DMET™ Plus genotyping and copy number methods

This document describes the genotyping analysis methodology used with data produced by the DMET™ Plus Product. This method is designed to analyze samples in one of two ways: single-sample (individual) analysis, and dynamic clustering analysis. Single-sample analysis produces conservative, accurate results that do not depend on any additional data within a batch of experiments. When single-sample analysis is not required, dynamic clustering analysis can improve the genotyping call rate by using information from other samples in the batch to adjust the pre-defined cluster properties relative to the experimental conditions of the batch.

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 depend only on the experiment done rather than on any individual batch of results.

Finally, these summary values are compared to the clusters, either pre-defined or adjusted, 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 provides technical detail on each step described.

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 the alternative alleles at each of them.

Section 2 is devoted to pre-processing the data prior to 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 that can affect the hybridization properties of probes.

Section 3 explores the expected signal distributions. For single-sample analysis, these are the reference distributions, but for dynamic clustering analysis, the sample signals from the analysis batch can modify these distributions.

Section 4 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 of the expected scatter for well-behaved experiments.

Section 5 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 6 discusses the typical behavior and performance of this methodology when run in single-sample and dynamic modes, including interpretation of quality control results and some suggestions for identifying suspicious batches of data.

Finally, in three appendices, we describe the construction of the standard references for genotyping (Appendices A and C) 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 distinct DNA samples run with a variety of operators and equipment, with high-value markers verified by independent genotyping methods for a subset of the training set.

Section 1: Design of the DMET™ Plus Array

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

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
	Chromosome X	46
	Chromosome Y	0

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

Markers on DMET Plus are interrogated using molecular inversion probe (MIP) technology[1–4]. One or more MIPs is included for each of the genotyping markers in the assay probe pools. For some of the more important markers that have adjacent secondary polymorphisms, multiple MIPs were designed specifically for each of the

possible sequence variants (or contexts) in which the polymorphism of interest is located. For example, if a biallelic marker of interest has three adjacent biallelic SNPs, it would have a MIP for each of the eight possible sequence contexts. For indels, at least one MIP is designed for each allele, with each MIP using a different tag. This type of design has the additional benefit of allowing the opportunity to discriminate genotypes by the use of allele-specific tags. Each of the other markers shares a common tag across all MIPs used.

Each of the five CN regions has some genotyping markers contained within them 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 is 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 (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.

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

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

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 3 lengths from both strands with 3 replicates each, for a total of 3 x 2 x 3 = 18 probes. Allele-specific genomic sequences are interrogated with probes from 2 strands, up to 5 probe lengths, up to 9 offsets relative to the interrogation base and up to 3 identical replicates on the array— as many as 270 probes per allele-context (Table 3). Factoring in that the ASO probe set for each genotyping marker comprises 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.

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

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 equal to this value multiplied by the number of alleles and by the number of sequence contexts. For example, the number of ASO probes for a biallelic SNP with two adjacent biallelic SNPs would be 2,160 (= 2 x 2 x 2 x 270). Markers of critical importance (as determined by the ADME consortium) receive the fullest extent of interrogation; other markers are interrogated with fewer combinations of strands and lengths but with at least 132 ASO probes per allele-context.

Section 2: Preprocessing of data

A single DMET Plus Array contains slightly more than 1 million features, with a diverse assortment of probe sequences, including control probes of various kinds. For each marker, the signal intensities of 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 four-stage process of standardizing the data:

1. Normalizing global assay effects
2. Summarizing individual allele-specific probe sets
3. Reducing multiple probe sets to the biallelic case
4. 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 normalization[2] 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, and 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. 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 designed to hybridize with a particular target sequence; however, the probes may be of differing lengths and may overlap a marker at different offsets (i.e., the SNP is not necessarily in the central position of the probe). These differences lead to systematic intensity differences among the 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 calculation 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[5–7], 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 biallelic markers, the probe set summarization process described above results in two numbers, one associated with each allele. However, unlike previous SNP products, the DMET™ Plus Array contains markers with more complexity associated with them, such as trialleles and markers with additional nearby mutations. Trialleles are similar to biallelic 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 within 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 biallelic marker with 8 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 represent 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 biallelics, these two values are always the summary values for each allele. For trialleles, 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 biallelic 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 among 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 biallelic 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 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" = $\log_2(A) - \log_2(B)$ is constructed, 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: Expected signal distributions

Genotype calls are made by comparing the observed signal values for a marker with the expected signal distributions for different genotypes and choosing the genotype that yields the maximum likelihood for the observed data. The source of the expected distribution of signal values depends on whether the analysis mode is single sample or dynamic clustering. Single-sample analysis utilizes a fixed reference distribution for each genotype based on the results of genotyping a large training set at Affymetrix. Dynamic clustering, on the other hand, modifies the reference probe set intensity distributions based on observed signal values for the samples in the batch. Appendix A describes the creation of the reference distributions, which are based on the results of genotyping a large training set at Affymetrix.

The dynamic clustering analysis uses a Bayesian approach to combine the observed signal values with a prior based on the reference distributions to form a posterior

expected signal distribution. The prior and posterior distributions for each genotype are two-dimensional normal-inverse-Wishart distributions in contrast-strength space. For a given marker, the samples in the analysis batch are assigned genotypes using a one-dimensional contrast-based Gaussian maximum likelihood approach[8]. For each genotype, the associated samples are combined with the prior to generate the posterior distributions.

To allow better adaptation to different experimental conditions when the number of samples in a batch is relatively small, the number of observations for a given cluster saturates at a parametric value set to a small number. Some markers have close-lying clusters, and allowing the dynamic adjustment of cluster locations can lead to undesirable results. Therefore, for the original dynamic method, the number of observations for these markers is kept at their full values to minimize cluster adaptation. Appendix C lists these special markers.

Section 4: Genotyping calls and confidences

Although different approaches are used to set signal ranges, genotype calling operates in the same manner for both single-sample mode and dynamic-clustering mode. The genotyping process compares the observed signal values for a marker with the expected signal values for appropriate genotypes and chooses 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 of 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 unambiguous genotype assignment 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 trialleles, there are three separate potential pairs of alleles and a reference model for each pair. The model also depends on the copy number state of the marker. For typical markers with copy number 2, there are three possible genotypes—AA, AB, and BB—each of which is modeled by a Gaussian. For markers that are definitely known to have copy number 1 (such as on sex chromosomes in appropriate individuals), there are only two genotypes—A and B—hence, only two clusters. For markers definitely known to have zero copy number (as in samples in which a region is deleted), the model outputs only CN=0 with no genotype.

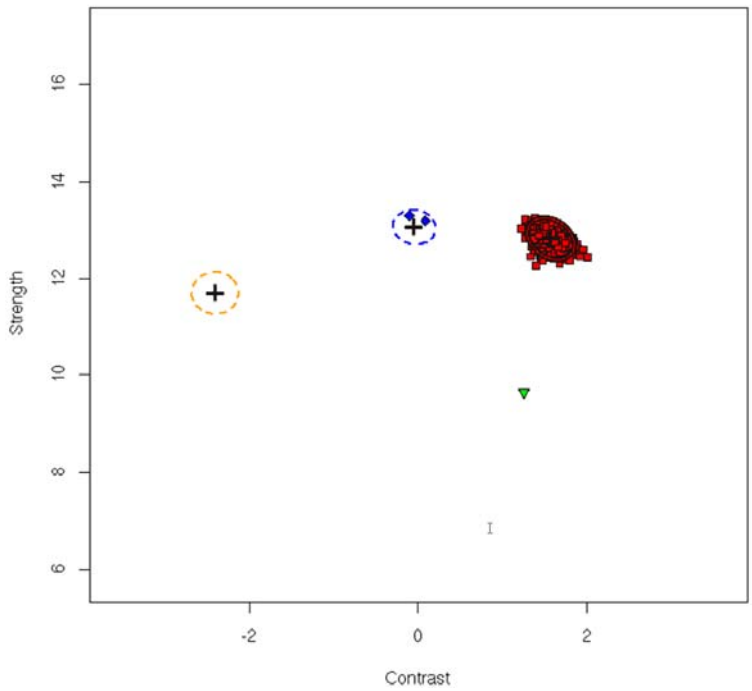

Figure 1: A marker showing copy number variation. The green triangle represents a call of CN=0 for the region and it is not carried through for further genotyping analysis. The ellipses are drawn at two standard deviations away from the center of the cluster and generally encircle the majority of the 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 in which the copy number is greater than zero, a prior model contains a two-dimensional Gaussian for each genotype and 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 (otherwise, untrained clusters would have a zero frequency and would 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 and strength values for a sample at a marker are compared to all clusters for the marker in the model. For each cluster, the likelihood of the data point is calculated assuming that the associated genotype is the true genotype and assuming 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 sum (likelihood of belonging to other clusters)/sum(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 also screens out 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, the genotype call is converted to a no-call and suppressed.

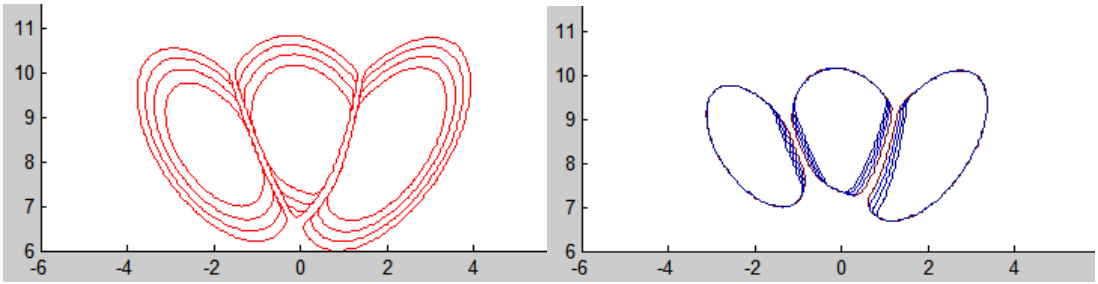

Figure 2: The left plot shows the possible calling regions for one marker when only the "ocean" parameter is tuned. The right plot shows the possible calling regions when both the ocean and confidence parameters are adjusted by the same amount. Judicious selection of these global parameters can be used to improve call rate without significantly degrading accuracy. The default values of these parameters vary across the DMET Plus genotyping methods used by DMET Console™ Software.

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 3 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 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. Data points that are located in ambiguous positions or 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 that were 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 5: Predicting chromosome copy number

The DMET Plus platform detects homozygous deletions (copy number = 0) in the five regions listed in Table 4 below.

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

Table 4: This table details the five copy number regions analyzed by DMET Plus. Coordinates refer to base positions in build 36 of the human genome. The gene min and max fields give the transcription footprints, including the untranslated region (UTR). The region min and max include the deletion footprint, which is often longer than the transcription footprint.

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 there is 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 3. 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. The LD score is defined as the ratio of the separation between cluster means and their pooled standard deviation (see Appendix B for exact definition).

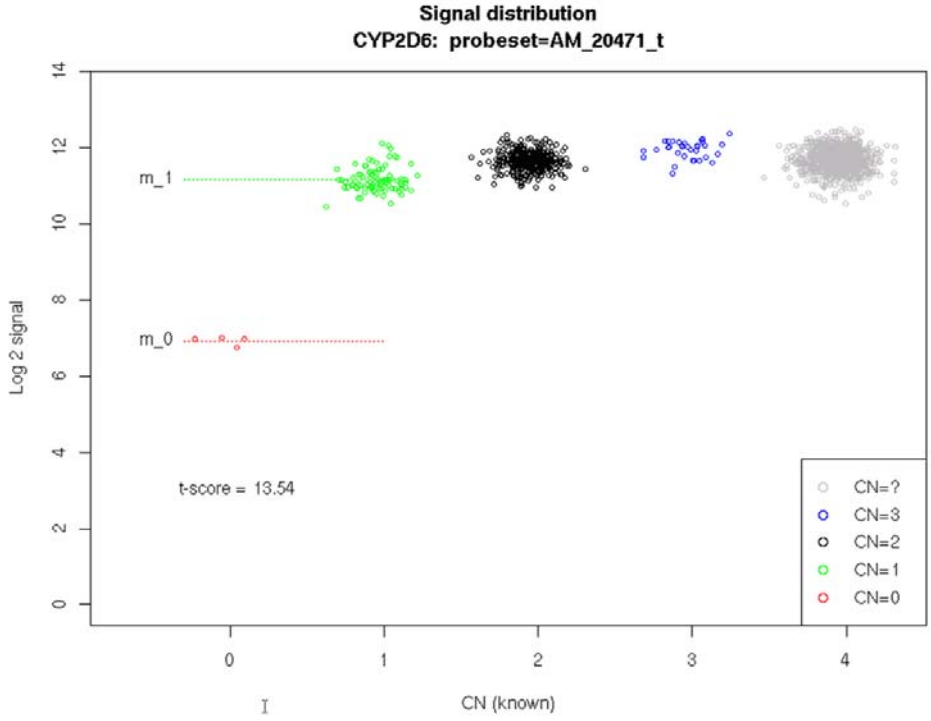

Figure 3: Plot of the signal values for probe set AM_20471_t, a tag probe set for a CN marker in the CYP2D6 region. The CN=0 samples form the red cluster with mean value m_0. The CN=1 samples form the green cluster with mean value m_1. The separation between these clusters has an LD score of 13.5. The 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 LD score quantifying separation. If these quantities are defined as m_0, m_1, and l, respectively, 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 LD scores, so 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 LD 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 4.

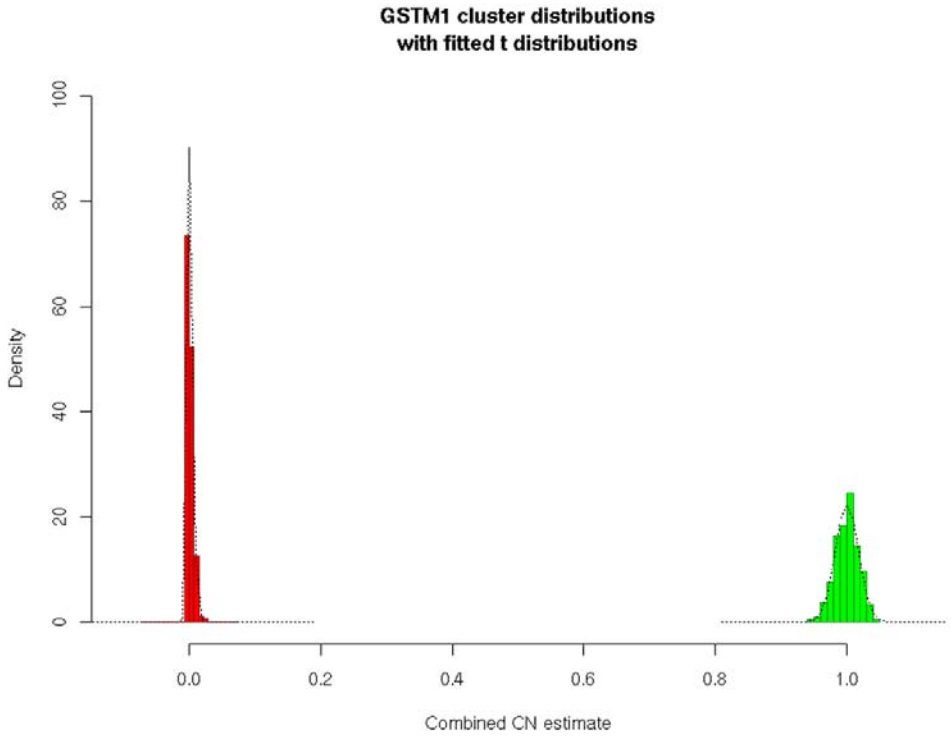

Figure 4: Histogram of combined CN estimates for region GSTM1 in a large sample collection. Only samples known to have CN=0 (red) or CN=1 (green) are shown; there is complete separation between the two.

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.

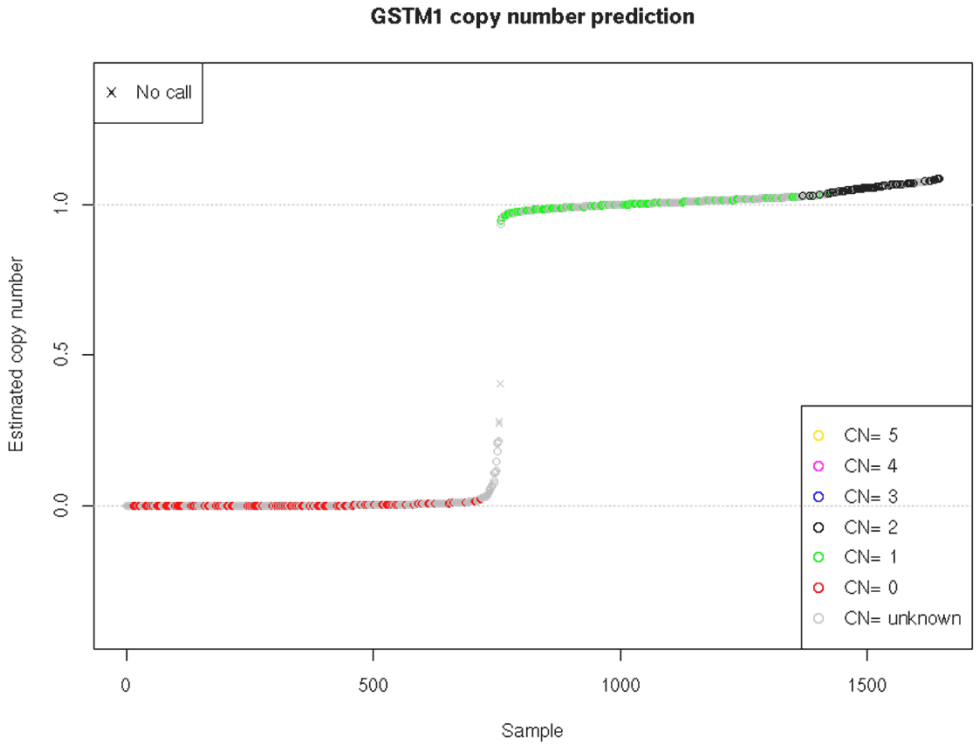

Figure 5: CN classification for roughly 1600 samples in the GSTM1 region. The y-axis shows the weighted CN estimate and the shape of points indicates how they were called. There are two no-calls in the region in between the two clusters; everything else is confidently assigned as either CN=0 or CN>0. Samples with known copy number are indicated by colored points as indicated in the legend.

Section 6: Outcomes of design decisions

The DMET™ Plus platform genotyping methodology supports two modes: single-sample and dynamic clustering. The conservative single-sample mode utilizes only the model provided and the data for an individual sample. Thus, the genotyping calls are completely independent of other concurrently analyzed samples. The parameters for the cluster model are therefore held invariant as a fixed standard for comparison. In contrast, dynamic clustering mode adapts to any shifts in genotype cluster positions specific to the experimental batch. Allowing the model to track the data causes the resulting genotypes to be influenced by what other samples were

analyzed at the same time because cluster boundaries are shifted with observations in the analysis set. In exchange for this dependence, the call rate generally improves as the expected cluster positions adjust to the properties of the batch.

In both modes, 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.

Although single-sample mode provides some tolerance for systematic shifts through 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, the confidence score will become poorer. Such an 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. As shown in Figure 6, this conservative behavior still results in average call rates in excess of 99.0% and average concordance to HapMap genotypes greater than 99.5% for development data sets. This data was genotyped using the original single-sample reference model and settings ("Fixed Genotype Boundaries" method in DMET Console Software).

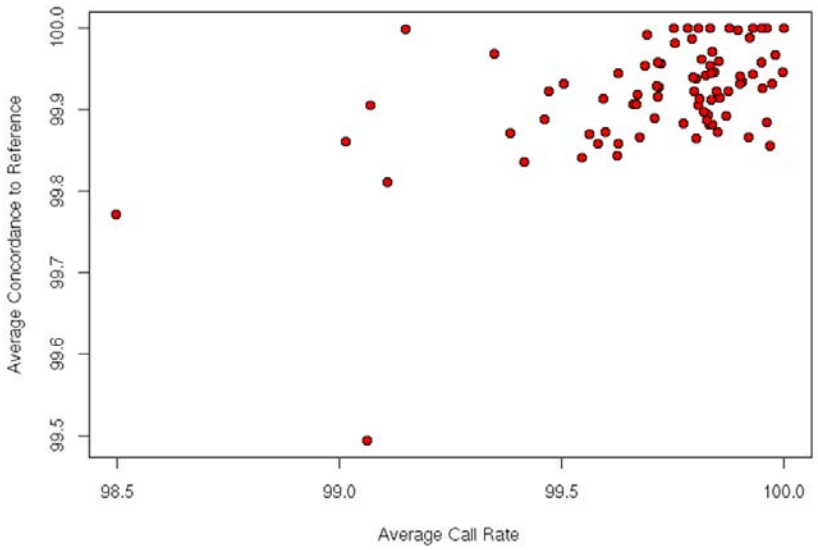

Figure 6: Observed performance of DMET Plus Array during product development. The DMET Plus Array was extensively tested, both at Affymetrix and at beta sites. There was a total of 82 batches of data (where run is defined as a single operator processing one week's worth of samples) with 23 of the batches coming from sites external to Affymetrix product development. The 82 batches include over 3,500 samples. Within each batch, any sample with a call rate less than 98% was rejected as a potentially problematic sample, leading to the exclusion of 4.4% of the samples attempted. For all remaining samples, concordance of the DMET Plus 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.

Dynamic clustering analysis enables the genotyping procedure to adapt to systematic variation from the training data, which can prove useful in improving call rates in certain data sets. The lower plot shows that allowing the clusters to adapt to the experimental data can dramatically improve the call rate.

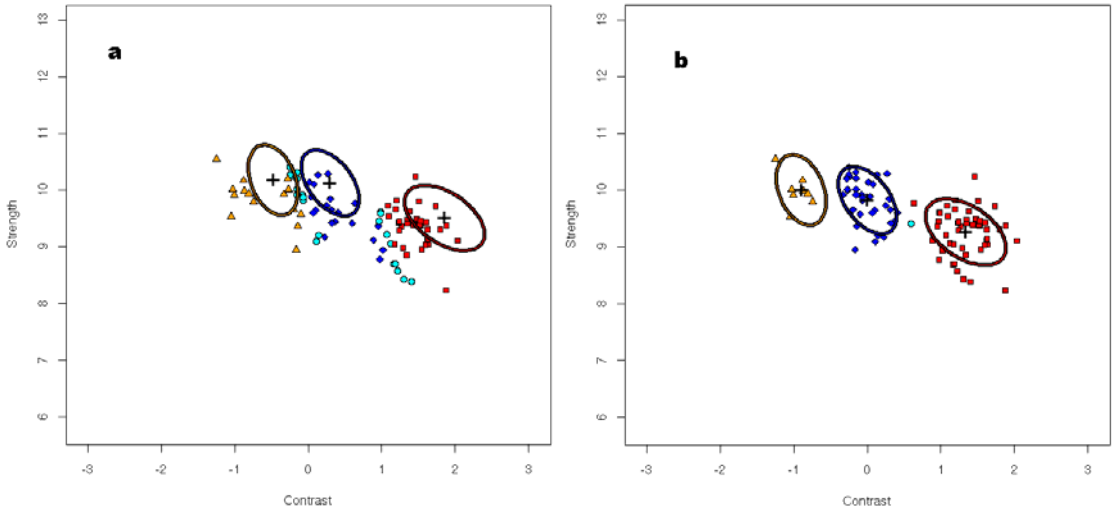

Figure 7: Comparison of the cluster plots for the UGT1A1*28 and co-located polymorphisms in an example data set colored by genotype call. The red cluster on the right represents combinations of $(TA)_6$ and shorter repeat lengths. The orange cluster on the left represents combinations of $(TA)_7$ and longer repeat lengths. The central blue cluster represents one allele that is $(TA)_6$ or shorter and another allele that is $(TA)_7$ or longer. The ellipses show the one standard deviation contour of the clusters. Panel a shows the single-sample genotyping analysis results; panel b shows the dynamic clustering analysis results. Allowing the clusters to adapt to the data substantially improves the call rate in this example.

Two years after the initial release of the DMET Plus Array, Affymetrix had the opportunity to evaluate the performance of the product in customers' hands. After reviewing four large data sets from customers, as well as replicate runs of an Affymetrix training plate by 19 customer sites, it was decided that customers would benefit from a tuning of the reference models used for both single-sample and dynamic clustering analysis methods. The changes involve updates to the global analysis parameters and to the reference models.

Analysis configuration	DMET Console Version			Comments
	1.2	1.1	1.0	
Fixed Genotype Boundaries – version 2	x			Recommended for general use
Dynamic Genotype Boundaries – version 2	x			Alternate method if conditions require its use
Fixed Genotype Boundaries	x	x	x	Legacy method
Dynamic Genotype Boundaries	x	x		Legacy method

Table 5: Supported genotyping methods in DMET Console Software.

Changes to the reference models for version 2 of the genotyping methods are discussed in Appendix A2. The global analysis parameters "ocean" and "confidence" (see Section 4) were adjusted to increase the calling region in areas with no nearby clusters, while also removing more of the ambiguous calls between neighboring clusters. A performance comparison of the available genotyping methods within DMET Console Software 1.2 is shown in Figure 8.

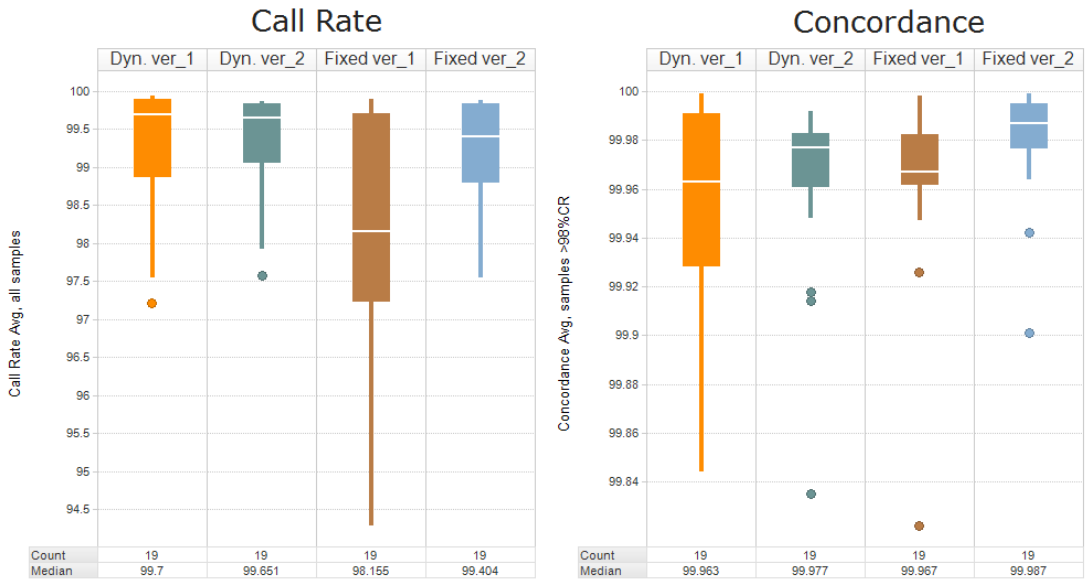

Figure 8: Nineteen runs of training plate samples were collected from customers and the call rate and concordance of each to the expected genotype calls were tallied using four genotyping methods. No samples were excluded for the average call rate metric. For the concordance metric, samples with a call rate less than 98% were excluded. The 98% cutoff is typically used to discriminate "in bounds" from "out of bounds" samples. "Dyn" is the dynamic boundaries method, "Fixed" is the single-sample method, and "ver_2" refers to the 2nd version of the models and settings that became available as of DMET Console Software 1.2.

References

1. Karlin-Neumann G., *et al.* Molecular inversion probes and universal tag arrays: Application to highplex targeted SNP genotyping. Cold Spring Harbor Lab, "Genetic Variation: A Laboratory Manual", p.199–211, Weiner M. P., Gabriel S. B. and Stephens J. C., eds. (2007).

2. Wang Y., *et al.* Analysis of molecular inversion probe performance for allele copy number determination. *Genome Biol* **8**:R246 (2007).

3 Wang Y., *et al.* Allele quantification using molecular inversion probes (MIP). *Nucleic Acids Res* **33**:e183 (2005).

4. Moorhead M., *et al.* Optimal genotype determination in highly multiplexed SNP data. *Eur J Human Genet* **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. *Nuc Acids Res* **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 whitepaper "BRLMM-P: A genotype calling method for the SNP 5.0 array". http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf

Appendix A1: Training original reference models for genotype calling

The reference models used with the DMET Plus platform are critical determinants of the genotypes and copy number estimates delivered by the product. For single-sample mode, the reference models serve as fixed distributions that determine the relative likelihood of the different genotypes. For dynamic clustering mode, the reference models for normalization and feature summarization are used in the same manner as in single-sample analysis; for genotyping, they represent the starting point or prior for the Bayesian update based on samples in the processing batch. This appendix provides an overview of the process followed during the development of the product that 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 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 5.0[8]) 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 some 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. Stated another way, the method is provided with a general description of where clusters should be located (e.g., BB genotypes should have higher B signal than AA genotypes). It then 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 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 is the typical scatter around the expected signal location, what is the rough relative frequency of a given genotype, and to what precision is this information 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, that is, 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.

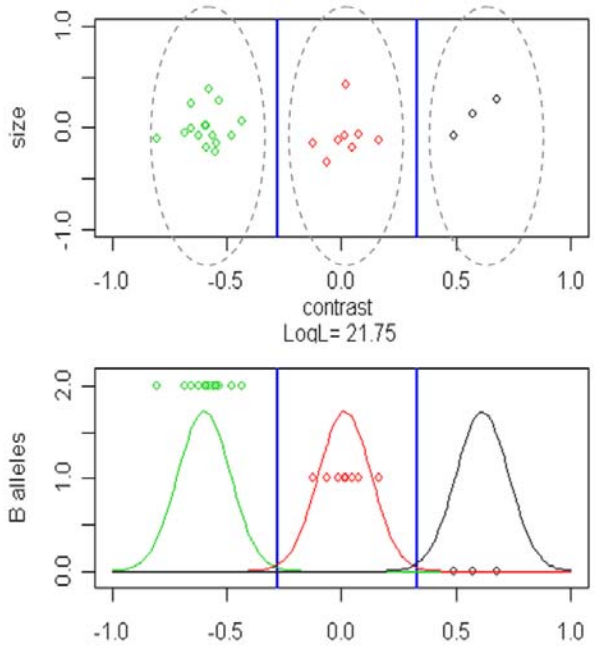

Figure 9: Tentative genotypes are computed by finding a hard labeling of the data with maximum likelihood under the posterior Gaussian model. Because plausible genotypes must have contrast(BB) < contrast(AB) < contrast(AA), an assignment of genotypes to data is exactly described by two transition points where the number of B alleles changes. All (n+2) x (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 non-zero 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 the 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 that sets 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. Because 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 platform 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, as well as producing a model that is consistently applicable to further data sets.

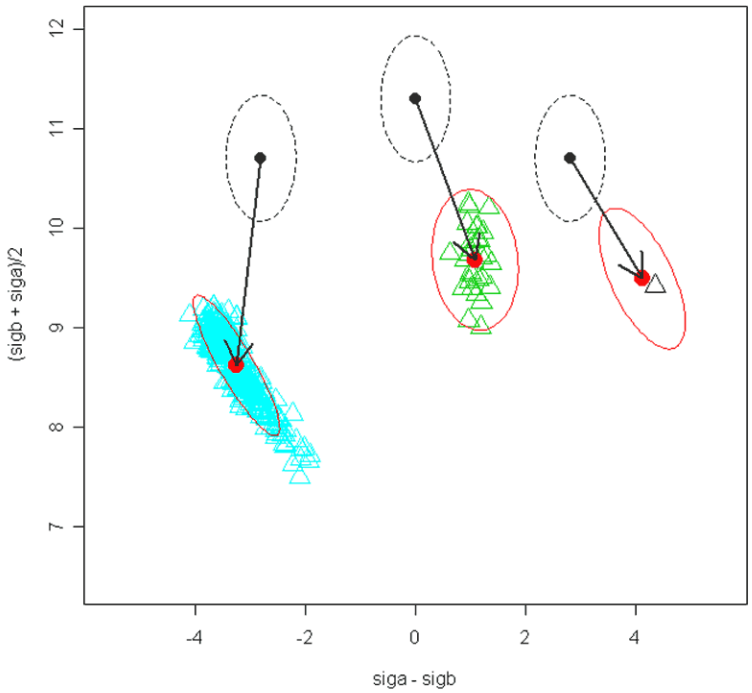

Figure 10: Visualizing a Bayesian update to the variance. Here the black dashed ellipses indicate the prior variances for clusters, and the solid ellipses indicate the posterior variance. Variances are shrunk to have common scale in X and Y directions, borrowing information from highly observed clusters 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 updated 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 $V(p+n) = pV0 + SS(observed) + kn(u-m)^2/(k+n)$, that is, the variance is the prior variance, plus the observed scatter within a cluster, plus the 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 that shrinks 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 analysis or used as a prior model for successive instances of training data to accumulate more details of marker behavior. For the DMETPlus platform, single-sample analysis uses the posterior model produced by the training runs (after manual curation), allowing the product to provide genotype calls relative to the fixed training set without depending on other samples. Dynamic clustering analysis uses the reference models as priors for additional adaption given the samples in the processing batch, enabling successful analysis for a wider range of assay conditions.

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 curated to ensure proper cluster labeling and proper positioning of unobserved clusters. Second, there were markers with unusual cluster properties due to the model being inapplicable, e.g., 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 curated 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 reclustered, allowing cluster centers to be close in absolute terms, and the resulting models were manually combined with the standard models generated by the automation.

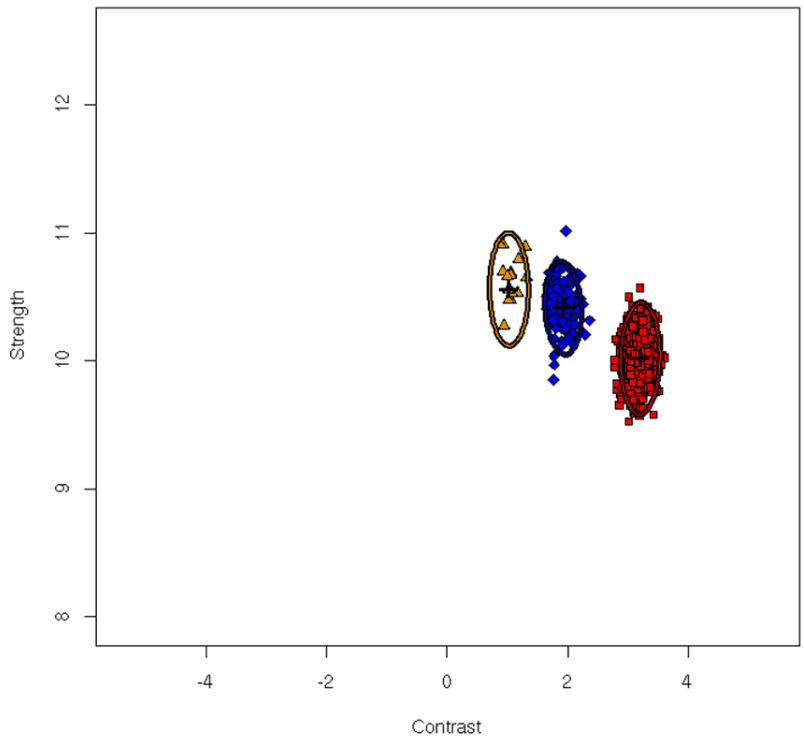

Figure 11: An example of an unusual marker. The location of the BB homozygous cluster has shifted to a position more typical of an AB heterozygous cluster. Further, the BB and AB clusters are unusually close together in absolute terms, even though their resolution is decent (separation of cluster means 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 dynamic clustering analysis 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. Large systematic shifts for individual markers can still lead to an increase in no-calls, indicating the inapplicability of the training data for such a marker and flagging 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 the least to correct classification of genotypes). For high-value, difficult markers, the probe set content was hand-edited 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, allowing for verification of concordance as well as providing 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 A2: Revision of reference models for genotype calling

Two years after the initial release of the DMET Plus Array, Affymetrix had the opportunity to evaluate the performance of the product in customers' hands. After reviewing four large customer data sets from customers, as well as replicate runs of an Affymetrix training plate by 19 customer sites, it was decided that customers would benefit from a tuning of the reference models used for both single-sample and dynamic methods. The updated genotyping methods, referred to as "version 2" in DMET Console Software 1.2, are discussed below.

First, the cluster models for all markers were visually inspected for a data set comprising ~2,400 samples. As part of the review, one-third of the markers had their cluster models adjusted. In most cases the adjustments did not result in any changed calls for these samples. However, some markers benefitted from reference model tuning (Figure 12).

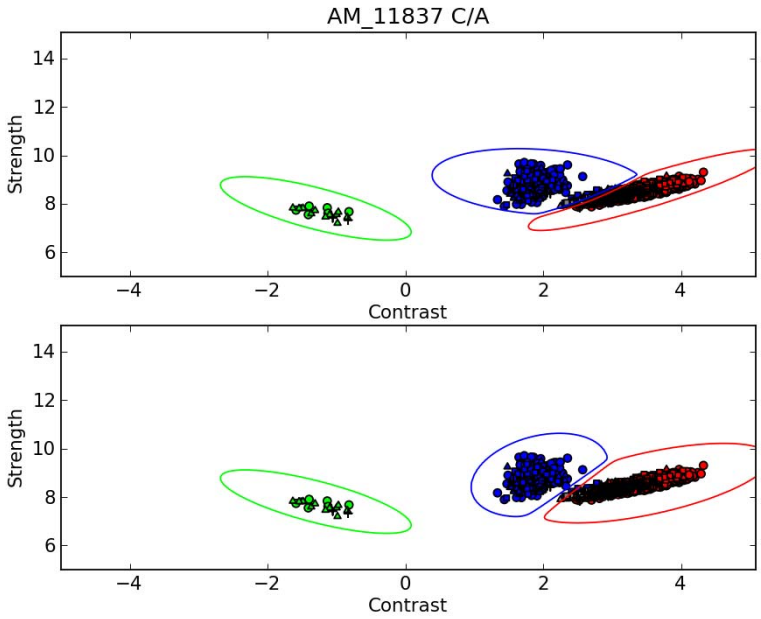

Figure 12: Cluster plots for one marker. The top plot shows the cluster boundaries used by the original fixed genotyping method available from DMET Console Software. It appears that some samples that should be assigned to the right (red) cluster are instead being captured by the middle (blue) cluster. The bottom plot shows the adjusted boundaries after changing only the reference models. Note that the final boundaries are a function of both the models (the "priors") and the analysis parameters, which are discussed in Section 4.

In the original data set used by Affymetrix to train cluster locations, some markers did not have sufficient samples representing the variant genotypes to properly define the cluster positions of the variant genotypes. When presented with samples that had these rare variant genotypes, the software would only report a possible rare allele (PRA) call. Fortunately, the additional data recently shared with Affymetrix includes samples with rare variant genotypes. In 47 cases in which there is now enough supporting data to confidently assign a genotype, the former PRA calls are now assigned a full call. An example is shown in Figure 13.

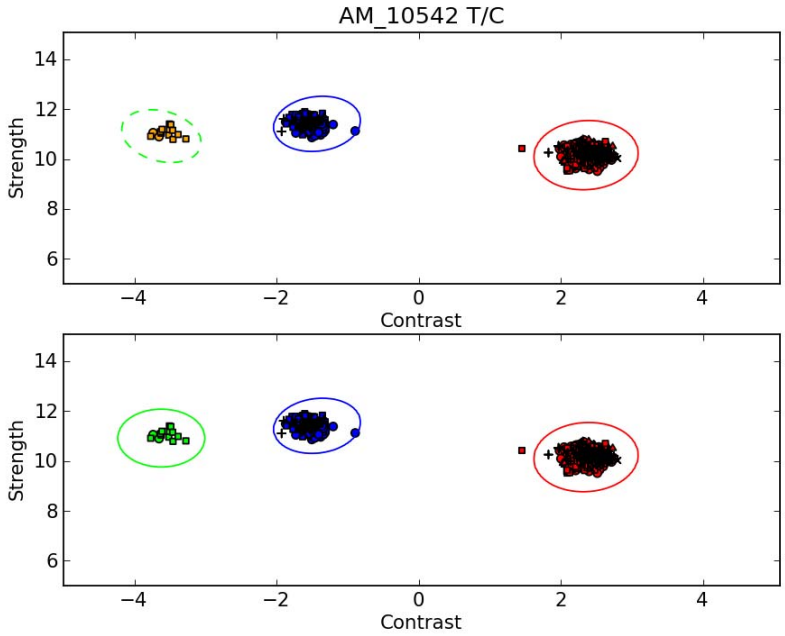

Figure 13: The top cluster plot shows the original model for one marker, in which the left cluster with the dashed boundaries would only report a PRA genotype if a sample was more likely to belong to it than to another cluster. This is because the original sample training set only had two samples with this rare genotype, fewer than the minimum of three required to report it as a full call (this minimum prior observations threshold is configurable in DMET Console Software). In the augmented data set, there are more samples with this rare genotype. As a result, the number of homozygous variant observations in the models file was adjusted upwards to reflect the increased confidence in the assignment of this call.

Table 6 summarizes the genotypes whose number of observations increased enough that they can now be fully called, instead of being reported as PRA.

Probe set ID	Common name	Reportable genotypes using version 2 genotyping methods (formerly PRA)
AM_14631	ABCB1_c.-1G>A	A/A
AM_10915	ABCC1_c.275C>T(S92F)	C/T
AM_10172	ABCC2_c.3396T>C(I1132I)	C/T
AM_14942	ABP1_c.-4132C>T	C/T
AM_13718	ADH5_c.-422G>C	C/C
AM_14467	AHR_c.65+125C>A	A/A
AM_11173	ALDH3A2_c.28C>T(Q10X)	C/T
AM_11808	ARNT_c.-60G>T	T/T

Probe set ID	Common name	Reportable genotypes using version 2 genotyping methods (formerly PRA)
AM_10591	ATP7B_c.2973G>A(T991T)	A/A
AM_10002	CHST3_c.*1155G>CorGG	C/G
AM_12527	CHST10_c.*39T>C	C/C
AM_10775	CYP1A1_1412T>C(I286T)	C/T
AM_10803	CYP1A2*6_5090C>T(R431W)	C/T
AM_11414	CYP2B6*3_18045C>A(S259R)	A/C
AM_11405	CYP2B6*8_13072A>G(K139E)	A/G
AM_11402	CYP2B6_12740G>C(P72P)	C/C
AM_10121	CYP2C9*12_50338C>T(P489S)	C/T
AM_10093	CYP2C9*13_3276T>C(L90P)	C/T
AM_12264	CYP2D6*9_2615delAAG	-/-
AM_10249	CYP2E1*2_1132G>A(R76H)	A/G
AM_11462	CYP2F1*3_11887C>T(P490L)	T/T
AM_11463	CYP2S1_1300G>A(P66P)	A/G
AM_14826	CYP3A4*10_14304G>C(D174H)	C/G
AM_14812	CYP3A4*11_21867C>T(T363M)	C/T
AM_14749	CYP3A5*4_14665A>G(Q200R)	A/G
AM_14790	CYP3A7*1D_-91G>A(Promoter)	A/G
AM_11302	CYP4F2_7207G>T(G185V)	T/T
AM_12051	FMO4_c.843C>T(F281F)	T/T
AM_14231	GSTA2_c.*149T>A	A/A
AM_10725	GSTZ1_c.124G>A(G42R)	A/A
AM_14982	NAT1*22_c.752A>T(D251V)	A/T
AM_14998	NAT2*19_c.190C>T(R64W)	C/T
AM_15335	ORM2_c.421G>C(G141R)	C/C
AM_14119	PPARD_c.-101-28005G>A	A/A
AM_10648	SLC15A1_c.1352C>A(T451N)	A/A
AM_10664	SLC15A1_c.22-40G>C	C/G
AM_10658	SLC15A1_c.351C>A(S117R)	A/C
AM_11743	SLC16A1_c.*145T>G	G/G
AM_14345	SLC22A1_c.113G>A(G38D)	A/A
AM_14350	SLC22A1_c.262T>C(C88R)	C/T
AM_10357	SLC22A8_c.913A>T(I305F)	T/T
AM_10731	SLC28A2_c.488T>G(L163W)	G/G
AM_12342	SLC5A6_c.282A>G(R94R)	G/G
AM_10542	SLCO1A2_c.38T>C(I13T)	C/C
AM_13591	SULT1B1_c.612A>C(E204D)	A/C
AM_12557	SULT1C2_c.179A>C(D60A)	A/C
AM_12414	XDH_c.837C>T(V279V)	T/T

Table 6: The genotypes that can now be reported because they appear to be present in at least three samples of an augmented data set. Previously, they were reported as PRA (when using the default setting "minimum prior observations = 3").

The second version of the dynamic genotyping method uses a reference model derived from the model used for the fixed boundaries genotyping method. Refer to Appendix C for a description of how the genotype cluster weights were edited.

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 Affymetrix 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 7.

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%

Table 7: Counts of known CN levels among samples in the training set.

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 sets available: CN ASO, genotyping ASO, CN tag, and genotyping tag. Table 8 counts the SNP and tag probe sets within the region minimum and maximum for each copy number 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

Table 8: Available MIP probe sets per region.

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 (LD) score to quantify capacity for reliably discriminating CN.

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

$$ld - score = \frac{m_1 - m_0}{\sqrt{\frac{n_1 s_1^2 + n_0 s_0^2}{n_1 + n_0}}}$$

Figure 14 shows the LD scores as a function of position for the CYP2D6 region.

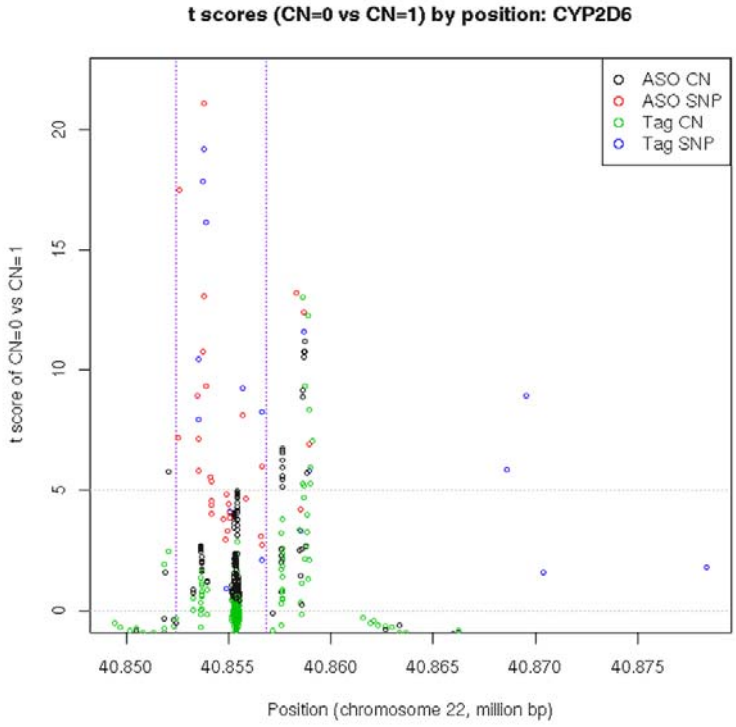

Figure 14: LD score of each probe set in the CYP2D6 region plotted as a function of genomic position of the probe set. The LD score indicates how well the probe set signal clusters by CN. High LD scores > 5 are good. The dotted vertical lines show the transcription footprint of CYP2D6.

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 LD 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 LD scores are used – experimenting with different numbers of probe sets to use had no discernable 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 LD score to quantify the CN discrimination ability for each region. These region-level LD scores are summarized in Table 9.

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 (5) 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.

Region	Samples with CN=0	Samples with CN=1	Chrom	LD 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

Table 9: Region-level LD scores in the training set.

Appendix C: Reference model changes for dynamic clustering analysis

The ability of samples in the analysis batch to update the reference signal distributions depends on the number of observations in the reference clusters. To facilitate reasonable updates for batch sizes of tens of samples, the number of observations per genotype is saturated at a small number. The strategy for selecting how to edit the number of observations depends on the dynamic method.

The first version of the dynamic genotyping method saturates the number of observations per genotype at 20. There are some markers with relatively close-lying clusters, however, that are better served by limiting the amount of update. For these markers, the number of observations remains unsaturated, in effect dramatically reducing the weight of samples in the analysis batch and limiting the cluster update. The markers for which the observations do not saturate were selected if they had a pair of genotype distributions for which the Fisher's linear determinant score was less than 20, with a small number of special exemptions. Table 10 lists the markers whose number of observations do not saturate.

The second version of the dynamic genotyping method uses a reference model derived from the updated single-sample model. For this method, the number of observations was usually linearly scaled down from its single-sample amount, with a maximum number of observations at 48. Genotypes having fewer than six observations were not scaled down for the dynamic method. The only markers whose number of observations were not scaled down were apparently monomorphic markers that were strongly tilted and had a nearby PRA cluster. These markers were

handled specially to reduce the likelihood of the PRA cluster drifting into the major homozygous cluster as part of the dynamic models adaptation process. Table 10 lists the markers whose number of observations do not saturate.

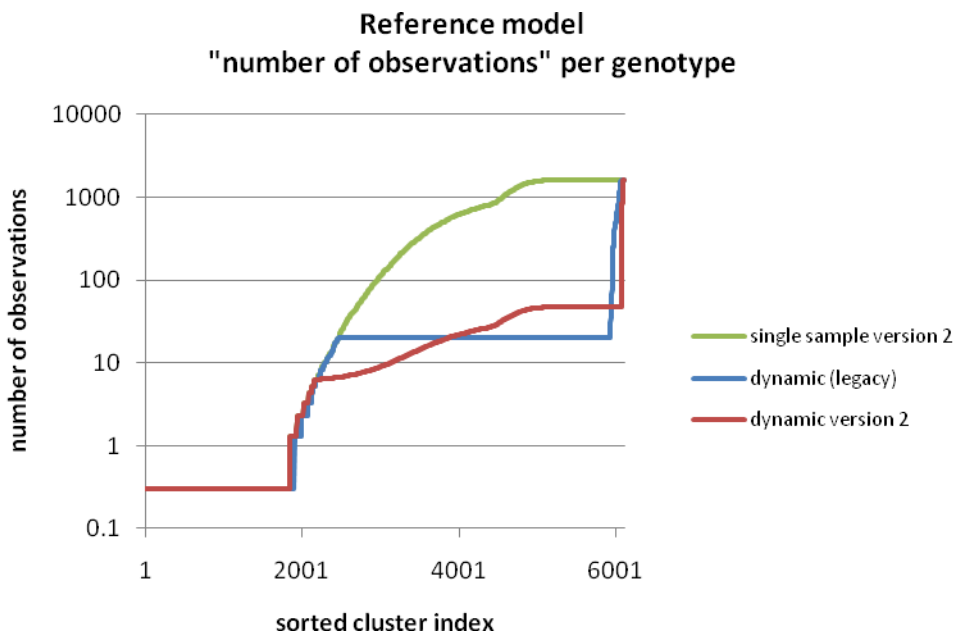

Figure 15: Comparison of cluster weights among reference model files. The original dynamic model file usually saturates weights at 20. Dynamic version 2 aims to preserve the relative weights of clusters. This change has a slight effect on call assignment of data between neighbor clusters of strongly dissimilar weights.

Table 10: Markers whose number of observations do not saturate in dynamic clustering mode.

Probe set ID	Category	Common name	Unsaturated model weight	
			Dynamic version 1	Dynamic version 2
AM_10100	Core	CYP2C9*2_3608C>T(R144C)	X	
AM_10135	Core	CYP2C8*3_2130G>A(R139K)	X	
AM_11647	Core	DPYD*2_c.1905+1G>A	X	
AM_11052	Pharma	VKORC1_c.85G>T(V29L)	X	
AM_14620	Non-core	ABCB1_c.729A>G(E243E)	X	
AM_10915	Non-core	ABCC1_c.275C>T(S92F)	X	
AM_10932	Non-core	ABCC1_c.2168G>A(R723Q)	X	
AM_10260	Non-core	ABCC8_c.4714G>A(V1572I)	X	
AM_13810	Non-core	ADH1C_c.1054C>A(P352T)	X	X
AM_14465	Non-core	AHR_c.-464G>A	X	
AM_14467	Non-core	AHR_c.65+125C>A	X	
AM_11202	Non-core	ALDH3A1_c.914A>T(Y305F)	X	
AM_11203	Non-core	ALDH3A1_c.741T>A(P247P)	X	

Probe set ID	Category	Common name	Unsaturated model weight	
			Dynamic version 1	Dynamic version 2
AM_15472	Non-core	ATP7A_c.4201G>C(V1401L)	X	
AM_15472	Non-core	ATP7A_c.4201G>C(V1401L)	X	
AM_10588	Non-core	ATP7B_c.*1172G>A		X
AM_11501	Non-core	CDA_c.154+3136T>C	X	
AM_13324	Non-core	CHST13_c.98-5237A>C	X	
AM_13342	Non-core	CHST2_-346G>C	X	X
AM_11126	Non-core	CHST5_c.510A>G(V170V)	X	X
AM_11127	Non-core	CHST5_c.490A>G(T164A)		X
AM_11128	Non-core	CHST5_c.-5A>G	X	
AM_15226	Non-core	CYP11B1_5573T>C	X	
AM_15258	Non-core	CYP11B2_4451T>C(I339T)	X	
AM_12498	Non-core	CYP1B1_81G>C(L27L)	X	X
AM_14021	Non-core	CYP21A2_G>C(E351D)	X	X
AM_14022	Non-core	CYP21A2_C>T(R379C)	X	X
AM_14023	Non-core	CYP21A2_G>A(G395S)		X
AM_10564	Non-core	CYP27B1_2595G>A(S356N)		X
AM_11349	Non-core	CYP2A6_1874G>T	X	
AM_11364	Non-core	CYP2A6*1D_-1013A>G	X	
AM_12278	Non-core	CYP2D6*29_1659G>A(V136I)	X	
AM_11459	Non-core	CYP2F1_5308G>C(V175L)	X	
AM_14826	Non-core	CYP3A4*10_14304G>C(D174H)		X
AM_14856	Non-core	CYP3A43_14956C>T	X	X
AM_14781	Non-core	CYP3A7*2_26041C>G(T409R)	X	
AM_11310	Non-core	CYP4F2*2_34T>G(W12G)	X	
AM_11279	Non-core	CYP4F3_11466G>A(P348P)	X	
AM_11280	Non-core	CYP4F3_11496A>G(V358V)	X	
AM_11608	Non-core	CYP4Z1_c.876+394T>G	X	
AM_11609	Non-core	CYP4Z1_c.1170T>C(I390I)	X	
AM_11611	Non-core	CYP4Z1_c.1202-2730A>C	X	
AM_15086	Non-core	CYP7B1_1678T>C	X	
AM_12087	Non-core	EPHX1_c.128G>C(R43T)	X	
AM_12112	Non-core	EPHX1_c.1216T>C(L406L)	X	
AM_12057	Non-core	FMO4_c.1250+591C>T	X	
AM_11913	Non-core	FMO6_1232G>A	X	
AM_15492	Non-core	G6PD_c.1466G>T(R489L)	X	
AM_15492	Non-core	G6PD_c.1466G>T(R489L)	X	
AM_15493	Non-core	G6PD_c.1093G>A(A365T)	X	X
AM_15493	Non-core	G6PD_c.1093G>A(A365T)	X	X
AM_15494	Non-core	G6PD_c.653C>T(S218F)	X	

Probe set ID	Category	Common name	Unsaturated model weight	
			Dynamic version 1	Dynamic version 2
AM_15498	Non-core	G6PD_c.319G>T(V107L)		X
AM_14253	Non-core	GSTA1_c.504G>A(E168E)	X	
AM_14254	Non-core	GSTA1_c.501C>G(V167V)	X	
AM_14287	Non-core	GSTA1_c.-5184G>T	X	
AM_14244	Non-core	GSTA2_c.-10G>C	X	
AM_11702	Non-core	GSTM1_c.84T>C(Y28Y)	X	
AM_11705	Non-core	GSTM1_c.178-78T>C		X
AM_11710	Non-core	GSTM5_c.-524C>T	X	
AM_10434	Non-core	GSTP1_c.21C>G(V7V)	X	
AM_13243	Non-core	NR1I2_c.418G>A(V140M)	X	
AM_11837	Non-core	NR1I3_C>A(rs11265572)	X	
AM_15317	Non-core	ORM1_c.199T>C(F67L)	X	
AM_11219	Non-core	PGAP3_c.*560C>T	X	
AM_11209	Non-core	PNMT_c.26A>G(N9S)		X
AM_11218	Non-core	PNMT_c.826T>A(W276R)		X
AM_14675	Non-core	PON1_c.582G>A(W194X)	X	
AM_14519	Non-core	POR_c.*372G>A	X	
AM_14128	Non-core	PPARD_c.-101-15034G>A		X
AM_11032	Non-core	PRSS53_c.89A>G(Q30R)	X	
AM_15349	Non-core	RXRA_c.1242-27G>A	X	
AM_15353	Non-core	RXRA_c.*2102G>TorC	X	X
AM_15353	Non-core	RXRA_c.*2102G>TorC	X	X
AM_12214	Non-core	SLC19A1_c.696T>C(P232P)	X	
AM_12215	Non-core	SLC19A1_c.246C>G(P82P)	X	
AM_12216	Non-core	SLC19A1_c.80A>G(H27R)	X	
AM_14377	Non-core	SLC22A1_c.*15G>A(3'UTR)	X	
AM_10379	Non-core	SLC22A11_c.91A>G(I31V)	X	X
AM_10384	Non-core	SLC22A11_c.464T>G(V155G)		X
AM_14410	Non-core	SLC22A3_c.360C>T>G(R120R)	X	
AM_13921	Non-core	SLC22A5_c.12C>G(Y4X)	X	X
AM_10360	Non-core	SLC22A8_c.779T>G(I260R)	X	X
AM_10810	Non-core	SLC28A1_c.124T>C(L42L)	X	X
AM_14172	Non-core	SLC29A1_c.687G>A(L229L)		X
AM_10417	Non-core	SLC29A2_c.288G>A(T96T)	X	
AM_10418	Non-core	SLC29A2_c.204C>A(N68K)	X	
AM_10518	Non-core	SLCO1A2_c.2003C>G(T668S)	X	
AM_10525	Non-core	SLCO1A2_c.968T>C(L323P)	X	X
AM_10497	Non-core	SLCO1B1*16_c.452A>G(N151S)		X
AM_10851	Non-core	SLCO3A1_c.604_605AT>TA(I202Y)	X	X

Probe set ID	Category	Common name	Unsaturated model weight	
			Dynamic version 1	Dynamic version 2
AM_12159	Non-core	SLCO4A1_c.617G>A(R206K)	X	X
AM_12160	Non-core	SLCO4A1_c.797-286T>C	X	
AM_10985	Non-core	SULT1A2_c.888A>G(R296R)	X	
AM_12588	Non-core	SULT1C4_c.-483C>T	X	
AM_14921	Non-core	TBXAS1_c.487C>A(L163I)	X	
AM_14924	Non-core	TBXAS1_c.772A>G(K258E)	X	
AM_14933	Non-core	TBXAS1_c.1273C>T(R425C)	X	
AM_13005	Non-core	UGT1A3_c.31T>C(W11R)	X	
AM_13007	Non-core	UGT1A3_c.81G>A(E27E)	X	
AM_13008	Non-core	UGT1A3*4_c.133C>T(R45W)	X	
AM_13011	Non-core	UGT1A3_c.477A>G(A159A)	X	
AM_13425	Non-core	UGT2B17_c.1313+840A>G	X	
AM_13464	Non-core	UGT2B7*2_c.801T>A(P267P)	X	
AM_11061	Non-core	VKORC1_c.-5014T>C(Promoter)	X	X