Anal. Chem. **2004,** *76,* 6928−6934

NIST Mixed Stain Study 3: Signal Intensity Balance in Commercial Short Tandem Repeat Multiplexes

David L. Duewer,* Margaret C. Kline, Janette W. Redman, and John M. Butler

Chemical Science and Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8394

Short-tandem repeat (STR) allelic intensities were collected from more than 60 forensic laboratories for a suite of seven samples as part of the National Institute of Standards and Technology−coordinated 2001 Mixed Stain Study 3 (MSS3). These interlaboratory challenge data illuminate the relative importance of intrinsic and user-determined factors affecting the locus-to-locus balance of signal intensities for currently used STR multiplexes. To varying degrees, seven of the eight commercially produced multiplexes used by MSS3 participants displayed very similar patterns of intensity differences among the different loci probed by the multiplexes for all samples, in the hands of multiple analysts, with a variety of supplies and instruments. These systematic differences reflect intrinsic properties of the individual multiplexes, not user-controllable measurement practices. To the extent that quality systems specify minimum and maximum absolute intensities for data acceptability and data interpretation schema require among-locus balance, these intrinsic intensity differences may decrease the utility of multiplex results and surely increase the cost of analysis.

The National Institute of Standards and Technology (NIST) periodically conducts interlaboratory challenge exercises to characterize the field performance of multiplexed short-tandem repeat (STR) assays designed for human identification.[1−3] The most recent of these, the 2001 Mixed Stain Study 3 (MSS3), was designed to explore the relationship between DNA quantitation and STR signal intensity.[3] Some of the data collected for this study also illuminate the relative importance of intrinsic and user-determined factors affecting the locus-to-locus balance in signal intensities for currently used STR multiplexes.

Forensic laboratories require signal intensities for probative samples to be above some threshold but below some saturation level.[4] Achieving approximate among-locus signal balance (i.e., having the same average signal intensity at all loci probed) within these bounds helps to ensure and demonstrate assay quality. Effective among-locus balance also facilitates the automation of STR interpretation, especially for multiple-source samples.[5,6]

However, the relative intensity of alleles observed at different genetic loci is a complex function of intrinsic, sample-specific, and user-determined factors.[7−9] Intrinsic factors include (1) the nature of the dyes used to facilitate detection of the polymerase chain reaction (PCR) products and (2) design considerations influencing the relative efficiencies of PCR amplification such as the nature of the primer binding sites, size of the product, and primer concentrations.[9,10] Sample-specific factors include (1) degree of sample degradation, (2) presence of PCR inhibitors, and (3) possible mutations in the primer-binding regions.[9,11,12] Factors under the user's influence or control include (1) the selection and maintenance of instrumentation, (2) PCR amplification parameters such as the amount of template DNA, type of polymerase, amount and volume of reactants, and number and time/temperature profile of amplification cycles, (3) electrophoretic separation parameters influencing the resolution of the PCR products such as amount of sample loaded, electric field applied, and temperature control detection of the separated products, and (4) data analysis parameters related to resolving the individual signals from the different dyes used in the STR multiplex from the observed total signal.

The following sections describe the relevant MSS3 data and how it has been processed, the observed among-loci relative intensities for all of the STR multiplexes used by participants in the MSS3, and the relative importance of user-influenced and

* Corresponding author. Tel: (301)-975-3935. Fax: (301)-977-0587. E-mail: david.duewer@nist.gov.

(1) Kline, M. C.; Duewer, D. L.; Newall, P.; Redman, J. W.; Reeder, D. J. J. *Forensic Sci.* **1997**, *42* (5), 897−906.
(2) Duewer, D. L.; Kline, M. C.; Redman, J. W.; Newall, P. J.; Reeder, D. J. J. *Forensic Sci.* **2001**, *46* (5), 1199−1210.
(3) Kline, M. C.; Duewer, D. L.; Redman, J. W.; Butler, J. M. *Anal. Chem.* **2003**, *75*, 2463−2469.
(4) Scientific Working Group on DNA Analysis Methods (SWGDAM). *Forensic Sci. Commun.* **2000**, *2* (3), 1-4 (http://www.fbi.gov/hq/lab/fsc/backissu/july2000/strig.htm).
(5) Kimpton, C.; Fisher, D.; Watson, S.; Adams, M.; Urquhart, A.; Lygo, J.; Gill, P. *Int. J. Legal Med.* **1994**, *106* (6), 302−311.
(6) Gill, P.; Kimpton, C. P.; Urquhart, A.; Oldroyd, N.; Millican, E. S.; Watson, S. K.; Downes, T. J. *Electrophoresis* **1995**, *16* (9), 1543−1552.
(7) Butler, J. M. *Forensic DNA Typing: Biology and Technology behind STR Markers*; Academic Press: London, 2001.
(8) Gill, P.; Sparkes, R.; Kimpton, C. *Forensic Sci. Int.* **1997**, *89* (3), 185−197.
(9) Wallin, J. M.; Holt, C. L.; Lazaruk, K. D.; Nguyen, T. H.; Walsh, P. S. J. *Forensic Sci.* **2002**, *47* (1), 52−65.
(10) Krenke, B. E.; Tereba, A.; Anderson, S. J.; Buel, E.; Culhane, S.; Finis, C. J.; Tomsey, C. S.; Zachetti, J. M.; Masibay, A.; Rabbach, D. R.; Amiott, E. A.; Sprecher, C. J. *J. Forensic Sci.* **2002**, *47* (4), 773−785.
(11) Lazaruk, K.; Wallin, J.; Holt, C.; Nguyen, T.; Walsh, P. S. *Forensic Sci. Int.* **2001**, *119* (1), 1−10.
(12) Leibelt, C.; Budowle, B.; Collins, P.; Daoudi, Y.; Moretti, T.; Nunn, G.; Reeder, D.; Roby, R. *Forensic Sci Int.* **2003**, *133* (3), 220−227.

10.1021/ac049178k
Published on Web 10/30/2004

Table 1. Number of Complete (7 Samples) Data Sets by Multiplex and Electrophoretic Instrumentation

multiplex	producer[a]	code	data sets/instrument 310[b]	377[c]	FMBIO[d]
AmpF*l*STR COfiler	ABI	Cof	42	7	
AmpF*l*STR Profiler Plus	ABI	Pro+	44	10	
AmpF*l*STR Identifiler	ABI	Idf	1		
AmpF*l*STR SGM Plus	ABI	SGM+		2	
PowerPlex 1.1	Promega	PP1.1			1
PowerPlex 1.1 + Amelogenin	Promega	PP1.1+			4
PowerPlex 2.1	Promega	PP2.1			3
PowerPlex 16	Promega	PP16	4	1	
total			91	20	8

[a] ABI, Applied Biosystems, Inc., Foster City, CA; Promega, Promega Corp., Madison, WI. [b] ABI PRISM 310 Genetic Analyzer. [c] ABI PRISM 377 Genetic Sequencer. [d] Hitachi FMBIO II Fluorescence Imaging System (MiraiBio Inc., Alameda, CA).

intrinsic factors affecting the among-locus signal intensity balance. More than 20,000 allelic heights for 833 STR multiplex analyses (129 complete sets of one control and 6 multiple-source DNA extracts) were assembled from 62 different participants in the MSS3. Each participant evaluated the samples with one or more STR multiplexes of their own choosing and provision. Sample-specific differences were controlled by the analysis of a common set of multiple-source DNA extracts. Comparison of within- and among-participant variation in the relative among-loci intensity balance identifies factors affected by short-term changes in the user's technique and instrumentation. Variation in the average intensity balance among participants using the same multiplex and similar instrumentation identifies factors affected by longer term differences. Regularities in the average balance among participants identify factors that are intrinsic to the realizations of the STR multiplexes used in the MSS3.

MATERIALS AND METHODS

Samples. A complete description of sample design and preparation is provided in ref 3. Briefly, seven different sets of human DNA extracts in tris−EDTA buffer were prepared at NIST. A single-donor extract, labeled "R", was prepared to have a DNA concentration ([DNA]) of 1 ng/μL. Participants were requested to evaluate this sample at the beginning and end of every set of analyses performed for the MSS3. Five two-donor and one three-donor samples, labeled "S" through "X", were prepared to have total [DNA] of from 1 to 4 ng/μL. The major/minor donor ratios in the two-donor samples were designed to fall in the range of 3:1−10:1; the three-donor sample was designed to have a major/minor$_1$/minor$_2$ ratio of 4:2:1.

Participants. A complete list of participating laboratories is provided in ref 3. Briefly, samples were distributed to 83 volunteer forensic laboratories. Seventy-four laboratories participated in the study by providing results for one or more of the study goals. Sixty-two of these participants reported quantitative allelic intensity data for all seven samples.

Multiplexes and Instrumentation. Table 1 lists the STR multiplexes and electrophoretic instrumentation used by MSS3 participants, along with the number of sets of allelic intensities reported for each unique combination of multiplex and instrumentation. Many participants reported intensities for two or more multiplexes. Table 2 lists the genetic loci evaluated in the various multiplexes and the fluorescent dye ("color") used to visualize the amplification products in each of the multiplexes.

Table 2. STR Multiplex Composition[a]

		Applied Biosystems, Inc.				Promega Corporation			
locus	code	Cof	Pro+	SGM+	Idf	PP1.1	PP1.1+	PP2.1	PP16
D3S1358[b]	D3	5-FAM	5-FAM	5-FAM	VIC			FL	FL
D5S818[b]	D5		NED		PET	FL	FL		JOE
D7S820[b]	D7	NED	NED		6-FAM	FL	FL		JOE
D8S1179[b]	D8		JOE	JOE	6-FAM			TMR	TMR
D13S317[b]	D13		NED		VIC	FL	FL		JOE
D16S539[b]	D16	5-FAM		5-FAM	VIC	FL	FL		JOE
D18S51[b]	D18		JOE	JOE	NED			FL	FL
D21S11[b]	D21		JOE	JOE	6-FAM			FL	FL
CSF1PO[b]	CSF	JOE			6-FAM	TMR	TMR		JOE
FGA[b]	FGA		5-FAM	NED	PET			TMR	TMR
TH01[b]	TH0	JOE		NED	VIC	TMR	TMR	FL	FL
TPOX[b]	TPO	JOE			NED	TMR	TMR	TMR	TMR
vWA[b]	vWA		5-FAM	5-FAM	NED	TMR	TMR	TMR	TMR
D2S1338	D2			5-FAM	VIC				
D19S433	D19			NED	NED				
Penta D	PnD								JOE
Penta E	PnE							FL	FL
Amelogenin[c]	AM	JOE	JOE	JOE	PET		TMR		TMR
no. loci		7	10	11	16	8	9	9	16

[a] Locus-specific primer labels are nominally detected as blue, green, yellow, and red. Dyes used: (blue) 5-FAM, G-FAM, and FL; (green) JOE and VIC; (yellow) NED and TMR; and (red) PET. See ref 7 for further information. [b] CODIS core locus. See ref 7 for further information. [c] A segment of the X−Y homologous gene used for gender identification.

Allelic Intensities. Participants were requested to report the intensities of all identified donor alleles. The nature of the intensity metric and the report format was purposely left unspecified to better survey actual forensic practice.

The proprietary ABI GeneScan analysis software used to construct and evaluate sample electropherograms from the intensity versus time data of capillary electrophoretic systems provides either or both the maximum signal (height) and total integrated signal (area) of an identified allelic peak. These metrics are normally expressed as relative fluorescence units. While peak areas have been reported to be more robust,[13,14] a large majority of the participants using these systems reported intensities only as allelic heights. Similarly, most of the participants using Hitachi or MiraiBio FMBIO image analysis software to evaluate intensity versus migration distance data of slab gel systems reported only the maximum optical density of the recognized allelic bands although one such participant provided integrated optical densities.

RESULTS AND DISCUSSION

Data Preprocessing. *Combining Signals from Alleles at a Given Locus.* DNA from genetically different donors yields different STR profiles. Evaluation of signal intensities of DNA from different donors must therefore use some abstraction of the measurements rather than direct same-allele comparison.

We symbolize the allelic heights for a given allele i at genetic locus j of replicate analysis k of sample l reported by participant

(13) Evett, I. W.; Gill, P. D.; Lambert, J. A. *J. Forensic Sci.* **1998**, *43* (1), 62−69.
(14) Gill, P.; Sparkes, R.; Pinchin, R.; Clayton, T.; Whitaker, J.; Buckleton, J. *Forensic Sci. Int.* **1998**, *91* (1), 41−53.

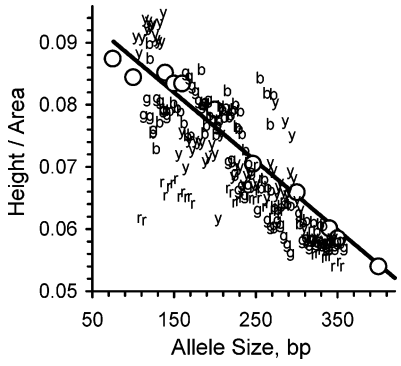

Figure 1. Peak height/area as a function of allele size. Each symbol denotes the height/area for one element of a routine AmpF*l*STR Identifiler allelic ladder assessment performed at NIST. Open circles represent the values observed for the included sizing ladder components. The line is a linear best-fit to these data. The symbols b, g, y, and r represent the values observed for the blue, green, yellow, and red allelic ladders, respectively.

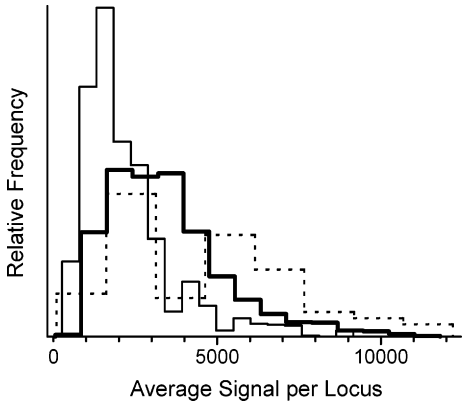

Figure 2. Signal intensity probability distributions. The histograms summarize the area-normalized PDFs of the average signal per locus reported for any STR multiplex using an ABI Prism 310 Genetic Analyzer (dark solid line), ABI Prism 377 DNA Sequencer (light solid line), or Hitachi FMBIO II Fluorescence Imaging System (dashed line). The signal for both ABI instruments is the electopherographic allelic heights in relative fluorescence units. The signal for the Hitachi FMBIO is the slab gel allelic optical density.

m as h_{ijklm}. A sample-independent metric for a given genetic locus should be proportional to the integral of all signals attributable to the sample. Since most participants reported only heights and only for identified donor alleles, we characterize the signal for a given genetic locus of a given analysis as the sum of the heights reported for the I alleles of the given sample: $h_{jklm} = \sum_i^I h_{ijklm}$.

Systematic Biases of the h_{jklm} Metric. Several factors may cause allelic signals to decline systematically with increasing fragment size, such as differential amplification and electrokinetic injection inefficiencies. However, for a given allelic area, allelic heights decline over the course of electrophoretic separation primarily due to diffusion-related peak broadening. Figure 1 displays the approximate linear decline in the allelic height/area ratio with increasing allelic size as observed in one routine AmpF*l*STR Identifiler (Applied Biosystems, Inc., Foster City, CA) allelic ladder analysis performed at NIST. While these ladder data may not adequately capture sample preparation and amplification biases, they do reflect both differential injection and diffusion effects.

The heights of the largest alleles are ~30% lower than those of the smallest alleles. Since this allelic size range accommodates all the alleles of three or more genetic loci per color, the within-locus differences in height/area ratio due to diffusion-related effects are expected to be no more than ~10%. This pattern is typical of other multiplexes we have analyzed using capillary electrophoretic systems (data not shown). Thus, the among-sample differences in major- and minor-donor allele heights will be rather small within a given genetic locus. However, the lowering of signal height with increasing allelic size will affect among-locus signal balance.

Most participants reported intensities only for signals that were identified as donor alleles. Virtually all participants correctly identified all major-donor alleles in all samples.[3] However, minor-donor allele recognition varied widely both by sample and by participant (data not shown). Whenever a minor allele of the six mixed-source samples had less than threshold intensity or was misidentified as stutter, the h_{jklm} for that locus is reduced. (Since the major and minor donors in all of the mixed-source samples shared some alleles, analysis of just major-donor signals is not practical.)

Since the intensity of below-threshold signals is—by definition—small, excluding these heights from the summation generally has little impact. However, because of participant-specific policies with regard to acceptable threshold and maximum signal levels, the exclusions may contribute to systematic differences among participants.

Most nonreported but above-threshold minor-donor alleles were interpreted as stutter. Participant-specific reporting policies strongly affected the number of such exclusions. Further, participant-specific choices of amplification parameters affect stutter intensity. While potentially leading to greater among-participant differences than below-threshold signals, the effect is bounded by the intensity of true stutter signals at the given locus. The heights of stutter artifacts are seldom more than 20% of the height of the parent allele and are typically much less.[10,15−18]

Transformations. To the extent that the h_{jklm} metric validly summarizes signal intensity at a given locus of a given analysis, the average height of the J loci probed, $\bar{h}_{klm} = \sum_j^J h_{jklm}/J$, provides a simple estimate of the signal intensity for that analysis. We elsewhere use this metric to explore the relationships between DNA quantitation and STR signal intensity.[3]

Figure 2 displays the probability density functions (PDFs) of $\bar{h}_{klm}$ for all analyses performed on the three types of instrument used in the MSS3. Where sufficient data are available to enable estimation, the PDFs for different multiplexes evaluated on the same instruments are quite similar (data not shown). There is well over a 10-fold range in signal intensities among the analyses. This wide variability in absolute intensity must be mathematically isolated before more subtle among-loci signal intensity differences can be evaluated. This is accomplished by normalizing the height-based metric for each locus by the average summed height, $h'_{jklm} = h_{jklm}/\bar{h}_{klm}$.

(15) Meldgaard, M.; Morling, N. *Electrophoresis* **1997**, *18* (11), 1928−1935.
(16) Holt, C. L.; Buoncristiani, M.; Wallin, J. M.; Nguyen, T.; Lazaruk, K. D.; Walsh, P. S. *J. Forensic Sci.* **2002**, *47* (1), 66−96.
(17) Buse, E. L.; Putinier, J. C.; Hong, M. M.; Yap, A. E.; Hartmann, J. M. J. *Forensic Sci.* **2003**, *48* (2), 348−357.
(18) Leclair, B.; Frégeau, C. J.; Bowen, K. L.; Fourney, R. M. *J. Forensic Sci.* **2004**, *49* (5), 968−980.

The resulting relative locus heights must be compared with caution. By construction, the average value h'_{jklm} is 1.0. If two loci have h'_{jklm} values of 0.1 and 10.0, they both differ from the average by the same, equally important multiplicative factor (10-fold). However, their arithmetic differences from the average (−0.9 and 9.0, respectively) are not equal. Logarithmic transformation of the ratios, $h''_{jklm} = \log(h'_{jklm})$, equalizes the magnitude of the differences. For the example, $\log(0.1) = -1$, $\log(1) = 0$, and $\log(10) = 1$, and the arithmetic differences from the average h''_{jklm} (−1.0 and 1.0, respectively) are equal. Since many data analysis systems assume arithmetic rather than geometric relationships, logarithmic transformation facilitates further analysis. However, results from such transformed data are often more interpretable when transformed back into a more familiar form, in this case, the multiplicative-factor intensity ratios: e.g., $h'_{jklm} = 10^{(h''_{jklm})}$. This back-transformation is readily accomplished in graphical displays through appropriate axis tics and labels.

Combining Replicates. Nearly 70% of participants reported allelic heights for two or more replicate analyses of at least one sample. To ensure equal representation of each participant in the data, each set of "K" replicate log-transformed relative heights has been reduced to its arithmetic average, $h''_{jlm} = \sum_k^K h''_{jklm}/K$.

Data Analysis. *Models.* There are many ways of summarizing a complete set of h''_{jlm} for the $L = 7$ samples and J loci for every set of $M \geq 2$ participants using the same nominal measurement systems. Since we are most interested in evaluating the relative importance of sample, participant, and intrinsic factors to the observed intensity differences among the genetic loci evaluated in the different multiplexes, we chose to model the h''_{jlm} as a three-level nested hierarchy: samples by participants by loci.

If it is assumed that the h''_{jlm} from a given participant at a given locus are independent random draws from an approximately normally distributed (at minimum, symmetric and unimodal) population of relative intensities, then the expected relative intensity balance for that participant is the among-sample average $h''_{jm} = \sum_l^L h''_{jlm}/L$ and the among-sample variance at a given locus for a given participant is $s^2_{sam,jm} = \sum_l^L (h''_{jlm} - h''_{jm})^2/(L - 1)$. The expected intensity variance attributable to the samples is related to the average of the individual variances $s^2_{sam} \cong \sum_j^J \sum_m^M s^2_{sam,jm}/(J \times M)$.

If it is assumed that the M h''_{jm} values for a given locus are independent random draws from an approximately normally distributed population of participants, then the expected intrinsic relative intensity for the locus is the among-participant average $h''_j = \sum_m^M h''_{jm}/M$ and the among-participant variance at a given locus is $s^2_{\mathrm{part},j} = \sum_m^M (h''_{jm} - h''_j)^2/(M - 1)$. The expected intensity variance attributable to the participants is related to the average of the per-locus variances $s^2_{\mathrm{part}} \cong \sum_j^J s^2_{\mathrm{part},j}/J$.

Since the h''_{jlm} are normalized to the average among-locus intensity, the among-locus average $\sum_j^J h''_j/J$ is zero by construction. If it is assumed that the J h''_j values for a given multiplex are independent random draws from a normally distributed population of loci, the variability attributable to intrinsic differences in intensity balance is related to the among-locus variance: $s^2_{\mathrm{loc}} \cong \sum_j^J (h''_j)^2/(J - 1)$.

Analytical Tools. While related to the desired quantities, these simple sums of squares are not themselves the most appropriate estimates of the conceptually disjoint variance components. The calculations that appropriately allocate degrees of freedom and correct for the nested dependencies are well known but somewhat involved.[19] Numerous software systems are available that implement some form of the required calculations; however, it typically requires specialist knowledge to specify the intended model and to convert "statistical" outputs into more chemically tractable information.

The WinBUGS empirical Bayesian analysis system is a powerful data modeling system with a fairly transparent model specification language that can greatly simplify result interpretation.[20] Rather than directly solving a set of equations valid under particular (often only implicitly specified) circumstances, empirical Bayesian systems estimate the entire distribution of values that are consistent with the observed data for each parameter of a given model. These distributions can be visualized as PDFs; graphical examination of these PDFs provides a powerful and chemically familiar way of assessing the model's validity. For example, if the data are not random draws from a nested set of unimodal, symmetrical distributions but rather are drawn from two or more distinct populations, then at least some of the WinBUGS parameter distributions will be multimodal or skewed. In contrast, the simple mean and variance calculations described above will indicate only that the putative normal distribution is quite broad.

Given a model that adequately describes the data, the median of the PDF for a given parameter is an appropriate point estimate. The interval that bounds $X/2$ percentage of the PDF area on either side of the median directly defines the X percentage confidence interval on the parameter.

Two nested hierarchical models were evaluated. The initial model assumed the same among-participant variance at each locus, $s^2_{\mathrm{part},j} = s^2_{\mathrm{part}}$. While numerically efficient because it reduces the number of parameters that must be estimated, this constraint causes the confidence intervals for all h''_j to be of identical length. Removing this constraint enabled calculation of more realistic confidence intervals at the cost of increased computational time: typically, a few wall clock minutes on a 1-GHz desktop processor. Both models produce very similar values of the h''_j point estimates. Particularly for the data sets having relatively few participants, the two models do result in different allocations of the among-participant and among-locus variance, although the confidence intervals of the estimates overlap. Given its more believable confidence limits, we report here only the results from the less constrained model.

Intrinsic Balance Differences. Figures 3−7 display the WinBUGS medians and (for graphical clarity) their 80% confidence intervals for the h''_j of the various multiplexes evaluated in MSS3. Loci are grouped by dye label and sorted within groups in order of increasing allelic size.

Figure 3 displays the relative intensity balance of the ABI Pro+ and Cof multiplexes evaluated on ABI 310 and 377 instruments. The patterns for the different instruments are quite similar. There is a general decline in signal intensity with increasing locus size. For Pro+, the signal intensities are about the same regardless of the dye label used; i.e., the loci are well balanced across the Pro+ three colors. Cof is less well color-balanced, with the blue loci

(19) Hocking, R. R. *Methods and applications of linear models: regression and the analysis of variance*, 2nd ed.; Wiley-Interscience. Hoboken, NJ, 2003.

(20) WinBUGS Version 1.4. MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK, http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml.

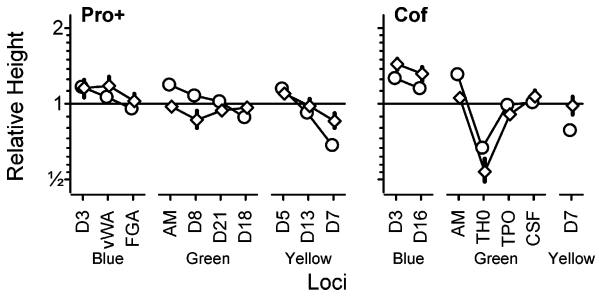

Figure 3. Among-locus normalized signal intensities for AmpF*l*STR Profiler Plus and COfiler. The left-hand graphical segment displays the among-participant expected value of the normalized allelic height for each locus of the Pro+ multiplex. The right-hand segment displays the expected values for each locus of the Cof multiplex. The loci are grouped by their detection color; within each color group the loci are arranged in order of increasing bp size. Open circles represent data from ABI 310 systems; open diamonds represent data from ABI 377 systems. Uncertainty bars denote approximate 80% confidence intervals about the parameter values.

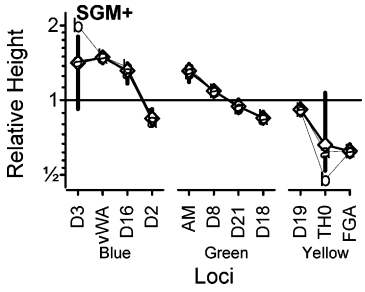

Figure 4. Among-locus normalized signal intensities for AmpFlSTR SGM Plus. The graphical elements are as in Figure 3, with the addition of symbols (a, b) representing both sets of within-participant expected values reported for this multiplex.

most intense and the yellow least. The THO1 locus is consistently much less intense than are the other Cof loci.

Figure 4 displays the balance of the ABI SGM+ multiplex, evaluated on ABI 377 instruments. While only two participants provided complete data for this multiplex, the pattern and magnitude of balance differences observed by both participants are quite similar. Loci are much less color-balanced than Pro+, again with the blue loci most intense and the yellow least. The decline of signal intensity with increasing locus size is more pronounced than for Pro+.

Only one participant reported quantitative ABI Idf multiplex results for the MSS3 samples, using an ABI 310. However, 371 unique single-donor samples had been typed at NIST with the Idf multiplex shortly after the MSS3 closing date using an ABI PRISM 3100 Genetic Analyzer.[21] Assuming that among-sample and between-instrument variation is relatively unimportant, a simple modification of the WinBUGS model enabled evaluation of the most interesting parameters from the 7-sample MSS3 and the 371-sample NIST sets. Figure 5 compares the relative heights for the two groups of samples. While the within-color pattern of differences vary somewhat, the pattern and magnitude of the among-color differences are quite similar.

Figure 6 displays the patterns for the Promega PP1.1+ and PP2.1 multiplexes, evaluated on Hitachi FMBIO II imaging

(21) Butler, J. M.; Schoske, R.; Vallone, P. M.; Redman, J. W.; Kline, M. C. *J. Forensic Sci.* **2003**, *48* (4), 908−911.

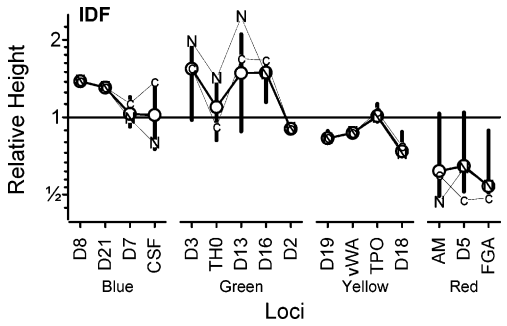

Figure 5. Among-locus normalized signal intensities for AmpF*l*STR Identifiler. The graphical elements are as in Figure 3, with the addition of symbols (c) representing the within-participant expected values reported for this multiplex in the MSS3 study and (N) representing within-analyst expected values for 371 samples analyzed at NIST on an ABI 3100 instrument.

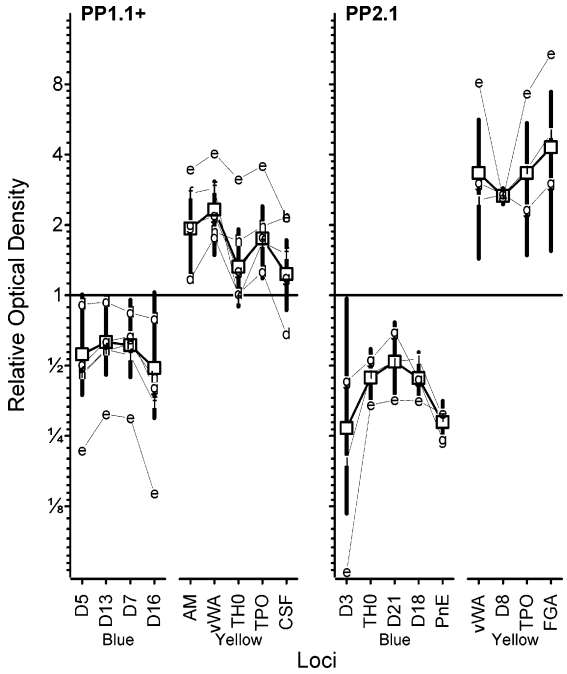

Figure 6. Among-locus normalized signal intensities for PowerPlex 1.1 and 2.1. The left-hand graphical segment displays the expected values of the normalized optical densities for each locus of PP1.1 and PP1.1+ multiplexes. The right-hand segment displays the expected values for each locus of the PP2.1 multiplex. Open squares represent the among-participant expected values. Uncertainty bars denote approximate 80% confidence intervals about the parameter values. The symbols d−g represent individual participant values for PP1.1+ and PP2.1. The h symbols represent the values reported by the one participant who used PP1.1.

systems. While the within-color patterns are similar among the different participants, the between-color difference is much more variable than with the other multiplexes. The within-color regularities suggest that they are intrinsic properties of the multiplexes; the between-color differences suggest that they reflect the effect of individual participants' choices of gel properties, electrophoretic settings, and analysis parameters.

Figure 7 displays the patterns for the Promega PP16 multiplex, evaluated on ABI 310 and 377 instruments. There is little consistency in either the within- or among-color group differences among the participants. This suggests that the loci in this multiplex are intrinsically well balanced but that there is considerable

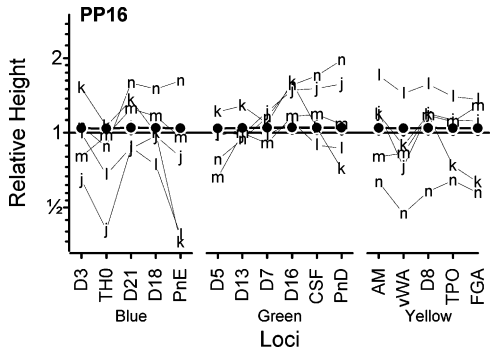

Figure 7. Among-locus normalized signal intensities for PowerPlex 16. Open diamonds represent the expected values for PP16 as analyzed on ABI 310 and ABI 377 systems. Uncertainty bars denote approximate 80% confidence intervals about the parameter values. The symbols j−m represent individual participant values for data collected on ABI 310s. The symbols n represent the values reported by the one participant who used an ABI 377.

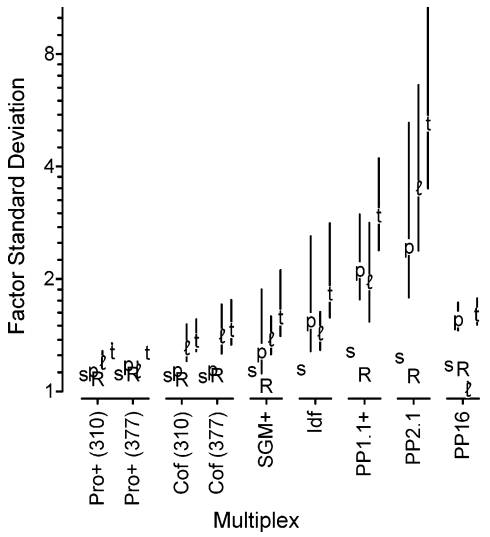

Figure 8. Sources of variability in the normalized signal intensities. Each symbol denotes the multiplicative standard deviation for a defined variance component estimated for a particular multiplex or, where sufficient data are available, multiplex (instrument) combination. The repeatability standard deviation for the control sample is denoted R. The WinBUGS-calculated among-sample, among-participant, among-locus, and total components are denoted s, p, l, and t, respectively. Uncertainty bars on the WinBUGS parameters denote approximate 80% confidence intervals.

variation induced by the participants' choices of analysis parameters.

Components of Variance. When appropriately defined, each data set's total variance is just the sum of the component variances: $s^2_{\text{total}} = s^2_{\text{sam}} + s^2_{\text{part}} + s^2_{\text{loc}}$. However, this model confounds the variance expected from random, short-term measurement changes, s^2_{repeat}, with that from the analysis of different samples. Since 30% of the participants did not report replicate analyses—even for the control sample—and the number of replicates that were reported varied widely among the participants, direct evaluation of s^2_{repeat} within WinBUGS would require specification of quite a complex model. It is more convenient to approximate the value for this variance component from the replicate analyses of control sample R, $s^2_{\text{repeat}} \approx s^2_R = \sum_j^J \sum_m^{M^*} \sum_k^{K_m} (h''_{jklm} - h''_{jlm})^2/(K_m - 1)/(J \times M^*)$, where M^* is the number of participants who reported at least two analyses and K_m is the number of replicates reported by the mth participant. With the exception of Idf, at least one participant reported replicates for control sample R for all of the multiplexes.

Figure 8 displays the square root of these variances (i.e., standard deviations) for all of the multiplexes evaluated in MSS3. Where sufficient data are available, the unique combinations of multiplex and instrument have been separately evaluated. The estimates are much the same for both the ABI 310 and 377 detection systems.

As expected, the s_R are consistently somewhat smaller than s_{sam} in all the multiplexes where both could be estimated. The values are sufficiently similar to conclude that there is little or no among-sample variance in the relative heights. The WinBUGS-estimated s_{sam} are thus measures of short-term random measurement variability. The s_{sam} are small (multiplicative factors of 1.1× to 1.2×) for all multiplexes, indicating that the relative allelic heights are not much affected by short-term events.

For the ABI-produced Pro+ and Cof multiplexes, the s_{part} are very nearly as small (~1.15×) as s_{sam} with both the ABI 310 and 377 instrument systems. This strongly suggests that the relative allelic intensities for these multiplexes are not much influenced by among-participant differences. For the Pro+ systems, s_{loc} is also small (~1.2×) as can be seen in the nearly uniform h''_j for these systems displayed in Figure 3. The s_{loc} for the Cof systems are still small (~1.4×) but are larger than for Pro+.

Too few participants reported data for the ABI-produced SGM+ and Idf and the Promega-produced PP1.1+ and PP2.1 multiplexes to reliably differentiate among-participant from intrinsic differences. The similarities in the patterns for different participants shown in Figures 4−6 do suggest that, whatever the relative allocation of variance, there are intrinsic intensity balance differences in these systems. The relative optical densities of the two slab-gel Promega systems are much more variable (s_{total} of 3× to 5×) than are the relative heights of all other systems.

As shown in Figure 7, the Promega-produced PP16 system appears to be remarkably well intensity-balanced across all genetic loci (s_{loc} of 1.01×). However, there is significant among-participant contribution to the relative intensity variance (s_{part} of 1.6×).

CONCLUSIONS AND RECOMMENDATIONS

All eight of the commercial STR multiplexes provided fit-for-purpose results even for the designedly difficult samples distributed in the MSS3 challenge study. Some among-loci differences observed in seven of the multiplexes can be attributed to effects at least partly under analyst control. However, these multiplexes displayed similar patterns of intensity differences for several different samples, in the hands of different analysts, using a variety of measurement platforms. These observed differences may be intrinsic properties of the individual multiplexes.

To the extent that quality systems specify minimum and maximum absolute intensities for data acceptability, intrinsic intensity differences may decrease the utility of the multiplex results and surely increase the cost of analysis. While the systematic decline in signal height with increasing allelic size may be partially offset by interpretation of allelic areas rather than heights, many of the observed among-locus differences are locus or color-group specific. Intrinsic differences are addressable only by the multiplex manufacturers, most likely through minor

adjustments of PCR primer concentrations. Unfortunately, if such modifications proceed by reducing the sensitivity of one or more colors, the PCR amplification protocol may need to be modified (and revalidated) to maintain sensitivity. If rebalance proves impractical, both manual and automatic STR interpretation schema could benefit from the explicit recognition of, and quantitative correction for, any intrinsic signal intensity differences.

ACKNOWLEDGMENT

This work was supported in part by the National Institute of Justice through an interagency agreement with the NIST Office of Law Enforcement Standards. Certain commercial equipment, instruments, or materials are identified in this report to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose.

Received for review June 4, 2004. Accepted September 15, 2004.

AC049178K