

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.