International Standard

ISO 24480

First edition
2024-11

Biotechnology — Validation of database used for nucleotide sequence evaluation

Biotechnologie — Validation de la base de données utilisée pour l'évaluation de la séquence nucléotidique

Reference number
ISO 24480:2024(en)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 276, *Biotechnology*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

A valid database is important for nucleotide sequence evaluation. The development of inclusivity and exclusivity panels for diagnostics and surveillance using community genomic databases, e.g., Genbank, has been evaluated[1],[2],[3]. However, a specific validation procedure for the databases has yet to be provided.

Considering the current database quality, inclusivity and exclusivity are almost impossible to be validated with ideal accuracy. Therefore, in this document a practical procedure for evaluating the quality of nucleotide sequence database to be used for the development of inclusivity and exclusivity panels is comprehensively described. The degree of data accuracy to be used is determined according to the user's intended test purpose. This evaluation can become a part of the validated diagnostic or surveillance method. Ensuring the quality of the database improves its sufficiency for validating the whole measuring system.

In polymerase chain reaction (PCR) and DNA microarray technologies, nucleotide sequence is used as primers or probes to detect the target nucleic acids. Those technologies utilize initially the hybridization of two single strand DNA molecules with complementary sequences. During the design process of the primers or probes, nucleotide sequence database is used for evaluating specificity and exclusivity of probes or primers. In general, target DNA sequences can be confirmed to match the intended sequences but not others by similarity (homology) search on nucleotide databases with computer tools, for example BLAST.

The validated databases can be used for evaluating specificity of probe or primer sequences and ensuring the selectivity of the qualification and quantification measurement system.

Validation of the entire nucleotide sequence database is not appropriate for the database providers because there are wide varieties of purpose of uses by users. It is almost impossible for the users, however, to evaluate the quality of each data entry especially in huge sequence databases. The database can reflect the fitness for the intended test purpose of users.

This document provides the minimum requirements of a practical procedure for the validation of database used for nucleotide sequence evaluation.

Biotechnology — Validation of database used for nucleotide sequence evaluation

1 Scope

This document describes a practical procedure for nucleotide sequence database evaluation and validation. This document describes minimum requirements for the validation of a nucleotide sequence database. This document is applicable only for databases consisting of entries of nucleotide sequences.

This document is not applicable to the general evaluation of the entire database quality including the quality of each data entry.

EXAMPLE The use of the validated database is for confirming a representative sequence specificity including primers or probes for qualification and quantification of target nucleic acids by conventional polymerase chain reaction (PCR), quantitative polymerase chain reaction (qPCR), digital polymerase chain reaction (dPCR) and microarray technologies.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 20691, *Biotechnology — Requirements for data formatting and description in the life sciences*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
nucleotide sequence specificity
capacity to exclusively recognize a specific nucleic acid target sequence, distinguishing it from other nucleic acids and contaminants

Note 1 to entry: It describes the degree of similarity to specifically match to the nucleotide sequence to be searched by distinguishing it from other nucleotide sequences, and the tendency for a primer or probe with the matched nucleotide sequence to hybridize with its intended target and not hybridize with other non-target sequences.

Note 2 to entry: “sequence specificity” can be considered to be the combination of *inclusivity* (3.4) and *exclusivity* (3.5)

3.2
selectivity
extent to which a method can determine particular analyte(s) in a mixture(s) or matrice(s) without interferences from other components of similar behaviour

Note 1 to entry: Selectivity is the recommended term in analytical chemistry to express the extent to which a particular method can determine analyte(s) in the presence other components. Selectivity can be graded. The use of the term “specificity” for the same concept is to be discouraged as this often leads to confusion.

Note 2 to entry: Sequence specificity in molecular biomarker analysis is differentiated from chemical analyte selectivity.

[SOURCE: ISO 16577:2022, 3.3.73]

3.3
sequence similarity
proportion of matched number of units, including nucleotides and amino acids, to the number of units in specified regions between two nucleic acids or proteins

Note 1 to entry: gap or deletion can be considered to compare the unit sequence of the regions between two nucleic acids or proteins.

Note 2 to entry: "sequence similarity" can be evaluated simply based on a proportion of matched number of units, whereas the term "homology" contains biological meaning in the comparison of two nucleic acids or proteins.

3.4
inclusivity
property of a nucleotide sequence to show high *sequence similarity* (3.3) specifically with intended target nucleotide sequence

Note 1 to entry: The term "inclusivity" is used as same meaning of "sensitivity" in some cases [4].

3.5
exclusivity
property of a nucleotide sequence to show low *sequence similarity* (3.3) with those excluding intended target nucleotide sequences

3.6
nucleic acid test
NAT
technique used to detect or quantify a target nucleic acid with specific sequence, by using of oligonucleotide as a primer or probe

3.7
representative sequence
group of nucleotide sequence data containing one or more target sequences in a complete or partial sequence intended for detection or quantification

3.8
undesirable sequence
group of nucleotide sequence data containing one or more nucleotide sequences, which are potentially either influencing or intentionally excluded, or both, for detection or quantification

3.9
intended test purpose
purpose of nucleic acid detection or quantification using oligonucleotide, e.g., primers or probes, whose design is evaluated by using validated nucleotide sequence databases containing representative and *undesirable sequences* (3.8)

3.10
exploring key
data used for examining data entries stored in database

EXAMPLE Key words, sequence data, taxon data, tissue name etc.

3.11
inclusivity database
database used for evaluating *inclusivity* (3.4) of a specified nucleotide sequence

3.12
exclusivity database
database used for evaluating *exclusivity* (3.5) of a specified nucleotide sequence

3.13
provenance information
information that documents the history of a described object and related described activities, and that contains information about the origin or source of the described object, any changes that can have taken place since it was originated, and who has had custody of it since it was originated

[SOURCE: ISO/TS 23494-1:2023, 3.13]

3.14
finalized provenance information
provenance information transformed into a representation specified by the common provenance model, and which is prepared to be conserved or archived and which is considered as being immutable

Note 1 to entry: Finalized provenance information is a subset of provenance information.

[SOURCE: ISO/TS 23494-1:2023, 3.5]

3.15
basic local alignment search tool
BLAST
sequence comparison algorithm optimized for speed that is used to search sequence databases for optimal local alignments to a query

[SOURCE: ISO 20813:2019, 3.1 modified — Notes to entry have been deleted.]

3.16
massively parallel nucleotide sequencing
next generation sequencing
NGS
high throughput nucleotide sequencing method capable of determining multiple DNA sequences simultaneously and in parallel

Note 1 to entry: The data from a single massively parallel sequencing analysis comprises of millions of sequences and the output is a file containing all sequences.

[SOURCE: ISO 16577:2022, 3.7.10, modified — "whole genome sequencing" and "WGS" have been deleted.]

4 General

There are significant differences between the inclusivity and exclusivity confirmation roles. A database used for the inclusivity analysis shall cover all intended sequence entries. In addition, other recognized unintended sequence entries should be contained to show the sequence similarity is specific to the intended sequence entries in a recognized extent, although exclusivity cannot be confirmed only with the recognized undesirable sequences. Entries with high and reliable quality of both intended sequences and recognized undesirable sequences should be included in the database[3]. Quality of the sequence entries for inclusivity analysis is described in the validation of inclusivity database.

On the other hand, a database used for exclusivity analysis should include as many sequence entries as possible including related and not likely related ones. Even though the entries are not likely related sequences, the database should present those sequences because they can contain sequences that can be unintentionally hybridized by the primers or probes in non-specific manner and rise the amplification background. The quality of the sequence entries for the exclusivity analysis is described in the validation of the exclusivity database (for details, see 7.3). One example for uses of a validated database is for an *in silico* analysis evaluating nucleotide sequence specificity in designing primers or probes for nucleic acid measurement including various PCR-based methods and microarray analysis, i.e., inclusivity and exclusivity analyses. Databases for the inclusivity analysis are used for evaluation on how the designed primers and probes can work to distinguish a target in the nucleic acid measurement methods. A database for the exclusivity analysis is used for evaluating the possibility of the designed primers and probes showing non-specific detection or quantification in the nucleic acid measurement methods. Therefore, when users validate a database, i.e., confirm that the requirements of the database for a specific intended use have been fulfilled,

they shall validate the database by fulfilling the requirements with those two roles, namely inclusivity and exclusivity confirmations.

Thus, users can specify two independent databases fitting to the two roles in many cases. There are specific requirements for databases used for inclusivity evaluation (inclusivity database) and databases used for exclusivity evaluation (exclusivity database) depending on the roles (see Clause 6 and Clause 7). It does not limit, however, to place both roles in one database. In some cases, the inclusivity database and the exclusivity database can be the same, for example, when the nucleic acid measurement method is used in specified environment where the available nucleotide sequences are well characterized and limited.

5 Common requirements of the database

Databases should implement the FAIR principles[8]. When constructing databases, they shall be constructed in accordance with ISO 20691. Each entry shall be identified with appropriate identifier(s), for example, scientific name, accession number for registered sequence in public database, unique number with authorship for the non-registered sequence. The data format of the database shall be machine-readable and generated in as accessible format[7] for nucleotide sequence analysis, for example, fasta or fastq format (see Annex A). The database shall be accessible by search functions, for example, local BLAST search, taxon search, text search of gene names.

NOTE In some cases, inter-jurisdictional consideration is important, depending on the characteristics of data entries.

6 Inclusivity database: database for inclusivity evaluation

6.1 Quality criteria

Quality criteria for each data entry in inclusivity database shall be determined by the user, considering the intended test purpose of the NAT.

An example of a whole validation process for an inclusivity database is described in Annex B.

6.2 Requirements of an inclusivity database

High quality target sequences, which are intended to be detected or quantified by the NAT measurement system, shall be sufficiently populated in the inclusivity database with numbers of the representative sequences to cover sequence discrimination by the measurement system.

Entries shall include representative sequences of the target and sequences that are undesirable to be detected by the measurement system, for showing that the sequence similarity is specific to the intended sequence entries to a recognized extent (see Annex E as an example for dataset verification commands). Representative sequences of the target should be stored in multiple entries, for example, the sequences of a target analysed by several different laboratories.

NOTE Redundancy of the sequences in the database can be allowed.

The inclusivity database shall contain sequences with high sequence similarity (species specific and non-specific; target to be included or excluded) based on criteria determined by users.

Users should take into account to include sequences with point mutations when applicable. In some cases, sequences from the same taxonomic rank such as genus, species, subspecies, homologous genes, or variants can be selected as sequences with high sequence similarity.

6.3 Individual data quality indicators

6.3.1 Data provenance and updates

The inclusivity database can be updated periodically. The updated database shall be validated by following the procedures described in this document. Date and time of the update shall be documented.

When database entries are documented using finalized provenance information according to the ISO 23494 series, the collected finalized provenance information can be used for quality assessment of the entries. The finalized provenance information can be the most significant indicator for the quality assessment.

6.3.2 Length of the entries

The entries in the database shall have appropriate length, which needs to be longer than the minimum length of the nucleotide sequence specified by the database user within the quality criteria.

NOTE Data entries with short sequences, such as raw data of NGS can be impeditive to the interpretation of results of the evaluation for validation and actual use of the database, for example, primer and probe design.

Long nucleotide sequences, such as long contigs and whole genome sequence data, which are outside of the maximum quality criteria should be excluded. Although longer sequences are useful for the exclusivity evaluation, which is the main purpose of exclusivity database usage (see below), they are less applicable for validation of inclusivity database, primer and probe design and the evaluation of their inclusivity.

6.3.3 Number of unidentified nucleotides (N)

The nucleotide sequence quality should meet certain criteria when incorporating NGS data into the inclusivity database because NGS data, especially raw data, have not only a short length of the nucleotide sequence (see 6.3.2) but also a higher possibility of deletion and ambiguous data (for example, unassigned data marked as “N”).

NOTE Quality Value in FASTQ data format can be used for estimating whether database entries can be used in the inclusivity database.

6.4 Validation of the inclusivity database

The validation plan of the inclusivity database shall be established, implemented and documented.

The validation of the inclusivity database is the result of confirming whether each data entry is appropriate or not to be included in the database. Therefore, each data entry in the inclusivity database should be confirmed by human curation to ensure that it fulfils the determined quality criteria. In cases where human curation is used, it shall be performed at the early stage of the inclusivity database validation to verify data entries, i.e., target genes and species, entries format, length of sequence, and literature references. When the human curation is eliminated in the validation, an alternative procedure to confirm the conformity of each data entry shall be used and documented.

During validation, the quality of the inclusivity database can be confirmed by searching it with representative sequences and evaluating the number of correct and incorrect best matches[5]. Some quality indicators and methods for the evaluation are described in the previous report[6].

The validation search procedure for the inclusivity database shall be able to retrieve correct best matches of the representative sequence and related entries.

The inclusivity database shall be accessible to users and can be retrieved in a popular format with correct header, nucleotide sequence, length of sequence which can be analysed using tools such as local BLAST search. Thereby, the quality of sequence entries, i.e., perfect match and mismatch of the representative sequence such as highly similar sequences (>99 % similarity), can be confirmed.

7 Exclusivity database: Database for the exclusivity evaluation

7.1 Quality criteria

The user shall determine the criteria for the exclusivity database to evaluate its coverage, i.e., the representative and undesirable data and how the database sufficient for the exclusivity evaluation. Quality criteria should be indicated with consistent information (e.g., number of entries, sequence length as many as possible to draw conclusions of the validation results).

The user shall also document exploring keys to confirm whether sufficient entries for the exclusivity evaluation are contained in the database.

EXAMPLE Nucleotide sequence, gene name, variant identification including isoform names, gene family name, functional ontology, and taxon data are examples that can be set for the exploring keys.

These criteria and exploring keys vary depending on the purpose of database.

The example of a whole validation process of an exclusivity database is described in Annex C.

7.2 Requirements of the exclusivity database

For the exclusivity database, as many as possible database(s) containing sufficient sequence data related to representative and undesirable data shall be selected. The sufficient number of entries varies depending on the purpose of exclusivity database usage. Taxonomic variation shall be included as much as possible in the exclusivity database depending on the purpose.

The developer(s) or user(s) shall determine the data quality parameters, including but not limited to the following:

a) sequence length;

b) predicted sequence;

c) number of sequences;

d) redundancy;

e) number of unidentified nucleotides (N).

NGS data, especially raw data, have short and higher possibility of deletion and ambiguous data (for example, unassigned data marked "N"). Nucleotide sequence quality should meet certain criteria when incorporating NGS data into the exclusivity database in common with the inclusivity database (see 6.3.2).

For the exclusivity database, preparing a subset database or narrowing down by searching with a specific selection of parameters, e.g., excluding short sequences or synthetic oligo nucleotides, or key words such as the gene family name or disease name can be specified for validation, verification and actual use for designing primers or probes.

The exclusivity database shall contain nucleotide sequences of species that are presumably relevant to the representative sequence and purpose of the intended test purpose of the NAT.

EXAMPLE 1 Sequence from the same taxonomic rank such as class "mammal", and genus / species "homo sapiens"

The exclusivity database shall contain nucleotide sequence(s) of unrelated species documented in the inclusivity database. These entries are critical for evaluating exclusivity.

EXAMPLE 2 Sequence from different taxonomic rank such as kingdom (plant or fungus, phylogenetically) far different from target species.

7.3 Validation of the exclusivity database

The validation plan of the exclusivity database shall be established, implemented and documented.

The validation search procedure for the exclusivity database shall be performed by executing a query with representative sequences and determined exploring keys. The selected type and version of database including date and time of execution shall be documented.

By querying with representative sequences, the target sequences and its closely related sequences, for example, variant, isoforms, and homologues, can be retrieved and confirmed as entries in the exclusivity database. Hence, the quality of sequence entries, i.e., perfect match and mismatch of the representative sequence such as highly similar sequences (>99 % similarity) or less similar sequences (<99 % similarity) can be observed and verified.

By querying with exploring keys, the related sequences that need to be discriminated can be retrieved and confirmed as entries in the exclusivity database. The number of these entries is critical for the purpose of the exclusivity database usage, namely the evaluation of exclusivity.

In contrast with the validation of the inclusivity database, human curation comes at the later stage prior to validation of the exclusivity database. A human curation should be performed to verify data entries, i.e., the exclusivity of target genes and species, the format of entries, and literature references.

Throughout the validation processes, the interpretation of the search results is most important. The strategy for the interpretation shall be described and documented in the validation plan, and the results of the interpretation shall be documented with objective evidence sufficient for the conclusion. Scientific references describing the representative sequences and circumjacent information, e.g., in the field of comparative biology, can be useful for showing the reason of the conclusion of interpretation.

EXAMPLE To develop the NAT for detection of subtype H1N1 of influenza virus, an exclusivity database is valid if the database includes not only target sequences, e.g., influenza H1N1, but also related genes undesirable to be detected, e.g., H3 for HA, N2 for NA sequences of influenza, rhinovirus and coronavirus.

For the NAT for the detection of animal species targeting rRNA, the exclusivity database shall contain all available sequences of rRNA.

The exclusivity database can be validated if the database includes not only major animal sequences, but also contains other eukaryotes such as fungus and plant rRNA.

8 Validation report

The validation report shall include, but not be limited to, the following:

a) the date of the report;

b) the intended test purpose of the NAT;

c) the inclusivity database validation plan including quality criteria, and representative sequences used;

d) the results of the inclusivity database validation;

e) the exclusivity database validation plan including the procedure to confirm the conformity of each data entry, when human curation cannot be applied, quality criteria, exploring keys used, the version of database used and date and time of execution;

f) the results of the exclusivity database validation;

g) a reference to this document, i.e., ISO 24480:2024;

h) any deviations from the procedure;

i) any unusual features observed;

j) the date of the validation.

The validation report can be stored in a digital format in the same location with the validated inclusivity database and validation files for exclusivity database.

The example for whole processes to use inclusivity and exclusivity databases that is helpful for verifying the requirements of the validation report is described in Annex D.

Annex A
(informative)

Example of a data entry format

A.1 Basics of nucleotide sequence databases

In the nucleotide sequence databases that can be used for conforming a representative sequence specificity including primers or probes for qualification and quantification of target nucleic acids, each entry is identified with appropriate identifier(s), such as scientific name, accession number for registered sequence in public database, unique number with authorship for the non-registered sequence. The data format of the database is generated in an accessible format for nucleotide sequence analysis, e.g., fasta or fastq format. A database can have search interfaces for users and can be searchable by search functions, for example, local BLAST search, taxon search, text search of gene names.

A.2 Examples of data entry

A.2.1 Valid example

>NM_001001491.2 Mus musculus tropomyosin 4 (Tpm4), mRNA

ACCGCAAGTATGAGGAGGTTGCTCGTAAGTTGGTCATCCTGGAGGGTGAGCT GAAGAGAGCAGAGGAG
AGGGCGGAGGTATCTGAACT AAAGTGTGGTGACCTGGAAGAAGAGCTCAAGA ATGTAACTAACAATCT
GAAATCACTGGAGGCTGCTTCTGAA

A.2.2 Non-valid example

The following sequences has no header or not in an applicable format (for example, fasta)

>

ACCGCAAGTATGAGGAGGTTGCTCGTAAGTTGGTCATCCTGGAGGGTGAGCT GAAGAGAGCAGAGGAG
AGGGCGGAGGTATCTGAACT AAAGTGTGGTGACCTGGAAGAAGAGCTCAAGA ATGTAACTAACAATCT
GAAATCACTGGAGGCTGCTTCTGAA

>ACCGCAAGTATGAGGAGGTTGCTCGTAAGTTGGTCATCCTGGAGGGTGAGCT GAAGAGAGCAGAGGA
GAGGGCGGAGGTATCTGAACT AAAGTGTGGTGACCTGGAAGAAGAGCTCAAGA ATGTAACTAACAATC
TGAAATCACTGGAGGCTGCTTCTGAA

Annex B
(informative)

Example for the validation of an inclusivity database (Inclusivity)

B.1 General

This annex describes an example for the validation of the inclusivity database for an inclusivity evaluation. This example starts with raw data processing, including human curation and validation, ends with validated inclusivity database as final output.

B.2 Inclusivity database validation workflow

The workflow for inclusivity database validation is shown in Figure B.1.

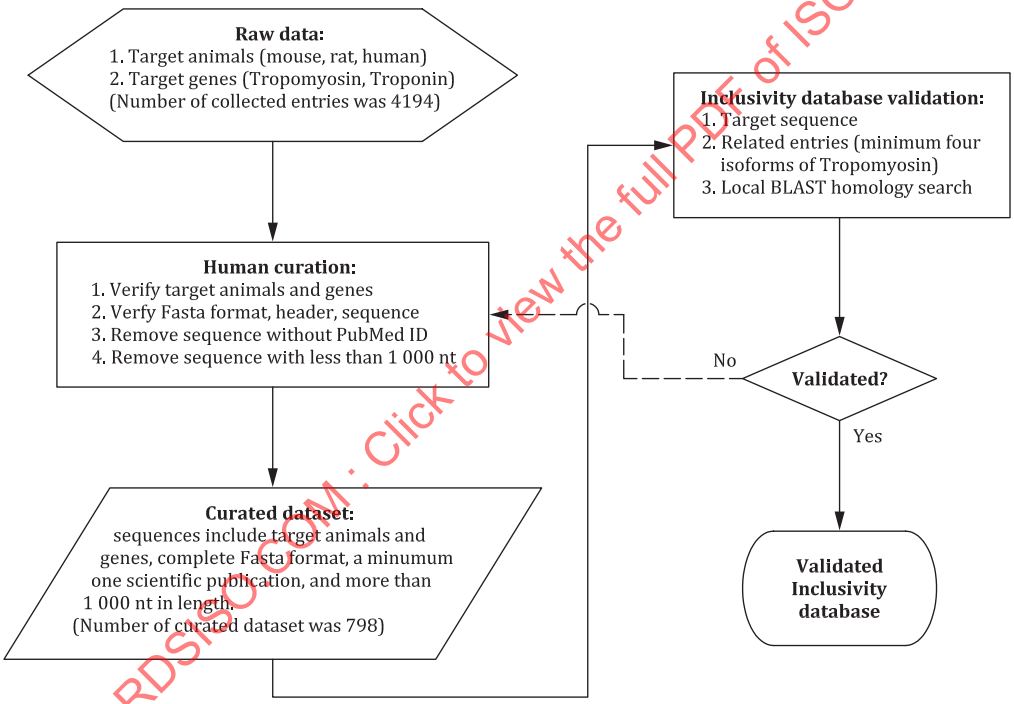

Figure B.1 — Inclusivity database validation workflow (adopted from Methodology for data validation 1.0 [6])

B.3 Inclusivity database validation plan

B.3.1 Intended test purpose of NAT

Detection and identification of Tropomyosin (TPM) isoforms in rodent.

B.3.2 Acceptance criteria

— Database contains the target and its homologous sequences.

— Database contains more than four isoforms of Tropomyosin.

— Data entry is in fasta format. Entry should contain header.

— Length of sequence should be more than 1 000 bases.

— Random search target nt sequence is not required.

— Local BLAST search should be applicable.

B.4 Purpose of database usage

Evaluate inclusivity of primer sequence to detect specific TPM gene in rodent.

B.5 Data entry

Below is the example of Tropomyosin and Troponin genes dataset from mouse, rat and human.

>NM_001164255.1 Mus musculus tropomyosin 1, alpha (Tpm1), transcript variant Tpm1.10, mRNA

>NM_001164256.1 Mus musculus tropomyosin 1, alpha (Tpm1), transcript variant Tpm1.12, mRNA

>NM_001293748.1 Mus musculus tropomyosin 3, gamma (Tpm3), transcript variant 2, mRNA

>NM_001253738.1 Mus musculus tropomyosin 3, gamma (Tpm3), transcript variant 3, mRNA

>NM_001001491.2 Mus musculus tropomyosin 4 (Tpm4), mRNA

>NM_001277903.1 Mus musculus troponin T1, skeletal, slow (Tnnt1), transcript variant 1, mRNA

>NM_001130178.2 Mus musculus troponin T2, cardiac (Tnnt2), transcript variant 5, mRNA

>NM_011620.3 Mus musculus troponin T3, skeletal, fast (Tnnt3), transcript variant 7, mRNA

>NM_001301336.1 Rattus norvegicus tropomyosin 1 (Tpm1), transcript variant Tpm1.1, mRNA

>NM_001301736.1 Rattus norvegicus tropomyosin 1 (Tpm1), transcript variant Tpm1.12, mRNA

>NM_173111.1 Rattus norvegicus tropomyosin 3 (Tpm3), transcript variant Tpm3.1, mRNA

>NM_001301285.1 Rattus norvegicus tropomyosin 3 (Tpm3), transcript variant Tpm3.12, mRNA

>M34136.1 Rat brain alpha-tropomyosin (TMBr-3) mRNA, 3' end

>NM_012676.1 Rattus norvegicus troponin T2, cardiac type (Tnnt2), mRNA

>NM_001270673.1 Rattus norvegicus troponin T3, fast skeletal type (Tnnt3), transcript variant 7, mRNA

>NM_001018006.2 Homo sapiens tropomyosin 1 (TPM1), transcript variant Tpm1.7, mRNA

>NM_003289.4 Homo sapiens tropomyosin 2 (TPM2), transcript variant Tpm2.2, mRNA

>NM_001278188.2 Homo sapiens tropomyosin 3 (TPM3), transcript variant 6, mRNA

>NM_001367837.2 Homo sapiens tropomyosin 4 (TPM4), transcript variant 4, mRNA

>NM_001126133.3 Homo sapiens troponin T1, slow skeletal type (TNNT1), transcript variant 3, mRNA

>NM_001276345.2 Homo sapiens troponin T2, cardiac type (TNNT2), transcript variant 5, mRNA

>NM_001367850.1 Homo sapiens troponin T3, fast skeletal type (TNNT3), transcript variant 15, mRNA

B.6 Human curation

Example for a list of scientific publication for each entry.

>NM_001164255.1 Mus musculus tropomyosin 1, alpha (Tpm1), transcript variant Tpm1.10, mRNA

PubMed ID: 32376900, 30700554, 30642949, 30567734, 30242109, 25369766, 7522680, 1631061, 2521606, 3244365

>NM_001001491.2 Mus musculus tropomyosin 4 (Tpm4), mRNA

PubMed ID: 28134622, 25369766, 19118250, 18270576, 18036591, 17968984, 16765662, 16236705

>NM_001277903.1 Mus musculus troponin T1, skeletal, slow (Tnnt1), transcript variant 1, mRNA

PubMed ID: 32059926, 31148174, 30979776, 29931346, 11003710, 10594179, 10449439, 10095098, 9651500, 9107680

>NM_001301336.1 Rattus norvegicus tropomyosin 1 (Tpm1), transcript variant Tpm1.1, mRNA

PubMed ID: 30462572, 28002632, 25369766, 24362038, 23609439, 23420843, 2022655, 2320008, 3352602, 3558368

>NM_001301285.1 Rattus norvegicus tropomyosin 3 (Tpm3), transcript variant Tpm3.12, mRNA

PubMed ID: 28677753, 25369766, 22749829, 22114352, 21036167, 20458337, 9473354, 8674141, 7704029, 8206382

>NM_012676.1 Rattus norvegicus troponin T2, cardiac type (Tnnt2), mRNA

PubMed ID: 32297828, 25771144, 24364879, 23357173, 23012479, 7898523, 8205619, 1433301, 2530435, 2760070

>NM_003289.4 Homo sapiens tropomyosin 2 (TPM2), transcript variant Tpm2.2, mRNA

PubMed ID: 32957762, 31487691, 30545627, 30535593, 25369766, 20301465, 20301436, 1631061, 1304342, 2059197

>NM_001367837.2 Homo sapiens tropomyosin 4 (TPM4), transcript variant 4, mRNA

PubMed ID: 32296183, 29455030, 28431393, 28330616, 28134622, 25369766, 1286667, 1836432, 3612796, 3865200

>NM_001367850.1 Homo sapiens troponin T3, fast skeletal type (TNNT3), transcript variant 15, mRNA

PubMed ID: 29596868, 26774798, 26915936, 25342443, 23936387, 18629027, 8681137, 8062920, 8172653, 7118902

B.7 Inclusivity database validation

The inclusivity database is validated according to the following criteria:

— Target sequence with regards to target animals and genes (passed/failed)

— Related entries with minimum four isoforms of Tropomyosin (passed/failed)

— Applicability for local BLAST search (passed/failed)

When all criteria are passed, the inclusivity database is valid and can be used for the inclusivity evaluation. An example of use of inclusivity database is shown in Annex D.

Annex C
(informative)

Example on validation of an exclusivity database (Exclusivity)

C.1 General

This annex describes the example on validation of exclusivity database for exclusivity evaluation. This example includes quality criteria determination, database selection, validation, and validated exclusivity database as final output.

C.2 Exclusivity database validation workflow

Detection and identification of Tropomyosin (TPM) isoforms in rodent.

The workflow for exclusivity database validation is shown in Figure C.1.

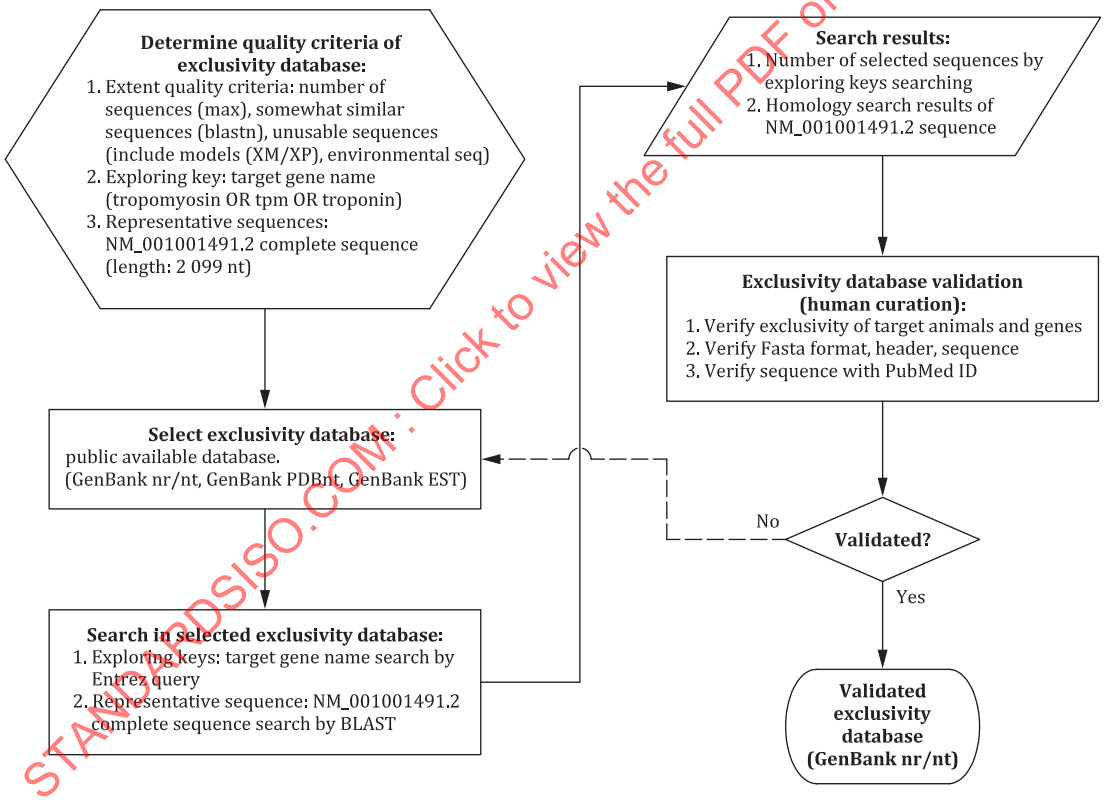

Figure C.1 — Exclusivity database validation workflow

C.3 Exclusivity database validation plan

C.3.1 Intended test purpose of NAT

— Detection and identification of Tropomyosin family from various animals, including rodents and human.

— Homologous genes, i.e., Troponin are included.

C.3.2 Acceptance criteria

— Database contains the representative sequence and its homologous sequences from various species and taxon.

— Database contains all four isoforms of Tropomyosin as set in the exploring keys.

C.4 Purpose of database usage

Evaluate exclusivity of primer sequence to detect specific TPM 4 gene in rodent.

C.5 Database selection

Below are the examples of selected databases:

— GenBank nucleotide collection (nr/nt)

— GenBank Protein Data Bank nucleotide (PDBnt)

— GenBank Expressed Sequence Tags (dbEST)

C.6 Exclusivity database validation

C.6.1 General

Database selection, exploring key search, and BLAST search were performed on November 26, 2020. All data were obtained on the same date.

C.6.2 Results of exploring key search

Number of selected sequences using Entrez nucleotide query (exploring key: tropomyosin OR tpm OR troponin) resulted in 137 474 entries including from animals, plants, fungi, protists, bacteria, archaea, and viruses. According to the acceptance criteria, all isoforms of Tropomyosin, i.e., Tropomyosin 1, Tropomyosin 2, Tropomyosin 3, Tropomyosin 4 and other genes, i.e., Troponin were confirmed and therefore, the exploring key was acceptable.

C.6.3 Results of representative sequence search in BLAST

C.6.3.1 Results using GenBank nucleotide collection (nr/nt) database

The result of sequence similarity verified by GenBank nucleotide collection (nr/nt) database with NM_001001491.2 complete sequence is shown in Table C.1.

Table C.1 — Sequence similarity of NM_001001491.2 complete sequence

Description	Query cover	E Value	Per. Ident	Acc. Len	Accession
Mus musculus tropomyosin 4 (Tpm4), mRNA	100 %	0,0	100	2 099	NM_001001491.2
Mus musculus tropomyosin 4, mRNA (cDNA clone MGC:95675...	99 %	0,0	100	2 109	BC070421.1
Mus musculus tropomyosin 4, mRNA (cDNA clone MGC:38384...	99 %	0,0	99,33	2 118	BC023701.1
Mus musculus cDNA clone IMAGE:5322101, containing frame-shift...	99 %	0,0	99,28	2 107	BC032174.1
Mus musculus cDNA clone IMAGE:5322155	99 %	0,0	99,28	2 106	BC032175.1
NOTE Explanations of query cover, E value and percent identity can be found in[10][11].					

Table C.1 *(continued)*

Description	Query cover	E Value	Per. Ident	Acc. Len	Accession
Mus musculus tropomyosin 4, mRNA (cDNA clone MGC:38326...	98 %	0,0	99,42	2 082	BC023827.2
PREDICTED: Mus caroli tropomyosin 4 (Tpm4), mRNA	99 %	0,0	96,82	2 138	XM_021169413.2
Mus musculus predicted gene 7809 (Gm7809) pseudogene on...	99 %	0,0	95,26	2 304	NG_060233.1
Mus musculus BAC clone RP24-113B3 from chromosome 17, complete...	99 %	0,0	95,26	182 923	AC154483.2
PREDICTED: Mus pahari tropomyosin 4 (Tpm4), mRNA	91 %	0,0	94,11	2 170	XM_029531890.1
NOTE Explanations of query cover, E value and percent identity can be found in[10][11].					

C.6.3.2 Results using GenBank protein data bank nucleotide (PDBnt) database

No hit or significant similarity was found for the complete sequence of NM_001001491.2 *Mus musculus* tropomyosin 4 (Tpm4), mRNA.

C.6.3.3 Results using GenBank Expressed Sequence Tags (dbEST) database

The result of sequence similarity verified by dbEST database with NM_001001491.2 complete sequence is shown in Table C.2.

Table C.2 — Sequence similarity of NM_001001491.2 complete sequence

Description	Query cover	E Value	Per. Ident	Acc. Len	Accession
AV122663 Mus musculus C57BL/6J 10-day embryo Mus musculus cDNA...	49 %	0,0	97,08	1 013	AV122663.2
AGENCOURT_10109747 NIH_MGC_134 Mus musculus cDNA clone...	45 %	0,0	98,14	969	BU516716.1
AGENCOURT_10128972 NIH_MGC_134 Mus musculus cDNA clone...	55 %	0,0	91,7	1 228	BU511567.1
AGENCOURT_10171374 NIH_MGC_134 Mus musculus cDNA clone...	45 %	0,0	97,9	979	BU519236.1
BB610989 RIKEN full-length enriched, 10 days embryo Mus muscul...	49 %	0,0	95,08	1 039	BB610989.1
AGENCOURT_10157382 NIH_MGC_134 Mus musculus cDNA clone...	45 %	0,0	97,47	974	BU520254.1
AGENCOURT_8747206 NIH_MGC_130 Mus musculus cDNA clone...	44 %	0,0	97,56	940	BU151829.1
AGENCOURT_11260355 NIH_MGC_135 Mus musculus cDNA clone...	42 %	0,0	98,99	892	CB194275.1
AGENCOURT_11294100 NIH_MGC_164 Mus musculus cDNA clone...	42 %	0,0	99,33	891	CA980249.1
AGENCOURT_8747517 NIH_MGC_130 Mus musculus cDNA clone...	44 %	0,0	97,87	943	BQ885763.1
NOTE Explanations of query cover, E value and percent identity can be found in[10][11].					

C.7 Human curation

C.7.1 GenBank nucleotide collection (nr/nt) database

GenBank nucleotide collection (nr/nt) database was validated and curated according to the following requirements.

— A total of 5 000 entries (maximum) were resulted from the complete sequence of NM_001001491.2 *Mus musculus* tropomyosin 4 (Tpm4), mRNA.

— GenBank nr/nt contained all isoforms of Tropomyosin and Troponin.

— GenBank nr/nt contained various taxon such as *Mus musculus, Rattus norvegicus, Macaca fascicularis, Homo sapiens, Salmo salar, Danio rerio, Pongo abelii, Gadus chalcogrammus, Pennahia argentata, Sus scrofa, Oryctolagus cuniculus, Scyliorhinus rotifer, Takifugu rubripes, Epinephelus coioides, Siniperca chuatsi, Oreochromis mossambicus, Xenopus laevis, Gallus gallus, Ambystoma mexicanum, Taeniopygia guttata, Ovis aries, Bos taurus, Canis lupus*.

— Top hit in both Tables showed the same accession number with confidence query cover, E value and percent identity[10,11].

— Top hits were matched, and PubMed ID was confirmed.

All requirements were confirmed. Therefore, GenBank nucleotide collection (nr/nt) database was valid and could be used as exclusivity database for exclusivity evaluation. Example of use of exclusivity database is shown in Annex D.3.

C.7.2 GenBank Expressed Sequence Tag (dbEST) database

GenBank Expressed Sequence Tag (dbEST) was validated and curated according to the following requirements.

— A total of 2 054 sequences were resulted from the complete sequence of NM_001001491.2 *Mus musculus* tropomyosin 4 (Tpm4), mRNA.

— No perfect match to the description of the complete sequence of NM_001001491.2 *Mus musculus* tropomyosin 4 (Tpm4), mRNA despite greater confidence of E value and percent identity.

— For complete sequence, 2 099 nt, query cover was 41 % at maximum because EST are typically short with less than 1 000 nt.

— Pubmed IDs were found for the top hits (no scientific publication).

All requirements could not be confirmed. Therefore, GenBank Expressed Sequence Tag (dbEST) database was not valid and could not be used as exclusivity database for exclusivity evaluation.

Annex D
(informative)

Example of use of an inclusivity database and an exclusivity database

D.1 General

This annex describes the example of use of inclusivity database and exclusivity database for evaluation of PCR primer specificity and exclusivity, respectively.

D.2 Evaluation of PCR primer specificity using an inclusivity database

The inclusivity database was used to evaluate specificity of PCR primer to detect mouse Tropomyosin 4. Specific PCR primer was designed from complete sequence of NM_001001491.2 *Mus musculus* tropomyosin 4 (Tpm4), mRNA. Examples of the selected PCR primers were the following:

— Forward primer : ACCGCAAGTATGAGGAGGTTG

— Reverse primer : TTCAGAAGCAGCCTCCAGTG

Example of the results from local BLAST search for each primer are shown in Table D.1 and Table D.2.

Table D.1 — Local BLAST search results of forward primer of mouse Tropomyosin 4

Description	Score	E
Sequences producing significant alignments:	(Bits)	Value
BC023701.1_Mus_musculus_tropomyosin_4,_mRNA_(cDNA_clone_MGC:38384...	42,1	0,005
BC070421.1_Mus_musculus_tropomyosin_4,_mRNA_(cDNA_clone_MGC:95675...	421	0,005
BC023827.2_Mus_musculus_tropomyosin_4,_mRNA_(cDNA_clone_MGC:38326...	42,1	0,005
NM_001001491.2_Mus_musculus_tropomyosin_4_(Tpm4),_mRNA	42,1	0,005
CM000216.2_Mus_musculus_chromosome_8,_whole_genome_shotgun_sequence	38,2	0,081
CH466525.1_Mus_musculus_232000009761530_genomic_scaffold,_whole_g...	38,2	0,081
NC_000074.7_Mus_musculus_strain_C57BL/6J_chromosome_8,_GRCm39	38,2	0,081
NC_005100.4_Rattus_norvegicus_strain_mixed_chromosome_1,_Rnor_6.0	32,2	5
CM000210.1_Mus_musculus_chromosome_2,_whole_genome_shotgun_sequence	32,2	5
CM000212.2_Mus_musculus_chromosome_4,_whole_genome_shotgun_sequence	32,2	5
NC_000068.8_Mus_musculus_strain_C57BL/6J_chromosome_2,_GRCm39	32,2	5
NC_000070.7_Mus_musculus_strain_C57BL/6J_chromosome_4,_GRCm39	32,2	5
CM000254.1_Homo_sapiens_chromosome_3,_whole_genome_shotgun_sequence	32,2	5
CH471055.1_Homo_sapiens_211000035832932_genomic_scaffold,_whole_g...	32,2	5
NC_000002.12_Homo_sapiens_chromosome_2,_GRCh38.p13_Primary_Assembly	32,2	5
NC_000003.12_Homo_sapiens_chromosome_3,_GRCh38.p13_Primary_Assembly	32,2	5
NOTE Explanations of E value can be found in[10][11].		

Table D.2 — Local BLAST search results of reverse primer of mouse Tropomyosin 4

Description	Score	E
Sequences producing significant alignments:	(Bits)	Value
CH474031.2_Rattus_norvegicus_342000077098238_genomic_scaffold,_wh...	40,1	0,021
AC_000084.1_Rattus_norvegicus_strain_BN;_Sprague-Dawley_chromosom...	40,1	0,021
NC_005115.4_Rattus_norvegicus_strain_mixed_chromosome_16,_Rnor_6.0	40,1	0,021
CM000215.2_Mus_musculus_chromosome_7,_whole_genome_shotgun_sequence	40,1	0,021
Y00169.1_Rat_TM-4_gene_for_fibroblast_tropomyosin_4	40,1	0,021
CM000216.2_Mus_musculus_chromosome_8,_whole_genome_shotgun_sequence	40,1	0,021
CH466525.1_Mus_musculus_232000009761530_genomic_scaffold,_whole_g...	40,1	0,021
NC_000074.7_Mus_musculus_strain_C57BL/6J_chromosome_8,_GRCm39	40,1	0,021
BC023701.1_Mus_musculus_tropomyosin_4,_mRNA_(cDNA_clone_MGC:38384...	40,1	0,021
BC070421.1_Mus_musculus_tropomyosin_4,_mRNA_(cDNA_clone_MGC:95675...	40,1	0,021
BC023827.2_Mus_musculus_tropomyosin_4,_mRNA_(cDNA_clone_MGC:38326...	40,1	0,021
NM_001001491.2_Mus_musculus_tropomyosin_4_(Tpm4),_mRNA	40,1	0,021
CH466531.1_Mus_musculus_232000009776713_genomic_scaffold,_whole_g...	38,2	0,081
NC_000073.7_Mus_musculus_strain_C57BL/6J_chromosome_7,_GRCm39	38,2	0,081
CM000210.1_Mus_musculus_chromosome_2,_whole_genome_shotgun_sequence	36,2	0,32
CH466551.1_Mus_musculus_232000009748833_genomic_scaffold,_whole_g...	36,2	0,32
NC_000068.8_Mus_musculus_strain_C57BL/6J_chromosome_2,_GRCm39	36,2	0,32
NC_000007.14_Homo_sapiens_chromosome_7,_GRCh38.p13_Primary_Assembly	36,2	0,32
AC_000081.1_Rattus_norvegicus_strain_BN;_Sprague-Dawley_chromosom...	36,2	0,32
NC_005112.4_Rattus_norvegicus_strain_mixed_chromosome_13,_Rnor_6.0	36,2	0,32
CM000209.2_Mus_musculus_chromosome_1,_whole_genome_shotgun_sequence	36,2	0,32
NC_000067.7_Mus_musculus_strain_C57BL/6J_chromosome_1,_GRCm39	36,2	0,32
CM000252.1_Homo_sapiens_chromosome_1,_whole_genome_shotgun_sequence	36,2	0,32
NC_000001.11_Homo_sapiens_chromosome_1,_GRCh38.p13_Primary_Assembly	36,2	0,32
NC_000072.7_Mus_musculus_strain_C57BL/6J_chromosome_6,_GRCm39	34,2	1,3
NC_000015.10_Homo_sapiens_chromosome_15,_GRCh38.p13_Primary_Assembly	34,2	1,3
AC_000073.1_Rattus_norvegicus_strain_BN;_Sprague-Dawley_chromosom...	34,2	1,3
NC_005104.4_Rattus_norvegicus_strain_mixed_chromosome_5,_Rnor_6.0	34,2	1,3
NC_000008.11_Homo_sapiens_chromosome_8,_GRCh38.p13_Primary_Assembly	34,2	1,3
CM000270.1_Homo_sapiens_chromosome_19,_whole_genome_shotgun_sequence	34,2	1,3
CH471106.1_Homo_sapiens_211000035837249_genomic_scaffold,_whole_g...	34,2	1,3
NC_000019.10_Homo_sapiens_chromosome_19,_GRCh38.p13_Primary_Assembly	34,2	1,3
AC_000070.1_Rattus_norvegicus_strain_BN;_Sprague-Dawley_chromosom...	32,2	5
CH474041.1_Rattus_norvegicus_342000077088087_genomic_scaffold,_wh...	32,2	5
AC_000076.1_Rattus_norvegicus_strain_BN;_Sprague-Dawley_chromosom...	32,2	5
AC_000069.1_Rattus_norvegicus_strain_BN;_Sprague-Dawley_chromosom...	32,2	5
NC_005107.4_Rattus_norvegicus_strain_mixed_chromosome_8,_Rnor_6.0	32,2	5
NC_005101.4_Rattus_norvegicus_strain_mixed_chromosome_2,_Rnor_6.0	32,2	5
NC_005100.4_Rattus_norvegicus_strain_mixed_chromosome_1,_Rnor_6.0	32,2	5
CM000212.2_Mus_musculus_chromosome_4,_whole_genome_shotgun_sequence	32,2	5
CM000217.2_Mus_musculus_chromosome_9,_whole_genome_shotgun_sequence	32,2	5
CH466522.1_Mus_musculus_232000009795437_genomic_scaffold,_whole_g...	32,2	5
NOTE Explanations of E value can be found in[10][11].		

Table D.2 *(continued)*

Description	Score	E
NC_000070.7_Mus_musculus_strain_C57BL/6J_chromosome_4,_GRCm39	32,2	5
NC_000069.7_Mus_musculus_strain_C57BL/6J_chromosome_3,_GRCm39	32,2	5
NC_000075.7_Mus_musculus_strain_C57BL/6J_chromosome_9,_GRCm39	32,2	5
NC_000016.10_Homo_sapiens_chromosome_16,_GRCh38.p13_Primary_Assembly	32,2	5
CM000260.1_Homo_sapiens_chromosome_9,_whole_genome_shotgun_sequence	32,2	5
CM000266.1_Homo_sapiens_chromosome_15,_whole_genome_shotgun_sequence	32,2	5
CH471082.1_Homo_sapiens_211000035835546_genomic_scaffold,_whole_g...	32,2	5
NC_000009.12_Homo_sapiens_chromosome_9,_GRCh38.p13_Primary_Assembly	32,2	5
NOTE Explanations of E value can be found in[10][11].		

D.3 Evaluation of PCR primer exclusivity using an exclusivity database

The exclusivity database was used to evaluate exclusivity of PCR primer to detect mouse Tropomyosin 4. PCR primer are the same as evaluated in D.2. All data were obtained on November 26, 2020.

BLAST search was set in order to observe the exclusivity including predicted sequences as resulted in Table D.3 and Table D.4. Total number of sequences resulted from forward and reverse primer sequences were 5 000 and 1 559, respectively.

Table D.3 — BLAST search results of forward primer of mouse Tropomyosin 4

Description	Query cover	E Value	Per. Ident	Acc. Len	Accession
PREDICTED: Ailuropoda melanoleuca tropomyosin beta chain-like...	100 %	0,14	100	516	XM_034650814.1
PREDICTED: Thalassophryne amazonica tropomyosin alpha-1...	100 %	0,14	100	1 061	XM_034183045.1
PREDICTED: Thalassophryne amazonica tropomyosin alpha-1...	100 %	0,14	100	1 176	XM_034183044.1
PREDICTED: Thalassophryne amazonica tropomyosin alpha-1...	100 %	0,14	100	1 272	XM_034183043.1
PREDICTED: Thalassophryne amazonica tropomyosin alpha-1...	100 %	0,14	100	1 386	XM_034183042.1
PREDICTED: Catharus ustulatus tropomyosin 4 (TPM4), transcript...	100 %	0,14	100	1 161	XM_033082299.1
PREDICTED: Catharus ustulatus tropomyosin 4 (TPM4), transcript...	100 %	0,14	100	1 101	XM_033082298.1
PREDICTED: Catharus ustulatus tropomyosin 4 (TPM4), transcript...	100 %	0,14	100	964	XM_033082297.1
PREDICTED: Xenopus tropicalis tropomyosin 2 (tpm2), transcript...	100 %	0,14	100	1 744	XM_031900880.1
PREDICTED: Xenopus tropicalis tropomyosin 2 (tpm2), transcript...	100 %	0,14	100	1 507	XM_031900876.1
PREDICTED: Xenopus tropicalis tropomyosin 2 (tpm2), transcript...	100 %	0,14	100	1 738	XM_031900872.1
PREDICTED: Lonchura striata domestica tropomyosin 4 (TPM4),...	100 %	0,14	100	1 082	XM_021540082.2
PREDICTED: Lonchura striata domestica tropomyosin 4 (TPM4),...	100 %	0,14	100	943	XM_021540083.2
NOTE Explanations of query cover, E value and percent identity can be found in[10][11].					