The Cancer Genome Atlas Pilot Project
Memorial Sloan Kettering Cancer Center
Raw Data Description

Platform: Agilent 244K Array

Raw data

The raw data format for the Agilent aCGH arrays is determined by the Feature Extraction program provided by Agilent. The files are tab delimited text files with one sample per file (called here a FE-file). Each FE-file has a total of 10 header lines broken into triplets of type, name, value lines for several variables and the first two blocks are delimited by a line with a '*' character in column 1. The first triplet contains the type, name, and values for the Feature Extraction Parameters. The second triplet contains the type, name and values for a number of statistical or quality measures for that slide. After the '*' break line the next two lines give the type and name of the raw data lines which are then followed by approx 240,000 lines of data for the various probes. The key column is labeled LogRatio. Agilent uses common logs (log base 10) which we leave untouched. The genomic coordinates for the probes are in the SystematicName column in the following format: chrN:<startpos>-<endpos>. Coordinates refer to build UCSC hg17 or NCBI build 35 (May 2004).

The Cancer Genome Atlas Pilot Project
Memorial Sloan Kettering Cancer Center
Normalized Data Description

Platform: Agilent 244K Array

Normalized data

Although the Feature Extraction program from Agilent does normalize the two channels we have found that this simple normalization misses an effect related to the local GC content near the probe region on the genome. These genomic-based artifacts can cause problems in the downstream analysis so we have implemented a more complex normalization procedure, which removes this effect. After our GC-normalization we take the LogRatio (which we convert to the usual log base 2 convention) for each sample in a given batch and create one tab delimited text file, which has each sample in a different column. The initial columns of the file contain the probe ids, chr, and genomic location of the probe and a pseudo-genome wide coordinate to facility graphing and analysis of the data. Finally we will filter out any samples that fail the QC metrics that have been established. Level 2 will only contain samples that pass QC.

The Cancer Genome Atlas Pilot Project
Memorial Sloan Kettering Cancer Center
Data Descriptions

Platform: Agilent 244K Array

Segmented data

The level 2 normalized data is then run through the Circular Binary Segmentation (CBS) algorithm developed by investigators within the MSK-CGCC (Olshen, Venkatraman et al. 2004; Venkatraman and Olshen 2007). The output is again a tab delimited file containing the following columns:

1. ID of sample
2. Chromosome of segment
3. Start of segment
4. End of segment
5. Number of probes in segment (Flagged probes not counted)

Page 2 of 3 $Revision: 2.1 $

6. Mean of segment

There is an additional level of normalization done at the segment level. The histogram of segment means, weighted by number of probes per segment, is shifted such the *mode* is set to 0. Therefore, an additional file is provided for convenience: the mode centered normalized log2 ratios for each probe and the mode-centered segment mean for each probe location. In addition there is one more column that indicates whether that probe sits in a known copy number polymorphic (CNP) region. (0 means no known CNP, n>=1 indicates the number of studies that have found this CNP). The file contains the following columns:

1. Probe ID
2. Chromosome
3. Chromosomal Position
4. Genomic Position
5. CNP.Flag
6. Centered, normalized log2 ratio for Sample #1
7. Centered, normalized log2 ratio for Sample #2
8. Segment mean at at probe position for Sample #1
9. Segment mean at at probe position for Sample #2
10. … etc.

The CNP flag is derived from a database of reported and locally curated CNPs. A tab-delimited form of the database is available from our website (www.cbio.mskcc.org/CGCC).

References

Olshen, A. B., E. S. Venkatraman, et al. (2004). "Circular binary segmentation for the analysis of array-based DNA copy number data." Biostatistics **5**(4): 557-72.

Venkatraman, E. S. and A. B. Olshen (2007). "A faster circular binary segmentation algorithm for the analysis of array CGH data." Bioinformatics **23**(6): 657-63.