illumina®

CytoChip[™] Algorithm Summary

An overview of array post-processing algorithms in BlueFuse® Multi software.

Introduction

BlueFuse software is used in conjunction with different array platforms, including Illumina bacterial artificial chromosome (BAC) platforms—24sure™, 24sure+, and CytoChip Focus Constitutional. The software estimates genomic copy number imbalance at differing resolutions, depending upon the sample and application.

In BlueFuse software, these microarrays are processed using the CytoChip module. This technical note describes post-processing for Illumina BAC arrays, with particular emphasis on the ideas behind the algorithms and key assumptions. The post-processing steps within the CytoChip module are fixed to ensure consistency among users; therefore, tuning parameters do not need to be set by the user.

Normalization

Using DNA hybridized against itself on a microarray, a spread in the log₂ ratios of the probes is expected, centered on the zero line. This spread is partly due to irreducible noise and partly due to the influence of several factors that systematically bias the log₂ ratio data. Systematic biases can be modeled and their effects subtracted, causing a reduction in the spread of log₂ ratio data and making profiles appear less noisy.

Spatial Bias Correction

Probes are allocated at random to addresses on the array. In most experiments, probes have a log₂ ratio close to zero. Therefore, large areas where all the adjacent spots have ratios above or below zero would not be expected. However, in practice, these regions can be observed. Some reasons include pin bias in printed arrays and uneven hybridization. These issues can lead to areas of the array where the spots have higher or lower log₂ ratio values than expected.

The spatial bias correction algorithm models the log₂ ratio data from an array as a smooth non-linear function of spot position. This estimate is subtracted from the observed log₂ ratio of each spot to remove bias.

Intensity Bias Correction

Due to variability in scanner laser power, probe labeling, and dye efficiencies, the measured log₂ ratios may be biased away from zero. In addition, the degree of bias can vary with the overall spot intensity. This bias can be visualized in a plot of log₂ ratio against log₂ product. Ideally, this plot shows no trends. In practice, however, it often has a slope or even a curve. Intensity bias correction models the trend between log₂ ratio and log₂ product and subtracts it from the data.

GC Content Bias Correction

GC-rich and GC-poor sequences have different hybridization characteristics. In array comparative genomic hybridization (arrayCGH) data, a characteristic waviness can sometimes be seen in the profile of a sample. This bias can be visualized by plotting log₂ ratio against GC content. Ideally, this plot shows no trends. In practice, however, it often has a slope or even a curve. GC bias correction models the trend between log₂ ratio and GC content and subtracts it from the data.

Exclusion

Illumina BAC arrays are manufactured to a consistent standard. Sequentially applied exclusion rules have been developed to use knowledge of this consistency to eliminate outlier replicates more efficiently. These rules are applied early in the process to improve robustness of standard deviation (SD)–based exclusion. Clone inclusion rates and data quality are typically improved. The exclusion rules are summarized as follows:

- Detection of outlier replicates for each clone. Individual replicates that show a substantially different log₂ ratio to the remainder are excluded as outliers.
- Removal of low-intensity spots. Low-intensity spots with an inadequate signal to noise ratio are removed. The confidence measure, which could lead to a small number of homozygous deletions being excluded, is not used in this calculation.
- Exclusion of replicates by robust SD for each individual subarray. Replicates in an individual subarray that are inconsistent according to SD are removed (log₂ SD threshold: 0.15). The SD threshold is doubled for clones where all replicates show a potentially genuine copy-number imbalance (> 0.3 or < -0.3). This method significantly reduces the probability of excluding a clone showing a genuine imbalance because it is just outside the regular SD threshold.
- Exclusion of clones with insufficient replicates. Where the number of replicates remaining for a clone is less than two, the clone is excluded.
- Removal of control spots. All control spots on the array are known and are removed from further consideration.

Fusion

Probe replication can be an effective method for reducing the noise in an arrayCGH profile. When more than one replicate result remains for a probe after exclusion, the data from the replicates are fused to produce a single final result for that probe. This fusion is done by taking a robust average over the replicates.

Smoothing

The log₂ ratio measurements made by individual probes can be noisy but, in general, neighboring probes produce similar values. Smoothing is a technique that allows reduction of the noise in an arrayCGH profile by pooling information from neighboring probes. Smoothing is controlled by picking a window size, k. To calculate the smooth log₂ ratio for a probe, the software calculates the median log₂ ratio of the probe and its k neighbors on either side. As k is increased, the profile becomes less noisy; however, the trade-off is a corresponding decrease in resolution. A helpful rule is that after smoothing with a window size k, only regions with k+1 or more probes are visible, and any smaller regions are suppressed.

CGH Region Calling

Region-calling algorithms have been optimized with respect to both the array technology and the application.

Aneuploidy (24sure+, 24sure)

24sure arrays have been optimized for the detection of wholechromosome aneuploidy. The region-calling algorithm works by determining the majority behavior of each chromosome. If most probes in a chromosome have a log₂ ratio consistent with typical log₂ ratios expected for gains or losses, the chromosome is called as a gain or loss. A confidence metric is also calculated for each call. Confidence values range from 0.5 to 1.

A value of 0.5 indicates that the chromosome is on the boundary between no change and either gain or loss. Values close to 1 indicate that the affected chromosome has shifted away from the expected ratio.

CGH Region Calling (CytoChip Focus Constitutional)

For high-resolution copy-number analysis, BlueFuse Multi uses a hierarchy of rules to detect regions of copy-number change. These rules weigh the factors involved in assessing the evidence for a call. These factors include the number of probes in the interval, the magnitude of the log₂ ratio change, and the level of noise. This approach allows for robust detection of large-magnitude changes in short intervals with few probes, as well as more subtle low-magnitude changes in log₂ ratio associated with mosaicism, which can involve larger intervals containing many probes.

Region Assessment (CytoChip Focus Constitutional)

Fully assessing the significance of a copy number imbalance requires both a high level of expertise and full information of the context.

BlueFuse Multi aims to support these expert decision makers by providing useful summaries of data relevant to the decision. BlueFuse Multi also applies simple rules to that data to suggest a region assessment in certain cases where the evidence appears to be clear.

To make an automated region assessment, the software first looks for clear evidence that the region is benign.

A region is classified as "Benign?" if it is:

- 1. Fully covered by either:
 - a. Significant copy-number variants (CNVs) from the Database of Genomic Variants (DGV)
 - b. Benign CNVs from the International Standard for Cytogenomic Arrays (ISCA) database
- 2. and not overlapping a known disease region
- 3. and not overlapping an ISCA pathogenic CNV

If the region cannot be classified as benign, it is assessed for evidence that it is pathogenic.

A region is classified "Pathogenic?" if it is:

- 1. Less than 80% covered by DGV regions of the same type
- 2. and at least one of the following criteria applies:
 - a. 1 Mb or greater in size
 - b. or overlaps a known disease region
 - c. or fully covered by ISCA pathogenic regions of the same type

If there is no evidence that the region is pathogenic, it is classified as "Unknown?"

Automated software assessments are always marked with a "?" symbol to indicate that these assessments are suggestions, not definitive results. Users should not rely on these results directly and should always investigate and confirm the results using full clinical information regarding a case.

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

FOR RESEARCH USE ONLY

© 2014 Illumina, Inc. All rights reserved.

Illumina, BlueFuse, CytoChip, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 1570-2014-036 Current as of 16 December 2014

