

The Cancer Genome Atlas Pilot Project
University of North Carolina
Raw Data Description

Platform: Agilent 44K Array

Raw Data

Raw data for UNC gene expression and microRNA is in the form of a tagged image file format (.tiff). All file information for each file is listed in the File Info dialog box. The TIFF file is compliant with Adobe version 6.0 file format. A full description of our raw data and the (the tiff images themselves) is provided at the following url:

<http://partners.adobe.com/asn/developer/PDFS/TN/TIFF6.pdf>.

There are two sets of custom TIFF tags in the Agilent file format. Genetic Analysis Technology Consortium (GATC) TIFF Tags Agilent Technologies is not a member of GATC or otherwise connected to this organization, and makes no internal use of these tags. They are included for the convenience of customers who use software that requires them. Custom TIFF Tags Agilent Technologies uses its own custom TIFF tags for storing additional file information. TIFF Tag 37701 This tag points to a data structure. This data structure is not public, but information stored in the data structure is available to customers in the MATLAB file format. TIFF Tag 37702 This tag points to a string containing the file description. The usual TIFF description tags (tag 270) are used to hold the color name, red or green, for each image. This allows programs that interpret only standard TIFF tags to determine image colors. The Page Name tag (tag 285) also contains the color names.

The Cancer Genome Atlas Pilot Project
University of North Carolina
Normalized Data Description

Platform: Agilent 44K Array

Normalized Data

Feature extraction and normalization is performed using the commercial Agilent feature extraction application. A complete description of the application is available at:

http://bioinfostore.unc.edu/sai/tcga/Agilent_Files/ReferenceGuide.pdf

Sample output file:

http://bioinfostore.unc.edu/sai/tcga/sample/US45102955_251584710099_S01_GE2-v5_91_0806.txt

Complete description of the column headers is available at:

<http://bioinfostore.unc.edu/sai/tcga/sample/headers.description>

The Cancer Genome Atlas Pilot Project
University of North Carolina
Segmented Data Description

Platform: Agilent 44K Array

Segmented Data

The following format is more appropriate for DNA copy number platforms, but could apply to expression arrays as well.

We use the following general format:

Columns 1-12 – See description below – process data and annotation

Columns 13-(n+13) (where n=number of samples) – sample data

Rows – Individual Probes

Columns 1-12

1. Probe ID
2. Genome build
3. Platform (chip name number version)
4. Either Probe ID or unique ID for summed/averaged gene or locus (such as an affy probe set).
5. Probes comprising the uniquely identified target described above (delimited by “/” or some other appropriate delimiter)
6. Method by which probes described in #4 were combined to give the data in #3.
7. Chromosome #
8. base position according to the genome build described in #1 above
9. Copy number estimate
10. Range, if the algorithm for range
11. p value if the algorithm
12. Method by which the copy number estimate was arrived at

Column 13-(13+n) where n=# of samples

Gene expression as Loess normalized log₂ ration of signal channel to reference. From the feature extraction application we prefer rBGSubSignal as the signal channel and gBGSubSignal as the reference. Loess normalization is implemented through the UMD microarray server, but is identical to the Loess normalization as implemented through SMA available through the Bioconductor project in the R statistical programming language.

Please find a sample of this file structure at
[http://bioinfostore.unc.edu/sai/tcga/sample/
DataCategoryRequest_level3dataSample_20070515.xls](http://bioinfostore.unc.edu/sai/tcga/sample/DataCategoryRequest_level3dataSample_20070515.xls)

When the analysis is anything other than DNA copy number (possibly methylation or genotype), a similar format is applicable although the columns change.