Using the myvariant package to annotate variants

Dave Tang

3 November 2016

Install necessary packages

```
source("https://bioconductor.org/biocLite.R")
biocLite('myvariant')
biocLite('VariantAnnotation')
```

Load libraries

```
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, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit

## Loading required package: GenomeInfoDb

## Loading required package: stats4

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     colMeans, colSums, expand.grid, rowMeans, rowSums
## 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: Biostrings
```

```
## Loading required package: XVector
```

```
##
## Attaching package: 'VariantAnnotation'
```

```
## The following object is masked from 'package:base':
##
##     tabulate
```

```r
library(myvariant)
```

Download example file

```r
download.file(url = 'http://davetang.org/eg/Pfeiffer.vcf', destfile = 'Pfeiffer.vcf')
my_vcf <- readVcf('Pfeiffer.vcf', genome = 'hg19')
my_hgvs <- formatHgvs(my_vcf)
head(my_hgvs)
```

```
## [1] "chr1:g.879317C>T" "chr1:g.879482G>C" "chr1:g.880390C>A"
## [4] "chr1:g.881627G>A" "chr1:g.884091C>G" "chr1:g.884101A>C"
```

```r
length(my_hgvs)
```

```
## [1] 37709
```

Obtain annotations for your variants

```r
my_var <- getVariants(my_hgvs)
```

```
## Querying chunk 1 of 38
```

```
## Querying chunk 2 of 38
```

```
## Querying chunk 3 of 38
```

```
## Querying chunk 4 of 38
```

```
## Querying chunk 5 of 38
```

```
## Querying chunk 6 of 38
## Querying chunk 7 of 38
## Querying chunk 8 of 38
## Querying chunk 9 of 38
## Querying chunk 10 of 38
## Querying chunk 11 of 38
## Querying chunk 12 of 38
## Querying chunk 13 of 38
## Querying chunk 14 of 38
## Querying chunk 15 of 38
## Querying chunk 16 of 38
## Querying chunk 17 of 38
## Querying chunk 18 of 38
## Querying chunk 19 of 38
## Querying chunk 20 of 38
## Querying chunk 21 of 38
## Querying chunk 22 of 38
## Querying chunk 23 of 38
## Querying chunk 24 of 38
## Querying chunk 25 of 38
## Querying chunk 26 of 38
## Querying chunk 27 of 38
## Querying chunk 28 of 38
## Querying chunk 29 of 38
## Querying chunk 30 of 38
## Querying chunk 31 of 38
## Querying chunk 32 of 38
## Querying chunk 33 of 38
## Querying chunk 34 of 38
## Querying chunk 35 of 38
## Querying chunk 36 of 38
## Querying chunk 37 of 38
## Querying chunk 38 of 38
## Concatenating data, please be patient.
```

Checking out the variant annotations

```
class(my_var)
```

```
## [1] "DataFrame"
## attr(,"package")
## [1] "S4Vectors"
```

```
dim(my_var)
```

```
## [1] 37709   695
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:VariantAnnotation':
##
##     select

## The following objects are masked from 'package:Biostrings':
##
##     collapse, intersect, setdiff, setequal, union

## The following object is masked from 'package:XVector':
##
##     slice

## The following object is masked from 'package:Biobase':
##
##     combine

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, regroup, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
my_var_tbl <- tbl_df(my_var)
```

```
## Warning in as.data.frame(x, row.names = NULL, optional = optional, ...):
## Arguments in '...' ignored
```

```
dim(my_var_tbl)
```

```
## [1] 37709   695
```

```
my_var_tbl %>% select(notfound) %>% count(notfound)
```

```
## # A tibble: 2 × 2
##   notfound     n
##      <lgl> <int>
## 1     TRUE  4240
## 2       NA 33469
```

Filtering cases that were not found

```
my_var_tbl %>% filter (is.na(notfound)) %>% select(query, starts_with('evs')) %>% dim()
```

```
## [1] 33469   122
```

```
my_var_tbl %>% filter (is.na(notfound)) %>% select(query, starts_with('cadd')) %>% dim()
```

```
## [1] 33469   120
```

ClinVar

```
my_var_tbl %>% filter (is.na(notfound), !is.na(clinvar.omim)) %>% select(query, starts_with('clinvar'))
```

```
## # A tibble: 94 × 33
##                   query clinvar.allele_id clinvar.alt clinvar.chrom
##                   <chr>             <int>       <chr>         <chr>
## 1    chr1:g.9323910G>A             31170           A             1
## 2   chr1:g.11854476T>G             18560           G             1
## 3   chr1:g.66058513A>G             23560           G             1
## 4   chr1:g.70904800G>T             17980           T             1
## 5   chr1:g.76199277A>C             98174           C             1
## 6   chr1:g.94512565C>T             22952           T             1
## 7  chr1:g.196642233G>A             31589           A             1
## 8  chr1:g.197031021C>T             31559           T             1
## 9  chr1:g.201079235C>T             32667           T             1
## 10 chr1:g.223284599T>C             21698           C             1
## # ... with 84 more rows, and 29 more variables: clinvar.cytogenic <chr>,
## #   clinvar.rcv <S3: AsIs>, clinvar.ref <chr>, clinvar.rsid <chr>,
## #   clinvar.type <chr>, clinvar.variant_id <int>, clinvar.omim <chr>,
## #   clinvar.gene.id <chr>, clinvar.gene.symbol <chr>,
## #   clinvar.hg19.end <int>, clinvar.hg19.start <int>,
## #   clinvar.hg38.end <int>, clinvar.hg38.start <int>,
## #   clinvar.hgvs.coding <S3: AsIs>, clinvar.hgvs.genomic <S3: AsIs>,
## #   clinvar.uniprot <chr>, clinvar.rcv.accession <chr>,
## #   clinvar.rcv.clinical_significance <chr>,
## #   clinvar.rcv.last_evaluated <chr>, clinvar.rcv.number_submitters <int>,
## #   clinvar.rcv.origin <chr>, clinvar.rcv.preferred_name <chr>,
## #   clinvar.rcv.review_status <chr>,
## #   clinvar.rcv.conditions.age_of_onset <chr>,
## #   clinvar.rcv.conditions.name <chr>,
## #   clinvar.rcv.conditions.synonyms <S3: AsIs>,
## #   clinvar.rcv.conditions.identifiers.medgen <chr>,
## #   clinvar.rcv.conditions.identifiers.omim <chr>,
## #   clinvar.rcv.conditions.identifiers.orphanet <chr>
```

dbSNP

```
my_var_tbl %>% filter (is.na(notfound), dbsnp.validated == 'TRUE') %>% select(query, starts_with('dbsnp
```

```
## # A tibble: 32,662 × 19
##                 query dbsnp.allele_origin dbsnp.alleles dbsnp.alt
##                 <chr>               <chr>    <S3: AsIs>     <chr>
## 1   chr1:g.879317C>T         unspecified    <S3: AsIs>         T
## 2   chr1:g.879482G>C         unspecified    <S3: AsIs>         C
## 3   chr1:g.880390C>A         unspecified    <S3: AsIs>         A
## 4   chr1:g.881627G>A         unspecified    <S3: AsIs>         A
## 5   chr1:g.884091C>G         unspecified    <S3: AsIs>         G
## 6   chr1:g.892460G>C         unspecified    <S3: AsIs>         C
## 7   chr1:g.897730C>T         unspecified    <S3: AsIs>         T
## 8   chr1:g.909238G>C         unspecified    <S3: AsIs>         C
## 9   chr1:g.948921T>C         unspecified    <S3: AsIs>         C
## 10 chr1:g.1021346A>G         unspecified    <S3: AsIs>         G
## # ... with 32,652 more rows, and 15 more variables: dbsnp.chrom <chr>,
## #   dbsnp.class <chr>, dbsnp.dbsnp_build <int>, dbsnp.flags <S3: AsIs>,
## #   dbsnp.gene <S3: AsIs>, dbsnp.gmaf <dbl>, dbsnp.ref <chr>,
## #   dbsnp.rsid <chr>, dbsnp.validated <lgl>, dbsnp.var_subtype <chr>,
## #   dbsnp.vartype <chr>, dbsnp.hg19.end <int>, dbsnp.hg19.start <int>,
## #   dbsnp.gene.geneid <chr>, dbsnp.gene.symbol <chr>
```