Novel normalization algorithms and QA measures for array CGH.

Normalization of arrayCGH data differs from that of expression data in many respects including that: (1) hybridization signals are typically smaller and with higher background due to the excess of non-target labeling products. (2) small changes in copy number may mark biologically significant events, such as single-copy gains or losses that may accompany unbalanced translocations or chromosomal rearrangements. (3) log2 ratios are expected to be highly correlated between probes targeting adjacent regions in the genome. There is no commercial normalization software, which addresses the unique aspects of aCGH data. We have developed and refined novel approaches to aCGH-specific normalization and QA measures based on more than 1000 hybridizations on Agilent cDNA and oligo expression arrays while in development at Dana-Farber, and several hundred hybridizations on oligo aCGH arrays recently at MSKCC. An additional unique feature of aCGH microarray data is that aCGH profiles are subject to artifacts related to genomic regional labeling efficiency. These artifacts may result from regional variation in DNA degradation, fragmentation, methylation or other features affecting sample DNA differently than reference DNA. Because the artifacts are regionally distributed, they can appear as regional increases and decreases in probe log2 ratios, mimicking copy number aberrations (CNA). CBS and other segmentation approaches cannot distinguish such artifacts, they are typically segmented as altered regions. We are actively investigating artifact models, which may permit artifact measurement in aCGH profiles and potential removal by local regression and other means. One effective artifact model has been the regional %GC in the genome assembly averaged over a series of base pair windows. Based on analysis of over 1000 aCGH profiles obtained on cDNA and long oligonucleotide arrays from different institutions we estimate that genomic regional artifacts significant enough to trigger aberrant segmentation appear sporadically in approximately 5% of sample hybridizations, representing a significant potential source of false-positive data. We have incorporated genomic artifact models based on %GC over 50Kb, 2Kb and 200bp windows for quantitation of artifact during QA and for removal using loess during normalization. Highest magnitude correlations with artifact are found with 50KB-100KB windows, while negative correlation with the 2KB window is the most specific predictor of an artifactually fragmented profile (see Figure 1). We have shared our algorithms with Agilent and they confirm that approximately 5% of profiles from their own core facility and from other institutions display genomic regional artifacts, and they are currently incorporating our measurements and normalization approach in their commercial software. Genomic regional artifacts appear to be equally prevalent in Affymetrix SNP array estimates of copy number (see Figure 1), suggesting that the artifacts are platform-independent and more likely related to sample DNA quality or preparation/labeling. We are currently testing more advanced artifact models based on digest fragment size and relation of oligo probe binding site to fragment ends, as well as other genome assembly parameters. Currently, multivariate loess can improve most profiles, which demonstrate artifacts but residual fragmentation is common, suggesting additional improvement in artifact models may improve correction.
1.6.1.3 Non-negative matrix factorization is common, suggesting additional improvement in sample DNA quality or preparation/labeling. We are currently testing more advanced artifact models based on number (see Figure 1.6), suggesting that the artifacts are platform-independent and more likely related to software. Genomic regional artifacts appear to be equally prevalent in different institutions we estimate to be effective in elucidating meaningful structure of false-positive data resulting from regional variation in DNA degradation, fragmentation, methylation or other features affecting sample hybridizations, representative of the 2KB window is the most specific predictor of an artifact model has been factually fragmented profile (see Figure 1.6). We have shared our findings of regional %GC in the genome assembly averaged over a series of base pair windows. Based on these correlations, genomic regional artifacts are found with 50KB-100KB windows, while negative correlations are closely reviewed during QC. Samples in the boxed region show artifact correlations. A similar pattern of artifactual fragmentation is seen during segmentation of copy number derived by Affymetrix CNAT software. 67 samples are similarly rank-ordered by the mean correlation of log2 copy number with artifact models derived by averaging %GC in the genomic region of each probe’s target sequence. Mean %GC is calculated over 50 kb, 2 kb and 200 bp windows (blue, pink and yellow, respectively). Left: Agilent 44K CGH array. 240 tumor samples are rank-ordered by the mean correlation with the 3 artifact models. Samples with high positive or negative correlations are closely reviewed during QC. Samples in the boxed region show artifactual fragmentation during segmentation. Right: Affymetrix 100K SNP platform, single 50K array, copy number derived from Affymetrix CNAT software. 67 samples are similarly rank-ordered by artifact correlation. A similar pattern of artifactual fragmentation is seen during segmentation of copy number derived by CNAT.

**Figure 1. Genomic Regional Artifact measured during QC.** Spearman rank correlation of log2 copy number with artifact models derived by averaging %GC in the genomic region of each probe’s target sequence. Mean %GC is calculated over 50 kb, 2 kb and 200 bp windows (blue, pink and yellow, respectively). Left: Agilent 44K CGH array. 240 tumor samples are rank-ordered by the mean correlation with the 3 artifact models. Samples with high positive or negative correlations are closely reviewed during QC. Samples in the boxed region show artifactual fragmentation during segmentation. Right: Affymetrix 100K SNP platform, single 50K array, copy number derived from Affymetrix CNAT software. 67 samples are similarly rank-ordered by artifact correlation. A similar pattern of artifactual fragmentation is seen during segmentation of copy number derived by CNAT.