**Gene ID ontology & Mapping of gene IDs across different DBs**

**Introduction:**

Different gene catalogs use different identification schemes. In each catalog, every gene has an ID and a symbol. The gene ID is typically a unique identifier within the catalog while the gene symbol is an abbreviation for the gene name. In most gene catalogs, gene symbols are not guaranteed to be unique, however they are the commonly used handles for gene reference in literature and study summaries. In addition to gene symbols ambiguity, mapping of gene IDs across different catalogs has more obstacles. Different catalogs are built on different gene annotations and thus several genes in one catalog might not exist or partially overlap with genes in another catalog. Several catalogs try to maintain cross-references of their gene IDs to the those in other resources. However, discrepancies of the cross-referencing information between different catalogs add more confusion to the topic.

As an example, these are gene IDs for the ACE2 gene in different gene catalogs:

- HGNC id (e.g. HGNC:13557)

- NCBI Gene ID (Entrez Gene) (e.g. 59272)

- Ensemble/Gencode gene ID (e.g. ENSG00000130234)

- refseq\_accession (e.g. NM\_001371415)

- vega\_id (e.g. OTTHUMG00000021177)

- ucsc\_id (e.g. uc004cxb.2)

- European Nucleotide Archive (ENA) id (e.g. AF291820)

- OMIM: 300335

- UniProtKB: Q9BYF1

- PubMed ID: 10969042

- MGI:1917258

- RGD:728890

**Why do we have different gene catalogs?**

Each gene catalog has a different focus on one or more of gene features. For example:

- **HGNC** is responsible for approving unique symbols and names for human loci, including protein coding genes, ncRNA genes and pseudogenes, to allow unambiguous scientific communication.

- **GENCODE** project is responsible for annotation of gene structure and features. It is made by merging the manual gene annotation produced by the Ensembl-Havana team and the Ensembl-genebuild automated gene annotation. GENCODE gene sets are used by many other projects (eg. Genotype-Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), NIH Roadmap Epigenomics Mapping Consortium, Blueprint Epigenome Project, Exome Aggregation Consortium (EXAC), Genome Aggregation Database (gnomAD), 1000 Genomes Project and the Human Cell Atlas (HCA)) as reference gene sets.

- **The Reference Sequence (RefSeq)** collection provides a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

- **NCBI gene database** integrates information from a wide range of species. It supplies gene-specific connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

- **UniProt** provides a comprehensive, high-quality resource of protein sequence

- Mouse Genome Informatics (**MGI**) and Rat genome database (**RGD**) are examples for organism specific genomic databases

**Problems:**

1. Multiple genes in the same database might have the same gene symbol. The problem is much worse when we consider the alternative gene alias. Moreover, shuffling of symbols between genes add another hard-to-fix source of errors for meta-analysis studies. Some gene catalogs report the previous gene symbols but without clear versioning scheme, it hard to resolve ambiguities.

- HGNC has 43942 genes, all of them has unique symbols even it is among the 1761 withdrawn entries. However, there are 2120 gene alias or previous symbols that associate to more than one gene and/or match official gene symbol. For example, there are 11 Metallothionein genes have the same gene alias “MT1”. Also, there are 10 HOX genes sharing the alias HOX1 and another 9 HOX genes sharing the alias HOX2. There more confusing examples for current symbols which are alias and/or previous symbols for other genes. For example, “GALNT15” is a current approved gene symbol but alias for “GALNT18” and previous symbol for “GALNTL5”. Similarly, “HAP1” is a current symbol but alias for “APEX1” and previous symbol for “HAPP”.

- NCBI gene database (Entrez genes) has 61622 genes on August 30, 2020 but they have only 61563 unique gene symbols (There are 41 symbols assigned to 100 genes). Also, there are 4668 gene alias that associate to more than one gene and/or match official gene symbol. For example, there are 14 histone genes have the same gene alias “H4-16”. Also, there are 3 genes having the gene symbol "COX1" while another gene uses this symbol as an alternative alias. Third example for the repeated alias can be seen in HOX genes where there are 7 genes sharing the alias HOX3 and another 7 genes sharing the alias HOX4. To assess the effect of gene symbols change and/or discontinuation, searching the NCBI gene database was done to retrieves discontinued or replaced human Entrez IDs in the last 20 years using the search terms: “(homo Sapiens[Organism]) AND ("2000"[Date Discontinued] : "2020"[Date Discontinued])”. On August 30 2020, there were 136668 discontinued and 27178 replaced IDs. Usually with replaced ID, the new Entrez ID will be associated with new gene symbol e.g. Gene ID: 388289 (C16orf47) was replaced on 19-Jul-2019 with Gene ID: 463 (ZFHX3)

- Gencode has 60656 genes in version 35 but they have only 59609 unique gene symbols (There are 119 symbols assigned to 1166 genes). Gencode annotation files does have gene information about gene alias or previous gene symbols. To test the effect of gene ID and symbol change over time in Gencode, we compared the current version of Gencode annotation (v35) to version 7 from Dec, 2010. In 10 years, Gencode discontinued 7530 ID and added 17104 new IDs. The remaining 43552 ID survived but 22402 changed their symbols!

2. Different catalogs are built on different gene annotations: There is a significant non-overlapping between different gene catalogs with inconsistent cross-referencing

Gene content in different catalogs on 8/30/2020

* HGNC has 43942 gene IDs (42181 Approved, 1761 Entry Withdrawn)
* NCBI has 61622 active Entrez IDs
* Ensemble/Gencode(v35) has 60656 genes

Cross-referencing stats

* HGNC offers the most comprehensive conversion between different gene naming schemes. Approved HGNC IDs include 42110 mapping to Entrez IDs and 39199 mapping to Gencode. There are 39167 genes mapping to both IDs.
* According to NIH gene database, active IDs include 35063 mapping to Gencode ID and 42106 mapping to HGNC. There are 33165 genes mapping to both IDs.
* In Gencode v35 annotation, Gencode IDs include 25600 mapping to Entrez ID and 38596 mapping to HGNC. There are 24528 genes mapping to both IDs.

3. Discrepancies in gene mapping between different resources (cross-referencing)

Details of discrepancies between NCBI and HGNC (48 records in total)

* Entrez ids with HGNC ids and symbols in NCBI different from those in HGNC database:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Entrez\_id** | **ncbi-Symbol** | **ncbi-hgnc\_id** | **hgnc-hgnc\_id** | **hgnc-Symbol** |
| 6395 | SEA | HGNC:10695 |  |  |
| 107984570 | LINC00444 | HGNC:42781 |  |  |
| 100287084 | ZNF72P | HGNC:13142 |  |  |
| 100533637 | LOC100533637 |  | HGNC:13142 | ZNF72P |
| 100130620 | LOC100130620 |  | HGNC:55137 | TPM2P1 |
| 101929959 | LOC101929959 |  | HGNC:49159 | BMS1P14 |
| 728034 | BMS1P14 | HGNC:49159 | HGNC:55088 | BMS1P23 |

\*\* HGNC:10695 & HGNC:42781 do not exist in the HGNC database!

* Entrez ids with the same HGNC ids in NCBI and HGNC database but with different symbols: There are 37 mitochondrial genes e.g.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Entrez\_id** | **ncbi-Symbol** | **ncbi-hgnc\_id** | **hgnc-hgnc\_id** | **hgnc-Symbol** |
| 4508 | ATP6 | HGNC:7414 | HGNC:7414 | MT-ATP6 |
| 4509 | ATP8 | HGNC:7415 | HGNC:7415 | MT-ATP8 |
| 4511 | TRNC | HGNC:7477 | HGNC:7477 | MT-TC |
| 4512 | COX1 | HGNC:7419 | HGNC:7419 | MT-CO1 |
| 4513 | COX2 | HGNC:7421 | HGNC:7421 | MT-CO2 |
| 4514 | COX3 | HGNC:7422 | HGNC:7422 | MT-CO3 |
| 4519 | CYTB | HGNC:7427 | HGNC:7427 | MT-CYB |

* Entrez ids without HGNC ids in NCBI database but have HGNC ids in HGNC database: There are 4 entries in HGNC with their Entrez IDs discontinued from NCBI.

|  |
| --- |
| HGNC:28452 maps to PGM5P1 (Entrez:653394) ## discontinued 28-Mar-2014, withdrawn |
| HGNC:29192 maps to KIAA1107 (Entrez:23285) ## discontinued 9-Nov-2019, replaced with BTBD8: 284697 |
| HGNC:35189 maps to DNM1P39 (Entrez:100216518) ## discontinued on 9-Jul-2020, replaced with DNM1P36: 100216515 |
| HGNC:48720 maps to FAM83H-AS1 (Entrez:100128338) ## discontinued on 18-Apr-2020, replaced with IQANK1: 642574 |

Details of discrepancies between Gencode and HGNC

* In HGNC database, there are 3 pairs of HGNC IDs where each pair maps to one gencode IDs:

|  |  |  |  |
| --- | --- | --- | --- |
| HGNC:31430 | LINC00595 | Entrez:414243 | ENSG00000230417 |
| HGNC:45111 | LINC00856 | Entrez:100132987 | ENSG00000230417 |
| HGNC:53830 | STRA6LP | Entrez:112272565 | ENSG00000254876 |
| HGNC:53834 | SUGT1P4-STRA6LP | Entrez:100499484 | ENSG00000254876 |
| HGNC:18875 | TAS2R43 | Entrez:259289 | ENSG00000255374 |
| HGNC:18876 | TAS2R45 | Entrez:259291 | ENSG00000255374 |

* In Gencode, there are 17 pairs of gencode IDs where each pair maps to one HGNC IDs. E.g.

|  |  |  |  |
| --- | --- | --- | --- |
| ENSG00000269900.3 | RMRP | HGNC:10031 | Entrez:6023 |
| ENSG00000277027.1 | RMRP | HGNC:10031 |  |
| ENSG00000284770.2 | TBCE | HGNC:11582 | Entrez:6905 |
| ENSG00000285053.1 | TBCE | HGNC:11582 | Entrez:6905 |
| ENSG00000145075.13 | CCDC39 | HGNC:25244 |  |
| ENSG00000284862.3 | CCDC39 | HGNC:25244 | Entrez:339829 |

\*\* Each pair has the gene symbol as well. Similarly, each pair has the same Entrez ID

but - unexpectedly - one member in 12 (out of the 17) pairs is missing the Entrez ID.

* Gencode ids with HGNC ids and symbols in Gencode different from those in HGNC (6 records)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gencode\_id** | **Gen-Symbol** | **Gen-hgnc\_id** | **hgnc-hgnc\_id** | **hgnc-Symbol** |
| ENSG00000197927 | C2orf27A | HGNC:25077 | HGNC:40003 | NBEAP2 |
| ENSG00000164112 | TMEM155 | HGNC:26418 | HGNC:55077 | SMIM43 |
| ENSG00000233828 | LINC01949 | HGNC:25650 | HGNC:54975 | MIR4280HG |
| ENSG00000255036 | STRA6LP | HGNC:53830 | HGNC:53835 | SUGT1P4-STRA6LP-CCDC180 |
| ENSG00000259324 | OR11K1P | HGNC:15371 | HGNC:55117 | OR11K1BP |
| ENSG00000226444 | ACTR3BP6 | HGNC:38683 | HGNC:54971 | ACTR3BP7 |

* Gencode ids with the same HGNC ids in Gencode and HGNC but with different symbols (48 genes). E.g.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Gencode\_id** | **Gen-Symbol** | **Gen-hgnc\_id** | **hgnc-hgnc\_id** | **hgnc-Symbol** |
| ENSG00000152763 | WDR78 | HGNC:26252 | HGNC:26252 | DNAI4 |
| ENSG00000162643 | WDR63 | HGNC:30711 | HGNC:30711 | DNAI3 |
| ENSG00000125462 | C1orf61 | HGNC:30780 | HGNC:30780 | MIR9-1HG |
| ENSG00000198842 | DUSP27 | HGNC:25034 | HGNC:25034 | STYXL2 |

* Gencode ids without HGNC ids in Gencode database but have HGNC ids in HGNC database (84 records). E.g.

|  |  |  |  |
| --- | --- | --- | --- |
| **hgnc\_id** | **symbol** | **entrez\_id** | **ensembl\_gene\_id** |
| HGNC:20988 | AGPAT4-IT1 | 79992 | ENSG00000279355 |
| HGNC:39662 | ATP8A2P1 | 100422505 | ENSG00000240800 |
| HGNC:49153 | BMS1P20 | 96610 | ENSG00000236850 |
| HGNC:42398 | C4B\_2 | 100293534 | ENSG00000233312 |

Details of discrepancies between NCBI and Gencode

* In NCBI gene database, there are 66 sets (65 pairs and an additional set of 3) of Gencode IDs where each set maps to one Entrez IDs with one gene symbol and one HGNC ID.

|  |  |  |  |
| --- | --- | --- | --- |
| GeneID | Symbol | HGNC | Ensembl |
| 221178 | SPATA13 | HGNC:23222 | ENSG00000228741,ENSG00000273167,ENSG00000182957 |
| 221468 | TMEM217 | HGNC:21238 | ENSG00000172738,ENSG00000286105 |
| 253970 | SFTA3 | HGNC:18387 | ENSG00000257520,ENSG00000229415 |

* In Gencode, there are only 45 sets (44 pairs and an additional set of 3) of Gencode IDs where each set maps to one Entrez IDs. However, in most sets, the genes have different symbols and HGNC IDs

4. Ambiguity of gene symbols across multiple species: Mapping genes across species has to accept some facts:

* Not every gene will have an orthologous match in every species.
* Ambiguity is inevitable because of possible one-to-many relationship between orthologous genes.

Example: Among the 41821 Approved gene IDs in HGNC on March 25, 2020, there are 18206 gene has corresponding ID in the Mouse genome database (MGI) as supplied by MGI but only 17976. This is because there are 413 HGNC gene where each 2 or more genes share the same MGI ID. For example, there are 8 HGNC genes have the same MGI ID:894323. On the other hand, one of 416 HGNC genes could be assigned to several MGI IDs. For example, HGNC:22970 has 15 corresponding MGI IDs. In comparison to the 18206 genes in HGNC with corresponding MGI ID as supplied by MGI, there are only 17692 genes in HGNC with corresponding curated MGI ID. There are 17515 record identical in the 2 map files. The curated MGI IDs has 3 records not reported in the MGI ID as supplied by MGI and 174 records with modified IDs.