Hannover Medical School

RCU Transcriptomics

Research Core Unit Transcriptomics

The Excel standard result file for dual-color studies

Version: 19.03.2012

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 dual-color studies

Our standardized Excel file format for dual color studies is named "MXXX1-MXXX4_FBOIFNUO_DC_DataExtractShort_InterArrayNorm_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, do contain 4 identical microarrays that are loaded independently with 4 pairs of samples, giving rise to 4 data sets , finally integrated into one Excel result file.

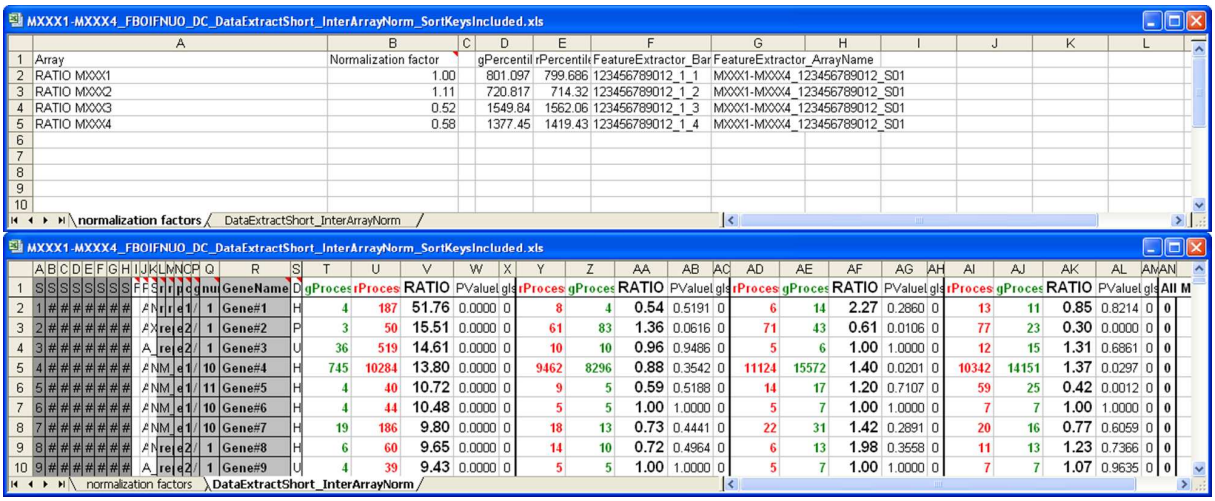

Figure 1: The two sheets of the standardized Excel file: "normalization factors" (upper panel) and "DataExtractShort_InterArrayNorm" (lower panel).

The "DataExtractShort_InterArrayNorm" sheet

This sheet is divided into three data blocks with sorting keys in the first, probe and gene annotation data in the second and experimental microarray data in the third block (Figure 2). Additionally, a higher-ranking flag column integrating individual flag columns of the 4 individual microarrays into one resulting entry is added after the right-most column of the Excel sheet (column AN).

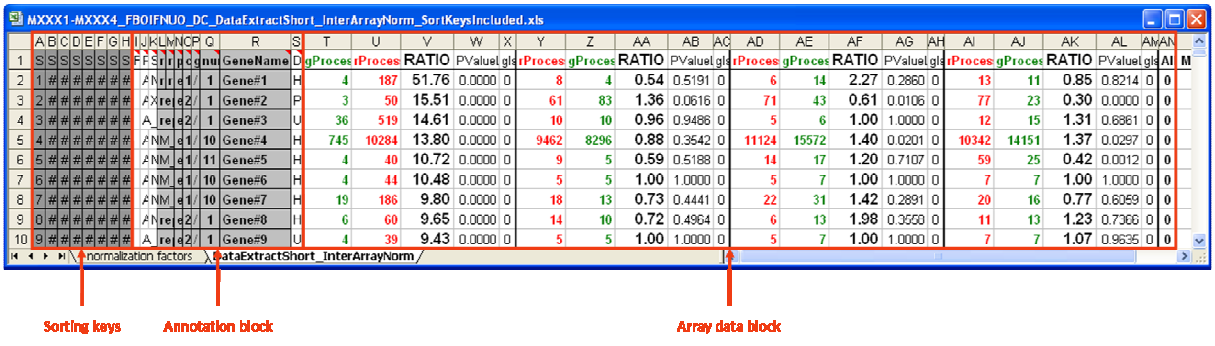

Figure 2: The "DataExtractShort_InterArrayNorm" sheet is subdivided in three blocks.

The Array data block

The array data block (Figure 3) summarizes the most important part of the data of the 4 microarrays each of which represented by 5 carefully selected columns.

These columns contain:

- the processed fluorescence intensity of the green and the red channel "gProcessedSignal" and "rProcessedSignal"
- the ratio of both intensities "RATIO MXXXX"
- a p-value "PValueLogRatio" and
- a flag, indicating technical performance of the respective spot/feature "gIsFeatNonUnifOL OR rIsFeatNonUnifOL" which stands for: green (or red) Is Feature a Non Uniformity OutLier).

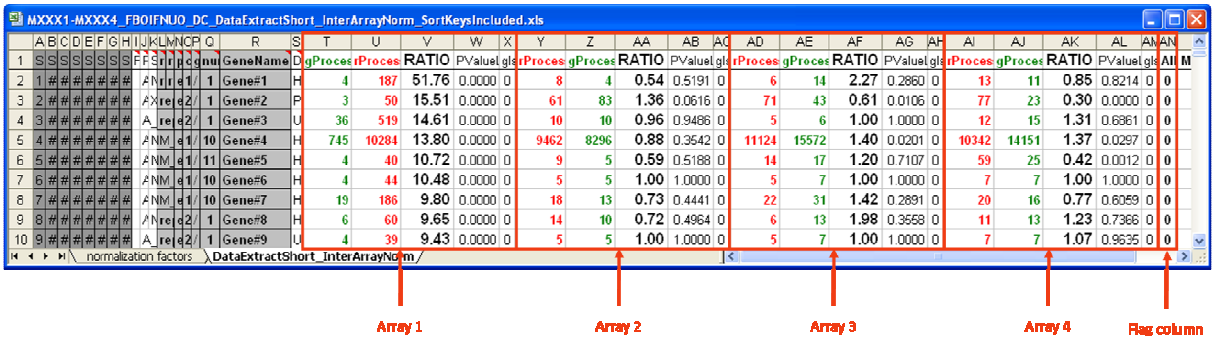

Figure 3: The Array data block and the summary flag column.

The fluorescence intensity values "gProcessedSignal" and "rProcessedSignal" of the first microarray dataset are directly taken over from the Feature Extraction Result Text File. Intensities of arrays 2 to 4 also originate from the same result file, but are subjected to inter-array normalization by global linear scaling prior to the integration into our Excel file format (see also Excursion: Normalization).

The depicted ratio values are always expressed as processed fluorescence intensity of a stimulated (or specifically treated) sample divided by the intensity of the basal (or reference) sample. Thus, ratio values indicate a fold induction (or upregulation) by the treatment, if they are (far enough) above 1. Otherwise, if they are below 1 they indicate a fold reduction (repression, downregulation) by the treatment.

The p-value denotes the statistical significance based on the log ratio of red and green fluorescence intensities. It is a measure of the confidence, that the underlying gene is not differentially expressed. Thus, the lower the p-value, the higher the probability that the measured difference in intensity reflects a true difference in abundance of the two competing labeled RNA species for a given spot. The p-value is calculated by complex algorithms in the course of data extraction with feature extraction software. The fold difference, the intensity level and the uniformity of pixel signals per spot are the most strongly weighed components constituting p-value calculations. Regarding "housekeeping" genes, for instance, the p-value should be near 1 as there is no differential expression expected between any two samples to be compared.

The flag column indicates technical performance of individual spots (features). If the uniformity of pixel signals in both channels (i.e. green and red) of a spot is high, a 0 is given. Otherwise, if there might be a spurious impairment in at least one channel it is documented by the entry 1.

Excursion: Normalization in dual-color mode

The fluorescence intensity values of the first microarray dataset are directly taken over from the Feature Extraction Result Text File. Intensities of arrays 2 to 4 are subjected to inter-array normalization by global linear scaling. For this, intensity values of the green or the red channel of these microarrays are multiplied by an array-specific scaling factor. This factor is calculated by dividing the 75th percentile of the green channel of Array#1 by the 75th percentile value of the particular microarray (Array i in the formula shown below). Accordingly, inter-array normalized processed signal (PS) values for all samples (microarray data sets) are calculated by the following formula:

Inter-array normalized $PS_{Array\ i} = PS_{Array\ i} \times (75^{th}\ Percentile_{Array\ \#1} / 75^{th}\ Percentile_{Array\ i})$

The individually calculated normalization factors for arrays 2 to 4 can be inspected on the "normalization factors" data sheet (Figure 6). The normalization step adjusts the global brightness (in our case the brightness is estimated by the 75th percentile of the intensity distribution) of each array to the global brightness of array 1. This allows for calculating ratios with reasonable confidence even across different arrays (and channels) within this Excel file. Such global scaling approaches are feasible, whenever the number of differentially expressed genes is low compared to the number of genes measured in total.

Flag column

The very last column of the array data block “All Microarrays gIsFeatNonUnifOL OR rIsFeatNonUnifOL” (column AN) integrates the flag entries from all 4 microarray datasets into one entry, representative for the series (a 4 Arrays) on the whole. If any of the array-specific flag entries for a given feature is 1, then this global flag is set to 1. It is set to 0, if none of the entries for individual microarrays is 1.

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”

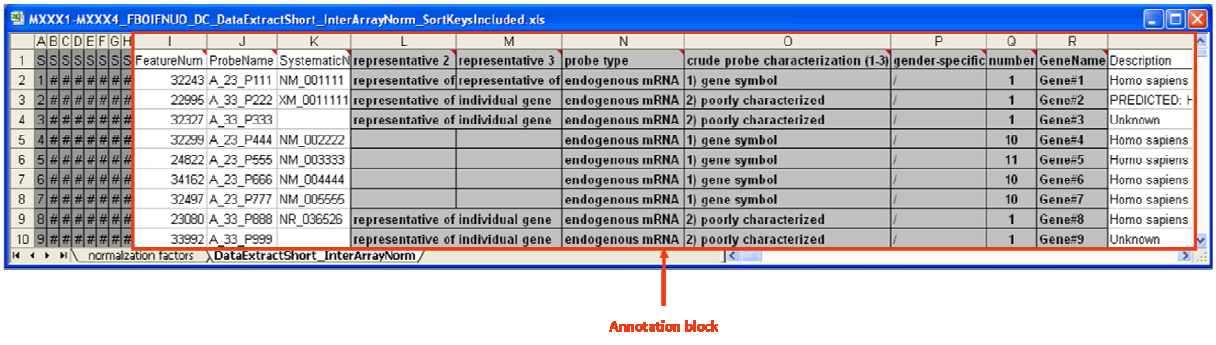

Figure 4: The annotation block

The columns “FeatureNum”, “ProbeName”, “SystematicName” and “Description” (Figure 4 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 (see excursion at the next page).

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.

[*] 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).*

Sorting keys

To rapidly inspect your data according to various aspects, the data sheet contains predefined sorting keys – two for each array.

The first sorting key was formerly generated by a 3-way sorting process.

1. Sorting of ratio values of array 1 descending from strongest up-regulation to strongest down-regulation
2. Sorting by the flag column of array 1 ascending to yield all the technically unimpaired features at the top of the table
3. 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. technically unimpaired and are 3. no controls.

The second sorting key also accounts for array 1 but sorts the ratios in an ascending order to get the strongest down-regulation events at the top of the table.

The 6 remaining sorting keys are made up the same way regarding the respective arrays 2-4.

Delivered data are finally sorted by the first sorting key.

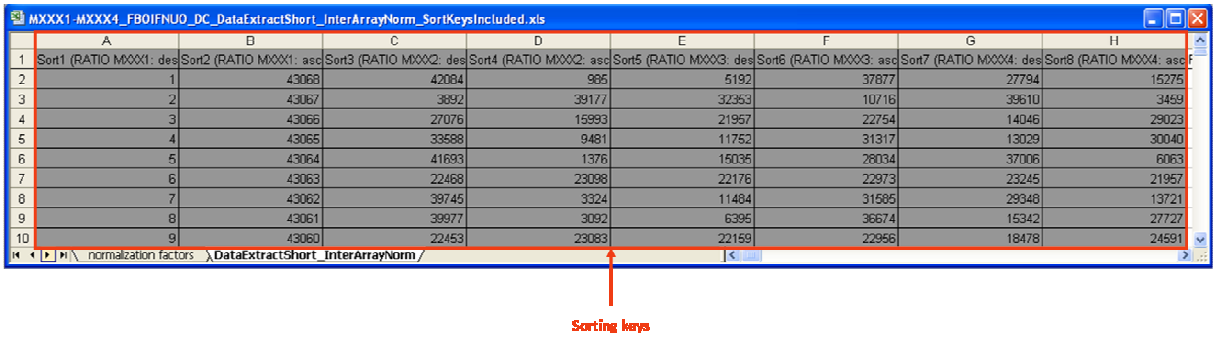

Figure 5: Sorting keys

Normalization sheet

The “normalization factors” sheet summarizes calculated and utilized normalization factors for each array, the 75th percentile of the intensity distribution (green and red channel) as well as the barcode and array name (taken over from Feature Extraction).

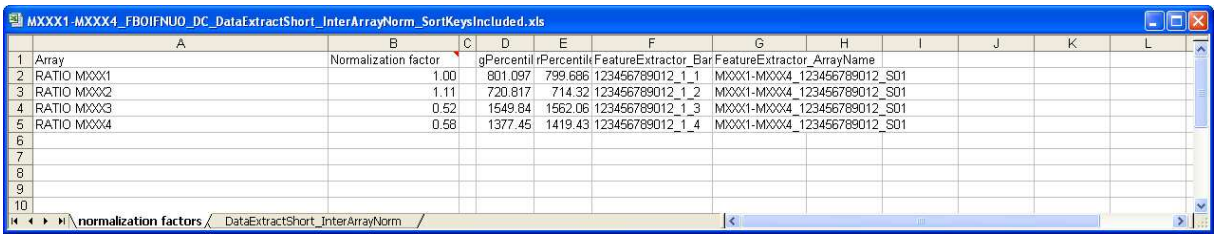

Figure 6: The "Normalization factors" sheet

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