Updating WormExp

Table of Contents:

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

1 Goal of the documentation

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

2 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.
Project Owner	Can allow and decline access to the database and controls the quality of the update.

3 Term/Definition/Abbreviation

GEO	Gene Expression Omnibus

4 Procedure/ Workflow

4.1 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/
PuTTY (or similar)	PuTTY is an SSH and telnet client, developed originally by Simon Tatham for the Windows platform.	0.77	https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html

4.2 Security details

N/A

4.3 Method procedure

4.3.1 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 curated 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- or an employee number. The access to the server lies with the project owner.

Feel free to update this documentation if some things are described unclearly or new additions are implemented to make updating WormExp easier.

4.3.2 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, but also the underlying JAVA project.
02_ServerFiles	A copy of the current files on the active server plus former versions.
03_Documentation	All documentation and information about his project and its structure.
04_Scripts	contains all scripts that have been used for updating the database, but also helpful scripts for visualization purposes.
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. Please update this repository with your updates but always keep in mind that it is available to the public! Exclude sensible data from your uploads and always keep the repo clean and representable. For example, supplementary data that you downloaded to curate gene sets do not need to be uploaded as well as all the intermediate processes while updating.

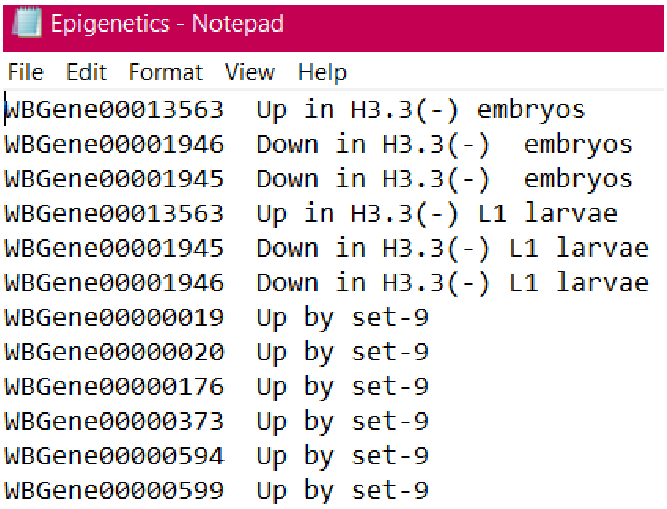
4.3.3 WormExp

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 and 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, currently based on Wormbase version WS283.
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 Epigenetics
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

Most important for the update are the mentioned category text files. They contain all current gene sets and all curated gene sets need to be added to these files, to make them available



```
WBGene00013563  Up in H3.3(-) embryos
WBGene00001946  Down in H3.3(-)  embryos
WBGene00001945  Down in H3.3(-)  embryos
WBGene00013563  Up in H3.3(-) L1 larvae
WBGene00001945  Down in H3.3(-) L1 larvae
WBGene00001946  Down in H3.3(-) L1 larvae
WBGene00000019  Up by set-9
WBGene00000020  Up by set-9
WBGene00000176  Up by set-9
WBGene00000373  Up by set-9
WBGene00000594  Up by set-9
WBGene00000599  Up by set-9
```

Figure 1: Exempt from Epigenetics category

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.

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! 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, instead count up the current version (e.g., v2.1, v2.2, etc.). A new WormExp version should only be done when changes on the code have been added.

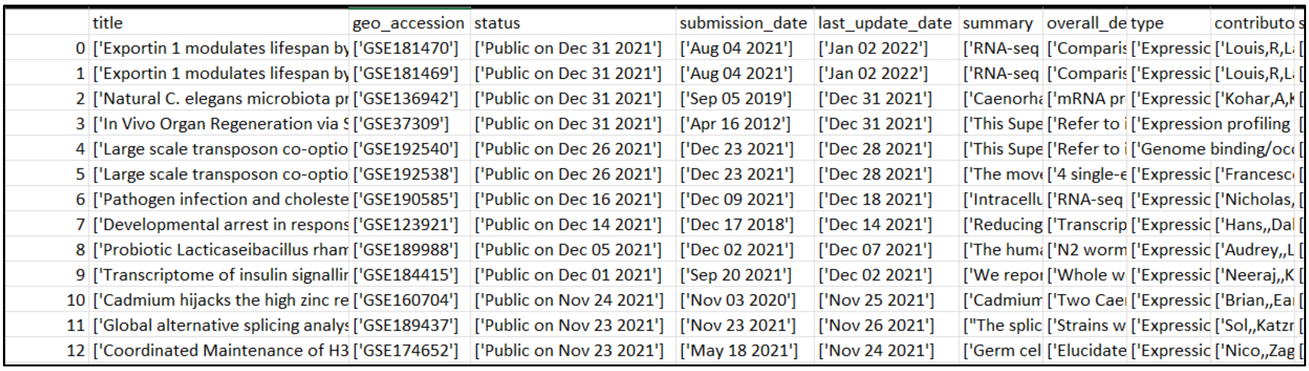
4.3.4 Finding new GEO datasets

To find new GEO datasets, the jupyter notebook "datafinder" can be used. In order to use the notebook, the software from section 4.1 has to be installed. The notebook 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.

Adaptions to the query should only be made in respect to the publication date. As of 02/2022 the database contains datasets until approx. 2019. The script will create a separate Excel file called "GEO_database_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".

4.3.5 Transcriptomics File

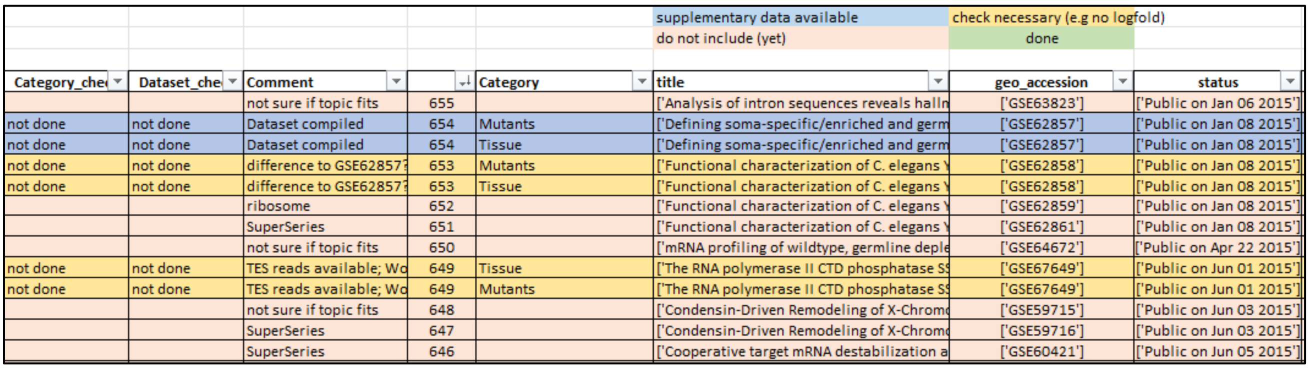
The datafinder script produces an Excel file that contains detailed information about every dataset found with GEOparse. This file can be found in “06_Datafinder/GEO_database_results.xlsx”. It shows not only the exact title of the dataset, but also its geo_accession number, publication date, contributors, etc.



	title	geo_accession	status	submission_date	last_update_date	summary	overall_de	type	contributo	s
0	['Exportin 1 modulates lifespan by	['GSE181470']	['Public on Dec 31 2021']	['Aug 04 2021']	['Jan 02 2022']	['RNA-seq	['Comparis	['Expressic	['Louis,R,L:	[
1	['Exportin 1 modulates lifespan by	['GSE181469']	['Public on Dec 31 2021']	['Aug 04 2021']	['Jan 02 2022']	['RNA-seq	['Comparis	['Expressic	['Louis,R,L:	[
2	['Natural C. elegans microbiota pr	['GSE136942']	['Public on Dec 31 2021']	['Sep 05 2019']	['Dec 31 2021']	['Caenorha	['mRNA pr	['Expressic	['Kohar,A,K	[
3	['In Vivo Organ Regeneration via S	['GSE37309']	['Public on Dec 31 2021']	['Apr 16 2012']	['Dec 31 2021']	['This Supe	['Refer to i	['Expression profiling		[
4	['Large scale transposon co-optio	['GSE192540']	['Public on Dec 26 2021']	['Dec 23 2021']	['Dec 28 2021']	['This Supe	['Refer to i	['Genome binding/occ		[
5	['Large scale transposon co-optio	['GSE192538']	['Public on Dec 26 2021']	['Dec 23 2021']	['Dec 28 2021']	['The move	['4 single-e	['Expressic	['Francesc	[
6	['Pathogen infection and choleste	['GSE190585']	['Public on Dec 16 2021']	['Dec 09 2021']	['Dec 18 2021']	['Intracellu	['RNA-seq	['Expressic	['Nicholas,	[
7	['Developmental arrest in respons	['GSE123921']	['Public on Dec 14 2021']	['Dec 17 2018']	['Dec 14 2021']	['Reducing	['Transcrip	['Expressic	['Hans,,Dal	[
8	['Probiotic Lacticaseibacillus rham	['GSE189988']	['Public on Dec 05 2021']	['Dec 02 2021']	['Dec 07 2021']	['The huma	['N2 worm	['Expressic	['Audrey,,L	[
9	['Transcriptome of insulin signallir	['GSE184415']	['Public on Dec 01 2021']	['Sep 20 2021']	['Dec 02 2021']	['We repo	['Whole w	['Expressic	['Neeraj,,K	[
10	['Cadmium hijacks the high zinc re	['GSE160704']	['Public on Nov 24 2021']	['Nov 03 2020']	['Nov 25 2021']	['Cadmium	['Two Cae	['Expressic	['Brian,,Ea	[
11	['Global alternative splicing analys	['GSE189437']	['Public on Nov 23 2021']	['Nov 23 2021']	['Nov 26 2021']	["The splic	['Strains w	['Expressic	['Sol,,Katzr	[
12	['Coordinated Maintenance of H3	['GSE174652']	['Public on Nov 23 2021']	['May 18 2021']	['Nov 24 2021']	['Germ cel	['Elucidate	['Expressic	['Nico,,Zag	[

Figure 2: Exempt from original Transcriptomics file

The main task here is to sort through the found datasets and find 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. The latter can happen, as the GEO query focuses mainly on experiments using high-throughput sequencing data, whict mostly (but not necessarily) conducts differential gene analysis.



					supplementary data available	check necessary (e.g no logfold)	
					do not include (yet)	done	
Category_che	Dataset_che	Comment		Category	title	geo_accession	status
		not sure if topic fits	655		['Analysis of intron sequences reveals halln	['GSE63823']	['Public on Jan 06 2015']
not done	not done	Dataset compiled	654	Mutants	['Defining soma-specific/enriched and germ	['GSE62857']	['Public on Jan 08 2015']
not done	not done	Dataset compiled	654	Tissue	['Defining soma-specific/enriched and germ	['GSE62857']	['Public on Jan 08 2015']
not done	not done	difference to GSE62857?	653	Mutants	['Functional characterization of C. elegans Y	['GSE62858']	['Public on Jan 08 2015']
not done	not done	difference to GSE62857?	653	Tissue	['Functional characterization of C. elegans Y	['GSE62858']	['Public on Jan 08 2015']
		ribosome	652		['Functional characterization of C. elegans Y	['GSE62859']	['Public on Jan 08 2015']
		SuperSeries	651		['Functional characterization of C. elegans Y	['GSE62861']	['Public on Jan 08 2015']
		not sure if topic fits	650		['mRNA profiling of wildtype, germline deple	['GSE64672']	['Public on Apr 22 2015']
not done	not done	TES reads available; Wo	649	Tissue	['The RNA polymerase II CTD phosphatase S	['GSE67649']	['Public on Jun 01 2015']
not done	not done	TES reads available; Wo	649	Mutants	['The RNA polymerase II CTD phosphatase S	['GSE67649']	['Public on Jun 01 2015']
		not sure if topic fits	648		['Condensin-Driven Remodeling of X-Chrom	['GSE59715']	['Public on Jun 03 2015']
		SuperSeries	647		['Condensin-Driven Remodeling of X-Chrom	['GSE59716']	['Public on Jun 03 2015']
		SuperSeries	646		['Cooperative target mRNA destabilization a	['GSE60421']	['Public on Jun 05 2015']

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

Additionally, new columns have been introduced. Columns “Category_check”, “Dataset_check”, and “Comment” were added for supervision purposes. 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.

4.3.6 Categorizing datasets

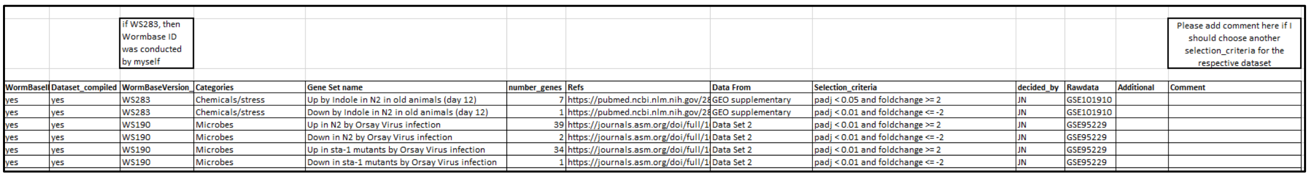
Datasets are categorized according to the scientific research topic of interest. The choice should always be validated by a supervisor, but general rules are the following:

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.

If a dataset can be sorted to more than one category, the respective row in the Transcriptomics file is copied and added directly underneath.

4.3.7 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.



if WS283, then Wormbase ID was conducted by myself

Please add comment here if I should choose another selection_criteria for the respective dataset

WormBaseI	Dataset_compiled	WormBaseVersion_	Categories	Gene Set name	number_genes	Refs	Data From	Selection_criteria	decided_by	Rawdata	Additional	Comment
yes	yes	WS283	Chemicals/stress	Up by Indole in N2 in old animals (day 12)	7	https://pubmed.ncbi.nlm.nih.gov/28	GEO supplementary	padj < 0.05 and foldchange >= 2	JN	GSE101910		
yes	yes	WS283	Chemicals/stress	Down by Indole in N2 in old animals (day 12)	1	https://pubmed.ncbi.nlm.nih.gov/28	GEO supplementary	padj < 0.01 and foldchange <= -2	JN	GSE101910		
yes	yes	WS190	Microbes	Up in N2 by Orsay Virus infection	39	https://journals.asm.org/doi/full/1	Data Set 2	padj < 0.01 and foldchange >= 2	JN	GSE95229		
yes	yes	WS190	Microbes	Down in N2 by Orsay Virus infection	2	https://journals.asm.org/doi/full/1	Data Set 2	padj < 0.01 and foldchange <= -2	JN	GSE95229		
yes	yes	WS190	Microbes	Up in sta-1 mutants by Orsay Virus infection	34	https://journals.asm.org/doi/full/1	Data Set 2	padj < 0.01 and foldchange >= 2	JN	GSE95229		
yes	yes	WS190	Microbes	Down in sta-1 mutants by Orsay Virus infection	1	https://journals.asm.org/doi/full/1	Data Set 2	padj < 0.01 and foldchange <= -2	JN	GSE95229		

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 implemented for consistency.

For every GSE number an overall gene set name should be chosen that describes sufficiently the experiment conducted (see 4.3.7.2 for good practice in naming gene sets). 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.

4.3.7.1 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" (you or author).

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.

4.3.7.2 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! Use a current version of WormExp_info to check for duplicates.

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’s name should be added.
3. If the *C.elegans* strain is of importance, write strains in Upper Case.
4. If gene alleles are of importance, write in lower case

4.3.8 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 collect all curated gene sets in its respective categories, still isolated from old gene sets. 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. 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 is helpful information for scientists who used WormExp in former versions and want to re-run some of their data.

4.3.9 Update database

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

4.3.9.1 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 WormExp folder.

4.3.9.2 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.

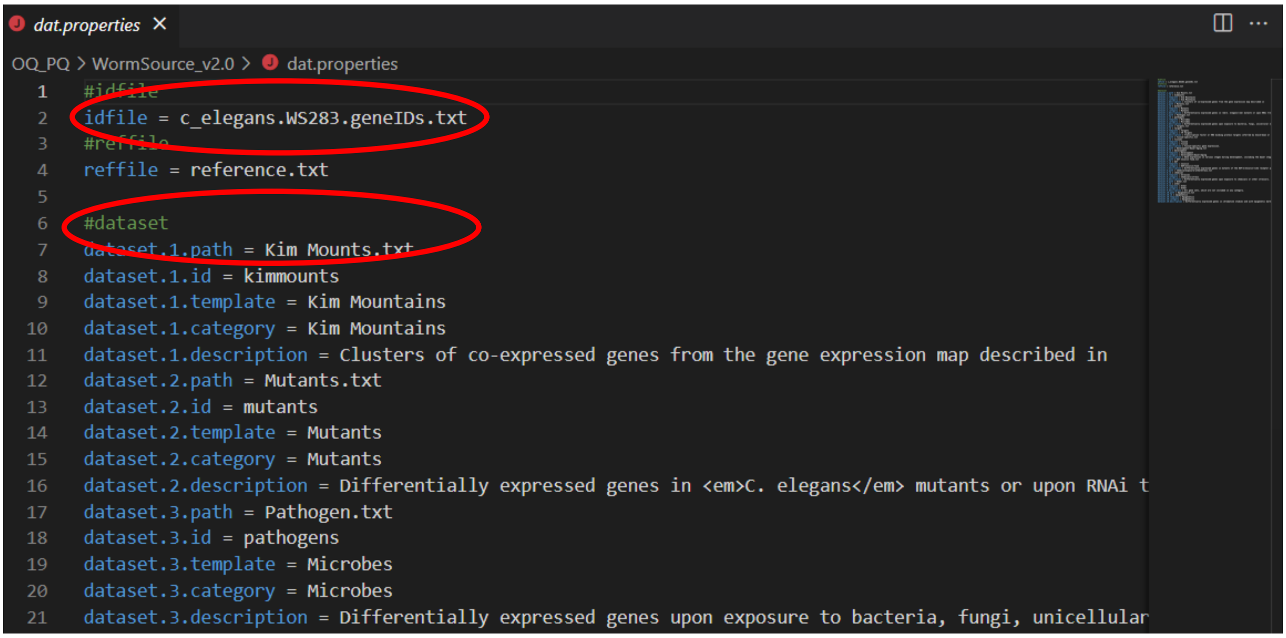
NB: WormExp uses gene set names to find hits and references in its analysis. This means, that gene set names in the category have to match EXACTLY the gene set names in the reference file. Additional spaces, other symbols or lower/upper case all lead to problems if they do not match exactly. Therefore, 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.

4.3.9.3 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.

4.3.9.4 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.

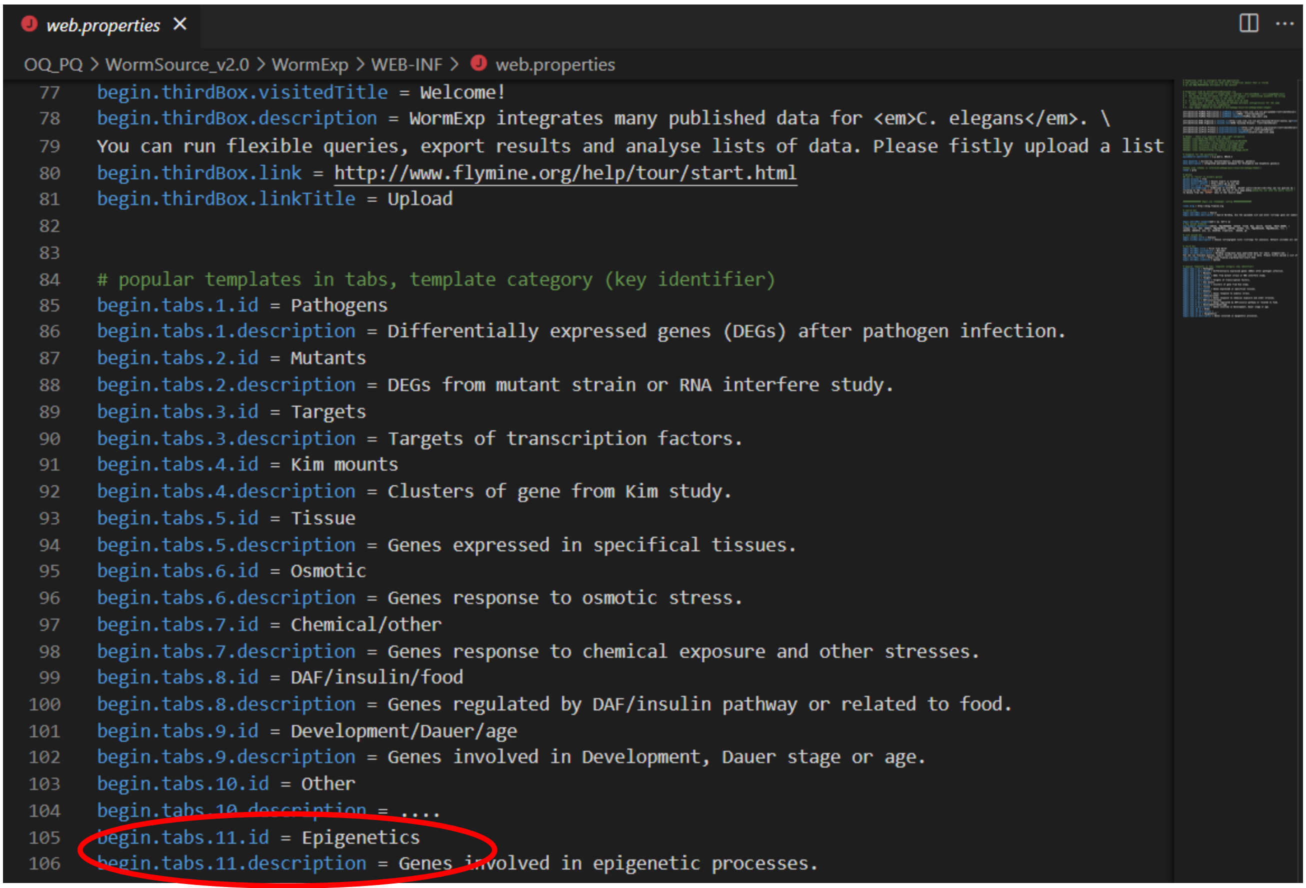


```
#idfile
idfile = c_elegans.WS283.geneIDs.txt
#reffile
reffile = reference.txt

#dataset
dataset.1.path = Kim Mounts.txt
dataset.1.id = kimmounts
dataset.1.template = Kim Mountains
dataset.1.category = Kim Mountains
dataset.1.description = Clusters of co-expressed genes from the gene expression map described in
dataset.2.path = Mutants.txt
dataset.2.id = mutants
dataset.2.template = Mutants
dataset.2.category = Mutants
dataset.2.description = Differentially expressed genes in <em>C. elegans</em> mutants or upon RNAi t
dataset.3.path = Pathogen.txt
dataset.3.id = pathogens
dataset.3.template = Microbes
dataset.3.category = Microbes
dataset.3.description = Differentially expressed genes upon exposure to bacteria, fungi, unicellular
```

Figure 5: Updating dat.properties

4.3.9.5 “tomcat” folder

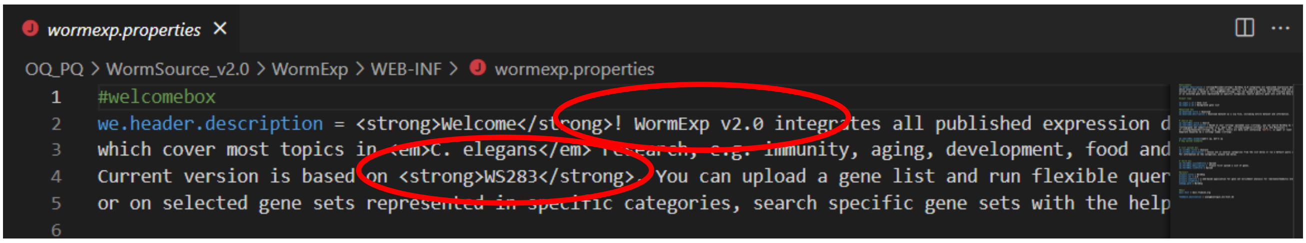
In the at the beginning mentioned “tomcat” folder, some small changes have to be made. Go into “./tomcat/webapps/wormexp/WEB-INF” and find the file web.properties. In web.properties some texts of the front end can be changed if so wished. Additionally, add a new data set category, if wished. Keep the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


web.properties
OQ_PQ > WormSource_v2.0 > WormExp > WEB-INF > web.properties

```
begin.thirdBox.visitedTitle = Welcome!
begin.thirdBox.description = WormExp integrates many published data for <em>C. elegans</em>. \
You can run flexible queries, export results and analyse lists of data. Please fistly upload a list
begin.thirdBox.link = http://www.flymine.org/help/tour/start.html
begin.thirdBox.linkTitle = Upload


# popular templates in tabs, template category (key identifier)
begin.tabs.1.id = Pathogens
begin.tabs.1.description = Differentially expressed genes (DEGs) after pathogen infection.
begin.tabs.2.id = Mutants
begin.tabs.2.description = DEGs from mutant strain or RNA interfere study.
begin.tabs.3.id = Targets
begin.tabs.3.description = Targets of transcription factors.
begin.tabs.4.id = Kim mounts
begin.tabs.4.description = Clusters of gene from Kim study.
begin.tabs.5.id = Tissue
begin.tabs.5.description = Genes expressed in specifical tissues.
begin.tabs.6.id = Osmotic
begin.tabs.6.description = Genes response to osmotic stress.
begin.tabs.7.id = Chemical/other
begin.tabs.7.description = Genes response to chemical exposure and other stresses.
begin.tabs.8.id = DAF/insulin/food
begin.tabs.8.description = Genes regulated by DAF/insulin pathway or related to food.
begin.tabs.9.id = Development/Dauer/age
begin.tabs.9.description = Genes involved in Development, Dauer stage or age.
begin.tabs.10.id = Other
begin.tabs.10.description = ....
begin.tabs.11.id = Epigenetics
begin.tabs.11.description = Genes involved in epigenetic processes.
```

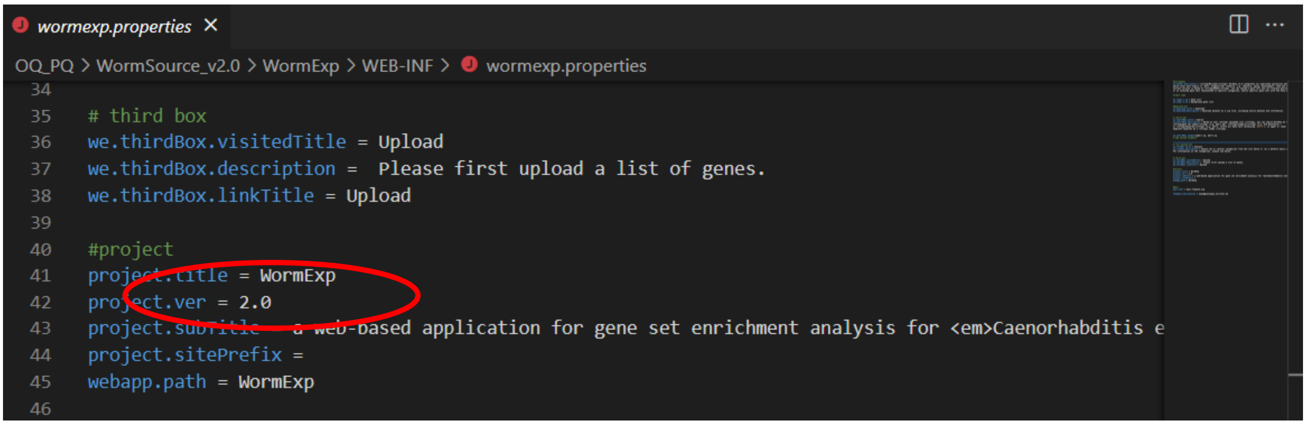
Figure 6: Updating web.properties

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).


wormexp.properties
OQ_PQ > WormSource_v2.0 > WormExp > WEB-INF > wormexp.properties

```
#welcomebox
we.header.description = <strong>Welcome</strong>! WormExp v2.0 integrates all published expression d
which cover most topics in <em>C. elegans</em> research, e.g. immunity, aging, development, food and
Current version is based on <strong>WS283</strong>. You can upload a gene list and run flexible quer
or on selected gene sets represented in specific categories, search specific gene sets with the help
```

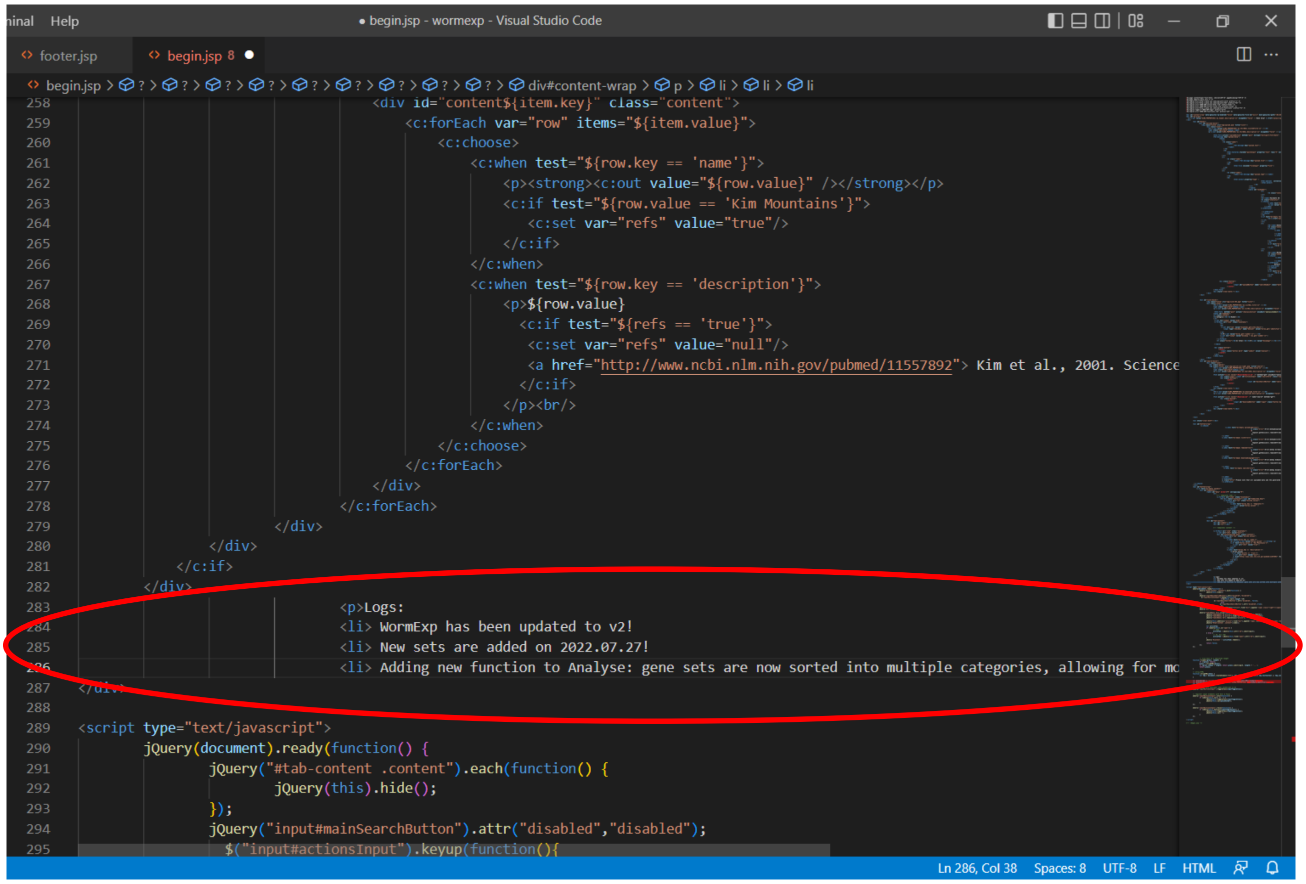
Figure 7: Updating wormexp.properties I


wormexp.properties ×
OQ_PQ > WormSource_v2.0 > WormExp > WEB-INF > wormexp.properties

```

# third box
we.thirdBox.visitedTitle = Upload
we.thirdBox.description =  Please first upload a list of genes.
we.thirdBox.linkTitle = Upload

#project
project.title = WormExp
project.ver = 2.0
project.subTitle = a web-based application for gene set enrichment analysis for <em>Caenorhabditis e
project.sitePrefix =
webapp.path = WormExp

```

Figure 8: Updating wormexp.properties II

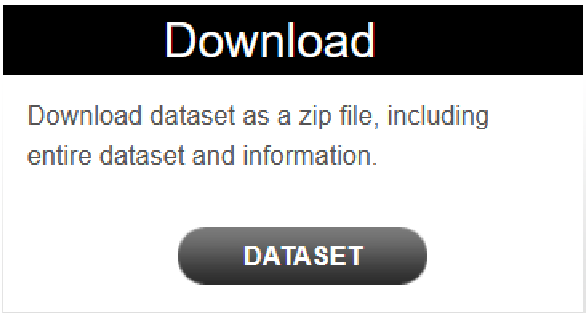
On the bottom of the WormExp website is a paragraph called “Logs”. It contains short news about the website. Depending on the update, add newest changes there. You can find the location to make the changes in “./tomcat/webapps/wormexp/begin.jsp”, line 283f:


inal Help
• begin.jsp - wormexp - Visual Studio Code
footer.jsp
begin.jsp 8
begin.jsp > ? > ? > ? > ? > ? > ? > ? > ? > ? > div#content-wrap > p > li > li > li
Ln 286, Col 38 Spaces: 8 UTF-8 LF HTML

```
258                                     <div id="content${item.key}" class="content">
259                                         <c:forEach var="row" items="${item.value}">
260                                             <c:choose>
261                                                 <c:when test="${row.key == 'name'}">
262                                                     <p><strong><c:out value="${row.value}" /></strong></p>
263                                                     <c:if test="${row.value == 'Kim Mountains'}">
264                                                        <c:set var="refs" value="true"/>
265                                                     </c:if>
266                                                 </c:when>
267                                                 <c:when test="${row.key == 'description'}">
268                                                     <p>${row.value}
269                                                       <c:if test="${refs == 'true'}">
270                                                        <c:set var="refs" value="null"/>
271                                                        <a href="http://www.ncbi.nlm.nih.gov/pubmed/11557892"> Kim et al., 2001. Science
272                                                       </c:if>
273                                                     </p><br/>
274                                                 </c:when>
275                                             </c:choose>
276                                         </c:forEach>
277                                     </div>
278                                 </c:forEach>
279                         </div>
280                 </div>
281             </c:if>
282         </div>
283                                 <p>Logs:
284                                 <li> WormExp has been updated to v2!
285                                 <li> New sets are added on 2022.07.27!
286                                 <li> Adding new function to Analyse: gene sets are now sorted into multiple categories, allowing for mo
287 </div>
288
289 <script type="text/javascript">
290         jQuery(document).ready(function() {
291                 jQuery("#tab-content .content").each(function() {
292                         jQuery(this).hide();
293                 });
294                 jQuery("input#mainSearchButton").attr("disabled","disabled");
295                   $("input#actionsInput").keyup(function(){
```

Figure 9: Updating logs

Another 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:



Download

Download dataset as a zip file, including entire dataset and information.

Figure 10: Download Button in WormExp

4.3.10 Test Runs

4.3.10.1 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 software at the start of the documentation. The instructions here are written for Windows.

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 folder “tomcat/webapps/wormexp” 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.

4.3.10.2 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.

4.3.11 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

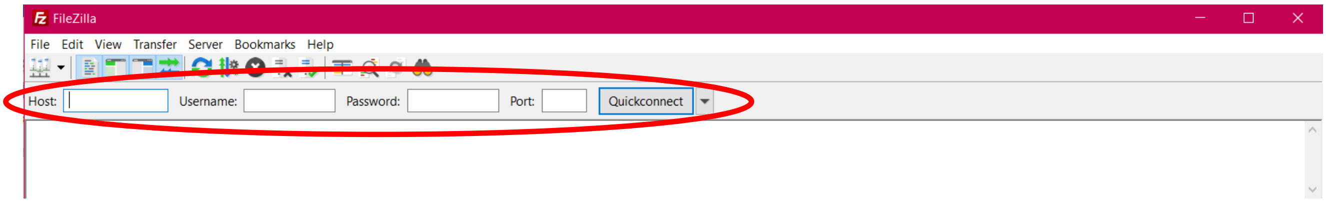
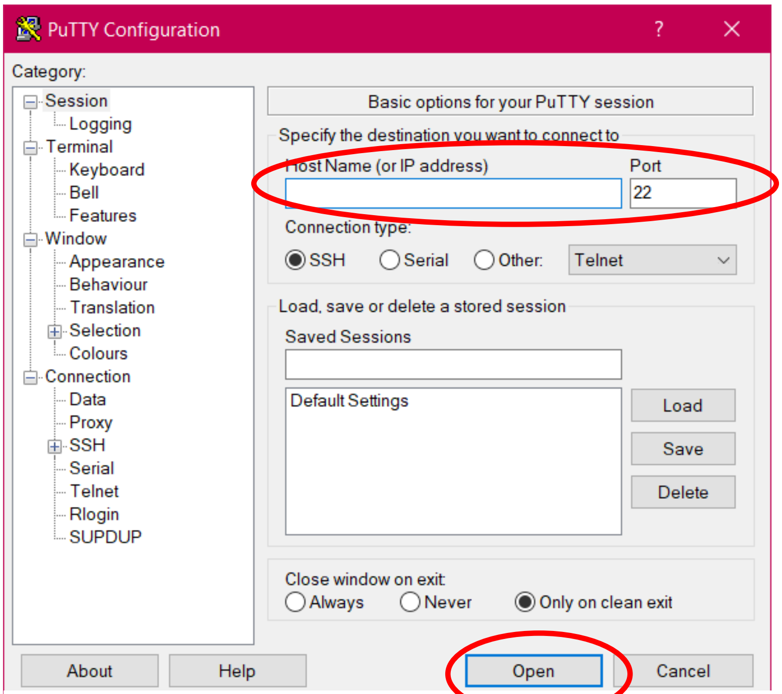


Figure 11: Exempt from FileZilla

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/webapps/wormexp” and “WormExpData” folders, but you can also update by a folder-by-folder basis. Always make sure to make a complete backup of the Server before you change anything! You can use the folder “./02_ServerFiles” for that.

After changing/adding folders, the tomcat service needs to restart. For that, you need to access the server via an SSH client. Here, PuTTY is used, but you can use any client you favor. After installing, add the host address in Host Name and click “Open”.



Specify the destination you want to connect to

Host Name (or IP address):

Port: 22

Connection type: ● SSH ○ Serial ○ Other: Telnet

Close window on exit: ○ Always ○ Never ● Only on clean exit

Figure 12: PuTTY layout

A second window opens, and you are asked to enter your username and password. Don’t worry, when you can’t see the typed password, that is normal for PuTTY. Afterwards you should see a command line terminal. There, enter the following commands:

- sudo /sbin/service tomcat@sunzm471 stop
- sudo /sbin/service tomcat@sunzm471 start

This way, Tomcat is stopped and restarted and should incorporate all your changes. Check if the website still works after the update by making a small test run (see 4.3.10).

4.3.12 Checklist

Use this checklist to keep track of all files that need to be updated and/or fused. Empty it if you want to use it yourself.

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	Yes
Dat.properties	Yes	Not necessary
Web.properties	Yes	Not necessary
Wormexp.properties	Yes	Not necessary

5 Document History

Reason for change

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

6 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).