Annotating genes, genomes, and variants

Martin Morgan (martin.morgan@roswellpark.org)
Roswell Park Cancer Institute
Buffalo, NY, USA

14 July, 2016

What is ‘Annotation’?

- Genes – classification schemes (e.g., Entrez, Ensembl), pathway membership, . . .
- Genomes – reference genomes; exons, transcripts, coding sequence; coding consequences
- System / network biology – pathways, biochemical reactions, . . .
- ‘Consortium’ resources, TCGA, ENCODE, dbSNP, GTEx, . . .

Other defintions (not covered here)

- SNP (and similar) consequences (*VariantAnnotation*, *VariantFiltering*, *ensemblVEP*)
- Assign function to novel sequences
- . . .

Bioconductor Annotation Resources – Packages

Model organism annotation packages

- *org.** – gene names and pathways
- *TxDb.** – gene models
- *BSgenome.** – whole-genome sequences

*org.** packages

The ‘select’ interface:

- Discovery: `keytypes`, `columns`, `keys`
- Retrieval: `select`, `mapIds`

```r
library(org.Hs.eg.db)
keytypes(org.Hs.eg.db)
columns(org.Hs.eg.db)
egid <-
  select(org.Hs.eg.db, "BRCA1", "ENTREZID", "SYMBOL")
```

*org.** (and other annotation) packages – Under the hood...

SQL (sqlite) data bases

- `org.Hs.eg_dbconn()` to query using *RSQLite* package
- `org.Hs.eg_dbfile()` to discover location and query outside *R*.

TxDb.* packages

- Gene models for common model organsisms / genome builds / known gene schemes
- Supports the ‘select’ interface (`keytypes`, `columns`, `keys`, `select`)
- ‘Easy’ to build custom packages when gene model exist

Retrieving genomic ranges

- `transcripts`, `exons`, `cds`,
- `transcriptsBy` , `exonsBy`, `cdsBy` – group by gene, transcirpt, etc.

```r
library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
cdsByTx <- cdsBy(txdb, "tx")
```

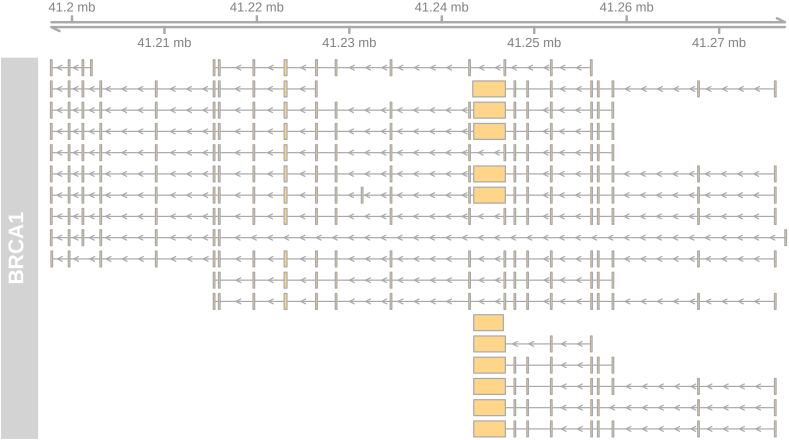
Example: Visualize BRCA1 Transcripts

```r
library(org.Hs.eg.db)
eid <- mapIds(org.Hs.eg.db, "BRCA1", "ENTREZID",
  "SYMBOL")

library(TxDb.Hsapiens.UCSC.hg19.knownGene)
txdb <- TxDb.Hsapiens.UCSC.hg19.knownGene
txid <- select(txdb, eid, "TXNAME", "GENEID")[["TXNAME"]]
cds <- cdsBy(txdb, by="tx", use.names=TRUE)
brca1cds <- cds[names(cds) %in% txid]

library(Gviz)
tx <- rep(names(brca1cds), lengths(brca1cds))
id <- unlist(brca1cds)$cds_id
grt <- GeneRegionTrack(brca1cds, name="BRCA1", id=tx,
  gene="BRCA1", feature=tx, transcript=tx, exon=id)
plotTracks(list(GenomeAxisTrack(), grt))
```

Example: Visualize BRCA1 Transcripts



*BSgenome.** Packages: Whole-Genome Sequences

- ‘Masks’ when available, e.g., repeat regions
- Load chromosomes, range-based queries: `getSeq`, `extactTranscriptSeqs`

```
library(BSgenome.Hsapiens.UCSC.hg19)
extractTranscriptSeqs(Hsapiens, brca1cds)

##   A DNAStringSet instance of length 20
##      width seq                                          names
##  [1]  2280 ATGGATTTATCTG...AGCCACTACTGA uc010whl.2
##  [2]  5379 ATGAGCCTACAAG...AGCCACTACTGA uc002icp.4
##  [3]   522 ATGGATGCTGAGT...AGCCACTACTGA uc010whm.2
##  ...   ... ...
## [18]  3954 ATGCTGAAACTTC...GATTCAAACTTA uc010cyz.2
## [19]  4017 ATGGATTTATCTG...GATTCAAACTTA uc010cza.2
## [20]  3207 ATGAATGTAGAAA...GATTCAAACTTA uc010wht.1
```

Web-based resources

AnnotationHub	Ensembl, Encode, dbSNP, UCSC data objects, . . .
biomaRt	Ensembl and other annotations, url
PSICQUIC	Protein interactions, url
uniprot.ws	Protein annotations, url
KEGGREST	KEGG pathways, url
SRAdb	Sequencing experiments, url
rtracklayer	genome tracks, url
GEOquery	Array and other data, url
ArrayExpress	Array and other data, url

Web-based resources

Demo

Summary

Genes

- *org.** packages, `columns()`, `keys()`, `mapIds()`, `select()`.

Genomes

- *TxDb.** packages. `select()`, `exons()`, `exonsBy()` & friends.
- *BSgenome.** packages. `FaFile`, `TwoBitFile` files.

Variants

- *VariantAnnotation*, *VariantFiltering*, *ensemblVEP*.

Web-based resources

- *biomaRt*, *AnnotationHub*, and others.

Acknowledgments

- Core: Valerie Obenchain, Hervé Pagès, (Dan Tenenbaum), Lori Shepherd, Marcel Ramos, Yubo Cheng.
- The research reported in this presentation was supported by the National Cancer Institute and the National Human Genome Research Institute of the National Institutes of Health under Award numbers U24CA180996 and U41HG004059. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

https://bioconductor.org,
https://support.bioconductor.org