White Paper: Unpaired Copy Number Estimation

The first step in estimating unpaired copy number is to normalize each chip so that the global chip intensity of the non-gender chromosomes are consistent for all samples. Partek creates a relative measure of global chip intensity for SNP and CN probes separately.

$\text{ChipIntensity}_{SNP} = \text{median}\{A_i + B_i\}$
$\text{ChipIntensity}_{CNV} = \text{median}\{C_j\}$

After the global chip intensity has been estimated, each probeset is scaled as:

$A_i' = A_i / \text{ChipIntensity}_{SNP}$
$B_i' = B_i / \text{ChipIntensity}_{SNP}$
$C_j' = C_j / \text{ChipIntensity}_{CNV}$

A_i and B_i are the intensities for allele A and B of SNP i respectively. C_j is the intensity for CN probe set j. Partek uses these adjusted probeset intensities to create log ratio estimates for each SNP. Log ratio are calculated as

$LR_i = \log((A_i' + B_i') / R_i)$,
$LR_j = \log(C_j' / R_j)$

The reference intensity for CN probesets, R_j, is estimated as the median intensity across all of the reference samples. For genotypes, the estimate of reference intensity, R_i, is estimated to adjust for the relative abundance of the A and B allele intensities. If the A and B intensities do not respond similarly, there may be genotype bias when using a simple sum of A and B intensities to represent overall SNP intensity. To reduce the impact of genotype bias a linear model is robustly estimated with of a covariate, θ_i.

$\theta_i = \text{atan}(B_i'/ A_i')$
$\log(A_i' + B_i') = \beta_0 + \beta_1 \theta_i + e$
$R_i(\theta_i) = \beta_0 + \beta_1 \theta_i + e$

The log ratio for each sample is then estimated as:

$LR_i = \log((A_i' + B_i') / R_i(\theta_i))$

The figures below show model and log ratios estimated for a selected SNP. The genotype bias has been reduced by accounting for the relative balance of A_i' and B_i' through the θ_i covariate.

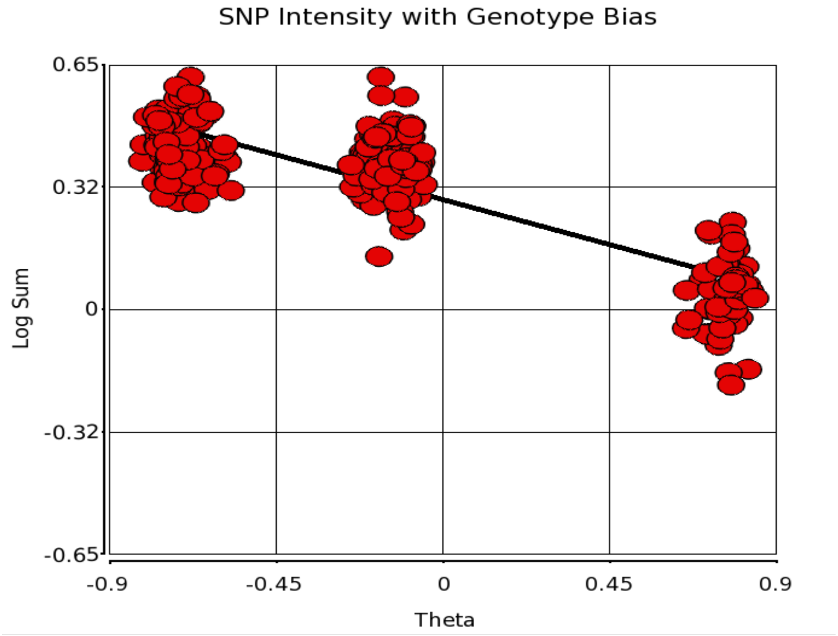

Figure 1: The estimated model of a selected SNP that exhibits a strong genotype bias

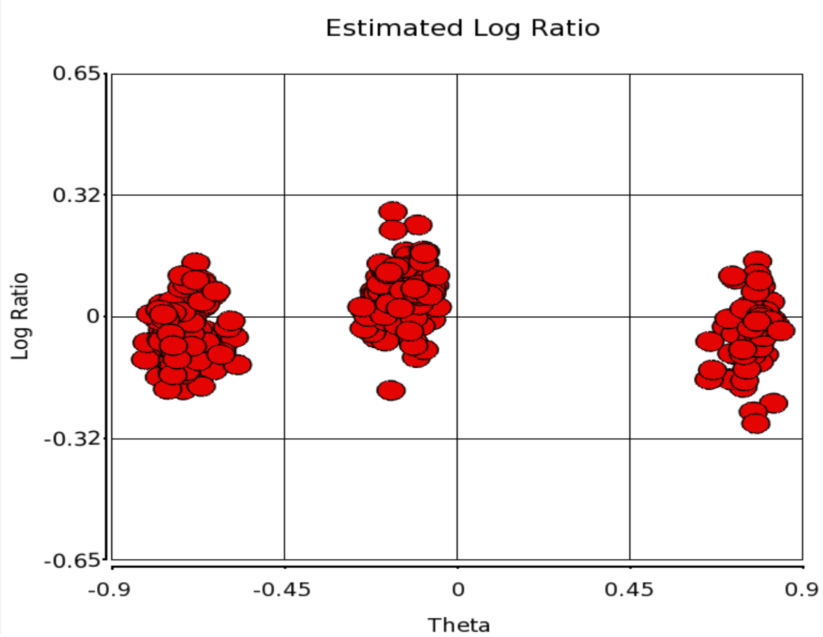

Figure 2: Log ratios after applying the estimated reference model. A genotype bias is still present, but has been significantly reduced

Allele Ratio Estimation

All SNP allele probesets do not respond identically to different DNA abundances during hybridization. To improve consistency of the allele ratio values for each SNP, a model is estimated by clustering θ_i values into three clusters to create an allele ratio model for that SNP. New AR_i values are then transformed such that the allele ratio of the middle cluster is .5, the AA cluster is 0, and the BB cluster is 1.

When A_i' and B_i 'are from a heterozygous SNP, AR_i will be centered at .5. Homozygous AA and BB genotypes will lead to values of 0 and 1 respectively. So, in a diploid region of the genome, we would expect to see three modes representing AA, AB, and BB calls in a region.

When all three genotypes are not present in the reference dataset, or the clustering algorithm fails to faithfully separate the clusters, there may be SNP models that do not faithfully represent one or more of the genotypes. These are identified using minimum cluster separation distances and produce missing values in the final allele ratio result.

GC Wave Adjustment

A new post processing step has been added in Partek 6.5 to adjust for local GC content in the DNA in a window around the probe. The GC Wave adjustment may improve analysis when the amount of input DNA, PCA characteristics, or other GC correlated technical artifacts vary between samples. The method is based on the published method from Diskin, et. al. with the slight difference of using a robust estimator rather than ordinary least squares.

You must specify a genomic window size to use when calculating the median probeset intensity and GC content around every probe. Partek uses genome reference data in a 2bit file format to calculate the local GC content. Tools for creating and manipulating 2bit files for non-human genomes are available from UCSC.

References

Sharon J. Diskin, Mingyao Li, Cuiping Hou, Shuzhang Yang, Joseph Glessner, Hakon Hakonarson, Maja Bucan, John M. Maris, and Kai Wang. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms, *Nucleic Acids Res., 2008, 36:19.*