

Import GEO Experiment into Partek® Genomics Suite™

This tutorial will illustrate how to:

- Import a gene expression experiment from GEO SOFT files
- Specify annotations
- Import RAW data from GEO for gene expression or other kinds of experiments

As this tutorial will not discuss analysis of the imported data, the user is expected to be familiar with data analysis using Partek® Genomics Suite™ (PGS) version 6.6. This tutorial also assumes the user is familiar with the Gene Expression Omnibus (GEO) features needed to search the GEO depository:

<http://www.ncbi.nlm.nih.gov/geo/>.

This tutorial will show you how to import GSE32924 (Suárez-Fariñas et al.) from GEO. This dataset contains three groups of samples: normal, with lesional atopic dermatitis, and non-lesional atopic dermatitis using the Affymetrix HG-U133_Plus2 microarray platform.

Overview of NCBI GEO experiment types

GEO contains many kinds of experiments from multiple vendors including, but not limited to, microarray gene expression, miRNA expression, exon expression, copy number, ChIP-on-ChIP, and Next Generation Sequencing (RNA-Seq) reads. Data can be deposited in either RAW format where the raw files from the scanner (CEL, CHP, gpr, Agilent® Feature Extraction format, etc.) have been uploaded and/or in SOFT (Simple Omnibus Format in Text) format. SOFT files contain **normalized** data and annotations in a simple tab-separated format. The PGS GEO importer can only import **gene expression** data in SOFT format. To use data in other formats or from other experiment types, please download the RAW data files.

This tutorial will demonstrate how to load GSE32924 into PGS version 6.6. As the PGS software evolves rapidly, the screenshots shown in this tutorial may be different if using a different version of the software.

Importing gene expression SOFT files

GEO SOFT gene expression experiments are imported using the NCBI GEO importer available from the *File > Import* menu.

- Set *Workflows* to **Gene Expression**
- Select **File > Import > NCBI GEO Downloader** as shown in Figure 1

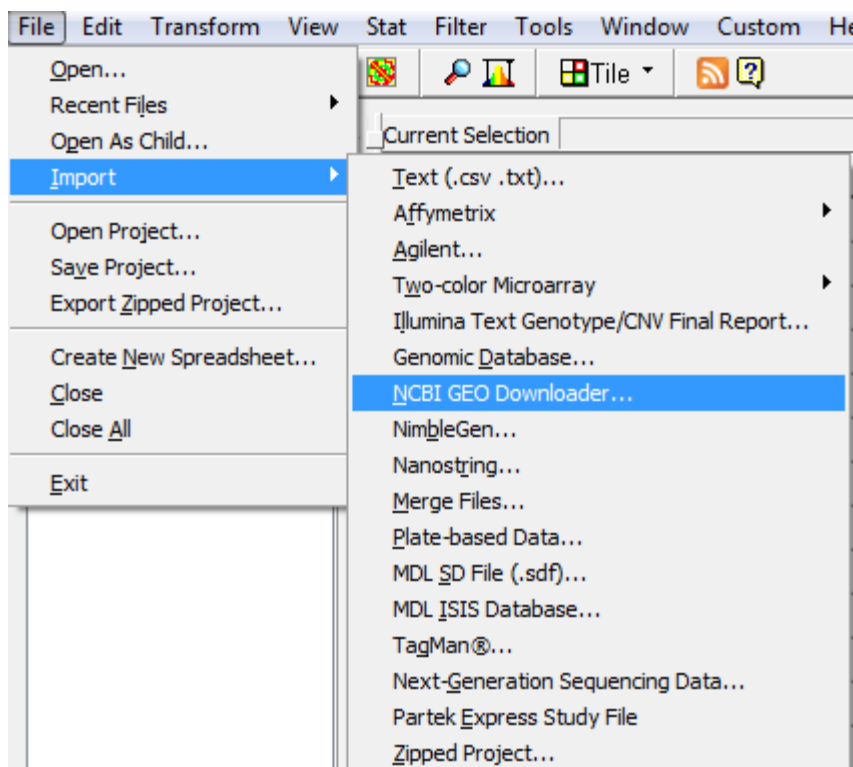


Figure 1: Invoking the GEO SOFT file importer for gene expression experiments

This invokes the *Download NCBI GEO Sample Files* dialog box shown in Figure 2. You have the ability to specify whether you want to download an entire experiment (specify the *GSE number*) or a series of samples (specify a range of *GSM numbers*).

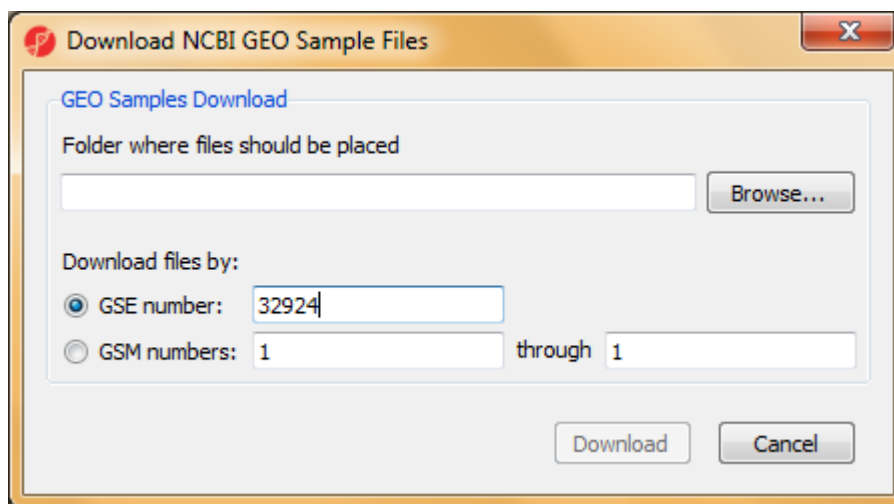


Figure 2: Specifying whether to download an experiment or particular samples

- Specify where the output files should be downloaded by selecting the **Browse** button. You may choose any directory on your system
- Select **GSE number** and type in **32924**. You do not need to specify the GSE in front of the experiment number. Then select **Download**. The

dialog box will list all of the samples as they are being download (not shown)

- When the download is complete, a dialog box will pop up that asks, *Would you like to merge these files?* Since the goal is to combine all of the data from the separate samples into one experiment, select **Yes**
- This invokes the *File Merge – Path Selection* dialog shown in Figure 3. To change where the merged file is stored (different location than the downloaded files), select *Browse* next to *Folder where files to be merged reside*. To change the name of the spreadsheet and the location of the PGS experiment files, select *Browse* next to *Folder where files to be merged reside*. Both of these steps are optional. However, in the *File Format Selection* panel, select the radio button next to **Identifying terms in columns (for example, gene IDs in columns)**. This is the usual format of an experiment (samples in rows; genes in columns). However, if you wish to download two-color data, you must use the top option, *Identify terms in rows (for example one gene ID per row)*. For this tutorial (one-color data), select **Next**

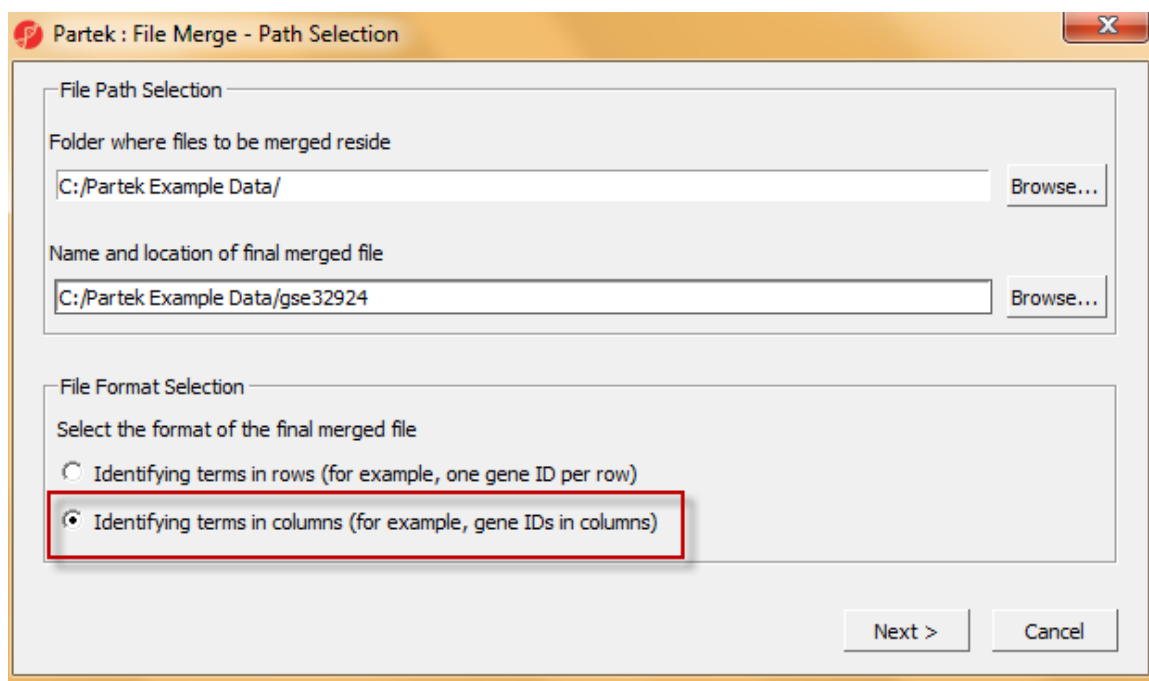


Figure 3: Specifying the format of the merged experiment

- The next dialog box (Figure 4) allows you to preview the sample files that were just downloaded by the NCBI GEO importer. If there were other .txt files in this folder, they could be added by selecting them in the *Files in Folder* pane and moving them to the *Files to Merge* pane by selecting the move button ->. As no other samples will be added to this experiment, select **Next**

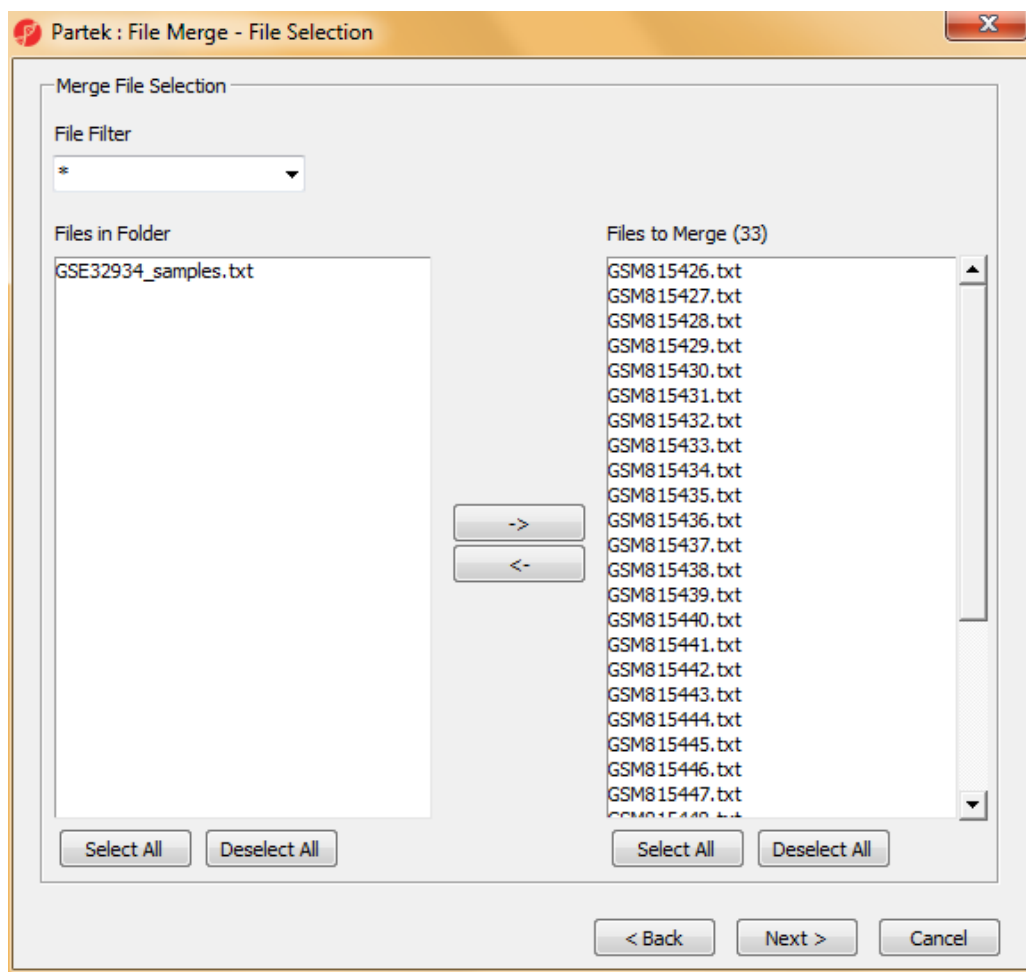


Figure 4: File Merge - File Selection dialog showing the samples that were just downloaded as part of the experiment. Samples from other GEO experiments could be included at this step

- Each sample will contain columns of normalized intensity data (*VALUE*), and each column/probeset has a header (*ID_REF*). PGS must be instructed what to do with these two fields (*VALUE* and *ID_REF*). As shown in Figure 5, set *Columns to Include In Merge* to **Value** and *Columns used to identify records common between each file (term columns)* to **ID_REF** and select **Next**
- You must next tell PGS what kind of data object is contained in the files. Set *ID_REF* to **text** and *VALUE* to **numeric** as shown in Figure 6. Furthermore, PGS must be directed how to handle duplicate terms. The defaults (*Compute mean* and *Show a list of values*) are appropriate here. Select **Next**

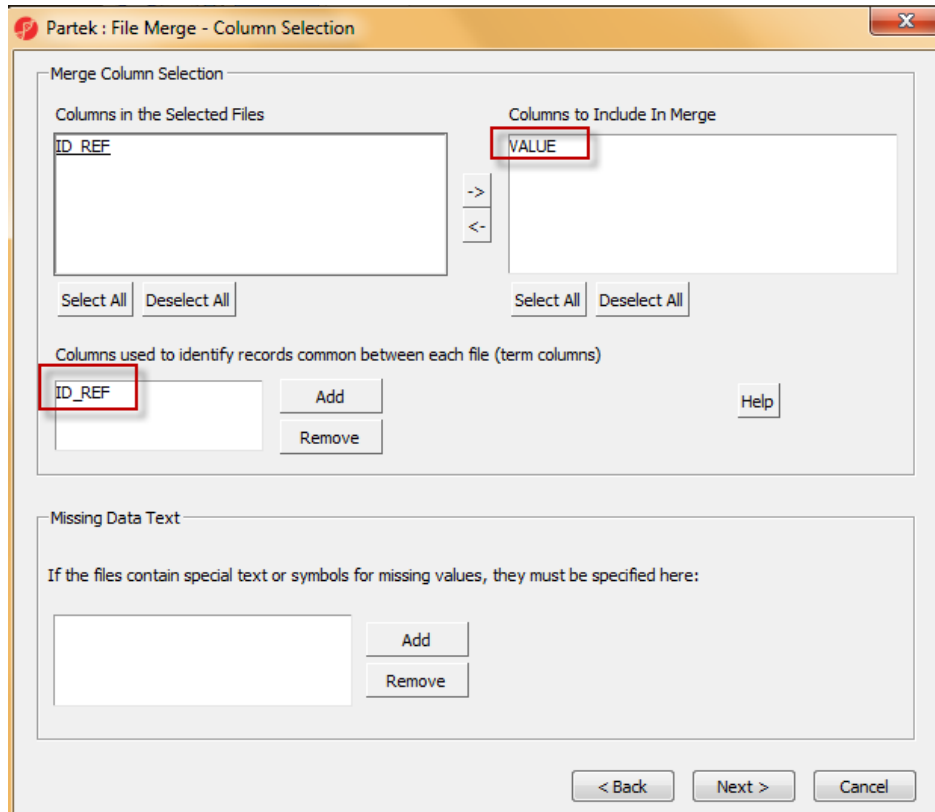


Figure 5: Identifying the values (intensities) as well as the probeset names

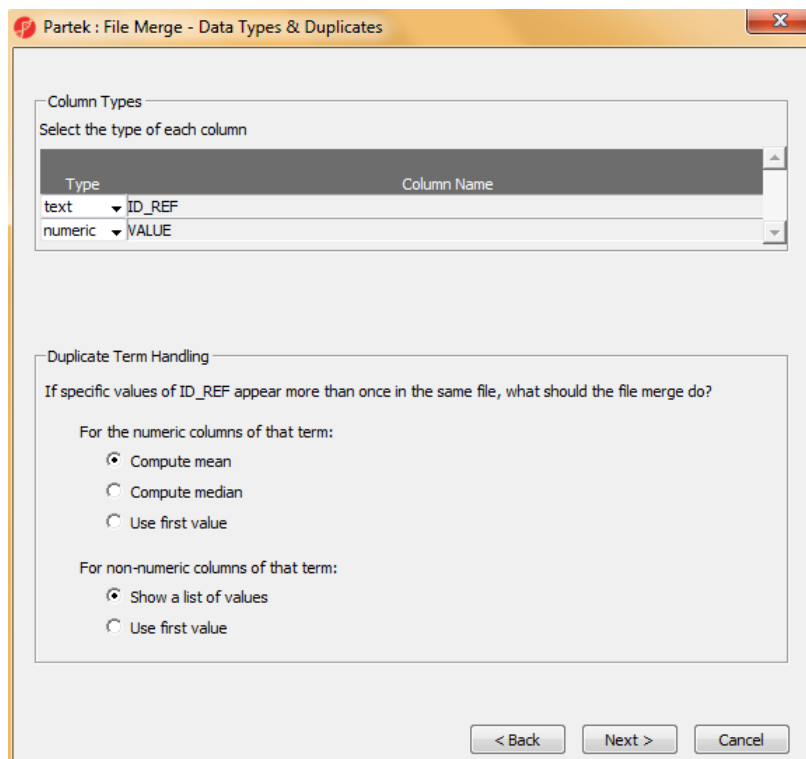


Figure 6: Identifying the type of data each column represents and how PGS should handle multiple values for the same probeset

- The *NCBI GEO Downloader* does not automatically import the sample attributes (sample groups, phenotype information, etc.). The properties can be specified in a file (*Load categories/properties from a file*) or entered manually (*Enter categories/properties manually*). As it is easier to describe the samples after the import, select **Next** as shown in Figure 7

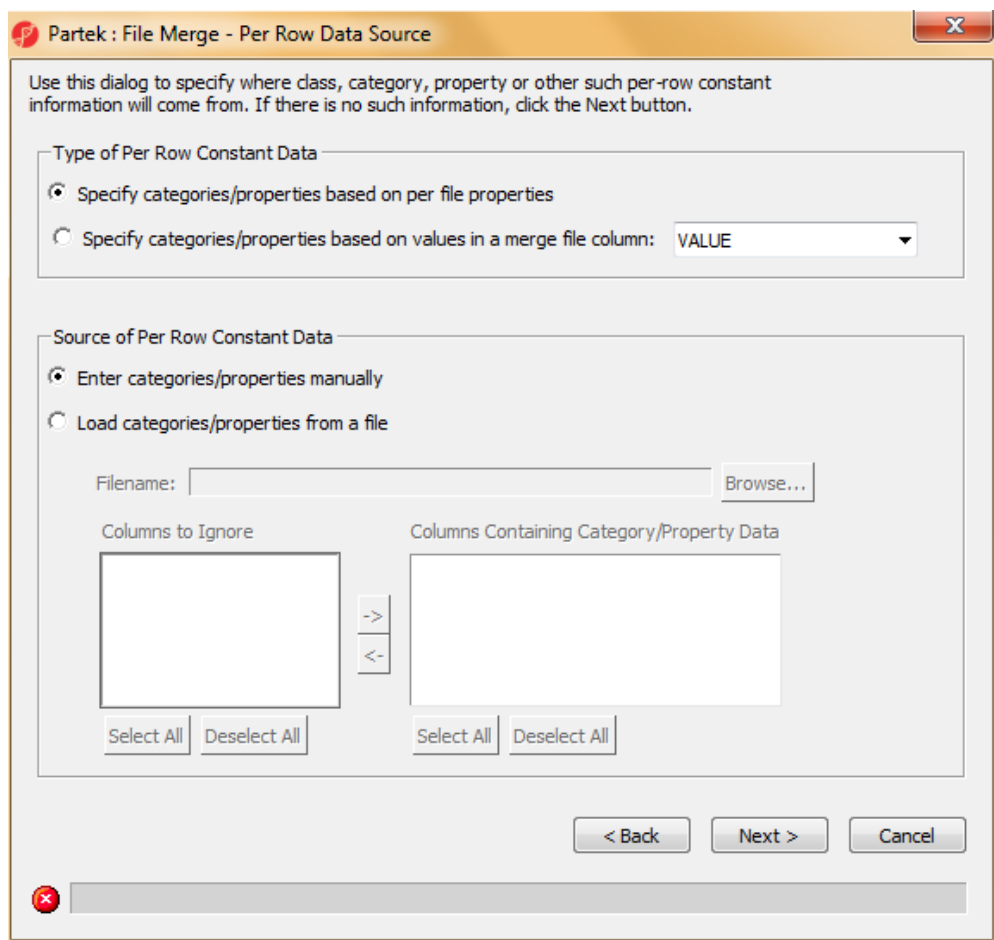


Figure 7: Optional step to add sample attributes

- Since this interface is not as friendly as *Add sample attributes*, select the *Name 1* column shown in Figure 8 and select **Delete Selected Property Columns**, **Yes**, and **Next**

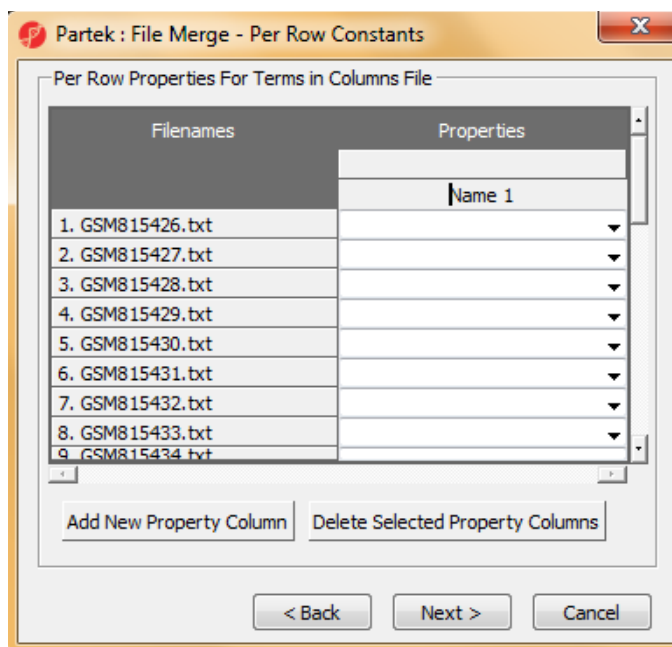


Figure 8: Delete the Name 1 properties column

The sample files will be read in, and the progress updated as shown in Figure 9.

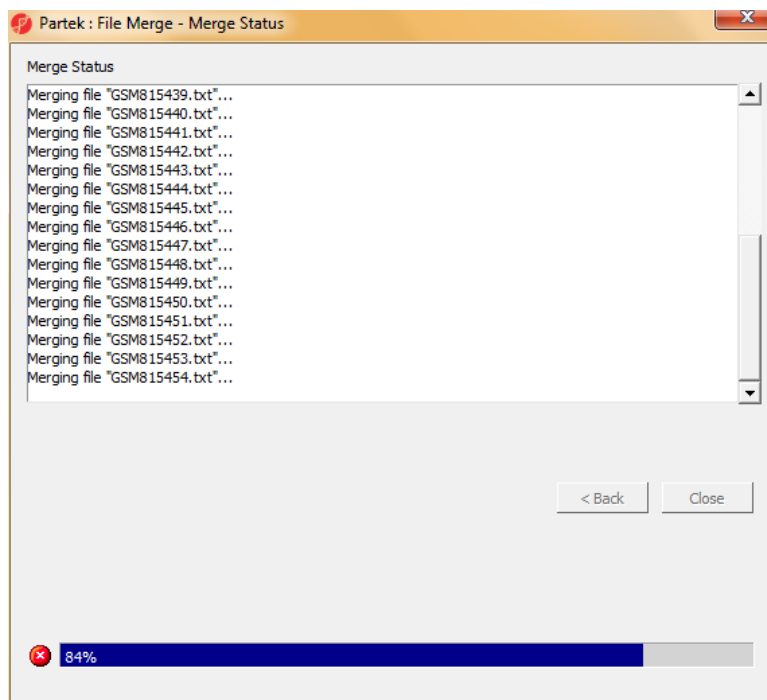


Figure 9: Progress showing the samples are they are imported into PGS

- When the *Merge Status* dialog box shows *Done*, select **Close**

The main spreadsheet will be created (Figure 10). Notice that samples are shown on rows and probesets are shown on columns. Column 2 shows an extraneous column which may be deleted by right-clicking on the column header and selecting **Delete**.

Current Selection		GSM815426.txt							
	1. Filename	2. !sample_table_end	3. 117_at	4. 121_at	5. 177_at	6. 179_at	7. 243_g_at	8. 266_s_at	9. 320_at
1.	GSM815426.txt	?	6.48148	4.42126	6.11105	2.1487	11.2166	10.7846	4.55006
2.	GSM815427.txt	?	6.14175	4.4157	6.74082	2.1487	10.9695	10.3923	4.33998
3.	GSM815428.txt	?	6.06948	4.26438	6.33314	2.1487	11.3825	12.295	4.57797
4.	GSM815429.txt	?	7.28279	4.29721	6.0275	2.15716	11.1987	12.2491	4.23161
5.	GSM815430.txt	?	5.91387	4.42126	7.35161	2.1487	11.7516	12.2783	4.25446
6.	GSM815431.txt	?	6.39249	4.43239	7.35491	2.1487	12.1206	13.3553	4.21115
7.	GSM815432.txt	?	5.72716	4.38957	5.10321	2.1487	10.5467	12.7407	3.97823
8.	GSM815433.txt	?	6.05792	4.18503	6.33505	2.1487	11.0207	12.71	3.82105
9.	GSM815434.txt	?	4.6146	4.42126	5.59659	2.1487	10.9568	13.0054	4.23161
10.	GSM815435.txt	?	6.06716	4.42126	7.66661	2.24512	11.4629	11.7218	3.98413
11.	GSM815436.txt	?	6.05644	4.26052	5.92805	2.1487	11.426	13.008	4.52146
12.	GSM815437.txt	?	6.06716	4.42126	7.78475	2.1487	11.9194	12.961	4.47771
13.	GSM815438.txt	?	5.06375	4.39303	6.19395	2.1487	11.6537	12.3299	4.81002
14.	GSM815439.txt	?	6.09791	4.42126	6.71247	2.1487	11.4399	11.6015	4.23161
15.	GSM815440.txt	?	6.15384	4.42126	6.69544	2.1487	11.7992	11.8093	4.23161

Figure 10: Top-level spreadsheet shown after importing GSE32924

Add sample attributes to describe the sample groups

As mentioned earlier, PGS does automatically import the sample groups from GEO. This information can be easily added in PGS by using the information from the GEO description of the experiment (Figure 11) and *Add sample attributes* from the *Import* section of the *Gene Expression* workflow.

- Select **Add sample attributes** from the *Import* section of the *Gene Expression* workflow
- Select **Add a categorical attribute** then **OK**
- Change *Attribute name* to something meaningful like **Disease**
- Since there are 3 groups in this experiment, select **New Group**. Change the names *Group 1*, *Group 2*, and *Group 3* to something more meaningful like **AN**, **ANL**, and **Normal**
- Drag and drop the files into their correct groups according to the information in the GEO description of the experiment as shown in Figure 12

Platforms (1)	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Samples (33)	GSM815426 ANL Chip#23
Less...	GSM815427 AL Chip#24
	GSM815428 ANL Chip#25
	GSM815429 AL Chip#26
	GSM815430 ANL Chip#27
	GSM815431 AL Chip#28
	GSM815432 ANL Chip#29
	GSM815433 AL Chip#30
	GSM815434 ANL Chip#31
	GSM815435 AL Chip#32
	GSM815436 ANL Chip#33
	GSM815437 AL Chip#34
	GSM815438 AL Chip#35
	GSM815439 ANL Chip#36

Figure 11: GEO sample descriptions

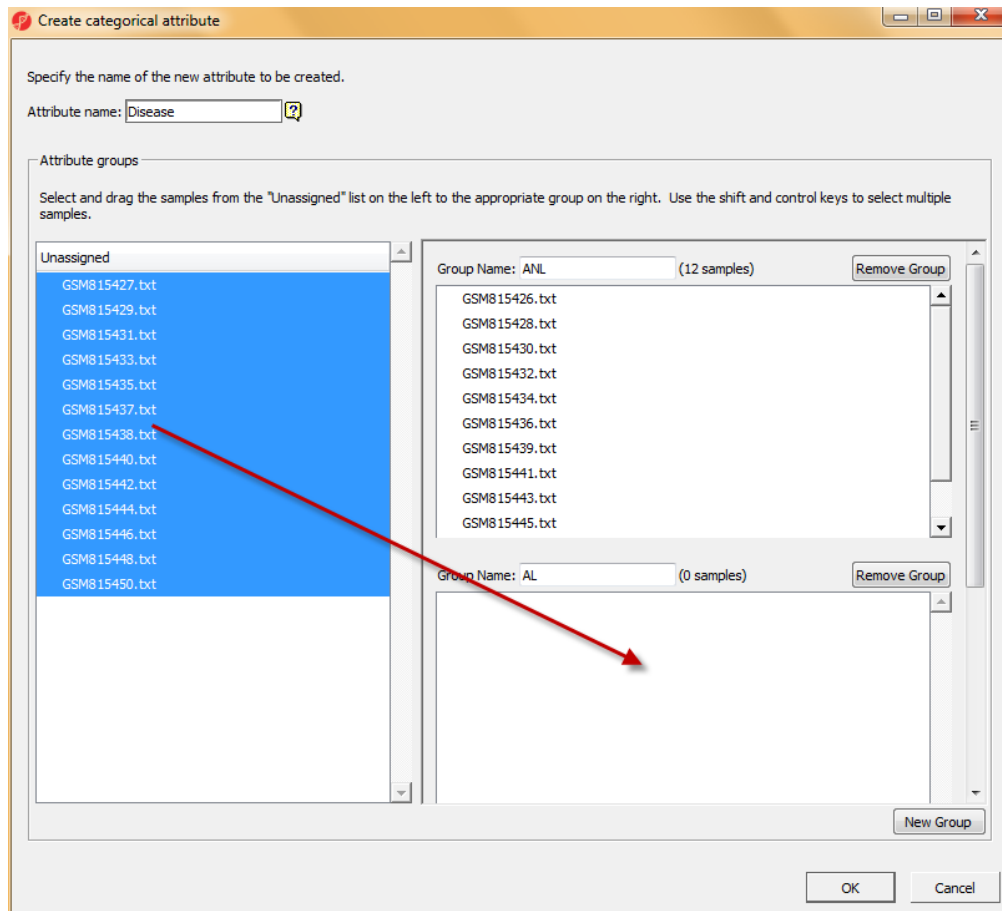


Figure 12: Dragging the last set of samples into the AL group

Adding Annotations

As imported, the data can be analyzed, but to extract the most meaning from the data, the data will need to be annotated.

Affymetrix annotations

Affymetrix® annotations are the easiest annotations to add since PGS can download the annotations directly from Affymetrix. This tutorial uses the U133_Plus2 array.

- Right-click on the spreadsheet in the spreadsheet navigator (pane) on the left and select **Properties**. The dialog box shown in Figure 13 will appear

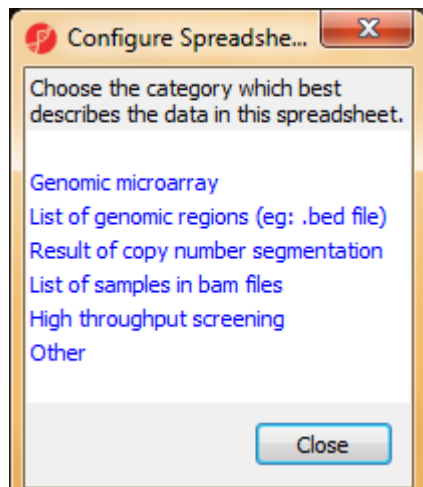


Figure 13: Adding annotations to an Affymetrix microarray

- Select **Genomic microarray**. This invokes the dialog box shown in Figure 14. Select **Gene Expression** from the *Choose the type of genomic data* pull-down menu. Under *Chip*, type **HG-U133_Plus_2** and select **Download**. If the Affymetrix microarray has already been used, you may be able to select *HG-U133_Plus_2* from the *Chip* drop-down menu
- Make sure that *Species* is set to **Homo sapiens**. Select **Edit Genome** and make sure that *Species Name* and *Genome Version* are set appropriately. For na32 annotations, *Genome Version* is *hg19*. Select **OK** to exit *Edit Genome*
- Select **OK** to exit *Configure Genomic Properties*

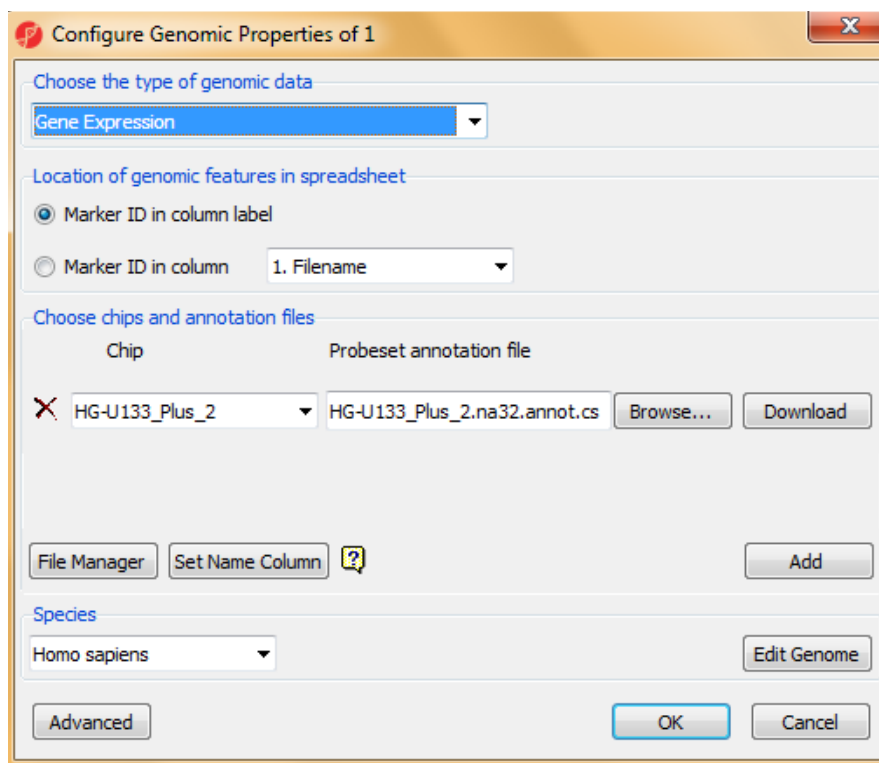


Figure 14: Specifying the Affymetrix microarray type

Other vendor annotations

The import of annotations for other vendors (Agilent[®], Illumina[®], and Nimblegen[®]) is handled differently in PGS. Instead of an automatic download of annotations, you will need to use a tab-separated file. These annotation files may be obtained from the vendor or from GEO. Each GEO experiment has a platform associated with it (GPL); GPLs will contain annotations that can be downloaded as SOFT file or as a table. The GEO experiment used for this tutorial is associated with GPL570.

Requirements for tab-delimited or comma-separated annotation files

The annotation files cannot have blank lines or comment lines at the top or bottom of the file, a header line describing the columns is optional. There may not be comments interspersed with annotations; the annotations must be contiguous within the file. Each annotation line should have the data arranged consistently (in the same column order) although missing data is allowed.

For illustrative purposes, the annotations will be downloaded from GEO for this tutorial. The Affymetrix download described in the previous section is easier and is less likely to result in older annotations being used.

- Download GPL570 annotations from GEO. Prepare the file (unzip, remove comments, etc.) as described in the *Requirements for tab-delimited or comma-separated annotation files*

- Right-click on the spreadsheet in the spreadsheet navigator (pane) on the left and select **Properties**. The dialog box shown in Figure 13 will appear
- Select **Genomic microarray**. This invokes the dialog box shown in Figure 14. Select **Gene Expression** from the *Choose the type of genomic data* pull-down menu
- Select **Browse** and choose the text file that contains the annotations which invokes the *Configure Annotation* dialog box shown in Figure 15

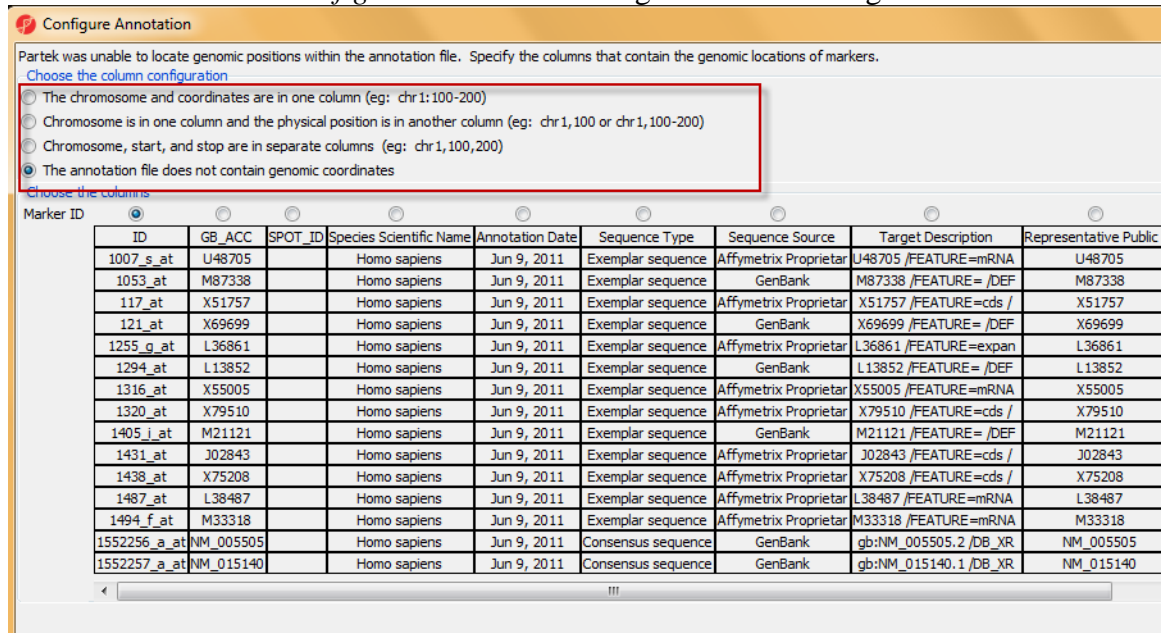


Figure 15: Configuring genomic coordinates from an annotation file

- Use the scroll bar at the bottom of the screen to examine the columns of the annotation file to look for chromosomal locations (chromosome number, start, and stop locations). If the chromosomal locations of the probes are present in the annotation file, take note of the layout of the annotations. Choose the radio button from the choices highlighted in Figure 15 that corresponds to the layout of the chromosomal locations in the annotation file. If any of the first three radio buttons are selected, the menu will be expanded as shown in Figure 16. The expanded options are used to specify which columns contain the chromosome number, start location, and stop locations. *Marker ID* must denote to the column headers shown in the imported data spreadsheet and does not have to be in the first column. Specify the correct format radio button and specify the chromosomal locations accordingly.

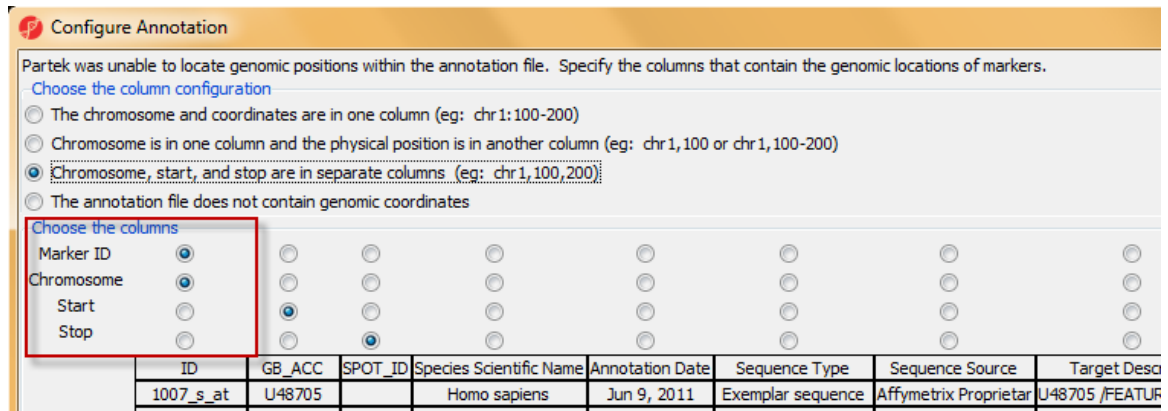


Figure 16: Expanded radio buttons for specifying the chromosomal locations in the annotation file

- Since the GEO annotations did not contain any genomic coordinates, select **The annotation file does not contain genomic coordinates** and **Close**
- Select **Set Name Column** and choose the annotation field that contains the gene symbol and select **OK**
- Choose the correct species under the *Species* drop-down list (**Homo sapiens** in this case) or type in the species if it is not present in the drop-down list
- Select **Edit Genome** and make sure that *Species Name* and *Genome Version* are set appropriately. For na32 annotations, *Genome Version* is **hg19**. Select **OK** to exit *Edit Genome*
- Select **OK** to exit *Configure Genomic Properties*
- If the message shown in Figure 17 appears, this indicates that no genomic coordinates were specified. In this case, this is correct because the annotation file did not contain genomic coordinates, so select **No**. However, if the annotation file contained annotations and you attempted to specify them as shown in Figure 16, this message is an indicator that something was wrong with the specification, likely the *Marker ID* column was not correct or the incorrect radio buttons were specified. In order to be able to reconfigure the annotations, you must go to the directory that contains the annotation file (outside of PGS) and delete the index file that corresponds to this index file (file will have the same name as the annotation file but will have an .idx suffix). Delete this file before selecting **Yes**

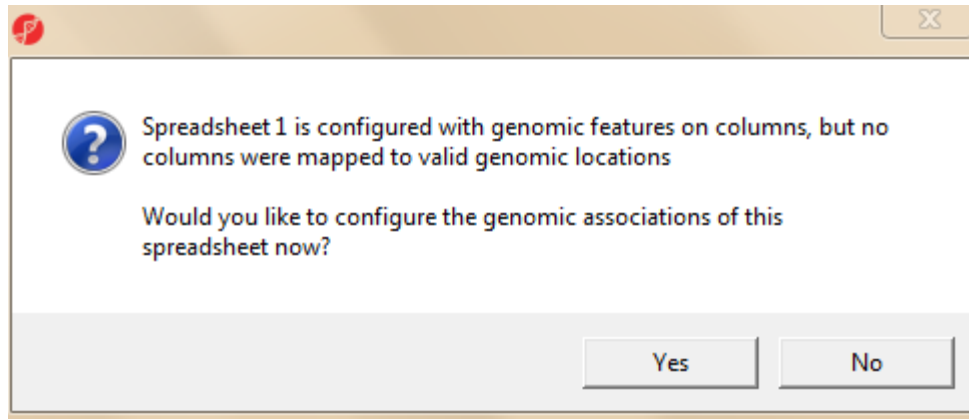


Figure 17: Message indicates that no genomic coordinates were specified

Analyze Data

Now that the data has been imported, annotated, and sample attributes have been added, the data may be analyzed exactly the same way as any other gene expression dataset in PGS.

Importing RAW data from GEO

The previous sections have described how to import normalized gene expression data from GEO into Partek® Genomics Suite™ using SOFT files. However, other gene expression data and other kinds of experimental data like copy number, CHIP-on-Chip, methylation, etc. have been deposited into GEO. If the RAW data (.gpr files, .CEL, CHP, Feature Extraction output from Agilent™, or Nimblegen .pairs files) have been deposited to GEO, then you may download the data from GEO outside of PGS and import it into PGS following the workflow's steps for importing that vendor's raw data. This would facilitate combining GEO data using the same array with your data (pay attention to batch effects when combining datasets from different sources).

If you wish to import other assay data in SOFT files, you will should download the data and import the data as a text file (not discussed in this tutorial).

End of Tutorial

This is the end of the tutorial. If you need additional assistance with this data set, you may call our technical support staff at +1-314-878-2329 or email support@partek.com.

References

Suárez-Fariñas M, Tintle SJ, Shemer A, Chiricozzi A et al. Nonlesional atopic dermatitis skin is characterized by broad terminal differentiation defects and variable immune abnormalities. *J Allergy Clin Immunol* 2011 Apr;127(4):954-64.e1-4.

Last revision: September 13, 2012