

AshkenazimSonChr21: Annotated variants on the chromosome 21, human genome 19, Ashkenazim Trio son sample

Tomasz Stokowy

November 5, 2019

Introduction

This vignette describes AshkenazimSonChr21 dataset, example input for RareVariantVis package. This dataset is CompleteGenomics whole genome sequencing dataset, coming from Stanford Genome in a Bottle Consortium. This dataset was made fully available for public, without restrictions. This particular data refer to sample HG002- NA24385 - huAA53E0 (son). Original data can be found at: <https://sites.stanford.edu/abms/content/giab-reference-materials-and-data>

Preprocessing

Original whole genome sequencing sample was (HG002-son) was too big for purpose of R/Bioconductor test data, therefore only chromosome 21 variants were selected. Complete Genomics output provides 3 types of variants: homozygous reference, heterozygous and homozygous alternative. To minimize data size and make it similar to Illumina X Ten output homozygous reference were excluded. Finally, small indels were filtered out, since they introduced a lot of noise into visualization. This noise was not observed in Illumina X Ten samples that we analyzed in our laboratory.

Possible usage of data

Data aims to work well with RareVariantVis package, however it can be used also in other packages that aim for whole genome sequencing data analysis. Dataset includes two types of files: txt file with rare variants and vcf file obtained from sequencing, very similar to one from Illumina X Ten output. Examples of data usage and file structure are listed below.

```
## text file
library(AshkenazimSonChr21)
head(SonVariantsChr21)

## Chromosome Start.position End.position Reference Variant
## 1 chr21 9411318 9411318 C T
```

```

## 2      chr21      9411327      9411327      C      G
## 3      chr21      9411410      9411410      C      T
## 4      chr21      9411500      9411500      G      T
## 5      chr21      9411602      9411602      T      C
## 6      chr21      9411609      9411609      G      T
##      Quality.by.Depth Variant.type      SNP.id SNP.Frequency Gene.name
## 1              313.61 Substitution rs373567667      -1
## 2              720.44 Substitution rs75025155      -1
## 3             1128.86 Substitution rs78200054      -1
## 4             1241.14 Substitution rs71235073      -1
## 5              615.72 Substitution rs368646645      -1
## 6              603.02 Substitution rs76676778      -1
##      Gene.component phyloP DP      AD  GT
## 1              -0.177 38 25,13 0/1
## 2              -0.307 37 13,24 0/1
## 3               0.717 49 15,34 0/1
## 4               0.717 62 24,38 0/1
## 5               0.624 57 35,22 0/1
## 6              -0.163 56 35,21 0/1

## vcf file
library(VariantAnnotation)

## 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, sd, var, xtabs
## The following objects are masked from 'package:base':
##
##      Filter, Find, Map, Position, Reduce, anyDuplicated, append,
##      as.data.frame, basename, cbind, colnames, dirname, do.call,
##      duplicated, eval, evalq, get, grep, grepl, intersect,
##      is.unsorted, lapply, mapply, match, mget, order, paste, pmax,
##      pmax.int, pmin, pmin.int, rank, rbind, rounames, sapply,
##      setdiff, sort, table, tapply, union, unique, unsplit, which,
##      which.max, which.min
## Loading required package: GenomeInfoDb
## Loading required package: S4Vectors
## Loading required package: stats4
##
## Attaching package: 'S4Vectors'

```

```

## The following object is masked from 'package:base':
##
##   expand.grid
## 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: DelayedArray
## Loading required package: matrixStats
##
## Attaching package: 'matrixStats'
## The following objects are masked from 'package:Biobase':
##
##   anyMissing, rowMedians
## Loading required package: BiocParallel
##
## Attaching package: 'DelayedArray'
## The following objects are masked from 'package:matrixStats':
##
##   colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges
## The following objects are masked from 'package:base':
##
##   aperm, apply, rowsum
## Loading required package: Rsamtools
## Loading required package: Biostrings
## Loading required package: XVector
##
## Attaching package: 'Biostrings'
## The following object is masked from 'package:base':
##
##   strsplit
##
## Attaching package: 'VariantAnnotation'
## The following object is masked from 'package:base':
##
##   tabulate

fl <- system.file("extdata", "SonVariantsChr21.vcf.gz",
                  package="AshkenazimSonChr21")
vcf <- readVcf(fl, genome="hg19")
geno(vcf)

## List of length 8
## names(8): GT GQX AD DP GQ MQ PL VF

```

```

info(vcf)
## DataFrame with 94527 rows and 35 columns
##           AC           AF           AN           DP           QD
##   <IntegerList> <character> <integer> <integer> <numeric>
## 1           1           0.50           2           38           8.25
## 2           1           0.50           2           37           19.47
## 3           1           0.50           2           49           23.04
## 4           1           0.50           2           62           20.02
## 5           1           0.50           2           57           10.8
## ...           ...           ...           ...           ...           ...
## 94523        1           0.50           2           101          2.04
## 94524        1           0.50           2           113          2.12
## 94525        1           0.50           2           115          2.01
## 94526        1           0.50           2           155          0.14
## 94527        1           0.50           2           169          0.02
##   BLOCKAVG_min30p3a BaseQRankSum           DS           Dels           END
##   <logical> <numeric> <logical> <numeric> <integer>
## 1           FALSE          -0.923          FALSE           0           NA
## 2           FALSE          -0.334          FALSE           0           NA
## 3           FALSE          -0.683          FALSE           0           NA
## 4           FALSE           1.395          FALSE           0           NA
## 5           FALSE          -1.436          FALSE           0           NA
## ...           ...           ...           ...           ...           ...
## 94523        FALSE           1.834          FALSE           0.01          NA
## 94524        FALSE           2.439          FALSE           0.06          NA
## 94525        FALSE           1.499          FALSE           0.01          NA
## 94526        FALSE           1.67           FALSE           0           NA
## 94527        FALSE           1.448          FALSE           0.01          NA
##           FS           HRun HaplotypeScore InbreedingCoeff           MQ
##   <numeric> <integer> <numeric> <numeric> <numeric>
## 1           0           0           1.9783           NA           51
## 2           1.443         1           0.9995           NA           52
## 3           11.788        1           0.8667           NA           50
## 4           1.005         0           0           NA           52
## 5           0           0           0           NA           53
## ...           ...           ...           ...           ...           ...
## 94523        0           1           128.0372          NA           25
## 94524        0           1           205.8792          NA           24
## 94525        0           1           250.5937          NA           22
## 94526        6.16         0           184.0491          NA           19
## 94527        2.884        3           195.0513          NA           18
##           MQ0 MQRankSum ReadPosRankSum           SB           VQSLOD
##   <integer> <numeric> <numeric> <numeric> <numeric>
## 1           0          -0.031          -0.154          -55.94          2.0206
## 2           0           0.016           0.97           -261.36          4.3216
## 3           0          -0.597          -0.011          -414.78          2.9995
## 4           0           1.322          -1.192          -535.11          2.156
## 5           6           0.086           0.276          -178.59          2.1432
## ...           ...           ...           ...           ...           ...

```

```

## 94523      3   -3.844      -0.805   -88.65  -27.4198
## 94524      4   -1.997      -1.33   -89.77  -60.7511
## 94525      5   -3.745      -0.59  -110.6  -89.2046
## 94526     37   -1.952       3.132   -0.01  -63.3093
## 94527     56   -1.775       2.138   -0.01  -70.4434
##          culprit          set          CSQT
##          <character> <character> <CharacterList>
## 1          QD FilteredInAll          NA
## 2          MQ          variant          NA
## 3          MQ FilteredInAll          NA
## 4          MQ FilteredInAll          NA
## 5          QD FilteredInAll          NA
## ...          ...          ...          ...
## 94523 HaplotypeScore FilteredInAll          NA
## 94524 HaplotypeScore FilteredInAll          NA
## 94525 HaplotypeScore FilteredInAll          NA
## 94526          DP FilteredInAll          NA
## 94527          DP FilteredInAll          NA
##          CSQR          AA
##          <CharacterList> <character>
## 1          NA          NA
## 2          NA          NA
## 3          NA          NA
## 4          NA          NA
## 5          NA          NA
## ...          ...          ...
## 94523 ENSR00000684572|regulatory_region_variant          NA
## 94524 ENSR00000684572|regulatory_region_variant          NA
## 94525 ENSR00000684572|regulatory_region_variant          NA
## 94526 ENSR00000684572|regulatory_region_variant          NA
## 94527 ENSR00000684572|regulatory_region_variant          NA
##          GMAF          EVS          cosmic          clinvar
##          <CharacterList> <CharacterList> <CharacterList> <CharacterList>
## 1          NA          NA          NA          NA
## 2          NA          NA          NA          NA
## 3          NA          NA          NA          NA
## 4          NA          NA          NA          NA
## 5          NA          NA          NA          NA
## ...          ...          ...          ...          ...
## 94523          NA          NA          NA          NA
## 94524          NA          NA          NA          NA
## 94525          NA          NA          NA          NA
## 94526          NA          NA          NA          NA
## 94527          NA          NA          NA          NA
##          phastCons          Variant.type          Gene.name          Gene.component          phyloP
##          <logical> <CharacterList> <CharacterList> <CharacterList> <numeric>
## 1          FALSE          Substitution          NA          NA          -0.177
## 2          FALSE          Substitution          NA          NA          -0.307
## 3          FALSE          Substitution          NA          NA          0.717

```

```

## 4      FALSE      Substitution      NA      NA      0.717
## 5      FALSE      Substitution      NA      NA      0.624
## ...      ...      ...      ...      ...      ...
## 94523  FALSE      Substitution      NA      NA      -100
## 94524  FALSE      Substitution      NA      NA      -100
## 94525  FALSE      Substitution      NA      NA      -100
## 94526  FALSE      Substitution      NA      NA      -100
## 94527  FALSE      Substitution      NA      NA      -100
##      SNP.Frequency
##      <numeric>
## 1      -1
## 2      -1
## 3      -1
## 4      -1
## 5      -1
## ...      ...
## 94523  -1
## 94524  -1
## 94525  -1
## 94526  -1
## 94527  -1

```