This is the page for Curation Standards, are you looking for the related page How to Curate GeneSets?

Secondary functional genomics data consists of the results of analyzed experiments in functional genomics. In contrast to primary data stores such as Gene Expression Omnibus (GEO) in which raw experimental data are stored, a secondary data store attempts to collect the results of experimental design and decision making process of the researcher so that one may interpret and integrate the gene set centered outcomes of the studies. Controlling the quality and validity of the large-scale analysis of secondary data requires the enforcement of interpretable standards for gene set construction and description. GeneWeaver?s use of discrete analysis eliminates many barriers to the integration of heterogeneous data sets across species and experiments. However, it is important for users to be able to rapidly interpret the nature of gene sets retrieved from the site, requiring a minimal standard for metadata associated with secondary data. For this purpose, both unstructured textual descriptions of the data and structured ontology annotations to the terms in these descriptions are used to define gene sets. In the interest of encouraging submission we are cautious not to be too prescriptive or burdensome to users, but rather to provide guidelines on standards used by internal curators to assess data quality and clarity to enable rapid acceptance of community submissions to the data repository.

Curation Tiers

Tier	Name	Curator	Description
Tier I	Public Resource Grade	Resource GeneWeaver	Large data sets primarily curated by their parent resource. GeneWeaver ensures consistency of metadata (gene annotations to KEGG, MP and GO, curated functional associations in the Neuroinformatics Framework, Comparative Toxicogenomics Database)
Tier II	Machine-Generated from public sources	GeneWeaver	Gene sets resulting from genome analysis, not otherwise published in total, e.g. gene co-expression to behavior from GeneNetwork.org, QTL positional candidates from MGI. GeneWeaver curators examine data and metadata.
Tier III	Human-Curated	GeneWeaver	Curated user-deposited data and publication supplements in domains of

			interest.
Tier IV	Submitted to Public- Provisional	User	User-deposited data made available to the public. All Tier IV is examined for promotion to Tier III
Tier V	Private User and Group data- Uncurated	User	Data sets deposited for private or group-only analysis

Description of GeneWeaver Curation Tiers

title here

Tier I: Public Resource Data	Tier I data are professionally curated into another major database and are imported into GeneWeaver, which ensures consistency of metadata. Resource grade data is updated on a six-month cycle. These include: gene annotations to KEGG, MP and GO, curated functional associations in Neuroinformatics Framework, and Comparative Toxicogenomics Database.
Tier II: Machine-Generated from public sources	Tier II data are computationally generated from data in public sources. These include empirical data obtained from public sources and their associated analytical tools, e.g. bulk analysis of gene co-expression to phenotypes across mouse strains from GeneNetwork.org, or QTL positional candidates from MGI. In contrast to Tier I in which the individual gene annotations to function are manually curated, Tier II includes machine generated gene annotations to functions from curated experimental data. GeneWeaver curators examine data and metadata.
Tier III: Human-Curated Data	Tier III data are directly entered or reviewed by a professional curator for redundancy with existing records and adherence to documentation standards. Users who submit data under Tier IV have the option of sharing their data to the public. These data will be marked provisional until reviewed by the curator for data entry errors, compliance to metadata standards and redundancy with existing data. The submitter of the data will have the opportunity to approve the curators modifications them prior to upgrade to Tier III status. For some research areas, a professional curator has identified and entered gene expression, quantitative trait locus and genomewide association studies (GWAS). Where possible, the curator has obtained results directly from the study authors, supplements, or data repositories such as GEO, in addition to the often highly-filtered set of results reported in publications.

Tier IV: Submitted to Public-Provisional	Tier IV consists of user submitted data that has been shared to the public prior to review. This data is indicated as provisional, but can be used in all analyses. Curatorial review is required to remove the provisional label.
Tier V: Private User and Group Data, Uncurated	Data in user accounts that is assigned private or group level access is confidential is not exposed to analyses by users outside of the group to whom it is shared, and is therefore not reviewed by the professional curator.

Subtitle here

Gene Set Field Definitions

***Gene Set Name**:*	A brief title for the gene set, approximately sentence length, that should provide a clear and concise description of the contents of a gene set interpretable to most users of GeneWeaver, but with sufficient detail to satisfy a domain expert. This is the major gene set name that is displayed in all search results, project directory and table views of analysis results. Standards for specific gene set types are given in the following section.
***Gene Set Figure Label**:*	A brief 23 character abbreviation to facilitate recognition of the gene set in a graph or other display.
***Gene Set Description**:*	A detailed description of the gene set, including rules for its construction, experimental methods and analyses used to generate data, anatomical terms, and traceable references to source data including accession information and date. Abbreviations should be avoided.
Ontology Annotations	Relevant terms from Disease Ontology, Mammalian Ontology and other OBO ontologies supplied by curators or identified through the application of the NCBO Annotator to textual descriptions including publication abstracts.
Publication Information	PubMed ID, Title, authors, publication information and full-text of the abstract.

General Definitions of GeneSet Fields

Type of Data: Differential Expression Profiling

***Gene Set Name**:*	Genes [upregulated/downregulated/differentially expressed] in [tissue] [comparison].
Example:	*Genes differentially expressed in striatum of C57Bl/6J compared to C57Bl/6C.*
	Note: spell out anatomical terms as nouns, e.g. striatum, not

	striatal. Include complete strain names, e.g. C57BL/6J not B6.
***Gene Set Figure Label**:*	B6JvsB6CStriatum
***Gene Set Description**:*	Indicate which samples were compared. What experimental manipulations or tissue differences are being examined. Indicate statistical methodology, significance thresholds and which changes are reported here. Indicate if uploaded p-value, q-value, effect size or fold change and fold change reference.
Example:	*Striatum gene expression differences between naive C57BL/6J and C57BL/6C substrains corresponding to a 5% FDR. A small number of genes are highly differentially expressed between B6 substrains, C57BL/6J (high alcohol consumption preference) and C57BL/6C (low alcohol consumption preference). Fold expression change are relative to B6/J.*
***Gene Set Contents**:*	Gene identifier and statistical score for differential expression, e.g. p-value, q-value, correlation coefficient, binary score, effect size or fold change.

Type of Data: Published QTL Candidate Gene List

***Gene Set Name**:*	Description (name, Published QT Chr # MGI:#)
Example:	*cocaine related behavior 10 (Cocrb10, Published QTL Chr #)*
***Gene Set Figure Label**:*	(QTL-name-Organism-Chr #)
Example:	*QTL-Cocrb10-Mouse-Chr 9*
***Gene Set Description**:*	QTL Name Definition, candidate gene selection method (e.g. 1.5 LOD drop; inter-marker interval). Exact description of phenotype. Strains used for mapping should be included.
Example:	*Rats were subjected to a forced swim test (FST) procedure in which they are placed in water for 5 min, and their behavior was scored every 5 sec as immobility, climbing, or swimming. Data were analyzed for each activity with consideration given to their non-independence. p-value:0.0002, Variance: 3.6, Peak Marker: D5Rat40 (BLAT 16538053) Spans 1-41538053. This interval was obtained by using a fixed interval width of 25 Mbp around the peak marker. Strains were WKY/NHsd and F344/NHsd. Also defined as Imm3*
***Gene Set Contents**:*	Gene identifier and binary score.

Type of Data: Co-Expression to Phenotype

***Gene Set Name**:*	Describe tissue and phenotype correlated.
Example:	*Cerebellum gene expression correlates of acetic acid writhing behavior in BXD recombinant inbred mice.*
***Gene Set Figure Label**:*	Co-expression writhing
***Gene Set Description**:*	Indicate what the comparison was that was made and any statistical cut-offs that were used.
Example:	*Cerebellum gene co-expression with acetic acid writhing in BXD RI mice. Gene expression data was obtained from genenetwork.org SJUT Cerebellum mRNA M430 (Mar05) RMA data set. Behavioral phenotype data was collected by RMQ and consisted of the number of writhes in response to 0.6% acetic acid i.p.*
***Gene Set Contents**:*	Gene identifier and statistical score for co-expression. e.g. R-squared, p-value, q-value, binary threshold.

Type of Data: Reference Ontology

Gene Set Name:	Term # and name
Example:	*MP:XXXXXXX Abnormal?*
***Gene Set Figure Label**:*	Term #
Example:	*Term #*
***Gene Set Description**:*	Term Definition.
Example:	*?Increase in the dose or concentration of a foreign compound required to induce a specific level of response? www.informatics.jax.org, 2010-12-01*
***Gene Set Contents**:*	All gene sets include genes, mutant alleles or gene products annotated to an ontology term by a professional curator. Each gene directly annotated to the term is given a score of 1, each gene connected to a term through annotations to its higher order parents is given a score of 2. To use only direct annotations in an analysis assign a threshold of < 2 to each Gene Set.

Type of Data: Co-Expression Clusters

***Gene Set Name**:*	Co-Expression clusters?
Example:	*Co-expression cluster of nicotine Dependence genes significantly expressed in the adolescent PFC, VS and Hippocampus.*
	Abbreviated description

***Gene Set Figure Label**:*	
Example:	*Adolesc Rat Nic Dependence*
***Gene Set Description**:*	Indicate what samples were compared and what was clustered.
Example:	*Studies analyzing brain samples from female rats that had been injected with nicotine at four different ages show that nicotine exerts the greatest influence during adolescence. Using DNA microarrays, gene expression correlates were obtained from the prefrontal cortex (PFC), ventral striatum (VS), and hippocampus. Principal cluster analysis was then used to identify 76 genes that changed significantly in at least one of these three brain regions during the experiment.*
***Gene Set Contents**:*	Gene identifier and statistical score for cluster analysis or binary threshold.

Type of Data: Genome Wide Association Study

***Gene Set Name**:*	GWAS of ?
Example:	*GWAS of Alcohol and Nicotine Dependence in Australian DNA-Pools.*
***Gene Set Figure Label**:*	Abbreviated description
Example:	*GWAS Alcohol Nicotine*
***Gene Set Description**:*	List of positional candidate genes after correcting for multiple testing and controlling the false discovery rate from genome wide association study. Represents genes associated with a linked cytological region or genes ?near? an associated SNP.
Example:	*Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis.*
***Gene Set Contents**:*	Gene identifier and binary threshold.