EntrezGene – Thesaurus and Concept generation

This initial activity for EntrezGene is to ensure all relevant gene concepts are added to the thesaurus and additional synonyms are added to new and existing Gene concepts.

September 2014: Two changes:

1. rules to include genes has changed

2. for the new genes, a mapping table to organism name is extended.

\\end of 2014 change overview.

June 2015 overview: two changes

1. Include additional TaxonomyID to the taxonomyID list: {559292}
2. for new genes, the mapping table extended with name: "Saccharomyces cerevisiae s288c"

September 2016: Rules to include genes - additional taxonomies and mapping table, see page

GeneInfo is a tab-delimited file, with 1 line per entry. It is released/updated daily.

File location: ftp://ftp.ncbi.nih.gov/gene//DATA/gene_info.gz

It is a very large file with over 24 million genes ; however we only need to load the genes with a status "official" (o) or "interim" (i) , which amount to around 450k gene concepts. The majority of the genes is expected to be present in Solr.

CHANGE SEPTEMBER 2014: We load the genes based on the following rule:

- Status = " official" (o) OR "interim" (i),

OR

- TaxonomyID = {3702, 39947, 39946, 3708, 4577, 4113, 4006, 3983, 4565, 4081, 3712, 3711, 3707,

71323, 4513, 4558, 559292}

This will add 150k entries and result in 600k gene concepts. \\end change september 2014 nr 1.

June 2015: This will add 6385 genes \\ end change June 2015

What we need to explain high level;

1) *How to find the subject term and associated measures*
   Only concepts, names and synonyms, and classification measures.

2) *How to find the object term and associated measures (if applicable)*
   Not applicable

3) *How to find the triple mapping and associated measures (if applicable)*
   Not applicable

4) *How to find the publication and associated measures (if applicable)*
   Not applicable

5) *The good and bad weather conditions for each step above*

In detail we expect the following information (without ambiguity);

1) *Where can we find the data source(s)?*

File location: ftp://ftp.ncbi.nih.gov/gene//DATA/gene_info.gz

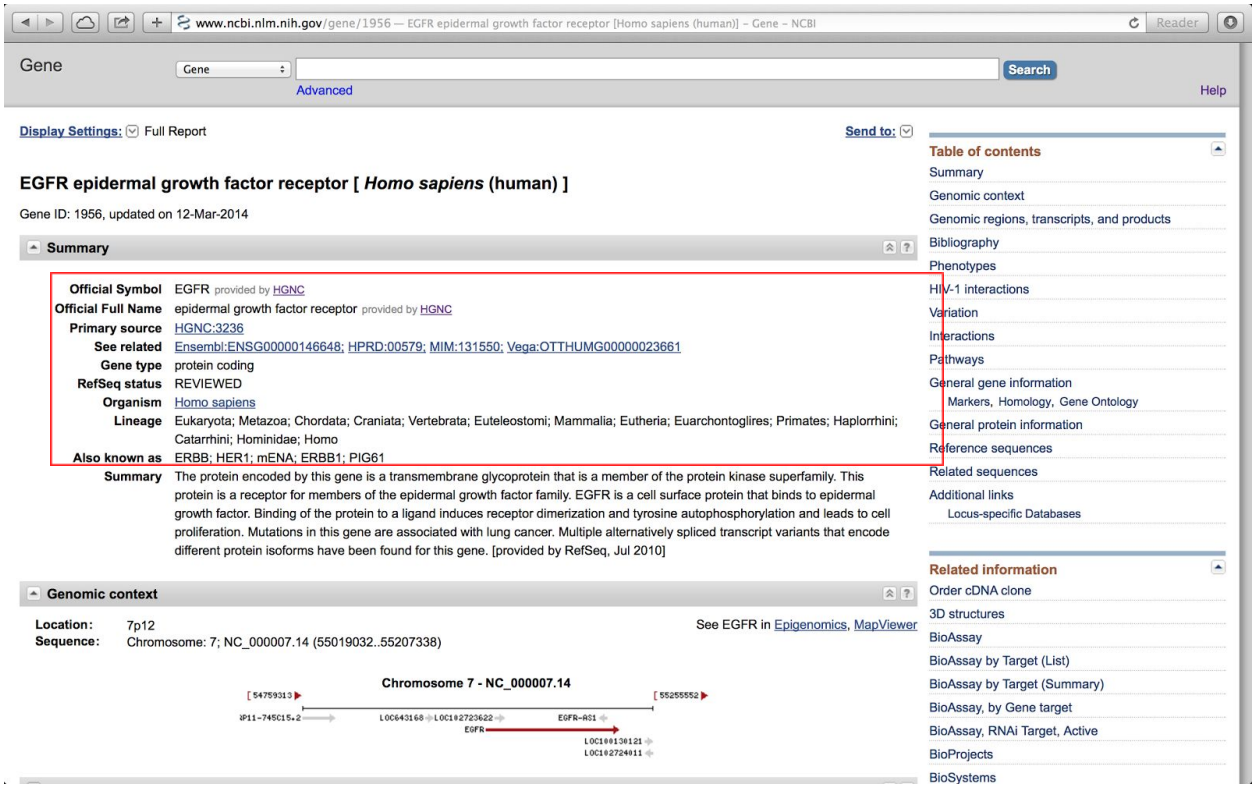
2) *Which tools / urls are available to validate / test content?*

http://www.ncbi.nlm.nih.gov/gene

3) *Which example screen dumps can we share to visualize a subject, object, predicate, publication? Can we validate against an example screen dump (perhaps based on tools/urls)?*

Example EGFR ; gene ID 1956

Information in this step includes summary part (red bracket).



4) *Are data source schema's available? And if so, where can they be found? Is it possible to mark the fields we need to extract /use?*

Location of the schema:

ftp://ftp.ncbi.nih.gov/gene//README

It is a tab delimited file, with some fileds containing multiple values separated by "|".

Gene_info: recalculated daily

---

tab-delimited ; one line per GeneID; Column header line is the first line in the file.

Key	Opt/Mand	Use	When absent	Value type	Modifications
tax_id	M	Use to filter cases to be processed or not ; AND identify organism name; append as suffix to	Do not process entry	Number ; single value	Map to organism name (string)

		create Preferred Name			
Gene_id	M	Add as Synonym to Solr	Do not process entry	Number ; single value	Add prefix [geneid] ; kb= geneid
Symbol	M	Synonym	Continue	String, single value	Add (organisms name) as suffix
LocusTag	O	Synonym	Continue	String, single value	
Synonyms	O	Synonym	Continue	String; multiple values separated by "|"	Add (organism name) as suffix
dbXrefs	O	Synonym	Continue	<Key:value> pairs, separated by "|"	Add prefix [ ] ; Solr kb= key (dbname)
Chromosome	n/a	donot process	n/a		
map location	n/a	donot process	n/a		
description	O	definition	Continue	String	None
type of gene	O	measure/classification	Continue	String, single value	None
Symbol from nomenclature authority	O	Preferred Name	Do not process entry	String, single value	Add (organism name) as suffix
Full name from nomenclature authority	O	Synonym	Continue	String, single value	None
Nomenclatus status	M	Filter to include entry, only include when "I" or "O" OR taxID part of list	Do not process entry	String, single value. Possible values {-, I, O}	None
Other designations	n/a	Donot process	n/a		
Modification date	n/a	Donot process	n/a		

EXAMPLES

tax_id:

the unique identifier provided by NCBI Taxonomy for the species.

MANDATORY

Use this value to identify the organism name, and append this organism name {eg (homo sapiens) } to the Symbol to create the unique gene name.

Mapping:

the unique identifier provided by NCBI Taxonomy for the species. This identified the organism source for the gene. Example: “9606” for “homo sapiens”. The table below translates the tax_id to the organism name:

taxID

	value	Frequency	Percent	Valid Percent	Cumulative Percent	ORGANISM NAME
	4577	4091	,9	,9	,9	zea mays
	7029	8873	2,0	2,0	2,9	acyrthosiphon pisum
	7070	670	,1	,1	3,1	tribolium castaneum
	7217	15978	3,6	3,6	6,6	drosophila ananassae
	7220	15810	3,5	3,5	10,2	drosophila erecta
	7222	15585	3,5	3,5	13,7	drosophila grimshawi
	7227	15874	3,6	3,6	17,2	drosophila melanogaster
	7230	15179	3,4	3,4	20,6	drosophila mojavensis
	7234	17573	3,9	3,9	24,5	drosophila persimilis
	7238	17286	3,9	3,9	28,4	drosophila sechellia
	7240	16117	3,6	3,6	32,0	drosophila simulans
	7244	15343	3,4	3,4	35,4	drosophila virilis
	7245	16904	3,8	3,8	39,2	drosophila yakuba
	7260	16385	3,7	3,7	42,9	drosophila willistoni
	7425	8792	2,0	2,0	44,9	nasonia vitripennis
	7460	9344	2,1	2,1	46,9	apis mellifera
	7955	24021	5,4	5,4	52,3	danio rerio
	8355	9651	2,2	2,2	54,5	xenopus laevis
	8364	9511	2,1	2,1	56,6	xenopus (silurana) tropicalis
	9606	34418	7,7	7,7	67,9	homo sapiens
	9913	15389	3,4	3,4	71,3	bos taurus
	10090	55998	12,5	12,5	83,9	mus musculus
	10116	55373	12,4	12,4	96,3	rattus norvegicus
	46245	16756	3,7	3,7	100,0	drosophila pseudoobscura pseudoobscura

September 2014 / June 2015 ADDITIONAL Mapping values organism - Tax ID

organism name	Tax ID value	nr of genes	gene example	gene synonym id example
arabidopsis thaliana	3702	33584	ORC4	AT2G01120
oryza sativa japonica	39947	30535	ndhA	OrsajCp093
oryza sativa indica	39946	161	mat-r	OrsaiPp41
brassica napus	3708	238	clpP	BRNAC_p047
zea mays	4577	26182	hp2	ZEAMMB73_076549
solanum tuberosum	4113	29032	psaJ	SotuCp041
linum usitatissimum	4006	-	-	-
manihot esculenta	3983	131	rpoC2	MaesCp011
triticum aestivum	4565	2218	War7.2	-
solanum lycopersicum	4081	27067	sos1	SISOS1
brassica oleracea	3712	98	tatC	BroleMp019
brassica rapa	3711	-	-	
brassica juncea	3707	99	trnP	BrjunMt002
camelina	71323	-	-	
hordeum vulgare	4513	476	DRF2	HvDRF2

sorghum bicolor (vulgare)	4558	33081	rpl16	SobioMp31
Saccharomyces cerevisiae s288c	559292	6385	RKM4	SGD:S000002665

\\End September 2104 \June 2015 additional mapping values.

September 2016: ADDITIONAL Mapping values organism - Tax ID

organism name	Tax ID value
caenorhabditis elegans	6239
escherichia coli	562
aspergillus niger	5061
clostridium acetobutylicum	1488
sus scrofa	9823
bos indicus	9915
capra hircus	9925
ovis gries	9940
Gallus gallus	9031
brachypodium distachyon	*15368*

triticum aestivum	*4565*
solanum lycopersicum var. cerasiforme	*195583*
cucumis sativus var. sativus	869827
brassica oleracea var. oleracea	*109376*
solanum melongena	4111
capsicum annuum	4072
capsicum chinense	80379
lactobacillus plantarum WCFS1	220668
bifidobacterium longum subsp. infantis	565040
bifidobacterium longum subsp. longum JCM 1217	565042
bifidobacterium longum subsp. infantis JCM 1222	391904
lactobacillus reuteri SD2112	491077
lactobacillus reuteri JCM 1112	557433
lactobacillus fermentum IFO 3956	334390

the organism name will be used together with the “symbol from nomenclature authority” or “symbol” to search for a term in Solr; For the lookup, always use ST=28 .

Example:
taxID = “10090” Symbol: “EGFR”, then lookup: “egfr (mus musculus)” & semtype=28

GeneID:

the unique numeric database identifier for a gene in ENTREZGENE. Example : 145270 MANDATORY

When adding to Solr, append prefix: [geneid]
Example key, values:

term, [geneid]145270

knowledgebase, geneid

source, entrezgene

Use this field to identify the proper Mongo entry to identify the concept details.

Symbol:

the default symbol (name) for the gene. Example: PRIMA1 Optional

When adding to Solr, append the organism name as mapped from tax id, to create the Solr entry.
Example key,value:

Symbol, prima1 (homo sapiens)

LocusTag:

the LocusTag value is a database reference of the supplier of the sequence provider. Example: hCG_2028654. For the purpose of Brain, this can be considered a synonym in Solr. Note. The file may contain "-" values – this is equivalent of <empty> ; ie no value. OPTIONAL

Synonyms:

bar-delimited set of unofficial symbols for the gene, for example: LCA13|LCA3|RP53|SDR7C2. The file may contain "-" values – this is equivalent of <empty> ; ie no value. OPTIONAL

dbXrefs:

bar-delimited set of identifiers (key,values pairs) in other databases for this gene, example: HGNC:20097|ENSEMBL:ENSG00000165807|HPRD:08505|Vega:OTTHUMG00000141316. The file may contain "-" values – this is equivalent of <empty> ; ie no value. This can be considered as synonyms for the genes. OPTIONAL

chromosome:

donot process

map location:

donot process

description:

a descriptive name for this gene. This is can be seen as input for the "Definition" for the gene concept. OPTIONAL

type of gene:

The type assigned to the gene; this is used as categorization value for the gene. OPTIONAL
Can be one of the following values:

- unknown
- tRNA
- rRNA
- snRNA
- scRNA
- snoRNA
- protein-coding
- pseudo
- transposon
- miscRNA
- ncRNA
- other

Symbol from nomenclature authority:

when not '-', indicates that this symbol is from a nomenclature authority. This is the Official symbol/name.

THIS IS USED TO CREATE THE PREFERRED NAME FOR THE GENE.

To create the entry, append the (organism name), as mapped from the TAXid to the symbol.

OPTIONAL ; when not present

Full name from nomenclature authority:

when not '-', indicates that this full name is from the nomenclature authority. OPTIONAL

Nomenclature status:

when not '-', indicates the status of the name from the

nomenclature authority (O for official, I for interim). ONLY PROCESS WHEN value = O or I. "-" values are not processed.
MANDATORY

Other designations:

Donot process

Modification date:

the last date a gene record was updated, in YYYYMMDD format.

Donot process

5) *Is the data source an authority?*

Yes, this is the authority for all genes.

a. *If so, do we expect SOLR create events? (new UUID)*

Yes, New UUID's can be created

b. *If so, do we expect SOLR update events? (additional synonym to existing UUID)*

Yes, new synonyms can be expected.

6) *What is the new element (subject, object, predicate, triple, publication) to be introduced?*

This activity does not include triples.

7) *For each data source reference (key / field) we need to describe if they are mandatory or optional.*

a. *And what is the expected action in case a mandatory reference is not available?*

DONOT PROCESS THE ENTRY

b. *And what is the expected action in case an optional reference is not available?*

PROCESS THE ENTRY, but donot include this key

8) *What makes a subject concept complete?*

GeneID and a preferred name

a. *And what is the expected action in case a subject concept is not complete?*

Concept Generation

Trigger for concept generation

Source	Knowledgebase	Semantic Type	Semantic Group	Preferred Y/N
Entrezgene	geneid	St=28		N

Identify Source-Mongo document and KEY

Solr Criteria	Mongo Document Key	Concept attribute	Value type	Error handling
St=28; source=entrezgene kb=geneid				

Mapping criteria

Measure / Classification	Source value	Concept target value
classification	typeofgene	Genetype

Access parameter

Source Mongo collection	RD	RT
entrezgene	0	57

Data Processing System