Evaluation of Degraded Human DNA Samples Using the Illumina® Global Screening Array

David A. Russell, MS* • Carmen Reedy, PhD • Elayna Moreithi, MS • Christina Neal, MS • Mary Heaton, MS • Stephen D. Turner, PhD

ABSTRACT

The Infinium® assay workflow is a genome-wide microarray genotyping assay that utilizes the BeadChip platform.[1] This accurate and flexible microarray technology allows for the ability to interrogate a large number of single nucleotide polymorphisms (SNP) through unlimited loci multiplexing.[2,3,4] Previous work to optimize sensitivity of the assay demonstrated a total DNA input <1.0 ng successfully generated high quality genotyping data.

DNA degradation is a common factor affecting forensic DNA samples that are not stored properly, that have been stored for extended periods of time or that have been exposed to the elements (i.e., sunlight or heat). Current short tandem repeat (STR) typing kits can handle DNA degradation providing, in most cases, partial DNA profiles with moderately degraded DNA. For single nucleotide polymorphism (SNP) genotyping analysis, the quality of the data is critical and loss of SNPs due to degradation will affect the ability to accurately search genealogical databases (GEDmatch, FamilyTreeDNA®, etc.). In addition to a larger goal of assessing the Illumina Global Screening Array (GSA), a whole-genome SNP genotyping method, assessment of the array's ability to handle degraded DNA is of main interest and is the focus for this project.

For this study, genomic DNA was experimentally degraded using ultraviolet C (UV-C) light at defined intervals up to 1.0 J/cm^2. Samples were quantified pre- and post-treatments and typed using GlobalFiler™ to confirm degradation. The STR profiles coupled with the Degradation Index (DI) from the quant method demonstrated that as the dosage increased, so too did the amount of degradation (quantified via visual pattern in the profile and increase in DI). Five (5) samples expressing specific degradation patterns based on STR results were selected for genome-wide SNP genotyping, in duplicate, on the iScan® to evaluate the performance of the Infinium GSA in its ability to accurately type low copy number (LCN) degraded DNA sample types.

Analysis of the SNP data showed similar trends in most aspects. Concordance of the degraded samples did trend down as the dosage amount increased. Concordance was performed between treated samples and a control sample run alongside. One metric used to evaluate the performance of the assay is the call rate, which is a percentage of the total number of SNPs genotyped over the total number of SNP targets in the assay. When evaluating the samples using this metric, the trend was unexpectedly parabolic. As the dosage level increased, the curve trended down then re-bounded and began increasing, resulting in the 0.0 mJ/cm^2 and the 1.0 J/cm^2 samples having very similar call rates.

METHODS

Two Coriell standard DNA samples (NIST: HG001, HG005) were prepared with PCR grade water in 0.25 – 1.5 µL tubes. Samples were normalized to 0.05 ng/µL in a volume of 20 µL, of which 2 µL was used for pre-UV quantification. Once prepared, samples were covered in foil to initiate control treatment of UV-C exposure. Samples were stored at 4°C until UV-C treatments. Samples were removed from 4°C, the foil was removed, and the dosage applied.

Each of the prepared samples were quantified prior to and then following treatment using the Applied Biosystems 7500 real-time PCR system with the Qiagen Investigator Quantiplex Pro quantification kit. The samples were initially diluted to a concentration below the recommended target of 1 ng/µL for the GlobalFiler kit.

Genomic DNA (15 µL) was amplified with GlobalFiler with a 1 µL aliquot used for STR typing on the SeqStudio and 4 µL for the SNP-genotyping. Electrophoresis was conducted as a confirmation that UV-C treatments caused degradation to the DNA. STR data was analyzed using GeneMapper® ID-X v1.6 analysis software. Treated samples that displayed degradation in STR profile interpretation were genotyped in duplicate.

Additionally, one sample from each standard was taken forward with no treatment at 20 ng/µL as a reference for concordance. Genotyping took place using the Infinium HTS assay workflow using the GSA kit and scanned on the Illumina iScan. The data generated was analyzed in GenomeStudio (GS). Concordance and call rate analysis was performed using BCFtools and R.

RESULTS

Sample Quantification and STR Analysis

Quantification of the pretreatment group was mostly consistent with the targeted dilution of 0.05 ng/µL. The degradation index (DI) is also consistent between the DNA standards in the pre-treatment group. Following UV-C treatment, quantification results showed an increase in the DI that, when graphed, appeared exponential (Figure 2). This was true for both samples. Additionally, the quantity of DNA appeared to decrease as the level of UV-C exposure increased (data not shown).

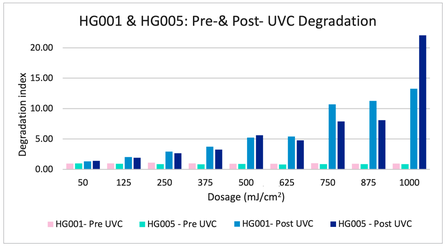

Figure 2: Pre- and Post- UV-C DI values. The graph demonstrates the pre- and post-UV-C treatment DI values for the two DNA standards (HG001 & HG005) across all the determined dosage intervals.

The peak morphology observed in the electropherograms (egrams) was as expected (Figure 3). The smaller amplicons exhibited moderate to high relative fluorescence units (RFU) throughout dye channels, then decreased in RFU with noticeable allelic drop-out in the larger amplicons. The data was analyzed at a minimum analytical threshold (AT) of 50 RFU which maintained a high number of total loci being called. Based on the observed degradation patterns, treatments 0, 125, 375, 625, and 1000 mJ/cm^2 were selected for SNP Genotyping.

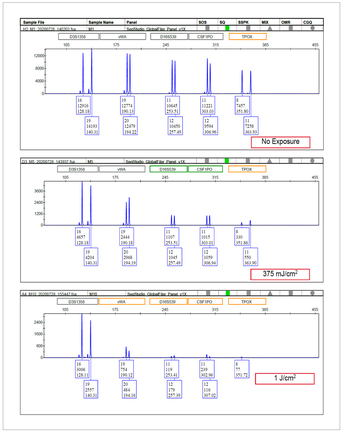

Figure 3: STR Egrams. The egrams show the peak morphology in a single dye channel at various treatment dosages (0, 375, & 1000 mJ/cm^2). No exposure shows a 'normal' distribution of peak heights. 375 mJ/cm^2 treatment shows a 'ski-slope' pattern usually observed with degraded DNA samples. 1000 mJ/cm^2 treatment shows the same sloped pattern in addition to allelic drop-out; demonstrating greater DNA degradation.

SNP Genotyping

The raw genotype call (GTC) files generated by the iScan Data Collection software were converted to variant call format (VCF) files and compiled together for concordance assessment. Each sample (0.2 ng input at 0, 125, 375, 625, and 1000 mJ/cm^2 dosage) was compared to its respective control (20 ng input at 0 mJ/cm^2 dosage). As expected, there was 100% concordance between the 0 mJ/cm^2 samples and the control sample. As the dosage increased, the concordance rate decreased.

The call rates appeared to decrease with the increase in the amount of dosage; however, the call rates rebounded and increased resulting in similar call rates being observed between the untreated (0 mJ/cm^2) and the highest dosed samples (1000 mJ/cm^2) (Figure 4). This was due to an excess of heterozygote to homozygote miscalls.

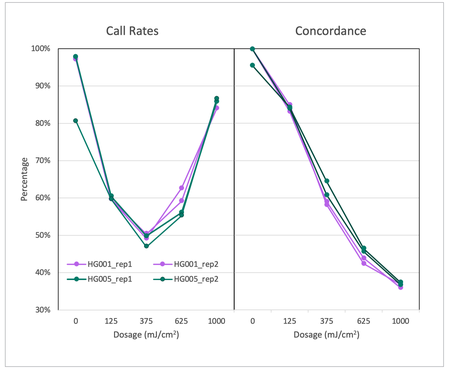

Figure 4: The line graphs show the plotted call rates and concordance as a percent over dosage in mJ/cm^2. Each line represents a replicate of the two DNA standards used in the study.

WORKFLOW

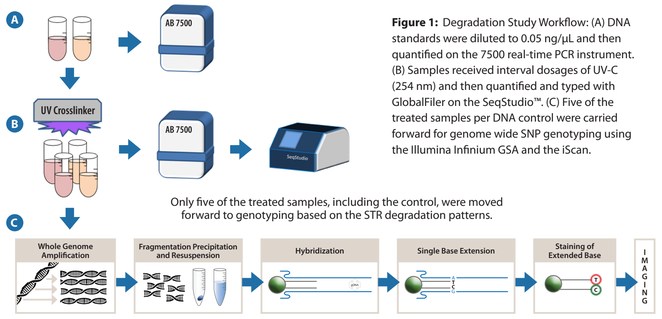

Figure 1: Degradation Study Workflow: (A) DNA standards were diluted to 0.05 ng/µL and then quantified on the 7500 real-time PCR instrument. (B) Samples received interval dosages of UV-C (254 nm) and then quantified and typed with GlobalFiler on the SeqStudio™. (C) Five of the treated samples per DNA control were carried forward for genome wide SNP genotyping using the Illumina Infinium GSA and the iScan.

Only five of the treated samples, including the control, were moved forward to genotyping based on the STR degradation patterns.

DISCUSSION

The severity of DNA degradation in traditional STR analysis can be visualized as it presents an identifiable sloped pattern demonstrating the interruption of the polymerase during amplification. Although this sloped pattern is indicative of degraded DNA, the form of degradation, whether it be fragmentation or DNA lesion (dimers, conformation changes, single-strand breaks, etc.), is not easily known. As the amplification chemistries are fundamentally different between STR and SNP typing, it was hypothesized that the form of DNA degradation could have less of an impact on the genotyping data.

Following a SNP-genotyping run there are many metrics to evaluate, but the call rates typically provide a good assessment of the data quality. Quality of the data is important for downstream analysis specifically when using third-party genealogical databases. The information returned is only as good as the data uploaded. With this study it showed that with degraded DNA the call rate alone is not a way to assess the overall data quality, because call rates were high in severely degraded samples resulting from excess heterozygote to homozygote miscalls.

Using call rates in conjunction with concordance data, the evaluation becomes clearer, demonstrating that the data quality is compromised with more degraded DNA samples. Concordance is a measure of the accuracy of the SNP genotype (GT) calls. The challenge faced in a forensic situation is that unlike experimental tests, there are limited opportunities to have a reference to provide concordance information.

SNP heterozygosity was observed as another potential metric to assess the sample run quality. However, the proportion is dependent on the SNP sites on the array and will vary across arrays. Given that most of the SNPs in the human genome are invariant, the heterozygosity of this array for the control samples was at ~17%. Additionally, ancestry does play a factor in the heterozygosity of the sample. Without confidence in the GT accuracy (concordance) there is lower confidence in the accuracy in the ancestry prediction; therefore, using any established correlations to percent heterozygosity in the GSA array and a certain population is not feasible.

CONCLUSIONS

A greater number of SNP sites that are called allows for more comparable sites when doing downstream analyses (ancestry, phenotype and kinship inferences). Call rates are the number of called sites over the total number of sites available. A call rate of >95% tends to produce high quality and accurate data. However, the data generated here were unexpectedly atypical given the known elements of the experiment. Samples known to have been severely degraded were observed to produce GT data with call rates relatively similar to those of pristine samples. The concordance data validated the trend that was expected, showing that the GT calls were less concordant as the samples became more degraded. Reliance on this analysis is problematic as with forensic cases, most often the DNA evidence is from an unknown individual.

One of the variables that was controlled for in this experiment was the manner of degradation. With the multiple ways a DNA sample can become degraded, controlling the manner for this study was considered an important element in order to understand if there were any potential effects the assay chemistry had on the data outcome. The particular manner of degradation in this experiment could be a possible explanation for the unexpected trend in the call rates.

Further evaluation is warranted, but in the absence of additional metrics, independent of a reference, the call rate alone is not dependable to assess overall sample quality if the quantification data suggests sample degradation given low DNA quantity (0.05 ng/µL).

PATH FORWARD

As an immediate priority, establishing a Degradation Index (DI) threshold under our current quant method will provide a preliminary and sufficient screening method to ensure that samples that are going to generate data unacceptable for SNP genotyping are not carried forward without communicating the potential outcome limitations.

In addition to a DI threshold, more experiments looking at other specific forms of degradation could provide useful information on the appropriate analysis, as well as seeing if the unexpected observations with the call rates are reproducible. From these studies, methods can be evaluated to repair DNA degradation in order to achieve better SNP results.

Ultimately, the continuation of what has been accomplished and any future studies will be performed on the NextSeq™ 2000. Whole genome sequencing can provide a greater depth and breadth of assessment of the SNPs specific to the GSA array and, in its own right, provide a new data set to evaluate in order to establish additional service opportunities when it comes to forensic genetic genealogy.

REFERENCES

[1] Kevin L. Gunderson et al., "Whole-Genome Genotyping," in Methods in Enzymology, vol. 410 (Elsevier, 2006), 359–76, https://doi.org/10.1016/S0076-6879(06)10017-8.

[2] Jian-Bing Fan et al., "[3] Illumina Universal Bead Arrays," in Methods in Enzymology, vol. 410 (Elsevier, 2006), 57–73, https://doi.org/10.1016/S0076-6879(06)10003-8.

[3] Frank J Steemers et al., "Whole-Genome Genotyping with the Single-Base Extension Assay," Nature Methods 3, no. 1 (January 2006): 31–33, https://doi.org/10.1038/nmeth842.

[4] Illumina, "Infinium Assay Workflow," Technology Spotlight: SNP genotyping, 2012.

ACKNOWLEDGEMENTS and CONTACT INFORMATION

This work is supported by Signature Science, LLC through the Center for Advanced Genomics.

David Russell, MS
Signature Science, LLC
1670 Discovery Drive, Suite 240
Charlottesville, VA 22911
drussell@signaturescience.com

Scan for more information