

1. General introduction

Each project is divided into replica groups that correspond to each pair of samples being compared. Each of the replica groups is hybridised to four microarrays. Two of these hybridisations are dye swaps, meaning that the sample that was labelled with Cy3 is now labelled with Cy5 and vice versa. This is to allow compensation for the differential labelling efficiencies of each dye.

The project index file describes the groups of replicate hybridisations and how each sample has been labelled. FlyChip projects codes are in the form Pnnnnn, where nnnnn is a unique integer. FlyChip replica group codes are in the form Rnnnnn, where nnnnn is a unique integer. Additionally, some files are named based on which microarray slide they represent. FlyChip microarray slide codes are in the form Snnnnnn, where nnnnn is a unique integer.

For your convenience we provide normalised data and we recommend that you examine this first. The tab-delimited text files are suitable for further processing, e.g. clustering with other tools. Raw data are also provided in case you wish to normalise the data yourself. If you wish to perform spot-finding and all of the data analysis yourself, we have provided the raw grey-scale TIF images.

Each of the terms and codes defined above will be used throughout this text and many of the files we have sent to you have names derived from these codes. We will now explain what each of the file types we have sent to you contains. You will find it easier to understand this help text if you have a copy of the file being described open at the same time. This will enable you to compare our descriptions with the file of interest.

2. Project index file: Pnnnnn_info.text description

Pnnnnn_info.text files contain a description of how your project was organised and what sample has been labelled in each channel for each slide within each replica group. You may want to refer to this file when analysing your results. This tab-delimited file is best viewed with the freely available spreadsheet program within [OpenOffice](#) or else Microsoft Excel because some text editors will have problems with the variable column spacing.

Important note; dye swaps and gene expression ratios:

- **Normalised data** have had the dye swap taken into account and are presented as one file for each replica group, and a separate column for each slide. The dye-swap slide column data are unswapped during the analysis and are presented as a ratio of Cy5 over Cy3. This means that all fn:M numbers should be treated as though there has been no dye swap (meaning Swap_status = 0 within Pnnnnn_info.text); for further analysis you do not need to make any compensation for the dye-swaps. We use Cy5/Cy3 ratios as this is the form typically presented in publications.
- **Raw data** have been presented as one file per slide and if they are dye-swap slides the dye swap has not been taken into account. This is the case for slides marked swap_status = 1 within the Pnnnnn_info.text file. So, if you wish to normalise this data yourself you will need to take into account the dye-swap. The ratios are presented as a background subtracted ratio of Cy3 over Cy5 because this is what the spot-finding tool does by default, and these are only included for historical reasons.
- **Project_Number:** FlyChip assigned project number in the form Pnnnnn, where nnnnn is a unique integer
- **Replicate_Group:** FlyChip assigned replica group in the form Rnnnnn, where nnnnn is a unique integer

- **Slide_Number:** FlyChip assigned slide number in the form Snnnnnn, where nnnnnn is a unique integer
- **Hyb_Number:** FlyChip assigned hybridisation batch number in the form Hnnnnn, where nnnnn is a unique integer
- **Cy3_Image:** Cy3 images name, derived from the Slide_Number and denoted by the w595 suffix
- **Cy5_Image:** Cy5 images name, derived from the Slide_Number and denoted by the w685 suffix
- **Cy3_Sample_Name:** the name you provided for this sample
- **Cy5_Sample_Name:** the name you provided for this sample
- **Swap_Status:** denotes whether the slide is a dye swap (1) or not (0)
- **Comments:** problems with the slide (broken, high background, poor signal etc.) will be described here, if any

3. What images have you sent to me?

We have dispatched two different types of images. The first type are the raw grey-scale 16-bit TIF images that we analyse to quantify how much labelled sample and control has bound to each spot. The second type of image are false colour PNG images. These provide you with a visual (not normalised) representation of your results. Such images can be used to check for slide-specific problems, and for presentations.

- **Snnnnnn_w595.tif:** grey-scale 16-bit TIF image of the Cy3 channel
- **Snnnnnn_w685.tif:** grey-scale 16-bit TIF image of the Cy5 channel
- **Snnnnnn_c.png::** false colour image where Cy3 is green, Cy5 is red, and equal red and green is yellow

4. Raw data

This file is named based on which microarray slide they represent. The spot signals within this file has NOT been normalised, and so this file contains raw unprocessed data. If this represents data from dye-swap slides, the dye swap has not been taken into account. This is the case for slides marked swap_status = 1 within the Pnnnnn_info.text file. If you wish to normalise the data yourself, you will need to take into account the dye swap; we recommend using the single channel data for this purpose.

- **Snnnnnn.state.dat:** spot quantification file from dapple with associated spot identities

We recommend viewing this tab-delimited file with the freely available spreadsheet program within [OpenOffice](#) or else Microsoft Excel, because some text editors will have problems with the variable column spacing. Header information is denoted by # at the beginning of the line and all other columns are defined below. However, please note that the first number within the grid_x column is the total spot number, and the first number within the grid_y column is the number of channels.

Column definitions for Snnnnnn.state.dat

These first few columns denote where the spot is located within the microarray. Such locations are provided using a system of Cartesian co-ordinates. The x-axis corresponds to the width of the image (the shortest side) and the y-axis corresponds to the length of the image (the longest side). The reference point for these co-ordinates (0,0) is the top left spot in each image.

- **tool_x:** x-axis co-ordinate for the sub-grid (sometimes called block or pin-patch)
- **tool_y:** y-axis co-ordinate for the sub-grid
- **sgrid_x:** x-axis co-ordinate for the spot within the sub-grid
- **sgrid_y:** y-axis co-ordinate for the spot within the sub-grid

The following columns provide a description of what each spot is. This description includes the '*Drosophila* Gene Collection' clones and the predicted gene for each spot. The last of these columns defines whether the spot should be included in any normalisation, should you choose to do this yourself.

- **CloneID:** '*Drosophila* Gene Collection' cDNA clone identity
- **Pred_Gene:** FlyBase gene assigned at the time of '*Drosophila* Gene Collection' construction
- **FlyBase_gene:** FlyBase gene assignment for the current predicted gene
- **FlyBaseID:** FlyBase ID (FBgn code) for the current gene assignment of the clone
- **norm_ignore:** value of 1 indicates that the spot should not be included in any normalisation
- **show:** flag is set to 0 when data should not be included in any downstream analysis after normalisation, e.g., the spot maps to a failure PCR reaction or a control. The data should be included when the flag is set to 1, i.e., this is a high quality *Drosophila* cDNA PCR amplicon.

We then have further columns that provide details about the spot status, signal and a pixel count for the foreground (i.e., the spot) and background (i.e., the local area around the spot). Spots with very few pixels in the foreground are probably unreliable because they contain too few pixels for reliable spot signal estimate.

- **StatusN:** status of each spot for channel N; where A = [A]ccepted, R = [R]ejected, and S = [S]uspicious
- **fgMedianN:** foreground (spot) median pixel intensity of channel N
- **fgAdjMADn:** foreground (spot) pixel intensity variability of channel n
- **bgMedianN:** background (local area around spot) median pixel intensity of channel N
- **bgAdjMADn:** background (local area around spot) pixel intensity variability of channel n
- **fgN:** number of pixels in the foreground
- **bgN:** number of pixels in the background

5. Normalised data

Measured fluorescent spot signals will differ systematically between different microarray hybridizations and dyes: there will be differences in background fluorescence as well as differences in overall brightness with, e.g., one dye being twice as bright as another one. The process of correcting for such systematic differences is called normalization. The normalization method employed is closely based on the work published by Huber *et al.* (2002). We only normalise spots with an [A]ccepted spot status that also have the norm_ignore flag set to 0 (see above).

For each dye and microarray, the background fluorescence and a factor reflecting overall brightness are inferred to make identical the signals for this subset of non-differentially expressed genes. A necessary assumption is that more than half the genes are NOT differentially expressed. For further (technical) information on the normalisation please refer to Huber *et al.* (2002) *Bioinformatics* 18(1), S96-104 ([abstract](#)).

Normalised data have had the dye swap taken into account and are presented as one file for each replica group and as a separate column for each slide. The dye-swap slide column data are unswapped during the analysis and are presented as a ratio of Cy5 over Cy3, without any background subtraction. This means that all fn:M numbers should be treated as though there has been no dye swap (meaning Swap_status = 0 within Pnnnnn_info.text); for further analysis you do not need to make any compensation for dye swaps. Within the replica group files the order of the slide data columns (e.g. f0:M, f1:M..., f0:A, f1:A...) are the same as within the replica group within the Pnnnnn_info.text file.

The normalised data are very similar to a log[2] scale, i.e

- +4 = Cy5 is 16-fold higher than Cy3
- +3 = Cy5 is 8-fold higher than Cy3
- +2 = Cy5 is 4-fold higher than Cy3
- +1 = Cy5 is 2-fold higher than Cy3
- 0 = no change

- -1 = Cy5 is 2-fold lower than Cy3
- -2 = Cy5 is 4-fold lower than Cy3
- -3 = Cy5 is 8-fold lower than Cy3
- -4 = Cy5 is 16-fold lower than Cy3

After normalisation, we first calculate 'M' (similar to a log-ratio) for each spot on each slide. This allows partial self-normalization of spatial effects, e.g., variations in hybridisation efficiency across the slide surface. We then calculate an uncertainty of 'M' from the 'pixel intensity fluctuations' reported by dapple. This assumes a flat disc spot model. Then, separately for 'normal' and 'dye-swapped' slides, we calculate a weighted average, using the certainties of 'M' as weights. Lastly, we form a simple average of 'M' for 'normal' and 'dye-swapped' slides. The average is unweighted to allow partial compensation of dye-swap effects.

What do Rnnnnn_vsn.tab files contain?

For each replica group within your project we produce a summary file that contains a description of what each spot is, the transformed normalised intensity differences between the Cy5 and Cy3 channels for the replicate slides (f0:M, f1:M, f2:M, f3:M), and the transformed normalised average intensities of the replicate slides (f0:A, f1:A, f2:A, f3:A). These file can be used for further downstream processing, for example, clustering.

The first group of columns within the Rnnnnn_vsn.tab file identify which gene is represented by the spot. The genes are named using FlyChip and FlyBase identifiers.

- **FlyBase_ID:** FBgn number for the current gene assignment of the clone
- **ID:** FlyChip well identifier
- **CloneID:** 'Drosophila Gene Collection' cDNA clone identity
- **FlyBase_symbol:** FlyBase symbol for the current gene assignment of the clone, if available.

The following columns contain the transformed normalised intensity differences between the Cy5 and Cy3 channels for the replicate slides (f0:M, f1:M, f2:M, f3:M), and the transformed normalised average intensities of the channels (f0:A, f1:A, f2:A, f3:A). You will probably be most interested in f0:M, f1:M, f2:M and f3:M. Positive numbers indicate an increase in relative intensity (Cy5 greater than Cy3), and negative numbers indicate a decrease in relative intensity (Cy5 less than Cy3). Numbers of equal size but opposite sign indicate equivalent fold changes up and down respectively. Please note, dye-swaps have been taken into account so that the numbers across all replicate slides are comparable, and should ideally change in the same direction.

- **PavgM:** a quantitative measure of deviation from avgM=0 (where 1 means no deviation). This is not a P value.
- **avgM:** average transformed intensity difference of all slides (f0, f1, f2 and f3)
- **f0:M:** slide 1, transformed intensity difference
- **f1:M:** slide 2, transformed intensity difference
- **f2:M:** slide 3, transformed intensity difference
- **f3:M:** slide 4, transformed intensity difference
- **f0:A:** slide 1, transformed average intensity
- **f1:A:** slide 2, transformed average intensity
- **f2:A:** slide 3, transformed average intensity
- **f3:A:** slide 4, transformed average intensity

The last column provides an indication of how good each spot looked and can be used to determine if the reported expression changes are reliable or subject to error due to problems with the either the printing, hybridisation, or spot-finding.

- **all:spotfindStatus:** this is the spot-finding status for each channel of each slide in the form Cy3-Cy5. A = accepted. R = rejected. S = suspect.

Ranking of microarray data is hard and an active area of research. The data provided should be treated as unranked.

6. Glossary of terms

- ***Drosophila* Gene Collection (DGC)** - cDNA clone-set from the BDGP
- **FBgn** - FlyBase unique gene identification number
- **Dye swap** - Sample or control that was labelled with Cy3 is now labelled with Cy5 and the sample or control that was labelled with Cy5 is now labelled with Cy3
- **Gene Ontology (GO)** - Conserved vocabulary used to define gene structure, function, and expression
- **Project(s)** - The experiment to be performed on your behalf by FlyChip
- **Replicate group(s)** - A replica groups corresponds to each user defined sample-control pair
- **Replicate slide(s)** - Slides that have been hybridised with the same sample-control
- **Sample(s)** - The biological material that was submitted to FlyChip for analysis
- **Transformed intensity difference** - Normalised difference in signal intensity between the two channels
- **Transformed average intensity** - Normalised average signal intensity for the two channels