Annotation

Mikhail Dozmorov
Fall 2016

Gene identifiers

Gene

- **Ensembl** ENSG00000139618
- **Entrez** Gene 675
- **Unigene** Hs.34012

RNA transcript

- **GenBank** BC026160.1
- **RefSeq** NM_000059
- **Ensembl** ENST00000380152

ID cross-mapping

- There are many IDs
- Software tools recognize only a handful
- Humans better recognize gene names

ID challenges

- Avoid errors: map IDs correctly
  - Beware of 1-to-many mappings
- Gene name ambiguity – not a good ID
  - e.g. FLJ92943, LFS1, TRP53, p53
  - Better to use the standard gene symbol, not aliases: TP53
- Excel error-introduction
  - OCT4 is changed to October-4 (open file/paste as text)
- Problems reaching 100% cross-mapping
  - E.g. due to version issues
  - Use multiple sources to increase coverage

BiomaRt

http://www.biomart.org/

https://bioconductor.org/packages/release/bioc/html/biomaRt.html

BiomaRt

The `getBM()` function has three arguments that need to be introduced: `filters`, `attributes` and `values`.

- `Filters` define a restriction on the query. Tell BiomaRt what kind of IDs do you have, so it will look for it. The `listFilters()` function shows you all available filters in the selected dataset.
- `Attributes` define the values we are interested in to retrieve. Which IDs associated with your IDs you want to get. The `listAttributes()` function displays all available attributes in the selected dataset.
- `Values` is a vector of IDs you want to convert

BiomaRt gotchas

- `host` is the database version. For gene ID conversion, use the latest database.

For genomic coordinates, use database that corresponds to genome assembly version you are interested in

Other options

Annotation data as R dataframes

R data package for annotating/converting Gene IDs

- `annotables` R package by Stephen Turner, https://github.com/stephenturner/annotables

http://www.gettinggeneticsdone.com/2015/11/annotables-convert-gene-ids.html