TECHNICAL NOTE

David L. Duewer, Ph.D.; Margaret C. Kline, M.S.; Janette W. Redman; Pamela J. Newall, M.A.; and Dennis J. Reeder, Ph.D.

NIST Mixed Stain Studies #1 and #2: Interlaboratory Comparison of DNA Quantification Practice and Short Tandem Repeat Multiplex Performance with Multiple-Source Samples

ABSTRACT: The Mixed Stain Study 1 (MSS1, Apr.–Nov. 1997) and Mixed Stain Study 2 (MSS2, Jan.–May 1999) evaluated multiplexed short-tandem repeat (STR) DNA typing systems with samples containing DNA from more than one source. These interlaboratory challenge studies evaluated forensic STR measurement, interpretation, and reporting practice using well-characterized samples of very different analytical difficulty. None of the relatively few errors reported in either exercise resulted in a false identification of a reference source; several errors in evaluating the unknown source in three-source samples would hinder matching the profile in any archival database. None of the measurement anomalies reported is associated with any particular STR multiplex; all DNA amplification anomalies are associated with inefficient DNA extraction, inaccurate DNA quantitation, and/or analytical threshold policies.

KEYWORDS: forensic science, blood stains, DNA fingerprinting, DNA typing, evaluation studies, interlaboratory comparison, polymerase chain reaction, reproducibility of results, semen stains, short tandem repeat (STR) alleles

Here, we report results from two interlaboratory comparison exercises conducted by the National Institute of Standards and Technology (NIST). These studies explored the performance of multiplexed short-tandem repeat (STR) DNA typing systems with samples containing DNA from more than one source. Both the Mixed Stain Study #1 (MSS1) and #2 (MSS2) were motivated by concerns that multiple-source samples could present measurement and interpretation challenges to systems requiring the simultaneous amplification of DNA at many different STR loci: i.e., STR multiplexes (1,2).

Validation studies addressing the above concerns have and continue to be performed by individual forensic laboratories as well as by the STR systems’ manufacturers. Further, forensic DNA analysts regularly participate in proficiency tests that, while primarily intended to evaluate and document the sample evaluation skills of analysts, also evaluate analytical methods. However, both MSS1 and MSS2 were “challenge studies” designed to illuminate potential measurement weaknesses. By recruiting users of diverse analytical systems and analysts of varying experience, and challenging them with unusual and difficult samples, errors and analytical failures were elicited. Most measurement abnormalities were recognized as such by the performing analysts; these failures waste analytical resources. There were a very few instances of true alleles excluded from the “match profile” of an unknown donor in a mixed-source sample. While unlikely to affect (or go unrecognized) in any direct casework comparison, such errors could lead to less efficient lead prioritization by the Federal Bureau of Investigation’s CODIS, the Forensic Science Service’s National DNA Database, or other DNA database systems (3,4).

None of the clerical, measurement, or interpretation anomalies observed in MSS1 or MSS2 can be attributed to particular STR systems or instrumentation. We attribute the few clerical errors to analyst inattention. All allele measurement failures (true alleles not called, stutter called as an allele) are attributable to inefficient extraction of DNA from the samples, inaccurate estimation of the amount of DNA used in the PCR mixture, and/or analytical threshold policies. To the extent that simultaneous amplification of multiple alleles demands better control of initial conditions as the complexity of the system increases, highly multiplexed STR systems may well require improved DNA quantification technology.


Copyright © 2001 by ASTM International
Summary results of both the MSS1 and MSS2 studies were provided to all participants as each study was completed. Preliminary interpretations of the studies’ results were presented as soon thereafter as possible (5,6). This report describes the critical aspects of the study designs and results, with a focus on the performance and importance of DNA extraction and DNA quantification technologies.

Materials and Methods

To better anticipate the reference material and measurement quality assurance needs of the human identity community, NIST assesses real-world performance of DNA typing technologies via interlaboratory comparison exercises. In collaboration with members of the Scientific Working Group for DNA Analysis Methods (SWGDAM, previously termed T(echnical)WGDAM), commercial “CTT” (CFS1PO, TH01, TPOX) triplex and “CTTv” (CTT + VWA) quadruplex STR systems were evaluated by 34 participants during the period of Jan.–May, 1996 (7). Several participants in this CTT study reported unequal amplification of alleles at one or more loci. Since our laboratories also occasionally observed unequal allele amplification with these and other STR systems, MSS1 and then MSS2 were designed to document the extent and forensic implications of this potential measurement problem. The MSS1 samples were distributed to 28 laboratories beginning in Apr. 1997; the final data set was received in Nov. 1997. The MSS2 samples were distributed to 52 laboratories beginning in Jan. 1999; the final data set was received in May 1999. Table 1 lists the participants in the MSS1 and MSS2 studies.

Mixed Stain Study #1 (MSS1)

This study was designed around a standard forensic casework problem: identification of all sources of DNA in multiple-source samples given a complete reference set of potential sources. Eleven sets of samples were prepared on S&S 903 paper (Schleicher & Schuell, Inc., Keene, NH) from human Buffy coat cells (QC Products, Inc., Pompano Beach, FL). Six of the samples were single-source reference samples, four were two-source mixtures, and one was a three-source mixture. Only the reference sources were...
present in the multiple-source samples; all of the reference sources were used in one or more of the multiple-source samples. The particular source combinations were selected to minimize allelic overlap; i.e., to provide as uncomplicated a set of samples as possible given the source materials available. Table 2 lists the composition of all MSS1 samples.

Each stain was targeted to contain a minimum of 30 ng DNA per source. The concentration of DNA extractable from the Buffy coat cells for each of the six DNA sources was estimated as the average of replicate slot blot evaluations (8) of organic (9) and Chelex® (Bio-Rad Industries, Hercules, CA) extracts (10). All stains were prepared from volumetric aliquots of continuously stirred suspensions of the individual source materials in Phosphate Buffered Saline (PBS) (11). All stains for each sample set were thoroughly air-dried in a laminar-flow hood, labeled, and stored at −20°C before preparing the next sample. The multiple-source samples were prepared by sequential addition of the individual sources, with thorough drying under laminar flow between additions.

Two stains of each of the 11 samples were supplied to each participant. Several participants who did not achieve satisfactory amplification for one or more of the samples in the initial distribution were provided with a complete second set of stains. Participants were asked to identify, using their standard casework protocol, all of the reference sources represented in all of the multiple-source samples.

**Mixed Stain Study #2 (MSS2)**

This study was designed around an emerging forensic opportunity: defining a “searchable profile” for an unknown DNA source in a multiple-source stain given an incomplete reference set of potential sources. It was also designed to provide quantitative information on pre-amplification (DNA extraction and quantity determination) stages of the STR measurement protocols. Two qualitatively different kinds of quantitatively well-defined samples were used to achieve these goals.

Sets 1 and 2, Stains on Cotton Cloth—Two sample sets of three stains each were prepared to present different unknown-source scenarios. Both sets consisted of a female reference prepared from whole blood, a male reference prepared from commercially obtained semen (Cryogenics Laboratories, Inc., North Roseville, MN), and a “questioned-sample” prepared from a mixture of whole blood and semen. The three-source sample of the first scenario contained DNA from both the female and the male reference sources plus DNA from a second “unknown” male. The two-source sample of the second scenario contained DNA from the reference female and a male source different from that of the male reference. All stains were prepared under laminar flow on 4 cm by 4 cm squares of white cotton cloth (Jo Ann Fabrics, Hagerstown, MD) that had been bleached, twice washed, and UV sterilized. After preparation, stains were dried for 2 h at ambient temperature, labeled, sealed under vacuum in aluminized Mylar® bags (MIL-B-131F Class I, Columbus Packaging Co., Columbus, GA), and stored at −20°C. Table 2 lists the composition of these samples.

Each stain was targeted to contain about 1 μg (1000 ng) of DNA per source. The DNA concentration of each whole blood was estimated from a white cell count obtained shortly before blood donation. The DNA concentration of each semen was estimated from the original sperm count; the number of intact sperm at the time of sample preparation was not determined. Unlike the MSS1 samples, the multiple-source MSS2 samples were prepared from single aliquots of a continuously stirred PBS mixture of blood and semen. There were no detected qualitative or quantitative differences between stains produced at the beginning of production and those produced near the end for any of the six samples. Complete sample preparation details are provided elsewhere (6).

For these six samples, participants were asked to: 1) specify all possible profiles for all sources in each sample, 2) provide a “CODIS search profile” for the unknown source in the two multiple-source samples, and 3) estimate the total amount of recoverable DNA in each sample (ng/stain). We did not specify the stringency or the format of the search profile. Participants were also asked to provide all “relevant details” of their extraction and DNA quantitation protocols, again without further specification.

Table 2—MSS1 and MSS2 samples.

<table>
<thead>
<tr>
<th>Study</th>
<th>Samples</th>
<th>Matrix</th>
<th>Composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSS1</td>
<td>349</td>
<td>Buffy coats, S&amp;S 903 paper</td>
<td>30–50 ng ( \bar{\varphi}_{349} )</td>
</tr>
<tr>
<td></td>
<td>350</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{350} )</td>
</tr>
<tr>
<td></td>
<td>351</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{351} )</td>
</tr>
<tr>
<td></td>
<td>352</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{352} )</td>
</tr>
<tr>
<td></td>
<td>353</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{353} )</td>
</tr>
<tr>
<td></td>
<td>354</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{354} )</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{349} )</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{350} )</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{351} )</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{352} )</td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>&quot;</td>
<td>30–50 ng ( \bar{\varphi}_{353} )</td>
</tr>
<tr>
<td>MSS2, Set 1</td>
<td>F</td>
<td>Whole blood, cotton cloth</td>
<td>0.9(2) μg ( \bar{\varphi}_{N01} )</td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>Semen, cotton cloth</td>
<td>1.2(1) μg ( \bar{\varphi}_{1037} )</td>
</tr>
<tr>
<td></td>
<td>H</td>
<td>Blood &amp; semen, cotton cloth</td>
<td>0.8(2) μg ( \bar{\varphi}_{N01} )</td>
</tr>
<tr>
<td>MSS2, Set 2</td>
<td>J</td>
<td>Whole blood, cotton cloth</td>
<td>1.7(3) μg ( \bar{\varphi}_{N02} )</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>Semen, cotton cloth</td>
<td>0.8(1) μg ( \bar{\varphi}_{1131} )</td>
</tr>
<tr>
<td>MSS2, Set 3</td>
<td>M</td>
<td>Extracted DNA, TE buffer</td>
<td>1.7(3) μg ( \bar{\varphi}_{N01} )</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>&quot;</td>
<td>2.5(3) ng/μL { ( \bar{\varphi}<em>{N03}, \bar{\varphi}</em>{N04} ) }</td>
</tr>
<tr>
<td></td>
<td>O</td>
<td>&quot;</td>
<td>1.0(1) ng/μL { ( \bar{\varphi}<em>{N03}, \bar{\varphi}</em>{N04} ) }</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>&quot;</td>
<td>5.0(5) ng/μL { ( \bar{\varphi}<em>{N03}, \bar{\varphi}</em>{N04} ) }</td>
</tr>
<tr>
<td></td>
<td>Q</td>
<td>&quot;</td>
<td>1.0(1) ng/μL { ( \bar{\varphi}<em>{N03}, \bar{\varphi}</em>{N04} ) }</td>
</tr>
<tr>
<td></td>
<td>L</td>
<td>Blood &amp; semen, cotton cloth</td>
<td>0.5(1) μg ( \bar{\varphi}_{1146} )</td>
</tr>
</tbody>
</table>

Q
Set 3, Buffered DNA Solutions in Sealed Vials—A four-level concentration series was prepared to evaluate the accuracy of each participant’s quantitative DNA measurements. All samples were prepared from a two-source stock solution of extracted DNA in tris-EDTA (TE) buffer (12). Aliquots of this stock solution were individually diluted with TE buffer to produce sample solutions at the desired DNA concentrations. Approximately 30 μL of the continuously-stirred sample solutions were aliquoted into 500 μL limited-volume vials (SARSTEDT, Inc., Newton, NC), sealed under argon, labeled, and stored at −20°C. Table 2 lists the composition of these samples.

The DNA used in the stock solution was an approximately equal blend of material extracted from blood donated by a female and a male source. The mixed DNA was repurified (13). The final purity blend of material extracted from blood donated by a female and a male source. The mixed DNA was repurified (13). The final purity blend of material extracted from blood donated by a female and a male source. The mixed DNA was repurified (13). The final purity blend of material extracted from blood donated by a female and a male source. The mixed DNA was repurified (13). The final purity blend of material extracted from blood donated by a female and a male source. The mixed DNA was repurified (13). The final purity blend of material extracted from blood donated by a female and a male source. The mixed DNA was repurified (13).

AmpF

Promega PowerPlex™ 2.1

Promega FFv 3

BHO Quad 4

Table 3 lists the STR multiplexes used in the MSS1 and MSS2 and the number of participants using each. Many participants used more than one multiplex, particularly in MSS2. More than half of the MSS2 participants reported alleles at all 13 CODIS core loci (14) plus amelogenin; only one MSS2 participant reported alleles for fewer than seven of these core loci. Table 4 lists the instrumentation employed. Several participants in MSS1 used more than one instrument. None of the multiplexes or any of the instruments were systematically associated with any measurement artifact.

Results

Multiplexes and Instrumentation

Table 3 lists the STR multiplexes used in the MSS1 and MSS2 and the number of participants using each. Many participants used more than one multiplex, particularly in MSS2. More than half of the MSS2 participants reported alleles at all 13 CODIS core loci (14) plus amelogenin; only one MSS2 participant reported alleles for fewer than seven of these core loci. Table 4 lists the instrumentation employed. Several participants in MSS1 used more than one instrument. None of the multiplexes or any of the instruments were systematically associated with any measurement artifact.

STR Profiles

Two participants in MSS1 and four in MSS2 incorrectly specified one or more alleles for one or more samples in their initial data submission. One of the MSS1 and three of the MSS2 errors were made copying correct data from a worksheet into the final report, with the one and only error involving multiple alleles arising from a misaligned spreadsheet column. Two participants misassigned one allele each while analyzing their gel images. In all cases, the analysts involved found and corrected the error after reexamination of their primary data. Additionally, five MSS2 participants called to our attention one or more errors we made when transcribing their data into our database—and two errors made in the entry of MSS1 data were identified while preparing this manuscript. Given that MSS1 and MSS2 both required atypical data evaluations and report formats, this does not represent an error rate for casework or other “routine” analyses. Table 5 summarizes profiling performance for all nondifferentially extracted samples (i.e., all MSS1 samples and the single sources samples of MSS2) after correction of all known errors.

Table 4—Instrumentation used by participants.

<table>
<thead>
<tr>
<th>Instruments</th>
<th>MSS1</th>
<th>MSS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABI 310</td>
<td>5</td>
<td>21</td>
</tr>
<tr>
<td>ABI 373</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ABI 377</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>Hitachi FMBio</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>MD FLoImager</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Silver stain</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>25</td>
<td>45</td>
</tr>
</tbody>
</table>

Table 5—Profiling performance for nondifferentially extracted samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Complete*</th>
<th>Partial†</th>
<th>No Result‡</th>
<th>No Signal§</th>
<th>Extral‖</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>349</td>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>350</td>
<td>33</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>351</td>
<td>32</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>34</td>
</tr>
<tr>
<td>352</td>
<td>31</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>353</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>354</td>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>35</td>
</tr>
<tr>
<td>A</td>
<td>23</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>B</td>
<td>36</td>
<td>8</td>
<td></td>
<td>1</td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>C</td>
<td>27</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td>44</td>
</tr>
<tr>
<td>D</td>
<td>42</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>E</td>
<td>8</td>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td>45</td>
</tr>
<tr>
<td>MSS1 Total</td>
<td>336</td>
<td>84</td>
<td>3</td>
<td>7</td>
<td>2</td>
<td>432</td>
</tr>
<tr>
<td>F</td>
<td>41</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>G**</td>
<td>41</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>J</td>
<td>40</td>
<td>2‖</td>
<td></td>
<td></td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>K**</td>
<td>42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>42</td>
</tr>
<tr>
<td>MSS2 Total</td>
<td>164</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>168</td>
</tr>
</tbody>
</table>

* Exact specification plus profiles with extra possible alleles clearly defined as weak (stutter) bands.
† At least one allele not specified for at least one locus.
‡ No result reported for one locus.
§ No result reported for any locus.
‖ One excess allele reported without comment for one locus of profile.
¶ Minor band of three-handed pattern not reported.
** Approximately 15% of the MSS2 participants differentially extracted these samples; only male fraction profiles were reported.
TABLE 6—Profiling performance for differentially extracted samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Exact*</th>
<th>Extra†</th>
<th>Partial‡</th>
<th>No Result§</th>
<th>No Signal‖</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>H(_{\text{sperm}})</td>
<td>30</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>L(_{\text{sperm}})</td>
<td>37</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>H(_{\text{Nonsperm}})</td>
<td>3</td>
<td>35</td>
<td>2</td>
<td>2</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>L(_{\text{Nonsperm}})</td>
<td>4</td>
<td>31</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>42</td>
</tr>
</tbody>
</table>

* Exact specification.
† All true alleles specified plus one or more alleles from incompletely differentiated source.
‡ At least one true allele not specified for at least one locus.
§ No result reported for one or more locus.
‖ No result reported for any locus.

A participant and NIST clerical errors. Table 6 likewise summarizes performance for the sperm and nonsperm components of the two samples requiring differential extraction (i.e., the multiple source samples of MSS2).

A number of MSS1 participants reported multiple profiles. While some are replicate assays by different analysts of the same laboratory, a number of participants did not obtain sufficient signal for one or more STR loci with their initial extractions. All but two of these participants obtained sufficient signal after reamplification, reextraction of their original samples, or complete reanalysis of a second set of MSS1 samples.

Only true alleles were reported for all single source samples. A number of participants in both studies reported extra alleles but appropriately labeled them as “stutter” or weak minor-component alleles. Two participants reported a “stutter” allele for one or another of the multiple-source samples without explicitly noting that the peak was of relatively low intensity or otherwise unlikely to be a true allele. Two participants in MSS2 did not report a true minor allele of a known three-banded pattern. About half of the MSS1 participants did not explicitly report all alleles at all loci for at least one of the two-source samples and most did not report all alleles for the three-source sample.

The majority of participants reported only male-source alleles for the sperm-fraction of the differential extracts of the MSS2 multiple-source samples. Several participants reported minor alleles attributable to the female source for one or both samples. Several participants failed to observe some true alleles in the sample having two male-sources. Nearly all participants reported male-source alleles in the nonsperm fractions of both multiple-source samples. (We attribute the strong male-source signal in the nonsperm fraction to sperm lysis prior to sample preparation.) A number of participants reported that male-source contamination of the nonsperm fraction is not typical of casework samples, with further note that casework sperm-fractions are seldom as free of female-source contamination as the MSS2 samples.

The reported signals ranged continuously from essentially all above baseline events to just major component electropherogram peaks or gel-image bands. The majority of participants who did report some “probably stutter” signals used a variety of methods to indicate relative intensities: listing order, nested parentheses, and/or a wide selection of footnotes.

Identification of Known Sources

Table 7 summarizes source identification results for the five multiple-source MSS1 samples. All identified sources were true contributors. Due to incomplete profiling of the multiple-source samples, one participant chose not to specify one of the sources in one two-source sample and five participants chose not to specify one source in the three-source sample. Due to a single incomplete reference profile, one participant chose not to specify the unrecognized source in two of the samples although all other samples were successfully excluded as possible sources. Two participants chose not to attempt source identification, one due to multiple incomplete reference profiles and one due to inexperience with casework samples.

Most of the MSS1 profiling difficulties were encountered with the only male source of the six used to prepare the samples. Since many participants reported difficulty in obtaining good signal for this source’s reference stain, it is probable that less than the target amount of this DNA was actually delivered to the stains involved.

A number of participants who identified all sources in the three-source sample noted that the three nonexcludable source profiles did not fully account for all features of the multiple source profile. Many of the differences between the expected and the observed multiple-source profiles are attributable to incomplete profiles for the multiple-source samples. However, several participants noted among-loci inconsistencies in the relative intensities of the allelic signals.

Specification of Unknown Sources

MSS2 participants were asked to provide “CODIS profile(s) to search for the suspect(s)” for the two multiple-source samples. The responses to this request were quite diverse, with many participants stating that they had little or no experience with CODIS. We have grouped the responses into the categories summarized in Table 8. These categories do not necessarily reflect CODIS nomenclature or practice.

About 25% of the participants chose not to profile the unknown male source in either sample, explaining that: 1) their laboratory did not perform this type of analysis; 2) it was against their laboratory’s policy to profile a source in the absence of a reference; or 3) they were inexperienced in this type of analysis. A number of participants who did specify a profile for one or both unknown sources also noted that it was against their laboratory’s policy to perform this type of analysis on casework samples.

All participants who profiled the unknown male in the two-source (known female, unknown male) sample specified all true
alleles; all but one participant specified only true alleles. The one participant who did not specify the exact profile for the unknown source explicitly identified all alleles unique to the unknown-source and implicitly identified all possible alleles for the loci with less than two unique alleles. One participant did not achieve a sufficiently good signal for the non-sperm fraction of the sample to confirm that the reference female was a true source and chose not to proceed with the analysis.

Specification of the unknown male in the three-source (known female, known male, unknown male) sample proved more problematic. Only four participants provided the exact profile. Eleven participants specified exact alleles at most loci and narrowly defined the possible allele combinations at the other loci. Eight participants chose to specify most-to-all signals observed in the sperm-fraction. Five participants found the problem too complex and chose not to proceed with the analysis. Several participants who specified large numbers of possible multiple allele pairs for one or more loci did not exhaustively specify all of the possibilities implied by the pairs that were specified.

Four participants did not specify all true alleles for the unknown profile. Two heterozygous loci were specified as homozygous and one homozygous locus was explicitly specified as heterozygous. None of these miss-specifications is attributable to unusual peak shape or intensities. Both alleles at one heterozygous locus were mis-specified due to stochastic “dropout” of one of the two alleles contributed by the unknown male. This event occurred in a sperm-fraction characterized by generally low-intensity signals at a locus with four male-source alleles. While present and not a possible “stutter” peak, the height of the “missing” allele peak was less than 25% of that expected and was very similar to the heights of true “stutter” at other loci.

A number of participants who specified a less than exact profile for the three-source sample, and at least one of the participants who chose not to attempt specification, noted among-loci inconsistencies in the relative intensities of the allelic signals. These inconsistent allelic signal intensity ratios for the different loci, also noted by a few MSS1 participants, were not related to any specific STR multiplex or manufacturer.

**DNA Concentration Estimates**

Figure 1 presents both the consensus and individual participant results reported for the MSS2 Set 3 extracted-DNA samples. The DNA concentrations in this four-level series ranged from 5.0 ng DNA per μL solution to 0.5 ng/μL. The results for all levels are well described as lognormal distributions. Given both the 10-fold span of concentration among the samples and the lognormal distribution of results at each level, all calculations have been performed on logarithmically transformed (Y = log10X) concentrations. All summary statistics have been back-transformed (X = 10Y) to report the results in units of concentration. Measures of location, like the median, are qualitatively unchanged by this manipulation. However, measures of dispersion, like the robust standard deviation (SD), change from symmetrically additive terms to symmetrically multiplicative factors. That is, about 68% of normally distributed concentrations are expected to be in the interval (median ± SD ≤ median ± median + SD); 68% of lognormally distributed concentrations are expected to be within the interval (median/SD ≤ median ± median × SD) (15).

There is excellent agreement between the nominal concentration (what we believe went into the tubes) and the median of the estimated concentrations (a robust consensus estimate of what came out (16)) for the highest three levels. At the lowest level of the series, the median is about 50% of the nominal; this difference may be due to DNA binding to the sample tube.

The concentration series reported by individual participants are typically collinear with the consensus values (i.e., straight lines with different intercepts but generally unit slope). The majority of participants reported concentrations clustered close to the consensus values, with a robust estimate for the SD being a factor of 1.8 for all four levels (16). Three participants reported quite similar values that are about five-fold higher than consensus, four reported values 4- to 40-fold lower than consensus for two or more levels, and three reported both very high and very low values. We use the term “concordance” to characterize the average difference of measurements from the consensus values; those measurement series that are consistently higher or lower than the consensus values are thus very positively or negatively discordant (17,18). We use the term “apparent precision” to characterize the SD among the differences; those measurement series that are inconsistently higher and lower than the consensus values are thus very “apparently imprecise” (18).

Figure 2 is a Youden plot (19) detailing all results reported for the duplicate samples. As with the Fig. 1 concentration series, participants who reported one very high or very low value tend to be consistently high or low. There is, however, little correlation between duplicate results for the concordant majority of participants.

The robust SD for these samples is the same factor of 1.8. Since these two samples are true independently analyzed duplicates, this factor of 1.8 SD represents the intrinsic precision of the DNA concentration measurements. That is, among-participant apparent precision is as good as possible given the current measurement technologies.
Figure 3 is a “target” plot (18) displaying the concordance and apparent precision characteristics of all MSS2 Set 3 sample measurements for each participant. The innermost “ring” of the target represents a combined discordance and apparent imprecision of one SD (here, a factor of 1.8) from complete agreement with the consensus values. The middle ring likewise represents the two SD
Total DNA Quantity Estimates

The total recoverable DNA quantities reported for the six MSS2 stains are presented in Fig. 4, a Youden variant plotting results for the blood, semen, and multiple-source stains of Set 1 against their analogues in Set 2. While different extraction protocols were typically used for blood and multiple-source stains (most, but not all, participants extracted the semen stains in the same manner as they did the blood stains), all participants used the same protocols for the Set 2 samples as they did for the Set 1 samples. Since all samples were prepared, randomized, and packaged independently, the strong correlation between the Set 1 and Set 2 estimated DNA quantities is a function of the participants and not of sample preparation.

The quantity estimates for the blood stains averaged about 30% of the nominal amount that we believe was actually present; the semen stains averaged about 20% of nominal; and the mixed-source stains about 10%. Three participants re-extracted the multiple-source stain matrix (cotton cloth) after differential extraction and estimated the residual quantity of DNA. All three reported about as much DNA in the residual as in the combined sperm and nonsperm fractions.

The variation in the estimated quantity of DNA in the stains, ranging from an average factor of 2.6 for the blood reference samples to a factor of 3.2 for the multiple-source stains, is much larger than the factor of 1.8 characteristic of DNA concentration measurements. Some of this variation may be attributed to stain subsampling. A few

\[ (1.8^2 = 3.2) \text{ and the outermost ring the three SD (} 1.8^3 = 5.8 \text{) disagreement factors. Roughly, 95\% of all participants with measurement characteristics “similar to the consensus” should plot within the two SD ring.} \]
participants directly estimated DNA quantities from the extracts of entire stains. However, most participants extracted a fraction of the stain and adjusted their estimate by the relative area of the subsample to the entire stain. Unless care is taken to proportionally sample all regions of the stain, this adjustment is valid only if the DNA is uniformly distributed over the entire stain. Given that leukocytes, intact sperm, and free DNA doubtless have different affinities for the cloth matrix, they may differentially concentrate at the stain center or the stain edge. Participants who subsampled from the center of each stain may well have sampled less representatively than those who divided the stain into halves or quarters.

Comparison of DNA Measurement Methods

The majority of MSS2 participants used one of two commercial human DNA quantitation kits with either colorimetric or chemiluminescence detection. The top section of Fig. 5 displays the concordance and apparent precision characteristics for these participants for both the Set 3 DNA concentration and the Set 1 and 2 DNA quantity measurements. There is no clear measurement-performance difference between the kits or between the detection methods.

Comparison of DNA Extraction and Concentration Methods

The lower section of Fig. 5 likewise displays the concordance and apparent precision characteristics for the participants that do not use a commercial DNA quantification kit. While the Set 3 results for one participant using an in-house method are extremely discordant, they are very similar to the results reported by two participants using commercial kits. Since the Sets 1 and 2 results for this participant are quite concordant, we believe that the extreme Set 3 discordance is analyst—rather than method—related. There is no other clear measurement-performance difference among the in-house methods or between the in-house and commercial methods.

Influence of Measurement Discordance on STR Profiles

Thirty-seven participants reported both Set 3 DNA concentrations and Sets 1 and 2 DNA quantities. Figure 7 displays the characteristic concordances for these participants, again as a Youden variant plotting the Sets 1 and 2 total DNA quantity concordance as a function of the Set 3 DNA concentration concordance. There is no systematic relationship between the two measurement characteristics. However, the one participant with highly positive con-
centation and quantity discordance is the only laboratory experi-
encing allele dropout. This participant’s consistent five-fold over-
estimation of available DNA may have resulted in too-little DNA
being used in the amplification reaction mixture.

Most of the participants who reported amplification failure at
one or more loci reported much less than the consensus quantity of
DNA in the Set 1 and Set 2 samples. If these highly negative quant-
ty discordances truly represent very inefficient DNA extractions,
the failures may reflect too little DNA being available in the am-
plification reaction. Alternatively, if the low quantitation results
are erroneous and the DNA extractions were “typically” efficient,
the failures may reflect too much DNA being present in the initial
reaction mixture for optimal amplification.

Discussion and Recommendations

The MSS1 and MSS2 interlaboratory comparisons challenged
multiple STR measurement systems with intentionally difficult
samples. About 10% of the participants in both studies experienced
measurement failures, misinterpreted some part of their analytical
signals, or made clerical errors. Most of these problems occurred
with the most challenging samples of each exercise—those con-
taining DNA from three sources. Furthermore, few participants
specified unique profiles for all three components of these complex
samples.

While none of the measurement anomalies in either exercise re-
sulted in a false identification of a reference source, measurement
difficulties waste laboratory resources. While few of the anomalies
reported could result in a false exclusion if they occurred in cas-
ework, database matching is most efficient when the target profile is
uniquely defined and complete. Thus, while the amplification and
detection stages of STR multiplex DNA profiling appear robust,
we believe that the efficiency of the overall measurement systems
can be further improved. The following recommendations address
the potential measurement difficulties suggested by the MSS1 and
MSS2 results.

Clerical Errors

The number of clerical errors reported in the MSS1 and MSS2
challenge studies is larger than we expect for “routine” information
transfers. However, by definition, any clerical error represents a
failure (or circumvention) of the laboratory’s internal technical re-
view process. Clerical errors in specialized reports prepared for in-
terlaboratory comparisons should be regarded as an early warning
sign of a laboratory’s need to review its review process (or the need
to apply it to all external reports).

Transcription errors can be minimized by direct electronic trans-
fer of single source profiles from instrument through local database
to final report. This can be best achieved by insisting that such link-
ages among information systems be a design goal for all forensic
software.

Independent data analysis by two or more analysts followed by
rigorous comparison and immediate resolution of differences al-
though time consuming can minimize interpretation as well as tran-
scription errors. Use of standardized report templates for casework
scenarios would minimize the clerical perils intrinsic to the pro-
duction of “once only” reports. Rigorous internal technical review
of all data, interpretation, and final reports as is done routinely for
casework will minimize the number of errors.

Turnabout—Clerical errors by interlaboratory study providers
must also be minimized, using the same tools as above. To better
approach the desired zero error rate, the NIST analyst-authors re-
cently switched from review of each other’s results to completely
independent analysis of the raw data.

Whenever information must be manually transcribed, trans-
formed, and/or interpreted before analysis, it is essential that the
analysts responsible for the original data have the opportunity to re-
view their modified results.

Terminology and Communication Issues

To enable assessment of current forensic STR reporting practice,
the MSS1 and MSS2 exercises did not specify that the requested in-
formation be provided in any particular format. While all reports
were quite interpretable, given the solid experimental background
of most of the authors, the terminology and style of reports were
unexpectedly diverse. This diversity includes descriptions of ana-
lytical techniques, materials and equipment, differential extraction
fractions, stutter and minor-component peaks, homozygous loci,
and nonuniquely determined profiles.

Information exchange among all members of the forensic com-
munity (especially those without a strong technical background)
would be facilitated if the forensic community adopted consensus
terms and symbols. Consensus development of report templates ap-
propriate for the different casework scenarios could also facilitate
communication and potentially lower the clerical error rate for less-
common situations.

Turnabout—Development of consensus standards and report
templates can only be accomplished through collaborative effort of
the (already overburdened) forensic analysts directly involved.
Should the community start such a project, interlaboratory studies
could be used to help evaluate and refine the proposed communi-
cation tools.

Extraction, Quantitation, and Amplification Issues

Several participants did not obtain complete amplification of all
components of all samples. These analytical failures are almost all
associated with difficulties in preamplification stages of the STR
measurement process: sample extraction, DNA concentration, and
DNA quantitation. With one exception, the MSS2 study results
indicate that these difficulties are not attributable to specific tech-
niques or methods but rather reflect possible variability within the
quantitation systems and/or analyst experience and ability. Chelex
extraction systems appear to be less efficient for mixed-stains than
the organic systems.

The participants’ estimates of DNA quantity were quite discor-
dant for many of the MSS2 stains. However, nearly all partici-
pants reported the same relative discordance for the three {Set 1,
Set 2} pairs of stains: blood, semen, and mixed. The between-set
uniformity of within-participant results suggests large and consis-
tent among-participant differences in DNA recovery (extraction
and/or concentration) efficiency. No DNA extraction or concen-
tration method was consistently associated with either higher or
lower results. The few participants who re-extracted the mixed
stains after differential extraction reported significant amounts
of DNA in the second extract (20). We speculate that some of the
recovery efficiency differences are related to sample digestion
conditions, particularly agitation (“poke it with a toothpick”) and
duration (“long enough”). Since nearly all participants who
reported total or multiple-locus amplification failure for any sam-
ple also reported very low DNA recovery, analysts experiencing
amplification failures with pristine samples (such as used in MSS1 and MSS2) should re-evaluate their sample preparation techniques.

Some of the recovery differences can also be attributed to how the stains were subsampled. At least for interlaboratory comparisons, stains should be divided into roughly equal “pie slices.” This will help ensure that all pieces of the sample contain about the same quantity and distribution of tissues; it will also greatly simplify estimating the quantity of DNA/stain.

The consensus DNA concentrations for the MSS2 Set 3 samples are nearly identical to the nominal values for the 1 nL, 2.5 nL, and 5 ng/μL samples. No DNA quantitation method consistently produced results very different from the consensus result. The SDs of the intra- and all interlaboratory DNA quantitation results are of nearly identical magnitude, a factor of 1.8 about the consensus results. Thus, all of the current quantitation methods appear to be capable of providing very comparable and accurate results, with expected bias of zero and precision of about a factor of two. The close agreement between the intra- and interlaboratory SD factors suggests that this 2-fold imprecision is intrinsic to the process and cannot be easily improved.

While the majority were in remarkable agreement, more than 10% of participants consistently reported DNA concentration results much higher or much lower than consensus. Few of these participants reported similarly discordant Sets 1 and 2 DNA quantities. Many of the anomalous Set 3 results may not indicate routinely biased DNA quantitation but rather pipetting or dilution errors specific to these unfamiliar samples. However, the only serious amplification anomaly reported in either MSS1 or MSS2 was experienced by a participant who consistently reported 4 to 10 fold more DNA than consensus. We believe that this resulted in too little DNA being added to the reaction mixture for reliable amplification of all alleles from all sources.

**Turnabout**—There are no control or reference materials currently available for documenting forensic DNA quantitation performance or evaluating DNA extraction and recovery efficiencies. Since these pre-amplification measurements impact the over-all performance of STR measurement systems, this lack should be rectified.

The consensus result for the nominal 0.5 ng/μL MSS2 Set 3 quantitation sample is 2-fold low. This suggests a sample preparation or storage problem for low-DNA-concentration samples that must be solved before producing any further DNA quantitation materials.

Several participants noted that the MSS2 multiple-source samples provided “cleaner” sperm fractions and “less clean” nonsperm fractions than typical of casework. We speculate that the lack of female-source contamination of the sperm fraction is due to our use of fresh whole blood rather than epithelial cells. We speculate that the male-source contamination of the nonsperm fraction is due to our use of over-age fertility-clinic semen. Since the nonsperm fraction profile primarily provides confirmation of the female-source, male-source contamination of this profile may be unrealistic but of little forensic consequence.

The MSS2 established that DNA extraction efficiency and quantitation accuracy do at least qualitatively (signal/no signal) affect “real life” STR measurement systems, even with pristine samples. There should be quantitative relationships among all stages of the measurement process from the DNA amount to the allelic signal intensities; it should be possible to design a study to evaluate these linkages.

**Policy Issues**

As noted above, there was considerable variation in how stutter peaks were described. More importantly, there was considerable variation in whether stutter peaks were described. While stutter is undesirable, it is frequently observed and may not be avoidable. We believe that peaks unambiguously attributable to stutter do not require notation in the final report. If the interpretation is ambiguous and stutter cannot be excluded as the source of a peak (that is, the peak could be either a true allele or a stutter peak) that peak should be reported, if reported at all, only as a potential allele. A notation that distinguishes between true and potential alleles would facilitate communication and help ensure consideration of all relevant possibilities. In any case, we believe some consensus policy on the evaluation and reporting of stutter would benefit the entire forensic community.

Again as noted above, we ascribe most observed differential allele amplification to inefficient DNA extraction and inaccurate quantitation (competitive kinetics with too much DNA in the reaction mixture, stochastic dropout with too little). However, we occasionally encounter profiles having unbalanced allelic signals at some loci, up to and including complete absence of one (“null”) allele. While of little importance when evaluating multiple-source profiles against reference profiles, it could well complicate specification of an “unknown” profile. When multiple interpretations cannot be excluded, “moderate stringency” profiles—unique where possible and explicitly describing the ambiguities where required—should be specified.

The _raison d’être_ of any forensic DNA database system is identification of leads to perpetrator identity from evidence containing DNA that is not accounted for otherwise. Given the utility of such information, a surprisingly large fraction of MSS2 participants noted that specifying a profile in the absence of a reference profile was against their laboratory’s policy. To ensure that such evidence can be recognized and efficiently exploited when needed, the evaluation of unknown-source profiles and use of DNA database match systems should become a routine component of forensic training and competency evaluation.

**Turnabout**—Complete and unique specification of an “unknown” profile will not always be possible. A comparative evaluation of how current DNA database matching systems process incomplete and ambiguous profiles could enable more effective use of these forensic tools.

**Conclusion**

The MSS1 and MSS2 interlaboratory comparisons challenged multiplex STR measurement systems with difficult samples representing unusual forensic scenarios. These studies were explicitly designed to elicit measurement problems related to unbalanced amplification of DNA from multiple-source samples. None of the relatively few analysis problems encountered can be attributed to abnormal STR multiplex performance; all DNA amplification anomalies reported are associated with inefficient DNA extraction, inaccurate DNA quantitation, and/or analytical threshold policies. Given an appropriate total amount of DNA in the reaction mixture, current STR multiplex systems reliably amplify multiple-source DNA.

---

The MSS2 established that DNA extraction efficiency and quantitation accuracy do at least qualitatively (signal/no signal) affect “real life” STR measurement systems, even with pristine samples. There should be quantitative relationships among all stages of the measurement process from the DNA amount to the allelic signal intensities; it should be possible to design a study to evaluate these linkages.

**Policy Issues**

As noted above, there was considerable variation in how stutter peaks were described. More importantly, there was considerable variation in whether stutter peaks were described. While stutter is undesirable, it is frequently observed and may not be avoidable. We believe that peaks unambiguously attributable to stutter do not require notation in the final report. If the interpretation is ambiguous and stutter cannot be excluded as the source of a peak (that is, the peak could be either a true allele or a stutter peak) that peak should be reported, if reported at all, only as a potential allele. A notation that distinguishes between true and potential alleles would facilitate communication and help ensure consideration of all relevant possibilities. In any case, we believe some consensus policy on the evaluation and reporting of stutter would benefit the entire forensic community.

Again as noted above, we ascribe most observed differential allele amplification to inefficient DNA extraction and inaccurate quantitation (competitive kinetics with too much DNA in the reaction mixture, stochastic dropout with too little). However, we occasionally encounter profiles having unbalanced allelic signals at some loci, up to and including complete absence of one (“null”) allele. While of little importance when evaluating multiple-source profiles against reference profiles, it could well complicate specification of an “unknown” profile. When multiple interpretations cannot be excluded, “moderate stringency” profiles—unique where possible and explicitly describing the ambiguities where required—should be specified.

The _raison d’être_ of any forensic DNA database system is identification of leads to perpetrator identity from evidence containing DNA that is not accounted for otherwise. Given the utility of such information, a surprisingly large fraction of MSS2 participants noted that specifying a profile in the absence of a reference profile was against their laboratory’s policy. To ensure that such evidence can be recognized and efficiently exploited when needed, the evaluation of unknown-source profiles and use of DNA database match systems should become a routine component of forensic training and competency evaluation.

**Turnabout**—Complete and unique specification of an “unknown” profile will not always be possible. A comparative evaluation of how current DNA database matching systems process incomplete and ambiguous profiles could enable more effective use of these forensic tools.

**Conclusion**

The MSS1 and MSS2 interlaboratory comparisons challenged multiplex STR measurement systems with difficult samples representing unusual forensic scenarios. These studies were explicitly designed to elicit measurement problems related to unbalanced amplification of DNA from multiple-source samples. None of the relatively few analysis problems encountered can be attributed to abnormal STR multiplex performance; all DNA amplification anomalies reported are associated with inefficient DNA extraction, inaccurate DNA quantitation, and/or analytical threshold policies. Given an appropriate total amount of DNA in the reaction mixture, current STR multiplex systems reliably amplify multiple-source DNA.
Acknowledgments

Interlaboratory studies are made possible by the cooperation of many analysts and laboratory supervisors. We thank them for sharing with us their time and resources, and for their willingness to tackle our somewhat contrived samples. Their voluntary and open participation in challenge exercises speaks as much to pursuit of analytical truth as to confidence in analytical systems.

References


Additional information and reprint requests:
Margaret C. Kline
NIST
100 Bureau Drive, Stop 8311
Gaithersburg, MD 20899-8311
Margaret.Kline@nist.gov