Toy data example – How To

1 The experiment

This example uses data (unpublished) of a chromatin immunoprecipitation experiment (ChIP) performed on samples from human cancer patients. ChIP was done using antibodies against acetylation on histone 3 (AcH3). We use data for samples from three patients.

A microarray was designed to study promotor regions of a large number of genes known or suspected to be involved in cancerogenesis. Each patient's ChIP sample was hybridized to a microarray (ChIP-Chip). An annotation file linking each array probe to a genomic locus was prepared.

Furthermore, the three samples were pooled and sequenced using a next-generation sequencing platform (ChIP-seq). The sequenced reads (36bp) were truncated to the first 24 bases and mapped using the SOAP aligner[1]. To make the data files smaller, we removed all unneeded and redundant columns (such as the read sequence), creating an input file with only three columns (strand, chromosome, starting position).

2 Input files

We provide 5 input files:

- Three array scans in GenePix (GPR) format (`patient*.gpr`)
- The reduced file of mapped reads (`mappedreads.tsv`)
- The genomic coordinates for the microarray probes (`coordinates.tsv`).

3 Aim and Caveat

Chip-Seq data is usually very large, containing many millions of mappable reads. In order to keep the size of this example small, we decided to only use a subset of all reads. This results in very low coverage of the genome, and as a result, the "expression values" created from the sequencing data can not be used for a meaningful analysis. The aim of this example is to show how different types of high-throughput experimental data can be imported into MAYDAY.

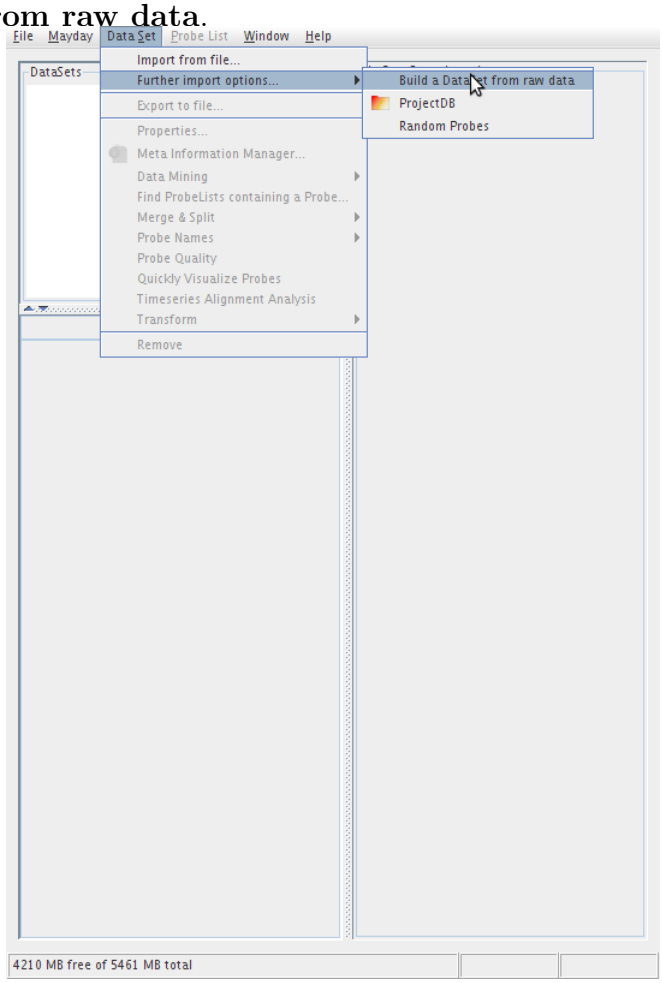
In this document, we describe how to integrate the different data types into one data set in MAYDAY and show how to create some first visualizations of the data.

4 Starting Mayday

Start MAYDAY[1] from our website at
`http://www-ps.informatik.uni-tuebingen.de/mayday/exp/webstart/Mayday.jnlp`
using Sun's Java WebStart (usually all you have to do is open the address in your web browser).

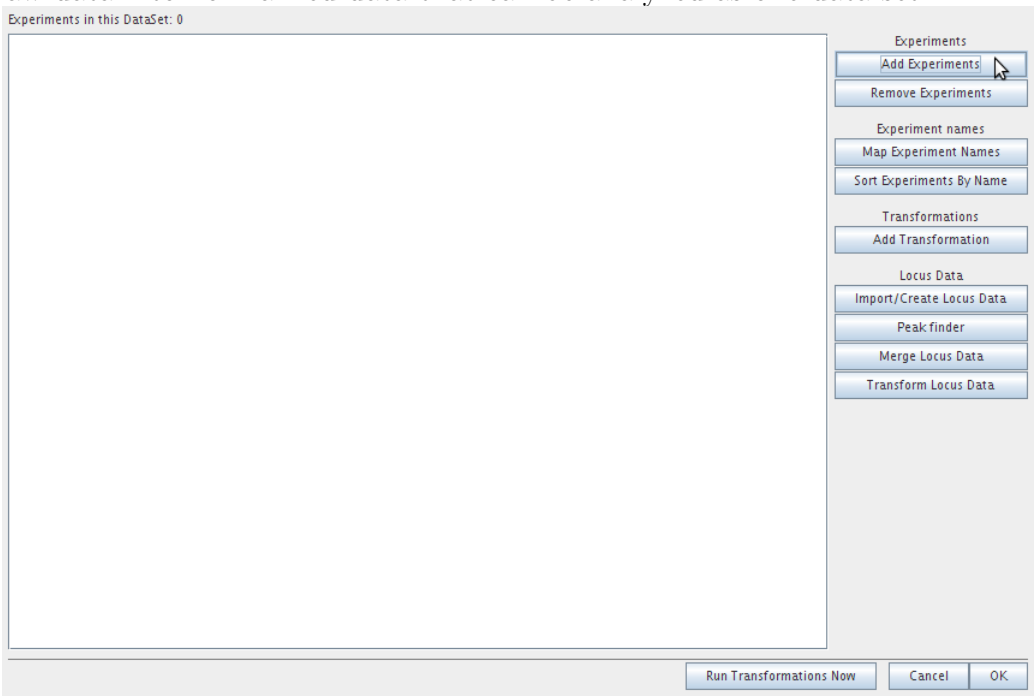
5 Importing the data

From MAYDAY's menu, select **Data Set→Further import options→Build a DataSet from raw data**.

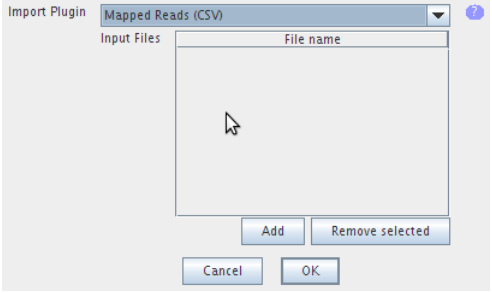


[1] Use the experimental version, not the stable build.

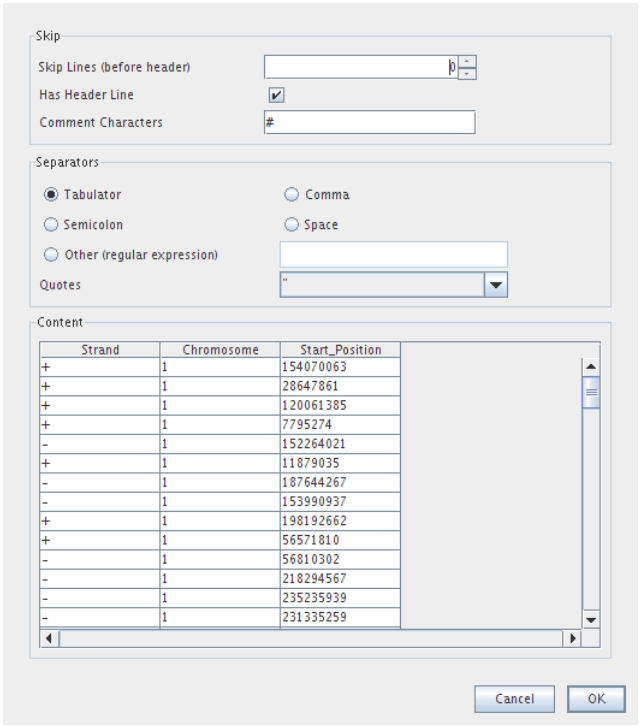
This will open the "Add and configure Experiments" dialog. We will now add the raw data files. After that, we will add transformation steps to convert the raw data into normalized data that can be analyzed as one data set.



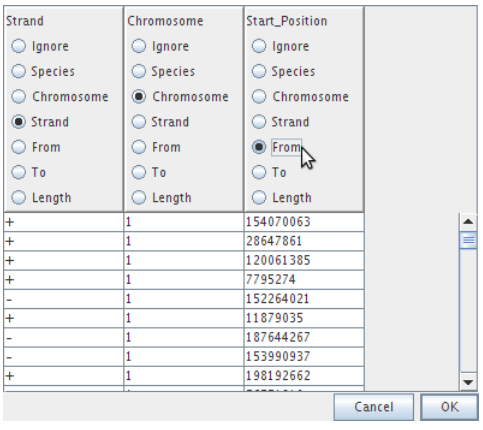
Click **Add Experiments** and select **Mapped Reads** from the drop-down list. Click **Add** and selected the example file `mappedreads.tsv`. Click **OK**.



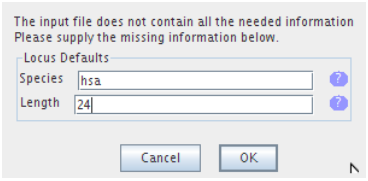
Make sure all settings are correct to parse the tabular file: The file is tab separated, and has a header line. Click **OK**.



MAYDAY needs to know what the columns contain. Select the appropriate options: The first column contains the strand information, the second denotes the cromosome, the first column contains the mapped start position of the read. Click **OK**.

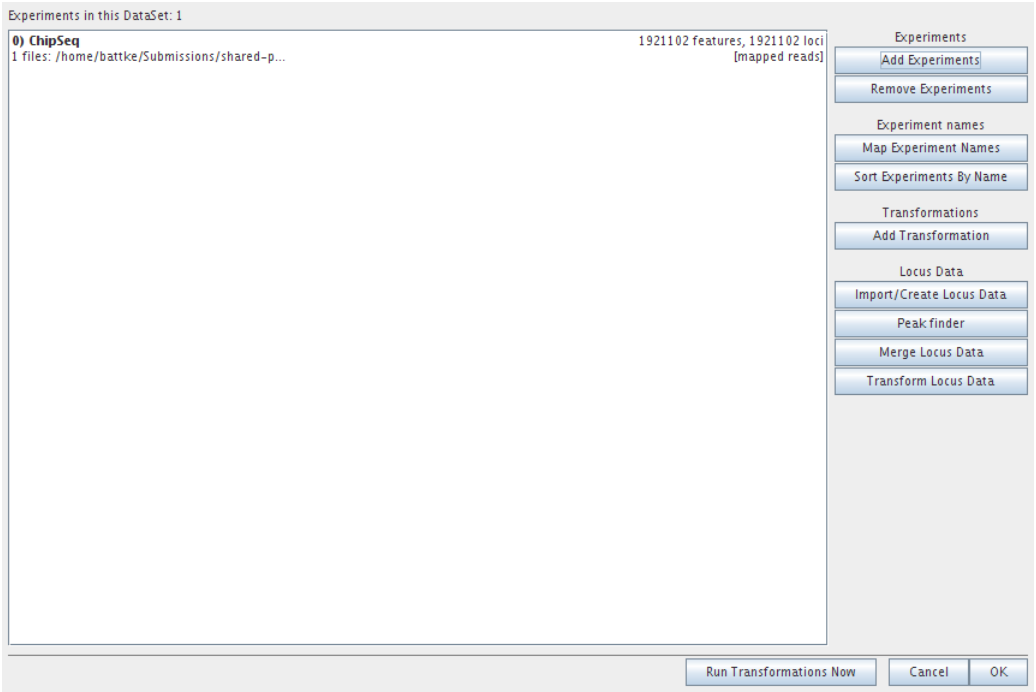


Since we removed redundant information to make the example files smaller, you need to supply the organism name ("hsa") and the read length (24bp) manually. Click **OK**.

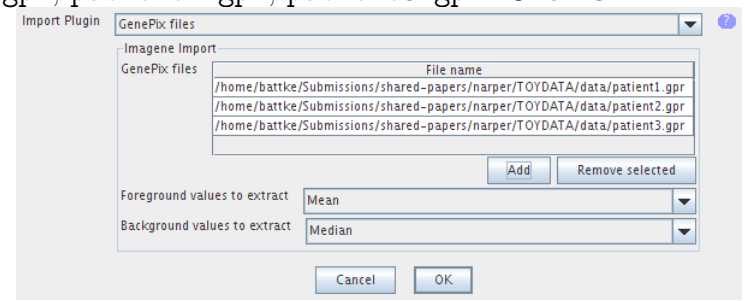


You can enter any name you like for this experiment. Click **OK**.

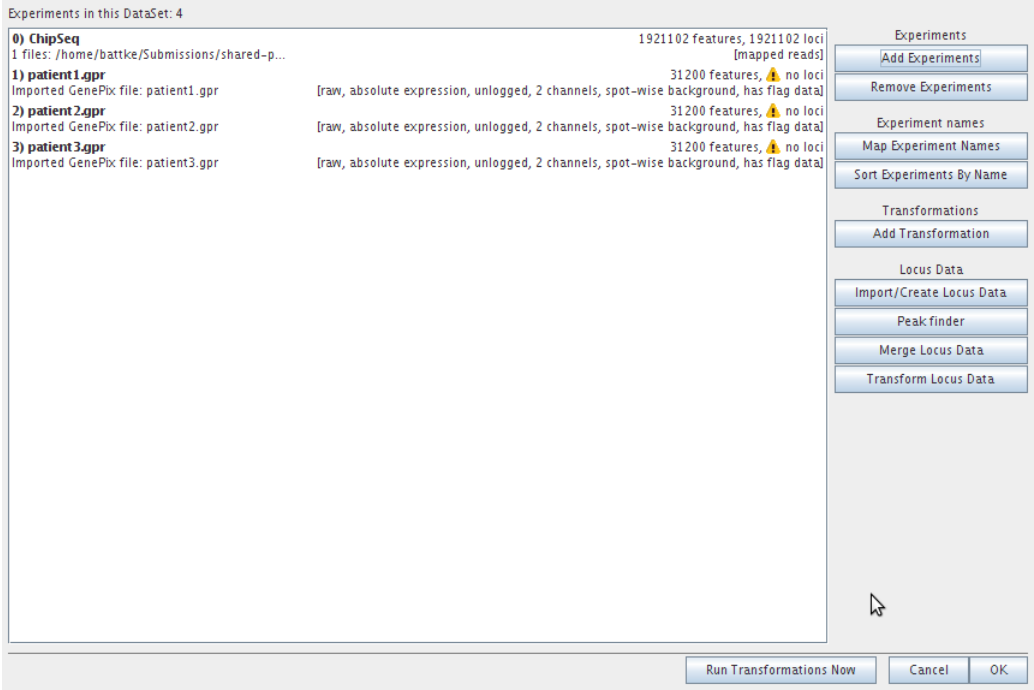
We have now imported the ChIP-seq data.



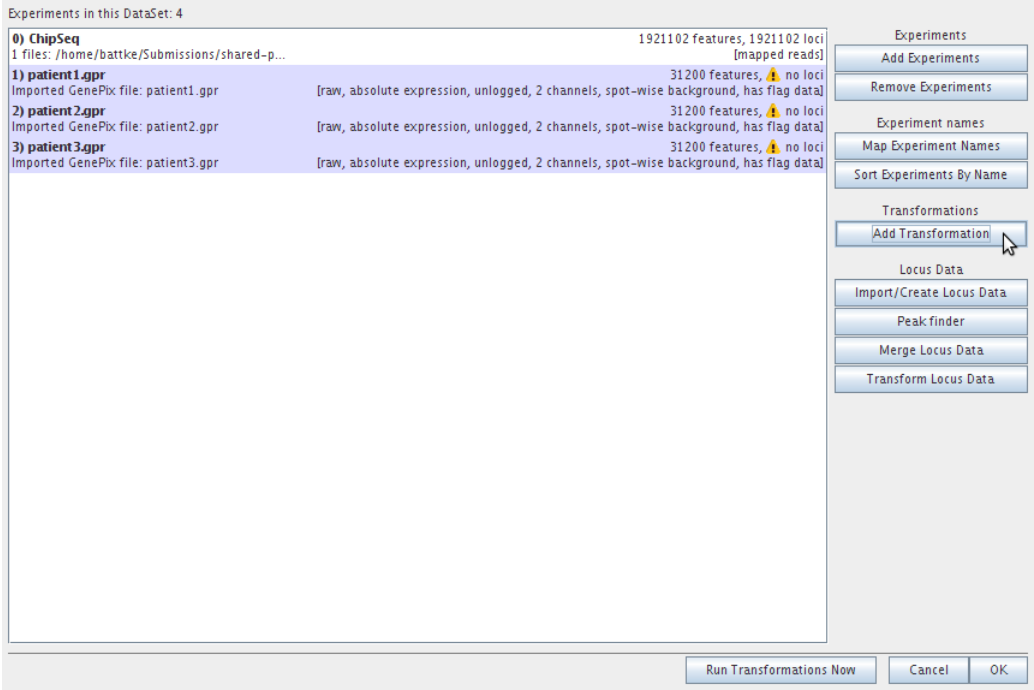
To import the ChIP-chip data, again click**Add Experiments**. Select **GenePix** from the drop-down list, click **Add** and select the example files `patient1.gpr`, `patient2.gpr`, `patient3.gpr`. Click **OK**.



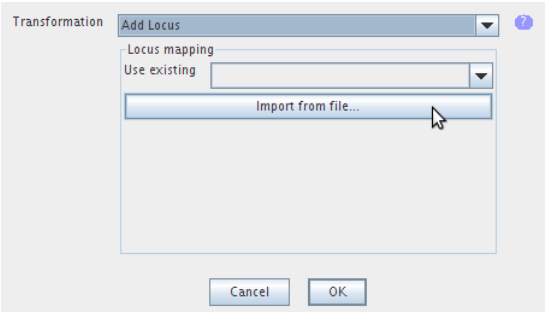
We have now imported the ChIP-chip data. In the next steps, we will add transformations to the data to make the experiments compatible and to apply normalization.



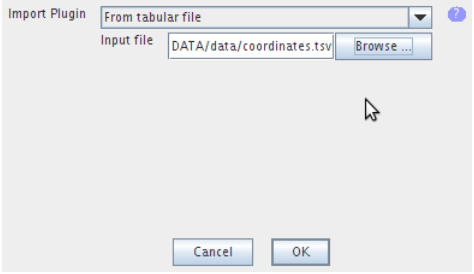
We first add the locus annotation to the ChIP-chip experiments. select the three patients (click on patient1, then shift-click on patient3). Then click **Add Transformation**.



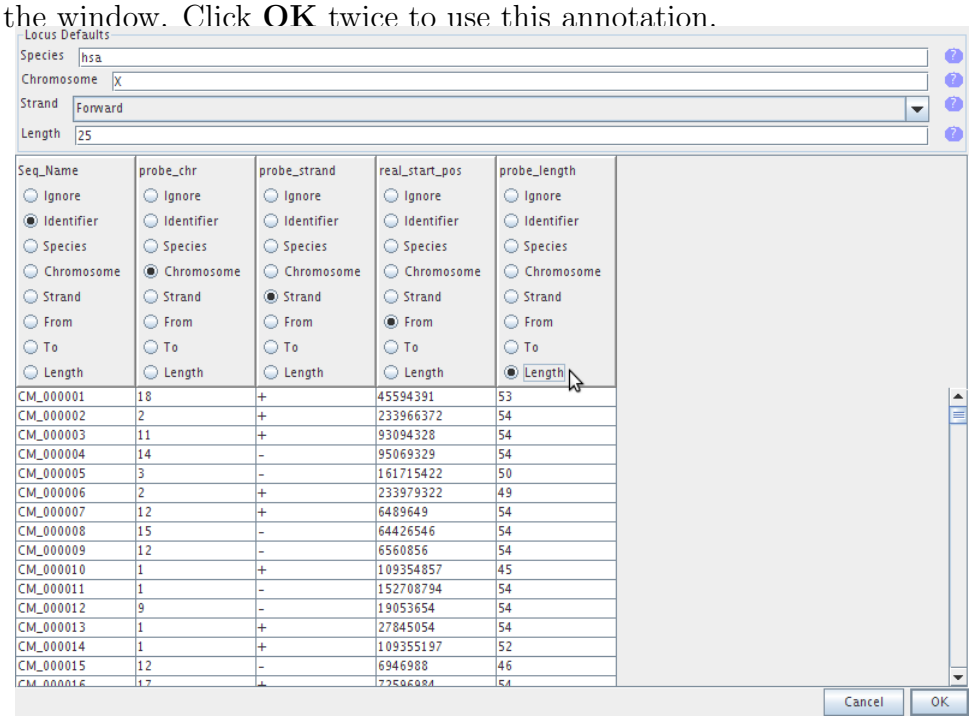
Select **Add Locus** and click on **Import from File**.



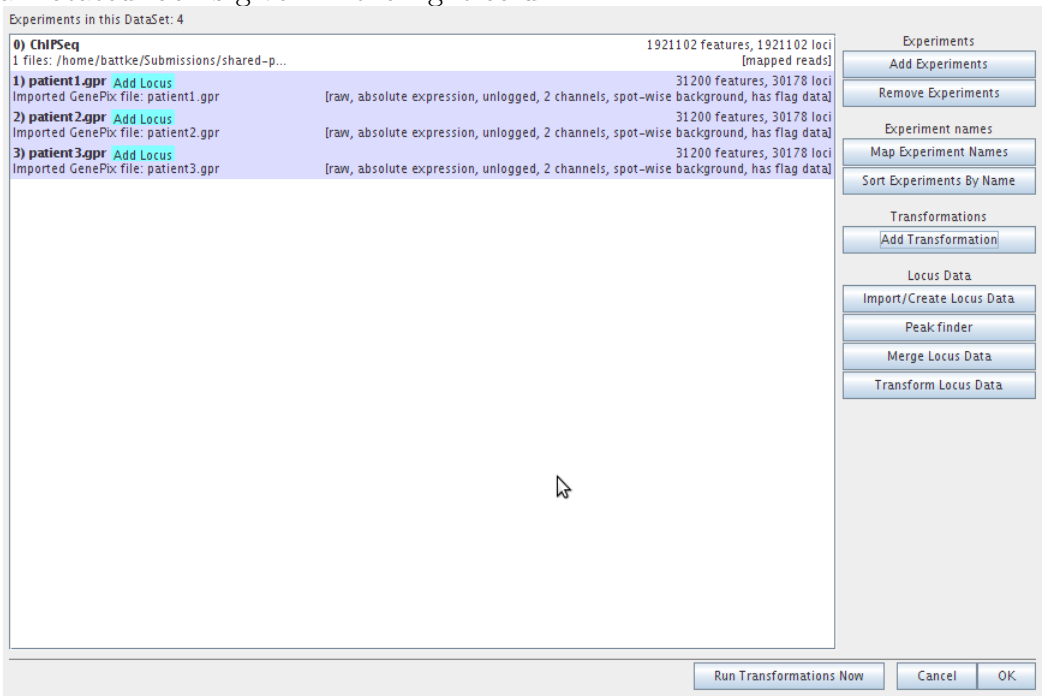
Select **From tabular file**, click on **Browse** and select the example file `coordinates.tsv`. Click **OK**.



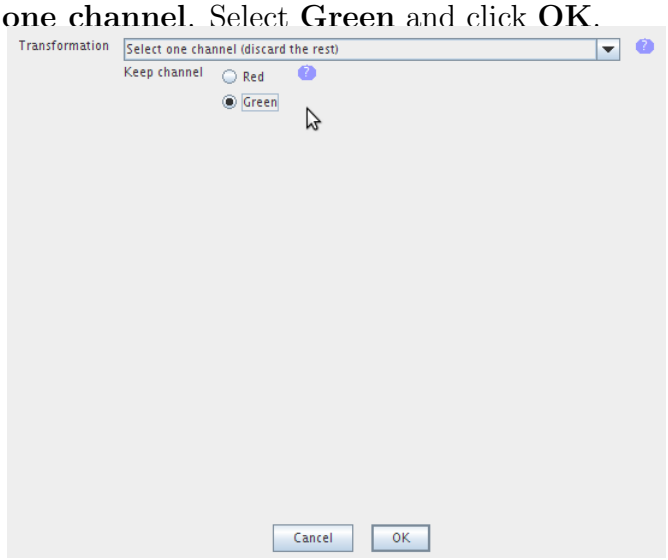
The parser settings from before will be re-used and are also correct for this file (tab-separated, with header line). After clicking **OK**, you will have to define the columns of the annotation file. MAYDAY tries to automatically determine the columns' contents, all you have to do is set the rightmost column to "Length" and provide the species name ("hsa") in the input field at the top of the window. Click **OK** twice to use this annotation.



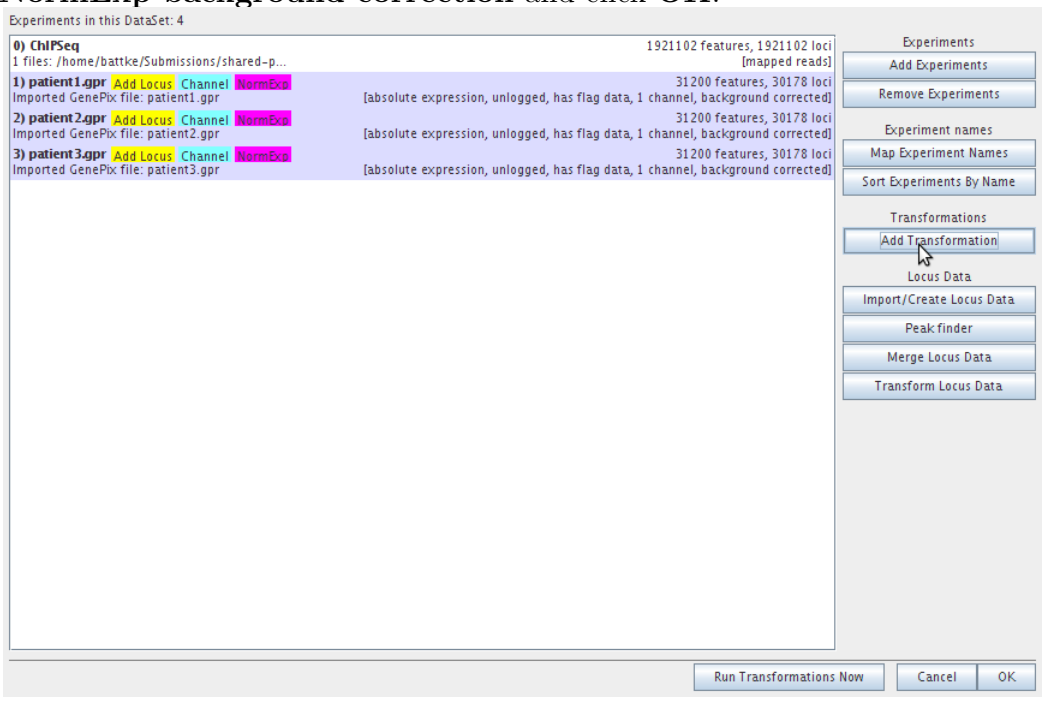
We have now added the locus annotation. This is reflected by the "Add locus" box in each experiment's transformation list. Furthermore, the number of annotated loci is given in the right column.



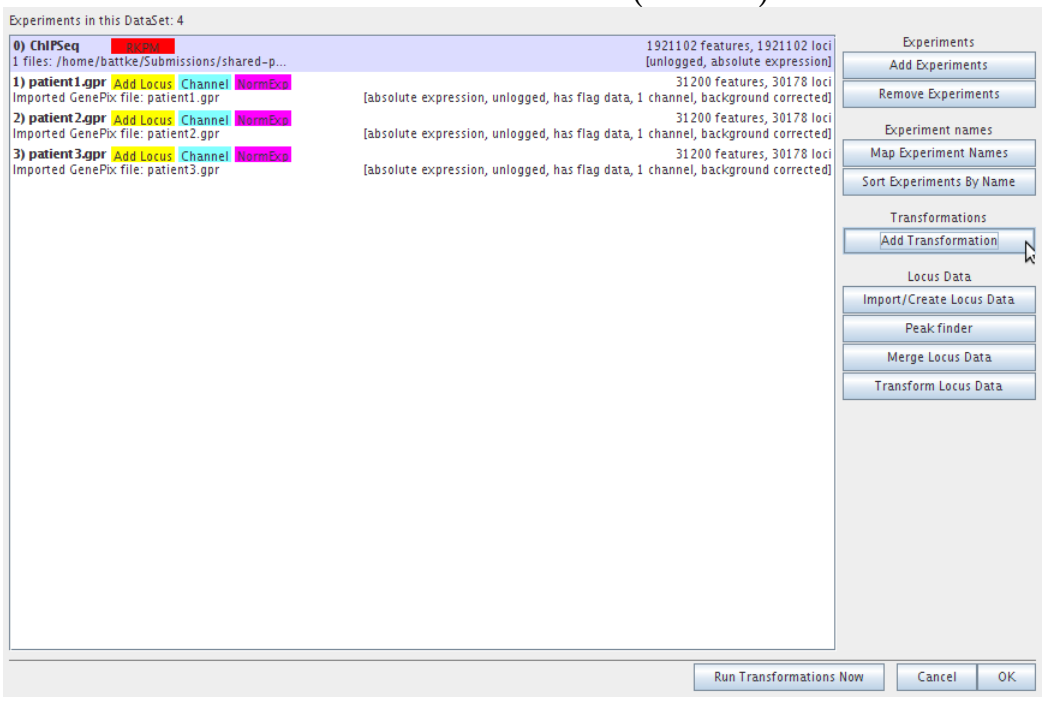
We now add further transformations to the array experiments. The ChIP sample was analyzed in the green channel of the arrays, the red channel is of no interest to us for this experiment. Thus, click **Add transformation** and select **Select one channel**. Select **Green** and click **OK**.



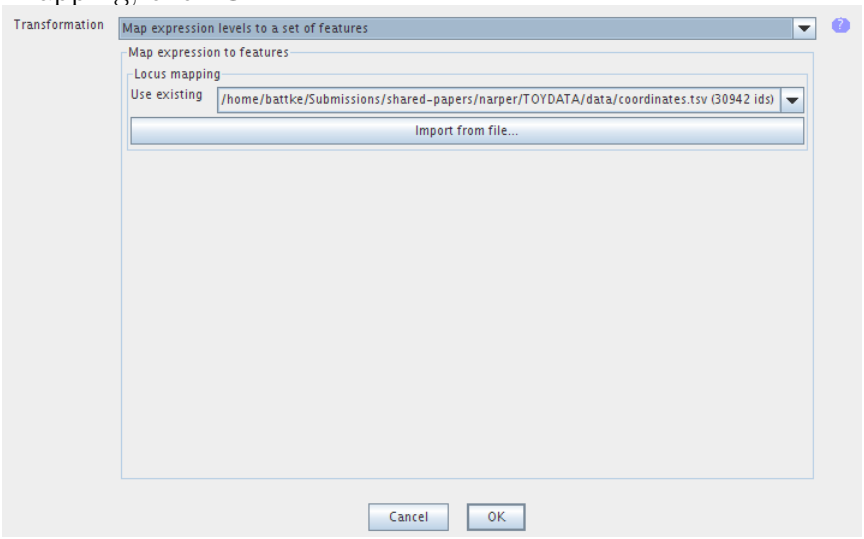
To add a background correction step, again click **Add transformation**, select **NormExp background correction** and click **OK**.



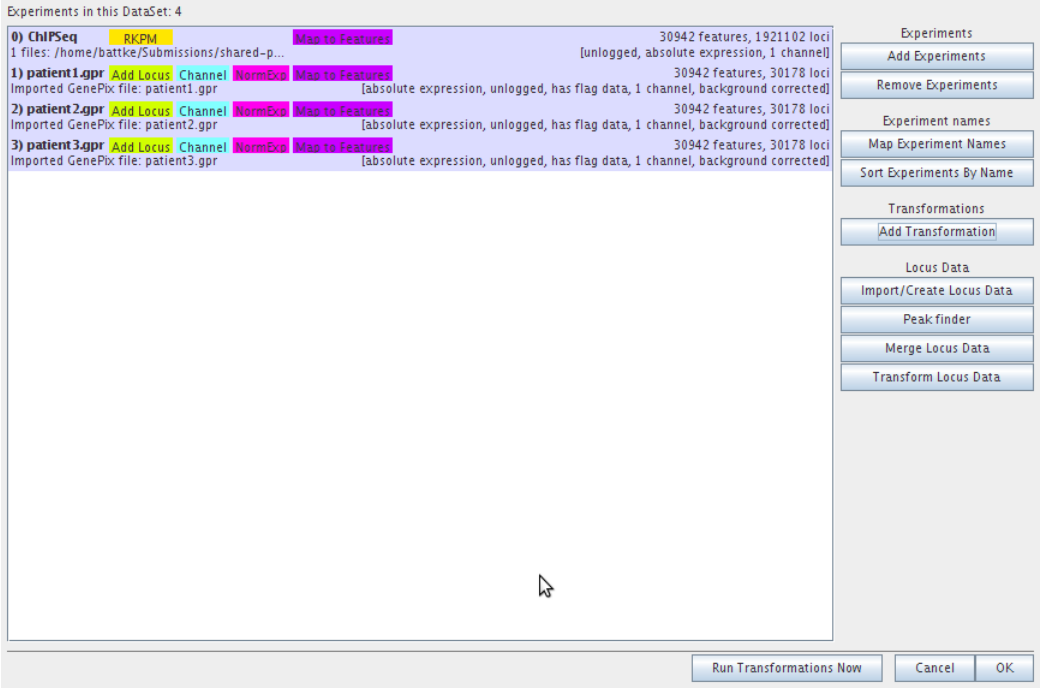
We now turn to the ChIP-seq data. Currently all we have are reads mapped to the genome. To compute an "expression" value from this data, we will use the RPKM measure[2]. Select the ChIP-Seq experiment, click **Add transformation** and select **Combine Read Counts (RPKM)**. Click **OK**.



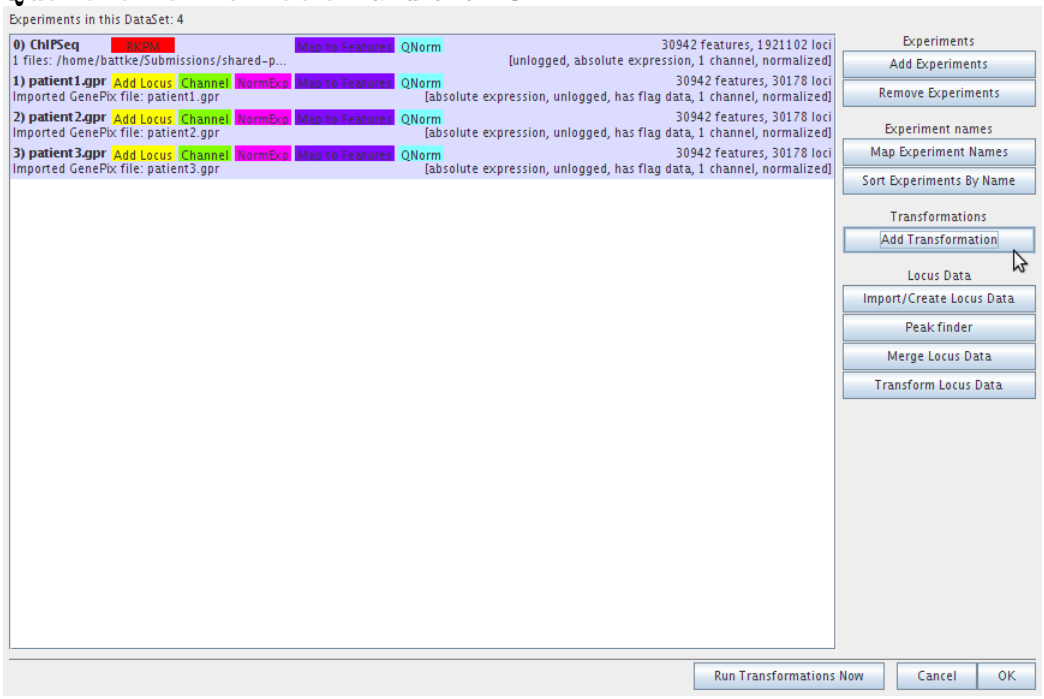
We now have all the data needed to combine the different experiment types. Select all four experiments, click *Add transformation* and select **Map expression levels to a set of features**. MAYDAY will already have selected the correct mapping, click **OK**.



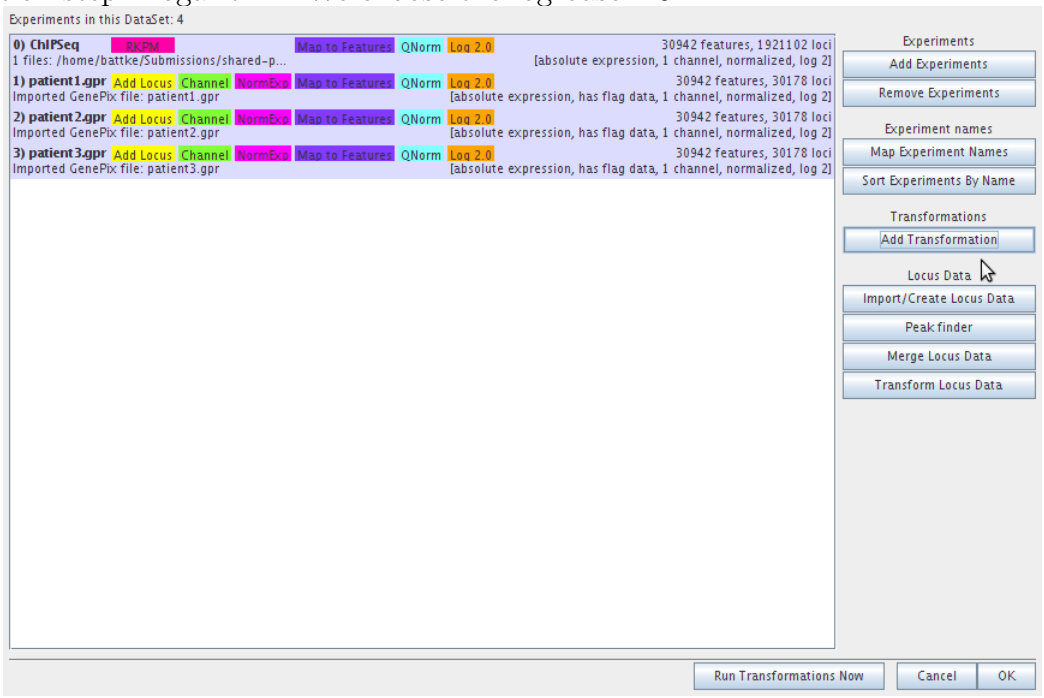
Now all experiments cover the same set of features. However, the values have a very different distribution (not least because of the caveat discussed above).



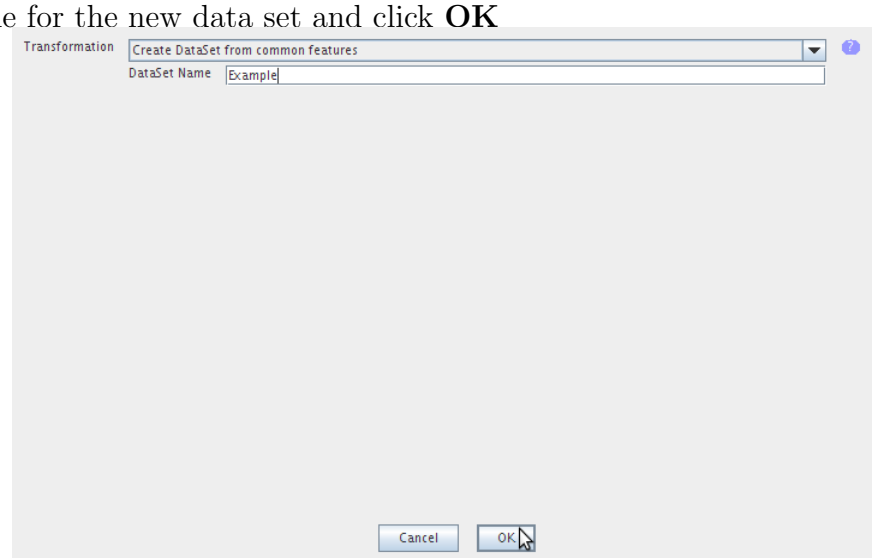
Quantile normalization is a simple way to reconcile the different distributions of the data. With all experiments selected, click **Add transformation**, select **Quantile normalization** and click **OK**.



The data is still not log-transformed, so we need to add a further transformation step: Logarithm. We choose the log base 2.0.

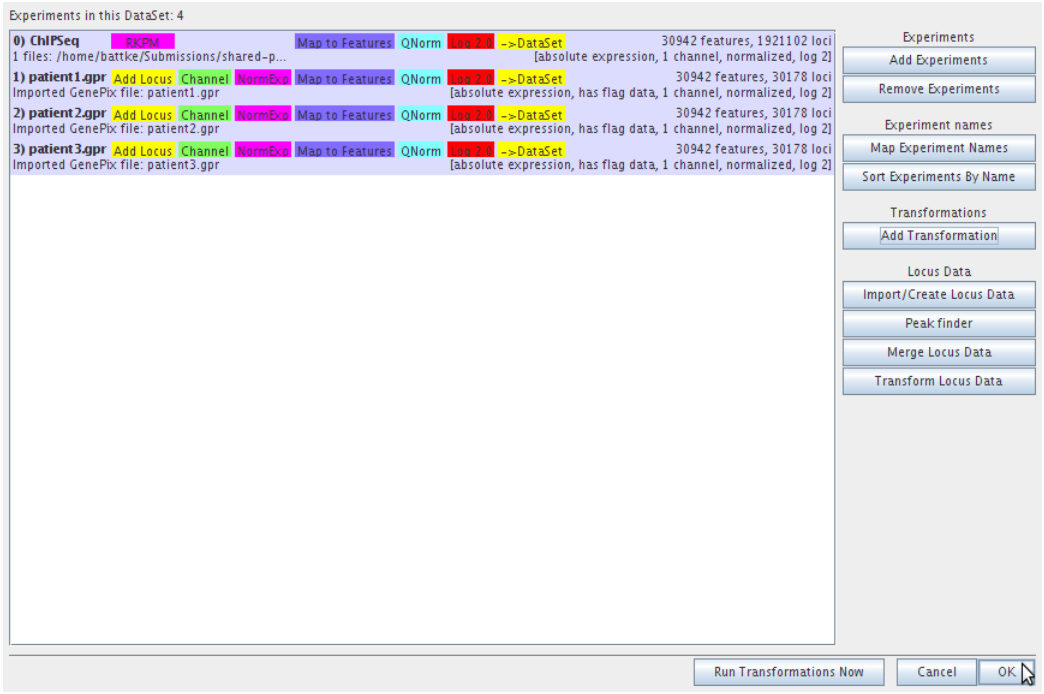


Finally we can bring the data into MAYDAY for further analysis. Click **Add transformation** and select **Create dataset from common features**. Enter a name for the new data set and click **OK**

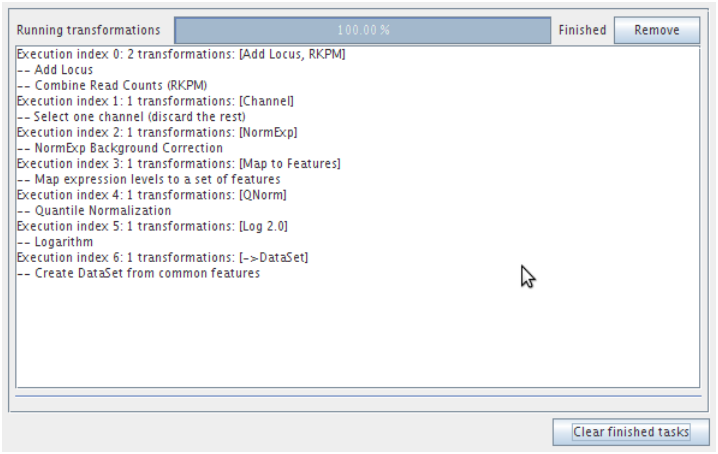


.

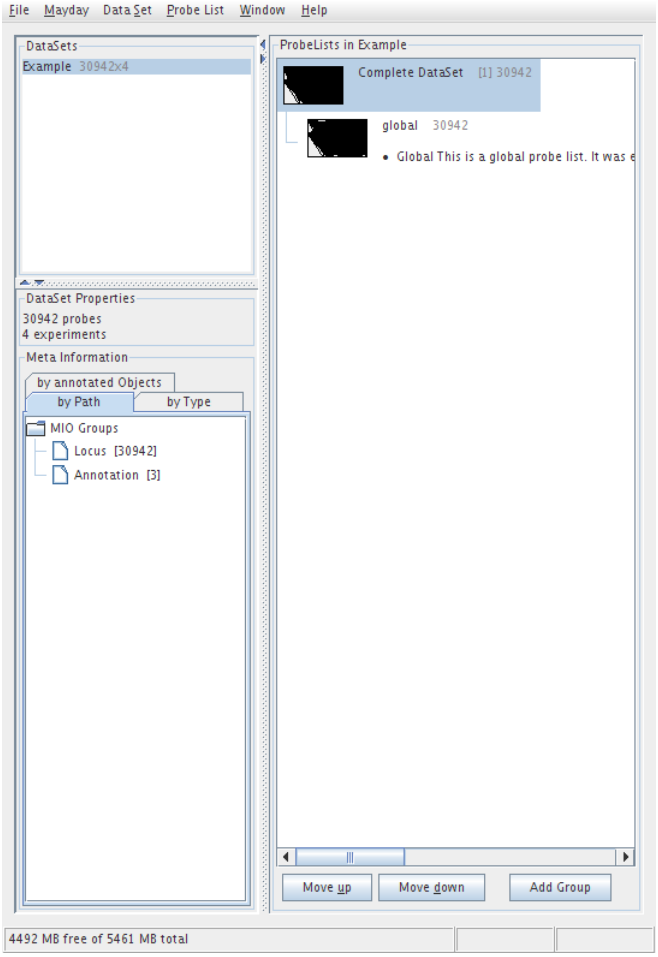
Click **OK** to import the data into MAYDAY.



After a few moments, the import will be done.



You can now work on the data in MAYDAY.



References

[1] R Li, Y Li, K Kristiansen, and J Wang. Soap: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, Mar 2008.

[2] A Mortazavi, B A Williams, K McCue, L Schaeffer, and B Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628, Jul 2008.