caArray 089 - Uploading Very Large Datasets

Problem: When uploading data from large microarrays, the size of your data archive may exceed the individual file size limit of 2 GB.

Topic: caArray Usage

Release: caArray 2.0 and above

Date entered: 10/17/2011

Solution

This article presents a workaround which allows you to break down your dataset into smaller, more manageable chunks that can be individually uploaded without violating the 2 GB limit.

Overview

Your experiment dataset consists of an IDF metadata file and its corresponding SDRF metadata file, which, in turn, is associated with one or more raw and derived array data files. In this tutorial, the array files we will use are in the Agilent TXT (raw) and TSV (derived) formats; the file formats for your data may differ.

Depending on the size of your array, the combined size of these files may exceed several gigabytes, even after they are compressed into the ZIP archive format required for uploading to caArray. Since the maximum size of a ZIP file that can be uploaded is 2 GB, any dataset which exceeds this limit must be broken down into smaller chunks, each of which contains a subset of the original data.

The general procedure for breaking down the dataset is as follows:

1. Divide the array data files into smaller batches, each of which will be no larger than 2 GB following ZIP compression.
2. Split the original SDRF file into multiple SDRF files, each corresponding to a single batch and referencing only the array data files from that batch.
3. Create multiple IDF files derived from the original IDF, with each one uniquely referencing one of the SDRF files created in the previous step.
4. Create a ZIP archive for each batch, containing a single IDF and its associated SDRF and raw and array data files.
5. Upload each ZIP archive individually, then validate and import the files from each.

Prerequisites

This tutorial assumes that you have past experience and basic familiarity with uploading data into caArray. Specifically, it assumes that you have already created an experiment for your data, uploaded the corresponding array design, and associated the experiment with that design. In case you lack a basic background on uploading caArray data, please refer to the official caArray User's Guide on the NCI wiki at https://wiki.nci.nih.gov/x/LBo9Ag.

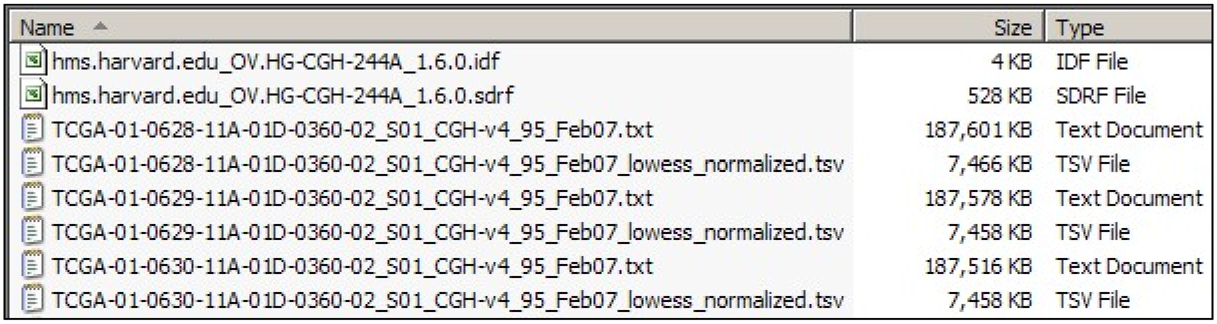
You must have all your experiment data readily accessible on your computer (i.e., not archived or compressed). The data should preferably be consolidated into a single location (i.e., a folder containing every single IDF, SDRF, raw and derived array data file from the experiment). You will also need an archive creation utility installed on your computer. In this tutorial, we will use WinZip (www.winzip.com), but any comparable utility with support for the ZIP format will do.

Reference Information

The experiment data used in this tutorial was not generated *de novo;* it came from an existing experiment whose data is publicly available on the official NCI instance of caArray at https://array.nci.nih.gov/caarray/home.action (Note that you can download this data without registering for an account on the site.) The experiment, entitled "TCGA Ovarian: Comparative Genome Hybridization Analysis Using the Agilent Human Genome CGH 244A Platform", was conducted at Harvard Medical School in Boston, MA. It can be accessed via the URL https://array.nci.nih.gov/caarray/project/EXP-498 or by searching for the experiment ID 'EXP-498' on the NCI caArray instance. The array design used was TCGA-Agilent_HG-CGH-244A; the array design files can be downloaded from the experiment in ADF format, as can all the experiment data, including the IDF and SDRF metadata files, the Agilent TXT raw array data files, and the TSV derived array data files.

Getting Started -- Dividing the Array Data Into Batches

The screenshot below shows a portion of the dataset from our sample experiment, including the IDF and SDRF files, as well as some TXT and TSV files.

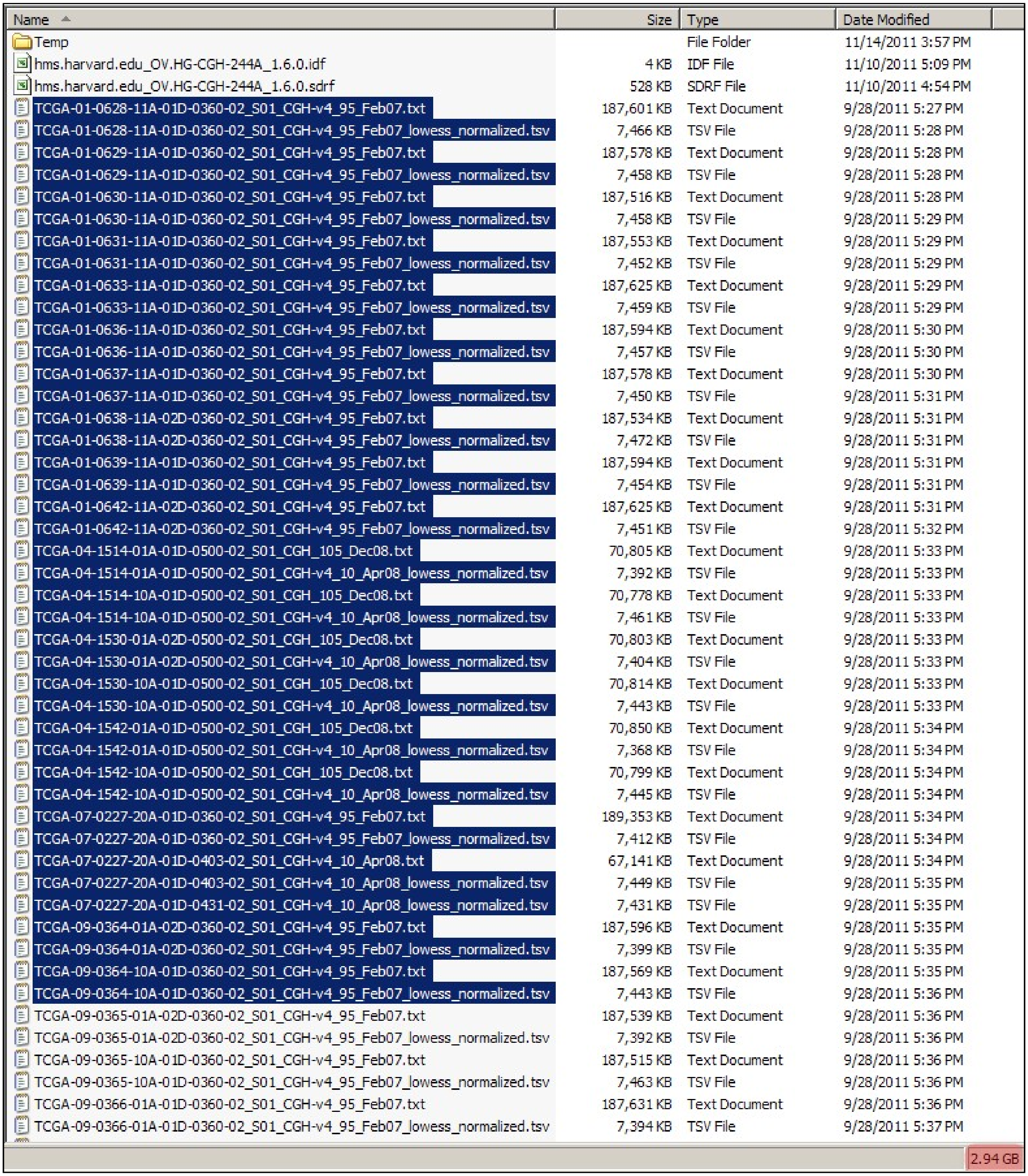


Name	Size	Type
hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	4 KB	IDF File
hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	528 KB	SDRF File
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,601 KB	Text Document
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,466 KB	TSV File
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,578 KB	Text Document
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,458 KB	TSV File
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,516 KB	Text Document
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,458 KB	TSV File

This dataset comprises IDF and SDRF metadata files, as well as the TXT raw array and TSV derived array data files they reference.

The total combined size of all the files in this dataset is a whopping 26.8 GB, which is way too large to be uploaded to caArray at once, even when archived into a single file. Our first step, then, is to break down the dataset into smaller batches, each of which will be no larger than 2 GB following ZIP compression. Since the average ZIP compression ratio of array data is about 2.5:1, we may safely assume that any batch smaller than 5 GB before compression will come out to less than 2 GB after compression.

Before creating the batches, first create a subfolder named 'Batches' in your experiment folder, then create individual subfolders ('batch1', 'batch2', etc.) within that folder for each batch. Now, select multiple TXT and TSV files in your file manager (Windows Explorer in this tutorial), taking care to keep the size of the selection below 5 GB, as shown below:



Name	Size	Type	Date Modified
Temp		File Folder	11/14/2011 3:57 PM
hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	4 KB	IDF File	11/10/2011 5:09 PM
hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	528 KB	SDRF File	11/10/2011 4:54 PM
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,601 KB	Text Document	9/28/2011 5:27 PM
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,466 KB	TSV File	9/28/2011 5:28 PM
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,578 KB	Text Document	9/28/2011 5:28 PM
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,458 KB	TSV File	9/28/2011 5:28 PM
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,516 KB	Text Document	9/28/2011 5:28 PM
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,458 KB	TSV File	9/28/2011 5:29 PM
TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,553 KB	Text Document	9/28/2011 5:29 PM
TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,452 KB	TSV File	9/28/2011 5:29 PM
TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,625 KB	Text Document	9/28/2011 5:29 PM
TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,459 KB	TSV File	9/28/2011 5:29 PM
TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,594 KB	Text Document	9/28/2011 5:30 PM
TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,457 KB	TSV File	9/28/2011 5:30 PM
TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,578 KB	Text Document	9/28/2011 5:30 PM
TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,450 KB	TSV File	9/28/2011 5:31 PM
TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	187,534 KB	Text Document	9/28/2011 5:31 PM
TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,472 KB	TSV File	9/28/2011 5:31 PM
TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,594 KB	Text Document	9/28/2011 5:31 PM
TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,454 KB	TSV File	9/28/2011 5:31 PM
TCGA-01-0642-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	187,625 KB	Text Document	9/28/2011 5:31 PM
TCGA-01-0642-11A-02D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,451 KB	TSV File	9/28/2011 5:32 PM
TCGA-04-1514-01A-01D-0500-02_S01_CGH_105_Dec08.txt	70,805 KB	Text Document	9/28/2011 5:33 PM
TCGA-04-1514-01A-01D-0500-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,392 KB	TSV File	9/28/2011 5:33 PM
TCGA-04-1514-10A-01D-0500-02_S01_CGH_105_Dec08.txt	70,778 KB	Text Document	9/28/2011 5:33 PM
TCGA-04-1514-10A-01D-0500-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,461 KB	TSV File	9/28/2011 5:33 PM
TCGA-04-1530-01A-02D-0500-02_S01_CGH_105_Dec08.txt	70,803 KB	Text Document	9/28/2011 5:33 PM
TCGA-04-1530-01A-02D-0500-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,404 KB	TSV File	9/28/2011 5:33 PM
TCGA-04-1530-10A-01D-0500-02_S01_CGH_105_Dec08.txt	70,814 KB	Text Document	9/28/2011 5:33 PM
TCGA-04-1530-10A-01D-0500-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,443 KB	TSV File	9/28/2011 5:33 PM
TCGA-04-1542-01A-01D-0500-02_S01_CGH_105_Dec08.txt	70,850 KB	Text Document	9/28/2011 5:34 PM
TCGA-04-1542-01A-01D-0500-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,368 KB	TSV File	9/28/2011 5:34 PM
TCGA-04-1542-10A-01D-0500-02_S01_CGH_105_Dec08.txt	70,799 KB	Text Document	9/28/2011 5:34 PM
TCGA-04-1542-10A-01D-0500-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,445 KB	TSV File	9/28/2011 5:34 PM
TCGA-07-0227-20A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	189,353 KB	Text Document	9/28/2011 5:34 PM
TCGA-07-0227-20A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,412 KB	TSV File	9/28/2011 5:34 PM
TCGA-07-0227-20A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	67,141 KB	Text Document	9/28/2011 5:34 PM
TCGA-07-0227-20A-01D-0403-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,449 KB	TSV File	9/28/2011 5:35 PM
TCGA-07-0227-20A-01D-0431-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	7,431 KB	TSV File	9/28/2011 5:35 PM
TCGA-09-0364-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	187,596 KB	Text Document	9/28/2011 5:35 PM
TCGA-09-0364-01A-02D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,399 KB	TSV File	9/28/2011 5:35 PM
TCGA-09-0364-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,569 KB	Text Document	9/28/2011 5:35 PM
TCGA-09-0364-10A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,443 KB	TSV File	9/28/2011 5:36 PM
TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	187,539 KB	Text Document	9/28/2011 5:36 PM
TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,392 KB	TSV File	9/28/2011 5:36 PM
TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,515 KB	Text Document	9/28/2011 5:36 PM
TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,463 KB	TSV File	9/28/2011 5:36 PM
TCGA-09-0366-01A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	187,631 KB	Text Document	9/28/2011 5:36 PM
TCGA-09-0366-01A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	7,394 KB	TSV File	9/28/2011 5:37 PM

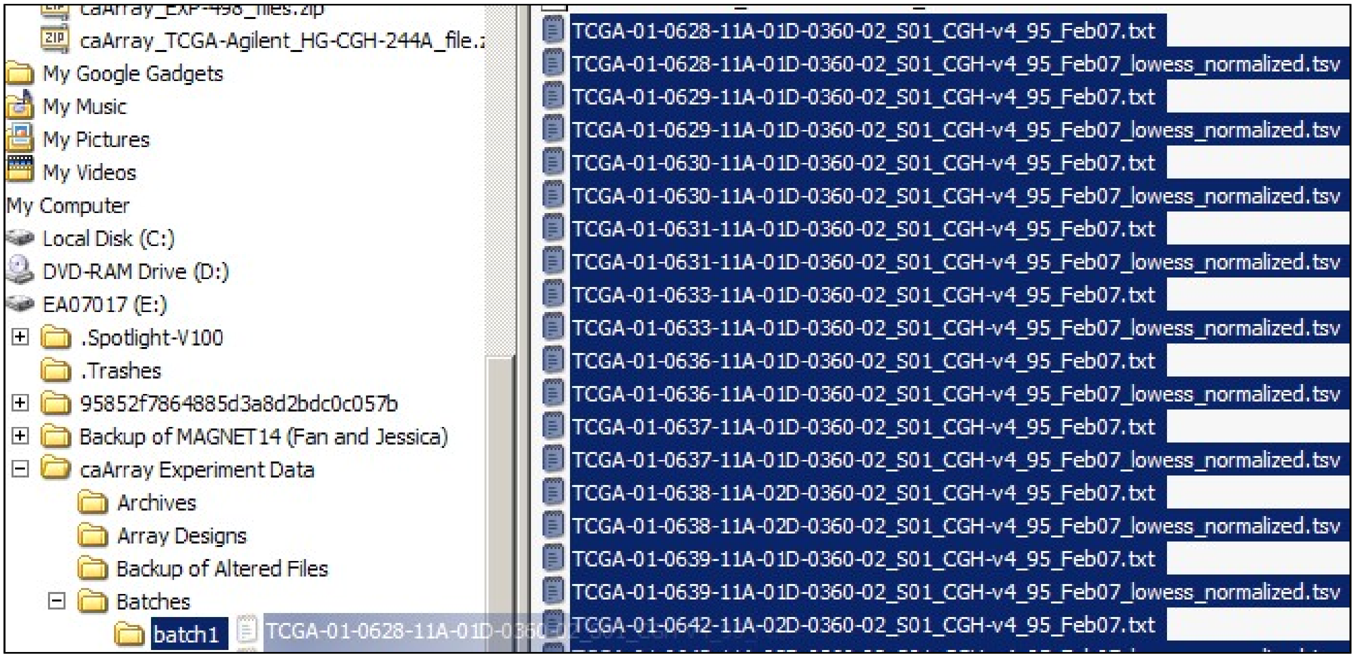
2.94 GB

When selecting a subset of your TXT and TSV files in your file manager, make sure the combined size of the selected files is below 5 GB, as anything larger may compress to greater than the 2 GB upload limit caArray imposes for a single ZIP archive.

Note

Even though caArray allows archives as large as 2 GB to be uploaded, in this tutorial we will keep the size of archives to approximately 1 GB each to facilitate rapid uploads on slow network connections.

You can now move the file selection to the 'batch1' subfolder we created earlier, as shown below:



```
caArray_EXP-498_files.zip
caArray_TCGA-Agilent_HG-CGH-244A_file.z
My Google Gadgets
My Music
My Pictures
My Videos
My Computer
Local Disk (C:)
DVD-RAM Drive (D:)
EA07017 (E:)
.Spotlight-V100
.Trashes
95852f7864885d3a8d2bdc0c057b
Backup of MAGNET14 (Fan and Jessica)
caArray Experiment Data
Archives
Array Designs
Backup of Altered Files
Batches
batch1
```

```
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv
TCGA-01-0642-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt
```

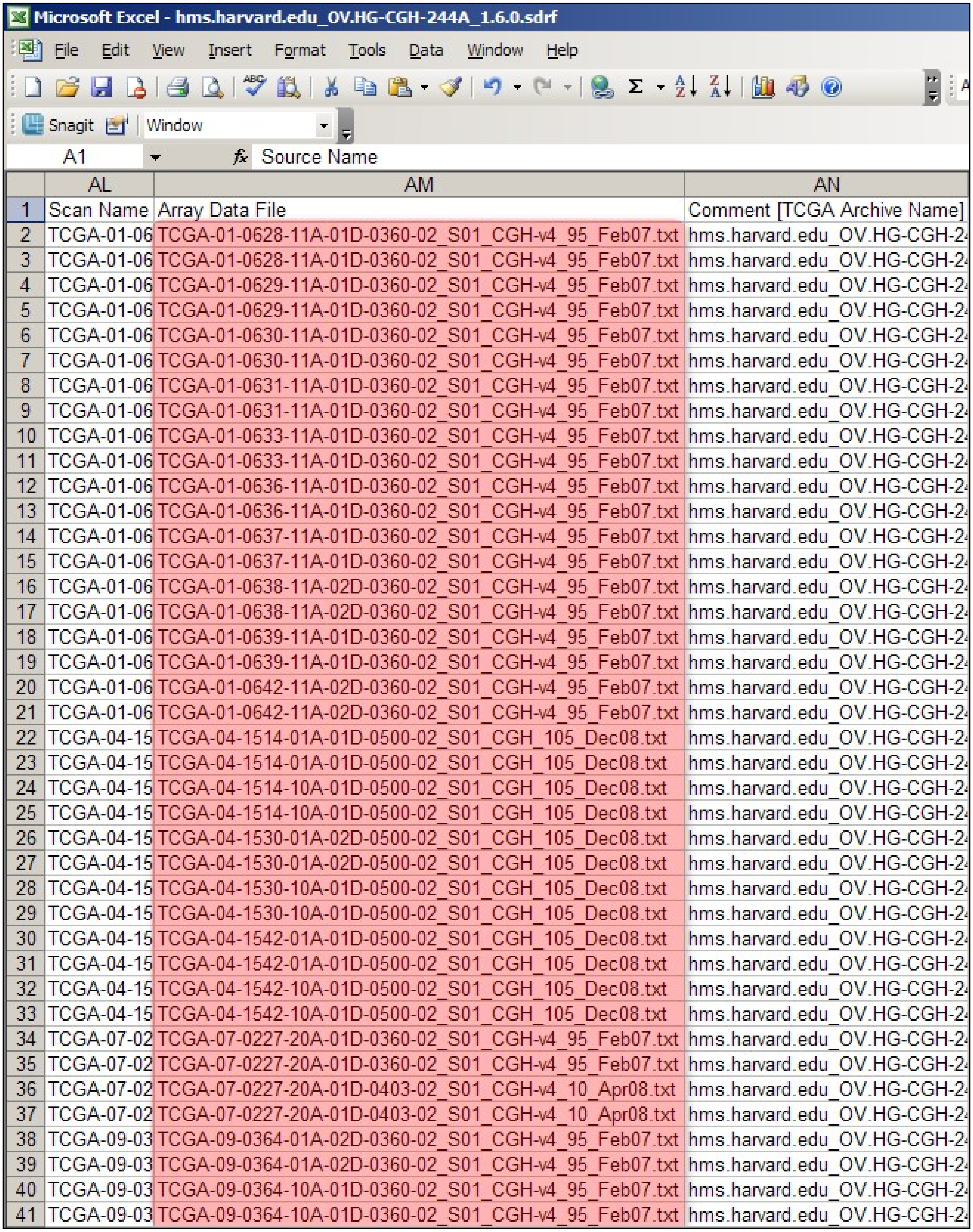
Move the selected files to the subfolder you created for this batch.

You can repeat this procedure to create the remaining batches, as summarized below, until every single file in the dataset has been accounted for:

1. Create a separate subfolder for each new batch
2. Select multiple data files in your file manager, taking care to keep selection size below 5 GB uncompressed (2 GB compressed)
3. Move selected files to respective batch folder

Splitting The Original SDRF File

Now that we've created batches of our array data files, our next step is to split the original SDRF file into multiple SDRFs, each corresponding to a single batch and referencing only the array data files from that batch. To do so, first open the original SDRF file in Microsoft Excel or another tab-limited data viewer, as shown below:



	AL	AM	AN
1	Scan Name	Array Data File	Comment [TCGA Archive Name]
2	TCGA-01-06	TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
3	TCGA-01-06	TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
4	TCGA-01-06	TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
5	TCGA-01-06	TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
6	TCGA-01-06	TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
7	TCGA-01-06	TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
8	TCGA-01-06	TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
9	TCGA-01-06	TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
10	TCGA-01-06	TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
11	TCGA-01-06	TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
12	TCGA-01-06	TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
13	TCGA-01-06	TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
14	TCGA-01-06	TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
15	TCGA-01-06	TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
16	TCGA-01-06	TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
17	TCGA-01-06	TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
18	TCGA-01-06	TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
19	TCGA-01-06	TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
20	TCGA-01-06	TCGA-01-0642-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
21	TCGA-01-06	TCGA-01-0642-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
22	TCGA-04-15	TCGA-04-1514-01A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
23	TCGA-04-15	TCGA-04-1514-01A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
24	TCGA-04-15	TCGA-04-1514-10A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
25	TCGA-04-15	TCGA-04-1514-10A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
26	TCGA-04-15	TCGA-04-1530-01A-02D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
27	TCGA-04-15	TCGA-04-1530-01A-02D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
28	TCGA-04-15	TCGA-04-1530-10A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
29	TCGA-04-15	TCGA-04-1530-10A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
30	TCGA-04-15	TCGA-04-1542-01A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
31	TCGA-04-15	TCGA-04-1542-01A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
32	TCGA-04-15	TCGA-04-1542-10A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
33	TCGA-04-15	TCGA-04-1542-10A-01D-0500-02_S01_CGH_105_Dec08.txt	hms.harvard.edu_OV.HG-CGH-2
34	TCGA-07-02	TCGA-07-0227-20A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
35	TCGA-07-02	TCGA-07-0227-20A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
36	TCGA-07-02	TCGA-07-0227-20A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	hms.harvard.edu_OV.HG-CGH-2
37	TCGA-07-02	TCGA-07-0227-20A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	hms.harvard.edu_OV.HG-CGH-2
38	TCGA-09-03	TCGA-09-0364-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
39	TCGA-09-03	TCGA-09-0364-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
40	TCGA-09-03	TCGA-09-0364-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2
41	TCGA-09-03	TCGA-09-0364-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	hms.harvard.edu_OV.HG-CGH-2

The SDRF file from your experiment lists all the associated raw array data files under the column headed 'Array Data File'.

As you can see, the column headed 'Array Data File' lists the filenames of all the raw array data files from the experiment. The first 40 rows correspond to all the data files from the first batch we created in the previous section, *Getting Started*. We can generate a unique SDRF file for this batch by deleting all the other rows from the file -- except, of course, for the top header row -- and saving the modified file as a new SDRF with a different filename from the original. *(The convention used in this tutorial is to prefix the original SDRF filename with a number representing the batch, followed by a period. For example, if the original SDRF filename is 'hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf', then the filename of the first, or 'zeroeth' batch, would be '0. hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf'.)*

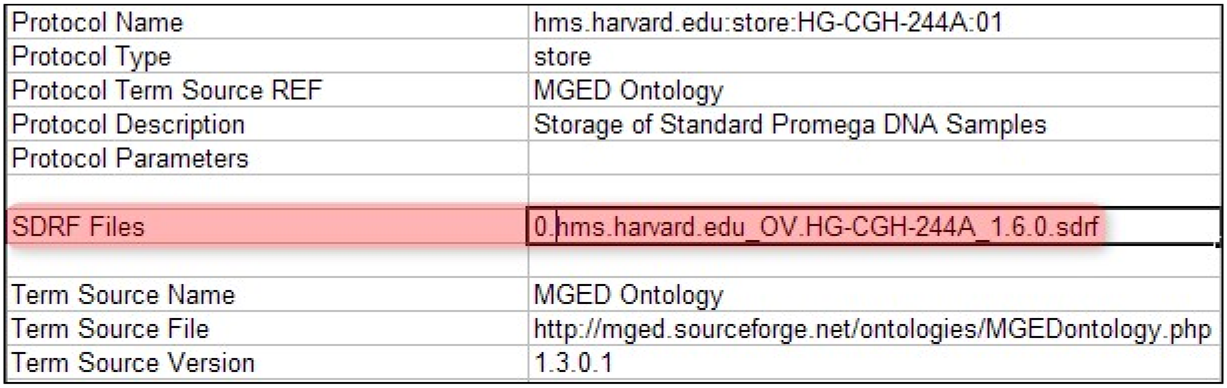
Once you've generated a new SDRF file, copy it over to its respective batch's folder. For example, '0.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf' would be copied to the 'batch1' folder containing all the array data files from the first batch we created.

You can generalize this procedure to the entire original SDRF file, and thus all the batches from your dataset, by following these steps:

1. Open the original SDRF file and locate the rows referencing the array files for the next batch
2. Delete all other rows except for the top header row
3. Save the modified SDRF as a new file with a filename unique to its respective batch
4. Copy the newly generated SDRF to its respective batch's folder

Creating a Unique IDF File For Each Batch

Once you've generated a unique SDRF file for each batch, you must also generate a unique IDF file which references that SDRF file. You can do so simply by opening the original IDF file and editing the field 'SDRF Files' with the filename of the SDRF you wish to reference, as shown below:



Protocol Name	hms.harvard.edu:store:HG-CGH-244A:01
Protocol Type	store
Protocol Term Source REF	MGED Ontology
Protocol Description	Storage of Standard Promega DNA Samples
Protocol Parameters	
SDRF Files	0.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
Term Source Name	MGED Ontology
Term Source File	http://mged.sourceforge.net/ontologies/MGEDontology.php
Term Source Version	1.3.0.1

Edit the field 'SDRF Files' field in your IDF file to reflect the file name of the new SDRF file you generated previously.

In this example, the originally referenced SDRF filename 'hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf' has been changed to '0.hms.harvard.edu_OV. HG-CGH-244A_1.6.0.sdrf', which is the SDRF for the first batch we created.

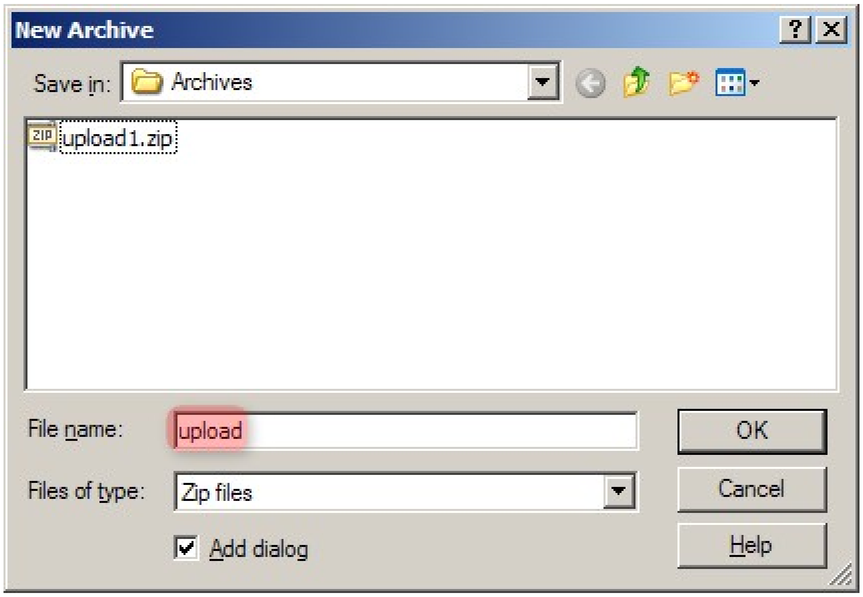
As with the SDRF files we modified in the previous section, be sure to save the modified IDF file as a new IDF with the same filename as its referenced SDRF, but with the 'IDF' extension instead of 'SDRF'. For example, the file which references '0.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf' would be named '0.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf'. Finally, copy this IDF file over to its respective batch's folder containing the referenced SDRF file and all the associated array data files.

You can repeat this procedure for all your batches. In summary:

1. Open the original IDF file and locate the 'SDRF Files' field.
2. Edit this field to reflect the file name of the SDRF file you wish to reference.
3. Save the modified IDF file with a unique filename that is parallel to the referenced SDRF's filename.
4. Copy the newly generated IDF to its respective batch's folder

Creating the Archives

Now that we've divided our dataset into batches and generated the corresponding IDF and SDRF files for each, our next step is to create a ZIP archive of each batch. Launch WinZip, click the 'New' toolbar button, and enter a name for your archive in the 'New Archive' dialog. We'll call ours 'upload.zip', as shown below.



Save in: Archives

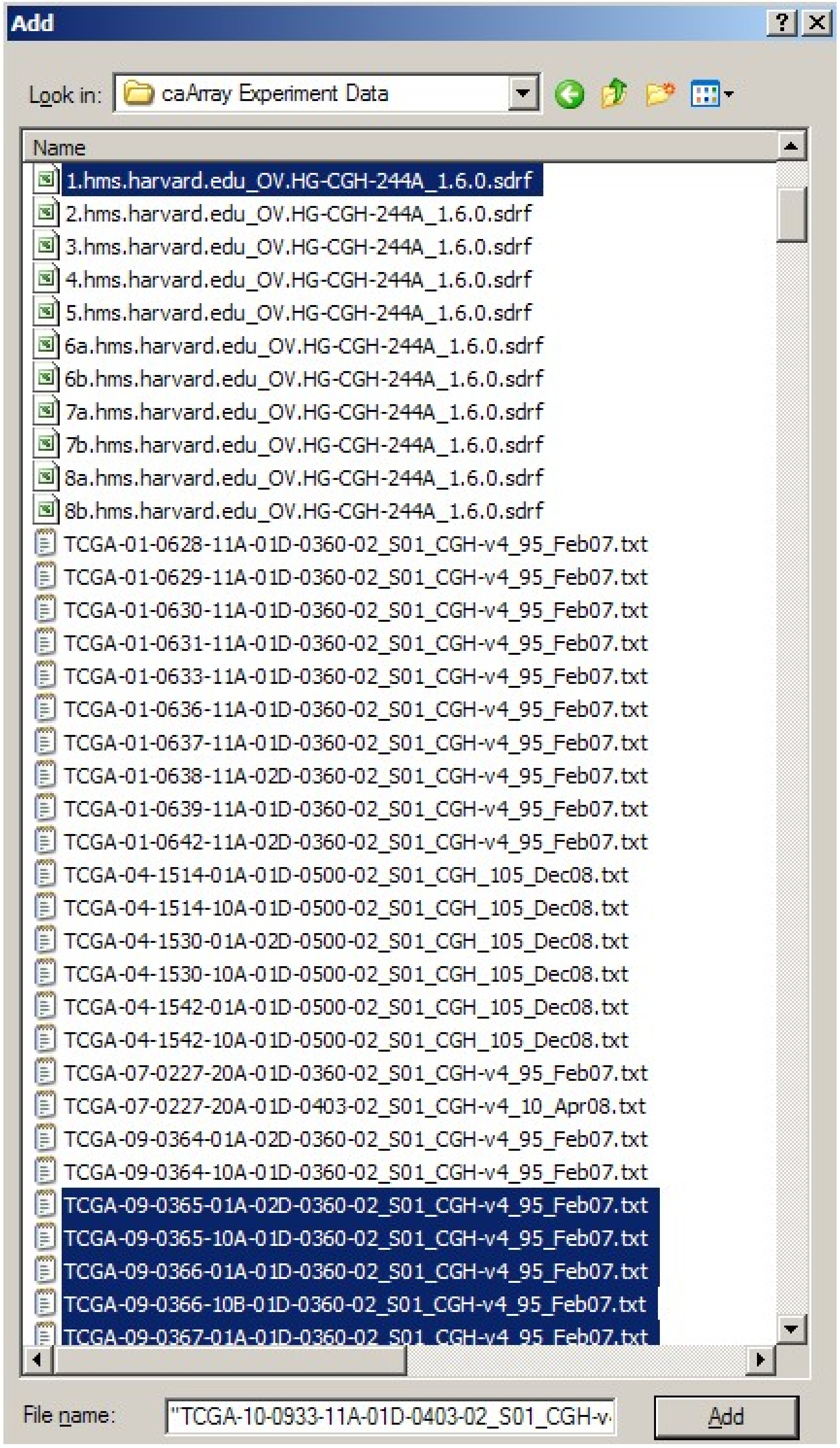
File name: upload

Files of type: Zip files

☑ Add dialog

In WinZip's 'New Archive' dialog, specify a filename for the data archive to be created ('upload.zip in our example').

Once we've created the archive, we can now add files to it. We can refer to our previous notes of all the filenames associated with our IDF file. In our example, the archive will consist of a total of 42 files: one IDF, one SDRF, 20 TXT, and 20 TSV files. We can select these files in the 'Add' dialog as shown below, then click the 'Add' button at the bottom to begin creating the archive. *(Hint: Hold down the CTRL key to select multiple files.)*



Look in: caArray Experiment Data

Name

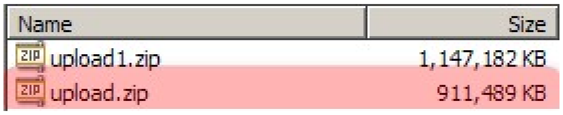
1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
2.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
3.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
4.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
5.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
6a.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
6b.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
7a.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
7b.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
8a.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
8b.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf
TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0631-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0633-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0636-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0637-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0638-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0639-11A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-01-0642-11A-02D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-04-1514-01A-01D-0500-02_S01_CGH_105_Dec08.txt
TCGA-04-1514-10A-01D-0500-02_S01_CGH_105_Dec08.txt
TCGA-04-1530-01A-02D-0500-02_S01_CGH_105_Dec08.txt
TCGA-04-1530-10A-01D-0500-02_S01_CGH_105_Dec08.txt
TCGA-04-1542-01A-01D-0500-02_S01_CGH_105_Dec08.txt
TCGA-04-1542-10A-01D-0500-02_S01_CGH_105_Dec08.txt
TCGA-07-0227-20A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-07-0227-20A-01D-0403-02_S01_CGH-v4_10_Apr08.txt
TCGA-09-0364-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-09-0364-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-09-0366-01A-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-09-0366-10B-01D-0360-02_S01_CGH-v4_95_Feb07.txt
TCGA-09-0367-01A-01D-0360-02_S01_CGH-v4_95_Feb07.txt

File name: "TCGA-10-0933-11A-01D-0403-02_S01_CGH-v

In WinZip's 'Add' dialog, select all the related IDF, SDRF, raw data, and derived data files (a total of 42 files in our example), then click the 'Add' button below to begin creating the archive.

Warning

After you've created the archive, ensure that the resulting file size is less than 2 GB. If it isn't, you will either have to re-create the archive with a higher compression ratio, or subdivide the batch into smaller batches. In our example, the size of the 'upload.zip' archive came out to approximately 900 MB, as shown below, so the file is ready to upload as is.

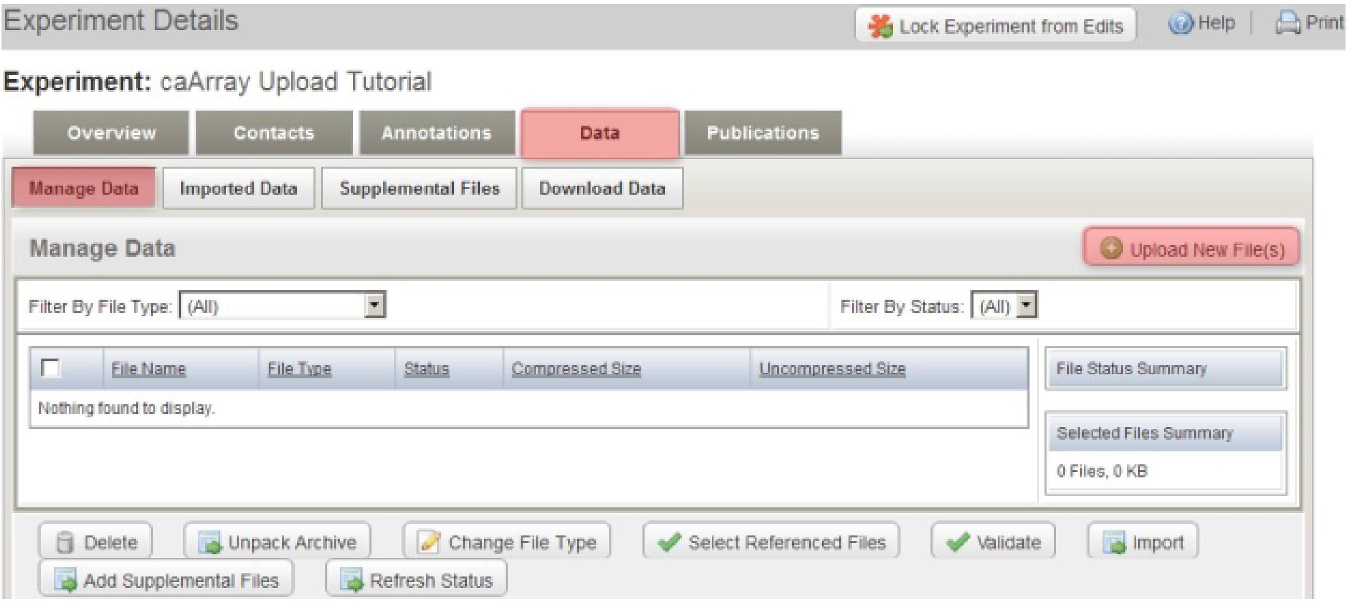


Name	Size
upload1.zip	1,147,182 KB
upload.zip	911,489 KB

In our example, the 'upload.zip' data archive we created is approximately 900 MB in size, which is below the 2 GB upload limit. If your data archive turns out to be larger than 2 GB, you will not be able to upload it until you re-create it with a higher compression ratio.

Uploading the Archive

To upload the archive, first log in to caArray and navigate to the experiment you will be upload your data into, then select the 'Data' tab, followed by the 'Manage Data' tab beneath it. Now click on the 'Upload New Files' button as shown below.



Experiment Details

Experiment: caArray Upload Tutorial

Manage Data

Filter By File Type: (All)

Filter By Status: (All)

☐	File Name	File Type	Status	Compressed Size	Uncompressed Size
Nothing found to display.					

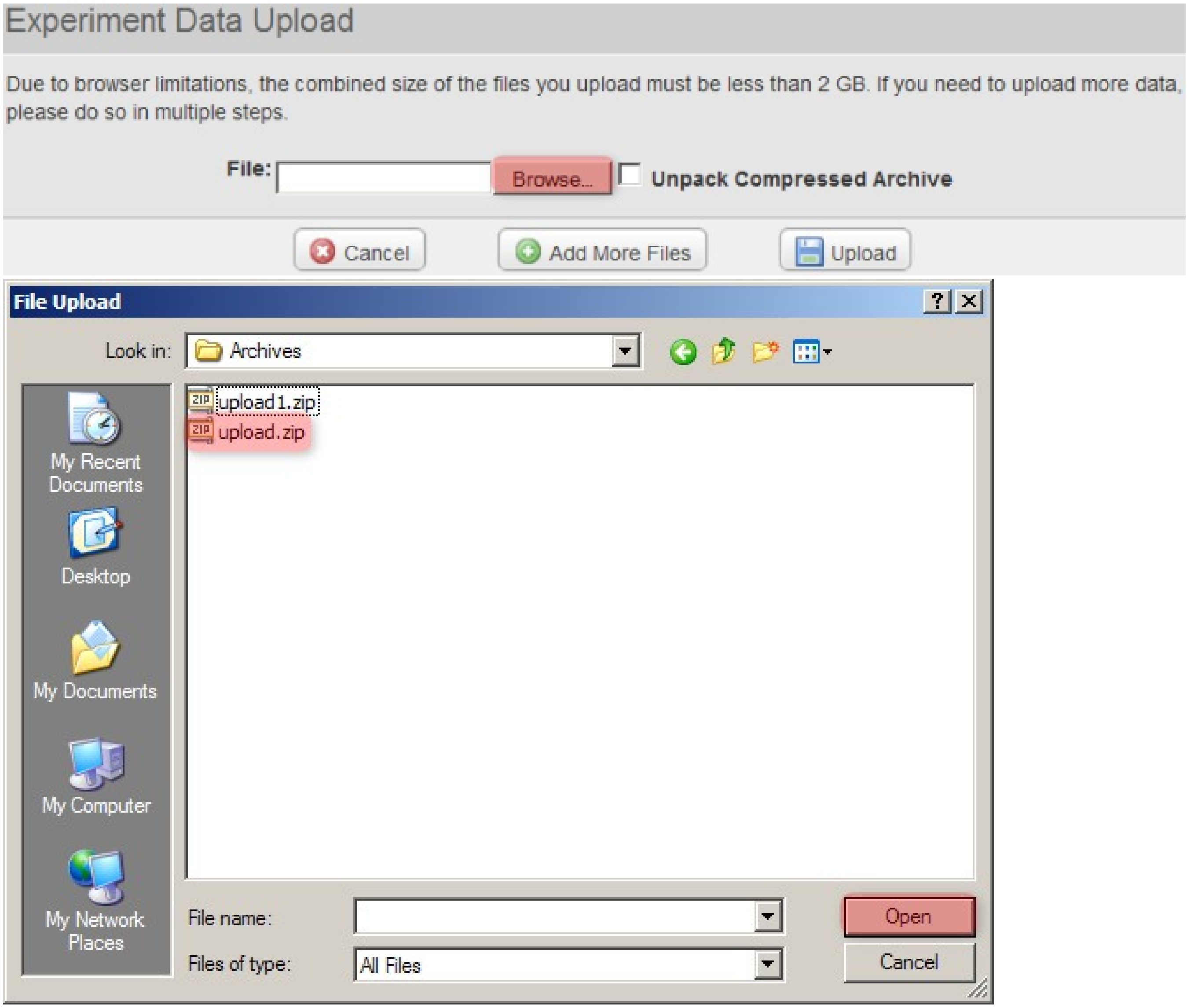
File Status Summary

Selected Files Summary

0 Files, 0 KB

Click the 'Upload New Files' button under the 'Manage Data' tab to specify the location of your data archive.

A new pop-up window entitled 'Experiment Data Upload' will appear in your Web browser, prompting you to upload files. Click on the 'Browse' button, then select the 'upload.zip' archive we created previously from the Open dialog as shown below.

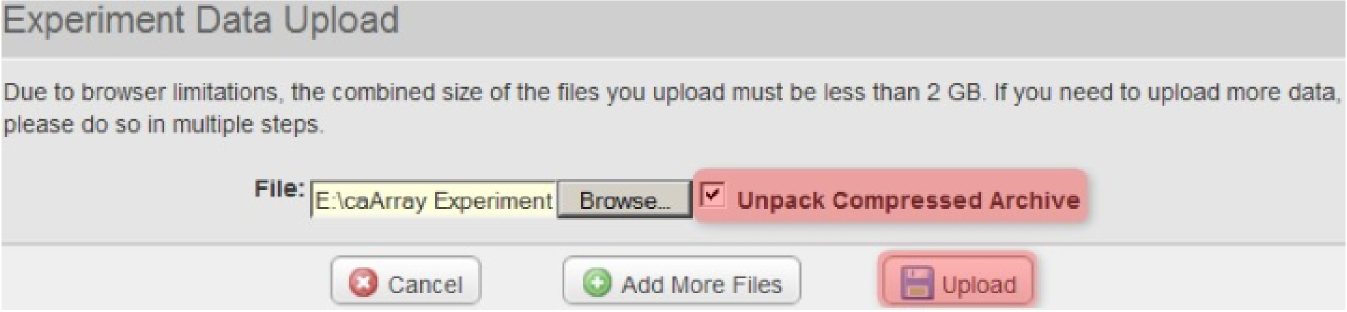


Experiment Data Upload

Due to browser limitations, the combined size of the files you upload must be less than 2 GB. If you need to upload more data, please do so in multiple steps.

In the 'Experiment Data Upload' pop-up window, click the 'Browse' button, then in the 'File Upload' dialog, navigate to the ZIP data archive we created previously and click on the 'Open' button.

Back in the 'Experiment Data Upload' window, make sure that the box labeled 'Unpack Compressed Archive' is checked, then click on the 'Upload' button to begin uploading the file.



Experiment Data Upload

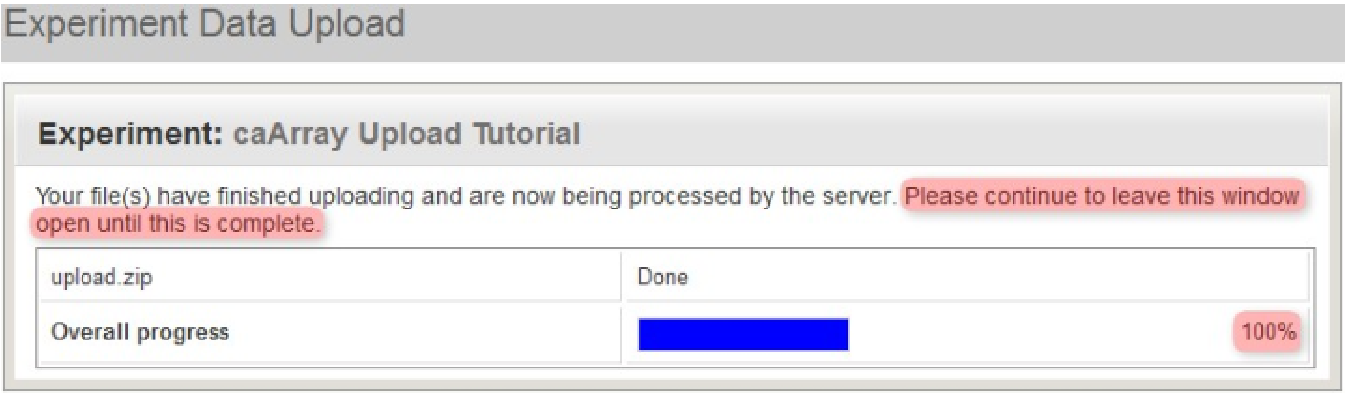
Due to browser limitations, the combined size of the files you upload must be less than 2 GB. If you need to upload more data, please do so in multiple steps.

File: E:\caArray Experiment

☑ Unpack Compressed Archive

Back in the 'Experiment Data Upload' window, make sure that the box labeled 'Unpack Compressed Archive' is checked, then click on the 'Upload' button to begin uploading the file.

Depending on the size of the archive, the performance of your caArray server, and your network bandwidth, it may take anywhere from five to 30 minutes -- and possibly longer -- for the archive to upload. Remember to keep the upload window open throughout the entire upload process, even after the blue progress bar has reached 100%. (For reference, on a caArray server running a quad-core 2.33 Ghz Intel(R) Xeon(R) 5148 CPU with 16 GB of memory, the total time required to extract and process a 1.1 GB upload after the progress bar had reached 100% was about 13 minutes and 30 seconds.)



Experiment Data Upload

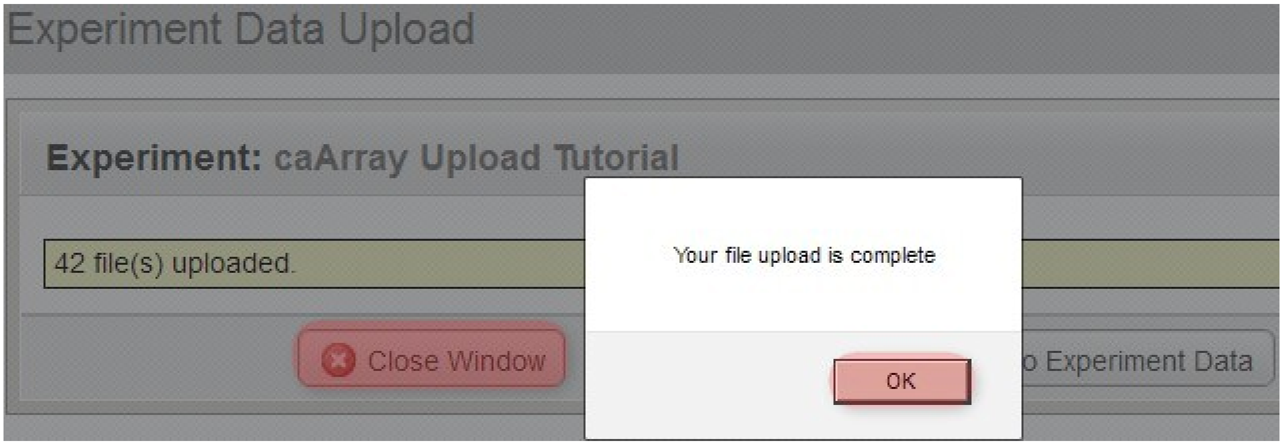
Experiment: caArray Upload Tutorial

Your file(s) have finished uploading and are now being processed by the server. Please continue to leave this window open until this is complete.

upload.zip	Done
Overall progress	100%

Even when the blue upload progress bar reaches 100%, do not close the 'Experiment Data Upload' window. You will be notified when the upload is complete.

You'll know when the upload is complete when you see a new window overlaid over the upload window with the message 'Your file upload is complete', as shown below. Click the 'OK' button below this message, then click on the 'Close Window' button behind it to return to the main experiment window.



Experiment Data Upload

Experiment: caArray Upload Tutorial

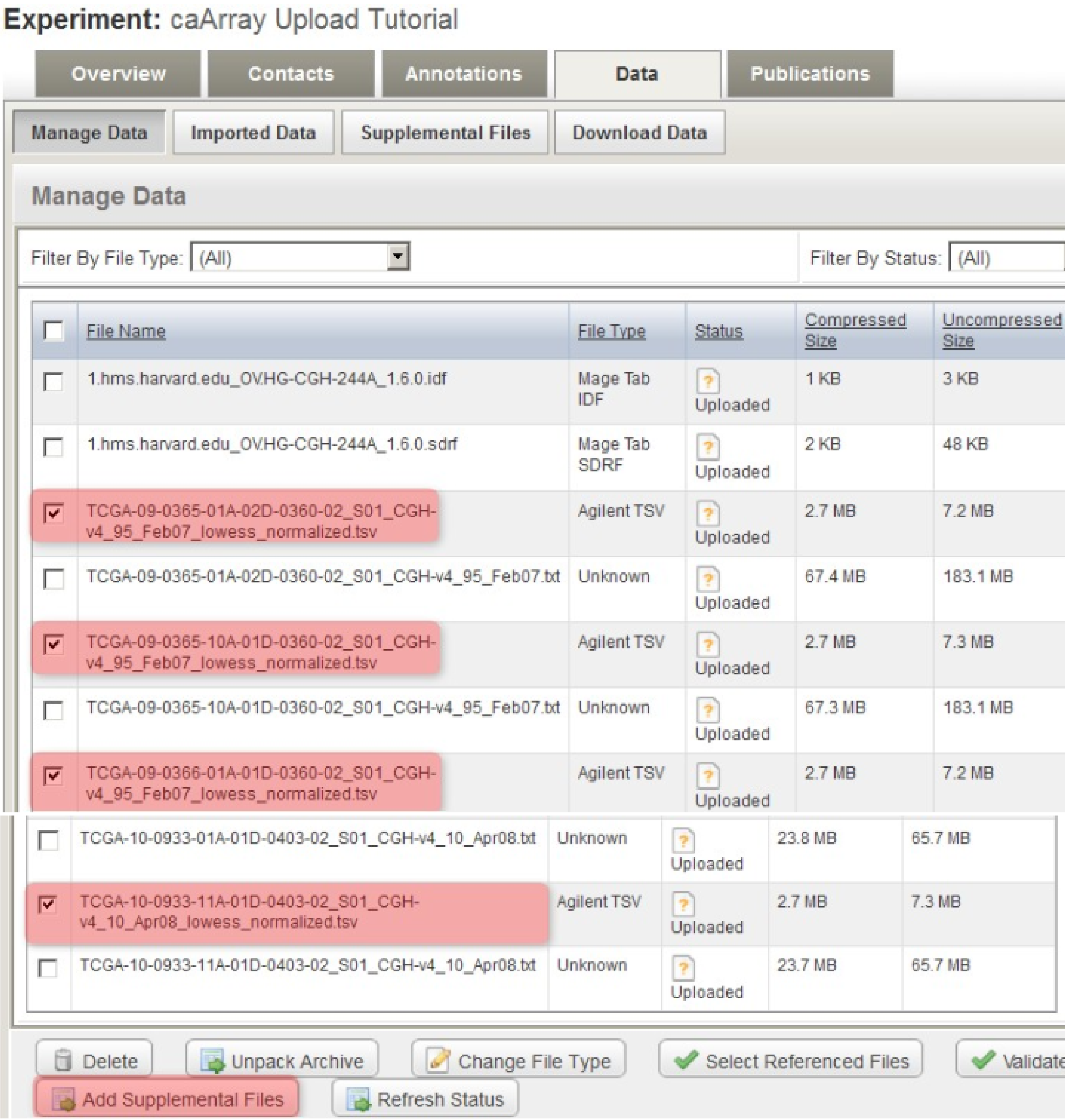
42 file(s) uploaded.

Your file upload is complete

You'll know when the upload is complete when you see a new window overlaid over the upload window with the message, 'Your file upload is complete'.

Validating the Archive

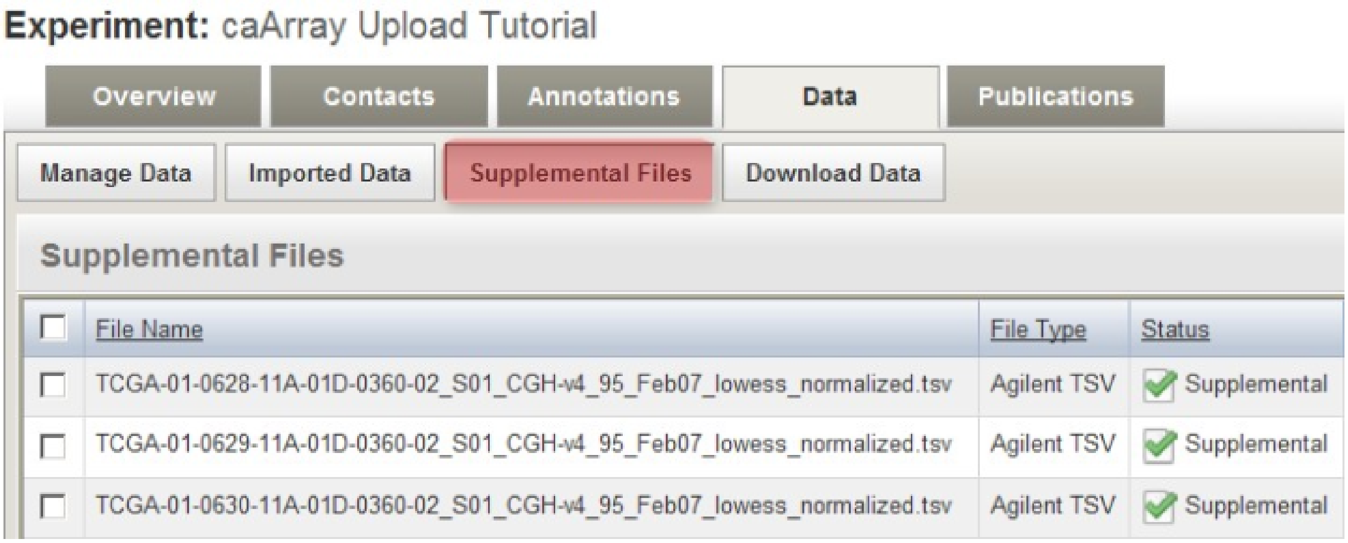
Back in the main experiment window, the contents of the archive we just uploaded are now listed under the 'Manage Data' tab. The TSV matrix files are considered supplemental, so we will move them to the 'Supplemental Files' tab by first using the 'Filter By File Type' drop-down to show only TSV files, then checking off all the TSV files in the list, and finally clicking on the 'Add Supplemental Files' button below.



□	File Name	File Type	Status	Compressed Size	Uncompressed Size
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF	? Uploaded	1 KB	3 KB
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF	? Uploaded	2 KB	48 KB
☑	TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	Agilent TSV	? Uploaded	2.7 MB	7.2 MB
□	TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	Unknown	? Uploaded	67.4 MB	183.1 MB
☑	TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	Agilent TSV	? Uploaded	2.7 MB	7.3 MB
□	TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	Unknown	? Uploaded	67.3 MB	183.1 MB
☑	TCGA-09-0366-01A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	Agilent TSV	? Uploaded	2.7 MB	7.2 MB
□	TCGA-10-0933-01A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Unknown	? Uploaded	23.8 MB	65.7 MB
☑	TCGA-10-0933-11A-01D-0403-02_S01_CGH-v4_10_Apr08_lowess_normalized.tsv	Agilent TSV	? Uploaded	2.7 MB	7.3 MB
□	TCGA-10-0933-11A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Unknown	? Uploaded	23.7 MB	65.7 MB

You can mark the derived array data files as supplemental by checking them off under the 'Manage Data' tab, then clicking the 'Add Supplemental Files' button.

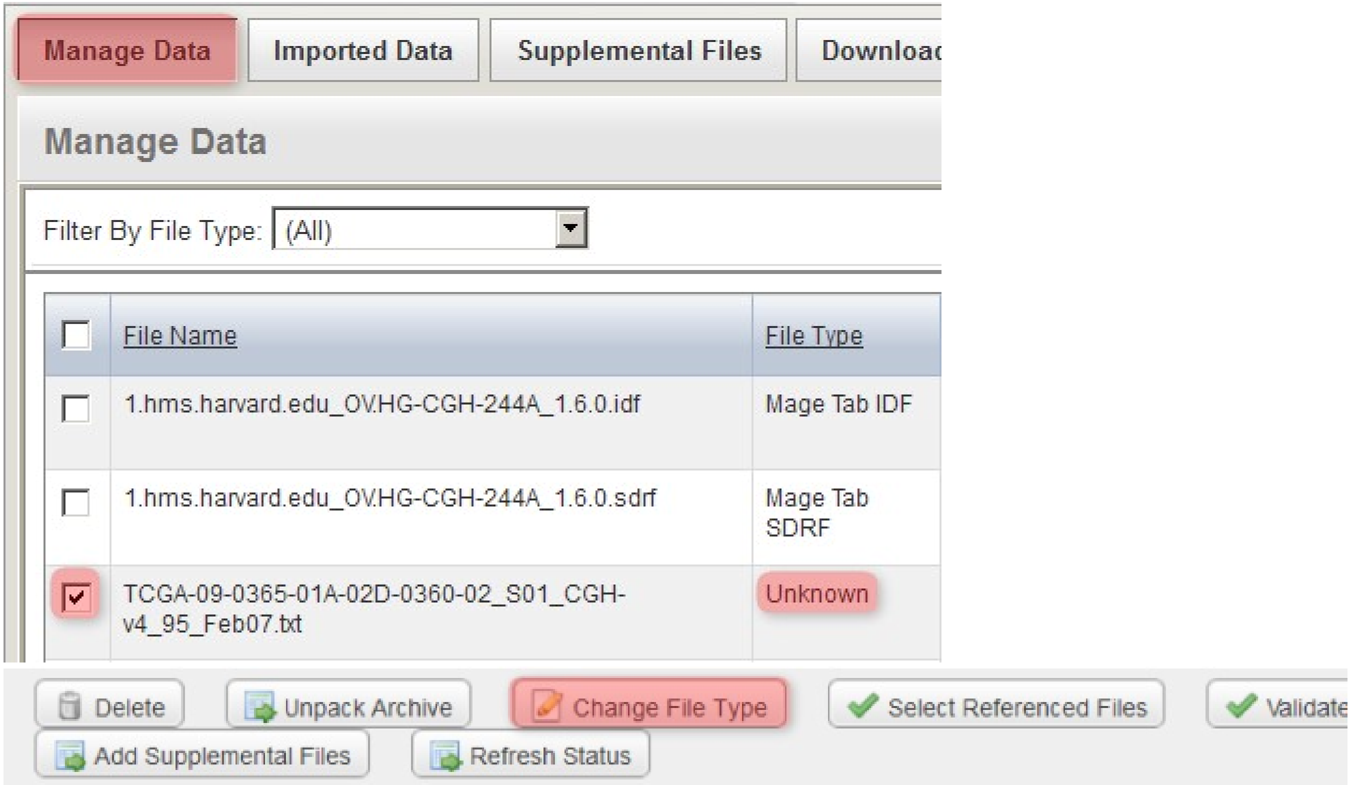
These TSV files now appear under the 'Supplemental Files' tab, alongside other TSV files from a previous upload to the same experiment.



□	File Name	File Type	Status
□	TCGA-01-0628-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	Agilent TSV	✓ Supplemental
□	TCGA-01-0629-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	Agilent TSV	✓ Supplemental
□	TCGA-01-0630-11A-01D-0360-02_S01_CGH-v4_95_Feb07_lowess_normalized.tsv	Agilent TSV	✓ Supplemental

The derived array data files we checked off under the 'Manage Data' tab now appear under the 'Supplemental Files' tab, alongside other such files from a previous upload to the same experiment.

Back on the 'Manage Data' tab, the remaining files from our upload are one IDF, one SDRF, and 20 TXTs (only the first three of these files is shown below due to space constraints). Note that the status of the TXT file from the screenshot (and of all other TXT files in the list) shows as 'Unknown', which means that caArray did not automatically recognize the file type in this particular case. As a result, we will have to manually specify the file type ourselves by first using the 'Filter By File Type' drop-down to show only TXT files, then checking off all the TXT files in the list, and finally clicking the 'Change File Type' button below.



Filter By File Type: (All)

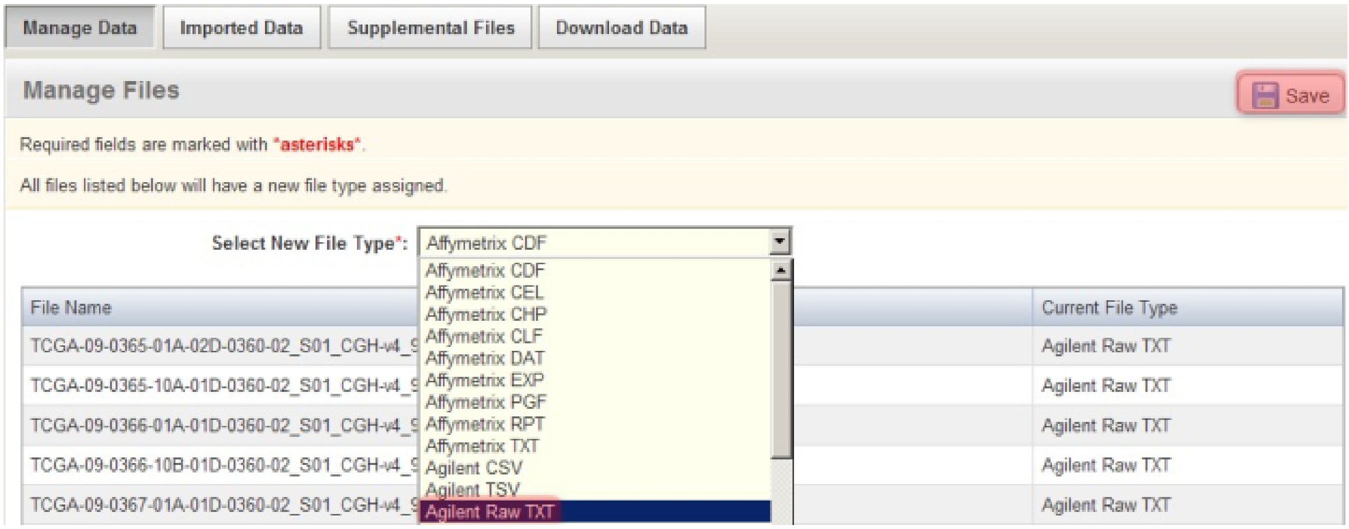
□	File Name	File Type
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF
☑	TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	Unknown

Since caArray didn't automatically recognize the format of the array data files we uploaded, we must manually specify the format ourselves by selecting the files under the 'Manage Data' tab, then clicking the 'Change File Type' button.

For the particular data in this example, the array data files are in the Agilent Raw TXT format. To specify this, in the 'Manage Files' window shown below, select 'Agilent Raw TXT' from the 'Select New File Type' drop-down list, then click on the 'Save' button above it.

Note

Depending on the assay type and array design used in your own experiment, your data may be in a different format, in which you will have to select that format from the drop-down list, or the file type may be automatically recognized by caArray, in which case you won't have to manually specify it yourself.



Manage Files

Required fields are marked with *asterisks*.

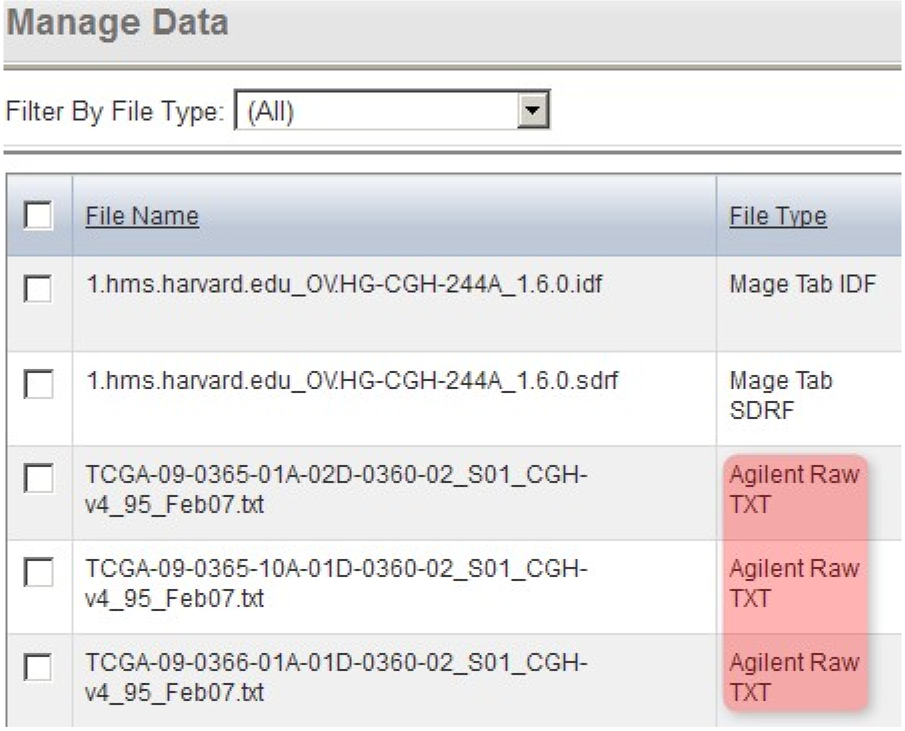
All files listed below will have a new file type assigned.

Select New File Type*: Affymetrix CDF

File Name	Current File Type
TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_9	Agilent Raw TXT
TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_9	Agilent Raw TXT
TCGA-09-0366-01A-01D-0360-02_S01_CGH-v4_9	Agilent Raw TXT
TCGA-09-0366-10B-01D-0360-02_S01_CGH-v4_9	Agilent Raw TXT
TCGA-09-0367-01A-01D-0360-02_S01_CGH-v4_9	Agilent Raw TXT

Manually specify the format of the uploaded array data files by selecting the appropriate format (Agilent Raw TXT in this example) from the 'Select New File Type' drop-down list.

Back on the 'Manage Data' window, the status of all the TXT files now shows as 'Agilent Raw TXT', indicating that caArray now correctly recognizes the file type.



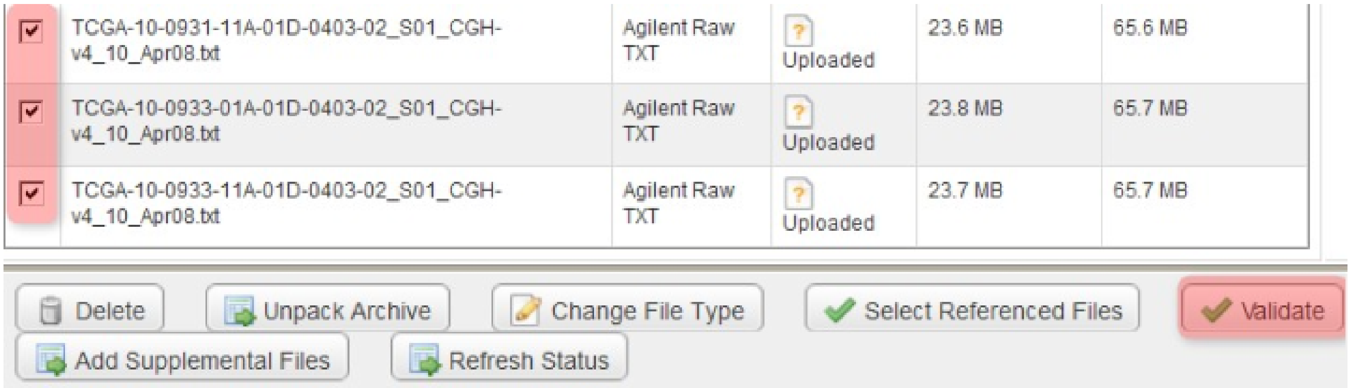
Manage Data

Filter By File Type: (All)

□	File Name	File Type
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF
□	TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	Agilent Raw TXT
□	TCGA-09-0365-10A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	Agilent Raw TXT
□	TCGA-09-0366-01A-01D-0360-02_S01_CGH-v4_95_Feb07.txt	Agilent Raw TXT

Back on the 'Manage Data' window, the format of all the originally unrecognized array data files now shows under the 'File Type' column (as Agilent Raw TXT in our example), indicating that caArray now correctly recognizes the file type.

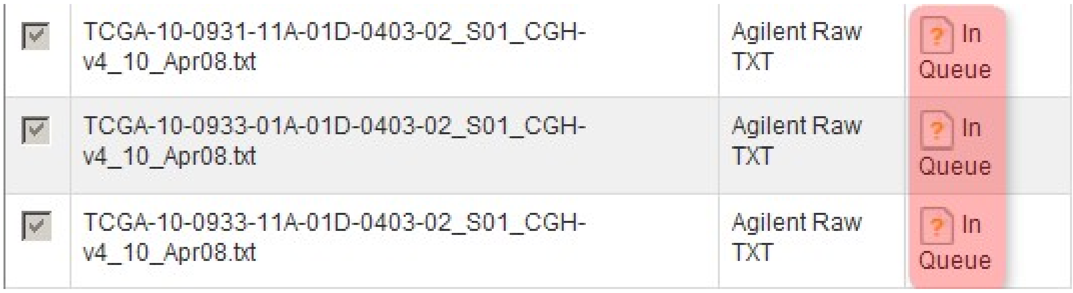
Our next step is to validate all the files, which we will do so by checking off every single file in the list (IDF, SDRF, and TXT), then clicking the 'Validate' button below.



☑	TCGA-10-0931-11A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Agilent Raw TXT	? Uploaded	23.6 MB	65.6 MB
☑	TCGA-10-0933-01A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Agilent Raw TXT	? Uploaded	23.8 MB	65.7 MB
☑	TCGA-10-0933-11A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Agilent Raw TXT	? Uploaded	23.7 MB	65.7 MB

To begin verifying the uploaded data, check off all the array data files under the 'Manage Data' tab, then click the 'Validate' button.

The page will now refresh with the updated status of the selected files showing as 'In Queue'. Depending on the size of the files and the performance of your server, the TXT files may take several minutes to validate, so be patient. Note that the page will not automatically refresh once the files have finished validating, so you will have to manually refresh the page yourself by periodically clicking on the 'Refresh Status' at the bottom of the window until the file status updates again.



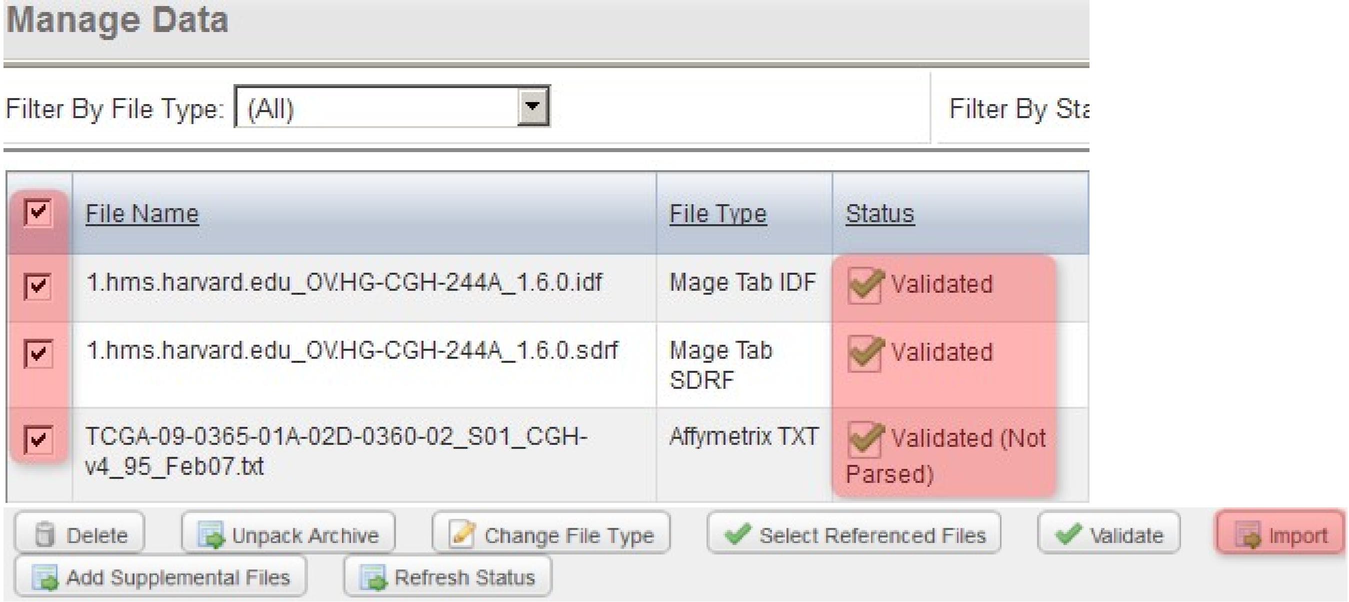
☑	TCGA-10-0931-11A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Agilent Raw TXT	? In Queue
☑	TCGA-10-0933-01A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Agilent Raw TXT	? In Queue
☑	TCGA-10-0933-11A-01D-0403-02_S01_CGH-v4_10_Apr08.txt	Agilent Raw TXT	? In Queue

The 'Manage Data' tab now refreshes with the status of the array data files showing as 'In Queue'.

You'll know when the validation is successful when the status of the files shows as 'Validated' or 'Validated, Not Parsed'.

> ⓘ **NOTE**
>
> The 'Not Parsed' status would only show in versions of caArray prior to v2.4.0 which had not yet implemented a parser for the Agilent TXT format and were thus unable to parse these files. Either way, these files can still be imported into your experiment with or without having been parsed beforehand.

Once the files have been validated, you can import them into the study by checking all the files in the list, then clicking on the 'Import' button below.



Manage Data

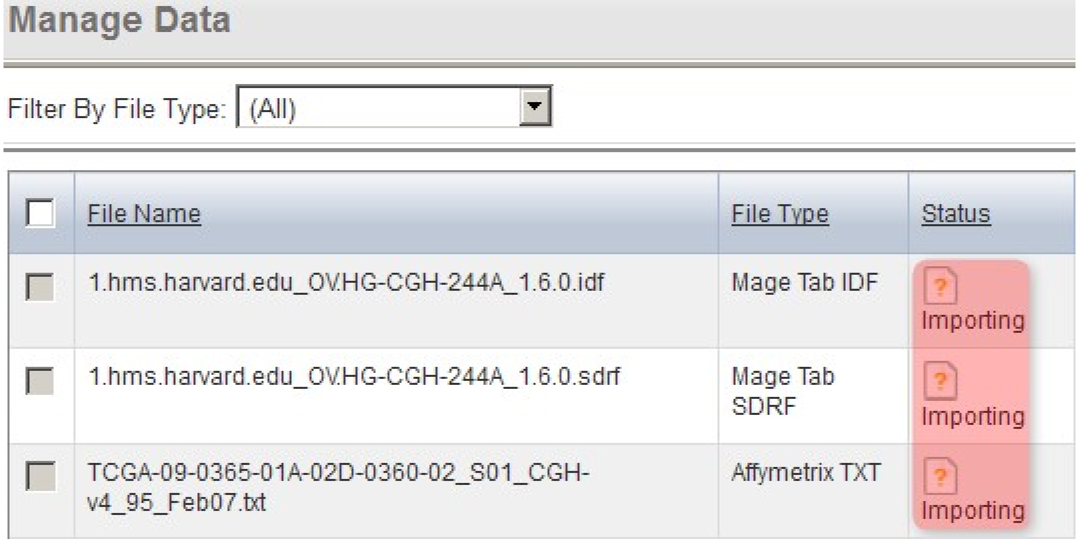
Filter By File Type: (All)

Filter By Sta

☑	File Name	File Type	Status
☑	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF	✓ Validated
☑	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF	✓ Validated
☑	TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	Affymetrix TXT	✓ Validated (Not Parsed)

Once the data finishes validating, the 'Manage Data' tab will appear with the status of the array data files showing as 'Validated' or 'Validated (Not Parsed)', depending on the version of caArray you're running. To import the files, select them all, then click the 'Import' button.

The page will again refresh with the files' status showing as 'Importing'. After a few minutes, click the 'Refresh Status' until the file status updates again.



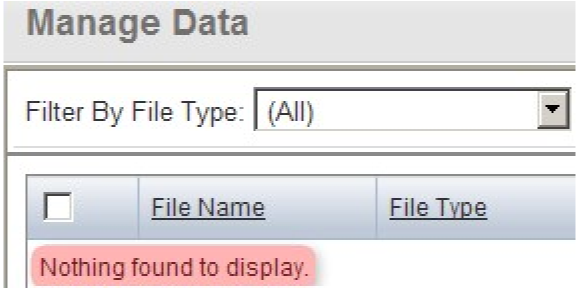
Manage Data

Filter By File Type: (All)

☐	File Name	File Type	Status
☐	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF	? Importing
☐	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF	? Importing
☐	TCGA-09-0365-01A-02D-0360-02_S01_CGH-v4_95_Feb07.txt	Affymetrix TXT	? Importing

The 'Manage Data' tab now refreshes with the status of all the selected files showing as 'Importing'.

You'll know when the importing is successful when the uploaded files no longer appear under the 'Manage Data' tab, with a message stating, 'Nothing Found To Display' in their place, as shown below.

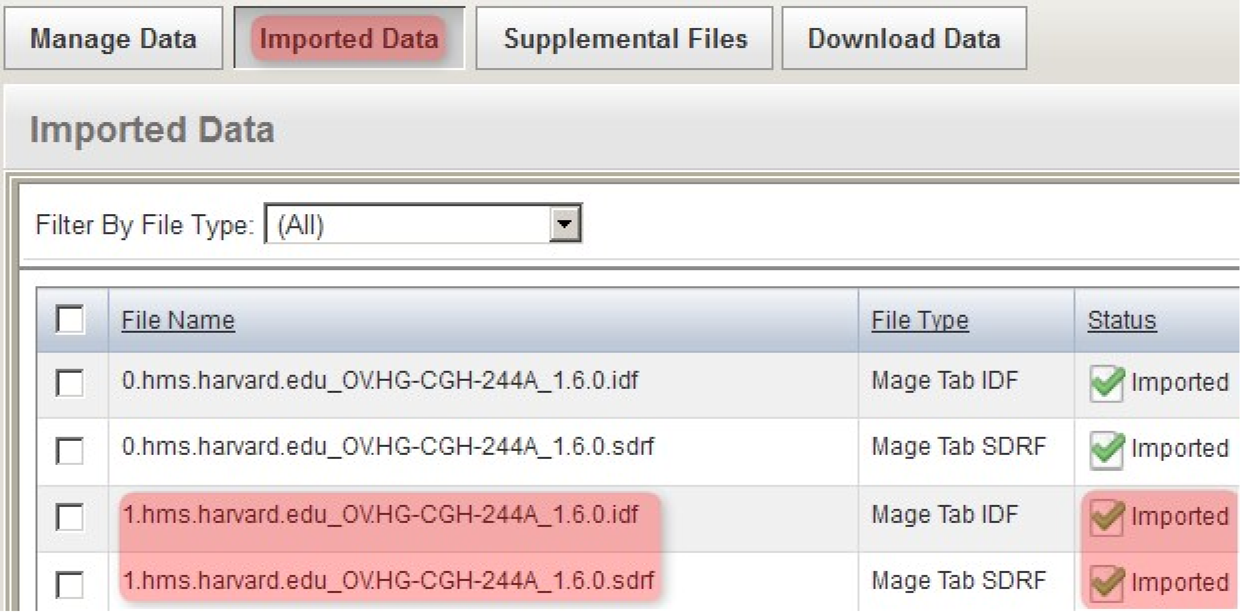


Manage Data

Filter By File Type: (All)

☐	File Name	File Type
Nothing found to display.		

The files now appear under the 'Imported Data' tab, as shown below, with a status of 'Imported'. Note that other, previously uploaded files from the same experiment appear under this tab as well alongside the files we just imported.



Imported Data

Filter By File Type: (All)

□	File Name	File Type	Status
□	0.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF	✓ Imported
□	0.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF	✓ Imported
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.idf	Mage Tab IDF	✓ Imported
□	1.hms.harvard.edu_OV.HG-CGH-244A_1.6.0.sdrf	Mage Tab SDRF	✓ Imported

The imported files now appear under the 'Imported Data' tab with a status of 'Imported' alongside other files from a previous upload to the same experiment.

Reproducing the Procedure

So far, only one-sixth of the data has been uploaded. You can reproduce the procedure we followed so far to upload the data from your experiment. The procedure, summarized below, is as follows:

- Create a ZIP archive for each batch which contains the IDF, SDRF, and all the associated TXT files, ensuring that the size of the archive is less than 2 GB following compression.
- Upload the ZIP archive to your caArray instance
- Depending on the format of your raw array data, manually specify the file type for the array data files, as they may not automatically recognized by caArray
- Validate the uploaded files
- Import the validated files into the experiment

Have a comment?

Please leave your comment in the caArray End User Forum.