

Single-sample analysis methodology for the DMET™ Plus Product

This document describes the single-sample analysis methodology used with data produced by the DMET™ Plus Product. This method is designed to analyze individual samples and produce conservative, accurate results that do not depend on any additional data within a batch of experiments.

To this end, the data from an assay is normalized to a fixed standard constructed at Affymetrix, removing typical experimental variation. Each individual probe set associated with a given marker is summarized using a robust measure (median) after removing invariant probe-probe differences. These summary values are thus resistant to outliers and do not depend on any individual batch of results, but only on the experiment done.

Finally, these summary values are compared to pre-defined clusters for each possible genotype or copy number for a polymorphism, as well as a universal absolute standard for detecting data points too far away from the reference. The result is a collection of genotype calls and confidence values. At each step, the most conservative interpretation of the data is attempted to ensure high accuracy in the genotyping calls produced.

The remainder of this document describes each step in detail.

Section 1 provides an overview of how the DMET Plus Array was designed, illustrating the kinds of markers interrogated and the kinds of array probes that were designed to detect them.

Section 2 is devoted to preprocessing of the data before genotyping and copy number (CN) determination: normalization, summarization, and the handling of unusual markers. Unusual markers include those with more than two alleles, markers in copy number variation regions, and markers with additional mutations in the vicinity which can affect hybridization properties of probes.

Section 3 concentrates on the mechanics of making genotyping calls. In essence, this is a simple matter of comparing data points to the expected summary values for each genotype and accounting for the typical variation seen in the data. However, this becomes slightly more complex for cases in which rare alleles have not been seen in the training data, or when the data falls outside the expected scatter for well-behaved experiments.

Section 4 provides details on the determination of copy number calls. The general concept applied for calling copy number is similar to calling genotypes, though there are differences in how the pre-defined copy number models are built.

Section 5 discusses the typical behavior and performance of this methodology when run in single-sample mode, including interpretation of quality control results and some suggestions for identifying suspicious batches of data.

Finally, in two appendices, we describe the construction of the standard references for genotyping (Appendix A) and copy number markers (Appendix B). These operations include computational elements (active clustering of data), as well as manual curation of reference genotypes and cluster properties. This standard reference was trained on more than 1,300 samples run with a variety of operators and equipment, with high-value markers sequence-validated for a subset of the training set.

Section 1: Design of the DMET™ Plus Array

The DMET Plus Array interrogates a variety of types of markers. They can be roughly categorized into genotyping markers and regions of copy number variation. The genotyping markers can be further classified as bi-allelic SNPs, tri-allelic SNPs, and insertions/deletions (indels) of varying length. Additionally, some of the genotyping markers are themselves located in copy number regions and/or may have nearby secondary polymorphisms that can interfere with genotyping. The frequencies of these classes are summarized in Table 1.

Table 1: A breakdown of the 1,931 genotyping markers with respect to number of alleles, the presence or absence of potentially interfering secondary polymorphisms, whether or not the marker is located in an autosomal CN region or on a sex chromosome.

Marker property		Number of genotyping markers
Number of alleles	2	1,902
	3	29
Secondary polymorphisms (within 10 bases)	none	1,502
	≥1	429
CN status	In a region assumed to always have CN = 2	1,869
	In an autosomal region of CN variation	62
Insertion/deletion	Not an indel	1,854
	Indel	77
Autosomal/sex chromosome	Autosomal	1,885
	ChrX	46
	ChrY	0

Markers on the DMET Plus Product are interrogated using Molecular Inversion Probe (MIP) technology^{1,2,3,4}. For each of the genotyping markers, there is at least one MIP designed. For some of the more important markers that have adjacent secondary polymorphisms, multiple MIPs were designed against the possible sequence variants (or contexts) in which the polymorphism of interest is located.

For example, if a bi-allelic marker of interest has three adjacent bi-allelic SNPs, it would have a MIP for each of the eight possible contexts. For indels, at least one MIP is designed for each allele, with each MIP using a different tag, allowing for an additional opportunity to discriminate genotypes by the use of allele-specific tags. Each of the other markers shares a common tag across all MIPs used.

The five copy number regions each contain some genotyping markers that can also be used for copy number estimation. In addition, MIPs were designed against unique regions contained within the CN region and not overlapping any other known polymorphisms. These are referred to as CN MIPs, as opposed to the genotyping MIPs described above.

For each MIP in the panel, a collection of probes are tiled on the DMET™ Plus Array to read out the signal. There are two kinds of probe sets for each MIP: one that is complementary to the genomic region targeted by the MIP and one that is complementary to the unique tag that is part of the MIP itself. These are referred to as ASO (allele-specific oligonucleotide) and tag probe sets, respectively. Table 2 gives the counts of ASO and tag sequences for various marker types.

Table 2: Number of distinct sequences interrogated for each of the various types of polymorphisms represented on the DMET Plus Array.

Marker type	Number of distinct sequences interrogated for each type of polymorphism	
	ASO	Tag
Copy number	1	1
Bi-allelic SNP	2	1
Tri-allele SNP	3	1
Bi-allelic indel	2	2
Wobble SNP	One for each allele in each context	1
Wobble indel	One for each allele in each context	2

An extensive collection of array probes is used to interrogate each targeted sequence to maximize the chance of successful signal detection. MIP tags are interrogated with array probes of up to three lengths from both strands with three replicates each, for a total of 18 probes (3 x 2 x 3). Allele-specific genomic sequences are interrogated with probes from two strands, up to five probe lengths, up to nine offsets relative to the interrogation base, and up to three identical replicates on the array—altogether as many as 270 probes per allele-context (see Table 3, following page). Factoring in that the ASO probe set for each genotyping marker is formed of multiple collections of probes (one per allele and increasing exponentially in the presence of adjacent secondary polymorphism); some markers are interrogated by thousands of array probes.

Table 3: Maximum possible number of array probes used to interrogate each unique sequence. The maximum number of array probes for a given marker is then equal to this value times the number of alleles, further multiplied by the number of sequence contexts. For example, the number of ASO probes for a bi-allelic SNP with two adjacent bi-allelic SNPs would be 2,160 (2 x 2 x 2 x 270). Markers of critical importance (as determined by the ADME consortium) are given the “full treatment,” while other markers are interrogated with fewer combinations of strands and lengths but with at least 132 ASO probes per allele-context.

Probe set Type	Array probe counts for each interrogated sequence						
	Alleles	Offsets	Strands	Length	Replicates	Contexts	Total
ASO	1	9	2	5	3	Varies	270
Tag	1	1	2	3	3	Varies	18

Section 2: Preprocessing of data in single-sample analysis

A single DMET™ Plus Array contains slightly over one million features, with a diverse assortment of probe sequences, including control probes of various kinds. For each marker, the relevant probes need to be extracted, normalized, and summarized to remove irrelevant effects on raw intensity and produce values that can be reasonably compared to the reference clusters. This section describes the process of standardizing the data: normalizing global assay effects, summarizing individual allele-specific probe sets, reducing multiple probe sets to the bi-allelic case, and transforming the data into an appropriate clustering space.

It is standard practice in genotyping assays to remove global intensity effects by a nonlinear transformation that makes the distribution of intensities observed in an experiment identical to a standard intensity distribution. This process is known as quantile normalization² because it transforms the intensity of all “quantiles” of the input distribution (median, 75th percentile, 97th percentile, etc.) to the intensity of the equivalent quantile in a standard distribution. The full transformation is memory and time intensive, so we approximate the distribution by 50,000 points within the distribution and linearly interpolate the intensity transformation between modeled points. This approximation is known as sketch normalization because it uses a set of representative points to “sketch” the distribution. Because this software runs in single-sample mode, the standard intensity distribution is a fixed distribution constructed at Affymetrix from a large training set. This transformation removes irrelevant global effects from the raw intensities on the array (overall brightness, etc.), allowing them to be compared with those experiments in the training set.

After removing global intensity effects, the next step is to summarize probe sets associated with each allele. Each probe set consists of a number of probes that associate with a particular target sequence; however, the probes may consist of differing lengths and may overlap a marker at different offsets. These differences lead to systematic intensity differences between probes. These multiplicative differences are removed by applying an individual multiplicative effect to every probe. These feature effects are read from a standard file and do not depend on the sample or samples in a batch. The probe set is then summarized by taking a median, which is a robust summary of the intensity values. The median is actually performed on the logarithmic scale, and the feature effects are removed from the additive model by subtraction on this scale. This procedure is essentially the median-polish summarization from the well-known Robust Multichip Analysis (RMA) used extensively with expression microarrays^{5,6,7}, but with fixed feature effects. After this

step, each probe set associated with a unique genomic target sequence is represented by a single number (the signal for that probe set).

For simple bi-allelic markers, the probe set summarization process described above results in two numbers, one associated with each allele. However, unlike previous SNP Array products, the DMET™ Plus Array contains markers with more complexity associated with them, such as tri-alleles and markers with additional nearby mutations. Tri-alleles are similar to bi-allelic markers in that each of the three alleles has a summary value. Markers with additional nearby sequence complexity have a probe set associated with each possible variant of the local sequence. We refer to such local sequence haplotypes as “contexts” for a marker. For such markers, we have a probe set for each allele and each context, which can lead to a large number of summary values—a bi-allelic marker with eight SNPs in the nearby region would have 256 contexts, each of which has a probe set for each allele, resulting in 512 summary values. Because humans are diploid organisms, only one or two summary values for each marker represents the true sequences in the individual assayed (at least for markers in a region with a copy number of two).

Therefore, for a given marker in a particular sample, the collection of summary values is reduced to only two values. For simple bi-allelics, this is always the two summary values (one for each allele). For tri-alleles, the two alleles that are most likely present are chosen to represent the marker in a given sample. This decision is made by choosing the two probe-sets with the highest signals as those most likely to represent perfect hybridization to a target. For multi-context bi-allelic markers, the contexts most likely to represent true hybridization for each allele are chosen. This decision is also made by choosing for each allele the probe set with the highest signal amongst all the contexts available. In this manner, every marker in a given sample is assigned two summary values, each representing one allele type for a marker. This assignment allows all markers to be mathematically handled as though they were simple bi-allelic markers for genotyping purposes.

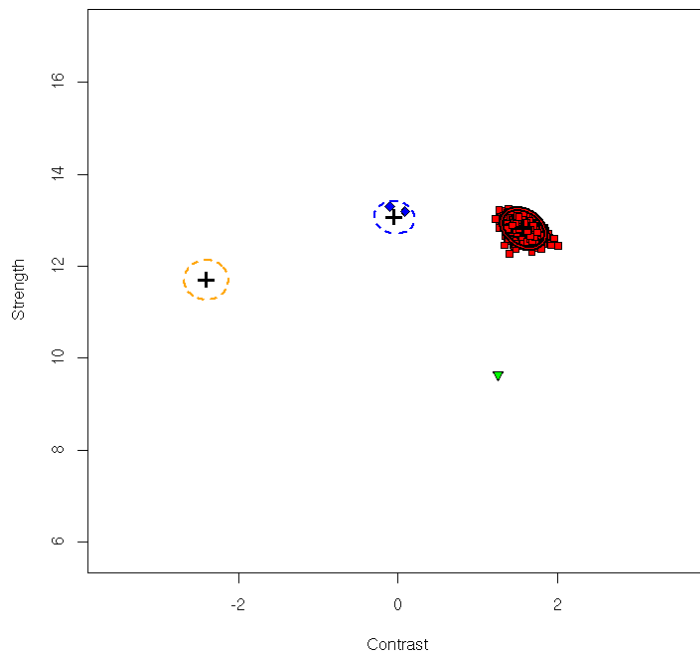
Prior to genotyping, the two summary values, one per allele, are transformed for mathematical convenience. Call these values alleles A and B. Because a typical scatter is multiplicative in intensity, the values are log-transformed. Because the difference in allelic content is of primary interest, a value known as “contrast” is constructed, $\log_2(A) - \log_2(B)$, which contains most of the variation between genotypes. The value “strength,” $0.5 * (\log_2(A) + \log_2(B))$, is also constructed. The paired value (contrast, strength) will be compared with typical values for various genotypes to determine the actual genotype call.

Section 3: Genotyping calls and confidences

Genotype calls are made by comparing the observed signal values for a marker with the expected signal values for appropriate genotypes and choosing the genotype that yields the maximum likelihood for the observed data. This likelihood function takes into account the expected observational scatter of the data, the typical frequency of a genotype, and the uncertainty caused by residual batch-batch effects. All these parameters are described by a two-dimensional Gaussian cluster for each genotype. Data points that do not fall into a particular cluster are assigned a confidence value reflecting this uncertainty. Clusters with sufficiently large uncertainty or low frequency are classified as possible rare allele (PRA) clusters because of the lack of observational data to support the genotype with confidence. In this way, signal values are converted to genotype calls with associated confidences or PRA calls.

The first step is to determine what reference model is most appropriate for evaluating the signal. This model depends on the pair of alleles that are potentially present. For example, in tri-alleles, there are three separate potential pairs of alleles and a reference model for each pair. It also depends on the copy number state of the marker. For typical markers with a copy number of 2, there are three possible genotypes—AA, AB, and BB—each of which is modeled by a Gaussian. For markers known definitely to have a copy number of 1 (such as on sex chromosomes in appropriate individuals), there are only two genotypes, A and B, and hence only two clusters. For markers known definitely to have zero copy number (as in samples where a region is deleted), the model outputs only CN = 0 with no genotype.

Figure 1: A marker showing copy number variation. The green triangle represents a call of CN = 0 for the region and is not carried through for further genotyping analysis. The ellipses are drawn at two standard deviations away from the center of the cluster and generally encircle the majority of the calls associated with each genotype. The dashed ellipses indicate that the cluster represents a possible rare allele (PRA). There were two heterozygote genotypes observed among ~1,200 distinct samples used to train the clusters and there were no observations of the rare homozygote, so any future calls of these genotypes will be classified as PRA given the uncertainty in the locations of the clusters for these two genotypes.



Assuming the case where the copy number is greater than zero, a prior model contains a two-dimensional Gaussian for each genotype, along with a “frequency” for each genotype. Each cluster has a mean and variance for contrast and strength values, along with a covariance between the two axes. Because this model was trained in a Bayesian fashion, the “frequency” of a cluster reflects the amount of training data found in that cluster and the prior knowledge of the approximate location of that cluster, scaled by a number of pseudo-observations. Thus, the “frequencies” are not exactly the population rate of a genotype in the training data (or untrained clusters would have a zero frequency and never be called), although they are approximately the same (a typical untrained cluster has the equivalent of approximately 0.3 observations in the reference set). This frequency is important when evaluating how unusual a data point is relative to a given cluster center. This

allows for more accurate placement of decision boundaries. The frequency of a cluster is also used when evaluating whether to assign a PRA call to a given data point.

To compute a call given a genotype model, the (contrast, strength) value for a marker is compared to all clusters in the model. For each cluster, the likelihood of the data point is calculated assuming that the associated genotype is the true genotype and that a Gaussian scatter with variance and covariance as described, along with the frequency of the cluster. The genotype of the data point is assigned to be the highest likelihood cluster. The confidence of this data point is the relative probability that the data point belongs to any of the other clusters, or belongs to an "ocean" of uniform probability density representing outlier behavior. The confidence is computed by $\text{sum}(\text{likelihood of belonging to other clusters}) / \text{sum}(\text{likelihood of any cluster})$, so lower confidence values indicate more confident genotype calls. This confidence value screens out ambiguous data points that lie between two clusters and unusual data points that are not well represented by the training set. Such points are not necessarily wrong, but are conservatively assessed as being outside the region the training data supports, and are marked as less confident. If the confidence rises above a threshold (0.1 by default), the genotype call is converted to a no-call and suppressed.

An exception to this conservative logic is used when constructing a PRA call. Clusters with few to zero reliable data points observed in training are necessarily uncertain in position, as they are derived by extrapolation from typical relationships between clusters (and occasionally by manual adjustment). If a genotype was not seen in training, then the location of the cluster cannot be learned from the actual data.

Therefore, when the data indicates that a cluster of low frequency is the most likely cluster, the call is set to PRA even if the confidence is poor, precisely because the uncertainty in cluster location is large. A PRA frequency cutoff (set to three by default) may be configured by the user such that a cluster is treated as a PRA if the number of observations of the genotype among the ~1,300 samples in the training set falls below the cutoff.

In summary, the genotyping logic is straightforward: signal values are compared to prototype clusters for each possible genotype, and the most likely cluster is chosen as the genotype. Points that are located in ambiguous positions, or those that are unusual compared to the data used as a reference, are marked as having poor confidences to conservatively screen out data for which the trained model may be in error or does not apply. To allow for discovery of rare alleles not seen in training, PRA calls are made liberally when data points appear to be compatible with a rare allele, although the confidences may still be poor.

Section 4: Predicting chromosomal copy number

The DMET™ Plus Product detects homozygous deletions (CN = 0) in the five regions listed in Table 4 below.

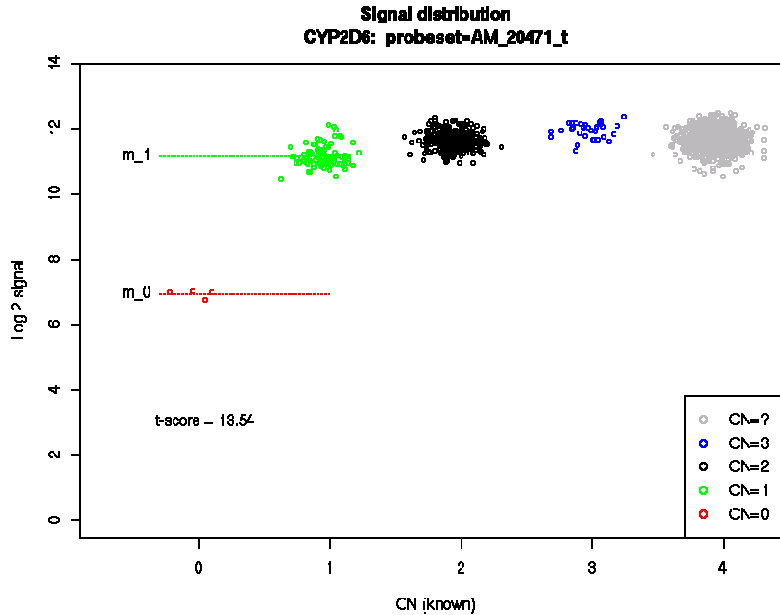
Table 4: The five copy number regions analyzed by the DMET Plus Product. Coordinates refer to build 36 of the human genome. The gene minimum and maximum fields give the transcription footprints, including the untranslated region (UTR). The region minimum and maximum include the deletion footprint, which is often wider than the transcription footprint.

Gene region	Chromosome	Gene min.	Gene max.	Region min.	Region max.
CYP2A6	19	46,041,284	46,048,180	46,039,000	46,073,000
CYP2D6	22	40,852,445	40,856,827	40,849,000	40,867,000
GSTM1	1	110,031,965	110,037,890	110,029,000	110,042,000
GSTT1	22	22,706,141	22,714,231	22,680,000	22,727,000
UGT2B17	4	69,085,497	69,116,840	69,057,000	69,170,000

Normalization and derivation of a signal value for each allele and context of the genotyping markers are performed in the same manner as described in Section 2. CN prediction also makes use of CN probe sets, which are similar to genotyping probe sets except that they contain just one allele.

The CN analysis becomes different immediately after the probe set summarization. The derived signal values are \log_2 -transformed and summed across all alleles and contexts, resulting in a single signal value for each marker. An example of the signal values for a single probe set is shown in Figure 2. During the training process, the utility of each probe set for discriminating between CN = 0 and CN > 0 is quantified by a linear discriminant (ld) score, defined as the ratio of the separation between cluster means and their pooled standard deviation (see Appendix B for exact definition).

Figure 2: Plot of the signal values for probe set AM_20471_t, a tag probe set for a CN marker in the CYP2D6 region. The CN = 0 samples formed the red cluster with mean value m_0 . The CN = 1 samples formed the green cluster with mean value m_1 . The separation between these clusters has an l -score of 13.5. The separation between the other known CN levels is poor for this probe set, as is the case for all five interrogated CN regions. As a result, no attempt is made to distinguish between any copy numbers larger than zero.



Each CN region is assessed using 10 probe sets. The process by which the probe sets were selected and trained is described in Appendix B. One of the results of this training process is an estimate for each probe set of the typical mean signal value for CN = 0 and CN = 1, as well as the l -score quantifying separation. If these quantities are defined as m_0 , m_1 , and l , respectively, then the CN estimate for a probe set signal s is defined as:

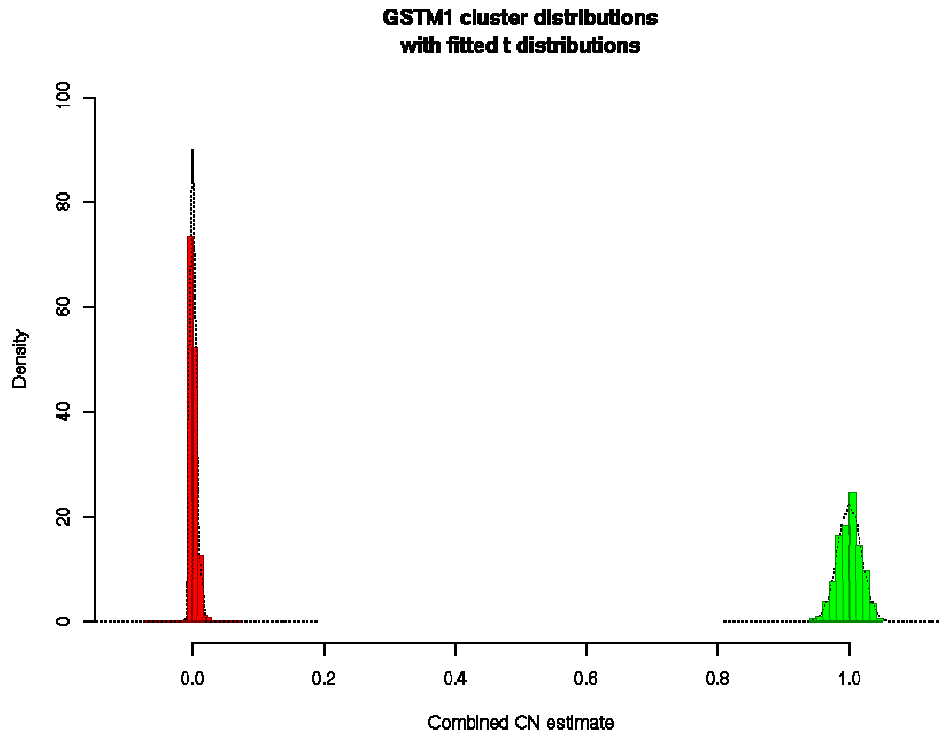
$$cn_estimate = \frac{s - m_0}{m_1 - m_0}$$

This is just the linear interpolation of the summary value between the CN = 0 cluster and the CN = 1 cluster at the probe set.

The CN estimate is more accurate for probe sets with high l -scores. Therefore, when the 10 probe set-specific CN estimates are combined to produce a region-specific CN estimate, it is done as a weighted average, with each probe set's weight proportional to its l -score.

The resulting weighted average is called the *weighted CN estimate*. An example of the distribution of weighted CN estimates for the CN region GSTM1 is shown in Figure 3.

Figure 3: Histogram of combined CN estimates for region GSTM1 in a large sample collection. Only samples known to have CN = 0 (red) or CN = 1 (green) are shown. There is perfect separation between the two.

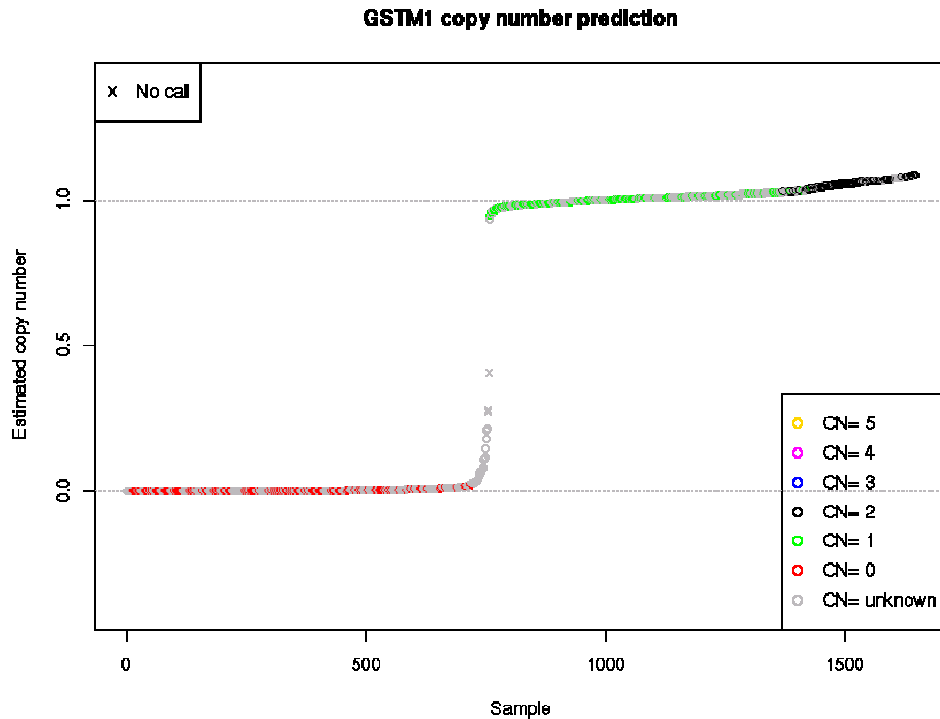


The final step is to derive from the weighted CN estimate a classification of the CN in the region. Each of the two CN levels is modeled by a t-distribution, with mean and standard deviation as estimated in the training set and degrees of freedom equal to the number of observations of the CN level in the training set. If the number of samples is large, then the t-distribution is very close to Gaussian. However, CN = 0 samples are relatively rare for the CYP2A6 and CYP2D6 regions, so the use of t-distributions with small degrees of freedom helps produce a heavier-tailed distribution that better reflects the uncertainty in the variance attributed to the cluster.

The t-distributions are used to make a maximum-likelihood prediction of the copy number for each sample. The probability of the observed weighted CN estimate is computed under the assumption that it comes from each of the two clusters. This is compared with a third "no-call" cluster with a uniform low probability. Each probability is multiplied by a cluster-specific prior probability estimated from the training set and the CN call is assigned to the cluster with the largest posterior probability.

A confidence value is assigned to each call, computed as $1 - p_{\max}$ where p_{\max} is the posterior probability for the most likely call (which is either CN = 0 or CN > 0). A lower value corresponds to greater confidence in the call. The confidence value is compared to a fixed threshold (0.1 by default). If it is too high, the CN region is classified as a no-call; otherwise it is classified as the cluster with maximum posterior likelihood.

Figure 4: CN classification for roughly 1,600 samples in the GSTM1 region. The y-axis plots the weighted CN estimate and the shape of points indicates how they were called. There are two no-calls in the region in between the two clusters and all others are confidently assigned as either CN = 0 or CN > 0. Samples with known copy number are indicated by colored points as indicated in the legend.



Section 5: Outcomes of design decisions

The DMET™ Plus Product single-sample genotyping methodology has been chosen to be conservative in the calls it makes. The design decisions support a true single-sample methodology, depending only on the model provided and the data for an individual experiment.

To this end, an experiment is normalized to a fixed distribution, making the intensity values comparable to the corresponding values in the training set on which the model is built. The systematic differences in feature-level intensity are held to be fixed at the values that were fitted to the training set, allowing summarized signal values to be directly compared to the invariant model. The preprocessing is extremely conservative.

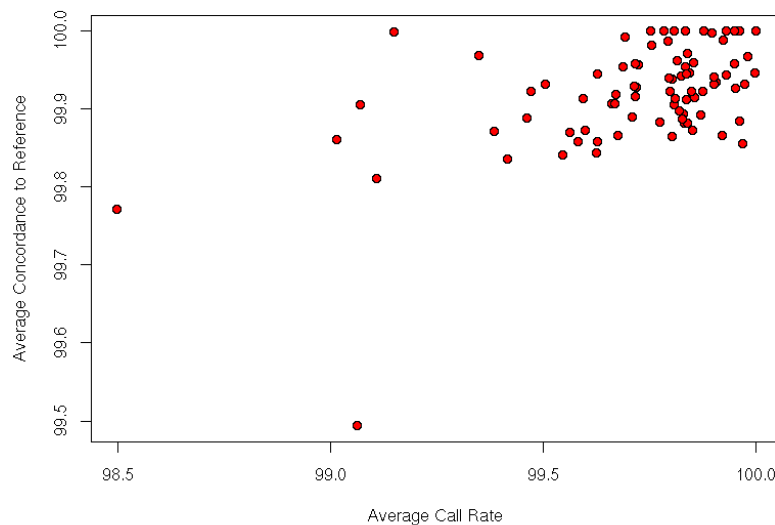
The genotyping model itself is not allowed to adapt to any shifts in clusters with experimental batches. While this would be convenient in cases when systematic deviations occur, allowing the model to track the data would cause genotypes to vary slightly depending on what other samples were analyzed at the same time. The parameters for the cluster model are therefore held invariant as a fixed standard for comparison.

Although some tolerance for systematic shifts is accommodated in the variances assigned to the clusters, it is generally the case that typical data are directly comparable to the training set. In cases where experimental results are not directly

comparable, even when dynamic clustering that adapts to the data would provide high-quality results, the confidence score will become poorer. Such incompatibility flags data points that deviate: samples or markers with unusually low call rates have changed in some way from the training data. The model itself cannot distinguish whether the outlier results are due to noise or systematic deviations, but low call rates imply that the training set does not effectively represent the data.

This conservative behavior still results in average call rates in excess of 99 percent and average concordance to HapMap genotypes greater than 99.5 percent.

Figure 5: Observed performance of the DMET™ Plus Product during development. The DMET Plus Array was extensively tested at Affymetrix and beta sites, with a total of 82 batches of data (where “run” is defined as a single operator processing one week’s worth of samples); 23 of the batches came from external sites. In total, the 82 batches include more than 3,500 samples. Within each batch, any sample with a call rate less than 98 percent was rejected as a potential problematic sample, leading to the exclusion of 4.4 percent of the samples attempted. For the remaining samples, concordance of genotype calls was compared to a reference data set made up of calls from HapMap, TaqMan, sequencing, and the DMET 2.0 Early Access Product. Combined, this provided reference calls for almost 1,200 of the 1,931 genotyping markers on the product. The plot contains a point for each of the 82 batches, showing the average call rate and average concordance to reference computed across all the passing samples in the batch.



References

- ¹ George Karlin-Neumann, et al. (edited by M. P. Weiner, S. B. Gabriel, and J. C. Stephens). Molecular Inversion Probes and Universal Tag Arrays: Application to Highplex Targeted SNP Genotyping. *Genetic Variation: A Laboratory Manual*, published by Cold Spring Harbor Laboratory Press (2007).
- ² Wang, Y., et al. Analysis of molecular inversion probe performance for allele copy number determination. *Genome Biology* 8:R246 (2007).
- ³ Wang, Y., et al. Allele quantification using molecular inversion probes (MIP). *Nucleic Acids Research* 33:e183 (2005).
- ⁴ Moorhead, M., et al. Optimal genotype determination in highly multiplexed SNP data. *European Journal of Human Genetics* 14, 207-215 (2006).
- ⁵ Bolstad, B. M., et al. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193 (2003).
- ⁶ Irizarry, R. A., et al. Summaries of Affymetrix® GeneChip® probe-level data. *Nucleic Acids Research* 31(4):e15 (2003).
- ⁷ Irizarry, R. A., et al. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* 4(2):249-64 (2003).
- ⁸ Affymetrix, Inc. White paper: BRLMM-P: A Genotype Calling Method for the SNP Array 5.0.
http://www.affymetrix.com/support/technical/whitepapers/brlmm_p_whitepaper.pdf

Appendix A: Training reference models for genotype calling

The reference models used with the DMET™ Plus Product are critical determinants of the genotypes and copy number estimates delivered by the product. This appendix provides an overview of the process followed during development, which resulted in the creation of the reference models.

The reference models used for standard normalization, feature summarization, and genotype comparison were created primarily with automatic genotyping methods. There was, however, some manual curation required, both of the reference genotype calls used in training and of the marker-specific cluster models. A modified version of BRLMM-P (as used on the SNP Array 5.0⁸) was utilized for the automated clustering step, with parameters altered to reflect the DMET Plus conditions. All markers were visualized and checked by expert analysts to verify that the results were in accordance with expectation. In a minority of cases, the cluster models for genotyping were manually altered to improve unusual markers that were not well served by the automated procedures.

The basic dynamic clustering adapts the genotyping models to represent each marker's unique distribution of signal values under each genotype. BRLMM-P is a likelihood-based clustering method that updates a prior distribution of genotype clusters with tentative genotypes that maximize the likelihood of the observed data under the posterior distribution. It then uses the posterior distribution to make the reported genotype calls.

In less technical terms, the method is provided with a general description of which clusters should be where (e.g., BB genotypes should have higher B signal than AA genotypes). Then it looks at the observed data to find where clusters actually appear for a marker. Finally, it compares the updated cluster information with individual data points to assign genotypes and calls. During this procedure, reference data is used to penalize cluster assignments that contradict pre-existing knowledge of marker genotypes in samples. There are additional penalties against undesired cluster properties, such as clusters that are poorly separated. These global penalties and reference data provide good average-case performance; however, unusual markers still require manual intervention for improved outcomes.

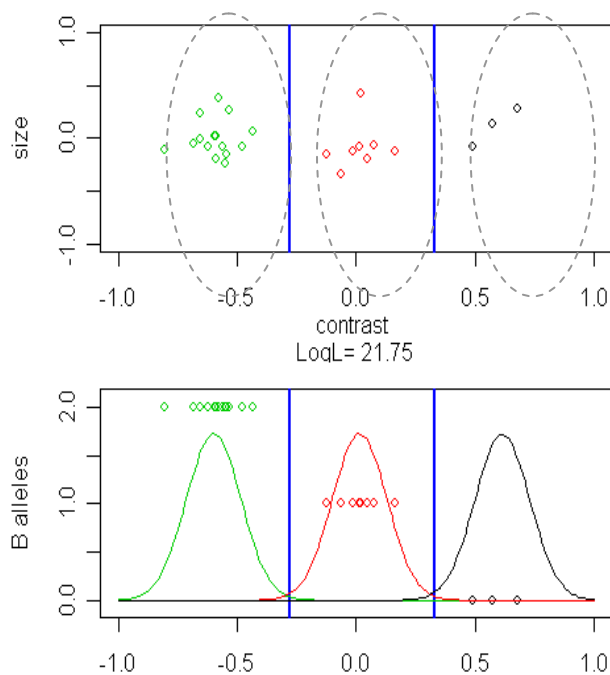
The model used to represent the genotype signal values is a Gaussian mixture model. Every genotype state corresponds to a two-dimensional Gaussian in the clustering space, with an associated frequency. In the case of BRLMM-P, this model is Bayesian with a fully conjugate normal-inverse-gamma (technically, normal-inverse-Wishart, because the clusters are multidimensional) prior on the mean of the cluster center and the variance of the cluster. Both the mean and the variance of a cluster have a precision to which they are known. Because of the conjugate prior, this precision is naturally scaled in "number of pseudo-observations" for a given cluster. After training, clusters with large uncertainty in their position have a small precision, and clusters with small uncertainty have a large precision (prior precision plus the number of observations of that genotype in the training data).

Thus, every individual genotype cluster has seven parameters: meanX, meanY, varX, varY, covXY, precisionMean, and precisionVariance, where the precisionMean also represents an approximate frequency for that cluster. The full model also includes correlations between cluster center locations: there are 12 parameters for correlations between the three cluster centers. This prior model represents the important facts about a marker: where signals should be found for each genotype in

a typical marker, what the typical scatter is around the expected signal location, a rough relative frequency of a given genotype, and to what precision this information is known.

To update a prior model, tentative genotypes are assigned to data points provided in dynamic clustering mode. The tentative genotypes are generated by trying all plausible assignments of genotypes to data points, following the rule that BB genotypes must have smaller contrasts ($\log_2(A) - \log_2(B)$) than AB genotypes, and AB genotypes in turn must have contrasts smaller than AA genotypes (i.e., more of a given allele should correspond to more signal for the probe set specific to a given allele). The posterior likelihood of each plausible assignment is then evaluated (using a one-dimensional Gaussian model in the contrast dimension alone) to find the maximum likelihood assignment of tentative genotypes.

Figure 6: Tentative genotypes are computed by finding a hard labeling of the data with maximum likelihood under the posterior Gaussian model. Because plausible genotypes must have contrast (BB) < contrast(AB) < contrast(AA), an assignment of genotypes to data is exactly described by two transition points where the number of B alleles changes. All $(n+2) \times (n+1)$ plausible assignments of genotypes are evaluated using the posterior Gaussian model, with additional penalties for contradiction of known references and bonuses for well-separated clusters. The tentative genotypes are then used to update the two-dimensional Gaussian model used in genotyping.

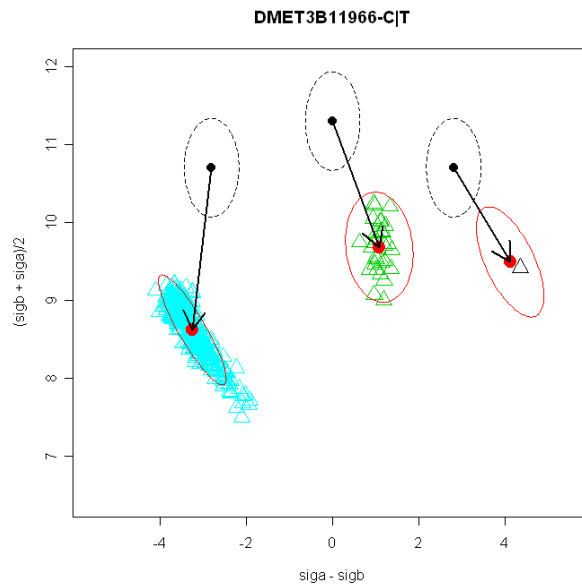


This likelihood function penalizes each assignment that contradicts a reference genotype, thereby reducing the likelihood of clusterings that are inconsistent with the reference data. Because even very good reference data sources have a nonzero error rate, this penalty is large, but not infinite. There are additional modifiers to the likelihood for clusters that are close, either in absolute distance between cluster means, or the distance between cluster means squared, scaled by the variance. The

former modification is implemented by means of an isotonic regression that forces the posterior cluster means to be separated by at least a specified distance. The latter modification is implemented as a bonus to the likelihood for well-separated clusters, which is smoothly thresholded by Geman-McClure transformation: for a given scaled separation F , the bonus per data point in such clusters is $B = F/(1+F/z)$, where z is a tuning parameter setting the threshold. Both of these modifications bias the clustering towards finding well-separated clusters even in the case of non-Gaussian behavior within a cluster or clusters. However, in the case of unusual markers in which the clusters are not well separated, this bias will work against finding the true division of the data. Since this is rarely the case for markers that are wanted for a high-accuracy product, the bias is useful. The few unusual exceptions in the DMET™ Plus Product have been handled by manual curation.

Thus, the tentative genotypes reflect the prior information about cluster locations, the observed data, and the reference data used for training. They are reasonably accurate in themselves, but can be improved by using the full two-dimensional posterior model and producing a model that is consistently applicable to further data sets.

Figure 7: Visualizing a Bayesian update to the variance. The black dashed ellipses indicate the prior variances for clusters and the solid ellipses indicate the posterior variance. Variances are shrunk to have common scale in x and y directions, borrowing information from highly observed clusters to estimate scale for poorly observed clusters such as the AA genotype. The correlation between the scatter in x and y is allowed to vary by cluster so that the major trend in the cluster is captured. The updated variance is a combination of the prior variance, the scatter of the observed data around the cluster center, and residual uncertainty from shifting the location of the cluster. The posterior precision is the prior precision plus the number of observations in a cluster.



The prior model is updated by tentative genotypes to produce a posterior model capturing the information provided by the observed data. The posterior model is of exactly the same form as the prior model, but with the means, variances, and precisions updated to reflect the increase in cluster information. The correlations between cluster centers are also taken into account in the update equations, which is

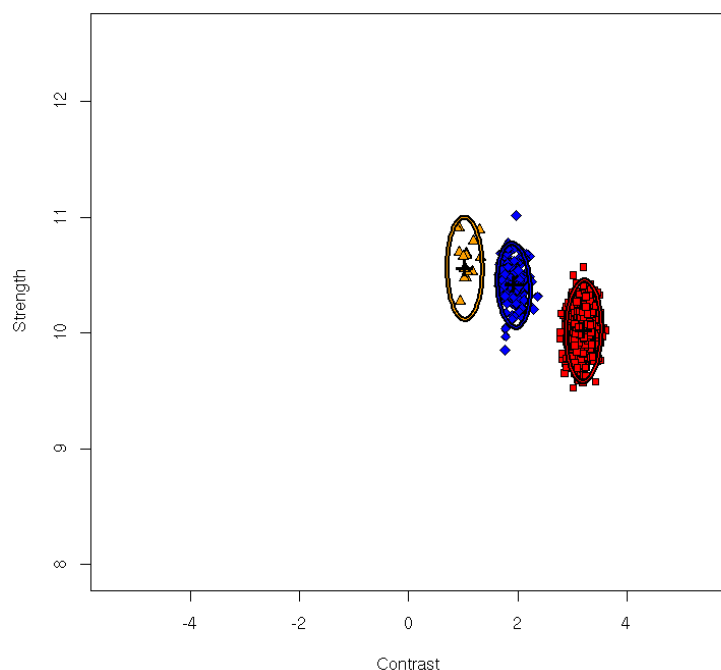
the standard $M = (K+N)^{-1} * (Ku+Nm)^{-1}$ for means, where K is the prior precision matrix, N is the matrix assigning tentative genotype observations to clusters, u is the prior mean locations, and m is the observed mean locations. The update equation for the variance is the typical full conjugate update $V*(p+n) = p*V0 + SS(\text{observed}) + k*n*(u-m)^2/(k+n)$ —the variance is the prior variance plus the observed scatter within a cluster, plus uncertainty in location due to moving the cluster center. This variance update is performed for each cluster independently and then a shrinkage term is applied to shrink the scale of the within-cluster variances to be similar. This ad hoc shrinkage improves the behavior of clusters with few data points.

After this update, the posterior model is used to evaluate the genotype calls and confidences for each sample. The call is made by assigning the genotype associated with the cluster to which the observed signals belong with highest relative probability. This likelihood is evaluated as the normal likelihood with cluster means and variances as in the posterior, and relative frequency as provided by the precision of the mean. The confidence is assigned as the relative probability that the data point belongs to one of the other clusters or to an “outlier” cluster with a small uniform probability density. This last cluster controls for data points unusual relative to the typical observed data where the model assumptions may not apply.

This posterior model can then be preserved for future single-sample runs or as a prior model for successive instances of training data to accumulate more details of marker behavior. For the DMET™ Plus Product, the posterior model produced by the training runs (after manual curation) is used as the single-sample model. This allows the product to provide genotype calls relative to the fixed training set without depending on other samples.

The automated clustering operates as above: global information about cluster properties is combined with observed data points and reference data to produce genotype models for each marker. However, there were several categories of important markers for which the automated cluster models were insufficiently resolved. First, there were multiple markers with unusual cluster locations due to idiosyncratic hybridization or amplification of probes. These markers were manually curate to ensure proper cluster labeling and proper positioning of unobserved clusters. Secondly, there were markers with unusual cluster properties due to the model being inapplicable: two or more clusters within a given genotype due to artifacts, copy number variations in some samples not reported by the literature, and so forth. These markers were manually curate to ensure the cluster model covered the appropriate samples. Third, there were multiple markers in which the clusters were well separated though in absolute terms not far from one another in contrast space, such that the global biases against solutions with closely located clusters were counterproductive in producing accurate genotypes. Such markers were re-clustered, allowing cluster centers to be close in absolute terms, and the resulting models were manually combined with the standard models generated by the automation.

Figure 8: An example of an unusual marker. The location of the BB homozygous cluster has shifted to a position more typical of an AB heterozygous cluster. Further, the BB and AB clusters are unusually close together in absolute terms, even though their resolution is decent (separation of cluster means is small, but so is the cluster variance). This marker was better clustered by relaxing the constraints requiring large separation of cluster means, and the model for this marker used in the DMET™ Plus Product was taken from a dynamic clustering with minimal constraints.



A final step in the manual modification of the marker-specific models was to inflate the variance for all clusters to account for shifting of cluster centers due to various experimental batch effects. That is, it is expected for clusters to wander slightly from the estimate of the true location provided by the training data. The variance was additively increased based on an estimate of the mean variance added by such cluster shifts in both X and Y. This increase allows for the confidence in such calls to be reasonably estimated under the real-world conditions in which markers vary slightly from the training data. Systematic, large shifts for individual markers can still lead to an increase in no-calls, which indicates the inapplicability of the training data for such a marker and flags the data points as suspicious, though importantly in such cases the concordance usually remains high. This behavior was chosen as a conservative option to avoid over-training to any individual marker and highlight unusual circumstances.

Manual intervention also occurred within the pre-processing of the data. An automated pass was done to remove the worst half of the probes within each probe set (those that contributed least to correct classification of genotypes). For high-value difficult markers, the probe set content was edited by hand to select only probes that specifically responded to genotype differences with good signal. Each probe was processed separately to yield call rate, concordance, and other measures of cluster quality. The selection process identified the largest number of probes consistent with a desired quality metric. This process was naturally subjective, difficult to automate, and marker-dependent, but allowed many important markers to be included in this product.

All of the training described above was done on a set of more than 1,300 samples (of which more than 1,200 were distinct DNAs) prepared and run at Affymetrix. The sample set included standard reference samples available from Coriell as well as samples from the extended HapMap collection. Sequencing results for high-importance markers were obtained in a number of samples to allow for verification of concordance and to provide reference data for constructing accurate genotype models. To ensure the capture of as much variability as possible, this data set included multiple types of naturally occurring variation in the experimental runs: multiple operators, different lab equipment, and multiple reagent lots.

Appendix B: Training reference models for copy number estimation

This section describes the process whereby the model parameters required for copy number estimation were derived.

A key requirement for this process was known examples for each of the CN levels for each region. Independent reference calls were obtained from a combination of CN estimates from the Genome-Wide Human SNP Array 6.0, Cogenics CYP2D6 commercial assays, and TaqMan® Assays. The sample set used included HapMap Caucasian, Asian, and Yoruban ethnicities, as well as some non-HapMap genomic DNAs. The number of known references for each region is summarized in Table 5.

Table 5: Counts of known CN levels among samples in the training set.

Region	Counts of unique samples			Frequency of CN=0
	CN=0	CN=1	CN≥2	
CYP2A6	5	52	441	1.0%
CYP2D6	4	89	424	0.8%
GSTM1	231	208	75	45.0%
GSTT1	155	234	124	30.2%
UGT2B17	145	182	192	27.9%

Probe set selection

The first step in training SNP-specific models is to use the reference genotype calls to determine the probe sets of maximal use for discriminating CN = 0 from CN > 0. There are four kinds of probe set available: CN ASO, genotyping ASO, CN tag, and genotyping tag.

Table 6 counts the SNP and tag probe sets within the region minimum and maximum for each copy number region.

Table 6: Available MIP probe sets per region.

Region	CN ASO	Genotyping ASO	CN Tag	Genotyping Tag
CYP2A6	32	19	32	24
CYP2D6	226	30	241	17
GSTM1	52	3	52	5
GSTT1	39	6	39	7
UGT2B17	71	4	71	6

Several genomic regions are subjected to mPCR amplification to disambiguate them from other, similar regions in the genome. It is possible to use markers in these amplified regions to estimate copy number, but the repeatability of mPCR introduces a new variance component into the prediction process. For this reason, markers located within mPCR amplicons are excluded from consideration for use in predicting CN.

A sample collection with as many examples as possible of CN = 0 and CN = 1 is used for training. All samples are normalized and summary values are computed for each CN probe set as described in Section 4. For each probe set, the samples with known CN are used to compute a linear discriminant, or Id-score, to quantify capacity for reliably discriminating CN.

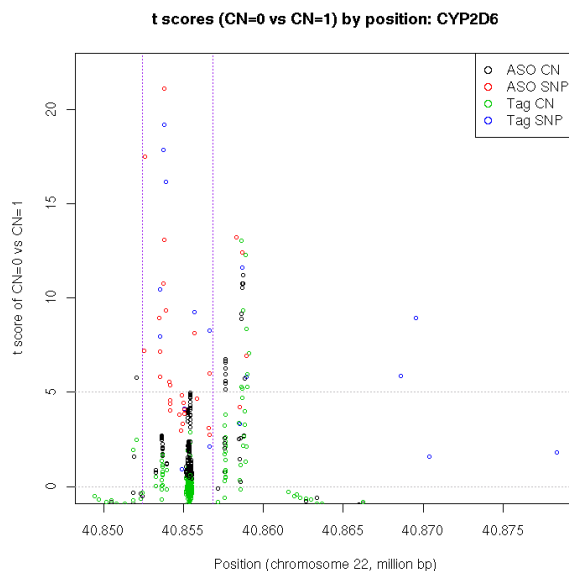
The number of samples, average, and standard deviation are computed based on all samples known to have CN = 0. Call these values n_0 , m_0 , and s_0 , respectively. Similarly, n_1 , m_1 , and s_1 are computed for the samples known to have CN = 1. The Id-score is then defined as

$$Id\text{-score} = \frac{(m_1 - m_0)}{\sqrt{\text{pooled variance}}}$$

$$\text{pooled variance} = \frac{n_1 \cdot s_1^2 + n_0 \cdot s_0^2}{n_1 + n_0}$$

Figure 9 shows the Id-scores as a function of position for the CYP2D6 region.

Figure 9: Id-score of each probe set in the CYP2D6 region plotted as a function of genomic position of the probe set. The Id-score indicates how well the probe set signal clusters by CN. High Id-scores, greater than five, are good. The dotted vertical lines show the transcription footprint of CYP2D6.



Distinguishing between CN = 0 and CN = 1 is the most important task, as the CN = 2 cluster is further from the CN = 0 cluster than the CN = 1 cluster. For this reason, the CN = 2 samples are not included with the CN = 1 samples when computing the Id-score, as this would shift the mean and inflate the variance of the CN > 0 cluster.

For the final CN model, the 10 probe sets with the highest Id-scores are used—experimenting with different numbers of probe sets to use had no discernible effect on performance. There are many criteria for choosing probes, but this simple method does about as well as any. The parameters for these 10 probe sets are then used to determine the region-level parameters by calculating for each sample of known CN the weighted CN estimate as described in Section 4, then computing a new set of means and variances based on these region-summarized values. These region-level values can also be used to compute a region-level Id-score to quantify the CN discrimination ability for each region. These region-level Id-scores are summarized in Table 7.

Note that CYP2A6 and CYP2D6 have particularly small representation for CN = 0, as these homozygous deletions are relatively rare. This makes estimates of their distribution parameters less precise. To reduce the risk of downstream problems from having few observations, a small number of pseudo-counts (five) is added to the number of observations for each cluster. Additionally, if the cluster variance is below a minimal threshold, it is brought up to the minimum value.

Table 7: Region-level Id-scores in the training set.

Region	# samples with CN=0	# samples with CN=1	Chrom	Id-score
CYP2A6	5	52	19	31.0
CYP2D6	4	89	22	8.7
GSTM1	231	208	1	47.5
GSTT1	155	234	22	22.5
UGT2B17	145	182	4	21.0