Data, Metadata, and Annotations

Page Contents

Documentation Table of Contents

To Print the Guide

We recommend you print one wiki page of the guide at a time. To do this, click the printer icon at the top right of the page; then from the browser File menu, choose Print. Printing multiple pages at one time is more complex. For instructions, refer to Printing multiple pages .

Having Trouble Reading the Text?

Resizing the text for any web page is easy. For information on how to do this in your web browser, refer to this W3C tutorial

Data, Metadata, and Annotations Overview

All of the Gene-Disease and Gene-Compound Cancer Gene Index data, annotations, and metadata are available in the two XML documents and the caBIO interfaces, excluding the caBIO Portlet Templated Search. These data include NCI Thesaurus disease and compound terms, NCI Thesaurus disease and compound concept identifiers, HUGO Gene Symbols, LocusLink Gene Symbols, UniProt Identifiers, the sentences that contained evidence of the gene-disease or gene-compound associations, and the PubMed identifier of the abstract from which each sentence was extracted. The gene-disease and gene-compound data were then annotated by human curators with Evidence Codes, Role Codes, and Role Details; Cell line and Negation Indicators; Sentence and Gene Status Flags; and Comments.

For additional information on how these data were collected, validated, and annotated, refer to the Creation of the Cancer Gene Index section.

Evidence Codes

Evidence codes qualify the assertions with respect to the association of a gene to a disease or compound term by telling how the assertions were made (for example, through inference or experimental data). The curators may have identified the means by which an assertion using the extracted sentence, alone, or through careful reading of the abstract from which the sentence originated. These codes follow the suggestions of Karp *et al.*

for ontologies used in pathway and genome databases. The Evidence Code associated with a specific gene-disease or gene-compound pair is found in the text contents of the XML `EvidenceCode` element and is the `EvidenceCode` attribute of the caBIO `Evidence Code class`, `gov.nih.nci.cabio.domain.EvidenceCode`.

Evidence Code	Description
EV-IC	Inferred by curator. An assertion was inferred by a curator from relevant information such as other assertions in a database.
EV-COMP	Inferred from computation. The evidence for an assertion comes from a computational analysis. The assertion itself might have been made by an author or by a computer, that is, EV-COMP does not specify whether manual interpretation of the computation occurred.
EV-COMP-HINF	Human inference. A curator or author inferred this assertion after review of one or more possible types of computational evidence such as sequence similarity, recognized motifs or consensus sequence, etc. When the inference was made by a computer in an automated fashion, use EV-AINF.
EV-COMP-HINF-SIMILAR-TO-CONSENSUS	An author inferred, or reviewed a computer inference of, sequence function based on similarity to a consensus sequence.
EV-COMP-HINF-POSITIONAL-IDENTIFICATION	An author inferred, or reviewed a computer inference of, promoter position relative to the -10 and -35 boxes.
EV-COMP-HINF-FN-FROM-SEQ	An author inferred, or reviewed a computer inference of, gene function based on sequence, profile, or structural similarity (as computed from sequence) to one or more other sequences.
EV-COMP-AINF	Automated inference. A computer inferred this assertion through one of many possible methods such as sequence similarity, recognized motifs or consensus sequence, etc. When a person made the inference from computational evidence, use EV-HINF.
EV-COMP-AINF-SINGLE-DIRECTON	Automated inference of transcription unit based on single-gene direction. Existence of a single-gene transcription unit for gene G is inferred computationally by the existence of upstream and downstream genes transcribed in the opposite direction of G.
EV-COMP-AINF-SIMILAR-TO-CONSENSUS	A DNA sequence similar to previously known consensus sequences is computationally identified.
EV-COMP-AINF-POSITIONAL-IDENTIFICATION	Automated inference of promoter position relative to the -10 and -35 boxes.
EV-COMP-AINF-FN-FROM-SEQ	Automated inference of function from sequence. A computer inferred a gene function based on sequence, profile, or structural similarity (as computed from sequence) to one or more other sequences.
EV-AS-TAS	Traceable author statement. The assertion was made in a publication – such as a review – that itself did not describe an experiment supporting the assertion. The statement referenced another publication that supported the assertion, but it is unclear whether that publication described an experiment that supported the assertion.
EV-AS-NAS	Non-traceable author statement. The assertion was made in a publication such as a review, without a reference to a publication describing an experiment that supports the assertion.

EV-EXP	Inferred from experiment. The evidence for an assertion comes from a wet-lab experiment of some type.
EV-EXP-IPI	IPI inferred from physical interaction The assertion was inferred from a physical interaction such as 2-hybrid interactions, Co-purification, Co-immunoprecipitation, Ion/protein binding experiments This code covers physical interactions between the gene product of interest and another molecule (or ion, or complex). For functions such as protein binding or nucleic acid binding, a binding assay is simultaneously IPI and IDA; IDA is preferred because the assay directly detects the binding.
EV-EXP-IDA	IDA inferred from direct assay. The assertion was inferred from a direct experimental assay such as Enzyme assays, In vitro reconstitution (for example, transcription), Immunofluorescence, Cell fractionation, etc.
EV-EXP-IDA-UNPURIFIED-PROTEIN	Direct assay of unpurified protein. Presence of a protein activity is indicated by an assay. However, the precise identity of the protein with that activity is not established by this experiment (protein has not been purified).
EV-EXP-IDA-TRANSCRIPTION-INIT-MAPPING	The transcription start site is identified by primer extension.
EV-EXP-IDA-TRANSCRIPT-LEN-DETERMINATION	The length of the (transcribed) RNA is experimentally determined. The length of the mRNA is compared with that of the DNA sequence and by this means the number of genes transcribed are established.
EV-EXP-IDA-RNA-POLYMERASE-FOOTPRINTING	The binding of RNA polymerase to a DNA region (the promoter) is shown by footprinting.
EV-EXP-IDA-PURIFIED-PROTEIN-MULTSPECIES	Protein purified from mixed culture or other multispecies environment (such as, infected plant or animal tissue), and activity measured through in vitro assay.
EV-EXP-IDA-PURIFIED-PROTEIN	Protein purified to homogeneity from specific species (or from heterologous expression vector), and activity measured through in vitro assay.
EV-EXP-IDA-BOUNDARIES-DEFINED	Sites or genes bounding the transcription unit are experimentally identified. Several possible cases exist, such as defining the boundaries of a transcription unit with an experimentally identified promoter and terminator, or with a promoter and a downstream gene that is transcribed in the opposite direction, or with a terminator and an upstream gene that is transcribed in the opposite direction.
EV-EXP-IDA-BINDING-OF-PURIFIED-PROTEINS	IDA inferred from direct assay. The assertion was inferred from a direct experimental assay such as Enzyme assays, In vitro reconstitution (for example, transcription), Immunofluorescence, Cell fractionation.
EV-EXP-IDA-BINDING-OF-CELLULAR-EXTRACTS	There exists physical evidence of the binding of cellular extracts containing a regulatory protein to its DNA binding site. This can be either by footprinting or mobility shift assays.
EV-EXP-IEP	IEP inferred from expression pattern. The assertion was inferred from a pattern of expression data such as Transcript levels (for example, Northerns, microarray data), Protein levels (for example, Western blots).
EV-EXP-IEP-GENE-EXPRESSION-ANALYSIS	The expression of the gene is analyzed through a transcriptional fusion (that is, lacZ), and a difference in expression levels is observed when the regulatory protein is present (wild type) vs in its absence. Note that this evidence does not eliminate the possibility of an indirect effect of the regulator on the regulated gene.

EV-EXP-IGI	IGI inferred from genetic interaction. The assertion was inferred from a genetic interaction such as "Traditional" genetic interactions such as suppressors, synthetic lethals, etc., Functional complementation, Inference about one gene drawn from the phenotype of a mutation in a different gene. This category includes any combination of alterations in the sequence (mutation) or expression of more than one gene /gene product. This category can therefore cover any of the IMP experiments that are done in a non-wild-type background, although we prefer to use it only when all mutations are documented.
EV-EXP-IGI-FUNC-COMPLEMENTATION	Protein activity inferred by isolating its gene and performing functional complementation of a well characterized heterologous mutant for the protein.
EV-EXP-IMP	IMP inferred from mutant phenotype. The assertion was inferred from a mutant phenotype such as Any gene mutation/knockout, Overexpression/ectopic expression of wild-type or mutant genes, Anti-sense experiments, RNA interference experiments, Specific protein inhibitors, Complementation. Inferences made from examining mutations or abnormal levels of only the product(s) of the gene of interest are covered by code EV-IMP (compare to code EV-IGI). Use this code for experiments that use antibodies or other specific inhibitors of RNA or protein activity, even though no gene may be mutated (the rationale is that EV-IMP is used where an abnormal situation prevails in a cell or organism).
EV-EXP-IMP-REACTION-ENHANCED	Gene is isolated and over-expressed, and increased accumulation of reaction product is observed.
EV-EXP-IMP-POLAR-MUTATION	If a mutation in a gene or promoter prevents expression of the downstream genes due to a polar effect, the mutated gene is clearly part of the transcription unit.
EV-EXP-IMP-REACTION-BLOCKED	Mutant is characterized, and blocking of reaction is demonstrated.
EV-EXP-IMP-SITE-MUTATION	A cis-mutation in the DNA sequence of the transcription-factor binding site interferes with the operation of the regulatory function. This is considered strong evidence for the existence and functional role of the DNA binding site.
not_assigned	Evidence Code was not assigned.
based on abstract	Determinations of whether the gene-disease or gene-compound association from a sentence was factual based upon an expert's interpretation of the abstract from which the sentence originated.
	No Evidence Code was assigned because the sentence did not contain the expected gene-disease or gene-compound association evidence.

Role Code and Role Detail Similarities

The Cancer Gene Index Role Codes and Role Details are derived from NCI Role Codes. Both describe the semantic associations between gene concept and either a disease or compound concept (that is, concept pairs). Whereas the Evidence Codes describe how the association was inferred or the type of experiment upon which the inference was made, Role Codes and Role Details give information about the actual gene-disease or gene-compound association. Multiple Role Codes and Role Details can be used for the same sentence.

> **Note**
>
> A concept is the actual compound, disease, or gene to which the various names, acronyms, alternate spellings, and abbreviations refer.

Role Codes

Gene-Disease and Gene-Compound Role Codes most often describe that a gene is associated with a disease or compound (for example, `GENE_ASSOCIATED_WITH_DISEASE`) or how the concepts are associated (for example, Chemical_or_Drug_Is_Metabolized_By_Enzyme), but they also may describe relevant features of the role of a particular gene (for example, `GENE_HAS_FUNCTION`). For the former, the gene name, Role Code, and disease or compound often can form a sentence, such as "BRCA1 `GENE_ASSOCIATED_WITH_DISEASE` BREAST CANCER." The Role Code `not_assigned` indicates that the curator did not or could not assign a specific code. The Role Code associated describing evidence of a gene-disease or gene-compound pair is found in the text contents of the XML `PrimaryNCIRoleCode` element and is the role attribute of the caBIO `GeneDiseaseAssociation` and `GeneAgentAssociation` classes, `gov.nih.nci.cabio.domain.GeneDiseaseAssociation` and `gov.nih.nci.cabio.domain.GeneAgentAssociation`.

> **Note**
>
> Although pharmacological substances are referred to as "compounds" in the Cancer Gene Index, the NCI Thesaurus and caBIO use the term "agent."

- **Gene-Disease Role Codes**
  - Gene_Associated_With_Disease
  - Gene_Product_Anormaly_Affects_Pathway
  - Gene_Product_Anomaly_Related_To_Gene_Anormaly
  - Gene_Product_Encoded_By_Gene
  - Gene_Product_Expressed_In_Tissue
  - Gene_Product_Has_Associated_Anatomie
  - Gene_Product_Has_Biochemical_Function
  - Gene_Product_Has_Chemical_Classification
  - Gene_Product_Has_Malfunction_Type
  - Gene_Product_Has_Organism_Source
  - Gene_Product_Has_Structural_Domain_Or_Motif
  - Gene_Product_is_Biomarker_of
  - Gene_Product_is_Biomarker_Type
  - Gene_Product_is_Pathway_Element
  - Gene_Product_is_Physical_Part_Of
  - Gene_Product_Malfunction_Associated_With_Disease
  - Gene_Product_Plays_Role_In_Biological_Process
  - Gene_Malfunction_Associated_With_Disease
  - Gene_Expressed_In_Tissue
  - Gene_Found_In_Organism
  - Gene_Has_Anormally
  - Gene_Has_Clone
  - Gene_Has_Expression_Measurement
  - Gene_Has_Function
  - Gene_In_Chromosomal_Location
  - Gene_is_Biomarker_of
  - Gene_Is_Pathway_Element
  - Gene_Plays_Role_In_Process
  - Disease_Has_Cytogenetic_Abnormality
  - Disease_May_Have_Cytogenetic_Abnormality
  - Disease_Has_Molecular_Abnormality
  - Disease_May_Have_Molecular_Abnormality
- **Gene-Compound Role Codes**
  - Chemical_or_Drug_Affects_Cell_Type_or_Tissue
  - Chemical_or_Drug_Plays_Role_in_Biological_Process
  - Chemical_or_Drug_FDA_Approved_for_Disease
  - Chemical_or_Drug_Is_Metabolized_By_Enzyme
  - Chemical_or_Drug_Has_Accepted_Therapeutic_Use_For
  - Chemical_or_Drug_Has_Study_Therapeutic_Use_For
  - Chemical_or_Drug_Has_Mechanism_Of_Action
  - Chemical_or_Drug_Affects_Gene_Product
  - Chemical_or_Drug_Has_Target_Gene_Product

Role Details

Gene-Disease and Gene-Compound Role Details most often provide precise descriptions of the association of a gene term and a corresponding disease or compound term. These Details can also describe relevant features of the role of a particular gene (for example, `Chemical_or_Drug_Represses_Gene_Product_Expression`). While similar to Role Codes, Role Details give more specific semantic descriptions. For example, a Role Detail for a particular gene-disease concept pair association may be `GENE_PRODUCT_UPREGULATED_IN_DISEASE`, whereas a similar role code may be `GENE_ASSOCIATED_WITH_DISEASE`. The Role Detail `not_assigned` indicates that the curator did not or could not assign a specific semantic detail. The Role Detail associated with a specific gene-disease or gene-compound pair is found in the text contents of the XML `OtherRole` element and is, like Role Codes, the role attribute of the caBIO `GeneDiseaseAssociation` and `GeneAgentAssociation` classes.

- **Gene-Disease Role Details**
  - Gene_Product_Affects_Disease
  - Gene_Product_Affects_Disease_Process
  - Gene_Product_Expressed_in_Disease
  - Gene_Product_Decreased_in_Disease
  - Gene_Product_Increased_in_Disease
  - Gene_Product_Level_Changed_in_Disease
  - Gene_Expressed_in_Disease
  - Gene_Expression_Downregulated_in_Disease
  - Gene_Expression_Upregulated_in_Disease
  - Gene_Expression_Changed_in_Disease
  - Gene_May_Be_Associated_With_Disease
  - Gene_Anormaly_has_Disease-Related_Function

    - Gene_Anormaly_May_have_Disease-Related_Function
    - Gene_Product_Anormaly_has_Disease-Related_Function
    - Gene_Product_Anormaly_May_have_Disease-Related_Function
    - Gene_has_Therapeutic_Relevance
    - Gene_May_have_Therapeutic_Relevance
    - Gene_Product_has_Therapeutic_Relevance
    - Gene_Product_May_have_Therapeutic_Relevance
- **Gene-Compound Role Details**
    - Chemical_or_Drug_in_Clinical_Study
    - Chemical_or_Drug_May_Affect_Gene_Product
    - Chemical_or_Drug_May_Affect_Gene
    - Chemical_or_Drug_Affects_Gene
    - Chemical_or_Drug_Regulates_Gene
    - Chemical_or_Drug_Regulates_Gene_Product
    - Chemical_or_Drug_Activates_Gene_Product
    - Chemical_or_Drug_Inhibits_Gene_Product
    - Chemical_or_Drug_Affects_Gene_Product_Function
    - Chemical_or_Drug_Binds_to_Gene_Product
    - Chemical_or_Drug_Affects_Expression
    - Chemical_or_Drug_Affects_Gene_Expression
    - Chemical_or_Drug_Affects_Gene_Product_Expression
    - Chemical_or_Drug_Changes_Expression
    - Chemical_or_Drug_Induces_Gene_Expression
    - Chemical_or_Drug_Induces_Gene_Product_Expression
    - Chemical_or_Drug_Regulates_Expression
    - Chemical_or_Drug_Represses_Gene_Expression
    - Chemical_or_Drug_Represses_Gene_Product_Expression
    - Chemical_or_Drug_Mediates_Pathway_Activity
    - Chemical_or_Drug_Increases_Pathway_Activity
    - Chemical_or_Drug_Decreases_Pathway_Activity
    - Chemical_or_Drug_Mediates_Metabolic_Status
    - Chemical_or_Drug_Increases_Metabolic_Status
    - Chemical_or_Drug_Decreases_Metabolic_Status
    - Chemical_or_Drug_Has_Physiologic_Effect
    - Gene_Product_Affects_Compound
    - Gene_Product_May_Affect_Compound
    - Gene_Affects_Compound
    - Gene_May_Affect_Compound
    - Gene_Product_Antagonizes_Chemical_or_Drug
    - Gene_Product_Transports_Compound
    - Gene_Anomaly_Effects_Resistance_to_Chemical_or_Drug
    - Gene_Product_Anomaly_Effects_Resistance_to_Chemical_or_Drug
    - Gene_Anomaly_May_Effect_Resistance_to_Chemical_or_Drug
    - Gene_Product_Anomaly_May_Effect_Resistance_to_Chemical_or_Drug
    - Gene_is_Associated_with_Resistance_to_Chemical_or_Drug
    - Gene_Product_is_Associated_with_Resistance_to_Chemical_or_Drug
    - Gene_May_be_Associated_with_Resistance_to_Chemical_or_Drug
    - Gene_Product_May_be_Associated_with_Resistance_to_Chemical_or_Drug

Cell Line and Negation Indicators

The binary indicators `yes` or `no` were set for cell line and negation annotations of each gene-disease and gene-compound association. Cell line indicators denote whether the evidence came from a cell line (`yes`) or other source, such a human subject, animal model, or primary cells (`no`). The cell line indicator is the text contents of the XML `CelllineIndicator` element and the `celllineStatus` attribute of the caBIO `Evidence` class. Negation indicators specify whether the evidence actually described a lack of association between the candidate binary concept pair (`yes`), or whether there was a true relationship between them (`no`). The curators may have deduced the negation indicator by the extracted sentence, alone, or through careful reading of the abstract from which the sentence originated. Occasionally, the curators did not set a negation indicator (`-`). The negation indicator is the text contents of the XML `NegationIndicator` element and the `negationStatus` attribute of the caBIO `Evidence` class.

Status Flags

The curators set flags to denote the state of annotations of a gene term or of a particular sentence.

Gene Status Flags

Gene Status Flags describe whether annotations for all of the sentences for a given gene term are complete and whether the gene term has been withdrawn from EntrezGene. All low frequency sentence count genes were finished, whereas some high frequency sentence count genes were not. The status of a specific gene is found in the text contents of the XML `GeneStatusFlag` element.

Gene Status Flag	Status Flag Description
Finished	All sentences with this gene have been annotated
New	Not all sentences with this high frequency sentence count gene have been annotated.
Withdrawn	Gene term has been withdrawn from EntrezGene and, where possible, the term has been mapped to a valid EntrezGene term.
Entry withdrawn	Gene term has been withdrawn from EntrezGene and, where possible, the term has been mapped to a valid EntrezGene term.

Sentence Status Flags

Unlike gene status flags, which can cover many sentences associated with a single gene, Sentence Status Flags describe the curator's findings for a specific sentence. Sentences can be true positives, false positives, unclear, or redundant. The status of a specific gene is found in the text contents of the XML `SentenceStatusFlag` element and is the `sentenceStatus` attribute of the caBIO `Evidence` class.

Sentence Status Flag	Status Flag Description
Finished	Sentence validation and annotation complete
No_fact	Invalid sentence or false positive
Unclear	Sentence included both a gene and disease or gene and compound term, but the relationship between the gene-disease or gene-compound pair was not obvious from the sentence.
Redundant	Identical gene-disease or gene-compound associations were captured from multiple sentences originating from the same abstract.

Comments

Often, the expert curators made free-text comments on records within the Gene-Disease or Gene-Compound databases. Comments included, but were not limited to, notations of genetic anomalies (for example, loss of heterozygosity, polymorphisms, or aberrant methylation), additional disease information, name of the non-human organism from which the experimental data were collected, information on the cell line or other notable reagents used in the execution of the experiment, and other miscellaneous information. Any comments on a sentence are found in the text contents of the XML `Comments` element and the `comment` attribute of the caBIO `Evidence` class.