

The Cancer Genome Atlas Pilot Project
Harvard Medical School and Brigham and Women's Hospital
Segmented Data Description

Platform: Agilent 244K Array

Segmented Data

An in-house R package, aCGHNorm, is used for the normalization of the log₂ ratio data described in the previous section. The normalization procedure involves the application of invariant set LOWESS normalization algorithm to log₂ ratio data. The algorithm assumes, in this case, that the majority of probe log₂ ratios do not change and are independent of the background corrected intensities of the probes. To build the LOWESS model, the log₂ ratios and the background corrected intensities of the sample and reference channels are used and a big window (21 probes) smoothing process is applied to log₂ ratio after sorting them by chromosome position. After mode-centering based on median-smoothed log₂ ratio, unchanged probes (median-smoothed log₂ ratio around zero) are then used to build LOWESS model. The invariant set LOWESS normalization is applied iteratively to the log₂ ratio data set until the sum of difference of LOWESS input and output log₂ ratio is zero or stabilized.

The artifact of the differences in probe GC content on log₂ ratios is corrected by applying LOWESS using probe GC percent, regional GC percent (GC percent of 20 KB genome residing the probe), and log₂ ratio.

Data generated by the normalization process are then merged with in-house annotation data to form a data set containing probe name, chromosomal location, and normalized log₂ ratio for each sample.

The square root of the mean sum squares of variance in log₂ ratios between consecutive probes arranged along chromosomes are calculated and used as another measurement of array quality. Arrays with values over 0.30 are considered noisy but may still contain useful information.

Data will be processed in the Belfer Genomics Center at the Dana Farber Cancer Institute.

Submission package will include the following items for each sample:

1. DNA quality measurements by nanodroping (tsv).
2. DNA quality measurements using Agilent's nucleic acid analyzer (tsv).
3. Raw image generated by a scanner (tif).
4. Compressed version of the raw image (jpeg).
5. Agilent's feature extraction QA report (pdf).
6. Additional QA statistics including percent quality probes and standard deviation (tsv).
7. Data from feature extraction (txt).
8. MEGAML version of extracted data (xml).
9. Normalized log₂ ratio (sample/reference) data (tsv).
10. Plot image of median smoothed normalized data (window size = 3) along chromosomes (png).