Genotyping by quantitative heteroduplex analysis: Theoretical derivation and experimental verification of optimal DNA mixing.

RA Palais, MA Liew, and CT Wittwer

Abstract

Detection of heteroduplexes effectively screens for heterozygotes, but usually does not distinguish between different homozygotes. If DNA of known genotype is added to PCR products of a sample of unknown genotype, and the mixture is heated then cooled, the absence or presence of heteroduplexes will indicate whether what was added had the same or different genotype as the sample. Depending upon the proportions mixed, the heterozygote still could be distinguished from both homozygotes by quantifying the amount of heteroduplexes produced. Instead of mixing after PCR, which requires two analyses and increases the risk of product contamination, we propose adding DNA of known homozygous genotype to each unknown before PCR. Our analysis suggests that when homozygotes are most similar, the quantity of additional DNA which optimizes the separation of the three genotypes of bi-allelic, diploid DNA comprises one-seventh of the resulting mixture. Experimental verification with both high-resolution melting analysis and quantitative temperature gradient capillary electrophoresis (qTGCE) confirmed this prediction, and the importance of the correct proportion. If the proportion is one-third or one-half, for instance, some genotypes are virtually indistinguishable. When combined with high-resolution melting analysis, this technique requires only one close-tube analysis for full genotyping.

Introduction

Heteroduplex analysis is a popular technique to screen for sequence variants in diploid DNA. After PCR, heteroduplexes are usually separated by conventional gel electrophoresis (1), although denaturing high pressure liquid chromatography (DHPLC, 2) and temperature gradient capillary electrophoresis (TGCE, 3) can be used. Recently, heteroduplexes have been detected in solution without separation by high-resolution melting analysis. Either labeled primers (4) or a saturating DNA dye (5) were used to detect a change in shape of the fluorescent melting curve when heteroduplexes were produced. High-resolution melting of PCR products from diploid DNA has been used for mutation scanning (6-8), HLA matching (9), and genotyping (5, 10).

Heteroduplex analysis is seldom used for genotyping because different homozygotes are usually not separated. In some cases, DHPLC may separate PCR products by size (11). However, both DHPLC and TGCE usually fail to detect homozygous single base changes, small insertions and deletions. If suspected, these homozygous changes can be detected by mixing the PCR product with a known homozygous PCR product. However, two sequential analyses are required and the concentrated PCR product is exposed to the laboratory, increasing the chance of PCR product contamination of subsequent reactions.

In contrast to DHPLC and TGCE, different homozygotes can usually be distinguished by high-resolution melting analysis. Complete genotyping of human SNPs is possible

in over 90% of cases because different homozygotes differ in melting temperature (10). However, in some cases the two homozygotes cannot be distinguished and mixing studies are necessary. When samples are mixed after PCR, equal volumes of PCR products are combined, denatured, annealed, and melted. Alternatively, unknown DNA can be mixed with known homozygous DNA before PCR. If the mixed samples have the same genotype, no heteroduplexes will be produced.

If the mixed samples are not the same, different amounts of heteroduplexes will be produced, depending on the genotype and the amount of homozygous DNA added.

Previously, we empirically determined that the optimum amount of known homozygous DNA to distinguish all SNP genotypes was approximately 15%. We now present a rigorous derivation of this optimum, by analyzing the theoretical heteroduplex content and its contribution to melting curves and TGCE measurements across a full spectrum of genotype mixing proportions, which are also of interest in pooled sample studies. We then verify the close agreement of the theory with experiment using both high-resolution melting analysis and TGCE. The experiments emphasize the importance of using the correct proportion of added DNA: If the proportion is one-third or one-half rather than one-seventh, for instance, some genotypes can be virtually indistinguishable.

Mathematical model for melting curves and heteroduplex content

In Figure 1, we show high-resolution melting curves of amplicons from DNA exhibiting three SNP genotypes. The melting curve corresponding to samples with a homozygous mutation is indistinguishable from that of the wild type, due to nearest-neighbor thermodynamic symmetry. (This says that the bases immediately surrounding the mutation are identical when the strands are interchanged, e.g.,

5'-TCA-3' 5'-TGA-3'
3'-AGT-5' 3'-ACT-5'

The melting curves corresponding to heterozygous samples, which are initially equal parts wild-type homoduplexes and mutant homoduplexes are quite different than the melting curves of either of these species of duplex. Even though PCR amplifies all strands in the form of homoduplexes, by the time it plateaus, strands are reassociating randomly into homoduplexes and heteroduplexes instead of extending. The DNA that is melting in the heterozygous case is an equal mixture of four species of duplex, the wild-type and mutant heteroduplexes, and two types of nearly complementary heteroduplexes, so the total heteroduplex content is $\frac{1}{2}$ (Figure 2-Rob's).

There are only complementary strands amplified from wild-type and homozygous mutant samples, so even at the end of PCR, there are no heteroduplexes present. After PCR and analysis has determined a sample to be indistinguishable wild-type or mutant homozygous, a procedure sometimes known as spiking can be performed, which consists of of adding DNA of known genotype to determine genotypic identity or difference from the absence or presence of heteroduplexes.

Our goal is to find a method requiring no post-PCR mixing (which is suceptible to contamination) which can distinguish wild-type and homozygous mutant from each other, as well as from heterozygous samples, in one-step. To do so, we seek the optimal fraction of wild-type DNA to be added to samples before PCR, so that after amplification and random reassociation of strands, the heteroduplex content of mixtures with the different genotypes will make the resulting melting curves most distinguishable.

The mixture with a wild-type samples will still have no heteroduplex content regardless of the amount of the identical DNA which is added. In contrast, if wild-type DNA is mixed with a homozygous mutant sample, even though PCR amplifies all strands of the mixture as homoduplexes, by the time it plateaus (or after heating then cooling the mixture to promote random reassociation) a fraction of heteroduplexes will be formed, depending on the amount of wild-type DNA added. If wild-type DNA is mixed with a heterozygous mutant sample, the mixture will now consist of unequal parts of wild-type homoduplexes and mutant homoduplexes. At the end of PCR, or after dissociating by heating and annealing by cooling, a reduced fraction of heteroduplexes will be formed depending on the amount of wild-type DNA added. As the heteroduplex enhanced homozygous mutant melting curve moves away from the wild-type melting curve, the heteroduplex reduced heterozygous melting curve moves toward them both. We seek the point where the three are best separated.

Therefore, we develop a model for the melting curve of a mixture of genotypes after amplification and reassociation, in terms of the melting curves of the constituent duplex types and the mixture fraction. In the case described above when the homozygous mutant and wild-type curves are indistinguishable, the result of our analysis shows that the difference between melting curves of different genotypes is given by difference in the heteroduplex content of the mixtures times the difference between the common homoduplex melting curve and the average heteroduplex melting curve. (Although the two homoduplexes have the same melting curves, the two heteroduplex species have different thermodynamic melting behavior, even with respect to nearest-neighbor approximation.) We then optimize the separation of heteroduplex contents of the three genotypes in terms of mixture fraction before testing the model experimentally.

We begin in Table 1a which lists the type and fraction of homoduplexes are amplified when we mix wild-type and unknown DNA in fraction x of wild-type to $1 - x$ of either wild-type, homozygous mutant, or heterozygous mutant. We will refer to x, which measures the fraction of wild-type DNA added as a proportion of the mixture of sample plus added DNA, as the mixture fraction. When $x = 0$, the mixture consists of just the original genotype. When $x = 1$, it describes a pure wild-type sample regardless of the initial unknown genotype. Table 1b lists the type and fraction of duplexes which result when complementary and nearly complementary strands reassociate independently to form duplexes when PCR has plateaued, or after subsequent heating and cooling. The assumption of independence states that complementary and nearly complementary strands associate in proportion to their concentration, so we obtain the fractions in Table 1b simply by multiplying the corresponding fractions from Table 1a.

Corresponding to each duplex species d_j will be its normalized fluorescence curve, $F_j(T)$. The function $F_j(T)$ refers to the curve with background fluorescence removed and accounts for factors such as lower fluorescence per duplex behavior of heteroduplexes. Table 1c gives the normalized melting curves for the mixture fraction x of wild-type DNA and $1-x$ DNA sample of each of the genotypes, Assuming fluorescence is additive, We use $W(T,x)$ for wild-type sample, $M(T,x)$ for mutant homozygous sample, and $H(T,x)$ for heterozygous sample (T is temperature).

Our goal is to maximize our ability to distinguish these three melting curves, as measured by the minimum separation between any two of them, defined as the maximum absolute value of their difference:

$$\max_{x\in[0,1]} \min\{\max_T |W(T,x)-M(T,x)|, \max_T |W(T,x)-H(T,x)|, \max_T |H(T,x)-M(T,x)|\}. \tag{1}$$

Table 1d gives expressions for the differences $W-M$ and $W-H$, and after simplifying, we will determine $H-M$ as the difference of these differences.

Since we are most interested in the situation where $F_1(T)$ and $F_4(T)$ are difficult to distinguish, or we would not be adding DNA of one type to do so, we assume $F_1(T) = F_4(T)$. This condition is predicted by the nearest-neighbor approximation for SNPs with the symmetry described in the previous section. This symmetry does not make F_2 and F_3 agree, but this is not necessary for our derivation. Table 1e gives simplified forms for the difference curves in this case, indicated by the subscript $_=$. They are obtained by first replacing F_4 by F_1, combining terms removing a common factor from the resulting coefficients of F_1, F_2, and F_3, then writing $F_1 = \frac{F_1+F_4}{2}$.

The result justifies our earlier statement that when the homozygous mutant and wild-type curves are indistinguishable, the difference between melting curves of different genotypes is given by difference in the heteroduplex content of the mixtures times the difference between the common homoduplex melting curve and the average heteroduplex melting curve:

$$W_=(T,x) - M_=(T,x) = m(x)F(T) \tag{2a}$$

$$W_=(T,x) - H_=(T,x) = h(x)F(T) \tag{2b}$$

and therefore

$$H_=(T,x) - M_=(T,x) = (h(x)-m(x))F(T), \tag{2c}$$

where

$$F(T) = \frac{F_1(T)+F_4(T)}{2} - \frac{F_2(T)+F_3(T)}{2} \tag{3}$$

is the difference between the average homoduplex melting curve and the average heteroduplex melting curve, and

$$m(x) = 2x(1-x) \tag{4a}$$

$$h(x) = \frac{1-x^2}{2} \tag{4b}$$

$$h(x) - m(x) = \frac{3x^2 - 4x + 1}{2}. \tag{4c}$$

represent the total heteroduplex content difference between the various genotype mixtures with wild-type DNA. The graphs of the absolute values of the three heteroduplex content difference functions are given in Figure 3a, annotated with key features derived below. The graphs of the theoretical duplex melting curves $F_j(T)$ and the corresponding $F(T)$ are given in Figure 3b.

These expressions uncouple the x (mixture fraction of added DNA) and T (temperature) dependence of the differences among fluorescence curves of different genotypes, and identify the x dependence of the separation of melting curves of different genotypes as proportional solely to the difference in heteroduplex content of the mixtures. This allows us to optimize the mixture fraction independently, regardless of the specific nature of individual duplex curves contributing to the superpositions. This is perhaps surprising, since in general, different combinations of different functions result in varying shapes and locations of their extrema. Since all expressions are linear, we can obtain the difference of the negative derivative curves of the different genotypes simply by replacing F_j in the expressions above with their negative derivatives, and all of the results we will obtain regarding maxima and minima will still hold.

Our original genotype separation problem has become

$$\max_{x\in[0,1]} \min\{m(x), h(x), |h(x) - m(x)|\}F, \tag{5}$$

where

$$F = \max_{T} F(T) \tag{6}$$

represents the twice the separation of the original (no added wild-type DNA) heterozygous curve from the common original wild-type and homozygous SNP curves, since $h(0) = \frac{1}{2}$. So all that remains to optimize the mixture fraction x is to maximize the separation by maximizing the smallest of the three heteroduplex content differences:

$$s(x) = \min\{m(x), h(x), |h(x) - m(x)|\}. \tag{7}$$

In Figure 3a, this corresponds to finding the highest point on the lowest of the three graphs of m, h, and $|m - h|$. Table 1f identifies three intervals on which the ordering of these functions is preserved. We can find these intervals by finding the points where two of them

become equal: $h(x) = m(x)$ when $h(x) - m(x) = \frac{3x^2-4x+1}{2} = \frac{1}{2}(3x-1)(x-1) = 0$, at $x = 1/3$ and $x = 1$, and $m(x) = h(x) - m(x)$ when $h(x) - 2m(x) = \frac{7}{2}x^2 - 4x + \frac{1}{2} = \frac{1}{2}(7x-1)(x-1) = 0$, at $x = \frac{1}{7}$ and $x = 1$.

Calculus confirms our visual intuition that the maximum of of the lowest graph can only occur where the slope of its tangent is zero, or it does not have a well-defined tangent. (If $s(x)$ has a local extremum at $x = a$, then $s'(a) = 0$ or $s'(a)$ does not exist.) The only place $s'(x) = 0$ is halfway between its roots $\frac{1}{3}$ and 1 (since it is quadratic in this interval), i.e., at $x = \frac{2}{3}$. This corresponds to adding twice as much wild-type DNA as there was unknown DNA and gives a separation of $\frac{1}{6}F$, between the heterogygous and homozygous SNP curves, or $\frac{1}{3}$ of the original separation between the heterozygous melting curve and the other two. The separation between the wild-type melting curve and the other two melting curves will be larger.

The only places $s(x)$ is not differentiable is where it changes form, i.e., at $x = \frac{1}{7}$ and $x = \frac{1}{3}$. Comparing the values at these points and $x = \frac{2}{3}$, we find the optimal mixture fraction occurs at $x = x_* = \frac{1}{7}$, as indicated in Figure 3a. Here, the melting curve of the mixture with homozygous DNA is halfway between the other curves at the temperature of maximum separation, where its separation from each is $\frac{12}{49}F$. This is only barely less than half of the original separation of $\frac{1}{2}F = \frac{24}{48}F$ between the heterozygous melting curve and the other two genotype curves.

A simple heuristic explanation for this value, corresponding to adding one part wild-type DNA to six parts unknown sample is based upon the observation that the melting curves will be optimally separated when the homozygous mutant curve becomes equidistant from both the wild-type and heterozygous melting curves, so the heteroduplex content of the wild-type-heterozygous mixture must be exactly twice that of the wild-type-homozygous mixture: $h(x) = 2m(x)$ The ratio of 1 part wild-type to 6 parts unknown is optimal because 6 is the unique number which can be divided in equal parts ($3 + 3$, the heterozygous sample strands), and when one of the parts (3, SNP strands) is multiplied by the other plus 1 ($4 = 3 + 1$, wild-type sample plus added wild-type strands) one obtains exactly twice the original number (6, the homozygous SNP strands) multiplied by 1 (added wild-type strands.) At the simplest level, it is because $(\frac{6}{2})(\frac{6}{2} + 1) = 2(6)(1)$ that $x_* = \frac{1}{7}$ of the total unknown plus added DNA is the optimal mixture fraction. This is visualized in the animation

http://www.math.utah.edu/∼palais/pcr/michael/spike/spike.html

When the wild-type and homozygous SNP curves are not the same, the max-min problem does not separate into individual mixture fraction and temperature problems. In this case, the optimal mixture fraction still may be characterized by the two-dimensional generalization of the above criteria, but it must be found numerically depending upon the specific curves $F_j(T)$. That is, the mixture fraction-temperature rectangle $[0, 1] \times [T_1, T_2]$ must be divided into regions on which the corresponding $h(x, T) - m(x, t)$ is either positive or negative, and within these regions, local extrema are characterized by when the gradient

of the smaller of $h(x,T)$ and $m(x,T)$, denoted $s(x,T)$ or of the larger minus the smaller, denoted $l(x,T) - s(x,T)$ is zero, or when $s(x,T) = l(x,T) - s(x,T)$. It seems reasonable to expect that the best separation could be obtained by including the possibility of adding whichever type of homozygous SNP or wild-type curve is already closer to the heterozygous curve.

Experimental tests of the quantitative model.

To test the model of the previous section, we performed several experiments to genotype DNA for the presence of a homozygous or heterozygous SNP, 187C>G, which is found in the hemochromatosis gene (HFE). This mutation we analyzed

5'-TCA-3' 5'-TGA-3'
3'-AGT-5' 3'-ACT-5'

has the nearest-neighbor symmetry described above which results in a melting curve for the homozygous case which is theoretically identical to and experimentally indistinguishable from that of wild-type DNA.

(See Table 1 for the complete sequence of melting analysis and TGCE amplicons with SNP and primers highlighted. H63D_sequence_031010.doc)

We examined mixtures of known quantities of DNA of each genotype and additional wild-type DNA, in 21 ratios of the added DNA to total DNA (initial plus added.) The ratios, from 1/28 to 14/28 by increments of 1/28, and from 15/28 to 27/28 in steps of 2/28, allowed us to include the theoretically optimal value, 1/7, and observe the behavior of the process in some detail over a wide range of interest for pooled samples as well. The ratio j/28 of additional wild-type DNA to total DNA (unknown sample plus additional) corresponds to the ratio j/(28-j) of additional DNA to unknown, so for instance, the optimal value of 4/28 of total DNA corresponds to adding 4/24=1/6 as much wild-type as there is unknown sample DNA.

We mixed wild-type DNA in three replicates for each of the three genotypes (denoted WT, MUT, and HET) before PCR with the appropriate mixture fraction of wild-type DNA. All samples with a common mixture fraction were amplified together in the presence of a high-resolution fluorescent dye, along with two control samples containing heterozygous DNA with no wild-type added. Since the extension step of each PCR cycle generates only homoduplexes, after amplification is complete, we perform a final additional melt and reannealing to produce the heteroduplexes. Samples with the same mixture fraction of additional DNA were also analyzed simultaneously.

Detailed protocols for the amplification and subsequent heating and cooling before melting analysis are in the appendix.

Validation with melting curve data

Next, we performed high-resolution melting analysis on all of the resulting samples to produce actual fluorescence vs. temperature melting curves corresponding to the model functions of the previous section.

This is a closed-tube process which avoids risk of contamination and leaves the sample undisturbed for further types of analysis. It provides a fast, economical, and accurate method of genotyping and mutation scanning which has been described and studied in a variety of contexts (5-10).

Melting curves are first standardized by removal of background fluorescence. Next, they are temperature shifted to adjust for small variations in reported temperature, by superimposing the 'toe' feature common to all curves, where only the most stable homoduplexes are left to melt. Difference plots were used to highlight relative variation between genotypes.

The value and location of the maximum difference and the area between curves of the different genotypes and the average of the wild-type replicates were recorded for analysis and comparison with the theory. According to the theory, location of maximum difference is constant, and magnitude of maximum difference, and area under difference are directly proportional to the heteroduplex concentration of the samples.

Figure 4 shows the average calculated values of the maximum difference between the three replicate melting curves of mixtures with homozygous and heterozygous genotypes and the average of the wild-type melting curves, as a function of the mixture fraction. The values are normalized by a scaling which makes the value of pure heterozygous control samples equal to 0.5. This corresponds to the concentration of heteroduplexes in the theoretical model, which is superimposed on the figures. The squared correlation between the experiment and the model, $R^2 = \frac{(\mathbf{X}\cdot\mathbf{M})^2}{(\mathbf{X}\cdot\mathbf{X})(\mathbf{M}\cdot\mathbf{M})}$ where $\mathbf{X}$ is the vector of experimental values and $\mathbf{M}$ is the vector of model values at the measured mixture fractions. (We do not use the usual formula for R^2 in terms of sums of squares, $1 - \frac{\mathbf{E}\cdot\mathbf{E}}{\mathbf{X}\cdot\mathbf{X}}$ where $\mathbf{X} + \mathbf{E} = \mathbf{M}$, which requires $\mathbf{M}$ to be the best fit of the experimental data satisfying the orthogonality condition $||\mathbf{M}||^2 + ||\mathbf{E}||^2 = ||\mathbf{X}||^2$, or $(\mathbf{M} \cdot \mathbf{E}) = 0$. Even though $\mathbf{M}$ is not the best fit, the squared correlation coefficient values satisfy $R^2 > .99$ for the heterozygous samples, and $R^2 > .98$ for the homozygous samples.

Figure 5a shows the standardized melting curves corresponding to the optimal 4/28 mixture fraction. The replicates cluster indistinguishably, appearing as one curve, and the three genotypes are equally separated. They may easily be classified by the observer's eye or by automatic classification software. This is a vast improvement from the initial figure in which replicates of the homozygous SNP and the wild-type samples overlapped each other completely.

For comparison, Figure 5b and 5c show the standardized melting curves corresponding to mixture fractions 9/28 and 14/28, in which it is again difficult to distinguish the homozygous and heterozygous samples as our model predicts. Figure 5d show the melting curves at mixture fraction 19/28 near where they are again best separated, albeit in a different order. This demonstrates the importance of the correct mixture fraction. Using equal proportions, or just any small proportion such as 1/3 gives no improvement on the original results with no mixing at all!

Figure 6 shows experimental difference curves between wild-type melting curves and the other two genotype melting curves at two values of x, $x = 0$, or no added DNA (Figure 6a) and at the optimal value $x = 4/28$ (Figure 6b). In both case, the shape and magnitude agree nicely with the theoretical melting curves and the predictions of our model.

Validation with quantitative TGCE data

We also performed temperature-gradient capillary electrophoresis (TGCE) on each sample to provide an independent and more direct analysis of heteroduplex content. In this technique, the arrival of duplexes in a sample is detected after they are drawn through a gel. Each species of duplex has a characteristic arrival time distribution depending on its spatial conformation. This corresponds to the frequency of duplex arrivals per frame, the quantity which is actually measured. The center of heteroduplex peaks are significantly delayed and separated from each other in comparison to homoduplex peaks. This is due to the distinct 'bubbles' formed by different mismatched base pairs of the two heteroduplexes. The two species of homoduplex have no bubbles and appear as one peak. The two heteroduplex peaks are easily separated from the homoduplex peak and from each other. These peaks exhibit simple mathematical behavior which makes it possible to separate and quantify the relative contributions of the heteroduplexes. We used this to validate our theoretical model of melting curve separation, which was based upon relative concentration of heteroduplexes in the samples. Figure 7a shows typical raw TGCE data for three replicates of mixtures with each genotype with mixture fraction 9/28, over a range of frames beginning with frame 1000 and containing all of the peaks. Figure 7b shows the same data normalized by shifting all peaks to the same frame number and scaling to the same height. This figure demonstrates the common heteroduplex content of the homozygous mutant and heterozygous genotypes at this mixture fraction, which is responsible for their overlapping melting curves seen in figure 5b, reemphasizing the sensitivity of the results to choosing the added DNA proportion suboptimally.

The TGCE data was analyzed quantitatively as follows. Individual TGCE peaks were approximated by exponential distributions of the form $F(t) = Ae^{-kt}, t \geq t_0; F(t) = 0, t < t_0$. Higher resolution data might be amenable to closer fit by higher order gamma distributions of which the exponential distribution is a special case, but since the peaks are only resolved by on the order of 10 data points, the simplest version must suffice. Some additional evidence that this is reasonable is provided by the fact that the fit parameters of each peak remained nearly invariant when the window of points used for the fit was varied in size and distance from the peak. The observed arrival frequency before each peak did not have the strict cutoff behavior of the exponential distribution, as some increase above background was seen one frame before the maximum of the first arrival peak. However, no increase above background could be seen two frames before the first arrival peak. Based upon this model, we could solve for the combined amplitudes and decay rate of homoduplex concentrations contributing to the first arrival peak, and by successive subtraction, iteratively solve for the amplitudes of subsequent peaks. Because of the large dynamic range of the peaks and their narrow extent in terms of data acquisition frames, the quantitative results might be expected to be sensitive to the fitting process. For example, we

approximated the start of each exponential sub-distribution with the frame of the maximum measured value, even though actual peak is located somewhere between this frame and an adjacent one. In spite of this sensitivity, we found that the decay rates of different peaks were nearly independent of duplex species, peak amplitude, fitting window and method, which provided additional confidence in the model. We are investigating more sophisticated gamma fits of the data, and corresponding deconvolution techniques which could reduce these sources of error.

Once the constituent peak amplitudes were quantified, the heteroduplex proportion was determined by first computing the ratio of the sum of the derived amplitudes of the two heteroduplex peaks to the sum of these plus the amplitude of the combined homoduplex peak. This ratio was then adjusted by a factor close to 1 which made the same ratio determined from heteroduplex control samples equal to the expected value of 0.5.

Figure 8 shows the average of the calculated values of the heteroduplex proportion of of three replicate of mixtures with homozygous and heterozygous samples as a function of the mixture fraction. Once again, the squared correlation coefficients between the experimental data and the model are high, with $R^2 > .97$ for both heterozygous and homozygous samples. This agreement between the results of a fairly simple analysis and those of the melting curve experiments and the theory suggest that quantitative TGCE (qTGCE) estimation of heteroduplex content of mixed or pooled samples is indeed feasible and informative.

Discussion of the results

The experimental results confirm the main points of the theory. The maximum difference between melting curves and the heteroduplex concentrations inferred from TGCE experiments agree with each other and with the theoretical predictions of heteroduplex concentration with considerable accuracy over a wide range of mixture fractions. The area between melting curves and the location of the maximum difference between curves also behave as predicted. The plots of these quantities follow the quadratic behavior of the model qualitatively over the entire range, and are quantitatively close over a range of mixture fractions up to one-half (14/28) of the total.

Where the data deviates from the model above this mixture fraction, there is a definite trend for heteroduplex concentrations estimated from TGCE and corresponding melting curve differences to be larger than those predicted by the theory for a given mixture fraction. Because the heteroduplex concentration vs. mixture fraction curves for both the heterozygous and heterozygous unknowns are decreasing for mixture fractions greater than 14/28, the inferred experimental values correspond to mixture fractions lower than those we prepared experimentally. So one possible source of such a trend could be that the actual fraction of additional wild-type DNA fell short of the intended value as that value grew beyond one-half. Selective amplification (unequal efficiencies) in PCR or amplification of initial variations that diminish final concentration of wild-type DNA at higher

concentrations could have such an effect. If complementary and nearly-complementary strands anneal preferentially rather than independent of their differences, the assumptions of the model would be violated, although it would be surprising if such a bias favored more rather than less heteroduplex formation.

Regardless of these subtle deviations from close agreement with a fairly simple model, the ultimate test of our method is given by the ease with which the simple melting curve approach to can be used to genotype the optimally mixed samples, in contrast to unmixed or non-optimally mixed ones.

Appendix: Experimental protocols

DNA was extracted using a QIAamp DNA Blood Kit (QIAGEN, Inc., Valencia, CA), concentrated by ethanol precipitation and quantified by A260.

For high-resolution melting analysis, we used small amplicon melting with primers as close to the SNP as dimer and misprime constraints permit, as described in Liew et al. (2004) (12). The PCR protocol followed here was modified slightly from the protocol described in (12). The amplicon was 40bp long. PCR was performed in a LightCycler with reagents commonly used in clinical laboratories. Ten microliter reaction mixtures consisted of 25ng of genomic DNA, 3 mM MgCl2, 1x LightCycler FastStart DNA Master Hybridization Probes master mix, 1x LCGreen Plus, 0.5 μM forward (CCAGCTGTTCGTGTTC-TATGAT) and reverse (CACACGGCGACTCTCAT) primers and 0.01U/μl Escherichia coli (E. coli) uracil N-glycosylase (UNG, Roche). The PCR was initiated with a 10 min hold at 50°C for contamination control by UNG and a 10 min hold at 95°C for activation of the polymerase. Rapid thermal cycling was performed between 85°C and the annealing temperature at a programmed transition rate of 20°C/s for 40 cycles. Samples were then rapidly heated to 94°C and cooled to 40°C followed by melting curve analysis between 60°C and 85°C to confirm the presence of amplicon.

Prior to analysis on the HR1, samples were again rapidly heated to 94°C and cooled to 40°C to promote heteroduplex formation.

Following amplification, an additional melting was performed to denature the perfectly complementary post-extension duplexes after which the temperature was rapidly decreased to re-anneal strands independent of the presence or absence of a single mismatched base-pair.

PCR protocol for HFE amplification for Temperature Gradient Capillary Electrophoresis (TGCE) analysis

The PCR protocol followed here was modified slightly from the protocol described in Bernard et al. (1998) (13). The amplicon was 242bp long. PCR was performed in a Perkin Elmer 9700 block cycler with similar reagents to those used for amplification in the Light-Cycler. Ten microliter reaction mixtures consisted of 25ng of genomic DNA, 3 mM MgCl2, 1x LightCycler FastStart DNA Master Hybridization Probes master mix, 0.4 μM forward (CACATGGTTAAGGCCTGTTG) and reverse (GATCCCACCCTTTCAGACTC) primers and 0.01U/μl Escherichia coli (E. coli) uracil N-glycosylase (UNG, Roche). All samples were then overlayed with mineral oil to prevent evaporation. The PCR was initiated with a 10 min hold at 25°C for contamination control by UNG and a 6 min hold at 95°C for activation of the polymerase. Thermal cycling consisted of a 30s hold at 94°C, a 30s hold at 62°C and a 1min hold at 72°C for 40 cycles followed by a 7min hold at 72°C for final elongation.

Upon completion of these thermal cycles the samples were then heated to 95°C for 5min followed by a slow cool over approximately 60min to 25°C.

TGCE analysis

The protocol followed here is similar to that described in Margraf et al. (2004) (14). To prepare samples for TGCE analysis, PCR amplicons were transferred to 24 well TGCE trays and diluted 1:1 with 1xFastStart Taq polymerase PCR buffer (Roche). These samples were then overlayed with mineral oil and the trays loaded into the TGCE instrument. TGCE was performed on a commercial instrument (Reveal

(TM-melting temperature or trademark?)

mutation discovery system, reagents and Revelation software by SpectruMedix LLC, State College, PA) (6). DNA samples were injected electro-kinetically at 2 kV for 45 seconds, resulting in peak heights ranging from 5,000-40,000 intensity units with ethidium bromide staining. Optimal results were obtained when the temperature was ramped from 60-65°C over 21 minutes and data was acquired over 35 minutes. Sequential camera images were converted to plots of image frame number (time) versus intensity units (DNA concentration).

References

1. Highsmith WE Jr, Jin Q, Nataraj AJ, O'Connor JM, Burland VD, Baubonis WR, Curtis FP, Kusukawa N, Garner MM. Use of a DNA toolbox for the characterization of mutation scanning methods. I: construction of the toolbox and evaluation of heteroduplex analysis. Electrophoresis. 1999 Jun;20(6):1186-94.

2. Xiao W, Oefner PJ. Denaturing high-performance liquid chromatography: A review. Hum Mutat. 2001 Jun;17(6):439-74.

3. Li Q, Liu Z, Monroe H, Culiat CT. Integrated platform for detection of DNA sequence variants using capillary array electrophoresis. Electrophoresis. 2002 May;23(10):1499-511.

4. Gundry CN, Vandersteen JG, Reed GH, Pryor RJ, Chen J, Wittwer CT. Amplicon melting analysis with labeled primers: a closed-tube method for differentiating homozygotes and heterozygotes. Clin Chem. 2003 Mar;49(3):396-406.

5. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ. High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem. 2003 Jun;49(6 Pt 1):853-60.

6. Reed GH, Wittwer CT. Sensitivity and specificity of single-nucleotide polymorphism scanning by high-resolution melting analysis. Clin Chem. 2004 Oct;50(10):1748-54.

7. McKinney JT, Longo N, Hahn SH, Matern D, Rinaldo P, Strauss AW, Dobrowolski SF. Rapid, comprehensive screening of the human medium chain acyl-CoA dehydrogenase gene. Mol Genet Metab. 2004 Jun;82(2):112-20.

8. Willmore C, Holden JA, Zhou L, Tripp S, Wittwer CT, Layfield LJ. Detection of c-kit-activating mutations in gastrointestinal stromal tumors by high-resolution amplicon melting analysis. Am J Clin Pathol. 2004 Aug;122(2):206-16.

9. Zhou L, Vandersteen J, Wang L, Fuller T, Taylor M, Palais B, Wittwer CT. High-resolution DNA melting curve analysis to establish HLA genotypic identity. Tissue Antigens. 2004 Aug;64(2):156-64.

10. Liew M, Pryor R, Palais R, Meadows C, Erali M, Lyon E, Wittwer C. Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. Clin Chem. 2004 Jul;50(7):1156-64.

11. Kuklin A, Davis AP, Hecker KH, Gjerde DT, Taylor PD. A novel technique for rapid automated genotyping of DNA polymorphisms in the mouse. Mol Cell Probes. 1999 Jun;13(3):239-42.

12. Liew, M., et al., Genotyping of single-nucleotide polymorphisms by high-resolution melting of small amplicons. Clin Chem, 2004. 50(7): p. 1156-64.

13. Bernard, P.S., et al., Homogeneous multiplex genotyping of hemochromatosis mutations with fluorescent hybridization probes. Am J Pathol, 1998. 153(4): p. 1055-61.

14. Margraf, R.L., et al., Genotyping hepatitis C virus by heteroduplex mobility analysis using temperature gradient capillary electrophoresis. J Clin Microbiol, 2004. 42(10): p. 4545-51.