

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This example uses data (unpublished) of a chromatin immunoprecipitation experiment (ChIP) performed on samples from human cancer patients.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.

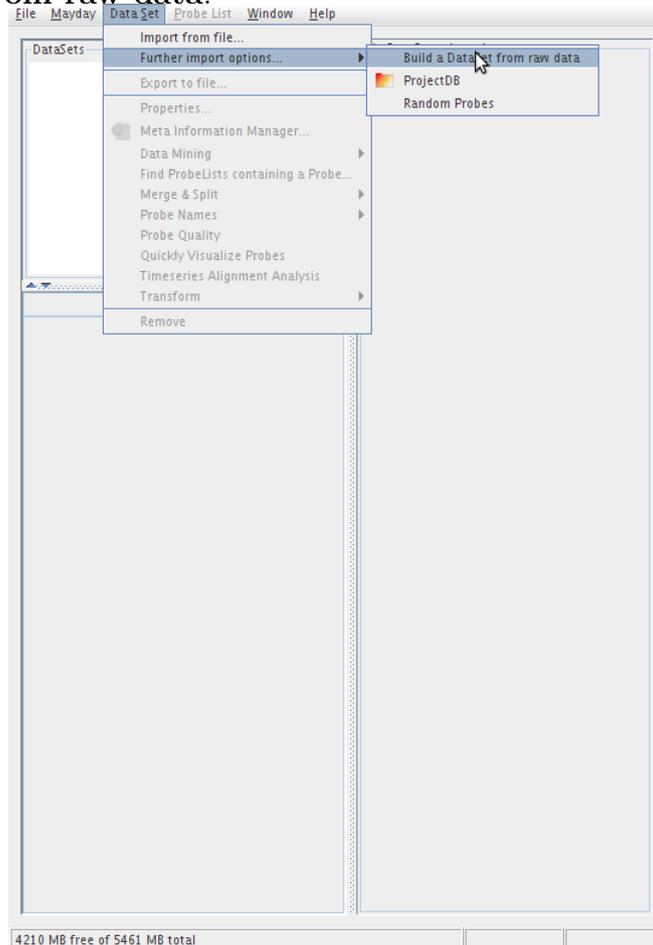
4 Starting Mayday

Start MAYDAY¹ 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).

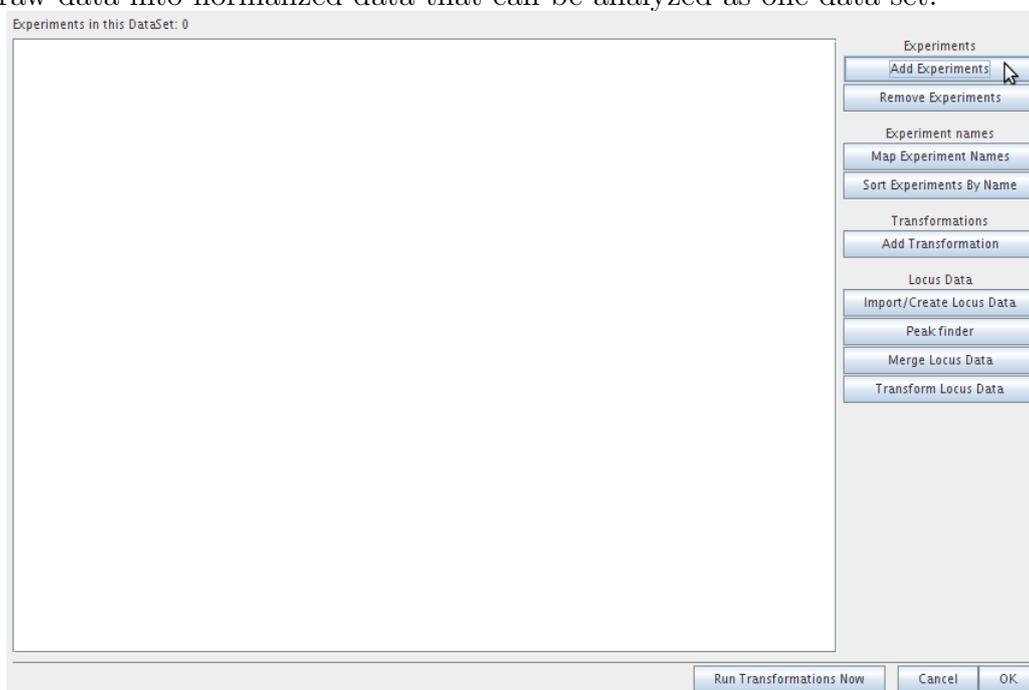
5 Importing the data

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

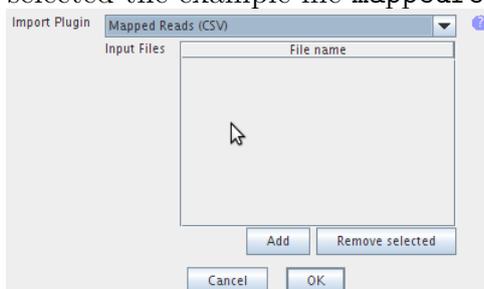


¹Use the experimental version, not the stable build.

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**.

Skip

Skip Lines (before header)

Has Header Line

Comment Characters

Separators

Tabulator Comma

Semicolon Space

Other (regular expression)

Quotes

Content

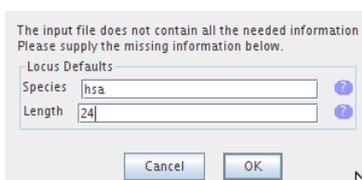
Strand	Chromosome	Start_Position
+	1	154070063
+	1	28647861
+	1	120061385
+	1	7795274
-	1	152264021
+	1	11879035
-	1	187644267
-	1	153990937
+	1	198192662
+	1	56571810
-	1	56810302
-	1	218294567
-	1	235235939
-	1	231335259

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

Strand	Chromosome	Start_Position
<input type="radio"/> Ignore	<input type="radio"/> Ignore	<input type="radio"/> Ignore
<input type="radio"/> Species	<input type="radio"/> Species	<input type="radio"/> Species
<input type="radio"/> Chromosome	<input checked="" type="radio"/> Chromosome	<input type="radio"/> Chromosome
<input checked="" type="radio"/> Strand	<input type="radio"/> Strand	<input type="radio"/> Strand
<input type="radio"/> From	<input type="radio"/> From	<input checked="" type="radio"/> From
<input type="radio"/> To	<input type="radio"/> To	<input type="radio"/> To
<input type="radio"/> Length	<input type="radio"/> Length	<input type="radio"/> Length

Strand	Chromosome	Start_Position
+	1	154070063
+	1	28647861
+	1	120061385
+	1	7795274
-	1	152264021
+	1	11879035
-	1	187644267
-	1	153990937
+	1	198192662

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**.



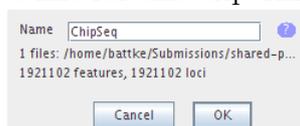
The input file does not contain all the needed information
Please supply the missing information below.

Locus Defaults

Species ?

Length ?

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

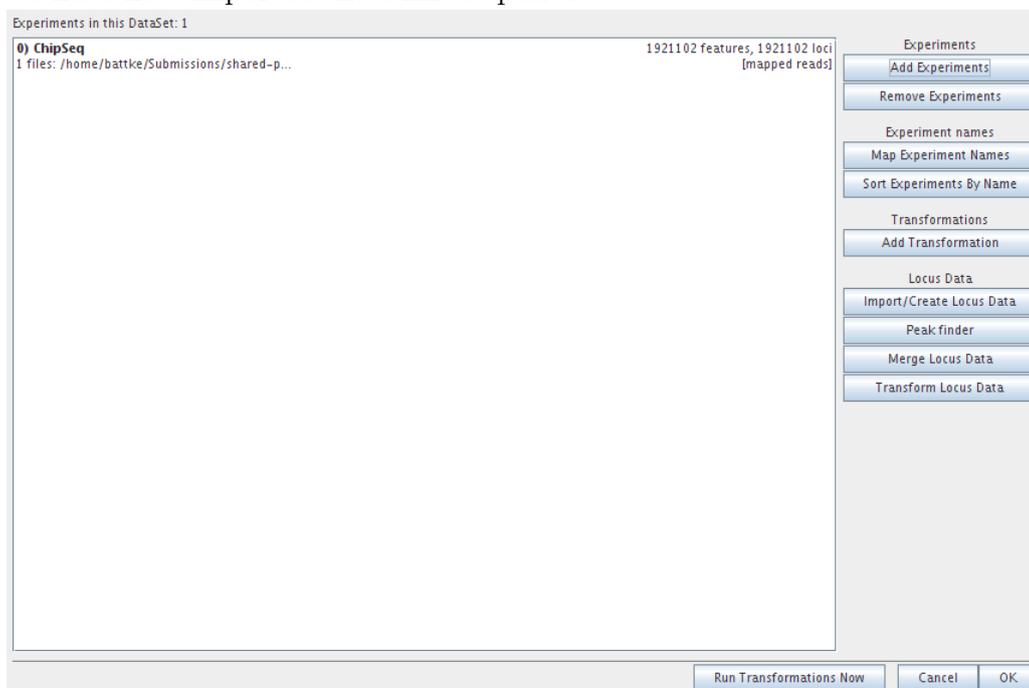


Name ?

1 files: /home/battke/Submissions/shared-p...

1921102 features, 1921102 loci

We have now imported the ChIP-seq data.



Experiments in this DataSet: 1

0 **ChIPseq** 1921102 features, 1921102 loci
1 files: /home/battke/Submissions/shared-p... [mapped reads]

Experiments

-
-

Experiment names

-
-

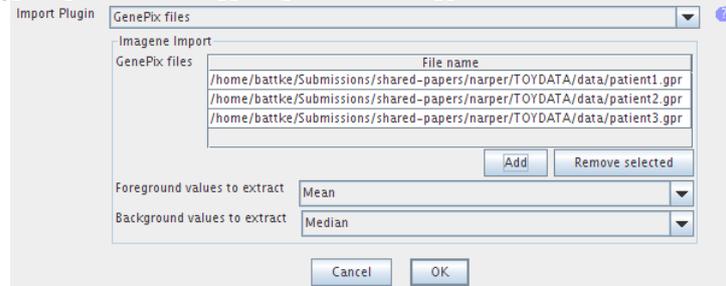
Transformations

-

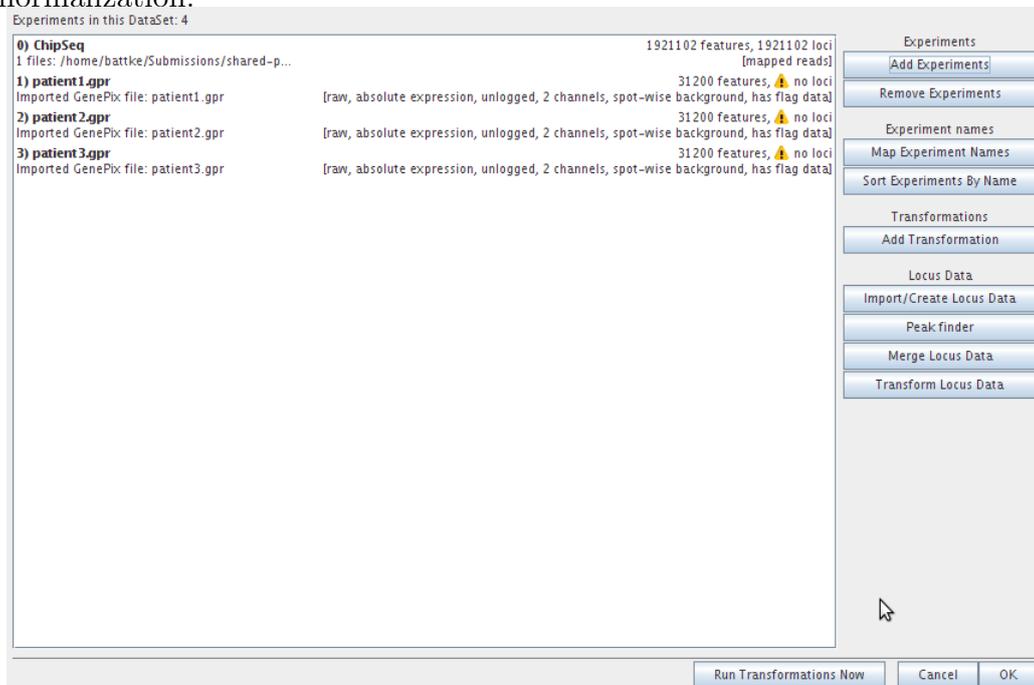
Locus Data

-
-
-
-

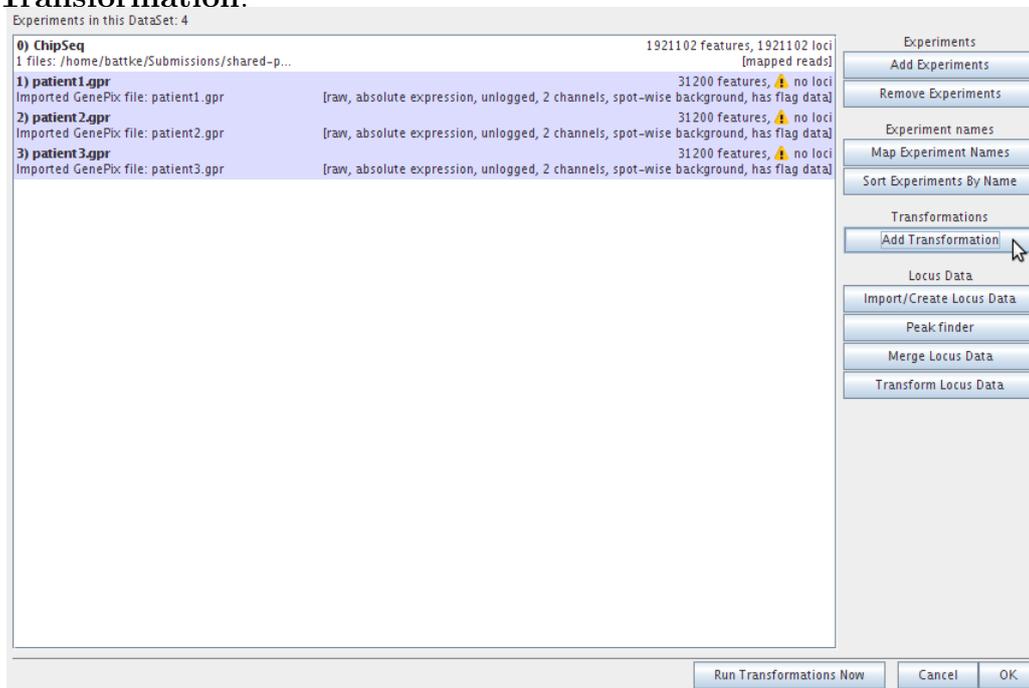
To import the ChIP-chip data, again click **Add Experiments**. Select **GenePix** from the drop-down list, click **Add** and select the example files `patient1.gpr`, `patient2.gpr`, `patient3.gpr`. Click **OK**.



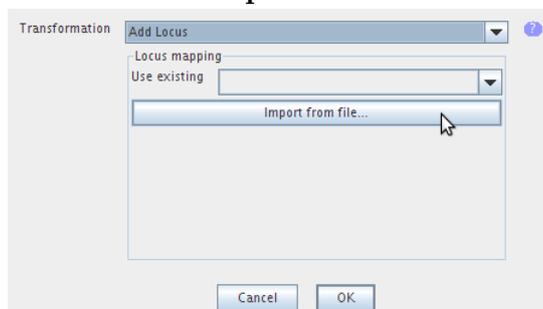
We have now imported the ChIP-chip data. In the next steps, we will add transformations to the data to make the experiments compatible and to apply normalization.



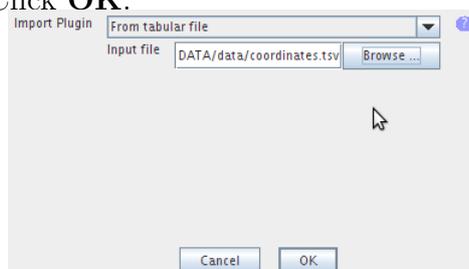
We first add the locus annotation to the ChIP-chip experiments. select the three patients (click on patient1, then shift-click on patient3). Then click **Add Transformation**.



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.

Locus Defaults

Species:

Chromosome:

Strand:

Length:

Seq_Name	probe_chr	probe_strand	real_start_pos	probe_length
<input type="radio"/> Ignore	<input type="radio"/> Ignore	<input type="radio"/> Ignore	<input type="radio"/> Ignore	<input type="radio"/> Ignore
<input checked="" type="radio"/> Identifier	<input type="radio"/> Identifier	<input type="radio"/> Identifier	<input type="radio"/> Identifier	<input type="radio"/> Identifier
<input type="radio"/> Species	<input type="radio"/> Species	<input type="radio"/> Species	<input type="radio"/> Species	<input type="radio"/> Species
<input type="radio"/> Chromosome	<input checked="" type="radio"/> Chromosome	<input type="radio"/> Chromosome	<input type="radio"/> Chromosome	<input type="radio"/> Chromosome
<input type="radio"/> Strand	<input type="radio"/> Strand	<input checked="" type="radio"/> Strand	<input type="radio"/> Strand	<input type="radio"/> Strand
<input type="radio"/> From	<input type="radio"/> From	<input type="radio"/> From	<input checked="" type="radio"/> From	<input type="radio"/> From
<input type="radio"/> To	<input type="radio"/> To	<input type="radio"/> To	<input type="radio"/> To	<input type="radio"/> To
<input type="radio"/> Length	<input type="radio"/> Length	<input type="radio"/> Length	<input type="radio"/> Length	<input checked="" type="radio"/> Length
CM_000001	18	+	45594391	53
CM_000002	2	+	233966372	54
CM_000003	11	+	93094328	54
CM_000004	14	-	95069329	54
CM_000005	3	-	161715422	50
CM_000006	2	+	233979322	49
CM_000007	12	+	6489649	54
CM_000008	15	-	64426546	54
CM_000009	12	-	6560856	54
CM_000010	1	+	109354857	45
CM_000011	1	-	152708794	54
CM_000012	9	-	19053654	54
CM_000013	1	+	27845054	54
CM_000014	1	+	109355197	52
CM_000015	12	-	6946988	46
CM_000016	17	+	72596984	54

Cancel OK

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.

The screenshot shows a software interface with a main window titled "Experiments in this DataSet: 4". The main window contains a list of experiments:

Experiment Name	Files	Features	Loci
0) ChIPSeq	1 files: /home/battke/Submissions/shared-p...	1921102 features	1921102 loci [mapped reads]
1) patient1.gpr	Imported GenePix file: patient1.gpr	31200 features	30178 loci
2) patient2.gpr	Imported GenePix file: patient2.gpr	31200 features	30178 loci
3) patient3.gpr	Imported GenePix file: patient3.gpr	31200 features	30178 loci

Each experiment entry includes an "Add Locus" link and a description of the data format: "[raw, absolute expression, unlogged, 2 channels, spot-wise background, has flag data]".

On the right side, there is a sidebar with several sections of buttons:

- Experiments:** Add Experiments, Remove Experiments
- Experiment names:** Map Experiment Names, Sort Experiments By Name
- Transformations:** Add Transformation
- Locus Data:** Import/Create Locus Data, Peak finder, Merge Locus Data, Transform Locus Data

At the bottom of the main window, there are three buttons: "Run Transformations Now", "Cancel", and "OK".

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**.

The screenshot shows a "Transformation" dialog box with the following content:

- Transformation: Select one channel (discard the rest)
- Keep channel:
 - Red
 - Green

At the bottom of the dialog box, there are two buttons: "Cancel" and "OK".

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

The screenshot shows a software interface with a main window titled "Experiments in this DataSet: 4". The main window contains a list of experiments:

Experiment	Files	Features	Loci	Description
0) ChIPSeq	1 files: /home/battke/Submissions/shared-p...	1921102 features	1921102 loci	[mapped reads]
1) patient1.gpr	Imported GenePix file: patient1.gpr	31200 features	30178 loci	[absolute expression, unlogged, has flag data, 1 channel, background corrected]
2) patient2.gpr	Imported GenePix file: patient2.gpr	31200 features	30178 loci	[absolute expression, unlogged, has flag data, 1 channel, background corrected]
3) patient3.gpr	Imported GenePix file: patient3.gpr	31200 features	30178 loci	[absolute expression, unlogged, has flag data, 1 channel, background corrected]

On the right side, there is a sidebar with several sections of buttons:

- Experiments:** Add Experiments, Remove Experiments
- Experiment names:** Map Experiment Names, Sort Experiments By Name
- Transformations:** Add Transformation (highlighted with a mouse cursor)
- Locus Data:** Import/Create Locus Data, Peak finder, Merge Locus Data, Transform Locus Data

At the bottom of the main window, there are three buttons: "Run Transformations Now", "Cancel", and "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**.

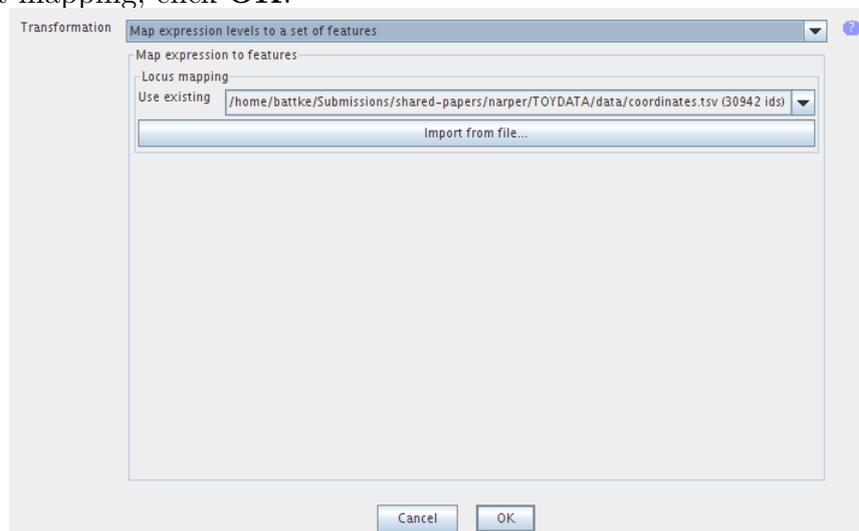
This screenshot is similar to the previous one, but the "0) ChIPSeq" experiment is now selected (highlighted in blue). The description for this experiment has changed to "[unlogged, absolute expression]". The "Add Transformation" button in the sidebar is also highlighted with a mouse cursor.

The main window still shows the same list of experiments as in the previous screenshot.

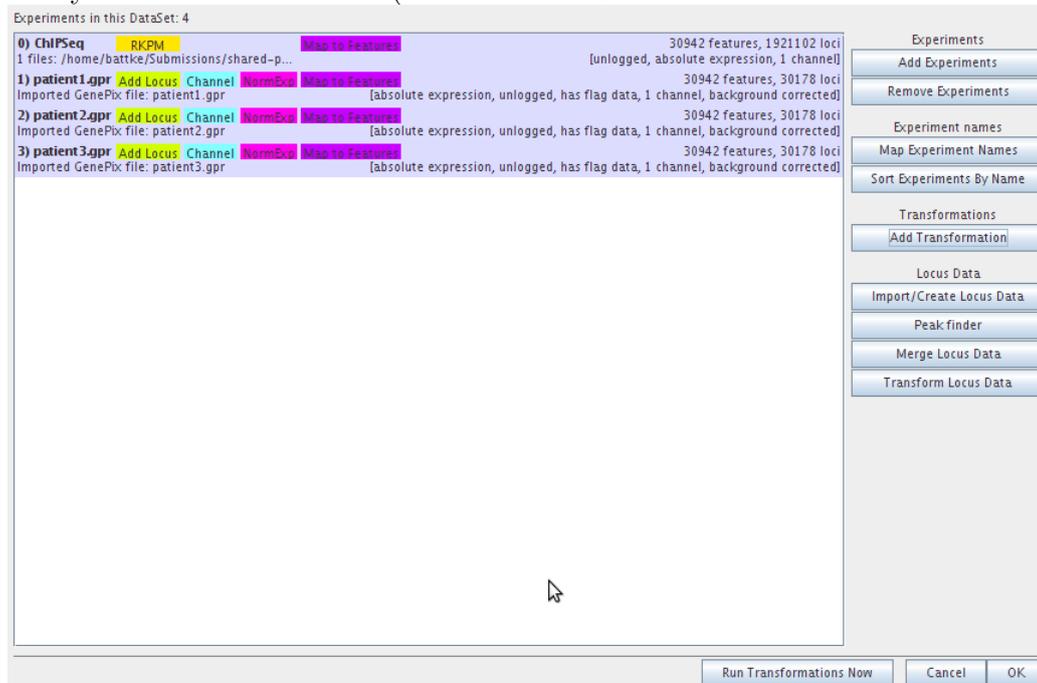
The sidebar buttons remain the same, with "Add Transformation" being the active one.

The bottom buttons "Run Transformations Now", "Cancel", and "OK" are still present.

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 screenshot shows a software interface with a list of experiments on the left and a sidebar on the right. The list contains four experiments:

- 0) ChIPSeq [unlogged, absolute expression, 1 channel, normalized] 30942 features, 1921102 loci
- 1) patient1.gpr [absolute expression, unlogged, has flag data, 1 channel, normalized] 30942 features, 30178 loci
- 2) patient2.gpr [absolute expression, unlogged, has flag data, 1 channel, normalized] 30942 features, 30178 loci
- 3) patient3.gpr [absolute expression, unlogged, has flag data, 1 channel, normalized] 30942 features, 30178 loci

The sidebar on the right has the following sections:

- Experiments: Add Experiments, Remove Experiments
- Experiment names: Map Experiment Names, Sort Experiments By Name
- Transformations: Add Transformation (highlighted by a mouse cursor)
- Locus Data: Import/Create Locus Data, Peak finder, Merge Locus Data, Transform Locus Data

At the bottom of the interface are buttons for "Run Transformations Now", "Cancel", and "OK".

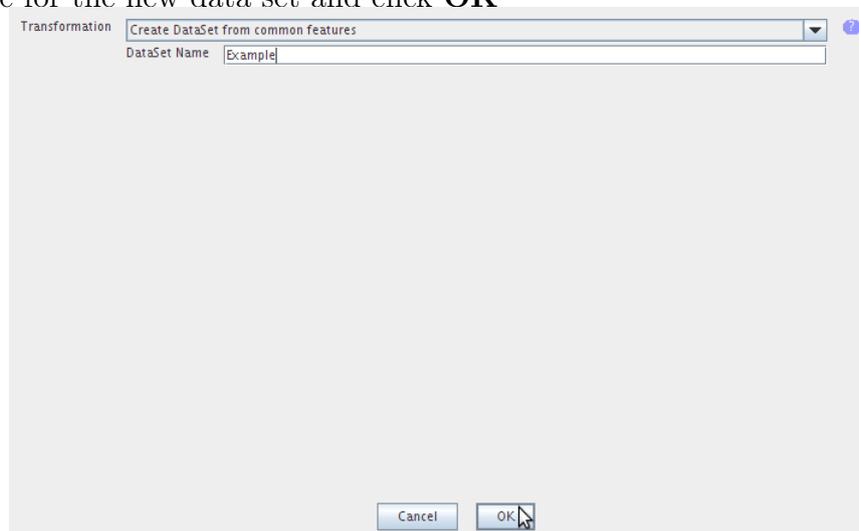
The data is still not log-transformed, so we need to add a further transformation step: Logarithm. We choose the log base 2.0.

This screenshot is similar to the previous one, but the transformation list for each experiment now includes "Log 2.0".

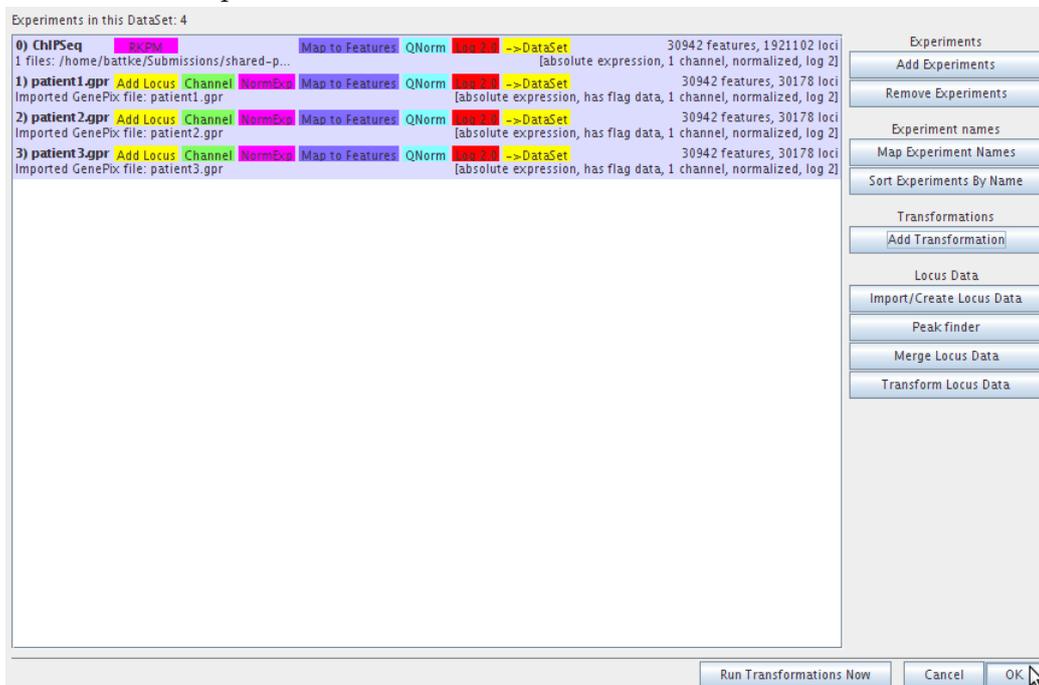
- 0) ChIPSeq [absolute expression, 1 channel, normalized, log 2] 30942 features, 1921102 loci
- 1) patient1.gpr [absolute expression, has flag data, 1 channel, normalized, log 2] 30942 features, 30178 loci
- 2) patient2.gpr [absolute expression, has flag data, 1 channel, normalized, log 2] 30942 features, 30178 loci
- 3) patient3.gpr [absolute expression, has flag data, 1 channel, normalized, log 2] 30942 features, 30178 loci

The sidebar and bottom buttons remain the same as in the previous screenshot.

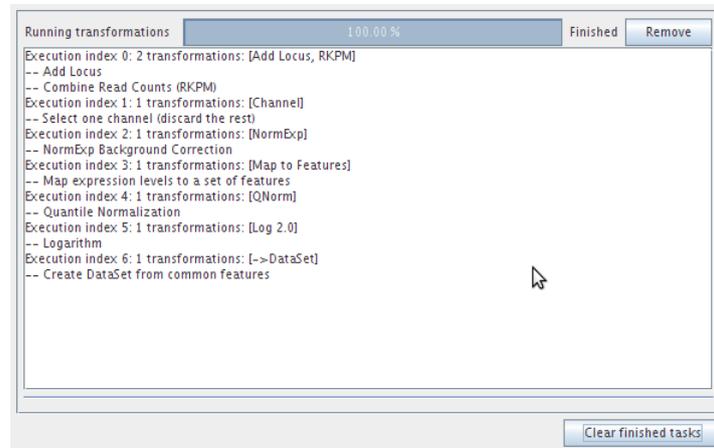
Finally we can bring the data into MAYDAY for further analysis. Click **Add transformation** and select **Create dataset from common features**. Enter a name for the new data set and click **OK**



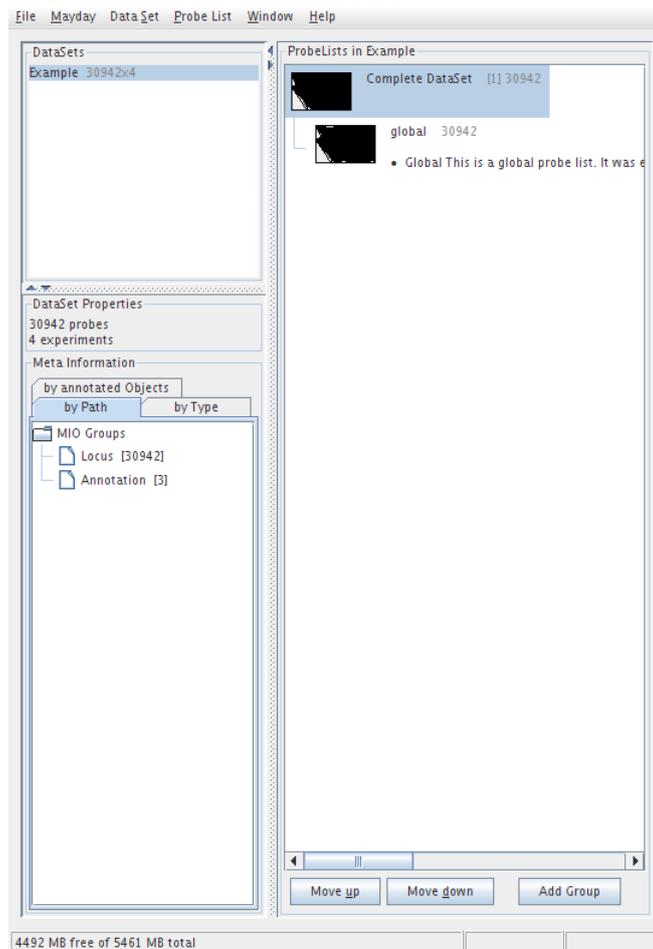
Click **OK** to import the data into MAYDAY.



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.