EntrezGene Load/Mouse

Author: Lori, Richard, Sophia

Created: September 22, 2004

Last Modified: March 24, 2006 13:09

1 Purpose of Document

To define requirements for the EntrezGene/MGI Mouse Marker association load.

2 Definitions

- **EntrezGene** - *From the EntrezGene site*: Gene provides a unified query environment for genes defined by sequence and/or in NCBI's Map Viewer. You can query on names, symbols, accessions, publications, GO terms, chromosome numbers, E.C. numbers, and many other attributes associated with genes and the products they encode.

  Abbreviated as EG in this document.
- **Load Reference** - The reference for the EG load, J:63101.
- All references to "MGI marker" implicitly means "mouse" only.
- All references to "EG gene" implicitly means "mouse" only.

3 Overview

The EntrezGene load (mouse) is used to make associations between EntrezGene genes and MGI markers.

We use many criteria to determine whether an EG association is appropriate. Our current criteria include:

1. the MGI marker must be of type Gene or PseudoGene
2. there is a match between the MGI id in EG and in MGI
3. there is at least one GenBank id in common between the EG gene and the MGI marker

These associations are annotated using the EntrezGene identifier (EG id) and the load reference. Once an annotation has been made between an EG id and a MGI marker, we load these additional annotations to the MGI marker:

1. NM RefSeq ids
2. RNA GenBank ids

   Since GenBank ids can be MGI-curatable, only those GenBank ids that are not already annotated to the MGI marker are loaded.

The current EntrezGene load deterimines approximately 27,500 associations.

The MGS group believes that this criteria is too restrictive and that by defining new criteria we can make more MGI-to-EG associations without compromising the integrity of these associations.

After doing some careful analysis, we have defined a new EG load algorithm, described in this document. This algorithm has resulted in approximately 12,300 new MGI-to-EGassociations (from 27,500 to 39,800). This was accomplished by eliminating some of the current 1:1 criteria and by relaxing others.

4 Brief Overview of Bucketization

Before we continue with the new EG load requirements, let's review how we determine a valid association between 2 data providers.

Our 2 data providers are:

1. MGI
2. EntrezGene

For each data provider, we generate a set of data records where each data record represents a unique data provider object (either a MGI marker or an EG gene). For example:

1. MGI data record = primary MGI id, secondary MGI ids, GenBank seq ids, XM RefSeq ids.
2. EG data record = EG id, MGI id, GenBank seq ids, XM RefSeq ids.

Our bucketization algorithm takes these data records and analyzes them by comparing the MGI ids from each data provider, the GenBank ids from each data provider, etc.. This produces the following types of correspondences between MGI markers and EG genes:

- 1:1; there is a one-to-one correspondence
- 1:N; there is a one-to-many correspondence
- N:1; there is a many-to-one correspondence
- N:N; there is a many-to-many correspondence.
- 1:0; there is a one-to-zero correspondence
- 0:1; there is a zero-to-one correspondence

The criteria used for a 1:1 correspondence between any 2 data providers is application-dependent. For the existing EG load, the 1:1 criteria is a match on the MGI id and at least one sequence id. However the correspondence criteria could be a match on just one id or on N ids.

The records that fall into the 1:1 bucket represent the valid associations we load into MGI.

5 Data Requirements

5.1 Load Strategy

The EG load is implemented as a delete/reload. That is, all of the data loaded from the previous execution of the load is deleted (using the load reference) from MGI before the new data is analyzed and processed.

1. Delete all associations to J:63103.
2. Delete all GenBank associations to J:53168. These GenBank associations are picked up as part of the SwissProt load. We want the EntrezGene load to pick up as many additional GenBank annotations as possible. And we want the SwissProt load to pick up any GenBank annotations that EntrezGene does not have. So, EntrezGene is the "master" source of GenBank annotations and SwissProt is a secondary source of GenBank annotations. This also avoids circular data dependency because the SwissProt GenBank annotations are made based on the EG load.

5.2 Excluded Sequence Sets

Exclude unwanted GenBank (RNA and genomic), XM_ RefSeq and XR_ RefSeq sequences from the MGI and EG sets:

1. Sequences that are associated with multiple MGI markers.
2. Sequences that are associated with multiple EG genes.
3. ~~Sequences associated with MGI markers via another association loader (for example, the SwissProt load, J:53168).~~

This eliminates:

- most of the large insert DNA sequences and rare cases of RNA sequences that are annotated to mulitple records in either MGI or EG or in both MGI and EG.

Note: Transcript sequences can be associated with multiple genes for various reasons, including bicistronic transcripts, chimeric transcripts, transcripts (usually partial) derived from duplicated genomic regions where the transcript is identical to more than one genomic region, etc.).

5.3 MGI and EG Sets

The MGI set is defined as all of the MGI markers defined by a primary MGI id where each MGI marker data record contains any of the following:

- MGI ids (both primary and secondary)

  because EG may have either the primary or secondary MGI id
- GenBank ids (RNA and genomic) not present in any Exclude Sequence sets
- XM_ RefSeq ids not present in any Exclude Sequence sets
- XR_ RefSeq ids not present in any Exclude Sequence sets

- MGI chromosome

The EG set is defined as all of the EG genes defined by a EG id where each EG gene data record contains any of the following:

- MGI ids

  Note that some EG genes do not have a MGI id. These EG gene records should still be included in the EG set.
- MGI-curatableGenBank ids (RNA and genomic) not present in any Exclude Sequence sets
- XM_ RefSeq ids not present in any Exclude Sequence sets
- XR_ RefSeq ids not present in any Exclude Sequence sets
- EG chromosome

The bucketizer will use these attributes when running its comparison algorithm.

5.4 1:1 Bucketization Criteria

The records that fall into the 1:1 bucket represent the EG/Marker associations we load into MGI.

The criteria for a 1:1 correspondence between an EG gene and a MGI marker is a match on *at least one* of the following id types:

- MGI id
- GenBank (RNA or genomic) id
- XM_ RefSeq id
- XR_ RefSeq id

and a chromosome match on the EG gene and MGI marker where:

- a chromosome of UN matches a known chromosome (for example, UN matches “1”) (there are approximately 110 MGI markers with a chromosome discrepency with EG)

5.5 Rules for Associating Additional Sequence Ids to MGI Markers (TRs 4451, 5656)

Once a 1:1 correspondence has been determined for an EG gene and a MGI marker, there are additional sequence ids in EG that we want to associate with the MGI marker.

Only sequence ids that:

1. are **not** present in the Excluded Sequence set **AND**
2. are **not** associated with a problem Clone (one that contains a note like “%staff have found evidence of artifact in the sequence of this molecular%”)

can be associated with a MGI marker during this process.

An MGI-curatable sequence category is one that MGI curators can annotate to via the Marker EI using a non-load reference.

The additional sequence categories are:

1. GenBank RNA, MGI-curatable
2. RefSeq XM, MGI-curatable
3. RefSeq XR, MGI-curatable
4. RefSeq XP
5. RefSeq NM
6. RefSeq NR
7. RefSeq NP
8. NCBI Gene Model

   That is, if the EG gene ID has a corresponding NCBI Gene Model sequence, then associate the NCBI Gene Model Sequence with the MGI Marker.

   However, this can only be done if there is an *exact* match between the MGI Marker Chromosome and the NCBI Gene Model Chromosome. In other words, MGI Markers with a chromosome of UN do not get NCBI Gene Model annotations (because you can't have a coordinate without a chromosome).

Note that the RefSeq loader currently loads all of the RefSeq categories into MGI, so there are no software changes to this loader (yeah!).

Rules for associating MGI-curatable and non-MGI-curatable sequence categories to MGI markers:

1. Delete/reload strategy for those annotated to the EG load reference. That is, the sequence categories associated with MGI markers via the load reference are deleted before the new load data is processed (See "Load Strategy" on page 3.).
2. If the sequence id is already curated to the MGI marker, then do not load it.

Sequence categories that are not designated as MGI-curatable (NM, NR, NP) require EI restrictions to prevent curators from annotating Markers to these types.

5.6 Rules for Associating Additional Non-Sequence Ids to MGI Markers

Once a 1:1 correspondence has been determined for an EG gene and a MGI marker, there are additional non-sequence ids that we want to associate with the MGI marker.

The additional categories are:

1. HomolGene Group ID

5.7 Caching Representative Sequences

Now that we are loading XM, XR, XP, NM, NR and NP sequence associations, we need to adjust the algorithms for determining the repesentative transcript and polypeptide sequences for a Marker.

The representative sequences are displayed in the WI Marker detail page.

Representative sequence assignments are cached in the Sequence/Marker cache table (*SEQ_Marker_Cache*) and loaded via the *seqcacheload* product. This cache is refreshed daily in order to reflect the most current Sequence/Marker associations and their representative sequence "qualifiers".

5.7.1 Representative Transcript

1. longest NM_ or NR_ RefSeq
2. longest non-EST GenBank sequence
3. longest XM_ or XR_ RefSeq
4. longest TIGR, DoTS, or NIA Mouse Gene Index Sequence (in that order of preference).
5. longest EST GenBank sequence

5.7.2 Representative Polypeptide

1. longest SWISS-PROT
2. longest NP_ RefSeq
3. longest TrEMBL
4. longest XP_ RefSeq

Dummy sequences are derived as representative sequences if there are no other ones from the appropriate provider with defined lengths.

6 EI Requirements

6.1 Restrictions on Curating RefSeq Sequences (TR 5654)

1. RefSeq XM, XR are MGI-curatable (adds, updates, deletions).
2. RefSeq NM, NR are MGI-curatable (delete only) if the association reference is not the EG load reference. These are RefSeq NM./NR that have been automatically upgraded from XM/XR due to a sequence merge (see TR5821).
3. All other RefSeq (XP, NP) are not MGI-curatable (no adds, no updates, no deletes).

7 Reports

1. Exclude Set Reports

Each exclude set (1-3) requires a tab-delimited report listing these fields.

- GenBank id
- Sequence type (DNA, RNA)
- Associated MGI ids (comma-delimited, sets 1 & 3 only)
- Associated EG ids (comma-delimited, sets 2 & 3 only)

Sorting:

- exclude set 1: sort by sequence type, Sequence id
- exclude set 2: sort by seuqence type, Sequence id
- exclude set 3: sort by sequence type, MGI id

2. Bucket Reports

   Each correspondence bucket (1:1, 1:N, etc.) requires a tab-delimited report listing the following fields.

   - MGI id
   - MGI symbol
   - MGI chromosome
   - EG id
   - EG symbol
   - EG chromosome

   Sorting:

   - 1:1, 1:0, 1:N, N:N: sort by MGI id.
   - 0:1, N:1: sort by EG id.

3. Chromosome Difference Report

   A row in this report describes a MGI marker and a EG gene that made it into the 1:1 bucket but have different chromosomes (where chromosome UN matches a known chromosome). The fields are:

   - MGI id
   - MGI symbol
   - MGI chromosome
   - EG id
   - EG symbol
   - EG chromosome

   Sort by MGI chromosome, MGI id.

4. XM Discrepency Report

   A row in this report describes

7.1 Public Reports (reports_db)

No changes.

~~These reports require modification to include only certain types of RefSeq ids.~~

~~1. Phenotype report (csmith)~~
   ~~ftp://ftp.informatics.jax.org/pub/reports/MGI_PhenotypicAllele.rpt~~
~~2. Sequence report:~~
   ~~ftp://ftp.informatics.jax.org/pub/reports/MRK_Sequence.rpt~~

7.2 QC Reports (qcreports_db)

No changes.

~~These reports require modification to include only certain types of RefSeq ids.~~

~~1. All Sequences with Marker/Molecular Segment (bobs)~~
   ~~http://shire.informatics.jax.org/data/reports/qcreports_db/output/SEQ_GenBank.rpt~~
~~2. Sequences Annotated to Mouse Markers (dq)~~
   ~~http://shire.informatics.jax.org/data/reports/qcreports_db/output/SEQ_Marker.rpt~~