Third Party Annotation (TPA) Sequence Database

What is a Third Party Annotation (TPA) Sequence?

TPA: A database designed to capture experimental or inferential results that support submitter-provided annotation for sequence data that the submitter did not directly determine but derived from GenBank primary data.

TPA records are divided into two categories:

TPA:experimental: Annotation of sequence data is supported by peer-reviewed wet-lab experimental evidence.
TPA:inferential: Annotation of sequence data by inference (where the source molecule or its product(s) have not been the subject of direct experimentation)

TPA database records differ from GenBank and RefSeq records:

GenBank: An archival database of primary nucleotide sequences that were directly sequenced by the submitter.
RefSeq: A curated, non-redundant database that includes genomic DNA, transcript (RNA), and protein products, for major organisms. The sequence data are derived from GenBank primary data, and the annotation is computational, from published literature, or from domain experts.

A TPA sequence is derived or assembled from primary sequence data currently found in the DDBJ/EMBL/GenBank International Nucleotide Sequence Collaboration Databases. It can be genomic or mRNA sequence and can be assembled or derived from primary genomic and/or mRNA sequences. TPA sequences are submitted to DDBJ/EMBL/GenBank as part of the process of publishing biological experiments that include the annotation of existing, primary nucleotide sequences.

Examples of TPA sequences are:

- mRNA assembled from overlapping EST sequences.
- mRNA derived from an unannotated section of genomic sequence by comparison with another known mRNA from a different organism.
- mRNA assembled from overlapping EST sequences, other partial mRNAs, and/or genomic sequences.
- previously unannotated genomic sequence now described with the exons, introns, and coding region information (CDS) of a new gene.

Note: It is required that all new annotations will be experimentally determined to exist, directly or indirectly.

What is a primary sequence?

'Primary' sequences used to assemble a TPA sequence are those that have been experimentally determined and are now publicly available in the GenBank/EMBL/DDBJ databases. These may also be trace data sequences and Whole Genome Shotgun (WGS) sequences. They may not be from a proprietary database. Each primary sequence used to assemble a TPA sequence must be identified by an Accession Number in the submission of the TPA sequence.

Reference sequences may not be cited as data used to build TPA sequences since RefSeqs are not primary data. For example, sequences with Accession Numbers such as NT_112233 or NW_123456 represent contig sequences; the sequences used to assemble these contigs, which can be found at the bottom of contig records, should be cited in a TPA sequence submission. Sequences with Accession Numbers such as XM_345678 or NM_123456 are RefSeqs representing mRNAs that are not experimentally determined and therefore cannot be cited as primary data.

How Do TPA Sequence Records Differ from Other GenBank/EMBL/DDBJ Records?

The display of a TPA sequence is similar to other Collaboration records, but includes the following:

- Keywords:
  TPA;THIRD PARTY ANNOTATION; TPA:experimental
  TPA;THIRD PARTY ANNOTATION; TPA:inferential
- The label 'TPA_exp: or TPA_inf:' at the beginning of each Definition Line.
- PRIMARY field providing the base pair spans of the primary sequences that contribute to the TPA sequence.

Other Features and References are similar to those displayed in other GenBank/EMBL/DDBJ records.

An example of a TPA:experimental is BK000016

An example of a TPA:inferential is BK000554

TPA sequence records are shared by all three Collaboration databases and can be found using typical search methods in EntrezNuc and EntrezProt (ie, submitter name, gene/protein name, Accession Number, etc)

How to Submit TPA Sequence Data

Sequence can be submitted to the TPA database through either BankIt or Sequin:

- BankIt

- Check 'No' to answer the question 'Is This Primary Sequence Data?'.
- Input list of Accession Numbers of all the primary sequences used to assemble or derive the submitted sequence.
- Provide explanation of all experimental evidence or other supporting evidence.
- Complete standard submission process, being sure to annotate all new descriptive information (CDS, protein name, gene name, etc) for the TPA sequence.
- Sequence submission will be labeled as a TPA sequence and will be processed accordingly.

- Sequin

  - Follow standard procedure for Sequin submission.
  - Choose Third Party Annotation from the Sequence Format window under Submission category
  - The Assembly Tracking box will appear with the flatfile display. The primary Accession Number(s) used to assemble/derive the TPA sequence should be entered into this box.
  - Click on Accept; new COMMENT field will appear in the flatfile, which will list the primary sequence Accession Numbers.
  - It is recommended that the submitter note in the email that contains the sequence submission that this is intended for TPA.

General Information

- The entire submitted sequence must be covered by primary sequence data.
- There is no limit on the number of overlapping/adjoining primary sequences that can be cited for a TPA submission.
- If sections of a sequence submitted to TPA have been newly determined by the submitter, those sequences (if they are more than 50 nt) must first be submitted to GenBank, processed, and released to the public before they can be cited as primary sequences
- TPA sequences must cite the same organism as the primary sequence data.

When are TPA sequences released?

- TPA sequences are held confidential until their Accession Numbers or sequence data and/or annotation appear in a peer-reviewed publication in a biological journal.
- No sequence accepted for the TPA database will be released to the public until the submitter notifies us of its publication or we determine independently that such information was published.

What should not be submitted to TPA

- Synthetic constructs such as cloning vectors that use well characterized, publicly available genes, promoters, or terminators; these should be submitted as synthetic sequences for GenBank.

- Microsatellites and related types of repeat regions
- New sequence that updates or changes existing sequence data from another submitter; these should be submitted as new sequences for GenBank.
- Annotation that has arisen from an automated tool, such as GeneMark, tRNA scan or ORF finder, where no further evidence, experimental or otherwise, is presented for the annotation.
- Annotation from in vivo, in vitro, or in silico experimentation that will not be submitted for publication in a peer-reviewed journal.

TPA:experimental

What is a Third Party Annotation:experimental (TPA:experimental) Sequence?

TPA:experimental A database designed to capture experimental results that support submitter-provided annotation for sequence data that the submitter did not directly determine but derived from GenBank primary data.

A TPA sequence is derived or assembled from primary sequence data currently found in the DDBJ/EMBL/GenBank databases. It can be genomic or mRNA sequence, and can be assembled or derived from primary genomic and/or mRNA sequences. TPA sequences are submitted as part of the process of publishing biological experiments that include the annotation of existing nucleotide sequences in the primary sequence database.

Published wet bench experiments are the most critical and valuable data to validate, add, or correct annotation on records in GenBank, particularly for those annotations on genome records that are largely computational results. The TPA:experimental database was established specifically to capture these experimental results in a familiar format. In addition, a publicly accessible TPA record will be linked to a publication that documents that the data are supported by biological experimentation.

Examples of TPA:experimental

- CDS and related annotation applied to a sequence derived from existing genomic and/or mRNA primary data with wet-lab experimental evidence for existence of all or part of the transcript. (For example, RT-PCR, Northern blot experiments)
- CDS and related annotation applied to a sequence derived from existing genomic and/or mRNA primary data with novel sequencing and wet-lab experimental evidence for existence of all or part of the transcript. (For example, RT-PCR, Northern blot experiments)
- CDS and related annotation applied to a sequence derived from existing genomic and/or mRNA primary data with experimental evidence for the presence of the product. (For example, antibody staining or biochemical assays.)
- CDS and related annotation applied to a sequence derived from existing genomic and/or mRNA primary data, with novel sequencing and experimental evidence for the presence of the product. (For example antibody staining or biochemical assays.)

- Confirmation of a new product name and/or function for a coding gene where there is no change to existing annotated exon, mRNA and CDS locations and wet-lab experimental evidence is presented.
- Annotation of non-coding transcripts, such as antisense regulators, with wet-lab experimental evidence for their existence and/or function.
- Annotation of repeat features in association with transposon, retrotransposon, integron, iteron, and insertion sequences with wet-lab experimental evidence.
- Annotation of functional RNA genes, such as tRNAs, scRNAs, etc. with wet-lab experimental evidence.
- Sequences that are part of a collection of annotated members of a gene family, where wet-lab experimental evidence exists for the annotation.

Note: It is required that all predicted annotations will be experimentally supported.

How Do TPA:experimental Sequence Records Differ from TPA:inferential and Other GenBank/EMBL/DDBJ Records?

The display of a TPA record is similar to other Collaboration records, but includes the following:

- Keywords:
  TPA;THIRD PARTY ANNOTATION; TPA:experimental.
- The label 'TPA_exp:' at the beginning of each Definition Line.
- PRIMARY field providing the base pair spans of the primary sequences that contribute to the TPA sequence.

Other Features and References are similar to those displayed in other GenBank/EMBL/DDBJ records.

An example of a TPA:experimental submission is BK000016

TPA sequence records are shared by all three Collaboration databases and can be found using typical search methods in EntrezNuc and EntrezProt (ie, submitter name, gene/protein name, Accession Number, etc)

How to Submit TPA Sequence Data

Sequence can be submitted to the TPA database through either BankIt or Sequin:

- BankIt

  - Check 'No' to answer the question 'Is This Primary Sequence Data?'.
  - Input list of Accession Numbers of all the primary sequences used to assemble or derive the submitted sequence.
  - Provide explanation of all experimental evidence.
  - Complete standard submission process, being sure to annotate all new descriptive information (CDS, protein name, gene name, etc) for the TPA sequence.

- Sequence submission will be labeled as a TPA sequence and will be processed accordingly.

- Sequin

  - Follow standard procedure for Sequin submission.
  - Choose Third Party Annotation from the Sequence Format window under Submission category
  - The Assembly Tracking box will appear with the flatfile display. The primary Accession Number(s) used to assemble/derive the TPA sequence should be entered into this box.
  - Click on Accept; new COMMENT field will appear in the flatfile, which will list the primary sequence Accession Numbers.
  - It is recommended that the submitter note in the email that contains the sequence submission that this is intended for TPA.

What should not be submitted as TPA:experimental

- Sequences with annotation not supported by experimental evidence.
- Synthetic constructs such as cloning vectors that use well characterized, publicly available genes, promoters, or terminators; these should be submitted as synthetic sequences for GenBank.
- Microsatellites and related types of repeat regions
- New sequences that updates or changes existing sequence data from another submitter; these should be submitted as new sequences for GenBank.
- Annotation that has arisen from an automated tool, such as GeneMark, tRNA scan or ORF finder, where no further evidence, experimental or otherwise, is presented for the annotation.
- Annotation from in vivo, in vitro, or in silico experimentation that will not be submitted for publication in a peer-reviewed journal.

TPA:inferential

What is a Third Party Annotation:inferential (TPA:inferential) Sequence?

TPA:inferential A database of sequences annotated by inference, where the source molecule or its product(s) have not been the subject of direct experimentation.

A TPA sequence is derived or assembled from primary sequence data currently found in the DDBJ/EMBL/GenBank databases. It can be genomic or mRNA sequence, and can be assembled or derived from primary genomic and/or mRNA sequences. These sequences are submitted to DDBJ/EMBL/GenBank as part of the process of publishing biological experiments that include the annotation of existing nucleotide sequences in the primary sequence database.

Examples of TPA:inferential

- CDS and related annotation applied to a sequence derived from existing genomic and/or mRNA primary data with reported wet-lab experimental evidence for a homologous molecule but no direct wet-lab experimental evidence. The reported experimental evidence must have been generated by the submission group and must be published in a peer-reviewed journal.
- CDS and related annotation applied to a sequence derived from existing genomic and/or mRNA primary data in addition to novel sequencing with no wet-lab experimental evidence. If the novel sequence was only used to bridge two pieces of sequence, there must be reported wet-lab experimental evidence for a homologous molecule.
- Sequence and annotation covered in a review paper or discussion section, where wet-lab experimental evidence is reported, but not generated by the TPA submitter. The experimental evidence should be reported directly in the review paper or be from a paper by the author of the review paper.
- Annotation of non-coding genes and transcripts with no wet-lab experimental evidence for their existence and/or function but are submitted as part of a study. One or more of the study's sequences should be supported by experimental evidence and be in TPA:experimental or DDBJ/EMBL/GenBank. For an example of this type of study see PubMed 14681587. The annotations cannot be generated by an annotation program such as tRNAscan.
- Annotation of pseudogenes with no wet-lab experimental evidence, when submitted as part of a study that includes sequences of functional homologs of the pseudogene. One or more of the study's sequences should be supported by experimental evidence and be in TPA:experimental or DDBJ/EMBL/GenBank.
- Annotation of pseudogenes that are not part of a gene study but there is experimental evidence. An example of experimental work done to support the description of a pseudogene can be found in PubMed 15908099.
- A sequence submitted as part of a collection of annotated members of a gene family, where wet-lab experimental evidence does not exist for the annotation. One or more members of the set should be supported by experimental evidence and be in TPA:experimental or DDBJ/EMBL/GenBank.
- A sequence representing an assembled genome or naturally occurring plasmid that includes features with assigned gene symbols or product identifiers, where the annotated features may be a mix of experimentally and inferentially determined data.

How Do TPA:inferential Sequence Records Differ from TPA:experimental and Other GenBank/EMBL/DDBJ Records?

The display of a TPA record is similar to other Collaboration records, but includes the following:

- Keywords: TPA; Third Party Annotation; TPA:inferential.
- The label 'TPA_inf:' at the beginning of each Definition Line.

- PRIMARY field providing the base pair spans of the primary sequences that contribute to the TPA sequence.

Other Features and References are similar to those displayed in other GenBank/EMBL/DDBJ records.

An example of a TPA:inferential submission is BK0000554

TPA sequence records are shared by all three Collaboration databases and can be found using typical search methods in EntrezNuc and EntrezProt (ie, submitter name, gene/protein name, Accession Number, etc)

How to Submit TPA Sequence Data

Sequence can be submitted to the TPA database through either BankIt or Sequin:

- BankIt
  - Check 'No' to answer the question 'Is This Primary Sequence Data?'.
  - Input list of Accession Numbers of all the primary sequences used to assemble or derive the submitted sequence.
  - Provide explanation of all experimental evidence.
  - Complete standard submission process, being sure to annotate all new descriptive information (CDS, protein name, gene name, etc) for the TPA sequence.
  - Sequence submission will be labeled as a TPA sequence and will be processed accordingly.

- Sequin
  - Follow standard procedure for Sequin submission.
  - Choose Third Party Annotation from the Sequence Format window under Submission category
  - The Assembly Tracking box will appear with the flatfile display. The primary Accession Number(s) used to assemble/derive the TPA sequence should be entered into this box.
  - Click on Accept; new COMMENT field will appear in the flatfile, which will list the primary sequence Accession Numbers.
  - It is recommended that the submitter note in the email that contains the sequence submission that this is intended for TPA.

What should not be submitted to TPA:inferential

- Sequences with annotation supported by experimental evidence. See TPA:experimental
- Synthetic constructs such as cloning vectors that use well characterized, publicly available genes, promoters, or terminators; these should be submitted as synthetic sequences for GenBank.

- Microsatellites and related types of repeat regions
- New sequence updates or changes existing sequence data from another submitter; these should be submitted as new sequences for GenBank.
- Annotation that has arisen from an automated tool, such as GeneMark, tRNA scan or ORF finder, where no further evidence, experimental or otherwise, is presented for the annotation.
- Annotation from in vivo, in vitro, or in silico experimentation that will not be submitted for publication in a peer-reviewed journal.

FAQs

What is the difference between TPA:experimental and TPA:inferential?

Sequence records in the TPA:experimental database are supported directly by experimental evidence while sequence data and annotation in the TPA:inferential database is indirectly supported by experimental evidence.

What is the difference between TPA and GenBank?

The TPA database consists of sequences that are derived/assembled from primary genomic and/or mRNA sequence that are already represented in the DDBJ/EMBL/GenBank Database. These sequence records include directly or indirectly experimentally supported new annotation, which has been published in a peer-reviewed scientific journal.

The GenBank archival sequence database includes publicly available DNA and RNA sequences submitted from individual laboratories and large-scale sequencing projects. GenBank accession numbers are assigned to these submitted sequences. Submitted sequence data is exchanged between NCBI's GenBank, EMBL Nucleotide Sequence Database (EMBL) and the DNA Data Bank of Japan (DDBJ) to achieve comprehensive coverage. As an archival database, GenBank can be redundant for some loci. GenBank sequence records are owned by the original submitter and can not be altered by a third party.

The major difference between TPA and GenBank is that GenBank records represent nucleotide sequences that have been directly determined by the submitter, whereas TPA records represent nucleotide sequences built from this primary sequence data with new annotation that is directly or indirectly supported by experimental evidence.

What is the difference between TPA and RefSeq?

Both TPA and RefSeq sequences are derived from primary sequence data found in the DDBJ/EMBL/GenBank International Nucleotide Sequence Collaboration Databases. However, while RefSeq annotation is based on additional information available in the

literature and/or an automated computation method, TPA annotation must be supported, directly or indirectly, with experimental evidence by the submitter.

What is primary sequence?

Primary nucleotide sequences have been experimentally determined by their submitter. To be used to build a TPA sequence submission, a primary sequence must be currently publicly available in DDBJ/EMBL/GenBank, trace archive, or Whole Genome Shotgun (WGS) sequence databases.

RefSeq sequences are not considered a primary sequence since they are derived from primary sequence. For example, NCBI Accession numbers that begin with the prefix XM_ (mRNA) and XR_ (non-coding transcript) are model reference sequences produced by NCBI's Genome Annotation project.

Should pseudogenes be submitted to TPA?

Though pseudogenes are common throughout the genomes of organisms, it is often difficult to prove a sequence truly represents a pseudogene. Therefore experimentally supported pseudogenes are acceptable only for TPA:inferential. An example of experimental work done to support the description of a pseudogene can be found in PubMed:15908099 . In addition, pseudogenes may be submitted as part of a study that includes TPA:experimental and.or DDBJ/EMBL/GenBank functional homologs as comparison sequences.

What constitutes an experiment?

The TPA database was created as a repository for annotations that are derived as a result of wet-bench experiments based on existing nucleotide sequence deposited in the DDBJ/EMBL/GenBank databases.

Supporting evidence for valid TPA sequences includes using the nucleotide sequence information to study expression of the RNA in various tissues and to study promoter elements. Conceptual translations have been used for a variety of protein studies, including GST-fusion analyses and enzyme kinetics. These are just some examples of the studies that can be done.

Computational studies on their own do not constitute experimental evidence and must be accompanied by biological experiments that support the new annotation.

Should phylogenetic or population studies be submitted to TPA?

Phylogenetic or population studies that describe a collection of annotated members of a gene family may be submitted as TPA:inferential providing that at least one member of the set is supported by experimental evidence determined by the submitter of the set. The

supporting evidence submission should qualify for either TPA:experimental or DDBJ/EMBL/GenBank.

Can a complete genome created from sequences in the database or improved annotation of a complete genome be submitted to TPA?

A sequence record representing a complete genome can be submitted to TPA:inferential provided the annotated features have been assigned gene symbols or product identifiers. An entire genome's sequence should not be submitted to TPA if only some of the annotation has been improved; just the new annotation's relevant sequence should be submitted.

Can gene families from a single organism or across multiple organisms be submitted to TPA?

Phylogenetic and BLAST (similarity) analysis on their own are not sufficient to support new annotation for either TPA:inferential or TPA:experimental. However, if at least one of the members of the set is supported by experimental evidence determined by the submitter of the set this may be submitted to TPA:inferential. The supporting evidence submission should qualify for either TPA:experimental or DDBJ/EMBL/GenBank.

Can I update my TPA record with additional sequence and/or annotation?

To update an existing record with additional sequence and/or annotation the following requirements should be met:

- Any new sequence must be covered by primary sequence. If the current primaries do not cover this additional sequence new primary accession numbers must be provided.
- Any new annotation must be covered in the existing publication or a new peer-reviewed publication.

Can a TPA:inferential record be changed to a TPA-experimental record?

For a sequence to move from TPA:inferential to TPA:experimental, wet-bench experimental work that supports the annotation presented in the TPA records must be performed and be published in a peer-reviewed journal.

Can annotation of existing data be appropriate elsewhere?

The NCBI RefSeq group does welcome expert comment, critique, and advice from others in the scientific community. There are many individuals and groups contributing to RefSeq on the organism, gene, or gene family level. If you are interested, please contact the NCBI RefSeq group at
refseq-admin@ncbi.nlm.nih.gov or see Refseq.