ONTOFORCE

Remote Data Subscriptions

Gene

MAIN OFFICE
Moutstraat 108
9000 Ghent
Belgium

Ontoforce.com
+32 9 396 80 07
info@ontoforce.com

Table of contents

RDS Data Sets

GENE

About the data set

The Gene data set offers annotated and reviewed species-specific genes. Additionally, it also contains data about cross-references between multiple genomic databases. Gene function, homology (between species) and genomic location data is also included.

Secondly, the data set offers annotated transcript data across different species.

Finally, the data set contains info about genes that are homologous across distinct species.

Integration strategy

- Three data sources contribute to the Transcript Data set: Ensembl, UCSC, NCBI Gene.
- For the Gene Data set, the following sources are contributing: Ensembl, Generif, NCBI Gene, HGNC and UCSC
- The Homology Data set consists of data provided by Homologene.
- The Ontology Data set consists of data provided by Gene Ontology (GO).

Quality Control

- For each potential mapping identifier in each source, a check is done to ensure that there are no internal overmappings:
    - Detection of overmapping is done by checking for every distinct source that the respective unique identifier is present only once
        - For the resources having non-unique IDs, the issue is investigated in the respective source pipeline.
    - This way, the Gene RDS Data set can not be built as long as non-unique ids are present.

Integrated data model

Four RDS data sets will be made available:

- Gene Data set:
    - Contains all gene resources from the listed data sources.
    - Ensembl as the main resource.
    - Contains:
        - Gene ids across multiple data sources

- Homology info
- Xrefs to other data sources
- Location in genome
- Gene family info
- Transcripts IDs
- Minimal genetic expression data and protein data
- Annotations about gene function
- Links to external pages dedicated to information on the gene and to genome browsers
- Nucleotide sequences associated with the gene.
- Proteins encoded by the gene in question
- Links to orthologs of the gene
- Links to associated phenotypes, diseases and mutations associated with the gene
- Data about publications in Pubmed about the gene

  - 2 kinds of Typed Links are created between the Genes and their GO annotations, ready for configuration:
    - Based on the functional aspects of the gene
      - I.e., "acts upstream of …", "involved in … ", "located in … ", etc.
    - Based on evidence (as seen in NCBI-Gene and Ensembl)
      - I.e., "Inferred from Experiment (EXP)", "Traceable Author Statement (TAS)", "No biological Data available (ND)", etc.
    - Some of the functional typed links indicate a NON-relation
      - I.e., "NOT located in …"
    - Example from NCBI-Gene:

Gene Ontology Provided by TAIR

Function	Evidence Code
enables RNA binding	IBA
enables mRNA binding	HDA
enables mRNA binding	IDA
enables structural constituent of ribosome	IBA

Process	Evidence Code
involved_in maturation of LSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA)	IBA
acts_upstream_of_or_within response to inorganic substance	IEA

Component	Evidence Code
NOT located_in Golgi apparatus	RCA
located_in chloroplast	HDA
located_in cytoplasm	ISM

- Transcript Data set:

  - Contains all transcript resources from the listed data sources.
  - UCSC as the main resource.
  - Contains:
    - Transcript genomic provenance
    - Transcript protein product
    - Xrefs to other data sources
    - Alternative identifiers
    - Corresponding mRNA sequence
    - Pathway info
    - Corresponding organism

- Homology Data set
  - Contains gene homologies between multiple species from the listed data sources
  - Homologene as the main resource
  - Contains:
    - Species and related genes
    - Their protein product
    - Conserved domains across species

- Ontology Data set
  - Contains gene-related Ontology information
  - GeneOntology as the main resource
  - Contains:
    - DNA sequence issue from transcription (including mRNA, tRNA...).
    - Link to the corresponding gene

RDS data sets made available in this package:

Data set	Instance Count	Disk Size	Contains	Run Time
Gene_Homology	*~45.000*	*~17 MB*	*The homologous gene for across of different species*	

Gene_Gene	*~44.773.000*	*~58 GB*	*The main gene instances*	
Gene_Ontology	*~44.000*	*~1.4 GB*	*The instances required to form the ontology/tree facets*	
Gene_Transcript	~66.200.000	*~15 GB*	*Transcript variant coding instances*	*1d01h*

Source Data Sets

UCSC

Description

The University of California Santa Cruz (UCSC) Genome Browser is a web-based tool serving as a multi-powered microscope that allows researchers to view all 23 chromosomes of the human genome at any scale from a full chromosome down to an individual nucleotide. The browser integrates the work of countless scientists in laboratories worldwide, including work generated at UCSC, in an interactive, graphical display. The Browser also affords access to the genomes of more than one hundred other organisms.

Download strategy

UCSC provides regular dumps of the database underlying the genome browser with folders divided by organisms.

Source update frequency

Frequent. Almost on a daily basis.

Download frequency

Every week on Friday a check is made for new data and downloaded if available.

Instances URI strategy

For genes were an Ensembl identifier is provided the URI is set as following:

http://www.ensembl.org/id/<ID>

- Example: http://www.ensembl.org/id/ENSG00000177519

When there is no Ensembl identifier but there is a Locus ID identifier, the NCBI gene URI form is adopted:

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

- Example: https://www.ncbi.nlm.nih.gov/gene/101059452

Original data model

UCSC provides a snapshot of the data underlying their genome browser.

For our RDS package, we partially ingest this data, and we publish only the Genes and the Transcript where an Ensembl identifier is provided plus some additional gene for which a unique NCBI Locus ID identifier is provided.

The full data contains the following files:

File Name	Comment
Imported files	
knowGene.tsv	These files are provided only for Homo Sapiens and Mus Musculus. Act as central resource for gene of these species
knowAttrs.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They complete the set of attributes of knowGene.tsv
knowToRefseq.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They provide a mapping between Ensembl transcript identifiers and NCBI RefSeq identifiers.
knownToEnsembl.tsv	These helper files are provided only for Homo Sapiens and Mus Musculus. They help understand what is currently mapped in GenCode V38.
kgXref.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They are references for transcripts like Gene Symbol, RefSeq ID, description of the gene, protein accession.
knownToPfam.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to Protein Families identifiers
wgEncodeGencodePubMedV38.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to PubMed ids.
kgAlias.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcript to a set of synonyms for the gene.
knownToMrna.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to GenBank ids.

knownToWikipedia.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to Wikipedia ids.
knownToKeggEntrez.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to Kegg/Entrez ids.
knownGeneMrna.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to Mrna Sequences.
bioCycPathway.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to BioCyc ids.
ensGene.tsv	These files are available for more organisms. They provide the main Ensembl transcript and gene identifiers.
ensemblToGeneName.tsv	These files are available for more organisms. They map Transcripts to Gene Name.
ensPep.tsv	These files are available for more organisms. They map Transcripts to their peptide sequence.
ensGtp.tsv	These files are available for more organisms. They map Transcripts to Gene and Proteins.
chromAlias.tsv	They provide the correspondence of UCSC chromosome names to refseq, genbank, and ensembl names
ucscScop.tsv	These files are provided only for Homo Sapiens and Mus Musculus. They map transcripts to SCOPe ids.
ncbiRefSeq.tsv	These files are available for more organisms. They provide the main resource for transcripts.
ncbiRefSeqLink.tsv	These files are available for more organisms. They provide attributes and links to genes for the transcripts.
keggMapDesc.tsv	Provides description for every Kegg id in the database.
scopDesc.tsv	Provides description for every SCOPe id in the database.

pfamDesc.tsv	Provides description for every Pfam id in the database.
bioCycMapDesc.tsv	Provides description for every ByoCic id in the database.
malacards.tsv	Provide malacards information by gene including the scores.
go_subset.tsv	Custom generated file including the GO ids associated with transcripts in the DB.
term.tsv	Terms and description for Gene Ontology ids
asmEquivalent.tsv	Correspondence table between UCSC organism databases name and RefSeq, Ensembl and Genbank assembly names.
databases.tsv	Custom generated file which helps generate external links from DISQOVER to UCSC genome browser.

Alignment efforts

- The many graphs as seen on the UCSC website are not easily reproducible in DISQOVER. Instead, every Transcript instance has URLs that link to their genome browser result based on chromosome position.
- Ensembl Transcript IDs (ENSTxxx), Ensembl Gene IDs (ENSGxxx), NCBI RefSeq IDs (NM_xxx, NP_xxx, XP_xxx) are the main identifiers used to link data and they are typically versioned, e.g., ENSMUST00000000010.9. Version number is stripped of when creating the instance URI in order to facilitate linking from sources using no or a differing version number.
- Labels are presented in the style *<Gene Name> [<Organism>]*:
    - Example: SLC39A1 [Homo sapiens]

ENSEMBL GENE

Description

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function, and collects disease data.

Download strategy

Ensembl provides regular dumps of the database underlying the genome browser with folders divided by organisms.

Source update frequency

The frequency of releases is on average every three months.

Download frequency

Every week on Friday a check is made for new data and downloaded if available.

Instances URI strategy

For gene: http://www.ensembl.org/id/<ID>

- Example: http://www.ensembl.org/id/ENSLSDG00000004508

For transcript: http://www.ensembl.org/id/<ID>

- Example: http://www.ensembl.org/id/ENSLSDT00000006935

Some Ensembl Transcript identifiers may sometimes correspond to the Gene identifier. About one hundred instances have been found in the data. In this case the gene URI is kept, but the transcript assumes this variant:

http://ns.ontoforce.com/instance/ensembl/transcript/<ID>

- Example:
    - Gene/Transcript identifier: YDR261W-A
    - Gene URI: http://www.ensembl.org/id/
    - Transcript URI: http://ns.ontoforce.com/instance/ensembl/transcript/

Original data model

For every organism, a json file is available with all the relevant data.

The full data contains the following files:

File Name	Comment
Imported files	
<organism_name>.json *(e.g. homo_sapiens.json)*	Main resource file containing the summarized data for each gene in the specified organism.

Data ingested comprises all organisms, but not all strains. For the following organisms: Canis Lupus Familiaris, Capra Hircus, Cricetulus Griseus, Cyprinus Carpio, Mus Musculus, Sus Scrofa. Only the reference strain was ingested.

For every organism, the following data is available in the corresponding file:

- Gene data:
    - Gene id
    - Homologue info
    - Xrefs to other data sources
    - Location in genome
    - Gene family
    - Transcripts
        - A gene can code for multiple transcripts
- Organism data
    - Name and aliases
    - Lineage
    - Species taxonomy

Alignment efforts

Ensembl contains numerous IDs to other data sources. Several of those are also part of the DISQOVER knowledge graph.

For these sources we provide linkouts to the respective websites. Additionally, we add the IDs as identifiers so that if a user uses one of these IDs in search, they can find the Ensembl instance. This allows a user to find exact matches when searching on ID's.

These IDs pointing to other data sources were also added as a label. The end result is 2 Canonical Types:

- Ensembl Genes
- Ensembl Transcripts

NCBI-GENE

Description

NCBI-Gene (f.k.a. Entrez) is the NCBI's database for gene-specific information, focusing on completely sequenced genomes, those with an active research community to contribute gene-specific information, or those that are scheduled for intense sequence analysis.

Download strategy

NCBI-Gene provides a main dump of its gene data across all of its listed organisms alongside a series of accompanying annotation files.

Source update frequency

Frequent. Almost on a daily basis.

Download frequency

Every week on Friday a check is made for new data and downloaded if available.

Instances URI strategy

http://www.ncbi.nlm.nih.gov/gene/<ID>

- Example: http://www.ncbi.nlm.nih.gov/gene/100006826

Original data model

NCBI-Gene provides a summary of Genetic information as well as displaying links to other sources (such as GeneRIF, Reactome, PubMed, etc.). They also provide minimal genetic expression data and protein data.

For our RDS package, we only ingest the summary and linkage data.

NCBI-Gene offers many alternative IDs for the given gene resources. The full list can be seen here. Out of these, a subset was selected to be turned into separate properties. The remaining IDs are conglomerated into a single property called “other_id”.

The selected IDs are: AnimalQTLdb, APHIDBASE, ApiDB_CryptoDB, Araport, BEEBASE, BEETLEBASE, BGD, CGNC, dictyBase, EcoGene, ENSEMBL, FLYBASE, HGNC, MGI, miRBase, NASONIABASE, MIM, RGD, SGD, TAIR, VectorBase, VGNC, WormBase, Xenbase and ZFIN.

The full data contains the following files:

File Name	Comment
Imported files	

gene_info.gz	Main resource file containing the summarized data for each gene.
gene2accession.gz	External accessions that are related to a GeneID. It includes sequences from the international sequence collaboration, Swiss-Prot, and RefSeq. The RefSeq subset of this file is also available as gene2refseq (Obsolete since we import this one).
gene2ensembl.gz	Matches between NCBI and Ensembl annotation based on comparison of rna and protein features.
gene2vega.gz	No longer being updated. The last update was on December 3, 2018. This file reports matches between NCBI and Vega annotation.
gene2go.gz	GO terms that have been associated with Genes in Entrez Gene.
gene2pubmed.gz	PubMed accessions that have been associated with Genes in Entrez Gene.
gene_group.gz	Genes and their relationships to other genes.
gene_orthologs.gz	Orthologous gene annotations.
gene_neighbors.gz	Neighboring gene annotations for all genes placed on a given genomic sequence.
mim2gene_medgen.gz	Relationship between MIM numbers (OMIM), GeneIDs, and Records in MedGen

Alignment efforts

No special alignment efforts were needed.

GENERIF

Description

A GeneRIF or Gene Reference Into Function is a short statement about the function of a gene. GeneRIFs provide a simple mechanism for allowing scientists to add to the functional annotation of genes described in the Entrez Gene database. In practice, function is constructed quite broadly.

Gene Reference Into Function is a gene annotation project of the National Library of Medicine, available through the Entrez Gene website of the National Center for Biotechnology Information. A GeneRIF is a short (425 characters or less) statement describing a function of a gene. Each GeneRIF is tagged with the Entrez Gene ID of the described gene and the PubMed ID of the reference supporting the asserted function.

Download strategy

The html from the https://ftp.ncbi.nih.gov/gene/GeneRIF/ is parsed and all links to files are retrieved.

Source update frequency

Frequent. Almost on a daily basis.

Download frequency

Every week on Friday a check is made for new data and downloaded if available.

Instances URI strategy

For genes ID:

http://www.ncbi.nlm.nih.gov/gene/<ID>

- Example: http://www.ncbi.nlm.nih.gov/gene/4355

Original data model

The creation of GeneRIF entries involves the identification of the genes mentioned in MEDLINE citations and the citation sentences describing a novel function.

For our RDS package, we partially ingest this data.

The full data contains the following files:

File Name	Comment
Imported files	
generifs_basic.gz	This file extracts the equivalent of the GRIF tag in the LL_tmpl file and reports text of the GeneRIF and the associated PubMed ids.

interactions.gz	These files are a compilation of interactions from BIND, HPRD, BioGRID and EcoCyc provided by NCBI.
Excluded files	
hiv_interactions.gz	Only human protein interaction data was extracted in these two files.
hiv_siRNA_interactions.gz	

Collection of protein and genetic interactions in different web pages:

Symbol: BIND

Web Page URL: http://www.bind.ca

Template URL: http://bind.ca/Action?idsearch=

Symbol: BioGRID

Web Page URL: http://thebiogrid.org/

Template URL: http://www.thebiogrid.org/SearchResults/summary/

Symbol: EcoCyc

Web Page URL: http://www.ecocyc.org

Template URL: http://biocyc.org/ecoli/new-image?object=

Symbol: HPRD

Web Page URL: http://www.hprd.org

Template URL: http://www.hprd.org/protein/

Alignment efforts

- The evidence data is agglomerated into JSON predicates and used to create subinstance.

HGNC

Description

The HGNC approves a unique and meaningful name for every known human gene, based on a query of experts. In addition to the name, which is usually 1 to 10 words long, the HGNC also assigns a symbol (a short group of characters) to every gene. As with an SI symbol, a gene symbol is like an abbreviation but is more than that, being a second unique name that can stand on its own just as much as a substitute for the longer name. It may not necessarily "stand for" the initials of the name, although many gene symbols do reflect that origin.

Source update frequency

The monthly files are produced on the 1st of every month while the quarterly files are produced on the 1st of Jan, Apr, Jul & Oct.

Download frequency

Every week on Friday a check is made for new data and downloaded if available.

Instances URI strategy

For genes and gene families:

http://identifiers.org/hgnc/<ID>

- Example: http://identifiers.org/hgnc/5

Original data model

HGNC contains info about:

- Genes
- Gene families.

For every gene, data is provided about the naming and symbols. They also provide many xrefs to other databases for every gene:

- Gene resources: links to external pages dedicated to information on the gene and to genome browsers.
- Nucleotide resources: a list of links to nucleotide sequences associated with the gene.
- Protein resources: information on proteins encoded by the gene in question. Links are made via UniProt protein accessions.
- Orthologs: links to orthologs of the gene in selected species.
- Specialist resources: if the gene in question is listed in an external database which is specific to certain classes of genes.

- Clinical resources: provides links to associated phenotypes, diseases and mutations associated with the gene.
- Other resources: Links to other external resources that provide useful information on the gene.
- Publication resources: Displays the title, (first) author, journal information and links to PubMed and Europe PubMed Central.

For the gene families, the following info is provided:

- Genes contained within the family
- Their symbols, names, and aliases
- Their location in the genome
- A description

The full data contains the following files:

File Name	Comment
Imported files	
hgnc_complete_set.txt	This file contains all the info on a gene level, as outlined above.
genefamilies.txt	This file contains all info about the gene families, as outlined above.

Alignment efforts

- Using the IDs in the cross-references to other data sources, labels were added as extra synonyms next to the HGNC Id
- These ids were also used to create uris, so that these uris can be used to map onto other data sources available in DISQOVER federation.

NCBI HOMOLOGENE

Description

HomoloGene, a tool of NCBI, is a system for automated detection of homologs (similarity attributable to descent from a common ancestor) among the annotated genes of several completely sequenced eukaryotic genomes.

The HomoloGene processing consists of the protein analysis from the input organisms. Sequences are compared using blastp, then matched up and put into groups, using a taxonomic tree built from sequence similarity, where closer related organisms are matched up first, and then further organisms are added to the tree. The protein alignments are mapped back to their corresponding DNA sequences, and then distance metrics as molecular distances can be calculated.

Source update frequency

The last update of the data source was in 2014, no new updates are planned.

Download frequency

Not applicable.

Instances URI strategy

For Homology ID:

http://www.ncbi.nlm.nih.gov//homologene/<ID>

- Example: http://identifiers.org/homologene/100065

Original data model

The creation of homologene entries involves the identification of the homologene ID, the highest taxonomy ID, the related organism to taxonomy ID and all the related gene and protein IDs to one homologene ID.

For our RDS package, we partially ingest this data.

The full data contains the following files:

File Name	Comment
Imported files	
homologene.xml.gz	homologene.xml.gz is a compressed file that contains a complete XML version of the HomoloGene build and includes the information available on the public webpage.

The HomoloGene is linked to all Entrez databases and based on homology and phenotype information of these links:

- Mouse Genome Informatics (MGI)
- Zebrafish Information Network (ZFIN)

- Saccharomyces Genome Database (SGD)
- Clusters of Orthologous Groups (COG)
- FlyBase
- Online Mendelian Inheritance in Man (OMIM)

As a result, HomoloGene displays information about Genes, Proteins, Phenotypes, and Conserved Domains.

Alignment efforts

- The organism's name is added to the homologene gene.
- Homologene label is a combination of:
    - Homologene ID
    - Info about gene conserved in which species

Integration Diagram