Annotation

Martin T. Morgan\textsuperscript{1}

27-28 February 2014
What is ‘Annotation’?

- Genes – classification schemes (e.g., Entrez, Ensembl), pathway membership, . . .
- Genomes – reference genomes; exons, transcripts, coding sequence; coding consequences

Other definitions (not covered here): assigning function to novel sequence assemblies, . . .
Model organism annotation packages

- `org.*` – gene names and pathways
- `TxDB.*` – gene models
- `BSgenome.*` – whole-genome sequences
The ‘select’ interface:

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

```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")
```
Within-vector or `data.frame`

- Finding and removing duplicates: `duplicated`, `unique`
- `any`, `all`

Between-vector or `data.frame`

- Matching `%in%`, `match`
- Set operations: `setdiff`, `union`, `intersect`
- `merge` Join two `data.frames` based on shared column.
SQL (sqlite) data bases

- `org.Hs,eg_dbconn()` to query using _RSQLite_ package
- `org.Hs,eg_dbfile()` to discover location and query outside _R_.

* pacakges – Under the hood...
**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.

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

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

```r
library(BSgenome.Hsapiens.UCSC.hg19)
library(GenomicFeatures)
dna <- extractTranscriptsFromGenome(Hsapiens, cdsByTx)
```
Bioconductor Annotation Resources – Web-based

Rich web resources

- `biomaRt (http://biomart.org), rtracklayer (UCSC genome browser)`
- `ArrayExpress, GEOquery, BiocpkgSRAdb`
- `PSICQUIC, KEGGREST, uniprot.ws, . . .`
- `AnnotationHub`
biomaRt

- http://biomart.org
- Drill-down discovery: listMarts, listDatasets, listFilters, listAttributes
- Retrieval: getBM

```r
library(biomaRt)
ensembl <-
  useMart("ensembl", dataset="hsapiens_gene_ensembl")
head(listFilters(ensembl), 3)
myFilter <- "chromosome_name"
myValues <- c("21", "22")
myAttributes <- c("ensembl_gene_id","chromosome_name")
res <-
  getBM(attributes=myAttributes, filters=myFilter, values=myValues, mart=ensembl)
```
PSICQUIC

- **Proteomics Standard Initiative Common QUery Interface**
- Programmatic access to molecular interaction data bases.
- [https://code.google.com/p/psicquic/](https://code.google.com/p/psicquic/)

```r
library(PSICQUIC)
## Query web service for available providers
psicquic <- PSICQUIC()

providers(psicquic)  # 25 available providers

## Interactions between TP53 and MYC

tbl <-
    interactions(psicquic, c("TP53", "MYC"), "9606")

nrow(tbl)  # 7 interactions

See the package vignette for additional detail.
```
AnnotationHub

- Large-scale genome resources, lightly curated for easy access from R.
- Supports tab-completion, metadata discovery, selection and filtering.

```r
library(AnnotationHub)
hub <- AnnotationHub()
hub # 10511 resources
```
Conclusions

Rich annotation resources

- Model organism and custom org.*, TxDb.*, BSgenome.* packages
- Web-based access to public (e.g., biomaRt and Bioconductor-specific (e.g., AnnotationHub) resources