DataHarmonizer Manual

I. **Purpose:** To structure SARS-CoV-2 contextual data according to the PHA4GE SARS-CoV-2 data standard for future-proofing contextual data and for improved harmonization and data exchange between laboratories.

   a. Data providers will extract and curate lab-specific contextual data according to the steps outlined in the procedure below.
   b. Laboratories will populate the harmonized template with information from their datasets using the *DataHarmonizer* application.
   c. Data providers will share the harmonized data according to organizational policies in consultation with the data steward.

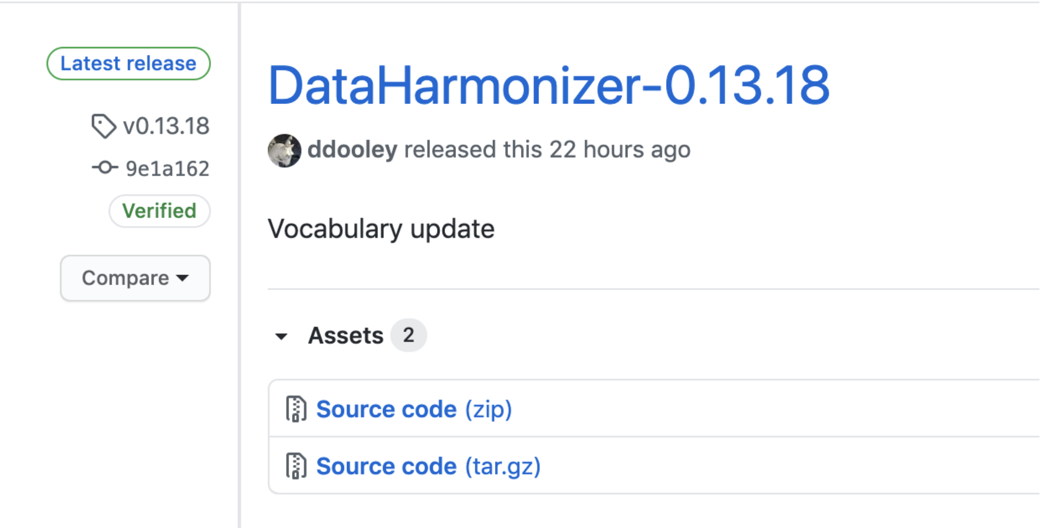
II. **Data:** The contextual data describing sample collection and processing, host information, host vaccination information, host exposure information, host reinfection information, lineage and variant information, sequencing, and bioinformatics and QC metrics, pathogen diagnostic testing information, and author attribution information as supplied by the data provider.

III. **Additional documentation:** Use the PHA4GE Contextual Data Curation SOP for further guidance on populating the contextual data collection template in the DataHarmonizer application, which can be found at https://www.protocols.io/view/pha4ge-contextual-metadata-sop-btpznmp6.

IV. **Procedure:**

	Action
1	Download the zip file (“Source code (zip))” containing The DataHarmonizer application from the following link: https://github.com/Public-Health-Bioinformatics/covid19ValidationGrid/releases

	Action



Latest release

v0.13.18

9e1a162

Verified

Compare

DataHarmonizer-0.13.18

ddooley released this 22 hours ago

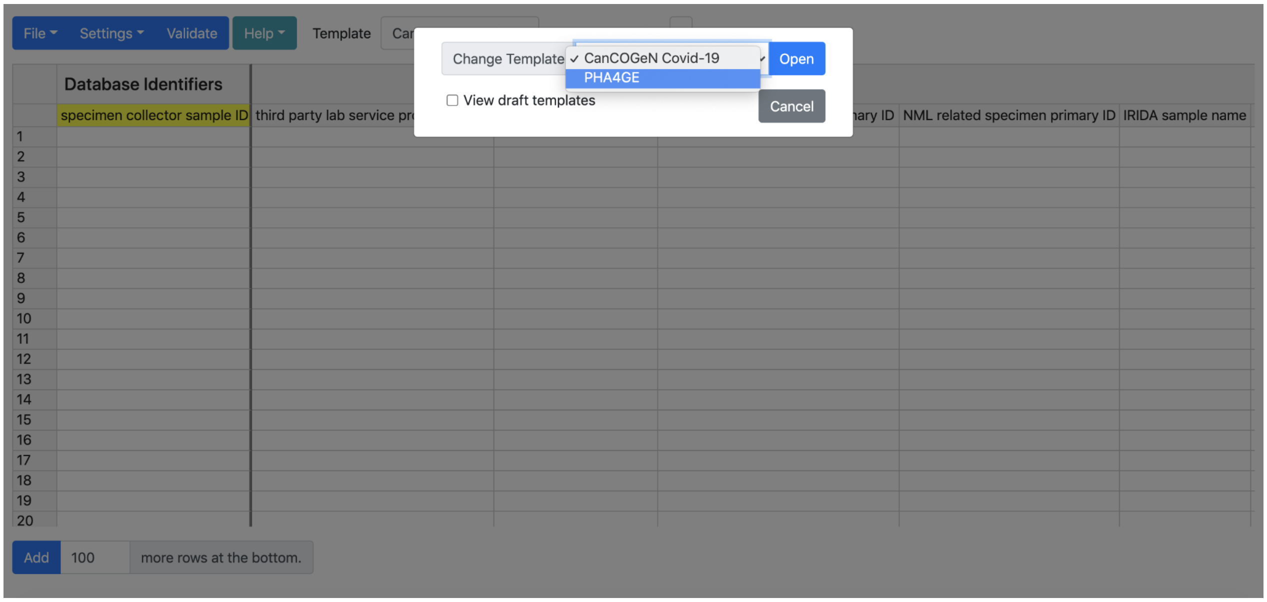
Vocabulary update

Assets 2

Source code (zip)

Source code (tar.gz)

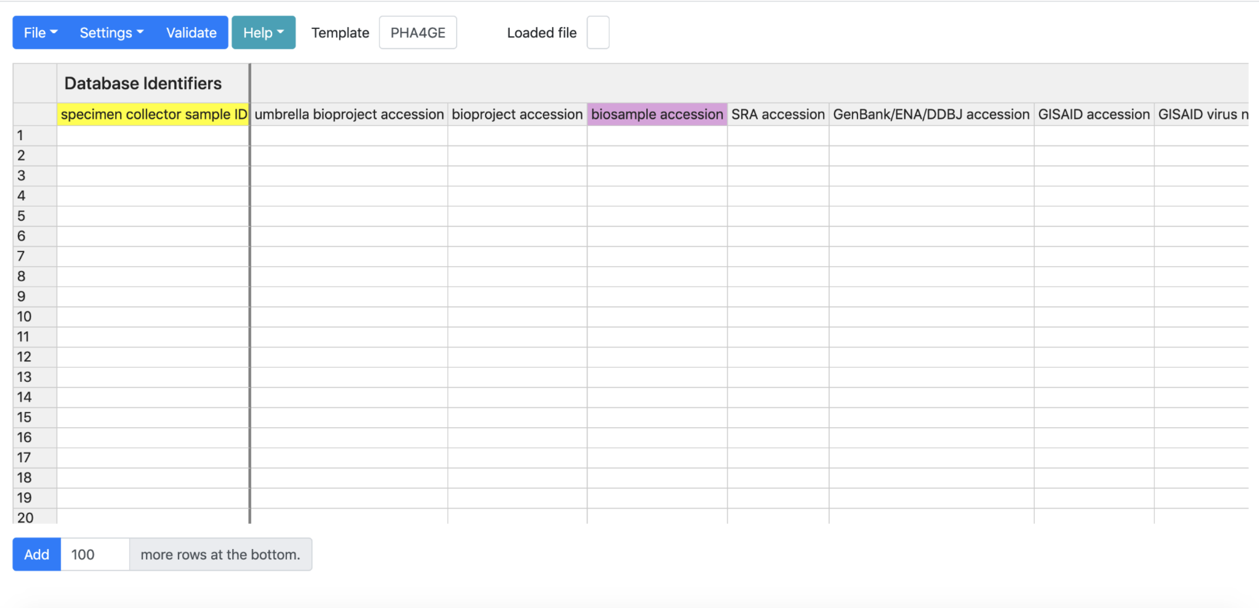
Extract the zip file’s contents, and navigate into the extracted folder. Open **main.html**. The validator application will open in your default browser. Click **File** on the top left toolbar, and then click **Change Template**. By the blue Open button, click the downwards arrow to see the template menu. It should look like this:



Select **PHA4GE** from the drop down pick list, and then click **Open**.

The PHA4GE template should look like this:

	Action

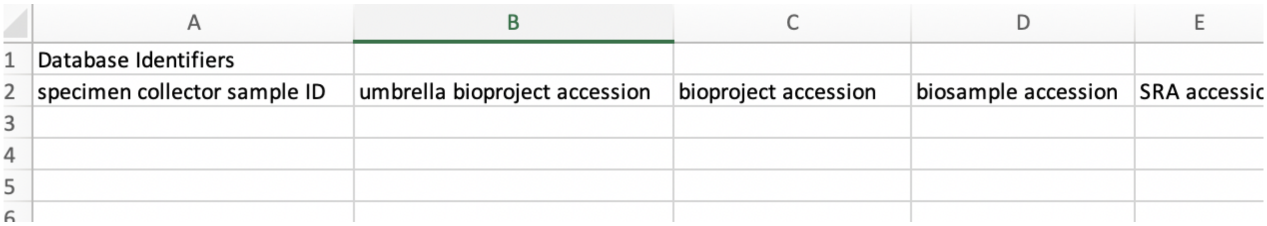


	Database Identifiers							
	specimen collector sample ID	umbrella bioproject accession	bioproject accession	biosample accession	SRA accession	GenBank/ENA/DDBJ accession	GISAID accession	GISAID virus n
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								

Data can be entered into the validator application manually, by typing values into the application’s spreadsheet, or data can be imported from local `xlsx`, `xls, tsv` and `csv` files.

To import local data, click **File** on the top-left toolbar, and then click **Open**. To enter data in a new file, click **File** on the top-left toolbar, and then click **New**. Data entered into the spreadsheet can be copied and pasted.

Note: Only files containing the headers expected by the DataHarmonizer can be opened in the application. Example:



	A	B	C	D	E
1	Database Identifiers				
2	specimen collector sample ID	umbrella bioproject accession	bioproject accession	biosample accession	SRA accessic
3					
4					
5					
6					

If you are missing the first row, you will get a warning. It will ask you which row has your column headers. Resolve by declaring “1” as the row in which your column headers reside.

2 Before you begin to curate sample metadata:

- Review your dataset
- Review the fields in the template of the DataHarmonizer application
- Review the field descriptions in the SOP Appendix

	Action
3	Familiarize yourself with DataHarmonizer functionality by reviewing the “**Getting Started**”. To access "Getting Started", click on the green **Help** button on the top-left toolbar, then click **Getting Started**. Definitions, examples and further guidance are available by double clicking on the field headers, or by using the “**Reference Guide**”. To access the “Reference Guide” click on the **Help** button, then click **Reference Guide**.
4	Confirm mapping of your data fields to those in the harmonized template with the data steward. *Note: If you intend to share the entered data, confirm the level of granularity of information that can be shared with the data steward and/or your privacy officer. The most detailed information allowable should be included here.*
5	Enter data into the collection template spreadsheet. ● Hide non-required fields (colour-coded purple 🟪 and white/grey) by clicking **Settings** on the top-left toolbar, followed by clicking on **Show Required Columns** (colour-coded in yellow 🟨). ● Double click in the field headers to see definitions and detailed guidance as needed (or consult Appendix A). ● Jump to a specific field header by clicking **Settings** on the top-left toolbar, followed by clicking on **Jump to**, then select the field header of the column you would like to view from the drop down list. ● Populate the template with the information from your dataset. ● Use picklists when provided. ● A value should be entered for every <u>*required field*</u> in each row. If data is missing or not collected, **choose a null value from the picklist**. ○ Not Applicable ○ Missing ○ Not Collected ○ Not Provided ○ Restricted Access ● Free text can be provided when picklists are not available. ● To autofill columns with the same values for every sample, click **Settings**, then **Fill Column**. Enter the name of the field being filled under “Fill cells in column”, then enter the value in “with value” and click **Enter**. **If a desired term is not present in a picklist, contact Emma Griffiths at** ega12@sfu.ca. *Note: Sometimes there will be constraints on what information can be shared, other times a field may not be applicable to your sample. You do not have to fill in every field, only the ones that pertain to your sample. If you are missing information for a required field, use the null values (controlled vocabulary indicating the reason why information is not provided) in the picklist to report missing data. If you are missing information for a recommended or optional field, or those fields simply do not pertain to your sample, leave them blank.*

<table>
<tr><th></th><th>Action</th></tr>
<tr><td></td><td>
<p>Required fields are organized into subsections (see <b>Appendix A</b> for required field definitions and guidance, and <b>Appendix B</b> for examples of how to structure sample descriptions):</p>
<table>
<tr><th>Subsection</th><th>Required Fields</th></tr>
<tr><td><b>Database Identifiers</b></td><td>specimen collector sample ID</td></tr>
<tr><td><b>Sample Collection and Processing</b>

<i>Note: Evaluate whether the specimen collector sample ID is considered identifiable by your institutional policies. If not, copy the sample ID into the sample ID field in the validator spreadsheet. If yes, provide the alternative sample ID as specified by the lab. Be sure to keep a copy of the key.</i></td><td>sample collected by
sequence submitted by
sample collection date
geo_loc (country)
geo_loc (state/province/territory)
organism
isolate</td></tr>
<tr><td><b>Host Information</b></td><td>host (scientific name)
host disease</td></tr>
<tr><td><b>Sequencing</b></td><td>sequencing instrument
purpose of sequencing
purpose of sequencing details</td></tr>
<tr><td><b>Bioinformatics and QC Metrics</b></td><td>consensus sequence method name
consensus sequence method version</td></tr>
</table>
</td></tr>
<tr><td>6</td><td>
<p>Validate the entered data by clicking on the <b>Validate</b> button on the top-left toolbar.</p>
<p>Missing information and invalid entries in required fields will be highlighted in red.</p>
<ul>
<li>Observe invalid rows by clicking <b>Settings</b> in the top-left toolbar, and then clicking on <b>Show invalid rows.</b></li>
<li>Observe valid rows by clicking <b>Settings</b> in the top-left toolbar, and then clicking on <b>Show valid rows.</b></li>
<li>Return view to all rows by clicking <b>Settings</b> in the top-left toolbar, and then clicking on <b>Show all rows.</b></li>
<li>Use the <b>Next Error</b> button on the far right to systematically move from one error to the next.</li>
</ul>
</td></tr>
</table>

	Action
	- Duplicated sample IDs will be highlighted in red in the “specimen collector sample ID” field. *Note: Row viewing options only appear after a validation attempt has been made.*
7	Address any invalid data that was flagged in red in the template. - 🟥 Pale Red = Incorrect data format - 🟥 Dark Red = Required data missing *Note: It is possible to export incomplete or invalid data. Make sure to review any errors prior to exporting.*
8	You can save your work at any time. Save data by clicking **File** on the top-left toolbar, and then clicking on **Save as**. Enter the file name and press **Save**. Files saved in the DataHarmonizer are (.xlsx) format and can be re-opened directly.
10	Optional: The DataHarmonizer enables automated transformations of entered data into GISAID, NCBI BioSample, NCBI SRA, NCBI GenBank, and NCBI GenBank-source-modifiers submission formats. Export validated data by clicking **File** on the top-left toolbar, and then clicking **Export to**.Type in the file name, and select the desired format from the Format picklist. Then click **Export**. - Have the validated data reviewed by the data steward
12	Additional Information: A local copy of the **Standard Operating Procedure (SOP)** is included in every download of the DataHarmonizer. To access it, click on the green **Help** button on the top-left toolbar, then click **SOP**. The latest version of the SOP is published online and accessible via a web browser at all times.

V. Appendix A: Required Field Definitions and Guidance

Field definitions for required fields, as well as guidance and examples, are provided below. This information has been sourced from the DataHarmonizer reference guide. Guidance for strongly recommended and optional fields can be found in the reference guide.

Database Identifiers

specimen collector sample ID
The user-defined name for the sample.
Store the collector sample ID. If this number is considered identifiable information, provide an alternative ID. Make sure to store the key between this alternative ID and the original ID for traceability. Every collector sample ID from a single submitter must be unique. It can have any format, but we suggest that you make it concise, unique and consistent within your lab.
e.g. prov_rona_99

Sample Collection and Processing

sample collected by
The name of the agency that collected the original sample.
The name of the sample collector should be written out in full, (with minor exceptions) and be consistent across multiple submissions e.g. Public Health Agency of Canada, Public Health Ontario, BC Centre for Disease Control. The sample collector specified is at the discretion of the data provider (i.e. may be hospital, provincial public health lab, or other).
e.g. BC Centre for Disease Control

sequence submitted by
The name of the agency that generated the sequence.
The name of the agency should be written out in full, (with minor exceptions) and be consistent across multiple submissions. If submitting specimens rather than sequencing data, please put the "National Microbiology Laboratory (NML)".
e.g. Public Health Ontario (PHO)

sample collection date
The date on which the sample was collected.
Sample collection date is critical for surveillance and many types of analyses. Required granularity includes year, month and day. Record the collection date accurately in the template. Before sharing this data, ensure you have consulted the data steward and/or your privacy officer regarding whether they consider this date to be identifiable information. If this date is considered identifiable, it is acceptable to add "jitter" to the collection date you share by adding or subtracting a calendar day (acceptable by GISAID). Do not change the collection date in your original records. Alternatively, ”received date” may be used as a substitute in the data you share. The date should be provided in ISO 8601 standard format "YYYY-MM-DD".
e.g. 2020-03-16

geo_loc_name (country)
The country where the sample was collected.
Provide the country name from the controlled vocabulary provided.
e.g. Canada

geo_loc_name (state/province/territory)
The state/province/territory where the sample was collected.
Provide the state/province/territory name from the GAZ geography ontology. Search for geography terms here: https://www.ebi.ac.uk/ols/ontologies/gaz
e.g. Western Cape

organism
Taxonomic name of the organism.
Select "Severe acute respiratory syndrome coronavirus 2" if sequencing SARS-CoV-2. If another Coronaviridae is being sequenced, provide the taxonomic name from NCBITaxon. Search for taxonomy terms at https://www.ebi.ac.uk/ols/ontologies/ncbitaxon.
e.g. Severe acute respiratory coronavirus 2

isolate
Identifier of the specific isolate.
This identifier should be an unique, indexed, alpha-numeric ID within your laboratory. If submitted to the INSDC, the "isolate" name is propagated throughout different databases. As such, structure the "isolate" name to be ICTV/INSDC compliant in the following format: "SARS-CoV-2/host/country/sampleID/date".
e.g. SARS-CoV-2/human/USA/CA-CDPH-001/2020

Host Information

host (scientific name)
The taxonomic, or scientific name of the host.
Common name or scientific name are required if there was a host. Both can be provided, if known. Use terms from the pick lists in the template. Scientific name e.g. Homo sapiens, If the sample was environmental, put "Not Applicable”.
e.g. Homo sapiens

host disease
The name of the disease experienced by the host.
This field is only required if there was a host. If the host was a human select COVID-19 from the pick list. If the host was asymptomatic, this can be recorded under “host health state details”. If the host is not human, and the disease state is not known or the host appears healthy, put “Not Applicable”.
e.g. COVID-19

Sequencing

sequencing instrument
The model of the sequencing instrument used.

Select a sequencing instrument from the picklist provided in the template.
e.g. MinIon

purpose of sequencing
The reason that the isolate was sequenced.
The reason an isolate was sequenced may provide information about potential biases in sequencing strategy. Provide the purpose of sequencing from the picklist in the template. Most likely, the sample was collected for Surveillance or Research. The reason why a sample was originally collected may differ from the reason why it was selected for sequencing. The reason for sample collection should be indicated in the "purpose of sampling" field.

purpose of sequencing details
The description of why the sample was sequenced providing specific details.
Provide an expanded description of why the sample was sequenced using free text. The description may include the importance of the sequences for a particular public health investigation/surveillance activity/research question. If details are not available, provide a null value.
e.g. The sample was sequenced to investigate the differences in lineages circulating in Canada during the spring and fall waves of the pandemic.

<u>Bioinformatics and QC Metrics</u>
consensus sequence method name
The name of the protocol used to produce the consensus sequence. Provide the software name followed by the version.
e.g. iVar

consensus sequence method version
The name and version number of the protocol used to produce the consensus sequence.
Provide the software name followed by the version.
e.g. 1.2

VI. Appendix B: Structuring Sample Descriptions (Examples)

To improve how sample information is structured, seven additional fields have been introduced to capture different kinds of anatomical and environmental samples, as well as collection methods (i.e. anatomical material, anatomical part, body product, environmental site, environmental material, collection device, and collection method). Not all of these fields need to be populated - only the fields that pertain to your sample. Provide the most granular information available and use the pick lists provided. If a term is missing from the list, contact Emma Griffiths at ega12@sfu.ca.

Definitions and guidance have been provided below to highlight differences in the meanings of sample type fields.

anatomical material
A substance obtained from an anatomical part of an organism e.g. tissue, blood.
Provide a descriptor if an anatomical material was sampled. Use the picklist provided in the template. If a desired term is missing from the picklis. If not applicable, leave blank.
e.g. Blood

anatomical part
An anatomical part/location of an organism e.g. oropharynx.
Provide a descriptor if an anatomical part was sampled. Use the picklist provided in the template. If not applicable, leave blank.
e.g. Nasopharynx (NP)

body product
A substance excreted/secreted from an organism e.g. feces, urine, sweat.
Provide a descriptor if a body product was sampled. Use the picklist provided in the template. If not applicable, leave blank.
e.g. Feces

environmental material
A substance or object obtained from the natural or man-made environment e.g. soil, water, sewage.
Provide a descriptor if an environmental material was sampled. Use the picklist provided in the template. If not applicable, leave blank.
e.g. Face Mask

environmental site
An environmental location may describe a site in the natural or built environment e.g. contact surface, metal can, hospital, wet market, bat cave.
Provide a descriptor if an environmental site was sampled. Use the picklist provided in the template. If not applicable, leave blank.
e.g. Building floor

collection device
The instrument or container used to collect the sample e.g. swab.

Provide a descriptor if a device was used for sampling. Use the picklist provided in the template. If not applicable, leave blank.
e.g. Swab

collection method
The process used to collect the sample e.g. phlebotomy, necropsy.
Provide a descriptor if a collection method was used for sampling. Use the picklist provided in the template. If not applicable, leave blank.
e.g. Bronchoalveolar Lavage (BAL)

Several examples are provided below which illustrate how to structure common sample descriptions.

e.g. nasal swab should be recorded:

host (scientific name)	host (common name)	host disease	anatomical part	collection device
Homo sapiens	Human	COVID-19	Nasopharynx (NP)	Swab

e.g. throat swab should be recorded:

host (scientific name)	host (common name)	host disease	anatomical part	collection device
Homo sapiens	Human	COVID-19	Oropharynx (OP)	Swab

e.g. saliva should be recorded:

host (scientific name)	host (common name)	host disease	anatomical material
Homo sapiens	Human	COVID-19	Saliva

e.g. salt water gargle should be recorded:

host (scientific name)	host (common name)	host disease	collection method
Homo sapiens	Human	COVID-19	Saline gargle (mouth rinse and gargle)

e.g. human feces should be recorded:

host (scientific name)	host (common name)	host disease	body product

Homo sapiens	Human	COVID-19	Feces

e.g. swab of a hospital bed rail should be recorded:

environmental site	environmental material	collection device
Hospital	Bed Rail	Swab

e.g. tissue from a bat (Platyrrhinus lineatus) in a cave should be recorded:

Host (common name)	Host (scientific name)	host disease	anatomical_part	environmental_site
Bat	Platyrrhinus lineatus	Not applicable	Tissue	Cave

e.g. particulates from air filter should be recorded:

environmental material	collection method
Particulate Matter	Air Filtration

VII. **Appendix C: Null Value Definitions**

Not Applicable
Information is inappropriate to report, can indicate that the standard itself fails to model or represent the information appropriately.

Missing
Information was known to be recorded in the past, but the observed value cannot be located or retrieved for some reason.

Not Collected
Information of an expected format was not given because it has not been collected.

Not Provided
Information of an expected format was not given, a value may be given at the later stage.

Restricted Access
Information exists but can not be released openly because of privacy concerns.

Source:
International Nucleotide Sequence Database Collaboration (INSDC) Missing Value Reporting Terms (2017-2018).
ENA Training Modules: https://ena-docs.readthedocs.io/en/latest/submit/samples/missing-values.html

Revision History

Version	Date	Writer	Description of Change
1.0	March 18, 2021	Emma Griffiths	Protocol adapted from CanCOGeN
2.0	June 16, 2021	Emma Griffiths	PHA4GE template moved from draft, new feature instructions added