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

RA Palais, MA Liew, and CT Wittwer

Abstract

Heteroduplex analysis effectively screens for heterozygotes, but usually does not distinguish between different homozygotes. Specific homozygotes can be identified by mixing PCR products from the unknown and a known homozygote and repeating the heteroduplex analysis. Disadvantages of mixing post-PCR include the need for a second analysis and an increased risk of PCR product contamination of subsequent reactions. Alternatively, DNA of known homozygous genotype can be added to each unknown before PCR. Depending upon the proportions mixed, different genotypes might then be distinguished by the different amounts of heteroduplexes produced. Our theoretical derivation suggests that the best separation of the three genotypes of bi-allelic, diploid DNA occurs when the known genotype is one-seventh of the total DNA. 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 can be virtually indistinguishable. Genotyping by high resolution melting analysis appeared more precise than temperature gradient capillary electrophoresis. By mixing unknowns with a known genotype before PCR, only one analysis is needed for full genotyping and, in the case of high-resolution melting analysis, the procedure is entirely closed-tube after the start of PCR, removing concerns of PCR product contamination.

(Would it be right to clarify that heteroduplex analysis means "heteroduplex content analysis" or heteroduplex detection or quantification, because it seems to be heteroduplex analysis in the sense of how much heteroduplex content as opposed to what kind of heteroduplexes?)

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, a homozygous difference will produce more heteroduplexes than a heterozygous difference.

(?? I don't think this is always the case? Only when the hom curve has crossed the het curve)

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 with experiment using both high-resolution melting analysis and TGCE. The experiments emphasize the importance of the using the correct proportion for those who might doubt that it matters: 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 the DNA amplicon from three genotypes of a SNP for which the homozygous mutation is indistinguishable from that of the wild type, due to nearest-neighbor thermodynamic symmetry of the mutation. (This says that the bases immediately surrounding the mutation are identical when the strands are interchanged, e.g., TCA/AGT ¡-¿ TGA/ACT.) Our goal is to add the right proportion of wild-type DNA to each genotype before PCR, so that after amplification, melting, and reannealing, the mixture with the homozygous mutant sample will develop heteroduplex content but the mixture with the wild-type sample will not, making the curves distinguishable. The mixture with the heterozygous sample will have its heteroduplex content reduced from its natural 50 % value, so we must be careful that its curve remains distinct from the heteroduplex-enhanced homozygous sample.

Therefore, in this section, we develop a model for the melting curve of a mixture of genotypes in terms of the melting curves of the consituent duplex types and the mixture proportions. In the case above that the homozygous mutant and wild-type curves are indistinguishable, we show that the difference among curves of different genotypes depends solely on the heteroduplex content of the mixture. Finally, we model and optimize the separation of heteroduplex contents of the three genotypes in terms of mixture proportion.

In what follows, $f_j(T)$ is the standardized fluorescence curve for a fixed concentration of the jth duplex species, where $j = 1$ corresponds to the wild-type homoduplex, $j = 2$ corresponds to one heteroduplex $j = 3$ corresponds to the other heteroduplex, and $j = 4$ corresponds to the SNP homoduplex. $f_j(T)$ refers to the curve with background fluorescence removed, and automatically accounts for whatever fluorescence per duplex variation there may be among the different species types, i.e., unequal contributions to the superposed curves. We will use $g_j(T) = -f_j(T)$ to describe the corresponding standard negative derivative curves.

If the wild-type spike is present in proportion x of the total sample, then if the unknown sample is wild-type, the entire sample is wild-type, and the resulting negative derivative of the melting curve is described by

$$W(T, x) = g_1(T).$$

If the unknown sample is a homozyogous SNP, then strands from the homozygous SNP are present in proportion $1 - x$ of the total sample, so assuming strands anneal independently, after melting and re-annealing, species concentrations correspond to coefficients of

$$(xs_1+(1-x)s_2)\otimes(xs_1+(1-x)s_2) = x^2(s_1\otimes s_1)+x(1-x)(s_1\otimes s_2)+(1-x)x(s_2\otimes s_1)+(1-x)^2(s_2\otimes s_2),$$

i.e., wild-type homoduplexes will be present in proportion x^2, homozygous SNP homoduplexes will be present in proportion $(1 - x)^2$, and each type of heteroduplex will be present in proportion $x(1 - x)$.

The negative derivative of the resulting melting curve is described by

$$M(T, x) = x^2 g_1(T) + x(1 - x)g_2(T) + x(1 - x)g_3(T)(1 - x)^2 g_4(T).$$

If the unknown sample is heterozyogous, then wild-type and homozygous SNP strands from the original sample are each present in proportion $\frac{1-x}{2}$. This is the entire contribution of homozygous SNP strands, while the spike contributes an additional proportion x of wild-type strands, for a total proportion of $x + \frac{1-x}{2} = \frac{1+x}{2}$ wild-type strands.

Again assuming strands anneal independently, after melting and re-annealing,

$$(\frac{1+x}{2}s_1 + \frac{1-x}{2}s_2) \otimes (\frac{1+x}{2}s_1 + \frac{1-x}{2}s_2) =$$

$$(\frac{1+x}{2})^2(s_1 \otimes s_1) + (\frac{1+x}{2})(\frac{1-x}{2})(s_1 \otimes s_2) + (\frac{1-x}{2})(\frac{1+x}{2})(s_2 \otimes s_1) + (\frac{1-x}{2})^2(s_2 \otimes s_2),$$

i.e., wild-type homoduplexes will be present in proportion $\frac{(1+x)^2}{4}$, homozygous SNP homoduplexes will be present in proportion $\frac{(1-x)^2}{4}$, and each type of heteroduplex will be present in proportion $\frac{(1+x)(1-x)}{4} = \frac{1-x^2}{4}$.

The resulting negative derivative of the melting curve is described by

$$H(T,x) = \frac{(1+x)^2}{4}g_1(T) + \frac{1-x^2}{4}g_2(T) + \frac{1-x^2}{4}g_3(T) + \frac{(1-x)^2}{4}g_4(T).$$

We may write

$$W(T,x) = 1g_1(T)$$

$$M(T,x) = m_1(x)g_1(T) + m_{23}(x)(g_2(T) + g_3(T)) + m_4(x)g_4(T)$$

$$H(T,x) = h_1(x)g_1(T) + h_{23}(x)(g_2(T) + g_3(T)) + h_4(x)g_4(T)$$

where $m_1(x) = x^2$, $m_{23}(x) = x(1-x)$, $m_4(x) = (1-x)^2$, and $h_1(x) = \frac{(1+x)^2}{4}$, $h_{23}(x) = \frac{1-x^2}{4}$, $h_4(x) = \frac{(1-x)^2}{4}$.

Note that

$$m_1(x) + 2m_{23}(x) + m_4(x) = 1 = h_1(x) + 2h_{23}(x) + h_4(x).$$

Our goal is to maximize our ability to distinguish these three curves, as measured by the minimum separation between any two of them. The separation will be defined by the maximum absolute value of their difference. So our goal is to find

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

For this reason we compute

$$W(T,x) - M(T,x) = (1 - m_1(x))g_1(T) - m_{23}(x)(g_2(T) + g_3(T))) - m_4(x)g_4(T).$$

$$W(T,x) - H(T,x) = (1 - h_1(x))g_1(T) - h_{23}(x)(g_2(T) + g_3(T))) - h_4(x)g_4(T).$$

$$H(T,x)-M(T,x) = (h_1(x)-m_1(x))g_1(T)+(h_{23}(x)-m_{23}(x)(g_2(T)+g_3(T)))+(h_4(x)-m_4(x))g_4(T).$$

In the situation where the nearest-neighbor model of the homozygous SNP has the same thermodynamic parameters as the wild-type and the corresponding melting curves and their (negative) derivatives are identical, (experimental curves indistinguishable) i.e., $g_1(T) = g_4(T)$, requiring us to use the spiking protocol, we may simplify these differences considerably. (Even when $g_1(T) = g_4(T)$ according to nearest neighbor theory, the thermodynamic parameters that determine g_2 and $g_3(x)$ are not identical. The differences are reported in Table x, but do not affect the subsequent analysis.)

Combining the coefficients of the g_1 and g_4 in the first two expressions, using

$$1 - (m_1(x) + m_4(x)) = 2m_{23}(x)$$

$$1 - (h_1(x) + h_4(x)) = 2h_{23}(x)$$

then distributing the result back equally between g_1 and g_4 for symmetry, we obtain

$$W_=(T,x) - M_=(T,x) = m_{23}(x)(g_1(T) + g_2(T) + g_3(T) + g_4(T))$$

$$W_=(T,x) - H_=(T,x) = h_{23}(x)(g_1(T) + g_2(T) + g_3(T) + g_4(T))$$

and writing the third difference as the difference of these differences,

$$H_=(T,x) - M_=(T,x) = (m_{23}(x) - h_{23}(x))(g_1(T) + g_2(T) + g_3(T) + g_4(T))$$

The graphs of these three functions are given in Figure 2, annotated with key features derived below.

These expressions have two important consequences. First, they show that the pointwise separation of the curves is proportional to the difference in heteroduplex fraction of the mixtures. Second, they uncouple the x (spike proportion) and T (temperature) dependence of the differences among fluorescence curves of different genotypes. This allows us to optimize the spike proportion independently, regardless of the specific nature of individual duplex curves contributing to the superpositions. Our problem reduces to

$$\begin{aligned}
&\max_{x\in[0,1]} \min\{\max_T |W_=(T,x) - M_=(T,x)|, \max_T |W_=(T,x) - H_=(T,x)|, \max_T |M_=(T,x) - H_=(T,x)|\} \\
&= \max_{x\in[0,1]} \min\{m_{23}(x), h_{23}(x), |m_{23}(x) - h_{23}(x)|\} \max_T |g_1(T) + g_2(T) + g_3(T) + g_4(T)| \\
&= G \max_{x\in[0,1]} \min\{m(x), h(x), |m(x) - h(x)|\},
\end{aligned}$$

where

$$m(x) = 2m_{23}(x) = 2x(1-x)$$

and

$$h(x) = 2h_{23}(x) = \frac{1-x^2}{2},$$

make $m(x)$ and $h(x)$ the (non-negative) total heteroduplex proportion in the spiked homozygous SNP and heterozygous samples, respectively, and

$$G = \frac{1}{2}|g_1(T) + g_2(T) + g_3(T) + g_4(T)|,$$

For example, when $x = 0$, $h(x) = \frac{1}{2}$, and $m(x) = 0$, $\frac{1}{2}G$ is the separation of the unspiked heterozygous curve from the common unspiked wild-type and homozygous SNP curves. What remains is to solve the spike dependent optimization problem

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

Graphically, we want to find the largest value of either the height of smaller of the first two functions or the vertical distance between them, whichever is smaller.

Now, we develop a condition characterizing the desired extreme values over all possible spiking proportions x, of the minimum among two non-negative functions of x and the magnitude of their difference.

First, we solve $h(x) - m(x) = \frac{3}{2}x^2 - 2x + \frac{1}{2} = \frac{1}{2}(3x-1)(x-1)$, we divide the interval $[0, 1]$ into two subintervals on which $h(x) \geq m(x)$ for $[0, x_=]$ and $m(x) \geq h(x)$ for $[x_=, 1)$, where $x_= = \frac{1}{3}$. The minimum of the three functions is zero at all endpoints of these intervals, $0, x_*$, and 1 where at least one of the three values is zero. We may replace $|m(x) - h(x)|$ by $h(x) - m(x)$ on $[0, \frac{1}{3}]$ and by $m(x) - h(x)$ on $[\frac{1}{3}, 1]$ and also eliminate the larger of m and h from the competition for minimum on these intervals. This further reduces our problem to

$$\max\{\max_{x\in[0,\frac{1}{3}]} \min\{m(x), h(x) - m(x)\} \max_{x\in[\frac{1}{3},1]} \min\{h(x), m(x) - h(x)\}.$$

A local extremum of a minimum among two differentiable functions can only occur when one is strictly and its derivative is zero) *or* where the two functions are equal.

If they are not equal at some point, then they are not equal in a neighborhood of that point, and the minimum of the two will be just the value of the smaller function on that neighborhood. If the derivative of the smaller function is not zero, that function cannot have a local extremum, for the usual calculus reason that it must increase in one direction and decrease in another. That argument by itself fails when the two functions are equal because the minimum may be transferred from one function to the other at such points.

To finish, we identify the points on $[0, \frac{1}{3}]$ where $m'(x) = 0$, $h'(x) - m'(x) = 0$ and $m(x) = h(x) - m(x)$. and on $[\frac{1}{3}, 1]$ where $h'(x) = 0$, $m'(x) - h'(x) = 0$ and $h(x) = m(x) - h(x)$. and compare the resulting values to obtain the global maximum, the optimal spiking fraction, x.

Using $h'(x) = -x = 0$ only when $x = 0$, $m'(x) = 2 - 4x = 0$ only when $x = \frac{1}{2}$, outside of the interval in which it is in competition, these type of points are eliminated. Using $h'(x) - m'(x) = 3x - 2 = 0$ when $x = \frac{2}{3}$ locates the local maximum of $\min\{h(x), m(x) - h(x)\}$ in $[\frac{1}{3}, 1]$ (where $m'(x) - h'(x) = -(h'(x) - m'(x)) = 0$.) This corresponds to adding twice as much spike to a given amount of unknown sample and gives a separation of $\frac{1}{6}G$, between the heterogygous and homozygous SNP curves, or $\frac{1}{3}$ of the original separation between the heterozygous curve and the other two, while the separation between each of these curves and the wild-type are both even greater.

The only point on $[\frac{1}{3}, 1]$ where $m(x) - h(x) = h(x)$ or $(1 - x)^2 = 0$ is at $x = 1$. However, saving the best for last, on $[0, \frac{1}{3}]$, $h(x) - m(x) = m(x)$ or $\frac{7}{2}x^2 - 4x + \frac{1}{2} = \frac{1}{2}(7x - 1)(x - 1) = 0$ gives us our global maximum value at $x = x_* = \frac{1}{7}$, where at the temperature of maximum separation, the homozygous SNP curve is halfway between the other curves, and its separation from each is $\frac{12}{49}G$. This is only barely less than half of the separation of $\frac{1}{2}G = \frac{24}{48}G$ between the unspiked heterozygous curve and the other two unspiked curves.

A simple heuristic explanation for this value, corresponding to one part spike to six parts unknown sample is based upon restating the optimality condition $h(x) - m(x) = m(x)$ as $h(x) = 2m(x)$ and asking what number (6, parts of unknown) when divided in equal parts ($3 + 3$, the heterozygous sample strands) and one part (3, SNP strands) is multiplied by the other plus one (3+1, wild-type sample plus spike strands) is exactly twice the original number (6, the homozygous SNP strands) multiplied by one (wild-type spike 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 spike plus unknown is the optimal spiking proportion. 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 spike and temperature problems. In this case, the optimal spike proportion still may be characterized by the two-dimensional generalization of the above criteria, but it must be found numerically depending upon the specific curves $g_j(T)$. That is, the spike-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 spike of 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 a range of 21 different ratios of wild-type spike to spike plus unknown, from 1/28 to 14/28 by increments of 1/28, and from 15/28 to 27/28 in steps of 2/28. This 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 spike to spike plus unknown corresponds to the ratio j/(28-j) of spike to unknown, so for instance, the optimal value of 4/28 is the same as 4/24=1/6 spike to unknown.

We spiked three replicates of each of the three genotypes (denoted WT, MUT, and HET) before PCR with the appropriate fraction of additional wild-type DNA to achieve the desired spike as a proportion of the total. All samples with a common spike fraction were amplified together in the presence of a high-resolution fluorescent dye, along with two control samples containing unspiked heterozygous DNA. 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 a common spike proportion were also analyzed simultaneously. Detailed protocols for the amplification and subsequent melting and reannealing 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 that 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 3 shows the average calculated values of the maximum difference between the three replicate spiked homozygous and heterozygous SNP melting curves and the average of the spiked wild-type melting curves as a function of the proportion of the spike in the total mixture. The values are normalized by a scaling which makes the value of unspiked 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 differences between the homozygous and heterozygous replicates can be obtained by taking the difference of their individual differences with the mean wild-type curve, as in the theoretical analysis.) The locations of these maxima were nearly independent of temperature, as predicted by the model. 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 spiking values. (Note that we cannot in this situation rewrite R^2 in terms of sums of squares such as $1 - \frac{(\mathbf{E}\cdot\mathbf{E})}{\mathbf{X}\cdot\mathbf{X}}$ where $\mathbf{X} + \mathbf{E} = \mathbf{M}$, This 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 4a shows the standardized melting curves corresponding to the optimal 4/28 spiking value. 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 our 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 4b and 4c show the standardized melting curves corresponding to spiking values 9/28 and 14/28, in which it is again difficult to distinguish the homozygous and heterozygous samples as our model predicts. Figure 4d show the melting curves at spiking value 19/28 near where they are again best separated, albeit in a different order. This demonstrates the importance of correct spiking. Using equal proportions, or just any small proportion such as 1/3 gives no better results than no spiking at all!

Validation with quantitative TGCE data

We also performed temperature-gradient capillary electrophoresis (TGCE) on each sample to provide to make an independent and more direct analysis of heteroduplex content. In this technique, we detect the arrival of duplexes in a sample after they are drawn through a gel. Each species of duplex has a characteristic arrival frequency distribution depending on its spatial conformation. The center of heteroduplex arrival frequency 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 their arrival frequency peaks which superpose indistinguishably. 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 5a shows typical raw TGCE data for three replicates of each genotype spiked with 9/28 spike. over a range of frames beginning with frame 1000 and containing all of the peaks. Figure 5b shows the same data normalized by shifting all peaks to the same frame number and scaled to the same height. This figure demonstrates the common heteroduplex content of the homozygous mutant and heterozygous genotypes at this spike proportion, which is responsible for their overlapping melting curves seen in figure 4b.

to emphasize the importance of optimality for the purposes of classification by heteroduplex content analysis. Once again, at this spiking value near 1/3, it is the heterozygous sample which is indistinguishable from the homozygous sample.

The TGCE data was analyzed quantitatively as follows. Individual TGCE arrival frequency peaks may be 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 6 shows the average of the calculated values of the heteroduplex proportion of three replicate spiked homozygous and heterozygous samples as a function of the proportion of the spike in the total mixture. 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 spiked 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 spiking proportions. 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 spike proportions up to one-half (14/28) of the total.

Where the data deviates from the model above this spiking value, 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 spike proportion. Not only do the overall melting curve and TGCE values follow each other, the individually labeled replicates have a high degree of correlation, both of which indicate that the measured values are indeed higher and not merely artifacts. Because the heteroduplex concentration vs. spike proportion curves for both the heterozygous and heterozygous unknowns are decreasing for spike proportions greater than 14/28, the inferred experimental values correspond to spike proportions lower than those we prepared experimentally. So one possible source of such a trend could be that the actual proportion of wild-type spike 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 spike at higher concentrations could have such an effect. Ways in which the experiments could deviate from the assumptions of the model include non-independent re-annealing of duplexes after the final post-extension melting, although

it would be surprising if this favored formation of more heteroduplexes than would be produced by random association.

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 spiked samples, in contrast to unspiked or non-optimally spiked 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(3?), 0.5 μM forward (CCAGCTGTTCGTGTTCTATGAT) 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 LightCycler. 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 to promote heteroduplex formation. (Why slow?)

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 (RevealTM 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.