biomaRt

BioC Developers’ Forum - 15/08/2019
Mike Smith
@grimbough
BioMart & biomaRt
BioMart

- BioMart & Ensembl
BioMart

- BioMart ≠ Ensembl
BioMart

- BioMart ≠ Ensembl
- Vast majority of the time BioMart = Ensembl
- Existing databases are either:
  - Dieing
  - Being absorbed into Ensembl e.g. COSMIC
  - Moving to other platforms e.g. Intermine
biomaRt

- R package for querying BioMart instance programmatically
  - Originally developed by Steffen Durinck in 2005
- Split between Ensembl specific & generic functionality
- 118 BioC packages list biomaRt in DESCRIPTION
biomaRt is still used

BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.
S Durinck, et al - Bioinformatics, 2005

Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt
Recent Issues

- Increasing server instability since Ensembl 96 release (Apr 2019)
  - Queries fail with ‘Error 500’
  - Larger queries fail more frequently - user has to resume from start
  - Mirror sites generally more responsive
- Unexpect renaming of attributes e.g. entrezgene to entrezgene_id
- These affect downstream packages too
Package changes

- Queries are broken down into batches, results of each batch written to tempdir
  - Re-running query picks up where it left off
  - Not persistent

- Results of complete queries cached using BiocFileCache
  - Identified via hash based on query parameters / function arguments

- Re-enable automatic mirror redirection
  - If a host is specified this is obeyed

- Doesn’t solve build failures - will the cache remain between builds?
Conversations with Ensembl

- Problems affecting web interface too
- Infrastructure struggling with scaling up number of genomes
- BioMart is effectively dead - no updates here
- ‘Massive’ R queries cause issues
  - IPs banned
  - Unclear to me how frequent this is vs other scaling problems
- Replacement is ‘coming’ - Spring 2020
  - I’ve been hearing this for a long time....

- Hopefully will know about variable renaming before release time!
Suggestions for the mid-term

● Impose hard or soft limits on query size?
● Actively promote other resources?
  ○ ensembldb, org.*.db packages, Ensembl resources e.g. REST API, VEP, Bulk download
  ○ How/Where? Messages, vignette?
● Use other resources ‘under the hood’?
● Develop something new for most common use cases?