Downloading Data from the CEBS FTP Site

Page Navigation (Quick Links):

Navigating the CEBS FTP Site

Pease use these instructions to navigate the CEBS FTP site and locate desired data. The CEBS FTP site allows users to download study data for many studies at once and raw data for individual studies. Data on the CEBS FTP site is divided by the data type, individual studies and raw data in file lists. Within the data types, or CEBS data domains (e.g. In Life Observations, Histopathology, and Genetic Toxicity), users can download the individual animal data for similar assays. For example, the In-Life Observation data from the NTP contains information about the treatment group, body weight, clinical observations, and food consumption for all subjects in all completed public NTP studies in CEBS for which in-life data was collected. The downloadable data on the CEBS FTP site allows for independent analyses, or identification of results of interest.

Most of the CEBS data domain folders on the CEBS FTP site contain the following file types:

- PDF document describing that CEBS data domain
- Files for each institution that collected the data
    - **Readme File**: explanation of file types and data columns
    - ***.zip File**: a compressed file containing data and sample data files in tab delimited text File (*.txt) format, in addition to a keys file containing column header definitions and assay descriptions (for NTP data)

Sample File: small sample dataset of the tab delimited text file, provided as an example format for the full dataset

Access Data by Data Domain

a) Select a Data Domain

From the CEBS homepage (https://cebs.niehs.nih.gov/cebs) select the link "Download Data" (https://cebs.niehs.nih.gov/ftp) located in the "Additional Resources" section of the CEBS homepage. The user is navigated to a page titled "CEBS Study Data Downloads" that displays a list of CEBS data domains and a brief description of the data. The displayed list of data domains is subject to change. Use the table below for descriptions of some of the common data domains. To navigate to downloadable data for a particular data domain on the CEBS FTP site, select the CEBS data domain of interest (please see important Note below prior to navigating to the CEBS FTP site). The user is navigated to the CEBS FTP site displaying a list of subfolders and/or files to allow the user to browse, open or download (depending on the user's browser settings).

Note: *The following is important information that applies to both users of Safari (Mac OS) in addition to users of other browsers on Windows and Mac OS. For users of Safari (Mac OS), please see the Appendix below for special instructions before accessing the CEBS FTP site. For all other browsers, the user may be prompted to enter a username or password prior to accessing the CEBS FTP site for data domains. The directories described herein are publicly available and do not require a username or password. If the user is prompted to enter a username and password, please select the "Guest" option and proceed to the CEBS FTP site.*

CEBS Data Domains:

CEBS Data Domain	Description
Clinical Chemistry	Clinical chemistry refers to the processes used to measure levels of biochemical components in bodily fluids such as serum and plasma. This data domain contains parameters related to lipid metabolism in the liver and kidneys, and to hormone (e.g., thyroid) and electrolyte (e.g., potassium) levels.
Developmental	Developmental toxicity refers to the adverse effects on the developing organism. This data domain contains information relating to prenatal and developmental assessments. Examples of endpoints: fetal examination, markers of puberty (e.g. vaginal opening), and neuromuscular (e.g., grip strength) and neurobehavioral (e.g., motor activity) endpoints.
Genetic Toxicology: Bacterial Mutagenicity	The bacterial mutagenicity test, or Ames Assay, is used to assess whether a test article causes genetic damage. These data are bacterial colony counts.

CEBS Data Domain	Description
Genetic Toxicology: In Vitro Micronucleus	The in vitro micronucleus assay provides insight into the genotoxic damage potential of test articles by measuring the frequency of micronuclei in mammalian cell cultures. Flow cytometry is commonly used to measure DNA damage in cell cultures by counting harvested cells to determine the percentage of cells that are micronucleated. In vitro micronucleus data can include cell counts, percentages of micronucleated cells, and percentages of apoptotic cells.
Genetic Toxicology: In Vivo Comet	The in vivo comet assay examines the ability of substances to cause DNA damage in cells from a variety of different tissues in an organism, such as stomach, liver, lung, or brain. The DNA damage detected in the comet assay may be in the form of breaks or adducts, as well as transient damage resulting from normal DNA repair processes.
Genetic Toxicology: Micronucleus	The peripheral blood erythrocyte micronucleus test provides an indication of whether a test article is capable of inducing structural and/or numerical chromosomal damage, specifically micronuclei damage. This data domain contains the frequency of micronucleated erythrocytes (red blood cells) in either bone marrow or peripheral blood in response to exposure to a test article. Examples of endpoints include cell counts, percentages of micronucleated cells, and percentages of immature erythrocytes (polychromatic erythrocytes or PCEs) and mature erythrocytes (normochromatic erythrocytes or NCE) within the total erythrocyte population.
Gross Pathology Observation	Gross observations identify the morphologic alterations in tissues that are visible without the aid of a microscope. This data domain includes changes in organs observed, terminal body weight, and absolute and relative organ weights at necropsy.
Hematology	Hematology is the analysis of cells that circulate in the bloodstream – namely red blood cells, white blood cells, and platelets. Hematology analysis is done on whole blood and includes a complete blood count (CBC) and blood smear evaluation. Examples of endpoints: cell counts, red blood cell size, and hemoglobin concentration.
High Throughput Screening Data (Tox21)	The Toxicology in the 21st Century (Tox21) program is a federal collaboration that uses automated high throughput screening (HTS) methods to quickly and efficiently test chemicals for activity across a battery of assays that target cellular processes. These assays rapidly evaluate large numbers of chemicals to provide insight on potential human health effects.
Histopathology	Histopathology is the study of microscopic changes or abnormalities in tissues. This dataset domain contains descriptions of tissue abnormalities (diagnosis), site of abnormality (location), and severity of lesions. Examples of endpoints: diagnosis, distribution, and severity of lesions, and indication of whether the observed lesion is neoplastic, metastatic, systemic, or malignant.
Immunology	Immunotoxicology identifies the potential for a test article to modulate immune function in mice and/or rats. This data domain contains studies that evaluate the basic functional aspects of diverse immune system components including innate immune function and adaptive immune function (humoral mediated immunity, cell-mediated immunity).

CEBS Data Domain	Description
In Life Observations	In-life observations are recorded during the exposure period of a study and describe individual animal's response to test article. Examples of endpoints: body weight measurements, clinical observations, food and water consumption, and body temperature measurements.
Microarray	Toxicogenomics studies that employ microarrays evaluate the biological response to a toxicological challenge at the level of the genome. Results from these studies are typically reported as gene and molecular pathway read outs, which can be further analyzed for their relationship to toxicity and disease outcomes. Datasets are provided as raw, probe-level files including CEL files (Affymetrix arrays) and TXT files (CodeLink or Agilent arrays).
PCR	Toxicogenomics studies that employ the polymerase chain reaction (PCR) technique evaluate the biological response to a toxicological challenge at the level of individual genes or a small set of genes. PCR is often used to test specific gene-level hypotheses related to a toxicological challenge or validate results from larger scale microarray studies. Data are reported as the expression of a number of genes related to toxicology under different conditions.
Reproductive	Reproductive toxicity describes the adverse effect of test articles on fertility and fecundity. This data domain contains data relating to the fertility, gestation, and number of live offspring. Examples of endpoints: estrous cyclicity, sperm parameters, live pup counts per litter, and pup sex ratio.
Tox21 Phase 2 Purity	The purity data is for the NTP Tox21 Phase 2 chemicals provided to NCATS for the Tox21 Phase 2 program. This data results from the analysis of neat chemicals, prior to the preparation of the DMSO solutions sent to NCATS and should not be equated with NCATS QC Day 0 or QC Day 4 data. Tox21 Phase 2 Purity Data is also available at NTP Data Collections Guided Search: https://cebs.niehs.nih.gov/datasets/search/tox21
Urinalysis	Urinalysis measures the physical, biochemical, and microscopic properties of urine. A urinalysis includes gross examination of urine for color and clarity and measurements of urine volume and specific gravity. This data domain contains data relating to levels of electrolytes, proteins, and enzymes. Examples of biochemical endpoints: levels of glucose, creatinine, protein, and sodium.

b) Select an Institution

Most data domain folders on the CEBS FTP site contain a PDF file that includes a description of domain and also includes direct links to one or more institution folders on the CEBS FTP site.

Some exceptions to this format are:

- **GAC Database**: The available data for this domain are located in two compressed (.zip) files for gene and chromosome data from this database.

- **Genetox-Drosophila**: The only data available for this data domain is a text file with the conclusion for each study and a folder with legacy sift files. These items are located within the NTP subfolder.
- **Genetox-Mouse Lymphoma**: The only data available for this data domain is a text file with the conclusion for each study and an NTP subfolder that contains another subfolder of legacy sift files.
- **Genetox-Rodent Cytogenetics**: The only data available for this data domain is a text file with the conclusion for each study and a folder with legacy sift files. These items are located within the NTP subfolder.
- **Microarray**: This data domain features data from both institutions and specific laboratories. Additional information regarding this domain is described the in the microarray section below.

c) Download Data by Data Type

Common File Types

Most institution folders on the CEBS FTP site contain some or all of the following file types:

- **Readme File**: explanation of file types and data columns
- ***.zip File**: a compressed file containing data and sample data files in tab delimited text File (*.txt) format, in addition to a keys file containing column header definitions and assay descriptions (for NTP data)
- **Sample File**: small sample dataset of the tab delimited text file, provided as an example format for the full dataset
- **Conclusions File**: tab delimited text file of the conclusion for all studies in the data domain

The CEBS Support team recommends downloading and previewing the sample data file prior to downloading the compressed (.zip) file that contains all of the individual animal data. The sample file includes the data headings located in the individual data files located in the compressed (.zip) file. Description of column headings are described in the "keys" file for the particular data domain to assist the user in understanding the data. Select any file name on the CEBS FTP site to download the file. Please see Additional Information for full downloading and viewing instructions.

Column Headings in Data Files

The following column headings are found in most tab delimited text data files (e.g. sample data files and data files located in compressed (*.zip) files):

- STUDY_TITLE

- ACCESSION_NUMBER (a unique identification number for each individual CEBS study)
- ORGANIZATION_NAME
- FACTORS (study design variables tested in the study such as dose or time)
- CHEMICAL_NAME
- GROUP_NAME
- SUBJECT_NAME
- DOSE
- DOSE_UNIT

The remaining headings can be found in either the readme or keys files.

Exception: Microarray

The exception to this format is the microarray data domain. Within specific institution/laboratory subfolders of the microarray data domain, most data are organized by the following hierarchy:

- **Protocol Subfolder**: study data containing protocol information in .sift file format
- **Raw Data Subfolder**: a subfolder containing the CEBS accession number directory
  - **CEBS Accession Number Directory**: subfolder(s) organized by CEBS accession number containing the platform directory
    - **Platform Directory**: subfolder(s) of array type(s) / platform(s) used in each study with each subfolder containing the raw data files
      - **Raw Data Files**: Affymetrix array (*.CEL) or tab delimited text (*.txt) files containing raw, probe-level data

Additionally, three microarray datasets are provided separately from the institution of origin:

- **Elk River** – This dataset contains data for the NTP toxicogenomics study of the chemicals from the West Virginia, Elk River chemical spill. Included are an additional 3 test article subfolders and one old analysis subfolder. Each test article subfolder contains:
  - ***.zip File**: a compressed file containing Affymetrix array (*.CEL) files containing raw, probe-level data
  - **Annotation File**: Excel file containing details of data processing
  - **BMDExpress File**: Excel file containing benchmark dose analysis tools
- **MouseLiver** – This dataset contains additional data from Gene Expression Omnibus (GEO) for toxicogenomics studies of liver; dataset complied in 2012. Files provide study details and normalized data

- **DrugMatrix** – This dataset contains additional data from NTP DrugMatrix, a comprehensive rat toxicogenomics dataset, organized into individual compressed (*.zip) files by tissue and array type

Download Raw Data by Individual Study

Additionally, raw data files provided by depositors are available by selecting the parent directory link at the top of the page until the user arrives at the CEBS FTP parent directory (ftp://anonftp.niehs.nih.gov/ntp-cebs/).

These files are located in FileList and individualstudy folders. Both folders are organized by CEBS accession number, which can be found in CEBS when reviewing study data.

Example Section

The following examples demonstrate the use and applicability of the CEBS FTP site to address practical questions a user may have.

a) Example: Immunology Data for 3 Test Articles

This example addresses the question, "How can the user download the individual animal immunology data for three test articles of interest: n,n-Dimethyl-p-Toluidine, M-Nitrotoluene, and o-Nitrotoluene?". The user can access the CEBS FTP site to download all individual animal immunology data and then filter the worksheet in Excel to display only the test articles of interest.

1. From the CEBS homepage (https://cebs.niehs.nih.gov/cebs) select the link "Download Data" (https://cebs.niehs.nih.gov/ftp) to navigate to the list of CEBS data domains
2. Select the link "CEBS Immunology Data" (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/IMMUNOLOGY/) from the list of data domains.
3. Select the subfolder "NTP"
4. Select the IMMUNOLOGY_NTP.zip file to download a compressed file containing all NTP individual animal immunology data
5. In the user's file explorer on their computer, extract (unzip) the downloaded compressed file to the desired location
6. Open the extracted IMMUNOLOGY_NTP.txt file using Excel
    i. Opening the text file in Excel is performed by alternate (right) clicking the mouse on the text file, then select "Open with", then select "Excel". This particular data file does not contain Chemical Abstract Service Registry Numbers (CASRNs) and can be opened in Excel as described. However, any data file that contains CASRNs should be copied and pasted from the text file into Excel. Please see Additional Instructions for full downloading and viewing instructions.

7. Save Excel file in desired location
8. Use Excel to filter the Chemical_Name column to display only the desired three test articles
    i. Select a cell in row 1 of the worksheet
    ii. Select "Data" among the Excel menu choices (typically located at the top of the Excel application), then select the Filter icon. Dropdown menu arrows should appear in all heading cells in the first row
    iii. Select the dropdown filter menu in the Chemical_Name column (Column A). Click the top Select all option, to uncheck all chemicals. Then scroll through the list to check the chemicals of interest: n,n-Dimethyl-p-Toluidine, M-Nitrotoluene, and o-Nitrotoluene
    iv. The list is then filtered to display all immunology data for the selected three test articles

    Note: *Filtering in the manner does not display the data from the control groups. To include display of control groups data, please filter by the accession numbers (located in column C) that are associated with each selected test article of interest.*

b) Example: All PCE Micronucleus Data

This example addresses the question, "How can the user find all the genetic toxicity individual animal data related to the number of micronucleus cells found in the polychromatic erythrocytes?".

1. From the CEBS homepage (https://cebs.niehs.nih.gov/cebs) select the link "Download Data" (https://cebs.niehs.nih.gov/ftp) to navigate to the list of CEBS data domains
2. Select the link "CEBS Genetic Toxicology: Micronucleus Data" (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/GENETOX_Genetic%20Toxicology%20-%20Micronucleus/) from the list of data domains.
3. Select the subfolder "NTP"
4. Select the "GENETOX_Genetic Toxicology - Micronucleus_NTP.zip" file to download a compressed file containing all NTP individual animal micronucleus data
5. In the user's file explorer on their computer, extract (unzip) the downloaded compressed file to the desired location
6. Open the extracted text file "GENETOX_Genetic Toxicology - Micronucleus_NTP.txt" in the user's preferred text editor.

7. Press Ctrl + A (PC) or Command + A (Mac) on the keyboard to highlight the entire document
8. Alternate (right) click the mouse and select copy
9. Open or create a new spreadsheet in Excel.
10. Press Ctrl + A (PC) or Command + A (Mac) on the keyboard to highlight the entire spreadsheet
11. Under the Excel Home ribbon, in the Number section, select the dropdown menu that is displaying "General" and change to "Text"
12. Select Cell A1. Alternate (right) click the mouse and select paste
13. Save the Excel file to the desired location
14. To locate the data column header for the desired endpoint (number of micronucleus cells found in the polychromatic erythrocytes), search the "GENETOX Micronucleus NTP Keys.xlsx." file for the endpoint. The data column header for the desired endpoint is "BMPCEMNK".
15. In the Excel sheet the user previously created to view the micronucleus data, filter the BMPCEMNK column to display only results where the desired endpoint was reported.
    i. Select a cell in row 1 of the worksheet
    ii. Select "Data" among the Excel menu choices (typically located at the top of the Excel application), then select the Filter icon. Dropdown menu arrows should appear in all heading cells in the first row
    iii. Select the dropdown filter menu for the column BMPCEMNK
    iv. In the list of values, scroll down to the bottom to find (Blanks). Click to uncheck this value
    v. Only rows that contain a value for the BMPCEMNK column is displayed

Additional Instructions/Information

1. Selecting the link "Chemical Effects in Biological Systems" in the banner at the top of any CEBS webpage returns the user back to the CEBS Home page
2. Downloading and displaying data in Excel
    - Select the data file link in the CEBS FTP site to download the data
    - Depending on the user's browser setting, a screen may prompt user to open or save the file. Select "Save" and select location to save the file (if not pre-determined by browser setting). Some browsers require alternate (right) clicking the mouse on the data file link and then selecting "Save target as" to download the file.

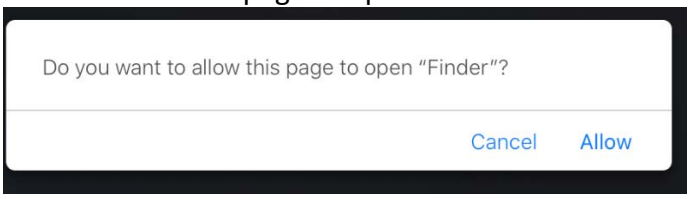
- In the user's file browser on their computer, navigate to the location the data is saved. If user downloaded a compressed (*.zip) file, extract (unzip) the file prior to proceeding
- The data is saved (or extracted) as a text file (*.txt)
- Open data file in Notepad or preferred text editor
- Press Ctrl + A (PC) or Command + A (Mac) on the keyboard to highlight the entire document
- Alternate (right) click the mouse and select copy
- Open or create a new spreadsheet in Excel
- Press Ctrl + A (PC) or Command + A (Mac) on the keyboard to highlight the entire spreadsheet
- Under the Excel Home ribbon, in the Number section, select the dropdown menu that is displaying "General" and change to "Text"
- Select cell A1
- Alternate (right) click the mouse and select paste. Formatting the entire sheet as "text" prior to pasting the data will prevent Excel from formatting some CASRNs as dates.
- Save the Excel document to the desired location
- To add filtering options, navigate to the Data ribbon in Excel and select Filter

3. The CEBS FTP directory (ftp://anonftp.niehs.nih.gov/ntp-cebs/datatype/) contains all of the NTP data domain folders in addition to other files and information
4. CEBS data may be cited by navigating to the CEBS Support page, and selecting the Citing CEBS document under FAQs. Here, instructions can be found on citing NTP Data as well as non-NTP data in CEBS
5. Please contact CEBS-Support@mail.nih.gov for additional assistance

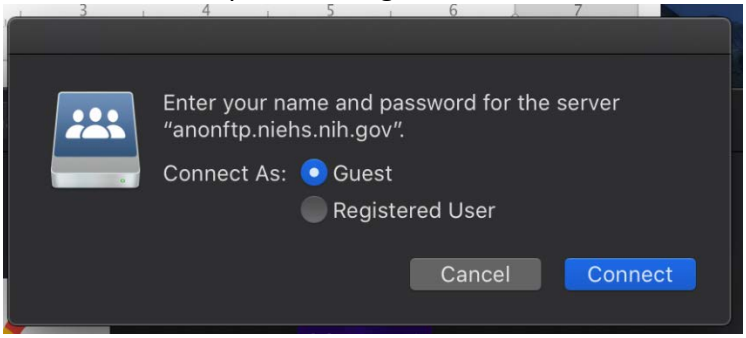
Appendix: Connecting to CEBS FTP site using Safari (Mac OS)

The following instructions only apply when connecting to the CEBS FTP site using the Safari browser and does not apply to other browsers in Mac OS.

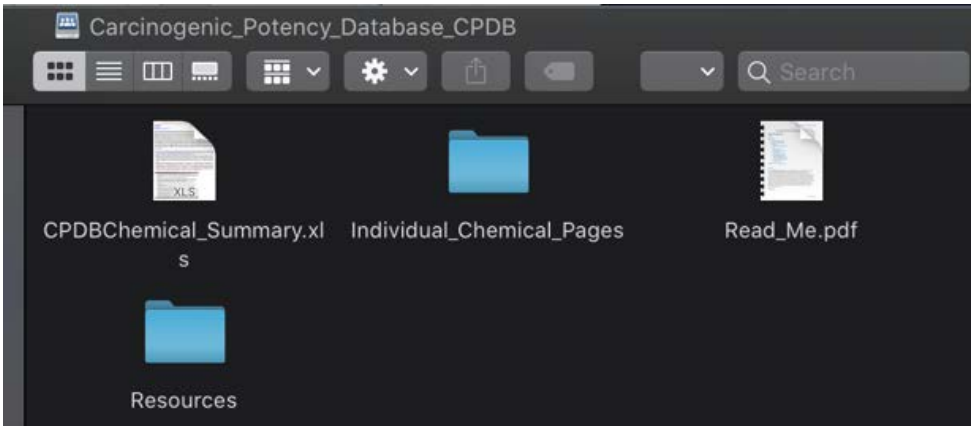
1. When selecting links from the list of data domains on the CEBS data domain page (https://cebs.niehs.nih.gov/ftp), Safari may prompt the user with a window "Do you want to allow this page to open in Finder?"



2. Please select “Allow”
3. Safari may then prompt the user with a window to “Enter your name and password for the server anonftp.niehs.nih.gov.”



4. Please select “Guest”, then select “Connect”. The directories described herein are publicly available and do not require a username or password
5. The Mac OS then opens a finder window and allows user to browse and obtain the desired data.



6. Please see Additional Instructions for additional downloading and viewing of data instructions