Final accepted manuscript, subsequently published as
Hicklin, R. A., Richetelli, N., Emerick, B. L., Bever, R. A., & Davoren, J. M. (2023). Variation in assessments of suitability and number of contributors for DNA mixtures. *Forensic Science International: Genetics, 65*, 102892.
https://doi.org/10.1016/j.fsigen.2023.102892

Variation in Assessments of Suitability and Number of Contributors for DNA Mixtures

R. Austin Hicklin[1,*], Nicole Richetelli[1], Brandi L. Emerick[1], Robert A. Bever[2], Jonathan M. Davoren[2]
[1]Noblis, Reston, VA, USA
[2]Bode Technology, Lorton, VA, USA
[*]Corresponding author (hicklin@noblis.org)

Abstract

The interpretation of a DNA mixture (a sample that contains DNA from two or more people) depends on a laboratory/analyst's assessment of the suitability of the sample for comparison/analysis, and an assessment of the number of contributors (NoC) present in the sample. In this study, 134 participants from 67 forensic laboratories provided a total of 2,272 assessments of 29 DNA mixtures (provided as electropherograms). The laboratories' responses were evaluated in terms of the variability of suitability assessments, and the accuracy and variability of NoC assessments. Policies and procedures related to suitability and NoC varied notably among labs.

We observed notable variation in whether labs would assess a given mixture as suitable or not, predominantly due to differences in lab policies: if two labs following their standard operating procedures (SOPs) were given the same mixture, they agreed on whether the mixture was suitable for comparison 66% of the time. Differences in suitability assessments have a direct effect on variability in interpretations among labs, since mixtures assessed as not suitable would not result in reported interpretations.

For labs following their SOPs, 79% of assessments of NoC were correct. When two different labs provided NoC responses, 63% of the time both labs were correct, and 7% of the time both labs were incorrect. Incorrect NoC assessments have been shown to affect statistical analyses in some cases, but do not necessarily imply inaccurate interpretations or conclusions. Most incorrect NoC estimates were overestimates, which previous research has shown have less of an effect on likelihood ratios than underestimates.

1 Introduction

Analysis of forensic DNA evidence often involves the interpretation of DNA mixtures (samples that contain DNA from two or more people). The interpretation of DNA mixtures is generally conditioned on two assessments routinely conducted by forensic laboratories: the suitability of a given sample for analysis, and the number of contributors (NoC) present in the sample. Here we present the results of a study in which forensic laboratories were provided electropherograms (EPGs) of DNA mixtures for interpretation; participants were asked to report on whether the mixtures were suitable for analysis and, if so, to provide the estimated number of contributors to the sample. The resulting responses were evaluated in terms of the variability of suitability assessments, and the accuracy and variability of NoC assessments.

These results are part of *DNAmix 2021,* a larger study that also evaluated the variability across laboratories for policies and procedures related to DNA mixture interpretation, as well as the extent of consistency and variation among forensic laboratories in interpretations, comparisons, and statistical analyses of DNA mixtures. This research was conducted to provide key empirical data to the legal and forensic science communities, expanding on the results and lessons learned from previous studies [1–3], and addressing concerns regarding complex DNA mixtures raised by the National Research Council (NRC) [4], the President's Council of Advisors on Science and Technology (PCAST) [5], the Government Accountability Office (GAO) [6], and the National Institute of Standards and Technology (NIST) [7]. Previous studies have been done to characterize the inter-laboratory variability and performance on DNA mixture samples (e.g., [1–3]); however, in the intervening years since those studies were conducted, probabilistic genotyping has advanced and seen more widespread adoption in forensic laboratories, a significant paradigm shift that impacted the practice of DNA mixture interpretation [7–9]. Mallinder, et al [10] report the results of a study on interpretations by laboratories using probabilistic genotyping, but which was limited to seven laboratories in the United Kingdom and Ireland. As such, *DNAmix 2021* serves as the first large-scale study evaluating the extent of variation in interpretation and statistical analysis of DNA mixtures, specifically to: include results from current state-of-the-practice probabilistic genotyping software (PGS), with samples selected to be representative of the range of attributes found in actual DNA casework, using only real human DNA samples, and not restricted to any specific product or statistical approach.

2 Background

DNA analysis is touted as the "gold standard" of forensic science, particularly for single-source samples and simple mixtures (samples with two contributors, relatively high DNA quantity, and no degradation or inhibition) [5,7]. However, several reports and reviews have called for additional research into the reliability of the interpretation of "complex" mixture samples [4–7]—those that include three or more contributors, have low amounts of DNA (low template), exhibit degradation or inhibition, and/or contain contributions from contributors with shared alleles (allele stacking). When a forensic laboratory receives a DNA sample in casework, it is not always known whether the sample is even a mixture (nor whether a potential mixture is simple or complex). As such, one of the first steps taken after STR fragment analysis is for a DNA analyst to evaluate the resulting profile for suitability and estimate the number of contributors to the sample. These initial assessments impact whether a DNA mixture is analyzed at all, and if so, how the interpretation, comparison, and statistical analysis proceeds given the profile characteristics [11]. Therefore, it is important to characterize the variability in each of these assessments and understand the accuracy of the number of contributors estimates because the remainder of mixture interpretation and analysis is rooted in these initial evaluations.

The assessment of suitability is one of the first steps in interpretation and is a key decision made in DNA mixture analysis—determining whether the content of the mixture is sufficient for interpretation and is worth continuing with comparison and statistical analysis [11]. Deciding whether a DNA mixture is suitable for comparison and statistical analysis requires expert judgement coupled with defined policies and validated procedures set by the laboratory. Suitability decisions are based upon a number of sample-specific characteristics (some of which are explicit and some which require additional judgement/discretion), including the total quantity of DNA required (template amount), estimated NoC, uncertainty of NoC, peak height ratios, presence of degradation/inhibition, presence of artifacts, and overall complexity/quality. Some variability in suitability decisions may be expected, even within laboratories, given that labs often differ in their standard operating procedures (SOPs) and analysts may differ in their own personal thresholds for making judgements (for additional details regarding the differences in laboratory SOPs related to suitability, refer to the companion paper [12]). The authors are not aware of any large-scale study characterizing the variability of suitability assessments for DNA mixtures, but this information has been collected as part of some previous studies (e.g., [1,13]). However, the importance of suitability assessments has been highlighted operationally via audits and re-analyses of evidence in several legal cases— there are a number of cases in which a sample was assessed as unsuitable for comparison/statistical analysis, but later found to yield useful data [14,15] (and conversely wherein a sample was interpreted that should not have been given quality issues [16,17]). Given the importance of the initial determination of suitability for DNA mixture interpretations (including whether or not a sample moves forward), it is imperative to evaluate the variability in these assessments in order to fully understand the reliability of DNA mixture interpretations.

As part of the assessment of suitability for a DNA mixture, analysts generally estimate the number of contributors to the sample (see *Appendix A2* for details regarding how NoC is estimated). While the number of contributors estimate certainly impacts the suitability determination (many laboratories have SOPs that specify a maximum number of contributors threshold [12]), this assessment can also have a notable influence on any further comparison or statistical analysis. In particular, NoC estimates may be used to aid in selecting a statistical method for analysis (e.g., using a binary approach for simple mixtures and PGS for more complex mixtures) and are also generally required as inputs for statistical analysis (either explicitly or implicitly). For probabilistic genotyping, the number of contributors is often set by the analyst and used as an assumption by the software for deconvolution and ultimately computing the likelihood ratio (LR) [8]. For binary approaches, such as random match probability (RMP) [18] and combined probability of inclusion/exclusion (CPI/CPE) [19], the estimated number of contributors is implicitly considered by an analyst when determining possible genotypes of contributors and doing computations. Several previous studies have examined the impact of incorrect NoC estimates on the resulting statistics (e.g., [2,13,20–25]). In general, overestimation of the true NoC produces less discriminating LRs (lower LRs for true contributors and higher LRs for non-contributors), whereas underestimation can result in false exclusion of true contributors to the sample [2,22,23,26]; LRs for minor contributors are more notably affected by variation in NoC estimates [13,25]. As such, it is critical to characterize the variability and accuracy of NoC estimates because they serve an important role in DNA mixture interpretations, both up front during suitability assessment and during any subsequent comparison and statistical analysis.

Overall, *DNAmix 2021* sought to characterize the extent of variability in interpretations of DNA mixture profiles starting from the electropherogram, including initial mixture assessments (e.g. suitability and number of contributors) and moving through comparison and statistical analysis. The results presented here address the first step of these interpretations by providing specific insight into the extent of variability between laboratories in their assessments of suitability and number of contributors. Given that these initial assessments have substantial effects on whether and how DNA mixtures are interpreted, these results provide estimates that may be used to assist in decision making, improving procedures and training, and highlighting areas for potential standardization.

3 Materials and Methods

This paper reports on a subset of the results of *DNAmix 2021*, a study consisting of four phases:

1. *Policies and Procedures (P&P) Questionnaire* — Online questionnaire to assess laboratory policies and procedures relevant to DNA mixture interpretation (notably systems, types of statistics reported, and parameter settings used).

2. *Casework Scenario Questionnaire* — Online questionnaire to assess analysis procedures or decisions that may vary depending upon the case scenario, and to assess the nature of mixture casework.
3. *Number of Contributors (NoC) Subtest* — Assessment of suitability and number of contributors, given electropherogram (EPG) data. A total of 21 mixtures were used in this subtest, out of which each participant was assigned 12 mixtures.
4. *Interpretation, Comparison, and Statistical Analysis (ICSA) Subtest* — Interpretations and statistical analyses, given EPG data for 8 mixtures, each provided with DNA profiles of potential contributors. All participants received the same 8 mixtures.

This paper reports the results regarding suitability and NoC assessments, which includes the relevant subset of *P&P Questionnaire* responses, all of the results from the *NoC Subtest*, and the suitability and NoC assessments from the *ICSA Subtest*. Each participant was presented with samples via custom web-based software that presented EPGs for download and recorded test responses. See *Appendix B* for a description of the overall study and its design. Participant instructions are summarized in *Appendix B2*.

3.1 DNA Mixtures

A total of 29 DNA mixtures were created for use in the study: 21 mixtures were used in the *NoC Subtest*, 12 of which were assigned to each participant; eight mixtures were used in the *ICSA Subtest*, all of which were assigned to each participant; each participant that completed *DNAmix 2021* was assigned 20 of the 29 mixtures. The mixtures were designed and created to be broadly representative of the range of attributes encountered in actual DNA mixture casework. The mixtures were created to vary with respect to the number of contributors, the amount of DNA (in total and for each contributor), the relative proportions of contributors, degradation, and the extent of allele sharing. Given the multitude of factors that influence DNA analysis, it is not feasible to exhaustively cover the factor space with a small number of mixtures. Recognizing that (significant) limitation [7], the experimental samples were selected to span a spectrum of the attributes encountered in actual DNA mixture casework [27], anticipating that most DNA mixtures of interest would either be similar to the provided mixtures or could expect performance interpolated between provided mixtures that are more complex and less complex. The DNA used to create the mixture profiles for this study came from various sources, including buccal, blood, and tissue samples. There were no simulated or contrived profiles: all DNA profiles in this study were from real people. Allele sharing was controlled through the selection of subjects from a broad pool of subjects from multiple sources: the mixtures include 102 subjects selected from a pool of 849 subjects from four sources.

In order to represent the SOPs of as many participating laboratories as feasible, EPGs were prepared using the four most commonly-used combinations of amplification (Amp) and capillary electrophoresis (CE) settings ("Amp/CE Settings"), as selected by registered participants. "Amp/CE Settings" (Table 1) refers to a specific combination of amplification kit, amplification cycles, volume of amplification reaction, CE instrument, and injection time and voltage. When selecting Amp/CE Settings, participants also indicated how the selected settings compared to their SOPs. When reporting results and analyses, we differentiate participants who indicated that the Amp/CE Settings exactly corresponded or were equivalent to their laboratory's SOPs ("SameSOP") from those participants who indicated the Amp/CE Settings differed from their laboratory's SOPs ("DiffSOP"). To verify that the amplifications produced consistent electropherograms and to minimize stochastic effects, each mixture was amplified at least twice for each Amp/CE setting; the mixture was only used if at least two amplifications were found to be replicable on a qualitative review, and any differences were within the normal variation expected for the DNA input level (considering dropout, dropin, and signal strength variation between the amplifications).

Abbreviation	***STR System***	***PCR Cycles***	***PCR Volume***	***CE Injection voltage***	***CE Injection Time***
ID28	Identifiler Plus	28	15 µL	1.2 kV	12 Seconds
GF28	Globalfiler		25 µL		24 Seconds
GF29	Globalfiler	29			
6C29	PowerPlex Fusion 6C				

Table 1. Amplification and capillary electrophoresis settings (Amp/CE Settings) used for mixture creation.

Table 2 summarizes the mixtures, showing how we varied NoCGT, the quantity of DNA, the ratios/proportions of contributors, degradation, and sharing of alleles. A variety of n-person mixtures were modeled using simulations in order to select subjects with a range of allelic sharing (as illustrated in the “Unique alleles” and “Alleles/locus” columns of Table 2). Note that one of the mixtures includes two brothers. See *Appendix C* for detailed discussion of mixture design and creation. EPGs of the mixtures are archived at OSF [28].

NoC$_{GT}$	Mixture ID [A]	DNA (Ng) [B]	Mix Ratio [C]	Smallest Minor [C]	DI [D]	Unique Alleles [E]	Alleles/ Locus [E]	Ref profiles [F]
2	ICSA_290/691	0.088	2.2 : 1	31%	<1	59	1 to 4	1
	NoC_52	0.054	1.3 : 1	44%	<1	75	2 to 4	
	NoC_24	0.043	2.1 : 1	34%	<1	54	1 to 4	
3	ICSA_192/680	0.341	14.9 : 1.8 : 1	6%	3.4[D1]	85	3 to 6	0
	NoC_49	0.191	1.6 : 1.4 : 1	26%	<1	78	2 to 5	
	NoC_74	0.186	1.3 : 1.2 : 1	29%	<1	91	2 to 6	
	NoC_28	0.180	1.5 : 1.2 : 1	28%	<1	79	2 to 5	
	ICSA_311/401	0.179	1.4 : 1.2 : 1	29%	<1	92	3 to 6	0
	ICSA_078/260	0.174	1.4 : 1.2 : 1	28%	<1	83	3 to 6	1(2)[F1]
	NoC_84	0.159	1.6 : 1.2 : 1	26%	<1	68	1 to 5	
	NoC_50	0.146	2.1 : 1.6 : 1	23%	<1	91	2 to 5	
	NoC_76	0.121	1.2 : 1.0 : 1	31%	<1	96	2 to 6	
	NoC_25	0.099	1.7 : 1.4 : 1	25%	<1	76	2 to 5	
	NoC_53	0.091	1.5 : 1.2 : 1	28%	<1	80	2 to 5	
	NoC_57	0.090	1.7 : 1.2 : 1	27%	<1	80	2 to 5	
4	NoC_29	0.872	1.5 : 1.5 : 1.4 : 1	19%	<1	82	2 to 6	
	NoC_93	0.580	2.4 : 1.5 : 1.2 : 1	17%	<1	89	2 to 6	
	NoC_15 [G]	0.580	1.5 : 1.3 : 1.1 : 1	21%	<1	90	3 to 6	
	ICSA_057/802	0.486	20.1 : 9.5 : 6.5 : 1	3%	6.1[D2]	86	2 to 7	1
	ICSA_671/828	0.481	16.8 : 14.8 : 1.3 : 1	3%	<1	99	3 to 6	0
	ICSA_370/530	0.479	13.5 : 9.6 : 7 : 1	3%	<1	87	2 to 6	0
	NoC_70	0.234	1.4 : 1.2 : 1.1 : 1	22%	<1	100	3 to 7	
	NoC_05	0.211	2.5 : 2.1 : 1.6 : 1	14%	<1	98	2 to 7	
	NoC_14	0.210	1.7 : 1.6 : 1.6 : 1	17%	<1	101	3 to 7	
	NoC_68	0.188	1.4 : 1.2 : 1.1 : 1	22%	<1	95	3 to 7	
	NoC_41	0.171	1.9 : 1.7 : 1.7 : 1	16%	<1	88	2 to 8	
5	NoC_31	0.720	1.8 : 1.3 : 1.2 : 1.1 : 1	16%	<1	92	2 to 9	
	ICSA_767/328	0.376	2.4 : 2.0 : 1.5 : 1.2 : 1	13%	1.2	108	3 to 9	0
6	NoC_71	0.801	3.0 : 2.8 : 2.4 : 2 : 1.5 : 1	8%	<1	115	3 to 9	

Table 2. Summary of the 29 mixtures included in the study. The *NoC Subtest* had a total of 21 mixtures, of which each participant was assigned 12. The *ICSA Subtest* had eight mixtures, which were assigned to all participants. Mixtures are sorted by the actual ("ground truth") number of contributors (NoC$_{GT}$), then DNA amount amplified (this sort order is used throughout this paper unless otherwise indicated). Notes: (A) Each mixture in the *ICSA Subtest* was assigned in a comparison packet with a person of interest (POI) profile; each mixture has two ID numbers to differentiate packets in which the POI was included in the mixture (C:contributor) vs. not included (N:non-contributor). (B) Amount of DNA amplified as determined by the small autosomal amplicon of Quantifiler Trio. (C) Mix ratios and proportion of the mixture for the smallest contributor are based on signal strength as determined by STRmix as the average across all alleles for each mixture profile; this table shows the average across the Amp/CE versions of each mixture (see Appendix I for explanation and results by Amp/CE). (D) Degradation index (DI) as determined by Quantifiler Trio (F1: only the major was degraded; F2: entire mix was degraded). (E) Allele values were determined for GF29 and assume no drop-out; Allele counts (throughout this paper) only consider autosomal STR loci, ignoring Amel, Y indel, DYS391, DYS576, and DYS570, which are not generally used for mixture interpretation. (F) Some ICSA mixtures were provided with reference profiles (F1: two reference profiles provided but only one present). (G) NoC_15 included two brothers.

Each ICSA mixture was provided in a comparison packet that included one DNA mixture profile, one person of interest (POI) reference profile, and up to two reference profiles (victim, consensual partner, and/or expected contributor). Each mixture was provided to participants either in a contributor packet in which the POI was present in the mixture ("C" in Table 2), or in a non-contributor packet, in which the POI was not present in the mixture ("N" in Table 2). Given the limited number of samples included in the ICSA Subtest, ICSA was mainly focused on the variation in the degrees of support for statistical analyses of contributor packets—the primary reason for the inclusion of the non-contributor packets was so that the participants could make no assumptions regarding the presence or absence of the POIs. For each ICSA mixture, the non-contributor version of the mixture was assigned to one-eighth of the participants, and therefore the non-contributor versions of the mixtures received far fewer responses than the contributor versions.

DNA mixtures were provided as EPGs in .HID format.* EPG images were not provided, because the creation of the image would be dependent on use of analytical threshold and stutter filters, and therefore would risk biasing participants.

The characteristics of the mixtures used in this study were consistent with what participants indicated they encounter in casework (based upon responses to *P&P* and *CSQ*). Only one of the mixtures included in the study had less than 0.05 ng of total DNA; in the *Casework Scenario Questionnaire (CS#23;* [12]*)*, the majority of participating labs indicated that they at least occasionally interpret mixtures with less than 0.05ng of DNA. The contributor ratios between the highest and second highest contributor for mixtures included in the study were generally less than 1.5:1 (20 mixtures); 5 mixtures were 1.5:1-2:1, 3 mixtures were 2:1-3:1, and 1 mixture was 7:1. In the *Casework Scenario Questionnaire (CS#22;* [12]*)*, a supermajority of participating labs indicated that they often interpret mixtures with proportions ranging from 2:1-10:1, and all but one participating lab interprets mixtures that are 1:1 at least occasionally. Overall, when looking at the nature of mixture casework reported by participating labs in the *CSQ*, there are very few labs who say they never encounter the types of samples that we included in our study (e.g., sexual assault kits (SAKs) vs trace samples, number of contributors, contributor ratios, DNA quantity, etc.); in other words, the vast majority of participating labs see mixtures similar to the ones included in the *NoC* and *ICSA Subtests* at least occasionally (often for some labs) [12].

3.2 Participation

Participation was open to forensic laboratories that conduct DNA mixture interpretation as part of their standard operating procedures (SOPs). Non-U.S. laboratories were welcome to participate if they report their interpretations in English. Laboratories were permitted to have multiple participants. Participants were required to use the same diligence in performing these analyses as used in operational casework, to use their laboratory's SOPs in performing these analyses, and to follow their laboratory's quality assurance procedures.

The results reported here are from 134 participants, representing 67 laboratories. The laboratories varied in the number of participants per lab: 48 labs had one participant each; 15 labs had two to five participants each (45 participants total); 4 labs had seven to 14 participants each (41 participants total). Because of the widely different number of participants per lab, results are generally weighted to 1 response per mixture per lab. Of the 67 laboratories, 57 were U.S. laboratories (28 local, 28 state, one private), and ten were non-U.S. laboratories (3 federal/national, 7 state/provincial). The U.S. state laboratories were from 24 states. The ten non-U.S. laboratories were from 7 countries.† See *Appendix D* for additional information regarding participants.

3.3 Response Data

Analyses are based on 2,272 responses from 134 participants representing 67 laboratories (mean 2.0 participants per lab; median 1). These include 1,507 responses for the 21 mixtures from the *NoC Subtest* and 765 responses for the 8 mixtures from the *ICSA Subtest*. Each mixture received an average of 78.3 responses (median 77; range 37-129). Out of 134 participants, each completed 16.9 trials on average (median 20 trials)—87 completed all 20 assigned trials; 33 completed 12-19 trials each; 14 completed 11 or fewer trials.

* *.HID refers to a file format used for files generated by the Applied Biosystems 3500 series genetic analyzers that has become a de facto interchange standard (replacing the earlier FSA file format).*

† *We are not releasing the names of the countries because we are enforcing k-anonymity* [50] *(for k=5) and no country other than the U.S. had five or more participating laboratories.*

General discussion of *P&P* and *CSQ* results (unless specifically associating these with suitability or NoC responses) use the results reported in the companion paper "DNAmix 2021: Variation in Laboratory Policies, Procedures, and Casework Scenario Decisions": *P&P* results are based on the majority responses for 86 labs, and *CSQ* results are based on the majority responses from 83 labs.

As discussed above, participants selected one of four combinations of Amp/CE settings (ID28, GF28, GF29, 6C29), and indicated how the selected settings compared to their SOPs. Table 3 shows the breakdown of participants, labs, and responses by Amp/CE settings, and also shows the response counts with respect to Amp/CE settings and correspondence with SOPs. Note that in a few cases, participants indicated in their comments that specific individual responses did not follow their SOPs, resulting in some trials that are treated as SameSOP for suitability analyses but DiffSOP for NoC analyses. Overall, 84.9% of suitability responses were SameSOP; 82.7% of NoC responses were SameSOP.

To accommodate the fact that there was notable variation in the number of participants per lab, we generally report results by lab: results by participant simply treat each response equally (we refer to this as the *AllResponse Dataset*); results weighted by lab weight each response so that each lab collectively has one response for each mixture (*WeightedResponse Dataset*); see *Appendix E2* for more details on the weighting of responses.

	Total	*ID28*	*GF28*	*GF29*	*6C29*	*Dataset*
Total Participants	134	19	22	43	50	
Total Labs	67	5	9	27	26	
All responses	2,272	380	360	690	842	*AllResponse*
Weighted responses	1,222	100	150	473	499	*WeightedResponse*
Suitability SameSOP weighted responses	970	60	104	377	429	*WeightedSuitSameSOP*
NoC SameSOP weighted responses	958	60	102	368	428	*WeightedNoCSameSOP*

Table 3. Summary of response data: counts of participants, labs, and responses by Amp/CE settings; counts of responses in terms of correspondence with SOPs. Weighted responses are weighted by lab so that each lab has a total of one response per mixture for the given category (if applicable). The dataset abbreviations defined here are used throughout the remainder of this paper. See Table 1 for Amp/CE abbreviations.

To assess reproducibility of responses we use a self join of the response data. For assessing inter-lab reproducibility on all responses, the 1,222 weighted responses in the *WeightedResponse dataset* are paired with every response from other labs on the same mixtures, resulting in 53,554 weighted inter-lab decision pairs (*Interlab dataset*). Most reproducibility analyses use subsets of this dataset: the *InterlabSuitSameSOP dataset* contains 33,280 weighted inter-lab decision pairs, limited to SameSOP suitability responses, and the *InterlabNoCSameSOP dataset* contains 32,258 weighted inter-lab decision pairs, limited to SameSOP NoC responses. Intra-lab reproducibility is limited to the 394 trials in which 19 labs had more than one response per mixture: the *Intralab dataset* contains 127 weighted intra-lab decision pairs; the *IntralabSameSOP dataset* contains 116 weighted intra-lab decision pairs.

4 Policies and Procedures Related to Suitability and NoC

The companion document ("DNAmix 2021: Variation in Laboratory Policies, Procedures, and Casework Scenario Decisions") [12] presents the results from the first phase of DNAmix 2021, the *Policies and Procedures Questionnaire*. Here we briefly summarize those results that are most relevant to suitability and NoC. All P&P results cited in this section are based on the majority responses for the 86 labs that responded to the *P&P Questionnaire*. Note that the results presented throughout the remainder of this paper are for the subset of laboratories that submitted responses for the *NoC* and/or *ICSA Subtests* (for additional details, see [12]). For any analyses that tie suitability or NoC performance to individual *P&P* responses, we use the specific values and explicitly state the totals.

The scope of this work discusses variability starting with the EPG — but do note that a variety of laboratory policies affect the contents of the EPG. As discussed with respect to the Amp/CE Settings (*Section 3.1*, *Appendix C2*), the participating labs reported a notable variety of settings (and systems) used in amplification and capillary electrophoresis, which can affect the EPG in major and minor ways. For example, amplification kits vary in which loci are included in an EPG, and the availability of a specific polymorphic locus may have a notable effect on the interpretability of a given mixture; the number of cycles directly affects peak height, which can increase the ability to interpret low signal but also increases the potential of artifacts. Analytical threshold (AT) and stochastic threshold (ST) settings directly affect how an EPG is interpreted. Of the 86 labs that responded to the *P&P Questionnaire*, 27 varied AT by dye channel; for the labs that had a single default AT value the settings ranged from 40-200 relative fluorescence units (RFUs) (mean 96 RFUs). For ST, 24 of those 86 labs do not use ST; for the labs that had a single default ST value the settings ranged from 150-1250 RFUs (mean 456 RFUs).

Several *P&P* questions define each lab's approach to suitability decisions (see Sections 2.4.1 and 2.4.8 in [12] for details):

- Most labs (64/86) have policies that terminate analysis prior to amplification based on total DNA quantity. These thresholds vary widely: 6 labs list a threshold of 0ng, 30 labs terminate analysis if there is 0.01ng or less, and 44 labs terminate analysis if there is 0.05ng or less (mean: 0.0432ng, std dev: 0.0598ng; median: 0.0125ng; range: [0ng, 0.24ng]).
- Just over half of the 86 labs terminate prior to amplification based upon the male proportion of DNA (if the POI is male)—20 labs terminate analysis if the male fraction is 1% or less and 44 labs terminate analysis if the male fraction is 5% or less (mean: 1.9%, std dev: 1.5%; median: 1.7%; range: [0%, 5.0%]).
- In general, lab SOPs do not require an analyst to terminate analysis based upon degradation index (DI)—the single lab that does uses a DI=2 threshold. (However, most labs indicated that DI may influence target input.)
- Many labs (42/86) require a minimum number of loci with data in order to interpret a mixed DNA profile. This threshold varied from 2 to 15 loci (mean 6.2). A few labs (6/86) require a minimum number of alleles called with data in order to interpret a mixed DNA profile (ranging from 2 to 16 alleles).
- Labs differ on the suitability of mixtures that have major contributors but an unknown number of minor contributors: 31/86 labs permit interpretation of the major contributor if a mixture has one major contributor and an unknown number of minor contributors, but only 20/86 labs permit interpretation if the mixture has 2 or more majors. (Note that 30/86 labs do not differentiate between major and minor contributors.)
- The majority of labs (71/86) do not permit a mixture to be considered suitable for exclusion, but not suitable for inclusion/statistical analysis. (Note this is particularly relevant given the recent review of the Smiley decision conducted by the Texas Forensic Science Commission, in which they state that even if a mixture is deemed not suitable, it should be reviewed for exclusionary purposes [14].)
- Many labs also used other factors in assessing suitability: see *Appendix F* for details.

Several *P&P* questions describe how suitability decisions are related to estimates of NoC (see Section 2.4.6 in [12] for details):

- Most labs limit interpretation and/or comparison based on a maximum total NoC: NoC=4 is the most common threshold (50/86 labs), but labs often have thresholds of 3 (12/86 labs) or 5 (12/86 labs). Some labs also limit interpretation based on the number of unknown contributors (21/86 labs), or the number of minor contributors (10/86 labs).
- Labs reported a variety of responses regarding the suitability of mixtures in which there is uncertainty in the number of contributors: 24/86 labs consider such mixtures unsuitable, and 15/86 labs will only report the major contributors. Labs frequently cited other treatment of such mixtures: see *Appendix F2* for details.

Almost all labs (83/86) reported assessing the number of contributors manually (not using software). Most labs reported taking multiple indicators into consideration during manual NoC determination:

- Maximum Allele Count (MAC) per locus (81/86 labs)

- Relative peak heights (peak height ratios and possible shared/stacked alleles) (80 labs)
- Peak heights (RFU) (79 labs)
- Sex determining markers (75 labs)
- Expected stutter ratios (69 labs)
- Information below the analytical threshold (69 labs)
- Presence of degradation (61 labs)
- Overall level of data (peak heights in relation to laboratory validated thresholds) (57 labs)
- Peak morphology (e.g., CE resolution; unresolved microvariants; peak shouldering) (54 labs)
- Presence of inhibition (50 labs)
- Discriminating potential/variability of loci (or allele frequency) (39 labs)
- Quantitation data (35 labs)
- Total allele count in profile (29 labs)

Labs vary in how NoC estimates are used in casework:

- Most labs (60/86) are permitted the option to evaluate a mixture under different assumed NoCs.
- Most labs (51/86) are required to determine and record NoC before comparison to the victim, consensual partner, and/or expected contributor; most of these (43 labs) are permitted to change the assumed NoC after such comparison.
- Almost all labs (74/86) are required to determine and record NoC before comparison to the POI; about half of these (38 labs) are permitted to change the assumed NoC after such comparison.
- Most labs (46/86) are permitted to change the assumed NoC after conducting statistical analyses.

For more information on policies and procedures related to suitability and NoC, see *Appendix F* and [12].

5 Results: Suitability

For each mixture, participants were asked this question regarding the suitability of the mixture: "Is this DNA mixture profile suitable for comparison and/or statistical analysis? In other words, did you determine that this DNA mixture profile can appropriately be used to conduct comparisons (i.e., comparison of the mixture to reference profiles of POIs, victims, consensual partners, and/or expected contributors) and/or statistical analyses (i.e., compute an LR, RMP, or CPI/CPE with respect to a POI)?" Table 4 summarizes the responses to this question (overall and limited to *SameSOP*, both by participant and weighted by lab. Note that for the analyses presented here, we group the three intermediate categories of responses as "Partial suitability" (*PartSuit*).

Answer	*Abbrev*	*Responses*				*SameSOP Responses*			
		By Participant (AllResponse)		*Weighted by Lab (WeightedResponse)*		*By Participant (SuitSame)*		*Weighted by Lab (WeightedSuitSame)*	
Yes (for the entire mixture and all contributors)	YesSuit	1,138	50.1%	702.4	57.5%	925	47.9%	541.2	55.8%
Yes, but only for a subset of the contributors (e.g., major(s))	PartSuit	144	6.3%	56.1	4.6%	136	7.0%	50.6	5.2%
Yes, but only for a subset of loci		58	2.6%	21.7	1.8%	51	2.6%	15.4	1.6%
Yes, but only for a subset of loci, and only for a subset of the contributors		14	0.6%	6.7	0.5%	12	0.6%	4.7	0.5%
No	NoSuit	918	40.4%	435.3	35.6%	806	41.8%	358.2	36.9%
Total		*2,272*		*1,222.0*		*1,930*		*970.0*	
Subtotal: PartSuit		*216*	*9.5%*	*84.4*	*6.9%*	*199*	*10.3%*	*70.6*	*7.3%*
Subtotal: Yes+PartSuit		*1,354*	*59.6%*	*786.8*	*64.4%*	*1,124*	*58.2%*	*611.8*	*63.1%*

Table 4. Suitability responses by participant, weighted by lab, limited to *SameSOP* by participant, and limited to *SameSOP* weighted by lab. See Table 3 for definitions of dataset names. "Group" and "Abbreviation" define the labels used in analyses.

Figure 1 shows how suitability assessments varied by mixture, by the actual (ground truth) number of contributors (hereafter "NoC$_{GT}$"), and by the amount of DNA. Note that suitability assessments cannot be assessed as correct or incorrect—whether a mixture is suitable for analysis is a decision by an analyst based on laboratory policies. Interestingly, no mixtures received unanimous suitability assessments; even the simplest mixtures (2 contributors) yielded 19-38% unsuitability rates (*WeightedSuitSameDS*). As shown in Figure 2, there was no obvious consensus on suitability based upon NoC, except for 5 and 6 person mixtures—a supermajority of labs indicated that these mixtures were not suitable for comparison and/or statistical analysis; there were no obvious trends for 2-4 person mixtures. See *Appendix G* for additional details.

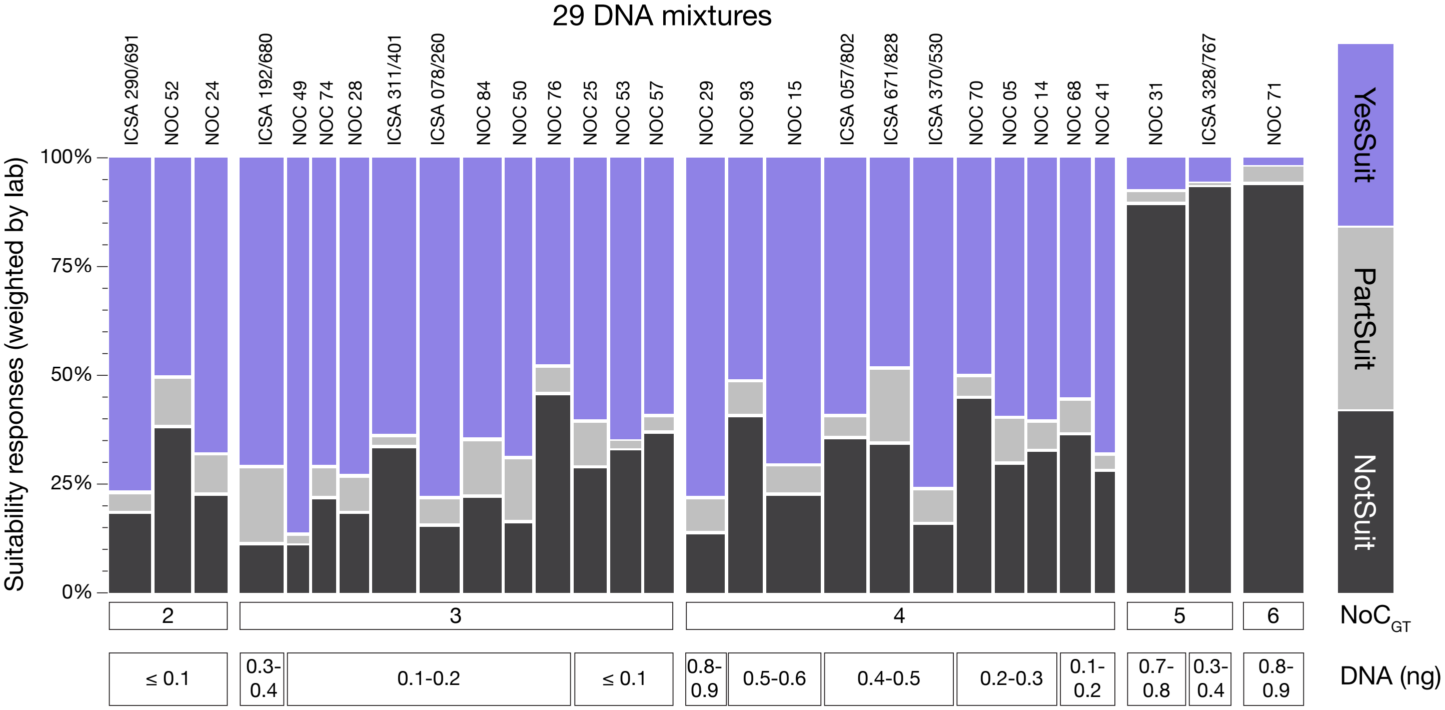

Figure 1. Suitability responses by mixture, weighted by lab, limited to *SameSOP.* See Table 2 for attributes of each mixture. Column widths are proportional to the number of weighted responses. See *Appendix G* for counts and *DiffSOP* responses. (*WeightedSuitSame dataset*: responses weighted to 1 response per lab)

Participants who responded *NotSuit* were asked to indicate the reason(s) for their decision. Figure 2 shows how the reasons for *NotSuit* decisions are associated with NoC_{GT} and with DNA template amount. The most-cited reason was "Too many contributors" (shown in red in Figure 2), which was cited as a reason for 63.5% of *NotSuit* responses (*sameSOP*, weighted) for NoC_{GT}=4, and 95.4% for NoC_{GT}≥5. Note that a majority of labs assessed four-person mixtures as *YesSuit*, but a majority of the labs that assessed four-person mixtures as NotSuit indicated that they contained too many contributors. Although the majority of labs assessed even the lowest template mixtures (≤0.1ng) as *YesSuit*, the majority of *NotSuit* responses for those mixtures cited "DNA template levels too low overall" (dark red in Figure 2), which was cited as a reason for the vast majority of NoC_{GT}=2. Other commonly-cited reasons were "Too much uncertainty in the number of contributors", and "Mixture proportions/contributor ratios" (see *Appendix G1* for combinations and less frequently-cited reasons).

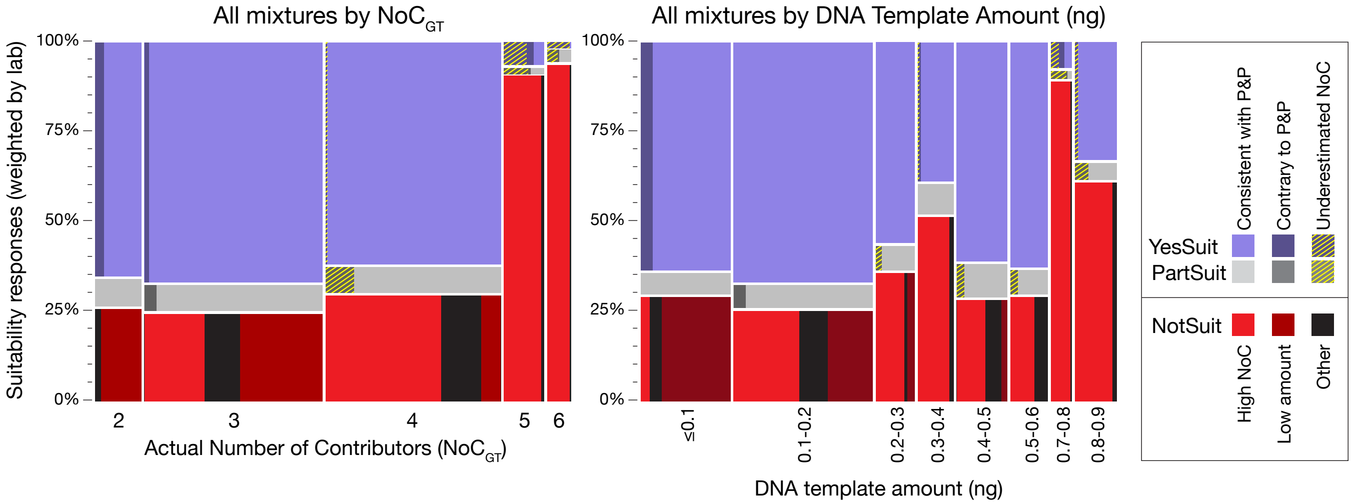

Figure 2. Suitability responses summarized by actual NoC (left), and by DNA template amount (right) weighted by lab, limited to *SameSOP*. (Summary of data shown in Figure 1)

As discussed in Section 4, a variety of policies guide suitability assessments for each lab. The suitability determinations made by a participant can often be predicted* based on two of the *P&P* settings for the lab: suitability thresholds based on NoC (as compared against NoC$_{GT}$) and suitability thresholds based on DNA amount (as compared against the actual DNA amount, which was provided to participants). These *P&P* thresholds are one-sided: if a mixture fails to meet one or both of these thresholds (minimum amount or maximum NoC), we would expect a *NotSuit* response — but a mixture passing both of these thresholds could be rejected for a variety of other reasons (note that participants provided explicit reasons for their *NotSuit* determinations in their responses). For weighted *SameSOP* responses, 80.8% of predicted *NotSuit* responses based on these two *P&P* settings were actual *NotSuit* responses (the remaining 19.2% are labeled "Contrary to P&P" in Figure 2). These unexpected results could often be explained by underestimates of NoC: the yellow hashed areas in Figure 2 indicate *YesSuit* and *PartSuit* trials in which NoC$_{GT}$ is greater than the participant's NoC threshold, but which can be explained by the participants' NoC responses. These are generally (65% of weighted *SameSuit* underestimates) due to reporting NoC as a valid range (such as a lab with a 4 contributor maximum assessing a 5-person mixture as ≥4), but the remaining trials are due to incorrect NoC estimates (such as the same lab assessing that mixture as exactly 4 contributors); see Appendix G2 for details.

Suitability determinations cannot be assessed as definitively correct or incorrect—whether a mixture is suitable for analysis is a decision made by an analyst based on laboratory policies. Therefore, the key point of interest with respect to suitability is the reproducibility of that assessment: if you give a mixture to two different labs, what portion of the time will they give the same (or different) suitability assessments? Since an assessment of *NotSuit* indicates that no further analysis would be conducted, different responses in an operational context would mean receiving interpretation/analysis from one lab, and nothing from another. Figure 3 summarizes the reproducibility of suitability determinations, considering various factors that may be of interest in a case. For example, top row of Figure 3 shows the chance that different participants from the same lab (SameLab) with *SameSOP* Amp/CE settings agreed on their suitability determinations for a given mixture is 86.3%. These intra-lab differences reflect the fact that laboratories were not always internally consistent in reporting the P&P settings relevant to suitability: of the 19 labs that had more than one participating subunit, six labs had differences in the minimum DNA thresholds they reported in the *P&P*

* *Note for clarity: "predictions" were not based upon regression or other modeling. Rather, predictions were developed by comparing the relevant P&P responses (e.g., maximum NoC, minimum DNA quantity, etc.) to the ground truth mixture characteristics (e.g., actual number of contributors, total DNA quantity amplified, etc.) in order to determine an expected suitability response.*

Questionnaire, and six labs had differences in the maximum NoC thresholds (one lab had differences in both). When considering responses from different labs (*DiffLab*), Figure 3 shows the agreement in suitability determinations is lower, generally ranging from 60.0%-68.8%. Not surprisingly based upon the results displayed in Figure 1 and Figure 2, the proportion of agreement on NoC_{GT} = 5-6 (*DiffLab SameSOP*) is much higher than that for NoC_{GT} = 2-4; this is driven by the general consensus that mixtures with high NoC are unsuitable for further analysis. See *Appendix G3* for additional details regarding reproducibility of suitability assessments.

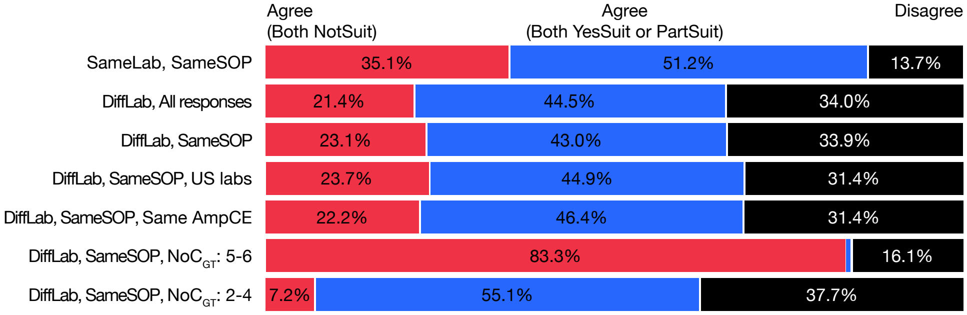

Figure 3. Reproducibility of suitability determinations considering various factors, including intra- and inter-laboratory consistency, the effect of including all responses vs limiting to SameSOP, the effects of limiting to US labs or to labs using the same Amp/CE settings, and the effects of different NoC_{GT} values. (*Intralab, Interlab, InterlabSuitSameSOP datasets*)

6 Results: Number of contributors

The actual number of contributors (NoC_{GT}) for the 29 mixtures in this study ranged from 2 to 6. For each mixture that a participant deemed suitable, they provided a response estimating the number of contributors (hereafter "NoC_{EST}")—they were allowed to provide NoC_{EST} as an exact value, as a minimum, or as a range. For the weighted *SameSOP* responses (n=958), 36% indicated that the mixture was not suitable for further analysis or too complex to determine NoC; of the remaining responses, 51.1% provided exact NoC values as estimates, 10.7% provided a NoC minimum, and 2.3% provided a NoC range. In evaluating minimum and range responses, we treat minima as ranges with an upper limit of 8 contributors.

Unlike suitability, NoC_{EST} can be evaluated in terms of accuracy, by comparing NoC_{EST} to NoC_{GT}. In this study, we know NoC_{GT} for all mixture samples. However, it is important to acknowledge that there may not be sufficient information in a given mixture to reasonably reach ground truth—both sample characteristics (e.g., DNA quantity, allele sharing, degradation, etc.) as well as laboratory SOPs (e.g., DNA quantity thresholds, AT, ST, etc.) may preclude an analyst from being able to readily observe NoC_{GT}. Given this, it is important to note that when we discuss "accuracy" of NoC_{EST} in this paper, we are considering ONLY consistency of NoC_{EST} with respect to NoC_{GT} — not whether the NoC_{GT} is apparent in the mixture (since this can be viewed as a judgement call). We use "correct" to indicate that NoC_{EST} is consistent with NoC_{GT} and "incorrect" to indicate that NoC_{EST} is not consistent with NoC_{GT}. We refrain from using the term "error" with respect to NoC_{EST} since there is ambiguity in whether NoC_{GT} is actually represented in the mixture—indeed, the variation in responses among labs provides an illustration of the extent to which NoC_{GT} is represented in the mixture.

Figure 4 shows the overall accuracy as well as accuracy by mixture: overall, 78.8% of *SameSOP* NoC_{EST} responses were correct. Even for 2-person mixtures— sometimes described as "simple mixtures"— 25% or more of NoC_{EST} responses were incorrect, suggesting that these are not necessarily always as straightforward as one might think. Note that for two mixtures (NOC_52 and NOC_74) more than half of the NoC responses were incorrect. Incorrect responses were disproportionately reported for exact NoC_{EST} (24.6% incorrect) as opposed to NoC_{EST} ranges (7.7% incorrect). On the four NoC responses in which NoC_{EST} and NoC_{GT} differed by two or more, we cannot rule out user error or carelessness. Note that all responses required participants to confirm responses before submitting. NoC_{EST} accuracy was very similar for *SameSOP* and *DiffSOP*; see *Appendix H* for *DiffSOP* results and additional details. For some mixtures, results varied by Amp/CE: see

Section 9 and *Appendix I* for details. NoC_{EST} accuracy for trials assessed as PartSuit was almost identical to that of trials assessed as YesSuit (*Appendix H2*).

Overestimations (NoC_{EST}> NoC_{GT}; red in Figure 4) almost always occurred on 2-3 person mixtures, whereas all underestimations (NoC_{EST}< NoC_{GT}; orange in Figure 4) occurred on 4-5 person mixtures. Note that in three of the 4-person mixtures, the proportion of DNA from the most minor contributor was less than 5% (Table 2). For the overestimates, we cannot be certain whether participants actually discerned too many contributors to a mixture or whether they intentionally increased their estimate because there was uncertainty, as a mechanism to be conservative in case of a low-level contributor. However, it is important to acknowledge that participants were offered the option of reporting a NoC minimum or NoC range but chose to report an exact value (potentially due to the requirements in their SOPs). With respect to underestimations, previous research has shown that as the number of contributors in a mixture increases, the chance of a "hidden" contributor also increases (e.g., [2]; *see Appendix A2* for additional details). Note that degradation or the inclusion of brothers did not have a striking effect on NoC rates (notes (c) and (d) in Figure 4, respectively).

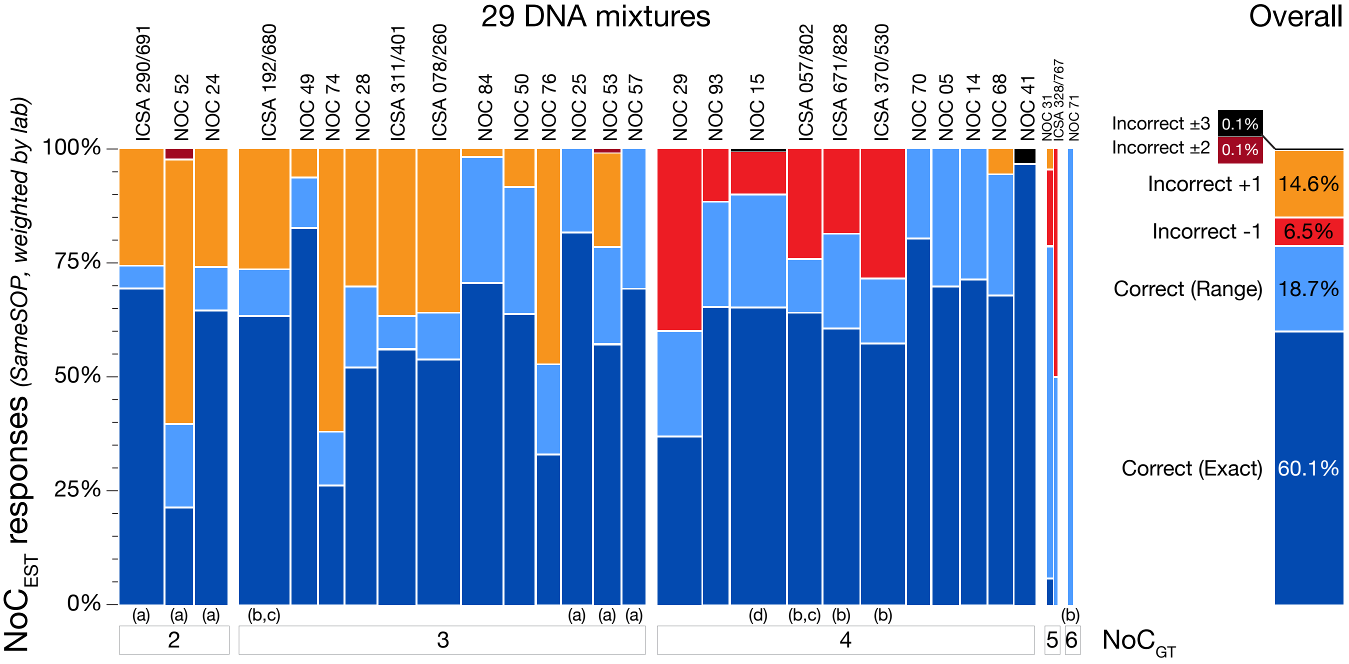

Figure 4. SameSOP NoC_{EST} responses with respect to NoC_{GT} by mixture. Note that column widths are proportional to the number of weighted responses; the columns for NoC_{GT}≥5 are thin because they had few responses other than NotSuit. The results for ICSA samples are summarized here, but presented in more detail in Section 7. Some mixtures vary by amplification: see Section 9. Notes (see Table 2 for details): (a) <0.1ng DNA; (b) smallest contributor is <10% of DNA; (c) DI >2; (d) 2 brothers. (*WeightedNoCSameSOP dataset*)

Participants generally assessed NoC_{EST} manually (88% of responses) or used diagnostics from a PGS (9%). Overall, NoC accuracy was not associated with the NoC method (see *Appendix J1*).

Table 5 summarizes the reproducibility of NoC_{EST} in terms of agreement and accuracy. For example, two different *SameSOP* labs provided identical NoC_{EST} responses 20% of the time but disagreed 11% of the time; 28% of the time both labs were correct, and 3% of the time both labs were incorrect. If considering only instances in which both labs provided NoC responses, different *SameSOP* labs provided identical NoC_{EST} responses 45% of the time and disagreed 25% of the time; 63% of the time both labs were correct, and 7% of the time both labs were incorrect. Note that intra-lab (same lab) responses were more likely to be identical than inter-lab responses (75% vs. 45% for SameSOP NoC responses), but were also more likely to both be incorrect (17% vs. 7%).

When considering the accuracy and reproducibility of NoC estimates in totality, the conditional probability of a different lab reproducing an incorrect NoC$_{\text{EST}}$ (i.e, two different labs agree in their NoC estimates, but both are incorrect) is higher than the independent probability of an incorrect NoC$_{\text{EST}}$ (i.e., a lab reporting an incorrect NoC estimate). Overall, the incorrect NoC$_{\text{EST}}$ rate is 21.2% (*SameSOP*, weighted), but given an incorrect NoC$_{\text{EST}}$ by a lab on a mixture, the conditional probability that different lab would make an incorrect NoC$_{\text{EST}}$ is 31.0%. Conditional probabilities of intra-lab reproducibility of incorrect NoC$_{\text{EST}}$ responses are even higher at 63.6%, likely due to the notably higher rates of reproducibility in responses. (See *Appendix H3* for details)

		All SameSOP responses				*SameSOP NoC responses*			
		Different labs		*Same lab*		*Different labs*		*Same lab*	
		AllSOP	*SameSOP*	*AllSOP*	*SameSOP*	*AllSOP*	*SameSOP*	*AllSOP*	*SameSOP*
Agreement	Agree (identical)	18%	20%	38%	39%	38%	45%	73%	75%
	Consistent	19%	13%	5%	4%	41%	30%	9%	8%
	Disagree	10%	11%	10%	9%	20%	25%	19%	17%
	1 NoNOC (NotSuit or too complex)	36%	34%	17%	14%				
	Both NoNOC (NotSuit or too complex)	18%	21%	31%	34%				
Accuracy	Both correct	30%	28%	33%	33%	65%	63%	63%	64%
	1 correct, 1 NoNOC	32%	31%	15%	12%				
	1 correct, 1 incorrect	14%	14%	11%	10%	29%	30%	21%	19%
	1 incorrect, 1 NoNOC	4%	4%	2%	2%				
	Both incorrect	3%	3%	8%	9%	6%	7%	16%	17%
	Both NoNOC	18%	21%	31%	34%				

Table 5. Reproducibility of NoC$_{\text{EST}}$ in terms of agreement and accuracy. "Consistent" indicates two NoC ranges that overlap, or an exact NoC value within NoC range. "NoNOC" refers to responses without NoC$_{\text{EST}}$ (*NotSuit* or too complex; this category is used to accommodate participants who provided a *DiffSOP* NoC$_{\text{EST}}$ on *NotSuit* trials). (*InterlabSameSOP* and *IntralabSameSOP datasets*. All results weighted by lab.)

7 Results: Suitability and NoC Estimates for Mixtures with Reference Profiles

As discussed in *Section 4*, the SOPs for most labs require that NoC be assessed before comparison to reference profiles or POIs, but many or most permit NoC to be changed after such comparisons. As discussed in *Section 3.1*, each of the eight mixtures in the *ICSA Subtest* was provided in a comparison packet that included one POI reference profile, and up to two reference profiles (victim, consensual partner, and/or expected contributor). For each mixture, seven-eighths of the assignments included a POI that was present in the mixture (contributor), and one-eighth included a POI that was not present in the mixture (non-contributor).

Figure 5 shows the *NotSuit* and NoC$_{\text{EST}}$ responses for the ICSA mixtures. The contributor and non-contributor (NC) versions of each mixture are paired: for example, ICSA_290 and ICSA_691 are the same mixture, but ICSA_691 used a different POI that was not in the mixture. "Additional contributors" indicates the number of contributors in a mixture after subtracting the contributor POI and reference samples: for example, ICSA_290 was a 2-person contributor mixture that was provided with a reference profile, and therefore (since both contributors were provided) is shown as having 0 additional contributors. Yellow hashed areas in Figure 5 indicate trials for participants whose NoC$_{\text{EST}}$ response was a range or minimum, but used a single NoC value as the basis for their statistical analyses/conclusions (see *Appendix B2i*); 69% of these trials resulted in a correct ICSA NoC basis.

A key takeaway from these results is that assessment of NoC was nontrivial even for ICSA_290, a 2-person 0.088 ng mixture in which a reference profile and the POI were both present (i.e. no additional contributors). Similarly, ICSA_691 and ICSA_078 each have only 1 unknown. We err on the side of caution in making observations for the noncontributor responses due to the very small N: for example, ICSA_691 has 11 *SameSOP* responses weighted down to 5. The single response labeled as "Incorrect -2" appears to have been a user error: this was a trial in which the NoC$_{\text{EST}}$ was a range of 3-8, but their ICSA NoC basis was 2.

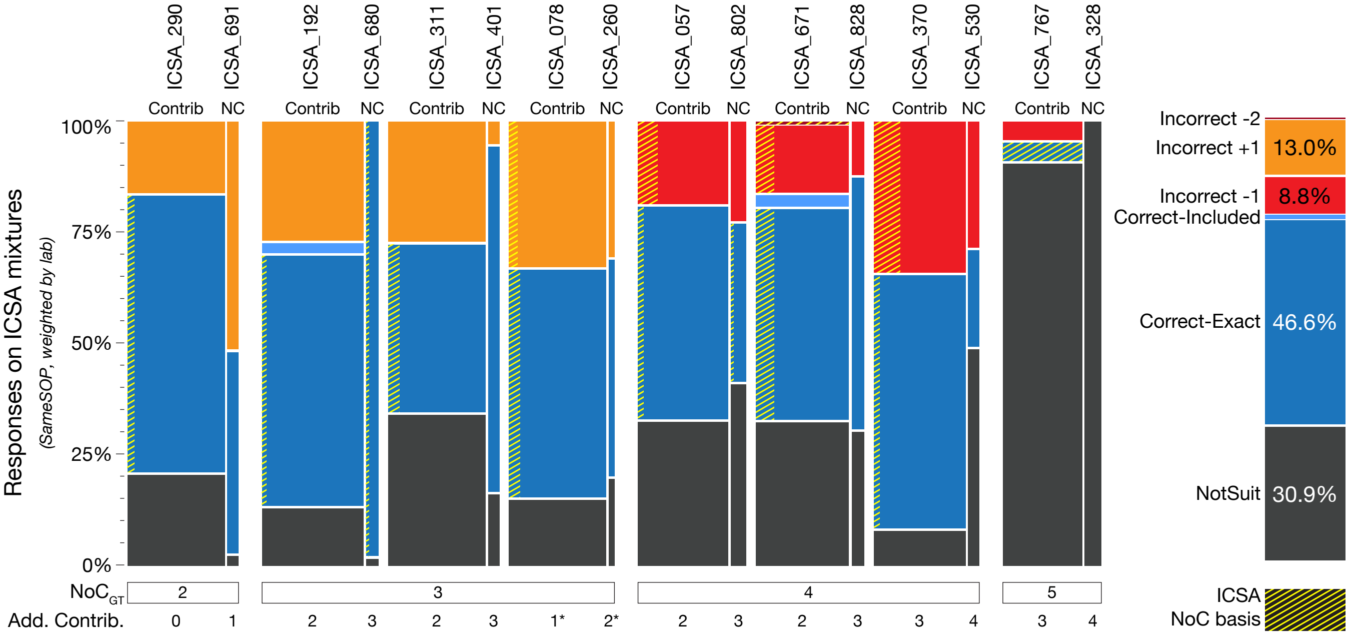

Figure 5. NoC responses for ICSA mixtures, which were provided with Person of interest (POI) profiles and reference profiles. POI profiles were present in the contributor mixtures, not the non-contributor (NC) mixtures. "Additional contributors" indicates the number of contributors in a mixture after subtracting the contributor POI and reference samples. "ICSA NoC basis" indicates responses for which labs provided NoC_{EST} ranges, but indicated that per their SOPs they would use a specific NoC value for statistical analyses. *For ICSA_078/260, 2 reference profiles were provided but only one was present in the mixture (as is common in sexual assault kit casework with reference profiles from the victim and consensual partner, but only the victim is present in the mixture). N=296 weighted responses by lab (1 response per lab per mixture), derived from 667 total SameSOP ICSA responses from 42 labs on 8 mixtures.

8 Results by Laboratory

Figure 6 shows the variation of suitability and NoC_{EST} accuracy by laboratory, grouped by lab type (US local, US state, other). Note the varied distribution of responses for all three lab types. Of the 52 labs that had *SameSOP* suitability responses, four labs replied *NotSuit* on every mixture (all black columns: two US Local and two US State), but none of them completed all 20 assigned mixtures; an additional 11 labs replied *NotSuit* on more than half of their responses. These differences in suitability assessments help to explain the disparity in reproducibility of suitability responses shown in Figure 3. There was a wide distribution of performance: nine labs were incorrect on 25% or more of their responses, and 11 labs had no incorrect NoC estimates; one lab was always correct and never reported *NotSuit* (but only completed 12 of the mixtures; all of that lab's NoC estimates were ranges).

The three lab types were similar in overall NoC_{EST} accuracy rates, but note that US labs overwhelmingly reported a single (exact) NoC_{EST} value (86% exact NoC_{EST} for both US local and US state labs), whereas Other labs overwhelmingly reported NoC_{EST} minima or ranges (33% exact NoC_{EST}).

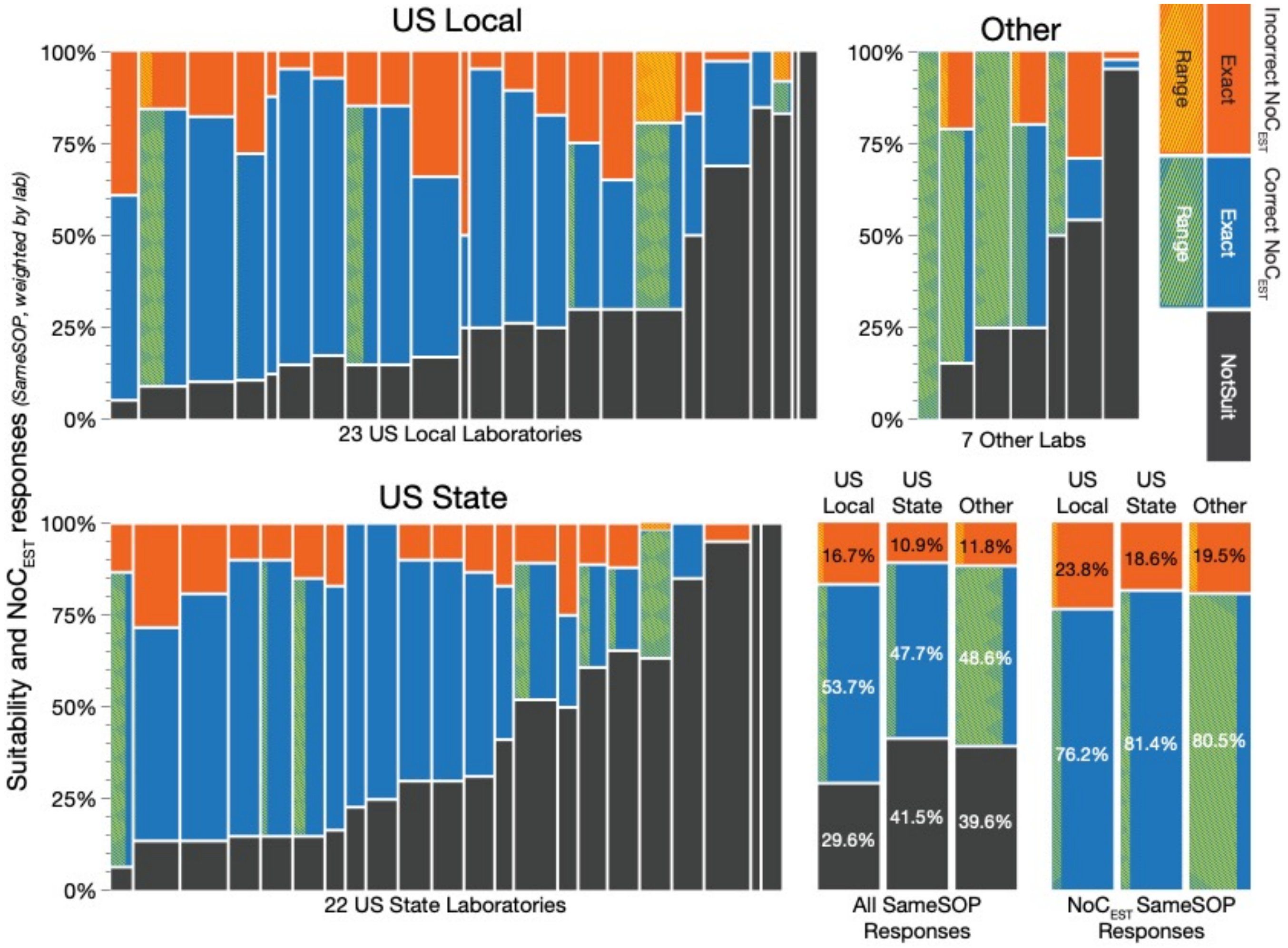

Figure 6. NoC$_{EST}$ accuracy (including *NotSuit* responses) by laboratory, grouped by laboratory type "Other" includes 1 U.S. private lab and 6 non-U.S. labs. (Limited to *SameSOP*; *WeightedNoCSame dataset*)

9 Results by Amplification

As discussed in Section 3.1, we distributed four amplifications of each mixture in order to accommodate the Amp/CE settings of as many participants as possible. Comparing the results between these Amp/CE versions of each mixture is a challenge because these confound stochastic effects, the intrinsic differences between amplification kits, the effects of other Amp/CE settings, and the differences between the labs that use each kit:

- Different amplifications of a given mixture can be expected to vary to some extent due to stochasticity. In this study, to verify that the amplifications were replicable and to minimize stochastic effects, each mixture was amplified at least twice for each Amp/CE setting; the mixture was only used if at least two amplifications were found to be replicable on review, and any differences were within the normal variation expected for the DNA input level. Note that these differences are not limited to this study: different amplifications of a single physical mixture will vary in casework and cannot be expected to be identical. See *Appendix I* for details of variability by Amp/CE.
- Amplifications of a single mixture created using different amplification kits will necessarily vary in content because of different loci. The loci used by Identifiler Plus are a subset of the loci used by GlobalFiler, which (with one exception, Y indel) uses a subset of the loci used in Fusion 6C. GlobalFiler

and Fusion 6C include SE33, which is of particular use in discriminating within mixtures; Fusion 6C includes Penta E and Penta D, which are also highly discriminating (see *Appendix J5* for details of the associations between specific loci and NoC accuracy).[*]

- Using different numbers of cycles (i.e., GF28 vs. GF29) notably affects the resulting levels. This difference in sensitivity may have both benefits and drawbacks—an increase in signal level may potentially be accompanied by an increase in artifacts (and conversely, reducing the number of cycles may reduce artifacts at the cost of signal strength).
- Laboratories that use the same Amp/CE settings may use different ATs.
- There are differences in the labs by Amp/CE: for example, most 6C29 labs (58%) were US state labs; most GF29 labs (63%) were US Local labs; most GF28 labs (67%) were US state labs; all ID28 labs were non-US labs. As discussed above, some—but not all—of the variability among labs can be explained in terms of different P&P responses.

Figure 7 shows the accuracy of NoC_{EST} responses (along with NotSuit responses) by Amp/CE setting. We reviewed nine of the mixtures in detail (starred in Figure 7), flagging mixtures for review if they exhibited unusually high overall incorrect NoC_{EST} rates, or notably different NoC_{EST} rates between Amp/CE versions of a given mixture. The detailed review of these mixtures revealed that they generally exhibited stochastic effects that could explain the differences in responses by Amp/CE setting, including stutter peaks (generally elevated or stacked stutter), heterozygous balance <60%, or dropouts. For example, the high incorrect NoC_{EST} rate for NOC_52 shown in Figure 4 is disproportionately from the GF29 version of that mixture, and may be explained by an elevated/stacked stutter on D8S1179 that is present only in GF29; of the responses to NOC_52 that indicated D8S1179 was a primary basis for their NoC assessment, 73% were incorrect.

The Amp/CE versions of each mixture did vary in terms of exact mixture ratios, proportions of the smallest contributors, and signal strength. However, we found no significant association between these values and rates of incorrect NoC_{EST} or *NotSuit*. See *Appendix I* for detailed results by mixture, and *Appendix I1* for review of individual mixtures.

[*] *Please note that this discussion is not intended to indicate recommendations or criticisms of amplification kits or any Amp/CE settings.*

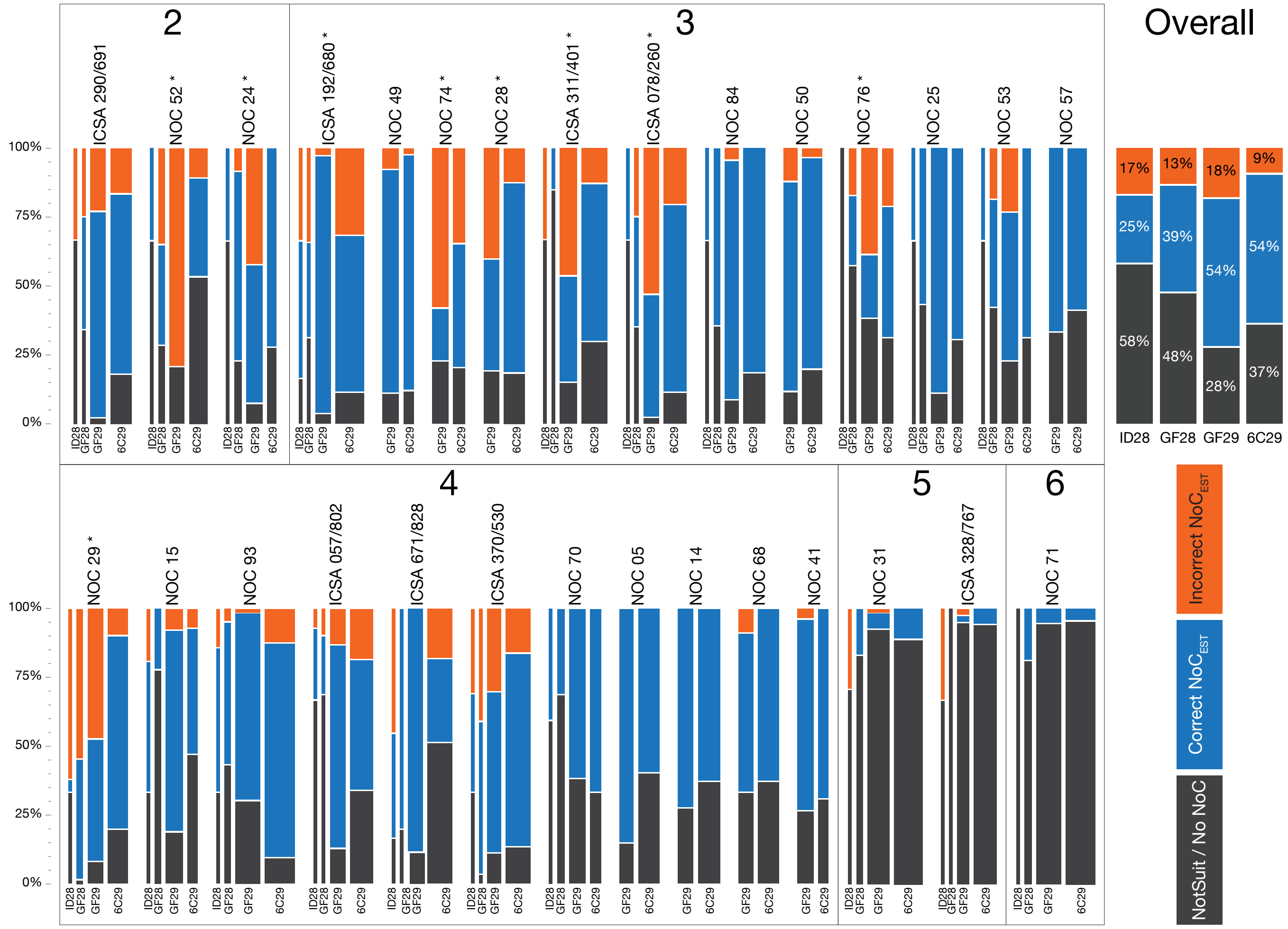

Figure 7. Accuracy of NoC$_{EST}$ shown with suitability responses, by mixture and by Amp/CE. If NotSuit is omitted (i.e., only NoC responses are considered), correct NoC$_{EST}$ rates are ID28:59%, GF28:74%, GF29:75%, 6C29:85%. Starred mixtures are reviewed in detail in *Appendix I1.* (*WeightedSuitSameSOP dataset*)

10 Additional Results

The main focus of this paper is the variability of suitability assessments and NoC estimates for DNA mixtures; however, a number of additional analyses were also performed, the results of which are summarized here and reported in detail in the appendices.

For each mixture, participants were asked if they used analytical thresholds (ATs) and/or stochastic thresholds (STs). Usage of ATs varied widely among participants: of the 67 labs, 35 used a single AT value in all of their responses, and 18 labs indicated their ATs varied by dye channel in all of their responses; the remaining 14 labs specified different AT values or different usage of ATs among their responses. Usage of STs also varied widely among participants: 20 used a single ST value in all of their responses, 2 labs indicated their STs varied by dye channel in all of their responses, and 29 labs indicated they did not use STs; the remaining 16 labs specified different ST values or different usage of STs among their responses. The specific values used for AT and ST varied widely. Trials in which no ST was used were more likely to assess mixtures as *YesSuit* than trials in which STs were used. We found no notable relationships between NoC accuracy and AT/ST usage or values. (See *Appendix J2* for additional details.)

For each assigned mixture, participants were asked if they could identify any major contributors. Responses that indicated major contributors were present were much more likely to make a suitability determination of *PartSuit* (generally for a subset of contributors, alone or in conjunction with loci). However, the ability to identify major contributors did not affect NoC accuracy. (See *Appendix J3* for additional details.)

When participants reported their NoC estimates, they were asked to indicate which factors affected their NoC assessments. Three factors were selected in the majority of trials: Maximum Allele Count (MAC) per locus (87% of *SameSOP* weighted responses), relative peak heights (82% of *SameSOP* weighted responses), and peak heights (RFU) (63% of *SameSOP* weighted responses). The only factor that exhibited a strong association with NoC accuracy was MAC per locus: responses indicating MAC per locus as a factor were 18% incorrect, whereas those not indicating MAC per locus as a factor were 44% incorrect. (See *Appendix J4* for additional details.)

Participants were also asked to indicate the primary loci used as the basis for determining NoC. On average, participants selected 4.9 loci (median 4, range 1-24). Some loci were cited much more frequently than others; this was weakly associated with the discriminating power of the selected locus (e.g., SE33 was selected the most overall, despite the fact that it is not included in ID28; Penta E was the second most commonly cited locus for 6C29). In comparing the primary loci selected to NoC accuracy, we detected support for varying degrees of association between some loci and the incorrect NoC_{EST} rate. For example, on weighted *SameSOP* trials, participants who reported that SE33 was used as a basis for estimating NoC had notably lower incorrect NoC_{EST} rates as compared to those who did not use the locus (17% versus 31%). (See *Appendix J5* for additional details).

11 Conclusions

DNAmix 2021 was conducted to evaluate the extent of consistency and variation among laboratories in the interpretation of DNA mixtures, and to assess the effects of various potential sources of variability. Here, we report on suitability assessments and number of contributors estimates for 29 DNA mixtures, reported by 134 participants representing 67 laboratories, and encompassing 2,272 total responses. In particular, this study sought to characterize the extent of variability in interpretations of DNA mixture profiles starting from the electropherogram; although every stage of the processing and analysis of DNA samples can result in variability, this study does not address variability in stages prior to the analysis of the electropherogram. While we made every effort to produce electropherograms that were as close as possible to those that a lab would typically analyze, the participating labs did not actually conduct the DNA workflow (from extraction through capillary electrophoresis); it is important to acknowledge this limitation because this does constitute a deviation from lab SOPs and could contribute additional variation than might be observed in casework. The differences in the Amp/CE versions of each mixture may explain some of the differences among labs in their suitability or NoC assessments, but note that this is also true in casework: different amplifications of a single physical mixture can be expected to vary in casework, and use of different amp kits or settings can be expected to result in differences.

Overall, the results show the extent of variation in suitability and NoC assessments among laboratories that conduct analyses of DNA mixtures. Labs exhibited a notable variation in the policies and procedures that govern suitability and NoC assessments. Suitability assessments cannot be assessed as correct or incorrect—whether a mixture is suitable for analysis is a decision by an analyst based on laboratory policies. Here we reported the extent to which participating laboratories vary in these policies related to suitability assessments. We observed notable variation in whether labs following their SOPs (*SameSOP*) would assess a given mixture as suitable or not: if two labs were given the same mixture, they agreed on whether the mixture was suitable 66% of the time; U.S. labs agreed 69% of the time. A decision that a given mixture is not suitable (*NotSuit*) means a laboratory would not interpret or report results for that mixture: disagreements among labs about one third of the time regarding whether or not a DNA mixture will be interpreted would be notable if these rates occurred operationally.

Unlike suitability, laboratory assessments of number of contributors (NoC_{EST}) can be evaluated in terms of accuracy against ground truth. Overall, 79% of *SameSOP* NoC_{EST} responses were correct—when NoC estimates were specific values, 24% were incorrect; when NoC estimates were ranges or minima, 8% were incorrect. When two different labs provided *SameSOP* NoC responses, 63% of the time both labs were correct, and 7% of the time both labs were incorrect (U.S. labs had the same rates). Incorrect NoC assessments do not necessarily imply inaccurate interpretations, conclusions, or statistical analyses for a mixture: since NoC estimates are used as a parameter in probabilistic genotyping, incorrect NoC estimates may have an effect on the resulting likelihood ratios. Most incorrect NoC estimates were overestimates, which previous research [2] has shown have less of an effect on likelihood ratios than underestimates.

Given the range of attributes of mixtures in the study, it is reasonable to consider whether NoC_{GT} is discernable for a given mixture. Although it may appear desirable to define the number of contributors that are discernable for each mixture, doing such would require multiple assumptions regarding settings that would be problematic. For example, detectability of a contributor depends (in part) on a lab's use of an analytical threshold (AT), but ATs vary by Amp/CE; labs may use a single AT, vary from sample to sample, vary by dye channel, and some labs are permitted to use peaks below AT in assessing NoC_{EST} (see Section 10). We cannot evaluate whether NoC_{GT} is discernable for a given mixture in an absolute sense without making inappropriate assumptions—we can, however, rely on the responses from participants to assess the practiciality of assessing NoC_{EST}. For the five- and six-person mixtures in this study, we can conclude that assessing NoC_{EST} precisely is not a reasonable expectation: these collectively received less than one weighted SameSOP correct exact NoC_{EST} response. However, every two- to four-person mixture received multiple correct exact NoC_{EST} responses: for 23 of those 26 mixtures, over 30% of weighted SameSOP responses were correct exact NoC_{EST}. If we consider each of the Amp/CE variations of each two- to four-person mixture (n=86), all but four had at least some correct exact NoC_{EST} SameSOP responses (and those could be attributed to small counts, as three of those four had six or fewer weighted responses).

This study was conducted for use by practitioners and laboratory managers to consider in assessing operational policies, training, and QA procedures, as well as by policy makers and the legal community in understanding these aspects of the DNA mixture interpretation process. The rates reported here are intended to serve as general estimates to assist in decision making and in determining how to improve operational procedures and standardization for DNA mixture analysis. The participating laboratories represent a cross-section of the forensic DNA community, acquired through convenience sampling (participating laboratories volunteered to participate and optionally enrolled more than one subunit). The samples included in this study were designed to be as broadly representative of casework as feasible; samples were selected to encompass a range of DNA mixture attributes, including number of contributors, contributor ratios, total quantity of DNA, presence of degradation, degree of allele sharing, and overall complexity. Participating laboratories were provided electropherograms as .HID files and did not process the DNA samples themselves. Therefore, it is important to note that these results may not be representative of all forensic DNA laboratories, analysts, or mixture casework.

Acknowledgements

We thank the laboratories and DNA analysts who participated in this study, as well as the DNAmix Working Group and Advisory Group for their expertise and detailed guidance in design of the study (see *Appendix B1* for the members of each); NIST, Maryland State Police, New York State Police, and the National Center for Forensic Science/University of Central Florida for providing letters in support of this proposal to NIJ; and NIST Applied Genetics Group for providing DNA samples. The opinions, findings, and conclusions or recommendations expressed in this program are those of the authors and do not necessarily reflect those of the Department of Justice. Names of commercial manufacturers and product names are included for the systems that are frequently used in the laboratories that registered for the DNAmix 2021 study; inclusion does not imply endorsement of the manufacturer or its products or services by the authors, Bode, Noblis, or the National Institute of Justice.

Funding: This project was supported by Award No. 2020-R2-CX-0049, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. **Competing interests:** Authors declare no competing interests. **Data and materials availability:** All response data and electropherogram data are available in the main text, supplementary materials, or archived at OSF [27]; because participants and laboratories were assured of anonymity, results by participant/lab are summarized or deidentified to prevent reidentification. **Author contributions:** Conceptualization: RAH,RAB,JMD; Methodology: All; Validation: RAH,NR,BLE,JMD; Formal analysis: RAH,NR,BLE; Investigation: RAH,NR,BLE,JMD; Resources (DNA sample collection): RAB,JMD; Data Curation: RAH,NR,JMD; Writing–Original Draft: RAH,NR,BLE,JMD; Writing–Review & Editing: All; Visualization: RAH,NR; Project administration: RAH,RAB; Funding acquisition: RAH,RAB,JMD.

References

[1] J.M. Butler, M.C. Kline, M.D. Coble, NIST interlaboratory studies involving DNA mixtures (MIX05 and MIX13): Variation observed and lessons learned, Forensic Sci Int Genet. 37 (2018) 81–94. https://doi.org/10.1016/j.fsigen.2018.07.024.

[2] J.-A. Bright, R. Richards, M. Kruijver, H. Kelly, C. McGovern, A. Magee, A. McWhorter, A. Ciecko, B. Peck, C. Baumgartner, C. Buettner, S. McWilliams, C. McKenna, C. Gallacher, B. Mallinder, D. Wright, D. Johnson, D. Catella, E. Lien, C. O'Connor, G. Duncan, J. Bundy, J. Echard, J. Lowe, J. Stewart, K. Corrado, S. Gentile, M. Kaplan, M. Hassler, N. McDonald, P. Hulme, R.H. Oefelein, S. Montpetit, M. Strong, S. Noël, S. Malsom, S. Myers, S. Welti, T. Moretti, T. McMahon, T. Grill, T. Kalafut, M.M. Greer-Ritzheimer, V. Beamer, D.A. Taylor, J.S. Buckleton, Internal validation of STRmix^{TM} – A multi laboratory response to PCAST, Forensic Sci Int Genet. 34 (2018) 11–24. https://doi.org/10.1016/j.fsigen.2018.01.003.

[3] E. Rogers, R. Aranda, P.M. Spencer, D.R. Myers, DNA Mixture Study: Novel Metrics to Quantify the Intra- and Inter-Laboratory Variability in Forensic DNA Mixture Interpretation (Report # 304317), 2022.

[4] National Research Council, Strengthening Forensic Science in the United States: A Path Forward, The National Academies Press, Washington, D.C., 2009.

[5] President's Council of Advisors on Science and Technology (PCAST), Report to the President. Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods, Executive Office of the President, Washington, D.C., 2016.

[6] U.S. Government Accountability Office (GAO), Science & Tech Spotlight: Probabilistic Genotyping Software, (2019).

[7] J.M. Butler, H. Iyer, R. Press, M.K. Taylor, P.M. Vallone, S. Willis, DNA Mixture Interpretation: A NIST Scientific Foundation Review (NISTIR 8351-DRAFT), 2021.

[8] M.D. Coble, J.-A. Bright, Probabilistic genotyping software: An overview, Forensic Sci Int Genet. 38 (2019) 219–224. https://doi.org/10.1016/j.fsigen.2018.11.009.

[9] G. Hampikian, Correcting forensic DNA errors, Forensic Sci Int Genet. 41 (2019) 32–33. https://doi.org/10.1016/j.fsigen.2019.03.005.

[10] B. Mallinder, S. Pope, J. Thomson, L.-A. Beck, A. McDonald, D. Ramsbottom, D.S. Court, D. Vanhinsbergh, M. Barber, I. Evett, K. Sullivan, J. Whitaker, Interpretation and reporting of mixed DNA profiles by seven forensic laboratories in the UK and Ireland, Forensic Science International: Genetics. 58 (2022) 102674. https://doi.org/https://doi.org/10.1016/j.fsigen.2022.102674.

[11] Scientific Working Group on DNA Analysis Methods (SWGDAM), SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories, (2017).

[12] L.M. Brinkac, N. Richetelli, J. Davoren, R. Bever, R.A. Hicklin, DNAmix 2021: Variation in Laboratory Policies, Procedures, and Casework Scenario Decisions, Data in Brief. 48 (2023). https://doi.org/10.1016/j.dib.2023.109150.

[13] T. Bille, S. Weitz, J.S. Buckleton, J.-A. Bright, Interpreting a major component from a mixed DNA profile with an unknown number of minor contributors, Forensic Sci Int Genet. 40 (2019) 150–159.

[14] Texas Forensic Science Commission, Final Report on Complaint No.21.54 James Smiley (Signature Science, LLC; Forensic Biology/DNA), 2022.

[15] E. Gillespie, Queensland is reviewing thousands of DNA samples connected to serious crimes. Here's why, The Guardian. (2022). https://www.theguardian.com/australia-news/2022/sep/22/queensland-is-reviewing-thousands-of-dna-samples-connected-to-serious-crimes-heres-why (accessed September 27, 2022).

[16] Texas Forensic Science Commission, Final Report National Medical Services, Inc. (NMS) DNA Analysis in Case of U.S. v. Torney, 2018.

[17] Texas Forensic Science Commission, Final Audit Report for Austin Police Department Forensic Services Division DNA Section, 2016.

[18] T. Bille, J.-A. Bright, J. Buckleton, Application of Random Match Probability Calculations to Mixed STR Profiles, J Forensic Sci. 58 (2013) 474–485. https://doi.org/10.1111/1556-4029.12067.

[19] F.R. Bieber, J.S. Buckleton, B. Budowle, J.M. Butler, M.D. Coble, Evaluation of forensic DNA mixture evidence: protocol for evaluation, interpretation, and statistical calculations using the combined probability of inclusion, BMC Genet. 17 (2016) 1–15.

[20] D. Taylor, J.-A. Bright, J. Buckleton, Interpreting forensic DNA profiling evidence without specifying the number of contributors, Forensic Sci Int Genet. 13 (2014) 269–280.

[21] J.S. Buckleton, J.-A. Bright, K. Cheng, H. Kelly, D.A. Taylor, The effect of varying the number of contributors in the prosecution and alternate propositions, Forensic Sci Int Genet. 38 (2019) 225–231.

[22] C.C.G. Benschop, H. Haned, L. Jeurissen, P.D. Gill, T. Sijen, The effect of varying the number of contributors on likelihood ratios for complex DNA mixtures, Forensic Sci Int Genet. 19 (2015) 92–99. https://doi.org/10.1016/j.fsigen.2015.07.003.

[23] J.-A. Bright, K. Cheng, Z. Kerr, C. McGovern, H. Kelly, T.R. Moretti, M.A. Smith, F.R. Bieber, B. Budowle, M.D. Coble, R. Alghafri, P.S. Allen, A. Barber, V. Beamer, C. Buettner, M. Russell, C. Gehrig, T. Hicks, J. Charak, K. Cheong-Wing, A. Ciecko, C.T. Davis, M. Donley, N. Pedersen, B. Gartside, D. Granger, M. Greer-Ritzheimer, E. Reisinger, J. Kennedy, E. Grammer, M. Kaplan, D. Hansen, H.J. Larsen, A. Laureano, C. Li, E. Lien, E. Lindberg, C. Kelly, B. Mallinder, S. Malsom, A. Yacovone-Margetts, A. McWhorter, S.M. Prajapati, T. Powell, G. Shutler, K. Stevenson, A.R. Stonehouse, L. Smith, J. Murakami, E. Halsing, D. Wright, L. Clark, D.A. Taylor, J. Buckleton, STRmix^{TM} collaborative exercise on DNA mixture interpretation, Forensic Sci Int Genet. 40 (2019) 1–8. https://doi.org/10.1016/j.fsigen.2019.01.006.

[24] J.S. Buckleton, J.M. Curran, P. Gill, Towards understanding the effect of uncertainty in the number of contributors to DNA stains, Forensic Sci Int Genet. 1 (2007) 20–28. https://doi.org/10.1016/j.fsigen.2006.09.002.

[25] J.-A. Bright, J.M. Curran, J.S. Buckleton, The effect of the uncertainty in the number of contributors to mixed DNA profiles on profile interpretation, Forensic Sci Int Genet. 12 (2014) 208–214.

[26] J.-A. Bright, D. Taylor, J. Curran, J. Buckleton, Searching mixed DNA profiles directly against profile databases, Forensic Sci Int Genet. 9 (2014) 102–110.

[27] Scientific Working Group on DNA Analysis Methods (SWGDAM), SWGDAM Comments on NISTIR 8351-DRAFT Entitled DNA Mixture Interpretation: A NIST Scientific Foundation Review, 2021.

[28] Noblis, DNAmix 2021 Archive (V1, Feb 2023), (2023). https://doi.org/http://doi.org/10.17605/OSF.IO/B3MZW.

[29] P. Gill, R. Sparkes, C. Kimpton, Development of guidelines to designate alleles using an STR multiplex system, Forensic Sci Int. 89 (1997) 185–197.

[30] D.R. Paoletti, T.E. Doom, C.M. Krane, M.L. Raymer, D.E. Krane, Empirical analysis of the STR profiles resulting from conceptual mixtures, J Forensic Sci. 50 (2005).

[31] I.W. Evett, P.D. Gill, J.A. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, Journal of Forensic Science. 43 (1998) 62–69.

[32] J.S. Buckleton, C.M. Triggs, S.J. Walsh, Forensic DNA Evidence Interpretation, CRC Press, Boca Raton, FL, 2004. https://doi.org/10.1201/9781420037920.

[33] M. Kruijver, H. Kelly, K. Cheng, M.-H. Lin, J. Morawitz, L. Russell, J. Buckleton, J.-A. Bright, Estimating the number of contributors to a DNA profile using decision trees, Forensic Sci Int Genet. 50 (2021) 102407. https://doi.org/10.1016/j.fsigen.2020.102407.

[34] N. Hu, B. Cong, S. Li, C. Ma, L. Fu, X. Zhang, Current developments in forensic interpretation of mixed DNA samples, Biomed Rep. 2 (2014) 309–316.

[35] A. Biedermann, S. Bozza, K. Konis, F. Taroni, Inference about the number of contributors to a DNA mixture: comparative analyses of a Bayesian network approach and the maximum allele count method, Forensic Sci Int Genet. 6 (2012) 689–696.

[36] H. Haned, L. Pène, F. Sauvage, D. Pontier, The predictive value of the maximum likelihood estimator of the number of contributors to a DNA mixture, Forensic Sci Int Genet. 5 (2011) 281–284.

[37] H. Haned, L. Pene, J.R. Lobry, A.B. Dufour, D. Pontier, Estimating the number of contributors to forensic DNA mixtures: does maximum likelihood perform better than maximum allele count?, J Forensic Sci. 56 (2011) 23–28.

[38] M.D. Coble, J.-A. Bright, J.S. Buckleton, J.M. Curran, Uncertainty in the number of contributors in the proposed new CODIS set, Forensic Sci Int Genet. 19 (2015) 207–211.

[39] L.A. Borsuk, K.B. Gettings, C.R. Steffen, K.M. Kiesler, P.M. Vallone, Sequence-based US population data for the SE33 locus, Electrophoresis. 39 (2018) 2694–2701. https://doi.org/10.1002/elps.201800091.

[40] H. Swaminathan, C.M. Grgicak, M. Medard, D.S. Lun, NOCIt : A computational method to infer the number of contributors to DNA samples analyzed by STR genotyping, Forensic Sci Int Genet. 16 (2015) 172–180. https://doi.org/10.1016/j.fsigen.2014.11.010.

[41] C.M. Grgicak, S. Karkar, X. Yearwood-Garcia, L.E. Alfonse, K.R. Duffy, D.S. Lun, A large-scale validation of NOCIt's a posteriori probability of the number of contributors and its integration into forensic interpretation pipelines, Forensic Sci Int Genet. 47 (2020) 102296. https://doi.org/10.1016/j.fsigen.2020.102296.

[42] M.A. Marciano, J.D. Adelman, PACE: Probabilistic Assessment for Contributor Estimation— A machine learning-based assessment of the number of contributors in DNA mixtures, Forensic Sci Int Genet. 27 (2017) 82–91. https://doi.org/10.1016/j.fsigen.2016.11.006.

[43] M.-H. Lin, S.-I. Lee, X. Zhang, L. Russell, H. Kelly, K. Cheng, S. Cooper, R. Wivell, Z. Kerr, J. Morawitz, J.-A. Bright, Developmental validation of FaSTR^{TM} DNA: Software for the analysis of forensic DNA profiles, Forensic Science International: Reports. 3 (2021) 100217. https://doi.org/10.1016/j.fsir.2021.100217.

[44] B. Budowle, A.J. Onorato, T.F. Callaghan, A. Della Manna, A.M. Gross, R.A. Guerrieri, J.C. Luttman, D.L. McClure, Mixture Interpretation: Defining the Relevant Features for Guidelines for the Assessment of Mixed DNA Profiles in Forensic Casework, J Forensic Sci. 54 (2009) 810–821. https://doi.org/10.1111/j.1556-4029.2009.01046.x.

[45] S. Norsworthy, D.S. Lun, C.M. Grgicak, Determining the number of contributors to DNA mixtures in the low-template regime: Exploring the impacts of sampling and detection effects, Leg Med. 32 (2018) 1–8. https://doi.org/10.1016/j.legalmed.2018.02.001.

[46] C. McGovern, K. Cheng, H. Kelly, A. Ciecko, D. Taylor, J.S. Buckleton, J.-A. Bright, Performance of a method for weighting a range in the number of contributors in probabilistic genotyping, Forensic Sci Int Genet. 48 (2020) 102352. https://doi.org/10.1016/j.fsigen.2020.102352.

[47] C.H. Brenner, DNA View User's Manual, (2019).

[48] C.R. Steffen, M.D. Coble, K.B. Gettings, P.M. Vallone, Corrigendum to "U.S. Population Data for 29 Autosomal STR Loci" [Forensic Sci. Int. Genet. 7 (2013) e82-e83]., Forensic Sci Int Genet. 31 (2017) e36–e40. https://doi.org/10.1016/j.fsigen.2017.08.011.

[49] NIST, STRBase Standard Reference Data 130, (n.d.).

[50] P. Samarati, L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression, 1998. https://epic.org/wp-content/uploads/privacy/reidentification/Samarati_Sweeney_paper.pdf.

Variation in Assessments of Suitability and Number of Contributors for DNA Mixtures

Supplemental Information — Appendices

This PDF file includes:

Other Supplementary Materials for this manuscript include the following:

- Supplementary Data 1: Suitability and NoC Response Data—Spreadsheet containing participants' de-identified responses.
- Supplementary Data 2: Instructions for DNAmix (all instruction documents, glossary, and frequently-asked questions)

The companion document "DNAmix 2021: Variation in Laboratory Policies, Procedures, and Casework Scenario Decisions" [12] provides a description of the study design, and results for the first two phases of DNAmix 2021.

The electropherograms for the 29 mixtures (including all Amp/CE versions of the mixtures) are archived as PDF images at OSF [28](http://doi.org/10.17605/OSF.IO/B3MZW). The EPGs in .HID format are available upon request, but require a signed data use agreement: see the OSF site for details.

Appendix A Background: DNA Mixture Assessments of Suitability and Number of Contributors

The primary purpose of the interpretation (and associated statistical analyses) of DNA mixtures is to evaluate whether the DNA profile for a given individual is (or is not) a contributor to a given mixture. Interpretation can only occur if the laboratory/analyst determines that the mixture is suitable for comparison, and therefore differences among laboratories in suitability assessments for a given mixture would be a fundamental source of variability, resulting in differences among laboratories in whether interpretations are reported at all. In some legal cases, crime scene DNA samples were assessed as not suitable and not tested by laboratories, but the initial suitability assessments were contested on later review (e.g., [14,15]).

In addition to being considered during suitability assessment, NoC estimates may also have a significant impact on interpretations and statistical analyses. NoC estimates may be used to aid in selecting a statistical method for analysis (e.g., in deciding to use a binary approach for simple mixtures and PGS for more complex mixtures) and are also generally required as inputs for statistical analysis (either explicitly or implicitly). For probabilistic genotyping, the number of contributors is often set by the analyst and used as an assumption by the software for deconvolution and ultimately computing the likelihood ratio (LR) [8]. For binary approaches, such as random match probability (RMP) [18] and combined probability of inclusion/exclusion (CPI/CPE) [19], the estimated number of contributors is implicitly considered by an analyst when determining possible genotypes of contributors and doing computations. Variations in NoC assessments have been shown to be a key factor explaining differences in LRs [23], and in some cases a difference in NoC may even change whether a person of interest is included or excluded [20].

While the analysis of single-source DNA is considered "gold standard" of forensic science [5], DNA mixtures are more problematic, particularly when mixtures include three or more contributors, have low total amounts of DNA (low template), are degraded, or the contributors in a mixture share alleles — generally referred to as complex mixtures [5,7] Assessments of suitability and NoC are notably more challenging for complex mixtures than for simple mixtures. For simple mixtures (two-person undegraded mixtures with a substantial amount of DNA available and minimal sharing of alleles), suitability has generally not been in question [5,7]; for complex mixtures, deciding whether a sample is suitable for analysis is dictated by lab policies and validated procedures, with the final decision made based on the judgement of experts. Assessing NoC in complex mixtures may be challenging because it can be difficult to differentiate artifacts from alleles in degraded or low template mixtures; this difficulty increases with additional contributors, due to issues such as allele stacking and determining if a peak is allele or stutter. Even for DNA samples that appear to have a single source, assessing NoC can be uncertain due to the potential for allele masking and drop-out [21]. With improved sensitivity for DNA testing and increased collection of trace samples (sometimes described as "touch DNA") from crime scenes, complex mixtures are expected to be encountered more frequently in casework. In addition, the availability of probabilistic genotyping software (PGS) now allows laboratories to report statistical analyses of more complex mixtures than would have been possible previously. Given these recent methodological and technological developments, the NRC, PCAST, GAO, and NIST reports all affirm that additional research is needed to establish the scientific validity of interpretations of complex DNA mixtures.

Appendix A1 Suitability

The assessment of suitability is an early key decision made in DNA mixture analysis: deciding whether a mixture is suitable is an assessment that the content of the mixture is sufficient for interpretation, and is worth moving forward to comparison and statistical analysis. Suitability decisions are based on a number of sample-specific characteristics, several of which are often specified explicitly in laboratories' standard operating procedures (SOPs), particularly the quantity of DNA (template amount), high estimated NoC, uncertainty of NoC, or the presence or absence of degradation or inhibition. Suitability decisions may also be based on the discretion of the analysts, as a holistic assessment of mixture complexity or quality. Note that since laboratory SOPs may differ in the thresholds used to make suitability decisions, we can expect some amount of variability in these decisions among laboratories. Some variability in suitability decisions may be expected even within laboratories, given that some suitability decisions are at the analysts' discretion. Note that suitability decisions can be evaluated with respect to the extent of variability, or a specific decision can be judged against the SOPs defined by that laboratory, but a decision cannot be judged as "correct" or "incorrect" in an absolute sense, since there is no overarching standard for assessing suitability.

Operationally, suitability decisions have occasionally been key aspects of legal cases. Audits and re-analyses of evidence in several legal cases revealed that there were a number of cases in which a sample was assessed as unsuitable for comparison/statistical analysis, but later found to yield useful data [14,15] (and conversely wherein a sample was interpreted that should not have been given quality issues [16,17]). For example, the Texas Forensic Science Commission recently investigated several court cases involving the suitability step of mixture interpretation. In the Smiley case, the Commission found that they analyst was correct for initially reporting a complex mixture profile as not suitable for comparison, based upon the lab's SOPs; however, the protocols failed to include an exclusionary assessment step for mixtures deemed not suitable, which ultimately led to failure to properly exculpate the suspect. The foundation did not cite the analyst as professionally negligent because the analyst followed the laboratory protocol, rather the Commission recommended that the laboratory establish protocols which indicate that an indistinguishable mixture of 4 or more contributors should be designated as unsuitable for comparison, but the profile should be examined for exclusionary data. Subsequent reanalysis of the complex mixture profile using STRmix resulted in the reported exclusion of the suspect; however, the commission found that the initial interpretation of the evidence should have resulted in the suspect being excluded (Texas Forensic Science Commission, 2022).

Appendix A2 Number of Contributors (NoC)

Appendix A2a How NoC is estimated

There are a variety of methods in which number of contributors to a DNA mixture are evaluated by an analyst, including both manual and software-based techniques (although the latter is rarely used by participants in this study [12]). With respect to manual methods, a variety of sample-specific indicators are generally considered by analysts (an average of nine factors were selected by participating laboratories for PP#38 in the *P&P Questionnaire* [12]). One of the most commonly-used methods for evaluating NoC is maximum allele count (MAC), which relies on an assumption that each heterozygote locus will contain two alleles. The maximum number of alleles at a given locus is divided by two and then rounded up to determine the minimum number of individuals necessary to explain the number of peaks in the electropherogram [29]. However, MAC suffers from limitations given that it is a purely mathematical/count-based estimate, and can be unreliable when there are several contributors, contributors with tri-alleles, loci with a small number of detectable alleles, allele sharing amongst contributors, or when laboratory policies allow for discarding anomalous data, such as an additional allele at a locus that seems inconsistent with the number of alleles observed at other loci [30]. When estimating contributors, many labs also manually examine peak heights for signs of stutter and consider peak ratios in addition to counting alleles [21]. Using quantitative information such as peak height or peak area may improve NoC estimates, though quantitative information can also be unreliable due to measurement errors and PCR amplification issues that can result in misleading proportions [29]. Quantitative data may also be uninformative when the mixture is comprised of individuals with similar contributions [30].

Some approaches that leverage additional quantitative/statistical data to determine NoC include normal approximation (e.g., [31]), Monte Carlo methods (e.g., [32]) and decision trees (e.g., [33]), however analysts often use a binary approach that relies on a set of rules and/or personal judgment and experience to include or exclude [24]. Hu et. al. (2014) [34] also described two alternatives to MAC that rely on Bayes' theorem: a probabilistic approach proposed by Biedermann et. al. (2012) [35] for inferring the number of contributors and a predictive value (PV) described by Haned et. al. (2011) [36,37] that could help characterize uncertainty.

However, these methods are more complicated than MAC, which may make them more difficult to explain in court and thereby limit their usage by examiners.

Several profile-specific factors can complicate the estimation of NoC. For example, degraded or low template DNA can lead to allele drop-out, which could cause underestimation, while artifacts and allele drop-in can cause overestimation. Degradation can occur when too much time elapses between a crime and collection of DNA samples, or due to environmental conditions such as exposure to light, heat, or high humidity; degradation can lead to incomplete profile information. Low template DNA occurs more frequently due to more sensitive methods and an increase in trace DNA collected from property crimes. In addition, as the number of contributors to a sample increases and/or the amount of allele stacking increases, underestimation may be more likely to occur.

Appendix A2b Variation in NoC Estimates

Given the importance of NoC and the challenges in estimating contributors, efforts have been made to characterize the likelihood of incorrect NoC estimates using the MAC method. Early studies showed that mischaracterization of NoC was relatively common with a limited number of loci, particularly as the true number of contributors increased [24,29,30]. However, the use of additional loci has been shown to reduce the underestimation of NoC. In a study conducted by Coble et. al., (2015) [38], simulated 4-person mixtures were found to appear as 3-person for ~79% of mixtures with the 13 CODIS loci, ~43% for the 20 CODIS loci, ~34% with the CODIS 20 + Penta D&E, and 16% with the CODIS 20 + SE33. Furthermore, the use of highly polymorphic loci, such as SE33 (0.9921 power of discrimination for Caucasians [39]), decreases the chance of NoC underestimation. Simulations using different populations produced slightly different percentages however the trend of more loci reducing underestimation was consistent.

It is worth noting that these studies did not include peak heights or area, which is often used by analysts in conjunction with allele counts to estimate NoC. Therefore, in order to determine how much estimates may vary in practice, it is useful to evaluate the frequency of overestimation and underestimation on actual DNA mixture samples, rather than via simulation. Bright et. al. (2018) [2] evaluated the probability of a "hidden contributor" (i.e., the apparent number of contributors being lower than the true number of contributors) on actual DNA mixture samples— they report that probability of a five-person mixture appearing as a five-person mixture in the profile (and therefore not having a hidden contributor) was 36%, which is much higher than the previous estimate of less than 1% reported by Coble et. al. (2015) [38] via simulation study considering only the MAC. With lower levels of DNA template, the difficulty in assigning the NoC increases, especially in cases where allelic drop-in and/or drop-out have occurred. The use of MAC can be improved by considering the signal for each allele, however as the number of contributors increases, the number of genotype combinations to consider can become impractical for manual interpretation. In recent years, several software packages have been developed and validated to assist in the estimation of NoC, such as NOCIt [40,41], PACE [42], and FaSTR [43]; however, based upon the results of the *P&P Questionnaire* in this study, these software tools are rarely used in practice [12].

Appendix A2c Impact of NoC uncertainty or errors

Several studies have examined the impact of NoC uncertainty or errors on the resulting statistics (e.g., [2,13,20–25]). Specifically, LRs are closer to neutral (one) and less likely to discriminate between true and false donors as the number of contributors to a mixture increases [2,22,23,25]. Overestimation of NoC can result in lower LRs for true donors and higher LRs for false donors, potentially leading to an adventitious match, though typically with low LR values [2,22,23]. Underestimation was determined to be more likely than overestimation, with three main causes: a "tiny minor" with insignificant contribution to the profile, a "hidden contributor" masked by other contributors (most likely relatives), or a "low level donors' scenario" in which both masking and dropout lead to underestimation [2]. Underestimation can lead to decreased LRs for true donors and false exclusions [2,22,23]. LRs for minor contributors are more notably affected by variation in NoC estimates [13,25].

In addition to evaluating the impact of overall estimation of number of contributors to a mixture, previous works have also examined the impact of assigning different NoC to each proposition in an LR [21,22,24]. Since it is not required that the two alternative hypotheses of an LR (which we refer to here as the prosecutor hypothesis (Hp) and the defense hypothesis (Hd)) to assume the same number of contributors, it is important to examine the potential impact on LR of varying each individually. In general, LRs may be inflated when there is an extra contributor added for both propositions, but the addition of an extra contributor for only Hp may have minimal consequences [21]. Conversely, assigning an unreasonable number of contributors (in either direction) to the defense hypothesis as an alternative explanation has been shown to generally increase the LR in favor of Hp [44].

While probabilistic genotyping can improve the objectivity and consistency of DNA analysis, there is still variation in estimated NoC, which can cause variation in interpretation and reported statistics, especially for minor contributors [2,13,22,23,25]. Much of this variation in practice arises for more complex mixtures, with factors such as high number of contributors, low template, and/or degradation or inhibition; however even samples with a single contributor may produce variation in NoC estimates. Research suggests that assessing the mixture under multiple assumed numbers of contributors or across a range of number of contributors may produce the most reasonable statistics [13,45]; several techniques and softwares have recently incorporated this approach (including STRmix [46] and DNAView [47]).

Appendix B DNAmix 2021 Study Overview

This appendix provides a brief summary of the DNAmix 2021 *study; see the companion document* [12] *for a detailed description of the overall study design.*

The *Inter-laboratory Variation in Interpretation of DNA Mixtures Study* (*DNAmix 2021*) was a large-scale independent study conducted to evaluate the extent of consistency and variation among forensic laboratories in interpretations, comparisons, and statistical analyses of DNA mixtures, and to assess the effects of numerous potential sources of variability. Noblis and Bode Technology conducted the study under National Institute of Justice (NIJ) grant award # 2020-R2-CX-0049.

The study was conducted in four phases:

1. *Policies and Procedures (P&P) Questionnaire* — Online questionnaire to assess laboratory policies and procedures relevant to DNA mixture interpretation (notably systems, types of statistics reported, and parameter settings used).
2. *Casework Scenario Questionnaire* — Online questionnaire to assess analysis procedures or decisions that may vary depending upon the case scenario, and to assess the nature of mixture casework.
3. *Number of Contributors (NoC) Subtest* — Assessment of suitability and number of contributors, given electropherogram data for 12 mixtures.
4. *Interpretation, Comparison, and Statistical Analysis (ICSA) Subtest* — Interpretations and statistical analyses, given electropherogram data for 8 mixtures, each provided with DNA profiles of potential contributors.

Each participant who completed both the *NoC* and *ICSA Subtests* received 20 packets total: 12 packets were assigned for the *NoC Subtest* and 8 packets were assigned for the *ICSA Subtest*. Each packet contained a DNA mixture profile provided as an electropherogram (EPG) in .HID analysis format, as well as positive and negative controls and allelic ladders. Participants were provided EPGs based the Amp/CE settings (including amplification kit, CE instrument, and amplification cycle) that they selected in a *Mixture Configuration Selection* questionnaire. The distribution of questionnaires and subtests and the collection of responses used web-based software developed by Noblis, which restricted access to the study to registered participants.

Participants were directed to conduct their assessments of the DNA mixtures based upon the policies and validated procedures in their Standard Operating Procedures (SOPs), using the same considerations and diligence that would be employed for operational casework.

The study was conducted with the approval of Bode Technology's Institutional Review Board (IRB) and NIJ's Human Subjects Protection Officer (HSPO), including the use of DNA samples, the informed consent, and the privacy certificate. With respect to participation in the study itself, the Bode IRB determined that participation by the participating laboratories does not meet the definition of Human Subjects Research based on 28 CFR Part 46 (pre-2018 Common Rule) as the participants are laboratories, not individuals.

Appendix B1 DNAmix Working Group and Advisory Group

In order to assure that the study design and details met with the approval with as wide a cross-section of the forensic DNA community as possible, we created two advisory groups to provide input:

- The ***DNAmix Working Group*** was a group of volunteers with a range of expertise in DNA mixture analysis who were invited to provide input and guidance to the Bode/Noblis team on study design and details, including review of instructions, questions and multiple-choice response categories for all four phases of the study. The information provided to working group members was limited so that they were able to participate in the study*; the information shared with the DNAmix Working Group was included in the instructions/FAQs provided to all participants to guarantee a level playing field. To avoid conflicts of interest, the DNAmix Working Group did not include anyone who works for any company that develops or sells PGS. The Working Group met (virtually in web conferences) from January-September 2021 (generally bi-weekly) and engaged in many email discussions. The Working Group included Jack Ballantyne (National Center for Forensic Science/University of Central Florida), Jen Breaux (Montgomery County Maryland), John Butler (NIST), Amber Carr (FBI Laboratory), Roger Frappier (Centre of Forensic Sciences, Toronto), Tim Goble (New York State Police), Bruce Heidebrecht (Maryland State Police), Kristy Kadash (Jefferson County (Colorado) Regional Crime Laboratory), Shawn Montpetit (San Diego Police Department), Steven Myers (California DOJ), Craig O'Connor (Office of Chief Medical Examiner of the City of New York), Robyn Ragsdale (Florida Department of Law Enforcement), Kristin Sasinouski (Bode Technologies), and Charlotte Word (Private Consultant). The Working Group members were selected to include input from US Federal, state, local, Canadian, private, and research perspectives. Working Group members included (but were not limited to) members of OSAC and SWGDAM.

- The ***DNAmix Advisory Group*** was created for discussions regarding mixture design decisions, detailed review of potential mixtures, and final approval of the mixtures used in the study; no potential participants were included. The Advisory Group was composed of the Bode/Noblis Study Team (Jonathan Davoren, Robert Bever, Austin Hicklin, Nicole Richetelli, Lauren Leone), the NIST Applied Genetics Group (Peter Vallone, Erica Romsos, Sarah Riman), and members of the DNAmix Working Group who were not eligible to participate in the study (Jack Ballantyne, John Butler, Charlotte Word). The Advisory Group met virtually as needed from February-November 2022.

Note the distinction that the Working Group was limited to issues related to the participant-facing portion of the study, whereas the Advisory Group focused on the design and selection of the mixtures. Many Working Group members were assumed to be participants; Advisory Group members included no participants.

Appendix B2 Summary of Participant Instructions

The following is summarized from the participant instructions for the NoC Subtest and the ICSA Subtest. The complete instructions are included as Supplemental Information S1. Note that the ICSA Subtest instructions summarized below are limited to the aspects of that subtest relevant to this paper.

Conduct your assessments of each DNA mixture profile (.HID file) and respond to each of these questions based upon the policies and validated procedures in your Standard Operating Procedures (SOPs), using the same considerations and diligence that you would employ for operational casework samples. Your responses should go through technical review/quality assurance as specified by your laboratory's SOPs.

Appendix B2a Mixture Configuration Selection

Mixture Configuration Selection is a two-question online survey accessed from the DNAmix 2021 website conducted for participants to:

- Indicate if you will participate in the NoC and/or ICSA Subtests
- Select the Amp/CE settings used to prepare the mixtures that will be assigned to you in the NoC/ ICSA Subtests
- Indicate how the selected Amp/CE Settings compare to your SOPs

Before you begin the *NoC Subtest*, you must complete *Mixture Configuration Selection*. The study team will use this information to assign mixtures to participants, and the DNAmix website software will use the information to determine which questions and response options are presented in the NoC and ICSA Subtests.

* *Note: The DNAmix Working Group was consulted regarding the general approach used in creating DNA mixtures (including the selection of Amp/CE Settings for the NoC and ICSA Subtests) but was shielded from details of selection and assignment so that they had no more information than any other participants.*

Mixture Configuration Selection includes the following questions:

- Select one of the following Amp/CE Settings that you will use to participate in the NoC and/or ICSA Subtests of this study:
  - 6C29: Promega PowerPlex Fusion 6C@29 cycles; Amp volume 25µL; ABI 3500xL injection at 1.2kV for 24 seconds (equivalent to ABI 3500 for 15 seconds)
  - GF28: Applied Biosystems GlobalFiler@28 cycles; Amp volume 25µL; ABI 3500xL injection at 1.2kV for 24 seconds (equivalent to ABI 3500 for 15 seconds)
  - GF29: Applied Biosystems GlobalFiler@29 cycles; Amp volume 25µL; ABI 3500xL injection at 1.2kV for 24 seconds (equivalent to ABI 3500 for 15 seconds)
  - ID28: Applied Biosystems AmpFLSTR Identifiler Plus@28 cycles; Amp volume 15µL; ABI 3500xL injection at 1.2kV for 12 seconds (equivalent to ABI 3500 for 7.5 seconds)
  - None of the above (We will not participate in the NoC or ICSA Subtests)
- [if not "none of the above"] Please indicate how the selected Amp/CE Settings compare to your SOPs:
  - *[EXACT]* This corresponds exactly to our lab's validated settings—we can use these settings for NoC and ICSA
  - *[EQUIVALENT]* This is equivalent to our lab's settings; this differs in details we consider minor or inconsequential (such as injection time of 15 vs 16 seconds)—we can use these settings for NoC and ICSA
  - *[DIFFERENT-BOTH NoC&ICSA]* This differs from our lab's validated settings, but we are willing to participate using these settings in both NoC and ICSA (Note: these results will be analyzed separately during analysis)
  - *[DIFFERENT-NoC ONLY]* This differs from our lab's validated settings; we are willing to participate using these settings in NoC, but not in ICSA (Note: these results will be analyzed separately during analysis)

Appendix B2b Packets (NoC Subtest)

In the *NoC Subtest* you will be assigned a total of 12 *NoC Packets*. Each *NoC Packet* includes one DNA mixture profile, positive and negative controls, and allelic ladders. No case information will be provided. All DNA mixture profiles are electropherograms, provided to participants as .HID files.

You will have access to only one *NoC Packet* at a time: to avoid the possibility of administrative errors or misunderstandings, you must submit your responses for a DNA mixture profile before downloading the next DNA mixture profile.

Appendix B2c Packets (ICSA Subtest)

In the *ICSA Subtest* you will be assigned a total of 8 *Comparison Packets*. Each *Comparison Packet* includes one DNA mixture profile, one or more reference profiles, positive and negative controls, and allelic ladders. No case information will be provided. All DNA mixture profiles and reference profiles are electropherograms, provided to participants as .HID files.

Each *Comparison Packet* includes one of more reference profiles:

- All comparison packets include 1 reference profile designated as "person of interest" (POI), which for the purposes of this study indicates an individual whose contribution to the mixture is in question (such as an alleged perpetrator).
- All comparison packets that are simulated sexual assault kits (SAKs) include 1 reference profile labeled "victim," which for the purposes of this study indicates the complainant in a sexual assault from whom simulated sexual assault kit samples are collected.
- Some comparison packets that are simulated SAKs may include 1 reference profile labeled "consensual partner," which for the purposes of this study indicates an individual known to have had consensual intimate contact with a victim of a sexual assault.
- Some non-SAK comparison packets may include 1 reference profile labeled "expected contributor," which for the purposes of this study indicates a known individual who is expected or assumed to be a contributor to a DNA mixture profile, such as the owner of an item or a member of a household.

You will have access to only one *Comparison Packet* at a time: to avoid the possibility of administrative errors or misunderstandings, you must submit your responses for a DNA mixture profile before downloading the next DNA mixture profile.

Appendix B2d Amp/CE Settings

In order to represent the SOPs of as many participating laboratories as feasible, electropherograms were prepared using four combinations of "Amp/CE Settings" (which refers to a specific combination of amplification kit, amplification cycles, volume of amplification reaction, CE instrument, and injection time and voltage). The combinations of Amp/CE Settings that have been implemented were the four most commonly-used Amp/CE settings selected by registered participants,* using the abbreviations [6C29, GF28, GF29, or ID28].

You will be assigned mixtures that were prepared using the Amp/CE setting option that you chose during *Mixture Configuration Selection*, prior to the *NoC Subtest*.

Appendix B2e Preparation of DNA Mixture Profiles

The DNA used to create the mixture profiles for this study came from various sources, including buccal, blood, and tissue samples. There were no simulated/contrived profiles; all DNA profiles in this study are from real people. DNA samples were extracted prior to mixing.

Mixtures were quantified using ABI Quantifiler Trio on an ABI 7500 real-time PCR instrument. The mixture quantification results (including the total amount of DNA amplified, amount of male DNA, and degradation index) will be included with the mixtures in the *NoC Subtest*.

Various volumes of DNA were pipetted into a single tube to make a large mixture stock. That stock was then aliquoted and amplified in each of the four amplification kits (see "Amp/CE settings" above for amplification volumes and cycles). The ABI 9700 thermocycler was used for amplification, using the specific Amp/CE settings and other standard manufacturer recommended settings. The ABI 3500xL was used for capillary electrophoresis (CE), using injection time and voltage settings specified in the Amp/CE Settings (see above); settings for run time, run voltage, capillary length, polymer type, etc. use the default settings specified for each amplification kit.

GeneMapper (v1.5; incorporated into ABI 3500xL) was used to create .HID files. We are not providing PDFs (images) of the electropherograms because creating such PDFs implements decisions regarding the analytical threshold (AT) value and the utilization of stutter filters, and we want all such decisions to be made by the participants.

Every effort was made with respect to quality assurance in creating these mixtures. Note that in some cases there may be artifacts (such as pull-up) present, as may be found in ordinary casework — please review the controls provided.

Appendix B2f Data Provided (NoC Subtest)

Each *NoC Packet* is numbered (NOC_01 through NOC_99, shown as "NOC_XX" in the table below). Participants are not necessarily assigned the same packets, and the order of assignments varies among participants.

Each *NoC Packet* is specific to the Amp/CE Settings previously selected by participants (shown as "YYYY" in the table below), using the abbreviations [6C29, GF28, GF29, or ID28].

** Registered participants were contacted by email and were given a deadline of 23 August 2021 to indicate preferences for Amp/CE settings.*

Each *NoC Packet* is contained in a Zip file, downloaded from the DNAmix 2021 website (https://dnamix.edgeaws.noblis.org/). Each *NoC Packet* includes the following files:

All Packets	1 DNA mixture profile (HID file)	NOC_XX_YYYY_Mixture.hid
	Amp/CE Settings used to create the electropherograms	AmpCESettings_YYYY.pdf
	Quantitation data for the mixture	NOC_XX_QuantResults.pdf
	2 Ladders	NOC_XX_YYYY_Ladder1.hid NOC_XX_YYYY_Ladder2.hid
	Positive and Negative controls	NOC_XX_YYYY_Pos.hid NOC_XX_YYYY_Neg.hid

In a few cases the positive or negative controls were re-injected, in which case they are in a subdirectory (named POS or NEG) with the associated ladders.

Appendix B2g Data Provided (ICSA Subtest)

In the ICSA Subtest you will be assigned a total of 8 *Comparison Packets*. You will have access to only one *Comparison Packet* at a time: to avoid the possibility of administrative errors or misunderstandings, you must submit your responses for a DNA mixture profile before downloading the next DNA mixture profile.

Each *Comparison Packet* is numbered (ICSA_001 through ICSA_999, shown as "ICSA_XXX" in the table below). Participants are not necessarily assigned the same packets, and the order of assignments varies among participants.

Each *Comparison Packet* is specific to the Amp/CE Settings previously selected by participants (shown as "YYYY" in the table below), using the abbreviations [6C29, GF28, GF29, or ID28].

Each *Comparison Packet* includes 1 DNA mixture profile and 1 person of interest (POI) reference profile. *Comparison Packets* that are simulated sexual assault kits (SAKs) include 1 victim (VIC) reference profile, and may include 1 consensual partner (CON) reference profile. Non-SAK comparison packets may include 1 expected contributor (EXP) reference profile. Each mixture or reference profile includes positive and negative controls, and 2 allelic ladders.

Each *Comparison Packet* is contained in a Zip file, downloaded from the DNAmix 2021 website (https://dnamix.edgeaws.noblis.org/). The mixture and reference profiles are included in separate subdirectories as shown in the table below.

	Subdirectory	*Files*	
All packets	ICSA_XXX_YYYY/Mixture/	**ICSA_XXX_YYYY_Mixture.HID**	***DNA mixture profile***
		ICSA_XXX_YYYY_Ladder1-Mixture.HID ICSA_XXX_YYYY_Ladder2-Mixture.HID ICSA_XXX_YYYY_NEG-Mixture.HID ICSA_XXX_YYYY_POS-Mixture.HID	*Controls for DNA mixture profile*
	ICSA_XXX_YYYY/POI/	**ICSA_XXX_YYYY_POI.HID**	***Person of interest (POI) reference profile***
		ICSA_XXX_YYYY_Ladder1-POI.HID ICSA_XXX_YYYY_Ladder2-POI.HID ICSA_XXX_YYYY_NEG-POI.HID ICSA_XXX_YYYY_POS-POI.HID	*Controls for POI reference profile*
Some packets	ICSA_XXX_YYYY/VIC/	**ICSA_XXX_YYYY_ VIC.HID**	***Victim (VIC) reference profile***
		ICSA_XXX_YYYY_Ladder1-VIC.HID ICSA_XXX_YYYY_Ladder2-VIC.HID ICSA_XXX_YYYY_NEG-VIC.HID ICSA_XXX_YYYY_POS-VIC.HID	*Controls for VIC reference profile*
	ICSA_XXX_YYYY/CON/	**ICSA_XXX_YYYY_ CON.HID**	***Consensual partner (CON) reference profile***
		ICSA_XXX_YYYY_Ladder1-CON.HID ICSA_XXX_YYYY_Ladder2-CON.HID ICSA_XXX_YYYY_NEG-CON.HID ICSA_XXX_YYYY_POS-CON.HID	*Controls for CON reference profile*
	ICSA_XXX_YYYY/EXP/	**ICSA_XXX_YYYY_ EXP.HID**	***Expected contributor (EXP) reference profile***
		ICSA_XXX_YYYY_Ladder1-EXP.HID ICSA_XXX_YYYY_Ladder2-EXP.HID ICSA_XXX_YYYY_NEG-EXP.HID ICSA_XXX_YYYY_POS-EXP.HID	*Controls for EXP reference profile*

In some *Comparison Packets*, the positive or negative controls were re-injected, in which case they are in a subdirectory (named POS or NEG) with the associated ladders.

Appendix B2h Subtest Questions (both NoC Subtest and ICSA Subtest)

Questions 1-13 were identical for the NoC Subtest *and the* ICSA Subtest.

On the DNAmix 2021 website, you will be asked to answer the following questions for ***each*** of the DNA mixtures that you are assigned. The subtest is completed online; this information is provided here as a reference.

As a quality assurance measure, the website will display an image of the electropherogram for the first several loci in the DNA mixture profile for the assigned packet. Please ensure that you are submitting your responses for the given mixture profile.

Note for each mixture profile assessed, you will be asked to review and confirm your responses prior to submission. After submission, your responses are considered final and cannot be changed.

Packet Assignment Details

1. Please re-enter the Participant ID shown at the top of the page (Dxxxx):______

   Note: this information will be used for quality assurance purposes only. The Participant ID (a 5 character alpha-numeric string starting with D) is located at the top right of the Number of Contributors (NoC) Page (right about the electropherogram preview image).

2. Please double-check the NoC Packet number: verify that the number of the HID mixture file you are assessing is the same as shown at the top of this page. Please enter that NoC Packet number here (For example, in the NoC Beta Test, you would enter the following NoC packet number: 99): ______

   The NoC Packet number is located in the HID filename, in the electropherogram preview image (located at the top of the NoC Subtest page of the DNAmix 2021 website), and embedded within the electropherogram data. You do not need to enter the "NOC_" portion; please only enter the two digit NoC Packet number.

Settings

3. Did you use an analytical threshold (AT) for this mixture?

   In other words, was there a minimum RFU value (either globally or per dye channel) to delineate signal (above the threshold) from noise (below the threshold)? This analytical threshold may have been utilized explicitly (by comparing measured RFU values to the threshold value(s)) or via general evaluation ("eye-balling" the data and comparing to the threshold value(s)).

   3.a Yes, I used a single AT (Please specify:_____)
   3.b Yes, but my ATs varied by dye channel
   3.c No

4. Did you use a stochastic threshold (ST) for this mixture?

   In other words, was there a minimum RFU value (either globally or per dye channel) to delineate peaks (above the threshold) from potential artifacts or stochastic effects (below the threshold)? This stochastic threshold may have been utilized explicitly (by comparing measured RFU values to the threshold value(s)) or via general evaluation ("eye-balling" the data and comparing to the threshold value(s)).

   4.a Yes, I used a single ST (Please specify:_____)
   4.b Yes, but my STs varied by dye channel
   4.c No

Replicate Amplifications

The next two questions both ask whether you would have conducted replicate amplifications if you had received this sample in actual casework, but please note the differences. Question 5 is regarding replicate amps based on the amount of DNA available: if only a small amount of DNA was available, would you have amplified all the DNA or divided it and done multiple amps? Question 6 is regarding replicate amps based on the review of the electropherogram: if additional DNA were available, after reviewing this EPG would you conduct another amp?

5. If you received this DNA mixture sample in casework and the total quantity of DNA available was the amount specified in the quantitation data, would you have divided the sample and conducted multiple replicate amplifications?

   In other words, given the DNA quantity alone (not based upon review of the electropherogram), would you have amplified the entire sample (as was done here), or would you have instead done 2 or more amplifications each using a part of the total sample? Assume that all DNA available in the entire (wet) sample was used here and there will not be additional DNA to permit a subsequent amplification after CE.

   5.a We do not ever conduct replicate amplifications in my laboratory (per our SOPs)
   5.b No: we would not have done replicate amplifications in this case (but we do in some cases)
   5.c Yes: we would do 2 replicate amplifications, each with 1/2 of the total amount, and increase sensitivity by adding 1 cycle
   5.d Yes: we would do 3 replicate amplifications, each with 1/3 of the total amount, and increase sensitivity by adding 1 cycle
   5.e Yes: we would do 3 replicate amplifications, each with 1/3 of the total amount, and increase sensitivity by adding 2 cycles
   5.f Other (Please explain:_____)

6. If there was sufficient DNA remaining for an additional amplification, would you do another amplification (re-amp) after seeing this mixture profile?

   In other words, if sufficient DNA remained would you conduct another amplification based upon the mixture profile/electropherogram provided in this NoC packet (e.g., to verify alleles, observe stochastic effects, etc.)?

6.a No: we are not permitted to re-amp per our SOPs
6.b No: we would interpret this profile
6.c Yes: we would re-amp using more DNA
6.d Yes: we would re-amp using less DNA
6.e Yes: we would re-amp using the same amount of DNA

Suitability

7. Is this DNA mixture profile suitable for comparison and/or statistical analysis?

In other words, did you determine that this DNA mixture profile can appropriately be used to conduct comparisons (i.e., comparison of the mixture to reference profiles of POIs, victims, consensual partners, and/or expected contributors) and/or statistical analyses (i.e., compute an LR, RMP, or CPI/CPE with respect to a POI)?

7.a Yes (for the entire mixture and all contributors)
7.b Yes, but only for a subset of the contributors (e.g., major(s))
7.c Yes, but only for a subset of loci
7.d Yes, but only for a subset of loci, and only for a subset of the contributors
7.e No

8. *[If the mixture is NOT suitable for comparison and/or statistical analysis]* Why is this profile unsuitable for comparison and statistical analysis? (check all that apply; check at least one)

In other words, if you indicated "No" in the previous question, what factor(s) informed your determination? Please select all factors that you considered in your determination that the DNA mixture profile was not suitable for comparison/statistical analysis.

8.a Not enough alleles or loci suitable for analysis
8.b DNA template levels too low overall
8.c Sample too degraded
8.d Sample too inhibited
8.e Too many contributors
8.f Too much uncertainty in the number of contributors
8.g Mixture proportions/contributor ratios
8.h Other (Please specify:_____)

Number of Contributors [Only shown if the mixture is suitable for comparison and/or statistical analysis]

9. How would you report the number of contributors in this profile?

In other words, how would you report the number of contributors to this DNA mixture profile if you encountered this mixture in casework? Would you be able to assess the number of contributors given this DNA mixture profile? If so, would you report an exact/single estimate of number of contributors (e.g., 3 contributors) or would you report a range of possible numbers of contributors (e.g., 3-4 contributors or minimum of 3 contributors/maximum of 4 contributors) or would you report a minimum number of contributors (e.g., at least 3 contributors)?

9.a I would report an exact number of contributors
9.b I would report a range of possible numbers of contributors
9.c I would report a minimum number of contributors
9.d The levels (overall quantity and/or peak heights) are not sufficient to determine the number of contributors [Go to Additional Comments]
9.e *The mixture is too complex to determine the number of contributors* [Go to Additional Comments]

9.1 Provide your estimate of NoC.

Note: you will only see the version of question 8.1 that is associated with your response to question 8 above. In other words, you will only see one of the three options in the question text below.

- *[if selected 8a: Exact NoC]* Select your estimate of the number of contributors:

In other words, select your single estimate for the number of contributors to this DNA mixture profile.

- *[if selected 8b: Range of NoC]* Select your estimate of the range of possible numbers of contributors. (For example, if you estimate that there are 3-5 possible contributors to this mixture profile, you must select 3, 4, and 5): (check all that apply; select at least two)

In other words, select all possible numbers of contributors included in your estimated range for this DNA mixture profile, which must include at least two options based upon your response to Q8 that you would report the NoC for this profile using a range of possible numbers of contributors.

- *[if selected 8c: Exact NoC]* Select your estimate of the minimum number of contributors:

In other words, select your single estimate for the minimum number of contributors to this DNA mixture profile.

9.1-a	*At least 1 contributor*
9.1-b	*At least 2 contributors*
9.1-c	*At least 3 contributors*
9.1-d	*At least 4 contributors*
9.1-e	*At least 5 contributors*
9.1-f	*At least 6 contributors*
9.1-g	*At least 7 contributors*
9.1-h	*At least 8 or more contributors*

10. What were the PRIMARY loci used as the basis for determining the number of contributors? In other words, indicate the loci that were most informative or most helpful. (check all that apply; select at least one)

Note: you will only see the set of loci included in the amplification kit that you previously selected in Mixture Configuration Selection.

Names of commercial manufacturers are included for the systems that are the most frequently used by registered participants; inclusion does not imply endorsement by the study team.

Applied Biosystems GlobalFiler (display order)	Promega PowerPlex Fusion 6C (display order)	Applied Biosystems AmpFLSTR Identifiler Plus (display order)
D3S1358	Amel	D8S1179
vWA	D3S1358	D21S11
D165539[*]	D1S1656	D7S820
CSF1PO	D2S441	CSF1PO
TPOX	D10S1248	D3S1358
Y indel	D13S317	TH01
Amel	Penta E	D13S317
D8S1179	D16S539	D16S539
D21S11	D18S51	D2S1338
D18551[†]	D2S1138[‡]	D19S433
DYS391	CSF1PO	vWA
D2S441	Penta D	TPOX
D19S433	TH01	D18S51
TH01	vWA	Amel
FGA	D21S11	D5S818
D22S1045	D7S820	FGA
D5S818	D5S818	
D13S317	TPOX	
D7S820	D8S1179	
SE33	D12S391	
D10S1248	D19S433	
D1S1656	SE33	
D12S391	D22S1045	
D2S1338	DYS391	
	FGA	
	DYS576	
	DYS570	

11. Which factors affected your assessment of number of contributors? (check all that apply; select at least one)

In other words, what factors did you consider when estimating the number of contributors for this DNA mixture sample? Please select all factors that informed your determination.

11.a Discriminating potential/variability of loci (or allele frequency)
11.b Expected stutter ratios
11.c Information below the analytical threshold
11.d Maximum Allele Count (MAC) per locus
11.e Overall level of data (peak heights in relation to laboratory validated thresholds)
11.f Peak heights (RFU)
11.g Peak morphology (e.g., CE resolution; unresolved microvariants; peak shouldering)
11.h Presence of degradation
11.i Presence of inhibition
11.j Quantitation data
11.k Relative peak heights (peak height ratios and possible shared/stacked alleles)
11.l Sex determining markers
11.m Total allele count in sample

[*] *Typo in the instructions: should be D16S539*

[†] *Typo in the instructions: should be D18S51*

[‡] *Typo in the instructions: should be D2S1338*

11.n Other (Please specify:_____)

12. Are you able to identify any major contributors?

In other words, would you consider one (or more) contributors to be major contributors according to the criteria outlined in your SOPs (e.g., based upon peak height ratios or RFU percentages). This separation of major contributor(s) may have been conducted explicitly (by computing peak height ratios/RFU percentages and comparing to a threshold, such as a 3:1 peak height ratio or 70% of the total RFUs) or via general evaluation (distinguished visually, without calculation). If your SOPs do not permit you to differentiate between major and minor contributors, please indicate as such.

12.a There are no contributors I would consider majors
12.b There is one major contributor
12.c There are two or more major contributors
12.d We do not differentiate between major and minor contributors

13. Did you use any software tool to assist in assessing the number of contributors?*

In other words, please indicate how you assessed number of contributors for this DNA mixture profile. If you used a combination of manual assessment and software, please select the option which most informed your assessment.

13.a No, I assessed the number of contributors manually
13.b Yes, I used NOCIt
13.c Yes, I used PACE
13.d Yes, I used FaSTR
13.e Yes, I used diagnostics from my ProbGen system (which does not directly tell the NoC, but aids in checking the NoC analysis)
13.f Yes, I used an internally developed tool
13.g Yes, I used another commercial or open-source tool (Please specify:_____)

Additional Comments

Additional comments: Please provide a comment ONLY if there is an issue or a limitation for this NoC packet that you could not adequately address using any of your responses above. (Please limit your responses to 75 words or less.)

Appendix B2i Subtest Questions (ICSA only)

The *ICSA Subtest* includes the following additional question (relevant to NoC), which is not in the *NoC Subtest*:

9.2 [if indicated a range or minimum NoC] (Given that you would report a range or minimum number of contributors) What NoC value did you use as a basis for your conclusions/analyses reported for this Comparison Packet regarding the person of interest (POI) as a potential contributor to the mixture sample?

In other words, when conducting your comparisons and statistical analyses for this Comparison Packet what value did you assume for number of contributors to the mixture sample?

Please report all statistical analyses based upon your response to this question.

9.2-a 1 contributor
9.2-b 2 contributors
9.2-c 3 contributors
9.2-d 4 contributors
9.2-e 5 contributors

* *The responses for question #13 are shown as they were presented to participants in the study; the wording for the responses in the instructions differed slightly.*

9.2-f 6 contributors
9.2-g 7 contributors
9.2-h 8 contributors
9.2-i I based analyses with respect to the POI on a range of number of contributors (not a single NoC value)

Appendix C DNA Samples and Mixtures

The DNA mixtures were designed and created to be broadly representative of the range of attributes encountered in actual DNA mixture casework. The mixtures were created to vary with respect to the number of contributors, the amount of DNA (in total and for each contributor), the relative proportions of contributors, degradation, and the extent of allele sharing.

Given the multitude of factors that influence DNA analysis, it is not feasible to exhaustively cover the factor space with a small number of mixtures. Given that (significant) limitation, the experimental samples were selected to span a spectrum of the attributes encountered in actual DNA mixture casework [27], anticipating that most DNA mixtures of interest would either be similar to the provided mixtures or could expect performance interpolated between provided mixtures that are more complex and less complex.

The following sections discuss the design and creation process in detail.

Appendix C1 Mixture design, sources, and selection of subjects

Allele sharing was controlled through the selection of subjects from a broad pool of subjects from multiple sources: the mixtures include 102 subjects selected from a pool of 849 subjects from four sources. A variety of n-person mixtures were modeled using simulations in order to select subjects with a range of allelic sharing. The DNA mixtures used in the study were created from DNA samples from four sources:

- ***Blood bank*** — Blood samples purchased from a commercial supplier of human blood.
- ***NIST*** —The publicly available genotypes from the NIST 1036 Revised U.S. Population Dataset NIST [48,49] were screened for downselection. Of the 1036 samples in the dataset, 623 were available in sufficient quantity for potential use in the study. Of the 623 candidate samples in the NIST 1036 dataset, 64 were identified for potential use in the study. These 64 DNA extracts (approximately 10 ng per sample) were provided by the Applied Genetics Group to support the study.
- ***Repository*** — DNA samples purchased from a genetic repository by Bode.
- ***Siblings*** — DNA samples from buccal swabs of three pairs of brothers (not twins), collected specifically for this study. (One pair of brothers was included in an *NoC* mixture; the other two pairs were used as contributor vs. noncontributor POIs in *ICSA*.)

Multiple amplifications were conducted for each of the samples in the initial pool; samples were omitted if they had notable genotyping differences between amplifications, high stutter, or trialleles. These samples were used after detailed quality control to detect and remove any samples with potential genetic anomalies. The mixtures included 102 subjects selected from a pool of 849 subjects from these four sources (Table S1). (An additional three subjects were used in the mixture provided for use in the pilot/Beta test.)

	# Subjects
Total available	849
Omitted due to anomalies	101
Used for modelling	748
Downselect pool	190
Used in mixtures	102
Used as noncontributor reference profiles (ICSA only)	9
Total used	111

Table S1. Sources of DNA samples with numbers of subjects.

Table S2 shows population distributions for the samples used in mixtures.

Population	*Total*	*Female*	*Male*
African American	27	10	17
Asian	18	10	8
Caucasian	39	9	30
Hispanic	15	6	9
Unknown	3	0	3
Total	*102*	*35*	*67*

Table S2. Population distributions for samples used in mixtures.

The design of mixtures followed the following steps:

- ***Preliminary modelling for downselection of NIST samples*** — The agreement with the NIST Applied Genetics Group was limited to providing 64 physical samples for use in the study. We downselected from 623 to 64 by doing a preliminary modelling step in which we simulated all possible two- and three-person virtual mixtures given the 623 NIST samples in combination with the Repository and Blood bank data, and selected the 64 NIST samples based on a range of allelic overlap (23 samples with high levels of allele sharing, 12 with low sharing, and 29 selected randomly).
- ***Modelling virtual mixtures*** — The 184 STR profiles in the downselect pool (other than the Sibling samples) were included in a relational database, all possible two-, three-, and four-person mixtures were simulated within the database, and the number of unique alleles was determined for each locus and summed up across all loci for each virtual mixture. Modelling of all possible five-person mixtures was in excess of the limits of the database so five-person mixtures were simulated by adding one sample to four-person mixtures.
- ***Target mixtures*** — In coordination with the Advisory Group, we developed conceptual "target mixtures" selected from the modelled virtual mixtures so that the samples for the NoC and ICSA subtests each spanned a range of attributes encountered in actual DNA mixture casework, with respect to the number of contributors, the amount of DNA (in total and for each contributor), the relative proportions of contributors, degradation, and the extent of allele sharing. In essence we created 20 mixtures designed to span a range of attributes found in DNA mixtures: all eight of the mixtures in the ICSA Subtest and 12 of the mixtures in the NoC Subtest were created based on target mixtures. The remaining nine mixtures in the NoC Subtest included randomly-selected subjects in order to collect additional data for three- and four-person mixtures.
- ***Acceptance review*** — On review of EPGs of physical mixtures by the Study Team and the Advisory Group, some virtual mixtures were revised and some physical mixtures were recreated/reamplified. In particular, the quantities of DNA included in several of the initial mixtures were reduced after review by the Advisory Group assessed them as not representative of casework. The final set of mixtures was acceptable to all on the Study Team and the Advisory Group as a reasonably representative range of mixtures.

Appendix C2 Amp/CE Settings

In order to represent the SOPs of as many participating laboratories as feasible, EPGs were prepared using four combinations of "Amp/CE Settings," which refers to a specific combination of amplification kit, amplification cycles, volume of amplification reaction, CE instrument, and injection time and voltage.

All registered participants were contacted by email and were given a deadline of 23 August 2021 to indicate preferences for Amp/CE settings. The combinations of Amp/CE Settings that were used to create the mixtures were the four most commonly-used Amp/CE settings selected by registered participants by that deadline (Table 1, main paper). The fifth most common combination of settings (Qiagen 24plex at 24 cycles, 25mcl, 24sec injection) would only have resulted in six participants (from three labs) who indicated the settings corresponded to or were equivalent to their SOPs.

In response to one prospective participant's question, we included the following statement in the Frequently Asked Questions (FAQs):

- Q: I was disappointed that the Qiagen Investigator 24plex kit was not included as one of the subset whilst two GlobalFiler settings were chosen instead. Why was this kit not included in the study?
- A: We wish we could replicate all the registered participants' STR laboratory protocols, but are unfortunately unable to provide all possible Amp/CE settings due to funding and time constraints. In selecting the specific Amp/CE settings options that we would provide, we used the four most commonly-used combinations of settings (as reported by registered participants by 23-Aug). Overall, there was a clear majority of participants who used GlobalFiler / 29 cycles or Fusion 6C / 29 cycles, but the remainder

of participants had a wide variety of settings. Relatively few participants selected Qiagen Investigator 24plex and those who did had a wide variety of cycle settings, so there was no single set of 24plex Amp/CE settings that would have captured a statistically-useable number of potential participants.

Please see Table S10 for the number of participants who used each of the Amp/CE Settings.

When selecting Amp/CE Settings, participants indicated how the selected settings compared to their SOPs, given these options:

- This corresponds exactly to our lab's validated settings-we can use these settings for NoC and ICSA
- This is equivalent to our lab's settings; this differs in details we consider minor or inconsequential (such as injection time of 15 vs 16 seconds)-we can use these settings for NoC and ICSA
- This differs from our lab's validated settings, but we are willing to participate using these settings in both NoC and ICSA (Note: these results will be analyzed separately during analysis)
- This differs from our lab's validated settings; we are willing to participate using these settings in NoC, but not in ICSA (Note: these results will be analyzed separately during analysis)

Results and analyses for participants who indicated that the Amp/CE Settings exactly corresponded or were equivalent to their laboratory's SOPs ("SameSOP") are separated from results for participants who indicated the Amp/CE Settings differed from their laboratory's SOPs ("DiffSOP").

Appendix C3 Mixture creation

Preparation of individual samples — There were no simulated/contrived profiles; all DNA profiles in this study are from real people. No sample was used in more than one mixture. The DNA for the Blood bank and Sibling samples was extracted using the QIAGEN EZ1 instrument and DNA Investigator kit. (NIST and Repository samples were already extracted.) The DNA for the Repository, Blood bank, and Sibling samples was quantified using Applied Biosystems Quantifiler Trio on an Applied Biosystems 7500 Real-Time PCR System with SDS Analysis Software. (The NIST samples were supplied with quantification values.) Quality control was performed on all individual samples by amplification with a Globalfiler kit. As part of the quality control testing a number of samples were eliminated due to higher than normal stutter, low intra locus balance or possible drop-in.

Degradation — The two degraded samples were degraded by exposure to UV light: DNA aliquots at the target concentration were added to 1.5 mL tubes, and subjected to UV exposure with the tube left open.

- For ICSA_192/680, one subject (the major donor) was degraded prior to creating the mixture. The donor sample diluted to 0.2ng/ul based on a previously obtained stock concentration, exposed to UV for 100 sec (total exposure of ~1 J/cm^2). Following exposure, the degraded sample was combined with undegraded contributor DNA. The mixture and the degraded DNA sample were both quantified with Trio to obtain quants/DIs for the mixture as a whole and for the degraded contributor DNA alone.
- For ICSA_057/802, each contributor was individually degraded prior to mixing (allowing us to control mixture ratios and degradation). Each donor was diluted to 0.4ng/ul based on previously obtained stock concentrations and exposed to UV for 80 secs (total exposure ~0.8 J/cm^2). Following exposure, the individual samples were combined to generate the mixture sample. Degraded donors individually and the final mixture sample were quantified with Trio to determine individual and combined mixture quant/DIs.

Mixture preparation — DNA samples were extracted prior to mixing. Various volumes of DNA were pipetted into a single tube to make a large mixture stock for each mixture. Each DNA mixture was first made to a specific target contributor ratio based on quantification data. Mixtures were initially amplified using the Globalfiler system at 29 cycles, separated on a 3500xL Genetic Analyzer, and analyzed using GeneMapper ID-X 1.5 software. Multiple rounds of testing and adjustments were performed to get mixtures close to the target contributor ratios.

Amp/CE versions of each mixture — Once the measured mixture ratios were reasonably close to the target ratios, the DNA mixture was then aliquoted and amplified for each of the Amp/CE settings (Table 1, main paper). Because far fewer participants used the GF28 and ID28 Amp/CE Settings than the GF29 and 6C29 settings, nine of the mixtures in the NoC Subtest were only amplified using GF29 and 6C29; all *ICSA Subtest* mixtures and the remaining 12 NoC Subtest mixtures were amplified using all of the Amp/CE settings. The ABI 9700 thermocycler was used for amplification, using the specific Amp/CE settings and other standard manufacturer recommended settings. The ABI 3500xL Genetic Analyzer was used for capillary electrophoresis (CE), using injection time and voltage settings specified in the Amp/CE Settings; settings for run time, run voltage, capillary length, polymer type, etc. use the default settings specified for each amplification kit. To verify that the amplifications were replicable and to minimize stochastic effects, each mixture was amplified at least twice for each Amp/CE setting; the mixture was only used if at least two amplifications were found to be replicable on review, and any differences were within the normal variation expected for the DNA input level.

.HID files. — GeneMapper ID-X v1.5 (incorporated into ABI 3500xL) was used to create .HID files. We did not provide PDFs (images) of the electropherograms because creating such PDFs implements decisions regarding the analytical threshold (AT) value and the utilization of stutter filters, and we wanted all such decisions to be made by the participants.

The following were provided for each mixture:

- Mixture .HID file specific to one combination of Amp/CE settings
- The total amount of DNA (as measured by Quantifiler Trio during quantitation)
- The degradation index (as measured by Quantifiler Trio during quantitation)
- Amp/CE settings (Amplification kit, Amplification cycles, Volume of amplification reaction, CE instrument, Injection time and voltage)
- For mixtures in the *ICSA Subtest*: whether or not the mixture was a simulated sexual assault kit (SAK) was also included

Appendix C4 Mixture Assignments

In the *ICSA Subtest* all participants were assigned the same 8 mixtures (selected to span the range of attributes discussed in Appendix C1).

In the *NoC Subtest* participants were assigned 12 out of 21 mixtures. In the initial study design, we planned to assign all *NoC Subtest* participants the same 12 mixtures ("initial 12" in Table S3, selected to span the range of attributes discussed in Appendix C1). However, when it became clear that we would get far more GF29/6C29 participants than ID28/GF28 participants (based upon the registration questionnaires as well as the Amp/CE preferences indicated by participants), we added 9 additional 3-4 person mixtures to the *NoC Subtest* ("additional 9" in Table S3) that were assigned only to GF29/6C29 participants. The purpose of doing so was primarily to collect responses from a broader collection of 3-4 person samples (based on the assumption that much of the interest regarding DNA mixtures would be focused on 3-4 person mixtures) because such a disproportionate number of responses from the GF29/6C29 on the initial 12 mixtures would be of marginal benefit. The ID28/GF28 participants were limited to the initial 12 mixtures to assure we would receive enough responses per mixture for analysis. Note that the initial 12 *NoC Subtest* mixtures were developed by modeling allele sharing among contributors, but the additional 9 mixtures were deliberately created as arbitrary combinations of contributors, to collect more information on such routine mixtures.

Table S3 illustrates how the *NoC Subtest* mixtures were assigned to participants, each of whom were assigned 12 mixtures out of the pool of 21 mixtures:

- All *NoC Subtest* participants were assigned the 3 most complex mixtures in the *NoC Subtest* (Assignment group "ALL" in Table S3): NOC_31 has NoC_{GT}=5, NOC_71 has NoC_{GT}=6, NOC_15 includes two brothers.
- All ID28/GF28 participants received the initial 12 mixtures (Assignment groups "28" and "ALL" in Table S3)

- GF29/6C29 participants were randomly assigned to two groups (“29a” and “29b”) and assigned a mix of the initial 12 and additional 9 mixtures.

PacketID	*NoC$_{GT}$*	*DNA Amplified (ng)*	*NoC Subtest Design Group*	*Assignment Group*
NOC_52	2	0.054	Initial 12	28, 29b
NOC_24	2	0.043	Initial 12	28, 29a
NOC_49	3	0.191	Additional 9	29a
NOC_74	3	0.186	Additional 9	29a
NOC_28	3	0.180	Additional 9	29b
NOC_84	3	0.159	Initial 12	28, 29b
NOC_50	3	0.146	Additional 9	29b
NOC_76	3	0.121	Initial 12	28, 29a
NOC_25	3	0.099	Initial 12	28, 29a
NOC_53	3	0.091	Initial 12	28, 29a
NOC_57	3	0.090	Additional 9	29b
NOC_29	4	0.872	Initial 12	28, 29b
NOC_93	4	0.580	Initial 12	28, 29a
NOC_15	4	0.580	Initial 12	ALL
NOC_70	4	0.234	Initial 12	28, 29a
NOC_05	4	0.211	Additional 9	29b
NOC_14	4	0.210	Additional 9	29b
NOC_68	4	0.188	Additional 9	29b
NOC_41	4	0.171	Additional 9	29a
NOC_31	5	0.720	Initial 12	ALL
NOC_71	6	0.801	Initial 12	ALL

Table S3. *NoC Subtest* mixture assignments by Amp/CE settings.

Appendix D Participation

Participation was open to all forensic laboratories that conduct DNA mixture interpretation as part of their SOPs. For the purposes of this study, participants are laboratories (or subunits within laboratories), not individuals. It was the discretion of participating laboratories to determine which analysts were involved in the study. The identities of the specific analysts were not known to the DNAmix Study Team. Analysts involved were required to be qualified by the laboratory for operational mixture casework (not trainees).

Non-U.S. laboratories were welcome to participate if they report interpretations in English.

Laboratories were permitted to register more than one participant (referred to as a "subunit"). Each subunit was required to complete each phase of the study completely independently from any other subunits within their laboratory or other laboratories. Technical reviews and quality assurance procedures as outlined in the laboratory's SOPs were to be conducted for each subunit independently, if feasible.

Participation was solicited via national and regional conference announcements, contact through professional organization rosters (including OSAC, SWGDAM, CODIS, AAFS, and ASCLD), and posts on multiple examiner forums. No participants who met the requirements were barred from participation[*]. Participants were required to complete an IRB-approved informed-consent form prior to starting the study.

Analyses are based on a total of 134 participants from 67 forensic laboratories.[†] Fig S1 shows the number of participants per laboratory. Table S4 shows the number of responses per participant and lab. Fig S2 shows completion vs. partial completion by lab type: partial completion participants were not notably different from completion participants in term of lab type. Table S5 details participating laboratory types, by participant and by laboratory.

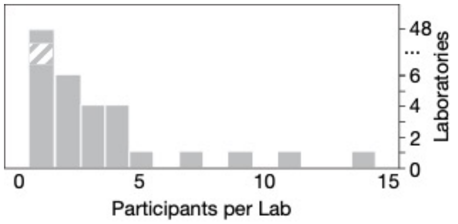

Fig S1. Number of participants (subunits) per laboratory. 48 labs had one participant each; 15 labs had two to five participants each (45 participants total); 4 labs had seven to 14 participants each (41 participants total). (mean 2.0 participants per lab; median 1)

	Participants	*Labs*	*Trials*
Complete (20)	87	45	1,740
Partial (12-19)	18	11	276
Partial (12, completed NoC only)	15	13	180
Partial (<12)	14	9	76

Table S4. Number of responses by participant and by lab. Mean 16.9, median 20 trials per participant. Note that 15 participants completed the *NoC Subtest* but did not participate in the *ICSA Subtest*. Mixtures were required to be completed in the order assigned: it was not possible for partial completion participants to select which mixtures to complete.

[*] *To avoid conflicts of interest, Bode analysts were not permitted to participate in the study.*

[†] *Note that in the overall DNAmix 2021 study, not all participants participated in all four phases, so the number of participants reported in the companion papers may vary depending on which aspect of the study is being reported.*

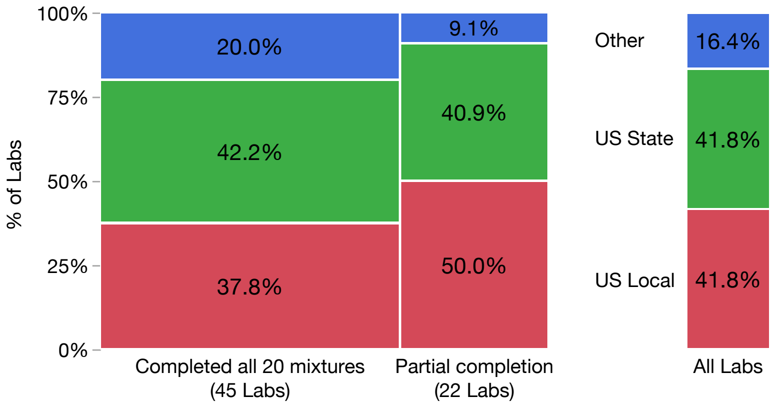

Fig S2. Completion vs. partial completion by lab type. (67 labs)

Lab Type	*Participants*		*Labs*	
Non-U.S. Federal/National laboratory	4	3%	3	4%
Non-U.S. State/Provincial laboratory	20	15%	7	10%
U.S. Local laboratory	56	42%	28	42%
U.S. Private laboratory	1	1%	1	1%
U.S. State laboratory	53	40%	28	42%
Totals	*134*		*67*	

Table S5. Participants and participating laboratories by type of laboratory.

Table S6 and Table S7 show the number of DNA analysts conducting mixture analysis in each participating laboratory (both overall by participant and weighted by laboratory).

Number of analysts in Lab	*Participants*		*Labs (raw)*		*Labs (weighted)*	
2 to 10	39	29%	25	34%	22.67	34%
11 to 25	55	41%	32	44%	29.67	44%
26 to 50	27	20%	11	15%	9.67	14%
51+	13	10%	5	7%	5	7%
Totals	134		73		67	

Table S6. Participants and participating laboratories by laboratory size. The participants for five of the labs with multiple participating subunits were not consistent in their responses. The weighted column weights the responses by the number of participants per lab: the resulting total equals the number of labs (67), but with fractional responses.

Lab Type	*Number of DNA analysts conducting mixture analysis per lab*				
	Total	*2 to 10*	*11 to 25*	*26 to 50*	*51+*
Non-U.S. Federal/National laboratory	3.0	2.0		1.0	
Non-U.S. State/Provincial laboratory	7.0	2.0	2.0	3.0	
U.S. Local laboratory	28.0	13.0	12.0	2.0	1.0
U.S. Private laboratory	1.0		1.0		
U.S. State laboratory	28.0	5.7	14.7	3.7	4.0
Total	67.0	22.7	29.7	9.7	5.0

Table S7. Participating labs by type of laboratory and laboratory size. The responses are weighted as described in Table S6.

Lab Type	*Total*	*ID28*	*GF28*	*GF29*	*6C29*
Non-U.S. Federal/National laboratory	3	1		1	1
Non-U.S. State/Provincial laboratory	7	4	1	1	1
U.S. Local laboratory	28		2	17	9
U.S. Private laboratory	1			1	
U.S. State laboratory	28		6	7	15
Total	*67*	*5*	*9*	*27*	*26*

Table S8. Participating labs by Amp/CE Settings. Note that ID28 was only used in non-U.S. labs.

Appendix E Test Yield

The 2,272 responses used for analyses include 1,507 responses from the *NoC Subtest* and 765 responses from the *ICSA Subtest*, reported by 134 participants from 67 laboratories on 29 distinct mixtures.

Appendix E1 Responses by Amp/CE settings and correspondence with SOPs

All participants completed a *Mixture Selection Questionnaire* in which they selected Amp/CE Settings (ID28, GF28, GF29, 6C29), and indicated how those settings corresponded to their SOPs (see *Appendix B2* for details). The participants were given these options to indicate how Amp/CE Settings corresponded with SOPs:

- This corresponds exactly to our lab's validated settings—we can use these settings for NoC and ICSA
- This is equivalent to our lab's settings; this differs in details we consider minor or inconsequential (such as injection time of 15 vs 16 seconds)—we can use these settings for NoC and ICSA
- This differs from our lab's validated settings, but we are willing to participate using these settings in both NoC and ICSA (Note: these results will be analyzed separately during analysis)
- This differs from our lab's validated settings; we are willing to participate using these settings in NoC, but not in ICSA (Note: these results will be analyzed separately during analysis)

For analyses, we group A and B (exact and equivalent) as "SameSOP", and C and D as "DiffSOP."

Table S9 shows the counts of participants and laboratories by Amp/CE settings, and by the correspondence of Amp/CE settings with their SOPs.

		Participants					*Labs*				
Amp/CE vs SOPs		*Total*	*ID28*	*GF28*	*GF29*	*6C29*	*Total*	*ID28*	*GF28*	*GF29*	*6C29*
Overall		134	19	22	43	50	67	5	9	27	26
SameSOP	A_EXACT	86	16	13	29	28	40	2	4	18	16
	B_EQUIV	23	1	5	8	9	18	1	3	6	8
DiffSOP	C_DIFFBOTH	23	2	3	5	13	15	2	3	5	5
	D_DIFFNOC	2	0	1	1	0	2	0	1	1	0
Subtotal (sameSOP)		109	17	18	37	37	58	3	7	24	24
Subtotal (diffSOP)		25	2	4	6	13	17	2	4	6	5

Table S9. Counts of participants and laboratories by Amp/CE settings and correspondence with SOPs. 81% of participants were *SameSOP* (64% exact, 17% equivalent); 87 % of labs were *SameSOP* (60 % exact, 27% equivalent). Note that the overall counts differ from the total of the other rows, because seven labs with multiple participants had responses in more than one category, two of which had responses in the *SameSOP* and *DiffSOP* categories.

In a few cases, participants indicated in their comments that specific individual responses did not follow their SOPs. These fell into two categories:

- In 13 trials, participants indicated the mixture was suitable but stated in comments that they would not have considered it suitable in casework (generally with comments indicating that the sample had less than their lab's template amount threshold) — these trials were flagged as *DiffSOP* even though the participant indicated that overall their Amp/CE settings were exact or equivalent to their SOPs.
- In 51 trials, participants indicated the mixture was not suitable but provided estimates of NoC in their comments — the NoC responses for these trials were flagged as *DiffSOP*.

Table S10 shows the resulting counts of responses used in analyses.

	Total	*ID28*	*GF28*	*GF29*	*6C29*	*Dataset Abbrev*
Total Participants	134	19	22	43	50	
Total Labs	67	5	9	27	26	
Total Responses	2,272	380	360	690	842	*AllResponse*
Responses Weighted by Lab	2,272	100	150	473	499	*WeightedResponse*

		Responses					*Dataset Abbrev*
		Total	*ID28*	*GF28*	*GF29*	*6C29*	
Suitability	DiffSOP	342	43	47	102	150	
	SameSOP	1,930	337	313	588	692	*SuitSame*
	Total	2272	380	360	690	842	
NoC	DiffSOP	393	53	59	121	160	
	SameSOP	1,879	327	301	569	682	*NoCSame*
	Total	2,272	380	360	690	842	

		Responses weighted by lab					*Dataset Abbrev*
		Total	*ID28*	*GF28*	*GF29*	*6C29*	
Suitability (weighted by lab)	DiffSOP	260.2	40.2	46.1	98.7	75.3	
	SameSOP	961.8	59.8	103.9	374.3	423.8	*WeightedSuitSame*
	Total	1,222.0	100.0	150.0	473.0	499.0	
NoC (weighted by lab)	DiffSOP	280.3	40.9	49.3	110.7	79.5	
	SameSOP	941.7	59.1	100.7	362.3	419.6	*WeightedNoCSame*
	Total	1,222.0	100.0	150.0	473.0	499.0	

Table S10. Counts of responses by Amp/CE settings and correspondence with SOPs.

Table S11 shows the counts of responses for each mixture, by Amp/CE.

PacketID	*NoC$_{GT}$*	*DNA Amplified (ng)*	*Assignment Group*	*Responses*				
				Total	*ID28*	*GF28*	*GF29*	*6C29*
ICSA_290/691	2	0.088	ICSA	94	19	15	25	35
NOC_52	2	0.054	NoC (28, 29b)	88	19	21	21	27
NOC_24	2	0.043	NoC (28, 29a)	76	19	21	17	19
ICSA_192/680	3	0.341	ICSA	99	19	16	26	38
NOC_49	3	0.191	NoC (29a)	37			18	19
NOC_74	3	0.186	NoC (29a)	38			18	20
NOC_28	3	0.180	NoC (29b)	51			23	28
ICSA_311/401	3	0.179	ICSA	97	19	16	27	35
ICSA_078/260	3	0.174	ICSA	93	19	15	27	32
NOC_84	3	0.159	NoC (28, 29b)	86	19	19	21	27
NOC_50	3	0.146	NoC (29b)	49			21	28
NOC_76	3	0.121	NoC (28, 29a)	75	19	18	18	20
NOC_25	3	0.099	NoC (28, 29a)	77	19	20	18	20
NOC_53	3	0.091	NoC (28, 29a)	75	19	19	18	19
NOC_57	3	0.090	NoC (29b)	48			21	27
NOC_29	4	0.872	NoC (28, 29b)	88	19	20	23	26
NOC_93	4	0.580	NoC (28, 29a)	77	19	20	18	20
NOC_15	4	0.580	NoC (ALL)	126	19	21	39	47
ICSA_057/802	4	0.486	ICSA	95	19	16	26	34
ICSA_671/828	4	0.481	ICSA	97	19	16	26	36
ICSA_370/530	4	0.479	ICSA	94	19	14	26	35
NOC_70	4	0.234	NoC (28, 29a)	76	19	19	18	20
NOC_05	4	0.211	NoC (29b)	49			22	27
NOC_14	4	0.210	NoC (29b)	49			22	27
NOC_68	4	0.188	NoC (29b)	50			23	27
NOC_41	4	0.171	NoC (29a)	37			18	19
NOC_31	5	0.720	NoC (ALL)	126	19	18	41	48
ICSA_328/767	5	0.376	ICSA	96	19	16	27	34
NOC_71	6	0.801	NoC (ALL)	129	19	20	42	48
Min				37	19	14	17	19
Max				129	19	21	42	48
Average				78.3	19.0	18.0	23.8	29.0

Table S11. Counts of responses per mixture by Amp/CE settings. Mixtures are sorted by NoC$_{GT}$ and amount of DNA. Assignment groups are discussed in Appendix C4.

Table S12 details the counts of responses for each mixture, by Amp/CE and correspondence with SOPs.

PacketID	*NoC$_{GT}$*	*ID28*			*GF28*			*GF29*			*6C29*		
		Diff	*S/D*	*Same*	*Diff*	*S/D*	*Same*	*Diff*	*S/D*	*Same*	*Diff*	*S/D*	*Same*
ICSA_290/691	2	2	0	17	2	0	13	3	0	22	2	0	33
NOC_52	2	2	0	17	3	0	18	6	0	15	9	0	18
NOC_24	2	2	0	17	3	0	18	1	0	16	3	1	15
ICSA_192/680	3	2	0	17	2	1	13	4	0	22	3	0	35
NOC_49	3							1	0	17	3	0	16
NOC_74	3							1	0	17	3	0	17
NOC_28	3							6	0	17	8	0	20
ICSA_311/401	3	2	0	17	2	1	13	4	0	23	2	0	33
ICSA_078/260	3	2	0	17	2	0	13	3	0	24	2	0	30
NOC_84	3	2	0	17	2	1	16	5	0	16	7	0	20
NOC_50	3							5	0	16	9	0	19
NOC_76	3	2	0	17	2	1	15	1	0	17	3	0	17
NOC_25	3	2	0	17	3	1	16	1	0	17	3	1	16
NOC_53	3	2	0	17	2	0	17	1	0	17	5	0	14
NOC_57	3							6	0	15	8	1	18
NOC_29	4	2	0	17	3	0	17	5	0	18	7	0	19
NOC_93	4	2	0	17	3	0	17	1	0	17	3	1	16
NOC_15	4	2	0	17	3	0	18	6	0	33	10	0	37
ICSA_057/802	4	2	0	17	2	1	13	3	1	22	2	1	31
ICSA_671/828	4	2	0	17	2	1	13	3	2	21	4	0	32
ICSA_370/530	4	2	0	17	2	1	11	4	1	21	3	0	32
NOC_70	4	5	1	13	2	1	16	1	0	17	3	0	17
NOC_05	4							5	0	17	7	0	20
NOC_14	4							5	0	17	7	0	20
NOC_68	4							5	1	17	7	0	20
NOC_41	4							1	0	17	3	0	16
NOC_31	5	2	1	16	2	0	16	6	5	30	11	3	34
ICSA_328/767	5	2	0	17	2	2	12	3	5	19	2	2	30
NOC_71	6	2	8	9	3	1	16	6	4	32	11	0	37
Min		2	0	9	2	0	11	1	0	15	2	0	14
Max		5	8	17	3	2	18	6	5	33	11	3	37
Average		2.2	0.5	16.4	2.4	0.6	15.1	3.5	0.7	19.6	5.2	0.3	23.5

Table S12. Counts of responses per mixture by Amp/CE settings (detail). Same data as Table S11 but with detail of counts by correspondence with SOPs. “S/D” indicates responses treated as *SameSOP* for suitability analyses but *DiffSOP* for NoC analyses.

Appendix E2 Weighting of responses by laboratory

Some labs had multiple subunits (we use “subunits” to refer to multiple participants within a lab), and there was notable variation in the number of subunits per lab: 48 labs had one participant each; 15 labs had two to five participants each; 4 labs had seven to 14 participants each. (See Appendix D for details)

To accommodate this, we report results both by participant and by lab:

- Results by participant simply treat each response equally, ignoring the number of participants per lab. N= 2,272 trials (mean of 1.9 responses per lab per mixture) — we refer to this as the *AllResponse dataset*
- Results by lab weight each response so that each lab collectively has one response for each mixture. Each response is weighted by 1/(responses by that lab for that mixture). N= 1,222 weighted trials (1 response per lab per mixture) — *WeightedResponse dataset*

As discussed in Section 3.3 (main paper), the datasets used in assessing reproducibility were created as a self join of the response data, pairing every response with every other response for the same mixtures. Weighting of the reproducibility datasets uses the product of the weights for the paired responses.

Appendix E3 Anomalies

During data curation and analysis of responses, we flagged any unusual responses for review by the study team. Upon review, each response was triaged into one of three groups:

- Anomaly: an answer or response that deviates from the norm or suggests an abnormality in the data that does not necessarily preclude its inclusion in analyses, but merits noting in reporting

- Modification: an answer or response that required manual updating by the study team based upon communication with the participant and/or review of any comments provided
- Omission: an answer or response that was flagged by the research team to be excluded from the analyses

Here, we summarize the anomalies, modifications, and omissions specific to the suitability/NoC analysis in an effort to provide complete transparency into the data curation and analysis process.

- We noted two anomalies in the suitability/NoC responses:
  - One participant did not complete the *P&P*, therefore precluding us from evaluating their responses versus their reported SOPs
  - One participant listed 0.268ng as the total quantity of DNA needed to move forward with amplification; the other two participants from their laboratory listed 0.1ng
- We flagged 17 responses that required modification of an answer to at least one question
  - Note that modifications were only made if they were specifically requested by the participant via email or indicated by the participant in their additional comments
  - All modifications were conducted by the study team to ensure that the responses accurately reflected what the participant intended to report
- We omitted 6 full responses (i.e., all answers to all questions for a given assignment)
  - One response was omitted because the participant indicated that they would report their NoC estimate as a range, but proceeded to select a single value
  - Four responses were omitted because participants entered a different packet number than the one they were supposed to be entering responses for, leading to concerns that the answers pertained to a different mixture
  - One response was omitted due to an apparent software error (all responses for the assignment were blank; all other assignment responses for that participant were saved properly)
- We omitted 6 individual answers (i.e., a single answer to a specific question)
  - Four comments were omitted due to a minor software bug in the additional comment box, which caused previous comment text to auto-fill in the comment box if left blank— this bug (which was limited to the comment box) was identified and fixed early in data collection, and all responses collected prior to the fix were reviewed for potential issues
  - One answer was omitted because a participant indicated that a question did not activate for them on one assignment
  - One answer was omitted because a participant indicated in their additional comments that the question did not apply

In addition to these anomalies, all comments were reviewed to assess whether they affected whether responses were *SameSOP* or *DiffSOP*:

- 13 suitability and $\mathrm{NoC}_{\mathrm{EST}}$ responses were assessed as *DiffSOP* based on comments in individual trials (for participants that were *SameSOP* in general).
- 51 $\mathrm{NoC}_{\mathrm{EST}}$ responses were provided as comments on *NotSuit SameSOP* trials. Note that if a participant indicated a trial was *NotSuit*, they were not prompted to provide $\mathrm{NoC}_{\mathrm{EST}}$ responses. These 51 $\mathrm{NoC}_{\mathrm{EST}}$ responses were treated as *DiffSOP*.

Note that many comments were unclear or otherwise hard to adjudicate. If in doubt, we flagged the response as *DiffSOP*.

Appendix F Details of P&P Results Relevant to Suitability

Appendix F1 Policies to terminate analysis prior to amplification

The DNAmix 2021 P&P Questionnaire [12] included three questions related to terminating analysis prior to amplification:

- PP#04. Do your SOPs include policies that terminate analysis prior to amplification based on total DNA quantity?
  - No (26% of labs)
  - Yes (74% of labs)
- PP#05. In cases where the person of interest is male, do your SOPs include policies that terminate analysis prior to amplification based on proportion of total DNA that is male?
  - No (43% of labs)
  - Yes (55% of labs; total does not sum to 100% due to inconsistent responses)
- PP#06. Do your SOPs include policies that terminate analysis prior to amplification based on other factors (other than DNA quantity or proportion of male DNA)? [57 participants from 30 labs indicated Yes and provided the following text responses]
  - "Redundant sample with higher quality, case cancellation"
  - "Ratio of male v. female in a sample"
  - "insufficient male"
  - "If the standard fails, IPC fails, or the degradation index shows inhibition"
  - "termination cut offs are modified based on Degradation Index from Quant Trio"
  - "In certain Sexual Assault (SA) cases, the non-probative fraction developed from a differential extraction can stop at quant."
  - "If the sample doesn't meet quality standards"
  - "If there are multiple samples in a case that have been screened for high throughput route in the lab, a sample may be stopped if a more probative sample can be tested. The other responses for #4 and #5 are also for the high throughput route."
  - "For single male subject, female victim sexual assault kit samples, if multiple samples present with enough male DNA (quantity and ratio) to amplify, the samples may be further triaged such that a single best sample from internal orifice swabs is amplified, followed by external genital and other body swabs if needed."
  - "When a sample yields a failed IPC at quantitation more than once, analysis may be terminated."
  - "If the person of interest is male and there is no male DNA the sample will be stopped prior to amplification. Also if the total quantity of male DNA is 0.001ng or less the sample will also be stopped prior to amplification."
  - ".005 Male ng/uL"
  - "sexual assaults with specified ratio and no suspect property crimes with undetected DNA; other case types all get amplified."
  - "In a direct to DNA approach for sexual assault cases, our SOPs allow an analyst to select the most informative sample(s) for amplification based on male DNA detected and total human to male DNA ratio. All samples are extracted and quanted but only the "best" 1-3 samples move forward to amplification in most cases."
  - "If the person of interest is male and 0.02 ng or less of Y is indicated at quant."
  - "Auto to Y ratio greater than 200:1"
  - "If the POI is male and no male DNA is detected, the sample analysis may be terminated."
  - "Sexual assault kit evidence consisting of a female complainant and male suspect is terminated prior to amplification if no male DNA is detected at quant."
  - "If the sample is inhibited (reflected by out of range IPC) and the inhibition does not get addressed despite attempts to get rid of the inhibition and no more/alternative sample is available to carry out re-extraction."
  - "Auto/Y ratio for forwarding to Y-STR processing is an option"
  - "Other probative and/or higher quantity samples going forward."
  - "If only one male suspect in Sexual Assault kit, and multiple samples are eligible to go forward, only one will go forward"

- “We test all samples in sexual assault kits as part of the sexual assault kit law. Will only bring best sample to amplification and once CODIS eligible profile is obtained, rest of sample will not be amplified (exceptions are multiple assailants, etc.).”
- “when male DNA is detected in a sample that is greater than or equal to 0.002 ng/µl, report the presence of male DNA in the sample. When the ratio of total human DNA to male DNA ([Auto]/[Y]) is unsuitable for autosomal STR analysis, this information will be reported. Generally, the laboratory uses a ratio of 40:1 (total human DNA [Auto]/ to male DNA [Y]) when determining the suitability for autosomal STR analysis, depending on case specifics.”
- “In cases where there is only sufficient DNA for Y-STRs but there is not a suspect, analysis would routinely be suspended.”
- “Auto/Y ratio for forwarding to Y-STR processing is an option.”
- “auto:Y of 100:1 or greater”
- “Terminate based upon proportion of male based upon dilution, not percentage. If the sample would require a dilution to the extent that the male portion would fall below the minimum quantity necessary to obtain a result, we would cease amplification.”
- “Auto/Y ratio for forwarding to Y-STR processing is an option”
- “If male value is <.0006 ng/ul in a sample in which the male is the probative value or there is not a value value present, insufficient male DNA or no male DNA is reported and the sample will not be analyzed. In instances in which the small autosomal value is "undetermined", the sample will not be amplified.”
- “Analysis can be stopped at any step if exams are canceled by the submitting agency.”
- “We test all samples in sexual assault kits as part of the sexual assault kit law. Will only bring best sample to amplification and once CODIS eligible profile is obtained, rest of sample will not be amplified (exceptions are multiple assailants, etc.)”
- “male screening - only choosing one sample to move forward even if all samples in the case are suitable for amplification”
- “Using case details the analyst determines if all samples need to be amplified or if "best sample" will suffice for the case. Examples would include SAKs - not amplifying both fractions or all items if only one potential donor, or for a B&E if similar samples submitted, select one for amplification”
- “Sexual Assault cases - limited number of samples are amplified based on case scenario”
- “For sexual assault cases, a limited number of samples are amplified based on the case scenarios.”
- “For sexual assault cases, a limited number of samples are amplified based on case scenario”
- “for sexual assault cases, samples are amplified based on case scenario”
- “auto/Y ratio for forwarding to y-str processing is an option”
- “Similar probative value to other samples being amplified.”
- “Analysts allowed discretion to terminate STR analysis in favor of Y-STR analysis based on amount or proportion of male DNA and degradation (however policy does not define specific thresholds), probative value, or availability of other probative/higher quality samples.”
- “We test all samples in sexual assault kits as part of the sexual assault kit law. We only bring the best sample to amplification and once a CODIS eligible profile is obtained, the rest of the samples will not be amplified, except when there are multiple assailants, etc.”
- “Auto/Y ratio for forwarding to Y-STR processing is an option”
- “Auto/Y ratio for forwarding to Y-STR processing is an option”
- “We terminate analysis where the POI is male based on the above portion of the sample being male and the quantity of male DNA exceeding 0.002 ng/uL.”
- “Based on the shape of the sample quantitation curve, a sample may contain possible animal DNA. Further testing, using Hematrace, can be used and termination of further analysis can occur.”
- “other probative/positive samples in case available”
- “We test all samples in sexual assault kits as part of the sexual assault kit law. Will only bring best sample to amplification and once CODIS eligible profile is obtained, rest of sample will not be amplified (exceptions are multiple assailants, etc.).”
- “Female to male ratio exceeds 200:1”
- “We use multiple DNA quantity values. 0.005 for all samples based on short autosomal target. Fore sexual assault cases with males suspect, all samples that obtain an undetected quantification value is terminated.”

- "We have a policy that says that if the male is the probative part of the sample, typically we need 50pg of male DNA to be able to be amplified - this is impacted by the amount of female DNA present. This has to do with both the human and male quantity but not strictly proportion, it is about the amount in the volume being added to amp."
- "We consider a Human:Male ratio and do not proceed with amplification above a H:M of 1:31."
- "Single male assailant, no consenting, the non-sperm fraction can be stopped at quant if the sperm fraction is robust on internal swabs."
- "We may not amplify samples that are suitable if other samples in the case are being amplified."
- "Male quant value 0.0001 if male probative"
- "If person of interest is male, analysis may also be terminated if male DNA quantity is <4.0E-04 ng/uL."
- "For sexual assault forensic evidence kit examinations, three samples are taken through quantitation but only one sample is taken forward to amplification provided certain case circumstances are met."

Appendix F2 Suitability of Unknown NoC

PP#37 asked "How do you report if there is uncertainty on the assumed number of contributors in a mixture?"

Participants selected among the following responses:

- The mixture is not suitable for comparison due to the unknown number of contributors (28% of labs)
- Report only with respect to major contributors (17% of labs)
- Report the statistical value that provides the lowest evidential weight (i.e., most conservative result: lowest LR, highest RMP, highest CPI) (10% of labs)
- Report multiple statistical values (3% of labs)
- Report one statistical value that accounts for variable number of contributors (e.g., VarNoC) (5% of labs)
- The following text responses were provided by the 49 participants from 24 labs that indicated "Other":
  - "Report only with respect to major contributors with TL approval"
  - "BP Sentry will use AIC to provide the best fit model considering a range of contributors, stutter, and degradation."
  - "complete uncertainty - uninterpretable if two sets of NOC are run, report the lowest LR"
  - "Declaring only part of the mixture suitable for comparison OR declaring the entire mixture not suitable for comparison"
  - "Deemed either too complex or can interpret under different NoCs"
  - "depending on the mixture, either the mixture is not suitable for comparison due to uncertainty or only certain components of the mixture are suitable for comparison"
  - "Depending on the mixture: mixture is not suitable for comparison due to the unknown number of contributors, or report only with respect to major contributors if the major contributors are clear"
  - "Depending on the obtained mixed profile, either the first or the second response."
  - "Depending on the profile, the mixture may not be suitable for comparison, multiple statistical values may be reported or VarNoC may be used with DNA TL approval."
  - "Depending on the sample- it could be that it is not suitable OR that multiple statistical values are reported based upon the different NOCs. We do not have VarNoC at this time."
  - "Depends on amount of uncertainty, if NoC cannot be determined at all, then not suitable for comparison, if NoC is within allowed range (2-3 or 3-4 contributors then VarNoC allowed with TL approval)"
  - "Either deemed the profile not suitable for comparison due to complexity OR report the LR for the various NoCs"
  - "either not suitable for comparison or report multiple statistical values"
  - "Either report multiple statistical values or report that the mixture is not suitable for comparison due to unknown NOC"
  - "entire mixture may be considered not suitable for comparison due to uncertainty, or the major contributor(s) may be deemed suitable with minor contributor(s) considered not suitable due to uncertainty"

- "If a clear major can be interpreted, then uncertainty will only be taken into account for the minor(s) - these will be reported as not suitable. If the entire mixture would be affected by uncertainty, it will be reported as not suitable due to the unknown number of contributors."
- "If it is determined during the interpretation phase of the sample that the NoC can't reasonably determined, then we report that the mixture is not suitable for comparisons due to uncertainty in the number of contributors. VarNoC may be considered for some samples, or if more than one NoC is considered and LRs calculated for each assumed NoC, then all LRs reported."
- "If multiple contributor assignments are being considered, comparisons and LR calculations will not be performed initially. LR calculations will only be performed for considerations that resulted in acceptable deconvolutions. All tested will be maintained in the case record deconvolutions will be retained in the case record."
- "if the mixture can be explained by both contributor assessments, then STRmix is run under both contributor scenarios; report the stat that provides the lowest evidential weight (i.e. most conservative)"
- "If there is uncertainty in NOC, the analyst may conclude the mixture is not suitable for comparisons, may report with respect to the major while saying additional observed DNA is insufficient for comparisons, or may interpret the mixture using STRmix at N-1 and/or N+1 and report the lowest LR value."
- "If unable to determine number of contributors (partial profile) - uninterpretable If 5 or more contributors - too complex for interpretation"
- "If uncertainty in overall mixture, reported as 'not suitable for comparison'. In some cases with major contributor(s), may report major contributors suitable for comparison while remainder of mixture not suitable due to uncertainty."
- "It depends... some mixtures may be wholly inconclusive (e.g., if all contributors are of equal intensity); some mixtures may be able to declare a single unambiguous major contributor while the total number of contributors is uncertain."
- "Mixture is not suitable for comparison due to unknown NOC or limited information. We also have the ability to use VarNOC."
- "Mixture may be suitable in part. One or more components may be deemed not suitable for comparison."
- "Options 1 & 4 - varNOC requires TL consult before use"
- "Options 1 and 4. VarNoc requires TL consult before use"
- "profile will be reported as 'at least X' contributor and more conservative assessment done"
- "Rare situations: report as inconclusive NOC (likely due to family/allele sharing)"
- "report a minimum NoC and no further statistical analysis unless a major can be reported"
- "Report one statistical value that accounts for variable number of contributors (e.g. RMNE) OR description (no statistical analysis)"
- "Report only with respect to contributors assessed to be suitable for comparison"
- "Report the lower end of the uncertain NoC range (i.e. if may be 3 or 4 contributors, attempt and report the value with 3)"
- "Report whether mixture is suitable for comparison, or only "in part" (some of the contributors). Report LRs only to suitable portions."
- "statistics for only interpretable components of the mixture, trace component not suitable for comparison"
- "The entire mixture can be deemed unsuitable or the major(s) can be interpreted (manually or with PG) and min(s) deemed unsuitable for interpretation/comparison."
- "The first three options are all allowed under lab procedure, and which is selected is mixture dependent. Additionally, as there is not a comment box for question 33, the question is not worded in a way that allows for a clear answer based on lab procedure. A person of interest (unless an assumed contributor) is not factored into a decision to interpret/not interpret or into NOC decisions. However, SOP allow that interpretation is a process and unintuitive statistics must be further evaluated."
- "The first three options in this list may be used depending on the nature of the profile"
- "The mixture is not suitable for comparison due to the complexity of the mixture."
- "The mixture is not suitable for comparison due to the unknown number of contributors UNLESS a major component is visible. That component can be used for comparison, and the minor component will not be used for comparison."
- "The mixture is uninterpretable due to limited data or complex data"

- "The mixture may be interpreted under more than one NOC assumption, resulting in multiple statistical values. The mixture may be deemed not suitable for interpretation if there were to be very limited information in the mixture and an assessment of the number of contributors was not feasible."
- "Typically would use "not suitable for comparisons due to..." but depending upon the mixture, sometimes a major can be reported out with NoC being uncertain"
- "Use TrueAllele with and report one statement that encompasses what could be the highest number of contributors present. "Assumed to be from a minimum of three contributors...""
- "We always report according to the minimal NoC"
- "we can either report that the mixture is not suitable for comparisons or report multiple statistical values"
- "We report a minimum number of contributors such as at least 2 donors, or at least 3 donors and perform statistics on that number."
- "We will utilize option 1 (not suitable for comparison) if no NOC can be reasonably assumed from the beginning. If we vary NOC during interpretation, however, we will only report the single NOC that best explains all of the available data."
- "Would choose what is determined to be the minimum and give additional wording in the reporting statement."

Appendix G Details of Suitability Results

The following tables present the suitability details for each of the 29 mixtures:

- Table S13 presents the suitability responses, weighted by laboratory to result a total of one response per lab per mixture. These are used for all suitability analyses by lab reported in this paper except for analyses limited to only *SameSOP* responses.
- Table S14 presents the *SameSOP* suitability responses, weighted by laboratory to result a total of one *SameSOP* response per lab per mixture. These are used for the suitability analyses limited to *SameSOP* responses. Separate weighting of *SameSOP* responses is necessary because some laboratories have participating subunits that differed overall on *SameSOP* vs. *DiffSOP*, and because some participating subunits that were *SameSOP* overall indicated in comments that specific suitability responses were *DiffSOP*.
- Table S15 presents the raw (unweighted) suitability responses. These are only shown here for completeness. Weighting is necessary because of the wide range of responses per lab (reporting raw responses would provide undue emphasis to those labs with multiple participants).

NoC_{GT}	*DNA (Ng)*	*Mixture*	*Weighted responses (1 response per lab per mixture)*												
				SameSOP						*DiffSOP*					
			Total (AllSOP)	*Subtotal (SameSOP)*	*NotSuit*	*PartSuit*			*YesSuit*	*Subtotal (DiffSOP)*	*NotSuit*	*PartSuit*			*YesSuit*
						Contrib	*Loci*	*Contrib & Loci*				*Contrib*	*Loci*	*Contrib & Loci*	
2	0.088	ICSA_290/691	56	38.0	7.2	0.6	1.1		29.0	9.0	1.0		1.0		7.0
	0.054	NOC_52	61	33.5	12.6	2.7	1.1	0.1	17.1	13.5		0.5	1.0		12.0
	0.043	NOC_24	47	29.5	6.4	1.2	1.6		20.3	8.5	4.5				4.0
3	0.341	ICSA_192/680	60	39.7	4.7	3.8	1.5	1.7	28.0	10.3	2.0	1.3	1.0		6.0
	0.191	NOC_49	28	20.5	2.0	0.4			18.1	3.5	2.5				1.0
	0.186	NOC_74	29	21.5	4.4	1.6			15.5	3.5	1.5		1.0		1.0
	0.180	NOC_28	42	27.7	5.3	2.2			20.2	7.3		0.3			7.0
	0.179	ICSA_311/401	58	38.5	13.3	0.8		0.1	24.3	9.5	2.5				7.0
	0.174	ICSA_078/260	54	36.0	5.8	0.8	1.5		28.0	9.0	2.0				7.0
	0.159	NOC_84	58	36.0	8.2	3.9	0.7		23.2	11.0	1.0				10.0
	0.146	NOC_50	42	26.0	4.3	3.8			17.8	8.0			1.0		7.0
	0.121	NOC_76	46	30.5	13.8	1.5	0.4		14.8	7.5	3.5	1.0			3.0
	0.099	NOC_25	48	30.5	8.7	1.9	1.1	0.2	18.6	8.5	3.5				5.0
	0.091	NOC_53	46	28.5	9.3	0.2	0.2		18.8	8.5	2.5				6.0
	0.090	NOC_57	41	25.5	9.2	0.5		0.5	15.3	7.5		0.5			7.0
4	0.872	NOC_29	60	36.0	5.1	1.7	1.2		28.0	12.0	1.0		1.0		10.0
	0.580	NOC_93	48	30.5	12.2	2.3		0.2	15.9	8.5	5.5				3.0
	0.580	NOC_15	79	48.8	11.1	1.2	2.2		34.3	15.3	2.2		0.3		12.8
	0.486	ICSA_057/802	55	37.0	13.3	1.2	0.4	0.3	21.8	9.0	3.0	1.0			5.0
	0.481	ICSA_671/828	57	37.0	12.8	5.0		1.3	17.9	10.0	2.5	1.0		2.0	4.5
	0.479	ICSA_370/530	58	36.5	5.5	0.8	2.1		28.1	10.5	4.5				6.0
	0.234	NOC_70	46	30.3	13.3	1.2	0.3		15.4	7.7	4.5				3.2
	0.211	NOC_05	41	27.0	8.2	2.8			16.0	7.0	1.5				5.5
	0.210	NOC_14	42	28.0	9.3	1.8			16.8	7.0	0.8				6.2
	0.188	NOC_68	42	28.0	10.3	2.2			15.5	7.0	1.2				5.8
	0.171	NOC_41	28	20.5	5.5	0.6		0.2	14.2	3.5	2.5				1.0
5	0.720	NOC_31	79	50.8	45.4	1.4			4.0	14.3	10.1				4.1
	0.376	ICSA_328/767	56	38.0	35.5	0.3			2.3	9.0	6.0				3.0
6	0.801	NOC_71	80	51.7	48.6	2.0	0.1		1.0	14.4	12.4				2.0

Table S13. Suitability responses, weighted to 1 response per lab per mixture. (Note that the total of weighted responses is an integer, the number of labs that provided responses for that mixture.)

NoC_{GT}	*DNA (ng)*	*Mixture*	*Weighted responses (1 SameSOP suitability response per lab per mixture)*					
			Total	*SameSOP*				
				NotSuit	*PartSuit*			*YesSuit*
					Contrib	*Loci*	*Contrib & Loci*	
2	0.088	ICSA_290/691	38	7.2	0.6	1.1		29.0
	0.054	NOC_52	34	13.1	2.7	1.1	0.1	17.1
	0.043	NOC_24	30	6.9	1.2	1.6		20.3
3	0.341	ICSA_192/680	40	4.8	3.8	1.5	1.7	28.2
	0.191	NOC_49	21	2.5	0.4			18.1
	0.186	NOC_74	22	4.9	1.6			15.5
	0.180	NOC_28	28	5.3	2.3			20.3
	0.179	ICSA_311/401	39	13.3	0.8		0.1	24.8
	0.174	ICSA_078/260	36	5.8	0.8	1.5		28.0
	0.159	NOC_84	36	8.2	3.9	0.7		23.2
	0.146	NOC_50	26	4.3	3.8			17.8
	0.121	NOC_76	31	14.3	1.5	0.4		14.8
	0.099	NOC_25	31	9.2	1.9	1.1	0.2	18.6
	0.091	NOC_53	29	9.8	0.2	0.2		18.8
	0.090	NOC_57	26	9.7	0.5		0.5	15.3
4	0.872	NOC_29	36	5.1	1.7	1.2		28.0
	0.580	NOC_93	31	12.7	2.3		0.2	15.9
	0.580	NOC_15	49	11.3	1.2	2.2		34.4
	0.486	ICSA_057/802	37	13.3	1.2	0.4	0.3	21.8
	0.481	ICSA_671/828	37	12.8	5.0		1.3	17.9
	0.479	ICSA_370/530	37	6.0	0.8	2.1		28.1
	0.234	NOC_70	31	14.0	1.2	0.3		15.5
	0.211	NOC_05	27	8.2	2.8			16.0
	0.210	NOC_14	28	9.3	1.8			16.8
	0.188	NOC_68	28	10.3	2.2			15.5
	0.171	NOC_41	21	6.0	0.6		0.2	14.2
5	0.720	NOC_31	51	45.6	1.4			4.0
	0.376	ICSA_328/767	38	35.5	0.3			2.3
6	0.801	NOC_71	52	48.9	2.0	0.1		1.0

Table S14. Suitability responses, weighted to 1 *SameSOP* suitability response per lab per mixture. These results are shown graphically in Figure 1 (main paper). (Note that the total of weighted responses is an integer, the number of labs that provided *SameSOP* suitability responses for that mixture.)

			Raw Responses												
				SameSOP						*DiffSOP*					
						PartSuit						*PartSuit*			
NoC_{GT}	*DNA (Ng)*	*Mixture*	*Total*	*Subtotal (SameSOP)*	*NotSuit*	*Contrib*	*Loci*	*Contrib & Loci*	*YesSuit*	*Subtotal (DiffSOP)*	*NotSuit*	*Contrib*	*Loci*	*Contrib & Loci*	*YesSuit*
2	0.088	ICSA_290/691	94	85	29	3	2		51	9	1		1		7
	0.054	NOC_52	88	68	33	4	6	1	24	20		1	1		18
	0.043	NOC_24	76	67	29	2	6		30	9	5				4
3	0.341	ICSA_192/680	99	88	16	13	2	3	54	11	2	2	1		6
	0.191	NOC_49	37	33	3	2			28	4	3				1
	0.186	NOC_74	38	34	8	4			22	4	2		1		1
	0.180	NOC_28	51	37	7	4			26	14		1			13
	0.179	ICSA_311/401	97	87	39	4		1	43	10	3				7
	0.174	ICSA_078/260	93	84	28	3	3		50	9	2				7
	0.159	NOC_84	86	70	29	7	4		30	16	1				15
	0.146	NOC_50	49	35	6	6			23	14			1		13
	0.121	NOC_76	75	67	36	4	3		24	8	4	1			3
	0.099	NOC_25	77	68	34	6	2	1	25	9	4				5
	0.091	NOC_53	75	65	34	1	2		28	10	3				7
	0.090	NOC_57	48	34	12	1		1	20	14		1			13
4	0.872	NOC_29	88	71	8	4	3		56	17	1		1		15
	0.580	NOC_93	77	68	27	4		1	36	9	6				3
	0.580	NOC_15	126	105	22	7	6		70	21	3		1		17
	0.486	ICSA_057/802	95	86	27	6	3	1	49	9	3	1			5
	0.481	ICSA_671/828	97	86	27	20		2	37	11	3	1		2	5
	0.479	ICSA_370/530	94	83	13	4	7		59	11	5				6
	0.234	NOC_70	76	65	36	2	1		26	11	5				6
	0.211	NOC_05	49	37	11	6			20	12	4				8
	0.210	NOC_14	49	37	11	4			22	12	5				7
	0.188	NOC_68	50	38	13	6			19	12	2				10
	0.171	NOC_41	37	33	9	3		1	20	4	3				1
5	0.720	NOC_31	126	105	86	3			16	21	16				5
	0.376	ICSA_328/767	96	87	70	1			16	9	6				3
6	0.801	NOC_71	129	107	103	2	1		1	22	20				2

Table S15. Raw (unweighted) suitability responses.

Appendix G1 Reasons for NotSuit responses

Table S16 shows the proportions of responses for each mixture that cited each of the given reasons for deciding a mixture was *NotSuit* (see question 8 in *Appendix B2h*). The table uses these abbreviations:

- Loci: Not enough alleles or loci suitable for analysis
- Levels: DNA template levels too low overall
- Degraded: Sample too degraded
- Inhibited: Sample too inhibited
- NoC: Too many contributors
- Uncertain NoC: Too much uncertainty in the number of contributors
- Ratios: Mixture proportions or contributor ratios

NoC$_{GT}$	*Mixture*	*All responses (weighted by lab))*							*SameSOP (weighted by lab)*						
		Loci	*Levels*	*Degraded*	*Inhibited*	*NoC*	*Uncertain NoC*	*Ratios*	*Loci*	*Levels*	*Degraded*	*Inhibited*	*NoC*	*Uncertain NoC*	*Ratios*
2	ICSA_290/691	3%	12%	1%			5%	3%	3%	15%	1%			3%	1%
	NOC_52	12%	25%	3%			14%	9%	16%	37%	4%			20%	13%
	NOC_24	16%	22%	1%	1%		10%	4%	10%	19%	1%	1%		6%	1%
3	ICSA_192/680	2%	1%	4%	1%	5%	3%	8%	2%	1%	3%	1%	6%	5%	5%
	NOC_49	2%	10%			6%	15%	10%		7%			5%	7%	10%
	NOC_74	2%	4%			12%	13%	5%		5%			9%	11%	6%
	NOC_28	1%	9%			3%	4%	7%	2%	11%			4%	5%	8%
	ICSA_311/401	3%	10%			16%	12%	16%	3%	10%			17%	11%	15%
	ICSA_078/260	2%	7%			8%	2%	9%	3%	8%			10%	3%	6%
	NOC_84	4%	10%			3%	3%	11%	5%	12%			4%	4%	12%
	NOC_50	7%	6%			1%	4%	4%	10%	8%			2%	5%	5%
	NOC_76	4%	11%			23%	15%	16%	4%	13%			23%	13%	14%
	NOC_25	5%	22%			5%	12%	14%	6%	22%			6%	7%	11%
	NOC_53	5%	19%			9%	7%	15%	5%	20%			11%	5%	15%
	NOC_57	11%	20%			1%	10%	8%	13%	28%			1%	14%	10%
4	NOC_29					5%	6%	8%					7%	9%	7%
	NOC_93	2%				29%	25%	23%	2%				25%	18%	21%
	NOC_15					14%	9%	11%					16%	9%	13%
	ICSA_057/802	4%	5%	17%	5%	15%	16%	21%	5%	6%	16%	6%	13%	15%	18%
	ICSA_671/828		2%			25%	16%	2%		3%			25%	18%	3%
	ICSA_370/530		2%			14%	11%	13%		3%			13%	8%	6%
	NOC_70		6%			37%	17%	20%		7%			32%	14%	19%
	NOC_05	4%	7%			20%	15%	11%	5%	9%			20%	19%	14%
	NOC_14	3%	6%			25%	10%	15%	4%	7%			29%	12%	19%
	NOC_68	3%	3%			30%	12%	15%	4%	4%			33%	15%	19%
	NOC_41	2%	4%			21%	17%	16%		5%			14%	12%	21%
5	NOC_31					83%	16%	11%					87%	15%	14%
	ICSA_328/767		2%			84%	9%	13%		3%			88%	8%	14%
6	NOC_71					88%	16%	13%					89%	18%	14%
% of all responses (weighted)		3%	7%	1%	0%	23%	11%	11%	3%	8%	1%	0%	23%	11%	11%
% of NotSuit (weighted)		8%	20%	3%	1%	64%	31%	32%	9%	23%	3%	1%	65%	30%	31%

Table S16. Reasons for NotSuit responses by mixture, weighted by lab. Other than the last row (which shows percentages of NotSuit responses), values are percentages of all responses. Red highlight indicates ≥50% of weighted responses; Yellow indicates ≥25% (<50%), Black text indicates ≥10% (<25%), Gray text indicates <10%.

Frequent combinations:

- Levels OR NoC: 79.6% of all responses (80.8% of *SameSOP*)
- Levels OR NoC OR Uncertain NoC OR Ratios: 93.5% of all responses (92.6% of *SameSOP*)

For 87 responses (78 *SameSOP* responses), participants indicated "Other" as a reason for *NotSuit*. The following were mentioned in at least 20 responses:

- Need replicate amplification
- Details about quantity of DNA needed[*]
- Lacking major contributors

[*] *Note that the metadata for each mixture provided to participants included the total amount of DNA amplified, DI, and the amount of male DNA.*

Appendix G2 Predicting Suitability Assessments based on P&P Settings

As discussed in *Section 4*, many labs have P&P settings related to suitability. With respect to this study, these can be split into these groups:

- Laboratory P&P settings that we can assess against specific values known to participants: suitability thresholds based on DNA amount, DI, or the male proportion of DNA (if the POI for an ICSA mixture is male).
- Laboratory P&P settings that we can assess against ground truth (not known to participants) OR against participants' assessments (if contrary to ground truth): suitability thresholds based on NoC.
- Laboratory P&P suitability settings that we cannot objectively assess: minimum number of loci with data, minimum number of alleles called with data, mixtures with an unknown number of minor contributors, or lab-specific suitability decision factors.

Table S17 shows the effectiveness of predictions based on P&P settings. Note that in practice, suitability thresholds based on DI or the male proportion of DNA had marginal value as predictors.

		Actual responses			
		SameSOP		AllSOP	
		NotSuit	YesSuit	NotSuit	YesSuit
Predictions based on DNA amount	predictNotSuit	80.8%	19.2%	71.3%	28.7%
and NoC threshholds	predictYesSuit	26.4%	73.6%	25.7%	74.3%
Predictions based on DNA amount, NoC, DI,	predictNotSuit	79.5%	20.5%	70.5%	29.5%
and male proportion threshholds	predictYesSuit	26.5%	73.5%	25.8%	74.2%

Table S17. Predicted vs actual suitability based on P&P responses.

Table S18 details the responses contrary to P&P settings, apparently due to inaccurate NoC$_{EST}$. These are shown as yellow hashed areas in Figure 2 (main paper).

				Raw responses			*Weighted responses*		
NoC$_{GT}$	*P&P NoC threshhold*	*NoC$_{EST}$*	*Suitability*	*Total*	*SameSOP*	*DiffSOP*	*Total*	*SameSOP*	*DiffSOP*
4	3	2-8	YesSuit	1	1	0	1	1	0
		3-8	PartSuit	3	3	0	2.1	2.1	0
		3-8	YesSuit	4	4	0	1.4	1.4	0
		1	YesSuit	1	1	0	0.1	0.1	0
		3	PartSuit	13	12	1	2.9	2.6	0.3
		3	YesSuit	10	7	3	3.9	0.9	3
5	3	3-8	PartSuit	1	1	0	1	1	0
	4	4-5	YesSuit	2	2	0	2	2	0
		4-8	YesSuit	3	0	3	3	0	3
		4	YesSuit	21	20	1	1.7	1.6	0.1
6	3	3-8	PartSuit	1	1	0	1	1	0
	5	4-5	YesSuit	1	0	1	1	0	1
		5-8	YesSuit	1	1	0	1	1	0

Table S18. *YesSuit* and *PartSuit* responses contrary to that lab's P&P settings, apparently due to inaccurate NoC$_{EST}$. Note that rows with large raw response counts but small weighted response counts highlight examples of multiple participants from a lab making the same responses; for example, 19 of the 21 raw responses in [NoC$_{GT}$=5, NoC$_{EST}$=4] are from a single lab (on 2 different mixtures).

Appendix G3 Reproducibility of Suitability Assessments

Reproducibility is the extent to which responses from different participants/labs agree when given the same mixture—each individual suitability assessment is paired with every other suitability assessment on the same mixtures, resulting in a summary of individual pairwise comparisons.

Table S19 details the reproducibility of suitability assessments weighted by laboratory.

			2-way (ignore PartSuit)		*2-way (treat PartSuit as YesSuit)*		*3-way*		
			Agree	*Disagree*	*Agree*	*Disagree*	*Agree*	*Partial*	*Disagree*
All	AllSOP	All	66%	34%	66%	34%	58%	12%	30%
DiffLab	AllSOP	All	65%	35%	66%	34%	58%	12%	30%
SameLab	AllSOP	All	90%	10%	86%	14%	79%	13%	8%
All	SameSOP (both)	All	66%	34%	66%	34%	58%	13%	29%
DiffLab	SameSOP (both)	All	66%	34%	66%	34%	57%	13%	29%
SameLab	SameSOP (both)	All	91%	9%	86%	14%	79%	13%	8%
DiffLab	SameSOP (both)	US lab (both)	68%	32%	69%	31%	60%	12%	28%
DiffLab	SameSOP (both)	NoC$_{GT}$ 2-4	59%	41%	61%	39%	51%	15%	34%
DiffLab	SameSOP (both)	NoC$_{GT}$ 5-6	90%	10%	85%	15%	85%	6%	9%
DiffLab	SameSOP (both)	NoC$_{GT}$=2	60%	40%	62%	38%	51%	15%	34%
DiffLab	SameSOP (both)	3	61%	39%	64%	36%	52%	15%	33%
DiffLab	SameSOP (both)	4	58%	42%	59%	41%	49%	15%	36%
DiffLab	SameSOP (both)	5	86%	14%	83%	17%	83%	4%	13%
DiffLab	SameSOP (both)	6	96%	4%	89%	11%	88%	8%	4%

Table S19. Reproducibility of Suitability Assessments (*Interlab and InterlabSuitSameSOP datasets*).

Fig S3 summarizes the reproducibility of conclusions, based on all pair-wise combinations of *SameSOP* responses from different labs on the same mixtures. The y-axis is associated with individual responses (reported by a lab for a given mixture), whereas the x-axis is associated with responses by all other labs (on the same mixtures). For example (in the top row of Fig S3), when one lab reported *YesSuit*, 64.0% of the other labs also responded *YesSuit*, 8.2%% responded *PartSuit*, and 27.8% responded *NotSuit*. Overall (if *PartSuit* is not distinguished from *YesSuit*), different labs agree 66% of the time (for both *AllSOP* and *SameSOP*); *SameSOP* US labs agree 69% of the time. Intra-lab reproducibility (i.e. responses within the same lab, for labs with multiple participating subunits) was higher: 86% agreement. Different labs are more likely to agree on the suitability of mixtures with NoC$_{GT}$≥5 (85% *SameSOP* agreement) than mixtures with NoC$_{GT}$≤4 (61% *SameSOP* agreement).

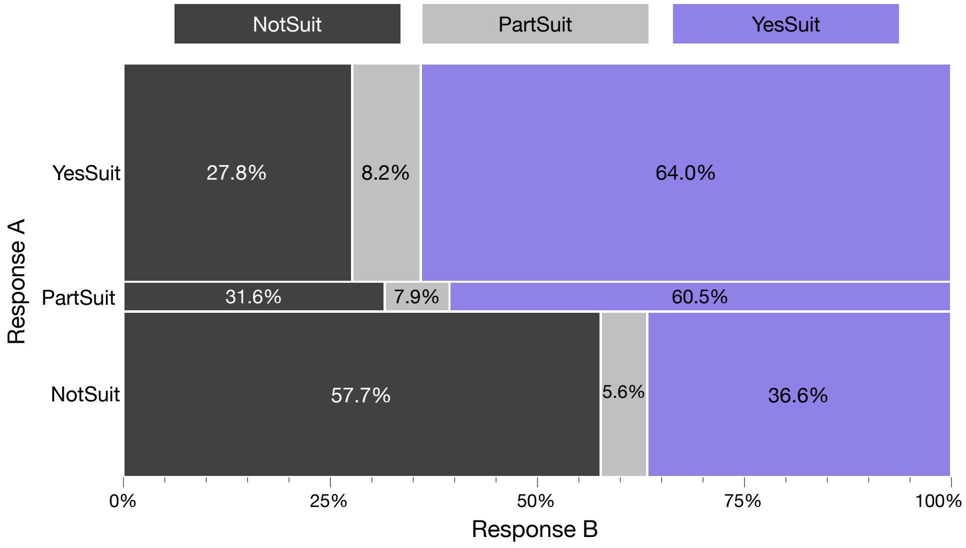

Fig S3. Reproducibility of suitability assessments illustrated as a mosaic display of the contingency table. All *SameSOP* responses are shown as rows; the x axis and color-coding show the proportions of each type of response among all other labs (*InterlabSuitSameSOP dataset*)

Appendix H Detailed NoC Results

Table S20 shows the distribution of NoC responses by type. (See *Appendix B2* for a summary of participant instructions with respect to reporting NoC.)

	Responses						*Weighted Responses*			
NoC type	*All*		*SameSOP*		*DiffSOP*		*All*		*SameSOP*	
Exact NoC value	1,069	47.1%	920	49.0%	149	37.9%	573.1	46.9%	489.2	51.1%
NoC range	47	2.1%	27	1.4%	20	5.1%	37.7	3.1%	21.8	2.3%
NoC minimum	304	13.4%	175	9.3%	129	32.8%	208.2	17.0%	102.5	10.7%
No NoC: too complex	3	0.1%	1	0.1%	2	0.5%	2.5	0.2%	0.5	0.1%
No NoC: not suitable	849	37.4%	756	40.2%	93	23.7%	400.5	32.8%	344.1	35.9%
	2,272		1,879		393		1,222.0		958.0	

Table S20. Distribution of NoC responses by type.

Appendix H1 NoC Accuracy details: DiffSOP vs SameSOP

NoC$_{\text{EST}}$ accuracy was very similar for *SameSOP* and *DiffSOP*. Table S21 shows the accuracy of NoC$_{\text{EST}}$ by *SameSOP* vs *DiffSOP*. Note that participants that indicated equivalent vs exact in *Mixture Configuration Selection* were very similar with respect to accuracy, and therefore they are not differentiated in analyses outside of this table. *DiffSOP* overall accuracy was somewhat higher than *SameSOP* accuracy, but that is explained by a higher proportion of NoC range responses, which have lower incorrect NoC$_{\text{EST}}$ rates. (For exact NoC$_{\text{EST}}$, % correct was 75% for *SameSOP*, 73% for *DiffSOP*; for NoC$_{\text{EST}}$ range, % correct was 92% for *SameSOP*, 93% for *DiffSOP*.)

NoC SameDiff SOP	*Mix Config SOP selection*	*% Correct-Exact*	*% Correct-Included*	*% Incorrect*	*NoC responses (raw)*	*NoC responses (weighted)*	*Correct-Exact (weighted)*	*Correct-Included (weighted)*	*Incorrect ±1 (weighted)*	*Incorrect ±2-3 (weighted)*
SameSOP	A_EXACT	60%	18%	22%	1,486	429.6	259.6	77.0	91.8	1.2
	B_EQUIV	59%	21%	20%	393	180.1	106.5	37.3	36.2	
DiffSOP	A_EXACT	19%	53%	28%	59	21.5	4.0	11.4	6.1	
	B_EQUIV				5	3.5	1.1	2.3		
	C_DIFFBOTH	34%	52%	14%	311	170.3	57.5	88.3	24.5	
	D_DIFFNOC				18	14.0	1.0	11.0	2.0	
Subtotal SameSOP		60%	19%	21%	1,879	609.7	366.1	114.3	128.0	1.2
Subtotal DiffSOP		30%	54%	16%	393	209.3	63.6	113.0	32.6	
Total		52%	28%	20%	2,272	819.0	429.7	227.3	160.6	1.2

Table S21. Accuracy of *SameSOP* vs *DiffSOP* NoC$_{\text{EST}}$ results. Grayed rows indicate low counts (for which percentages are not shown).

The following figures are variations of Figure 4 (in the main paper). Fig S4 shows *DiffSOP* responses in addition to *SameSOP* responses. Fig S5 shows combined suitability and NoC responses.

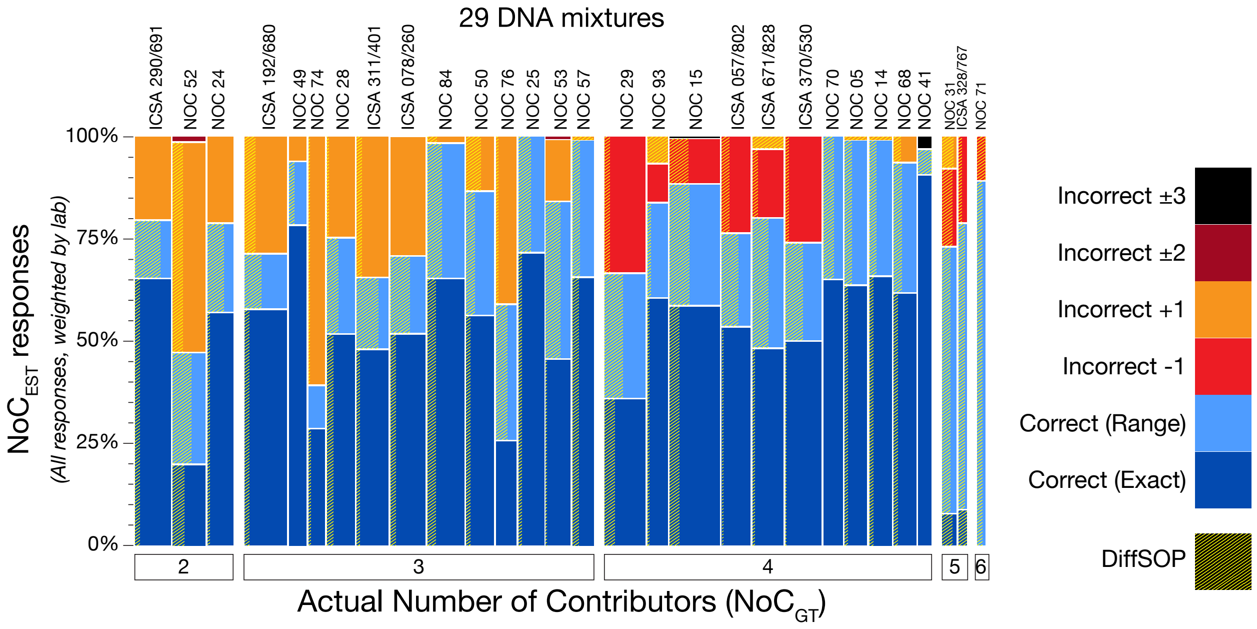

Fig S4. Accuracy of all NoC responses by mixture. Same as Figure 4 (main paper) but including all responses; hashing indicates *DiffSOP* responses. (*WeightedResponse* dataset)

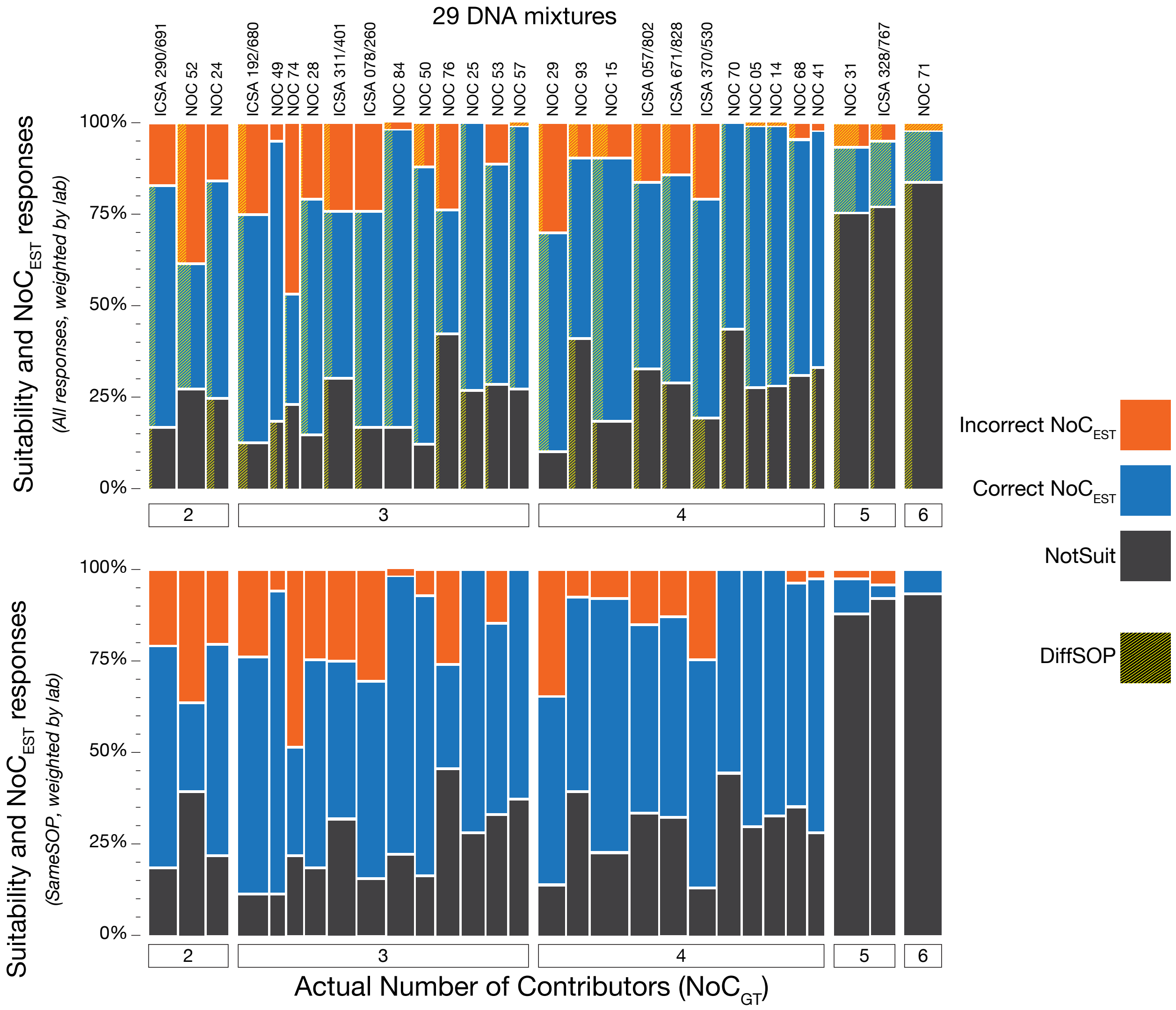

Fig S5. Accuracy of NoC responses combined with suitability responses: (top) all responses, with hashing indicating *DiffSOP* responses; (bottom) *SameSOP* responses. (Top: *WeightedResponse* dataset; bottom: *WeightedNoCSameSOP* dataset)

Appendix H2 NoC$_{EST}$ accuracy for PartSuit Trials

The NoC$_{EST}$ accuracy for trials assessed as *PartSuit* was almost identical to that of trials assessed as *YesSuit*. For trials with *SameSOP* NoC$_{EST}$ (*WeightedNoCSameSOP dataset*),

- *YesSuit*: 78.7% correct
- *PartSuit* for a subset of contributors: 78.7% correct
- *PartSuit* for a subset of loci: 80.0% correct
- (Only 5 trials were *PartSuit* for both a subset of contributors and loci; too few to report a rate)

Appendix H3 NoC_{EST} Accuracy and Reproducibility

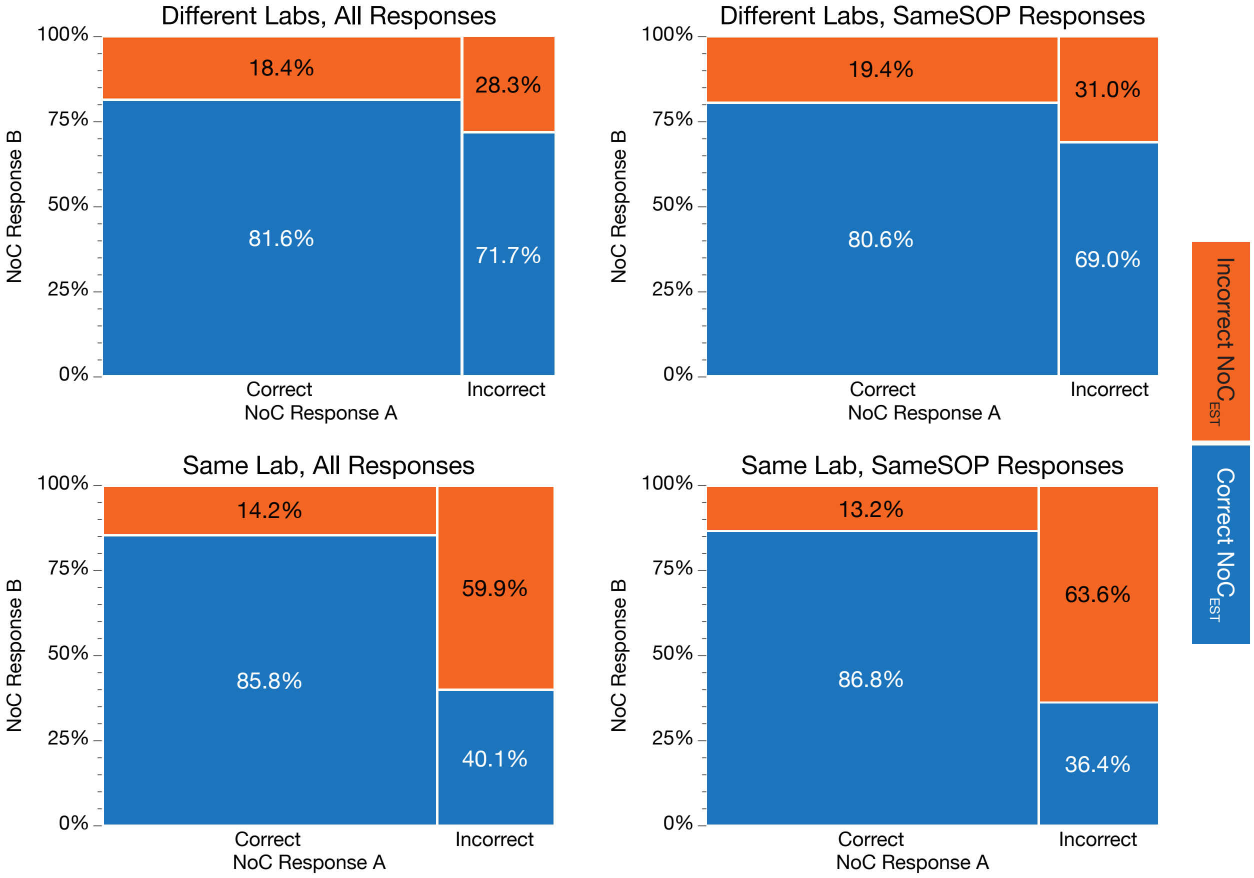

Fig S6. Mosaic displays of contingency tables for reproducibility of NoC correct and incorrect responses. (Different labs: *Interlab* and *InterlabNoCSameSOP datasets*. Same lab: *Intralab* and *IntralabSameSOP datasets*. All results weighted by lab.)

Appendix I Variability by Amp/CE versions of each mixture

Table S22 through Table S25 provide additional details regarding variation in the Amp/CE versions of each mixture (summarized in Table 2 in the main paper). All Amp/CE versions of a given mixture were created from the same physical mixture, and efforts were made to make the AmpCE variations as consistent as practical, these tables show that the Amp/CE versions of each mixture are not identical. This is to be expected: 1) any two amplifications of a sample will be different to some extent due to stochasticity, 2) differences in cycles can result in notable differences in the signal strength in the EPG, and 3) different amplification kits do not include the same loci and therefore contain different information. Note that this source of variability is not limited to this study: different amplifications of a single physical mixture will vary in casework and cannot be expected to be identical.

Notes regarding the content of Table S22 through Table S25:

- Mixture ratios and proportion of the mixture for the smallest contributor are based on signal strength as determined by STRmix[*] as the average across all alleles for each mixture profile, measured in RFUs. For example, for NOC_52 ID28 (Table S22), the two contributors have an average signal strength across all alleles of 416 RFUs and 293 RFUs, resulting in a ratio of 1.4 : 1, or 41% for the smaller contributor.
- The "NoC over mean threshold" columns are based on the same signal strength data that was used in calculating mixture ratios, but here we show the number of contributors for which the average signal strength was greater than [100,150,200] RFUs. For example, for NOC_24 GF28 (Table S22), the signal strength of the two contributors was 150 RFUs and 128 RFUs, so two contributors are over a threshold of 100 RFU, one meets a threshold of 150 RFU, and neither is over 200 RFU. Since this is an average across all alleles, this does not compare precisely to use of an analytical threshold (AT), but does provide an indication of the amount of information available at different RFU thresholds. For all mixtures and Amp/CE versions, all contributors had an average signal strength greater than 50 RFU. See Table S26 for a summary of these values. Caution should be used in comparing counts across Amp/CE at a given threshold, because different Amp/CE settings may imply different thresholds. For example, the tables show that 6C29 has fewer effective contributors at 150 or 200 RFU than the other Amp/CE mixtures — but the tables also show that 6C29 had the lowest incorrect NOC_{EST} rate.
- "No NoC_{EST}" is overwhelmingly NotSuit: of the responses without NoC_{EST}, there were 344.0 weighted *NotSuit* responses vs 0.5 responses of "too complex" (*WeightedNoCSameSOP dataset*).

Despite the variation by Amp/CE in mixture ratios, proportions of smallest contributors, and NoC by threshold, we found no significant association between these values and rates of incorrect NoC_{EST} or *NotSuit*.

[*] *The STRmix analysis of the ID28 data was done on STRmix 2.9 (using an ID+ model calibration from ESR) while the other three Amp/CE systems were done using STRmix 2.5.11, because Bode's installation of 2.5.11 did not have a calibration for ID+.*

Mixture	*Amp/CE*	*Mix Ratio*	*Smallest contrib.*	*NoC over mean threshold*			*No NoC$_{EST}$*	*Correct NoC$_{EST}$*	*Incorrect NoC$_{EST}$*
				100 rfu	*150 rfu*	*200 rfu*			
ICSA_290/691	(Overall)						19%	60%	21%
	ID28	1.9 : 1	35%	2			67%	0%	33%
	GF28	2.4 : 1	30%	2			34%	41%	25%
	GF29	2.5 : 1	28%	2			3%	74%	23%
	6C29	2.1 : 1	32%	2		1	18%	65%	17%
NOC_52 *	(Overall)						40%	24%	36%
	ID28	1.4 : 1	41%	2			67%	33%	0%
	GF28	1.1 : 1	47%	2			29%	36%	35%
	GF29	1.4 : 1	42%	2			21%	0%	79%
	6C29	1.2 : 1	46%	2	0		54%	36%	11%
NOC_24 *	(Overall)						22%	58%	20%
	ID28	2.6 : 1	27%	2		1	67%	33%	0%
	GF28	1.2 : 1	46%	2	1	0	23%	68%	8%
	GF29	2.1 : 1	33%	2			8%	50%	42%
	6C29	2.5 : 1	28%	1		0	28%	72%	0%

Table S22. Details of two-contributor mixtures by Amp/CE, shown with suitability and NoC response rates. (*):Packets reviewed in detail in Appendix I1. Yellow highlight indicates one contributor does not meet the given threshold; orange highlight indicates two or more contributors do not meet the threshold. (Rates based on *WeightedNoCSameSOP dataset*)

Mixture	Amp/CE	Mix Ratio	Smallest contrib.	NoC over mean threshold			No NoC_{EST}	Correct NoC_{EST}	Incorrect NoC_{EST}
				100 rfu	150 rfu	200 rfu			
ICSA_192/680 *	(Overall)						12%	65%	23%
	ID28	11.8 : 1.6 : 1	7%	3	2		17%	50%	33%
	GF28	14.4 : 1.5 : 1	6%	3			31%	34%	34%
	GF29	15.2 : 2.1 : 1	5%	3			4%	93%	3%
	6C29	18.2 : 2.1 : 1	5%	3		2	12%	57%	32%
NOC_49	(Overall)						12%	82%	6%
	GF29	1.7 : 1.5 : 1	24%	3			12%	81%	8%
	6C29	1.5 : 1.2 : 1	27%	3			13%	85%	3%
NOC_74 *	(Overall)						22%	30%	48%
	GF29	1.4 : 1.3 : 1	27%	3			23%	19%	58%
	6C29	1.2 : 1.0 : 1	31%	3			21%	45%	34%
NOC_28	(Overall)						19%	57%	24%
	GF29	1.4 : 1.2 : 1	28%	3			19%	40%	40%
	6C29	1.5 : 1.1 : 1	28%	3			19%	69%	13%
ICSA_311/401 *	(Overall)						32%	43%	25%
	ID28	1.5 : 1.3 : 1	27%	3			67%	0%	33%
	GF28	1.7 : 1.3 : 1	25%	3			85%	15%	0%
	GF29	1.2 : 1.0 : 1	31%	3			15%	39%	46%
	6C29	1.1 : 1.1 : 1	31%	3			30%	57%	13%
ICSA_078/260 *	(Overall)						16%	54%	30%
	ID28	1.7 : 1.3 : 1	25%	3			67%	33%	0%
	GF28	1.2 : 1.1 : 1	30%	3			35%	40%	25%
	GF29	1.5 : 1.1 : 1	27%	3			3%	44%	53%
	6C29	1.2 : 1.1 : 1	30%	3			12%	68%	21%
NOC_84	(Overall)						23%	76%	1%
	ID28	1.8 : 1.2 : 1	25%	3			67%	33%	0%
	GF28	1.5 : 1.1 : 1	28%	3			36%	64%	0%
	GF29	1.7 : 1.4 : 1	24%	3			9%	86%	5%
	6C29	1.4 : 1.1 : 1	28%	3			19%	81%	0%
NOC_50	(Overall)						17%	76%	7%
	GF29	1.3 : 1.2 : 1	28%	3			12%	76%	12%
	6C29	2.8 : 2.0 : 1	17%	3	2		20%	77%	3%
NOC_76 *	(Overall)						46%	28%	26%
	ID28	1.1 : 1.0 : 1	32%	3			100%	0%	0%
	GF28	1.3 : 1.0 : 1	30%	3			57%	26%	17%
	GF29	1.1 : 1.0 : 1	31%	3			39%	23%	39%
	6C29	1.3 : 1.1 : 1	29%	3		2	32%	47%	21%
NOC_25	(Overall)						29%	71%	0%
	ID28	1.8 : 1.3 : 1	25%	3			67%	33%	0%
	GF28	1.5 : 1.1 : 1	27%	3		1	43%	57%	0%
	GF29	2.1 : 1.9 : 1	20%	3		2	12%	89%	0%
	6C29	1.4 : 1.1 : 1	29%	3		1	31%	69%	0%
NOC_53	(Overall)						34%	52%	14%
	ID28	1.5 : 1.1 : 1	27%	3			67%	33%	0%
	GF28	1.9 : 1.4 : 1	24%	3		2	43%	39%	19%
	GF29	1.5 : 1.0 : 1	29%	3			23%	54%	23%
	6C29	1.1 : 1.1 : 1	31%	3		2	31%	69%	0%
NOC_57	(Overall)						38%	62%	0%
	GF29	1.4 : 1.1 : 1	29%	3			33%	67%	0%
	6C29	1.9 : 1.2 : 1	24%	2	1	0	41%	59%	0%

Table S23. Details of three-contributor mixtures by Amp/CE, shown with suitability and NoC response rates. (*):Packets reviewed in detail in Appendix I1. Yellow highlight indicates one contributor does not meet the given threshold; orange highlight indicates two or more contributors do not meet the threshold. (Rates based on *WeightedNoCSameSOP dataset*)

Mixture	*Amp/CE*	*Mix Ratio*	*Smallest contrib.*	*NoC over mean threshold*			*No*	*Correct*	*Incorrect*
				100 rfu	*150 rfu*	*200 rfu*	*NoC_{EST}*	*NoC_{EST}*	*NoC_{EST}*
NOC_29 *	(Overall)						14%	52%	34%
	ID28	1.5 : 1.4 : 1.3 : 1	19%	4			33%	5%	62%
	GF28	1.5 : 1.5 : 1.5 : 1	18%	4			2%	44%	55%
	GF29	1.7 : 1.6 : 1.5 : 1	17%	4			8%	44%	47%
	6C29	1.4 : 1.4 : 1.3 : 1	20%	4			20%	70%	10%
NOC_93	(Overall)						40%	53%	7%
	ID28	2.1 : 1.1 : 1.0 : 1	19%	4			33%	48%	19%
	GF28	2.2 : 1.4 : 1.4 : 1	17%	4			78%	22%	0%
	GF29	3.0 : 1.7 : 1.1 : 1	15%	4			19%	73%	8%
	6C29	2.3 : 1.6 : 1.1 : 1	17%	4			47%	46%	7%
NOC_15	(Overall)						23%	69%	8%
	ID28	1.7 : 1.5 : 1.0 : 1	19%	4			33%	52%	14%
	GF28	1.7 : 1.3 : 1.2 : 1	20%	4			43%	52%	5%
	GF29	1.5 : 1.1 : 1.1 : 1	22%	4			31%	68%	1%
	6C29	1.2 : 1.2 : 1.1 : 1	22%	4			10%	78%	13%
ICSA_057/802	(Overall)						34%	51%	15%
	ID28	18.3 : 6.5 : 4.6 : 1	3%	4	3		67%	26%	7%
	GF28	11.8 : 6.5 : 4.5 : 1	4%	4		3	69%	22%	10%
	GF29	20.1 : 8.3 : 4.9 : 1	3%	4		3	13%	74%	13%
	6C29	30.2 : 16.6 : 12 : 1	2%	3			34%	47%	18%
ICSA_671/828	(Overall)						33%	54%	13%
	ID28	16.6 : 13.8 : 1.4 : 1	3%	3	2		17%	38%	45%
	GF28	17.2 : 17.1 : 1.2 : 1	3%	2			20%	80%	0%
	GF29	16.7 : 14.1 : 1.3 : 1	3%	4		3	12%	88%	0%
	6C29	16.7 : 14.1 : 1.3 : 1	3%	4		3	51%	31%	18%
ICSA_370/530	(Overall)						14%	62%	24%
	ID28	12.6 : 9.6 : 6.2 : 1	3%	4		3	33%	36%	31%
	GF28	19.7 : 13.8 : 10.9 : 1	2%	3			4%	56%	41%
	GF29	13.3 : 8.8 : 7 : 1	3%	4			11%	58%	30%
	6C29	8.2 : 6.1 : 4 : 1	5%	4		3	14%	70%	16%
NOC_70	(Overall)						45%	55%	0%
	ID28	1.2 : 1.1 : 1.1 : 1	22%	4			59%	41%	0%
	GF28	1.3 : 1.1 : 1.0 : 1	22%	4			69%	31%	0%
	GF29	1.5 : 1.2 : 1.1 : 1	21%	4			39%	62%	0%
	6C29	1.4 : 1.2 : 1.1 : 1	21%	4			33%	67%	0%
NOC_05	(Overall)						30%	70%	0%
	GF29	2.4 : 2.2 : 1.9 : 1	13%	4			15%	85%	0%
	6C29	2.5 : 1.9 : 1.2 : 1	15%	4			41%	59%	0%
NOC_14	(Overall)						33%	67%	0%
	GF29	1.8 : 1.6 : 1.5 : 1	17%	4			28%	72%	0%
	6C29	1.6 : 1.6 : 1.6 : 1	17%	4		3	38%	63%	0%
NOC_68	(Overall)						36%	61%	4%
	GF29	1.4 : 1.3 : 1.2 : 1	21%	4			33%	58%	9%
	6C29	1.4 : 1.1 : 1.0 : 1	22%	4			38%	63%	0%
NOC_41	(Overall)						29%	69%	2%
	GF29	1.7 : 1.6 : 1.6 : 1	17%	4			27%	69%	4%
	6C29	2.0 : 1.8 : 1.7 : 1	15%	4		3	31%	69%	0%

Table S24. Details of four-contributor mixtures by Amp/CE, shown with suitability and NoC response rates. (*):Packets reviewed in detail in Appendix I1. Yellow highlight indicates one contributor does not meet the given threshold. (Rates based on *WeightedNoCSameSOP dataset*)

NoC$_{GT}$	*Mixture*	*Amp/CE*	*Mix Ratio*	*Smallest contrib.*	*NoC over mean threshold*			*No NoC$_{EST}$*	*Correct NoC$_{EST}$*	*Incorrect NoC$_{EST}$*
					100 rfu	*150 rfu*	*200 rfu*			
5	NOC_31	(Overall)						88%	9%	2%
		ID28	2.1 : 1.3 : 1.3 : 1.2 : 1	15%	5			71%	0%	29%
		GF28	1.8 : 1.1 : 1.1 : 1.0 : 1	17%	5			83%	17%	0%
		GF29	1.6 : 1.5 : 1.3 : 1.1 : 1	15%	5			92%	6%	2%
		6C29	1.6 : 1.2 : 1.1 : 1.1 : 1	17%	5			89%	11%	0%
5	ICSA_328/767	(Overall)						92%	4%	4%
		ID28	2.7 : 2.5 : 1.5 : 1.4 : 1	11%	5			67%	0%	33%
		GF28	2.4 : 1.9 : 1.6 : 1.0 : 1	13%	5		4	100%	0%	0%
		GF29	2.0 : 1.7 : 1.4 : 1.0 : 1	14%	5			95%	3%	3%
		6C29	2.5 : 1.8 : 1.6 : 1.2 : 1	12%	5		4	94%	6%	0%
6	NOC_71	(Overall)						94%	6%	0%
		ID28	3.2 : 3.1 : 2.4 : 2.2 : 1.5 : 1	7%	6			100%	0%	0%
		GF28	3.0 : 2.7 : 2.1 : 1.7 : 1.3 : 1	8%	6			81%	19%	0%
		GF29	2.7 : 2.5 : 2.2 : 1.7 : 1.1 : 1	9%	6			95%	6%	0%
		6C29	3.0 : 3.0 : 2.8 : 2.2 : 1.9 : 1	7%	6			96%	5%	0%

Table S25. Details of five and six-contributor mixtures by Amp/CE, shown with suitability and NoC response rates. Yellow highlight indicates one contributor does not meet the given threshold. (Rates based on *WeightedNoCSameSOP dataset*)

Table S26 provides a summary of the "NoC over mean threshold" columns in Table S22-Table S25. Caution should be used in comparing counts across Amp/CE at a given threshold, because different thresholds may be appropriate for different Amp/CE settings. Overall across all of the mixtures, 98% of contributors had an average signal strength of 100 RFU or higher (ranged by Amp/CE from 96-100%); 88% of contributors had an average signal strength of 200 RFU or higher (ranged by Amp/CE from 82-97%). However, these values were not significantly associated with *NotSuit* or incorrect NOC$_{EST}$ rates. For example, Table S26 shows that 6C29 had fewer effective contributors at 150 or 200 RFU than the other Amp/CE mixtures — but Table S22-Table S25 show that 6C29 had the lowest incorrect NOC$_{EST}$ rate.

RFU threshold	*Contributors over RFU threshold (average across all alleles)*									
	Overall		*ID28*		*GF28*		*GF29*		*6C29*	
0	275		71		71		102		102	
100	269	98%	70	99%	68	96%	102	100%	99	97%
150	264	96%	67	94%	67	94%	102	100%	95	93%
200	243	88%	65	92%	62	87%	99	97%	82	80%

Table S26. Proportion of contributors over various thresholds, summarizing the "NoC over mean threshold" columns in Table S22-Table S25.

Appendix I1 Detailed review of flagged mixtures

Of the 29 mixtures in the study, there were nine mixtures that we flagged for detailed review due to unusually high overall incorrect NoC$_{EST}$ rates, or notably different NoC$_{EST}$ rates between Amp/CE versions of a given mixture. The determination of NoC can be ambiguous for some mixtures and some level of variation is expected for different amplification systems. The nine mixtures with differing results were evaluated to determine if the details of the EPGs revealed information that could explain the responses.

Note in the following reviews that the number of drop-out alleles observed depend on the limit of detection and AT. The observations in the sections below were based on an AT of 100 RFU for 6C29, and 125 RFU for the others.

Appendix I1a Review of NOC_52

Flagged for review due to high incorrect NoC (overall, especially for GF29). Note also a high overall *NotSuit* rate for a 2P mixture.

- Mixture Details:
  - 2P mixture, close to equal contributions.
  - Low level of DNA - 0.052 ng.

- Relatively low levels of allele sharing.

- Mixture Observations:
  - D8S1179 has an elevated / stacked stutter ~137 RFU in GF29 only. Note that in Table S32 (*Appendix J5*) that of the responses to NOC_52 that indicated D8S1179 was a primary basis for their NoC assessment, 73% were incorrect.
  - Heterozygous balance <60% for some genotypes in loci with 4 alleles.
  - Some loci have drop-out.

Appendix I1b Review of NOC_24

Flagged for review due to high incorrect NoC for GF29.

- Mixture Details:
  - 2P mixture, ~2:1 ratio between contributors.
  - Low level of DNA - 0.043 ng.
  - Relatively high levels of allele sharing.
- Mixture Observations:
  - FGA has a trace peak in N+1 stutter position in GF29 only.
  - Heterozygous balance <60% for some genotypes in loci with 4 alleles.
  - Some loci have drop-out.

Appendix I1c Review of ICSA_192/680

Flagged for review due to high incorrect NoC for all Amp/CE versions except for GF29.

- Mixture Details:
  - 3P mixture, 1 major, 2 minor contributors ~14:2:1 ratio.
  - 0.34 ng of DNA, minor contributors are quite low.
  - Major contributor degraded.
- Mixture Observations:
  - 1 elevated/stacked stutter at SE33 for GF28 only.
  - Heterozygous balance <60% for some genotypes in loci with 6 alleles.
  - Some loci have drop-out.

Appendix I1d Review of NOC_74

Flagged for review due to high incorrect NoC overall, and disparate incorrect NoC rates for 6C29 and GF29.

- Mixture Details:
  - 3P mixture, close to equal contributions.
  - Low level of DNA - 0.186 ng.
  - Relatively low allele sharing.
- Mixture Observations:
  - Heterozygous balance <60% for some genotypes in loci with 6 alleles.
  - No allelic drop-out observed.

Appendix I1e Review of NOC_28

Flagged for review due to disparate incorrect NoC rates for 6C29 and GF29.

- Mixture Details:
  - 3P mixture, close to equal contributions.
  - Low level of DNA - 0.180 ng of DNA.
  - Higher levels of allele sharing.

- Mixture Observations:
  - D21S11 has an elevated / stacked stutter ~137 RFU in GF29 only.
  - FGA has an elevated / stacked stutter ~185 RFU in GF29 only.
  - D16S & D22S have alleles visible but <125 RFU for GF29.
  - D18S has allele 15 visible but <100 RFU for 6C29.
  - The 2 elevated/stacked stutters are the likely contributing factors in the GF29 profile being designated incorrectly: those 2 additional peaks (in GF29) could be interpreted as an additional low-level contributor.
  - In this mixture the Penta Loci (available in 6C29 but not GF29) may have been a help in designating as a 3P mixture.

Appendix I1f Review of ICSA_311/401

Flagged for review due to high incorrect NoC for GF29. Note also high *NotSuit* rates for GF28 and ID28.

- Mixture Details:
  - 3P mixture, close to equal contributions.
  - Low level of DNA - 0.179 ng.
  - Relatively low allele sharing.
- Mixture Observations:
  - Heterozygous balance <60% for some genotypes in loci with 6 alleles.
  - GF28 has 2 drop-outs, GF29 has none, 6C has 1, ID28 has none

Appendix I1g Review of ICSA_078/260

Flagged for review due to high incorrect NoC for GF29.

- Mixture Details:
  - 3P mix with a victim (in mix) and an Expected contributor (not in mix).
  - Lower levels of DNA (0.174 ng), similar contribution levels.
  - Typical levels of allele sharing.
- Mixture Observations:
  - 0-1 drop-outs in profiles.

Appendix I1h Review of NOC_76

Flagged for review due to high incorrect NoC overall, and high incorrect NoC for GF29. Note also a high overall *NotSuit* rate for a 3P mixture.

- Mixture Details:
  - 3P mixture with similar contribution levels.
  - Low level of DNA 0.12 ng.
  - Relatively low levels of allele sharing
- Mixture Observations:
  - Heterozygous balance <60% for some genotypes in loci with 6 alleles.
  - GF28 has 13 drop-outs, GF29 has 3, 6C has 7, ID28 has 3. The number of GF28 dropouts helps explain the high *NotSuit* rate for GF28.

Appendix I1i Review of NOC_29

Flagged for review due to high overall incorrect NoC rate, much lower for 6C29 than for the other Amp/CE versions.

- Mixture Details:
  - 4P mixture with contribution ratios <2:1.

  - 0.87 ng of DNA, good balance, strong allele peaks.
  - Higher levels of allele sharing.
- Mixture Observations:
  - D12S391 has an elevated/stacked stutter ~282 RFU in GF29 only. Note that in Table S32 (Appendix J5) that of the responses to NOC_29 that indicated D12S391 was a primary basis for their NoC assessment, 67% were incorrect.
  - D19S433 has an elevated/stacked stutter at 108 RFU in 6C only. (No responses indicated that D19S433 was a primary basis for their NoC assessment for NOC_29; see Table S31 (Appendix J5).)
  - No drop-outs in profiles.

Appendix J Secondary Assessments

Appendix J1 Use of Software to Assess NoC

NoC Question 13 asked “Did you use any software tool to assist in assessing the number of contributors?” summarizes the responses received.

NoC software	*Responses*		*Weighted responses*		*# Participants*		*# Labs*	
Assessed the number of contributors manually	1188	87.7%	648.4	82.4%	117	83.6%	57	74.0%
Diagnostics from ProbGen system	127	9.4%	99.3	12.6%	17	12.1%	14	18.2%
Internally developed tool	19	1.4%	19.0	2.4%	3	2.1%	3	3.9%
BP Sentry	16	1.2%	16.0	2.0%	1	0.7%	1	1.3%
FaSTR	2	0.1%	2.0	0.3%	1	0.7%	1	1.3%
LRmix studio	2	0.1%	2.0	0.3%	1	0.7%	1	1.3%

Table S27. Software tools used in assessing NoC. 47 labs always assessed manually; 14 labs assessed manually on some responses and used PGS on some responses. (Not asked if participants responded *NotSuit*.)

NoC_{EST} accuracy was not generally associated with the method used in assessing NoC: 79% of manual NoC_{EST} responses were correct (n=498.6 weighted SameSOP responses) vs 81% of responses based on PGS diagnostics (n=77.8 weighted SameSOP responses).

Appendix J2 Analytical and Stochastic Thresholds

For each mixture, participants were asked if they used an analytical threshold (AT) with the following options (*Appendix B2h*, question #3):

- Yes, I used a single AT (Please specify:_____)
- Yes, but my ATs varied by dye channel
- No

Usage of ATs varied widely among participants (Table S28):

- 40 labs specified “single AT” for all mixtures (35 of these labs specified the same AT value for all trials; 5 of these labs specified different AT values by mixture)
- 18 labs indicated “ATs varied by dye channel” for all mixtures
- 9 labs had different AT responses depending on the mixture
- (No labs indicated no AT for all mixtures)

Participants were also asked if they used a stochastic threshold (ST), with the same options as for AT (*Appendix B2h*, question #4). Usage of STs varied widely among participants (Table S28):

- 22 labs specified “single ST” for all mixtures (2 of these labs specified different ST value by mixture)
- 2 labs indicated “STs varied by dye channel” for all mixtures
- 14 labs had different ST responses depending on the mixture
- 29 labs indicated no ST for all mixtures

Amp/CE	Analytical thresholds						Stochastic thresholds					
	Number of labs			Number of labs (SameSOP)			Number of labs			Number of labs (SameSOP)		
	Single AT	AT varies	No AT	Single AT	AT varies	No AT	Single ST	ST varies	No ST	Single ST	ST varies	No ST
ID28	2	4	2	-	3	2	-	3	4	-	3	2
GF28	7	2	1	5	1	1	6	1	2	5	-	1
GF29	19	8	2	15	6	2	13	3	19	12	3	14
6C29	17	10	2	13	9	2	13	-	18	10	-	16
Total	45	24	7	33	19	7	32	7	43	27	6	33

Table S28. Analytical threshold (AT) and stochastic threshold (ST) usage. “Number of labs” indicates labs that ever indicated the given response: see text for the number of labs that always used a given response.

Fig S7 shows the distributions of AT and ST for the participants who specified single values in their responses.

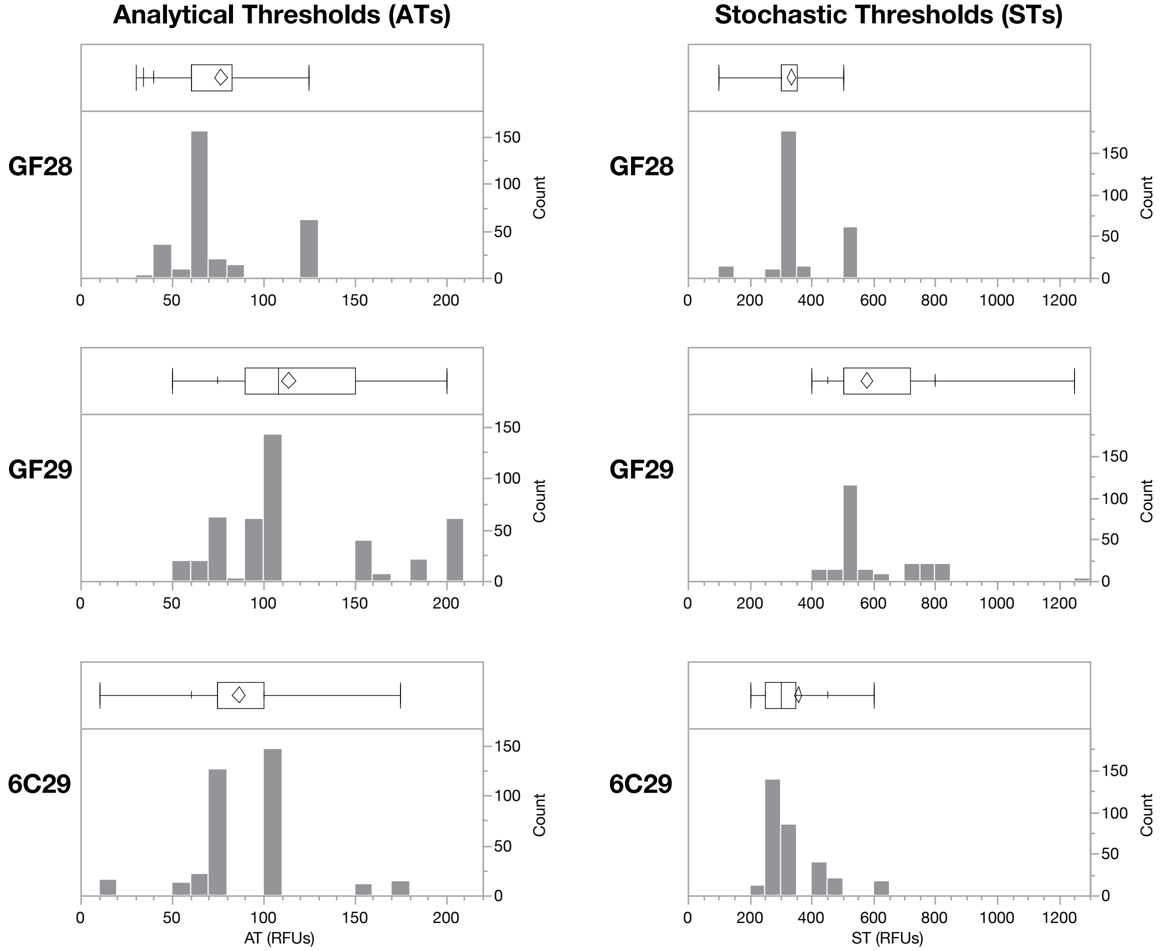

Fig S7. ATs and STs specified by participants for mixtures in the study. Limited to SameSOP. Counts are weighted to 1 response per lab per mixture. (*WeightedSuitSameSOP dataset*)

Trials indicating No ST were more likely than Single ST or Vary ST to assess YesSuit: 66.5% of weighted SameSOP No ST suitability responses were YesSuit, vs 43.1% of other responses. The (few) trials indicating No AT were almost entirely NotSuit.

We did not find any notable relationships between AT/ST and NoC accuracy, given that associations between NoC accuracy and AT/ST usage and values are confounded with Amp/CE settings.

Appendix J3 Identification of Major Contributors

NoC Question 12 asked "Are you able to identify any major contributors?"

- There are no contributors I would consider majors — 47.7% of weighted responses
- There is one major contributor — 3.5%
- There are two or more major contributors — 7.8%
- We do not differentiate between major and minor contributors — 41.1%

(Weighted responses by lab are fractional because of inconsistent responses within labs.)

Responses indicating major contributors were much more likely to make *PartSuit* responses:

- Overall, 9.5% of weighted responses were *PartSuit*
- For trials indicating one major contributor, 26.4% of weighted responses were *PartSuit*

- For trials indicating two or more major contributors, 44.9% of weighted responses were *PartSuit*

However, NoC$_{EST}$ accuracy was virtually identical regardless of the major contributor selection.

Appendix J4 Factors Reported to Affect Number of Contributors Assessment

NoC Question 11 asked "Which factors affected your assessment of number of contributors? (check all that apply; select at least one)." Table S29 summarizes the responses received. Of the NoC factors that were frequently used, the only factor with a strong association to NoC$_{EST}$ accuracy was Maximum Allele Count (MAC) per locus: trials indicating MAC per locus as a factor were 18% incorrect; trials not indicating MAC per locus as a factor were 44% incorrect. Less-used factors associated with NoC accuracy included sex determining markers (7% more accurate if selected) and total allele count in sample (11% less accurate if selected). "Other" was rarely cited but notably less accurate when cited (see list below Table S29).

NoC Factors	*% of Labs*	*% incorrect NoC$_{EST}$*		*Delta*
		if selected	*if NOT selected*	
Discriminating potential/variability of loci (or allele frequency)	15%	18%	22%	4%
Expected stutter ratios	30%	22%	21%	-1%
Information below the analytical threshold	23%	20%	22%	2%
Maximum Allele Count (MAC) per locus	91%	18%	44%	26%
Overall level of data (peak heights in relation to laboratory validated thresholds)	31%	22%	21%	-1%
Peak heights (RFU)	63%	21%	22%	2%
Peak morphology (e.g., CE resolution; unresolved microvariants; peak shouldering)	5%	21%	21%	0%
Presence of degradation	4%	21%	21%	0%
Presence of inhibition	1%	16%	21%	5%
Quantitation data	8%	20%	21%	1%
Relative peak heights (peak height ratios and possible shared/stacked alleles)	81%	22%	17%	-5%
Sex determining markers	25%	16%	23%	7%
Total allele count in sample	12%	31%	20%	-11%
Other	2%	49%	20%	-28%

Table S29. Factors selected by participants as affecting NoC$_{EST}$. Delta is the difference in incorrect NoC$_{EST}$ if the given factor is not selected vs. is selected: positive values indicate that responses selecting a factor were more accurate. (*WeightedNoCSameSOP*)

"Other" factors entered as affecting NoC$_{EST}$ assessments:

- Also used the sample type (intimate sexual assault sample) and the assumption of the victim's profile to determine NoC. (ICSA_290, correct NoCest)
- Assumed contributor (ICSA_290, correct NoCest)
- Based on our validation of STRmix, there is not enough loci supporting 3 contributors to increase the NoC to 4. (ICSA_192, correct NoCest)
- Conditioning on the intimate donor and assessing which alleles are foreign to the VICTIM (ICSA_290, correct NoCest)
- contributor proportions between loci, based on low peaks (NOC_84, correct NoCest)
- Exp donor (ICSA_057, correct NoCest)
- Expected contribution from EXP (ICSA_057, correct NoCest)
- Expected contributor profile (ICSA_057, correct NoCest)
- Genotypes from known/assumed individual (ICSA_057, correct NoCest)
- in house tac curves (NOC_53, correct NoCest)
- known contribution from expected contributor (ICSA_057, correct NoCest)
- locus specific amplification efficiency (NOC_52, correct NoCest)
- number of loci supporting 3 contributors, based on our validation this is an indication of 4 contributors (ICSA_671, correct NoCest)
- Number of loci supporting 3 contributors suggests that there are actually 4 contributors. (ICSA_370, correct NoCest)
- TAC curves (NOC_70, correct NoCest)
- TAC curves (NOC_93, correct NoCest)

- The number of loci indicating 3 contributors as well as peak height ratios at D18 and inconsistent proportions of contributors (see D19 vs D12) support that this is 4 contributors. (NOC_93, correct NoCest)
- the number of loci supporting 3 contributors either with minimum allele count or PHR indicates the mixture is actually 4 contributors (based on validation data) (NOC_29, correct NoCest)
- The number of loci supporting three contributors indicates (based on our validation) that the mixture is actually 4 contributors (ICSA_370, correct NoCest)
- The proposed genotype subsets for the minor contributor(s) at THO1. (ICSA_671, correct NoCest)
- The total number of loci supporting at least 3 contributors (NOC_15, correct NoCest)
- Use of conditioning upon EXP contributor. (ICSA_057, correct NoCest)
- evaluation of artefacts ‚Äì pull up @D18 (NOC_29, incorrect NoCest)
- Genotypes of expected contributor VIC. (ICSA_078, incorrect NoCest)
- number of loci supporting 2 contributors indicates that this mixture is likely 3 contributors, based on our validation and laboratory guidelines (NOC_52, incorrect NoCest)
- Number of loci supporting 3 contributors along with inconsistent proportions for 3 contributors support that this is 4 contributors. (ICSA_192, incorrect NoCest)
- Number of loci supporting 3 contributors indicates that this is 4 contributors. (ICSA_260, incorrect NoCest)
- Number of loci supporting two contributors in addition to the low peak heights observed support that there are actually three contributors. (NOC_24, incorrect NoCest)
- Peak height balance and using the BP Sentry AIC best fit model tool (NOC_52, incorrect NoCest)
- Peak heights for VIC (assumed in sexual assault cases) (ICSA_078, incorrect NoCest)
- Peak heights of alleles attributed to assumed contributor (17@SE33 vs 17@D22- would need additional contributor at SE33 to account for height difference). (ICSA_078, incorrect NoCest)
- Presence of pull-up in the positive control = discount 11.2 peak at D18 (NOC_29, incorrect NoCest)
- Relative peak heights at D18S51 and subthreshold peaks at FGA suggest possible mix of 4; however, peak heights may be a result of quant and subthreshold peaks may be baseline artefact. Anticipate little to no impact from a low level fourth. (NOC_93, incorrect NoCest)
- Relative peak heights at D21S11 and FGA suggest possible mix of 4; however, no evidence elsewhere, including the absence of peaks below the analytical threshold, taken together with quant. (NOC_15, incorrect NoCest)
- the number of loci supporting 3 contributors indicates there are actually 4 contributors (ICSA_311, incorrect NoCest)
- The number of loci supporting 3 contributors suggests that this is 4 contributors. (ICSA_311, incorrect NoCest)
- The number of loci supporting 3 contributors was used to help determine if it was 3 or 4 contributors. (NOC_76, incorrect NoCest)
- The number of loci supporting 3 contributors was used to help determine if it was a mixture of 3 or 4 contributors. (NOC_53, incorrect NoCest)
- the number of loci supporting at least 3 contributors (NOC_50, incorrect NoCest)
- The number of loci supporting at least 3 contributors, based on our validation data, indicates the mixture is 4 contributors (NOC_28, incorrect NoCest)
- the number of loci supporting at least three contributors (ICSA_078, incorrect NoCest)
- The number of unique alleles expected at these loci. (NOC_76, incorrect NoCest)
- We use the number of loci that support 3 contributors (e.g. 5-6 alleles) as a gauge to determine if the mixture is 3 contributors or 4. (NOC_49, incorrect NoCest)

Appendix J5 Primary Loci Used for Number of Contributors Assessments

NoC Question 10 asked "What were the PRIMARY loci used as the basis for determining the number of contributors? In other words, indicate the loci that were most informative or most helpful. (check all that apply; select at least one)." The options provided were the loci for the amplification kit selected as part of the Amp/CE settings; see *Appendix B2h* for the loci provided for each amplification kit.

Participants who provided NoC assessments selected a mean (unweighted) of 4.9 loci (median 4, range 1-24). Table S30 shows the proportions of weighted *SameSOP* trials a given locus was a primary basis for NoC_{EST}. For example, Penta E was selected as a primary locus for 25% of 6C29 responses; it is blank for the other columns indicating it is not used in those kits. Note that SE33 is the most cited locus overall, even though it is not used by all kits.

Table S30 also shows the associations between the selected loci and the incorrect NoC_{EST} rate. For example, on weighted *SameSOP* trials in which SE33 was cited as a primary locus, the incorrect NoC_{EST} rate was 17%, but on trials in which SE33 was not cited, the incorrect NoC_{EST} rate was 31%; the difference (14%) is highlighted in green. Note that SE33 stands out as both heavily used and beneficial. Of the other most-used loci, some (such as D18S51) appear to have been counterproductive. Some of the loci that were rarely used overall had a strong association with NoC accuracy (such as Amel, only used in 3% of trials but an 11% marginal improvement in accuracy); these may indicate the locus was useful for unusual or mixture-specific issues, or may be adventitious artifacts (e.g., a reflection of the labs that selected Amel rather than of Amel per se). Conversely, note some less-used loci that were inversely associated with accuracy (such as CSF1PO or TPOX): these may indicate mixture-specific situations in which that locus was misleading.

	% of Weighted Responses (SameSOP NoC)					*Incorrect NoC$_{EST}$ Rate*		*Delta*
	All	*ID28*	*GF28*	*GF29*	*6C29*	*If used*	*If NOT used*	
Amel	3%				6%	11%	22%	11%
CSF1PO	4%	5%	7%	4%	4%	31%	21%	-10%
D10S1248	7%		6%	7%	8%	15%	22%	7%
D12S391	20%		19%	23%	20%	23%	20%	-3%
D13S317	15%	8%	11%	15%	16%	25%	20%	-4%
D16S539	12%	10%	12%	14%	11%	24%	21%	-3%
D18S51	22%	23%	27%	23%	20%	25%	19%	-6%
D19S433	7%	7%	5%	11%	5%	14%	22%	9%
D1S1656	16%		13%	22%	13%	20%	22%	1%
D21S11	20%	19%	15%	21%	21%	15%	24%	9%
D22S1045	7%		3%	10%	6%	10%	23%	12%
D2S1338	17%	17%	18%	15%	20%	18%	22%	4%
D2S441	6%		5%	7%	6%	14%	22%	8%
D3S1358	11%	4%	8%	11%	12%	17%	22%	6%
D5S818	4%	4%	3%	4%	4%	23%	21%	-2%
D7S820	7%	3%	5%	8%	8%	28%	20%	-8%
D8S1179	16%	19%	13%	19%	13%	23%	21%	-3%
DYS391	4%		2%	5%	4%	14%	22%	7%
DYS570	3%				7%	6%	22%	16%
DYS576	5%				12%	8%	22%	15%
FGA	24%	13%	17%	31%	21%	23%	20%	-2%
Penta D	5%				12%	7%	22%	15%
Penta E	11%				25%	12%	23%	11%
SE33	45%		30%	53%	48%	17%	31%	14%
TH01	6%	1%	2%	6%	7%	24%	21%	-3%
TPOX	2%	3%	1%	2%	3%	28%	21%	-8%
vWA	5%				10%	12%	22%	10%
Y indel	0%		0%	0%				

Table S30. Loci used for NoC$_{EST}$. Left columns: percentage of all weighted trials for the given Amp/CE setting that indicated the specified locus was a primary basis for NoC$_{EST}$. Participants could select multiple loci, hence totals are greater than 100%. Percentages ≥30% are highlighted in red, ≥20% in yellow, <10% are grayed (highlighting is based on unrounded percentages). Right columns: incorrect NoC$_{EST}$ rates on trials citing and not citing the specified locus; the delta column is highlighted with percentages ≥10% in green, negative percentages in orange. (*WeightedNoCSameSOP dataset*)

Table S31 shows the usage of loci by mixture as a percentage of all NoCEST responses that indicated the specified locus was a primary basis for NoCEST. For example, on mixture NOC_28, 15% of responses that resulted in an NoCEST indicated Amel was a primary basis for the decision.

	Amel	CSF1PO	D10S1248	D12S391	D13S317	D16S539	D18S51	D19S433	D1S1656	D21S11	D22S1045	D2S1338	D2S441	D3S1358	D5S818	D7S820	D8S1179	DYS391	DYS570	DYS576	FGA	Penta D	Penta E	SE33	TH01	TPOX	vWA
ICSA_290/691	4%	0%	37%	65%	34%	5%	4%	3%	26%	65%	4%	40%	1%	3%	28%	25%	36%	0%	19%	19%	69%	3%	15%	76%	0%	25%	16%
NOC_52	0%	12%	5%	29%	8%	6%	56%	10%	35%	27%	11%	57%	23%	37%	13%	39%	58%	0%	0%	0%	47%	2%	15%	23%	5%	6%	15%
NOC_24	0%	32%	0%	29%	29%	36%	9%	0%	35%	28%	0%	9%	41%	72%	0%	0%	53%	9%	11%	0%	17%	0%	7%	68%	18%	10%	0%
ICSA_192/680	4%	0%	1%	33%	80%	4%	2%	5%	22%	16%	1%	0%	1%	4%	16%	1%	67%	1%	3%	0%	21%	0%	24%	66%	0%	0%	25%
NOC_49	13%	5%	0%	88%	0%	26%	87%	0%	0%	15%	94%	0%	86%	0%	0%	0%	0%	3%	13%	13%	0%	34%	3%	96%	0%	0%	0%
NOC_74	12%	18%	18%	68%	74%	15%	35%	6%	51%	32%	6%	13%	0%	6%	6%	68%	12%	6%	12%	6%	37%	13%	24%	80%	6%	0%	13%
NOC_28	15%	9%	28%	18%	70%	68%	31%	4%	0%	40%	0%	0%	7%	72%	0%	0%	19%	2%	2%	51%	34%	40%	44%	72%	0%	0%	0%
ICSA_311/401	0%	0%	1%	46%	41%	59%	84%	15%	33%	1%	1%	44%	2%	7%	0%	0%	50%	2%	6%	0%	53%	0%	50%	93%	49%	11%	6%
ICSA_078/260	5%	0%	54%	62%	56%	18%	83%	0%	4%	8%	19%	49%	3%	11%	3%	15%	7%	14%	7%	7%	17%	31%	32%	52%	3%	2%	2%
NOC_84	0%	0%	0%	4%	18%	0%	9%	0%	33%	15%	7%	93%	0%	29%	2%	52%	0%	0%	0%	34%	11%	4%	0%	84%	0%	0%	0%
NOC_50	5%	0%	16%	60%	12%	5%	5%	43%	40%	22%	7%	14%	7%	85%	7%	0%	85%	7%	0%	0%	78%	0%	0%	53%	59%	0%	9%
NOC_76	3%	33%	1%	38%	85%	29%	57%	2%	1%	24%	4%	9%	0%	0%	17%	35%	0%	0%	19%	20%	39%	16%	31%	77%	71%	18%	0%
NOC_25	11%	1%	21%	22%	5%	0%	41%	7%	56%	34%	33%	31%	5%	23%	5%	5%	41%	39%	0%	0%	46%	5%	12%	89%	5%	7%	12%
NOC_53	5%	33%	5%	74%	5%	6%	28%	30%	76%	71%	5%	62%	6%	6%	5%	8%	6%	0%	5%	0%	0%	20%	16%	70%	16%	5%	22%
NOC_57	0%	3%	5%	6%	12%	6%	0%	27%	11%	88%	45%	9%	0%	6%	3%	3%	39%	0%	0%	0%	77%	0%	43%	24%	0%	0%	9%
NOC_29	0%	0%	0%	46%	4%	12%	76%	0%	42%	9%	5%	0%	0%	0%	13%	1%	3%	4%	0%	0%	55%	0%	18%	88%	0%	0%	0%
NOC_93	11%	5%	5%	21%	19%	34%	77%	60%	15%	36%	37%	37%	14%	5%	12%	9%	21%	16%	16%	11%	60%	0%	0%	30%	5%	0%	12%
NOC_15	1%	6%	8%	14%	0%	65%	1%	1%	3%	69%	0%	24%	13%	20%	3%	3%	18%	0%	12%	13%	60%	3%	10%	85%	6%	0%	0%
ICSA_057/802	0%	11%	0%	1%	4%	0%	1%	43%	0%	57%	7%	9%	25%	21%	4%	0%	1%	12%	0%	0%	64%	4%	0%	30%	0%	0%	25%
ICSA_671/828	2%	4%	27%	27%	11%	16%	25%	10%	1%	22%	7%	5%	3%	15%	5%	4%	17%	10%	0%	0%	33%	0%	11%	85%	3%	1%	1%
ICSA_370/530	3%	17%	0%	10%	0%	43%	30%	5%	10%	7%	0%	59%	20%	3%	0%	9%	54%	10%	3%	0%	40%	0%	3%	82%	1%	0%	0%
NOC_70	0%	0%	9%	0%	0%	9%	22%	40%	84%	39%	0%	25%	0%	0%	0%	0%	6%	0%	0%	0%	60%	0%	9%	84%	0%	1%	0%
NOC_05	0%	0%	3%	0%	0%	0%	47%	0%	34%	3%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	21%	50%	50%	100%	0%	0%	5%
NOC_14	16%	0%	3%	33%	3%	0%	13%	0%	38%	89%	0%	95%	0%	3%	5%	23%	0%	3%	3%	27%	3%	5%	48%	21%	4%	0%	19%
NOC_68	0%	0%	0%	6%	0%	0%	98%	0%	0%	8%	0%	0%	0%	0%	0%	8%	0%	2%	0%	0%	0%	0%	6%	100%	0%	0%	0%
NOC_41	17%	10%	10%	3%	10%	3%	31%	38%	17%	10%	23%	10%	10%	10%	3%	10%	17%	30%	0%	13%	10%	7%	7%	100%	17%	10%	0%
NOC_31	0%	0%	0%	0%	0%	0%	18%	0%	0%	56%	0%	0%	0%	0%	0%	0%	0%	18%	0%	0%	4%	0%	0%	65%	0%	0%	0%
ICSA_328/767	0%	0%	40%	40%	0%	17%	80%	0%	60%	6%	0%	0%	0%	12%	0%	40%	0%	0%	0%	0%	37%	0%	40%	60%	0%	28%	0%
NOC_71	0%	4%	0%	0%	4%	0%	0%	0%	64%	32%	4%	0%	0%	0%	4%	4%	32%	0%	0%	0%	0%	0%	0%	32%	0%	0%	0%

Table S31. Loci used for NoCEST by mixture, showing the percentage of all NoCEST responses that indicated the specified locus was a primary basis for NoCEST. (Subset of *WeightedNoCSameSOP dataset* but omitting NoNOC responses)

Table S32 shows the relative proportions of NoC responses citing given loci that were incorrect. To avoid small-n effects, rates based on a denominator of less that 5 weighted responses are omitted. For example, of the NoC estimates on mixture NOC_52 that cited D8S1179 as a primary basis for the NoC assessment, 73%

were incorrect (8.8/11.9 weighted responses) — Appendix I noted that for that mixture, D8S1179 has an elevated / stacked stutter (~137 RFU) in the GF29 version of the mixture.

	Amel	*CSF1PO*	*D10S1248*	*D12S391*	*D13S317*	*D16S539*	*D18S51*	*D19S433*	*D1S1656*	*D21S11*	*D22S1045*	*D2S1338*	*D2S441*	*D3S1358*	*D5S818*	*D7S820*	*D8S1179*	*DYS391*	*DYS570*	*DYS576*	*FGA*	*Penta D*	*Penta E*	*SE33*	*TH01*	*TPOX*	*vWA*
ICSA_290/691			0%	16%	19%				13%	13%		2%			12%	18%	10%		0%	0%	9%			13%		13%	
NOC_52				61%			38%		55%	18%		51%		27%		54%	73%				52%						
NOC_24		7%		15%	37%	12%			25%	15%			5%	15%			25%							35%			
ICSA_192/680				0%	21%				0%	20%					6%		16%				24%		0%	18%			6%
NOC_49				4%			3%				7%		3%									0%		6%			
NOC_74				44%	49%		33%		39%	22%						53%					16%			60%			
NOC_28			21%		9%	16%	50%			4%				12%						13%	43%	6%	10%	21%			
ICSA_311/401				39%	16%	22%	26%		53%			32%					23%				26%		17%	34%	24%		
ICSA_078/260			22%	17%	21%	6%	40%				20%	23%									12%	11%	5%	21%			
NOC_84									5%			2%		6%		3%				0%				0%			
NOC_50				12%				11%	15%					5%			5%				6%			0%	12%		
NOC_76		54%		43%	45%		42%									47%					57%		23%	47%	41%		
NOC_25							0%		0%	0%	0%	0%		0%			0%	0%			0%			0%			
NOC_53		18%		25%			21%	17%	21%	12%		8%												22%			
NOC_57										0%	0%						0%				0%		0%				
NOC_29				67%			52%		65%												66%		27%	36%			
NOC_93						23%	14%	17%		25%	14%	21%									13%			25%			
NOC_15				25%		15%				14%		30%		26%			30%				12%			7%			
ICSA_057/802								21%		40%			18%	65%							27%			41%			59%
ICSA_671/828			33%	0%			27%			52%											16%			11%			
ICSA_370/530		57%				56%	46%					47%	28%				37%				52%			27%			
NOC_70								0%	0%	0%											0%			0%			
NOC_05							0%		0%													0%	0%	0%			
NOC_14				0%					0%	0%		0%								0%			0%				
NOC_68							4%																	6%			
NOC_41								9%																3%			
NOC_31																											
ICSA_328/767																											
NOC_71																											

Table S32. Percent of NoC$_{EST}$ responses that cited a given locus as a primary basis that were incorrect, by mixture. Omits rates calculated on less than 5 weighted responses. Values less than 50% were associated with more correct than incorrect NoC responses. Values 60-80% are highlighted in yellow, 40-60% not highlighted, 20-40% in green, 0-20% in blue. (Subset of *WeightedNoCSameSOP dataset*, omitting NoNOC responses, and any locus:mixture combination that was not cited in at least 5 weighted SameSOP responses.)

Appendix K Glossary

AllResponse dataset	In this study, the set of all responses, unweighted (i.e., one response per participant, regardless of the number of participants per lab)
Amp/CE settings	In this study, a specific combination of settings used in preparing an electropherogram, including the amplification kit, number of amplification cycles, volume of amplification reaction, CE instrument, and injection time/voltage. In the *NoC Subtest* and *ICSA Subtest*, the electropherograms will be created using the most popular combinations of Amp/CE settings (as reported during registration) in an attempt to accommodate the standard operating procedures used by participating laboratories.
Casework scenarios questionnaire	Subtest of this study intended to assess analysis procedures or decisions that may vary depending upon the case scenario, and the nature of participating laboratories' mixture casework.
Comparison	The examination of a *DNA mixture profile* with respect to a *reference profile* to assess the degree of similarity or difference. In the *ICSA Subtest*, the comparison of *DNA mixture profile* with respect to the *person of interest reference profile* results in a *comparison conclusion* and/or *statistical analysis* results.
Comparison conclusion	Categorical conclusion (e.g. *exclusion*, *inconclusive*, *inclusion*) resulting from a *comparison*, generally supported by *statistical analysis* results.
Comparison packet	In this study, data provided in the *ICSA Subtest*, containing • 1 *DNA mixture profile* • 1 *person of interest reference profile* • (for *SAK Comparison packets*): o 1 *victim reference profile* o 0 or 1 *consensual partner reference profile* • (for *Non-SAK Comparison packets*): o 0 or 1 *expected contributor reference profile* • Textual information: o *Amp/CE Settings* o Quantitation data (as measured by Quantifiler Trio during quantitation of the mixture) • Quality assurance files for each mixture mixture profile and each reference profile: o Amplification positive control o Amplification negative control o 2 allelic ladders
Complex mixture	In this study, a DNA mixture that includes three or more contributors, has low total amounts of DNA, is degraded, or includes subjects that share alleles
Consensual partner	For the purposes of this study, an individual known to have had consensual intimate contact with a *victim* of a sexual assault. *Comparison packets* that are *simulated sexual assault kits (SAKs)* include 0 or 1 *consensual partner reference profiles*.
DI	Degradation index
DiffSOP	In this study, responses from participants who indicated that the *Amp/CE Settings* differed from their laboratory's *SOPs.*
DNA mixture profile	A *profile* that contains more than one contributor.
EPG	Electropherogram
Exclusion	An analyst's *comparison conclusion*, based upon the results of *comparison* and/or *statistical analysis*, that a known individual is eliminated as a possible contributor to a *DNA mixture profile*.
Expected contributor	In this study, a known individual who is expected or assumed to be a contributor to a *DNA mixture profile*, such as the owner of an item or a member of a household. *Non-SAK comparison packets* include 0 or 1 *expected contributor reference profiles*. (*Reference profiles* in *SAK comparison packets* do not use the term "expected contributor" in order to explicitly label *victim* and *consensual partner reference profiles*.)
GAO	Government Accountability Office

.HID File	A “human ID” format used for files generated by the Applied Biosystems 3500 series genetic analyzers that has become a de facto interchange standard (replacing the earlier FSA file format). The *profiles* used in this study are electropherograms in .HID file format.
ICSA Subtest	*Interpretation*, *comparison*, and *statistical analysis* — a subtest of this study intended to assess the *categorical conclusions* and *statistical analysis* reported in response to a *comparison packet*. In the *ICSA Subtest* each participant was assigned 8 *comparison packets*.
Inclusion	An analyst’s *comparison conclusion*, based upon the results of a *comparison* and/or *statistical analysis*, that a known individual may be considered a possible contributor to a *DNA mixture profile*.
Inconclusive	An analyst’s *comparison conclusion*, based upon the results of a *comparison* and/or *statistical analysis*, that a known individual can neither be *excluded* nor *included* as a possible contributor to a *DNA mixture profile*.
Interlab dataset	In this study, the set of data to evaluate inter-lab reproducibility, in which the 1,222 weighted responses in the *WeightedResponse dataset* are paired with every response from other labs on the same mixtures, resulting in 53,554 weighted inter-lab decision pairs.
InterlabNoCSameSOP dataset	In this study, the subset of the *Interlab dataset* limited to those inter-lab decision pairs in which both NoC_{EST} assessments were *SameSOP*.
InterlabSuitSameSOP dataset	In this study, the subset of the *Interlab dataset* limited to those inter-lab decision pairs in which both suitability assessments were *SameSOP*.
Interpretation	The process of evaluating a *DNA mixture profile* for purposes including, but not limited to, defining assumptions related to the mixture profile, distinguishing between alleles and artifacts, assessing the possibility of degradation, inhibition, and stochastic effects, and determining whether the profile is suitable for comparison. This study evaluates interpretation in the *ICSA Subtest*, and some aspects of interpretation in the *NoC Subtest*.
LoD	Limit of detection
Low template	A sample with low total amounts of DNA
NIJ	National Institute of Justice
NIST	National Institute of Standards and Technology
NoC Packet	In this study, data provided in the *NoC Subtest*, containing • 1 *DNA mixture profile* • Textual information: o *Amp/CE Settings* o Quantitation data (as measured by Quantifiler Trio during quantitation of the mixture) • Quality assurance files: o Amplification positive control o Amplification negative control o 2 allelic ladders
NoC Subtest	The third phase of the DNAmix 2021 study. The *NoC Subtest* of this study was conducted to assess the variability of assessments of suitability and number of contributors.
Non-SAK Comparison packet	In this study, A *comparison packet* that is not a simulated sexual assault kit (SAK).
NotSuit	In this study, abbreviation for *Unsuitable*.
NRC	National Research Council (of the National Academy of Sciences)
Number of contributors (NoC)	Number of contributors (in a DNA mixture). In this study we distinguish between • NoC_{GT} — ground truth NoC, referring to NoC for the mixtures collected/created under controlled laboratory conditions in which the contributors are definitively known. • NoC_{EST} — estimated NoC, referring to an analyst/lab’s assessment of NoC in a *DNA mixture profile*.
PartSuit	In this study, abbreviation/group reference collectively referring to a determination that a mixture is only suitable for a subset of the contributors (e.g., majors) and/or only suitable for a subset of the loci.
PCAST	President’s Council of Advisors on Science and Technology
Person of interest (POI)	An individual whose contribution to the mixture is in question (such as an alleged perpetrator). In the *ICSA Subtest*, one *person of interest reference profile* will be provided in each *comparison packet*.

PGS	Probabilistic genotyping software
Policies and procedures (P&P) questionnaire	Subtest of this study intended to capture information pertaining to participating laboratories' standard operating procedures (SOPs) relevant to DNA mixture *interpretation*, *comparison*, and *statistical analysis*.
Profile	A DNA electropherogram provided in *.HID format*. A *DNA mixture profile* contains more than one contributor (containing known and/or unknown individuals). A *reference profile* contains one known contributor (*person of interest*, *victim*, *consensual partner*, or *expected contributor*).
Reference profile	A *profile* for a single subject.
SAK Comparison packet	In this study, a *comparison packet* that is a simulated sexual assault kit (SAK).
SameSOP	In this study, responses from participants who indicated that the *Amp/CE Settings* exactly corresponded or were equivalent to their laboratory's *SOPs.*
Single-source DNA	Sample that contains DNA from one person
SOPs	Standard operating procedures
Statistical analysis	The computation of weight of evidence for the comparison of a *reference profile* with a *DNA mixture profile*. Potential statistical analysis results include the combined probability of inclusion/exclusion (CPI/CPE), random match probability (RMP), modified random match probability (mRMP), or likelihood ratio (LR).
Subunits	Term used in this study to indicate multiple participants from a single laboratory.
Suitability	In this study, an analyst's assessment during interpretation to determine whether a *DNA mixture profile* should be considered *Suitable (YesSuit), Unsuitable (NotSuit), or PartSuit*.
Suitable / Suitable for comparison	An analyst's determination during *interpretation* that a *DNA mixture profile* is appropriate for use in *comparisons* and/or *statistical analyses*. (not *unsuitable*) In this study, abbreviated *YesSuit*.
Unsuitable / Unsuitable for comparison	An analyst's determination during *interpretation* that a *DNA mixture profile* cannot be used for *comparisons* and/or *statistical analyses* for reasons including (but not limited to) poor or limited data quality, mixture complexity, or a failure to meet laboratory quality assurance requirements. (not *suitable*) In this study, abbreviated *NotSuit*.
Victim	For the purposes of this study, the complainant in a sexual assault from whom *simulated sexual assault kit* samples are collected. In the *ICSA Subtest*, one *victim reference profile* will be provided in each *SAK comparison packet*.
WeightedResponse dataset	In this study, the set of all responses, weighted by lab so that each lab collectively has one response for each mixture.
WeightedNoCSameSOP dataset	In this study, the subset of the *WeightedResponse dataset* limited to those trials in which the NoC_{EST} was *SameSOP*.
WeightedSuitSameSOP dataset	In this study, the subset of the *WeightedResponse dataset* limited to those trials in which the suitability assessment was *SameSOP*.
YesSuit	In this study, abbreviation for *Suitable*.

Appendix L References for Appendices

Ed. Note: References for appendices will be separated out when the manuscript is finalized and the appendices are split into a separate document.