Annotation of Genetic Variants

Valerie Obenchain

Fred Hutchinson Cancer Research Center

27-28 February 2012

Read VCF Files

Structural location of variants

Amino acid coding changes

Extras

Outline

Read VCF Files

Structural location of variants

Amino acid coding changes

Extras

VCF (Variant Call Format)

Format description on 1000 Genomes Web site, `http://www.1000genomes.org`

- fixed : CHROM, POS, ID, REF, ALT, QUAL, FILTER
- info : data in INFO field, includes values such as allele count or frequency, membership in dbSNP or HapMap, etc.
- geno : genotype information for samples defined in FORMAT field

readVcf

VCF object

```
> fl <- system.file("extdata", "ex1.vcf", package = "VariantAnnotation"
> vcf <- readVcf(fl, "hg19")
> vcf
```

```
class: VCF
dim: 10 2
genome: hg19
exptData(1): HEADER
fixed(4): REF ALT QUAL FILTER
info(1): DP
geno(3): GT GQ DP
rownames(10): 16:97430 16:101558 ... 21:6765544
  21:9779122
rowData values names(1): rangeID
colnames(2): A B
colData names(1): Samples
```

readVcf

ScanVcfParam

- Specify subsets of data by genomic position (ranges) or VCF fields
- Ranges are specified with the `which` argument
- VCF elements are specified with `fixed`, `info` and `geno` arguments

```
> param <- ScanVcfParam(which = GRanges("chr1", IRanges(1, 1e8)),
+                       asGRanges = FALSE,
+                       fixed = c("ALT", "FILTER"),
+                       info = "DP",
+                       geno = c("DP", "GT"))
```

seqlevels

Helper functions to aid with renaming and subsetting seqlevels

renameSeqlevels

- `renameSeqlevels` accepts a named character vector in the format of oldname=newname

```
> library(TxDb.Hsapiens.UCSC.hg19.knownGene)
> txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
> vcf_mod <- renameSeqlevels(vcf, c("16"="chr16", "21"="chr21"))
> intersect(seqlevels(vcf_mod), seqlevels(txdb))
```

```
[1] "chr16" "chr21"
```

keepSeqlevels

- `keepSeqlevels` accepts a character vector of seqlevels to ''keep''

```
> vcf_21 <- keepSeqlevels(vcf_mod, "chr21")
> seqlevels(vcf_21)
```

```
[1] "chr21"
```

Outline

Read VCF Files

Structural location of variants

Amino acid coding changes

Extras

locateVariants

- *TranscriptDb* annotations are used to identify variants that fall in coding, intron, 5'UTR or 3'UTR regions
- Output is a *DataFrame* with one row for each variant-transcript match

```
> loc <- locateVariants(vcf_mod, txdb)
> head(loc)
DataFrame with 6 rows and 7 columns
    queryID location      txID     cdsID      geneID precedesID   followsID
  <integer> <factor> <integer> <integer> <character> <character> <character>
1         1   coding     58928    173190       51728          NA          NA
2         2   coding     58928    173190       51728          NA          NA
3         3   coding     58928    173190       51728          NA          NA
4         4   intron     58970    173318       55692          NA          NA
5         4   intron     58971    173328        8312          NA          NA
6         4   intron     58972    173328        8312          NA          NA
```

locateVariants

Intergenic variants have gene IDs for precedes and follows

```
> loc[loc$location == "intergenic",]
DataFrame with 4 rows and 7 columns
    queryID   location      txID     cdsID      geneID  precedesID   followsID
  <integer>   <factor> <integer> <integer> <character> <character> <character>
1         7 intergenic        NA        NA          NA   100500862   100132288
2         8 intergenic        NA        NA          NA   100500862   100132288
3         9 intergenic        NA        NA          NA   100500862   100132288
4        10 intergenic        NA        NA          NA   100500862   100132288
```

Outline

Read VCF Files

Structural location of variants

Amino acid coding changes

Extras

predictCoding

- Compute amino acid codes with *BSgenome* or fasta file reference and user supplied variant alleles
- Output is a *DataFrame* with one row for each variant-transcript match. Results for coding variants only.

```
> library(BSgenome.Hsapiens.UCSC.hg19)
> aa <- predictCoding(vcf_mod, txdb, Hsapiens)
> head(aa, 5)
DataFrame with 5 rows and 9 columns
    queryID   consequence         refSeq         varSeq         refAA
  <integer>      <factor> <DNAStringSet> <DNAStringSet> <AAStringSet>
1         1 nonsynonymous         GATTAG            GTA            D*
2         1    frameshift         GATTAG             GT            D*
3         2    frameshift         GCAGAG           GGAG            AE
4         2 nonsynonymous         GCAGAG         GGTAAG            AE
5         3     synonymous         AAGGTA         AAGGTA            KV
          varAA        txID   geneID     cdsID
  <AAStringSet> <character> <factor> <integer>
1             V       58928    51728    173188
2                     58928    51728    173188
3                     58928    51728    173189
4            GK       58928    51728    173189
5            KV       58928    51728    173190
```

Consequence of coding changes

SIFT (Sort Intolerant From Tolerant)

- predicts possible impact of amino acid substitution on protein function
- protein evolution is correlated with protein function; positions important for function are conserved
- uses multiple alignment information to predict tolerated and deleterious substitutions for every position of the query

PolyPhen (Polymorphism Phenotyping)

- predicts possible impact of amino acid substitution on protein function and structure
- applies empirical rules to the sequence, phylogenetic and structural information characterizing the substitution
- uses multiple alignment, UniProt features and structural databases

SIFT example

```
> library(SIFT.Hsapiens.dbSNP132)
> rsids <- c("rs2142947", "rs3026284")
> subst <- c("AACHANGE", "METHOD", "AA", "PREDICTION", "SCORE")
> select(SIFT.Hsapiens.dbSNP132, keys = rsids, cols = subst)

       RSID AACHANGE    METHOD AA  PREDICTION SCORE
1 rs2142947    F430L BEST HITS  L   TOLERATED  1.00
2 rs2142947    F430L BEST HITS  F   TOLERATED  0.74
3 rs2142947    F430L  ALL HITS  L   TOLERATED  0.72
4 rs2142947    F430L  ALL HITS  F   TOLERATED     1
5 rs3026284    G202D BEST HITS  D DELETERIOUS  0.03
6 rs3026284    G202D BEST HITS  G   TOLERATED  1.00
7 rs3026284    G202D  ALL HITS  D DELETERIOUS  0.00
8 rs3026284    G202D  ALL HITS  G   TOLERATED     1
```

Outline

Read VCF Files

Structural location of variants

Amino acid coding changes

Extras

Other functions in *VariantAnnotation*

- long form *GRanges* : set `asGRanges=TRUE` in *ScanVcfParam*
- `MatrixToSnpMatrix` converts 'GT' genotype in *VCF* object to a *SnpMatrix*
- `writeVcf` – in progress –