CytoChip™ Oligo Algorithms

An overview of CytoChip Oligo array analysis in BlueFuse® Multi software.

Introduction

BlueFuse Multi software uses a series of robust algorithms to automatically analyze CytoChip Oligo arrays, ranging from grid alignment to image quantification, normalization, and calling of regions of copy-number imbalance. Calling algorithms are fully automated and do not require user tuning of parameters, thus ensuring complete reproducibility among users.

All algorithms in v2.2 are identical to those in v2.1 of BlueFuse Multi, with the exception of calling algorithms which have been modified to reduce calls of low-level artefacts, and increase calls of smaller five-oligo changes.

This technical note provides a description of the algorithms, with particular emphasis on their underlying motivation and key assumptions.

Grid Alignment

Illumina has developed fully automated and robust grid alignment algorithms that have been tested on tens of thousands of arrays. Key stages include:

- A method deploying the Radon transform estimates any rotation of the array within the image, to correct for imperfect positioning of the slide on the scanner bed.
- A robust template-matching algorithm is deployed to automatically overlay the expected grid on the image at the correct location. The algorithm is robust to highly local and slide-level background effects, and is also effective for arrays of very low signal-to-background ratio.

Quantification

Illumina has refined and optimized its image quantification methods for oligo data. A hierarchical Bayesian model is used to represent statistical variability in the spot formation and noise processes. This model is then used with appropriate numerical methods to perform the following tasks jointly:

- Segment each individual spot from the background.
- Calculate robust measurements of intensity in the spot in the Cy3 and Cy5 channels.
- Calculate a background estimate on the basis of the region around the spot and spots in the local area in the Cy3 and Cy5 channels.
- Calculate background-subtracted amplitudes and ratios that are displayed in BlueFuse Multi as the **Raw Results**.

Normalization

Illumina has developed unique normalization algorithms to correct experimental bias in the data.

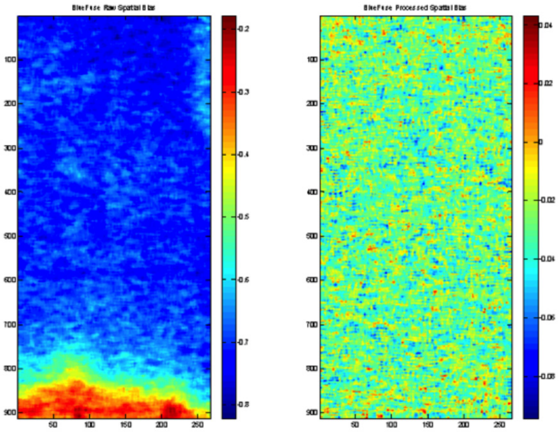
Spatial Correction

Spatial bias is estimated by applying a sliding two-dimensional (2D) median filter to the raw log ratios calculated in quantification, where the raw log ratios are first arranged in a matrix according to the physical location of their corresponding spots on the array. The estimated spatial bias is subtracted from the raw log ratios in order to remove slowly varying spatial ratio bias. This bias arises from hybridization and scanning; e.g., scanner bias causing one side of the array to have an increased average ratio relative to the other side.

To gain further insight, consider the effect of spatial correction on a single spot in the middle of the microarray. The 2D median filter calculates the median log ratio of all spots that are within 20 spots, either vertically or horizontally, of the spot (1,681 spots in total). This is effectively a robust average of the log ratios in the middle of the microarray. In the absence of technical spatial bias this value should be close to zero, provided that oligos are spatially randomized on the array with respect to genomic position (as they are on CytoChips) and the majority of oligos can be expected to show no copy-number change (as can be expected for constitutional samples). A non-zero value is therefore an estimate of technical spatial bias. This value is subtracted from the raw log ratio.

An example of spatial bias correction is shown in Figure 1.

Figure 1: Spatial Correction



Raw log ratios (left panel) and corrected values (right panel). The technical spatial artefact at the bottom of the microarray has been removed.

Intensity Bias Correction

A robust local polynomial regression fitting (Loess/Lowess) algorithm is applied to remove intensity-dependent bias from the spatially corrected log ratio. This bias can arise due to varying spot intensity relative to the background, dye bias, and non-linearity in scanner response.

Lowess regression is a technique for fitting a smoothing curve to a dataset. In this case, the regression is between the log product of the Cy3 and Cy5 amplitudes (a measure of intensity) and the log ratio. The log ratios implied by the Lowess regression are used to correct the log ratio data. The fundamental assumption of intensity bias correction is that spots of all amplitudes should, on average, have a log ratio of zero.

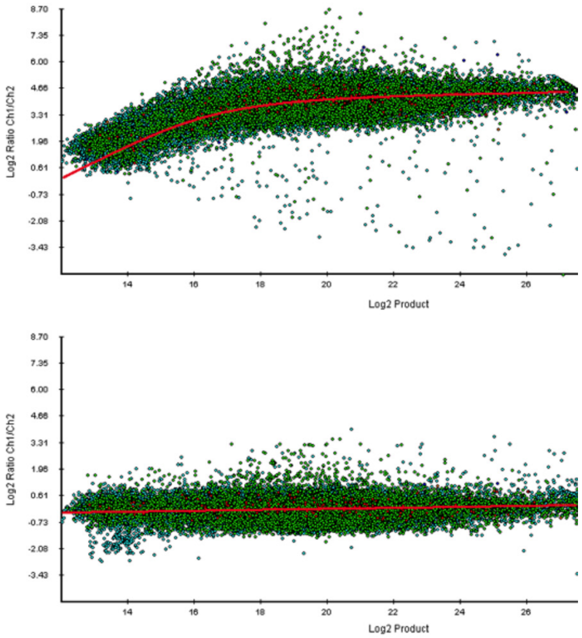
An example of intensity correction is shown in Figure 2.

GC Correction

A GC correction algorithm is applied to reduce GC bias that may arise when sample quality is low. Such GC bias typically manifests itself as spikes towards some telomeres, and raised portions of chromosome 1, 17, and 19, in the genomic profile.

As with intensity correction, a robust regression is performed, in this case between %GC content of each oligo sequence and its surrounding genomic region, and the log ratio following spatial and intensity bias correction. The resulting curve is an estimate of GC bias and is subtracted from the log ratio. The fundamental assumption of GC correction is that oligos of all GC contents should, on average, have a log ratio of zero. It is worth noting that different oligos will be corrected by the GC bias removal algorithm to differing degrees, depending on their GC content. Equally, the degree of GC correction for an oligo is not fixed among different samples; the algorithm measures overall GC-dependent bias, and corrects proportionally. Thus, if a sample shows little GC bias overall, negligible correction will be performed, even where the GC content of an individual oligo is high.

Figure 2: Intensity Correction



Raw data are shown in the top plot and the corrected data are shown in the bottom plot. In both plots, the Y axis is log ratio and the X axis is log product. The red line indicates the Lowess fit.

Smoothing

Results of normalization are smoothed using a sliding three-oligo median filter (the oligo in question, and the immediately adjacent oligos). This smoothing serves to reduce noise in the profile by removing any single oligo changes that are not supported by at least one adjacent oligo. This result reflects the assessment that a single oligo is insufficiently robust to reliably provide useful information. The normalized and smoothed ratios are seen in BlueFuse Multi as the **Fused Results** and form the basis of all comparative genomic hybridization (CGH) calling.

CGH Calling

A series of rules are applied to detect regions of copy-number imbalance in the data. To be called:

- An aberration must include sufficient oligos to reliably distinguish it from noise. At least three oligos are required for any call.
- Depending on the number of oligos involved, an aberration is required to be separated from the autosome by both a set number of robust standard deviations and a fixed $\log_2$ ratio threshold. The larger the number of oligos involved, the smaller the separation required. A $\log_2$ ratio of at least 0.3 is required for smaller aberrations.

Key thresholds are shown in Table 1.

Table 1: Key Thresholds for CGH Calling

No. of Oligos	Robust SDs	Abs. Threshold
3	5	0.3
4	4	0.3
5	3.5	0.3
8	3	0.3
15	2.5	0.12
100	2	0.12

It can be seen that a potential 50-oligo aberration requires less separation to be called as significant compared to a 3-oligo aberration. This allows reliable detection of larger, low-level mosaic aberrations.

Individual oligos that fail to show a substantial copy-number change, but are in the middle of a clearly significant region of clear copy-number change (at least five consecutive oligos) are included in the region. This calculation reduces inappropriate splitting of large single regions in noisier samples.

Finally, regions that are less than 50 oligos in length are tested to check that they are separated from both the immediately surrounding regions by more than two robust standard deviations (and that the immediately surrounding regions are not genuine changes). The immediately surrounding region on one side has length equal to 2× the size of the region in question, except that the immediately surrounding region must always be at least 10 oligos in length. This is to avoid calling the crest of large-scale "waves" that exceed standard deviation thresholds set out in Table 1, but which are transparently part of larger-scale DNA-related artefacts (e.g., shifting of chromosome 19 telomeres).

Reporting

The CGH calling algorithm calls any region that appears significant in the data, large or small. This calling does not necessarily correspond to which regions are interesting for follow-up. Therefore, regions detected in CGH calling are automatically marked as reported or not, depending on whether they exceed user-defined region size thresholds. Four thresholds are used (Table 2).

"Exclusively backbone" thresholds are applied to regions that do not include any oligos associated with a known disease region or Online Mendelian Inheritance in Man (OMIM) gene. "Including disease" thresholds are applied in all other cases. By default, all region thresholds are set to 0 Kb.

Table 2: Key Thresholds for Reporting

Threshold	Example Setting
Minimum del size - exclusively backbone (Kb)	100
Minimum del size – including disease (Kb)	50
Minimum dup size - exclusively backbone (Kb)	300
Minimum dup size – including disease (Kb)	200

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 1570-2014-035 Current as of 09 December 2014