Annotation

Martin Morgan[1]

June 23 – 28, 2013

[1] mtmorgan@fhcrc.org

Overview

1. Annotation
   - Genes
   - Gene models
   - Web / cloud resources
2. Called variants
   - VCF files
   - Variant annotation

Genes: *org.** & friends

Gene-centric annotation packages[2]

- Organism: *org.Hs.eg.db*, ...
- Pathway: *GO.db*, ...
- Microarray: *hgu95av2.db*, ...
- And: *MotifDb*, ...

The 'select' interface

```
library(org.Hs.eg.db)
keytypes(org.Hs.eg.db)        # available queries
keys(org.Hs.eg.db, "SYMBOL") # possible 'SYMBOL' values
cols(org.Hs.eg.db)            # available return values
select(org.Hs.eg.db, "TERT", keytype="SYMBOL",
    c("ENTREZID", "GENENAME", "GO"))
```

[2] http://bioconductor.org/packages/release/BiocViews.html

Gene models: *TxDb.**

- Exon and transcript coordinates for model organisms
- 'Select' interface – `keytypes`, `keys`, `cols`, `select`

```
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
select(txdb, "7015", keytype="GENEID", "TXNAME")
```

- *GRanges* queries: `transcripts`, `exons`, `cds`, `promoters`
- *GRangesList* queries: `transcriptsBy`, `exonsBy`, `cdsBy`, `intronsByTranscript`, `fiveUTRsByTranscript`, `threeUTRsByTranscript`

```
exonsBy(txdb, "tx")
exonsBy(txdb, "gene")["7015"]
```

Web resources: *biomaRt* & friends

Ensembl Biomart[3] queries

```
library(biomaRt)
listMarts()                              # many marts!
listDatasets(useMart("ensembl")) # many datasets!
ensembl <-
    useMart("ensembl", dataset="hsapiens_gene_ensembl")

listFilters(ensembl)                     # restrict query
listAttributes(ensembl)                  # specify return values

getBM(mart=ensembl,
    filters="chromosome_name", values=c("21", "22"),
    attributes=c("ensembl_gene_id", "chromosome_name"))
```

[3] http://www.ensembl.org/biomart

Web resources: *biomaRt* & friends

Also:

- *KEGGREST*, *uniprot.ws*, ...
- `GenomicFeatures::makeTranscriptDbFromBiomart`

*org.** vs. *biomaRt*?

- Scope: *biomaRt* much more extensive
- Convenience: *org.Hs.eg.db* passes the 'airplane' test
- Reproducibility: *org.Hs.eg.db* data consistent within each *Bioconductor* release

Cloud resources: *AnnotationHub*

Goal

- 'Mirror' large resources in *Bioconductor*-friendly formats
- Easily retrieve into *R* session
- Embody *Bioconductor* principles, e.g., metadata, versioning, ...

Available – examples

- Ensembl FASTA (*FaFile*), GTF (*GRanges*), release 69+
- ENCODE tracks at UCSC (*GRanges*)
- dbSNP VCF files (*GRanges*)

Use

```
library(AnnotationHub)
hub <- AnnotationHub()  # query for available resources
## metadata(hub) -- info. about available resources
## hub$<tab> -- tab completion to select resource
```

Variants: VCF files

- Text file describing called variants (SNPs or more complicated)
- Metadata, header, and then rows for each variant
- Exact content described by metadata, depends on source of VCF

9 fields plus fields for each sample

- `CHROM`, `POS`, `ID`
- `REF`, `ALT`: Reference and alternate (variant) sequence
- `QUAL`, `FILTER`
- `INFO`: Per-variant information
- `FORMAT`: Instructions for interpreting per-sample information

Variants: VCF files

- Example: 1000 genomes chr22, five samples.

Metadata:

- `##INFO=<ID=AN,Number=1,Type=Integer,`
  `Description="Total Allele Count">`

One row:

- `22 50300078 rs7410291`
- `A G`
- `100 PASS`
- `AN=2184;AVGPOST=0.9890;THETA=0.0005;VT=SNP;`
  `RSQ=0.9856;ERATE=0.0020;SNPSOURCE=LOWCOV;`
  `AA=N;AC=751;LDAF=0.3431;AF=0.34;ASN_AF=0.19;`
  `AMR_AF=0.20;AFR_AF=0.83;EUR_AF=0.22`
- `GT:DS:GL`
- `0|0:0.000:-0.00,-2.77,-5.00` (similar for each sample)

Variants: *VariantAnnotation*

Index

```
library(Rsamtools)
indexTabix("chr22.vcf.gz", type="vcf4")
```

Discovery

```
library(VariantAnnotation)
scanVcfHeader("chr22.vcf.gz")
```

Query: *SummarizedExperiment*-like *VCF* class

```
which <- GRanges("22",
    IRanges(c(321680, 14477080), c(321689, 14477090)))
param <- ScanVcfParam(info=c("AN", "LDAF"), geno=NA,
    which=which)
readVcf("chr22.vcf.gz", "hg19", param=param)
```

In devel

- `readInfo`, `readGeno`, `readGT` for simple queries

Variant annotation

Bioconductor

- `locateVariants` with *TranscriptDb* / *GRangesList*
- `predictCoding` with *TranscriptDb* and *BSgenome* / *FaFile*
- `find` / `countOverlaps`

SIFT / PolyPhen

- *SIFT.Hsapiens.dbSNP132*, *PolyPhen.Hapiens.dbSNP131.db*: rsid look-up via 'select' interface

ensemblVEP

- Interface to Ensembl Variant Effect Predictor using Ensembl perl script and web interface

Summary & Acknowledgments

Summary

1. Annotation
   - Genes
   - Gene models
   - Web / cloud resources
2. Called variants
   - VCF files
   - Variant annotation

Acknowledgments

- Marc Carlson (Genes, gene models)
- Valerie Obenchain (Variants)
- Dan Tenenbaum (*AnnotationHub*)
- Hervé Pagès (Gene models)
- Paul Shannon (*MotifDb*, *AnnotationHub*)

And

- Steffen Durink (*biomaRt*)