Hannover Medical School

RCU Transcriptomics

Research Core Unit Transcriptomics

The Excel standard result file for single-color studies

Version: 15.01.2013

Introduction

Within our Core Unit, we exclusively utilize the microarray platform of Agilent technologies. Raw data of hybridized and scanned microarrays are extracted, quality-controlled and preprocessed by Agilent's Feature Extraction software. All of the thereby produced data are integrated into a text-file format by default, which is referred to as "Feature Extraction Result Text File". This file can be imported into downstream data analysis programs (e.g. GeneSpring, Multi Experiment Viewer, Mayday, …) for subsequent data transformation, normalization, analysis and visualization. However, as long as a given study design is not too complex, we also consider Excel as a reasonable data analysis tool, at least for some basal filtering and sorting approaches. Therefore, microarray data are routinely integrated into standardized Excel file formats by us. These result files represent the study's data in a well-arranged format. We selected the most informative part of data initially acquired per gene in order to reduce complexity. Additionally, adequate sorting keys and supplemental information are integrated to facilitate navigation through the complex data. Even scientists, un-experienced with microarray data should get enabled to gain a first impression on most prominent gene expression changes and overall data reliability.

The Excel standard result file for single-color studies

Our standardized Excel file format for single-color studies is named “MXXX1-MXXX4_SC_ DataExtractShort_InterArrayNorm_SurrogateUsed_SortKeysIncluded.xls", where MXXX1 to MXXX4 denote the consecutive M-numbers of the arrays given in our lab (Figure 1). Note, that the Agilent microarray slides utilized for mRNA expression profiling, contain 4 identical microarrays that are loaded independently with 4 different samples, giving rise to 4 data sets, finally integrated into one Excel result file.

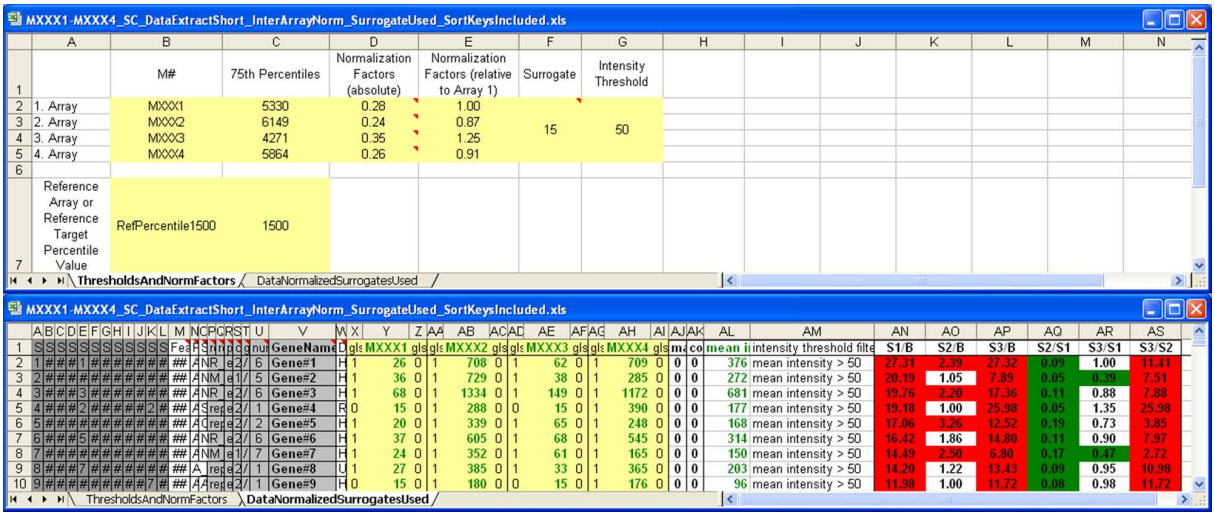

Figure 1: The two sheets of the standardized Excel file: “ThresholdsAndNormFactors” (upper panel) and “DataNormalizedSurrogatesUsed” (lower panel).

Data sheet

The “DataNormalizedSurrogatesUsed” data sheet is divided into five data blocks with sorting keys in the first, probe and gene annotation data in the second, microarray data in the third block, summarizing information in the 4th and calculated ratios in the 5th block (Figure 2).

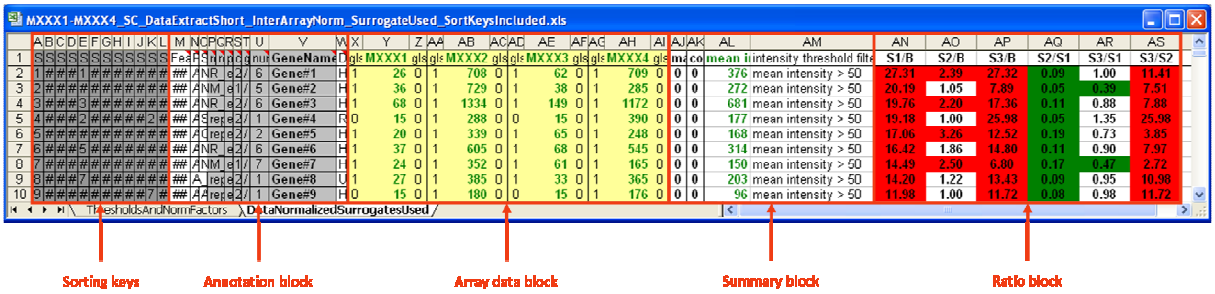

Figure 2: The “DataNormalizedSurrogatesUsed” sheet is subdivided in five blocks.

The Array data block

The array data block (Figure 2) summarizes the most important part of data of the 4 microarrays each of which represented by 3 carefully selected columns. These columns contain:

- an entry, that specifies, if a signal is above background noise “gIsWellAboveBG”
- the processed normalized fluorescence intensity values
  “MXXXX gProcessedSignal_NormalizedTo75thPercentileOfRefPercentile1500_Surrogate15Used”
- a flag, indicating technical performance of the respective spot/feature
  “gIsFeatNonUnifOL” which stands for: green Is Feature a Non Uniformity OutLier).”

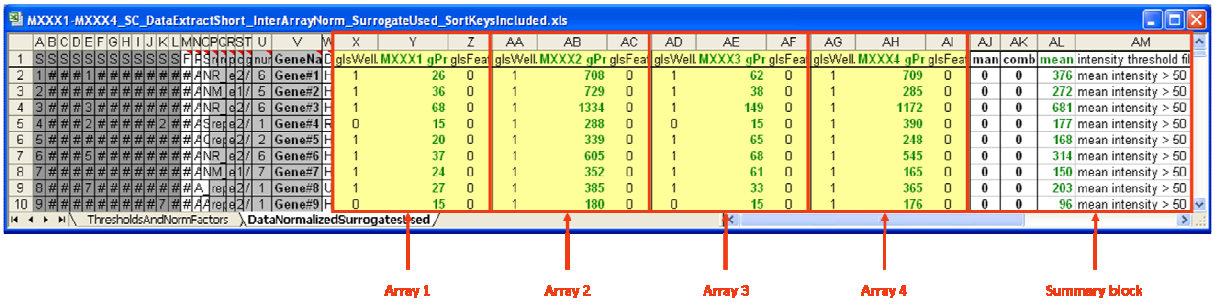

Figure 2: Array data block and summary block

The column “gIsWellAboveBG” indicates if a feature’s intensity is significant above background noise and is set to 1 in that case.

The fluorescence intensity values “gProcessedSignal_NormalizedTo75thPercentileOfRefPercentile1500_Surrogate15Used” originate from the Feature Extraction Result Text File but are subjected to inter-array normalization by global linear scaling and surrogate usage prior to the integration into our Excel file format. The 75th percentile of the intensity distribution of each array (and dataset) is fitted to the reference value of 1500 by linear scaling (see also Excursion: Normalization). The individually calculated normalization factors for all microarrays can be inspected on the “ThresholdsAndNormFactors” data sheet (Figure 7). Such global scaling approaches are feasible, whenever the number of differentially expressed genes is low compared to the number of genes measured in total.

Excursion: Normalization and Surrogate Usage in single-color mode

The fluorescence intensity values originate from the Feature Extraction Result Text File but are subjected to inter-array normalization by global linear scaling and surrogate usage prior to the integration into our Excel file format. For this, processed intensity values of the green channel ("gProcessedSignal" or "gPS") are globally normalized by a linear scaling approach: All gPS values of one sample are multiplied by an array-specific scaling factor. This scaling factor is calculated by dividing a "reference 75th Percentile value" (set to 1500 for the whole series) by the 75th Percentile value of the particular microarray ("Array i" in the formula shown below). Accordingly, normalized gPS values for all samples (microarray data sets) are calculated by the following formula:

$$\text{normalized } gPS_{Array\ i} = gPS_{Array\ i} \times (1500 / 75^{th}\ Percentile_{Array\ i})$$

A lower intensity threshold was defined as 1% of the reference 75th Percentile value (= 15). All of the normalized gPS values that fall below this intensity border, are substituted by the respective surrogate value of 15.

The flag column indicates technical performance of individual spots (features). If the uniformity of pixel signals of a spot is high, a 0 is given, otherwise, if there might be a spurious impairment this is documented by the entry 1.

Summary block

The entries in the summary block (Figure 2) give summarized information about the arrays as follows:

- "manual flags"
- "combined flag"
- "mean intensity" over the four arrays
- an entry if the mean intensity is above 50 "intensity threshold filter: mean intensity > 50"

The "manual flags" column is a relict of earlier versions and is not used anymore.

The "combined flag" column summarizes the "gIsFeatNonUnifOL" columns of the 4 arrays.

The "mean intensity" column contains the arithmetic mean over all 4 intensity values of a certain probe.

The last column of this block “intensity threshold filter: mean intensity > 50” is filled accordingly if the mean intensity across all 4 microarrays exceeds a threshold value of 50.

Ratio block

The ratio block (Figure 3) contains columns for all possible ratios that can be calculated from the 4 individual microarrays (datasets). The 4 arrays are represented by the following abbreviations:

- B = array 1
- S1 = array 2
- S2 = array 3
- S3 = array 4

B stands for ‘basal’ and S for ‘stimulated’. This notation has its rationale in a standard study design where three stimulated samples (S1 to S3) are compared to one basal reference sample (B). Anyway, in cases where the 4 analyzed samples do not match this interpretation, the 4 abbreviations should simply be regarded to explicitly define the position of a sample within our Excel file and to specify, which arrays’ intensities have been used to fill a specific ratio column.

The ratios are colored in red if the value is higher than or equal to 2. Ratios of 0.5 or below are colored in green. Thus, the given color code allows for an intuitive inspection of results.

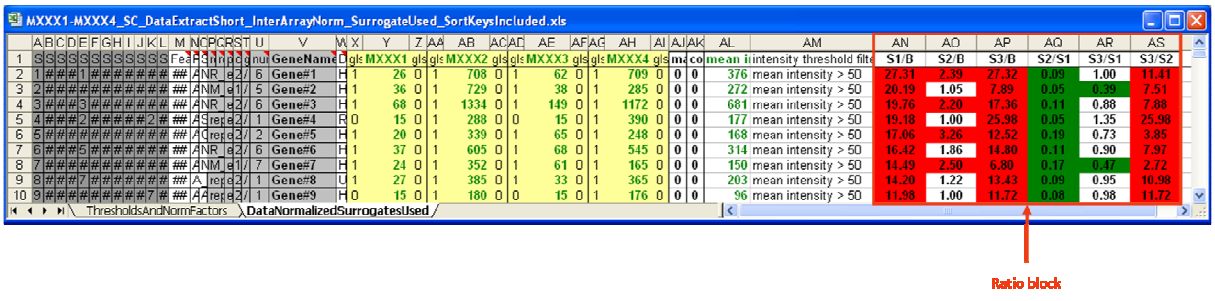

Figure 3: The ratio block

The Annotation block

The annotation block provides the following information on each feature:

- the feature number “FeatureNum” (location of the feature on the array)
- the unique probe identifier “ProbeName”
- the RefSeq accession number “SystematicName”

- "representative 2"
- "representative 3"
- the "probe type"
- a "crude probe characterization"
- a "gender-specific marker"
- the number of probes present on the microarrays allocated to a given gene name
- the gene name "GeneName"
- the gene "description"

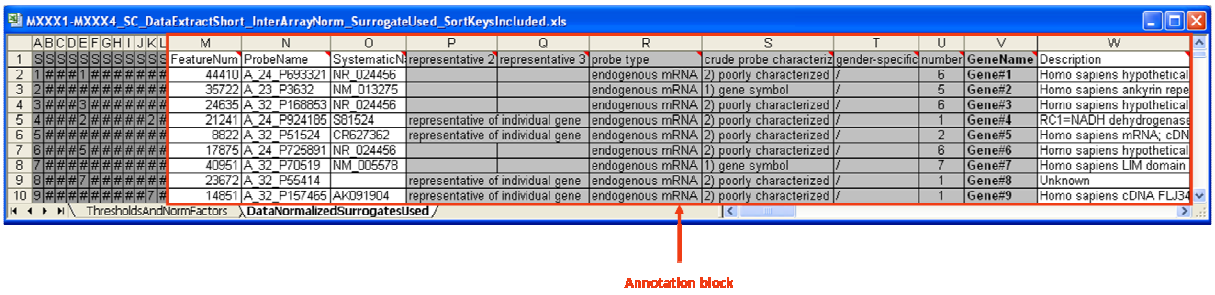

Figure 5: The annotation block

The columns "FeatureNum", "ProbeName", "SystematicName" and "Description" (Figure 5 in white) are taken over unchanged from the regularly updated annotation files provided by Agilent. The additional columns depicted in grey are generated by our own algorithms but also originate from Agilent's annotation files.

Each gene is represented by either one or several probes on the microarray. If the number of probes is exactly 10 then these 10 probes are identical (on-chip replicate probes). The information on the number of probes per gene is given in column Q: "number of probes present on microarrays allocated to individual gene name (symbol) entry"). Genes with less than 10 allocated probes are represented by distinct probes, whereas more than 10 probes per gene indicate a representation with 10 identical and additional distinct ones. Please note, that intensity measurements should be highly similar among identical probes. If this doesn't hold true, respective outlier features presumably indicate a locally restricted impairment in hybridization performance or (much more rarely) in spot quality.

Excursion: The concept of probe selection for mRNA expression arrays by Agilent

Agilent evaluated their mRNA expression microarrays after virtual design of a certain number of different probes per gene. These probes were synthesized, printed and finally evaluated in microarray experiments, using a panel of different tissues and cell systems. This gave rise to specific expression profiles for each gene across all cell systems tested.
If the probes of a gene indicated a common expression pattern, i.e. forming a single cluster, one probe was selected as representative. In the case of multiple clusters, one probe per cluster was selected for the final microarray design. Thus, more than one single probe per gene is contained on the microarray in the latter case. The multiple clusters could be an indication for different splice variants. But be always aware that the microarray types used have not been designed to explicitly discriminate between different splice variants of a gene.
Note, that within our standard Excel result files, the intensity values of probes, targeting identical transcripts are kept non-averaged to allow for a separate inspection of fluorescence intensity measurements.

The "probe type" column discriminates each probe in "endogenous mRNA" and "control", whereas the "crude probe characterization" divides the probes into three classes: "1) gene symbol", "2) poorly characterized" and "3) control". "1) gene symbol" is assigned to probes of genes with a real gene name entry in Agilent's annotation. Probes of genes that are less-well characterized are assigned by "2) poorly characterized" such as ESTs or non-protein coding genes whose function is uncertain. Technical control probes on the array are marked as "3) control".[*]

The columns "representative 2" and "representative 3" offer two shortcuts for data review to get a first impression.

[*] For more detailed information regarding our classification of Agilent probes please consult our manual: 'Crude probe characterization'.

Excursion: "representative 2" and "representative 3" columns

In column "representative 2", just one probe for each gene is pre-selected and marked as "representative of individual gene". This column can be used as a subsequent sorting criterion, if data has been sorted firstly based on a ratio column. While the first sorting orders complete data depending on the strength of differential expression (e.g. ratio value in descending order), it does not give a reliable estimate about the number of regulated genes (e.g. with ratio values above a certain threshold) unless probe replicates contained within this first selection are eliminated. At this stage, a second sorting, using the "representative 2" column in descending order, gives rise to an order, where the representative probe values get to the top of the Excel table, so that the number of regulated genes can be easily determined by counting rows (e.g. by counting all rows starting with a specific threshold ratio up to the top of the Excel table).
The "representative 3" column fulfills the same function as "representative 2", except that only characterized or annotated genes are marked. Thus, sorting complete data according to the entries in this column (in descending order, after a first ratio-based sorting) enables to count the number of rows and to conclude, that there are N well-characterized genes in the list that are constraint by a particular factor showing an X-fold induction.

The selection criteria that determine which probe is assigned to as a "representative" are briefly summarized below.

"representative 2" (rep2):

- *if just one probe per GeneName exists, assign this probe as "rep2"*
- *if 1-9 probes per GeneName exist, assign the probe with the lowest FeatureNum entry as "rep2"*
- *if >= 10 probes per GeneName exist, assign the probe with the lowest FeatureNum entry, that additionally possesses a RefSeq accession entry of the type "NM_*" as "rep2". If no such probe exists, select the probe with the lowest FeatureNum entry as "rep2"*

"representative 3" (rep3):

- *consider all probes, assigned as "rep2" probes and additionally assign those with "rep3" that have been allocated with the attribute "1) gene symbol" during crude probe characterization (see respective manual for details).*

Finally, a gender-specific marker column indicates male-specific transcripts, i.e. all genes expressed from the Y-chromosome, as well as Xist as the only female-specific marker transcript. In studies with heterogeneous (unmatched) tissue samples it should be kept in mind not to interpret gender-specific patterns as putative result of a given treatment.

Sorting keys

To rapidly inspect your data according to various aspects, the data sheet contains predefined sorting keys (Figure 6) – two for each pairwise array comparison (ratio column).

The first sorting key was formerly generated by a 4-way sorting process.

1. Sorting of ratios of the first array comparison (S1/B) descending from strongest up-regulation to strongest down-regulation
2. Sorting by the “intensity threshold filter: mean intensity > 50” column to get features with a mean intensity over all four arrays above a threshold of 50. Thus, features having a mean intensity in the range of background intensities are positioned at the bottom of the table
3. Sorting by the “combined flag” column ascending to get unimpaired features at the top of the table
4. Sorting by the “probe type” column to move technical controls to the bottom of the table

The idea behind this sorting is to arrange those measurements at the top of the table which indicate 1. the strongest up-regulation, that are 2. far enough from background, 3. technically unimpaired, and are 4. no controls.

The second sorting key also accounts for the first pairwise array comparison (S1/B) but sorts the ratios in an ascending order to get the strongest down-regulation events at the top of the table.

The 10 remaining sorting keys are made up the same way regarding the 5 remaining pairwise array comparisons. Delivered data are finally sorted by the first sorting key.

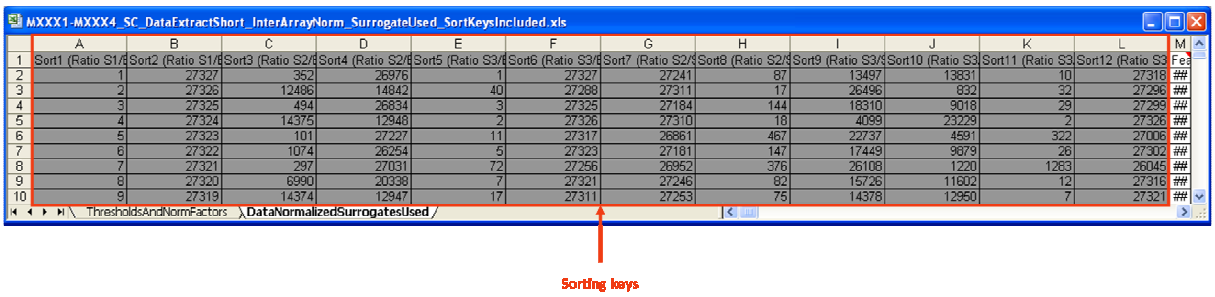

Figure 6: Sorting keys

Normalization sheet

The "ThresholdsAndNormFactors" data sheet (Figure 4) summarizes for each array its 75th percentile, the utilized normalization factor (absolute value), the normalization factor relative to array 1, the applied surrogate value and the intensity threshold value set. Additionally, the reference 75th percentile value (1500 by default) is specified.

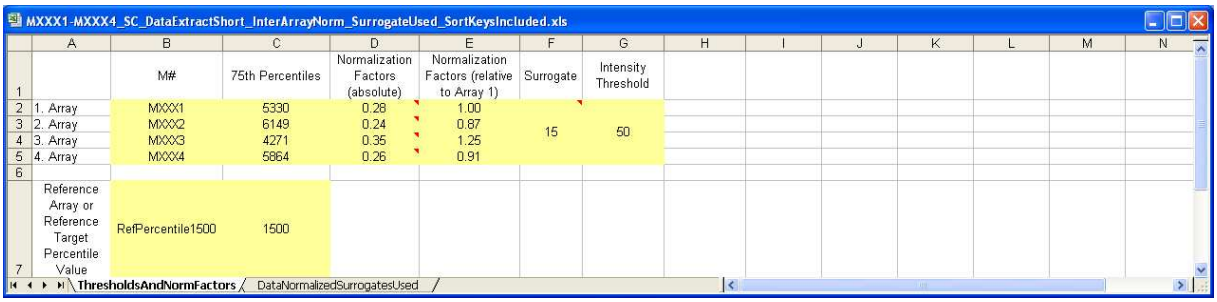

Figure 4: „ ThresholdsAndNormFactors" sheet.

Kontakt:

Heike Schneider
Technische Assistentin
Gebäude I3, Ebene 1, Raum 1080
Tel: +49 (0)511-532 2830
E-Mail: Schneider.Heike@mh-hannover.de

Torsten Glomb
Diplom-Bioinformatiker
Gebäude I3, Ebene 1, Raum 1300
Tel: +49 (0)511-532 2869
E-Mail: Glomb.Torsten@mh-hannover.de

Dr. rer. nat Oliver Dittrich-Breiholz
Leiter der Zentralen Forschungseinrichtung Transcriptomics
Gebäude I3, Ebene 1, Raum 1270
Tel: +49 (0)511-532 5814
E-Mail: Dittrich.Oliver@mh-hannover.de