White Paper: Allele Intensity Import

Introduction

Marker-level estimates of DNA intensity is dependent on the sequence of the marker (or "probe"), and for DNA that is fragmented prior to amplification (i.e. Affymetrix), which is also dependent on fragment length. In order to reduce bias and noise in the resulting data prior to analysis, Partek provides options to estimate and remove these effects.

When importing Affymetrix data from CEL files, the intensity of individual probes are first to adjust for bias due to fragment length (smaller fragments are sometimes amplified more than larger fragments). This bias is modeled using a nonlinear function and removed on a sample-by-sample basis, leaving the unbiased residual intensities.

After optional fragment length correction, the bias due to the probe sequence are estimated and removed. GC content is known to cause bias in hybridization intensity and is a simple, fast, and effective way to remove this bias. Partek's GC correction is described in the Partek GS User manual (chapter 4). Optionally, a full sequence-based correction ("sequence correction") may be applied, which includes removal of GC bias and other sequence-based hybridization bias. Sequence correction may give slightly better results than GC correction alone, but it takes substantially longer to perform the calculation and therefore leads to slower import. Since sequence correction also adjusts for GC content, only one of the two algorithms is applied. Interrogating probes and control probes are both used and treated identically during fitting and adjustment.

For best results, it is important to apply the same settings for bias correction to both the reference ("baseline") samples and the study samples.

Each allele is summarized using the geometric mean of the multiple probes for that allele. The output intensity spreadsheet contains the log intensities for each allele or CNV probe.

The total probeset intensities are defined as:

- $I_a + I_b$ for SNPs where I_i is the intensity for allele *i*
- CNV probes are defined as just the intensity, I_{CNV}.

After calculating total probeset intensities, samples are normalized by scaling the samples' geometric mean intensity to one (0 in log space).

Creating Copy Number from Allele Intensities

Creating copy number from the summarized intensities is accomplished by normalizing each sample to the reference - either paired references or a pooled reference depending on paired or unpaired workflow.

Unpaired Baseline Creation

Inputs:

- Gender for each sample (optional)
- Chromosome variation file (optional – it includes the probes that are uniquely on the non-autosomal regions of the X and Y chromosome – which used together with gender information, can further remove bias – see below.)
- Data spreadsheet containing log base 2 normalized intensities for alleles and CNVs

Outputs:

- A file with the reference copy number intensities

Each probeset is summarized to its total intensity as $I_a + I_b$ or I_{CNV} for SNPs and CNVs respectively. Each sample's overall ($\log_2$) intensity is adjusted such that the geometric mean intensity is 1 on the non-gender influenced chromosomes (as defined by the chromosome variation file).

The pooled reference intensity is calculated as the mean of all samples for the respective probe set.

Probes within each group specified in the chromosome variation file are considered as their own normalization group and their geometric mean intensity is adjusted to 1 (or to the geometric mean of females if available on chromosome X).

If there are multiple expected copy number levels on the X chromosome, these are remembered to be used as a bias adjustment during copy number creation.

Unpaired Copy Number Creation

Inputs:

- Spreadsheet containing log base 2 normalized intensities for alleles and CNVs.
- Reference intensity file created during baseline creation.

Each probeset is summarized to its total intensity. A log ratio calculation is carried out to find the intensity relative to the reference intensity.

A bias adjustment uses the X geometric mean intensities found during baseline creation (or a predefined coefficient based on the built-in 270 HapMap reference) to adjust the log ratio values to be unbiased (i.e. females should have 2 copies of X and 0 copies of Y, while should have 1 copy of X and 1 copy of Y). This operation is simply finding the proper coefficient that will remove copy number estimation bias.

Log ratios are turned into copy number space using the formula:

$$Copy\ Number = 2 \times 2^{(log2\ intensity)}$$

As a final step, outliers are removed in a method suggested Olshen et al. An estimate of the standard deviation is found for each chromosome using the average squared lag one distance to estimate 2*var. Each local neighborhood is defined by the current probe and the two next probes and two previous probes in genomic ordering for a total of 5 probes. All probes greater than four standard deviations from the neighborhood median are adjusted to two standard deviations.

References

Olshen, A.B., Venkatraman, E.S., Lucito, R., & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics, 2004, 5(4):557-72.