
Title: Comparison of Base Composition Analysis and Sanger Sequencing of Mitochondrial DNA for Four U.S. Population Groups

Author: Kevin M. Kiesler Michael D. Coble Thomas A. Hall Peter M. Vallone

PII: S1872-4973(13)00214-7
DOI: http://dx.doi.org/doi:10.1016/j.fsigen.2013.10.003
Reference: FSIGEN 1061

To appear in: *Forensic Science International: Genetics*

Received date: 8-8-2013
Revised date: 1-10-2013
Accepted date: 8-10-2013

Please cite this article as: K.M. Kiesler, M.D. Coble, T.A. Hall, P.M. Vallone, Comparison of Base Composition Analysis and Sanger Sequencing of Mitochondrial DNA for Four U.S. Population Groups, *Forensic Science International: Genetics* (2013), http://dx.doi.org/10.1016/j.fsigen.2013.10.003

# Comparison of Base Composition Analysis and Sanger Sequencing of Mitochondrial DNA for Four U.S. Population Groups

**Kevin M. Kiesler[a], Michael D. Coble[a], Thomas A. Hall[b], Peter M. Vallone[a]**

[a]National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD, 20899 USA

[b]Ibis Biosciences, a division of Abbott, 2251 Faraday Ave, Suite 150, Carlsbad, CA, 92008, USA

Corresponding author telephone: +1 (301) 975-4306 Fax: +1 (301) 975-8505

E-mail address: Kevin.Kiesler@nist.gov

## 1. Introduction

The use of human mitochondrial DNA (mtDNA) sequence content for human identification (HID) is well established in forensic casework [1, 2]. The high mtDNA copy number per cell relative to nuclear DNA [3] often makes amplification by polymerase chain reaction (PCR) possible when limitations in sample quantity cause nuclear markers such as short tandem repeats (STRs) to fail to amplify. Other qualities such as uniparental maternal inheritance [4] and high levels of sequence diversity [5] make mtDNA analysis a valuable tool for HID.

Forensic mtDNA analysis utilizes the non-coding control region (or D-loop), defined as nucleotide positions 16,024 through 16,569 and 1 through 576, spanning across the origin of replication of the circular molecule. In some instances only a subset of the 1,122 base pairs of the control region may be analyzed. The minimum mtDNA sequence coverage analyzed in typical case work includes Hypervariable region I (HV1) (positions 16,024 through 16,365) and hypervariable region II (HV2) (positions 73 through 340). Sequence data is generated using the method developed by Sanger *et al.* [6], with some modifications to accommodate contemporary automated sequencing instrumentation. The introduction of fluorescently labeled dideoxynucleotide triphosphates and automated capillary-based sequencing instruments has significantly improved the speed and accuracy of DNA sequencing while reducing the cost to generate sequence information. However, Sanger sequencing is still a labor intensive process which requires several steps to be performed by qualified specialists.

As an alternative to DNA sequencing, Ibis Biosciences, a division of Abbott Laboratories developed a fully automated electrospray ionization time-of-flight mass spectrometer (ESI-TOF MS) called the PLEX-ID. The instrument replaces the majority of steps in the Sanger sequencing workflow with one fully automated process, vastly reducing the amount of time and labor required to generate a forensic mtDNA profile. PCR product desalting, mass measurement, and primary data processing are performed by the PLEX-ID instrument without any operator interaction. The PLEX-ID mtDNA 2.0 assay characterizes positions 15,924 through 16,428 (excluding three highly conserved bases at positions 16,251 through 16,253 which fall in-between two adjacent amplicons) and 31 through 576 of the mtDNA control region. This is in contrast to the established control region positions typically used in sequencing. The assay uses PCR amplification of 24 primer pairs tiled across the HV1 and HV2 regions in overlapping fashion, which results in 24 short PCR fragments ranging in size from 85 to 140 base pairs that are analyzed by the mass spectrometer.

The combined base composition results for all 24 amplicons constitute a “base composition profile” for a sample. A complete profile may be submitted to the instrument’s searchable database for matching. Tools within the instrument’s analysis software, IbisTrack, allow for backwards compatibility with Sanger sequencing-derived mtDNA information by converting sequence data into base composition results which can then be searched for matches to other base composition results in the PLEX-ID database.

It is has been proposed that the PLEX-ID mtDNA 2.0 assay, “is more discriminating than sequencing the minimum HV1 and HV2 regions because it interrogates more of the mitochondrial genome” [7]. In this

study we compared the discriminatory power of the assay relative to Sanger sequencing by studying a set of 711 DNA samples from four population groups in the U.S.

Similar concordance studies have been published previously. Hall *et al.* [7] examined a set of 163 samples with the PLEX-ID and results were compared to Sanger sequence converted to base composition. Of that sample set, the 95 NIST samples are repeated in the present study. The study found that PLEX-ID results were in 100% agreement with converted Sanger sequence data. Another study by Warshauer *et al.* [8] analyzed 150 samples from three populations (50 African American, 50 Caucasians, and 50 Hispanics) and no discrepancies between PLEX-ID base composition results and Sanger sequencing were found. However, for three of the samples no products could be detected on the PLEX-ID for amplicon 2893, resulting in an overall concordance rate of 99.92% based on the number of successful mass measurements divided by the theoretical maximum number of measurements that could be made (3597/3600). A third study by Howard *et al.* (2013) [9] utilized the T5000 instrument from Abbott which is the predecessor to the PLEX-ID. The T5000 uses the same PCR amplification and ESI-TOF MS detection strategy as the PLEX-ID. The Howard study analyzed 250 bone samples from a World War I mass grave site believed to contain the remains of Australian and British soldiers who perished in the battle of Fromelles. Due to the limited quantity and degraded nature of the samples analyzed, the success rate was expected to be lower than what would be seen with pristine DNA extracts. Sanger sequencing and base composition results were achieved for 225 of the unidentified remains. Each of the 225 successfully sequenced samples were amplified and analyzed in duplicate on the T5000 in order to maximize the likelihood of obtaining a full base composition profile. The result of the Howard study was that base compositions were obtained for 10,435 amplicons, out of the maximum 10,800 possible, yielding a success rate of 96.6%. However, of these results there were two mass spectral base compositions which did not match the corresponding samples' converted Sanger sequence data, yielding an overall concordance rate of 96.6% and a false positive rate of 0.02% (2/10,435) in samples which represent typical casework specimens.

In the study below, the performance of the mass spectrometry base composition method was assessed by comparing the PLEX-ID and Sanger sequencing results for overall concordance rates and discrimination capacity using a population of 711 samples. We also compared the two methods across four population groups of self-identified ancestry to evaluate any potential performance bias in the PLEX-ID mtDNA 2.0 resulting from assay design.

## 2. Materials and Methods

### *2.1 DNA extraction and quantification*

Template DNA was extracted from whole blood using a modified salt-out procedure previously described [10]. DNA extracts were quantified using the Quantifiler Human DNA Quantification kit (Life Technologies, Carlsbad, CA, USA) under manufacturer's recommended conditions on an Applied Biosystems model 7500 (Life Technologies) instrument. Resultant DNA concentrations were used to normalize all DNA extracts to a final concentration of 0.5 ng/µL for DNA sequencing or 0.1 ng/µL for use with the PLEX-ID.

*2.2 Preparation of DNA template for sequencing*

Sanger sequencing is a multi-step process (see **Figure 1**) which requires manual preparation at each step. Time to complete the full process (not including DNA extraction) is approximately 10 to 12 hours, about three to four hours of which is hands-on time.

**Figure 1**: Steps in the Sanger sequencing process compared to the PLEX-ID workflow.

The majority (625 of 711) of the control region DNA sequencing for this study was performed at the Armed Forces DNA Identification Laboratory (AFDIL) using the PCR amplification and sequencing procedures described previously [11]. The remaining 86 samples were amplified and sequenced at NIST using the following procedure.

Sequencing template was generated first by PCR amplification using primers designed to amplify the mtDNA control region from position 15,898 to 628 (origin of replication inclusive) using the following primer sequences: forward primer F15878 - AAA TGG GCC TGT CCT TGT AG, and reverse primer R649 - TTT GTT TAT GGG GTG ATG TGA. Reaction conditions were 1x Amplitaq Gold Buffer (Life Technologies), 1.8 µmol/L $MgCl_2$ (Life Technologies), 0.5 µmol/L each primer (Eurofins MWG/Operon, Huntsville, AL, USA), 0.16 mg/mL BSA (Sigma-Aldrich Co., St. Louis, MO, USA), and 1.5 units of Amplitaq Gold polymerase (Life Technologies). Two microliters of DNA template described above were added to the reaction for a final reaction volume of 40 µL. Reactions were incubated in an Applied Biosystems GeneAmp 9700 PCR System (Life Technologies) thermal cycler using the following program: initial denaturation at 96 °C for 10 minutes; 36 cycles of denaturation at 94 °C for 30 seconds; annealing at 56 °C for 30 seconds; and extension at 72 °C for 60 seconds; followed by a final extension at 72 °C for 7 minutes and a hold step at 4 °C.

Reactions were verified for positive amplification on a 2.2% Lonza FlashGel DNA Cassette on the FlashGel System (Lonza Inc., Rockland, ME, USA) by loading 4 µL of PCR product mixed with 1 µL of loading dye. A size standard, FlashGel DNA Marker 100-4000bp (Lonza Inc.) was run on each gel tier to verify that the correct amplicon size was generated.

Unincorporated nucleotides and primers were inactivated using ExoSAP-IT (Affymetrix/USB, Santa Clara, CA, USA) following the manufacturer's recommended conditions, except incubation time at 37 °C was increased from 15 minutes to 90 minutes.

*2.3 DNA sequencing*

Sequencing reactions consisted of 4 µL of BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies), 4 µL of BigDye Terminator v3.1 5x Sequencing Buffer (Life Technologies), 3 µL of sequencing primer (1 µmol/L stock), and 2 µL of purified PCR product in a total volume of 20 µL.

Reactions were incubated in an Applied Biosystems GeneAmp 9700 PCR System (Life Technologies) thermal cycler using the following program: 25 cycles of denaturation at 96 °C for 10 seconds; annealing at 50 °C for 5 seconds; and extension at 60 °C for 4 minutes; followed by a hold at 4 °C. Several sequencing primers were utilized in order to achieve a minimum of double sequence coverage across the entire mtDNA control region. The primer sequences are listed in **Table 1** below.

**Table 1:** Control region Sequencing Primers. Eight primers were used to generate sequence to a minimum of 2x coverage across the mtDNA control region (nucleotide positions 16,024 through 576).

After cycle sequencing, reactions were cleaned up using a Performa DTR V3 96-Well Short Plate Kit (Edge Biosystems, Gaithersburg, MD, USA) according to the manufacturer's recommended protocol. Two microliters of purified sequencing reaction was added to 14 µL of Hi-Di Formamide (Life Technologies) in a 96-well optical plate (Phenix Research Products, Candler, NC, USA). Sequencing products were electrophoresed on an Applied Biosystems 3130xl Genetic Analyzer (Life Technologies) using an 80 cm array and POP-7 polymer under the long sequencing module conditions (injection at 2.4 kV for 20 seconds; run at 14.6 kV for 7400 seconds).

Resulting sequence electropherograms were trimmed and assembled into a contiguous sequence using Sequencher 5.1 software (Genecodes, Ann Arbor, MI, USA). Contiguous sequence spanning the entire control region with at least two sequence reads providing coverage of each base was considered to be the minimum quality criteria for acceptance of a sample's sequencing results into the study. Successfully completed sequences were converted into theoretical base composition using IbisTrack software.

*2.4 Amplification for detection by PLEX-ID mass spectrometer*

Mitochondrial DNA analysis on the PLEX-ID is a less labor intensive and faster means of achieving a profile for comparison to a forensic DNA database for haplotype frequency estimation. The PLEX-ID process replaces steps 3 through 7 of the Sanger sequencing process in **Figure 1** with a single, fully automated process. Time required to complete the process is approximately five hours, with about 30 to 60 minutes of hands-on time. Assay coverage is different for the mtDNA 2.0 kit when compared to Sanger sequencing of HV1 and HV2 or the full control region (see **Figure 2**). Primer sequences for PCR amplification and their associated target coordinates in human mtDNA have been described previously [7].

**Figure 2:** Coverage of the PLEX-ID mtDNA 2.0 assay relative to the region typically covered by Sanger sequencing. The PLEX-ID assays nucleotides outside the defined coordinates of the control region.

Nucleotides 16,251 through 16,253 are not analyzed by the mtDNA 2.0 kit.

DNA extracts described above were diluted to a concentration of 0.1 ng/µL for subsequent addition to eight wells in a column of a 96-well PLEX-ID mtDNA v2.0 assay plate (Abbott Laboratories, Des Plaines, IL, USA) prefabricated with all reagents required for amplification. The foil seal covering the wells of the plate was pierced with a pipette tip as 5 µL of template DNA was added. After template addition, the plate was heat sealed with PCR Foil seals (Abbott Laboratories) on an ALPS 50V Heat Sealer (ThermoFisher Scientific, Waltham, MA, USA) by compressing for four seconds at 180 °C. Following a brief centrifugation, the reaction plate was incubated in a Mastercycler Pro S (Eppendorf AG, Hamburg, Germany) thermal cycler according to the manufacturer's recommended thermal cycling program: denaturation at 95 °C for 10 min followed by 36 cycles of denaturation at 95 °C for 20 sec; annealing at 50 °C (ramp rate 5%) for 90 sec; and extension at 72 °C for 5 sec; with a final extension step of 72 °C for 4 min and a final heating step of 99 °C for 10 min followed by a 4 °C hold. On each reaction plate 10 samples plus a positive and negative control were run.

After completing the thermal cycling the plate was then loaded into the input stacker of the PLEX-ID which has capacity to hold up to 15 plates. PCR products were purified by the PLEX-ID using proprietary magnetic bead chemistry to remove salts, enzymes, unincorporated nucleotides, and any other PCR components which might interfere with generation of mass spectra. Purified PCR product was eluted in a buffer containing two peptide standards with masses of 727.4 Da and 1347.7 Da which act as calibrants to facilitate data processing. The electrospray ionization source operates in negative mode at approximately -4000 V (depending on the individual instrument's tuning parameters which are not user configurable) and 300 °C. PCR products are sprayed into the ionization source at a flow rate of 280 µL per hour with dry compressed air, generated onboard the instrument, used as a countercurrent to aid in analyte desolvation. The time-of-flight analyzer collects 5000 scans per second, for a period of approximately 28 seconds. Resultant mass spectra are processed by proprietary software which performs several steps to produce a background-subtracted, deconvolved spectrum that presents the data as if only the singly charged mass peak were detected. Mass spectra are represented in the software as traces on a graph with mass on the x-axis and signal on the y-axis. Successfully detected masses were stored on the Ibis Track database. Any peaks which were not automatically detected by the Ibis Track algorithm, but were determined to be valid masses of PCR products, were manually annotated and added to the database.

Sanger derived control region sequences were converted to theoretical base composition results and imported into the PLEX-ID database by Ibis Track software.

Sanger sequence data was analyzed in Sequencher software. Mitochondrial haplotypes were categorized by creating a multiple sequence alignment for comparison of all relevant sequences. Sequence analysis was limited to the 706 control region sequences for samples which corresponded to those with full 24-amplicon base composition profiles from the PLEX-ID.

Each of the samples with a complete 24-amplicon base composition profile from the PLEX-ID was compared pairwise with all other full profiles (n = 706) in the PLEX-ID database to determine whether there was a matching profile present in the study population. To do so, a custom utility was created to cross-compare base composition profiles to assess the number of profiles in the comparison set that cannot be differentiated from it and the number of differences that are present between the profile and other profiles in the comparison set. The purpose of this analysis was to assess discriminatory power of the base composition method compared to the sequencing method. The utility was designed to infer the minimum number of sequence differences that must be present between two sequences using the base compositions of overlapping PCR products as the input with the most conservative approach possible for handling of heteroplasmy and polymorphisms which occur within regions assayed by overlapping amplicons.

Discrimination capacity (DC) was calculated by summing the number of haplotypes observed in a population and dividing by the population sample size.

Control region sequences are available in the EMPOP database (www.empop.org [12]) under the following accession numbers: EMP00051 [13], EMP00053 [14], EMP00055, EMP00417, EMP00640, EMP00641, and EMP00642.

**3. Results and Discussion**

*3.1 Population samples*

In total 711 DNA samples from the NIST U.S. population dataset were processed on the PLEX-ID instrument. The number of samples for each of the U.S. sample groups was: African American (260), Asian American (49), Caucasian (262), and Hispanic American (140).

*3.2 Concordance*

Concordance was evaluated by comparison between the PLEX-ID measured base composition and theoretical sequence derived base composition for each amplicon. Because the mtDNA 2.0 assay interrogates areas outside the region targeted by a typical Sanger sequencing approach (see **Figure 2**), three amplicons upstream of the control region did not have sequence data available for comparison. We did not generate additional sequence data to verify performance of the three amplicons outside of the canonical control region. Those three amplicons are arbitrarily termed 2901, 2925, and 2899. They target mtDNA nucleotide positions 15,893 to 16,012, 15,937 to 16,041, and 15,895 to 16,073 respectively. A typical result from the Batch Pairwise Comparison tool appears in **Table 2** where each of the 24 amplicons' base compositions are compared to converted Sanger sequence data and the number of differences between the two results are listed.

**Table 2:** Results from the Batch Pairwise Comparison tool. For each of the 24 amplicons in the mtDNA 2.0 assay the amplicon name appears in the first column, the amplified coordinates (including primer binding

sites) in the second column, theoretical base compositions calculated from Sanger sequencing data appear in the third column, base compositions calculated from measured masses appear in the fourth column, and any differences between the base composition results are tallied in the fifth column.

There were no instances where measurements made on the PLEX-ID disagreed with base compositions calculated from Sanger sequencing data, yielding a concordance rate of 100% when a measurement was successfully made on the PLEX-ID. However, in the course of running the 711 samples, there were five templates which had a single amplicon that could not be detected by the PLEX-ID, resulting in a partial profile (23 of 24 amplicons). Sanger sequencing results were examined for the five samples with partial profiles. In two African American samples, there were three polymorphisms (89 C, 93 G, and 95 C) present in the reverse primer binding site for amplicon 2902, which spans nucleotide positions 76 through 97. These three mismatches with the primer sequence are likely sufficient to disrupt primer annealing and prevent PCR amplification. While information for amplicon 2902 was missing, the three polymorphisms listed above are assayed by the adjacent amplicon, 2903, which spans positions 42 through 113. Information from only ten bases not covered by amplicon 2903 was lost with the dropout of amplicon 2902. This demonstrates the utility of an overlapping amplicon strategy.

In the three remaining samples with incomplete profiles, mass spectral data showed evidence of extensive C-length heteroplasmy in the form of two or more products for an amplicon associated with the C-stretch regions in one of the hypervariable regions (HV1, HV2, or HV3). C-length heteroplasmy in the HV1 region resulted in the failure to detect amplicon 2893 (spanning positions 16,154 through 16,268) for one Asian sample and amplicon 2895 (spanning nucleotide positions 16,130 through 16,224) for one Hispanic sample. Similarly, C-length heteroplasmy was observed in the HV3 C-stretch for a second Asian sample, resulting in the failure to detect amplicon 2913 (spanning nucleotide positions 464 through 603).

For an additional 11 samples (8 Asian, 2 Caucasian, and 1 Hispanic), IbisTrack software's automated analysis failed to identify amplicons associated with the HV1 C-stretch region (primer pairs 2893 and 2895). These 11 samples were carefully inspected using the "Base Composition Browser" tool in the software. Appropriate peaks were identified, but were very low in signal intensity and/or had F:R peak height ratios above the default of 2.5. Manual assignment of the masses resulted in full concordance with converted Sanger sequence data.

Overall the concordance rate was 99.3% (706/711) based on the criterion that a full 24-amplicon profile is required in order to register the sample with the PLEX-ID database. When considering that 14,931 (711 x 21) measurements were made which could be verified with sequence data, results were obtained for 14,926, yielding a concordance rate of 99.97%.

*3.3 Heteroplasmy*

When evaluating the results for amplicons which span the poly-C tracts of the hypervariable regions, where C-length heteroplasmy is common (i.e. HV1 nucleotides 16,180 through 16,194 and HV2

nucleotides 302 through 310) mismatches in the number of C residues between PLEX-ID and Sanger results were not considered discordant because determining the exact number of poly-C nucleotides can often be challenging using a Sanger sequencing approach [15]. C-length heteroplasmy in the hypervariable regions was detected frequently, in 320 samples (or 45% of the population). An additional five samples had length heteroplasmy elsewhere in the control region, giving an overall length heteroplasmy rate of 46% (325/711), which is slightly lower than the 52% reported by Irwin *et al.* [16] in a large study of worldwide populations. The majority of C-length heteroplasmy was found in the HV2 poly-C stretch (35% of samples). This is slightly lower than the frequency of 45% in HV2 seen in the Irwin *et al.* [16] study. C-length heteroplasmy was observed in the HV1 poly-C stretch in 16% of samples (consistent with Irwin *et al.* [16] at 15%) and in the HV3 poly-C stretch (positions 568 to 573) in 6% of samples. Of the 325 total observations, 102 were observed in the Sanger sequence dataset but not in the PLEX-ID results. Further inspection of the mass spectra from a subset of these 102 samples revealed minor component peaks corresponding to +1 C at signal strengths which were barely distinguishable from noise at 8.8% (± 2.8%) of the major component. This is well below the Ibis Track algorithm's 20% cutoff for minor component mass peaks, and their resemblance to baseline noise makes it unlikely that they would have been manually identified. There were 93 cases where length heteroplasmy were observed in mass spectral results which were not seen in sequencing data.

Point heteroplasmy was detected 53 times in 47 samples by the PLEX-ID during this study giving an occurrence rate of 6.6%. In cases where point heteroplasmy was observed, two amplicons with masses consistent with their nucleotide content were detected by the PLEX-ID (e.g. four mass peaks present in the mass spectrum). Presence of heteroplasmy was not considered to be a case of discordance as long as one of the amplicons from the PLEX-ID matched the base composition of the converted Sanger sequence result. Due to inconsistencies in the ability of Sanger sequencing to detect heteroplasmy when the minor component approaches ≈ 20% or below, observation of point heteroplasmy by one method and not the other was not considered to be a criterion for discordance. Overall, the mass spectral results identified point heteroplasmy at an average minor component abundance of 29% (± 12%). The minimum minor component abundance was observed at 10%, while the maximum was 50%. Of the 53 observed heteroplasmic sites, more than half (27) occurred in regions which were covered by two amplicons. When two amplicons assayed the same heteroplasmic site, there was consensus on the quantification of the minor component. The average difference between two amplicons' minor component quantity was 3% (± 3%), with a minimum of 0% and a maximum of 10%. There were six samples which had two heteroplasmic sites that were detected in mass spectral data, giving an occurrence rate of 1%, similar to the rate described by Tully et al. [17]. In the 53 total observations of point heteroplasmy, most were transitions (50 of 53 = 94%), 43 of which were mixed C/T (Y) bases, seven were A/G (R) mixed bases. Of the remaining three transversion heteroplasmies, one was an A/C (M) mixed base (identified in sequence data at position 16,183 which is adjacent to the HV1 poly-C stretch), one was an A/T (W) mixed base, and one was a G/T (K) mixed base. Of the 53 point heteroplasmy observations in the mass spectral data, Sanger sequencing corroborated the presence of the heteroplasmy in 30 instances, while in 23 instances sequencing did not detect heteroplasmy. In one such case, the location was at a site outside the control region and therefore would not be detected in the sequencing data. The remaining 22 cases not detected by sequencing had an average minor

component of 21% (± 11%), with 15 of the observations below 20%, five between 21% and 30%, and two above 31%. The minimum and maximum for point heteroplasmy unique to the PLEX-ID results were 10% and 48% respectively. There were no cases where heteroplasmy was identified by sequencing that could not be identified by mass spectrometry. One possible explanation for the improved performance of the PLEX-ID system in identifying point heteroplasmy is that Sanger sequencing electropherograms frequently contain baseline noise which resembles low level heteroplasmy, making it difficult to identify heteroplasmy below 20% with high confidence in sequence data while the signature of heteroplasmy in mass spectral data is more easily differentiated from baseline noise.

*3.4 Discrimination capacity*

Using the 706 samples from the concordance study which had full 24-amplicon profiles, a comparison between the PLEX-ID and Sanger sequencing was made to determine the relative abilities of each method to discriminate between mitochondrial DNA species. The PLEX-ID database was searched against itself, comparing each of the 706 samples pairwise to identify matching base composition profiles using the custom utility described above.

Sequence data was analyzed using two methods: 1) with sequence restricted to just the HV1 and HV2 regions (16,024 – 16,365 and 73 – 340 respectively), which is the typical range analyzed for casework samples and 2) with sequence covering the full control region (16,024 – 576). The results of the three analyses are presented in **Table 3.**

**Table 3:** Results of database searches for common mitochondrial types. Percentage of samples uniquely identified, discrimination capacity, and number of nucleotides covered by each method are included.

From **Table 3**, the majority (> 70%) of the 706 samples in the analysis were unique to a single sample. Analysis of the full control region sequence had the highest number of uniquely identified individuals (549), the PLEX-ID had an intermediate number of uniquely identified individuals (522), and the fewest uniquely identified individuals were seen in the hypervariable region 1&2 sequencing results (499), as was expected. The PLEX-ID was able to uniquely identify more samples than HV1/HV2 sequencing by a difference of 4.5%, while the full control region sequence identified more samples uniquely by a difference of 5.0% relative to the PLEX-ID. Our results agree with those of Hall *et al.* [7] that the PLEX-ID mtDNA 2.0 assay is more discriminatory than the minimal HV1/HV2 sequencing strategy.

We also examined the discrimination capacity (DC) of each approach. The trend seen with uniquely identified samples holds with DC. Full control region sequencing gave the highest DC (0.85), the PLEX-ID was intermediate (0.83), and hypervariable region sequencing had the lowest DC (0.80). The PLEX-ID had a higher DC than HV1/HV2 sequencing by 2.4%. However, the PLEX-ID had a lower DC than full

control region sequencing by – 3.4%. Each of the three analysis methods assays a different number of nucleotides. If DC is normalized to account for differences in coverage by dividing DC by number of nucleotides we can compare the DC per nucleotide assayed (DC/nt). Using the normalized statistic, hypervariable region sequencing had the highest DC/nt at $1.31 \times 10^{-3}$, followed by the PLEX-ID at $7.92 \times 10^{-4}$, and control region sequencing was lowest at $7.58 \times 10^{-4}$. This analysis shows that the majority of information available for discrimination is found within the hypervariable regions 1 & 2 and that assaying additional sites modestly increases discrimination capacity.

*3.5 Performance within population groups*

Four distinct, self-identified, populations were represented in this study, Asian American, African American, Caucasian, and Hispanic. **Table 4** gives the breakdown of unique types and DC for each of the four groups.

**Table 4:** Results of database searches separated into the four populations in this study. Percentage of unique samples and discrimination capacity is shown at the bottom of each column.

From **Table 4**, all of the Asian samples were uniquely identified by all three assays, which could be explained by the small sample size in our study. For the remaining three populations, DC was lowest in the Caucasian population while African American and Hispanic populations were similar to each other in DC, but slightly higher than Caucasians. These results concur with other studies that show more shared, common types among Caucasians compared to other population groups [13, 14, 18].

In the concordance study, we noted a null allele at amplicon 2902 in two African American samples from mtDNA haplogroup L1c*, commonly found in Central and West Africa. Only a small amount of information was lost due to the failure of amplicon 2902 because an overlapping amplicon, 2903, provided redundant coverage which included the polymorphisms at positions 89, 93, and 95. However, we identified a third African American sample within the study population having the same set of SNPs at positions 89, 93, and 95 - yet this sample was successfully assayed by amplicon 2902. Further investigation into the performance of amplicon 2902 in mitochondrial haplogroup L1c* is planned.

The increased discrimination capacity in the Asian population group and two dropouts in the African American population group were the only variations in assay performance noted which could be attributed to analysis of different ethnographic population groups.

**4. Conclusions**

The concordance study presented here has shown that measurements made with the PLEX-ID mtDNA 2.0 assay are able to unequivocally determine the mtDNA base composition content 100% of the time. However, there were five instances where one or more amplicons were not detected resulting in partial base composition profiles and, therefore, reduced information content. Since there are no companion assays which might allow for amplification of just one of the primer pairs in the mtDNA 2.0 assay, there

are no other means to complete a partial profile using the PLEX-ID mass spectrometer other than repeating the analysis of the sample.

From the unique and common haplotypes and discrimination capacity studies, one can generalize that the performance of the PLEX-ID is intermediate in its information content relative to the two sequencing methods. The PLEX-ID's discriminatory power was slightly better than sequencing HV1 and HV2 and slightly less than full control region sequencing. Although the PLEX-ID and full control region sequencing assays contain nearly the same number of bases, there is a slight loss of discriminatory power in mass spectrometry analysis compared to sequencing over an equivalent coordinate range due to rare instances of compensatory sequence differences (e.g. a C->T and a T->C) within the same amplified product [7]. In addition, the inclusion of the 16,519 SNP between HV1 and HV2 in the full control region sequencing method, and not in the mass spectrometry method, will add to the slight difference in discriminatory capacity between the two approaches.

Performance in four populations of different ethnic background showed that the PLEX-ID was similar to sequencing in its performance as evidenced by DC comparisons. This observation leads to the conclusion that the PLEX-ID mtDNA 2.0 assay and supporting database were constructed in such a way as to not be biased in its ability to perform across the populations studied, although two instances where a specific set of primers failed to amplify was observed in two individuals of African lineage. Additional studies to characterize potential primer hybridization issues in other populations groups (e.g. among a larger Asian population sample size) are ideal for future studies.

There is a considerable savings in labor input when using the PLEX-ID system, relative to Sanger sequencing, primarily due to comprehensive automation of sample handling, DNA measurement, and primary data analysis. Increased automation of Sanger sequencing poses challenges in the laboratory because it requires sophisticated robotic liquid handling systems which are currently unable to perform all of the steps in Sanger sequencing in a hands-off process. Despite the caveat that mass spectrometry base composition analysis has slightly lower information content than Sanger sequencing, it may be worthwhile to exchange a small amount of discriminatory power for a significant reduction in labor by using the PLEX-ID automated ESI-TOF mass spectrometer for forensic mitochondrial DNA profiling.

Note to the reader regarding the PLEX-ID system recall: The PLEX-ID was recalled by Abbott in November of 2012 for reasons stated as, "reliability issues reported by clinical customers". All PLEX-ID instruments were removed from the field by Abbott. The instrument is currently undergoing redesign for a re-launch at an undetermined time in the future.

**Acknowledgements**

Thank you to Chantel Giamanco and John Fredericks from the Abbott technical support staff for their invaluable assistance.

**Disclaimer**

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

**References**

[1] Sullivan, K.M., Hopgood, R., and Gill P. (1992) Identification of human remains by amplification and automated sequencing of mitochondrial DNA. *Int. J. Legal Med.* **105**, 83-86.

[2] Wilson, M.R., Stoneking, M., Holland, M.M., DiZinno, J.A., and Budowle, B. (1993) Guidelines for the use of mitochondrial DNA sequencing in forensic science. *Crime Lab. Digest* **20** (4), 68-77.

[3] Bogenhagen, D., and Clayton, D. (1974). The number of mitochondrial deoxyribonucleic acid genomes in mouse L and human HeLa cells. *J. Biol. Chem.* **249**, 7991-7995.

[4] Giles, R.E., Blanc, H., Cann, H.M., and Wallace, D.C. (1980). Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U.S.A.* **77** (11), 6715-6719.

[5] Stoneking, M., Hedgecock, D., Higuchi, R.G., Vigilant, L., and Erlich, H.A. (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am. J. Hum. Genet.* **48** (2), 370-382.

[6] Sanger, F., Nicklen, S., and Coulson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74** (12), 5463-5467.

[7] Hall, T.A., Sannes-Lowery, K.A., McCurdy, L.D., Fisher, C., Anderson, T., Henthorne, A., Gioeni, L., Budowle, B., and Hofstadler, S.A. (2009). Base composition profiling of human mitochondrial DNA using polymerase chain reaction and direct automated electrospray ionization mass spectrometry. *Anal. Chem.* **81**, 7515-7526.

[8] Warshauer, D.H., King, J., Eisenberg, A.J., and Budowle, B. (2013) Validation of the PLEX-ID mass spectrometry mitochondrial DNA assay. *Int. J. Legal Med.* **127** (2), 277-286.

[9] Howard, R., Vesela, E., Thomson, J., Bache, K., Chan, Y., Cowen, S., Debenham, P., Dixon, P., Krause, J., Krishan, E., Moore, D., Moore, V., Ojo, M., Rodrigues, S., Stokes, P., Walker, J., Zimmerman, W., and Barallon, R. (2013) Comparative analysis of human mitochondrial DNA from World War I bone samples by DNA sequencing and ESI-TOF mass spectrometry. *Forensic Sci. Int. Genet.* **7** (1), 1-9.

[10] Miller S.A., Dykes D.D., Polesky H.F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res.* **16** (3), 1215.

[11] Irwin, J.A., Saunier, J.L., Strouss, K.M., Sturk, J.A., Diegoli, T.M., Just, R.S., Coble, M.D., Parson, W., and Parsons, T.J. (2007). Development and expansion of high-quality control region databases to improve forensic mtDNA evidence interpretation. *Forensic Sci. Int. Genet.* **1** (2), 154-157.

[12] Parson, W., Dür, A. (2007) EMPOP – a forensic mtDNA database. *Forensic Sci. Int. Genet.* **1**(2), 88-92.

[13] Saunier, J.L., Irwin, J.A., Just, R.S., O'Callaghan, J., and Parsons, T.J. (2008). Mitochondrial control region sequences from a U.S. "Hispanic" population sample. *Forensic Sci. Int. Genet.* **2** (2), e19-23.

[14] Diegoli, T.M., Irwin, J.A., Just, R.S., Saunier, J.L., O'Callaghan, J.E., and Parsons, T.J. (2009). Mitochondrial control region sequences from an African American population sample. *Forensic Sci. Int. Genet.* **4** (1), e45-52.

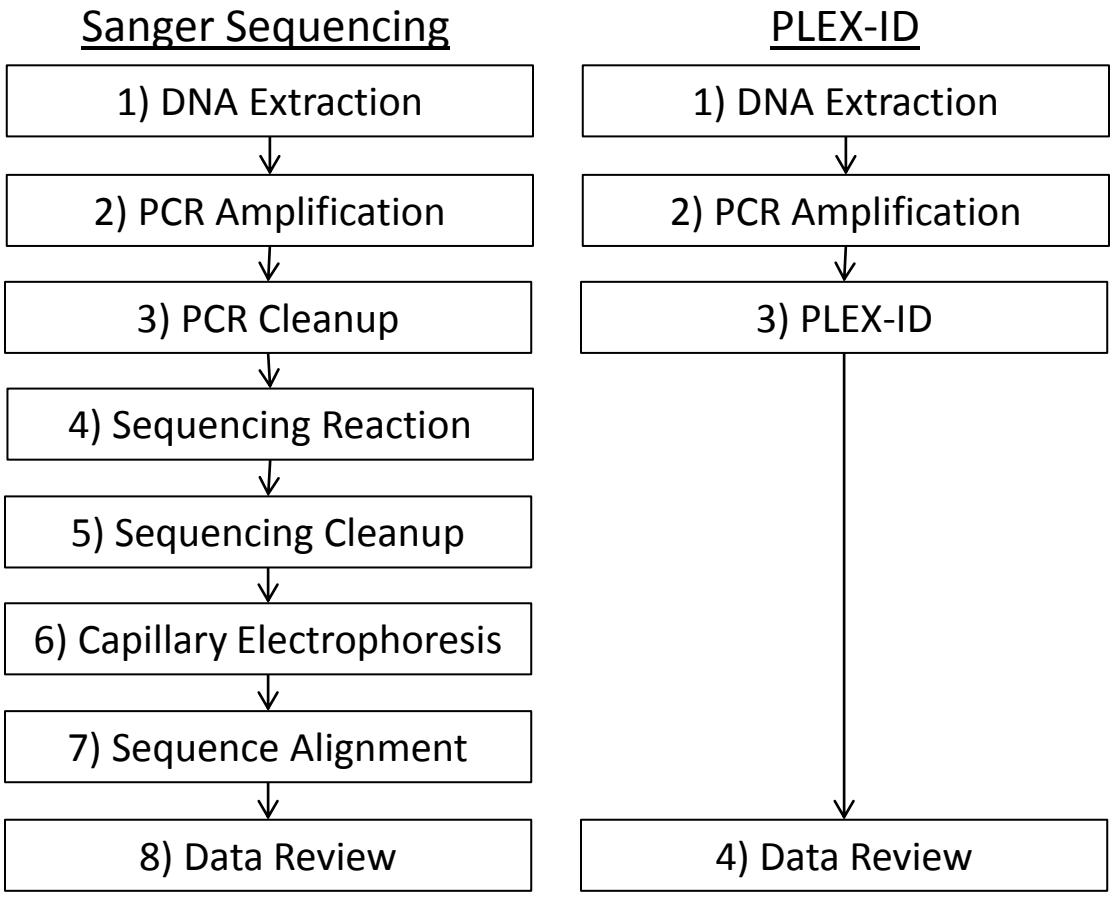
[15] Stewart, J.E., Fisher, C.L., Aagaard, P.J., Wilson, M.R., Isenberg, A.R., Polanskey, D., Pokorak, E, DiZinnio, J.A., and Budowle, B. (2001) Length variation in HV2 of the human mitochondrial DNA control region. *J. Forensic Sci.* 46(4), 862-70.

[16] Irwin, J.A., Saunier, J.L., Niederstätter, H., Strouss, K.M., Sturk, K.A., Diegoli, T.M., Brandstätter, A., Parson, W., and Parsons, T.J. (2009) Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. *J. Mol. Evol.* **68**, 516-527.

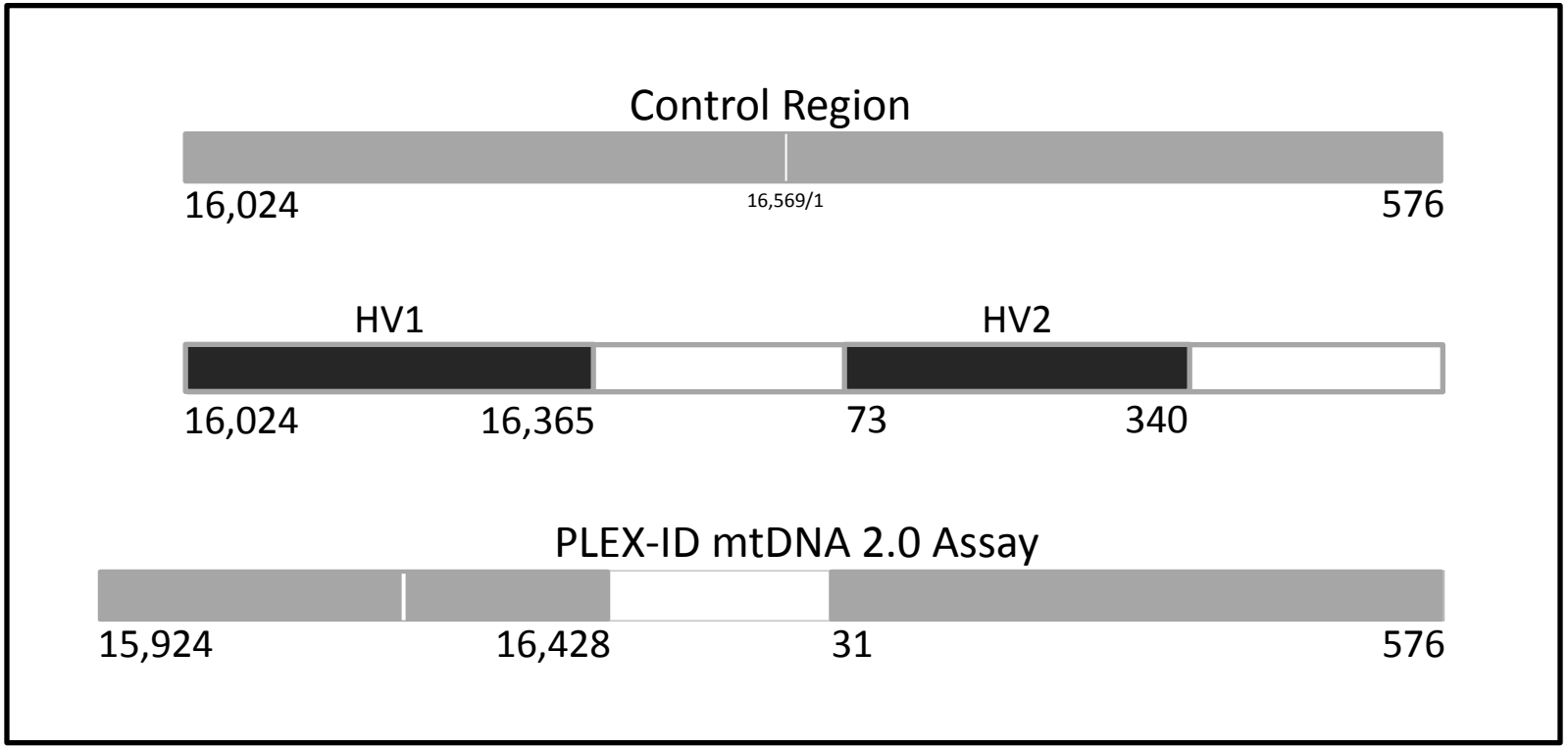
[17] Tully, L.A., Parsons, T.J., Steighner, R.J., Holland, M.M., Marino, M.A., and Prenger, V.L. (2000). A sensitive denaturing gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am. J. Hum. Genet.* **67**, 432-443.

[18] Coble, M.D., Just, R.S., O'Callaghan, J.E., Lemanyi, I.H., Peterson, C.T., Irwin, J.A., and Parsons, T.J. (2004). Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. *Int. J. Legal Med.* **118** (3), 137-146.

Figure


Sanger Sequencing
1) DNA Extraction
2) PCR Amplification
3) PCR Cleanup
4) Sequencing Reaction
5) Sequencing Cleanup
6) Capillary Electrophoresis
7) Sequence Alignment
8) Data Review
PLEX-ID
1) DNA Extraction
2) PCR Amplification
3) PLEX-ID
4) Data Review

Figure


Control Region
16,024
16,569/1
576
HV1
HV2
16,024
16,365
73
340
PLEX-ID mtDNA 2.0 Assay
15,924
16,428
31
576

| Primer Name | Sequence |
|---|---|
| F15971 | 5' TTA ACT CCA CCA TTA GCA CC 3' |
| F16190 | 5' CCC CAT GCT TAC AAG CAA GT 3' |
| R16400 | 5' GTC AAG GGA CCC CTA TCT GA 3' |
| F155 | 5' TAT TTA TCG CAC CTA CGT TC 3' |
| R285 | 5' GTT ATG ATG TCT GTG TGA A 3' |
| F314 | 5' CCG CTT CTG GCC ACA GCA CT 3' |
| R484 | 5' TGA GAT TAG TAG TAT GGG AG 3' |
| R599 | 5' TTG AGG AGG TAA GCT ACA TA 3' |

| pp | coords | Y8_seq | | Y8 | Diff |
|---|---|---|---|---|---|
| 2901: | 15893..16012: | NODATA | ---> | A48 G17 C25 T30 | 0 |
| 2925: | 15937..16041: | NODATA | ---> | A35 G14 C24 T32 | 0 |
| 2899: | 15985..16073: | NODATA | ---> | A26 G15 C21 T27 | 0 |
| 2898: | 16025..16119: | A26 G17 C27 T25 | ---> | A26 G17 C27 T25 | 0 |
| 2897: | 16055..16155: | A31 G13 C30 T27 | ---> | A31 G13 C30 T27 | 0 |
| 2896: | 16102..16224: | A45 G13 C42 T23 | ---> | A45 G13 C42 T23 | 0 |
| 2895: | 16130..16224: | A36 G7 C33 T19 | ---> | A36 G7 C33 T19 | 0 |
| 2893: | 16154..16268: | A44 G7 C46 T18 | ---> | A44 G7 C46 T18 | 0 |
| 2892: | 16231..16338: | A40 G9 C39 T20 | ---> | A40 G9 C39 T20 | 0 |
| 2891: | 16256..16366: | A37 G9 C40 T25 | ---> | A37 G9 C40 T25 | 0 |
| 2890: | 16318..16402: | A20 G14 C30 T21 | ---> | A20 G14 C30 T21 | 0 |
| 2889: | 16357..16451: | A21 G17 C36 T21 | ---> | A21 G17 C36 T21 | 0 |
| 2902: | 5..97: | A19 G24 C24 T26 | ---> | A19 G24 C24 T26 | 0 |
| 2903: | 20..139: | A24 G34 C29 T33 | ---> | A24 G34 C29 T33 | 0 |
| 2904: | 83..187: | A23 G21 C29 T32 | ---> | A23 G21 C29 T32 | 0 |
| 2905: | 113..245: | A39 G18 C28 T48 | ---> | A39 G18 C28 T48 | 0 |
| 2906: | 154..290: | A48 G18 C31 T40 | ---> | A48 G18 C31 T40 | 0 |
| 2908: | 204..330: | A42 G16 C39 T32 | ---> | A42 G16 C39 T32 | 0 |
| 2907: | 239..363: | A43 G11 C50 T23 | ---> | A43 G11 C50 T23 | 0 |
| 2923: | 262..390: | A47 G10 C54 T20 | ---> | A47 G10 C54 T20 | 0 |
| 2910: | 331..425: | A33 G9 C27 T26 | ---> | A33 G9 C27 T26 | 0 |
| 2916: | 367..463: | A27 G8 C32 T30 | ---> | A27 G8 C32 T30 | 0 |
| 2912: | 409..521: | A32 G7 C48 T26 | ---> | A32 G7 C48 T26 | 0 |
| 2913: | 464..603: | A44 G10 C63 T23 | ---> | A44 G10 C63 T23 | 0 |

Total differences: 0

| # of Times Haplotype Observed | PLEX-ID | Sequence | |
|---|---|---|---|
| | | HV1 & HV2 | 16024 - 576 |
| 1 | 522 | 499 | 549 |
| 2 | 36 | 32 | 31 |
| 3 | 12 | 14 | 12 |
| 4 | 4 | 4 | 4 |
| 5 | 5 | 6 | 4 |
| 6 | 1 | 4 | 2 |
| 7 | 1 | 1 | - |
| 8 | - | 1 | - |
| 9 | 1 | - | - |
| 10 | - | - | - |
| 11 | - | - | 1 |
| 12 | - | - | - |
| 13 | 1 | - | - |
| 14 | - | - | - |
| 15 | - | - | - |
| 16 | - | 1 | - |
| % unique | 73.9% | 70.7% | 77.8% |
| DC | 0.83 | 0.80 | 0.85 |
| Coverage | 1048 bp | 610 bp | 1122 bp |

| # of Times Haplotype Observed | Asian American n = 47 | | | African American n = 258 | | | Caucasian n = 262 | |
|---|---|---|---|---|---|---|---|---|
| | PLEX-ID | HV1/2 | Full C.R. | PLEX | HV1/2 | Full C.R. | PLEX | HV1/2 |
| 1 | 47 | 47 | 47 | 192 | 187 | 198 | 182 | 170 |
| 2 | - | - | - | 13 | 12 | 16 | 9 | 11 |
| 3 | - | - | - | 6 | 6 | 8 | 4 | 5 |
| 4 | - | - | - | 2 | 1 | - | 2 | 2 |
| 5 | - | - | - | 1 | 1 | 1 | 4 | 4 |
| 6 | - | - | - | - | 2 | - | - | 1 |
| 7 | - | - | - | - | - | - | 1 | 1 |
| 8 | - | - | - | - | 1 | - | - | - |
| 9 | - | - | - | 1 | - | - | - | - |
| 10 | - | - | - | - | - | - | - | - |
| 11 | - | - | - | - | - | - | - | - |
| 12 | - | - | - | - | - | - | - | - |
| 13 | - | - | - | - | - | - | 1 | - |
| 14 | - | - | - | - | - | - | - | - |
| 15 | - | - | - | - | - | - | - | - |
| 16 | - | - | - | - | - | - | - | 1 |
| % unique | 100% | 100% | 100% | 74% | 72% | 77% | 69% | 65% |
| DC | 1.00 | 1.00 | 1.00 | 0.83 | 0.81 | 0.86 | 0.77 | 0.74 |