosome,
- *ID*—optional variant identifier,
- *REF*—reference allele,
- *ALT*—alternative allele,
- *QUAL*—Phred-scaled probability that a REF/ALT polymorphism exists at this site given sequencing data,
- *SVTYPE*—variant type: DUP(duplication)/DEL(deletion),
- *GT*—genotype: 0/1 (heterozygous) or 1/1 (homozygous)
- *end_position*—genotype

If variant spans over coding regions, additional additional UniProt columns are filled analogously as in SNVs.

Per-chromosome visualisation

For each analyzed chromosome, discovered variants are visualized. The example visualisation file for NA12878 chromosome 19 is presented in Figure 1. Figure illustrates variants (blue dots) in their genomic coordinates (X axis). Ratio of alternative reads and depth (Y axis) gives information about type of variant: homozygous alternative (expected ratio 1) and heterozygous (expected ratio 0.5). Green dots represent rare variants that pass filters: coding/UTR, nonsynonymous variant with dbSNP frequency ¡ 0.01 and ExAC frequency ¡ 0.01. Orange vertical lines depict position of centromere. Orange dots depict structural and copy number variants that overlap with coding region and are of relatively good quality (QUAL ¿ 0). Red curve illustrates moving average of alternative reads/depth ratio. High values of this curve (exceeding 0.75) can suggest potential homozygous/deleterious regions.

Interactive visualisation

In order to perform interactive visualisation of NA12878 rare SNVs, the following piece of code has to be executed:

```
rareVariantVis("RareVariants_CoriellIndex_S1.txt",
               "RareVariants_CoriellIndex_S1.html",
               "CorielIndex")
```

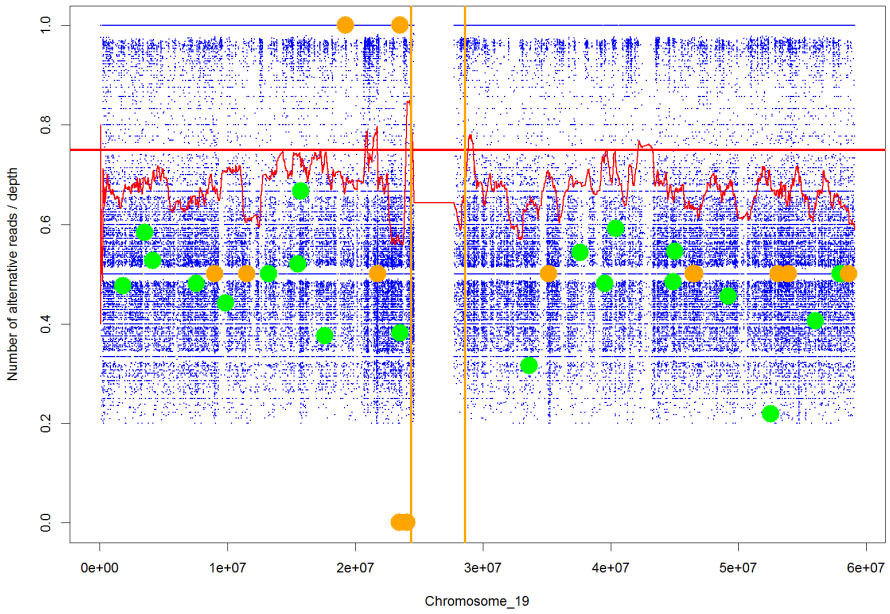

Figure 1: The visualisation file for NA12878 sample, chromosome 19.

It reads a file containing table of rare variants (obtained from `chromosomeVis` function) and provides an adequate visualization (Figure 2). Function outputs visualization html figure in the current working directory. Figure illustrates variants (dots) in their genomic coordinates (X axis). Ratio of alternative reads and depth (Y axis) gives information about type of variant: homozygous alternative (expected ratio 1) and heterozygous (expected ratio 0.5). Zooming the plot is also supported. Pointing on variants provides basic information about the variant—name of gene and the position on the chromosome. If variants from many chromosomes are present in the table, all chromosomes are added to the same html file.

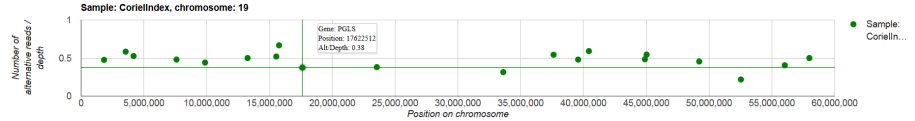

Figure 2: Interactive visualisation file for NA12878 sample, chromosome 19.

Multiple sample visualisation

Another feature of RareVariantVis2 package is the ability to perform visualisation of multiple samples. Let *RareVariants_CoriellIndex_S1.txt* and *RareVariants_Coriell_S2.txt* be variant tables from different geneomes. One can visualise

their 19 chromosome variants simulateously by running:

```
inputFiles = c("RareVariants_CoriellIndex_S1.txt",
               "RareVariants_Coriell_S2.txt")
sampleNames = c("CoriellIndex_S1", "Coriell_S2");
multipleVis(inputFiles, "CorielSamples.html", sampleNames, "19")
```