International Congress Series 1239 (2003) 825–829

Grading of qualitative and quantitative responses in the PI proficiency survey of the College of American Pathologists for mailings in 1997–2000

R.W. Allen*, A. Eisenberg, C. Harrison, R.H. Walker, C.T. Young, D.L. Zeagler, R. Roby, H.F. Polesky

Parentage Testing Committee, College of American Pathologists, Northfield, IL, USA

Abstract

The PI Survey offered jointly by the American Association of Blood Banks (AABB) and the College of American Pathologists (CAP) provides biological samples for proficiency testing that mirror paternity casework samples. A grading scheme providing for three levels of performance (good, acceptable, and unacceptable) has been developed for responses provided by subscribers. Reported here is a review of the performance of participants using RFLP and PCR/STR methods for the 1997–2000 period. In general, laboratories reporting phenotypes and paternity index values using RFLP methods exhibited a fairly constant rate of unacceptable responses over the 4-year period. In contrast, laboratories using PCR/STR systems improved dramatically in the frequency of unacceptable responses (an average of 3.5% for mailings in 1997 vs. 0.7% for mailings in 2000). A comparison of unacceptable responses for RFLP and PCR/STR methods over the 4-year period shows that whereas phenotype designations for PCR/STR systems is better than for RFLP systems, the opposite is true for reported paternity index values.

Keywords: PI Survey; Phenotype; Paternity index value

1. Introduction

Quality assurance programs in effect in clinical testing laboratories incorporate proficiency testing as an essential component to demonstrate competence in the policies, procedures and personnel involved with producing accurate test results. Identity testing is

[*] Corresponding author.

0531-5131/03 © 2003 Elsevier Science B.V. All rights reserved.
PII: S0531-5131(02)00622-2

no exception to this concept and proficiency testing programs targeted at laboratories engaged in parentage and forensic testing are available from several organizations. The parentage testing program offered jointly by the American Association of Blood Banks (AABB) and the College of American Pathologists (CAP) known as the PI Survey has existed since 1990 and began as a sample sharing program. Laboratories participating in the survey receive blood samples and/or buccal swabs obtained from volunteer trios and perform their test battery on the samples. Results are communicated back to the College for collation and comparison by the Parentage Testing Committee. Analysis of the results is peer based and consensus phenotypes and inclusion/exclusion answers are forwarded to subscribers along with a narrative of interesting findings and observations made by the Committee.

For a proficiency testing program to be fully utilized as a quality assurance tool, grading of responses demonstrates and documents that an individual subscriber has provided answers that were also reached by a consensus of peers. In 1997, the PI Survey began grading qualitative and quantitative responses and provided subscribers with individual graded reports of performance on each challenge in which they participated [1]. The grading criteria applied to responses are peer based and require a minimum of 10 members of the peer group, 9 of which report the same qualitative result. Examples of qualitative results include phenotypes for genetic markers consisting of discreet alleles (i.e. blood group systems and PCR/STR systems for example). Grading was also extended to selected quantitative responses by participants in 1997 as well. Grading of quantitative responses is also peer based and relies upon establishing a mean and standard deviation from among responses. Three standard deviations define the limits of acceptable responses thereby identifying outliers. The remaining acceptable responses (there must be a minimum of 10 remaining) are then used to establish a new mean and standard deviation. The three-standard deviation filter is again applied to the data identifying additional outliers should they exist. This grading method has been designated the 3SDX2 scheme [1,2]. Examples of quantitative data subjected to 3SDX2 grading included RFLP phenotypes (in 1997) and more recently paternity index values (in 1999).

This study summarizes the grading experience of the PI Survey over the 4 years of mailings (three times each year) from 1997 through 2000.

2. Study design

Graded responses from participating laboratories were collected over the three annual mailings from 1997 to 2000. Peer group-based grading was performed as described above. Graded responses incorporated into the study included PCR/STR phenotypes (graded as a qualitative response) and RFLP phenotypes and paternity index values for both RFLP and PCR/STR systems (quantitative responses). It should be noted that the quantitative responses for paternity index values for RFLP systems included all values for each locus, regardless of the restriction enzyme used. This approach increased the number of PI values that could be graded and earlier studies showed little, if any, effect of grouping responses by locus rather than by locus and enzyme on the unacceptable response percentage [2].

3. Results

The results of our analysis are summarized in Table 1. An examination of the unacceptable response rate for PCR/STR phenotypes (Table 1) reveals that reporting the correct PCR/STR phenotype has improved over the 4-year period to a level less than 1% for all of the mailings in the year 2000. In contrast, unacceptable grades for RFLP phenotypes have remained fairly constant, averaging about 3–4% for all years except 1999 in which rates were less than half the average of the other 3 years. Interestingly for

Table 1
Incorrect qualitative and quantitative responses from participants enrolled in the PI Survey for 1997–2000[a]

Mailing	STR[b]	PI-STR[c]	RFLP[c]	PI-RFLP[c]
	Phenotype	Systems	Phenotype	Systems
1997[d]				
PI-A	4/276 (1.5%)	2/62 (3.2%)	10/488 (2.1%)	2/66 (3.0%)
PI-B	12/309 (3.9%)	1/89 (1.1%)	25/503 (5.0%)	1/71 (1.4%)
PI-C	17/336 (5.1%)	5/78 (6.4%)	17/486 (3.5%)	0/51 (0%)
Mean	3.5%	3.6%	3.5%	1.5%
1998[d]				
PI-A	14/440 (3.2%)	6/125 (4.8%)	16/504 (3.2%)	1/60 (1.7%)
PI-B	13/360 (3.6%)	5/104 (4.8%)	17/540 (3.2%)	0/63 (0%)
PI-C	5/408 (1.2%)	3/100 (3.0%)	22/468 (4.7%)	0/60 (0%)
Mean	2.7%	4.2%	3.7%	0.6%
1999				
PI-A	45/1709 (2.6%)	29/413 (7.0%)	8/1320 (0.6%)	1/164 (0.6%)
PI-B	31/2155 (1.4%)	19/612 (3.1%)	14/1344 (1.0%)	2/101 (2.0%)
PI-C	44/2718 (1.6%)	24/712 (3.4%)	18/1029 (1.8%)	3/169 (1.8%)
Mean	1.9%	4.5%	1.5%	1.1%
2000				
PI-A	22/3104 (0.71%)	49/1269 (3.9%)	11/212 (4.7%)	1/34 (2.9%)
PI-B	25/3494 (0.72%)	46/1360 (3.4%)	6/224 (2.7%)	0/34 (0%)[e]
PI-C	20/3648 (0.55%)	20/1458 (1.4%)	10/214 (4.7%)	1/47 (2.1%)
Mean	0.7%	2.9%	4.0%	1.7%
Mean (all years)	2.2% ± 1.5%	3.8% ± 1.8%	3.2% ± 1.4%	1.2% ± 1.1%

[a] Incorrect responses are those qualitative data that do not match the target value established by a consensus of the peer group. Incorrect quantitative responses represent outliers identified by the 3SDX2 criteria (see footnote 3 for an explanation of the 3SDX2 method). In each category of graded response, the total number of unacceptable responses versus the total in that category are shown. The percentages in parentheses therefore represent the error rates in each category.

[b] Qualitative response graded against a target value established by the consensus of the peer group.

[c] Quantitative response graded by establishing a mean and standard deviation (S.D.) of a peer group of 10 or more participants. Each response outside of a range of ± 3 S.D. is classified as an outlier. A new mean and S.D. are calculated from the remaining responses and the outlier analysis is repeated (3SDX2 method).

[d] Data include responses for HUMCSF1P0, HUMTPOX, and HUMTHO1 for PCR/STR systems and D2S44/Hae III and D10S28/Hae III for phenotype responses and for these two loci regardless of enzyme used for the reported PI values.

[e] The alleged fathers were both excluded at the two RFLP loci capable of being graded.

Table 2
Summary of lab performance for PCR/STR and RFLP methods in the AABB/CAP PI Survey for 1997–2000

Year	STR	PI-STR	RFLP	PI-RFLP
	Phenotype	Systems	Phenotype	Systems
1997	3.5%	3.6%	3.5%	1.5%
1998	2.7%	4.2%	3.7%	0.6%
1999	1.9%	4.5%	1.5%	1.1%
2000	0.7%	2.9%	4.0%	1.7%
Mean (all years)	2.2%	3.8%	3.2%	1.2%

Data include responses for HUMCSF1P0, HUMTPOX, and HUMTHO1 for PCR/STR systems and D2S44/Hae III and D10S28/Hae III for RFLP phenotypes. PI grading for RFLP systems did not limit responses by restriction enzymes (i.e. all responses were considered together).

1999, the total number of RFLP phenotype responses was greatest reflecting a rather dramatic increase in the number of loci that could be graded. By the year 2000, the number of labs using PCR/STR technology increased at the expense of RFLP labs and so the number of graded RFLP responses decreased again.

If one compares the reported paternity index values for PCR/STR and RFLP systems, it is seen that unacceptable responses for PCR/STR systems have typically been about two-fold higher than that observed for RFLP systems. Table 2 summarizes the average unacceptable response rate by year for the four graded criteria. It is clear from Table 2 that whereas accuracy in assigning PCR/STR phenotypes has improved significantly over the 4-year period, the unacceptable response rate for PI values for PCR/STR systems has remained constant. Likewise, for RFLP systems, constancy in the rate of unacceptable responses for both phenotype and PI values was observed over the period examined. Thus, as of the completion of year 2000 mailings, PCR/STR laboratories are much better at assigning phenotypes than are RFLP labs, they are less able to report consistent paternity index values.

4. Discussion

This study extends prior studies [1,2] on the performance of laboratories subscribing to the PI Survey offered jointly by the AABB and CAP. The conclusions apparent from this study are that laboratories performing PCR/STR methods are more consistent in reporting phenotypes and have demonstrated this improvement over the 4-year period covered by this study. However, laboratories reporting PI values using PCR/STR methods are significantly less consistent in the values they report and are less consistent than labs reporting PI values using RFLP methods by a factor three- to four-fold. The reason for this discrepancy is unclear but may reflect the use of a more limited number of allele frequency databases by participants. Most laboratories using PCR/STR methods utilize the database provided by the supplier of the STR kit they use. If the databases are significantly different for particular alleles due to sampling or even population differences, those differences would perhaps be reflected in the frequency of unacceptable responses. In contrast, labs using RFLP methods typically develop their own databases from local populations and therefore the PI responses they provide represent somewhat of an amalgamation of allele frequencies among a number of related phenotypic designations and population groups. It is possible that one effect of this difference between RFLP and PCR/STR methodologies is to increase the magnitude of the standard deviations used for grading thereby broadening the range of acceptable responses and hence reducing the unacceptable response rate.

References

[1] H.F. Polesky, R.W. Allen, A.J. Eisenberg, C.R. Harrison, R. Roby, R.H. Walker, Scoring systems for DNA test results from parentage testing proficiency testing program, Prog. Forensic Genet. 7 (1998) 546–548.

[2] R.W. Allen, C.R. Harrison, A.J. Eisenber, R.K. Roby, R.H. Walker, R.E. Wenk, H.F. Polesky, Grading of quantitative data in the CAP/AABB parentage testing program, Prog. Forensic Genet. 8 (2000) 602–605.