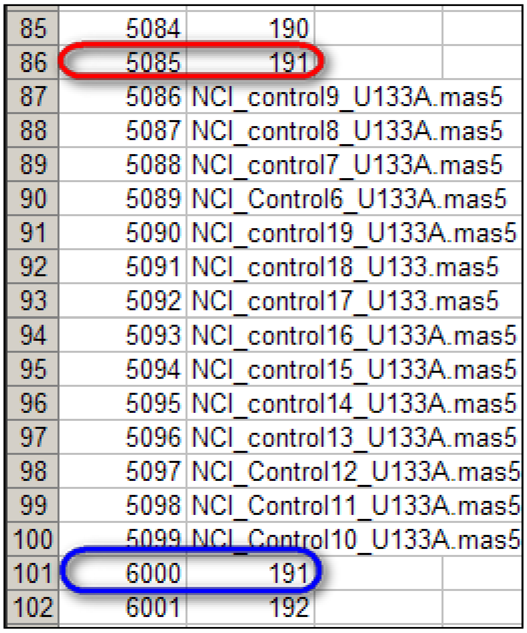
How To Add Data to an Existing Study in caIntegrator - Step 4.12

4.12. To see what obstacles may arise in the course of loading mapping data, let's try another file. This one, named 'duplicate_mapping_file_tutorial.CSV', will replace the one we loaded in steps 4.5 to 4.6. A partial screenshot of this file, taken from a Microsoft Excel 2007 window, is shown below.

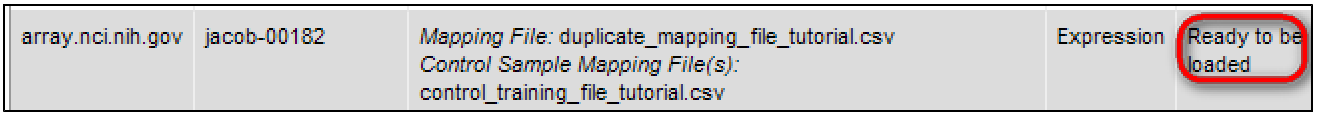


85	5084	190	
86	5085	191	
87	5086	NCI_control9_U133A.mas5	
88	5087	NCI_control8_U133A.mas5	
89	5088	NCI_control7_U133A.mas5	
90	5089	NCI_Control6_U133A.mas5	
91	5090	NCI_control19_U133A.mas5	
92	5091	NCI_control18_U133.mas5	
93	5092	NCI_control17_U133.mas5	
94	5093	NCI_control16_U133A.mas5	
95	5094	NCI_control15_U133A.mas5	
96	5095	NCI_control14_U133A.mas5	
97	5096	NCI_control13_U133A.mas5	
98	5097	NCI_Control12_U133A.mas5	
99	5098	NCI_Control11_U133A.mas5	
100	5099	NCI_Control10_U133A.mas5	
101	6000	191	
102	6001	192	

In this mapping file, the same sample (ID 191) is mapped twice, once to subject ID 5085 (highlighted in red) and again to subject ID 6000 (highlighted in blue).

You may notice something unusual about this mappings: the same sample ID (191) is mapped twice, and each mapping is to a different subject ID (5085 in one case, 6000 to another). This is obviously an error in the mappings, as each sample is taken from a single subject and must be unique to that subject. However, the question remains, what happens when we attempt to load these mappings into the study?

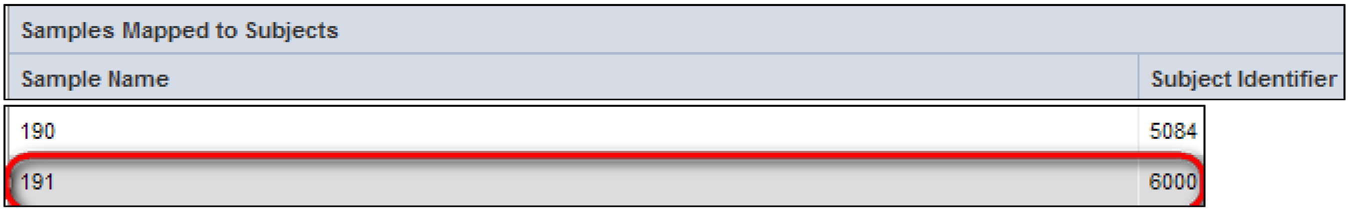
Surprisingly, when we repeat the procedure for loading mappings with the 'duplicate_mapping_file_tutorial.CSV', caIntegrator does not display any error message, and the source's status shows as 'Ready to be loaded' in the 'Genomic Data Sources' table, as was the case with the previous mapping file we loaded successfully. Does this mean that caIntegrator allows multiple mappings of the same sample to different subjects?



array.nci.nih.gov	jacob-00182	*Mapping File:* duplicate_mapping_file_tutorial.csv *Control Sample Mapping File(s):* control_training_file_tutorial.csv	Expression	Ready to be loaded

When loaded an invalid mapping file, caIntegrator does not display any error messages and shows the status of the invalidly mapped source as 'Ready to be loaded' (highlighted in red).

As it turns out, when caIntegrator parses a mapping file where the same sample is mapped to multiple subjects and encounters a sample ID that has already been mapped, it will overwrite the old mapping with the new one. We can confirm this by clicking on the 'Map Samples' button for the source we mapped and examining the 'Samples Mapped to Subjects' table on the 'Edit Sample Mappings' page.



Samples Mapped to Subjects

Sample Name	Subject Identifier
190	5084
191	6000

On the 'Edit Sample Mappings' page, sample ID 191 is only mapped to a single subject (highlighted in red), even though the mapping file we just loaded mapped that same sample twice.

As you can see, the mapping table shows only one mapping for sample ID 191, even though this sample was mapped to two different subjects in the new mapping file we just loaded. The subject ID it's mapped to is 6000 (the second one in the mapping file), not 5085 (the fist one in the mapping file). This means that caIntegrator overwrote the first mapping of sample ID 191 with the second one.

We've learned a valuable lesson from this exercise: be sure to check your mapping file for any duplicates before loading it into your study, as caIntegrator does not perform this check for you!