

Optimizing Sensitivity and Validating the Illumina Infinium Assay for Genotyping of Forensically Relevant Sample Types for Investigative Lead Generation

signature
science[®] LLC

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

INTRODUCTION

Forensic genealogy applies enhanced genetic processing techniques (array-based genome-wide SNP genotyping) combined with traditional genealogical research techniques to produce new leads in cases which have gone cold or where traditional investigative means have been exhausted. The use of this technology in investigative forensics has skyrocketed since the 2018 arrest of Joseph DeAngelo as the Golden State Killer.

Most microarray-based genome-wide SNP genotyping takes place under clinical research, providing services that are not adapted to forensically relevant sample types. Furthermore, direct to consumer (DTC) laboratory tests require high quality and quantity DNA. 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 SNPs through unlimited loci multiplexing.^{2,3,4} However, overcoming the 200 ng standard input for this assay is essential for forensic genomics, as it is rare to obtain DNA at such high quantities from forensic samples.

Described here is a study using Illumina's Infinium Global Screening Array (GSA) to show successful and accurate genotyping at very low DNA input levels; shifting the applicability from clinical laboratories to the forensic community. Additionally, to set a standard for validating forensic workflows for generating genome-wide SNP genotyping data, the study design, where applicable, was guided by the current Federal Bureau of Investigation (FBI) Quality Assurance Standards (QAS) for DNA testing laboratories⁵ and the Scientific Working Group on DNA Analysis Methods (SWGDM) Validation Guidelines for DNA Analysis Methods.⁶


The precision and sensitivity of the assay were evaluated using Coriell DNA /NIST standard reference material that has been extensively characterized.⁷ For the sensitivity study, a single reference sample was quantified and diluted to decreasing input amounts ranging from 200 ng—the manufacturer recommended input target, down to 0.2 ng total DNA input—an input amount more consistent with forensic samples. The precision study involved genotyping three Coriell references at 200 ng input for comparison to high quality sequencing data.

Results of the assessment highlight the ability of the array to produce very precise genotyping calls when compared to known reference data and even more so when looking at the concordance between replicates of the same sample. Sensitivity of the assay is crucial to allow for smaller DNA input amounts while remaining precise. The Sensitivity study conducted showed highly concordant (>99%) genotype calls for samples with DNA input greater than 1 ng, and >95% concordance for samples down to the 0.2 ng.

These studies have demonstrated, using a forensic workflow, the Illumina Infinium assay is capable of producing accurate SNP genotyping data for investigative lead generation with a higher sensitivity to lower DNA amounts and geared towards forensic-centric sample types.


ACKNOWLEDGEMENTS and CONTACT INFORMATION

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



Scan for more information

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



METHODS

Workflow and Assay Chemistry

Figure 1: Infinium assay workflow and chemistry.

Day 1
Whole Genome Amplification
gDNA

Day 2
Fragmentation Precipitation and Resuspension
Hybridization
gDNA

Day 3
X-Stain
Single Base Extension
Staining of Extended Base
Imaging
gDNA

Day 1: The Infinium assay workflow takes genomic DNA (200 ng) and amplifies the DNA in an isothermal reaction.

Day 2: The amplified DNA is enzymatically fragmented to an optimal length; a controlled process fragments the DNA into 300 to 600 base pair segments. Samples undergo purification via isopropanol precipitation and then are re-suspended in a buffer that provides the ideal conditions necessary to hybridize to the array. The Infinium array contains millions of beads, each studded with hundreds of the same, small oligonucleotide sequence (~50mer) unique to that bead and specific to a locus of interest in the target genome.

Day 3: The next step in the assay is X-Stain. It is the process of a single base extension and staining. The single base extension (SBE) allows for differentiation of genotypes. **A** The probe oligo sequence flanks the locus of interest on the gDNA. Chain-terminating dideoxynucleotides labeled with either dinitrophenol (DNP) for A and T nucleotides or Biotin for G and C nucleotides are incorporated. After SBE occurs and the ddNTP's are incorporated, the gDNA or target DNA is removed, and the probes are ready for staining. **B** The Biotin-labeled probes are tagged with green fluorescent streptavidin molecule, and the DNP-labeled probes are tagged with red fluorescent anti-DNP antibody. This process applies the specific fluorescent signal to the labeled probes. **C** Next Biotin- and DNP- labeled antibodies are applied to the array. Following multiple iterations of adding fluorescent molecules and antibodies, the fluorescent signal is amplified and becomes strong enough to be imaged by the iScan which uses red and green lasers to excite the fluorophores and measures the signal intensity of each bead to determine genotype.

Experimental Design

The design of this study focused on two metrics. These metrics are critical to assessing the efficacy of Illumina's Global Screening Array (GSA) for Investigative lead generation, which is dependent on the kit's ability to provide accurate genotyping below the recommended DNA input level. One focus area was precision. Samples were quantified using the QIAamp DNA Investigator Kit. Two 200 ng replicates from the three DNA standards (Table 1), were compared against the NIST / Genome-in-a-Bottle (GIAB) gold standard genotypes. Concordance/discordance statistics were calculated for each input amount separately for each replicate, then averaged across all replicates.

Another area of focus was on the sensitivity of the assay. Three replicates of a single sample, HG001/NA12878, were run at inputs of 200ng, 40ng, 20ng, 8ng, 2ng, 1ng, and 0.2ng. Call rate and concordance/discordance statistics were calculated for each input amount separately for each replicate, then averaged across all replicates at each input.

In both cases, "bad" SNPs that did not genotype in a single sample were removed from each analysis.

Table 1: Samples that were used in this study

nist_id	coriell_id	coriell_name
HG001	NA12878	NA12878
HG002	NA24385	AJSon
HG005	NA24631	ChineseSon

DEFINITIONS
Concordant: The genotype for the lower input sample is identical to the 200 ng genotype.
Discordant: The genotype for the lower input sample differs from the 200 ng genotype.
Concordance rate (conc.rate): The total number of concordant sites divided by the total.
Discordance rate (disc.rate): The total number of discordant sites divided by the total.
Missing either: This shows the number/percent of SNPs where genotypes were missing in either the 200 ng sample or the lower input sample. In these cases, the genotypes cannot be concordant or discordant.
Concordance at called sites (conc.calledsites):=(1-Discordance Rate). The concordance rate measures exactly how often the genotype for the lower input sample is identical to the genotype for the 200 ng sample. If one or the other is missing, the genotypes are not concordant. However, they are not discordant either (it is unknown whether they are concordant or discordant). The "Concordance at called sites" is a looser definition of concordance that measures how often two genotypes are the same conditioned on genotypes being called in both samples.

RESULTS

Sensitivity

Prior to calculating any call rate or sensitivity data, **8,345** "bad" SNPs were removed. Below are tables that show the calculated call rates and concordance/discordance statistics for each input level averaged across all three replicates.

input	total	called	missing	callrate
200.0	635930	635452	478	99.92%
40.0	635930	635454	476	99.93%
20.0	635930	635413	517	99.92%
8.0	635930	634782	1148	99.82%
2.0	635930	634380	1550	99.76%
1.0	635930	631943	3987	99.37%
0.2	635930	617901	18029	97.16%

Figure 2: Call rates for each input level, averaged across all three replicates. Called, missing, and callrate are averaged across replicates.

input	total	conc.n	disc.n	conc.rate	disc.rate	conc.calledsites	missing_either.n	missing_either.pct
40.0	635930	635112	2	99.8711%	0.000%	100.000%	816	0.128%
20.0	635930	635065	4	99.864%	0.001%	99.999%	861	0.135%
8.0	635930	634421	26	99.763%	0.004%	99.996%	1484	0.233%
2.0	635930	633999	34	99.696%	0.005%	99.995%	1896	0.298%
1.0	635930	631370	222	99.283%	0.035%	99.965%	4338	0.682%
0.2	635930	615564	2000	96.797%	0.315%	99.685%	18366	2.888%

Figure 3: Concordance, discordance, and concordance at called sites rates for each input level, separately for each replicate. Dashed line shows 98% concordance. Dotted line shows 95% concordance.

Precision and Accuracy

Results from the precision study comparing GSA genotypes to NIST/GIAB samples are shown below. Prior to calculating any precision data, **7,185** "bad" SNPs were removed. Further, 102 SNPs where either the GIAB genotype or the GSA genotype were not represented as diploid biallelic single nucleotide polymorphisms were removed.

id	total	conc.n	disc.n	conc.rate	disc.rate	conc.calledsites	missing.n	missing.pct
HG001	593783	589007	342	99.196%	0.058%	99.842%	379	0.064%
HG002	548279	543780	388	99.180%	0.071%	99.929%	299	0.055%
HG005	556487	552118	289	99.215%	0.052%	99.948%	286	0.051%

Table 4: Concordance statistics for each sample, averaged across replicates

The data above compared the GSA genotypes for each sample/replicate to its corresponding NIST/GIAB genotype. The table below shows within-sample precision, comparing genotypes for the two duplicate samples against each other for all SNPs used in the analysis here (recall, excluding SNPs that did not genotype in any sample, those not having a genotype in the NIST/GIAB sample, and those that were not biallelic in either of the callsets).

id	total	conc.n	disc.n	conc.rate	disc.rate	conc.calledsites	missing.n	missing.pct
HG001	593,783	593,146	3	99.8927%	0.0005%	99.9995%	634	0.1068%
HG002	548,279	547,776	4	99.9083%	0.0007%	99.9993%	499	0.0910%
HG005	556,487	556,008	3	99.9139%	0.0005%	99.9995%	476	0.0855%


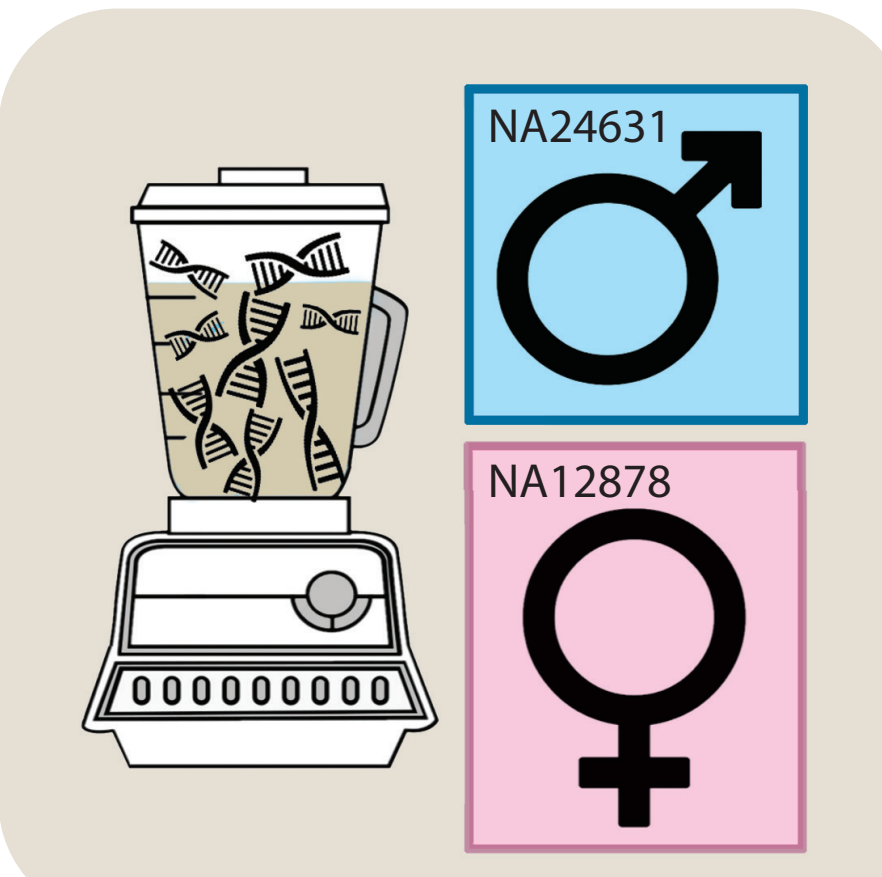

Table 5: Within-sample precision (reproducibility)

FUTURE WORK

SPECIES SPECIFICITY
This study will validate whether the SNP probes have any affinity to human targets. Genomic (gDNA) DNA from seven (7) non-human animals, bacteria and fungi will be diluted to a total DNA target amount of 200 ng and genotyped in triplicate.

MIXTURES
In casework the levels of DNA are varied high to low, often the latter. Contamination and mixtures are always a possibility. Although it is not possible to deconvolute mixed samples, it is still pertinent to be able to identify what a mixed sample would look like. A mixture study was performed using standards NA12878 and NA24631 at varying ratios (M:F) 1:0, 9:1, 3:1, 1:1, ...). Analysis of the mixture data is ongoing, and the goal is to identify patterns that may be used to recognize possible mixed samples at different ratios.

MOCK SAMPLE TYPES
Precision and sensitivity studies were performed using gDNA. Further evaluation of typical forensic sample types (blood, semen, saliva, touch etc.) to simulate potential casework collections is of interest to examine a forensic workflow. Understanding the limitations to produce useful genotyping data from potentially problematic samples is important information for agencies looking to utilize investigative genetic genealogy.



DISCUSSION

Sensitivity study call rates are >99% and >95% for inputs ≥1 ng and 0.2 ng, respectively. Additionally, results are highly concordant down to 0.2 ng: <0.001% discordance for all replicates down to 1 ng, and <0.5% discordance even down to 0.2 ng.

The precision study also exhibited excellent results. When comparing samples against the NIST/GIAB sequencing data, average concordance rates are 99.2% across all samples. Comparing duplicate samples to each other resulted in a concordance rate >99.8% across all three samples. The intra-sample reproducibility is extremely high. Among a total of 1,698,549 total genotypes evaluated across 3 samples, there were a total of only 10 total discordant genotypes.

REFERENCES

¹ Gunderson, Kevin L. Frank J Steemers, Hong Ren, Pauline Ng, Lixin Zhou, Chan Tzan, Weihua Chang, et al. 2006. "Whole-Genome Genotyping" Methods in Enzymology 359–376.

² Fan, J., K.L. Gunderson, M. Bibikova, J.M. Yeakley, J. Chen, E. Wickham Garcia, L.L. Lebruska, M. Laurent, R. Shen, and D. Barker. 2006. "Illumina Universal Bead Arrays." Methods in Enzymology 57–73.

³ Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen & Kevin L Gunderson. 2006. "Whole-genome genotyping with the single-base extension assay." Natur Methods 3 (1): 31–33.

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

⁵ Federal Bureau of Investigation. 2000. "Quality Assurance Standards for Forensic DNA Testing Laboratories." Forensic Science Communications 2 (3): 29.

⁶ Scientific Working Group on DNA Analysis Methods (SWGDM). 2016. "Validation Guidelines for DNA Analysis Methods." December 5.

⁷ Zook Justin M., Catoe David, McDaniel Jennifer, Vang Lindsay, Spies Noah, Sidow Arend, Weng Ziming, et al. 2016. "Extensive sequencing of seven human genomes to characterize benchmark reference materials." Scientific Data 1–26.