

# Annotation

Martin Morgan ([mtmorgan@fhcrc.org](mailto:mtmorgan@fhcrc.org))  
Fred Hutchinson Cancer Research Center  
Seattle, WA, USA

June 24, 2014

# 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, ...

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

# 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

```
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.\* packages – Useful R commands

### 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.

## *org.\** 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")
```

## BSgenome.\* packages

### Whole-genome sequences

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

```
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`

```
library(biomaRt)
ensembl <- ## discover & use
  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/>

```
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.

```
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