

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

Tomasz Stokowy

May 20, 2021

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 Quality.by.Depth
## 1         chr21      9411318     9411318         C         T           313.61
```

```

## 2 chr21 9411327 9411327 C G 720.44
## 3 chr21 9411410 9411410 C T 1128.86
## 4 chr21 9411500 9411500 G T 1241.14
## 5 chr21 9411602 9411602 T C 615.72
## 6 chr21 9411609 9411609 G T 603.02
## Variant.type SNP.id SNP.Frequency Gene.name Gene.component phyloP DP
## 1 Substitution rs373567667 -1 -0.177 38
## 2 Substitution rs75025155 -1 -0.307 37
## 3 Substitution rs78200054 -1 0.717 49
## 4 Substitution rs71235073 -1 0.717 62
## 5 Substitution rs368646645 -1 0.624 57
## 6 Substitution rs76676778 -1 -0.163 56
## AD GT
## 1 25,13 0/1
## 2 13,24 0/1
## 3 15,34 0/1
## 4 24,38 0/1
## 5 35,22 0/1
## 6 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, rownames, sapply, setdiff, sort, table,
## tapply, union, unique, unsplit, which.max, which.min
## Loading required package: MatrixGenerics
## Loading required package: matrixStats
##
## Attaching package: 'MatrixGenerics'
## The following objects are masked from 'package:matrixStats':
##
## colAlls, colAnyNAs, colAnys, colAugsPerRowSet, colCollapse,

```

```

##      colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##      colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##      colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##      colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##      colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##      colWeightedMeans, colWeightedMedians, colWeightedSds,
##      colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAugsPerColSet,
##      rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##      rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##      rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##      rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##      rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##      rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##      rowWeightedSds, rowWeightedVars
## Loading required package: GenomeInfoDb
## Loading required package: S4Vectors
## Loading required package: stats4
##
## Attaching package: 'S4Vectors'
## The following objects are masked from 'package:base':
##
##      I, expand.grid, unname
## 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)".
##
## Attaching package: 'Biobase'
## The following object is masked from 'package:MatrixGenerics':
##
##      rowMedians
## The following objects are masked from 'package:matrixStats':
##
##      anyMissing, rowMedians
## 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 BLOCKAVG_min30p3a
##   <IntegerList> <character> <integer> <integer> <numeric>          <logical>
## 1             1           0.50           2           38           8.25           FALSE
## 2             1           0.50           2           37           19.47          FALSE
## 3             1           0.50           2           49           23.04          FALSE
## 4             1           0.50           2           62           20.02          FALSE
## 5             1           0.50           2           57           10.80          FALSE
## ...           ...           ...           ...           ...           ...           ...
## 94523          1           0.50           2           101          2.04           FALSE
## 94524          1           0.50           2           113          2.12           FALSE
## 94525          1           0.50           2           115          2.01           FALSE
## 94526          1           0.50           2           155          0.14           FALSE
## 94527          1           0.50           2           169          0.02           FALSE
##           BaseQRankSum          DS          Dels          END          FS          HRun
##   <numeric> <logical> <numeric> <integer> <numeric> <integer>
## 1          -0.923          FALSE           0           NA           0.000           0
## 2          -0.334          FALSE           0           NA           1.443           1
## 3          -0.683          FALSE           0           NA           11.788          1
## 4           1.395          FALSE           0           NA           1.005           0
## 5          -1.436          FALSE           0           NA           0.000           0
## ...           ...           ...           ...           ...           ...           ...
## 94523          1.834          FALSE           0.01           NA           0.000           1
## 94524          2.439          FALSE           0.06           NA           0.000           1
## 94525          1.499          FALSE           0.01           NA           0.000           1
## 94526          1.670          FALSE           0.00           NA           6.160           0
## 94527          1.448          FALSE           0.01           NA           2.884           3
##           HaplotypeScore InbreedingCoeff          MQ          MQ0 MQRankSum
##   <numeric>          <numeric> <numeric> <integer> <numeric>
## 1           1.9783           NA           51           0          -0.031
## 2           0.9995           NA           52           0           0.016
## 3           0.8667           NA           50           0          -0.597
## 4           0.0000           NA           52           0           1.322
## 5           0.0000           NA           53           6           0.086
## ...           ...           ...           ...           ...           ...
## 94523          128.037           NA           25           3          -3.844

```

```

## 94524      205.879      NA      24      4      -1.997
## 94525      250.594      NA      22      5      -3.745
## 94526      184.049      NA      19      37     -1.952
## 94527      195.051      NA      18      56     -1.775
##      ReadPosRankSum      SB      VQSLOD      culprit      set
##      <numeric> <numeric> <numeric>      <character>      <character>
## 1          -0.154      -55.94      2.0206      QD FilteredInAll
## 2           0.970     -261.36      4.3216      MQ      variant
## 3          -0.011     -414.78      2.9995      MQ FilteredInAll
## 4          -1.192     -535.11      2.1560      MQ FilteredInAll
## 5           0.276     -178.59      2.1432      QD FilteredInAll
## ...      ...      ...      ...      ...      ...
## 94523      -0.805     -88.65     -27.4198 HaplotypeScore FilteredInAll
## 94524      -1.330     -89.77     -60.7511 HaplotypeScore FilteredInAll
## 94525      -0.590     -110.60    -89.2046 HaplotypeScore FilteredInAll
## 94526         3.132       -0.01    -63.3093      DP FilteredInAll
## 94527         2.138       -0.01    -70.4434      DP FilteredInAll
##      CSQT      CSQR      AA      GMAF
##      <CharacterList>      <CharacterList> <character> <CharacterList>
## 1      NA
## 2      NA
## 3      NA
## 4      NA
## 5      NA
## ...      ...      ...      ...      ...
## 94523      ENSR00000684572|regu..      NA
## 94524      ENSR00000684572|regu..      NA
## 94525      ENSR00000684572|regu..      NA
## 94526      ENSR00000684572|regu..      NA
## 94527      ENSR00000684572|regu..      NA
##      EVS      cosmic      clinvar phastCons      Variant.type
##      <CharacterList> <CharacterList> <CharacterList> <logical> <CharacterList>
## 1      FALSE      Substitution
## 2      FALSE      Substitution
## 3      FALSE      Substitution
## 4      FALSE      Substitution
## 5      FALSE      Substitution
## ...      ...      ...      ...      ...
## 94523      FALSE      Substitution
## 94524      FALSE      Substitution
## 94525      FALSE      Substitution
## 94526      FALSE      Substitution
## 94527      FALSE      Substitution
##      Gene.name      Gene.component      phyloP      SNP.Frequency
##      <CharacterList> <CharacterList> <numeric>      <numeric>
## 1      -0.177      -1
## 2      -0.307      -1
## 3      0.717      -1
## 4      0.717      -1

```

```
## 5 0.624 -1
## ... ...
## 94523 -100 -1
## 94524 -100 -1
## 94525 -100 -1
## 94526 -100 -1
## 94527 -100 -1
```