Median of the Absolute Values of all Pairwise Differences and Quality Control on Affymetrix Genome-Wide Human SNP Array 6.0

Introduction

For quality control purposes, we need to define a metric that demonstrates whether the assay will produce data that is useful for copy number (CN) analysis. This metric is Median of the Absolute values of all Pairwise Differences (MAPD).

MAPD is defined as the Median of the Absolute values of all Pairwise Differences between $\log_2$ ratios for a given chip. Each pair is defined as adjacent in terms of genomic distance, with SNP markers and CN markers being treated equally. Hence, any two markers that are adjacent in the genomic coordinates are a pair. Except at the beginning and the end of a chromosome, every marker belongs to two pairs as it is adjacent to a marker preceding it and a marker following it on the genome. Formally, if x_i: is the $\log_2$ ratio for marker i:

$$\mathrm{MAPD} = \mathrm{median}(|\, x_{i+1} - x_i \,|, i \text{ ordered by genomic position})$$

MAPD is a per-chip estimate of variability, like standard deviation (SD) or interquartile range (IQR). If the $\log_2$ ratios are distributed normally with a constant SD, then MAPD/0.96 is equal to SD and MAPD*1.41 is equal to IQR. However, unlike SD or IQR, using MAPD is robust against high biological variability in $\log_2$ ratios induced by conditions such as cancer.

Variability in $\log_2$ ratios in a chip arises from two distinct sources:

- Intrinsic variability in the starting material, hyb cocktail preparation, chip or scanner
- Apparent variability induced by the fact that the reference may have systematic differences from this chip

Regardless of the source of the variability, increased variability decreases the quality of CN calls.

MAPD and Data Quality Guidelines

However, variability in general will be reduced by using a reference set formed from chips run at the same lab using the same reagent lots.

As in genotyping, there can be substantial batch effects or lab to lab systematic effects. If a reference is formed of chips run in another lab, such systematic differences inflate apparent variability. Affymetrix has observed that using the supplied Affymetrix reference with chips run in different labs will inflate MAPD by around 50 percent, but a factor of two is possible.

If a chip with MAPD formed using the Affymetrix reference is greater than 0.4, we recommend against using that chip in an analysis.

If a chip with MAPD formed using a reference formed from chips run at the same lab with the same reagent lots as this chip is greater than 0.3, we recommend against using that chip in an analysis. The rationale for these cutoffs is described in the section, “MAPD and Functional Performance.”

The Effect of the Reference on MAPD

As mentioned earlier, MAPD is a measure of variability relative to a reference of samples. When systematic effects between a sample and the reference exist, then MAPD is inflated by this difference. Figure 1 illustrates the effect of different sites. MAPD of all samples is relatively constant when using their own internal reference, but varies up to a factor of two or higher using the Affymetrix reference. This shows that the data within each site is quite consistent, but that there are systematic differences between sites.

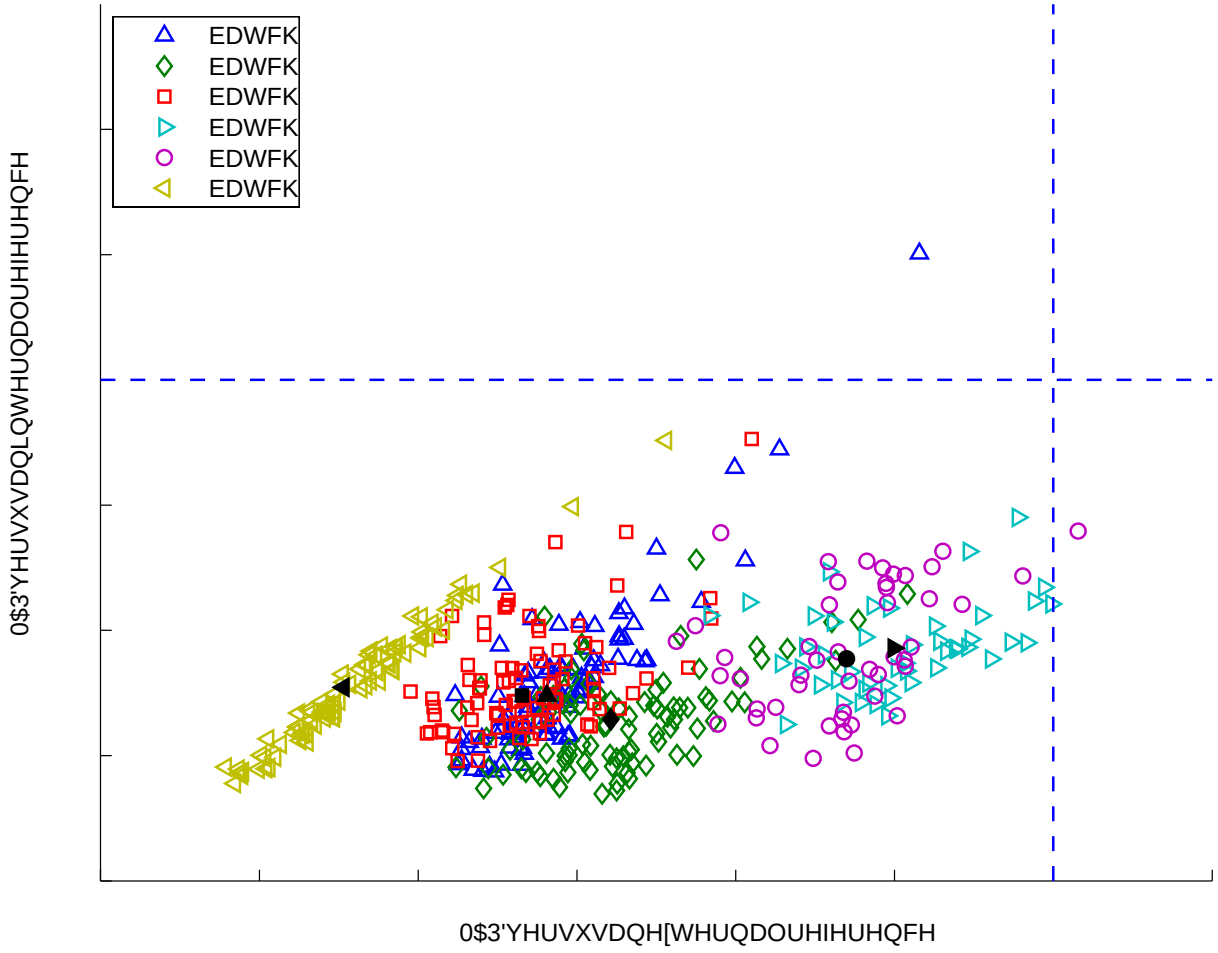

Figure 1: Per-chip MAPD calculated for six different batches of samples (four labs, two of which ran plates at different times). The x-axis shows MAPD per chip against the Affymetrix (“external”) reference and the y-axis shows MAPD per chip against its in-batch (“internal”) reference. The batches all used DNA derived from the same 84 samples. The external reference used was from a superset of batch six run at Affymetrix. Filled markers indicate the median MAPD of each batch. Dashed lines indicate QC cutoffs.

The underlying reason for the observed batch variation is unknown. It may be due to different lab practices, equipment or reagent variation. Hence, such variation may occur within the same lab over time and so we strongly recommend that MAPD be systematically tracked to capture this. For uses that require very high-resolution detection of copy number changes, the reference may have to recomputed.

Effect of MAPD on Functional Performance

As a measure of performance, we simulated copy number gain and loss using samples that have variation in chromosome X. The simulation substituted $\log_2$ ratios in a random region in chromosome X of specified size (e.g., 100 kilobases, or kb) into a random 40 megabase (Mb) region of chromosome 2. Samples where chromosome X has known copy number different than two were used. Pseudo-autosomal regions were excluded from the selection region. The simulation used the samples described in the previous section to ensure that lab-to-lab variability was captured.

The sample set used 26 HapMap male samples repeated in five different batches (CN=1) and five repetitions of chromosome X copy number gain aneuploidies (CN=3, 4 and 5) in five different labs. These are the same samples used in the prior section, “MAPD and the Effect of the Reference.”

The simulation repeated the random substitution 100 times per sample. To remove any real copy number changes in the 40 Mb region, $\log_2$ ratios of unsubstituted markers were randomized.

For each random substitution, the Hidden Markov Model (HMM) using all default parameters in GTC 2.0 was used to compute copy number for each marker in the 40 Mb region. Forty Mb was chosen by considering the size of the smallest chromosomes.

Sensitivity was calculated on a per-chip basis by calculating the proportion of calls that matched the expected. False positive rates were similarly determined by calculating the proportion of calls in the unsubstituted regions that did not match two. Finally, the median of the 100 simulations allowed a per-chip measurement of sensitivity and false positive rates.

There are two different situations: using an internal reference or using an external reference. Figure 2 shows sensitivity and false positive rates using a 100 kb substitution simulation where $\log^2$ ratios were calculated using a site-specific reference.

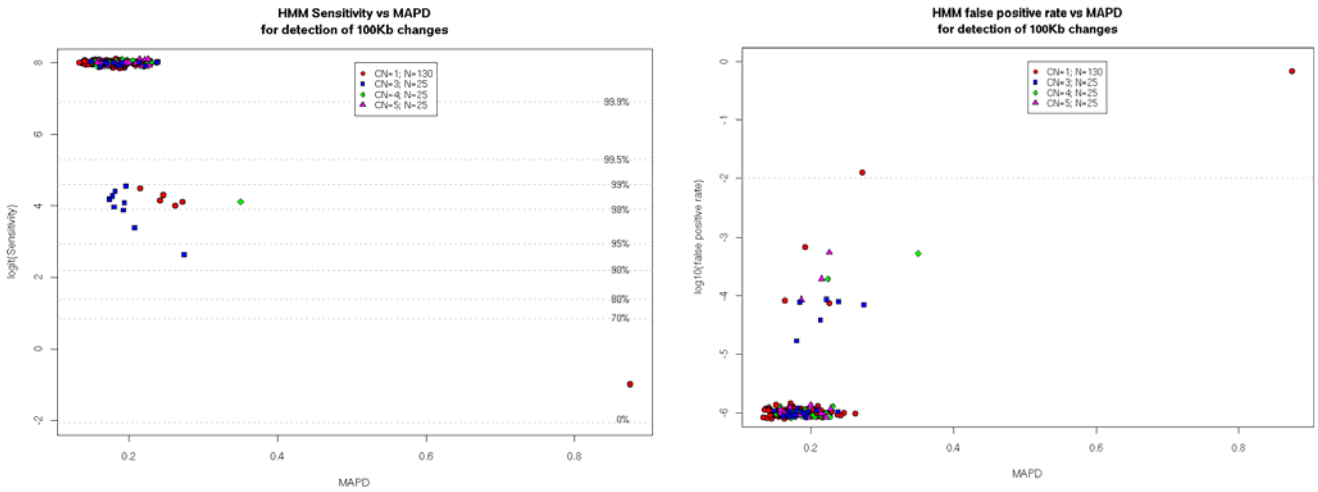

Figure 2: Logit (sensitivity) is plotted against MAPD on the left. Dashed lines show different levels of sensitivity. Log_{10} (false positive rate) is plotted against MAPD on the right, using an internal reference. The dashed line shows a false positive rate of 1 percent. Logit values greater than 8 were trimmed to 8 and the points jittered to show them. Log_{10} values less than -6 were trimmed to -6 and the points jittered to show them.

Figure 2 demonstrates high performance in the simulation. In general, distinguishing CN=3 from expected CN=2 is the most difficult case, but even here the sensitivity for the 9/25 chips ranges from almost 95 percent to almost 99 percent. Five out of 130 CN=1 vs. CN=2 chips have slightly lowered performance. All chips with MAPD less than 0.3 show good performance for CN=1 vs. CN=2, or CN>=4 vs. CN=2. There is one outlier with MAPD of about 0.9; this chip completely failed to show high sensitivity and or low false positive rates.

If an external reference is used to calculate the log_2 ratios, then for a fixed region the size sensitivity will be lower and false positive rates will be higher. The simulation described above was repeated for a region size of 500 kb and the GTC 2.0 supplied external reference. Figure 3 shows the results.

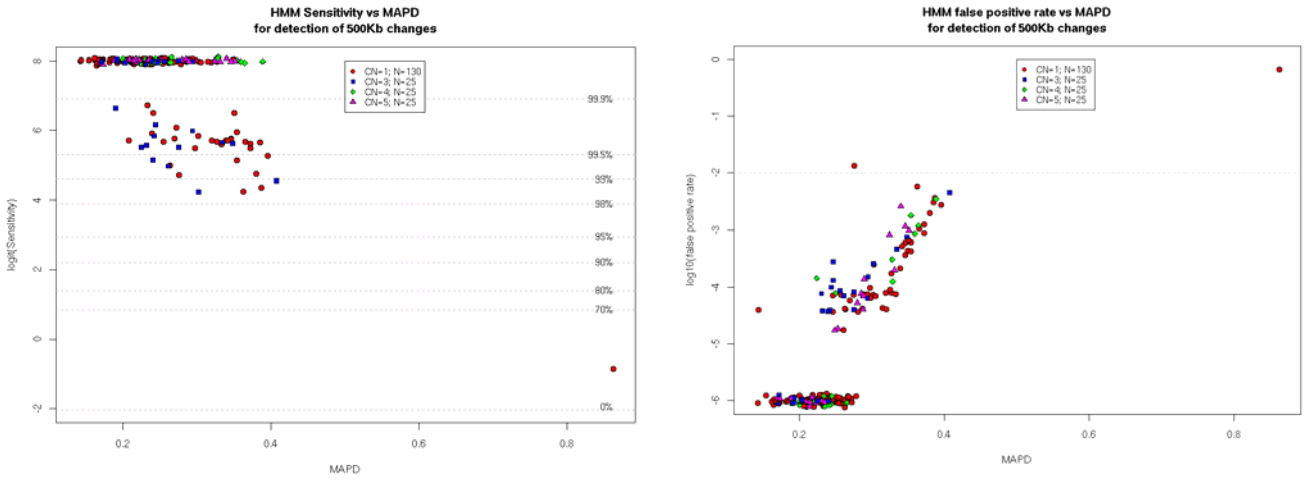

Figure 3: Logit is plotted against MAPD on the left. Dashed lines show different levels of sensitivity. Log_{10} is plotted against MAPD on the right, using an internal reference. The dashed line shows a false positive rate of 1 percent. Logit values greater than 8 were trimmed to 8 and the points jittered to show them. Log_{10} values less than -6 were trimmed to -6 and the points jittered to show them.

Figure 3 shows the performance using 500 kb regions with an external reference. As before, CN=3 vs. CN=2 is the most challenging case with 13 out of 25 chips ranging between 98 percent and 99.9 percent sensitivity. The false positive rates are considerably higher than in the previous simulation, but all except one chip plus the obvious outlier have false positive rates lower than 1 percent. Extrapolating the trend seen in the false positives, MAPDs above 0.4 are likely to have false positive rates greater than 1 percent.

Influence of Reference Size on MAPD

The log_2 ratio for each autosomal marker in a sample is calculated with respect to the reference signal, which itself is calculated as the median signal over all samples in the reference. X chromosomal markers only use samples determined as having at least two X chromosomes to make the reference; Y chromosomal markers only use samples determined to have at least one Y chromosome. The possibility of aneuploidy in the sex chromosomes is assumed to be low frequency and is ignored. Note that taking medians will be robust against such sample outliers.

However, since the reference itself is made up of samples, it too contributes to noise in log_2 ratios. We investigated the effect of sample size in the reference using a simulation in which the log_2 signal is normally distributed with a sample independent but the marker dependent is mean and a sample is specific to variance common to all markers. The assumption that the expected log_2 signal is independent of the sample is reasonable for internal references, which means that the log_2 ratio of each marker has an expectation of 0. The assumption of independent normals with common variance is consistent with observed overall distributions as discussed in the section, “MAPD related to IQR and genotyping call rate.”

In GTC 2.0, all samples are included in the reference when an internal reference is used. For small sample sizes, computed MAPD of each sample will underestimate the true variability. To see this: if the reference size is 1, then median signal in the reference will match the signal of markers in the sample, so MAPD will be 0. As part of the simulation, we also investigate the effect of using small sample sizes for internal references on MAPD.

Figure 4 shows a simulation of MAPD against internal reference size.

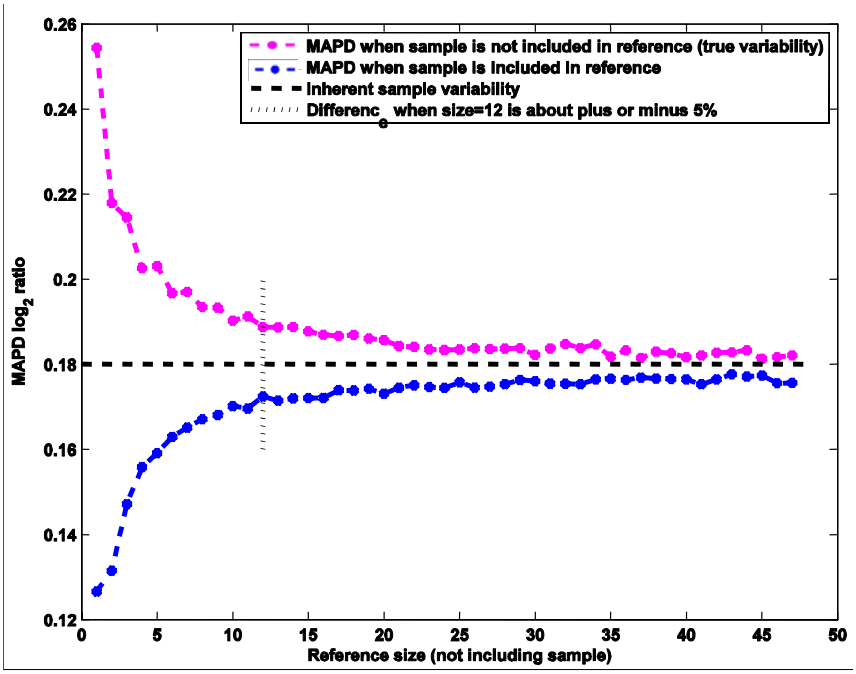

Figure 4: A simulation of sample-specific MAPD for a sample against reference size showing the relationship between reference size and MAPD. The simulation assumes that all markers are independent and have $\log_2$ signals normally distributed such that expected MAPD is 0.18; 0.18 is close to the typical sample variability in the GTC 2.0 reference.

In this simulation, for reference size 12, MAPD is underestimated by about 9 percent. True MAPD is a composite of residual variability in estimating the reference plus the inherent sample-specific variability. For larger reference sizes such as 22 and above, the underestimate is less than 5 percent. Accordingly, we recommend at least 12 of each gender when computing an internal reference as variability of markers in the sex chromosomes is determined by the number of samples of that gender.

For very small reference sizes such as 2-6, the underestimate of MAPD is quite severe, ranging from around 50 percent to around 20 percent. If small reference sizes are used, then MAPD should be treated with caution.

MAPD Related to IQR and Genotyping QC Call Rate

MAPD is closely related to the interquartile range of the $\log_2$ ratios (IQR). Under the assumption of $\log_2$ ratios being distributed as independent normals with a mean of 0, variance ^{2}, then IQR is 1.35 . If X_i is the $\log_2$ ratio for the marker at the i^{th} sequential position in the genome and if $Y_i=X_i - X_{i-1}$, Y is also distributed normally with mean 0 and variance 2 ^{2} and the ratio of IQR to MAPD is 1.41. Figure 2 illustrates the pairwise scatter plots between IQR, MAPD and logit transformed DM call rates using the QC SNPs (this is the single-chip metric used in SNP genotyping).

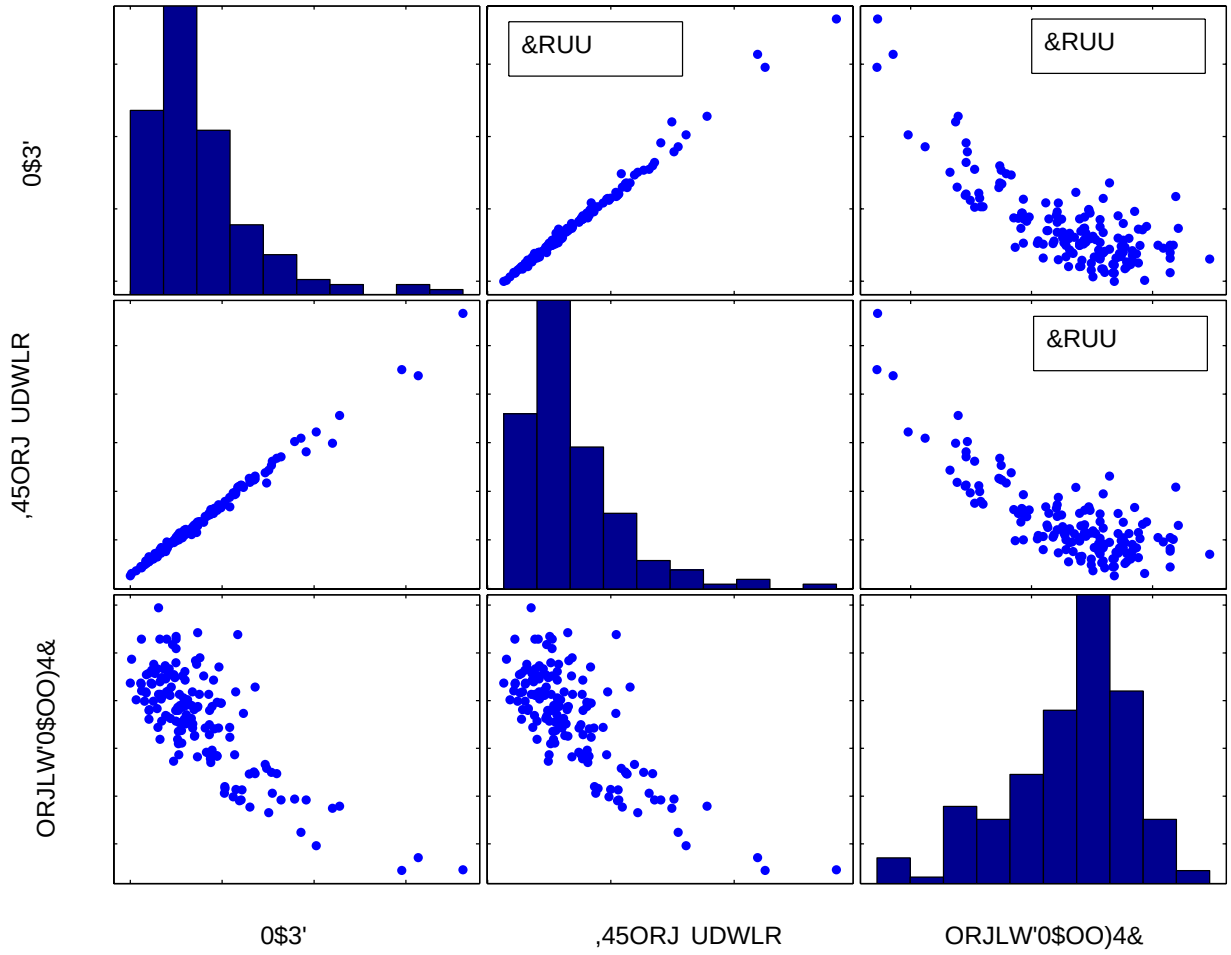

Figure 5: QC metrics for 158 samples used in a manufacturing study are shown here in a pairwise scatter plot. Observed correlation between each QC metric is shown in each panel. IQR and MAPD was calculated with respect to the HapMap 270 reference provided in GTC 2.0. All chips used the same DNA source (a normal sample with expected copy number 2) so sample source variability is not included. SNP Genotyping QC call rates are transformed using logit prior to being plotted.

Figure 5 illustrates that MAPD and IQR are both tightly correlated and linearly related. A regression of IQR on MAPD has a y intercept of -0.001 (not significantly different at the 0.05 level from 0), suggesting direct proportionality. Furthermore, the expected ratio of IQR/MAPD=1.41 under the assumption of independent, identically distributed normality is quite closely matched by 1.43=the empirical median across chips of these ratios, suggesting that overall distributional assumptions are roughly correct. Calculating per-chip lag-1 autocorrelations gives values of 0.1 or less, a fairly minor violation of the assumption of independence between X_i and X_{i-1}.

The SNP genotyping QC call rates range from around 85 percent (logit (0.85)=1.7) to around 99 percent (logit (0.99)=4.6). These call rates correlate quite well with MAPD and IQR as the quality goes down (roughly below 95 percent or logit (0.95)=2.9). The correlation in this range is -0.86. For high-quality data above this range, these metrics are less well correlated (correlation=-0.22). 75 percent of the chips fall into this high-quality range.

MAPD Related to Coefficient of Variation (CV)

A final comment about MAPD is in order. If the $\log_2$ ratios are normally distributed, then MAPD will be quite close to the SD of these ratios.

The Coefficient of Variation $CV = \frac{\sigma}{\mu}$ is a measure of overall relative error where is the SD and is the mean of each marker. If $\hat{\mu}$ is the estimate of typical signal from the reference then using a first-order Taylor expansion:

$$\log_2\left(\frac{x}{\hat{\mu}}\right) \approx \log_2\left(\frac{\hat{\mu}+(x-\hat{\mu})}{\hat{\mu}}\right) = \log_2\left(1+\frac{(x-\hat{\mu})}{\hat{\mu}}\right) \approx 1.44\frac{(x-\hat{\mu})}{\hat{\mu}}$$

(Which assumes that the difference from the reference is relatively small relative to the reference). In this case the $\log_2$ ratio is distributed similarly to the relative error inflated by 1.44. Hence, the SD of the $\log_2$ ratio will be closely related to 1.44 times the CV, and hence be similar to 1.38 times the MAPD. A typical MAPD of 0.18 in a high-quality experiment using an internal reference corresponds to a CV of around 13 percent.

The approximation described above is more inaccurate for large MAPDs, but using them gives a MAPD of around 0.3 corresponding to a CV of around 22 percent and a MAPD of 0.4 corresponding to a CV of around 30 percent.