Final accepted manuscript, subsequently published as two journal articles:

- Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2016). Interexaminer variation of minutia markup on latent fingerprints. *Forensic Science International 264*, 89–99. https://doi.org/10.1016/j.forsciint.2016.03.014
- Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2016). Data on the interexaminer variation of minutia markup on latent fingerprints. *Data in Brief 8*, 158–190. https://doi.org/10.1016/j.dib.2016.04.068

Interexaminer variation of minutia markup on latent fingerprints

Bradford T. Ulery,[a] R. Austin Hicklin,[a] Maria Antonia Roberts,[b] and JoAnn Buscaglia [c]

[a] Noblis, Falls Church, Virginia, USA; [b] Latent Print Support Unit, Federal Bureau of Investigation Laboratory Division, Quantico, Virginia, USA; [c] Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, Quantico, Virginia, USA

Corresponding author:

JoAnn Buscaglia, PhD
Counterterrorism and Forensic Science Research Unit
Federal Bureau of Investigation Laboratory
2501 Investigation Parkway
Quantico, Virginia 22135 USA
(703) 632-4553
(703) 632-4557 (fax)
joann.buscaglia@ic.fbi.gov

Keywords

Biometrics; Latent fingerprint examination; Fingermark; ACE-V; Repeatability; Reproducibility

Abstract

Latent print examiners often differ in the number of minutiae they mark during analysis of a latent, and also during comparison of a latent with an exemplar. Differences in minutia counts understate interexaminer variability: examiners' markups may have similar minutia counts but differ greatly in which specific minutiae were marked. We assessed variability in minutia markup among 170 volunteer latent print examiners. Each provided detailed markup documenting their examinations of 22 latent-exemplar pairs of prints randomly assigned from a pool of 320 pairs. An average of 12 examiners marked each latent.

The primary factors associated with minutia reproducibility were clarity, which regions of the prints examiners chose to mark, and agreement on value or comparison determinations. In clear areas (where the examiner was "certain of the location, presence, and absence of all minutiae"), median reproducibility was 82%; in unclear areas, median reproducibility was 46%. Differing interpretations regarding which regions should be marked (e.g., when there is ambiguity in the continuity of a print) contributed to variability in minutia markup: especially in unclear areas, marked minutiae were often far from the nearest minutia marked by a majority of examiners. Low reproducibility was also associated with differences in value or comparison determinations. Lack of standardization in minutia markup and unfamiliarity with test procedures presumably contribute to the variability we observed. We have identified factors accounting for interexaminer variability; implementing standards for detailed markup as part of documentation and focusing future training efforts on these factors may help to facilitate transparency and reduce subjectivity in the examination process.

1 Introduction

During the latent print[1] examination process, an examiner detects and interprets the features of a latent as the basis for determining whether the latent is of sufficient value for comparison — then, in comparing the latent with an exemplar, the examiner detects and interprets the corresponding or contradictory features as the basis for a determination of identification, exclusion, or inconclusive. We have previously seen that the predominant factor explaining examiners' value determinations is the count of minutiae in the latent [5,6], and the predominant factor explaining examiners' individualization determinations is the count of corresponding minutiae [6,9]. Several studies have noted substantial interexaminer variation in minutia counts [10,11,5,6]. However, differences in minutia ***counts*** understate the variability among examiners: examiners' markups may have similar minutia counts but differ greatly in which specific minutiae were marked. Here we report the results of a large-scale study describing how the ***markup*** of minutiae varies among examiners, during both the analysis of a latent and the comparison with an exemplar.

Why does it matter if examiners mark different minutiae? The conventional wisdom has been that it doesn't matter which features examiners use for their conclusions as long as they reach the same conclusion. However, because there is substantial interexaminer variation in determinations [2,3], there is reason for scrutiny of which features examiners use. In some legal cases [12,13,14,15,16,17], different conclusions among examiners have hinged on different interpretations regarding the presence or correspondence of features. Even if differences in interpretations of features do not result in differing conclusions, differences in the interpretation or markup of features underscore the subjectivity of the latent print examination process.

Friction ridge impressions contain various types of features [18,19]. Minutiae are of special importance in latent print examination, because they are the predominant features used in comparisons [18], and because they are strongly associated with value or comparison determinations [5,6]. Although several studies have noted the variation in minutia ***counts*** among examiners, few studies have described this variation in detail. Swofford [20] noted that detection and interpretation of minutiae is subjective, and therefore prone to variation. Dror et al. [21] stated "The apparent lack of consistency may reflect the absence of objective and quantifiable measures as to what constitutes a minutia, especially with latent marks that are of varying quality. However, these differences may also reflect individual differences between the examiners (arising from variations in eyesight, training, feature selection strategy, cognitive style, threshold criteria, etc.)." Langenburg [22] noted that as a group, Dutch experts were much more homogeneous than US examiners, marking fewer and more reproducible minutiae (as well as having more reproducible determinations), which he attributes to training and operational procedures that reward marking minutiae that will be reproduced by other examiners. We previously observed [6] that examiners themselves are not consistent in their selection and interpretation of minutiae over time: some of the ***inter***examiner variability in minutiae may be due to this ***intra***examiner variability.

Previously [5,6], we found that examiners' minutia counts were strongly associated with their determinations: when one examiner individualized and another was inconclusive on the same image pair, the examiner who individualized typically marked more corresponding minutiae than the examiner who was inconclusive. We also found that variability in minutia markup was not limited to cases where examiners disagreed on determinations: substantial interexaminer variability in minutia counts was the norm across a wide range of latent prints, even among examiners who reached the same determination. Here we are attempting to understand this variability more completely.

1.1 Interpretation and documentation of minutiae

A minutia is defined as a ridge ending, bifurcation (fork), or (less frequently) dot, as shown in Figure 1. Some definitions include dots as a third type of minutia, but terminology has shifted, in part, because dots are not readily detected by Automated Fingerprint Identification Systems (AFIS).

[1] Regarding the use of terminology — "latent print" is the preferred term in North America for a friction ridge impression from an unknown source, and "print" is used to refer generically to known or unknown impressions. We recognize that outside of North America, the preferred term for an impression from an unknown source is "mark" or "trace," and that "print" is used to refer only to known impressions. We are using the North American standard terminology to maintain consistency with our previous and future papers in this series [1,2,3,4,5,6,7,8]. See Glossary, Appendix A.1.

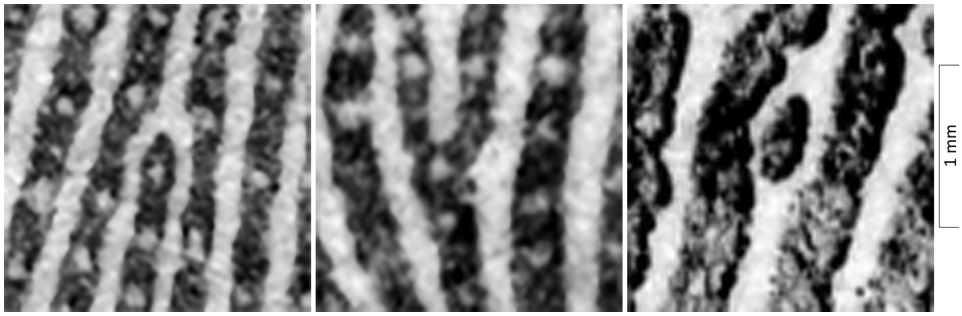


Figure 1: Examples of minutiae: (left) ridge ending, (middle) bifurcation, (right) dot. Ridges are shown in black, and valleys are shown in white.

However, not all ridge features are as readily classified as those shown in Figure 1. Disagreements among examiners may be due to actual differences in interpretation or merely to differences in how examiners document those interpretations. In this section we illustrate some of the potential causes for such disagreements: differences in interpretation may be due to ambiguous features, low clarity, or disagreements regarding the boundaries of the region of interest; differences in how examiners document minutiae may be due to human error, differences in criteria for marking minutiae, or unfamiliarity with instructions and tools.

Fingerprint examiners have not developed a standard and precise vocabulary for describing the extensive variety of friction ridge features. As a result, it can be ambiguous how to classify some features. Figure 2 shows examples where examiners might disagree on minutia markup due to the complex shapes and configurations of ridge patterns. In these instances, differences in markup may not imply actual differences in interpretation among examiners, but disagreements regarding the definition of a minutia and which features should be documented. Some features are not readily reduced to specific point locations of ridge endings and bifurcations, and one may expect examiners' minutia markup to vary in such areas (e.g., in Figure 2D, the notable "feature" is the scar).

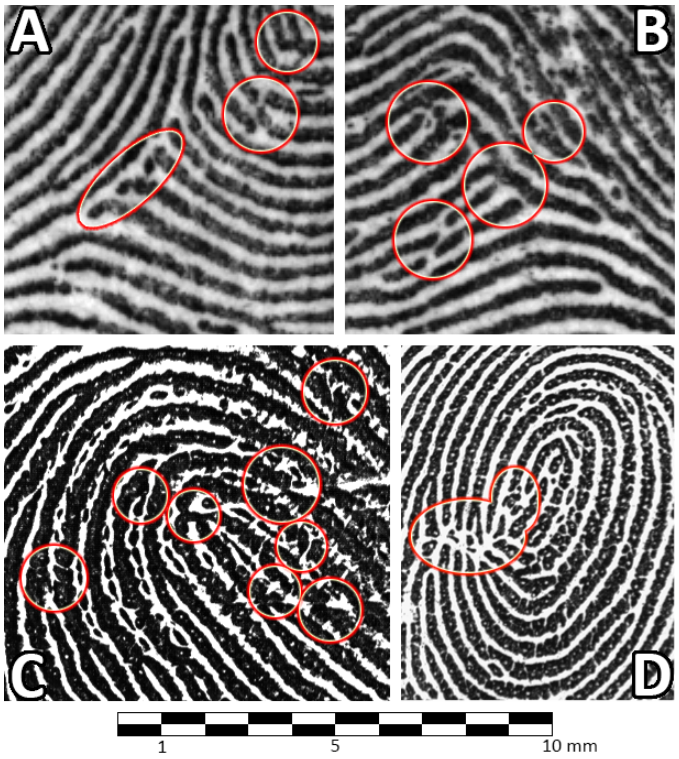


Figure 2: Examples of features that are intrinsically difficult to classify. (A) angular formations that might be described as minutiae; (B) short breaks, dots, and incipient ridges; (C) incipient ridges and irregular ridge edge details; (D) scar.

Latents are often poor quality (e.g., Figure 3), due to factors such as uncontrolled deposition (e.g., distortion, smearing, superimposed prints), substrate (surface on which the print is deposited), matrix (substance transferred to the surface), and development (physical or chemical process used to visualize the print). In practice, examiners often

differ in their assessments of whether the information in an unclear area is sufficient to determine that a minutia is present, and therefore we can expect that markup in unclear areas will be less reproducible than in clear areas. Differences in reproducibility by clarity are to be expected: examiners should generally agree on minutiae in clear areas, but may or may not agree in areas they consider unclear. Examiners also may differ in their interpretations of whether fine ridge details are persistent features, or could be artifacts in the impression.

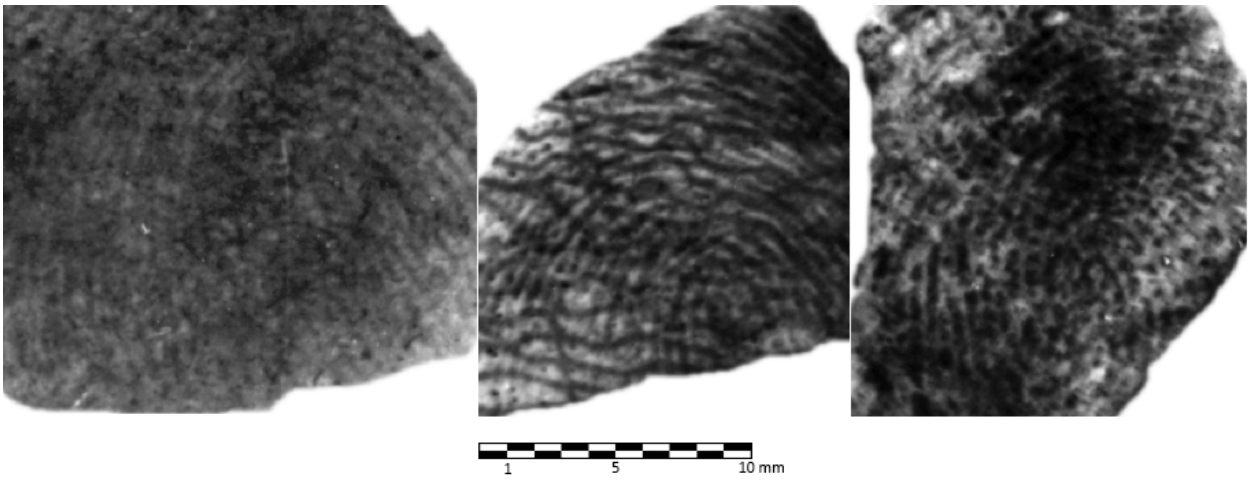


Figure 3: Low-clarity examples where the ***presence or absence*** of minutiae is ambiguous.

Even when examiners agree that a minutia is present and should be marked, clarity may affect their assessments of the exact locations and types of minutiae (e.g., Figure 4).

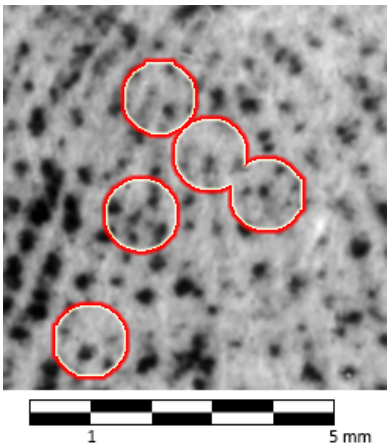


Figure 4: Ambiguous minutia ***locations and types***. Each circle indicates an area where three ridges converge to two ridges, so a minutia must be present, but cannot be located precisely, and the type (whether it is a ridge ending or bifurcation) is ambiguous.

Another source of disagreement in minutia markup stems from disagreements regarding the boundaries of the impression being considered (i.e., the region of interest). Generally, examiners are looking to compare a single contiguous impression, in which they can assess the relative positions and topological relationships of minutiae and other features. However, it is not apparent whether some images (e.g., Figure 5) contain a single impression or multiple superimposed impressions, and therefore examiners may disagree on whether specific minutiae are part of the impression of interest. Some of the disagreements regarding minutiae in the Madrid misidentification [12] were based on differing assessments of whether the image contained a single impression, a double touch (partially superimposed impressions from the same finger), or impressions from two fingers. A similar situation occurs even in clear impressions when examiners may differ in whether to consider the area below the crease (i.e., in the medial segment of the finger) as the same impression.

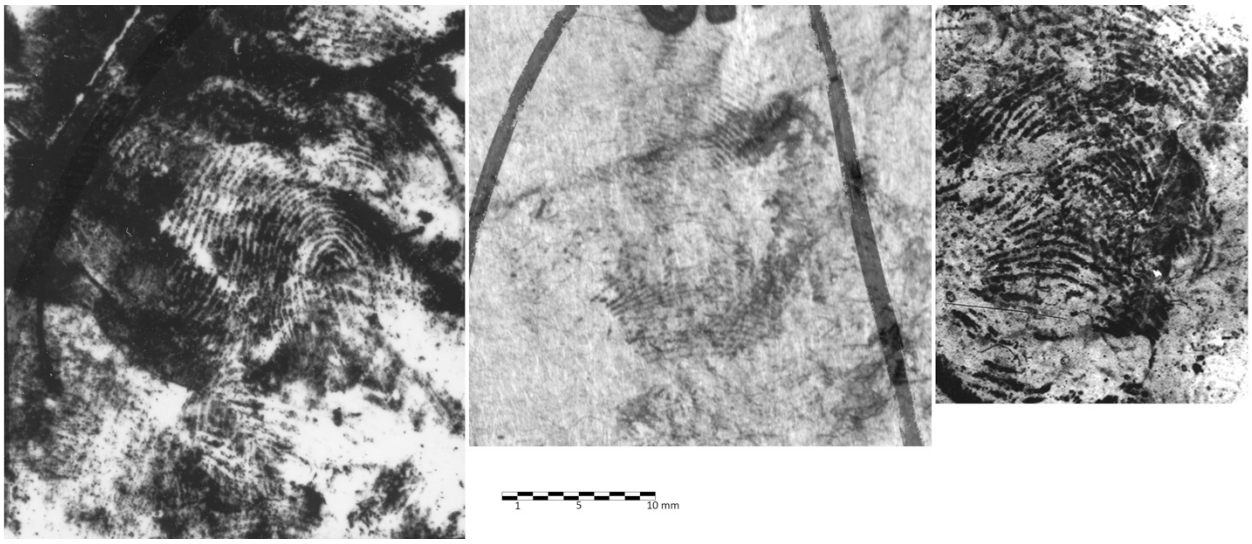


Figure 5: Examples where the ***region of interest*** is debatable because it is ambiguous which areas are from a single continuous impression. The example on the right is the latent from the Madrid misidentification [12].

Some of the variation in markup can be attributed to a lack of clear criteria specifying when and how to mark minutiae, and to a lack of standardization. While the Scientific Working Group on Friction Ridge Analysis, Study and Technology's (SWGFAST's) *Standard for the Documentation of ACE-V* [23] directs examiners to document the examination process, the details of how to document minutiae are mostly unspecified. Because documentation is not standardized in practice, it is difficult to ascertain the extent to which variation among examiners can be attributed to actual differences in interpretation, as opposed to differences in how examiners choose to document their work. Few agencies train examiners specifically on how to interpret, select, and record minutiae in a standard, reproducible manner, other than for AFIS searches, which generally require following proprietary rules. Those agencies that do require markup vary substantially on how that markup is effected, including pinpricks in physical photographs, color-coding approaches (e.g., GYRO [24], EFS [19]), software-based solutions specific to fingerprints (e.g., the FBI's Universal Latent Workstation (ULW) [25], Mideo Latentworks® [26], and PiAnOS [27]), and generic image processing software. Several authors [9,21,22,20,6,7] have stressed the need for standardization of minutia markup. In this study, we use the Extended Feature Set (EFS) format as defined in the ANSI/NIST-ITL standard [19] and supporting guidelines for examiners [28]. However, although EFS is broadly used as a non-proprietary format for searches of an AFIS, it is not yet frequently used for markup of non-AFIS casework.

The lack of standardization in definitions, training, and tools increases the likelihood of documentation errors. Human error may result in stray marks, or accidental omissions (especially in images with very large numbers of minutiae). Unfamiliarity may result in misuse of software tools or misunderstanding of instructions, especially in situations such as this study in which the examiners are using unfamiliar markup procedures and tools. Technical factors, particularly different quality computer screens or video processors, may also contribute to examiner differences.

Because of the various factors we have discussed that may result in interexaminer variation in minutia interpretation or markup, there is currently no means of establishing a "correct" minutia markup for any given latent: both in tests like this and in operational casework, we can compare examiners' markups against each other, or against a group consensus, but cannot judge whether or not they are correct in an absolute sense.

This is the third paper reporting different aspects of the "White Box" study, in which practicing latent print examiners annotated features, clarity, and correspondences in latent and exemplar fingerprints to document what they saw when performing examinations. The previous two papers presented analyses of the sufficiency of information for individualizations [6], and described changes in markup between the Analysis and Comparison phases of ACE [7]. Here we describe how the markup of minutiae varies among examiners and discuss the implications of this variation. The aim of our research is to strengthen the understanding of the latent print examination process, and provide data to assist the community in how to improve procedures, training, and standardization.

2 Materials and Methods

This paper presents analyses of data collected in the "White Box" study [6]; the test procedure, participants, and fingerprint data are summarized in Section 1 of the accompanying *Data in Brief* article [29] (abbreviated here as DiB-1).

The test procedure was designed to correspond to that part of casework in which an examiner compares a single latent to a single exemplar print (latent-exemplar image pair). The test software's workflow followed the Analysis, Comparison, Evaluation (ACE) method. In the Analysis phase, only the latent was presented, and the examiners provided the following markup: local clarity map (produced by "painting" the image using six colors denoting defined levels of clarity [19,4]); locations of features; types of features (minutia, core, delta, or "other" features such as incipient ridges, ridge edge features, or pores); and value determination (of value for individualization (VID), of value for exclusion only (VEO), or no value (NV)). If the latent print was determined to be VEO or VID, the exemplar was presented for side-by-side comparison with the latent. During this combined Comparison/Evaluation phase (henceforth "Comparison phase"), the examiner annotated the exemplar (clarity and features) and assessed its value (VID, VEO, NV); optionally revised the latent markup and value determination, further annotated the pair of images to indicate corresponding and discrepant features; reported the comparison determination (individualization, exclusion, or inconclusive); and assessed comparison difficulty (very easy, easy, moderate, difficult, very difficult). Examples of minutia and clarity markup are shown in Figure 6 and DiB-2 [TBD-29].

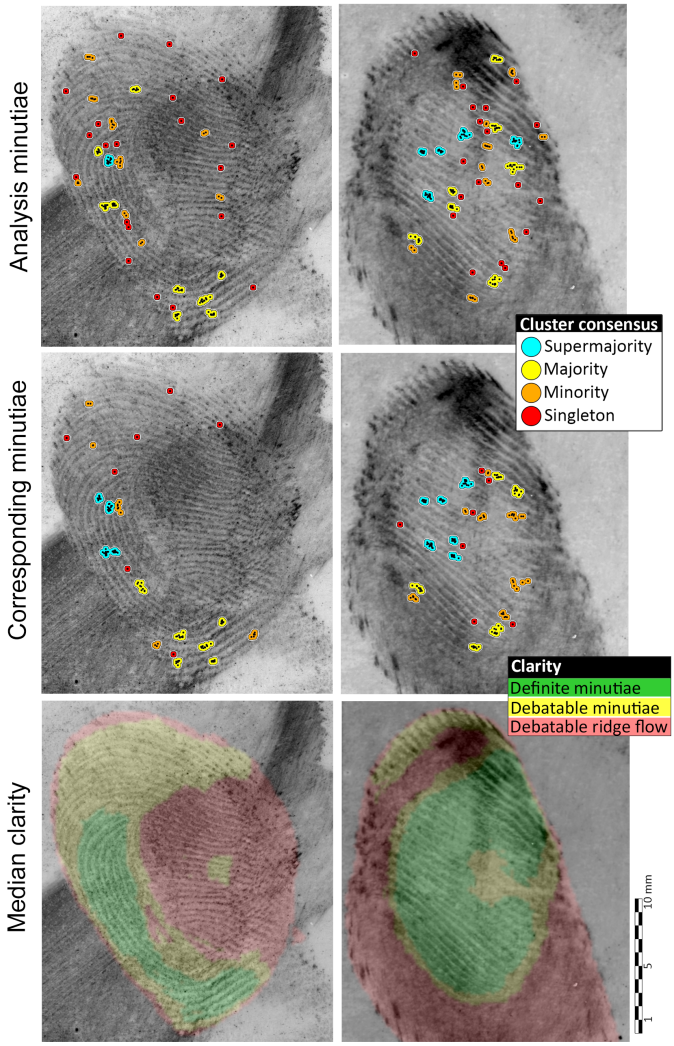


Figure 6: Two examples of latent markup. Marked minutiae are shown as small black dots inside color-coded clusters. Row 1: Analysis phase; cluster colors indicate the proportion of examiners who marked within that cluster. Row 2: Comparison phase; cluster colors indicate the proportion of comparing examiners who corresponded the minutia; only those minutiae marked as corresponding are shown. Row 3: Analysis phase; median clarity map, which combines clarity responses from all examiners. Unmarked latents, exemplars, and additional examples are included in DiB-2.

Examiners were instructed to mark all minutiae on the latent; on the exemplar, examiners were instructed to mark those minutiae that correspond with the latent. Examiners marked the location of each minutia; the software did not provide a means to indicate the minutia direction or type.

In this report, we generally summarize clarity results by aggregating the six levels specified by the examiners into two levels: clear and unclear. Clear areas are those where the examiner can follow individual friction ridges and is certain of the location, presence and absence of all minutiae. Unclear areas are those in which the presence or absence of any minutiae is debatable. Unless otherwise stated, we report the clarity of minutiae as marked by that examiner (sometimes "examiner clarity" to be explicit); we use "(un)clear minutia" to refer to a minutia marked by an examiner in an area the examiner marked as (un)clear. In some analyses we use "median clarity," calculated across multiple examiners. Clarity is explained more fully in DiB-1.3.

Since operational procedures vary among agencies on documentation of latent print examination, a single method of documentation cannot fully correspond to actual casework across multiple agencies. In this study, the fingerprint markup and value determinations complied with EFS, which is an international latent fingerprint data exchange standard [19]; the test instructions were derived from proposed standard instructions for the markup of latent prints [28]. The software application used for our experiment is a variant of the FBI's Universal Latent Workstation [25], which is widely used for operational casework by local, state, and federal agencies in the United States, as well as by

some international agencies. Participants were instructed in the test objectives, procedures, and software usage through a short video, a detailed instruction document, and practice exercises [6].

Participation was open to practicing latent print examiners and included a broad cross-section of the fingerprint community. A total of 170 latent print examiners participated: 33% were Certified Latent Print Examiners (an International Association for Identification certification); an additional 56% had other certifications or qualifications as latent print examiners, generally by their employers or non-US national accreditations; and 82% were from the United States. For further description of participants see [6].

The study included fingerprints collected under controlled conditions, and prints from operational casework. The fingerprint pairs were selected to vary broadly over a four-dimensional design space: number of corresponding minutiae, image clarity, presence or absence of corresponding cores and deltas, and complexity (based on distortion, background, or processing). The primary focus of the White Box study was to test the boundaries of sufficiency for individualization determinations, and therefore we deliberately limited the proportion of image pairs on which we expected unanimous determinations. The test dataset included 320 image pairs, 231 mated (from the same finger and subject) and 89 nonmated (from different fingers or subjects). The image pairs were constructed from 301 latents and 319 exemplars (DiB-1.2).

Each examiner was assigned 17 mated image pairs and 5 nonmated image pairs; these proportions were not revealed to participants. Results are based on 3730 responses, with a median of 12 examiners assigned to each image pair. Comparison-phase results are based on 2966 comparisons where neither the latent nor the exemplar was assessed to be NV. Results for corresponding minutiae are based on 3618 responses, excluding markups by five examiners who routinely did not annotate correspondences (details in DiB-1.4).

2.1 Clustering

Examiners' markups differed in whether or not individual minutiae were marked, and in the precise location where the minutiae were marked. In order to focus on whether examiners agree on the presence or absence of minutiae, we need to see past minor variations in minutia location. We use a commonly-used data clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [30], to classify minutiae marked by multiple examiners as representing the same minutia on the latent. The DBSCAN algorithm was parameterized with a reachability distance of 0.38mm (0.015 inch)[2] [4,18]; any marked minutiae within this distance of one another coalesce into a cluster (a cluster starts with an arbitrary marked minutia, grows to include any other marked minutiae (from all examiners) within that distance, and then iteratively grows to include any other marked minutiae within that distance of the cluster). As detailed in DiB-3, small changes to the reachability parameter had a large effect on the total number of resulting clusters, especially with respect to singletons (clusters containing only one marked minutia). The distance of 0.38mm was selected after extensively reviewing the algorithm's performance over a range of reachability settings. After performing this initial clustering, we then identified a relatively small number of clusters that had grown excessively large: for example, a single minutia mark located between what would otherwise have been two distinct ridge event locations would cause those two clusters to coalesce. These "overgrown" clusters were split using agglomerative hierarchical clustering to produce the final set of clusters for analysis (DiB-3).

For some potential uses, a composite or voted markup (based on multiple examiners' markups) is desirable. Such a composite markup could be constructed using the majority (or other consensus level) clusters, at the centroid of the minutia locations in that cluster, and clarity defined by the median.

2.2 Measuring interexaminer variation

Generally, "minutia" refers to an actual feature on the skin, or in an impression of the skin. However, in this study we have no special knowledge of the actual features beyond what we can learn from what was marked by examiners. To avoid ambiguity in what we are measuring, we define two terms:

- The annotation by an individual examiner at some location on the latent (***marked minutia***);
- A set of marked minutiae from multiple examiners that were grouped into the same cluster (***cluster***).

[2] The distance between ridges varies within an impression and between subjects, but average peak-to-peak distances are reported as varying between 0.43mm and 0.56mm [4, 18].

Our Analysis-phase results are based on 44,941 marked minutiae, which resulted in 10,324 clusters. We say that two examiners have marked the same minutia if both examiners marked within the same cluster. We define two closely related measures of interexaminer variation:

- For each marked minutia, we use the term ***reproducibility*** to refer to the proportion of other examiners who marked that minutia (i.e., marked within the same cluster).
- For each cluster, we use the term ***consensus*** to refer to the proportion of examiners who marked a minutia in that cluster.

Aggregate statistics for ***reproducibility*** are based on a sample of markings of minutiae (one event for each examiner who marked a minutia). Aggregate statistics for ***consensus*** are based on a sample of clusters (one event for all examiners who marked at a location). Therefore, those minutiae that were marked by a majority of examiners contribute more heavily to the aggregate reproducibility statistics than minutiae that were marked by a minority of examiners. We sometimes partition the minutiae or clusters according to the level of reproducibility or consensus: singleton (marked by only one examiner), minority (< 50% of examiners), majority (50-90%), and supermajority (>= 90%). Our measures of reproducibility and consensus are sensitive to the clustering parameters used: larger clusters would generally increase our measures of reproducibility, for example increasing the number of majority clusters and decreasing the number of singletons (DiB-3).

3 Results

Here we describe interexaminer variability in minutia markup on latent fingerprints and in marking correspondences to the exemplar. We discuss which factors account for the variability in minutia markup, the extent to which examiners agree when describing the clarity of ridge details, how examiners' changes to their markup from Analysis to Comparison relate to reproducibility, and how reproducibility relates to mating and determinations.

3.1 Reproducibility of Analysis-phase minutiae

Overall, the probability of randomly selected minutiae being reproduced (mean reproducibility) was 63% (DiB-4). However, as shown in Figure 7, clarity is a major determinant of whether examiners mark the same minutiae: reproducibility is lower in areas the examiner marked as unclear (47% mean reproducibility), and higher in areas marked as clear (70% mean reproducibility). Unclear minutiae were much less likely to be unanimously reproduced than clear (9% of unclear minutiae, 26% of clear), and much more likely to be singletons (17% of unclear, 7% of clear minutiae).

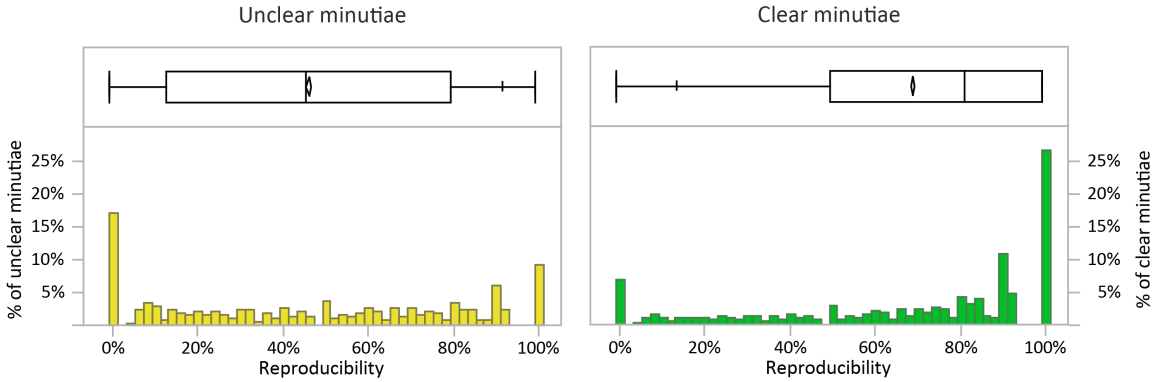


Figure 7. Reproducibility of Analysis-phase marked minutiae, by examiner clarity. The mean reproducibility was 63% (47% for unclear minutiae, 70% clear); median reproducibility was 75% (46% for unclear minutiae, 82% clear); 66% of minutiae were reproduced by the majority of other examiners, i.e., greater than 50% reproducibility (46% unclear, 73% clear). (n=44,941 minutiae: 12,782 unclear, 32,159 clear)

Figure 8 contrasts the two ways of measuring interexaminer variability: the ***reproducibility*** of marked minutiae (i.e., the 44,941 marked minutiae), and the extent of ***consensus*** among examiners that a minutia is present at a given location (i.e., the 10,324 minutia clusters). By counting each marked minutia equally, reproducibility gives more weight to minutiae marked by many examiners; consensus gives equal weight to each cluster regardless of how many examiners marked that minutia. A singleton is counted once in either case. As a result, the mean reproducibility (63%)

is higher than the mean consensus (36%). Most of the marked minutiae (68%) were reproduced by a majority of other examiners, but most of the clusters (coincidentally 68%) were marked by a minority of examiners.

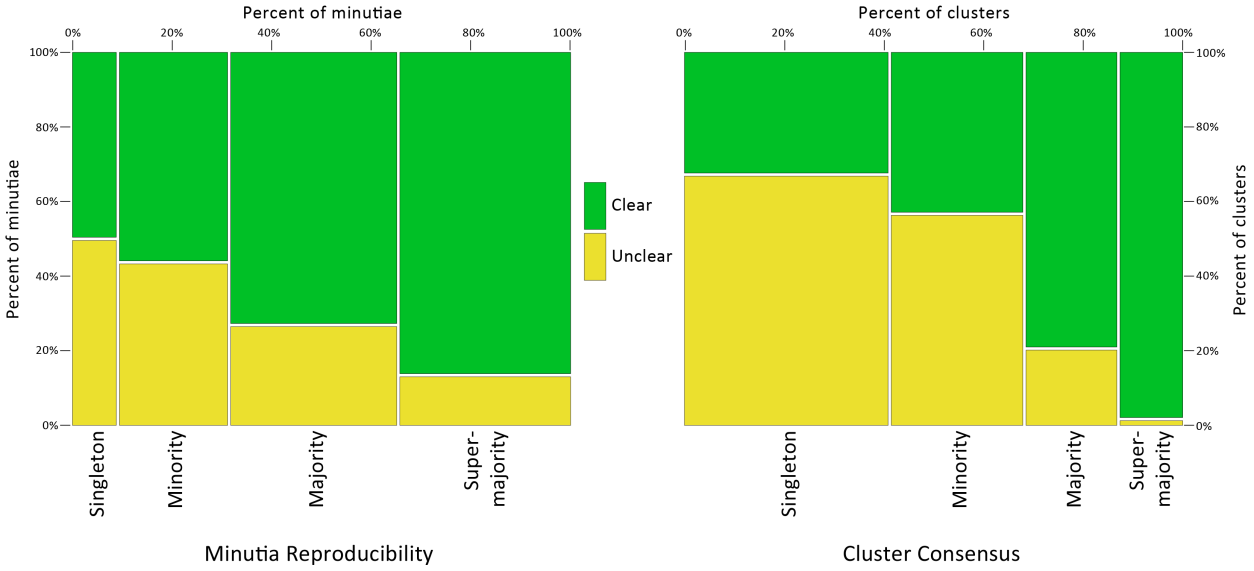


Figure 8: Mosaic plots showing the associations between clarity and interexaminer variability in minutia markup. (Left) minutia reproducibility by examiner clarity (n=44,941 minutiae); (Right) cluster consensus by median clarity (n=10,324 clusters). For example, there were 4269 singletons, accounting for 9% of marked minutiae and 41% of clusters.

The fact that an examiner marked a minutia, regardless of how that examiner marked clarity, indicates a high probability that a majority of examiners described the area as clear: even when examiners marked minutiae as unclear, on average about half of other examiners marked that area as clear (DiB-6). While marking a minutia as unclear effectively signaled low reproducibility, a voted description of clarity (median clarity map) provided an even better explanation of reproducibility (DiB-4.1). For example, 67% of the singletons were in median unclear areas, yet only 50% were marked as unclear by the examiner who marked the singleton; 98% of supermajorities were in median clear areas, but only 86% of those minutiae were marked as clear. Previously, we reported a similar result: median clarity predicted changes in minutia markup between Analysis and Comparison better than examiner clarity [7]. In general, we found that median clarity markups conform well to our expectations of proper and careful characterizations of latent clarity, by reducing the impact of outliers and imprecision found in the individual examiners' clarity markups. The (unexpected) result that median clarity was a better predictor of changes and reproducibility than examiner clarity suggests that greater consistency among examiners in describing clarity would make clarity markup more effective in flagging unreliable minutiae, and has the potential to make substantive disagreements among examiners more readily apparent.

There were many areas in the latents where there was no strong consensus among examiners on whether an area was clear or unclear; we refer to these areas as having "debatable clarity." Individual examiners were presumably uncertain how to mark clarity in some of these areas, but the test forced a choice between clear and unclear. Figure 9 indicates how these areas of debatable clarity contribute to our results. As the proportion of examiners describing an area as clear increased, both the number of minutiae marked and minutia reproducibility increased. Supermajorities sometimes occurred in areas where examiners did not agree on clarity (e.g., 20-80% voted clear). Even in areas that examiners agreed (90-100%) are clear, reproducibility was not unanimous: on review, the lack of unanimity usually could be attributed to adequate-but-difficult clarity, complex ridge flow, unclustered minutiae due to differences in location, or marking of features that were only debatably minutiae. Although reproducibility was lowest in areas that a large majority of examiners described as unclear, relatively few minutiae were marked in those areas: much of the lack of reproducibility therefore arose in areas of debatable clarity (e.g., 20-80% voted clear). This continuous voted measure of clarity provided a more complete explanation of the relationship between clarity and reproducibility than whether the majority of examiners described the location as clear or unclear (median clarity), which in turn provided a more complete explanation than whether the individual examiner described the location as clear or unclear.

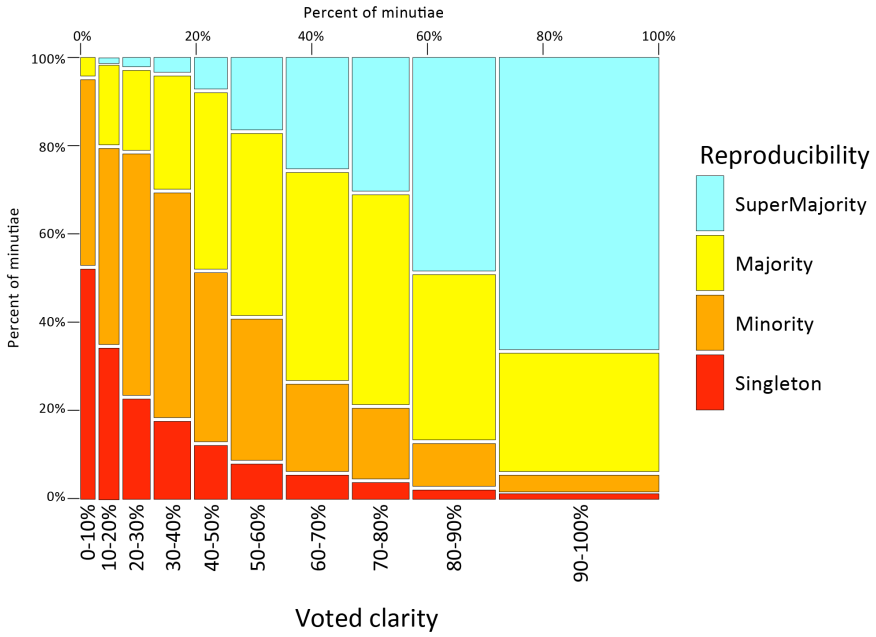


Figure 9: Voted clarity by reproducibility (n=44,941 minutiae). Voted clarity describes the percentage of examiners who described the location of that minutia as clear. 74% of minutiae were marked in areas described as clear by at least half of examiners.

As we discussed in the Introduction, one explanation for some lack of reproducibility is that examiners do not always agree on the region of interest. Additionally, examiners sometimes differ in whether they choose to mark minutiae in low-clarity areas. As a consequence, examiners often marked minutiae far away from those marked by other examiners. To quantify this effect, we measured the distance from each marked minutia to the nearest majority cluster (details in DiB-7). We can (somewhat arbitrarily) consider that a marked minutia is "relatively far" from a minutiae cluster if they are at least 2.5mm (0.1") apart; this would be about 5 ridge intervals on average. Similarly, marked minutiae are "very far" apart if they are at least 5.1mm (0.2"), about 10 ridges, apart. By that measure, 11.2% of marked minutiae are relatively far from the center of the nearest majority cluster (3.2% of median clear minutiae and 35.9% of median unclear minutiae); 3.5% of marked minutiae are very far from the nearest majority cluster (0.5% of median clear minutiae and 12.9% of median unclear minutiae). Disagreements among examiners regarding the regions in which to mark minutiae account for a substantial proportion of interexaminer variability, especially in unclear areas.

Another possible explanation for lack of reproducibility that we discussed in the Introduction is the potential ambiguity of whether a feature should be considered a minutia or a nonminutia feature, such as a dot or an event on an incipient ridge. Examiners were instructed to mark "other" (nonminutia) features when they were used as the basis for a Comparison determination; marking during Analysis was optional [6]. For this reason, markup of nonminutia features was incomplete in both phases, limiting our ability to measure disagreements on feature type. On review of the markups, singletons were often marked on incipient ridges, dots, or on nonminutia features in cores or deltas. In the Comparison phase, features other than minutiae were present in the area of only 4.5% of minutia clusters on the latents; not all of these represent potential disagreements regarding the type of the feature (DiB-5).

In addition to assessing interexaminer variability by marked minutiae (reproducibility) and by clusters (consensus), we can assess variability by entire markups. Based on the idea that examiners should agree on minutiae in clear areas and differences regarding unclear minutiae should be acceptable, we could define markups as being in "perfect" agreement with the majority when they satisfy two conditions: all minutiae marked by that examiner in clear areas are in majority clusters, and that examiner marked a minutia in each of the majority clusters (in any clarity). By that measure, 15% of the 3730 Analysis-phase markups of latents were in perfect agreement (including 9% with no clear minutiae or no majority clusters). If we loosen the requirements to "75% agreement" (the examiner marked at least 75% of the majority clusters, and at least 75% of the minutia that the examiner marked in clear areas coincided with majority clusters), 52% of markups were in agreement (DiB-4.2).

Minutia reproducibility tended to be higher on latents that examiners agreed are VID than on those that examiners agreed are not VID. Most of this association can be accounted for in terms of differences in clarity: those latents that examiners agreed are VID tend to have more minutiae marked in clear areas (DiB-4.4).

3.2 Reproducibility of Analysis-Comparison changes

In a previous White Box study report [7], we described the extent to which examiners changed their minutia markup of the latent from the Analysis phase (of the latent alone) to the Comparison phase (considering both the latent and exemplar). We reported that changes in markup were most prevalent on individualizations (minutiae were added or deleted on 90.3% of individualizations); for inconclusive and exclusion determinations, changes were more prevalent when the image pair was mated; a greater percentage of minutiae were deleted or added in unclear areas than in clear areas. Here, we see that the net effect of these changes was a small increase in minutia reproducibility on latents that were compared to mated exemplars; no net change in reproducibility was detected among the nonmate comparisons.

Deleted and added minutiae were each associated with low reproducibility. Examiners were more likely to delete minutiae that were marked by a minority of other examiners. Interestingly, the minutiae that they added (even those in clear areas) also tended to be marked by a minority of other examiners: this might be due in part to a motivation to thoroughly document individualization conclusions. These effects were particularly pronounced for singletons (e.g., among latents that were compared, 23% of singletons were deleted). The association of deleted and added minutiae with low reproducibility does not simply reflect higher volatility in unclear areas: a strong inverse association between changes and reproducibility remains after controlling for clarity. In other words, proportionally more minutiae were deleted and added in unclear areas than in clear areas and, after accounting for clarity, those minutiae with low reproducibility were more likely to be deleted or added than those with high reproducibility (details in DiB-9).

3.3 Reproducibility of corresponding minutiae from the Comparison phase

Comparisons between a latent and an exemplar introduce another dimension of interexaminer variation in minutia markup: the examiners may differ not only on whether they mark a given minutia in the latent, but also on whether those minutiae that they agree are present in the latent correspond to the exemplar. Interpreting interexaminer variability in marking minutia correspondences is complicated by the fact that marking of correspondences is strongly associated with determinations: comparison markup is only available from those examiners who agreed that the latent is suitable for comparison (the number of examiners varies considerably; see DiB-1.4), and examiners who individualize tend to mark more corresponding minutiae than those who exclude or are inconclusive [6]. For these reasons, we describe interexaminer variability for Comparison-phase results slightly differently than for Analysis-phase results, as shown in Table 1.

Table 1 describes the reproducibility of marked minutiae in the Comparison phase, categorized by whether the examiners corresponded the minutiae.[3] For each examiner ("Examiner A") the probability that a second examiner ("Examiner B") marked and corresponded a minutiae was measured by considering all other examiners, regardless of whether the other examiners compared the latent. On average, if an examiner marked a minutia on the latent and corresponded that minutia to the exemplar, the probability that a second examiner also marked and corresponded that minutia was 69% for clear minutiae and 47% for unclear. When two examiners ***both*** individualized, that probability increased to 76% for clear and 57% for unclear (Table D19 in DiB-10.2). Examiners marked few correspondences on nonmated pairs: the probability that a second examiner reproduced a correspondence on a nonmated pair was 8% regardless of clarity (Table D18 in DiB-10.2).

Clarity accounts for much of the difference in whether the second examiner marked the minutia, but little of the difference in whether the second examiner corresponded a marked minutia, as shown in the right column of Table 1. In cases where two examiners agreed that a minutia was present on the latent and one examiner corresponded the minutia, the probability that the second examiner would also correspond the minutia was approximately the same for clear and unclear minutiae (88% vs. 84%).

The probability of examiners corresponding marked minutiae was correlated with the reproducibility of those minutiae. On individualizations, examiners corresponded 60% of their singletons and 92% of minutiae that were

[3] To construct Table 1, clustering was performed on all marked minutiae, whether marked during Analysis or Comparison, including those that were deleted during Comparison; DiB-10.2 includes additional results related to whether minutiae were deleted or added.

unanimously marked by comparing examiners; when examiners did not individualize, they corresponded 10% of their singletons and 25% of minutiae unanimously marked by comparing examiners (DiB-10.1). Note that because the latent and exemplar do not always completely overlap, not all minutiae in the latent can be corresponded with a given exemplar.

				Examiner B				
				Did not mark	Marked			Marked and compared minutiae that were corresponded
					Did not compare (NV)	Compared		
All Minutiae			Minutiae			Unassoc.	Corresp.	
Examiner A	Clear minutiae	Unassociated	14,744	36%	5%	44%	15%	26%
		Corresponding	20,470	20%	2%	10%	69%	88%
	Unclear minutiae	Unassociated	8221	59%	6%	25%	11%	30%
		Corresponding	7459	42%	2%	9%	47%	84%
Examiner A individualized								
Examiner A	Clear minutiae	Unassociated	5507	41%	1%	39%	20%	34%
		Corresponding	18,823	20%	1%	9%	70%	89%
	Unclear minutiae	Unassociated	2600	66%	1%	20%	14%	41%
		Corresponding	6576	42%	1%	8%	49%	86%

Table 1: When examiner A marked a minutia, what examiner B did at that location, for all minutiae marked during Analysis (including deletions) or added during Comparison (n=50,894 minutiae, 3618 responses), and conditioned on examiner A having individualized (n=33,506 minutiae, 1654 responses). "Unassociated" includes all marked minutiae that were not corresponded. Percentages calculated as weighted sums over all other examiners who marked each latent, such that each minutia marked by examiner A is weighted equally. "Marked and compared minutiae that were corresponded" is the probability that examiner B corresponded a minutia given that examiner B marked that minutia and compared the latent to the exemplar.

Examiners were instructed to mark any discrepancies used to support their exclusion determinations. Reproducibility of discrepancies was not substantially greater than chance (see DiB-11). A likely explanation for the lack of reproducibility of discrepancies may simply be that perceived differences (e.g., in ridge flow) often cannot be localized to a single point. Only 29% of exclusions had any discrepancies marked; and most examiners never marked more than one discrepancy on a latent. Minutiae marked as discrepant by one examiner were often (11%) marked as correspondences by other examiners (DiB-10.2, Table D16).

Review of the markup provided another explanation for variation in minutia markup. The locations at which minutiae were marked often vary substantially among examiners. Marked minutiae in separate clusters on the latent were often corresponded to a single cluster on the exemplar: multiple examiners agreed that the minutia was present and agreed on the location in the exemplar, but differed substantially in where they marked the minutia on the latent. In order to better understand the extent of this issue, we clustered the minutiae marked on the exemplars, so that we could see how these exemplar clusters corresponded to latent clusters. Considering only those clusters in which corresponding minutiae were marked, there were 6% fewer clusters on the exemplars than on the latents. However, this effect was observed in both directions: 15% of exemplar clusters were corresponded to more than one latent cluster; 9% of latent clusters were corresponded to more than one exemplar cluster. Although some of these clustering issues might have been resolved with a different clustering algorithm, often the distance was large enough that we would not expect any clustering algorithm to group them (DiB-12).

4 Discussion

We identified several factors that affect minutia reproducibility: clarity, region of interest, feature type, and location. The fact that an examiner marked a minutia, regardless of how that examiner marked clarity, indicates a high probability that a majority of examiners described the area as clear. Marking a minutia as unclear was a good predictor that reproducibility would be low: in effect, by marking minutiae as unclear, examiners seem to anticipate low reproducibility. Differences in markup were most prevalent in areas where examiners did not agree on clarity, in part because relatively few minutiae were marked in areas that examiners agreed were unclear. Much of the variability, especially in unclear areas, can be attributed to differences in which areas of the prints examiners chose to mark: 36% of minutiae marked in median unclear areas during Analysis were relatively far away from the nearest majority cluster (at least 0.1 inch or approximately five or more ridge intervals). Some variability can be attributed to disagreements

regarding minutia type: singletons were often marked on incipient ridges, dots, or on nonminutia features in cores or deltas. Additionally, some of the reported variability can be attributed to uncertainty in the precise location at which to mark a minutia on the latent: marked minutiae that were singletons or in separate clusters in the latent were often corresponded to a single location in the exemplar.

Some of the reported variability can be attributed to our measurement techniques, including the clustering algorithm, fingerprint selection, and markup procedures. Clustering was sensitive to our choice of radius, and did not account for factors such as local ridge width and direction. The fingerprints were selected to test the boundaries of sufficiency for individualization determinations, deliberately limiting the proportion of image pairs on which we expected unanimous determinations. Because requirements and procedures for markup are not standardized in practice, the tools and procedures we used were novel to the participants, contributing to the variability.

In a separate study evaluating variation in examiners' determinations [3], we found that much of the lack of (interexaminer) reproducibility of value and comparison determinations was associated with images and image pairs on which we also observed low (intraexaminer) repeatability. We assume there is a similar association between reproducibility and repeatability of minutia markup, based on previously reported results [6] in which we saw a notable lack of repeatability in minutia markup (on a small sample of markups).

In our previous work [5,6], we found that the association between examiners' minutia counts and their determinations was ***not*** notably affected by minutia clarity. Here, however, we see that clarity has a notable effect on the reproducibility of marked minutiae. Thus, while the total minutia count (clear and unclear minutiae) is indicative of examiners' determinations, most of the variance accounting for examiner differences in marked minutiae arises in unclear areas: when examiners individualized (or assessed a latent to be VID) those examiners generally marked more minutiae in unclear areas than examiners whose comparison determinations were inconclusive (or who assessed the latent to be NV).

We should not assume that reducing variability in markup would necessarily improve reproducibility of determinations. There are some indications that the relationship between markup and determinations may not be a simple forward causality: we have previously reported that examiner determinations appear to influence markup, as evidenced by the tendency of examiners to modify their latent markup more extensively when individualizing than when inconclusive [7], and by a tendency not to mark just fewer than the minimum number of minutiae typically associated with individualization determinations [6]. It is possible that some of the variability in markup relates to processes motivated by the determination, such as reviewing unclear and peripheral areas to double-check one's work and document that nothing calls the conclusion into doubt.

There is not currently any method of defining a "correct" minutia markup for any given latent. An examiner's decision of whether a minutia is present in an unclear location is analogous to an examiner's decision as to whether the similarity of two prints is sufficient to make an individualization determination: in either case, the best information we have to evaluate the appropriateness of examiners' decisions is the collective judgment of other experts. Our method of clustering minutiae could be used to develop training sets in which an "ideal" markup would be based on a group consensus.

Differences in minutia markup are not always due to differences in interpretation, but often may be due merely to differences in how examiners document their interpretations. Examiners' clarity markup is a useful indicator of the reproducibility of the minutiae they marked, which suggests that greater consistency among examiners in describing clarity has the potential to appreciably limit the apparent disagreements among examiners in the interpretation of latent prints. We expect that standardizing markup of features and clarity (through formal specification, inclusion in training, and broad usage in operational casework) would facilitate greater transparency by making markup a more reliable description of examiners' interpretations.

Acknowledgments

We thank the latent print examiners who participated in this study, and Erik Stanford for his technical support with the clustering algorithms. This is publication number 15-15 of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. This work was funded in part under a contract award to Noblis, Inc. from the FBI Biometric Center of Excellence and in part by the FBI Laboratory Division. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

References

1 Hicklin, R.A., et al. (2011). Latent fingerprint quality: a survey of examiners. Journal of Forensic Identification, 61(4): 385-419.

2 Ulery, B.T., Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci, 108(19): 7733-7738. http://dx.doi.org/10.1073/pnas.1018707108

3 Ulery, B.T., Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. PloS ONE, 7(3), e32800. http://dx.doi.org/10.1371/journal.pone.0032800

4 Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2013). Assessing the clarity of friction ridge impressions. Forensic Sci Int, 226(1):106-117. http://dx.doi.org/10.1016/j.forsciint.2012.12.015

5 Ulery, B.T., Hicklin, R.A., Kiebuzinski, G.I., Roberts, M.A., Buscaglia, J. (2013). Understanding the sufficiency of information for latent fingerprint value determinations. Forensic Sci Int, 230(1): 99-106. http://dx.doi.org/10.1016/j.forsciint.2013.01.012

6 Ulery, B.T., Hicklin, R.A., Roberts, M.A., Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE, 9(11), e110179. http://dx.doi.org/10.1371/journal.pone.0110179

7 Ulery, B.T., Hicklin, R.A., Roberts, M.A., Buscaglia, J. (2014). Changes in latent fingerprint examiners' markup between Analysis and Comparison. Forensic Sci Int, 247: 54-61. http://dx.doi.org/10.1016/j.forsciint.2014.11.021

8 Kalka, N.D., Hicklin, R.A. (2014). On relative distortion in fingerprint comparison. Forensic Sci Int, 244: 78-84. http://dx.doi.org/10.1016/j.forsciint.2014.08.007

9 Neumann, C., Champod, C., Yoo, M., Genessay, T., & Langenburg, G. (2013). Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination. National Institute of Justice, Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pdf

10 Evett, I.W., Williams, R.L. (1996). A review of the sixteen point fingerprint standard in England and Wales, Journal of Forensic Identification, 46(1): 49–73. [Also published in Fingerprint Whorld, 21 (82), October, 1995]

11 Schiffer, B., Champod, C. (2007). The potential (negative) influence of observational biases at the analysis stage of fingermark individualisation. Forensic Sci Int, 167(2): 116-120. http://dx.doi.org/ 10.1016/j.forsciint.2006.06.036

12 Fine, G.A. (2006). A review of the FBI's handling of the Brandon Mayfield case. Washington, DC: US Department of Justice Office of the Inspector General.

13 Campbell, A. (2011). The fingerprint inquiry report. Edinburgh (Scotland): APS Group Scotland. Available: http://www.thefingerprintinquiryscotland.org.uk/inquiry/3127-2.html

14 State of Indiana v. Lana Canen (2002).

15 Texas vs Gregory Edward Wright (1997). Trial Cause No. F97-01215-PJ. (http://www.freegregwright.com/NewVol47.pdf)

16 German, E. (2014) Problem Idents [Website; accessed 5 June 2015] http://onin.com/fp/problemidents.html

17 Kim Jackson v. State of Florida (2014). Case No. SC13-2090.

18 Ashbaugh, D. (1999). Quantitative-Qualitative Friction Ridge Analysis: an Introduction to Basic and Advanced Ridgeology. Boca Raton, FLCRC Press.

19 National Institute of Standards (2013). American National Standard for Information Systems: Data format for the interchange of fingerprint, facial & other biometric information. ANSI/NIST-ITL 1-2011 Update:2013. (NIST Special Publication 500-290 Rev1) Gaithersburg, MD: National Institute of Standards and Technology. http://fingerprint.nist.gov/standard

20 Swofford, H., Steffan, S., Warner, G., Bridge, C., Salyards, J. (2013). Inter and intra-examiner variation in the detection of friction ridge skin minutiae. Journal of Forensic Identification, 63(5), 553-571. https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=265931

21 Dror, I.E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: inter-and intra-expert consistency and the effect of a 'target' comparison. Forensic Sci Int, 208(1): 10-17. http://dx.doi.org/10.1016/j.forsciint.2010.10.013

22 Langenburg, G. (2012). A critical analysis and study of the ACE-V process (unpublished doctoral dissertation). Université de Lausanne, Lausanne. http://www.unil.ch/files/live//sites/esc/files/shared/Langenburg_Thesis_Critical_Analysis_of_ACE-V_2012.pdf

23 SWGFAST (2012). Standard for the Documentation of Analysis, Comparison, Evaluation, and Verification (ACE-V) (Latent). http://www.swgfast.org/documents/documentation/121124_Standard-Documentation-ACE-V_2.0.pdf

24 Langenburg G., Champod C. (2011). The GYRO System - A Recommended Approach to More Transparent Documentation. Journal of Forensic Identification 61(4):373-384

25 Federal Bureau of Investigation; Universal Latent Workstation (ULW) [software]. https://www.fbibiospecs.org/Latent/LatentPrintServices.aspx

26 Mideo Latentworks [software]. http://www.mideosystems.com/latentworks.htm

27 PiAnoS Picture Annotation System [software]. https://ips-labs.unil.ch/pianos

28 Taylor, M.K., et al. (2013). Markup Instructions for Extended Friction Ridge Features. (NIST Special Publication 1151) Gaithersburg, MD: National Institute of Standards and Technology. http://dx.doi.org/10.6028/NIST.SP.1151

29 Ulery, B.T., Hicklin, R.A., Roberts, M.A., Buscaglia, J. (2016) Data describing interexaminer variation of minutia markup on latent fingerprints. Forensic Sci Int Data in Brief (submitted).

30 Ester, M, Kriegel, H.P., Sander, J. , Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9. CiteSeerX: 10.1.1.71.1980.

Appendix A. Supporting Information — Glossary

This section defines terms and acronyms as they are used in this paper.

ACE	The phases of ACE-V prior to verification: Analysis, Comparison, Evaluation.
ACE-V	The prevailing method for latent print examination: Analysis, Comparison, Evaluation, Verification.
AFIS	Automated Fingerprint Identification System (generic term)
Analysis phase	The first phase of the ACE-V method. In this test, the examiner annotated the latent and made a value determination before seeing the exemplar print.
ANSI/NIST-ITL	An electronic file and interchange format that is the basis for biometric and forensic standards used around the world, including the FBI's EBTS and Interpol's INT-I, among others. As of 2011, this incorporates the Extended Feature Set (EFS) definition of friction ridge features used in this study. [1]
Clarity	The clarity of a friction ridge impression refers to the fidelity with which anatomical details are represented in a 2D impression, and directly corresponds to an examiner's confidence that the presence, absence, and details of the anatomical friction ridge features in that area can be correctly discerned in that impression. (Note: The term "clarity" is used here instead of "quality" to avoid ambiguity, since the latter term as used in biometrics and forensic science is often used to include not only clarity but also the quantity or distinctiveness of features.) See Fig. D2.
Clarity map	A color-coded annotation of a friction ridge image indicating the clarity for every location in the print, as described in [2] and defined in EFS (in the ANSI/NIST-ITL standard [1]).
Clear area	Area of local clarity where the examiner can be certain of the location, presence and absence of all minutiae (see Fig. D2). This may be assessed by an individual examiner ("examiner clarity") or by all examiners who examined the print ("median clarity"). In the clarity map, Green, Blue, and Aqua are here all considered clear areas.
Cluster (Minutia cluster)	In this paper, a cluster refers to a set of examiner-marked minutiae that are algorithmically determined to be a single friction ridge event.
Comparison phase	In this test, there was no procedural demarcation between the second (Comparison) and third (Evaluation) phases of the ACE-V method; hence, this refers to the single combined phase during which both images were presented side-by-side.
Comparison determination	The determination of individualization, exclusion, or inconclusive reached in the Comparison phase of the test. SWGFAST [3] refers to this determination as the Evaluation Conclusion.
Consensus	The proportion of examiners who marked a minutia within a given cluster.
Corresponding minutia	Explicit annotation by an examiner associating a marked minutia in the latent with a marked minutia in the exemplar, as defined in EFS. Examiners were instructed to mark all such correspondences that they used to make their Comparison determinations. Also described as Definite correspondence.
DBSCAN	Density-Based Spatial Clustering of Applications with Noise, a clustering algorithm used to classify the minutiae marked by multiple participants into sets (clusters) representing the same friction ridge event.
Debatable correspondence	An explicitly marked relationship between a feature marked on a latent and a feature marked on the exemplar indicating an apparent correspondence between those features that does not rise to the threshold of (definite) correspondence. (Not to be confused with debatable ridge flow or debatable features, which were indicated by painting the image clarity.)
Definite correspondence	See Corresponding minutia.
Determination	An examiner's decision: the Analysis phase results in a latent value determination, and the Comparison phase results in a Comparison determination.
Discrepancy	A minutia that the examiner indicates exists in one print and is definitely not present in the other print. Participants were instructed to indicate points in one print that definitely do not exist in the other print as needed to support an exclusion determination. (Also known as noncorresponding minutia).
Dot	An isolated friction ridge unit whose length approximates its width in size. In this study, examiners were instructed to mark dots as "other" features, not as minutiae.
EFS	The Extended Feature Set — fingerprint and palmprint features as defined in ANSI/NIST-ITL [1].
Exclusion	The comparison determination that the latent and exemplar fingerprints did not come from the same finger. For our purposes, this is *exclusion of source*, which means the two impressions originated from different sources of friction ridge skin, but the subject cannot be excluded, whereas *exclusion of subject* means the two impressions originated from different subjects.
Exemplar	A fingerprint from a known source, intentionally recorded.
Feature	Minutia, core, delta, or "other" point in a print.
IAFIS	The FBI's Integrated Automated Fingerprint Identification System. In 2013, IAFIS latent print services were replaced by the FBI's Next Generation Identification (NGI) system.
Image	A fingerprint as presented on the computer screen to test participants. The test software permitted rotating, panning, zooming, tonal inversion, and grayscale adjustment of the image.
Incipient ridge	A friction ridge not fully formed that may appear shorter and thinner in appearance than fully developed friction ridges. In this study, examiners were instructed to mark incipient ridges or ridge endings as "other" features, not as minutiae.
Inconclusive	The comparison determination that neither individualization nor exclusion is possible.
Individualization	The comparison determination that the latent and exemplar fingerprints originated from the same source. Individualization is synonymous with identification for latent print determinations in the U.S. Both are defined as: "the decision by an examiner that there are sufficient discrimination friction ridge features in agreement to conclude that two areas of friction ridge impressions originated from the same source. Individualization of an impression to one source is the decision that the likelihood the impression was made by another (different) source is so remote that it is considered as a practical impossibility."[3,4]

Latent (or latent print)	A friction ridge impression from an unknown source. In North America, "print" is used to refer generically to known or unknown impressions [5]. Outside of North America, an impression from an unknown source (latent) is often described as a "mark" or "trace," and "print" is used to refer only to known impressions (exemplars).
Level-3 detail	Friction ridge dimensional attributes such as width, edge shapes, and pores.
Mated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of the same source (finger). Compare with "individualization," which is an examiner's ***determination*** that the prints are from the same source.
Median clarity map	A clarity map combining the annotations from multiple examiners, based on the median clarity at each location across the clarity maps from all examiners who annotated the clarity of an image.
Marked minutia	An annotation by an examiner on the print indicating the presence of a minutia at that location.
Minutia	An event along the path of a single friction ridge, such as a bifurcation or ridge ending. Examiners were instructed to mark features such as scars, dots, incipient ridges, creases and linear discontinuities, ridge edge features, or pores as "other" features, not as minutiae. In this study, examiners did not differentiate between bifurcations and ending ridges.
Nonmated	A pair of images (latent and exemplar) known *a priori* to derive from impressions of different sources (different fingers and/or different subjects).
NV (No value)	The impression is not of value for individualization and contains no usable friction ridge information. See also VEO and VID.
Other point	In this study, features such as scars, dots, incipient ridges, creases and linear discontinuities, ridge edge features, or pores (i.e., features other than minutiae, cores, and deltas).
Region of interest	Area of the image that includes the single, contiguous fingerprint impression being considered.
Reproducibility	The reproducibility of a minutia is measured as the proportion of other examiners who marked that minutia, as determined by the clustering algorithm.
Retained minutia	A minutia that was marked during the Analysis phase and was not deleted or moved in the Comparison phase.
Source	An area of friction ridge skin used to create an impression. Two impressions are said to be from the "same source" when they have in common a region of overlapping friction ridge skin.
Sufficient	An examiner's assessment that the quality and quantity of information in a print (or image pair) justifies a specific determination (especially used with respect to the decision between individualization and inconclusive).
ULW	The FBI's Universal Latent Workstation software. [6]
Unclear area	Area where the ridge flow may or may not be clear, but the examiner cannot be certain of the location, presence and absence of all minutiae (see Fig. D2). This may be assessed by an individual examiner ("examiner clarity") or by all examiners who examined the print ("median clarity"). In the clarity map, yellow and red are considered unclear areas. Black areas are outside the region of interest, but are considered unclear in those (few) instances in which minutiae were recorded in black areas, generally due to border conditions.
Value determination	An examiner's determination of the suitability of an impression for comparison: value for individualization (VID), value for exclusion only (VEO), or no value (NV). A latent value determination is made during the Analysis phase. Agency policy often reduces the three value categories into two, either by combining VID and VEO into a value for comparison category or by combining VEO with NV into a "not of value for individualization" (Not VID) category [survey in 7].
VEO	Value for exclusion only: Value determination based on the analysis of a latent that the impression is of value for exclusion only and contains some friction ridge information that may be appropriate for exclusion if an appropriate exemplar is available. See also NV and VID.
VID	Value for individualization: Determination based on the analysis of a latent that the impression is of value and is appropriate for potential individualization if an appropriate exemplar is available. See also VEO and NV.

1 National Institute of Standards (2013). American National Standard for Information Systems: Data format for the interchange of fingerprint, facial & other biometric information. ANSI/NIST-ITL 1-2011 Update:2013. (NIST Special Publication 500-290 Rev1) Gaithersburg, MD: National Institute of Standards and Technology. http://fingerprint.nist.gov/standard

2 Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2013). Assessing the clarity of friction ridge impressions. Forensic Sci Int, 226(1):106-117. http://dx.doi.org/10.1016/j.forsciint.2012.12.015 (preprint: http://www.noblis.org/media/4b209d60-a147-414e-8bba-90c3a1e22c18/docs/article_assessing_clarity_latent_friction_ridge_pdf)

3 SWGFAST (2013). Standards for Examining Friction Ridge Impressions and Resulting Conclusions, Version 2.0. http://www.swgfast.org/documents/examinations-conclusions/130427_Examinations-Conclusions_2.0.pdf

4 SWGFAST (2012). Individualization / Identification Position Statement, Version 1.0. http://swgfast.org/Comments-Positions/120306_Individualization-Identification.pdf

5 SWGFAST (2011). Standard terminology of friction ridge examination, Version 3.0. http://swgfast.org/documents/terminology/110323_Standard-Terminology_3.0.pdf

6 Federal Bureau of Investigation; Universal Latent Workstation (ULW) Software. https://www.fbibiospecs.org/Latent/LatentPrintServices.aspx

7 Ulery, B.T., Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. Proc Natl Acad Sci, 108(19): 7733-7738. http://dx.doi.org/10.1073/pnas.1018707108

Data article

Data supporting interexaminer variation of minutia markup on latent fingerprints

Authors

Bradford T. Ulery,[a] R. Austin Hicklin,[a] Maria Antonia Roberts,[b] and JoAnn Buscaglia [c]

Affiliations

[a] Noblis, Falls Church, Virginia, USA; [b] Latent Print Support Unit, Federal Bureau of Investigation Laboratory Division, Quantico, Virginia, USA; [c] Counterterrorism and Forensic Science Research Unit, Federal Bureau of Investigation Laboratory Division, Quantico, Virginia, USA

Contact email

JoAnn Buscaglia: joann.buscaglia@ic.fbi.gov

Abstract

The data in this article supports the research paper entitled "Interexaminer variation of minutia markup on latent fingerprints." The data in this article describes the variability in minutia markup during both analysis of the latents and comparison between latents and exemplars. The data was collected in the White Box latent print examiner study, in which each of 170 volunteer latent print examiners provided detailed markup documenting their examinations of latent-exemplar pairs of prints randomly assigned from a pool of 320 pairs. Each examiner examined 22 latent-exemplar pairs; an average of 12 examiners marked each latent.

Specifications Table

Subject area	*Forensic Science*
More specific subject area	*Latent fingerprints*
Type of data	*Tables, graphs, text descriptions*
How data was acquired	*Markup of latent fingerprints by latent print examiners under test conditions*
Data format	*Analyzed*
Experimental factors	*Feature types, locations, correspondences; local ridge clarity; examiner determinations*
Experimental features	*Automated clustering algorithms used to classify minutiae marked by multiple examiners as representing the same minutia*
Data source location	*N/A*
Data accessibility	*Within the Data in Brief article*

Value of the data

- Latent print examiners often differ in the features they use in the analysis and comparison of fingerprints. This data provides a wealth of information on how markup varies among examiners, how this relates to the quality of the fingerprints and to examiners' differing determinations.
- We provide this data in order to serve as a benchmark, to strengthen the community's understanding of the latent print examination process.
- This data provides greater visibility into the bases for examiners' decisions, and increases the community's understanding of subjectivity in latent print examination.
- This data may assist the community in deciding how to improve operational procedures, training, and standardization.
- This data may be of particular interest for automated fingerprint identification systems, which rely on human markup of minutiae.

Contents

Overview of Data

This paper presents data describing the variation in how minutiae are marked on latent fingerprints by latent print examiners, in support of the article "Interexaminer variation of minutia markup on latent fingerprints" [1]. The underlying data was collected in the "White Box" study [2]; the aspects of that data specific to interexaminer variation in minutiae markup have not been previously published.

Experimental Design, Materials and Methods

The test procedure, participants, and fingerprint data are summarized here, and are described in greater detail in [2].

DiB-1 Materials and methods

DiB-1.1 Test procedure

Fig. D1 summarizes the test workflow, which conforms broadly to the prevailing ACE methodology. This study did not address the Verification phase. Examiners could review and revise their work prior to submitting their results. Examiners were free to modify the markup and value determination for the latent after the exemplar was presented, but any such changes were recorded and could be compared with their Analysis responses. For a more complete description of the test procedure, including the complete test instructions and introductory video, see our previous report [2].

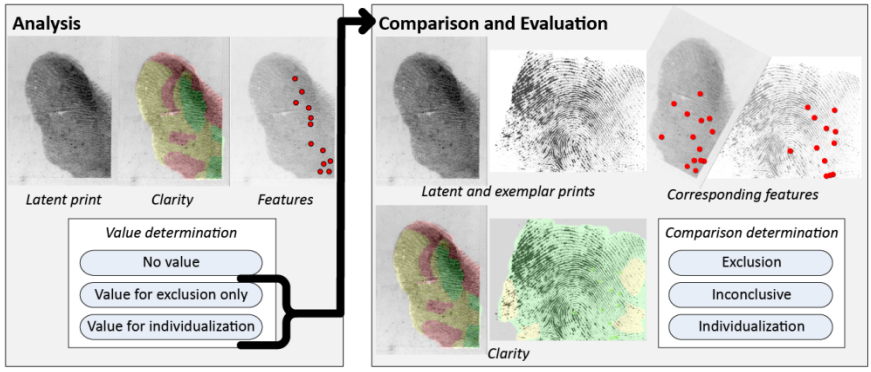


Fig. D1: Test workflow. Each examiner was assigned a distinct, randomized sequence of latent-exemplar image pairs. For each pair, the latent was presented first for a value determination. If the latent was determined to be no value, the test proceeded directly to the latent from the next image pair; otherwise, an exemplar was presented for comparison and evaluation.

DiB-1.2 Fingerprints

The fingerprints for the study were collected at the FBI Laboratory and at Noblis under controlled conditions, and from operational casework datasets collected by the FBI. We provide a detailed description of the fingerprint data selection process in [Appendix S.5 in 2]. All prints were impressions of distal segments of fingers, including some sides and tips.

The latents were processed using a variety of development techniques. The processed latents were captured electronically at 8-bit grayscale, uncompressed, at a resolution of 1000 pixels per inch.

The exemplars included both rolled and plain impressions captured as inked prints on paper cards or using FBI-certified livescan devices; they were captured at 8-bit grayscale, 1000 or 500 pixels per inch and either uncompressed or compressed using Wavelet Scalar Quantization [3].

The fingerprint pairs were selected to vary broadly over a four-dimensional design space: number of corresponding minutiae, image clarity, presence or absence of corresponding cores and deltas, and complexity (based on distortion, background, or processing). The primary focus of the White Box study was to test the boundaries of sufficiency for individualization determinations, and therefore we deliberately limited the proportion of image pairs on which we expected unanimous determinations.

We selected nonmated pairs to result in challenging comparisons either by down-selecting among exemplar prints returned by searches of the FBI's Integrated AFIS (IAFIS) or from among neighboring fingers from the same subject.

To ensure coverage of the design space and balance of image pairs across examiners, the assignments of fingerprint images to examiners were randomized based on an incomplete block design (with examiners as blocks, image pairs as factor levels), balanced to the extent possible (using the criterion of D-Optimality).

For each image pair assigned to an examiner, the test process saved two data files: one saved upon completion of the Analysis phase (before the exemplar print was presented) and a second upon completion of the Comparison phase. The files complied with the ANSI/NIST-ITL [4] standard, using the COMP transaction described in the Latent Interoperability Transmission Specification [5].

DiB-1.3 Local ridge clarity

The annotations of local ridge clarity complied with the Extended Feature Set (EFS), which is part of the ANSI/NIST-ITL standard [4]. Fig. D2 summarizes the color-coding method for describing clarity [6]. For minutiae, the primary distinction with regard to clarity is that for green or better areas, the examiner is "certain of the location, presence, and absence of all minutiae" (White Box Instructions, [Appendix S22 in 2]). Yellow areas indicate the opposite, that location, presence, and/or absence are not certain. Black or red areas should not have any marked minutia: when this occurs it is often due to imprecise painting of the clarity, or to not following instructions.§ For this analysis, we simplified the classification to clear (green or better) vs. unclear (yellow or worse).

Unless otherwise stated, we report the clarity as marked by that examiner. In some analyses we use the median clarity across multiple examiners, which combines the clarity maps from the examiners who were assigned that pair to

§ *1.9% of the 44,941 minutiae were marked in black areas, and 2.3% were in red areas.*

represent a group consensus. This reduces the impact of outlier opinions and imprecision. When constructing the median clarity maps, we excluded four examiners whose clarity markup did not comply with the test instructions.

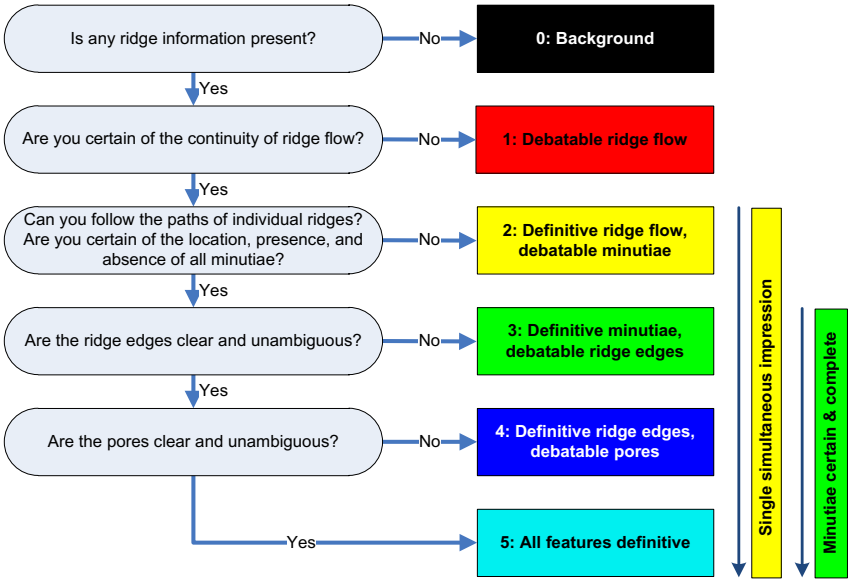


Fig. D2: Definitions of local image clarity. (from [4])

DiB-1.4 Test size

As detailed in [2], we received valid responses from 170 participants. Each participant was assigned 22 image pairs from a pool of 320 total pairs. Early in the testing process, a problem was identified in seven image pairs; ten responses on these image pairs were excluded, yielding a total of 3730 valid responses from the Analysis phase. Examiners marked 44,941 minutiae on 3550 latents (180 Analysis-phase markups included no minutiae).

Comparison-phase results are based on 2966 comparisons where neither the latent nor the exemplar was assessed to be NV. These results omit 2 invalid determinations (software issue) and 762 NV determinations (713 Analysis-phase latent NV, 43 Comparison-phase latent NV, and 6 Comparison-phase exemplar NV). Our previous report on changes made from Analysis to Comparison [7] omitted an additional nine responses whose Analysis-phase markup was not captured until after the exemplar had been presented. The number of valid responses per image pair is summarized in Fig. D3.

For our analyses of corresponding minutiae, we excluded markups by five examiners who routinely did not annotate correspondences, and two markups that were missing a Comparison determination. This resulted in 3618 valid markups for analyses of corresponding minutiae (45,130 Comparison-phase minutiae marked on the latent). For some analyses, we include all minutiae marked during Analysis (including deletions) or added during Comparison (52,155 minutiae, 50,894 of which are on the 3618 markups with valid corresponding minutiae).

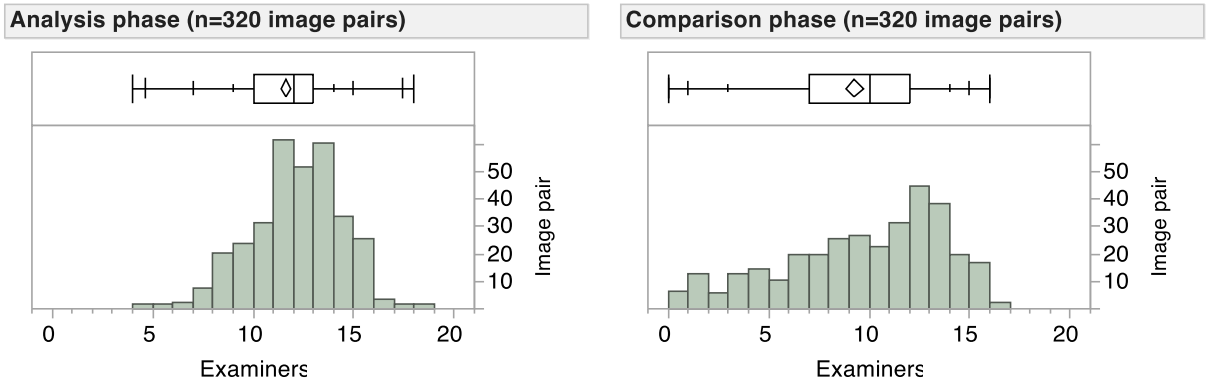


Fig. D3: Number of valid examiner markups per image pair. (Left) Analysis phase (median 12); (Right) Comparison phase (median 10). 314 image pairs were compared by one or more examiners; 271 were compared by five or more.

Data

DiB-2 Example markups

Fig. D4 shows four examples of latent-exemplar pairs (columns A-D); this expands on the examples (A and B) used in Figure 6 of [1]. Marked minutiae are shown as small black dots inside color-coded clusters. For the Analysis phase, cluster colors indicate the proportion of examiners who marked within that cluster; for the Comparison phase, colors indicate the proportion of comparing examiners who corresponded the minutia as marked on the latent. The third row of images ("Latent with Analysis minutiae") shows all minutiae as marked in the Analysis phase; the fourth row ("Latent with corresponding minutiae") shows markup from the Comparison phase limited to those minutiae that examiners marked as corresponding; the fifth row ("Exemplar with corresponding minutiae") shows the locations of the corresponding minutiae as marked on the exemplar. Because marked minutiae from one cluster on the latent did not always correspond to one cluster on the exemplar (either due to examiner disagreements or behavior of the clustering algorithm), the fifth row ("Exemplar with corresponding minutiae") uses the color-coding from the latent markup to help visualize the correspondences.

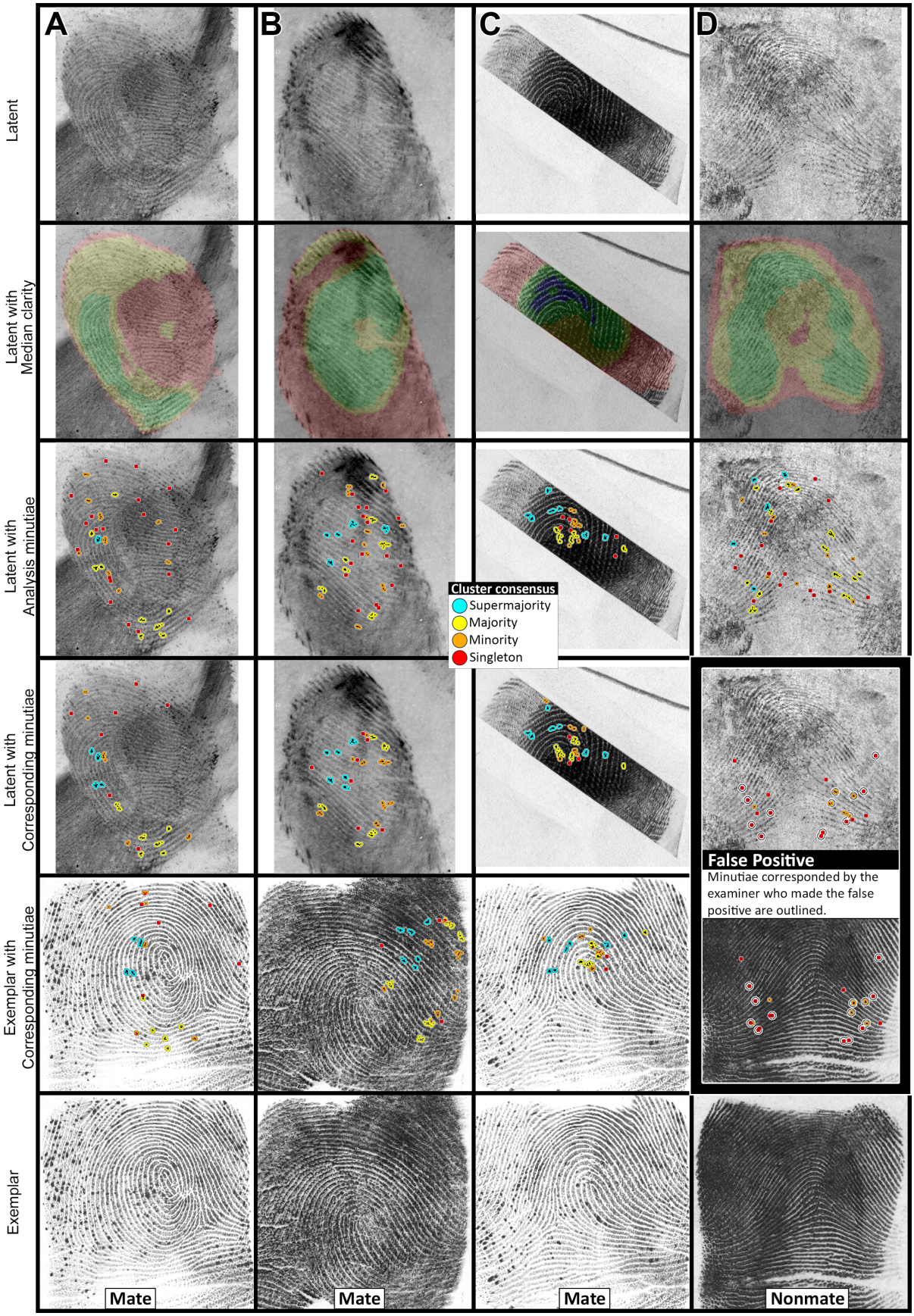


Fig. D4: Examples of markup for four comparisons (described in text).

Table D1 describes for each of the four examples shown in Fig. D4, the number of examiners contributing to the clusters, and their determinations.

	Number of Examiners								Mating
	Assigned	Value	VEO	NV	Compared	ID	Inc	Excl	
A	15	12	2	1	14	9	2	3	Mate
B	15	14	1	-	15	15	-	-	Mate
C	14	13	-	1	13	13	-	-	Mate
D	11	11	-	-	11	1	2	8	Nonmate

Table D1: Examiner determinations for the four examples shown in Fig. D4.

Note that example D is the one comparison on which an erroneous individualization occurred (also shown as an example in Figure 2 of [7]). Five examiners marked correspondences (two of whom also marked discrepancies), one additional examiner marked debatable correspondences, and one additional examiner marked discrepancies. Even after omitting the examiner who individualized, more correspondences were marked on this image pair (22, in 11 clusters) than on any other nonmated image pair in the test. Other top examples of nonmated image pairs with many correspondences marked included one with 18 correspondences (in 12 clusters, by two of ten comparing examiners), and another with 13 correspondences (in 8 clusters, by five of eight comparing examiners).

DiB-3 Effect of clustering parameters

Examiners' markups differed in whether or not individual minutiae were marked, and in the precise location where the minutiae were marked. In order to focus on whether examiners agree on the presence or absence of minutiae, we need to see past minor variations in minutia location. Neumann et al. [8] used ellipses to determine whether two minutiae should be considered the same, based on an expectation of more variation in location along the direction of the ridge than perpendicular to ridge flow; here we did not collect minutia direction, making this approach impractical. In [9], our technique of classifying features as retained, moved, added or deleted was based on a fixed radius of 0.5 mm (0.02 inch, or approximately the average inter-ridge distance) — although that approach was satisfactory for two markups where one was derived from the other, it is not well suited to comparing more than two markups.

We used automated clustering algorithms in order to classify minutiae marked by multiple examiners as representing the same minutia on the latent. Clustering was implemented in two stages as follows:

1. For each fingerprint, the set of all minutiae x,y coordinates (as marked by the examiners) was preliminarily clustered using DBSCAN with a given radius r, and no lower limit to the cluster size. That is, singletons were treated as valid clusters, not labeled as "noise."
2. Oversized preliminary clusters were split using agglomerative hierarchical clustering, with ceiling (mean number of marks per examiner) as the cutoff point. Hierarchical clustering assembles a tree of cluster relationships; there is no assumption of a fixed radius.

Neither algorithm makes use of any information from the fingerprint images themselves; they rely entirely on the x,y coordinates of the minutiae as marked by examiners. The implementation of Density-based Spatial Clustering of Applications with Noise (DBSCAN) we used was written by Michal Daszkyowski of the University of Silesia in 2004. [10,11][**] The DBSCAN radius was set to 0.015" (0.38mm) after extensively reviewing the algorithm's performance over a range of radius settings. In our review, we considered several standard clustering performance measures and visually assessed the resulting clusters as plotted superimposed over the latent prints. As shown in Fig. D5 and Table D2, any choice of radius substantially biases the reproducibility distributions: increasing the radius increases the measured mean reproducibility, and decreases the measured number of clusters. We selected a slightly large radius in order to aggregate some of the less precisely focused clusters; we then split many of the oversized clusters in the second step.

Oversized preliminary clusters were selected for subsequent splitting by agglomerative hierarchical clustering based on a criterion of (mean number of marked minutiae per examiner) > 1.5. This arbitrary threshold was selected because (1) automated splitting of clusters meeting this criterion was highly successful, and (2) for lower values (between 1 and 1.5), it was usually not apparent even to a human how to split correctly without careful interpretation of the fingerprint image. The oversized preliminary clusters often contained multiple, clearly distinct ridge events, but

[**] *The DBSCAN MATLAB source code was downloaded from http://www.chemometria.us.edu.pl/index.php?goto=downloads*

otherwise were difficult to resolve by visual inspection. We used MATLAB's implementation of agglomerative hierarchical clustering algorithm; Ward's method was selected for computing the distance between clusters.†† Ward's method helps overcome the main flaw of DBSCAN, which is that it tends to fail when faced with highly heteroskedastic data (data in which the variance differs among subsets).

Clustering was performed separately on Analysis markup (n=44,941 minutiae), Comparison markup (n=46,205 minutiae), and combined markup (n=52,155 minutiae – includes both deleted and added minutiae; limited to results presented in DiB-9 and DiB-10.2). 94% of the Analysis-phase clusters have a maximum radius less than 1mm; 99.2% less than 1.5mm; 99.95% less than 2mm.

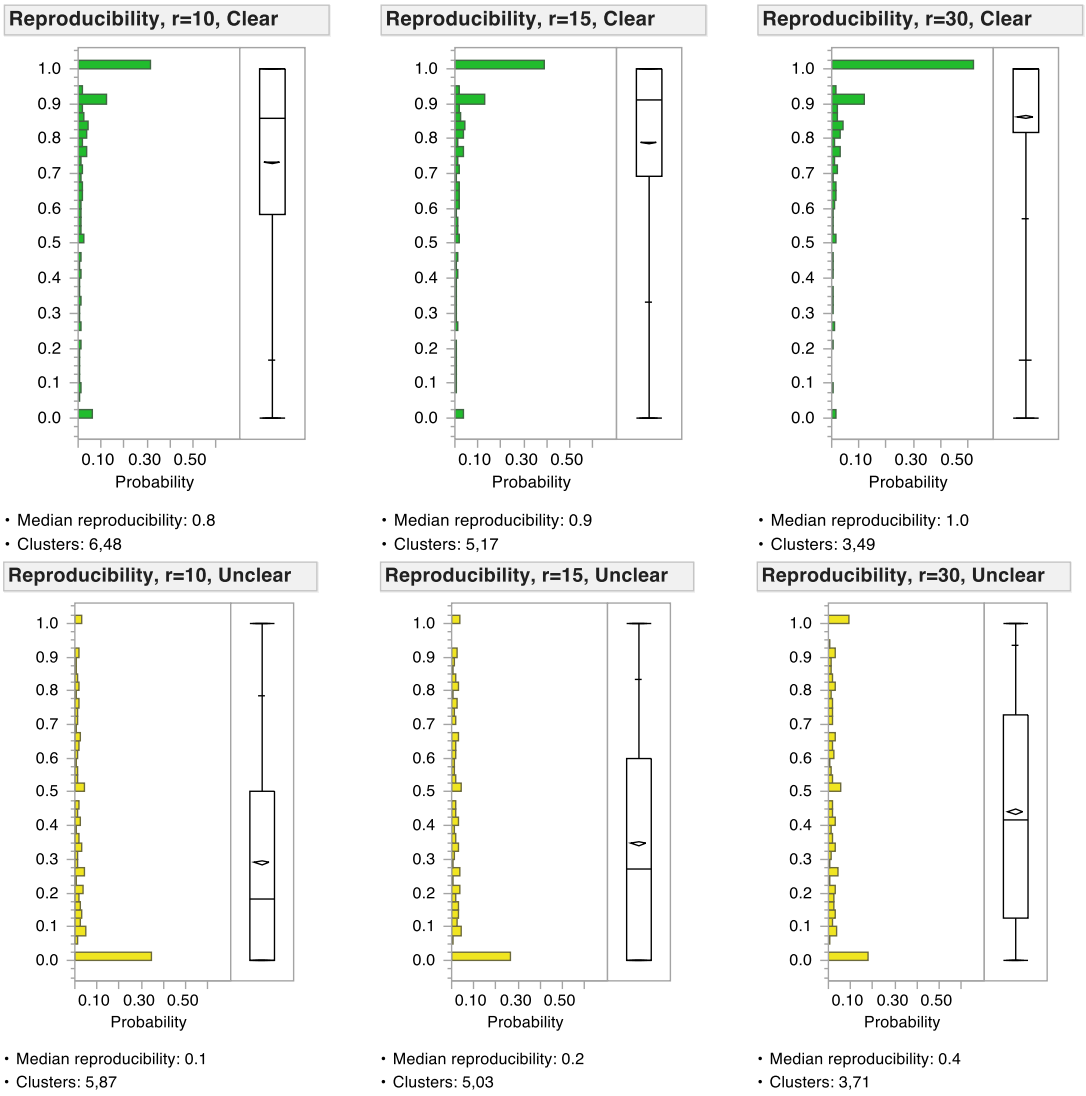


Fig. D5: Histograms showing effects of varying DBSCAN reachability distance (r = 0.010", 0.015", 0.030") on reproducibility measure. Comparison-phase minutia reproducibility distributions after DBSCAN clustering: oversized clusters were not split.

†† *MATLAB version R2014a. MATLAB's implementation of agglomerative hierarchical clustering is documented at [www.mathworks.com/help/stats/linkage.html.](www.mathworks.com/help/stats/linkage.html)*

		0.25mm (0.010")	0.38mm (0.015")	0.76mm (0.030")
Median reproducibility	Clear	86%	91%	100%
	Unclear	18%	27%	42%
# Clusters	Clear	6484	5174	3496
	Unclear	5874	5035	3711
% Singleton clusters	Clear	34%	23%	12%
	Unclear	67%	60%	49%
% Singleton minutiae	Clear	6%	3%	1%
	Unclear	34%	26%	17%

Table D2: Effects of varying DBSCAN reachability distance. Minutia reproducibility distributions after DBSCAN clustering: oversized clusters were not split. (n=46,205 minutiae)

DiB-4 Minutia reproducibility and consensus (Analysis phase)

DiB-4.1 Reproducibility and consensus by clarity

While clarity as painted by the examiners who marked the minutiae is a strong predictor of reproducibility, consensus descriptions of clarity provide a better explanation of interexaminer variation in minutiae markup (Table D3 through Fig. D8).

	Minutiae	Mean reproducibility	Median reproducibility	Mean consensus	Median consensus
Examiner clarity					
Unclear	32,159	46.9%	46.2%	N/A	N/A
Clear	12,782	69.7%	81.8%	N/A	N/A
Median clarity					
Unclear	33,846	29.8%	22.2%	19.0%	10.0%
Clear	11,095	74.1%	84.6%	51.8%	50.0%
Voted clarity					
0-10% clear	1543	10.8%	0.0%	18.4%	12.5%
10-20% clear	1780	23.3%	14.3%	29.9%	20.0%
20-30% clear	2419	26.9%	20.0%	33.1%	27.3%
30-40% clear	3022	33.3%	30.0%	39.0%	36.4%
40-50% clear	2866	44.8%	44.4%	49.4%	50.0%
50-60% clear	4297	54.4%	58.3%	58.3%	61.5%
60-70% clear	5003	63.0%	70.0%	66.1%	72.7%
70-80% clear	4755	68.8%	76.9%	71.4%	78.6%
80-90% clear	6675	77.7%	87.5%	79.7%	88.9%
90-100% clear	12,581	86.9%	92.3%	88.0%	92.9%
Overall	44,941	63.2%	75.0%	36.3%	20.0%

Table D3: Reproducibility and consensus by clarity (Analysis phase, n=44,941 minutiae; 10,324 clusters)

Minutiae that were more highly reproduced were more likely to be found in clear areas of the latent. Fig. D6 illustrates how median clarity explains this association better than examiner clarity.

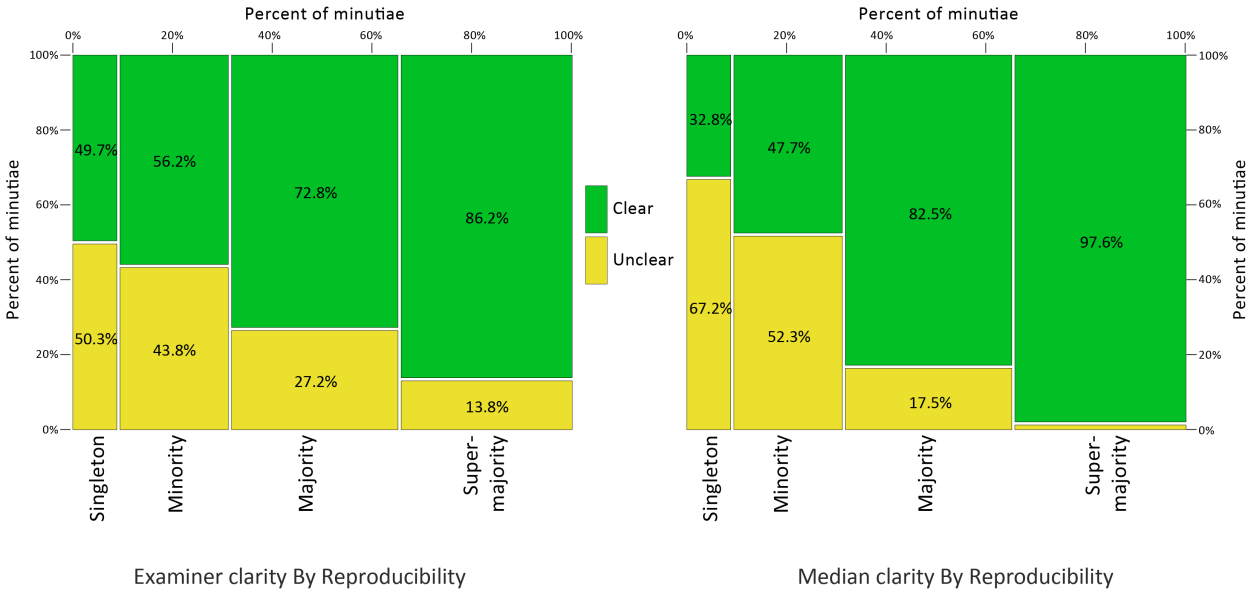


Fig. D6: Minutiae clarity by reproducibility. (Left) examiner clarity by reproducibility; (Right) median clarity by reproducibility. Cells labeled by percentages within each reproducibility level. The left figure is the same as Figure 8A in [1]. (Analysis phase, n=44,941 minutiae)

The latent prints included many areas where examiners did not agree on clarity ("debatable clarity"). Fig. D7 indicates how these areas of debatable clarity contribute to our results: Fig. D7A (reproducibility) is the same as Figure 9 in [1]; Fig. D7B shows the data in terms of cluster consensus.

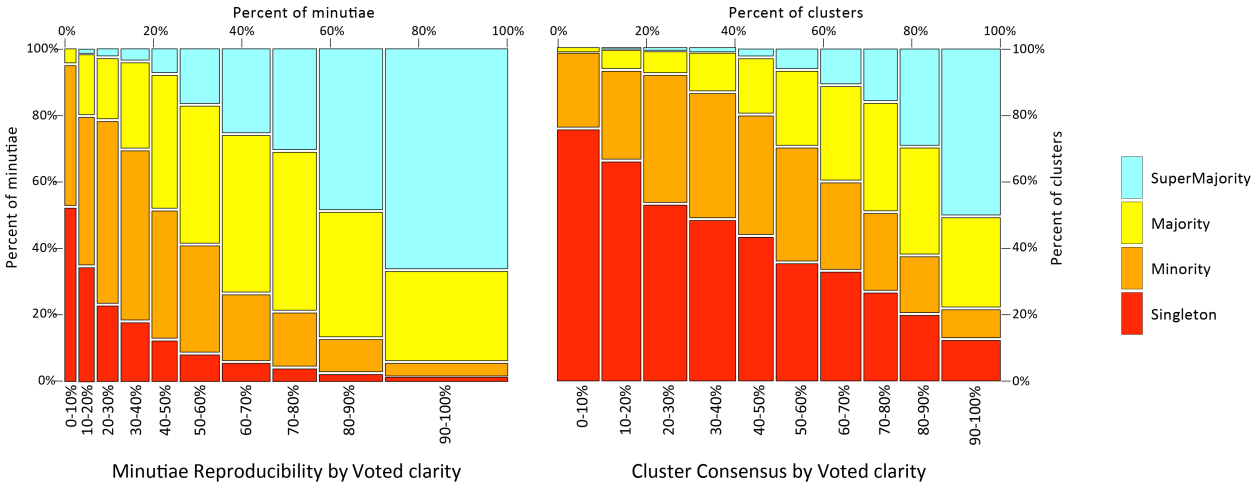


Fig. D7: (Left) reproducibility by voted clarity; (Right) consensus by voted clarity (Analysis phase, n=10,324 clusters).

Fig. D8 shows that the voted assessment of clarity is a strong predictor of minutia reproducibility: minutia reproducibility is very high when examiners concur that a location is clear, very low when examiners concur that a location is unclear, and varied when there is no concurrence on clarity. This can explain some of the lack of association seen in Fig. D7.

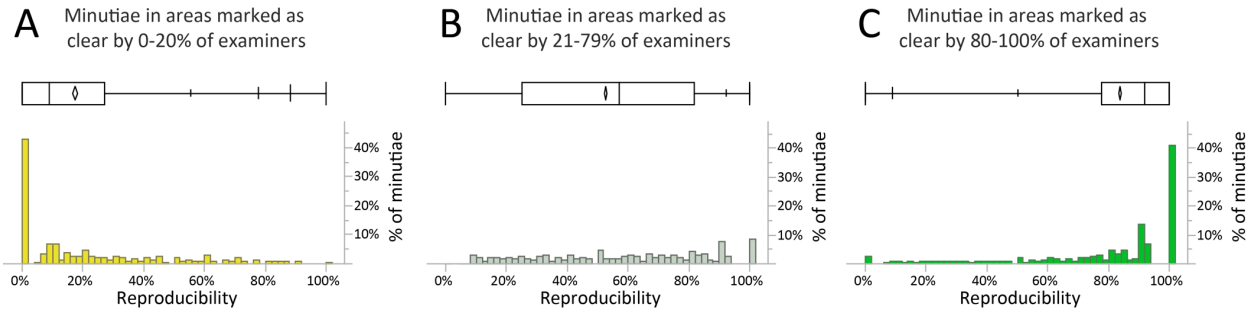


Fig. D8: Reproducibility by voted clarity in areas (A) that examiners agree are unclear; (B) where examiners do not agree on clarity; (C) that examiners agree are clear. (Analysis phase, n=44,941 minutiae). Mean reproducibility = (A) 17%; (B) 53%; (C) 84%.

DiB-4.2 Reproducibility of entire markups

In addition to assessing interexaminer variability by minutiae (reproducibility) and by clusters (consensus), we can assess variability by entire markups, as shown in Table D4.

Any clear minutiae	Any majority clusters	Markups	"Perfect" agreement		90% agreement		75% agreement	
Yes	Yes	2897	230	(8%)	479	(17%)	1462	(50%)
	No	18	0	(0%)	0	(0%)	0	(0%)
No	Yes	691	194	(28%)	220	(32%)	365	(53%)
	No	124	124	(100%)	124	(100%)	124	(100%)
Total		3730	548	(15%)	823	(22%)	1951	(52%)

Table D4: "Perfect" agreement counts those Analysis-phase markups in which (1) all minutiae that the examiner marked in clear areas were in majority clusters and (2) the examiner marked in all majority clusters (in any clarity). The 90% and 75% agreement columns require that at least 90% (75%) of the minutia that the examiner marked in clear areas were in majority clusters and the examiner marked at least 90% (75%) of the majority clusters.

DiB-4.3 Singletons and solo misses

As shown in Table D5, with a mean of 12 examiners per latent, 50% of the Analysis-phase markups had singletons. 15% of all markups had more than two singletons, and these markups accounted for 59% of all singletons. 6.6% of examiner clear minutiae were singletons; 16.8% of examiner unclear minutiae were singletons.

Category	Markups	Singletons	% markups	% singletons
No singletons	1883	0	50%	0%
1 or 2 singletons	1299	1761	35%	41%
>2 singletons	548	2508	15%	59%
Total	3730	4269	100%	100%

Table D5: Distribution of singletons per markup (Analysis phase, mean of 12 examiners per latent).

Analogous to singletons are "solo misses," i.e., minutiae that were marked by all but one of the examiners. Unlike singletons, solo misses occur primarily in clear areas: there were a total of 640 solo misses during Analysis (6% of clusters), 610 of which were in median clear areas. Although singletons are far more numerous than solo misses, solo misses disproportionately affect measures such as mean reproducibility, because reproducibility counts each singleton once (as reproducibility = 0) while it counts solo misses once for each examiner who marked that minutia (e.g., as mean reproducibility = 92% if 11 of 12 examiners marked a minutia).

DiB-4.4 Reproducibility of minutia with respect to value determinations

Minutia reproducibility tended to be higher on latents that examiners agreed are VID than those that examiners agreed are not VID. However, as shown in Fig. D9, most of this association can be accounted for in terms of differences in clarity: those latents that examiners agreed are VID tend to have more minutiae marked in clear areas.

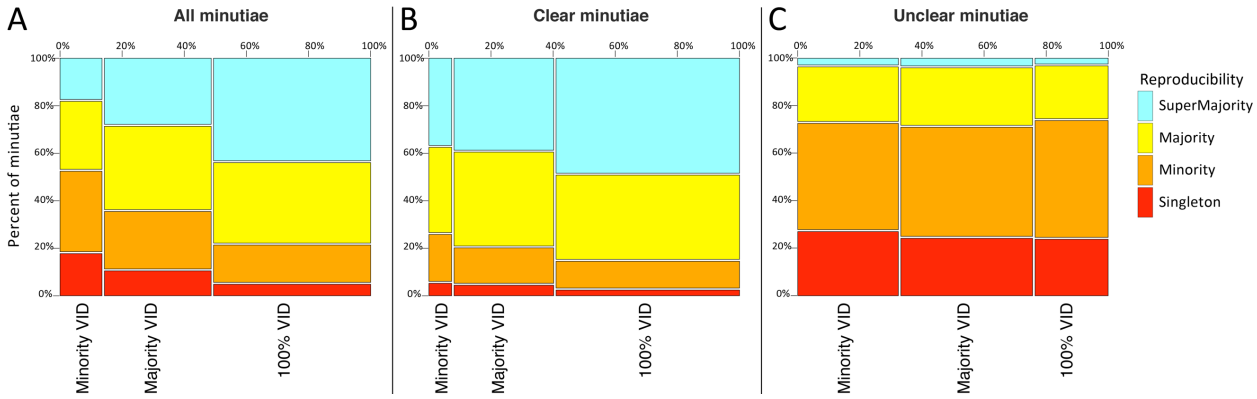


Fig. D9: Association between latent value determinations and reproducibility. (A) all minutiae (Analysis phase, n=44,941 minutiae); (B) median clear minutiae (n=33,846 minutiae); (C) median unclear minutiae (n=11,095 minutiae).

We have previously reported [2,7] that when one examiner assesses a latent to be VID and another examiner assesses that same latent to be NV, the examiner assessing the latent to be VID can be expected to mark more minutiae. Here we take a closer look at how differences in value assessments relate to whether examiners mark specific minutiae.

The following logistic regression model was used to estimate the probability that an examiner would mark a minutia given the level of consensus for that minutia and the examiner's value assessment. This model allows us to estimate how much effect is specifically associated with the value assessments as opposed to other factors such as clarity or which regions of the prints examiners chose to mark that are largely accounted for by conditioning on consensus:

$$\text{logit}(\pi) = \beta_0 + \beta_{\text{Value}}*\text{Value} + \beta_{\text{Consensus}}*\text{Consensus}, \qquad \text{(Eq 1)}$$

where π is the probability that this examiner marked the minutia given this examiner's value assessment of the latent and given the proportion of all examiners who marked this minutia. The results are summarized in Table D6.

Consensus	P(marking|NV)	P(marking|VEO)	P(marking|VID)
0.1	0.049	0.071	0.122
0.5	0.323	0.412	0.560
0.9	0.814	0.865	0.921

Table D6: Probability of marking minutiae conditioned on the examiner's value assessment (Analysis phase, n=10,324 clusters).

Table D6 shows that, even after accounting for the level of consensus on each minutia, examiners are more likely to mark minutiae when they assess a latent to be VID. Of course, the decisions to mark or not mark minutiae on a single latent are not independent events. For example, examiners occasionally mark no minutiae on latents assessed to be NV or VEO; this may contribute to the lower probability of examiners marking minutiae in majority clusters on these responses. Taking this lack of independence into account, we realize that conditioning on the level of consensus does not completely remove the confounding effects of factors such as clarity. Fig. D10 and Fig. D11 show that when examiners assessed latents to be VID, they almost always marked most of the majority clusters; when they assessed latents to be NV or VEO, they often marked fewer than half of the majority clusters.

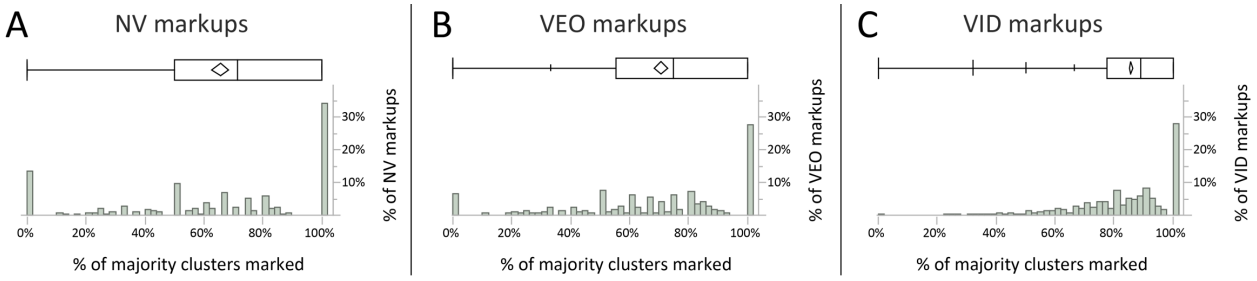


Fig. D10: Percentage of majority clusters marked, conditioned on value assessment (Analysis phase, n=3588 markups = (A) 602 NV + (B) 570 VEO + (C) 2416 VID; 142 of the 3730 markups had no majority clusters)

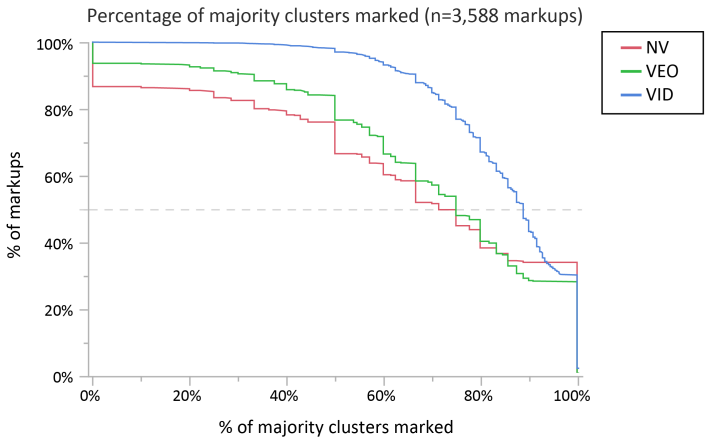


Fig. D11: Cumulative distribution functions of the percentage of majority clusters marked, conditioned on value assessment (same data as Fig. D10). The median number of majority clusters marked (dashed line) was 71% of NVs; 75% of VEOs; 89% of VIDs. No majority clusters were marked (left extreme) on 13% NV latents; 6% of VEO latents; and 0% of VID latents. All majority clusters were marked (right extreme) on 34% NVs; 27% VEOs; and 28% VIDs.

Table D7 and Table D8 summarize Analysis-phase reproducibility by latent value assessment and clarity.

	Mean reproducibility			Median reproducibility		
	Clear	**Unclear**	**Overall**	**Clear**	**Unclear**	**Overall**
All	0.697	0.469	0.632	0.818	0.462	0.750
VID	0.705	0.469	0.646	0.833	0.462	0.750
VEO	0.614	0.450	0.541	0.733	0.455	0.600
NV	0.655	0.490	0.568	0.750	0.500	0.636

Table D7: Mean and median reproducibility of minutiae by ***examiner clarity*** and latent value assessment (Analysis phase, n=44,941 minutiae).

	Mean reproducibility			Median reproducibility		
	Clear	**Unclear**	**Overall**	**Clear**	**Unclear**	**Overall**
All	0.741	0.298	0.632	0.846	0.222	0.750
VID	0.743	0.287	0.646	0.846	0.214	0.750
VEO	0.725	0.304	0.541	0.833	0.222	0.600
NV	0.742	0.369	0.568	0.846	0.357	0.636

Table D8: Mean and median reproducibility of minutiae by ***median clarity*** and latent value assessment (Analysis phase, n=44,941 minutiae).

DiB-5 Reproducibility of nonminutia features

Reproducibility of cores and deltas was low, and never unanimous (Fig. D12). Examiners were instructed to mark all cores and deltas on the latents, provided they could be located within approximately three ridge intervals. On those latents that had one or more cores or deltas marked by any examiners, typically only about half of the examiners marked them: no cores or deltas were unanimously marked.

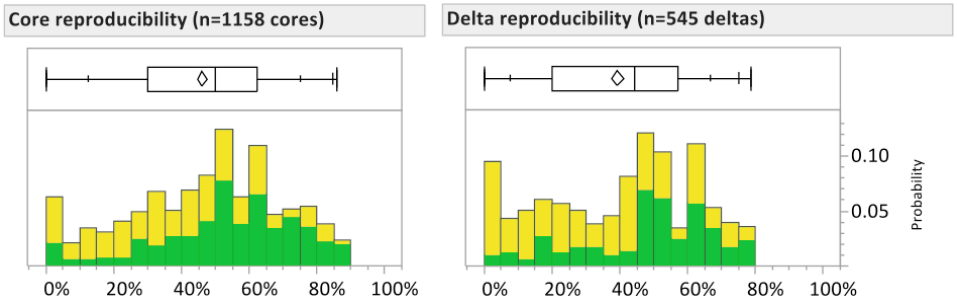


Fig. D12: Reproducibility of cores and deltas, Analysis phase. Here we gauge reproducibility based on a 1.5mm (0.06") radius (corresponding to our instructions that cores and deltas could be located within approximately three ridge intervals). Data is color-coded by examiner clarity: green=clear, yellow=unclear.

Features other than minutiae were sometimes present in or near minutia clusters, which could indicate a disagreement as to whether a feature should be marked as a minutia, a nonminutia feature, or both. However, this did not explain much of the interexaminer variability: only 4.5% of clusters contained features other than minutiae (Table D9).

	Features	Features in clusters		Clusters with nonminutia features	
Cores	1269	519	40.9%	174	1.7%
Deltas	621	180	29.0%	78	0.8%
Other nonminutia features	703	320	45.5%	223	2.1%
Total nonminutia features	2593	1019	39.3%	465	4.5%

Table D9: Prevalence of nonminutia features in the area of minutia clusters (Comparison phase, n=10,398 clusters). Here we consider a nonminutia feature as being in a minutia cluster if it is within 0.38mm (0.015”) of the cluster center. We report Comparison-phase counts because examiners were only instructed to mark “other” features during Comparison.

DiB-6 Agreement in clarity markup (Analysis phase)

Examiners often disagreed as to whether or not minutiae were present and as to whether the locations of minutiae were sufficiently clear to be certain of the presence or absence of minutiae.

Table D10 and Fig. D13 show for every minutia (n=44,941) the distribution of clarity assigned to that location by other examiners, regardless of whether the other examiners marked a minutia at that location. When an examiner marked a minutia in an area that that examiner described as unclear, other examiners were about equally likely to describe that area as clear or unclear.

Table D11 and Fig. D14 show for every cluster center (n=10,324) the distribution of clarity assigned to that location by pairs of examiners, regardless of whether those examiners marked a minutia at that location. Selecting examiner pairs and cluster centers at random, the probability of the two examiners agreeing whether to describe that location as clear vs. unclear was 65%.

Table D12 shows for every minutia marked (n=44,941) the distribution of clarity assigned to that location by other examiners, conditioned by whether the second examiner marked at that location. When a second examiner agreed on the presence of a minutia, that examiner was much more likely to describe the location as clear, whereas if the second examiner did not mark the minutia, that examiner was likely to describe the location as unclear.

Minutiae			Examiner B						Total minutiae
			Unclear			Clear			
			Black	Red	Yellow	Green	Blue	Aqua	
Examiner A	Unclear	Black	60	55	206	434	87	22	863
		Red	41	158	447	357	49	5	1056
		Yellow	324	1026	4258	4505	653	93	10,859
	Clear	Green	656	956	5858	14,608	3111	565	25,754
		Blue	119	86	701	3060	1085	220	5271
		Aqua	35	9	102	569	222	201	1138

Minutiae		Examiner B		Total minutiae
		Unclear	Clear	
Examiner A	Unclear	6574 (51%)	6204 (49%)	12,778
	Clear	8522 (26%)	23,641 (74%)	32,163

Table D10: Examiner B clarity by examiner A clarity for each ***minutia*** marked by examiner A. Data is constructed from all pairs of examiners on each latent; each minutia marked by examiner A is equally weighted (Analysis phase, n=44,941 minutiae). The tables summarize the clarity examiner B assigned to each location without regard to whether examiner B marked a minutia at that location.

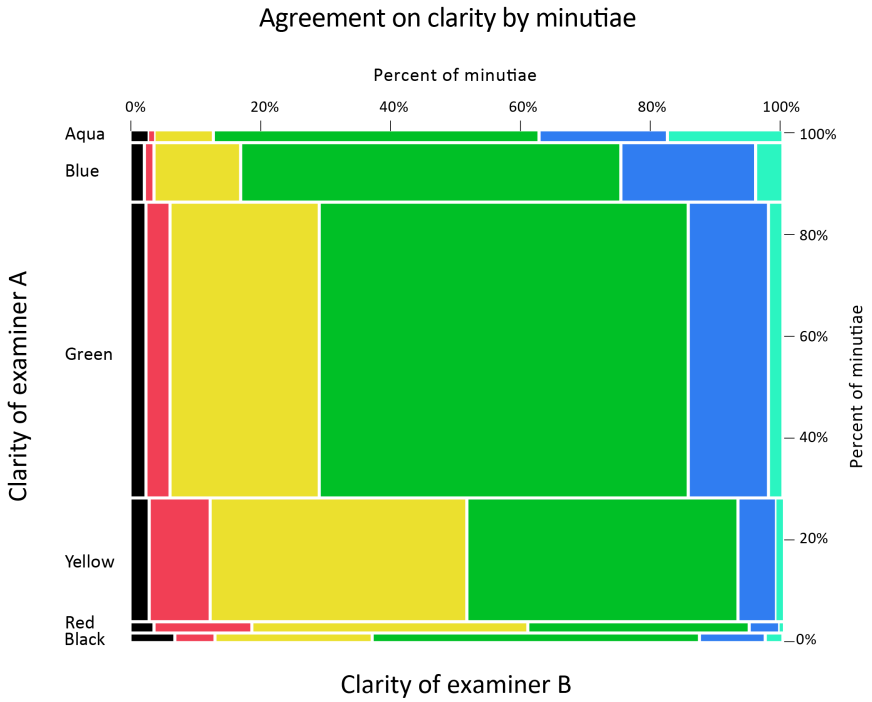


Fig. D13: Examiner B clarity by examiner A clarity for each ***minutia*** marked by examiner A. Same data as Table D10, shown graphically.

Clusters			Examiner B						Total clusters
			Unclear			Clear			
			Black	Red	Yellow	Green	Blue	Aqua	
Examiner A	Unclear	Black	86	57	127	124	21	5	420
		Red	57	238	484	233	25	2	1039
		Yellow	127	484	1648	1228	150	19	3657
	Clear	Green	124	233	1228	2216	418	71	4292
		Blue	21	25	150	418	129	26	770
		Aqua	5	2	19	71	26	23	147

Clusters		Examiner B		Total clusters
		Unclear	Clear	
Examiner A	Unclear	3308 (65%)	1808 (35%)	5116
	Clear	1808 (35%)	3400 (65%)	5208

Table D11: Examiner B clarity by examiner A clarity at each ***cluster*** center. Data is constructed from all pairs of examiners on each latent regardless of whether the examiners marked in the cluster; each cluster is weighted equally (n=10,324 clusters). The tables summarize the clarity examiners assigned to each cluster without regard to whether those examiners marked a minutia in the cluster.

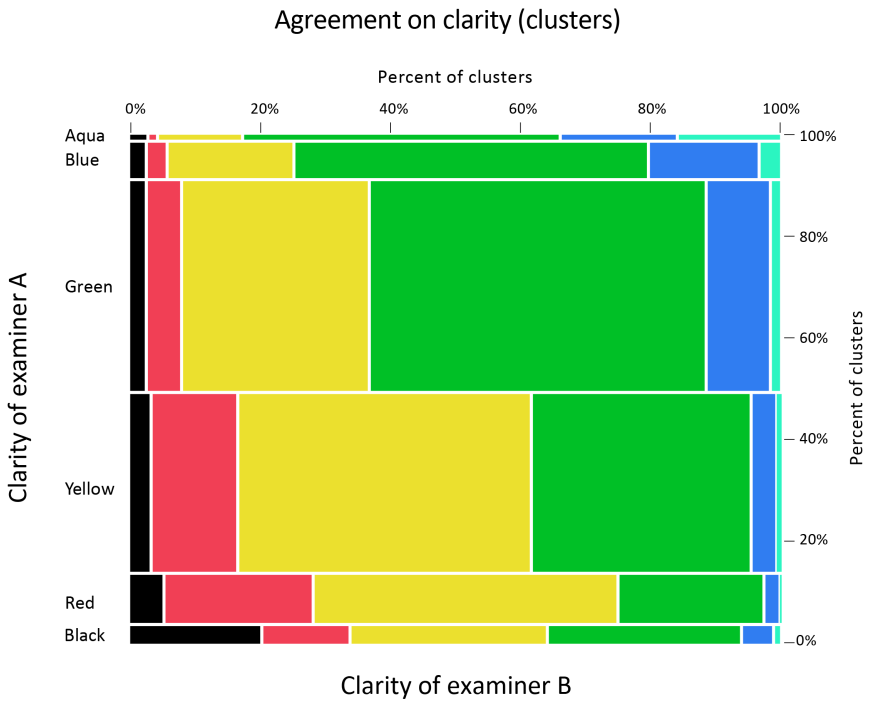


Fig. D14: Examiner B clarity by examiner A clarity at each ***cluster*** center. Same data as Table D11, shown graphically.

Minutiae		B marked			B not marked			Total minutiae
		Unclear	Clear	Subtotal	Unclear	Clear	Subtotal	
Examiner A	Unclear	2127 (35%)	4014 (65%)	6141	4384 (66%)	2253 (34%)	6637	12,778
	Clear	4016 (18%)	18,590 (82%)	22,606	4448 (47%)	5109 (53%)	9557	32,163

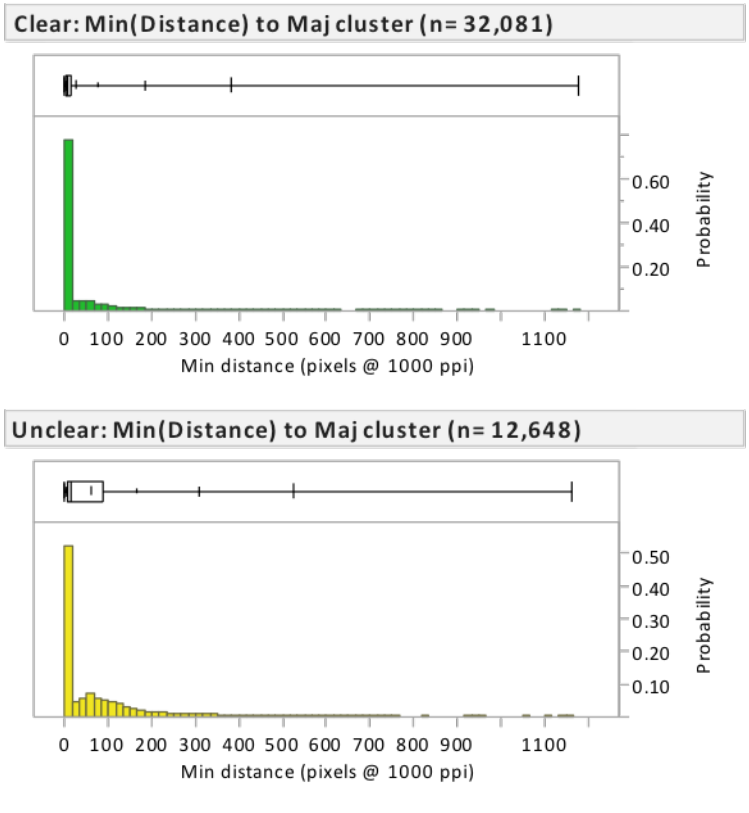
Table D12: Examiner B clarity by examiner A clarity for each minutia marked by examiner A, conditioned by whether examiner B marked a minutia at that location. Data constructed from all pairs of examiners on each latent; each minutia marked by examiner A is equally weighted (n=44,941 Analysis-phase minutiae).

DiB-7 Differences in regions with marked minutiae

Some examiners mark minutiae far away from those marked by other examiners. This may be due to disagreements regarding the boundaries of the impression being considered (i.e., the region of interest), or disagreements on which areas in the region of interest are of sufficient quality to mark minutiae. Table D13 describes what proportion of minutiae were marked far from the nearest majority cluster. Fig. D15 (Analysis phase) and Fig. D16 (corresponding minutiae, Comparison phase) show the distributions of the distances from marked minutiae to the nearest majority cluster.

		Minutiae	Relatively far (Distance > 0.1")		Very far (Distance > 0.2")	
			Minutiae	%	Minutiae	%
Marked minutiae (Analysis phase)	Total	44,729	5006	11.2%	1581	3.5%
	Examiner Clear	32,081	2250	7.0%	701	2.2%
	Examiner Unclear	12,648	2756	21.8%	880	7.0%
	Median Clear	33,840	1094	3.2%	176	0.5%
	Median Unclear	10,889	3912	35.9%	1405	12.9%
Corresponding minutiae (Comparison phase)	Total	27,486	2277	8.3%	632	2.3%
	Examiner Clear	20,271	1110	5.5%	317	1.6%
	Examiner Unclear	7215	1167	16.2%	315	4.4%

Table D13: Percentage of minutiae that are "relatively far" (more than 0.1", about 5 ridge intervals on average) or "very far" (more than 0.2", about 10 ridge intervals) from the nearest majority cluster, by phase and minutia clarity. The total minutia count is limited to latents that had at least one majority cluster. For corresponding minutiae, distance is measured to the nearest cluster that was marked and corresponded by a majority of comparing examiners. (Analysis phase, n=44,729; another 212 minutiae were marked on latents having no majority clusters).



Clear

100.0%	maximum	1177.31
99.5%		381.869
97.5%		184.434
90.0%		76.6551
75.0%	quartile	13.6015
50.0%	median	5.65685
25.0%	quartile	3
10.0%		1.41421
2.5%		1
0.5%		0
0.0%	minimum	0

Unclear

100.0%	maximum	1163.85
99.5%		524.547
97.5%		307.628
90.0%		164.63
75.0%	quartile	88.4081
50.0%	median	13.4536
25.0%	quartile	4.47214
10.0%		2.23607
2.5%		1
0.5%		0
0.0%	minimum	0

Fig. D15: Distance of ***Analysis-phase*** minutiae to nearest majority cluster by examiner clarity. Distance is measured in units of 0.001". (Analysis phase, n=44,729; another 212 minutiae were marked on latents having no majority clusters).

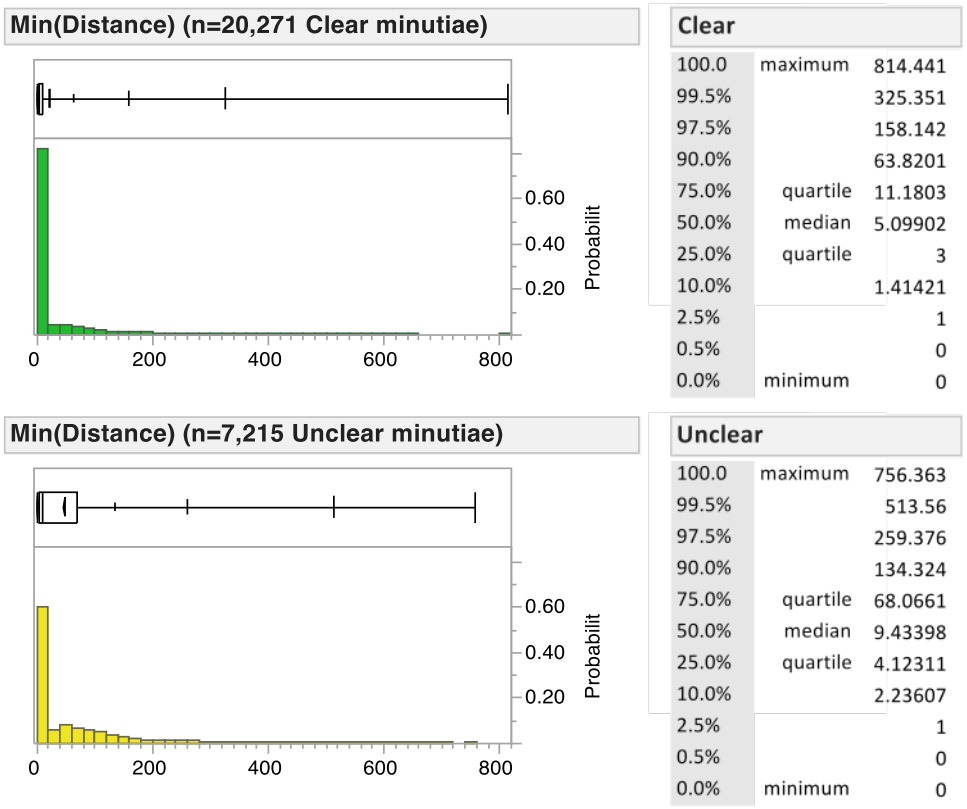


Fig. D16: Distance of ***corresponding*** minutiae to the nearest cluster corresponded by a majority of comparing examiners, by examiner latent clarity. Distance is measured in units of 0.001". The set of majority clusters was limited to those in which at least three examiners marked corresponding minutiae; "majority" was calculated among those examiners who marked at least one correspondence on the image pair. (Comparison phase, n=27,486; another 454 corresponding minutiae were marked on latents having no majority cluster).

DiB-8 Consensus and sufficiency (Analysis and Comparison phases)

Previously, we reported [2] that the number of minutiae annotated by examiners is strongly associated with their own value and comparison determinations, and that seven minutiae was an approximate "tipping point": "for any minutia count greater than seven, the majority of value determinations were VID, and for any corresponding minutia count greater than seven, the majority of comparison determinations were individualization." Across multiple examiners, a *mean* of seven corresponding minutiae was also the point at which approximately 50% of examiners individualized (approximately 50% of examiners assessed latents to be VID when the mean minutia count was seven).

Here we report similar thresholds as measured by consensus on minutia clusters. We find counts of majority clusters comparable to mean minutia counts as predictors of examiner determinations. For example, when predicting VID determinations using logistic regression, $r^2 = 0.4253$ for mean minutia counts vs. $r^2 = 0.4310$ for majority clusters. As shown in Fig. D17, these majority cluster statistics are highly correlated with the mean number of minutiae, which tends to be slightly larger than the number of majority clusters.

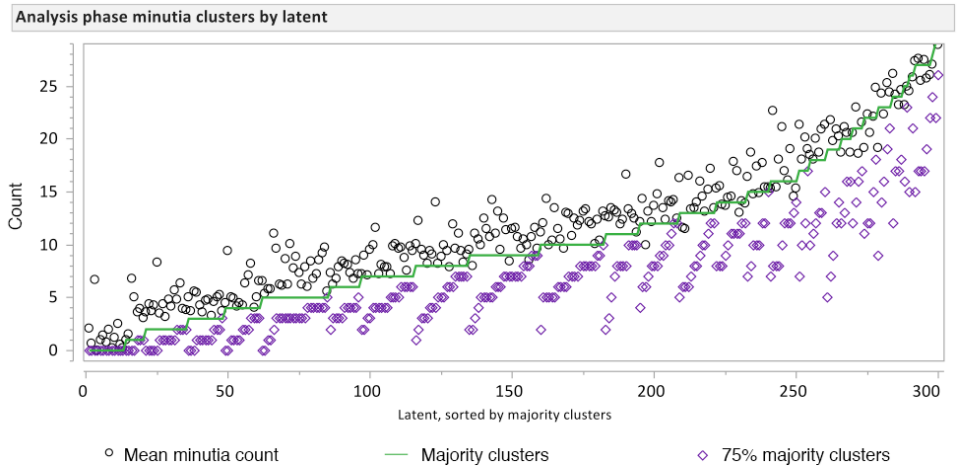


Fig. D17: Relation among mean minutia counts and majority clusters (Analysis phase, n=301 latents). Latents (x-axis) are sorted by the number of majority clusters. Shows the mean minutia count (black), number of majority clusters (green), and number of clusters marked by at least 75% of examiners (purple).

As shown in Fig. D18 and Fig. D19A, latents with fewer than 5 majority clusters were usually not assessed as VID; latents with 10 or more majority clusters were usually assessed to be VID.

Fig. D19B shows similar results for clusters corresponded by the majority of comparing examiners: almost all image pairs with 7 or more clusters that were corresponded by a majority of comparing examiners were individualized by the majority of examiners; almost no image pairs with 5 or fewer majority corresponding clusters were individualized by the majority of examiners.

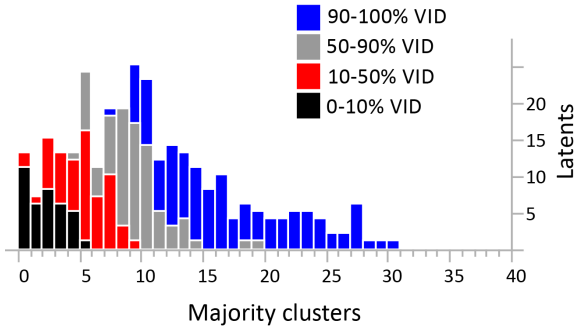


Fig. D18: Distribution of the number of majority clusters in latents, shaded to indicate percentages of examiners who assessed each latent as VID (n=301 latents). Overall distribution reflects data selection for the test.

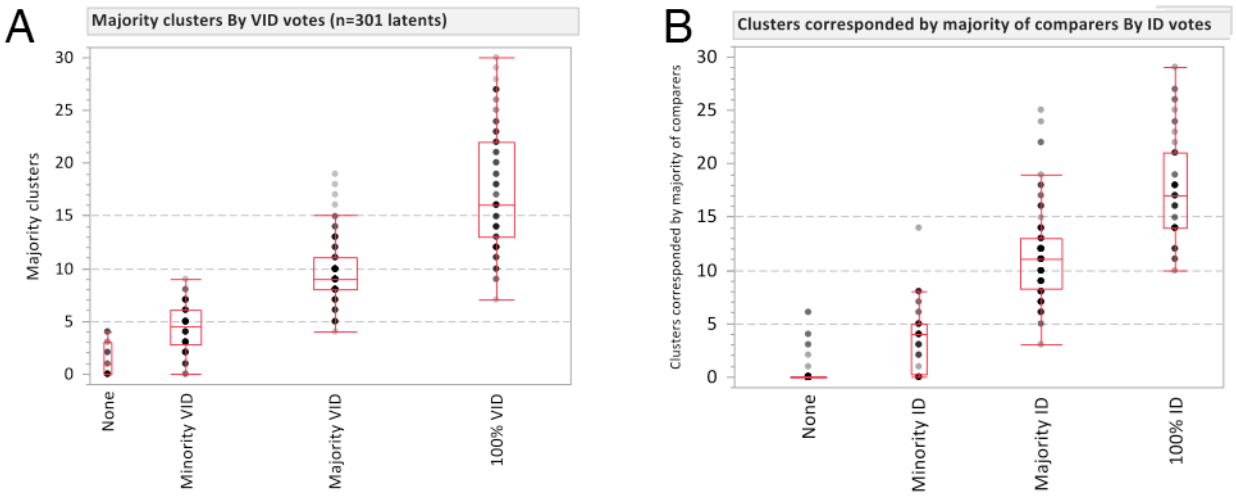


Fig. D19: Majority minutia clusters by proportion of examiners determining (A) value for individualization (n=301 latents), (B) individualization (n=271 image pairs). Y-axis in chart B is the number of clusters corresponded by a majority of comparers: (number of corresponding examiners / number of comparing examiners) ≥ 0.5. Data excludes image pairs with fewer than five Comparison markups. One data point at y=65 (100% ID) not shown in (A). One data point at y=42 (100% ID) not shown in (B).

In [2] we included several figures to show the association between minutia counts and value determinations, and between corresponding minutia counts and comparison determinations. Fig. D20 is comparable to Figure 5 of [2] except that it includes a data series for the number of clusters corresponded by a majority of examiners who compared

the image pair; it also includes data for both mated and nonmated image pairs. In general, the number of majority clusters tends to be approximately equal to the mean minutia count.

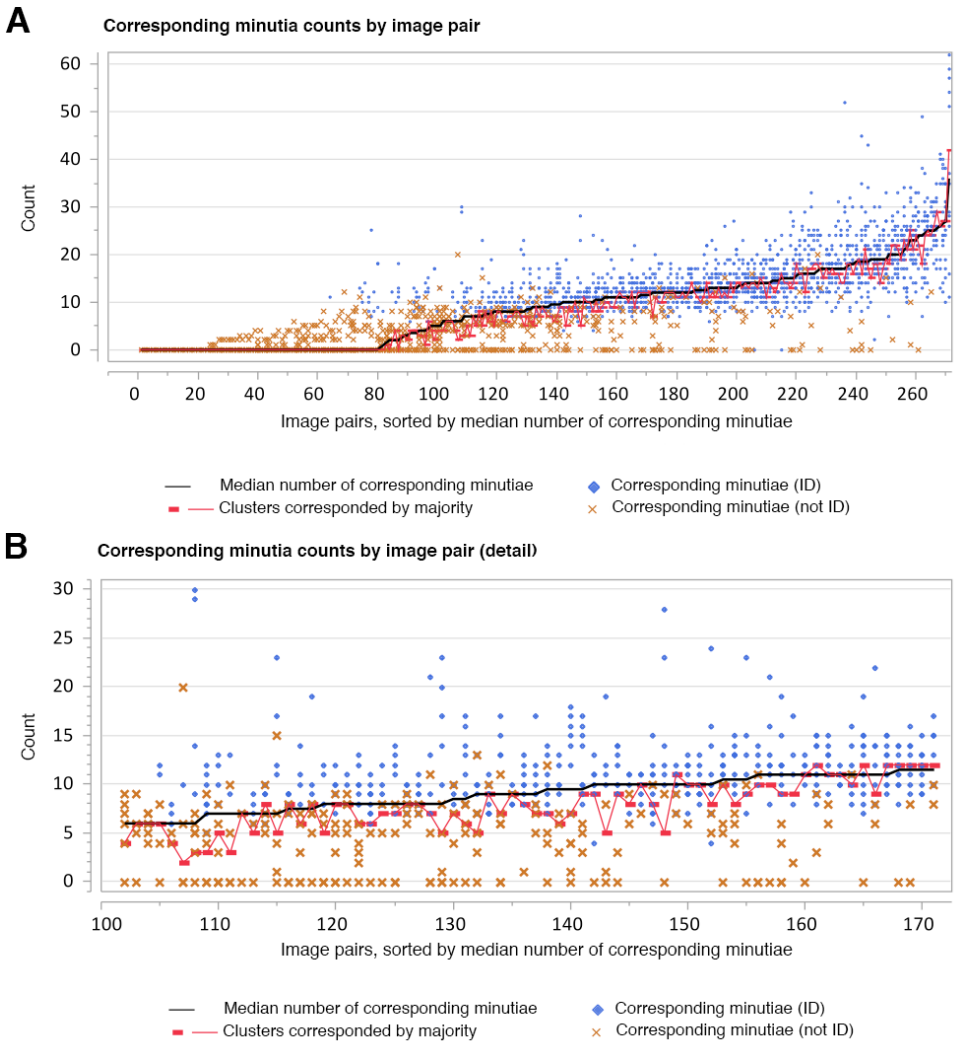


Fig. D20: Corresponding minutiae counts by image pair: median corresponding minutiae (black line); clusters corresponded by a majority of comparing examiners (red rectangle); counts by examiners who individualized (blue diamond); counts by examiners who did not individualize (orange x). (A) 271 image pairs compared by at least 5 examiners; (B) a subset of the data enlarged to reveal interexaminer variability on 70 image pairs having 6-10 median corresponding minutiae.

DiB-9 Reproducibility of Analysis-Comparison changes

As previously reported, examiners often modified their latent Analysis markup during the Comparison phase [7]. For each pair of latent markups (Analysis and Comparison phases), we classified features as retained, moved, deleted, or added. A retained feature is one that is present at exactly the same pixel location in both markups; a moved feature refers to one that was deleted during Comparison and replaced by another within 0.5 mm (approximately one ridge width); a deleted feature is one that was present in the Analysis markup only (no Comparison feature within 0.5 mm); an added feature is one that was present in the Comparison markup only (no Analysis feature within 0.5mm). Fig. D21 summarizes the extent of such changes, by clarity, showing that unclear minutiae were much more likely to be changed.

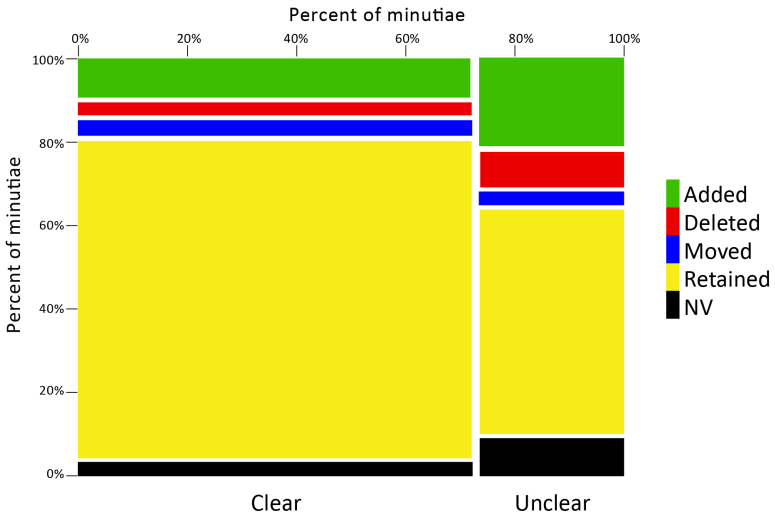


Fig. D21: Analysis-Comparison changes by examiner clarity. Chart represents all 52,155 minutiae marked during either the Analysis or Comparison phases.

Table D14 and Table D15 show that deleted and added minutiae are strongly associated with low reproducibility. This association is stronger in clear areas than unclear areas: using logistic regression to predict deletions and additions from minutia reproducibility, we find that for deleted minutiae, $r^2 = 0.1243$ (clear) and 0.0686 (unclear); for added minutiae, $r^2 = 0.0640$ (clear) and 0.0332 (unclear).

Having shown in this report that reproducibility and clarity are strongly associated, we now take a closer look at how reproducibility and clarity are associated with changes. We used logistic regression to model deleted and added minutiae as responses to reproducibility and clarity. Predicting deleted minutiae from reproducibility and examiner clarity ($r^2 = 0.1114$), only the reproducibility term is significant; clarity provides no additional information (using median clarity makes no meaningful improvement to the model: $r^2 = 0.1116$). Predicting added minutiae from reproducibility and examiner clarity ($r^2 = 0.0762$), both terms are significant, though the reproducibility term contributes much more than clarity (predicting added minutiae from reproducibility alone results in $r^2 = 0.0682$; from examiner clarity alone, $r^2 = 0.0271$; from median clarity alone, $r^2 = 0.0359$). Examiners are more likely to add minutiae in low-clarity areas even after accounting for reproducibility of those minutiae. Our ability to predict deleted minutiae is not further improved by knowing clarity after accounting for reproducibility.

Clarity	Reproducibility	Retained	Moved	Deleted	% Deleted
Clear	SuperMajority	11,953	701	236	1.8%
	Majority	9555	667	475	4.4%
	Minority	4274	361	646	12.2%
	Singleton	1410	108	515	25.3%
Unclear	SuperMajority	1707	132	53	2.8%
	Majority	3201	261	207	5.6%
	Minority	3203	230	448	11.5%
	Singleton	1439	82	415	21.4%
All	SuperMajority	13,660	833	289	2.0%
	Majority	12,756	928	682	4.7%
	Minority	7477	591	1094	11.9%
	Singleton	2849	190	930	23.4%

Table D14: Reproducibility of ***Analysis*** minutiae by clarity and change type (n=42,279 Analysis-phase minutiae). Data are limited to 3709 responses on 320 image pairs, which excludes 31 markups with data collection problems (detailed in [7]).

Clarity	ReproCategory	Retained	Moved	Added	% Added
Clear	SuperMajority	12,675	714	768	5.4%
	Majority	9095	686	1449	12.9%
	Minority	3966	303	1229	22.4%
	Singleton	1346	100	506	25.9%
Unclear	SuperMajority	1590	157	237	11.9%
	Majority	3198	289	933	21.1%
	Minority	3031	209	1380	29.9%
	Singleton	1443	73	742	32.9%
All	SuperMajority	14,265	871	1005	6.2%
	Majority	12,293	975	2382	15.2%
	Minority	6997	512	2609	25.8%
	Singleton	2789	173	1248	29.6%

Table D15: Reproducibility of ***Comparison*** minutiae by clarity and change type (n=46,119 Comparison-phase minutiae). Data are limited to 2957 comparisons of 313 image pairs, which excludes markups where either the latent or exemplar was assessed to be NV and some data collection problems (detailed in [7]).

The net effect on minutia reproducibility was to increase from the Analysis to Comparison phase, but only for those latents compared to mated exemplars (not for those compared to nonmated exemplars). Fig. D22 shows this effect on a subset of 19 latents, each of which was assigned in both mated and nonmated image pairs; this subset controls for any differences in how latents were selected for the mated and nonmated pairs. Minutia reproducibility for mated pairs increased in both clear and unclear areas. These results are generally representative of what was observed across all latents. For further discussion of how changes in markup relate to whether or not the exemplar was mated, see [7].

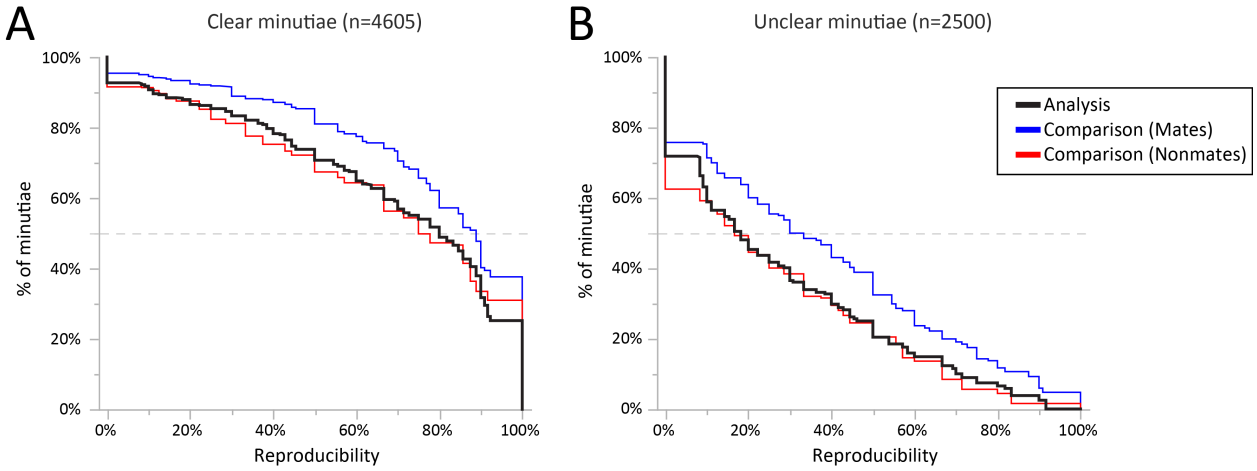
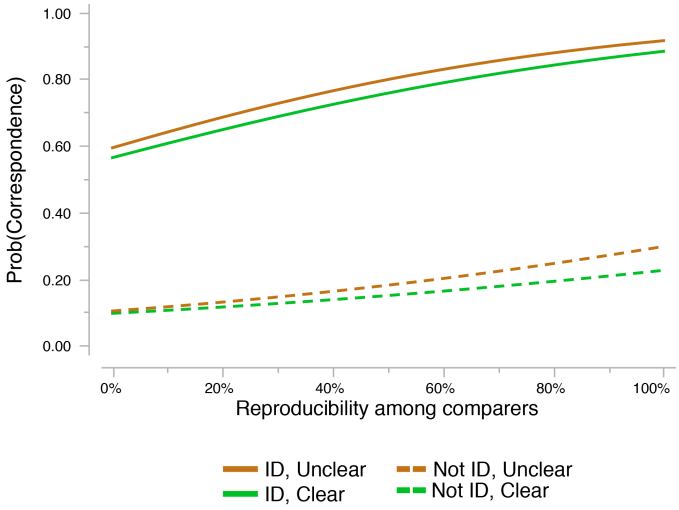


Fig. D22: Minutia reproducibility in Analysis to Comparison phases, by median clarity. Y-axis indicates the percentage of minutiae that meet or exceed the x-axis reproducibility level. Data is limited to 19 latents that were presented to examiners in both mated and nonmated pairings: 302 markups (179 mated, 173 nonmated) where the examiner proceeded to Comparison (latent was not assessed NV). On the mated pairs, median reproducibility (dashed line) increased in clear areas from 82% (A, black curve) to 89% (A, blue curve), and in unclear areas increased from 20% (B, black curve) to 32% (B, blue curve). On mated pairs, the percentage of minutiae marked by all examiners (unanimously marked) increased from 23% to 38% in median Clear areas (A, compare black and blue lines at reproducibility = 100%).

DiB-10 Corresponding minutiae

DiB-10.1 Probability of correspondence

The probability of examiners corresponding marked minutiae was correlated with the reproducibility of those minutiae. Fig. D23 shows the probability of examiners corresponding minutiae as estimated by four logistic regression models, one for each combination of clarity (as marked by that examiner) and whether the examiner individualized.



Reproducibility	Probability of corresponding			
	Not ID		ID	
	Clear	Unclear	Clear	Unclear
0%	0.097	0.104	0.564	0.594
50%	0.150	0.181	0.758	0.799
100%	0.226	0.297	0.883	0.915

Fig. D23: Probability of an examiner corresponding a minutia given the Comparison-phase reproducibility of that minutia among examiners who compared each image pair, conditioned on whether that examiner individualized, and whether that examiner said the minutia was clear. Probabilities calculated using logistic regression. (n=45,130 Comparison-phase minutiae; data from 11 latents that were each compared by only one examiner are excluded).

DiB-10.2 Reproducibility of corresponding minutiae

In our previous work [2], we noted "Disagreements on sufficiency for individualization tend to be associated with substantial disagreements on corresponding minutiae." Table D16 shows the very substantial interexaminer differences as to which minutiae are marked. Often when one examiner said a latent was NV, other examiners corresponded minutiae on that latent (recall that fingerprint comparisons in this test were selected to be borderline value). In addition to marking "definite" correspondences, examiners were instructed to indicate discrepancies (features in one print that definitely do not exist in the other print) as needed to support an exclusion determination. Examiners were also permitted to mark "debatable" correspondences: features "that potentially correspond, but do not meet your threshold for supporting an ID." The correspondences referred to in the body of this report include only "definite" correspondences.

Whereas definite correspondences occurred much more often in clear than unclear areas (3x), debatable correspondences occurred about equally in clear and unclear areas. After controlling for clarity, minutiae that were marked as debatable correspondences have a similar, but slightly lower, reproducibility distribution to all minutiae.

The following four tables (Table D16 through Table D19) describe reproducibility by type of correspondence markup as conditional probabilities: when examiner A marked a minutia, what did examiner B do? Table D16 summarizes these results across all data; Table D17 through Table D19 summarize these results on subsets of the data. The probabilities are calculated as weighted sums over all other examiners who marked each latent, such that each minutia marked by examiner A is weighted equally. The final column, "Marked and compared minutiae that were definitely corresponded," is the probability that examiner B definitely corresponded a minutia given that examiner B marked that minutia and compared the latent to the exemplar. For example, Table D16 shows that when examiners corresponded minutiae marked as clear, 68.8% of the time other examiners also corresponded those minutiae; 20.0% of the time other examiners did not mark those minutiae at all. The data in these tables is limited to 3618 markups as discussed in DiB-1.4.

ALL Minutiae				Minutiae	Examiner B						Marked and compared minutiae that were definitely corresponded
					Did not mark	Marked					
						Not Compared (NV)	Compared				
							Not corresponded		Corresponded		
							Unassoc.	Discrepant	Debatable	Definite	
Examiner A	Clear	NV		1379	33.4%	25.0%	20.2%	1.0%	1.7%	18.7%	45.0%
		Not corresponded	Unassoc.	12,231	36.8%	2.8%	43.7%	1.5%	1.4%	13.8%	22.9%
			Discrepant	457	32.7%	4.2%	41.9%	6.9%	1.6%	12.7%	20.2%
		Corresponded	Debatable	677	36.6%	4.2%	23.4%	1.0%	3.7%	30.9%	52.3%
			Definite	20,470	20.0%	1.5%	8.2%	0.3%	1.3%	68.8%	87.6%
	Unclear	NV		1447	49.7%	19.5%	16.8%	0.8%	1.4%	11.8%	38.4%
		Not corresponded	Unassoc.	5844	60.4%	3.0%	25.3%	0.9%	1.2%	9.2%	25.0%
			Discrepant	175	56.5%	3.4%	27.4%	5.4%	1.2%	6.0%	15.1%
		Corresponded	Debatable	755	63.2%	2.0%	10.2%	0.3%	2.3%	22.0%	63.3%
			Definite	7459	42.1%	1.8%	7.1%	0.2%	1.6%	47.3%	84.2%

Table D16: When examiner A marked a minutia, what examiner B did (n=50,894 minutiae marked during Analysis or added during Comparison). Without regard to clarity, 63.1% of the minutiae definitely corresponded by examiner A were also definitely corresponded by examiner B; 10.9% of examiner A's discrepancies were definitely corresponded by examiner B.

Mates				Minutiae	Examiner B						Marked and compared minutiae that were definitely corresponded
					Did not mark	Marked					
						Not Compared (NV)	Compared				
							Not corresponded		Corresponded		
							Unassoc.	Discrepant	Debatable	Definite	
Examiner A	Clear	NV		937	32.5%	23.2%	16.0%	0.1%	1.9%	26.3%	59.4%
		Not corresponded	Unassoc.	8613	38.6%	2.3%	38.9%	0.3%	1.4%	18.5%	31.2%
			Discrepant	137	34.8%	1.4%	24.0%	1.2%	1.3%	37.3%	58.4%
		Corresponded	Debatable	575	38.6%	3.8%	19.4%	0.2%	3.2%	35.0%	60.7%
			Definite	20,245	19.8%	1.4%	7.8%	0.2%	1.2%	69.5%	88.2%
	Unclear	NV		1013	48.7%	18.9%	14.3%	0.2%	1.4%	16.4%	50.7%
		Not corresponded	Unassoc.	4189	62.0%	2.4%	22.1%	0.2%	1.2%	12.1%	34.0%
			Discrepant	48	68.8%	1.7%	13.3%	0.9%	0.5%	14.8%	50.0%
		Corresponded	Debatable	672	63.6%	1.6%	8.2%	0.2%	2.1%	24.3%	70.0%
			Definite	7391	42.0%	1.7%	6.9%	0.2%	1.5%	47.6%	84.7%

Table D17: When examiner A marked a minutia, what examiner B did, limited to minutiae marked on ***mated pairs***. Without regard to clarity, 63.7% of the minutiae definitely corresponded by examiner A were also definitely corresponded by examiner B.

Nonmates				Minutiae	Examiner B						Marked and compared minutiae that were definitely corresponded
					Did not mark	Marked					
						Not Compared (NV)	Compared				
							Not corresponded		Corresponded		
							Unassoc.	Discrepant	Debatable	Definite	
Examiner A	Clear	NV		442	35.4%	28.9%	29.0%	3.0%	1.2%	2.5%	7.1%
		Not corresponded	Unassoc.	3618	32.4%	4.2%	55.0%	4.3%	1.3%	2.8%	4.4%
			Discrepant	320	31.7%	5.4%	49.6%	9.4%	1.7%	2.2%	3.6%
		Corresponded	Debatable	102	25.9%	6.9%	46.5%	5.9%	7.0%	7.9%	11.8%
			Definite	225	31.6%	5.8%	46.9%	3.9%	3.8%	8.0%	12.8%
	Unclear	NV		434	51.9%	20.8%	22.6%	2.1%	1.4%	1.2%	4.4%
		Not corresponded	Unassoc.	1655	56.5%	4.5%	33.2%	2.7%	1.3%	1.7%	4.4%
			Discrepant	127	51.9%	4.0%	32.8%	7.1%	1.5%	2.7%	6.2%
		Corresponded	Debatable	83	59.5%	5.4%	26.4%	1.5%	3.9%	3.3%	9.4%
			Definite	68	45.6%	4.8%	35.2%	2.9%	3.3%	8.3%	16.7%

Table D18: When examiner A marked a minutia, what examiner B did, limited to minutiae marked on ***nonmated pairs***. Without regard to clarity, 8.1% of the minutiae definitely corresponded by examiner A were also definitely corresponded by examiner B.

Both ID				Minutiae	Examiner B						Marked and compared minutiae that were definitely corresponded
					Did not mark	Marked					
						Not Compared (NV)	Compared				
							Not corresponded		Corresponded		
							Unassoc.	Discrepant	Debatable	Definite	
Examiner A	Clear	NV		N/A							
		Not corresponded	Unassoc.	5125	39.7%	N/A	38.5%	0.1%	1.2%	20.5%	34.0%
			Discrepant	8	48.1%	N/A	50.9%	0.0%	0.0%	1.0%	1.9%
		Corresponded	Debatable	317	35.1%	N/A	18.8%	0.0%	2.9%	43.2%	66.5%
			Definite	18,738	17.3%	N/A	5.5%	0.0%	0.9%	76.4%	92.4%
	Unclear	NV		N/A							
		Not corresponded	Unassoc.	2228	63.6%	N/A	20.8%	0.0%	0.9%	14.7%	40.5%
			Discrepant	7	83.3%	N/A	16.7%	0.0%	0.0%	0.0%	0.0%
		Corresponded	Debatable	356	62.8%	N/A	6.3%	0.0%	1.7%	29.2%	78.5%
			Definite	6558	36.6%	N/A	5.2%	0.0%	1.2%	57.0%	89.9%

Table D19: When examiner A marked a minutia, what examiner B did, limited to minutiae marked when ***both examiners individualized***; based on 185 image pairs that were individualized by at least two examiners (out of 231 mated pairs). Without regard to clarity, 69.4% of the minutiae definitely corresponded by examiner A were also definitely corresponded by examiner B.

Similar to the preceding tables, Table D20 and Table D21 describe reproducibility by type of correspondence markup and whether the examiners changed their Analysis markup during Comparison.

ALL Minutiae			Minutiae	Examiner B							Marked and compared minutiae that were definitely corresponded
				Did not mark	Marked						
					NV (Not Compared)	Compared					
						Not corresponded				Definite Corresp.	
						Retained	Moved	Deleted	Added		
Examiner A	NV		2826	41.8%	22.2%	17.6%	0.8%	1.6%	0.8%	15.2%	42.1%
	Not corresponded	Retained	15,384	39.4%	3.2%	41.2%	0.8%	2.2%	1.1%	12.0%	21.0%
		Moved	440	40.1%	5.2%	27.2%	1.3%	3.5%	1.2%	21.6%	39.5%
		Deleted	2895	63.4%	1.6%	12.0%	0.5%	5.3%	0.8%	16.4%	46.8%
		Added	1420	65.4%	1.6%	11.7%	0.4%	1.5%	1.5%	17.8%	54.1%
	Corresponded		27,929	25.9%	1.5%	6.6%	0.3%	1.7%	0.9%	63.1%	86.9%

Table D20: When examiner A marked a minutia, what examiner B did (n=50,894 minutiae marked during Analysis or added during Comparison).

CLEAR Minutiae			Minutiae	Examiner B							Marked and compared minutiae that were definitely corresponded
				Did not mark	Marked						
					NV (Not Compared)	Compared					
						Not corresponded				Definite Corresp.	
						Retained	Moved	Deleted	Added		
Examiner A	NV		1379	33.4%	25.0%	19.6%	0.7%	1.7%	0.9%	18.7%	45.0%
	Not corresponded	Retained	10,624	31.8%	3.1%	47.5%	0.8%	2.2%	1.1%	13.4%	20.6%
		Moved	307	36.3%	5.5%	30.4%	1.4%	3.7%	1.2%	21.5%	36.9%
		Deleted	1810	58.6%	1.8%	13.8%	0.6%	5.5%	0.8%	18.9%	47.8%
		Added	624	56.2%	2.1%	18.1%	0.4%	1.7%	1.4%	20.1%	48.2%
	Corresponded		20,470	20.0%	1.5%	7.1%	0.4%	1.6%	0.8%	68.8%	87.6%

Table D21: When examiner A marked a minutia, what examiner B did (n=35,214 minutiae marked by examiner A as Clear during Analysis or added during Comparison).

Fig. D24 shows the distribution of the proportion of examiners who corresponded each cluster by clarity among examiners who compared each image pair; Fig. D25 shows similar results limited to examiners who individualized the image pairs. These charts show that while consensus is generally low in unclear areas, results are mixed in clear areas: often a minority of examiners correspond minutiae in clear areas.

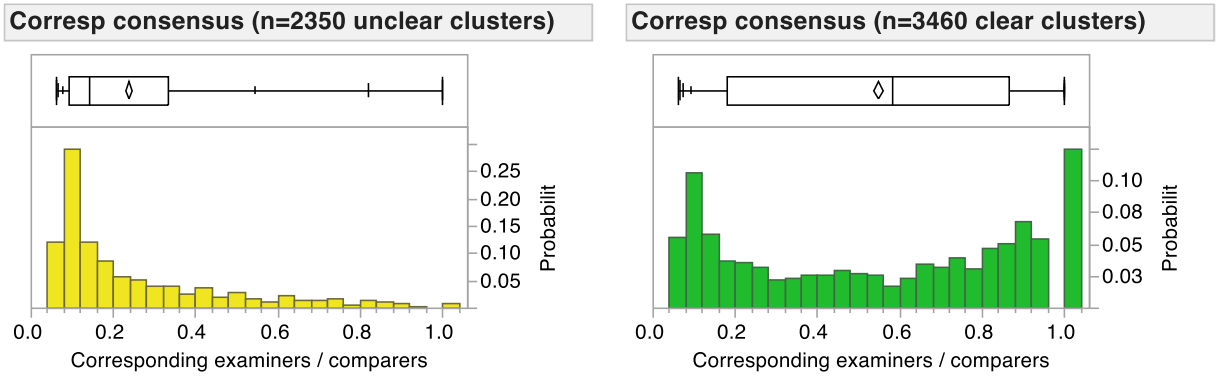


Fig. D24: Consensus on whether to correspond clusters by clarity, among examiners who ***compared*** each image pair. For each cluster, consensus is measured as (number of examiners who corresponded at least one marked minutia in the cluster) / (number examiners who compared). Excludes 5 image pairs that were compared by fewer than three examiners; also excludes clusters that no examiner corresponded. (3,126 comparisons of 263 image pairs, 215 mated)

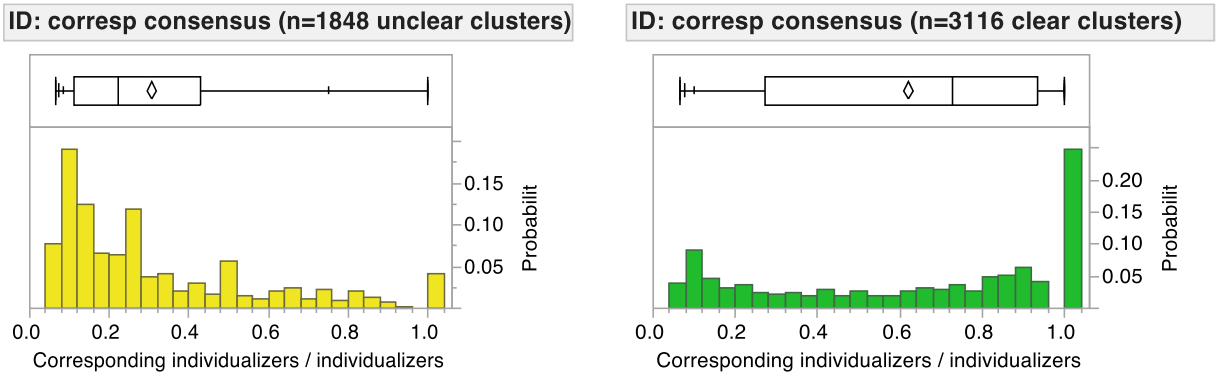


Fig. D25: Consensus on whether to correspond clusters by clarity, among examiners who ***individualized*** each image pair. For each cluster, consensus is measured as (number of individualizing examiners who corresponded at least one marked minutia in the cluster) / (number examiners who individualized). Excludes 140 image pairs that were individualized by fewer than three examiners (60/231 mated pairs excluded); also excludes clusters that no individualizer corresponded. (1662 comparisons)

A) Compared	Unclear		Clear		Total clusters
	Clusters	%	Clusters	%	
Singleton	990	68%	460	32%	1450
Minority	1037	49%	1058	51%	2095
Majority	297	21%	1119	79%	1416
SuperMajority	26	3%	823	97%	849

B) ID	Unclear		Clear		Total clusters
	Clusters	%	Clusters	%	
Singleton	753	65%	398	35%	1151
Minority	667	48%	720	52%	1387
Majority	347	27%	958	73%	1305
SuperMajority	81	7%	1040	93%	1121

Table D22: A) Cluster clarity by consensus on whether to correspond minutiae, among examiners who ***compared*** each image pair (same data as Fig. D24; n=5810 clusters); B) Cluster clarity by consensus on whether to correspond minutiae, among examiners who ***individualized*** each image pair (same data as Fig. D25; n=4975 clusters).

DiB-11 Reproducibility of minutia with respect to exclusion determinations

The test resulted in 561 exclusions on 81 mated and 75 nonmated pairs. When examiners determined that the latent and exemplar were not from the same source, they were asked to indicate a reason for the exclusion. The reasons given are summarized in Table D23. The distribution of reasons was not substantially different for nonmated and mated pairs (true and false exclusions). For 80% of exclusions, the reason given was "one or more minutiae differ."

There were 25 mated pairs and 70 nonmated pairs that more than one examiner excluded. Agreement on exclusion reasons was low (beyond chance). For example, the probability that examiner B said "minutiae differ" given that examiner A said "minutiae differ" was 67% for mated pairs and 48% for nonmated pairs (each image pair weighted equally).

Exclusion reason	Mates		Nonmates	
Pattern classes differ	12	9%	49	9%
Core or delta differences	8	6%	50	10%
One or more minutiae differ	104	80%	447	80%
Level-3 features differ	3	2%	6	1%
Other	3	2%	8	1%
Total	130	100%	430	100%

Table D23: Exclusion reasons. Examiners were instructed to select the first option that applied. The exclusion reason was missing for one comparison.

When examiners said "minutiae differ," discrepancies were not usually marked (34% of mates, 42% of nonmates, 40% overall). Agreement on discrepancies was greater than chance, but not substantially. There were 47 image pairs on which at least two examiners marked discrepancies.

Upon completing the examinations that resulted in exclusions, examiners had marked 1744 minutiae (in 1264 clusters) on mated latents, 123 (7.1%) as discrepant; and 4901 minutiae (in 1703 clusters) on nonmated latents, 425 (8.7%) as discrepant. As shown in Table D24, there were 18 clusters with 3 discrepancies marked and 8 clusters with 4 discrepancies marked on nonmated image pairs (vs. 7 and 1 predicted from simulations that randomly assigned the "discrepant" label throughout the minutiae at the average rates for mates and nonmates).

	Mates					Nonmates					
	Number of discrepancies					Number of discrepancies					
	0	**1**	**2**	**3**	**Total**	**0**	**1**	**2**	**3**	**4**	**Total**
Singleton	252	17	0	0	269	663	48	0	0	0	711
Not singleton	894	97	3	1	995	714	212	40	18	8	992
Total clusters	**1146**	**114**	**3**	**1**	**1264**	**1377**	**260**	**40**	**18**	**8**	**1703**

Table D24: Counts of discrepant minutiae among clusters on exclusion determinations by whether the cluster was a singleton. For example, 97 clusters on mated pairs that were marked by more than one examiner ("Not singleton") were marked as discrepant by exactly one examiner. In no case did more than four examiners mark a minutia as discrepant.

As shown in Table D25, when discrepancies were marked, they were more likely to be in clusters marked by many examiners: this pattern largely reflects chance (more opportunities for some examiner to note a discrepancy).

	Mates			Nonmates		
	Clusters	**Discrepancies**	**% Discrep**	**Clusters**	**Discrepancies**	**% Discrep**
Singleton	269	17	6%	711	48	7%
Minority	252	25	10%	354	72	20%
Majority	365	43	12%	406	178	44%
SuperMajority	378	38	10%	232	128	55%
Total	**1264**	**123**	**10%**	**1703**	**426**	**25%**

Table D25: Percentage of clusters marked as discrepant by any comparing examiner by Comparison-phase consensus.

DiB-12 Variation in minutia locations

In order to better understand the lack of reproducibility, we clustered minutiae marked on the exemplars and then looked to see how these exemplar clusters corresponded to latent clusters. We expected to find many examples of exemplar clusters whose corresponding minutiae on the latents had not been assigned to a single cluster because of variation in the precise location at which examiners marked minutiae in unclear areas on the latent.

Clustering was performed on the 3618 exemplar markups (Comparison phase) described in DiB-1.4 using the same clustering procedures and parameters as were used for the latents (DiB-3). Although clustering was performed on all minutiae marked on the exemplars, this analysis focuses on a subset of those minutiae that examiners marked as corresponding. In defining this subset, an additional 60 markups were omitted because of documentation errors in how the correspondences were marked. Most of these omitted markups were initially identified on the basis of having abnormally high bending energy (a measure of the non-linear component of the relative distortion between the minutiae marked on the latent and exemplar) [12,13]). Each of the omitted markups was manually reviewed and most were identified as having "crossed" correspondences that were clearly incorrect (and presumably inadvertent documentation errors).

13,397 clusters were constructed from the 41,071 minutiae on the 3618 markups; 27,159 of these minutiae were marked as corresponding (after omitting the documentation errors). The 27,159 corresponding minutiae were contained in 5470 clusters on the exemplars and corresponded to 5794 clusters on the latents.

Table D26 summarizes correspondences among latent and exemplar clusters. 15% (830/5470) of exemplar clusters were corresponded to more than one latent cluster; 9% (538/5794) of latent clusters were corresponded to more than one exemplar cluster. 31% (1672/5470) of exemplar clusters were corresponded to only one latent cluster simply because only one minutia within the cluster was corresponded; similarly, 35% (2015/5794) of latent clusters.

Just as most minutiae were marked in median clear areas, this variation in the location at which examiners marked minutiae was most often observed in median clear areas: although examiners could be confident in the presence of these minutiae, certain aspects of clarity can interfere more with determining the precise location of minutiae than with determining their presence or absence. Variation in location (together with the clustering criteria) accounts for most of the lack of one-to-one correspondence between latent and exemplar clusters; examples of incorrect alignment of the latent and exemplar were also noted.

	Latent clusters	Exemplar clusters
Only one minutia in the cluster was corresponded	2015	1672
More than one minutia in the cluster was corresponded	3779	3798
those minutiae corresponded to the same cluster	*3241*	*2968*
those minutiae corresponded to different clusters	*538*	*830*
Total	**5794**	**5470**

Table D26: Correspondences among latent and exemplar clusters

Acknowledgements

We thank the latent print examiners who participated in this study, and Erik Stanford for his technical support with the clustering algorithms. This is publication number 15-TBD of the FBI Laboratory Division. Names of commercial manufacturers are provided for identification purposes only and inclusion does not imply endorsement of the manufacturer or its products or services by the FBI. This work was funded in part under a contract award to Noblis, Inc. from the FBI Biometric Center of Excellence and in part by the FBI Laboratory Division. The views expressed are those of the authors and do not necessarily reflect the official policy or position of the FBI or the U.S. Government.

References

1 Ulery, B.T., Hicklin, R.A., Roberts, M.A., Buscaglia, J. (2016). Interexaminer variation of minutia markup on latent fingerprints. Forensic Sci Int (in press)

2 Ulery, B.T., Hicklin, R.A. , Roberts, M.A., Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. PLoS ONE, 9(11), e110179. http://dx.doi.org/10.1371/journal.pone.0110179

3 Criminal Justice Information Services (1997). Wavelet Scalar Quantization (WSQ) Gray-Scale Fingerprint Image Compression Specification, Version 3.1. https://www.fbibiospecs.org/docs/WSQ_Gray-scale_Specification_Version_3_1_Final.pdf

4 National Institute of Standards (2013). American National Standard for Information Systems: Data format for the interchange of fingerprint, facial & other biometric information. ANSI/NIST-ITL 1-2011 Update:2013. (NIST Special Publication 500-290 Rev1) Gaithersburg, MD: National Institute of Standards and Technology. http://fingerprint.nist.gov/standard

5 Chapman, W., et al. (2013). Latent Interoperability Transmission Specification. (NIST Special Publication 1152) Gaithersburg, MD: National Institute of Standards and Technology. http://dx.doi.org/10.6028/NIST.SP.1152

6 Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2013). Assessing the clarity of friction ridge impressions. Forensic Sci Int, 226(1):106-117. http://dx.doi.org/10.1016/j.forsciint.2012.12.015 (preprint: http://www.noblis.org/media/4b209d60-a147-414e-8bba-90c3a1e22c18/docs/article_assessing_clarity_latent_friction_ridge_pdf)

7 Ulery, B.T., Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2014). Changes in latent fingerprint examiners' markup between Analysis and Comparison. Forensic Sci Int, 247: 54-61. http://dx.doi.org/10.1016/j.forsciint.2014.11.021

8 Neumann, C., Champod, C., Yoo, M., Genessay, T., & Langenburg, G. (2013). Improving the understanding and the reliability of the concept of "sufficiency" in friction ridge examination. National Institute of Justice, Washington DC. https://www.ncjrs.gov/pdffiles1/nij/grants/244231.pdf

9 Ulery, B.T., Hicklin, R.A., Buscaglia, J., Roberts, M.A. (2014). Changes in latent fingerprint examiners' markup between Analysis and Comparison. Forensic Sci Int, 247: 54-61. http://dx.doi.org/10.1016/j.forsciint.2014.11.021

10 Daszykowski, M., Walczak, B., Massart, D.L. (2001). Looking for natural patterns in data. Part 1: Density based approach. Chemometr Intell Lab Syst, 56(2): 83-92. http://dx.doi.org/10.1016/S0169-7439(01)00111-3

11 Daszykowski, M., Walczak, B., Massart, D.L. (2002). Looking for natural patterns in analytical data. Part 2. Tracing local density with OPTICS. J Chem Inf Comp Sci, 42(3): 500-507. http://dx.doi.org/10.1021/ci010384s

12 Bookstein F.L. (1989). Principal warps: thin plate splines and the decomposition of deformations, IEEE T Pattern Anal, 11(6): 567-585. http://user.engineering.uiowa.edu/~aip/papers/bookstein-89.pdf

13 Kalka N.D., Hicklin R.A. (2014). On relative distortion in fingerprint comparison. Forensic Sci Int, 244:78-84, Nov 2014.