

Formalization of matching strategies for duplicated loci

L.A.D. Hutchison^{1,2}, N.M. Myres², K.H. Ritchie², J.E. Ekins², J.B. Ekins², L. Layton², M.L. Lunt², S.S. Maesek², A.A. Nelson², M.E. Nelson², K.L. Pennington², U.A. Perego², J.L. Peterson², T. Tolley², S.R. Woodward².

¹ Computational Biology Research Group, CSAIL, Massachusetts Institute of Technology, MA, USA ² Sorenson Molecular Genealogy Foundation (<http://smgf.org/>), Salt Lake City, UT, USA
Email: luke.hutch@mit.edu, scott@smgf.org

Abstract. Several commonly-used Y chromosome STR loci (including DYS385, DYS459 and YCAII) are referred to as *duplicated*, because PCR using the standard primers amplifies multiple regions of the chromosome, resulting in more than one allele at the locus. In general it is not possible to determine which of the possible locus copies each allele came from. This makes it difficult to reliably compare the correct alleles between two duplicated-locus genotypes. It is important for analysis purposes however to be able to determine the correct origin of each allele, to improve the accuracy of various genetic similarity calculations such as tMRCA (time to Most Recent Common Ancestor), by eliminating the chance of a spurious match or mismatch due to confusion of allele identity. In this research we describe a more formal analysis of the problem of matching of duplicated loci, with the goal of measuring error inherent in current strategies, and of producing a new matching strategy that minimizes error.

1 BACKGROUND

The least error-prone way to match haplotypes including duplicated loci has not previously been formalized, leading to a range of ad-hoc matching strategies in current use. No consensus currently exists as to the best way to count the number of matching loci between two duplicated-locus genotypes, and the relative benefits and disadvantages of different matching strategies has not previously been analyzed.

Examples of questions that arise while considering how to match duplicated loci include: Should a duplicated locus be handled as two loci or one? Should each match between the two genotypes (with alleles considered in either order) be counted as one match out of two, summed towards a total out of two? Alternatively, should a mismatch at either or both alleles in the genotype cause the whole duplicated locus to be treated as a “zero out of one” match?

Note that it is not always correct to treat a duplicated locus as two loci, because the identity of the two alleles at the locus is generally unknown. This can lead to either spurious matching or spurious mismatching if the duplicated locus is treated as two separate loci, because an arbitrary identity is assigned to each locus copy in this process.

2 DEFINITIONS

A duplicate locus consisting of alleles A and B will be written as $\overline{A \ B}$. Note that there is no specific order implied for the two alleles in this notation, meaning A can be greater than, less than, or equal to B. The homozygous duplicated-locus genotype consisting of just allele A will be written as $\overline{A \ A}$ for clarity.

We define the *observed number of matches* between two duplicated-locus genotypes $P = \overline{P_0 \ P_1}$ and $Q = \overline{Q_0 \ Q_1}$ as:

$$M_o = \sum_{i=0}^1 \begin{cases} 1, & P_i = Q_0 \text{ or } P_i = Q_1 \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7
Genotype 1	$\overline{A \ B}$	$\overline{A \ A}$	$\overline{A \ A}$	$\overline{A \ A}$	$\overline{A \ A}$	$\overline{A \ B}$	$\overline{A \ B}$
Genotype 2	$\overline{C \ D}$	$\overline{B \ C}$	$\overline{B \ B}$	$\overline{A \ B}$	$\overline{A \ A}$	$\overline{A \ D}$	$\overline{A \ B}$
Observed Matches M_o	0	0	0	1	2	1	2
Actual Matches M_a	0	0	0	1	2	0 or 1	0 or 2
	Trivial Cases			Ambiguous Cases			

Figure 1. A summary of all possible match situations between two duplicate-locus genotypes. Lines connecting the genotypes designate a potential match, while “=” designates homozygosity. M_a describes the number of matches that would be counted if the locus copy identities were known. Cases 1-3 trivially yield $M_a = 0$, while cases 4 and 5 are interesting in that, due to homozygosity, they can accurately predict M_a without knowing which locus copy the allele A actually belongs to.

Thus $M_o \in \{0, 1, 2\}$, corresponding to how many alleles of one genotype match one or the other of the alleles of the other genotype. The *actual number of matches* M_a describes the number of matches that would be counted if the true identity of each allele were known, and if matches were only counted between alleles from the same locus copy. It will always be true that $M_a \leq M_o$. If $M_a < M_o$, it is due to either (i) spurious matches between alleles of the same value at different locus copies, or (ii) spurious matches due to *Identity By State (IBS)*, where alleles match due to mutation rather than common descent. Initially, IBS matches will be ignored. If IBS matches are to be considered, then $M_a = x$ should be replaced by $M_a \leq x$.

The dataset used for analysis (the “SMGF dataset”) refers to a dataset of 13767 Y chromosome haplotypes produced by the Sorenson Molecular Genealogy Foundation. These samples were taken from all over the world (predominantly from the USA), and most samples are complete at 3 duplicated STR loci (DYS385, DYS459 and YCAII) and 33 non-duplicated loci. Pairwise comparison of all 94,758,261 pairs of these haplotypes was performed to derive the statistics for this research.

3 ANALYSIS

Figure 1 shows the complete list of unique situations that may occur in duplicated-locus comparisons. Due to various symmetries, the total number of these *match cases* collapses down to just seven.

It is clear from the comparison of M_o to M_a that in cases 1-5, a duplicated locus may indeed be treated unambiguously as two separate loci, because M_a may be determined directly from M_o . Cases 6 and 7 may however each be interpreted in two different ways, depending on the identity of the alleles. This would not be a big problem if the prevalences of case 6 and case 7 were not high, but they collectively comprise 61% of pairwise matches in the SMGF dataset (see Figure 2). Note also that cases 1-3 collectively make up only 21% of match cases, meaning there is a very high probability of at least one matching allele at a duplicated locus between random pairs drawn from the population, regardless of relatedness of the individuals. This indicates a high degree of spurious matching, and emphasizes the need to minimize error in the matching of duplicated loci.

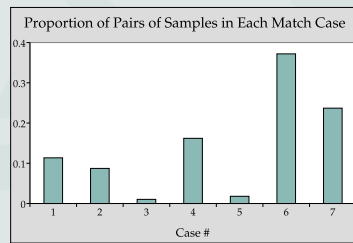


Figure 2. Proportion of pairs of duplicated loci that fall into each match case. Calculated from 256,911,555 pairs of duplicated loci (three duplicated loci total) between 94,758,261 pairs of samples in the SMGF dataset.

For the ambiguous cases (6 and 7), it would be possible to create an MLE classifier by measuring the most likely of the two values for M_a given M_o , and always choosing that value. However in order to produce a classifier that more accurately predicts M_a , we should base our prediction not just on M_o but upon the total number of matching non-duplicated loci between the two genotypes, n , because spurious matches are more likely to occur between more distantly-related samples, which can be approximately measured by counting the total number of matches across both haplotypes. In the SMGF dataset, we have $N = 33$ non-duplicated loci to assist in choosing M_a .

Consider the posterior likelihood of case j conditioned upon the total number of matching non-duplicated loci n (Figure 3). The shape of these likelihood distributions is intuitive for $n \geq 30$ (implying high likelihood of relatedness): the two cases with possible values of $M_a = 2$ (cases 5 and 7) spike up sharply, and all other cases (with $M_a < 2$) spike downward sharply.

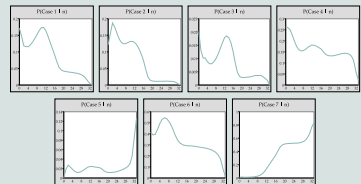


Figure 3. The posterior likelihood that a duplicated locus is in match class j , given that n out of 33 non-duplicated loci in the genotypes match.

The other strong features (local maxima and minima) in these posterior likelihood distributions are produced by nonuniform class membership w.r.t. n , which is due to the effects of mutation. This can be more clearly seen if the classes corresponding to each possible value of M_a are combined. This may be approximated by grouping together classes with equal values of M_o (Figure 4).

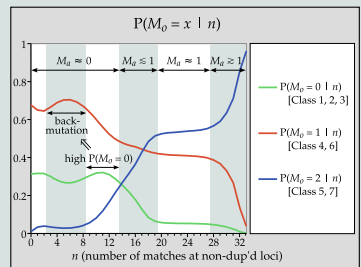


Figure 4. The posterior likelihood of M_o given that n out of 33 non-duplicated loci in the genotypes match. Values of n over which different reasonable hypotheses about the value of M_a are shown.

Based on the shapes of the distributions in Figure 4 and our knowledge of basic genetic inheritance, ranges of n may be clearly delineated over which M_a is most likely to be close to zero, between zero and one, close to one, and between one and two. Interestingly, a strong peak in $P(M_o = 1 | n)$ is also clearly discernible at low values of n , which is most plausibly explained by IBS matches due to back-mutation between distantly-related individuals.

Plausible predictions as to the value of M_a over these ranges allow us to approximate $E(M_a | n)$, and hence produce a classifier that yields $P(M_a | n, M_o)$. This classifier will be subject to the constraint $M_a \leq M_o$, meaning we are trying to determine how many of the observed matches are actual matches (and not due to spurious matches either due to locus identity confusion or IBS). Because the ranges of n delineated in Figure 4 are derived from a dataset subject to IBS matches, the resulting classifier should handle prediction of M_a even in the presence of matches by state.

We define a sigmoid function blending smoothly from $(1 - p)$ to p between x_0 and x_1 :

$$S(x_0, x_1, p) = \frac{1}{1 + \exp\left[\frac{-x_1 + x_0 - 2x}{x_1 - x_0} \ln\left(\frac{p}{1-p}\right)\right]} \quad (2)$$

To approximate $P(M_a | n, M_o)$, we require that $P(M_a = 2 | n, M_o)$ blends from 0.1 to 0.9 over the range [20,27], and that $P(M_a = 0 | n, M_o)$ blends from 0.9 to 0.1 over the range [14,19]. (Note that these parameters are determined by manual examination; it is left as future work to determine them optimally for a specific dataset.)

$$\begin{aligned} P(M_a = 2 | n, M_o = x) &= \alpha(2, x) S(20, 27, 0.9) & (3) \\ P(M_a = 0 | n, M_o = x) &= \alpha(0, x) S(14, 19, 0.9) - \text{Eq. (3)} & (4) \\ P(M_a = 1 | n, M_o = x) &= 1 - \text{Eq. (3)} - \text{Eq. (4)} & (5) \end{aligned}$$

where $\alpha(y, x) = 1$ if $y \leq x$, 0 otherwise. (6)

Eq. (6) ensures the constraint $M_a \leq M_o$ is satisfied. These likelihoods are plotted in Figure 5.

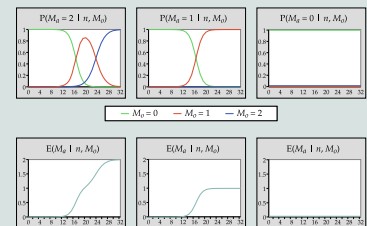


Figure 5. Top: The likelihood of M_a given M_o and n , subject to $M_a \leq M_o$, as predicted from Figure 4. Bottom: The expected value of M_a derived from each of these likelihood functions.

For simplicity’s sake, we use the MLE to predict M_a from these likelihood functions, yielding the straightforward classifier:

$$q_{01} = (14 + 19) / (2 \times 33) = 50\% \quad (7)$$

$$q_{12} = (20 + 27) / (2 \times 33) = 71\% \quad (8)$$

$$\hat{M}_a = \begin{cases} 2, & n/N \geq q_{01} \text{ and } M_o = 2 \\ 1, & n/N \geq q_{01} \text{ and } (M_o = 1 \text{ or } M_o = 2 \text{ and } n/N < q_{12}) \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

The threshold values q_{01} and q_{12} delimit classification regions for M_a based on n/N . These values were determined from our dataset using $N = 33$ loci, and although these thresholds will probably work reasonably well with other loci (assuming our loci are representative) and similar values of N , the amount of IBS matching will be too high for these thresholds to make sense for small N (e.g. $N < 12$). Determining whether a more optimal classifier could be produced by choosing other thresholds is left as future work.

In the absence of a set of augmented primers that would allow us to distinguish between alleles belonging to different locus copies, to test the error rate of this classifier we created a *pseudo duplicated locus* from two of the 33 non-duplicated alleles, leaving $N' = 31$ non-duplicated alleles for determining n . DYS390 and DYS456 were selected for this purpose, due to the Gaussian shape and wide spread of their distributions. Alleles were offset to align the mean of the two distributions. IBS matching was modeled as a random process for each match with $P(\text{IBS} | M_o = 1) = 1 - S(6, 13, 0.9)$. The correct assignment of alleles to the original locus were hidden from the classifier.

We compared our classifier against one of the most common matching strategies in current use, which assumes $M_a = M_o$. Average error rate of classification was measured across all pairs of samples in the SMGF dataset. The relative RMS/MSE error rates were: 1.42/2.02 for the standard classifier, and 1.04/1.09 for our new classifier, resulting in a 27%/46% decrease in error using the new classifier on the test dataset modeled as described above.

4 CONCLUSION

We have produced a simple classifier that improves prediction error for duplicated loci, based only upon the observed number of matching alleles at duplicated loci and the total number of matches at non-duplicated loci. The improvement is dramatic on our test dataset.