

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

**Platform: Agilent 244K Array**

**Raw Data**

Raw data refer to the tab-delimited text (with a .txt extension) file generated from a scanned image using Agilent Feature Extraction Software (we are currently using v9.5.11). A full version of the text files contains three sections namely FEPARAMS, STATS, and FEATURES.

FEPARAMS is the top most section containing the parameter and option settings under which the feature extraction software runs. Examples include the protocol name, version of the grid placement algorithm, offset values, etc.

STATS is the middle section that provides statistical descriptions of the results (usually on per channel or image basis). Examples include average number of saturated features per channel, standard deviation of the data points measured per channel, number of features that are flagged as population outliers, etc.

FEATURES is the section that concerns most users as it contains values/descriptions of the results for each individual feature (probe). Some of the important data include the physical locations of probes (features) on an array, intensity measurements, quality flags, and biological annotations.

The raw data (tab-delimited text file) are processed following the procedures listed below.

- When raw data are produced by feature extraction, the QA report is examined for various QA measurements for potential quality problems. Additional QA statistics will also be generated from the raw/normalized data.
- Each raw data file is read in by extracting only data in the FEATURES section that are needed for analysis and QA purposes (the file is huge and may crash a Windows machine if the complete file is read in). Data extracted include probe name, control type, median signals for both channels, and flags indicating whether a feature is non-uniform or whether a channel is saturated for that probe. Biological annotations by Agilent are not read in and will be replaced with in-house annotation data obtained by blasting the probe sequences against the genome of interest.
- The median background signal values of each channel are subtracted from the median signal values of the features (probes) of the corresponding channel to obtain the background corrected intensity values for each probe for both channels.
- Background corrected values from duplicated probes on an array are then merged by taking the median across the duplicating probes.

- Probes that are flagged out as non-uniform or saturated by Agilent feature extraction software are excluded. Probes whose median signal values are lower than that of the background are considered faint and are also excluded. The percentages of probes that are flagged out during feature extraction or faint are calculated and used as one of the QA measurements. Arrays with over 5% of probes flagged out or being faint are considered as low quality arrays.
- The log<sub>2</sub> ratios of background corrected values for the sample channel over reference channel are then calculated. The square root of the mean sum squares of variance in log<sub>2</sub> ratios between consecutive probes arranged along chromosomes are calculated and used as another measurement of array quality. An array with a value over 0.30 is considered as low quality.
- The log<sub>2</sub> ratios obtained above are ready to be normalized using an in-house normalization algorithm that will be described in the normalization section.

**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 log<sub>2</sub> ratio (sample/reference) data (tsv).
10. Plot image of median smoothed normalized data (window size = 3) along chromosomes (png).

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

**Platform: Agilent 244K Array**

**Normalized data**

An in-house R package, aCGHNorm, is used for the normalization of the log<sub>2</sub> ratio data described in the previous section. The normalization procedure involves the application of invariant set LOWESS normalization algorithm to log<sub>2</sub> ratio data. The algorithm assumes, in this case, that the majority of probe log<sub>2</sub> ratios do not change and are independent of the background corrected intensities of the probes. To build the LOWESS model, the log<sub>2</sub> 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 log<sub>2</sub> ratio after sorting them by chromosome position. After mode-centering based on median-smoothed log<sub>2</sub> ratio, unchanged probes (median-smoothed log<sub>2</sub> ratio around zero) are then used to build LOWESS model. The invariant set LOWESS normalization is applied iteratively to the log<sub>2</sub> ratio data set until the sum of difference of LOWESS input and output log<sub>2</sub> ratio is zero or stabilized.

The artifact of the differences in probe GC content on log<sub>2</sub> ratios is corrected by applying LOWESS using probe GC percent, regional GC percent (GC percent of 20 KB genome residing the probe), and log<sub>2</sub> 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 log<sub>2</sub> ratio for each sample.

The square root of the mean sum squares of variance in log<sub>2</sub> 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 nanodropping (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 log<sub>2</sub> ratio (sample/reference) data (tsv).
10. Plot image of median smoothed normalized data (window size = 3) along chromosomes (png).
11. Segment data (tsv).

**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 log<sub>2</sub> ratio data described in the previous section. The normalization procedure involves the application of invariant set LOWESS normalization algorithm to log<sub>2</sub> ratio data. The algorithm assumes, in this case, that the majority of probe log<sub>2</sub> ratios do not change and are independent of the background corrected intensities of the probes. To build the LOWESS model, the log<sub>2</sub> 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 log<sub>2</sub> ratio after sorting them by chromosome position. After mode-centering based on median-smoothed log<sub>2</sub> ratio, unchanged probes (median-smoothed log<sub>2</sub> ratio around zero) are then used to build LOWESS model. The invariant set LOWESS normalization is applied iteratively to the log<sub>2</sub> ratio data set until the sum of difference of LOWESS input and output log<sub>2</sub> ratio is zero or stabilized.

The artifact of the differences in probe GC content on log<sub>2</sub> ratios is corrected by applying LOWESS using probe GC percent, regional GC percent (GC percent of 20 KB genome residing the probe), and log<sub>2</sub> 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 log<sub>2</sub> ratio for each sample.

The square root of the mean sum squares of variance in log<sub>2</sub> 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 log<sub>2</sub> ratio (sample/reference) data (tsv).
10. Plot image of median smoothed normalized data (window size = 3) along chromosomes (png).