This document describes the criteria for calculation of B Allele Frequency and Log R Ratio from tQN normalized Infinium data

Calculation of B Allele Frequency from tQN normalized data

Calculation of B Allele Frequency (BAF) is performed per SNP using linear interpolation of the measured theta to genotype clusters. Genotype clusters are determined per SNP from analysis of HapMap samples. Genotype clusters ideally consists of 6 values per SNP:
* tAA : mean theta value for SNPs genotyped as AA
* tAB: mean theta value for SNPs genotyped as AB
* tBB: mean theta value for SNPs genotyped as BB
* rAA : mean R value for SNPs genotyped as AA (used for Log R Ratio)
* rAB: mean R value for SNPs genotyped as AB (used for Log R Ratio)
* rBB: mean R value for SNPs genotyped as BB (used for Log R Ratio)

Linear interpolation is performed as described in Peiffer et al. Genome Research. tQN reference genotype clusters are constructed from the data present in the approx 120 HapMap samples analyzed on each Infinium II product. In the following cases problems arise:

1. No genotype has been assigned to any cluster (AA, AB, BB) for all HapMap samples. None of the values for the cluster centers (above) can be computed. As a result, the B allele frequency is set to NaN for this SNP.

2. Only one genotype is present across all HapMap samples. In the case of all HapMap samples having an AA genotype, only tAA and rAA can be defined. For such SNPs:

- If the measured theta <= tAA then BAF = 0
- Otherwise BAF = NaN

Correspondingly for the case of only BB HapMap samples:

- If the measured theta >= tBB then BAF = 1
- Otherwise BAF = NaN

3. One homozygous genotype is not present in any HapMap sample. E.g. only AA and AB are present, in which case only tAA, tAB, rAA, and rAB can be defined. For such SNPs:

- If the measured theta is between tAA and tAB then linear interpolation is performed as in Peiffer et al.
- If the measured theta is above tAB (tBB does not exist) then tBB is estimated as the median of tBB for all SNPs in the reference genotype set and linear interpolation is performed as in Peiffer et al.

4. Only AA and BB are present as genotypes in the HapMap samples. This case generates only tAA, tBB, rAA, and rBB. If the measured theta <= tAA then BAF = 0. If the measured theta >= tBB then BAF = 1. Otherwise, BAF = NaN.

5. Genotyping failure in the HapMap samples. HapMap samples may be genotyped incorrectly for certain SNPs. For such SNPs inconsistencies may arise, e.g. that tAA > tAB or tAB>tBB. This creates artifacts in the linear interpolation. Such SNPs are set to BAF = NaN.

Calculation of Log R Ratio from tQN normalized data

Calculation of Log R Ratios is performed per SNP by linear interpolation as described in Peiffer et al. In the following cases problems arise:

1. No genotype has been assigned to any cluster (AA, AB, BB) for all HapMap samples. None of values for the cluster centers (above) can be computed. As a result, the Log R Ratio is set to NaN for this SNP.

2. Only one genotype is present across all HapMap samples. In the case of all HapMap samples having an AA genotype, only tAA and rAA can be defined. For such SNPs:
   - If the measured theta <= tAA then Log R ratio is calculated as logR = log2(R/rAA)
   - Otherwise Log R Ratio = NaN

   Correspondingly for the case of only BB HapMap samples:
   - If the measured theta >= tBB then Log R ratio is calculated as logR = log2(R/rBB)
   - Otherwise Log R Ratio = NaN

3. One homozygous genotype is not present in any HapMap sample. E.g. only AA and AB are present, in which case only tAA, tAB, rAA, and rAB can be defined. For such SNPs:
   - If the measured theta is between tAA and tAB then linear interpolation is performed as in Peiffer et al.
   - If the measured theta is above tAB (tBB does not exist) then Log R Ratio = NaN

4. Only AA and BB present as genotypes in the HapMap samples. This generates only tAA, tBB, rAA, and rBB. If the measured theta <= tAA or if the measured theta >= tBB then Log R Ratio is calculated as in 2. Otherwise, Log R Ratio = NaN.

5. Genotyping failure in the HapMap samples. HapMap samples may be genotyped incorrectly for certain SNPs. For such SNPs inconsistencies may arise, e.g. that tAA>tAB or tAB>tBB. This creates artifacts in the linear interpolation. Such SNPs are set to Log R Ratio = NaN.