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.