A Step-by-Step Guide to Probabilistic Genotype Calculation Logic

Table of Contents

[L/R Averseness 3](#_Toc416264140)

[Soapbox 4](#_Toc416264141)

[Helpful Tricks 4](#_Toc416264142)

[Example Scenario Defined 5](#_Toc416264143)

[Adjusting Allele Frequencies for FST 5](#_Toc416264144)

[2.3. Allele frequencies and shared ancestry 6](#_Toc416264145)

[Converting NIST Allele Counts to Frequencies 6](#_Toc416264146)

[Z Brevity 7](#_Toc416264147)

[RMP Matrix 7](#_Toc416264148)

[Constructing H1 8](#_Toc416264149)

[Genotype Combination RMP 8](#_Toc416264150)

[Assigning H1 Probabilities of Drop-out 9](#_Toc416264151)

[Pr(D) Modeling with Allele Stacking (Formula) 11](#_Toc416264152)

[H1 Pr(~~D~~) Probability of not Drop Out 11](#_Toc416264153)

[H1 Pr(C) 11](#_Toc416264154)

[Formula for Pr(~~C~~) 12](#_Toc416264155)

[Building the denominator H2 13](#_Toc416264156)

[Determining H2 Genotype Combinations 14](#_Toc416264157)

[Calculating the Modified RMP for the Genotype Combinations 15](#_Toc416264158)

[Finishing up 15](#_Toc416264159)

[Stress Test the Logic 18](#_Toc416264160)

[Plotted Results 18](#_Toc416264161)

[Same Data New Scenario 19](#_Toc416264162)

[Scenario Defined 19](#_Toc416264163)

[The One Row Numerator 20](#_Toc416264164)

[Did not “not” drop-out 20](#_Toc416264165)

[The Denominator 0 S | 1 UNK 20](#_Toc416264166)

[Stress Test the Logic 21](#_Toc416264167)

[Introducing an Assumed Sample 22](#_Toc416264168)

[Scenario Defined 22](#_Toc416264169)

[Handling the H1 Assumed 22](#_Toc416264170)

[Explaining the Unattributed Alleles 22](#_Toc416264171)

[Normalizing the Allele Frequencies 23](#_Toc416264172)

[H2 0 S | 1 UNK with an Assumed Sample 23](#_Toc416264173)

[Stress Testing the Logic 24](#_Toc416264174)

[Introducing IBD 24](#_Toc416264175)

[LabR Default IBD Values 24](#_Toc416264176)

[IBD Big Picture 24](#_Toc416264177)

[RMP KIN 25](#_Toc416264178)

[IBD Probabilities 26](#_Toc416264179)

[Parent-Offspring IBD [0|1|0] 26](#_Toc416264180)

[Hall Sib IBD [0.5|0.5|0] 27](#_Toc416264181)

[eDNA LIMS Integration 29](#_Toc416264182)

[Author Contact 29](#_Toc416264183)

Part One

Transitioning away from the current forensic DNA typing interpretation protocols—Random Match Probability (RMP), Combined Probability of Inclusion (CPI) and the Random Man not Excluded (RMNE)—to Likelihood Ratios incorporating Probabilistic Genotype (PG) models that factor probability of Drop-out (Pr(D)) and Drop-in (Pr(C)) constitutes major change, yet, it is inescapable. Let’s get in front of it.

## L/R Averseness

Binary models fail miserably on ambiguous profiles. So, what is there to be afraid of? Perhaps, the thought of not being able to understand or explain PG.

Concerns might include:

* Overly complex calculations
  + Conversely I think we will discover PG will greatly streamline interpretation processes while reducing subjectivity.
* Ability to adequately testify to the results
  + Remove the fuzzy logic of existing subjective Crime Scene Profile (CSP) interpretation, ignored loci, 2p and a host of other problems with stochastic data—testimony might be less stressful.
* Validation of the PG method
  + Currently there exists bastardization on how the modified RMP is used in interpreting CSPs…loci are haphazardly ignored, policy driven thresholds are implemented, 2p naively applied…wow, how did we ever validate that ludicrousness?

There are multitudes of anxieties in any transition—but let’s “eat this elephant, one bite at a time”.

Norah Rudin, Keith Inman and Ken Cheng have helped me and many others through the learning process—thanks folks…hopefully this and other treaties to follow will help “pay forward” and help flatten the learning curve for others that follow.

This document does not initially address methods of selecting a suitable Probability of Drop-out or Drop-in or selecting the best H1 / H2 hypothesis to calculate a particular scenario. But it does “step” you through the Probabilistic calculation logic. I did not select an overly simple scenario or an overly complex one to begin with—hopefully I picked one somewhere in between. The goal is to make the process understandable for the practitioner.

## Soapbox

I do not condone “Black Box” software applications (True Allele)—in my current mindset I believe an analyst should have a fundamental understanding of the calculations and that fundamental understanding should include the ability to successfully manually calculate simplistic scenarios using the PG model.

LabR incorporates a robust semi-continuous PG model that can be understood and appreciated by the typical DNA Analyst. This model has been extended within eDNA to allow for the consideration of differentially degraded samples and the use of RFU values that allow the model to approach a fully continuous model—but not crossing the “Black-Box” line. With that said let’s step into it…

## Validation

The LabR algorithm has been recoded from C++ to C# for the eDNA LIMS integration. Therefore using LabR as the reference algorithm we constructed the spreadsheet detailed below to manually validate LabR. From this point we additionally verify the eDNA PG Integration. LabR=Manual Calculations=eDNA PG Integration.

Considerable amount additional work has been performed analyzing mixtures of knowns and then comparing the results to known unknowns. Various ratios of 1:1, 1:10, and 1:20 for two person mixtures were analyzed then a third known was added creating a three party mixture.

The eDNA Integrated PG tool provides great power of discrimination between known contributors to a mixture when compared to known unknown contributor to a mixture.

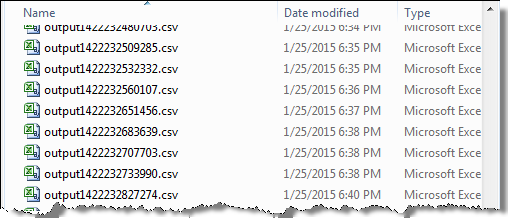
## Helpful Tricks

LabR runs on a stand-alone computer with even the most fanciful scenarios calculating in less time than it takes to enjoy a sip of coffee. Each time you “Run” a calculation using LabR a small, information packed file is written to your hard drive.

You will want to access this file, as it will serve in the validation process. This file can be found by navigating to this location on your hard drive…obviously substituting your User profile and Drive:



The folder will contain a .CSV file for each Input and Output file.



If you open the .csv with Excel the data is nicely presented.

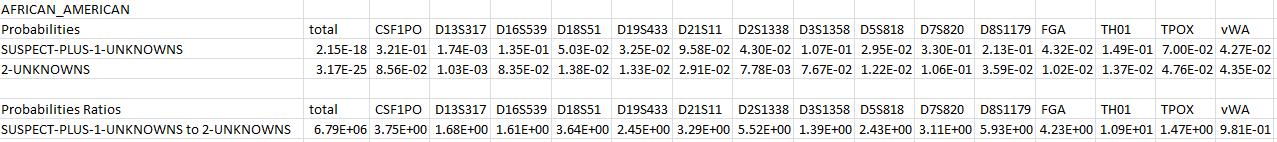


Figure LabR Output File

The reason I bring this to your attention is that now you can see the Numerator (H1) and the Denominator (H2)…and this will help you significantly when verifying your manual calculations.

## Example Scenario Defined

The .csv Output file was created when I calculated the following scenario:

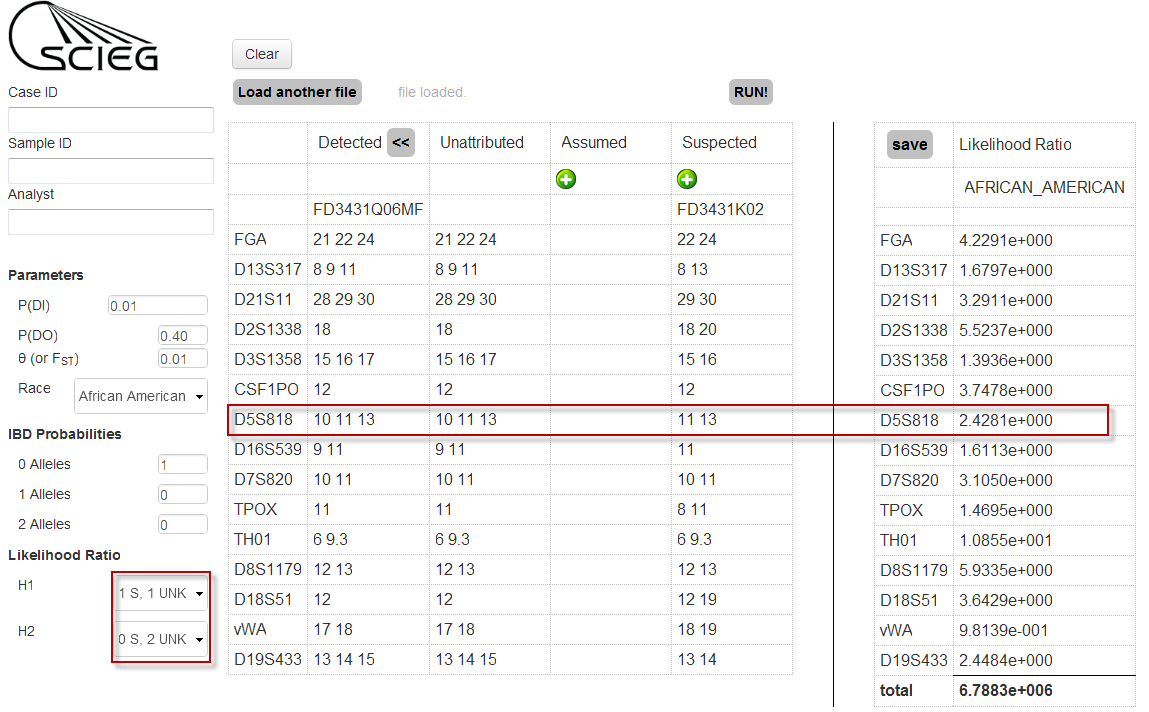


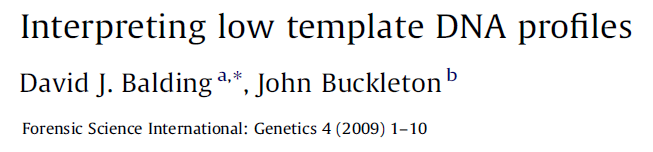
Figure Scenario Calculated

For the remainder of this exercise, I will calculate this same scenario over a range of Pr(D)—incrementally, from zero to one—looking only at the results at locus D5S818. So, by accessing the .csv file, our task has already gotten simpler; we can look-up each H1 & H2 over the range of Pr(D) and use the results to help us build and verify our manual calculations.

## Adjusting Allele Frequencies for FST

In the table showing the scenario data (Figure 2), you can see the CSP is (10,11,13) and the Suspected Sample is (11,13). My allele frequencies will be from the African American population, and I will use an FST of 0.01 and a Pr(C) of 0.01 throughout the calculations that follow.

One of the first things we need to do is adjust the allele frequencies. I am going to refer you to Balding and Buckleton’s 2009 paper:



Citation Balding 2009, FST Corrections

Note: LabR does not adjust for sampling error so I did not include that paragraph however each allele frequency is adjusted as described below in the highlighted section.

### 2.3. Allele frequencies and shared ancestry

**[…]**

We do not here consider alternative sources of the CSP that are direct relatives of s, even though we regard this as an important possibility in practice. To allow approximately for the effects of remote shared ancestry between s and other possible sources of the CSP, we replace the allele proportion P, adjusted for sampling error as described above, with (FST + (1 - FST)P)/(1 + FST) for each allele when s is heterozygote, and with (2FST + (1 - FST)P)/(1 + FST) in the homozygote case. For alleles not in the profile of S, the allele proportion P is replaced with (1 - FST) P/(1 + FST). See [20] for a discussion of the appropriate value of FST. We recommend using a value of at least 0.01 in all cases, and use 0.02 in numerical examples below. In the numerical examples below, genotype proportions are obtained from allele proportions assuming Hardy-Weinberg Equilibrium.

## Converting NIST Allele Counts to Frequencies

You can get to the LabR allele frequency (count) tables on your local computer by drilling down to the folder where it was installed. There is a .csv file for each locus in the folder titled “Allele Frequency Tables”. These frequencies are based on the NIST allele counts which are also online at <http://www.cstl.nist.gov/strbase/NISTpop.htm#Autosomal> .



In the FST Adjusted Allele Frequencies table below all of the alleles for D5 are displayed in row one. In the following rows, we have the allele counts (note that the minimum allele frequency is adjusted in LabR by 5/n but for only the CSP, Suspected (S), and Assumed (V) alleles), the raw frequency, and then the FST adjusted frequencies to be used from this point forward. Of course, we need to base the manual calculations on the same allele frequencies used in LabR to ensure concordant results—so, don’t cut any corners here.



***N*** = 684

Figure FST Adjusted Allele Frequencies

In choosing which alleles to incorporate (and adjust for the specified FST correction), I first note the allele(s) in the CSP but “Not in S”, (Blue 10) and then the homozygote/heterozygote allele(s) in the Suspected Sample (Red 11,13).

Per Balding 2009, the following FST adjustments were made.

**FST Non S** (Blue 10) was calculated as (1 - FST) P/(1 + FST)

**FST Het S** (Red 11,13).were calculated as (FST + (1 - FST)P)/(1 + FST)

Note: eDNA does not need to covert allele counts to allele frequencies. The third row in the Fig 3 table is populated directly from the eDNA Race Frequency database.

## Z Brevity

Here, I’m going to save you a bunch of work. Go ahead and roll-up all alleles not in the CSP as an allele designated ***Z***. Simply subtracting the FST adjusted frequency of 10,11,13 from 1.

***Z*** = 1 – (.0717 + .2392 + .2292) = .4600

Let’s call this “Z Brevity” so when we start looking at genotype combinations the list is much abbreviated. The number of genotype combinations for D5 just went from 66 to 10…please note if you are a “glutton for punishment” you can certainly work with the 66 combinations but the results will be indistinguishable out to the hundreds position. I will point out exponential savings later when we start combining genotypes of individuals to consider in H2.

## RMP Matrix

Next calculate the genotype matrix—this is simply RMP (2pq for the heterozygous system and p2 for the homozygous)—remember FST has already been factored into the allele frequencies in the initial step.



Table RMP Matrix

To recap the CSP is 10,11,13, the S is 11,13…and H1 is [1S,1UNK].

## Constructing H1

This tells us an 11,13 person and a second person of an unknown genotype must be included in H1 (numerator). Looking at the RMP Matrix above you can see there are ten genotype combinations. (10,10) (10,11) (10,13) (10,Z) (11,11) (11,13) (11,Z) (13,13) (13,Z) (Z,Z). The ten rows in the numerator will contain the S and each of these combinations.

Note: The general formula for calculating the number of combinations that must be considered is where x is the number alleles under consideration (CSP). Again, without Z Brevity the numerator would contain 66 combinations (rows).



Figure Suspected and Unknown Genotypes H1

### Genotype Combination RMP

Keep in mind that **S** is part of the H1 hypothesis—in other words the **S** is presumed and therefore calculated as 1. The Unknown contributor however will have its RMP calculated and included. This adds another column to the table. Welcome to the world of Probabilistic Genotypes.



Figure H1 with UNK RMP

### Assigning H1 Probabilities of Drop-out

The next step is to assign a probability of Drop-out Pr(D) to each row of the numerator. Since we used Z Brevity, these figures will jump out at you! The first “logic” question that must be asked is: How can we explain the CSP given the combination of contributor alleles? The answer: Row by row.

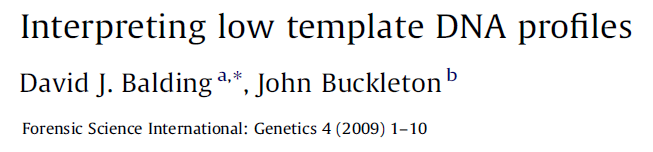
The first three rows of Figure 6 we see a 10,11,13 thus no Pr(D) necessary—but in row four we see a Z allele…and since the CSP is 10,11,13 the Z must have dropped out. In other words, if it hadn’t dropped out the CSP would have been 10,11,13, Z… So we will record a Pr(D) event for that row and for each row that there exists a Z allele…the last row has two Z alleles so that row would record a [Pr(D2) \* **α**] event. (Crap—where did that Greek thing come from?) I’ll explain next.

So if we use a Pr(D) of 0.4 the table will grow to:



Figure H1 with Pr(D)

Before we move on to the Probability of **not** Dropping out Pr(~~D~~) we need to explore what Balding has described as alpha α. (I’ve got to explain the 0.08 in the last row above.)



Citation Balding 2009 Introduces α

Balding writes:

We write D2 for the homozygote drop-out probability. It is natural to seek to express D2 in terms of D. One possibility is to assume that the allele on each homologous chromosome drops out independently with known probability D, so that D2 = D2 [2,8]. This is likely to overstate D2 if both alleles can generate partial signals that combine to reach the reporting standard, whereas each individual signal would fail to reach this standard. For example, suppose that under the prevailing conditions a heterozygous allele would generate a peak with mean height 40 RFU and that D = 0.7. If the signal from a homozygote is a superposition of two signals, with expected height close to 80 RFU, the corresponding drop-out probability is likely to be much less than D2 = 0.49. In 40 laboratory trials of 10 pg DNA samples, 9 instances of homozygote drop-out were observed, from which D2 ~ 0.225. The value of D estimated from 160 heterozygote alleles at the same three loci was 0.66, so D2 ~ 0.44 and hence D2 ~ 1/2D2 (Simon Cowen, personal communication). On the other hand, at very low template levels the epg peak for each allele may be closer to ‘‘all or nothing’’, in which case the independence assumption may be reasonable. Here, we assume that D = αD2. The appropriate value of a should be chosen on the basis of laboratory trials. Typically we expect that 0 << α ≤ 1…

The point here is that the homozygous system is less likely to drop out than the Pr(D2) formula suggests due to stacking—we know from experience that homozygous systems will typically have twice the RFU value of its near heterozygous buddies. So through experimentation Balding suggests a correction value of 0.5—that correction value has been designated as **α** (alpha) . LabR uses an alpha of 0.5.

Let us extend this concept. Given a Pr(D) of 0.4, a single copy of an allele (in a row of contributor combinations) that is not observed in the CSP must have dropped out and been assigned 0.4. If two copies of an allele in the row of contributors are not observed in the CSP, then the Pr(D) would be calculated as [(0.4)2 \* 0.5].

Now, you ponder what to do if three, four, or more copies of an allele in a row of contributors are not observed in the CSP. Well, there is a formula for that as well. It’s not in Balding’s paper, but in a document supplied by Ken Cheng, your LabR programmer:

## Pr(D) Modeling with Allele Stacking (Formula)

“The probability of copies of one allele dropping is modeled by for some value α (which is determined experimentally to be 0.5)”.

Using this model then if we had three copies of an allele in the row of contributors that was not observed in the CSP we would calculate —this will be used in the denominator on a few occasions and even once.

So the chance the allele did not dropout Pr(~~D~~) is (1- PrD) or more fancifully extended

to cover the range of possibilities.

### H1 Pr(~~D~~) Probability of not Drop Out

Let’s add three more columns to the table [Pr(~~D~~)AL10], [Pr(~~D~~)AL11], [Pr(~~D~~)AL13]. Now for each row we will look at the row’s contributors’ alleles and determine if they are in the CSP…if they are in the CSP they did **not** dropout Pr(~~D~~).

In the first row of H1 there is (10,10) & (11,13) genotype. So the Pr(~~D~~) of the 10 allele would be calculated as , the Pr(~~D~~) of 11 is simply 1- 0.4 = 0.6 and the same for Pr(~~D~~) of 13. The table fills out like this:



Figure H1 Pr(~~D~~)

### H1 Pr(C)

Guess what—the hard part of H1 is done…we just need to add columns for Pr(C) and Pr(~~C~~), multiply each row across then sum the results of each row. I’m not going to delve deeply into the Pr(Contamination) other than to say it is an allele in the CSP that cannot be explained by a row of contributors’ alleles. Balding says this:

Citation Balding 2009 Explanation of Drop-In

2.2. Drop-in

Here we assume that at most one drop-in occurs per locus, with probability C, but this restriction is easily relaxed. We also exclude the possibility of a drop-out followed by drop-in of the same allele. Following [8], we treat drop-ins at different loci as mutually independent, and independent of any drop-outs. Multiple drop-ins in a profile may, depending on the values of C and D, be better interpreted as an additional unknown contributor. Typically we use C ≤ 0.05 below, reflecting that drop-in is rare [8], and such low values of C automatically penalise a prosecution hypothesis requiring multiple drop-in events.”

In my manual calculations I used a Pr(C) of 0.01. The numerator is simple; we only have the bottom six rows proposing scenarios that lack a 10 allele. Stated another way, the 10 allele in the CSP cannot be explained with the proposed genotype combinations, and can only be explained with drop-in. Drop-in is calculated by multiplying the Pr(C) by the FST adjusted allele frequency. In this case it is 0.01 \* 0.0717 = 0.0007. We will use the Pr(C) 11 & Pr(C) 13 columns in the denominator; they will remain empty in the numerator.



Figure H1 with Pr(C)

## Formula for Pr(~~C~~)

We need to add one more column before multiplying everything across—the Pr(~~C~~). Pr(~~C~~) is the probability of an allele not dropping in. I’ll again refer to Ken Cheng’s document. Pr(~~C~~) is roughly 1- Pr(C). However more fancifully extended:

Where is the number of alleles included in the locus—since we used Z Brevity that number is 4. Our Pr(~~C~~) is then = 0.989899



Figure H1 with Pr(~~C~~)

Next multiply each row across and sum the results. There you have it; the Numerator calculates as 2.951E-02…Woo-hoo, so does the H1 in LabR!



Figure H1 completed

Time to take a break…when you come back we’ll build the denominator.

# Building the denominator H2

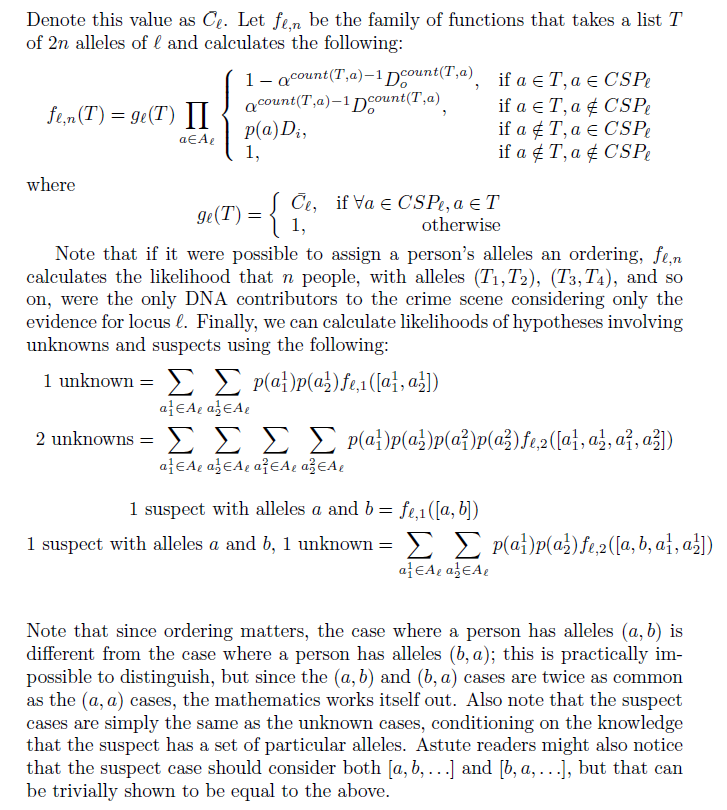
In the denominator we will use the exact logic for Pr(D), Pr(~~D~~), Pr(C), and Pr(~~C~~) as detailed above. What changes is the hypothesis wherein S (11,13) is no longer an assumed contributor. Now we must investigate the combined genotype combinations of two unknowns.

To recap the CSP is 10,11,13, and the H2 is [0 S| 2 UNK].

What we will do next is assemble the rows of all genotype combinations possible when considering two unknowns—this is where you will come to greatly appreciate Z Brevity.

Had we not used Z Brevity in the Numerator but considered all genotype combinations using all alleles in the locus we would be building 1596 genotype combinations (rows) in denominator. However with Z Brevity we will get the same result using 55 rows.

What follows is a screenshot of the document provided by Ken Cheng that formally details the process for which we are about to embark—I’ve included it for the mathematically gifted so you would not feel “left out”. For the rest of us I’ll just skip to the practical solution.



Citation Ken Cheng Constructing the Genotypes matrix

## Determining H2 Genotype Combinations

We have already calculated that we need 55 rows of 2 person genotype combinations…so let’s set up a simple matrix using the following genotypes (10,10) (10,11) (10,13) (10,Z) (11,11) (11,13) (11,Z) (13,13) (13,Z) (Z,Z)—remember we already built a matrix using alleles 10,11,13, and Z to determine these combinations.



Table Two Unknown person matrix

### Calculating the Modified RMP for the Genotype Combinations

Each intersection above provides the combined RMP for the combination of the two unknowns. For example, the (10,10) (10,10) combination is calculate where the frequency of the 10 allele. The (10,10) (10,11) combination is calculated where the frequency of the 10 allele and the frequency of the 11 allele.

What about the “blue cells”? We need to multiply each combination by a factor of two. So you will notice each blue cell is calculated as . This will account for the 10,13 but also the 13,10 combination….and of course the 11,Z vs Z,11 and all others that would have filled the blank cells below the brown diagonal.

When you are done they should sum across, then down, and equal 1.



# Finishing up

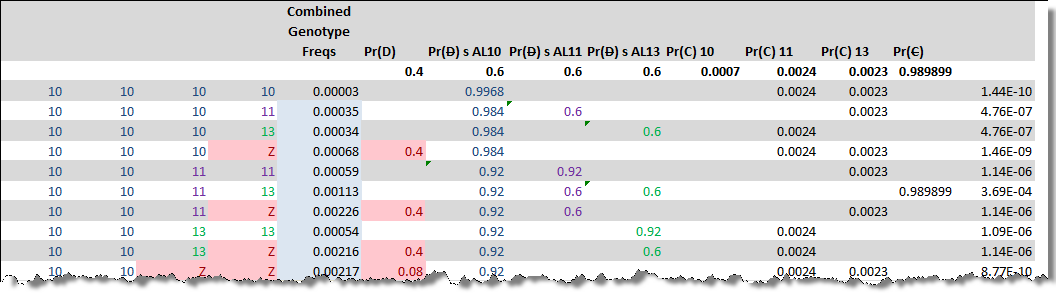
Now, you have all the logic and skills to complete the remainder of this algorithm, so I’ll step the pace up a bit. Let’s begin with the 55 rows of genotype combinations that represent the two unknowns.

The matrix at first glance looks similar to a plate of Christmas cookies but the multi-colored allele designations helped me match up the columns that follow. If you look at the previous matrix (Table 2) you’ll see where the rows and combined genotype frequencies originated. Don’t let the rounding differences throw you off…

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  | **Combined Freqs** |
|  |  |  |  |  |
| 10 | 10 | 10 | 10 | 0.00003 |
| 10 | 10 | 10 | 11 | 0.00035 |
| 10 | 10 | 10 | 13 | 0.00034 |
| 10 | 10 | 10 | Z | 0.00068 |
| 10 | 10 | 11 | 11 | 0.00059 |
| 10 | 10 | 11 | 13 | 0.00113 |
| 10 | 10 | 11 | Z | 0.00226 |
| 10 | 10 | 13 | 13 | 0.00054 |
| 10 | 10 | 13 | Z | 0.00216 |
| 10 | 10 | Z | Z | 0.00217 |
| 10 | 11 | 10 | 11 | 0.00117 |
| 10 | 11 | 10 | 13 | 0.00225 |
| 10 | 11 | 10 | Z | 0.00452 |
| 10 | 11 | 11 | 11 | 0.00392 |
| 10 | 11 | 11 | 13 | 0.00751 |
| 10 | 11 | 11 | Z | 0.01509 |
| 10 | 11 | 13 | 13 | 0.00360 |
| 10 | 11 | 13 | Z | 0.01445 |
| 10 | 11 | Z | Z | 0.01451 |
| 10 | 13 | 10 | 13 | 0.00108 |
| 10 | 13 | 10 | Z | 0.00433 |
| 10 | 13 | 11 | 11 | 0.00376 |
| 10 | 13 | 11 | 13 | 0.00720 |
| 10 | 13 | 11 | Z | 0.01445 |
| 10 | 13 | 13 | 13 | 0.00345 |
| 10 | 13 | 13 | Z | 0.01385 |
| 10 | 13 | Z | Z | 0.01390 |
| 10 | Z | 10 | Z | 0.00435 |
| 10 | Z | 11 | 11 | 0.00754 |
| 10 | Z | 11 | 13 | 0.01445 |
| 10 | Z | 11 | Z | 0.02901 |
| 10 | Z | 13 | 13 | 0.00692 |
| 10 | Z | 13 | Z | 0.02780 |
| 10 | Z | Z | Z | 0.02790 |
| 11 | 11 | 11 | 11 | 0.00327 |
| 11 | 11 | 11 | 13 | 0.01254 |
| 11 | 11 | 11 | Z | 0.02518 |
| 11 | 11 | 13 | 13 | 0.00601 |
| 11 | 11 | 13 | Z | 0.02412 |
| 11 | 11 | Z | Z | 0.02421 |
| 11 | 13 | 11 | 13 | 0.01202 |
| 11 | 13 | 11 | Z | 0.04825 |
| 11 | 13 | 13 | 13 | 0.01151 |
| 11 | 13 | 13 | Z | 0.04622 |
| 11 | 13 | Z | Z | 0.04639 |
| 11 | Z | 11 | Z | 0.04842 |
| 11 | Z | 13 | 13 | 0.02311 |
| 11 | Z | 13 | Z | 0.09279 |
| 11 | Z | Z | Z | 0.09313 |
| 13 | 13 | 13 | 13 | 0.00276 |
| 13 | 13 | 13 | Z | 0.02214 |
| 13 | 13 | Z | Z | 0.02222 |
| 13 | Z | 13 | Z | 0.04445 |
| 13 | Z | Z | Z | 0.08922 |
| Z | Z | Z | Z | 0.04478 |

Figure H2 Genotype Combinations

From this point forward I’ll truncate the matrix as we add columns—you already know what the next steps are anyway. Nothing new…nothing tricky—just remember we are comparing each row of genotype combinations to the CSP of 10,11,13 and explaining how that CSP can exist using probability of drop-in and drop-out if necessary to explain it. (The colors are a bit annoying but they help match up the column logic.)



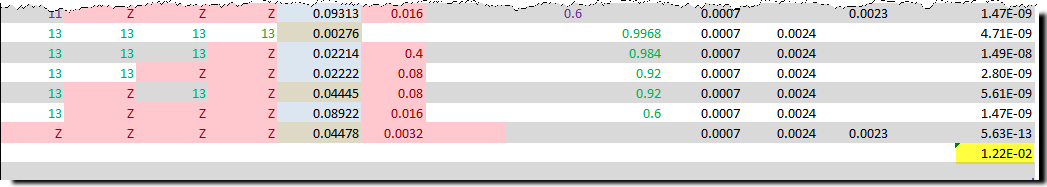


Figure H2 Completed

Multiply the rows across then sum down and you have the number to be used in the H2 position of the likelihood ratio.

Putting it all together you get a LR of 2.43 for the D5S818 locus using the NIST African American population database, a Pr(D) of 0.4, Pr(C) of 0.01 and the hypothesis of (1 S, 1 UNK) | (0 S, 2 UNK).

## Stress Test the Logic

By the time you get to this point you have probably named your spreadsheet—I called mine Abby (probabilistic genotypes was too long). But once it is built, you can “stress test” your logic. By this I mean run it through the ringer—calculated a Pr(D) of zero through 1 incrementally.



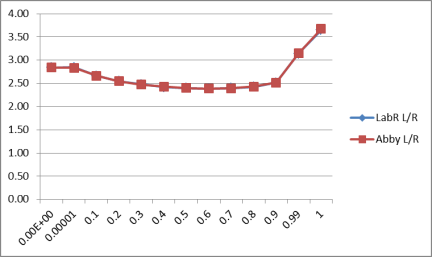
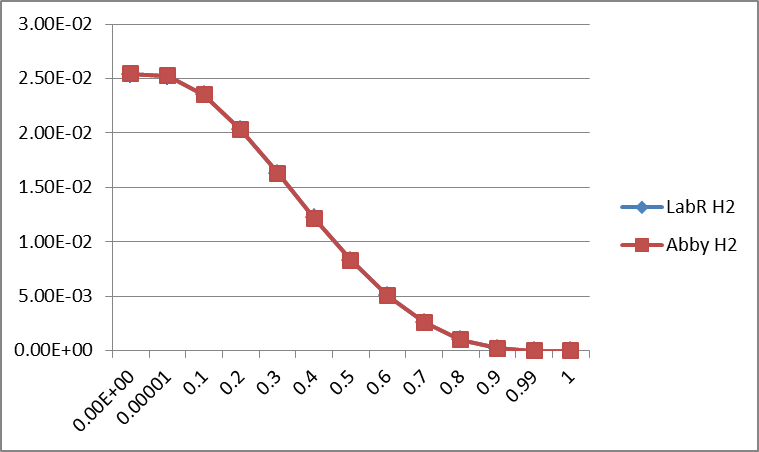
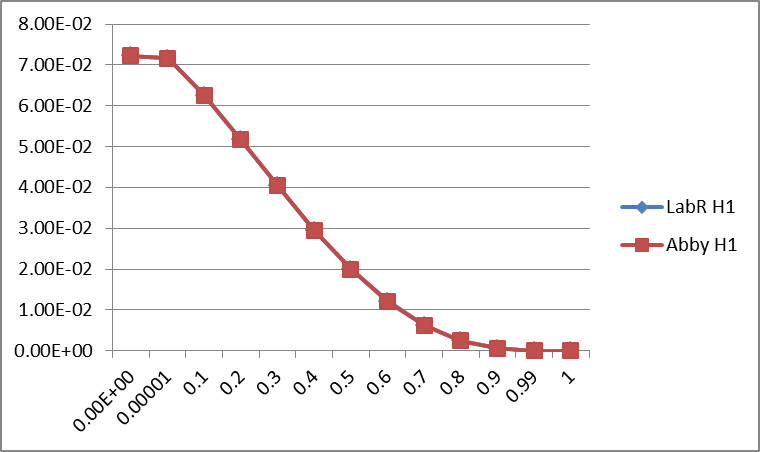
Figure Stress Test Results

The blue rows above are LabR and green are Abby...you see some “hundreds place” divergence at a Pr(D) of 0.99 and at 1. However the point being, multiple times I thought my logic was correct (because everything matched up nicely at a particular Pr(D)) but when stress tested it wildly fell apart.

In this scenario I won’t lose much sleep over the small divergence at 100% drop-out—**not really in the realm of plausibility anyway.**

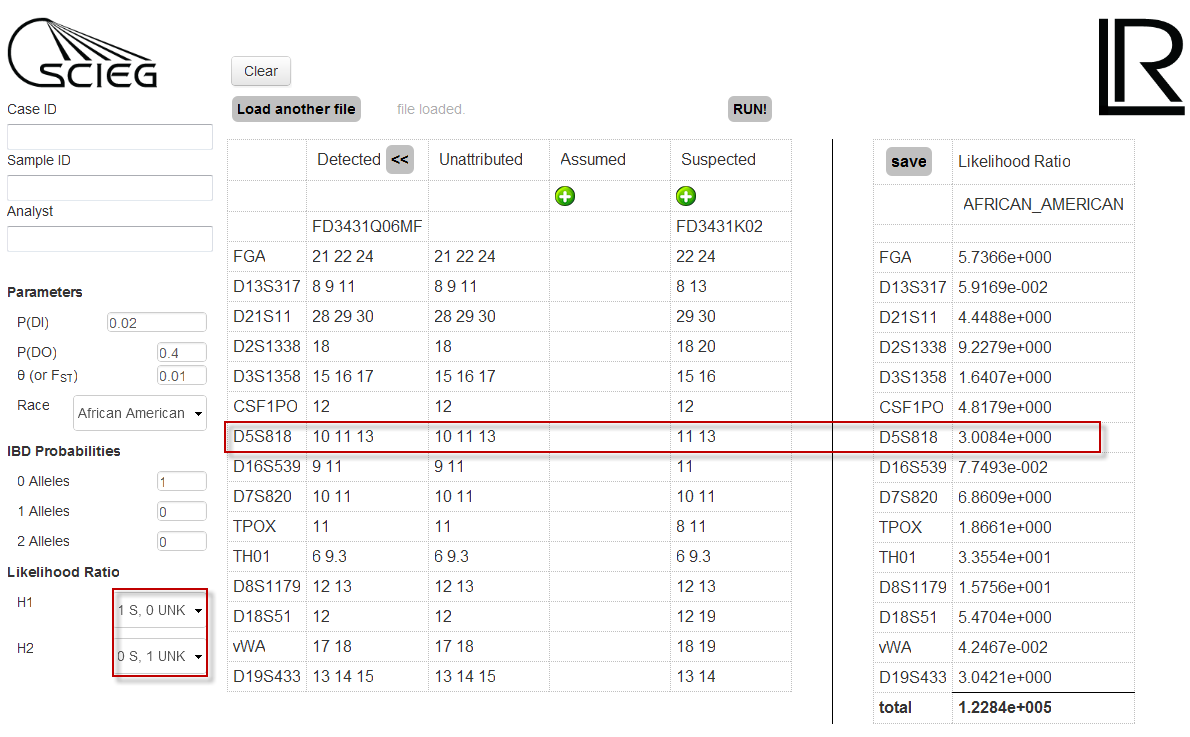
From experience I would suggest working on the Numerator first and getting that logic stressed tested—once you are satisfied then move onto the Denominator. Graphing the numerator and denominator while building each scenario helped me isolate my logic issues and pointed me in the general direction to focus my effort. For this scenario my graphs ultimately came into alignment.

## Plotted Results



Part Two

# Same Data New Scenario



## Scenario Defined

In this section the process is based on the previously defined Logic Rules. Uncertainty is accounted for using a probability or Drop-out and Drop-in. The Hypothesis defined for H1 and H2 are nonsensical given there is evidence of at least two contributors but this is merely an exercise in the calculation logic.

We will fix the Probability of Drop-in at 0.02 and calculate the range of Drop-out incrementally from zero to 100%. The numerator (H1) will be defined as 1 Suspected and Zero Unknown with the denominator (H2) defined as zero Suspected and 1 Unknown.

The allele frequency table will be the same as defined in Figure 3.

I randomly selected a Pr(D) of 0.4 for the following example.

## The One Row Numerator

H1 calculates the probability that the Suspected Sample (11,13) alone explains the Crime Scene Profile (10, 11, 13). The hypothesis dictates that the Suspected Sample is present therefor the “RMP” is set to 1.



Since 11 and 13 are present in the CSP, a Pr(D) need not be evoked to explain any missing alleles in the CSP. The next three columns Pr(~~D~~) AL10, Pr(~~D~~) AL11, Pr(~~D~~) AL13 are used to explain the CSP given the Suspected Sample (11, 13) in the following way:

The probability that an allele did not drop-out equals **1 - pr(D)** or in this example **1 – 0.4 = 0.6**. Therefore, the probability that allele 11 nor 13 did not drop-out is 0.6 for each.

The Pr(~~D~~) AL10 given S (11, 13) is not evoked and the 10 allele is explained as Drop-in (Pr(C)). The probability of Drop-in for an allele is calculated as the **Pr(C) multiplied by the FST adjusted allele frequency** of that allele. For allele 10 the probability of Drop-in is calculated as .**02 x 0.0717 = .0014**. Drop-in is not needed to explain the 11 or 13 allele.

The probability that an allele did not drop-in (Pr(~~C~~)) cannot be evoked because the CSP could not be explained without a drop-in of allele 10 given only a S (11,13).



Figure H1 (1 S | 0 UNK)

### Did not “not” drop-out

Keep in mind it is easy to forget we are explaining each allele of the CSP as a (Pr(D), Pr(~~D~~), Pr(C), or a Pr(~~C~~) event given the hypothesized contributor(s). From that perspective one can see (if they squint real hard) that the probability that an allele did not “not” drop out, it is best explained as drop-in.

## The Denominator 0 S | 1 UNK

H2 calculates the probability of one unknown and zero suspected samples explaining the CSP. This one unknown however can be one of many combinations. Using Z Brevity as described in section 1 we can bring together ten combinations. Each combination will be calculated as a separate row.

The RMP for each genotype is calculated using a matrix.



The same logic defined previously is applied to each row of the denominator. Keep in mind, more than one copy of an allele in a row will affect the Pr(D) and Pr(~~D~~) as described in Citation 2. Each row is multiplied across then summed.



## Stress Test the Logic

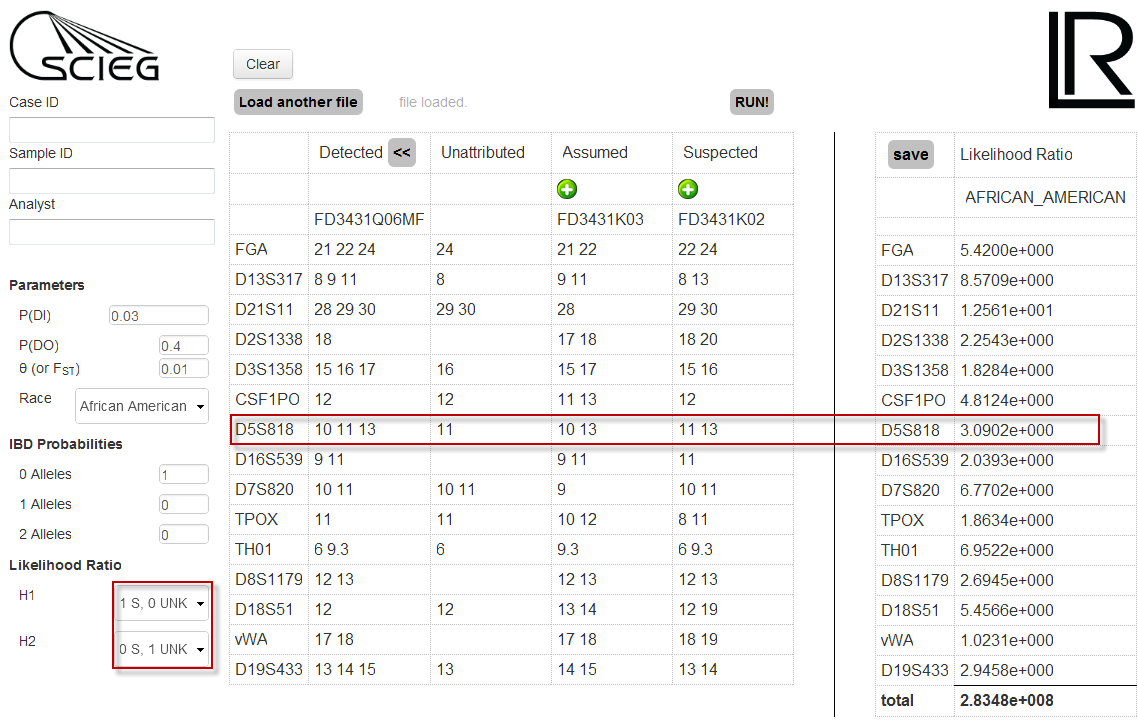
When testing the manually assembled logic for H1 and H2 across the range of drop-out probabilities we note tight concordance with the LabR output. The following table shows the results of the manual logic stress test.



Interestingly the ultimate L/R remains relatively flat across the range from Pr(D) of zero through 90%.

Part Three

# Introducing an Assumed Sample



## Scenario Defined

In this section we will build on the same logic detailed in the first two sections but include an Assumed Sample. An Assumed Sample might be a Victim (most papers label this sample V), or owner of some stolen item, maybe a co-defendant. When an Assumed Sample is assigned, the alleles of the Assumed are suppressed in the CSP—the remaining alleles are then listed as Unattributed.

## Handling the H1 Assumed

The Assumed Profile is now a “Given” in the numerator and denominator. The H1 hypothesis will be calculated as **1 S | 0 UNK.** Since both the Assumed and Suspected Samples are part of the H1 hypothesis the “RMP” is calculated as 1.

## Explaining the Unattributed Alleles

In the H1 table below notice the columns became fewer than used in the previous two scenarios. This is because once you explain the CSP using Pr(D) you merely use the Unattributed Allele(s) for explaining Pr(~~D~~), Pr(C), and Pr(~~C~~).



## Normalizing the Allele Frequencies

When an Assumed Sample is designated, the allele frequencies for calculating the Pr(C) come from a normalized table. The table is normalized after removing the Assumed Sample’s alleles. This is accomplished by computing a new *N* (the new count of alleles is now 481). The frequency for the 11 allele is now 160 / 481 or .3326. The normalized frequency is then adjusted for FST as described in Section 1.



The Pr(C) in this scenario was set at .03. Therefore, the probability that an 11 allele will drop-in is **.03 x .336 = .0101.**

## H2 0 S | 1 UNK with an Assumed Sample

The Assumed Sample is included in each row of the table below for clarity however other than calculating the Unattributed alleles it is not part of the calculations. The Unknown profile can be any combination of alleles 10, 11, 13, and Z. Each combination is provided a new row.

The same logic described in previous sections is again applied to each row, column by column.



## Stress Testing the Logic

The logic described in this section holds together throughout the range of Drop-out probabilities.



Part Four

# Introducing IBD

Each new section is getting shorter as new concepts and variables are introduced—primarily because we are building on previously established logic. I will introduce the Identical by Descent (IBD) logic by using the defined scenario in Part Three and modify it only for a new IBD.

The following is an IBD Table as it is typically displayed in the literature where k0 represents no alleles shared IBD, k1 one shared allele IBD, and k2 where both alleles are shared IBD.

|  |  |  |  |
| --- | --- | --- | --- |
| Relationship | k0 | k1 | k2 |
| Identical Twin | 0.0 | 0.0 | 1.0 |
| Parent-Offspring | 0.0 | 1.0 | 0.0 |
| Full Sibs | 0.25 | 0.5 | 0.25 |
| Half Sibs (or Uncle – Niece) | 0.5 | 0.5 | 0.0 |
| First Cousins | 0.75 | 0.25 | 0.0 |
| Unrelated (Default) | 1.0 | 0.0 | 0.0 |

Table IBD Table

## LabR Default IBD Values

LabR defaults to an IBD value of [1|0|0] which for two diploid individuals, that are not inbred, simply indicates that two alleles will not be shared IBD between two unrelated individuals.

## IBD Big Picture

Without getting lost in the “Forest” of calculations let’s explore the “big picture”.

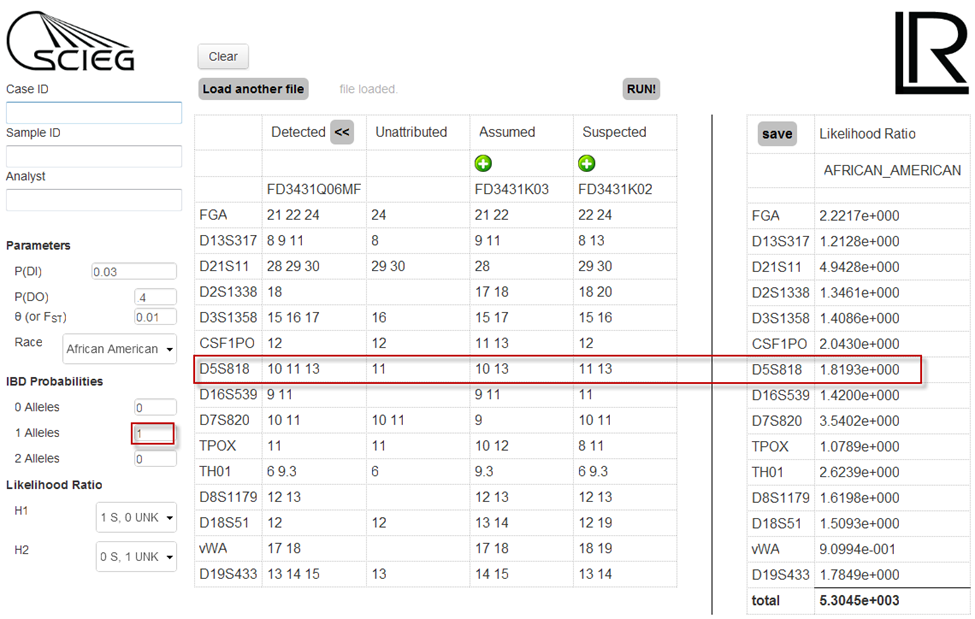
It should be recognized that if the Unknown in H2 is not a Random but a Related person—the allele pool is restricted to the family pedigree. In that light, it is easy to recognize the H2 random match probability becomes much less “random” and the ultimate L/R will be a become a smaller value.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Unrelated | 1st Cousins | Half Sibs & Avuncular | Full Sibs & Parent-Child | | Twins | |
|  | | Degree of Randomness | |  | | |
|  | | | | | | |
| High | Moderate | Low | | | | None |

This calculation is not as daunting as one might think. The fact is:

1. H1 remains unchanged
2. Pr(D), Pr(~~D~~), P(C), Pr(~~C~~) logic remains unchanged in H2.
3. The only change will be the RMP in H2

The following LabR screenshot is set to calculate the H2 alternate hypothesis using an Unknown 1st degree relative (Parent-Offspring) of the H1 Suspected profile. The IBD for this scenario will be [0|1|0].



## RMP KIN

The only metric that will change is the H2’s RMP of the Unknown. We can call this new probability RMP KIN. The RMP KIN is determined by selecting the genotype combinations in H2 that supports the hypothesis that a Parent–Offspring relationship between the H1 Suspected profile and H2 Unknown profile exist.

## IBD Probabilities

The RMP KIN for each possible combination in H2 should consider four probabilities:

1. Could both alleles be IBD (k2)
2. Could allele 1 be IBD (k1)
3. Could allele 2 be IBD (k1)
4. Could no alleles be IBD (k0)

For the Parent-Offspring relationship with an IBD [0|1|0] we see options of k2 andk0 arenot possible. The CSP in H2 is now explained by an Unknown Parent-Offspring person of the H1 Suspected profile. The RMP KIN for this scenario is shown for each combination below.

### Parent-Offspring IBD [0|1|0]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | 1st Degree Relative IBD ( 0 | 1 | 0) |  |
| Suspected | | Unknown | | Formula | RMP KIN |
| 11 | 13 | 10 | 10 | Not Possible |  |
| 11 | 13 | 10 | **11** | f(10)\*0.5 | 0.0358 |
| 11 | 13 | 10 | **13** | f(10)\*0.5 | 0.0358 |
| 11 | 13 | 10 | Z | Not Possible |  |
| 11 | 13 | 11 | **11** | f(11)\*0.5 | 0.1196 |
| 11 | 13 | **11** | **13** | [f(11)+f(13)]\*0.5 | 0.2342 |
| 11 | 13 | **11** | Z | f(Z)\*0.5 | 0.2300 |
| 11 | 13 | 13 | **13** | f(13)\*0.5 | 0.1146 |
| 11 | 13 | **13** | Z | f(Z)\*0.5 | 0.2300 |
| 11 | 13 | Z | Z | Not Possible |  |

For each genotype combination that supports the IBD hypothesis the RMP KIN is calculated using the FST adjusted allele frequencies and the possible IBD Probabilities. In the scenario above the combinations that do not support a Parent-Offspring relationship are zero / not possible.

The new RMP KIN column then replaces RMP in H2 as seen in the following table. The scenario depicted is based on a CSP 10,11,13, Assumed 10,13, Suspected 11,13, Pr(D) = 0.2, Pr(C) = 0.03, Theta = .01, α = 0.5 with L/R 1 S, 0 UNK | 0 UNK, 1 S and IBD [0|1|0].



Again the logic passes the Stress Test throughout the range of drop-out probabilities.



Calculating an IBD for Parent-Offspring was a simplistic scenario. But the same concepts apply to the more fanciful Half Sib RMP KIN. Just step through the 4 IBD Probabilities for each combination and the formula will present itself.

### Hall Sib IBD [0.5|0.5|0]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | Half Sibs IBD (.5 | .5 | 0) |  |
| Suspected | | Unknown | | Formula | RMP KIN |
| 11 | 13 | 10 | 10 | f(10)^2/2 | 0.0026 |
| 11 | 13 | 10 | 11 | f(10)/4+(2\*f(10)\*f(11)/2) | 0.0351 |
| 11 | 13 | 10 | 13 | f(10)/4+(2\*f(10)\*f(13)/2) | 0.0343 |
| 11 | 13 | 10 | Z | (2\*f(10)\*f(Z))/2 | 0.0330 |
| 11 | 13 | 11 | 11 | f(11)/4+f(11)^2/2 | 0.0884 |
| 11 | 13 | 11 | 13 | f(11)/4+f(13)/4+(2\*f(11)\*f(13)/2) | 0.1719 |
| 11 | 13 | 11 | Z | f(Z)/4+(2\*f(11)\*f(Z)/2) | 0.2250 |
| 11 | 13 | 13 | 13 | f(13)/4+(f(13)^2/2) | 0.0835 |
| 11 | 13 | 13 | Z | f(Z)/4+(2\*f(13)\*f(Z)/2) | 0.2204 |
| 11 | 13 | Z | Z | f(Z)^2/2 | 0.1058 |

Again by replacing the H2 RMP column with the Half Sib RMP KIN factors you will successfully calculate the IBD [0.5|0.5|0] modified hypothesis.



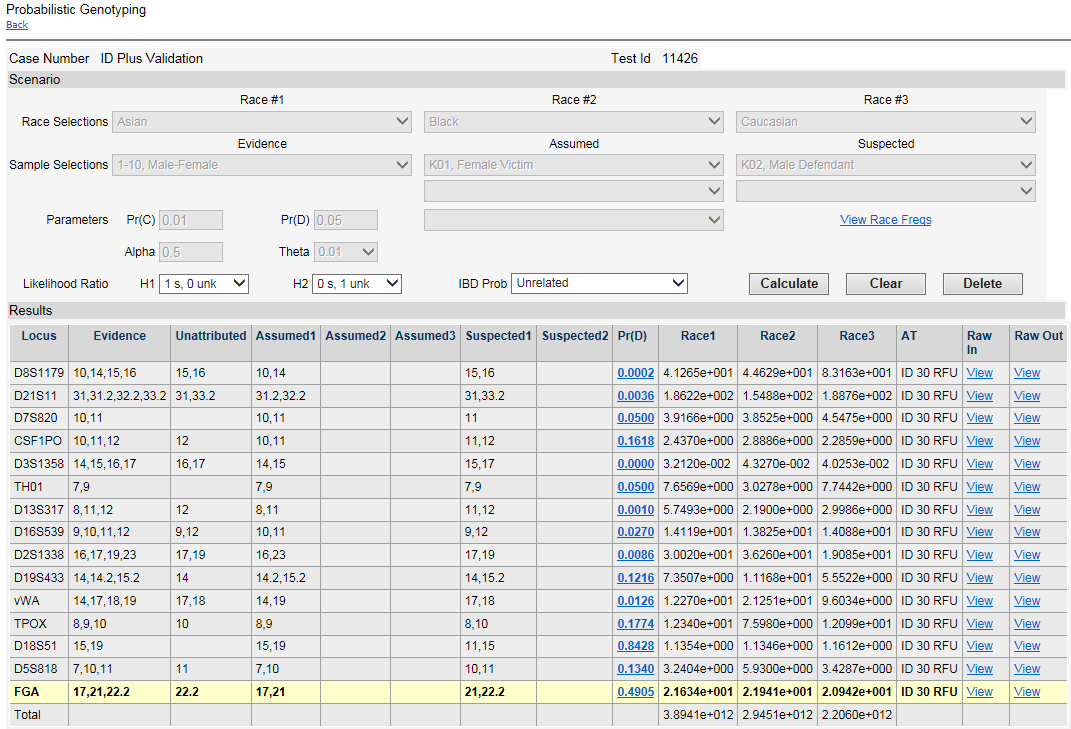
Our logic withstands the Stress Test.



The companion Excel spreadsheet named Abby will be made available to anyone who requests a copy.

# eDNA LIMS Integration

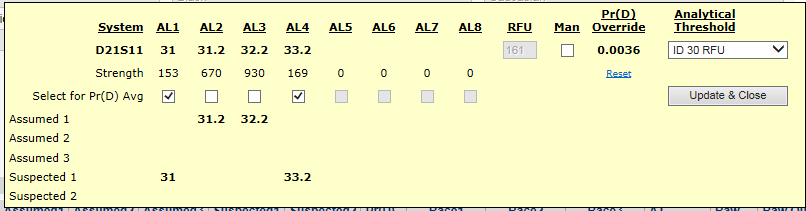
With permission from SCIEG the LabR calculation logic has been integrated into the eDNA LIMS. This integration allows for any number of Population Database selections, templatized forensic report generation and the ability to memorialize all case specific calculations within the electronic case.



The PG eDNA Integration includes the following additional features:

* Multiple Race Database options
* System by system Pr(D) selection (can vary with amplicon length)
* RFU values directly imported
* Easy access to input, output, and allele frequency data for validation
* Simplified IBD selection
* Calculation / scenario / results dependent report wording generation.
* Ability to include multiple Analytical Thresholds (Chemistry, equipment, dye lane specific…)

The following is an example of the granularity presented to the User when selecting the average RFU value(s) to be used to calculate a specific system’s probability of dropout.



## Author Contact

Feel free to contact me with Probabilistic Genotyping calculation logic questions or suggestions in making this paper more readable.

Kent M. Harman, 877-451-4363

[kharman@genetictechnologies.com](mailto:kharman@genetictechnologies.com)