U.S. National Library of Medicine
National Center for Biotechnology Information

RefSeq Select: a curated set of representative transcripts for mouse protein-coding genes

S Pujar, A. Astashyn, E Cox, O. Ermolaeva, C. Farrell, T Goldfarb, J Jackson, V. Joardar, V. Kodali, K. McGarvey, M Murphy, B Rajput, L Riddick, C Wallin, D Webb, T. Murphy
National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract: Alternative splicing results in multiple transcript variants within a gene. About half of the genes in the RefSeq's mouse annotation have multiple transcript forms. While the full complement of transcripts may be of interest to carry out in-depth studies about a gene, several other analyses, such as comparative genomics and evolutionary studies, often require the use of one transcript per gene. Several genomic databases offer a representative transcript, however, they are often based on simplistic criteria such as transcript length or the oldest accession. At NCBI's (National Center for Biotechnology Information) RefSeq group, we have developed a method to identify a representative or the 'RefSeq Select' based on functionally relevant properties of a gene, including conservation of the coding region and expression. The RefSeq Select pipeline leverages manual curation carried out by a group of expert curators. The Select transcript is usually well-supported by archived data and represents the biology of the gene. We initially released the RefSeq Select set for the human genome and have now expanded the set to include the mouse genome (GRCm38). The RefSeq Select transcripts can be accessed from multiple NCBI resources such as Gene, RefSeq, Genome Data Viewer and annotation files, and are available for bulk download from NCBI's FTP sites. In the long term, we hope to expand the RefSeq Select data further to include a set of key high-value taxa.

https://www.ncbi.nlm.nih.gov/refseq/refseq_select/

RATIONALE

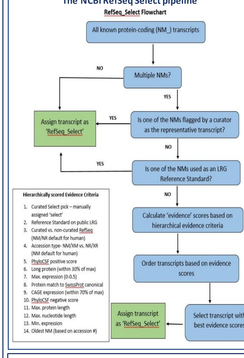
- **65% of mouse protein-coding genes have more than one protein-coding RefSeq, and 35% of genes have more than one curated protein-coding RefSeq (NM_).**

Number of curated (NM) RefSeqs only	Count of genes	Curated and predicted RefSeqs (NM and XM)	Count of genes
One	13327	One	8451
Up to 5	6956	Up to 5	9154
Up to 10	444	Up to 10	2877
More than 10	62	More than 10	1524

- **Biological analyses (such as evolutionary genomics and comparative genomics) may require a single transcript per gene.**
- **A default transcript across mouse genomic resources would facilitate exchange of data between scientists who use different data resources.**
- **RefSeq Select transcripts are based on multiple biological evidence data types (such as CAGE, PhyloCSF, polyA-Seq) not typically considered by automated genome annotation pipelines.**

METHOD

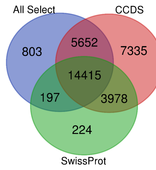
The NCBI RefSeq Select pipeline



SALIENT FEATURES OF REFSEQ SELECT

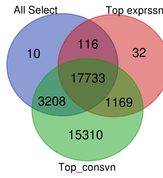
1. **Only curated (NM_) RefSeqs are included in this set.**
2. **The choice of the RefSeq Select is based on multiple criteria, particularly the conservation of the coding region (CDS) and overall expression levels of the transcript.**
3. **The mouse RefSeq Select set is largely generated by the computational pipeline. Curation support is provided in terms of creating new alternatively spliced transcripts that may be assigned as Select by the pipeline, or, updating existing Select transcripts.**
4. **In the long term, the RefSeq Select set is expected to be stable, with no updates unless there are compelling reasons.**

COMPARISON WITH OTHER DATASETS



- **95% of RefSeq Select transcripts are included in the Consensus Coding Sequence (CCDS) dataset, which is considered as a gold standard.**
- **69% of RefSeq Select proteins match the corresponding SwissProt canonical protein for the gene.**

REFSEQ SELECT QUALITY METRICS
Conservation and Expression



- **99% of RefSeq Select transcripts include the most conserved coding region, among all coding RefSeq transcripts for the gene.**
- **85% of RefSeq Select transcripts represent the most expressed protein-coding transcript for the gene.**
- **84% of RefSeq Select transcripts are included in both the above sets.**

DATA ACCESS

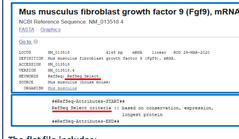
Entrez Query:
"Mus musculus"[Organism] AND refseq_select[filter]

in NCBI webpage (https://www.ncbi.nlm.nih.gov/)
or NCBI Nucleotide webpage
(https://www.ncbi.nlm.nih.gov/nucleotide/)
or NCBI Protein webpage
(https://www.ncbi.nlm.nih.gov/protein/)

Download: From the results page, using the "Send To" option

For individual genes:
Entrez Query: "Mus musculus"[Organism] AND Fgf9[Gene] AND refseq_select[keyword]

Result: This query brings up the flat file of the RefSeq Select transcript



The flat file includes:

- **The 'RefSeq Select' markup in the Keyword section**
- **RefSeq Select criteria in the RefSeq Attributes section, which list the major criteria based on which the transcript was picked as the RefSeq Select**

FUTURE DIRECTIONS

- **Additional markup of RefSeq Select transcripts in:**
  - **RefSeq annotation files (expected in the next annotation of the current mouse reference genome)**
  - **NCBI Gene pages**
- **Update of 5' and 3' ends of RefSeq Select transcripts based on CAGE and polyA-seq data, respectively, where available.**
- **Transition to the next version of the mouse reference genome assembly (GRCm39). The assembly update is expected to become available soon! RefSeq Select data and markups will be available when RefSeq annotates the updated assembly.**

USEFUL LINKS

NCBI RefSeq Select documentation:
https://www.ncbi.nlm.nih.gov/refseq/refseq_select/

Contact us for questions, comments or suggestions:
Gene and RefSeq Feedback Web Form

ACKNOWLEDGEMENTS

This work is supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.