# Annotation and down-stream analysis

Martin Morgan[1]
Fred Hutchinson Cancer Research Institute, Seattle, WA

June 27-July 1, 2011

[1] mtmorgan@fhcrc.org

# *AnnotationDbi*

The *org.** packages

- Curated data base of model organism annotations, e.g., *org.***Dm***.eg.db* annotates *Drosophila melanogaster*
- Gene-centric

Bimaps of 'Lkeys' and 'Rkeys' (values)

- Each package has a central 'Lkey': *org.Dm.***eg***.db* uses **e**ntrez **g**ene identifiers as the Lkey
- Each bimap describes the mapping between the Lkey and its Rkey / value. E.g., `org.Hs.egENSEMBL` maps between Entrez and Ensembl gene identifiers

Metadata describing the content, e.g., `org.Dm.eg()` and `?org.Dm.egENSEMBL`

# *AnnotationDbi*: how it works

Loading / available maps

- `library(org.Dm.eg.db)`
- `ls("package:org.Dm.eg.db")`

Common operations

- Subset `[`; subset-extract `[[`
- Interrogation: `mappedLkeys`, `mappedRkeys`
- Coercion: `toTable` (data frame), `as.list` (named list)
- Reverse mapping: `revmap`

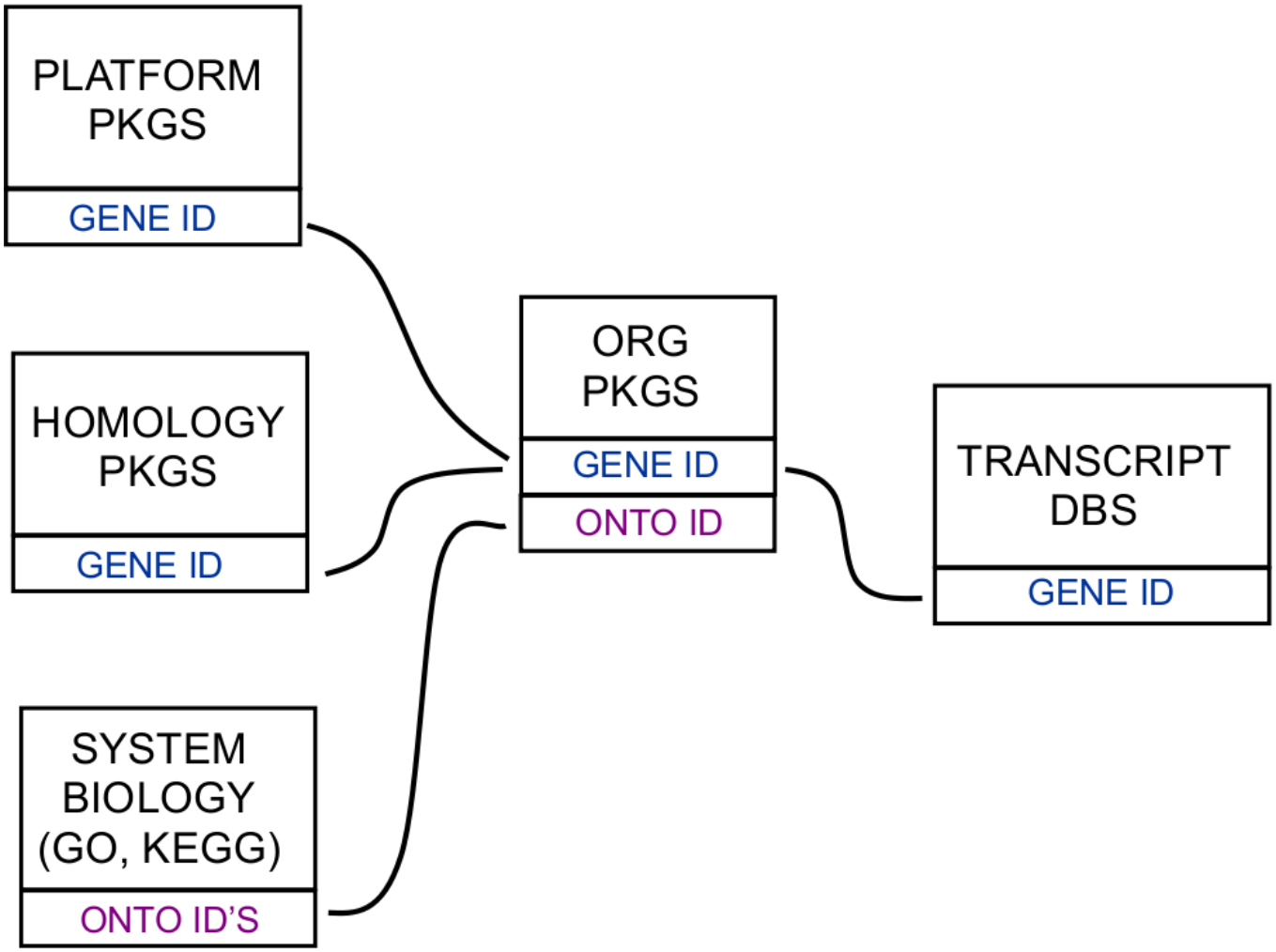
# *AnnotationDbi*

Other *AnnotationDbi* packages

- Pathways: KEGG, GO
- Homology
- Microarray

See `http://bioconductor.org/packages/release/data/annotation/`

# Under the hood: SQLite



# Biomart

Biomarts

- Collection of data bases with common interface
- Explorable at `http://biomart.org`

*biomaRt*

- Discover: `listMarts`, `listDatasets`, `listFilters`, `listAttributes`
- Select: `useMart`, `useDataset`, ...
- Retrieval: `getBM`

*AnnotationDbi* or *biomaRt*?

- Current, stable, versioned versus up-to-the-minute, extensive, whims of the internet

# UCSC

Via *rtracklayer*

- `import` and `export` common formats, e.g., `bed`, `wig`, from / to *GRanges* instances
- Start a browser session: `session <- browserSession("UCSC")`
- Lay a track: `track(session, "targets") <- targetTrack`
- Retrieve a track: `ensGene <- track(session, "ensGene")`
- See `browseVignettes("rtracklayer")`

Via *GenomicFeatures*

- Later in presentation

# GEO, ArrayExpress

- Previous experiments as very rich source of data

e.g., *GEOquery*

- Search & retrieve
- End result: *ExpressionSet*, a standard *Bioconductor* representation of a microarray experiment

# *GenomicFeatures*

- Structural information about genes: exon, transcript, coding sequence coordinates
- Uses *GenomicRanges*, so fits well with sequence analysis tools
- Created by querying, e.g., UCSC for *ensGene* track
- Saved as SQLite data bases
- 'Forge' to create packages, e.g., to share in a working group