WhitePaper

affymetrix
Biology for a better world

Microarray normalization using Signal Space Transformation with probe Guanine Cytosine Count Correction

Introduction
Gene expression analytical tools have been extensively developed over the past two decades for microarray data.

Standard expression microarray analysis consists of a number of different steps to enable signal comparisons between experimental conditions. The first step entails normalizing the signal intensity distributions of all probe features on all arrays. A probe feature is a location on the array that contains many copies of the same 25-mer DNA sequence. Normalizing these distributions enables comparison of probe signal between groups. Next, the signal intensities for all probes in a probe set that defines a gene or exon are aggregated into a single value for each array or sample. The aggregate signal value is used to compare gene-level or exon-level expression changes between sample groups or conditions.

A gene-level or exon-level probe set is determined as being differentially expressed when the normalized aggregate signal value is significantly different between two groups of samples. The significance of the change in signal for a gene-level or exon-level probe set between two conditions is often determined using a T-test. The T-test provides a measurement of confidence in the fold change estimate between conditions in units of standard deviation. If the replicate measurements for a probe set in a group of samples is more variable, the resulting T-test score will be lower. A lower T-test score indicates a lower confidence in the fold change estimation. The probability value (p-value) of the T-statistic can be calculated for every probe set using a T-distribution. This p-value provides a measure of significance in the estimated value, which is often used to sort and filter data. Therefore, it is important to consider both p-value and fold change when applying filter criteria to differential expression data.

The use of both fold change and p-value was reinforced in the summary of the Microarray Quality Control consortium publication in 2006.[1] There they noted that "fold change ranking plus a non-stringent p-value cutoff can be used as a baseline practice for generating more reproducible signature gene lists." This conclusion was reinforced by an article published in *The Scientist* that warned researchers of the dangers in relying on fold change alone and emphasized the importance of using robust statistical cutoffs such as p-value.[2] Utilizing fold change and p-value is extremely important when directly comparing expression results using different platforms. The direct comparison of statistically significant differential expression presents no problem. However, comparing the magnitude of fold changes, especially when the same cutoff is applied to both data sets, will confound the comparison.

Applying the same fold change cutoff to two different technologies confounds the comparison because the technology could relate differences on entirely different scales. A microarray uses globally amplified total RNA, which is then hybridized to probes on the array. The signal is measured from a scanned image of the fluorescent stain present at each feature and is summarized across the pixels for each probe area. This "system" is fundamentally different than other technologies such as RT-PCR and RNA sequencing (RNA-Seq), so the signal space utilized by each technology is fundamentally different. The Guanine Cytosine Count Normalization (GCCN) and Signal Space Transformation (SST) algorithms adjust CEL file intensities, allowing inter-platform comparisons. The GCCN algorithm normalizes the intensities with respect to the difference in probe affinity associated with Guanine and Cytosine (GC) content. The SST stretches the intensity distribution to a common range with a power law mapping that "decompresses" the fold change ratios. Together these transformations permit the more direct comparison of microarray expression results to those of other platforms.

Fold change compression has always existed with microarrays

A systematic property of most microarray expression results is the underestimation of overall fold change, often termed "fold change compression." As early as 2003, in Latin square spike-in studies, Rajagopalan[3] characterized three different signal summary methods and found that for all three, the ratio of signals resulting from microarrays was always lower than the corresponding ratio of actual transcript spikes. In 2005, Choe, *et al*.[4]

was one of the first to note the compression relative to RT-PCR [4] and that the cause was only partly the result of signal saturation. They also note that the compression is not a byproduct of the signal summarization.

Also in 2005, Dallas, *et al*.[5] published a detailed comparison between GeneChip® Human Genome U133A 2.0 Array and RT-PCR. In Figure 2 of their paper, they show a strong Pearson's correlation between fold changes; however, the magnitude of the fold change observed in the microarray data was compressed. In the 2006 MAQC paper,[6] a compression in the log ratio or fold change comparing GeneChip® Human Genome-U133 Plus 2.0 Array to TaqMan® data can be seen in the analysis, although it was never noted by the authors.

Using fold changes to compare technologies can be misleading if one does not correct for compression effects

Simply comparing fold changes using different technologies without adjusting the scale of measurement may result in over- or underestimation of fold changes. As described in the SEQC/MAQC-III Consortium study in 2014,[7] whole-transcriptome profiling technologies, such as microarray and RNA-Seq, provide relative, not absolute, expression measurements. As relative measurements, it is important to ensure that the scales used to compare these relative measurements are comparable.

Over- or underestimation of fold changes occurs depending on the technology used to measure expression. Affymetrix microarrays have traditionally underestimated the fold change, and conversely, other technologies such as RNA-Seq methods can overestimate fold changes. *"RNA-Seq produces a wide range of read counts per gene, and genes with a low coverage of reads can produce artificially high fold change values*."[8] Low-coverage results for RNA-Seq have high variance due to poor statistics.

A common caveat of comparing gene expression differences between technologies is observed when filtering results using a fold change cutoff. A fold change cutoff of 2 is often used to select a list of genes considered to be differentially expressed. This value of 2 is arbitrary and chosen based primarily on tradition and the unreasonable notion that low fold changes are not biologically relevant. As mentioned previously, it is more appropriate to select differentially expressed genes on a statistical basis, thus avoiding the need for fold change adjustments.

Figure 1a shows typical fold change correlation between RNA-Seq and microarrays which are well correlated (Pearson r = 0.879). The slope of the linear regression is less than 1 due to the compressed fold changes observed with the array data on the y-axis. By using the same fold change cutoff of 2 for each platform, you would conclude that RNA-Seq identifies far more differentially expressed genes than microarrays (as shown in the Venn diagram). For the purpose of illustration, having determined the slope to be 0.446, one can simply divide all of the array data values by the slope and correct for the differences in fold change observed between the platforms (Figure 1b). This correction results in both platforms identifying a comparable number of differentially expressed genes. This simple linear correction, while not useful in practice, demonstrates that the effects of the compression can be moderated.

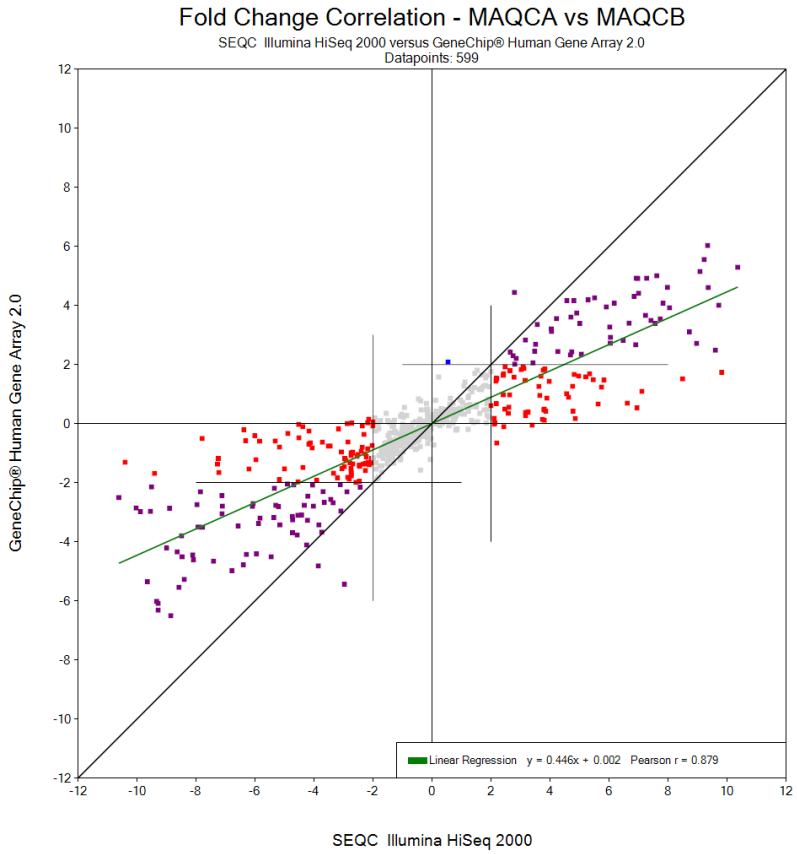
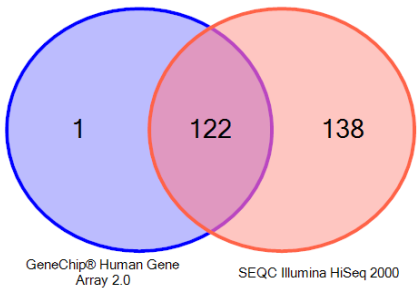


Figure 1a: Fold change correlation comparing RNA-Seq differences on the x-axis to microarray differences on the y-axis. Data generated from SEQC comparing MAQCA and MAQCB.[1]

- FC >2 for both RNA-Seq and microarray
- FC >2 for RNA-Seq only
- FC >2 for microarray only



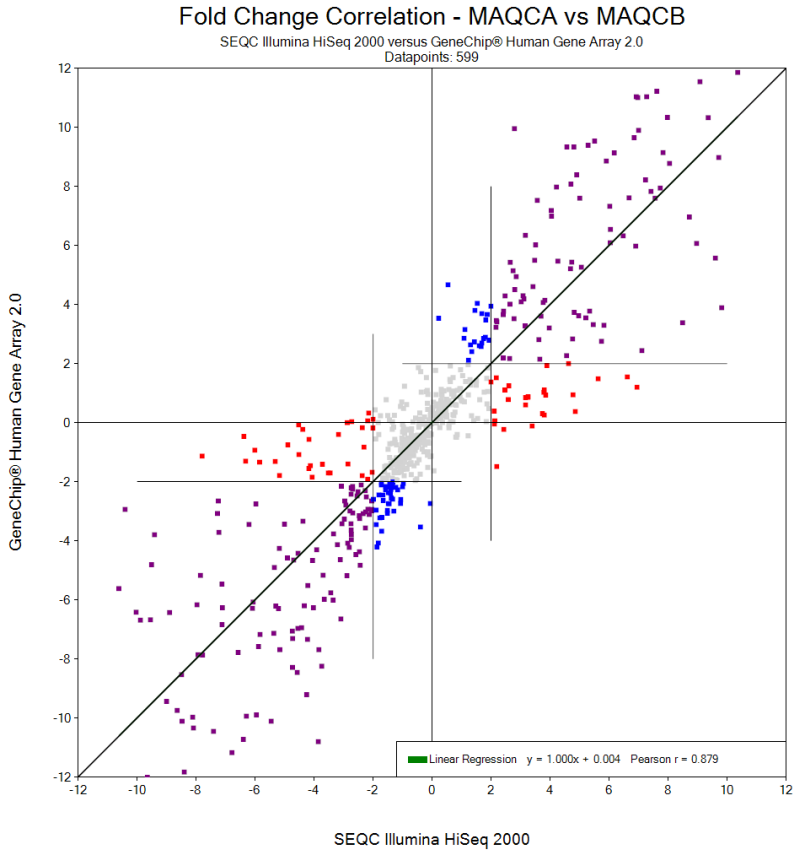
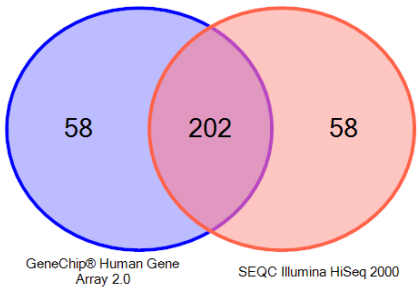


Figure 1b: After correcting for slope. Fold change correlation comparing RNA-Seq differences on the x-axis to microarray differences on the y-axis. Data generated from SEQC comparing MAQCA and MAQCB.[1]

- FC >2 for both RNA-Seq and microarray
- FC >2 for RNA-Seq only
- FC >2 for microarray only



Causes of fold change compression

Signal and fold change compression occur when a platform does not respond in a linearly proportional manner to concentration. The important issue is whether the response is sufficiently predictable that reliable and statistically significant measurements can be made.

Signal intensity values coming from microarrays are in fact a combination of signal and background. Signal is the sum of the "true" hybridization plus any cross-hybridizing signal. Background is typically thought of as non-specific localized effects. Summarizing pixel intensities will never yield a zero value for a non-hybridizing feature due to background effects. Factored into the signal measure are probe-specific effects. The largest measurable probe effect is guanine-cytosine (GC) content of the probe sequence. A given GC pair is bound by three hydrogen bonds, while AT pairs are bound by two hydrogen bonds. Probes with higher GC content are more stable than those with low GC content primarily due to stacking interactions. This resulting higher thermodynamic stability during hybridization results in faster capture rates and slower off rates compared to higher AT probes. The overall effect is that for the same molecular target, a higher GC probe will be brighter than an AT probe.[9] These effects are commonly observed in raw DNA copy number data before covariate adjusting removes waviness.[10] The combination of background signal and probe effects can reduce the overall fold change for a given comparison. This effect occurs primarily for cases where at least one of the values being compared is relatively close to the background measurement. Consider the examples in Figure 2. When we compare the difference seen in Example 1 where we have a strong background or probe effect, M^2 is 37% larger than M^1. If we reduce the background or probe effect, however, we get a larger difference. In Example 2, M^2 is 60% larger than M^1.

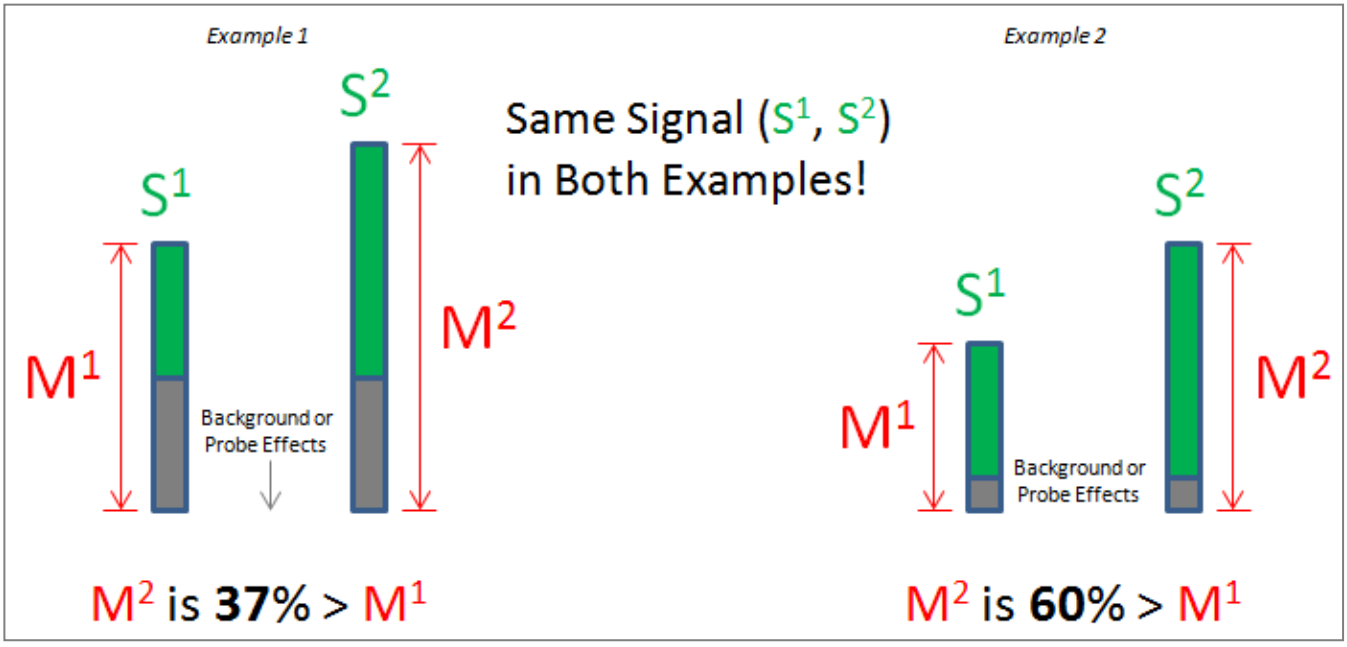


Figure 2: The overall measurement (M) is a combination of Signal (S) and background. Better removal of background and probe effects can reduce the difference observed when comparing measurements.

Saturation of the feature on the microarray or of the imaging system can also cause the response to no longer be linearly proportional to the concentration of the target. This cause of signal compression applies to high concentrations.

GCCN description

Probe designs for microarrays attempt to constrain the GC content of probes in order to optimize hybridization conditions. However, large-format, high-density microarrays typically contain millions of probes spanning a wide range of GC content.

GCCN was developed to normalize the signal of probes within a particular GC count to an intensity distribution that closely matches the intensity distribution for the set of probes with GC count of 12. Before GCCN is applied, probes with lower GC count tend to have lower intensities, and probes with higher GC count tend to have higher intensities. After GCCN, the set of probes with any given GC count has approximately the same distribution as the probes with GC count of 12 (Figure 3).

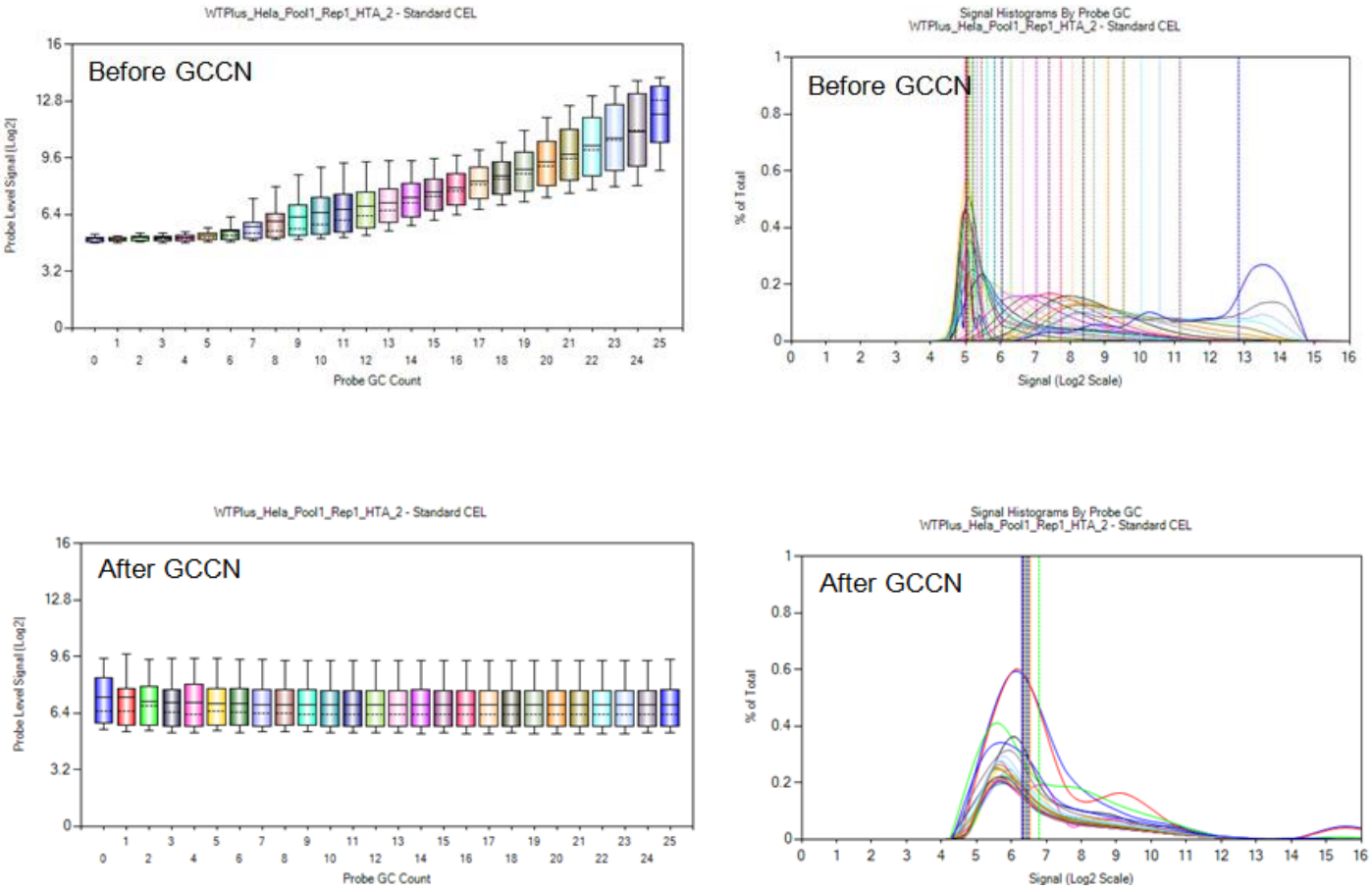


Figure 3: Box plots and histograms of raw probe intensity for probes by GC count, before and after GCCN processing. All probe intensities taken from a single GeneChip® Human Transcriptome Array (HTA) 2.0 CEL file.

GCCN algorithm description

The intensity distribution of probes in each bin is computed, and the intensity distribution for the GC content equal to 12 is selected as the reference distribution. For each probe, its percentile within the intensity distribution of its bin is assessed, and its intensity is matched to the equivalent percentile of the reference distribution. The bins representing extremely low or high GC contents have low occupation and require some special handling. The flowchart in Figure 4 depicts the algorithm.

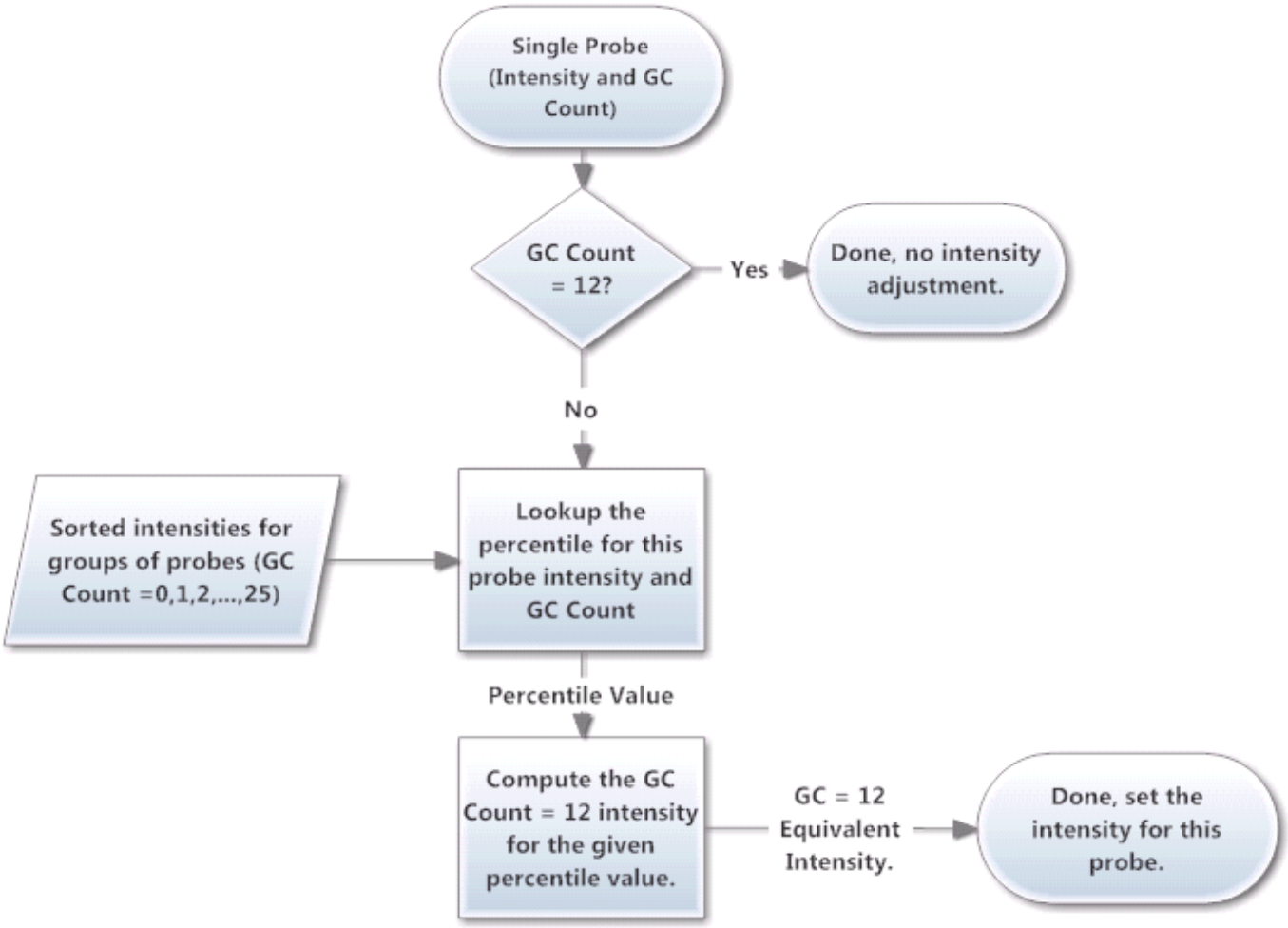


Figure 4: Flowchart for the processing of a single probe by the GCCN algorithm.

Figure 5 illustrates the concept of looking up a percentile value for a given probe intensity and GC count. In practice, we do not have the green intensity profile curve shown in Figure 5. In order to find the percentile associated with a given probe intensity, a binary search is undertaken. The lookup algorithm iteratively approaches the correct percentile value, as illustrated in Figure 6. In this manner, the percentile value may be calculated with an arbitrarily small error. With the percentile value in hand, it is a simple matter to compute the associated GC count = 12 intensity. This procedure is illustrated in Figure 7.

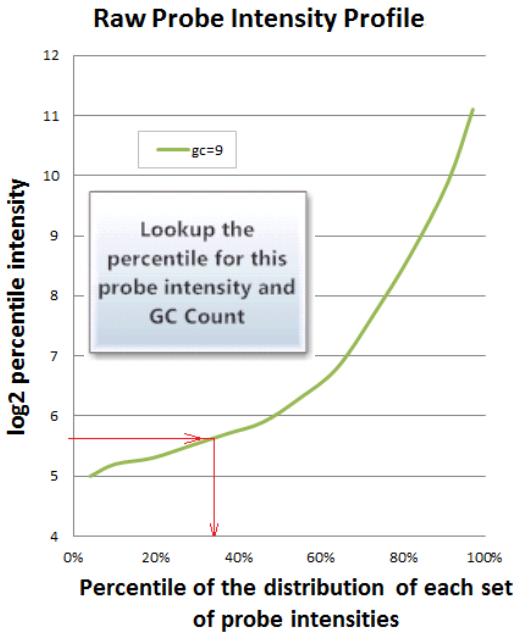


Figure 5: Illustration of the concept of looking up the percentile of a probe's intensity.

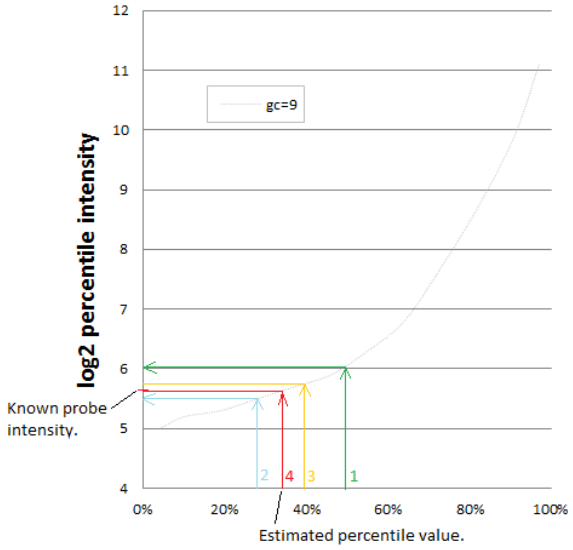


Figure 6: Actual binary search lookup of the percentile for the same probe.

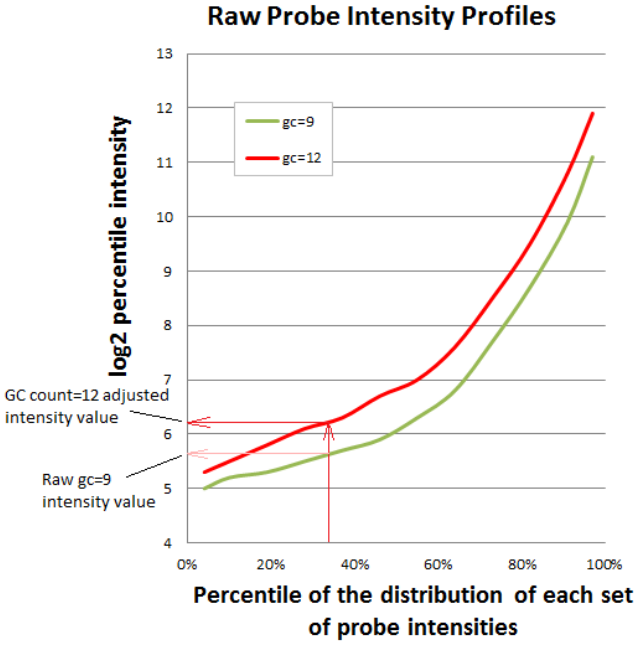


Figure 7: Computing the GC count = 12 adjusted intensity of a GC count = 9 probe.

The data for Figure 8 was created by sampling the intensity profiles for various GC counts for all probes in a single HTA 2.0 CEL file before and after GCCN. As expected, after GCCN, the intensity profiles for all sets of probes closely match the GC count = 12 intensity profile.

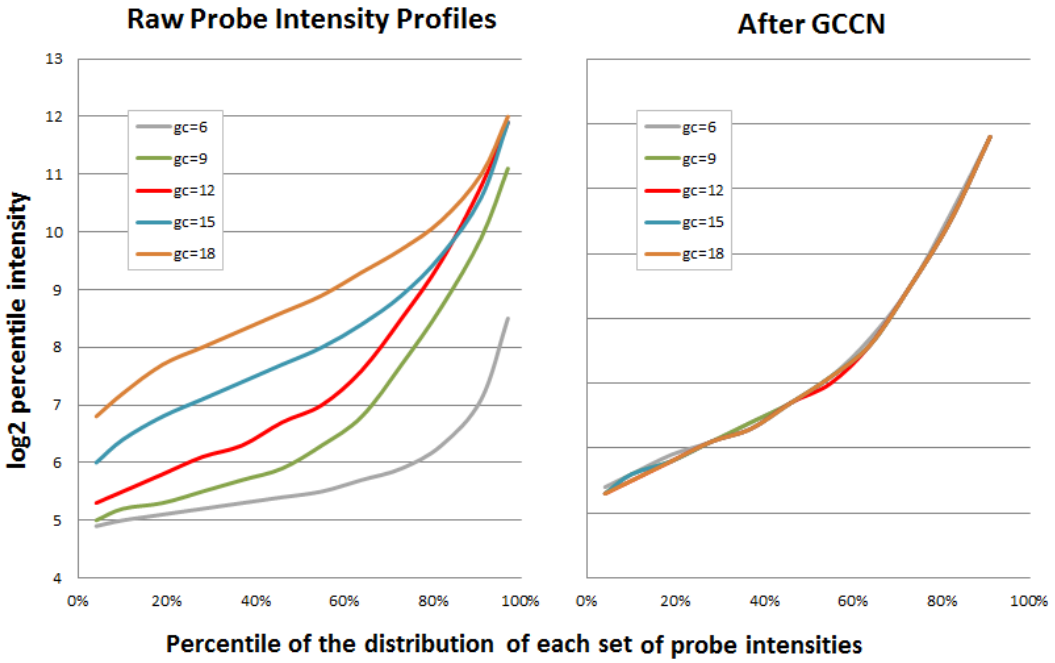


Figure 8: Probe intensity profiles before and after GCCN processing. All probe intensities taken from a single HTA 2.0 CEL file.

Performance of the GCCN algorithm can be evaluated by comparing the difference between two concentrations of known spike targets. Using a classic Latin square experimental design[11] containing 16 samples with 4 spike concentrations and 685 exon spikes, we can evaluate the absolute fold change detected. The spike concentrations are 0, 1:200K, 1:100K, 1:50K. These concentrations are equivalent to 0, 1, 2, and 4 copies per cell, respectively assuming 10 pg cell equivalents. Comparing the exon probe set difference between 1:100K spike sample vs. 1:200K spike sample should give a 2-fold difference, or in log base 2, a difference of 1. A 1:50K vs. 1:200K should give a 4-fold difference, or in log base 2, a difference of 2. While the GCCN correction does not by itself raise the fold change to the expected level, we do see an overall increase in the fold change (Figure 9).

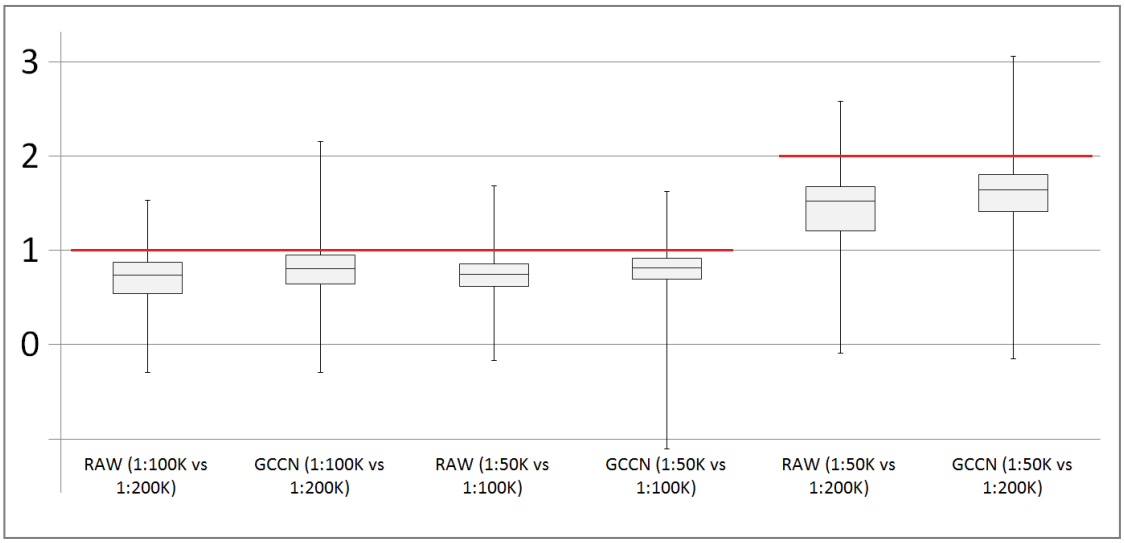


Figure 9: Distribution of exon fold change results in log2 space for various pairs of RNA target concentration in the Latin square. GCCN corrected CEL files were created from the RAW CEL files using Affymetrix® Power Tools: "apt-cel-transformer –c gc-correction" to produce a GCCN corrected CEL file followed by processing the RAW and GCCN corrected CEL files using "apt-probeset-summarize –a rma-sketch".

GCCN significantly reduces the number of exons exhibiting poor fold change response (Figure 10).

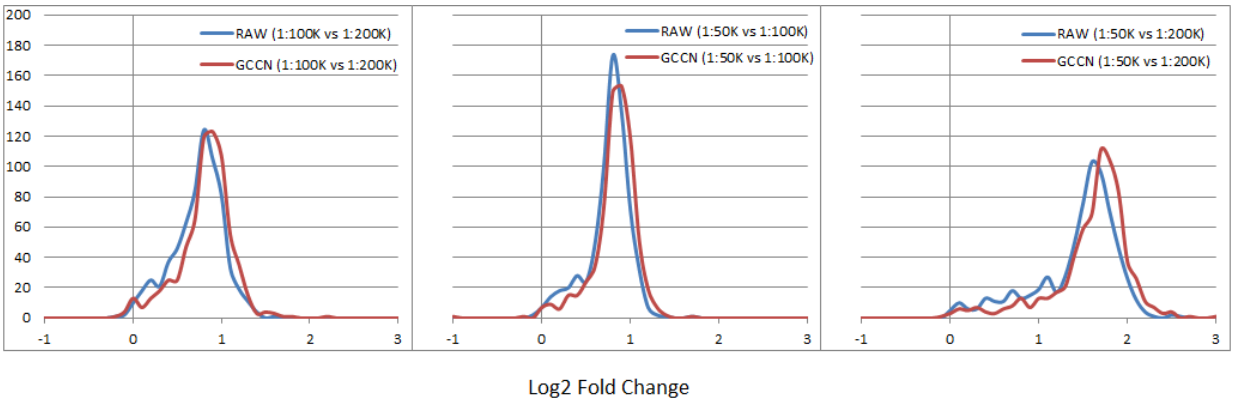


Figure 10: Histogram of exon fold change results in log2 space for various pairs of RNA target concentration in the Latin square.

In order to determine the effect of GCCN on fidelity or sensitivity and specificity of HTA 2.0, we can compare the distribution of T-statistics for the relative concentrations to that of a null distribution (in this case, everything else on the array). Using a Receiver Operator Characteristic (ROC) curve, we can express the fidelity as an Area Under the Curve (AUC). A perfect AUC of 1 would mean there is complete separation of the T-statistics for the spikes from the null distribution within the comparison. For all three cases shown in Figure 11, GCCN noticeably improves ROC AUC values.

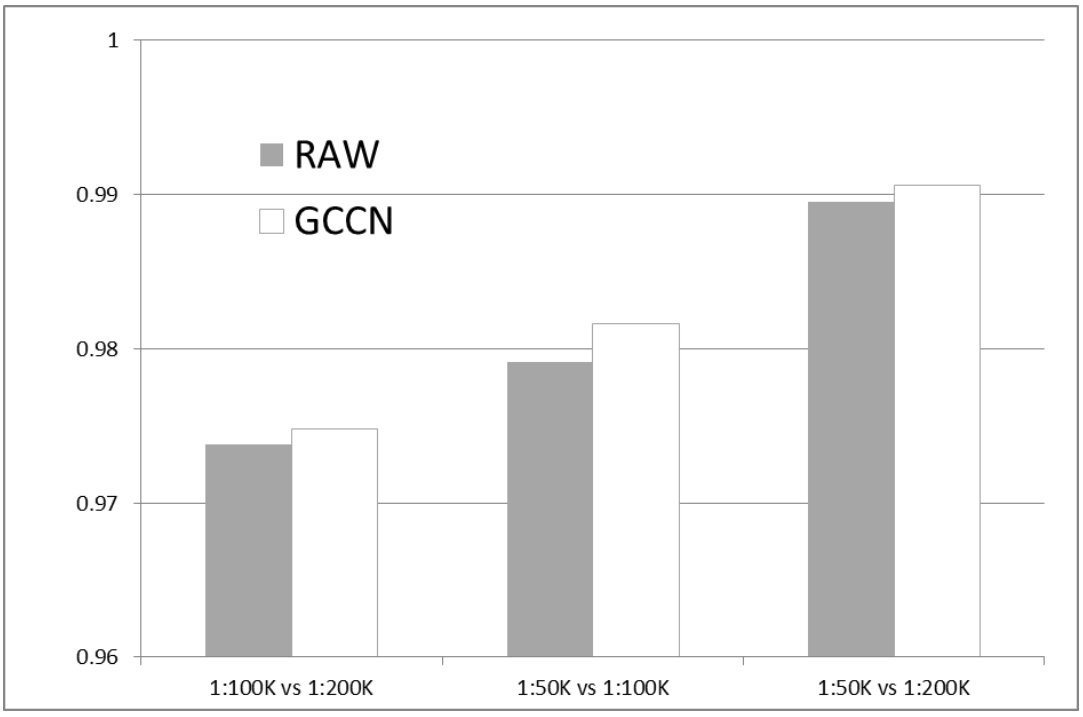


Figure 11: ROC AUC values for exon spike-ins, RAW and with GCCN.

SST description

SST is an algorithm that translates probe intensities in log space using a number of tuning parameters to simplify fold change comparisons between different technologies.

When properly tuned, SST will adjust Affymetrix' expression array probe intensities so as to eliminate significant fold change compression with minimal degradation of other array performance metrics.

The SST algorithm takes four parameters.

Parameter	Description
floor	Minimum allowed value for CEL intensities in the target distribution (default is 1).
ceiling	Maximum allowed value for CEL intensities in the target distribution (default is 1,000,000).
low	SST transforms the 2nd percentile intensity of the target distribution to this value (default is 20).
high	SST transforms the 98th percentile intensity of the target distribution to this value (default is 50,000).

To normalize the scale of measurement, the SST algorithm stretches the distribution according to a power law with an exponent greater than 1 while preserving the rank order of the intensities and the overall shape of the distribution (Figure 12). The intensities from the 2nd percentile to the 98th percentile will have matching ranges for different arrays. The power law nature of the transformation increases higher-intensity values more than lower-intensity values such that their ratio is uncompressed.

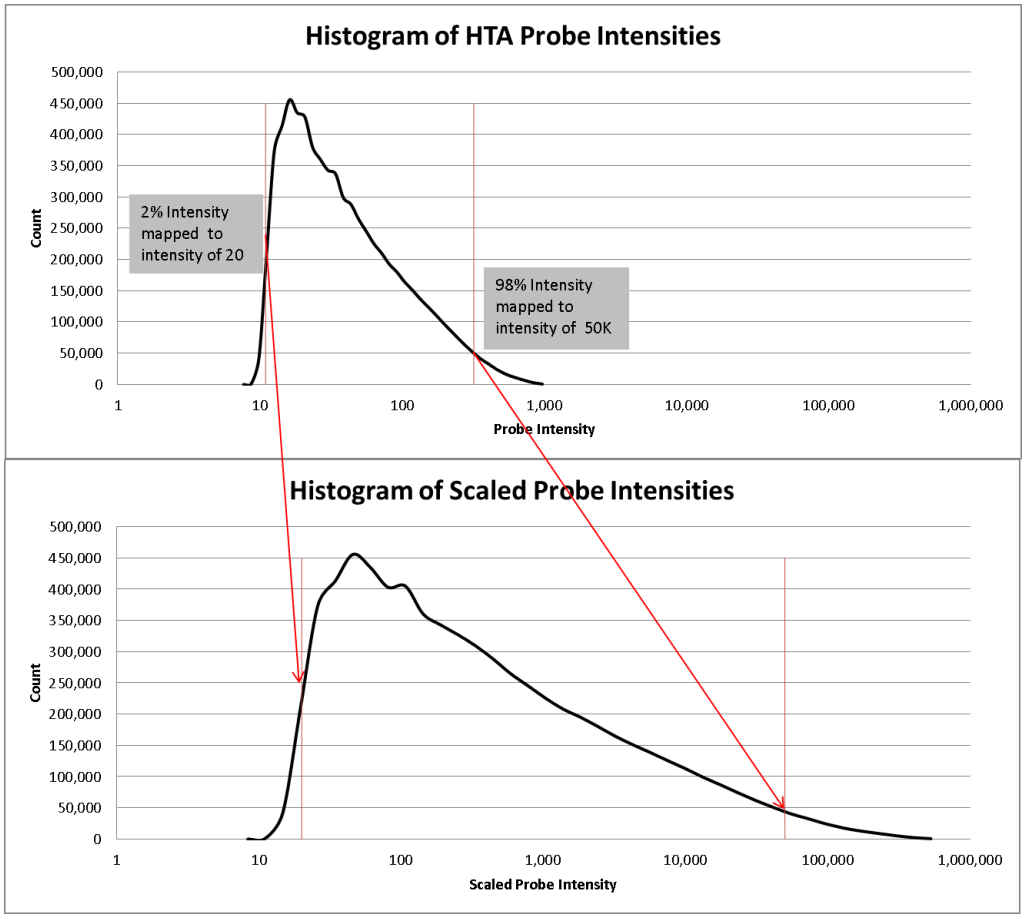


Figure 12: Illustration of the SST scaling algorithm.

Performance of the SST algorithm can be evaluated by comparing the difference between two concentrations of known spike targets. Using a classic Latin square experimental design containing 16 samples with 4 spike concentrations and 685 exon spikes, we can evaluate the absolute fold change detected. The spike concentrations are 0, 1:200K, 1:100K, 1:50K. Comparing the exon probe set difference between 1:100K vs. 1:200K spike samples should give a 2-fold difference, or in log base 2, a difference of 1. A 1:50K vs. 1:200K should give a 4-fold difference, or in log base 2 a difference of 2. In all comparisons, the SST correction raises the median fold change to the expected level (Figure 13).

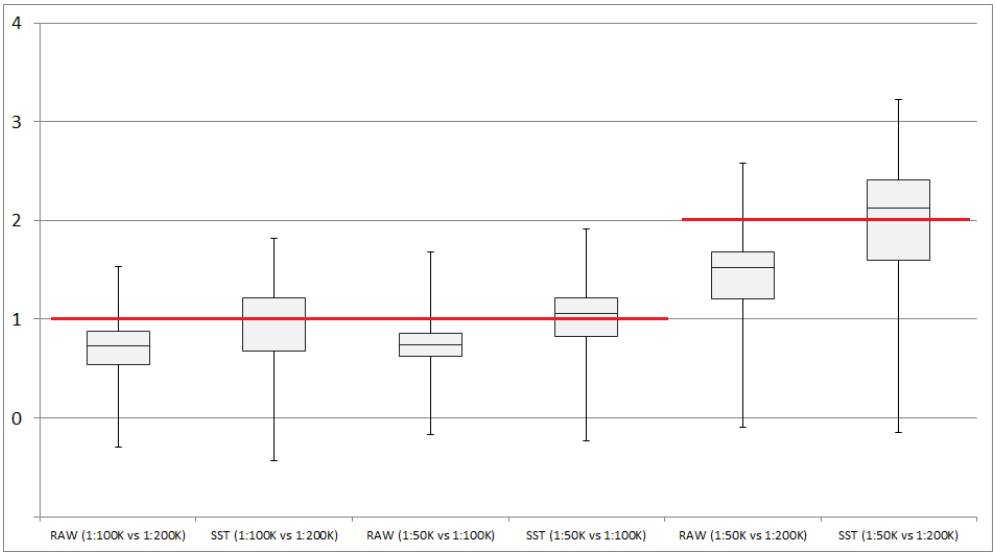


Figure 13: Distribution of exon fold change results in log2 space for various pairs of RNA target concentration in the Latin square. SST corrected CEL files were created from the RAW CEL files using Affymetrix Power Tools: "apt-cel-transformer –c scale-intensities" to produce an SST corrected CEL file. After the CEL files were transformed, both the original (RAW) and corrected (SST) CEL files were then processed using "apt-probeset-summarize –a rma-sketch". Red lines indicate the fold change expected for the associated pair of RNA target concentrations.

In order to determine the effect of SST on overall performance of HTA 2.0, we take the same Latin square of 16 CEL files and run ROC analysis. The 1:100K vs. 1:200K fold change is more difficult to discriminate from the background; hence, the ROC AUC for this case is the lowest. For all three cases shown in Figure 14, changes to ROC AUC values are insignificant. In summary, we are increasing the fold change and maintaining the fidelity and our ability to detect expression differences at low exon concentrations.

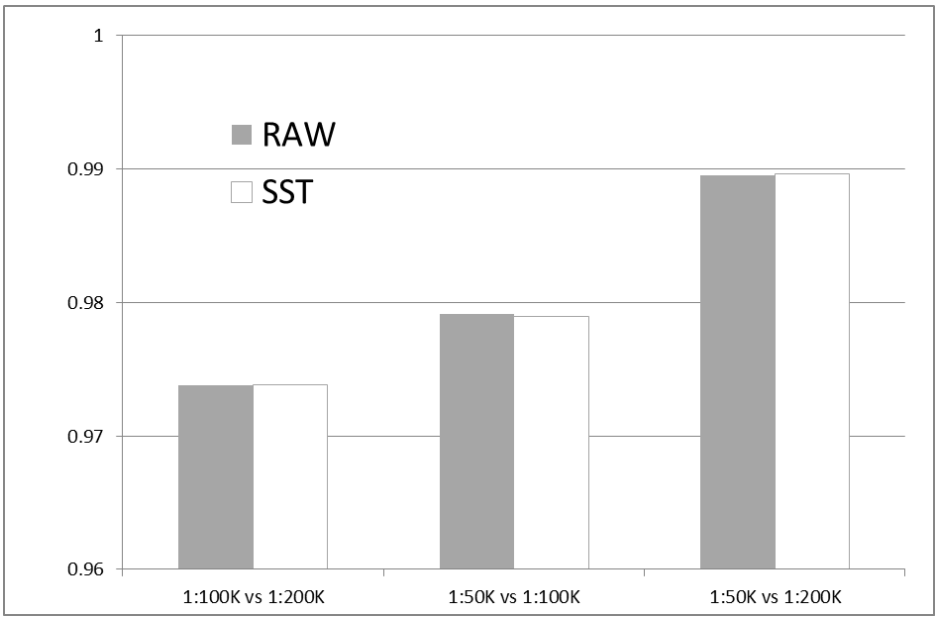


Figure 14: ROC AUC values for exon spike-ins, RAW and with SST.

SST algorithm

The intensity of each probe is adjusted using the following equation:

$$I_{out} = low \times B^{ln(I_{in})-ln(I_{2\%})}$$

where

I_{out} = Output probe intensity.
I_{in} = Input probe intensity.
B = Base for exponential expansion.
low = SST fixes the 2% intensity of the target distribution at this value. Default is 20.
$I_{2\%}$ = 2nd percentile intensity of the input intensities.

Clearly, $I_{out} = low$ when $I_{in} = I_{2\%}$.

Rewriting the equation as

$$I_{out} = low \times \left(\frac{I_{in}}{I_{2\%}}\right)^{lnB},$$

it becomes clear that the transformation reflects a power law. The floor and ceiling are applied to I_{out} as saturation limits: if $I_{out} < floor$, set I_{out} to $floor$ and if $I_{out} > ceiling$, set I_{out} to $ceiling$.

$\boldsymbol{B}$ is computed so that I_{out} is equal to $high$ when I_{in} is equal to $I_{98\%}$ (the 98th percentile intensity of the input distribution). Writing the equation in these terms we have

$$high - low = low \times B^{ln(I_{98\%})-ln(I_{2\%})}.$$

Solving for B we get

$$B = exp\left(\frac{ln\left(\frac{high-low}{low}\right)}{ln\left(\frac{I_{98\%}}{I_{2\%}}\right)}\right).$$

For each array, the intensity distribution is used to determine $I_{2\%}$ and $I_{98\%}$. Then, B is computed, and the power law transformation can be applied to each probe signal (with the exception of certain QC probes).

Combining GCCN and SST algorithms into the expression workflow

To minimize the impact on your current data analysis workflow, Affymetrix has provided two options for transforming data using the GCCN and SST algorithms. These options include the following:

- Expression Console™ Software (version 1.4.1 or later)
- Affymetrix® Power Tools (APT) (version 1.17.0 or later)

The GCCN and SST transformation algorithms have been built into Expression Console™ Software to minimize the impact to your analysis workflow. Researchers using third-party analysis tools such as Partek® Genomics Suite® or Bioconductor can take advantage of these algorithms by using APT to apply the GCCN and SST algorithms to their CEL files prior to importing them into a third-party analysis software package.

Figure 15: Expression array analysis workflow.

GCCN and SST can be combined upstream of a typical analysis workflow. APT includes "apt-cel-transformer.exe", which is a command line tool that transforms CEL files using chipstream algorithms. Transformed CEL files are written to a specified output directory. This command line tool may be used to run the GCCN and/or SST chipstreams on one or more input CEL files.

APT example: transform HTA CEL files using GCCN and SST using APT with default parameters
In this example we will apply the GC correction and the SST to CEL files, creating new ones with the same name in the "Output" directory.

```
apt-cel-transformer
 --pgf-file ..\HTA-2_0.r1.pgf
 --clf-file ..\HTA-2_0.r1.clf
 -c gc-correction,scale-intensities
 -o Output
 ..\CEL\*.CEL
 --qc-probesets HTA-2_0.r1.qcc
```

Note: The order of the transformations here is important. GCCN must precede SST. Changing the order will significantly degrade performance.

- The "apt-cel-transformer" is the APT command line function.
- The "--pgf-file" and "--clf-file" are required to identify array-specific library files.
- The "-c gc-correction, scale-intensities" adds the GC correction and the SST correction.
- The "-o Output" is optional and specifies the output directory so that the original files will not be overwritten. If the directory does not exist, it will be created.
- The "--qc-probesets HTA-2_0.r1.qcc" specifies the array-specific QCC file and is optional.

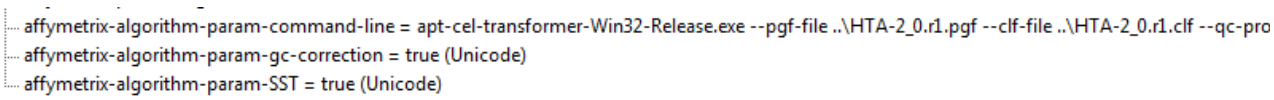
Using the "qc-probesets" option file is highly recommended. Control probes such as hybridization controls or target prep controls are useful because their relative abundance can be used as a performance metric. Historically these controls work because they have been designed to provide a ladder of response. Their design did not take into consideration SST transformation and the GC effects we can now remove based on the GC correction. Many QC pipelines will require these controls to remain unmodified so that the QC intensities remain comparable to legacy work. By using the "qc-probesets" option, you can leave the QC probes unadjusted. This allows for the expected ladder of target prep and hybridization controls to be maintained at the legacy levels. However, note that the reported probe and thus probe set values will be slightly different with and without using this option for the control probes. This is due to the downstream normalization and any additional background removal such as that applied when running GC-RMA. The command is utilizing the fourth column in the QCC file. If "quantification_in_header" is equal to "1," the probe set will not be adjusted by the GC correction of the SST transformation.

The QCC file is a tab delimited text file with the following columns:

- probeset_id (Matches the probe-set id in the pgf file)
- group_name
- probeset_name
- quantification_in_header

probeset_id	group_name	probeset_name	quantification_in_header
18678055	control->affx->bac_spike	AFFX-r2-Ec-bioB-5_at	1
18678061	control->affx->bac_spike	AFFX-r2-Ec-bioC-5_at	1
18678065	control->affx->bac_spike	AFFX-r2-Ec-bioD-5_at	1
18678069	control->affx->bac_spike	AFFX-r2-P1-cre-5_at	1
18678032	control->affx->polya_spike	AFFX-r2-Bs-dap-5_st	1
18678038	control->affx->polya_spike	AFFX-r2-Bs-lys-5_st	1
18678044	control->affx->polya_spike	AFFX-r2-Bs-phe-5_st	1
18678050	control->affx->polya_spike	AFFX-r2-Bs-thr-5_st	1
18678001	control->bgp->antigenomic	AFFX-BkGr-GC11_at	1
18678002	control->bgp->antigenomic	AFFX-BkGr-GC12_at	1
18678003	control->bgp->antigenomic	AFFX-BkGr-GC13_at	1

The apt-cel-transformer adds parameters to the CEL file header. The applied command line is actually written into affymetrix-algorithm-param-command-line. If GC-correction is included in the chipstream, then affymetrix-algorithm-param-gc-correction is added to the header. If scale-intensities is included in the chipstream, then affymetrix-algorithm-param-SST is added to the header. An excerpt from the CEL file header that includes both algorithms appears as follows:



If the transformations are attempted a second time on previously transformed CEL files, an error message will be displayed. Examples of the message appear below:

gc-correction has already been performed on your_data.CEL

sst has already been performed on your_data.CEL

Using GCCN and SST with the Expression Console™ Software workflow

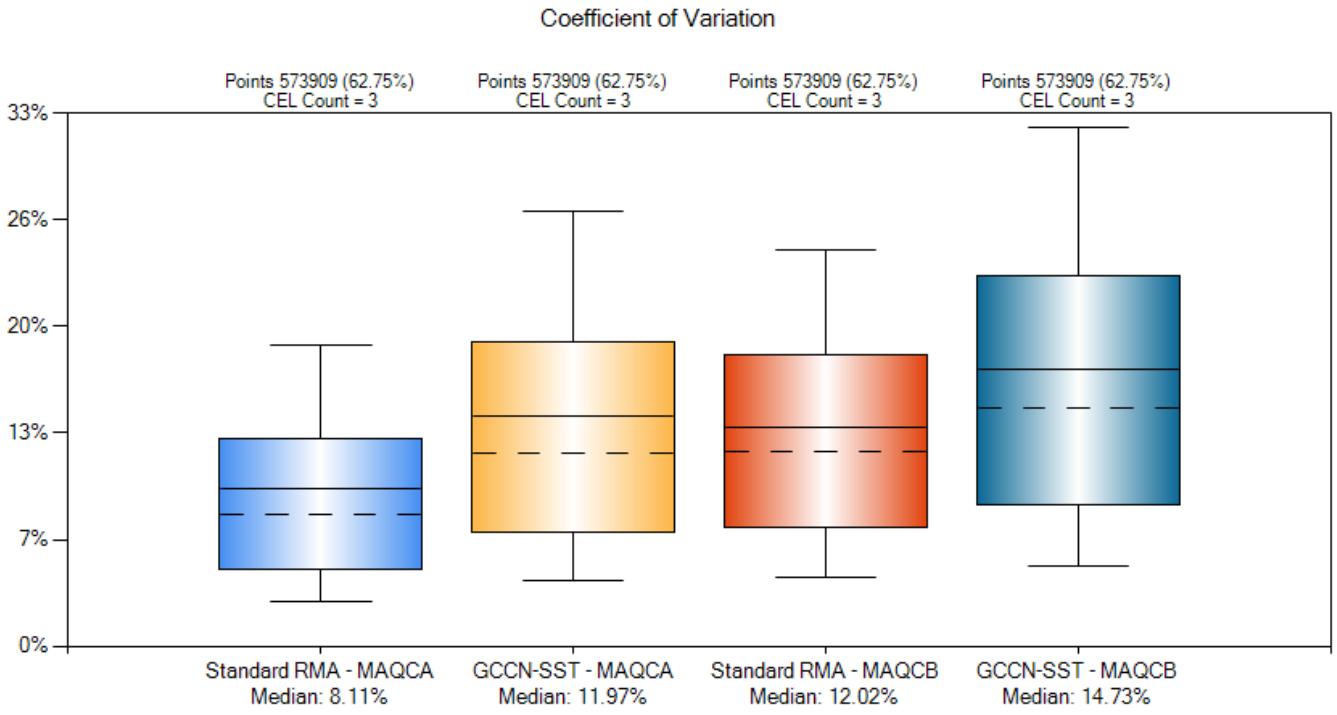
Adding GCCN and SST to the Expression Console Software workflow is managed through the array-specific analysis configuration file. For HTA 2.0, this file is "HTA-2_0.exon_analysis_configuration". Updated Expression Console array configuration files are available from Affymetrix for all transcriptome view arrays.
Note that the following parameters will exist in updated analysis configuration files and can be used with all Affymetrix' human, mouse, and rat transcriptome series of arrays:

#%affymetrix-algorithm-param-apt-opt-analysis-spec=gc-correction.algo_rev=4.txt_out=false,scale-intensities.floor=1.low=20.high=50000.ceiling=1000000.low_pct=0.02.high_pct=0.98.txt_out=false,rma-bg,quant-norm.sketch=-1.bioc=true.lowprecision=false. usepm=true.target=0. doavg=false.subsetmd5=cb592d36df3bc90aa197311f2735a579,pm-only,med-polish. FixFeatureEffect=false.UseInputModel=false.FitFeatureResponse=true.expon=false.attenuate=true.l=0.005. h=-1,expr.genotype=false.strand=false.allele-a=false

Expression Console Software will not create a new CEL file or replace the existing one. The GCCN and SST corrections are being handled in memory using an Affymetrix chipstream.

Comparing the effects of combining GCCN and SST

The coefficient of variation (CV) is a normalized measure of dispersion or variance defined as the ratio of the standard deviation to the mean. The SST increases the CV calculated across replicates relative to the untransformed signals (see Figure 16). This increase in CV, however, does not reflect a reduction in data quality. Instead, the change in CV is a product of the transformation process due to the fact that, as a function of the power law transformation exponent, the transformation increases the spread of a set of data faster than it increases the average leading to an increase in CV. The GCCN-SST transformations should be applied when comparison to other expression measurement platforms is required, but the transformations should not be applied when comparison to legacy data sets is required.



Condition	CEL Count	Analyzed Probesets	Median Coefficient of Variation
Standard RMA - MAQCA	3	573909 (Used 62.75 %)	8.11 %
GCCN-SST - MAQCA	3	573909 (Used 62.75 %)	11.97 %
Standard RMA - MAQCB	3	573909 (Used 62.75 %)	12.02 %
GCCN-SST - MAQCB	3	573909 (Used 62.75 %)	14.73 %

Figure 16: CV across replicates varies by tissue. GCCN and SST increase the median CV.

The reason for the increase of CV is due to the SST transformation. Consider the following. You have a set of intensities $\{x_i\}$ over which to compute the CV.

$$CV = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{n}\sum_i \left(x_i - \frac{1}{n}\sum_j x_j\right)^2}}{\frac{1}{n}\sum_i x_i} = \frac{\sqrt{\frac{1}{n}\sum_i x_i^2 - \frac{1}{n^2}\sum_{ij} x_i x_j}}{\frac{1}{n}\sum_i x_i} = \sqrt{\frac{\frac{1}{n}\sum_i x_i^2}{\frac{1}{n^2}\sum_{ij} x_i x_j} - 1}$$

The SST modifies the intensities as follows:

$$x_{out} = low \times \left(\frac{x_{in}}{x_{2\%}}\right)^{lnB}$$

Let $r = lnB$, where $r > 1$. The $low \times x_{2\%}^{-lnB}$ factors cancel out.

$$CV_{SST} = \frac{\sigma}{\mu} = \frac{\sqrt{\frac{1}{n}\sum_i \left(x_i^r - \frac{1}{n}\sum_j x_j^r\right)^2}}{\frac{1}{n}\sum_i x_i^r} = \frac{\sqrt{\frac{1}{n}\sum_i x_i^{2r} - \frac{1}{n^2}\sum_{ij} x_i^r x_j^r}}{\frac{1}{n}\sum_i x_i^r} = \sqrt{\frac{\frac{1}{n}\sum_i x_i^{2r}}{\frac{1}{n^2}\sum_{ij} x_i^r x_j^r} - 1}$$

For the first term under the square root, the numerator is the average of the squares of the intensities, and the denominator is the average of all possible quadratic pairs of the intensities. The thing to note is that the denominator becomes larger as the intensities reduce their range, and it becomes equal to the numerator when

all the intensities have the same value (yielding $CV = 0$). The SST exponentiates each intensity by some power greater than 1, in effect increasing the range of the intensities of the set, reducing the denominator relative to the untransformed case and ultimately increasing the CV. Therefore, the SST will always increase the CV, i.e., $CV_{SST} > CV$. In other words, the SST increases the range of the intensities faster than it increases the mean.

To illustrate this behavior of the CV, consider a hypothetical example with intensities equal to 2, 5, 6, and 8. Figure 17 shows the intensity statistics as a function of the power law exponent. Figure 18 shows the CV as a function of the power law exponent. The standard deviation increases faster than the mean with the result that the CV increases with the power law exponent.

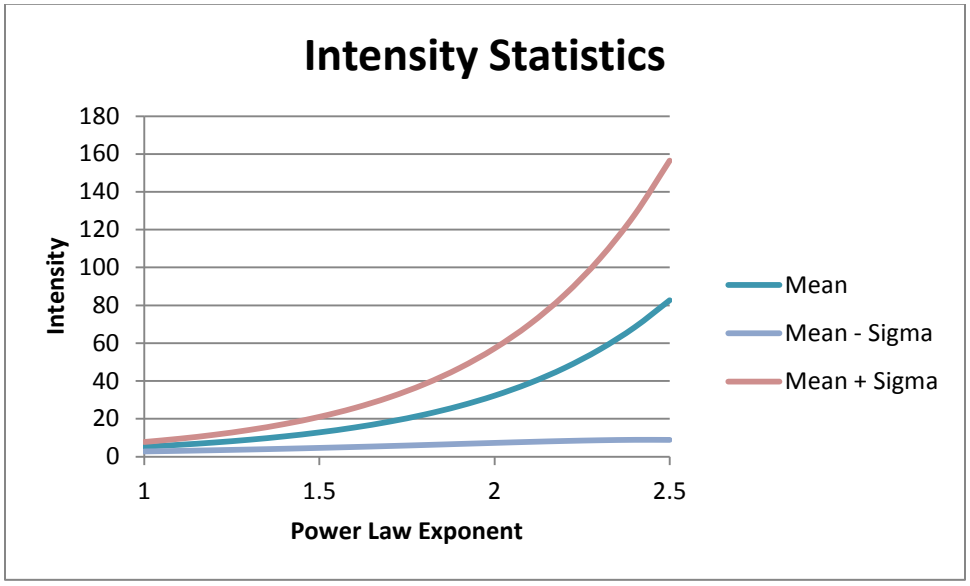


Figure 17: Intensity statistics for a toy model. The standard deviation increases faster than the mean.

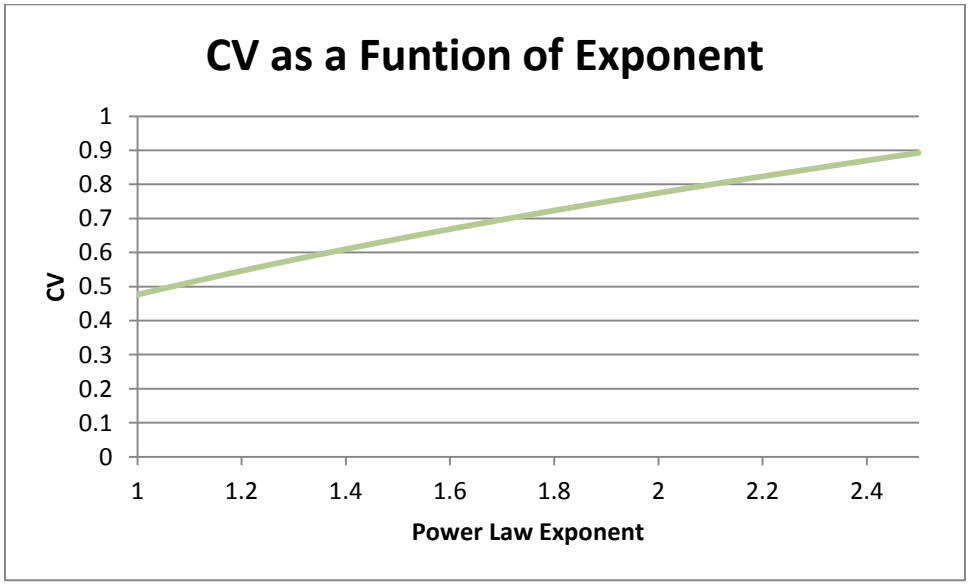


Figure 18: The CV for the range of power law exponents explored in Figure 1. The CV increases as expected as the exponent increases.

The effect of combining GCCN and SST on sensitivity and specificity can be evaluated in the Latin square study as seen in Figure 19. For small changes at low expressed probe sets, there is minimal change in the fidelity of the data. The median T-statistic and AUC values are minimally different.

	Median T stat		AUC	
Comparison	Standard RMA	GCCN-SST	Standard RMA	GCCN-SST
1:50000 versus 0	19.415	19.097	0.9907	0.9917
1:100000 versus 0	13.797	13.696	0.9846	0.9847
1:200000 versus 0	9.15	8.891	0.9649	0.9673
1:50000 versus 1:200000	10.646	11.324	0.9884	0.9897
1:50000 versus 1:100000	5.822	6.429	0.9723	0.9727
1:100000 versus 1:200000	4.958	5.287	0.9641	0.9651

Figure 19: Fidelity of Latin square spikes is maintained by GCCN and SST at the exon level on HTA 2.0.

The total effect of fold change can be observed by comparing statistics before and after GCCN and SST corrections from a fold change comparison between two tissues. In Figure 20 we see a comparison before and after correction. Notice that the slope of the line regression is now improved and reflects an overall increase in fold change.

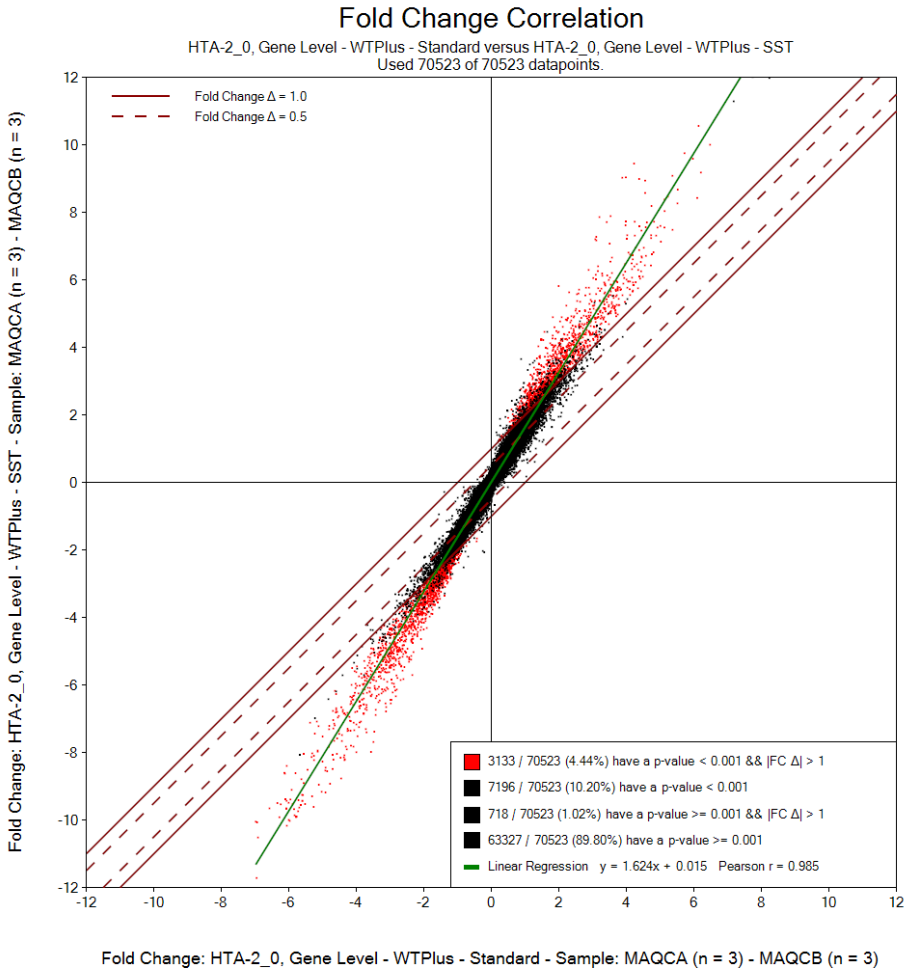


Figure 20: Fold change correlation with and without GCCN-SST corrections on HTA 2.0 using standard RMA (RMA-GC). Before (x) and after (y).

We can also evaluate the effect of the GCCN-SST transformation on fold change by comparing data generated with RNA sequencing as part of the SEQC/MAQC-III Consortium study[7] with data generated using the same source RNA on the HTA 2.0. For this correlation, 834 gene-level differences observed in the publication across several

technologies are selected and correlated to the array data. The ratio of fold changes between technologies improves to a 1:1 ratio. There is also a slight improvement in the Pearson r (Figure 21).

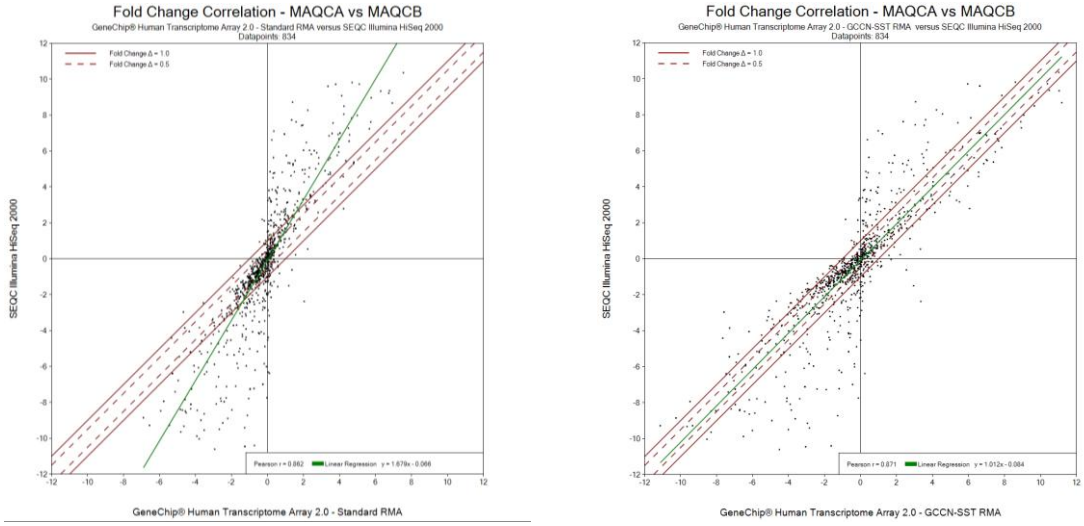


Figure 21: Fold change orrelation before and after GCCN-SST corrections on HTA 2.0 with RNA-seq. Before GCCN-SST on the left and after GCCN-SST on the right.

In addition to the effect we see with a standard RMA signal summarization, we can see similar tuning of the fold changes with other probe set summarization methods as well (Figure 22).

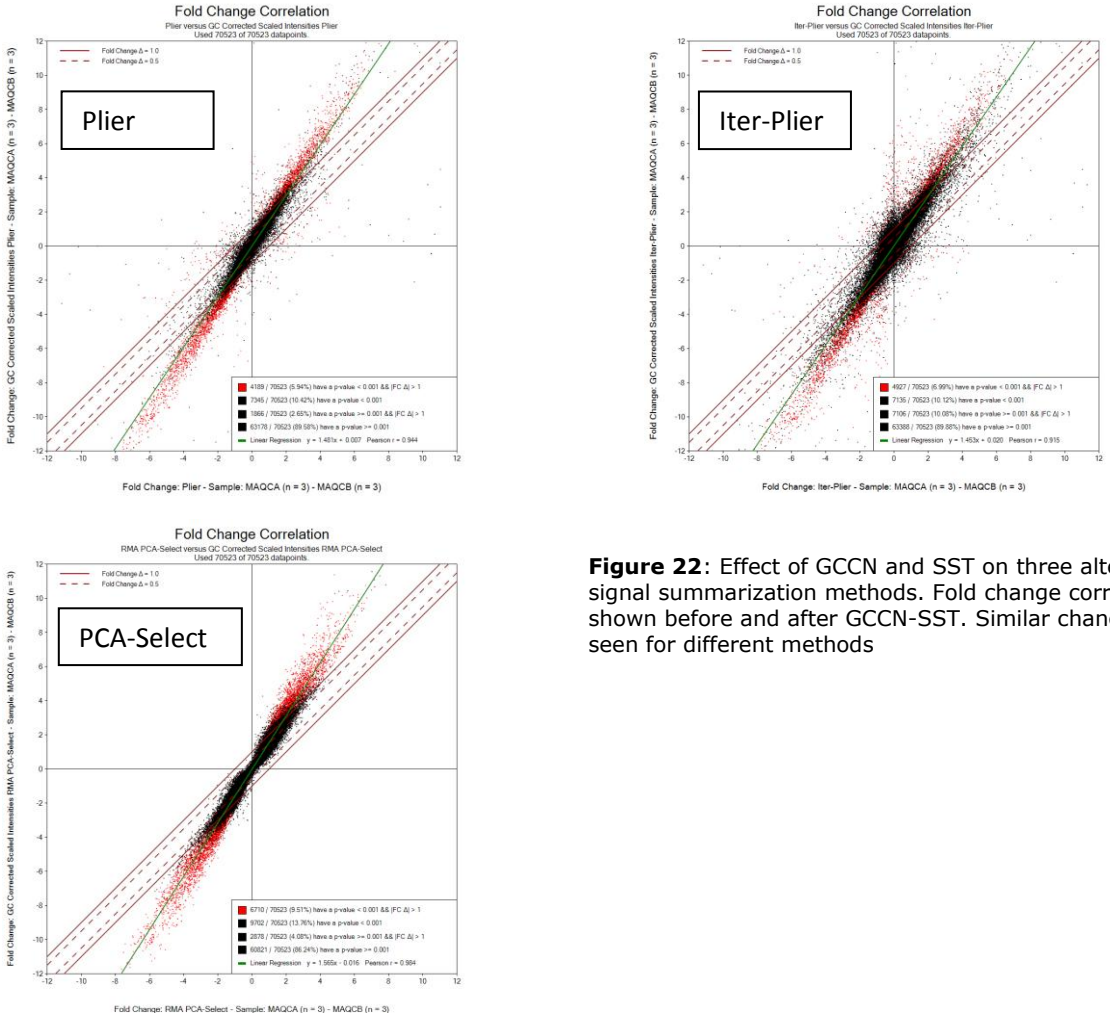


Figure 22: Effect of GCCN and SST on three alternative signal summarization methods. Fold change correlation is shown before and after GCCN-SST. Similar changes are seen for different methods

Combining GCCN and SST and using a fold change cutoff greatly increases the number of differentially expressed genes

Using Affymetrix' Transcriptome Analysis Console (TAC) Software we can see the practical outcome from combining the methods documented in this paper. For determining the genes that are differentially expressed, TAC Software uses both fold change and ANOVA p-value cutoff values (Figure 23).

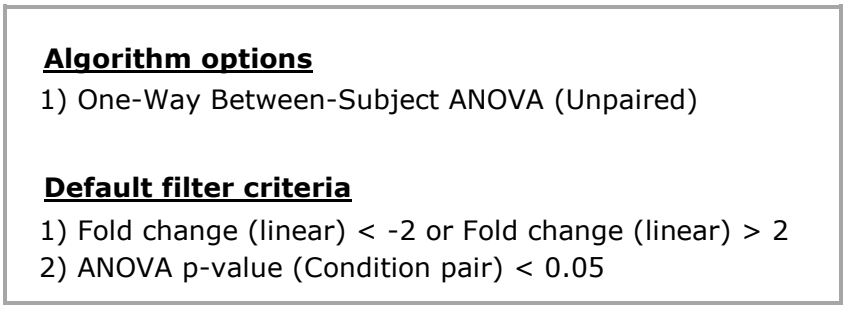


Figure 23: TAC Software default algorithm, fold change and p-value cutoff values.

When the default criteria are used before and after GCCN and SST, we see a dramatic increase in the number of differentially expressed genes (Figure 24).

	Standard Processing	Differentially expressed genes	Genes upregulated – standard processing	Genes downregulated – standard processing
Before	Coding	2,774	1,590	1,184
	Non-coding	837	457	380
	Total	**3,611**	2,047	1,564
After GCCN-SST	Coding	5,953	3,230	2,723
	Non-coding	1,647	851	796
	Total	**7,600**	4,081	3,519

Figure 24: Effect of combining GCCN and SST on TAC Software gene-level results.

Conclusion

SST and GCCN of microarray data can be applied to help adjust fold changes to levels consistent to what is seen using spike data. This overcomes the compression that has been observed and characterized historically in Affymetrix expression data. The compression itself, as well as these techniques, has little effect on the overall fidelity of the technology in terms of sensitivity and specificity.

The use of the GCCN and SST algorithms with RMA will be the default setting using Expression Console™ Software version 1.4.1 and higher with HTA 2.0, MTA 1.0, RTA 1.0, and future transcriptome array designs. To enable the use of the new default analysis method, please install Expression Console Software version 1.4.1 and higher, and install the up to date library files for HTA 2.0, MTA 1.0 and RTA 1.0. The default analysis method in Expression Console Software will be "GeneLevel-SST-RMA". The "GeneLevel-SST-RMA" analysis method executes both the GCCN and SST algorithms prior to normalizing the CEL files using RMA.

Expression Console Software also provides the options for additional normalization methods, which include the following:

Analysis method	Actions
Gene Level-SST-RMA	Performs GCCN and SST prior to RMA
Gene Level-RMA-Sketch	Performs RMA
Exon Level-SST-Alt Splice Analysis	Performs GCCN and SST prior to RMA
Exon Level-Alt Splice Analysis	Performs RMA

References

1. Shi L. Executive Summary. The Microarray Quality Control (MAQC) project. November 21, 2006. http://www.fda.gov/downloads/scienceresearch/bioinformaticstools/microarrayqualitycontrolproject/ucm132150.pdf
2. Perkel J. M. Six things you won't find in the MAQC. *The Scientist* (2006). http://www.the-scientist.com/?articles.view/articleNo/24438/title/Six-Things-You-Won-t-Find-in-the-MAQC/
3. Rajagopalan D. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics* **19**(12):1469–1476 (2003).
4. Choe S. E., *et al.* Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology* **6**:R16 (2005).
5. Dallas P. B., *et al.* Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR – how well do they correlate? *BMC Genomics* **6**:59 (2005).
6. MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**(9):1151-1161 (2006).
7. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology* **32**(9):903–914 (2014).
8. Warden C. D., Yuan Y-C., Wu X. Optimal calculation of RNA-Seq fold-change values. *International Journal of Computational Bioinformatics and In Silico Modeling* **2**(6):285–292 (2013). [Note: sRAP goes through an entire analysis for an example dataset provided with the sRAP package 10 (2013).]
9. Yakovchuk P., Protozanova E., Frank-Kamenetskii M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* **34**(2):564–574 (2006).
10. Affymetrix, Inc. Copy number algorithm with built-in GC waviness correction in Genotyping Console™ Software. *Affymetrix White Paper* P/N: WH103-2 (2009).
11. Abdueva D., *et al.* Experimental comparison and evaluation of the Affymetrix exon and U133Plus2 GeneChip® arrays. *PLoS ONE* **2**(9):e913. (2007).

Affymetrix, Inc. Tel: +1-888-362-2447 ▪ **Affymetrix UK Ltd.** Tel: +44-(0)1628-552550 ▪ **Affymetrix Japan K.K.** Tel: +81-(0)3-6430-4020
Panomics Solutions Tel: +1-877-726-6642 panomics.affymetrix.com ▪ **USB Products** Tel: +1-800-321-9322 usb.affymetrix.com

www.affymetrix.com Please visit our website for international distributor contact information.

For Research Use Only. Not for use in diagnostic procedures.

P/N EMI05133-2