Health and Environmental Sciences Institute

740 15th Street, NW
Suite 600
Washington, DC 20005
USA

1.202.659.8404 office
1.202.659.3859 fax
www.hesiglobal.org

COMPARE
Comprehensive Protein Allergen Resource

COMPARE 2020 Update Documentation

The COMPARE Team is pleased to present the COMPARE 2020 database update, released on 01/29/2020 (www.comparedatabase.org).

Contents

1. COMPARE 2020: General Overview

The COMPARE Team is pleased to present the release of the COMPARE 2020 database.

COMPARE 2020 includes results from the regular annual database update process (as in previous years), as well as two additional initiatives, conducted as part of COMPARE's continuous improvement strategies:

- "**Annual update screenings"** – applying the COMPARE process to sequence records submitted and/or published within the past 1-year time window.
- **"Historical screenings"** – a plan to undertake a historic screening by applying the COMPARE process to sequence records dated from 2016 and previous years was completed. This project was undertaken to harmonize the database content to up-to-date COMPARE processes, given that the COMPARE process was developed in 2016 and applied only to a "one-year" time window every year since then.
- **"Database audit"** – this process was conducted by an external Bioinformatics services provider to analyze the current content of the COMPARE database as a quality control measure.

Each of these additional processes will be described in more details in sections 2 to 4 below.

Overall, the updates to COMPARE 2020, resulting from the output of these three work streams, include:

- the **addition of 191 unique sequences**, from the annual update and historical screenings;
- the **removal of 24 obsolete sequences** from the COMPARE 2019 dataset (1 substituted by a newly entered sequence and 23 per audit results; exact accessions are listed in section 2.1 and 3.1 respectively).
- and **2 accession number amendments**, per audit results (section 3.1).

COMPARE 2020 is comprised of 2,248 sequences:

2,081 entries (COMPARE 2019) - 23 removed (audit) - 1 (substituted by new entry) + 191 new unique entries = 2,248 entries

2. New Additions in COMPARE 2020

The 191 new unique entries are the result of: (1) The **COMPARE Annual Download** (2) the **Historic Dataset review**. Entries derived from each process can be recognized in the database by unique identifiers present under the field "Year adopted"

(1) Annual Screening: 2020 and 2020MS
(2) Historic Screening: 2020H and 2020H_MS

In response to the increasing number of allergen identification studies using the combination of 2D western-blots to identify IgE-binding proteins, with mass spectrometry (MS) as a tool for peptides sequence identification, COMPARE's independent Peer-Review Panel (PRP) updated their acceptance criteria to accept the peptides (10 amino acids in length or above) explicitly identified by MS and which sequences are included in the supporting literature. As a

result, a same protein may be represented in the database by several peptide fragments, listed as individua entries. Note also that those MS sequence fragments may not necessarily represent the IgE binding area of the protein, but are part of a protein that has evidence of IgE binding.

Only in the (rare) cases where: a) the full protein is then produced (recombinant) based on the deducted MASCOT sequence, b) the full sequence published as part of the study, and 3) the recombinant protein is tested again for IgE binding, would PRP accept the full protein sequence (instead of the isolated peptides), if IgE binding is confirmed.

Sequences corresponding to short peptides derived from MS data, either sourced from the "annual screening" or the "historic screening" are labelled "2020MS" or "2020H_MS" respectively, for clarity.

2.1. COMPARE Annual Screening

The COMPARE Peer-Review Panel (PRP) reviewed a total of 201 sequences from 72 publications that were identified during the COMPARE annual screening process. The candidate sequences were sourced from NCBI, UniProt, Allergen Online (version 19), IUIS, as well as a targeted literature search, for the time window <01 March 2018 - 15 May 2019>. In total, there were **83 unique sequences approved by PRP to include in the COMPARE2020 database, from this data set.**

Note that one of these sequences (AAB35977.1) replaces an older entry (with entry date 2010) with the same accession # but which had the wrong species listed – reason why the 2010 entry has been removed, being replaced with the 2020 entry listing the correct information based on the article reviewed in this cycle.

2.2. Historic Screenings

A targeted historic data retrieval was completed to apply the COMPARE screening process to sequences/literature from years prior to COMPARE's first year of operation, 2016, in order to identify possible sequences not captured in the foundational dataset (AllergenOnline v.16) on which COMPARE built on. For this task, a sequence data retrieval for 2016 and previous years with no other date limit was conducted, querying UniProt, AllergenOnline & IUIS. Entries were required to have literature citation(s) and at least one of the literature citations was required to contain allerg* and/or IgE in its title and/or abstract. All candidates were deduplicated against COMPARE 2019. Duplicate checks were performed at the accession level and at the sequence level within this data set. This resulted in 327 unique articles for PRP consideration. From this dataset, **PRP approved 108 sequences.**

As mentioned above, following the release of COMPARE 2019 the COMPARE PRP elected to expand the acceptance criteria to include peptides identified via MS for proteins with compelling evidence of IgE binding. Due to this change, PRP revisited candidates linked to MS studies that were rejected in past review cycles to apply these new set of criteria consistently. Entries accepted during this exercise are included as part of the Historical dataset and labelled "2020H-MS", similarly to other MS-related sequences sourced from the "historic screenings".

3. Amendments to COMPARE 2019 entries

3.1. COMPARE Database Audit.

The COMPARE database audit was conducted by an external Bioinformatics services provider and PRP, over two consecutive cycles:

- The audit was conducted in 2018 by the Bioinformatics services specialists, on the existing database version at the time, COMPARE 2018, with the aim to identify sequences with quality issues or present in duplicates in the database.
- Findings were presented to PRP in the following year review cycle, (i.e., 2019' PRP review cycle, leading to the COMPARE 2020 update) for consideration and decision on eventual amendments needed.

From this exercise, 2 sequences were amended and 23 were removed. Details of the updates and removals are included in table 1 below, with the corresponding justifications from PRP.

Table 1: Itemized list of amendments and sequence removals in COMPARE 2019 entries, applied in COMPARE 2020, as a result from the audit task.

Accession # (as listed in COMPARE 2019)	Status	Comment
100|Putative Act d 9.0101 Manual Entry [Actinidia deliciosa]	Updated	Specific accession # now found for this sequence. Updated the accession # to: FG438715.1
XP_004232333.1|Allergen PREDICTED: non-specific lipid-transfer protein 2 [Solanum lycopersicum (Lycopersicon esculentum)]	Updated	Specific accession # now found for this sequence. Update the accession # to: NP_001306883
AAB26195.1|Putative major allergen ABA-1=TBA-1 allergen homolog {N- terminal} [Ascaris suum]	Removed	Remove. Justification: it is 100% identical to Q06811.2; except for the 4 "X"s.
AAB20452.1|Allergen Bet v I=major allergen [Betula pendula]	Removed	Remove. Justification: it is 100% identical to1B6F_A. 2; except for the 2 "X"s.
3F5V_B|Allergen Chain B, C2 Crystal Form of Mite Allergen Der P 1 [Dermatophagoides pteronyssinus]	Removed	Remove. Justification: X is a C in the sequence 2AS8_B.that is already in the database. For the rest of the sequence is identical
P49274.1|Putative Alpha-amylase (Allergen Der p 4) (Der p IV) [Dermatophagoides pteronyssinus]	Removed	Remove, completely contained in AAD38942.1 without X ambiguous AA
1UHG_D|Allergen Chain D, Crystal Structure Of S-Ovalbumin At 1.9 Angstrom Resolution [Gallus gallus]	Removed	Remove. Justification: it is 100% identical to P01012.2; except for five "X"s.

3EBW_A|Putative Chain A, Crystal Structure Of Major Allergens, Per A 4 From Cockroaches [Periplaneta americana]	Removed	Remove, it is fully contained (and 3 X's resolved) in ACJ37391.1 that is in COMPARE
1N10_A|Allergen Chain A, Crystal Structure Of Phl P 1, A Major Timothy Grass Pollen Allergen [Phleum pratense]	Removed	Remove. Justification. The CAA55390.1 sequence already in the database is one aa longer, i.e. the N-terminal methionine (start codon) which was not included in molecule used for crystallization.
A60372|Putative pollen allergen Poa-pl - Kentucky bluegrass (fragment) [Poa pratensis]	Removed	Remove, completely contained in CAA10520.1 without X
2ATM_A|Allergen Chain A, Crystal Structure Of The Recombinant Allergen Ves V 2 [Vespula vulgaris]	Removed	Keep P49370.1 (M instead of X) "retire"
MEHB2|Putative melittin, minor - honeybee [Apis mellifera]	Removed	Remove: same as CAA26038.1 (in COMPARE) - 100% identity, same length, same species. Keep any articles not mentioned in the sequence staying in COMPARE
AAB36117.1|Putative Sol i 1=antigen {N- terminal} [Solenopsis invicta]	Removed	RETIRE. Not in NCBI or Uniprot. (entry had same publications as in AAT95008.1, currently in COMPARE)
AAB36119.1|Putative Sol i 1=antigen {N- terminal} [Solenopsis invicta]	Removed	RETIRE. Obsolete Accession # (record removed by NCBI staff). Not in NCBI or Uniprot. (entry had same publications as in AAT95008.1, currently in COMPARE)
AAB36120.1|Putative Sol i 1=antigen {N- terminal} [Solenopsis invicta]	Removed	RETIRE. Obsolete Accession # (record removed by NCBI staff). Not in NCBI or Uniprot. (entry had same publications as in AAT95008.1, currently in COMPARE)
AAB36121.1|Putative Sol i 1=antigen {N- terminal} [Solenopsis invicta]	Removed	RETIRE. Obsolete Accession # (record removed by NCBI staff). Not in NCBI or Uniprot. (entry had same publications as in AAT95008.1, currently in COMPARE)
3SMH_A|Allergen Chain A, Crystal Structure Of Major Peanut Allergen Ara H 1 [Arachis hypogaea]	Removed	Remove, identical to 3S7E_A, except for resolved X
4AUD_B|Allergen Chain B, Crystal Structure Of Alternaria Alternata Major Allergen Alt A 1 [Alternaria alternata]	Removed	Remove: A longer identical sequence is in COMPARE in which the x's are identified (AAB47552.1)
4B9R_A|Allergen Chain A, Crystal Structure Of The Major Birch Pollen Allergen Bet V 1.0101 (isoform A) Nitrated In Vitro With Tetranitromethan. [Betula pendula]	Removed	Remove: identical to 4BKD_A which has 2/3 X's filled in.

4BKC_A|Allergen Chain A, Crystal Structure Of A Unusually Linked Dimeric Variant Of Bet V 1 [Betula pendula]	Removed	Remove already in COMPARE without X (CAA33887.1)
4BKD_A|Allergen Chain A, Crystal Structure Of An Unusually Linked Dimeric Variant Of Bet V 1 (b) [Betula pendula]	Removed	Remove. Justification. Same sequence as CAA33887.1 except for one X
4BTZ_A|Allergen Chain A, Crystal Structure Of Peroxinitrite Treated Major Birch Pollen Allergen Bet V 1.0101 (isoform A) [Betula pendula]	Removed	Remove. Justification. Same sequence as CAA33887.1 except for four X'es
AAB20453.1|Allergen Car b l=major allergen [Carpinus betulus]	Removed	Remove. The full sequence is in COMPARE, AB281040.1 - only differs in the terminal AA after X (XK here, SS in the full one)
CAA51204.1|Allergen gamma 3 hordein, partial [Hordeum vulgare]	Removed	There is a more complete sequence with AA residues to replace the XXX positions: Hor v 20.101; P80198 (HOG3_HORVU)
P83885.1|Allergen Allergen Ani s 4 [Anisakis simplex]	Removed	Remove, fully covered by CAK50389.1 and two X'es resolved

4. Other Improvements & New Features

4.1. 2020 Sequence Header Format

The COMPARE 2020 header format has been modified to adopt the standard FASTA format used in NCBI and other sequence databases and enhance compatibility with a number of bioinformatic sequence alignment tools, including the COMPASS (COMPare Analysis of Sequences with Software tool; launched June 7th 2019 – see section 4.2). Specifically, the term “accession” is no longer used as part of the header, nor are the vertical “pipe” separators.

Thus, an entry that would have had the form in previous version of COMPARE:

```
>accession|CAA33887.1|Allergen Bet v 1 - like [Betula pendula]
```

now instead has the form:

```
>CAA33887.1 Allergen Bet v 1 - like [Betula pendula]
```

This new header format has been applied to all the 2,048 entries in COMPARE 2020.

4.2. COMPARE’s bioinformatics companion tool, “COMPASS” (launched June 7th, 2019)

As of June 7th, 2019, the COMPARE database is equipped with its companion tool, **COMPASS (COMPare Analysis of Sequences with Software)**, as a built-in feature. COMPASS is a comparative sequence search tool, incorporating the open source FASTA software

package (FASTA v36). With this tool, COMPARE users can conduct website-based, real-time use of the COMPARE database to identify similarities between a protein sequence of interest and COMPARE's allergens via amino acid sequence alignments (between two or more amino acid sequences). To access the tool, go www.comparedatabase.org, click on the "Database" tab and click on the green button "Run COMPASS". For detailed information, instructions on how to use and supporting references, visit COMPASS' "About" page

COMPARE 2020 has otherwise retained specifications from COMPARE 2019, described in the COMPARE 2019 documentation file available on the online database page itself - direct link: here - *e.g.,* use of "COMPARE #" Accessions when no other public accession number is known for a specific sequence; consistent approach for determining "description" fields; information sharing via documentation and transparency files; as well as many user-friendly features implemented with COMPARE 2019.

5. Your Feedback is Appreciated - Contact Us

The HESI COMPARE database program is committed to transparency and open dialog. Individuals or organizations are invited to submit feedback, questions or inquiries via the "Contact us" portal in the COMPARE database website, or email to comparedatabase@hesiglobal.org. HESI staff will respond if the information is readily available or will relay the inquiries to PRP if a more in-depth response is required.

6. Support COMPARE!

Is COMPARE useful a as a resource and do you like its commitment to continuous improvement? If so, support COMPARE! We have other ideas to continue improving this resource and making it as comprehensive and thorough as possible.

The COMPARE database is a collaborative HESI program that combines programmatic support from the Joint Institute for Food Safety and Nutrition (JIFSAN), www.jifsan.umd.edu. The annual update of the database is a resource intensive process that involves many more partners and collaborators, rolling on a steady annual cycle schedule. The execution of the program relies on the contribution of scientific expertise as well as in-kind and direct financial support from both public sector and private sector scientific organizations to maintain this free, public resource. If you would like to learn more about how you or your organization can contribute, please contact us at the address listed above.

Your support is tax deductible in the US: HESI is non-profit, 501(c)(3) organization committed to generating science for a safer, more sustainable world. Financial support to HESI scientific programs is considered a tax-deductible charitable donation in the United States. Your support for our mission, through funding and participation, makes our scientific collaborations and outreach possible and helps to improve both human and environmental health across the globe.

We look forward to hearing from you!

About HESI (www.hesiglobal.org): The Health and Environmental Sciences Institute (HESI) is a non-profit institution whose mission is to collaboratively identify and help resolve global health and environmental challenges through the engagement of scientists from academia, government, industry, NGOs, and other strategic partners. Since 1989, HESI has provided the framework for scientists from public and private sectors to meaningfully collaborate in developing science for a safer, more sustainable world.

www.comparedatabase.org