

Annotation

Mikhail Dozmorov
Fall 2017

Gene identifiers

Gene

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

RNA transcript

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

NCBI Gene

<https://www.ncbi.nlm.nih.gov/gene>



The screenshot shows the top section of the NCBI Gene website. At the top, there is a search bar with the word "Gene" on the left, a dropdown menu set to "Gene", a search input field, and a "Search" button. Below the search bar is a dark blue banner with a microscopic image of cells on the left and the word "Gene" in white on the right. Underneath the banner, there is a paragraph of text: "Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotype and links to genome-, phenotype-, and locus-specific resources worldwide." Below this banner are three columns of navigation links: "Using Gene" with a link to "Gene Quick Start", "Gene Tools" with a link to "Submit GeneRIFs", and "Other Resources" with a link to "HomoloGene".

3/10

GeneCards

<http://www.genecards.org/>



The screenshot shows the top section of the GeneCards website. At the top, there is a navigation bar with several tabs: "GeneCardsSuite", "GeneCards" (which is highlighted), "MetaCards", "LifeMap Discovery", "PathCards", "TGex", and "VarElect". Below the navigation bar is a dark blue banner with the GeneCards logo (a stylized DNA double helix) and the text "GeneCards® HUMAN GENE DATABASE". To the right of the logo is a search bar with a "Keywords" dropdown menu and a "Search Term" input field. Below the banner is a navigation bar with several links: "Home", "User Guide", "Analysis Tools" (with a dropdown arrow), "News And Views", and "About" (with a dropdown arrow). Below the navigation bar is a large orange banner with the text "GeneCards®: The Human Gene Database".

4/10

ID cross-mapping

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

5/10

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

6/10

BiomaRt

<http://www.biomart.org/>

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

7/10

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

8/10

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

9/10

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>

10/10