

Annotating genes, genomes, and variants

Martin Morgan (mtmorgan@fredhutch.org)
Fred Hutchinson Cancer Research Center
Seattle, WA, USA

18 June, 2015

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 definitions (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

```
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 organisms / 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, transcript, etc.

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

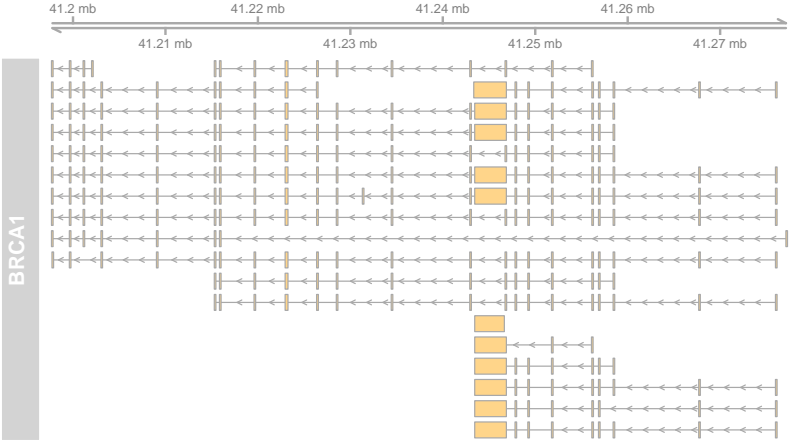
Example: Visualize BRCA1 Transcripts

```
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), elementLengths(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`, `extractTranscriptSeqs`

```
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...GATTCAAACCTTA uc010cyz.2
## [19] 4017 ATGGATTTATCTG...GATTCAAACCTTA uc010cza.2
## [20] 3207 ATGAATGTAGAAA...GATTCAAACCTTA uc010wht.1
```

Web-based resources

<i>AnnotationHub</i>	Ensembl, Encode, dbSNP, UCSC data objects, ...
<i>biomaRt</i>	Ensembl and other annotations, url
<i>PSICQUIC</i>	Protein interactions, url
<i>uniprot.ws</i>	Protein annotations, url
<i>KEGGREST</i>	KEGG pathways, url
<i>SRAdb</i>	Sequencing experiments, url
<i>rtracklayer</i>	genome tracks, url
<i>GEOquery</i>	Array and other data, url
<i>ArrayExpress</i>	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 (Seattle): **Sonali Arora**, Marc Carlson, Nate Hayden, Valerie Obenchain, Hervé Pagès, Paul Shannon, Dan Tenenbaum.
- ▶ 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, and the National Science Foundation under Award number 1247813. 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.

BioC 2015 Annual Conference, Seattle, WA, 20-22 July.