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).