On the Probability of Dizygotic Twins Being Concordant for Two Alleles at Multiple Polymorphic Loci

Dale R. Nyholt

Queensland Institute of Medical Research, Brisbane, Australia

Accurate determination of same-sex twin zygosity is important for medical, scientific and personal reasons. Determination may be based upon questionnaire data, blood group, enzyme isoforms and fetal membrane examination, but assignment of zygosity must ultimately be confirmed by genotypic data. Here methods are reviewed for calculating average probabilities of correctly concluding a twin pair is monozygotic, given they share the same genotypes across all loci for commonly utilized multiplex short tandem repeat (STR) kits. Numerous tools enabling convenient and accurate zygosity probability calculation may be accessed via the ZygProb homepage at http://genepi.qimr.edu.au/general/daleN/ZygProb/

Accurate determination of twin zygosity is important for medical, scientific and personal reasons (Derom et al., 2001). Typically, questionnaire-based assessment of same-sex twin pair zygosity is relied upon, using questions about physical similarity and confusion in childhood. Example questions utilized in our laboratory include: Do you have the same eye colour? Do you have similar height, weight, and natural hair colour and texture? Were you usually mistaken for one another by nonfamily members as children? If the answer to these questions is 'yes', the pair is almost certainly monozygotic (MZ).

While most such methods combine information from both twins from a pair, answers from a single twin were found to yield a misclassification rate below 5% (Magnus et al., 1983). However, an approach utilizing latent class analysis was recently introduced which further improved the accuracy of questionnaire-based zygosity assessment (Heath et al., 2003). Nonetheless, while MZ twins usually have closely similar phenotypes they are seldom absolutely 'identical' — indeed some MZ twin pairs show quite marked phenotypic discordances. Moreover, it is preferable to determine zygosity at birth; examination of fetal membranes by properly trained personnel enables diagnosis of only two thirds of MZ twin pairs since one third are dichorionic (Derom et al., 2001). Consequently, assignment of zygosity must ultimately be confirmed by genotypic data, for example by utilizing the AMP*Fl*STR Profiler Plus™ PCR Amplification Kit (Applied Biosystems), widely used in forensic and paternity applications.

The desire to gain more information from a sample, coupled with the need to limit consumption of a DNA sample where its availability may be limited, has led to the coamplification and typing of multiple short tandem repeat (STR) systems. Kits from commercial sources, namely Promega and Applied Biosystems, are preferred due to their ease of use. Commonly utilized multiplex STR kits are listed in Table 1.

Given the multiplex kits listed in Table 1 differ in both number and site of STR loci, it is important to examine their differences. More specifically, we are interested in their respective average probabilities of correctly concluding a twin pair is MZ given they share the same genotypes across all loci.

Following the notation of Li (1996), Mendel's law of segregation dictates parents with genotypes (b,c) and (d,e) can produce four types of children (b,d), (b,e), (c,d) and (c,e) with equal probabilities. The 4×4 array

Table 1

Commonly Utilized Multiplex STR Kits

Manufacturer	Kit name (abbreviation)	Number of STRs
Promega	PowerPlex 1.1 (PPlex1.1)	8
	PowerPlex 2.1 (PPlex2.1)	9
	PowerPlex 16 (PPlex16)	15
Applied Biosystems	AMP*Fl*STR COfiler (COfiler)	6
	AMP*Fl*STR Profiler (Profiler)	9
	AMP*Fl*STR Profiler Plus (ProfilerP)	9
	AMP*Fl*STR SGM Plus (SGMPlus)	10
	AMP*Fl*STR SEfiler (SEfiler)	11
	AMP*Fl*STR Identifier (Identifier)	15

Received 7 December, 2005; accepted 14 December, 2005.

Address for correspondence: Dale R. Nyholt, Queensland Institute of Medical Research, PO Royal Brisbane Hospital, Queensland 4029, Australia. E-mail: Dale.Nyholt@qimr.edu.au

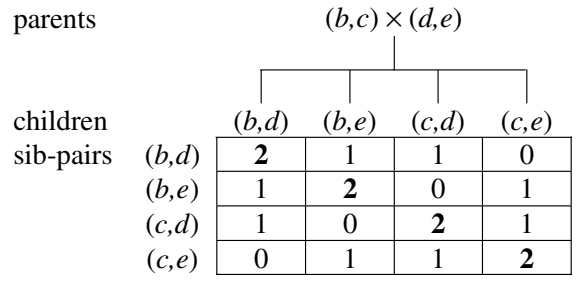

Figure 1

The three types of DZ twin pairs from a cross in which all parental alleles can be uniquely identified.

Note: Genotypes are represented by letters within brackets. Numbers represent the number of alleles shared identical by descent (IBD).

(Figure 1) shows the 16 possible dizygotic (DZ) twin pairs (/sib-pairs) arising from a fully informative mating.

Based on the number of alleles shared identical by descent (IBD) between the 16 possible pairs, there are three types of DZ pairs: (1) those sharing no alleles (z_0), with frequency 4/16 = 1/4; (2) those sharing one allele (z_1), with frequency 8/16 = 1/2; and (3) those sharing two alleles (z_2), with frequency 4/16 = 1/4. Therefore, the probability of a genotype match between two DZ twins, denoted M(DZ), is as follows:

$$M(\mathrm{DZ}) = 1/4\ \mathrm{Prob}(z_0) + 1/2\ \mathrm{Prob}(z_1) + 1/4\ \mathrm{Prob}(z_2)$$

When the two DZ twins have no alleles identical by descent (z_0), they are like two unrelated individuals. Assuming a population in Hardy-Weinberg equilibrium, the probability of a genotype match between two random individuals is the sum of the frequencies of all pairs of identical genotypes. For simplicity, we first consider a locus with two alleles: allele A_i and allele A_j, with frequencies p_i and p_j, and genotypes $A_i\,A_i$ and $A_i\,A_j$, with frequencies ($A_i\ A_i$) and ($A_i\ A_j$). The frequency of two random individuals both being A_iA_i is $(p_i)^2 \times (p_i)^2 = (p_i)^4$ and the frequency of two random individuals both being A_iA_j is $(2p_ip_j) \times (2p_ip_j) = (2p_ip_j)^2$. Therefore, the random match probability for two random individuals,

$$M_0 = \sum_{i=1}^{n}(A_iA_i)^2 + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(A_iA_j\right)^2$$

$$= \sum_{i=1}^{n}(p_i)^4 + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(2p_ip_j\right)^2$$

For loci with more than two alleles (n), i and j take all values from 1 to n.

When two DZ twins have one allele identical by descent (z_1), they are like a parent-child pair where the probability of a genotype match between a parent and child is the sum of the homozygous probabilities. The frequency of a random individual being A_i is (p_i) the frequency of both a parent and child both being A_i is $(p_i)^2$. Therefore, the random match probability for a parent and child,

$$a_2 = \sum_{i=1}^{n}(p_i)^2$$

When two DZ twins have both alleles identical by descent (z_2), they are like MZ twins with a match probability of 1. Therefore, the probability of a genotype match between two DZ twins, M(DZ), is as follows:

$$M(\mathrm{DZ}) = 1/4\ \mathrm{Prob}(z_0) + 1/2\ \mathrm{Prob}(z_1) + 1/4\ \mathrm{Prob}(z_2)$$

$$= 1/4\ M_0 + 1/2\ a_2 + 1/4$$

$$= 1/4\left(\sum_{i=1}^{n}(p_i)^4 + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\left(2p_ip_j\right)^2\right) + 1/2\sum_{i=1}^{n}(p_i)^2 + 1/4$$

Using Australian Caucasian allele frequency data (Bagdonavicius et al., 2002) and the above exact M(DZ) equation (Li, 1996), the author calculated the average probability of a DZ twin pair sharing both alleles at all markers and the resulting probability of correct zygosity assignment for the commonly used multiplex STR kits COfiler, PPlex1.1, Profiler and ProfilerP for which allele frequency data were available (Table 2). That is, when the loci are unlinked, the allele sharing at each locus is independent and one simply multiplies together all individual marker M(DZ) probabilities.

In addition to the exact equation presented above, Presciuttini and colleagues (Presciuttini et al., 2002) showed that the probabilities (z_i) depend on locus heterozygosity (H), and are scarcely affected by variation of the distribution of allele frequencies. This allowed them to obtain empirical curves relating z_is to H for a series of common relationships, so that the likelihood ratio of a pair of relationships between any two individuals, given their genotypes at a locus, is a function of a single parameter, H. Plotting the sharing probabilities for 19 STR

Table 2

Exact Probability of an Australian Caucasian DZ Pair Sharing the Same Two Alleles

STR locus	Common multiplex STR kits			
	PPlex1.1	COfiler	Profiler	ProfilerP
TPOX	.486	.486	.486	
D5S818	.431		.431	.431
CSF1PO	.422	.422	.422	
D16S539	.379	.379		
D13S317	.372		.372	.372
THO1	.383	.383	.383	
D3S1358		.374	.374	.374
VWA	.362		.362	.362
D8S1179				.357
D7S820	.360	.360	.360	.360
D21S11				.337
FGA			.326	.326
D18S51				.319
Number of STRs	8	6	9	9
Overall M(DZ)	6.21E-04	4.01E-03	2.00E-04	9.79E-05
$\mathrm{Odds}_{\mathrm{MZ:DZ}}$	1611	250	4993	10219
$\mathrm{MZ\ Certainty}_{\mathrm{avg}}$ (%)	99.93792	99.59927	99.97997	99.99021

Note: Utilizing allele frequency data from Bagdonavicius et al. (2002) and applying the exact M(DZ) formula of Li (1996).

Table 3

Approximate Probability of an Australian Caucasian DZ Pair Sharing the Same Two Alleles

STR locus	Common multiplex STR kits			
	PPlex1.1	COfiler	Profiler	ProfilerP
TPOX	.489	.489	.489	
D5S818	.431		.431	.431
CSF1PO	.419	.419	.419	
D16S539	.377	.377		
D13S317	.372		.372	.372
THO1	.381	.381	.381	
D3S1358		.374	.374	.374
VWA	.362		.362	.362
D8S1179				.357
D7S820	.360	.360	.360	.360
D21S11				.337
FGA			.326	.326
D18S51				.319
Number of STRs	8	6	9	9
Overall $M(DZ)_{approx}$	6.19E-04	3.96E-03	1.99E-04	9.75E-05
$Odds_{MZ:DZ}$	1618	252	5036	10,252
MZ $Certainty_{avg}$ (%)	99.93820	99.60378	99.98014	99.99025

Note: Utilizing expected heterozygosity (H) values reported by Bagdonavicius et al. (2002) and applying the $M(DZ)_{approx}$ formula of Presciuttini et al. (2002).

loci relating to H produced the following equation for a third order polynomial curve for the probability of a genotype match between two DZ twins, $M(DZ)_{approx}$:

$$M(DZ)_a = 0.7753 + 0.0358 \times H - 1.1771 \times H^2 + 0.6181 \times H^3$$

Table 3 contains approximate average probability of a DZ twin pair sharing both alleles at all markers and resulting probability of correct zygosity assignment, for the commonly used multiplex STR kits COfiler, PPlex1.1, Profiler and ProfilerP utilizing the expected heterozygosity (H) values reported by Bagdonavicius et al., (2002) and applying the $M(DZ)_{approx}$ formula of Presciuttini et al., (2002).

Upon comparing Table 2 and Table 3, it is clear the formula of Presciuttini et al. (2002) approximates the exact probabilities well. Furthermore, they report, because heterozygosity is a composite parameter it is inherently less variable among populations than individual allele frequencies. Hence heterozygosity is sufficiently homogeneous, at least among Caucasian populations, as to justify the adoption of a single common mean value, apart from special cases of historically isolated groups. Consequently researchers may simply utilize the marker $M(DZ)$ values reported by Presciuttini et al. (2002; and repeated in Table 4 below)

Table 4

Approximate Probability of a Caucasian DZ Twin Pair Sharing the Same Two Alleles

STR locus	Commonly utilized multiplex STR kits								
	PPlex1.1	PPlex2.1	PPlex16	COfiler	Profiler	ProfilerP	SGMPlus	SEfiler	Identifier
TPOX	.483	.483	.483	.483	.483				.483
D5S818	.422		.422		.422	.422			.422
CSF1PO	.414		.414	.414	.414				.414
D16S539	.386		.386	.386			.386	.386	.386
D13S317	.376		.376		.376	.376			.376
THO1	.376	.376	.376	.376	.376		.376	.376	.376
D3S1358		.374	.374	.374	.374	.374	.374.	374	.374
VWA	.362	.362	.362		.362	.362	.362	.362	.362
D8S1179		.358	.358			.358	.358	.358	.358
D7S820	.356		.356	.356	.356	.356			.356
D21S11		.336	.336			.336	.336	.336	.336
FGA		.328	.328		.328	.328	.328	.328	.328
D18S51		.318	.318			.318	.318	.318	.318
D19S433							.390	.390	.390
Penta D			.346						
D2S1338							.315	.315	.315
Penta E		.307	.307						
SE33								.283	
Number of STRs	8	9	15	6	9	9	10	11	15
Overall $M(DZ)_{approx}$	5.93E-04	9.47E-05	2.96E-07	3.86E-03	1.89E-04	9.60E-05	3.03E-05	8.57E-06	3.42E-07
$Odds_{MZ:DZ}$	1685	10,555	3,378,274	259	5302	10,422	33,013	116,653	2,922,758
MZ $Certainty_{avg}$ (%)	99.94066	99.99053	99.99997	99.61360	99.98114	99.99041	99.99697	99.99914	99.99997

Note: Utilizing approximate $M(DZ)$ values reported in Table 4 (FULL SIB, z_2 column) of Presciuttini et al. (2002).

to obtain probability of correct zygosity assignment for any combination of the 18 loci commonly used in the forensic practice. Table 4 contains $M(\mathrm{DZ})_{\mathrm{approx}}$ and resulting probability of correct zygosity assignment for the commonly utilized multiplex STR kits listed in Table 1. Moreover, because these markers are routinely used it should not be too difficult for researchers to obtain population-specific heterozygosity values appropriate for the twins under investigation.

Although the average certainty of correctly designating a twin pair sharing all alleles as MZ, is greater than 99% for all nine multiplex STR kits listed in Table 4, the chance of a DZ pair sharing all markers ranges from 1 in 259 (COfiler) to 1 in 3,378,274 (PPlex16). Indeed, as a result of determining accurate probabilities for the number of alleles being shared, researchers should keep in mind the possibility of genotyping errors and/or spontaneous mutations when determining zygosity. For example, the average probability of a DZ pair being z_2 at 8 of the 9 Profiler Plus loci range from 1.314E-04 to 2.058E-04 (odds of 1 in 7609 to 1 in 4859), the odds of being z_2 at 7 loci range from 1 in 4585 to 1 in 2372, the odds of being z_2 at 6 loci range from 1 in 2739 to 1 in 1200, while the odds of a DZ pair being z_2 at 5 loci range from 1 in 1554 to 1 in 663. However, error (including mutation) rates ranging from 0.25% to 2.38% (odds of 1 in 400 to 1 in 42) may be quite realistic (Ewen et al., 2000). Hence, it may be more likely that twin pairs sharing both alleles at 7 or 8 of the 9 Profiler Plus loci are MZ with genotyping errors than DZ.

To this end, one can either calculate the overall average probability for the observed number of loci for which the pair of individuals are z_2 (and not z_2) and compare this to an assumed error rate(s), or use a fully parametric approach such as that implemented in the ECLIPSE2 program (Sieberts et al., 2002) which can provide the exact probability for a pair of individuals sharing particular alleles (i.e., the chance of a pair sharing common alleles is higher than for sharing rare alleles) and also take into account any correlation in allele sharing due to the use of linked markers. Given that a wide range of researchers may be interested in the latter approach, the author has implemented a user-friendly www interface ('ZygProb') to ECLIPSE2, allowing users to simply upload three input files to obtain ECLIPSE2 likelihood results for a pair of individuals sharing the uploaded marker alleles. Likelihoods are given, assuming user-specified error rates, for being a DZ pair (/full-sibling pair), half-sibling, unrelated, and MZ pair. Additionally, the MZ/DZ likelihood ratio is given, which represents the odds in favour of the two individuals being an MZ pair compared to a DZ pair (i.e., odds greater than 1 indicate the pair is more likely to be MZ). Excel worksheets may also be downloaded from the ZygProb homepage (http://genepi.qimr.edu.au/general/daleN/ZygProb/) enabling easy calculation of exact and approximate random match probabilities $M(\mathrm{DZ})$ from allele frequency and heterozygosity data, respectively.

Acknowledgments

I thank Nick Martin and Peter Visscher for helpful suggestions. DRN was supported by an NHMRC RD Wright Fellowship #339462.

References

Bagdonavicius, A., Turbett, G. R., Buckleton, J. S., & Walsh, S. J. (2002). Western Australian sub-population data for the thirteen AMP*Fl*STR Profiler Plus and COfiler STR loci. *Journal of Forensic Sciences*, *47*, 1149–1153.

Derom, R., Bryan, E., Derom, C., Keith, L., & Vlietinck, R. (2001). Twins, chorionicity and zygosity. *Twin Research*, *4*, 134–136.

Ewen, K. R., Bahlo, M., Treloar, S. A., Levinson, D. F., Mowry, B., Barlow, J. W., & Foote, S. J. (2000). Identification and analysis of error types in high-throughput genotyping. *American Journal of Human Genetics*, *67*, 727–736.

Heath, A. C., Nyholt, D. R., Neuman, R., Madden, P. A., Bucholz, K. K., Todd, R. D., Nelson, E. C., Montgomery, G. W., & Martin, N. G. (2003). Zygosity diagnosis in the absence of genotypic data: An approach using latent class analysis. *Twin Research*, *6*, 22–26.

Li, C. C. (1996). Population genetics of coincidental DNA matches. *Human Biology*, *68*, 167–184.

Magnus, P., Berg, K., & Nance, W. E. (1983). Predicting zygosity in Norwegian twins born 1915-1960. *Clinical Genetics*, *24*, 103–112.

Presciuttini, S., Toni, C., Tempestini, E., Verdiani, S., Casarino, L., Spinetti, I., De Stefano, F., Domenici, R., & Bailey-Wilson, J. E. (2002). Inferring relationships between pairs of individuals from locus heterozygosities. *BMC Genetics*, *3*, 23–33.

Sieberts, S. K., Wijsman, E. M., & Thompson, E. A. (2002). Relationship inference from trios of individuals, in the presence of typing error. *American Journal of Human Genetics*, *70*, 170–180.