**Table of Contents:**

[1 Goal of the documentation 2](#_Toc108446450)

[2 Scope and Responsibilities 2](#_Toc108446451)

[3 Term/Definition/Abbreviation 2](#_Toc108446452)

[4 Procedure/ Workflow 2](#_Toc108446453)

[4.1 Software 2](#_Toc108446454)

[4.2 Security details 3](#_Toc108446455)

[4.3 Method procedure 4](#_Toc108446456)

[4.3.1 General Notes 4](#_Toc108446457)

[4.3.2 GitHub 4](#_Toc108446458)

[4.3.3 WormExp 4](#_Toc108446459)

[4.3.4 Finding new GEO datasets 6](#_Toc108446460)

[4.3.5 Transcriptomics File 7](#_Toc108446461)

[4.3.6 Categorizing datasets 8](#_Toc108446462)

[4.3.7 Assembling data sets 8](#_Toc108446463)

[4.3.8 WormBase ID 10](#_Toc108446464)

[4.3.9 Update database 11](#_Toc108446465)

[4.3.10 Test Runs 14](#_Toc108446466)

[4.3.11 Updating Server 15](#_Toc108446467)

[4.3.12 Checklist 15](#_Toc108446468)

[5 Document History 16](#_Toc108446469)

[6 Publication bibliography 16](#_Toc108446470)

**Further documents**:

(evogen 2015) Manual for WormExp

(Yang et al. 2016) WormExp :WormExp: a web-based application for a Caenorhabditis elegans-specific gene expression enrichment analysis

Website <https://wormexp.zoologie.uni-kiel.de/wormexp/>

(Dozmorov 2016) GEOparse documentation

datafinder jupyter notebook file

1. **Goal of the documentation**

This documentation describes the procedure to update the web-based application WormExp.

1. **Scope and Responsibilities**

This documentation is only valid for AG Schulenburg.

|  |  |
| --- | --- |
| **Function** | **Responsibilities** |
| Employee/User | Updating WormExp  Is responsible for the correct use and update of the application. |

1. **Term/Definition/Abbreviation**

|  |  |
| --- | --- |
| GEO | Gene Expression Omnibus |

1. **Procedure/** **Workflow**

## Software

|  |  |  |  |
| --- | --- | --- | --- |
| Software/Website | Specification | Version | Source/Link |
| WormExp | web-based application for a taxon-specific gene set exploration and enrichment analysis | WormExp v1.0 | https://wormexp.zoologie.uni-kiel.de/wormexp/ |
| Python | Programming language | 3.8.11 | https://www.python.org/ |
| Anaconda | Python package distribution and management | 2020.11 | https://www.anaconda.com/ |
| GEOparse | Python library to access Gene Expression Omnibus Database (GEO) | 2.0.3 | https://geoparse.readthedocs.io/en/latest/GEOparse.html |
| Jupyter Notebook | Web-based environment for working with notebooks | 6.4.0 | https://jupyter-notebook.readthedocs.io/en/stable/index.html |
| Matplotlib | Python data visualization tool | 3.4.2 | https://matplotlib.org/ |
| Numpy | Core package for scientific computing with Python. | 1.20.3 | https://numpy.org/ |
| Pandas | Library for tabular data structures | 1.3.2 | https://pandas.pydata.org/docs/index.html# |
| Biopython | Library for biological computation written in Python | 1.78 | https://biopython.org/ |
| Java DK | The JDK is a development environment for building applications using the Java programming language. | 17.0 | https://www.oracle.com/java/ |
| Apache TomCat | Apache Tomcat software powers numerous large-scale, mission-critical web applications. | 8.5.78 | https://tomcat.apache.org/ |
| R | R is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files. | 4.1.2 | https://cran.r-project.org/bin/windows/base/ |
| R Studio | IDE for R | 2021.09.2-382 | https://www.rstudio.com/products/rstudio/download/ |
| FortiClient | VPN to access servers from the university of Kiel. | 7.05 | https://www.rz.uni-kiel.de/de/tipps/vpn/Windows/index.html |
| FileZilla | FileZilla is an FTP Client. This is a program designed to transfer files between between a client and a server over the internet or any other TCP/IP network. | 3.59.0 | https://filezilla-project.org/ |

## Security details

N/A

## Method procedure

### General Notes

This documentation describes the general procedure on how to update the database WormExp. The scripts and procedures used here were based on a Python (version 3.8.11) script in a virtual environment managed by anaconda (version 2020.11). The virtual environment contained several additional libraries: GEOparse (version 2.0.3), Jupyter (version 6.4.0), Matplotlib (version 3.4.2), Numpy (version 1.20.3), Pandas (version 1.3.2) and Biopython (version 1.78).

These libraries are essential for a correct function of the provided script to find new GEO datasets uploaded into Pubmed GEO. However, GEOparse is also available for R and the whole procedure can therefore be transferred to R if wished. Here, R scripts have only been created to revise files and visualize results.

The data should be collected in a separate folder and new Excel file, and only merged with the existing database at the very end. Additionally, a pilot run should be included on a copy of the current database with new gene sets to make sure that everything works appropriately. The database will be tested locally via Apache tomcat, but other software to locally host servers are applicable.

To access the server at the end that hosts WormExp, access to the network of the University of Kiel is necessary. Either by stu-number or an employee number.

### GitHub

The current server files, as well as all documentation and files can be found in the following GitHub repository: *https://github.com/evoecogen/WormExp*.

The repository has the following structure:

|  |  |
| --- | --- |
| Folder | Description |
| 00\_Archive | Contains files no longer needed, but will also not be deleted |
| 01\_Background | contains background information and literature about the project, like e.g., contacts |
| 02\_ServerFiles | A copy of the current files on the active server. Current version is v2.0. |
| 03\_Documentation | All documentation and information about his project and its structure. |
| 04\_Scripts | contains all scripts that have been used (and can be used) |
| 05\_QualityManagement | Contains folders with test sets and test WormExp databases to test for errors |
| 06\_Datafinder | Contains folders and files with found GEO accession numbers |
| 07\_Wormbase | contains a collection of WormBase ID changes |

To update the Database almost all folders, have to be used during the procedure. This repo structure is not set in stone, although this way it should provide the best overview over the project and its contents.

### WormExp

To understand the structure of WormExp, go into “02\_Serverfiles/WormExp\_v2.0”. This folder is a mirror of the folders uploaded onto the server and the active WormExp database. Two folders here are of importance: “tomcat” and “WormExpData”. The “tomcat” folder contains the application “Apache Tomcat” which is essential to start WormExp locally on your computer and later on the server. If you have installed Apache Tomcat on your computer, the installation folder should look very similar to the folder “tomcat”. The folder will be explained later on in more detail, as some minor things have to be updated there, too.

The main update, however, is in the folder “WormExpData”. It holds the complete database information. contains several .txt files, as well as some Excel and Java files:

|  |  |
| --- | --- |
| File | Description |
| c\_elegans.WS283.geneIDs | contains Wormbase ID, gene name and gene ID for every gene |
| reference | contains dataset name and link to publication |
| Chemicalexposure-otherStress | contains datasets categorized to Chemicals/Stress |
| DAF Insulin food | contains datasets categorized to DAF/Insulin/food |
| Development-Dauer-Aging | contains datasets categorized to Development/Dauer/Aging |
| Kim Mounts | contains datasets categorized to Kim Mountains |
| Mutants | contains datasets categorized to Mutants |
| Other | contains datasets categorized to Other |
| Pathogen | contains datasets categorized to Microbes |
| Targets | contains datasets categorized to TF Targets |
| Tissue-specific | contains datasets categorized to Tissue |
| Epigenetics | contains datasets categorized to Tissue |
| WormExp\_info (several versions) | has key information about the datasets, like number of genes, links to publications and methodology |
| dat.properties | Contains text/information for the website |
| Test.jar | Essential for starting the website |
| Dataset.zip | Contains all .txt files mentioned before, plus the current WormExp version (used for download purposes of WormExp) |
| Py.py and WE\_startup | Can be ignored for the update, but need to be copied later on as well onto the server! |

Text

Description automatically generated with low confidenceMost important for the update are the mentioned text files. They contain all current gene sets and all curated gene sets need to be added to these files, to make them available for the database. When opened, they only contain two columns. One with a WormBaseID that codes for a specific gene, and a gene set name that indicates if this gene is up or downregulated by something.

Figure 1: Exempt from Epigenetics category

The following sections describe, how such gene sets can be found and collected. Furthermore, as these files are very big, it is of utmost importance to first work on a copy of these files and later on combine old and new gene sets! This and more procedures are described in the following sections. For all future steps, please copy the folder “WormExpData” from the newest WormExp version into “05\_QualityManagement/WormSource\_vX.X”, where vX.X depends on the newest version for this update. If only new data sets are added, no new version is necessary. A new WormExp version should only be done when changes on the code have been added.

### Finding new GEO datasets

To find new GEO datasets, the jupyter notebook “datafinder” can be used. It can be found in “04\_Scripts/datafinder.ipynb”. It contains scripts and instructions on how to use it. The script uses the API GEOparse and searches for datasets depending on the inserted query. During the start of this project an analysis was conducted to ensure the used query contains the correct datasets. The results are documented in the supporting PowerPoint document in “03\_Documentation”.

Adaptions to the query should only be made in respect to the publication date. As of 02/2022 the database contains datasets until approx. 2018.

The script will create a separate Excel file called “GSE\_query\_results” in which further (manual) work will be conducted. This Excel file is the backbone of all further investigation and will be described in detail in the following section “Transcriptomics File”. It must be noted that in order to use the datafinder notebook, the software from section 4.1 has to be installed.

### Transcriptomics File

The datafinder script produces an Excel file that contains detailed information about every dataset found with GEOparse. This file can be found in “06\_Datafinder/GEO\_database\_results.xlsx” (renamed from “GSE\_query\_results” to ensure that it will not be overwritten on accident). It shows not only the exact title of the dataset, but also its geo\_accession number, publication date, contributors, etc. etc. See Figure 2 for an example.

Calendar

Description automatically generated with low confidence

Figure 2: Exempt from original Transcriptomics file

The main task here is to start sorting through the found datasets and finding out which datasets are useful for the updates. It has been decided to concentrate mostly on datasets that already possess an in-depth transcriptomics analysis done by the respective scientists. Datasets that only possess raw data can be ignored until otherwise stated. For a better overview, the transcriptomics file was transformed, and colors have been introduced (see Figure 3) to show which data has supplementary data available (blue), which datasets contain only raw data (yellow), and which datasets can be excluded (red). Exclusion of datasets was mainly due to no available paper or dataset was not focused on differential gene expression.

Table

Description automatically generated

Figure 3: Excerpt from the Transcriptomics file, after initial preparation

Additionally, new columns have been introduced. One of the main tasks in updating the database is to sort found datasets to categories mentioned in 4.3.2. If a dataset can be sorted to more than one category, the respective row is copied and added directly underneath (see section 4.3.5 for more detailed information about categories). Furthermore, columns “Category\_check”, “Dataset\_check”, and “Comment” were added for supervision. Those categories will only be switched to “done” when categories and assembled dataset (see 4.3.6) were checked by supervisors and no problems occurred.

### Categorizing datasets

Datasets are categorized according to the scientific research topic of interest. The choice should always be validated by a supervisor, but general rules are the following:

*Kim Mountains* is a specific category reserved for results from Kim et al., 2001. In the category *Mutants* all differentially expressed genes that show up in mutants or upon RNA interference-silencing of a particular gene are ordered. Datasets that show exposure or feeding of various microorganism are categorized in *Microbes* (WormExp also calls this category *Pathogens*). *TF Targets* is for transcription factor targets inferred by knock-down/knock-out of the respective transcription factors. The category *Tissue* is for gene expressions in specific tissues. *Development/Dauer/Aging* includes differential expression in the various developmental stages and during aging. *DAF/Insulin/food* has differential expression in response to food, starvation, or insulin-like receptor activation/de-activation. The category *Chemicals/stress* incorporates exposure to chemical compounds or other stressors and *Other* includes all gene sets not categorized. A new category *Epigenetics* has been added. It includes all gene sets that came from chromatin studies or epigenetic markers. If a data set fits more than one category, it will be added to all of them.

### Assembling data sets

Next to the transcriptomics file, another file must be updated for the database. For the following dataset assembly, the main work will be conducted in the WormExp\_info file. Please make a copy of the newest WormExp\_info file and work in there for all next steps.

Chart, scatter chart

Description automatically generated

Figure 4: Excerpt from WormExp\_info file

All new and added datasets must be added in the same manner as can be seen in the existing file. The columns are described in detail in the following table:

|  |  |
| --- | --- |
| Column | Description |
| WormBaseVersion | contains information which WormBase Version was used to map Entrez IDs to gene IDs. If n.a. no information was given. More information in 4.3.7. |
| Category\_1 | same as in transcriptomics file. |
| Additional\_categories | If gene set is applicable for more than one data set, add all additional categories here, separated by “;”. |
| Gene Set name | Explanation for the assembled gene set. |
| number\_genes | number of genes collected for respective gene set |
| Refs | Reference to paper |
| Data From | gives information where the gene set was found in the paper |
| Selection\_criteria | Shows which selection criteria was applied when extracting the data set. More information in 4.3.6.1 |
| decided\_by | gives information if the selection\_criteria was given by the authors or if it was decided by the assembler |
| Rawdata | Information to GSE accession number |
| Additional | Column for additional information |
| Comment | Add here the date, when this data set has been added (roughly, can also be done at the end with the same date for all data sets) |

Assembling the dataset is the trickiest and most error-prone part in this work. This work cannot be streamlined, as every scientist analyzed their work differently and uploaded it in different places, and in various formats. However, most of the times a differential gene expression analysis is uploaded in a separate table and can be found in the supplementary of the respective paper. Depending on the authors, selection criteria are more or less strict. The assembler must decide in most cases which genes to extract. In 4.3.7.1 some guidelines for selection criteria have been decided.

For every GSE number an overall gene set name should be chosen that describes sufficiently the experiment conducted. Every gene set extracted from the respective experiment should be saved in a .txt. In this .txt file all gene sets belonging to this experiment should be saved. Please save all found gene sets in independent .txt files for now, as these gene sets have to be checked over by a supervisor before fusion.

#### Criteria for Supplementary Data Filtering

As mentioned, every scientist employed their own significance selection criteria. If possible, selection criteria from the authors should be used. These criteria will be added in the column „selection\_criteria” and “decided\_by”.

Exceptions come into play if the selection criteria are not strict enough (e.g., p-value > 0.1 without any adjustments and without corrections). In general, p-fdr/padj < 0.01 and log-foldchange >= 2 (or <= -2) should be used.

#### Good Practice for Gene Set Names

Some guidelines have been established to ensure a good practice of naming gene sets. Additionally, it is of utmost importance, that no gene set name occurs twice!

1. If a gene has been regulated by a mutation, choose “in” instead of “by” (e.g., “Up in daf-2”).
2. Add first author name if gene sets with similar experiments are already in the database. This is very important for the category *Mutants*. The database contains e.g., several gene sets about daf-2. To differentiate between those, the author name has been added.
3. If the C.elegans strain is of importance, write strains are in Upper Case.
4. If gene alleles are of importance, write in lower case

### WormBase ID

WormBase IDs should only be mapped, when all categories and datasets have been checked by the supervisor. In this state of the project, it is recommended to make a copy of the current WormSource database and delete all data sets in the respective category files. This way, all gene sets can be collected, and a test run (see 4.3.9) with only the new data sets can be conducted. Afterwards and if no errors are found in the test run, the category files can be fused at once with the old database.

If a gene set has been curated, it is possible that WormBase IDs (WBGeneXXXXXXXX) are still missing. As there is no standard methodology of uploading differential gene expression analysis, some datasets only have gene IDs or gene names, whereas other already possess WormBase IDs. If WormBase IDs are provided, add WormBase Version to the file WormExp\_info (column WormBaseVersion) and this gene set can be directly added to respective category files. This information should be provided in the paper (but not always…). If no information is given in the paper about a WormBaseVersion, add n.a.

For all gene sets, that are not in WormBase ID format, use the following procedure:

1. Go to <https://wormbase.org//tools/mine/simplemine.cgi>. It’s an online tool provided by WormBase to get WormBase ID for found genes.
2. In Step 1 choose Caenorhabditis elegans.
3. In Step 2 check “case insensitive input”, “download results as a tab-delimited file”, and “keep duplicate gene entries in results”.
4. In Step 3 uncheck everything besides WormBase Gene ID.

It is helpful to use an Excel file to update the curated gene set from gene IDs to WormBase IDs. As “keep duplicate gene entries in results” was marked, the downloaded file from SimpleMine can have entries such as “Multiple entries found” or also “not found”. In the first case, keep both WormBase IDs found for this gene. In the latter case, delete the found gene. In any case, adjust the number of curated genes in the WormExp\_info file when deleting or adding multiple genes. Also add the respective WormBase Version that has been used for mapping the WormBase IDs.

If all gene sets have a WormBase ID mapping, they can be added to the category files of the database copy. In “./07\_Wormbase/Wormbase\_version\_changes.xlsx” add changes that were made by Wormbase in the last updates. This Excel file should also be added to the dataset.zip package as reference for other scientists.

### Update database

If all category files have been updated, several other files have to be updated as well and the procedure is explained here.

#### c\_elegans.WSXXX.geneIDs

This file is a WormBaseID reference file for the database. It should contain the current Version of WormBase. To update this file, follow these steps:

1. Go to <https://wormbase.org//tools/mine/simplemine.cgi>.
2. In Step 1 choose Caenorhabditis elegans.
3. In Step 2 check “case insensitive input”, “download results as a tab-delimited file”, and “keep duplicate gene entries in results”.
4. In Step 3 uncheck everything besides WormBase Gene ID, Public Name and Sequence Name.
5. In Step 4 click “Query all genes in this species” to get a full list of the current WormBase version.

Change this content to the same format as is currently used in the file c\_elegans.WSXXX.geneIDs (separated by comma instead of tab). Change the name of the file accordingly and add it to the WormSource folder.

#### Reference file

WormExp has an additional reference file that only contains the gene set name and its respective source (PubMed link, etc.). As this information is automatically curated in WormExp\_info, updating the reference file is fairly straightforward. Copy the columns “gene set names”, “Refs” and “additional\_categories”. Then the new table can be saved as tab delimited .txt file. This file should also be added in the database copy and can be fused later on with the pre-existing reference file.

!**NB: WormExp uses gene set names to find hits and references in its analysis. This means, that gene set names in the category have to match EXACTLY the gene set names in the reference file. Additional spaces, other symbols or lower/upper case all lead to problems if they do not match exactly. Therefore, when curating the gene sets take care that the names are exactly the same. For leading/trailing white spaces, a script has been provided that takes care of those. Find it in “04\_Scripts/reference\_cleaning.R”, open it and follow the instructions in the comments.**

#### Category files

Fuse all newly collected gene sets with the pre-existing category files by simply adding them under the entries. Make sure that no empty lines are at the end of the file! Otherwise, an error will occur when booting the server.

#### dat.properties

In dat.properties, add the name of idfile, if a new wormbase version has been added to the database. Additionally, if a new data set category is added, add it to the bottom of dataset in the same format as all the other data sets.

Text

Description automatically generated

#### “tomcat” folder

In the at the beginning mentioned “tomcat” folder, some small changes have to be made. Go into “./tomcat/webapps/wormexp/WEB-INF” and find the file web.properties. In web.properties some texts of the front end can be changed if so wished. Additionally, add a new data set category, if wished. Keep the format as indicated by other categories.

Text

Description automatically generated

Here, you can also find wormexp.properties. In wormexp.properties more changes on the displayed text can be made. Important changes are the update of the current version as well as the Wormbase version (if necessary).

Text

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application, chat or text message

Description automatically generatedAnother important and necessary update is the change of downloadable data sets from the website. Zip all updated category files plus the newest WormExp.info and replace the file in “./tomcat/webapps/wormexp/upload” as well as in “WormExpData”. This changes the downloadable content here:

### Test Runs

#### Apache TomCat

To test if all changes were implemented correctly, test runs will be applied. In order to run the server locally and check all changes, Java needs to be installed, as well as Apache Tomcat. Find all links to the softwares at the start of the documentation.

The following steps describe how to locally run the server to conduct test runs:

1. Make sure, that the file test.jar is in the folder that also contains all text files with categories, references and WormBaseIDs. Run Windows Powershell and go into the project directory that contains test.jar.
2. Run: “java -classpath test.jar com.dem.test.HiServer”
3. Go to “.\Apache Software Foundation\Tomcat 8.5\webapps” and copy the “tomcat” folder from the database folder into it.
4. Run TomCat (you can find the .exe file for that in “./Tomcat 8.5/bin/Tomcat8”)
5. Go into your browser and write “localhost:8080/wormexp/”
6. There should now be an exact copy of the WormExp website in your browser.

With this locally run server, test runs can be conducted, and changes can be made.

#### Selecting Test sets

Test gene sets should be selected appropriate to the changes. This choice can be made in accordance with the supervisors. Test sets should be saved in the folder “./05\_QualityManagement/OQ\_PQ”.

As a baseline test, the newly curated gene sets should be tested alone on the current version of WormExp. If no errors occurred and all references can be found, then new gene sets can be fused with the old library and the tests repeated.

Goal of the test run is to compare outputs of old version and new database version and check that it correctly adds new gene sets. Additionally, if new categories were added, it should be checked that those are selectable, and the server correctly accesses the new data files.

### Updating Server

As a last step, all files need to be updated on the Server where WormExp is hosted. For that, one needs to be in the network of the university of Kiel. Either, directly by entering eduroam or using FortiClient VPN, provided by the Rechenzentrum.

Afterwards, open Filezilla and add the following information in the top row:

* Host: applux05.rz.uni-kiel.de
* Username: sunzm471
* Passwort: *(ask supervisor)*
* Port: 22

Graphical user interface, application

Description automatically generated

Figure 5: Exempt from FileZilla

At the bottom of FileZilla, you can now see the files that are hosted on this server. If everything has been done according to this documentation, you only need to update all mentioned files by overwriting them on the server. If structural updates have been added, best to replace both “tomcat” and “WormExpData” folders. Check if the website still works after the update by making a small test run (see 4.3.10).

### Checklist

Use this checklist to keep track of all files that need to be updated and/or fused.

|  |  |  |
| --- | --- | --- |
| **To-Do** | **Updated?** | **Fused?** |
| Wormbase\_version\_changes | yes | Not necessary |
| c.elegans.WSXXX.geneIDs | updated to from WS235 to WS283 | Not necessary |
| reference | Yes | Yes |
| Chemicalexposure-otherStress | Yes | Yes |
| DAF Insulin food | Yes | Yes |
| Development-Dauer-Aging | Yes | Yes |
| Kim Mounts | no | Not necessary |
| Mutants | Yes | Yes |
| Other | no | Not necessary |
| Pathogen (Microbes) | Yes | Yes |
| Targets | Yes | Yes |
| Tissue-specific | Yes | Yes |
| Epigenetics | Newly created; yes | Not necessary |
| WormExp\_info | Yes | No |
| Dat.properties | Yes | Not necessary |
| Web.properties | Yes | Not necessary |
| Wormexp.properties | Yes | Not necessary |

1. **Document History**

Reason for change

|  |  |  |
| --- | --- | --- |
| **Version number** | **Description of change** | **Valid from:** |
| 01 | 1. Version for establishing internal standards | 27.01.2022 |

Publication bibliography

Dozmorov, Mikhail (2016): GEOparse. Reading the NCBI’s GEO microarray SOFT files in R/BioConductor: public domain.

evogen (2015): Manual for WormExp, 2015. Available online at https://academic.oup.com/bioinformatics/article/32/6/943/1744078.

Yang, Wentao; Dierking, Katja; Schulenburg, Hinrich (2016): WormExp: a web-based application for a Caenorhabditis elegans-specific gene expression enrichnment analysis. In *Bioinformatics Advance Access* 32 (6).