



**The author(s) shown below used Federal funding provided by the U.S. Department of Justice to prepare the following resource:**

**Document Title:            Genome-Wide Forensic DNA Analysis**  
**Author(s):                    Lynn B. Jorde, Ph.D.**  
**Document Number:        251934**  
**Date Received:             August 2018**  
**Award Number:            2012-DN-BX-K037**

**This resource has not been published by the U.S. Department of Justice. This resource is being made publically available through the Office of Justice Programs' National Criminal Justice Reference Service.**

**Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.**

National Institute of Justice  
Final Report  
11/30/15

Genome-Wide Forensic DNA Analysis  
Award # 2012-DN-BX-K037

Prepared For:  
**National Institute of Justice**  
810 Seventh Street, NW  
Washington, DC 20001

**Principal Investigator**

Dr. Lynn B. Jorde, Ph.D, Professor.  
75 South 2000 East, Salt Lake City,  
Utah, 84112-5330.  
Phone: 801-581-4566.  
Fax: 801-585-9148.  
E-mail: lbj@genetics.utah.edu

**DUNS / EIN**

DUNS 009095365  
EIN 87-6000525

**Recipient Identifying Number**

56200018

## **Project Abstract**

The ability to identify individuals by comparison of DNA profiles from questioned biological samples to those of relatives of the presumed source of the sample is an application of central importance to forensics. Likewise, accurate analysis of kinship by DNA is a biometric tool of great potential importance to immigration and national security. While comparison to family references is a primary method for missing persons or disaster victim identification, its present utility is greatly constrained by the need for multiple immediate relatives to achieve a high certainty of identification. Lack of close family reference samples has proven a key limitation for the identification of victims of major US disasters and in the US national missing persons program. The ability to detect distant relationships between people with high certainty would resolve a fundamental weakness in major applications of forensic DNA analysis, including security/immigration biometrics.

We have developed a computational method (ERSA) that uses high-density SNP genotypes to link relatives at 6<sup>th</sup> degree relationships with high confidence. This method has proven effective on pristine DNA. The main obstacle limiting the application of ERSA to forensic DNA is that it requires high-density, high-quality SNP genotype data, and such data cannot currently be generated from many forensic DNA samples: they are frequently too fragmented, chemically damaged, inhibited, or limited in quantity. While progress has been made sequencing samples from formalin-fixed, paraffin-embedded (FFPE) tissue or ancient DNA (aDNA) using high-throughput next-generation sequencing, solutions are often prohibitively expensive for application in forensic DNA laboratories.

To meet this need, we adapted and improved techniques for amplifying degraded DNA specifically for SNP genotyping using microarray technology. We identified and quantified the error patterns in the genotyped high density SNP data from limited and fragmented DNA samples. We compared these error patterns to those observed from genotyping laboratory-grade high quality samples and we used a multivariable logistic regression model to optimize true calls from false genotypes. Our aim was to affordably allow genotyping of SNPs at high density in DNA samples present at concentrations and

fragmentation levels expected in a forensic setting, and do so at accuracies above and beyond those necessary for ERSA relationship detection.

After artificially fragmenting high quality DNA to 100bp or 300bp average size, we tested various sample pre-processing pipelines with the Infinium HD Assay on the Illumina HumanOmniExpress-FFPE or HumanOmni2.5 BeadChips. The Illumina-supported HD FFPE Sample Solution was explored, as was a novel single-stranded DNA sequencing library concatemerization protocol. From artificially fragmented DNA, very high call rates and accuracy were achieved at all inputs  $\geq 20$  ng. These results were equivalent to those using the suggested input of 200 ng standard DNA. At 2 ng quantities of fragmented DNA, on both BeadChips tested, our library preparation and concatemerization technique achieved comparable or better call rates and concordance rates with a high quality control than similarly low concentration samples processed through the standard Infinium workflow. As this input is two orders of magnitude less than is suggested by the Infinium genotyping microarray protocols, and as the resulting SNP data supports ancestry and relationship estimation as accurately as data from normal amounts of high quality DNA, this represents a substantial step forward in forensic high density genotyping.

At subnanogram quantities our novel methodology performed above expectations, with calls rates at  $\sim 58\%$  and concordance rates of 80%. These results represent at least 57% more calls and a substantial (up to 20%) increase in accuracy relative to the same input with the Illumina HD FFPE Sample Solution on the HumanOmni2.5 BeadChip. From 200 picograms of 100bp fragmented DNA, using our optimized laboratory protocol for genotyping of degraded samples and a multivariable logistic regression model to select true calls from false genotypes, we achieved concordance rates upwards of 88% while retaining approximately one million SNPs for downstream analysis.

## TABLE OF CONTENTS

<b>Title Page</b>	<b>1</b>
<b>Abstract</b>	<b>2</b>
<b>Table of Contents</b>	<b>4</b>
<b>Executive Summary</b>	<b>5</b>
<b>Introduction</b>	<b>14</b>
Statement of the Problem	<b>14</b>
Review	<b>14</b>
Objectives	<b>16</b>
<b>Methods</b>	<b>18</b>
Research Design	<b>18</b>
Materials and Procedures	<b>18</b>
<b>Results</b>	<b>23</b>
Statement of Results	<b>23</b>
Table 1	<b>32</b>
Table 2	<b>33</b>
Table 3	<b>34</b>
Table 4	<b>35</b>
Figure 1	<b>36</b>
Figure 2	<b>37</b>
Figure 3	<b>38</b>
Figure 4	<b>39</b>
Figure 5	<b>41</b>
Figure 6	<b>42</b>
Figure 7	<b>43</b>
Figure 8	<b>44</b>
Figure 9	<b>45</b>
Figure 10	<b>47</b>
Figure 11	<b>48</b>
Figure 12	<b>49</b>
Figure 13	<b>50</b>
<b>Conclusions</b>	<b>51</b>
Discussion of findings	<b>51</b>
Implications for policy and practice	<b>54</b>
Implications for further research	<b>56</b>
<b>References</b>	<b>59</b>
<b>Dissemination of Research Findings</b>	<b>62</b>

## **Executive Summary**

The ability to identify individuals from traces of their DNA is invaluable in forensic identification of the remains of missing persons, victims of mass disasters, and suspects in criminal investigations. This DNA methodology faces two major challenges to identify the source of the forensic sample: obtaining genotypic information from degraded forensic DNA samples, and matching that genetic information with similar information collected from individuals whose identities are known. Forensic scientists have generally met both challenges with success: genotypes at the CODIS STR loci are routinely retrieved from nanogram amounts of degraded DNA, and those genotypes are sufficient to connect the sample with a reference DNA sample, provided one exists (reviewed in Jobling and Gill 2004). Should preservation conditions be less optimal, however, STR loci fail to genotype and the power of the marker set to identify or exclude individuals with sufficient match probabilities is compromised.

In many situations a known reference DNA sample from the individual who left the forensic sample (a “direct” reference) is not available. It is sometimes possible to link forensic DNA to samples from their first- or second- degree relatives (i.e. parents, children, or full siblings), but this typically requires multiple such relatives to permit a strong inference of relatedness. For example, the International Commission for Missing Persons (ICMP) has had some success with this technique. Unfortunately, the requirement for multiple immediate relatives is often not met. Thus the ICMP has thousands of DNA profiles for which identifications cannot be made because of the lack of immediate family reference samples. This problem is common in situations where the scope of a disaster or armed

conflict, or the passage of time, has made it difficult to find potential close relatives. Our inability to detect distant genetic relationships with confidence renders many forensic DNA samples unidentifiable and leaves their sources anonymous.

If genotypic information could be used to reliably connect distant relatives, the number of people who could provide a useful reference DNA sample would increase dramatically, since people generally have many more distant relatives than first-degree relatives. We have developed a method that can identify 6<sup>th</sup> degree relationships with very high confidence, with accuracy through 9<sup>th</sup> degree relationships (ERSA; Huff et al. 2011, Li et al 2014). This represents a qualitative leap beyond previous capabilities, though it requires high-density, high-quality SNP genotype data, and such data cannot currently be generated from many forensic DNA samples: they are frequently too fragmented, chemically damaged, inhibited, or limited in quantity for this technology. Most forensic research on SNP markers has focused on PCR-targeted genotyping of modestly-sized marker panels (dozens to a few hundred) and/or better preserved samples.

SNP microarrays require large amounts of high quality DNA for genotyping. In general, whole-genome amplification (WGA) from nanogram amounts of high-quality DNA does not prevent genotyping calls of high quality, call rate, and concordance with the original genomic DNA sample (Jasmine et al. 2008; Maciejewska et al. 2013; Tzvetkov et al. 2005; Xing et al. 2008). Given initial high quality, but not quantity, Multiple Displacement Amplification (MDA) protocols ( $\Phi$ 29 DNA polymerase) seem to produce abundant amounts of DNA without systematic amplification bias in the subsequent genotypes of MDA sample.

When DNA fragments have been degraded to segments much less than 2Kb, however,  $\Phi$ 29 polymerase's high processivity is no longer as useful, with many fewer new strand branches being available. In addition, potentially numerous off-target strands are synthesized not from the template but from the primer pool (Mead et al. 2008). If true sample fragments are successfully enriched, residual unfragmented long strands in the sample might have benefited disproportionately from amplification and the genotypes typed might reflect this. The oligonucleotide hybridization probes used in modern high-density arrays are no more than 50 bp long. If there was a way to amplify the small DNA fragments without bias, these probes should be capable of assaying SNP genotypes even in DNA too damaged to support the PCR required for many STR loci.

We explored two options, one proprietary and one freely disseminated, available to amplify degraded DNA. The first is the Illumina Infinium HD FFPE Sample Solution, optimized for FFPE-extracted DNA. It is inherently a pre-processing step inserted before WGA. While the protocol is proprietary, documentation suggests that the DNA Restoration Solution assay involves a DNA repair step, a single round of DNA replication, and a DNA ligation step. Once repaired, the DNA fragments should be long enough for MDA protocols, namely the Illumina Infinium assay used in the BeadChip high-density SNP panels. The second option is to incorporate progress in next-generation sequencing library creation. As sequencing platforms utilize massively parallel short read technology, and this technology requires large amounts of DNA fragments for detection, PCR amplification of target fragments ligated to platform-specific adapters is routinely performed. Particular to our goals,



unbiased library amplification of low quantity and degraded ancient DNA samples has been successful (Dabney et al. 2013; Fu et al. 2013; Meyer et al. 2014; Rasmussen et al. 2014).

To understand how well high-density SNP genotyping tolerates fragmented and low concentration DNA samples, we selected the Illumina BeadArray Microarray technology as our platform of choice. Two BeadChips were chosen: the HumanOmniExpress-FFPE and the HumanOmni2.5. The HumanOmniExpress-FFPE BeadChip contains 693,543 loci determined by Illumina to yield optimum performance with FFPE samples using the Infinium HD FFPE Assay. The HumanOmni2.5 BeadChip is a higher-density array with 2,391,739 markers. While the HumanOmni2.5 SNP selections are not optimized for FFPE DNA, the protocol uses the Infinium LCG Assay that contains the same reagent improvements as the Infinium HD FFPE Assay. The required input DNA for both Assays is 200 ng.

For our samples, we initially split DNA samples from three human individuals into fragmented (mechanically sheared to 300 bp and 100 bp fragments with a Covaris S2 Focused-ultrasonicator) and non-fragmented (high molecular weight) categories. Within each category we created panels of human DNA samples varying in PicoGreen fluorescence-determined amounts (from 200 ng to 200 pg). A further subset was selected for either DNA repair via the Infinium FFPE DNA Restoration Solution (readily incorporated into both BeadChip Assays prior to WGA), or for processing in various double-stranded (dsDNA) or single-stranded (ssDNA) next-generation sequencing library protocols. After these sequencing libraries were created and amplified, we ligated together

the library products into long concatemers of alternating target-adaptor fragments for subsequent MDA amplification in the Infinium assays. Alternatively, we bypassed the WGA step and placed libraries we had amplified to plateau twice onto the BeadChips directly just prior to genotyping. We determined genotype call and concordance rates between experimentally limited and degraded samples versus original control DNA samples from the three individuals.

The dsDNA libraries were processed through a modified NEBNext ChIP-Seq protocol with an Illumina 'Y-adaptor' of full-length P5 adaptor and an indexed P7 adaptor. This protocol is already optimized for the low input amounts characteristic of chromatin immunoprecipitation experiments. dsDNA libraries were amplified with NEBNext High-Fidelity 2X PCR Master Mix. The ssDNA libraries were processed through a novel high-sensitivity single-stranded approach (Gansauge and Meyer 2013). This protocol offers many benefits when working with degraded genetic material – biotinylation of the DNA fragments avoids the loss of molecules during purification steps and double-stranded molecules with single-strand breaks or end modifications usually lost during dsDNA library preparation are saved here by rescuing the undamaged complementary strand. We used P5 and P7 indexed primers for primary library amplifications and we altered the spacer arms of the single-stranded adaptor CL78 to lower its cost. ssDNA libraries, and all double-amplified libraries, were amplified with the high-fidelity AccuPrime Pfx polymerase. All DNA samples were incubated with NEB USER enzyme prior to library construction, to facilitate future work with environmentally degraded samples. Deamination of cytosine to uracil and the presence of abasic sites is expected in such DNA samples (Briggs et al. 2007)

and the USER cocktail offers a level of repair for this damage. Post-amplified library molecules were ligated together with T4 DNA ligase and T4 PNK, followed by cleanup with spin columns.

For all three individuals, genotype call rates of  $\geq 99\%$  were observed for samples of  $\geq 20$  ng input restored with the Illumina DNA Restoration Solution, with concordance rates of  $>99\%$  to non-restored samples of 200 ng input, irrespective of fragmentation. This implies that the standard Illumina assay is robust down to one-tenth of the recommended input. For lower input amounts, we focused our attention on samples from just one individual. An initial input of 10 ng of DNA, followed by restoration, reduced call rates to 95% and 83 % for the FFPE and Omni2.5 Beadchips, respectively. Call rates dropped further to 84% and 77% at 2 ng restored input, and 56% and 38% at 0.2 ng restored input. Blank samples ('restored' ultrapure water) had call rates of 51% for the HumanOmniExpress-FFPE and 31% for HumanOmni2.5. Concordance rates with the high quality 200ng control were 94-99% for samples  $\geq 2.0$  ng.

As the Infinium assays appear to have a basement input amount of 2 ng, we explored our library preparation and concatemerization technique as an alternative at this and lower sample amounts. On the HumanOmniExpress-FFPE BeadChip, ligated sequencing libraries from 2 ng of fragmented DNA raised the call rate from 84% (for restored 2 ng input) to as high as 90% for 100 bp fragmented DNA made into a ssDNA library; the restored sample already had a high (98%) concordance rate and this was improved to 99%. On the HumanOmni2.5, similarly treated 2 ng samples with the ssDNA library technique raised the

call rate from 77% to 85%, and the concordance rate from 94% to 98%. When initial input DNA was 200 pg on the HumanOmni2.5 BeadChip, the concatemerized ssDNA libraries increased call rates from 38% to 58%, and concordance rates from 60% to 80%. In summary, when applying our ligated library method, we recovered more SNPs with higher accuracy with both 2 ng and 200 pg of fragmented DNA input when compared to the FFPE Restore Protocol. In particular, the ssDNA library method with ligation using 200 pg of DNA allowed genotyping of >1.3 million SNPs on the HumanOmni2.5 at 80% concordance with the control.

As an alternative metric to concordance rates, we performed various Principal Component Analyses of the shared called loci between the low-input and degraded samples and the high quality control samples. On both BeadChips tested, the concatemerized ssDNA libraries clustered closer to the control samples than either comparable dsDNA libraries, restored samples, or the blank sample(s). This indicates that the method, at least at this scale, is capturing the true genetic signature of the sample, particularly at levels where the FFPE Restore protocol does not.

The non-WGA library preparation method failed, with call and concordance metrics similar to the restored blanks. One possible reason is that our PCR-amplified library products never reached the DNA quantities produced from MDA. In our experience the Infinium Assay post precipitation and resuspension of the MSA5 plate generates ~100 µg of DNA, while post-cleanup double-amplified libraries produced at most ~1.5 µg of DNA. The

BeadChip oligo probes do not appear to find and hybridize targets properly at these low input amounts, even though these libraries were properly buffered before hybridization.

It may be possible to further discriminate between more and less reliable genotypes returned by Illumina's genotype calling software. We investigated two approaches – a general filtering based on the blank samples and a ranking of calls based on call-by-call quality metrics. On the hypothesis that loci called in the absence of target DNA are most likely to be erroneous at small sample concentrations, we examined the effect of excluding from analysis the genotypes of all SNPs that were also called in the restored water blanks. For samples on the FFPE-optimized HumanOmniExpress Bead Chip, this data reduction step did not result in any concordance rate increase but did produce a 50-51% drop in call rate. On the HumanOmni2.5 BeadChip, this data reduction approach fared slightly better. For samples at  $\geq 2$  ng input, concordance rates increased by 1%, with only a 30% drop in call rate. At 200 pg input, samples had a 2-3% increase in concordance, with a 32-37% drop in call rates. The ligated ssDNA library from 200pg continued to perform well, with a final 82% concordance out of 933,694 called SNPs.

Alternatively, we estimated a multivariable logistic regression model to discriminate correct vs. incorrect genotype calls based on quantitative scores that are produced by Illumina's GenomeStudio for every genotyped SNP. We compared the genotype calls and quality scores of the ligated dsDNA library from 200 pg (which had a 75% concordance rate to the control) to the quality scores of the control sample. We then used that model to compute a discriminator variable from the scores returned for the ligated ssDNA library

from 200 pg, which showed an 80% concordance rate among called genotypes. We were able to yield sets of approximately one million called SNPs with a substantially improved concordance rate of 88%. This result implies that the Illumina quality scores contain additional information that can discriminate accurate from inaccurate genotype calls for genotypes collected from forensic samples.

High-density SNP genotyping of forensic samples will open the door to many forensically useful inferences: precise estimates of a DNA donor's biogeographical ancestry, the ancestries of his/her parents, even the degree and timing of population admixture events in the donor's past; estimation of any phenotypes previously identified in genome-wide association analyses; and powerful, accurate inferences of even distant relationships, using a method we developed (Huff et al. 2011, Li et al. 2014). We foresee the techniques developed here to be integral when the passage of time or scope of the deaths has made identifying close relatives impossible or has caused the DNA to degrade to a point where typical STR markers fail to amplify. Work in progress is applying this methodology to five 70-year old forensic-quality WWII bone remains, with known modern relatives, and six remains from missing persons discovered between 1992-1995, with unreported provenance.

## **Introduction**

### **Statement of the problem**

Currently, high-quality and high-density SNP genotype data cannot be consistently obtained from small amounts of degraded DNA samples. DNA quality and concentration remain a bottleneck both in SNP microarray and next-generation sequencing library construction. While sequencing libraries can be generated from degraded formalin-fixed, paraffin-embedded (FFPE) tissue (Hadd et al. 2013; Menon et al. 2012; Yost et al. 2012) and ancient DNA samples (Dabney et al. 2013; Fu et al. 2013; Meyer et al. 2014; Rasmussen et al. 2014), success is not assured for each sample and sequencing costs remain prohibitively expensive for routine application in forensic DNA laboratories. Generating high-density SNP genotype data from such samples remains a goal in identification of human remains, since high-density SNP genotype data allow reliable and accurate detection of relationships beyond the first degree (Huff et al. 2011; Li et al. 2014). Many factors need to be taken into account when applying high-throughput genomic approaches to forensic or ancient DNA (Briggs et al. 2007). In particular, the amounts of DNA in forensic samples and DNA fragment sizes are often below the lower limits of current technologies.

### **Review**

Over the past decade, the need to obtain genotypes from small amounts of sample DNA has prompted the development of whole genome amplification (WGA) steps prior to input on genotyping arrays. In general, WGA from nanogram amounts of high-quality DNA does not prevent genotyping calls of high quality, call rate, and concordance with the original genomic DNA sample (Jasmine et al. 2008; Maciejewska et al. 2013; Tzvetkov et al. 2005; Xing et al. 2008). The WGA method of choice has been Multiple Displacement Amplification (MDA), which most commonly uses the  $\Phi$ 29 DNA polymerase, as in the GenomiPhi and REPLI-g kits (GE Healthcare, Piscataway, NJ and Qiagen, Inc., Hilden, Germany, respectively). While allelic imbalance and allele drop-in/drop-out (in particular loss of

heterozygosity) have been a concern, most studies have not detected any systematic amplification bias in the subsequent genotypes of MDA samples (though see Han et al. 2012 and references therein for non-uniformity discussions).

When DNA fragments have been degraded to segments much less than 2Kb, however,  $\Phi$ 29 polymerase's high processivity is no longer as useful, with many fewer new strand branches being available. In addition, potentially numerous off-target strands are synthesized not from the template but from the primer pool (Mead et al. 2008). This suggests that alternative methods of WGA would be more appropriate for fragmented DNA. Indeed, when typing naturally or artificially degraded DNA samples on high-density SNP arrays (Al Safar et al. 2011; Mead et al. 2008) or low-density polymorphic loci (Maciejewska et al. 2013), PCR-based WGA methods such as GenomePlex (Sigma-Aldrich) work well and perform better than MDA (Croft Jr et al. 2008). Unusually, in two studies, one typing STRs (Ballantyne et al. 2007) and the other typing low-density SNP TaqMan panels (Giardina et al. 2009), MDA-based WGA performed well not only on low copy-number DNA, but on artificially degraded DNA as well. In both cases, however, residual unfragmented long strands might have been the main targets of amplification, explaining their apparent success.

Novel MDA methods have been applied to fragmented DNA. Short DNA fragments can be ligated together first, creating long concatemers for subsequent  $\Phi$ 29 amplification. This tactic has been used with plasma-circulating DNA (Li et al. 2006) and is applied commercially by Qiagen, Inc. (e.g. the REPLI-g FFPE product). Alternatively, Rolling Circle Amplification (RCA) exploits circularized DNA fragments to generate very long repetitive strands of synthesized DNA. Tate et al. (2012) utilized a primase/helicase based isothermal WGA on short 100 bp and 1,200 bp ssDNA products that had been circularized. While amplification was successful, RCA did not perform better than other WGA methods. Additionally, when applying their circularization and RCA protocol to STR typing on experimentally degraded and low copy- number DNA samples, no STR loci successfully amplified. Tate et al. (2012) suggest that the multiple enzymatic cleanup steps required effectively eliminated any endogenous DNA remaining in their samples.



The Illumina Infinium HD FFPE Sample Solution is an Illumina-supported process for achieving high-quality genotypes from FFPE samples. It has been experimentally shown to improve the performance of FFPE samples on Human CytoSNP FFPE-12v2.1 arrays (Hosein et al. 2013). It incorporates an initial sample quality-control (QC) test based on qPCR of a known locus, a proprietary DNA repair protocol, and an FFPE-optimized Infinium HD assay. Samples that pass QC are processed through the FFPE Restore Protocol (which requires 100 ng input DNA), and a resulting 200 ng DNA is processed with the Infinium HD FFPE Assay. The assay combines a proprietary WGA step with direct, array-based hybridization and enzymatic scoring of SNP loci by allele-specific single-base extension. The assay is matched with FFPE-optimized BeadChips. The HumanOmniExpress-FFPE BeadChip contains the 693,543 loci that Illumina determined yielded optimum performance with FFPE samples. Illumina states that artificially degraded FFPE DNA that yielded call rates of 40-70% prior to application of the Infinium HD FFPE protocol led to call rates of >90% and concordance rates of >99% with genotypes obtained from high-quality control DNA of the same individuals (Illumina Pub No. 370-2010-009).

The Illumina HumanOmni2.5-8 BeadChip is a higher-density array with 2,391,739 markers. It offers coverage of common and rare SNPs from the 1000 Genomes Project (1kGP; minor allele frequency, MAF > 2.5%) at a lower price per SNP than the HumanOmniExpress-FFPE BeadChip. While it is not optimized for FFPE DNA, it uses the Infinium LCG Assay, which contains the same reagent and protocol improvements as the Infinium HD FFPE Assay. Additionally, the Infinium HD FFPE Sample Solution can be used to restore DNA prior to using the LCG Assay.

## **Objectives**

Both assays require 100 ng of DNA (or 200 ng input without restoration), which is still more than is found in many forensic samples. The SNP-specific hybridization oligonucleotides on Illumina BeadChips are just 50 nucleotides long, so they may be able to assay fragmented DNA, but it is unclear how well the overall assay tolerates fragmented

DNA. To better understand these limitations, we created panels of human DNA samples varying in amount (from 200 pg to 200 ng) and degree of fragmentation (from high molecular weight DNA down to 100 bp mean fragment size; Table 1). We determined genotype call and concordance rates between experimentally limited and degraded samples versus control DNA samples utilizing the Infinium HD FFPE and Infinium LCG Assays. We explored replacing the restore protocol with ligated next-generation sequencing libraries prior to WGA, as well as replacing the whole Infinium DNA preparation protocol with a sequencing library amplification method to produce DNA for BeadChip hybridization.

## Methods

### Research Design

For genotyping experiments on HumanOmniExpress-FFPE or HumanOmni2.5 BeadChips, we initially split DNA samples from three individuals into fragmented (sheared to 300 bp and 100 bp fragments) and non-fragmented categories. Within each category we varied DNA input amount in steps down from the manufacturer-suggested 200 ng and attempted DNA repair (via the Infinium FFPE DNA Restoration Solution). We then prepared the samples either by directly inserting them into the whole-genome amplification step of the Infinium Assay, by creating sequencing libraries that had been ligated together prior to the Infinium WGA step, or by bypassing the WGA step and placing amplified libraries onto the BeadChips for subsequent genotyping (see Table 1 for an overview).

### Materials and Procedures

#### *DNA Samples*

DNA for all three individuals was extracted between 1996 and 1997 from whole blood using Puregene kits (Gentra Systems Inc) and stored in TE buffer at -80°C. Individual A was a Brahmin of the Telegu language group, from Andhra Pradesh, India (Bamshad et al. 1998). Individuals B and C were Kattunayakans, of the Kannada language group, from a tribal hunter/gatherer population in Kerala, India (Watkins et al. 1999). Analyses of 45 unlinked autosomal STR loci indicated that the average genetic distance between these two populations was 2.3% (Watkins et al. 2005).

DNA from Individual A was chosen because this person has been historically used in our laboratory as a high quality reference sample typed in multiple genotyping systems. DNA from individuals B and C were chosen because the blood was collected during a field season in a tropical location and refrigeration was not available. These two DNA samples exhibited substantial smearing after whole DNA gel electrophoresis and we initially believed they would offer lower genotyping success than Individual A (even following artificial fragmentation – see below).

### DNA Input Measurements

We initially determined DNA concentration from a combination of a qPCR assay with a 202 bp amplicon and QuBit dsDNA HS Assay kits, which use the PicoGreen fluorescent nucleic acid staining dye. However, the qPCR assay is inaccurate for samples fragmented to 100 bp, so we relied on PicoGreen fluorescence for all further measurements.

### Fragmentation Protocol

To fragment DNA to 300 bp, we performed one round of shearing on a Covaris S2 Focused-ultrasonicator (intensity 5, duty cycle 10%, cycles/burst 200, treatment 60 seconds), followed by a double-sided SPRI selection (Rodrigue et al. 2010). We tested our fragmented DNA to confirm that PCR products shorter than 300 bp still amplified consistently, but ~900 bp and ~1800 bp products (arbitrarily targeting the human ALB gene) failed to amplify. To obtain 100 bp fragmentation, we performed one round of Covaris S2 shearing (intensity 5, duty cycle 20%, cycles/burst 200, treatment 156 seconds), one round of dSPRI size selection, and two more rounds of Covaris shearing at progressively longer treatment times (intensity 5, duty cycle 10%, cycles/burst 100, treatment 600 seconds). These samples sporadically amplified the 202 bp amplicon from the qPCR and never amplified larger products.

### Size Range Quality Control

All artificially fragmented genomic DNA samples and all sequencing libraries were checked to ensure correct size ranges on electropherograms with an Agilent 2100 Bioanalyzer using a DNA High Sensitivity Chip or an Agilent 2200 TapeStation with a High Sensitivity D1000 Screen Tape (Agilent, Santa Clara, CA).

### Infinium FFPE DNA Restoration Solution

The Restoration Solution consists of two kits: the Infinium HD FFPE DNA Restore Kit (Illumina # WG-321-1002) and the Illumina FFPE QC Kit (Illumina # WG-321-1001). The suggested workflow is as follows: Extracted DNA is quantified and processed with the FFPE QC Kit; samples that pass QC are used in the Restore Kit; and total restored DNA product is prepared for the WGA that is the first step of the Infinium Assay.

The FFPE QC Kit instructs the user to test 2 ng of DNA with a qPCR reaction that generates a ~90 bp product and to determine whether the observed  $C_t$  is more than 5 cycles beyond the expected value. This is essentially a test of the user's quantification of the sample and of any potential PCR inhibitors in the sample. Since the three stock DNA samples we used are high quality, previously quantified genomic DNA samples, we omitted this assay after an initial test.

We followed the Infinium HD FFPE Restore Protocol (Illumina # WG-901-2004) as described by the manufacturer, except 1.5 mL LoBind Tubes were used instead of a 96-well plate, and final restored DNA was eluted into 11  $\mu$ l instead of 10  $\mu$ l ultrapure H<sub>2</sub>O, to provide an extra microliter of sample for quality-control purposes. Although the Infinium HD FFPE Restore Protocol is proprietary, its documentation suggests that the assay involves a DNA repair step, a single round of DNA replication, and a DNA ligation step.

Fragmented and non-fragmented DNA samples were restored from samples of varying amounts based on QuBit measurements (Table 1). DNA from Individual B and Individual C was concentrated with DNA Clean & Concentrator™-5 columns (Zymo Research # D4013; Zymo Research Corporation, Irvine, CA) prior to restoration.

### Sequencing Library Preparation

We prepared sequencing libraries from 300 bp and 100 bp fragmented DNA samples of dsDNA and ssDNA (Table 1).

The dsDNA libraries were processed through modified NEBNext ChIP-Seq protocols (NEB #E6240) with an Illumina 'Y-adapter' of full-length P5 adapter and an indexed P7 adapter (Table 2). DNA samples were incubated with 0.5  $\mu$ l USER enzyme (NEB # M5505) for 37 °C for 1 hour in 1X NEBNext End Repair Reaction Buffer. 1  $\mu$ l NEBNext End Repair Enzyme Mix was then added and the end-repair reaction was incubated for 30 minutes at 20 °C. For 300 bp fragmented libraries, we reduced the elution volume of the 1.8X AMPure XP (Beckman Coulter) cleanup from end-repair to 26.25  $\mu$ l. The reaction volume of the dA-

tailing step was subsequently 30 µl and no cleanup step was performed prior to adaptor ligation. The quick ligation reaction was setup in an increased volume of 46 µl and two 1.1X AMPure XP cleanups were performed. PCR enrichment of the libraries was performed with oligos IS5 and IS6 (Table 2), with a number of cycles (typically 10-16 cycles) based on a qPCR quantification of the unamplified library (KAPA Library Quantification Kit (KAPA # KK4824). Final AMPure XP cleanup was at 1X. For 100 bp fragmented libraries, reaction conditions and volumes were the same as 300 bp fragmented libraries. However, post end-repair cleanup was done with Zymo DCC-5 spin columns (Zymo Research # D4013), and a 1.6X ratio of Sera-Mag Magnetic SpeedBeads (Fisher Scientific; Rohland and Reich 2012) to DNA was used for post-ligation and post-amplification cleanups. All libraries were amplified with NEBNext High-Fidelity 2X PCR Master Mix.

The ssDNA libraries were processed through a novel high-sensitivity single-stranded approach as described (Gansauge and Meyer 2013). We used P5 and P7 indexed primers from (Kircher et al. 2012) for primary library amplifications and we altered the spacer arms of single-stranded adapter CL78 to lower its cost (Table 2).

#### Library Ligation

Post-amplified library molecules were ligated together with T4 DNA ligase / T4 PNK, followed by cleanup with Zymo DCC-5 spin columns. 40 ng of each library was used as input. Final ligated DNA was eluted into 11 µl of ultrapure water, with 1 µl retained for QC purposes and the remaining 9 µl as input into the Infinium HD FFPE/LCG workflow.

#### Library Input Without WGA

A subset of libraries (Table 1; 'Double Amplification') were amplified twice to plateau with qPCR-determined cycle numbers and AccuPrime Pfx polymerase. Libraries were purified with 1.2X (300 bp libraries) or 1.6X (100 bp libraries) AMPureXP beads, eluted into 40 µl TE<sup>-4</sup>, and inserted at the "Precipitate the MSA5 Plate" step of the Infinium HD FFPE protocol, prior to PM1 addition.

#### Infinium HD FFPE and Infinium LCG Assays

The Infinium HD FFPE Assay and LCG Assay were run as described in their respective manual protocols, except that for all samples 9  $\mu$ l of DNA was added to 20  $\mu$ l MA1 to make the MSA5 and MSA6 plates, respectively. We note that both of these assays differ from the typical Infinium HD Assay in that reagent MA2 has been replaced with reagent RPM in the WGA step. Fluorescence intensity data were collected with an Illumina iScan Plus.

### Genotype Data Processing

We processed the fluorescence data using Illumina's GenomeStudio software package (ver 1.9.0), the Genotyping module (1.9.4), and the suggested cluster files for each BeadChip. We calculated call rate, concordance rate, principal components (flashpcaR), and various genotype metrics using R (ver 3.0.2).

Concordance was defined as: For two samples genotyped on the same Beadchip (HumanOmniExpress-FFPE or HumanOmni2.5), take each locus that is called in both samples, compare the Illumina standardized AA/AB/BB genotypes, and mark the locus as concordant if the two genotypes match exactly. If one genotype was NC in either or both of the two samples, that locus is discarded.

## Results

### Statement of Results

#### Fragmented Genomic DNA Profiles

Electropherograms for the 300 bp and 100 bp fragmented genomic DNA samples of Individuals A, B, and C are shown in Figure 1.

#### Sequencing Library Profiles

Electropherograms for the sequencing libraries genotyped in this study are shown in Figure 2 (for the initial libraries) and Figure 3 (for the secondary amplified libraries). Negative controls for all three sequencing library protocols (300 bp dsDNA, 100 bp dsDNA, and 100 bp ssDNA) showed the expected adapter dimer peaks, although the 300 bp dsDNA library blank had a short tail of multiple peaks as well, possibly indicating heteroduplex formation from over-amplification. All libraries showed increasing adapter dimer formation with decreasing input DNA amount (and therefore increasing primary amplification cycle numbers). The 300 bp dsDNA libraries showed the lowest dimer peak relative to the desired library product, likely because of the two restrictive 1.1X AMPure XP cleanups following the initial amplification.

The ssDNA library preparation was expected to have an adapter dimer peak of 129 bp for double indexed primers (Gansauge and Meyer 2013), yet ours showed a slightly larger peak of ~150-170bp. This peak was not efficiently removed in the 2.0 ng and 0.2 ng input preparations. Our single-stranded adapter (CL78\_2d) was not synthesized exactly as in the original paper, and it may contain unforeseen impurities. This did not seem to prevent the single-stranded DNA library prep from performing the best in subsequent analyses (see below).

#### Input DNA Observations and DNA Restoration Solution Recoveries

For non-fragmented DNA, the concentration of DNA measured using PicoGreen was approximately the same as that determined by qPCR, yielding an average ratio 0.87. For



fragmented DNA that ratio increased to 2.24-fold, presumably as fewer amplicon targets of the appropriate size (202 bp) exist in the fragmented DNA pool.

The Infinium HD FFPE DNA Restore Kit performed differently based on the fragmented status of the input DNA sample. For fragmented DNA, the restored DNA amount was  $91\% \pm 52\%$  (mean  $\pm$  s.d.) of the input, based on 16 qPCR observations, or  $99\% \pm 67\%$  based on 15 PicoGreen observations. For non-fragmented DNA, the yield of restored DNA was much higher:  $558\% \pm 219\%$  of the input based on 5 qPCR observations or  $278 \pm 115\%$  based on 5 PicoGreen observations.

### Call Rates and GenCall Scores

Illumina quality metrics for both the HumanOmniExpress-FFPE BeadChip and the HumanOmni2.5 BeadChip, called using default cluster files (see Methods), are shown in Figure 4, while sample descriptions can be found in Table 1. For all three individuals, genotype call rates of  $\geq 99\%$  were observed for restored samples of  $\geq 20$  ng input or non-restored samples of 200 ng input, irrespective of fragmentation. For Individual A, an initial input of 10 ng of DNA, followed by restoration, reduced call rates to 95% and 83 % for the FFPE and Omni2.5 Beadchips, respectively (Samples 13 and 26). Call rates dropped further to 84% and 77% at 2 ng restored input (Samples 12 and 27, respectively), and 56% and 38% at 0.2 ng restored input (Samples 14 and 28, respectively). Note that the 2 ng sample for the HumanOmniExpress-FFPE BeadChip was fragmented to 300 bp, while the rest were fragmented to 100 bp. Blank samples (ultrapure water) had call rates of 51% for the HumanOmniExpress-FFPE and 31% for HumanOmni2.5 (Samples 15 and 29).

On the HumanOmniExpress-FFPE BeadChip, ligated sequencing libraries from 2 ng of fragmented DNA (Samples 16, 17, and 18) raised the call rate from 84% (for restored 300 bp 2 ng input; Sample 12) to as high as 90% for 100 bp fragmented DNA made into a ssDNA library (Sample 18). On the HumanOmni2.5, similarly treated samples raised the call rate from 77% (Sample 27) to 85% (Sample 31), again with the ssDNA library technique. When initial input DNA was at 0.2 ng, our ligated library technique on the HumanOmni2.5 BeadChip increased call rates from 38% (for restored 100 bp, 0.2 ng input DNA; Sample 28)

to 49% (dsDNA library; Sample 30) or 58% (ssDNA library; Sample 32). Bypassing the Infinium Assay (including WGA) and directly placing sequencing libraries into the MSA5 plate led to call rates of 51-53%, irrespective of initial DNA quantity or fragmentation state (Samples 19-24). This was a similar call rate to the blank samples.

For similarly-treated samples, p10 GC values (the tenth percentile of the distribution of GenCall scores across all called genotypes, per sample) for the HumanOmniExpress-FFPE BeadChip were equal to or greater than those called on the HumanOmni2.5 BeadChip; p50 GC Values (the fiftieth percentile) were consistently greater for the same matched samples called on the HumanOmni2.5 BeadChip. The only exception were the controls (Samples 1 and 25), where the HumanOmni2.5 BeadChip outperformed on both measures. Negative controls (water; Samples 15 and 29) had p10 and p50 GC values that were the lowest scores for each BeadChip, although the actual values were notably high (p10 = .25 and .23, p50 = .68 and .71, for the HumanOmniExpress-FFPE and HumanOmni2.5, respectively; Figure 4).

#### Individual B and C Concordance Rates

For all samples of individuals B and C, concordance rates were calculated for all pairwise comparisons. As each sample from these individuals had call rates of >99% (Figure 4), it was unsurprising that concordance rates were nearly 100% for all comparisons of samples from the same individual. HumanOmniExpress-FFPE BeadChip genotyping calls were so consistent that the number of pairwise differences never exceeded 44 out of 688,534 total genotypes called. These results indicate that 300 bp fragmented samples, when repaired with the Restore Protocol, perform identically to high-quality DNA when input amounts are  $\geq 20$  ng.

Concordance rates for all pairwise comparisons between samples from individuals B and C (two members of the Kattunayakan tribal population) averaged 64%. The concordance rates between all samples of individuals B and C and Sample 1 (the 'control' sample of 200 ng non-restored, non-fragmented DNA; Table 1) of individual A (of Telegu Brahmin ancestry, genetically more similar to Europeans than to Kattunayakan Watkins et al. 2005)

averaged 57%. These rates give some idea of the concordance rate to be expected between unrelated individuals when using the HumanOmniExpress-FFPE BeadChip.

#### Individual A Concordance Rates

For the HumanOmniExpress-FFPE BeadChip, concordance rates for Samples 8-11 relative to Sample 1 were nearly 100%, with the number of pairwise differences at 96 or lower (out of 685,820 to 690,161 total loci called). As the DNA input amount dropped to 10 ng and below, concordance rates were reduced, as were call rates above. Concordance rates and the number of shared called SNPs between the remaining samples and Sample 1 (for the HumanOmniExpress-FFPE BeadChip) and Sample 25 (for the HumanOmni2.5 BeadChip) are shown in Tables 3 and 4.

At picogram amounts of restored fragmented DNA, both of the Infinium Assays performed poorly, with concordance rates near those of the water blanks (compare Sample 14 vs. Sample 15 and Sample 28 vs. Sample 29). While using the HumanOmni2.5 BeadChip allowed a greater concordance rate with the control (Sample 25) in comparison with similarly-treated samples on the HumanOmniExpress-FFPE BeadChip, concordance rates between the control and the blank was also greater (52% vs. 31%; Tables 3 and 4). When attempting to genotype sequencing libraries directly on a HumanOmniExpress-FFPE BeadChip, without utilizing the Infinium HD FFPE WGA+fragmentation step, samples performed similarly to the water blank (Samples 19-24). The amount of DNA produced by the PCR reamplification is likely far too low for adequate hybridization and signal generation on the BeadChip.

With 2 ng DNA input, the ligated library method produced concordance improvements on both Beadchips relative to the matched restored sample. With 200 pg DNA, the ssDNA and ligation method recovered 80% concordant SNPs (32) on the HumanOmni2.5 Beadchip, compared with 60% concordance in the FFPE-Restored sample (28).

#### Optimized Genotyping Calls for Individual A

Above, we have considered only the qualitative genotype calls (or no-call states) returned by Illumina's signal processing and genotype calling software. The genotype calls themselves are based on underlying quantitative scores which are also available for inspection. If the scores for a genotype indicate a low reliability according to Illumina's algorithms, a 'no-call' genotype (NC) is returned. The remaining called genotypes (typically more than 95% of SNPs tested) are very likely to be correct if high-quality DNA was used. However, with the challenging DNA samples examined here, many of the called genotypes (those passed by Illumina's algorithm) are incorrect, as evidenced by the 20% discordance rate for the ssDNA library analyzed on the HumanOmni2.5 BeadChip (Sample 32 vs. Sample 25 of Individual A). With such challenging samples, it may be possible to further discriminate between more and less reliable genotypes returned by Illumina's genotype calling software. We investigated two approaches to a solution – a general filtering based on the blank samples and a ranking of calls based on call-by-call quality metrics.

As the blank samples yielded a substantial number of unexpected genotyping calls, we investigated whether these called SNP loci (evidently incorrect, given the absence of sample DNA) were more likely to be incorrectly called in the low input samples. We performed a simple data reduction step by removing all loci called in each of the blanks (Sample 15 and Sample 29) in the relevant samples (for HumanOmniExpress-FFPE or HumanOmni2.5).

We present concordance rates and number of called shared SNPs between the remaining samples and Sample 1 (for the HumanOmniExpress-FFPE BeadChip) and Sample 25 (for the HumanOmni2.5 BeadChip) for this reduced data set in Tables 3 and 4. For samples with  $\geq 2$  ng DNA input on the FFPE-optimized HumanOmniExpress Bead Chip, this data reduction step did not result in any concordance rate increase but did produce a 51% drop in call rate. The single 200 pg sample (Sample 14) had a 1% increase in concordance (with a 50% drop in call rate), and the non-WGA samples showed 0-2% increases. On the HumanOmni2.5 BeadChip, this data reduction approach fared slightly better. For samples at  $\geq 2$ ng input, concordance rates increased by 1%, with only a 30% drop in call rate. The 200 pg samples had a 2-3% increase in concordance, with a 32-37% drop in call rates. The

best performing sample at this lowest input (the ligated ssDNA library of Sample 32) continued to perform well with a final 82% concordance to Sample 1, with 933,694 called SNPs (Table 4).

In an alternative approach, we estimated a multivariable logistic regression model to discriminate correct vs. incorrect genotype calls based on five quantitative scores that are produced by Illumina's GenomeStudio software for every genotyped SNP (independent variables GC\_SCORE, GT\_SCORE, LOG\_R\_RATIO, log(R), and RDEV). We estimated the model by comparing the genotype calls and scores of Sample 30 (0.2 ng input, 100 bp fragmented, ligated dsDNA library, with 75% concordance to control within called genotypes) vs. the control genotype calls of Sample 25. We then used that model to compute the discriminator variable  $y$  from the scores returned for Sample 32 (0.2 ng input, 100 bp fragmented, ligated ssDNA library), which showed an 80% concordance rate among called genotypes. Correct calls should be correlated with higher values of  $y$ , so eliminating all calls with values of  $y$  lower than some threshold  $t$  (ranging from 0 to 1) should improve the concordance rate. Figure 5 shows the numbers of correct vs. incorrect genotype calls that result for varying values of  $t$  (a receiver operating characteristic, or ROC, curve). Values of  $t$  near 0.96 yield sets of approximately one million called SNPs with a substantially improved concordance rate of 88%. While more data will be required to determine whether this pattern generalizes usefully, this result implies that the Illumina quality scores contain additional information that can discriminate accurate from inaccurate genotype calls for genotypes collected from forensic samples.

### *Principal Components Analysis of Genotype Calls*

In addition to the concordance rate comparisons, we performed principal component analyses (PCA) to explore the variation in our data between our high quality controls (1 and 25) and the various samples genotyped with less than expected call rates and concordance rates on the two BeadChips. Our first scatterplot (Figure 6) reports the first two principal components for genotyping calls in Sample 1 and Samples 13-24 in the HumanOmniExpress-FFPE BeadChip; these two principal components accounted for 86% of the variance in the data. Only those loci called in all 13 samples were included; this

dropped the available dataset down to 12,493 loci (mostly due to the low number of shared called loci in Samples 19-24; see following section). Even at the substantially reduced data set, the genotyped ligated sequencing libraries clearly cluster tightly with the high quality Sample 1 and the successful 10 ng restored Sample 13. These are separated along PC1 from Sample 14, the water blank (15), and the non-WGA samples (19-24). PC2 separates out the water blank from both the non-WGA samples and the failed 0.2 ng input restored DNA sample (14).

We next clustered the genotype calls from Samples 1 and 16-18 from Individual A, samples 2 and 5 from Individual B, and samples 6 and 7 from individual C; the number of shared called loci was much larger - 502,097 SNPs. The first two principal components accounted for 96.7% of the variance in the data (Figure 7). All the 2ng ligated sequencing libraries clustered tightly with the high quality control, separated along PC1 from the genotyped samples from Individual B and Individual C. As expected, given the high concordance rates, the lower input fragmented samples for Individual B (5) and C (7) clustered directly on top of the high quality controls for each Individual.

We finally clustered the genotype calls from all eight samples from the HumanOmni2.5 Beadchip. The number of shared called loci was limited to 73,917 SNPs (mostly due to the low number of shared called loci between Samples 28 and 29; see following section). The first two principal components only accounted for 69.5% of the variance, so we report principal components 3 and 4 as well, which brings the total proportion of variance to 96.2% (Figure 8A & B). Along PC1, Sample 28 (0.2 ng input restored DNA sample) and Sample 29 (restored water blank) are distinguished from the other samples. The two sub-nanogram ligated sequencing library samples (30 and 32) are clustered near the high quality control, though at PC2 and PC3 the dsDNA library is clearly removed. Sample 32 is not clearly distinguished from the 'truth' cluster until PC4.

### *Patterning in Genotypes*

The patterns of genotypes called varied across sample treatments. On the HumanOmniExpress-FFPE BeadChip, the water blank (15) and non-WGA samples (19-24)

both showed an excess of heterozygosity and a lack of 'BB' genotypes, while the 200 pg restored sample (14) showed an excess of 'AA' genotypes with a decrease in both 'AB' and 'BB' genotypes relative to the control (1) (Figure 9A). We note that this discordance in genotyping profiles might be reflected in the large distance along PC2 between Sample 14 and Sample 15 in the PCA of Figure 6. On the HumanOmni2.5 BeadChip the FFPE-Restored 200 pg (28) and the blank (29) samples shared a proportional excess of AA genotypes (Figure 9B), however they also do not cluster tightly in a reduced loci PCA (Figure 8A & B).

Called genotypes in the water blank and the poorly performing samples exhibited a 31-38% concordance rate on the HumanOmniExpress-FFPE BeadChip out of ~350,000 to 370,000 loci (Table 2). However, the number of shared called loci (regardless of genotype) between Samples 14, 15, and 19-24 was only 17,979, or 30,554 without Sample 14. The number of shared not-called loci between Samples 14, 15, and 19-24 was similarly small at 12,742, or 23,531 between Samples 15 and 19-24. This suggests that the called and concordant SNPs between these samples and the high quality control Sample 1 is the result of random chance – if a SNP is called in a failed sample or blank, it has a 1/3 chance of matching the correctly called genotype (AA,AB,BB). Any consistency in the loci that are called in a water blank sample on the HumanOmniExpress-FFPE BeadChip will need to wait for further blank measurements.

We also investigated the type of genotype error in the discordant genotype calls between Sample 1 and all samples with both DNA and lower than expected concordance (Samples 14, 16, and 19-24; Table 2). As shown in Figure 10, the majority of genotype changes are from a called genotype in Sample 1 to a NC (not called) genotype in the paired samples; this is expected given the previously determined low number of shared called loci (Table 2). Of the called genotypes in Samples 14, and 19-24 there is a high relative frequency of AA>AB, AB>AA, BB>AA, and BB>AB discordance, as well as virtually no BB>BB concordance. Low BB>BB concordance, and BB>AA and BB>AB error, was expected given the relative dearth of 'BB' genotypes called overall (Figure 9A). In Sample 16, with a concordance of 93% (Table 2), a clear signal of a loss of heterozygosity (AB>AA, AB>BB) appears.

On the HumanOmni2.5 BeadChip the number of shared called loci (regardless of genotype) between all 7 samples and the water blank (Sample 29) was 73,917. This is a similar proportion (~3%) to the number of shared called loci on the HumanOmniExpress-FFPE BeadChip between the failed samples, and suggests that multiple replications of the same type of blank sample are necessary to determine consistently refractory loci.

In Figure 11 we investigate the type of genotype error in the discordant genotype calls between Sample 25 and all samples with DNA (i.e. excluding Sample 29) on the HumanOmni2.5 BeadChip. As with the HumanOmniExpress-FFPE BeadChip, the majority of genotype changes were to a no called state in the test sample. Of the called loci, the low concordance samples (Sample 28, 30, 32; Table 3) all showed an excess of AB>AA and AB>BB genotype errors, as well as a low relative proportion of concordant AB>AB genotypes (this mimics the error pattern in Sample 16, on a different BeadChip). The next most prevalent genotyping error across all samples on the HumanOmni2.5 was BB>AB.

We can take the reverse stance by asking what the observed discordance rates are given a genotype called in an experimental sample. In Figure 12, for the same paired samples as Figure 11, relative frequencies of concordance and discordance are plotted for each called genotype (AA,AB,BB). It is clearly evident that AB genotypes are the most likely to be discordant, however we note that AB genotypes are the least plentiful calls given the Sample 25 results (Figure 13).



## Tables

**Table 1: Experimental Conditions**

Sample ID	Individual ID	BeadChip, Illumina HumanOmni-	Fragment Size (bp) <sup>a</sup>	DNA Input (ng)	FFPE Restoration <sup>a</sup>	Library Preparation <sup>a</sup>	Library Ligation <sup>a</sup>	Double PCR Amplification (w/o WGA) <sup>a, b</sup>	Input into Infinium Assay (ng)
1	A	Express-FFPE		200					200
2	B	Express-FFPE		100	+				288
3	B	Express-FFPE		50	+				173
4	B	Express-FFPE	300	100	+				50
5	B	Express-FFPE	300	20	+				32
6	C	Express-FFPE		100	+				180
7	C	Express-FFPE	300	100	+				63
8	A	Express-FFPE		100	+				324
9	A	Express-FFPE	300	200	+				74
10	A	Express-FFPE	300	100	+				67
11	A	Express-FFPE	300	20	+				23
12	A	Express-FFPE	300	2	+				too low
13	A	Express-FFPE	100	10	+				23
14	A	Express-FFPE	100	0.2	+				too low
15	Water	Express-FFPE		0	+				too low
16	A	Express-FFPE	300	2		dsDNA	+		20
17	A	Express-FFPE	100	2		dsDNA	+		14
18	A	Express-FFPE	100	2		ssDNA	+		18
19	A	Express-FFPE	300	2		dsDNA		+	1488
20	A	Express-FFPE	300	0.2		dsDNA		+	396
21	A	Express-FFPE	100	10		dsDNA		+	552
22	A	Express-FFPE	100	0.2		dsDNA		+	38
23	A	Express-FFPE	100	10		ssDNA		+	588
24	A	Express-FFPE	100	0.2		ssDNA		+	167
25	A	2.5		200					200
26	A	2.5	100	10	+				5
27	A	2.5	100	2	+				too low
28	A	2.5	100	0.2	+				too low
29	Water	2.5		0	+				too low
30	A	2.5	100	0.2		dsDNA	+		11
31	A	2.5	100	2		ssDNA	+		14
32	A	2.5	100	0.2		ssDNA	+		16

<sup>a</sup> Empty cells indicate no treatment (fragmentation, restoration, library preparation, ligation, PCR reamplification)

<sup>b</sup> Libraries under this treatment were amplified twice by PCR and the WGA step of the Infinium assay was not performed.

**Table 2: Oligos used in this study**

Adaptors/Primers (5'-3')	ID	Note	Purification
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T	Illumina P5		IDT Ultramer, Standard Desalting
/5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCACTCGGCAATCTCGTATGCCGTCTTCTGCTTG	Indexed Illumina P7	Index TSBC48 - 'TCGGCA'	IDT Ultramer, Standard Desalting
AATGATACGGCGACCACCGA	IS5	See (Meyer and Kircher 2010)	Standard Desalting
CAAGCAGAAGACGGCATAACGA	IS6	See (Meyer and Kircher 2010)	Standard Desalting
AATGATACGGCGACCACCGAGATCTACACCCTGCGAACACTCTTTCCCTACACGACGCTCTT	Kircher P5 Indexing Primer	Kircher P5_iPCR-LP-1, (Kircher et al. 2012)	IDT Ultramer, Standard Desalting
CAAGCAGAAGACGGCATAACGAGATGAAGATTGTGACTGGAGTTCAGACGTGT	Kircher P7 Indexing Primer	Sol_iPCR-MPI-97, (Kircher et al. 2012)	IDT Ultramer, Standard Desalting
/5Phos/AGATCGGAAG/iSp18//iSp18//3BioTEG/	CL78_2d	Modified CL78, (Gansauge and Meyer 2013)	Dual HPLC Purification

**Table 3:** Genotype Concordance for shared called SNPs, Individual A, HumanOmniExpress-FFPE BeadChip

Sample #	12	13	14	15 (Water)	16	17	18	19	20	21	22	23	24
Relative to #1	98%	99%	38%	31%	93%	97%	99%	38%	33%	34%	32%	35%	31%
# Shared SNPs	584,065	654,490	386,493	351,349	610,632	604,734	619,665	365,411	355,461	354,872	363,614	367,243	359,623

Excluding SNPs called in Sample 15:

Relative to #1	98%	99%	39%		93%	97%	99%	40%	34%	35%	32%	36%	31%
# Shared SNPs	287,611	321,579	192,722		300,252	296,936	304,595	166,995	161,000	155,870	162,853	169,507	160,577

**Table 4:** Genotype Concordance for shared called SNPs, Individual A, HumanOmni2.5 BeadChip

Sample #	26	27	28	29 (Water)	30	31	32
Relative to #25	97%	94%	60%	52%	75%	98%	80%
# Shared SNPs	1,981,453	1,832,307	899,845	738,719	1,162,521	2,035,168	1,375,326

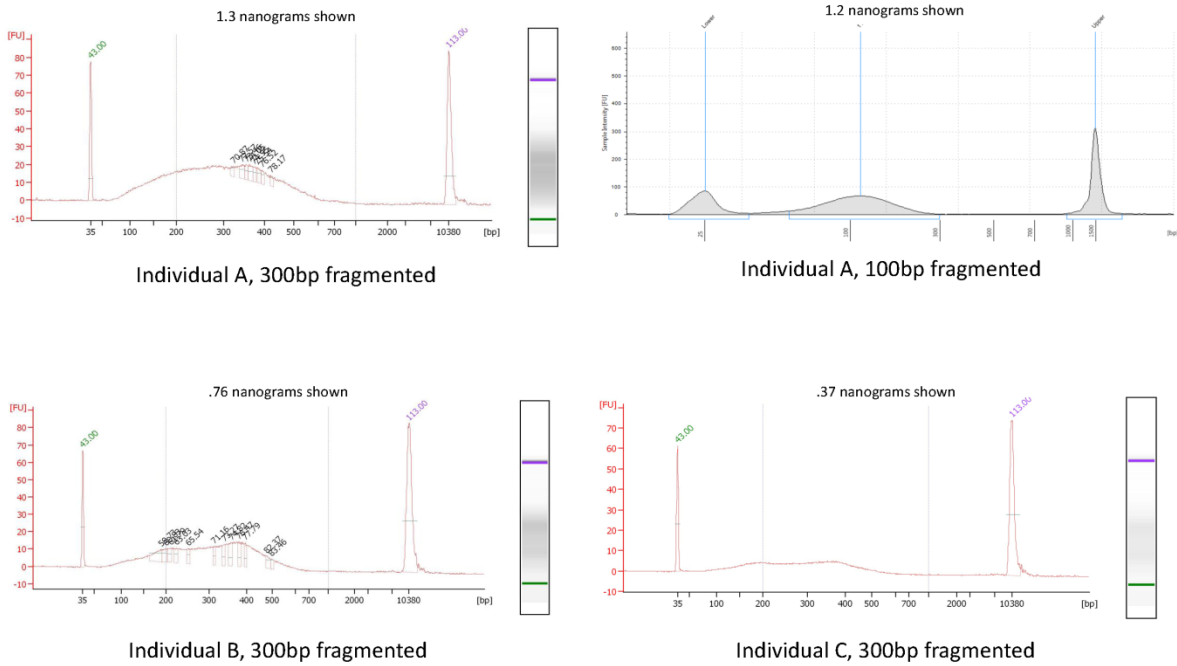
Excluding SNPs called in Sample 29:

Relative to #25	97%	95%	63%	77%	99%	82%
# Shared SNPs	1,376,542	1,271,468	565,231	770,630	1,420,094	933,694

## Figures

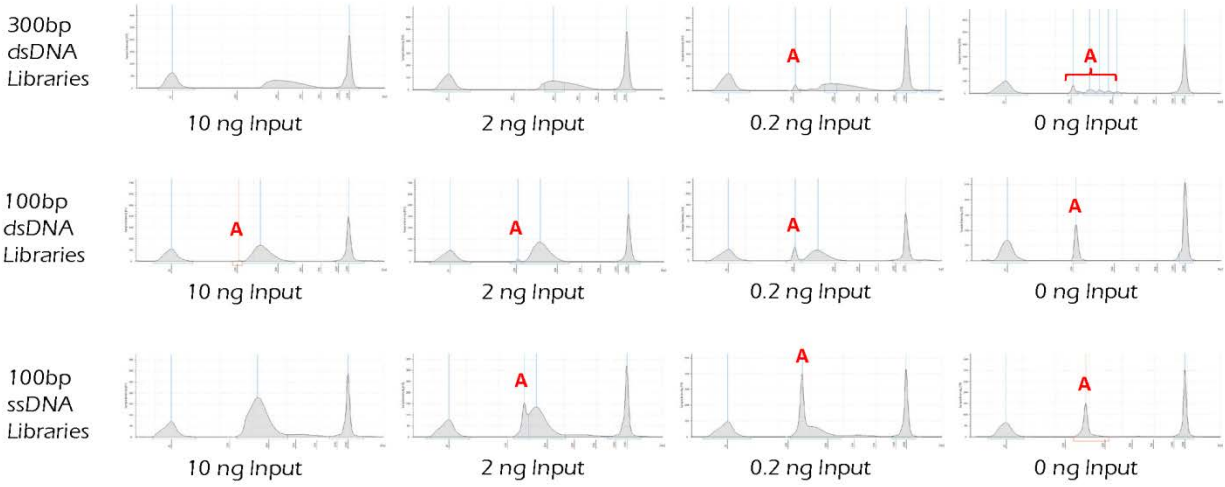
*Figure 1: Electropherograms of the fragmented genomic DNA samples of Individuals A, B, and C.*

The peak of the 100 bp fragmented sample from Individual A, used in the majority of this study, is at 113 bp (upper right panel, Agilent High Sensitivity D1000 Screen Tape; other panels show results from Agilent BioAnalyzer High Sensitivity DNA Chips).



*Figure 2: Electropherograms for the primary amplified next-generation sequencing libraries*

Electropherograms (Agilent High Sensitivity D1000 Screen Tape) are distinguished by fragmentation size, initial DNA input, and library method.



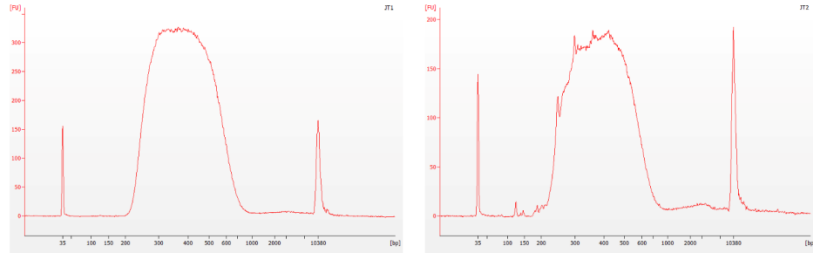
**A** – Adapter dimer/artifact peak(s)

25bp and 1500 bp peaks - Electropherogram size standards

*Figure 3: Electropherograms for the secondary amplified next-generation sequencing libraries*

Electropherograms (Agilent High Sensitivity D1000 Screen Tape) are distinguished by fragmentation size, initial DNA input, and library method.

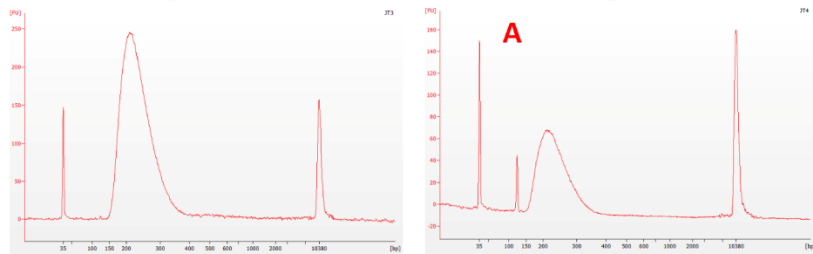
Double Amplified,  
300bp dsDNA Library



2 ng Input

0.2 ng Input

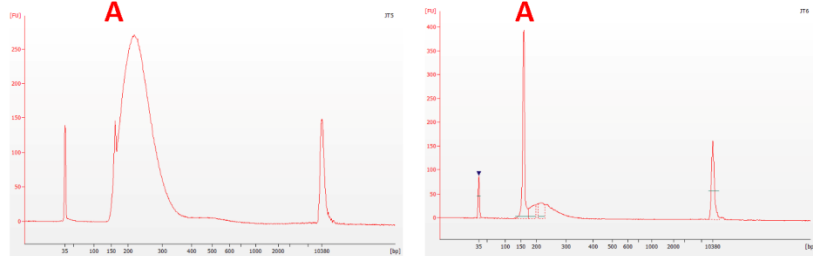
Double Amplified,  
100bp dsDNA Library



10 ng Input

0.2 ng Input

Double Amplified,  
100bp ssDNA Library



10 ng Input

0.2 ng Input

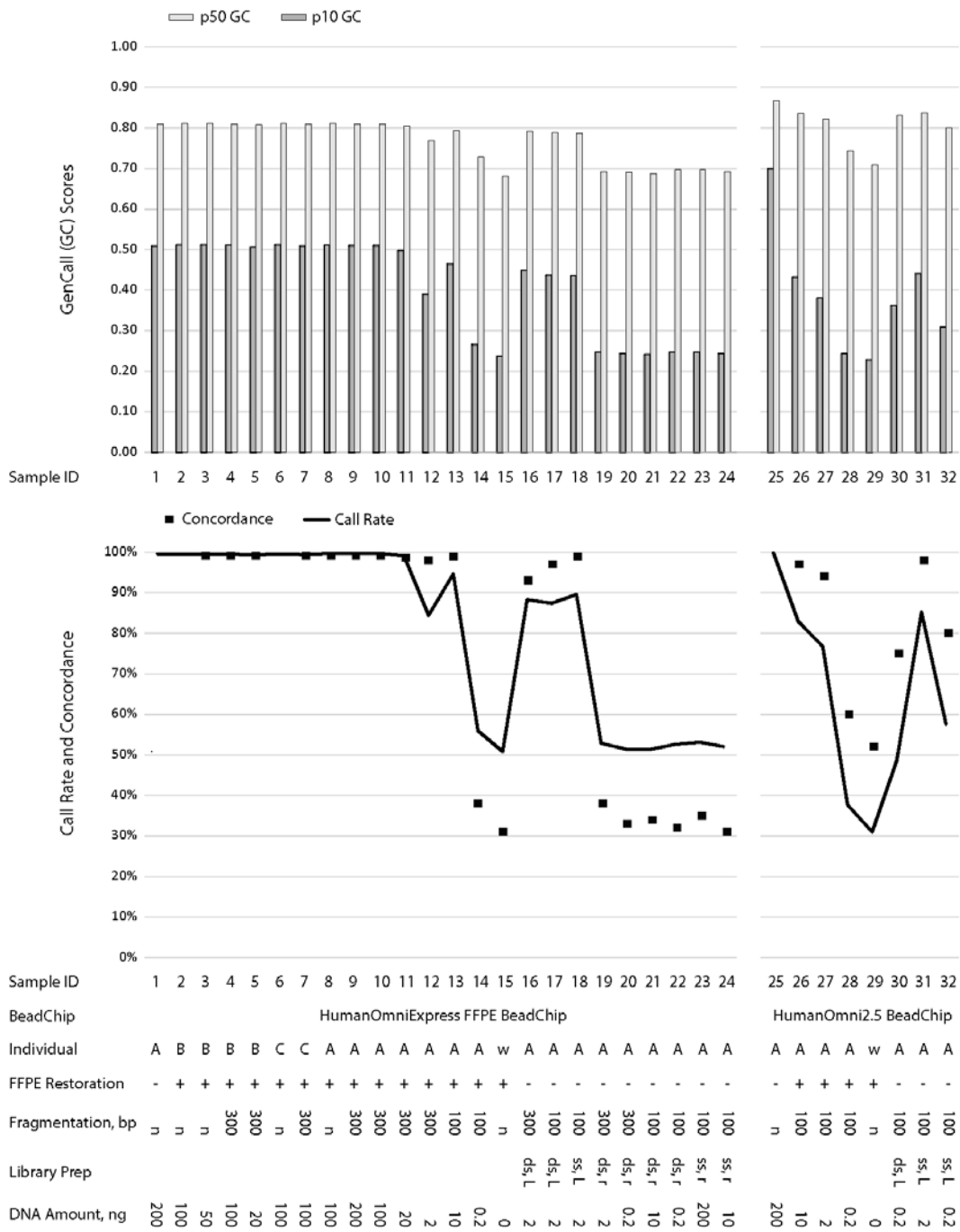
**A** – Adapter dimer/artifact peak(s)

35bp and 10380 bp peaks - Electropherogram size standards

*Figure 4: Sample Quality Metrics*

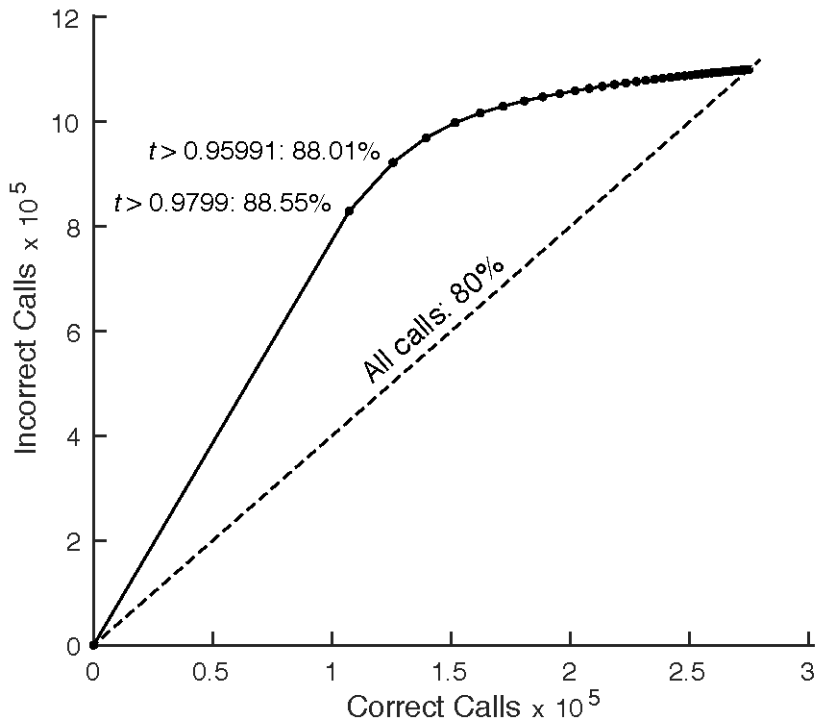
GenCall (upper panel), call rate, and concordance (lower panel) metrics for all 32 samples analyzed in this study. Experimental conditions are noted below the panels. Illumina GenCall scores are a product of the SNP GenTrain score and a data-to-model fit score; they are locus-specific and indicate the reliability of each genotype call. The p10 and p50 metrics are the GenCall score for a SNP at the lower 10<sup>th</sup> and 50<sup>th</sup> percentile when all SNPs for that sample are ranked by score. The call rates are relative to 693,543 loci for Samples 1-24 (HumanOmniExpress-FFPE BeadChip) and 2,391,739 loci for Samples 25-32 (HumanOmni2.5 BeadChip). Concordance of genotypes for treatment samples vs. high-quality, high-quantity control samples (for Individual A, B or C, as appropriate; Samples 1, 2, 6 and 25) are shown according to the right axis scale (see also Table 1). Sample ID, individual label, degree of fragmentation, amount of initial DNA (ng), whether or not the FFPE restoration protocol was applied (+/-), and the library preparation protocol used (ssDNA or dsDNA; subsequently ligated and whole-genome amplified, L, or re-amplified by PCR, r) are listed at bottom and apply to both panels.





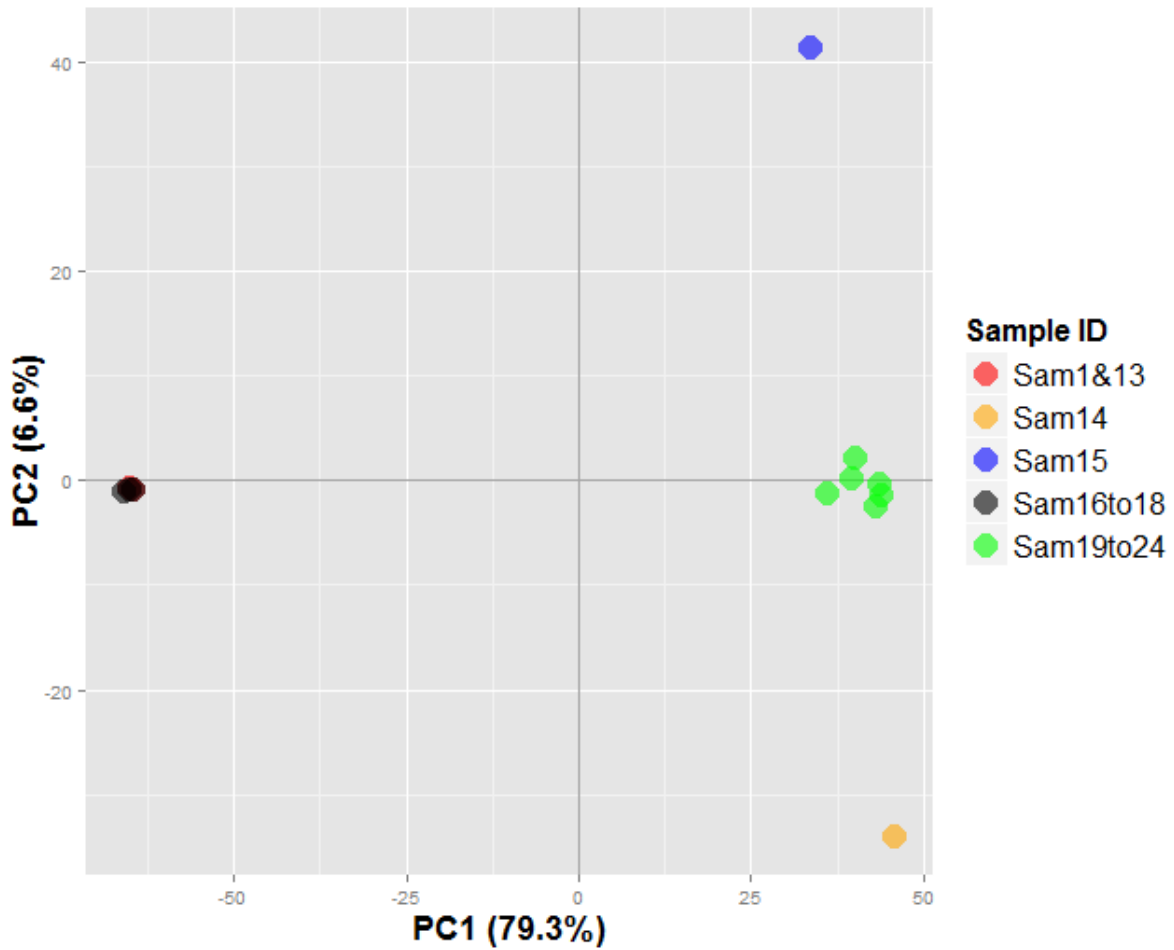
*Figure 5: Regression modeling of genotype calls*

Numbers of correct genotype calls vs. incorrect calls for values of a sensitivity threshold  $t$  on  $[0,1]$ . Each point on the solid line represents the set of genotype calls whose quality scores (from a logistic regression model, see text) exceed a different value of  $t$ . At upper right,  $t$  approaches 0, and all genotypes called by Illumina's software are included, yielding a concordance rate of 80% (determined by comparison with genotypes obtained from high-quality DNA for that individual.) The dashed line represents the 80% correct ratio. At lower right,  $t = 1$ , and all genotypes are excluded. Two choices of  $t$  are shown that raise the concordance rate from 80% to more than 88% (88.55% and 88.01% on sets of 936,365 and 1,047,802 SNPs, respectively).



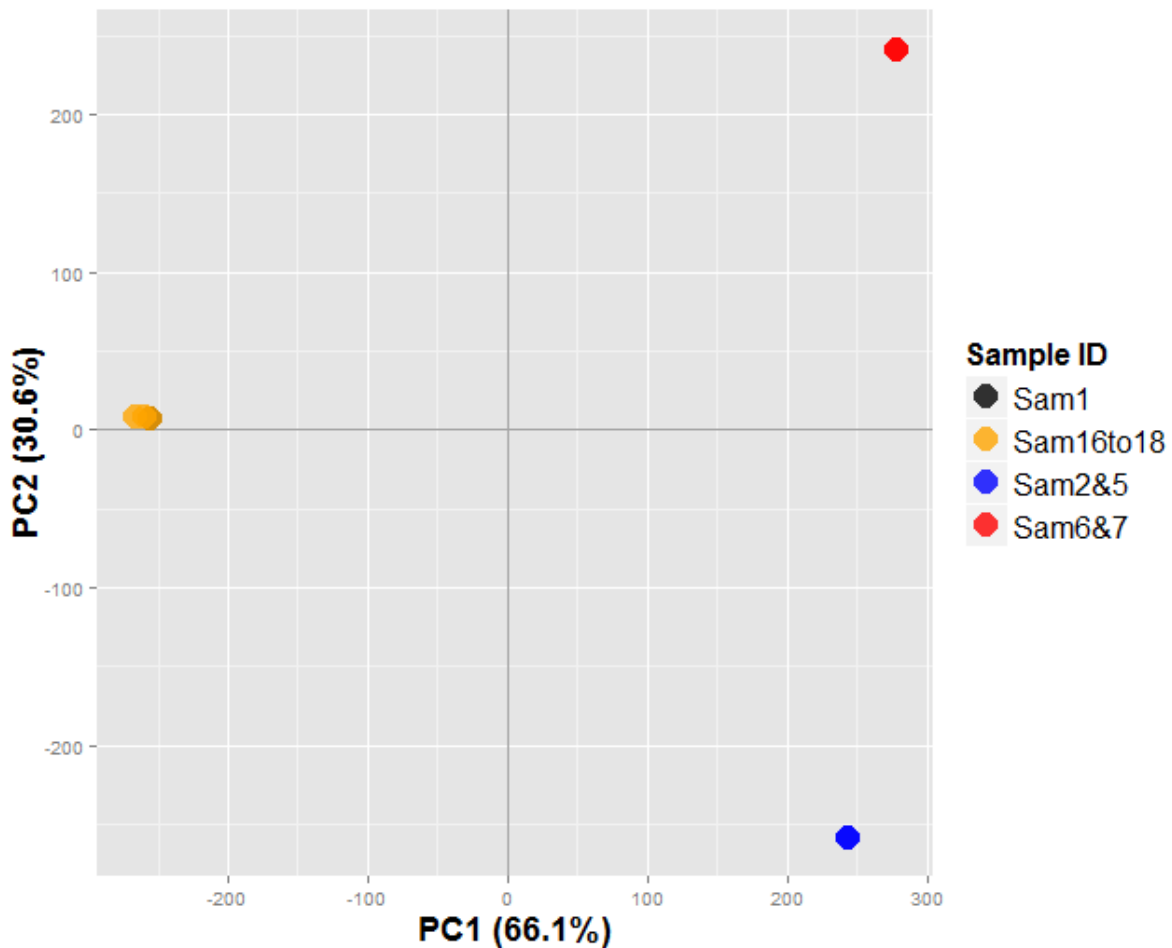
*Figure 6: Principal Component Analysis of low quality and failed samples on the HumanOmniExpress-FFPE BeadChip*

Plots of the first two principal components, with associated proportions of variance, for Samples 1 and 13 through 24. A reduced dataset of 12,493 SNPs were used. Note that Samples 1 and 13 are colored in red and are located underneath Samples 16 through 18.



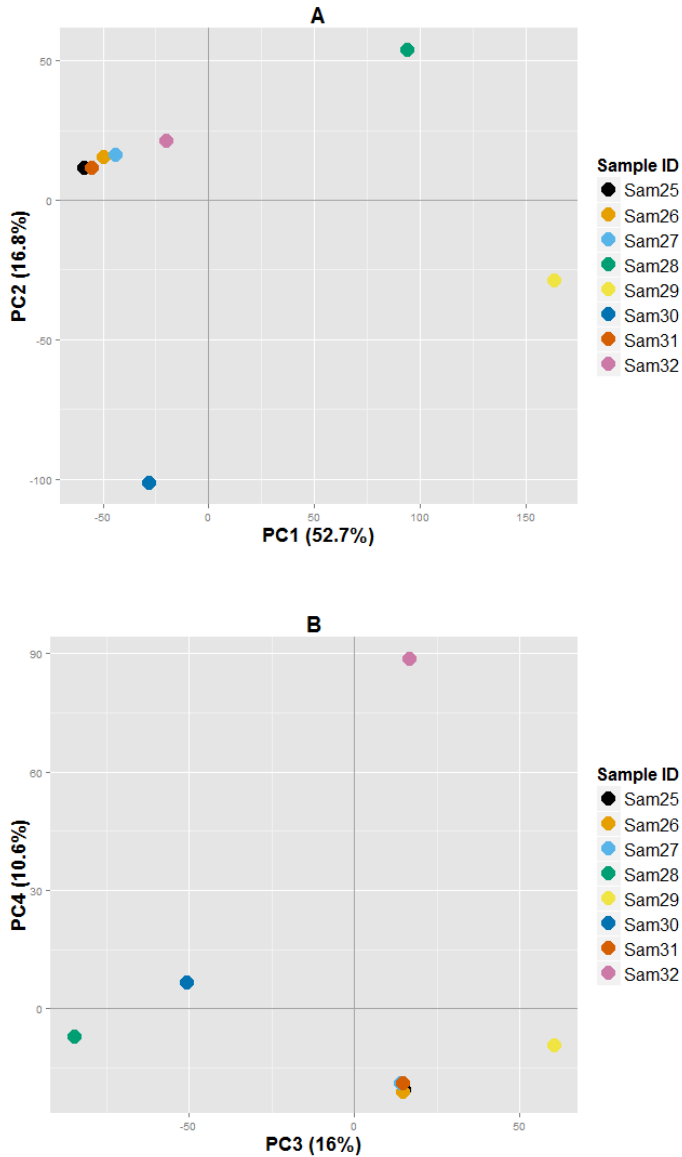
*Figure 7: Principal Component Analysis of ligated sequencing libraries of fragmented samples from Individual A and restored fragmented samples from Individuals B and C*

Plots of the first two principal components, with associated proportions of variance, for Samples 1 and 16-18 (Individual A), Samples 2 and 5 (Individual B), and Samples 6 and 7 (Individual C). A total dataset of 502,097 SNPs was used. Note that Sample 1, colored in black, is located directly under Samples 16 through 18, colored in orange.



*Figure 8A&B: Principal Component Analysis of samples genotyped on the HumanOmni2.5 BeadChip*

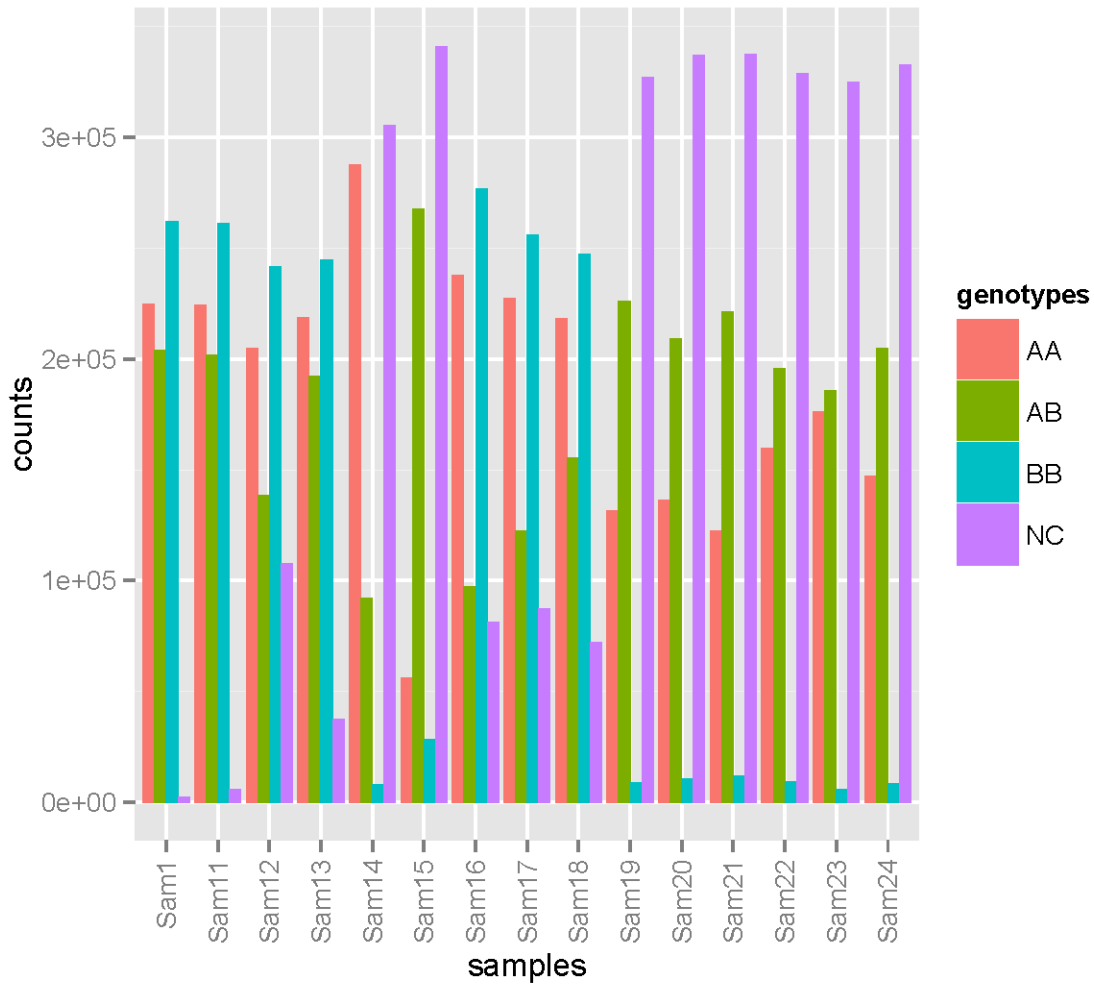
A reduced dataset of 73,917 SNPs were used. (A) Plots of the first two principal components, with associated proportions of variance, for Samples 25 to 32. (B) Plots of the 3<sup>rd</sup> and 4<sup>th</sup> principal components, with associated proportions of variance, for Samples 25 to 32.



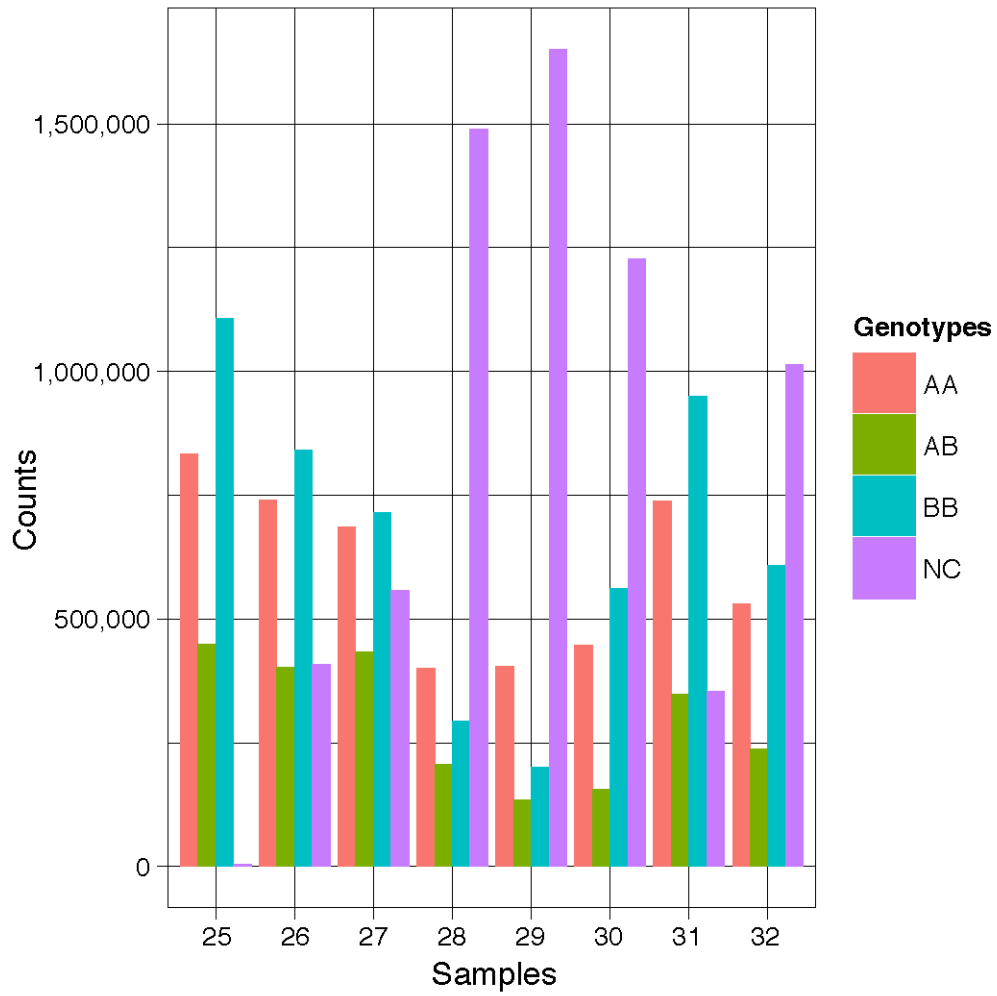
*Figure 9: Individual A Genotype Histograms*

Histograms of genotype counts for all samples of individual A typed on the HumanOmniExpress-FFPE BeadChip, with 693,543 loci (Figure 4A; Samples 1 and 11-24) and the HumanOmni2.5 BeadChip, with 2,391,739 loci (Figure 4B; Samples 25-32).

**(A)**

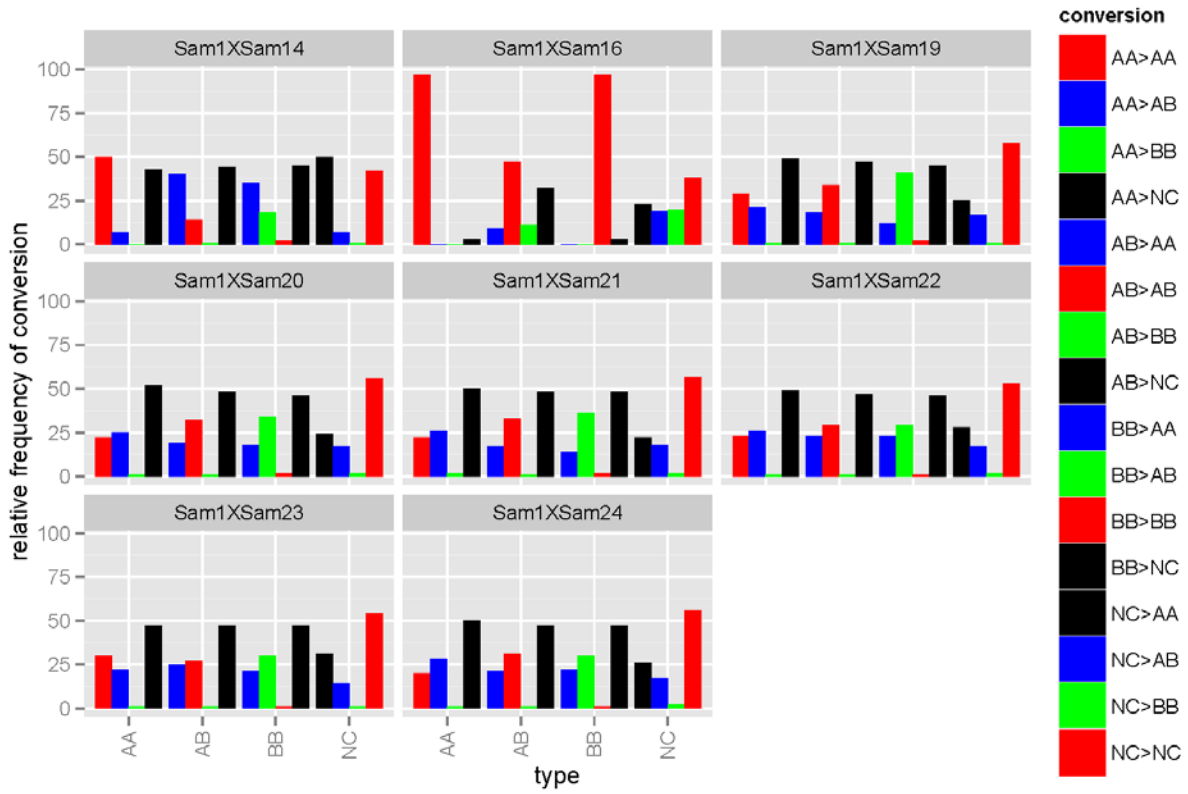


(B)



*Figure 10: HumanOmniExpress-FFPE BeadChip Genotyping Discordance*

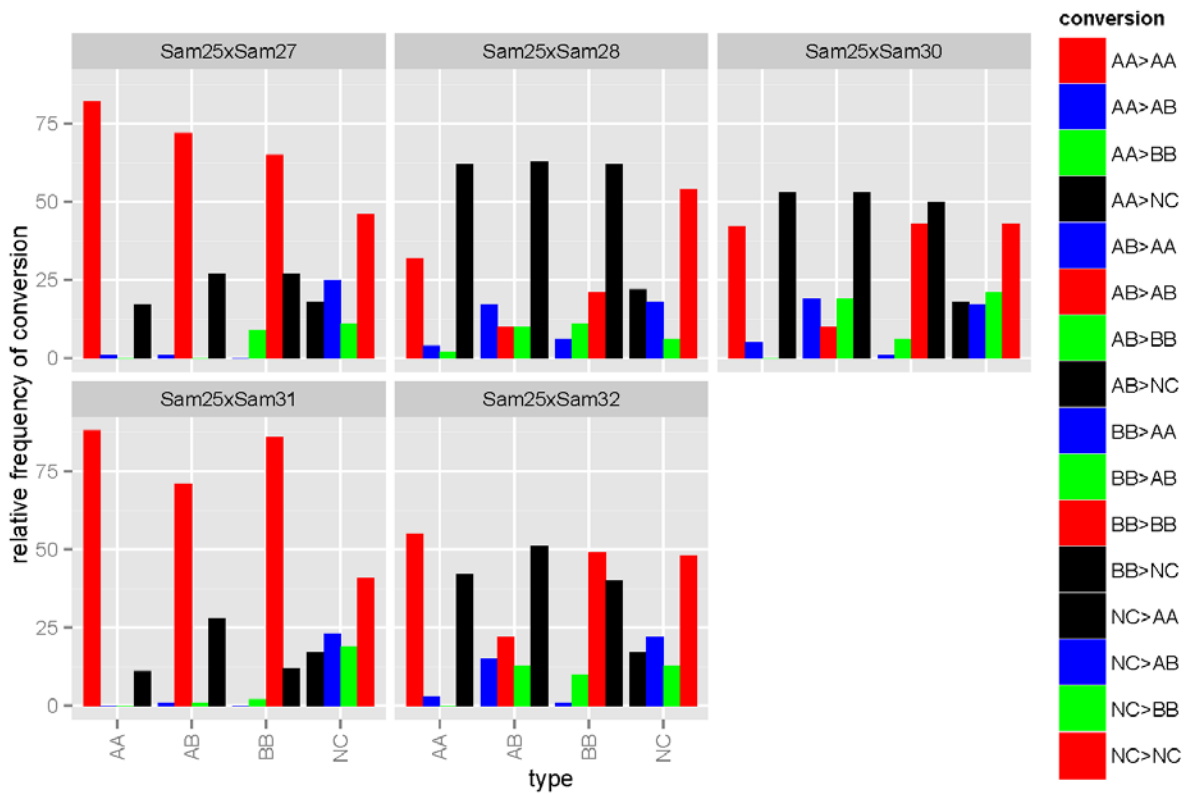
Relative frequencies of all possible genotype changes between Sample 1 and Samples 14, 16, and 19 to 24 for each possible genotype call (AA, AB, BB, NC) at each locus. Within each genotype (including NC) in Sample 1, relative frequencies of the genotype concordance or error were calculated from the total number of that genotype called in Sample 1. For simplicity all concordant genotype calls are shown in red and all genotypes called in Sample 1 and not called in the paired sample are shown in black; except for loci not called in Sample 1 (arbitrarily NC>AA is shown in black in this case).





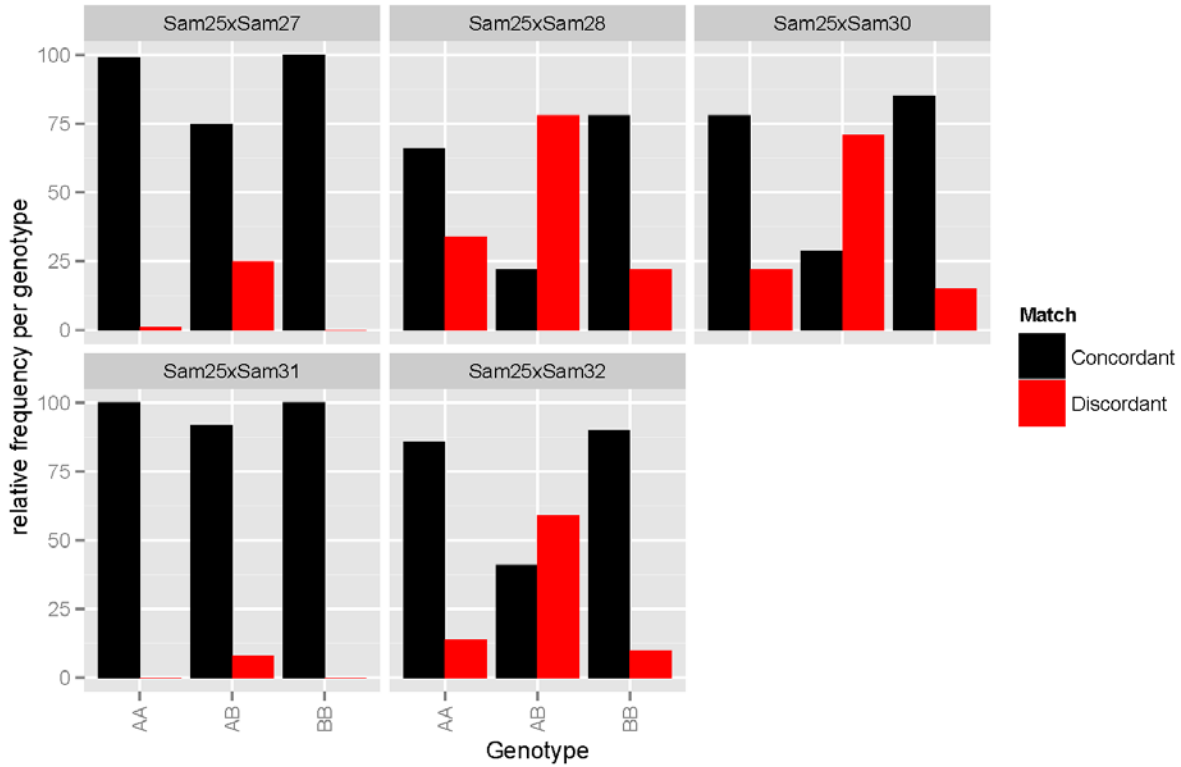
*Figure 11: HumanOmni2.5 BeadChip Genotyping Discordance*

Relative frequencies of all possible genotype changes between Sample 25 and Samples 27,28, 30,31, and 32 for each possible genotype call (AA, AB, BB, NC) at each locus. Within each genotype type called in Sample 25, relative frequencies of the genotype concordance or error were calculated from the total number of that genotype called in Sample 25. For simplicity all concordant genotype calls are shown in red and all genotypes called in Sample 25 and not called in the paired sample are shown in black; except for loci not called in Sample 25 (arbitrarily NC>AA is shown in black in this case).



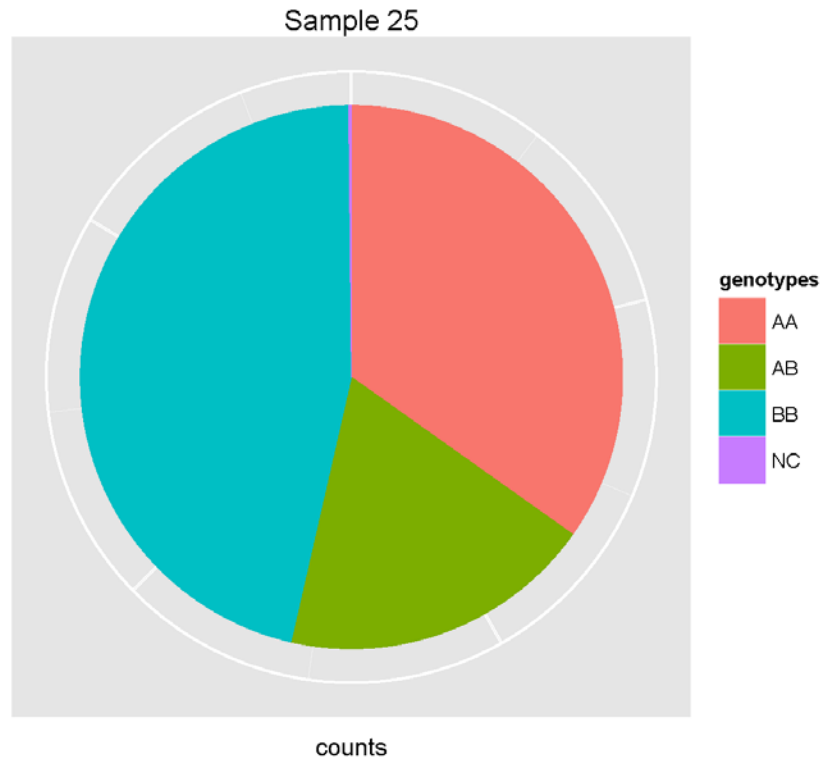
*Figure 12: HumanOmni2.5 BeadChip Observed Sample Discordance*

Relative frequencies of genotyping error given an AA, AB, or BB genotype call in Samples 27,28, 30,31, and 32. Only shared called loci are considered. Each red and blue column sums to 100% and represents all the loci called with each genotype in the experimental sample (Sample 25 can either be concordant with the same genotype, or discordant with a different, NC-excluded, genotype).



*Figure 13: Genotyping counts for Sample 25*

Counts of AA,AB,BB, and NC calls for the control sample of the HumanOmni2.5 BeadChip. Given the high call rate for Sample 25, NC only account for 3,970 loci out of 2,391,739 markers.



## Conclusions

### Discussion of findings

Illumina's Infinium HD FFPE Sample Solution is an officially supported workflow for Infinium BeadChip assays, and as such it is Illumina's default pipeline for clinical FFPE genotyping. It is unclear, however, how far DNA input limits can be pressed if an investigator has no option other than to genotype a degraded sample. In this study we tested those limits, designed an alternative solution for genotyping low copy number and degraded DNA, and optimized genotype accuracy of called SNPs on Infinium BeadChips. Samples prepared into sequencing libraries and amplified by PCR, without concatemerization and WGA (Samples 19-24), performed similarly to water blanks. For all other techniques tested, when using  $\geq 20$  ng DNA, fragmented to mean size of 100 or 300 bp, very high call rates and accuracy were achieved, equivalent to the suggested 200 ng input of high-quality DNA. With 10 ng DNA, performance was slightly decreased (Samples 13 and 26). With 2 ng quantities of fragmented DNA, on both BeadChips tested, our library preparation and concatemerization technique achieved call rates and concordance rates comparable or better than those processed through the Infinium FFPE DNA Restore protocol (Samples 16-18 vs. 12 and 31 vs. 27). With 200 pg of fragmented DNA, our library preparation and concatemerization technique yielded at least 57% more calls and a substantial (up to 20%) increase in accuracy relative to the FFPE Restore protocol on the HumanOmni2.5 BeadChip (Samples 30 and 32 vs. 28). The FFPE Restore protocol with 200 pg of DNA performed almost as poorly as blank samples (Samples 14 and 15, and Samples 28 and 29).

The Infinium assay uses an isothermal WGA reaction, likely similar to Qiagen's REPLI-g MDA (Gunderson et al. 2005), and as such it is not expected to perform well on short DNA fragments (Mead et al. 2008). To solve this issue, prior to the release of the Restoration Solution, Thirlwell et al. (2010) successfully ligated together fragments of DNA extracted from FFPE sections, using Qiagen's REPLI-g FFPE kit, then performed the Infinium Assay using HumanMethylation27 BeadChips and Illumina WGA reagents. Since the Infinium HD FFPE Restore Protocol integrates a ligation as a final proprietary step, it is possible that a

similar concatemerization strategy was chosen. Alternatively, fragments in the Restore Protocol may be circularized and the subsequent WGA in the Infinium HD assay is RCA (Tate et al. 2012).

In MDA WGA, irrespective of fragmentation levels, any long DNA molecules in a sample will contribute disproportionately to genotyping signals. This might bias the genotypes from a forensic sample with mixed DNA sizes towards those of the donor of the longer DNA fragments, or give the impression of a successful genotyping experiment when in reality only a small fraction of the DNA, the longest fragments, contributed the signal, while the fragmented DNA molecules did not contribute. Others have noted this concern when artificially degraded DNA was tested in genotyping protocols like ours (Ballantyne et al. 2007; Giardina et al. 2009). For this reason, we prepared fragmented DNA with long molecules specifically removed (see Methods).

While the Qiagen and Illumina solutions for handling fragmented DNA should allow MDA to proceed, unavoidable reaction cleanup steps cause substantial DNA loss and thus require higher initial DNA input amounts. Recognizing this, we took advantage of progress in the fields of next-generation sequencing of ChIP and ancient DNA, which now allow unbiased sequencing library creation from sub-nanogram amounts of DNA (Bowman et al. 2013; Gansauge and Meyer 2013; Kircher et al. 2012). We correctly hypothesized that these library techniques would allow genotyping of picograms of sample DNA, even if severely fragmented, by increasing PCR amplification cycles sufficiently. After capturing the sample in a library, we attempted two pipelines for genotyping: we ligated the library molecules together to form long concatemers that would support the Infinium WGA step (Samples 16-18, 30-32; Table 1), as described previously; and we PCR-amplified the libraries a second time to a plateau, in the hopes of generating enough DNA to bypass WGA and directly genotype on the BeadChip (Samples 19-24; Table 1).

The ligated library method, in particular using the ssDNA protocol (Gansauge and Meyer 2013), performed well. We view this as the optimal protocol for genotyping of degraded samples. PCA analysis of these genotyped samples indicated tight clustering with the high quality controls, with clear separation from both the blank sample and the various 'failed'

samples (Figures 7 and 8). At sub-nanogram quantities, using the high-density Omni2.5 Beadchip, the ligated ssDNA library clustered much closer to the high quality control and the various higher input samples, than did the ligated dsDNA library sample or the sub-nanogram restored sample (Figure 8A&B). This indicates to us that our method is capturing the true genetic signature of an artificially degraded sample, particularly at levels where the FFPE Restore protocol does not.

The non-WGA library preparation method failed (Samples 19-24). One possible reason is that our PCR-amplified library products never reached the DNA quantities usually produced from MDA. In our experience the Infinium Assay post precipitation and resuspension of the MSA5 plate generates ~100 µg of DNA. The BeadChip may not perform successfully without orders of magnitude more DNA than we provided.

When switching to the Infinium LCG Assay, FFPE-Restored, fragmented DNA samples from 10 ng to 0.2 ng input had lower call rates on the HumanOmni2.5 than did similarly treated samples genotyped with the Infinium HD FFPE method; however, given that the HumanOmni2.5 BeadChip carries 3.4X as many SNPs, the actual number of called SNPs was substantially higher. When applying our ligated library method, we recovered more SNPs with higher accuracy with both 2 ng and 0.2 ng of fragmented DNA input when compared to the FFPE Restore Protocol (Samples 27 and 28 vs. 30-32). In particular, the ssDNA library method with ligation using 200 pg of DNA allowed genotyping of >1.3 million SNPs on the HumanOmni2.5 at 80% concordance (Sample 32) with the control (Sample 25).

While the HumanOmniExpress-FFPE BeadChip has already been optimized for degraded samples, the HumanOmni2.5 BeadChip has not. The metrics above suggested further optimization of genotype calls for the HumanOmni2.5 BeadChip might be obtained. On the hypothesis that loci called in the absence of target DNA are most likely to be erroneous at small sample concentrations, we examined the effect of excluding from analysis the genotypes of all SNPs that were also called in the restored water blank (Sample 29). This increased the concordance rates by up to 3% (0-3% range) for individual A, with concomitant decreases in the number of SNPs called (30-37%; Samples 26-32, excluding 29). At the lowest input DNA amount, we were able to improve the results for Sample 32

(200 pg) to 82% concordance with the control, genotyping 933,694 SNPs (Tables 3 and 4). We also tested the ability of a multivariable logistic regression model to discriminate correct vs. incorrect genotypes. The model was estimated from a comparison of quality metrics from Sample 30 with control genotypes from Sample 25. We then applied the model to Sample 32 and increased the concordance rate from 80% to 88% while retaining approximately one million SNPs for analysis (Figure 5).

A further investigation of the type of genotyping error seen in the concatemerized sequencing library method suggests an abundance of AB>AA and AB>BB error when genotyping on the HumanOmni2.5 BeadChip (Figure 11) or the HumanOmniExpress-FFPE BeadChip (Figure 10; Sample 16). When an AB genotype is called in a fragmented and low input, greater than 50% of the time it is called in error (Figure 12). Potentially all AB calls could be masked in both the degraded sample and the reference sample, though further sampling and replications are necessary.

## **Implications for policy and practice**

High-density SNP genotyping microarrays are promising tools for forensic research, with the potential to deliver more than a million genome-wide SNP genotypes per sample. Such data could open the door to many forensically useful inferences: precise estimates of a DNA donor's biogeographical ancestry, the ancestries of his/her parents, even the degree and timing of population admixture events in the donor's past; estimation of any phenotypes previously identified in genome-wide association analyses; and powerful, accurate inferences of even distant relationships, using an algorithm we developed (Huff et al. 2011, Li et al. 2014).

The technique we developed and tested here allows upwards of 1 million SNPs to be genotyped from as little of 200 pg DNA, fragmented to 100bp. Although the resulting data are less complete and more error-prone than genotypes from high-quality DNA, they may allow estimation of biogeographic ancestry and extend the range of relatives whose DNA

might allow for positive identification of previously unidentifiable remains, especially for victims of disasters that have left no reference samples from the individuals or their first-degree relatives.

High density genotyping from degraded DNA samples has also been achieved by ‘shotgun’ sequencing multiple next-generation sequencing libraries from the same DNA sample, over multiple lanes of Illumina HiSeq flow cells (Rasmussen et al. 2014). Depending on the level of degradation and the desired sequencing coverage at each variant locus, required sequencing could vary anywhere from 3 lanes (for a modern whole genome at 30X coverage on an Illumina HiSeq 2500) to upwards of 20 lanes for an ancient DNA specimen. At current rates of \$800 (single end) to \$2000 (paired end) per lane, excluding library creation and QC costs, costs per individual using this methodology becomes prohibitively expensive for routine forensic cases.. By utilizing our novel library ligation method with the Infinium Beadchip technology, our costs are \$210 per sample for the Omni2.5 Beadchip with 8 samples, plus ~\$90 for library prep, QC, and ligation, for a grand total of \$300 per sample (The manufacturer suggested Infinium HD FFPE DNA Restore Kit costs ~\$80 a sample, so about the same price as the library ligation method but with much poorer performance.)

Genotyping so many loci from a single individual also brings with it potential privacy concerns, particularly as they might be linked to disease or private phenotypic traits. The forensic and/or legal communities might not want to explicitly genotype these loci by default and instead mask much of data unless such a need arises; we have previously shown that much less than a million SNPs are needed to reliably detect relationships through the 5<sup>th</sup> degree (Li et al. 2014). Additionally, one benefit of the BeadChips we have chosen is that the Omni line specifically target common and rare SNP content from the 1000 Genomes Project to maximally cover worldwide population diversity; they do not specifically target disease alleles or disease-linked loci or phenotype-informative loci. These Beadchips were chosen to help with ancestry and relationship estimations,



## Study Limitations and Further Research

The improved call rate and concordance rates of the concatemerized sequencing library method for low concentration and fragmented DNA samples suggests changes to our pipeline. Genotyping success from MDA amplified samples has been improved previously by either running replicates of the amplification reaction (Tzvetkov et al. 2005; Affymetrix GeneChip Arrays) or by using more than the recommended amount of DNA as input to the genotyping assay (Mead et al. 2008; Illumina GoldenGate Assay). In this work, we ligated 40 ng of library DNA and used all that remained following cleanup (11-20 ng; Samples 16-18 and 30-31, Table 1). We could increase this amount and potentially genotype more target library molecules following Infinium WGA.

After the BeadChips were scanned, our genotypes were called with the default clustering files for each chip. With the HumanOmniExpress-FFPE, this might not be a concern for degraded samples, but it could be a concern for the non-optimized HumanOmni2.5 BeadChip. Once a large set of degraded samples have been typed, manual re-clustering of SNP data, removal of markers that perform poorly relative to a control set (as suggested in Illumina Pub No. 970-2014-001), and calling algorithms other than the Illumina GenCall method (Ritchie et al. 2011) could be explored.

One bioinformatics pipeline we employed, following genotyping, used the information provided by our blank samples to adjust calls on target samples. This approach assumes that a consistent set of SNPs is either not called or falsely called in the blank samples (Loci most likely to be 'refractory'). In this report only one blank was run for each of the two types of Beadchips. Further measurements of blank samples will be made using multiple BeadChips to validate how many loci are consistently assayed incorrectly, particularly with the HumanOmni2.5 BeadChip. Furthermore, our blank samples were FFPE-Restored and not samples from negative library preparations. Comparing these two types of blanks would be informative, especially considering that our non-WGA experimental samples failed at different loci than our blank sample.

While DNA fragmentation and low quantities are two main characteristics of forensic and ancient DNA, decreased sample quality presents just as much a roadblock to successful genotyping. Nucleotide bases are damaged over time, somewhat predictably (Briggs et al. 2007); soil bacterial DNA can be co-extracted and co-amplified from a sample, which may interfere with probes on the BeadChips; and PCR inhibitors could be present in the DNA extracts. This study focused on artificially degraded DNA samples. Future research in progress will apply this methodology to five 70-year old forensic-quality remains, with modern 1<sup>st</sup> and 2<sup>nd</sup> degree relatives, and six remains from missing persons unearthed between 1992-1995, with unreported provenance.

From our results with fragmented 2 ng or greater input samples, we know that the high strength and specificity of ERSA's relationship detection could easily be applied; the error rates we observed here are consistent with those encountered from high quality samples (Huff et al. 2011, Li et al 2014). For the sub-nanogram input, increased error would require increased error tolerance when reconstructing shared identical-by-descent (IBD) segments, which is a necessary processing step before running ERSA. If error tolerance becomes too permissive, individuals might randomly share very large IBD segments and present as being more related than in reality (ERSA's sensitivity at 6<sup>th</sup> degree relationships and farther would work against it in these situations).

There are two potential, not mutually exclusive, solutions to this problem. First, there are alternative approaches for inferring kinship from high-density SNP data using genome-wide average metrics, such as RELPAIR (Epstein et al. 2000). These methods can reliably detect relationships as distant as first cousins with fewer than 10,000 markers (Huff et al. 2011). Given the higher observed error of AB genotype calls (Figure 12), genome-wide average models could be adapted to ignore heterozygote calls and treat diploid homozygote calls as haploid when computing the likelihood of marker genotype data. Second, IBD segments could be mapped with a shared genomic segment (SGS) (Thomas et al. 2008) analysis. We observed lower error rates for homozygote calls and a very low to non-existent reciprocally homozygote error rate (AA>BB, BB>AA; Figure 11). Because SGS only considers homozygote genotype calls, SGS should be far less sensitive to the error patterns observed in sub-nanogram samples. IBD segments inferred from SGS

would enable more distant relationship detection in ERSA. Both approaches will be the focus of future work.

## References

- Al Safar H, Abidi F, Khazanehdari K, Dadour I, and Tay G. 2011. Evaluation of different sources of DNA for use in genome wide studies and forensic application. *Appl Microbiol Biotechnol* 89(3):807-815.
- Ballantyne KN, van Oorschot RAH, and Mitchell RJ. 2007. Comparison of two whole genome amplification methods for STR genotyping of LCN and degraded DNA samples. *Forensic Science International* 166(1):35-41.
- Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BV, Rasanayagam A, and Hammer MF. 1998. Female gene flow stratifies Hindu castes. *Nature* 395(6703):651-652.
- Bowman SK, Simon MD, Deaton AM, Tolstorukov M, Borowsky ML, and Kingston RE. 2013. Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC genomics* 14(1):466.
- Briggs AW, Stenzel U, Johnson PL, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, and Lachmann M. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences* 104(37):14616-14621.
- Croft Jr DT, Jordan RM, Patney HL, Shriver CD, Vernalis MN, Orchard TJ, and Ellsworth DL. 2008. Performance of whole-genome amplified DNA isolated from serum and plasma on high-density single nucleotide polymorphism arrays. *The Journal of Molecular Diagnostics* 10(3):249-257.
- Dabney J, Knapp M, Glocke I, Gansauge M-T, Weihmann A, Nickel B, Valdiosera C, García N, Pääbo S, Arsuaga J-L et al. . 2013. Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proceedings of the National Academy of Sciences*.
- Epstein, M.P., Duren, W.L., and Boehnke, M. 2000. Improved inference of relationship for pairs of individuals. *Am J Hum Genet* 67: 1219-1231.
- Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, and Pääbo S. 2013. DNA analysis of an early modern human from Tianyuan Cave, China. *Proceedings of the National Academy of Sciences* 110(6):2223-2227.
- Gansauge M-T, and Meyer M. 2013. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature protocols* 8(4):737-748.
- Giardina E, Pietrangeli I, Martone C, Zampatti S, Marsala P, Gabriele L, Ricci O, Solla G, Asili P, Arcudi G et al. . 2009. Whole genome amplification and real-time PCR in forensic casework. *BMC genomics* 10(1):159.
- Gunderson KL, Steemers FJ, Lee G, Mendoza LG, and Chee MS. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature genetics* 37(5):549-554.
- Hadd AG, Houghton J, Choudhary A, Sah S, Chen L, Marko AC, Sanford T, Buddavarapu K, Krosting J, Garmire L et al. . 2013. Targeted, High-Depth, Next-Generation Sequencing of Cancer Genes in Formalin-Fixed, Paraffin-Embedded and Fine-Needle Aspiration Tumor Specimens. *The Journal of Molecular Diagnostics* 15(2):234-247.
- Han T, Chang C-W, Kwekel J, Chen Y, Ge Y, Martinez-Murillo F, Roscoe D, Tezak Z, Philip R, Bijwaard K et al. . 2012. Characterization of whole genome amplified (WGA) DNA for use in genotyping assay development. *BMC genomics* 13(1):217.

- Hosein AN, Song S, McCart Reed AE, Jayanthan J, Reid LE, Kutasovic JR, Cummings MC, Waddell N, Lakhani SR, Chenevix-Trench G et al. . 2013. Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP-CGH analysis. *Lab Invest* 93(6):701-710.
- Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, Tuohy TM, Neklason DW, Burt RW, Guthery SL et al. . 2011. Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome Research* 21(5):768-774.
- Jasmine F, Ahsan H, Andrulis IL, John EM, Chang-Claude J, and Kibriya MG. 2008. Whole-Genome Amplification Enables Accurate Genotyping for Microarray-Based High-Density Single Nucleotide Polymorphism Array. *Cancer Epidemiology Biomarkers & Prevention* 17(12):3499-3508.
- Jobling MA and Gill P. 2004. Encoded evidence: DNA in forensic analysis. *Nat Rev Genet* 5: 739-751.
- Kircher M, Sawyer S, and Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research* 40(1):e3.
- Li H, Glusman G, Hu H, Shankaracharya, Caballero J, Hubley R, Witherspoon D, Guthery SL, Mauldin DE, Jorde LB et al. 2014. Relationship Estimation from Whole-Genome Sequence Data. *PLoS Genet* 10(1):e1004144.
- Li J, Harris L, Mamon H, Kulke MH, Liu W-H, Zhu P, and Mike Makrigiorgos G. 2006. Whole genome amplification of plasma-circulating DNA enables expanded screening for allelic imbalance in plasma. *The Journal of Molecular Diagnostics* 8(1):22-30.
- Maciejewska A, Jakubowska J, and Pawlowski R. 2013. Whole genome amplification of degraded and nondegraded DNA for forensic purposes. *International journal of legal medicine* 127(2):309-319.
- Mead S, Poulter M, Beck J, Uphill J, Jones C, Ang CE, Mein CA, and Collinge J. 2008. Successful amplification of degraded DNA for use with high-throughput SNP genotyping platforms. *Human Mutation* 29(12):1452-1458.
- Menon R, Deng M, Boehm D, Braun M, Fend F, Boehm D, Biskup S, and Perner S. 2012. Exome Enrichment and SOLiD Sequencing of Formalin Fixed Paraffin Embedded (FFPE) Prostate Cancer Tissue. *International Journal of Molecular Sciences* 13(7):8933-8942.
- Meyer M, Fu Q, Aximu-Petri A, Glocke I, Nickel B, Arsuaga J-L, Martinez I, Gracia A, de Castro JMB, Carbonell E et al. . 2014. A mitochondrial genome sequence of a hominin from Sima de los Huesos. *Nature* 505(7483):403-406.
- Meyer M, and Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols* 2010(6):pdb. prot5448.
- Musgrave-Brown E, Ballard D, Balogh K, Bender K, Berger B, Bogus M, Borsting C, Brion M, Fondevila M, Harrison C et al. . 2007. Forensic validation of the SNPforID 52-plex assay. *Forensic Sci Int Genet* 1(2):186-190.
- Pardo-Seco J, Martinon-Torres F, and Salas A. 2014. Evaluating the accuracy of AIM panels at quantifying genome ancestry. *BMC Genomics* 15:543.

- Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford Jr TW, Rasmussen S, Moltke I, Albrechtsen A, Doyle SM et al. . 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* 506(7487):225-229.
- Ritchie M, Liu R, Carvalho B, Australia T, Consortium NZMSG, and Irizarry R. 2011. Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. *BMC Bioinformatics* 12(1):68.
- Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, and Chisholm SW. 2010. Unlocking short read sequencing for metagenomics. *PLoS One* 5(7):e11840.
- Rohland N, and Reich D. 2012. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Research*.
- Tate CM, Nuñez AN, Goldstein CA, Gomes I, Robertson JM, Kavlick MF, and Budowle B. 2012. Evaluation of circular DNA substrates for whole genome amplification prior to forensic analysis. *Forensic Science International: Genetics* 6(2):185-190.
- Thirlwell C, Eymard M, Feber A, Teschendorff A, Pearce K, Lechner M, Widschwendter M, and Beck S. 2010. Genome-wide DNA methylation analysis of archival formalin-fixed paraffin-embedded tissue using the Illumina Infinium HumanMethylation27 BeadChip. *Methods* 52(3):248-254.
- Thomas A, Camp NJ, Farnham JM, Allen-Brady K, Cannon-Albright LA. 2008. Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann Hum Genet* 2008, 72(Pt 2):279-87.
- Tzvetkov MV, Becker C, Kulle B, Nürnberg P, Brockmöller J, and Wojnowski L. 2005. Genome-wide single-nucleotide polymorphism arrays demonstrate high fidelity of multiple displacement-based whole-genome amplification. *ELECTROPHORESIS* 26(3):710-715.
- Watkins W, Prasad B, Naidu J, Rao B, Bhanu B, Ramachandran B, Das P, Gai P, Reddy P, and Reddy P. 2005. Diversity and divergence among the tribal populations of India. *Annals of human genetics* 69(6):680-692.
- Watkins WS, Bamshad M, Dixon ME, Bhaskara Rao B, Naidu JM, Reddy PG, Prasad BV, Das PK, Reddy PC, Gai PB et al. . 1999. Multiple origins of the mtDNA 9-bp deletion in populations of South India. *Am J Phys Anthropol* 109(2):147-158.
- Xing J, Watkins WS, Zhang Y, Witherspoon DJ, and Jorde LB. 2008. High fidelity of whole-genome amplified DNA on high-density single nucleotide polymorphism arrays. *Genomics* 92(6):452-456.
- Yost SE, Smith EN, Schwab RB, Bao L, Jung H, Wang X, Voest E, Pierce JP, Messer K, Parker BA et al. . 2012. Identification of high-confidence somatic mutations in whole genome sequence of formalin-fixed breast cancer specimens. *Nucleic acids research* 40(14):e107.

## Dissemination of Research Findings

### Publications

Upcoming publication (In Review November 2015): Tackney JC, Jorde LB, and Witherspoon DJ. High-density microarray SNP genotyping from artificially fragmented and low concentration DNA samples. *Journal of Forensic Science*

Related Publication: Li H, Glusman G, Hu H, Shankaracharya , Caballero J, Hubley R, Witherspoon D, Guthery SL, Mauldin DE, Jorde LB, Hood L, Roach JC, and Huff CD. (2014) Relationship Estimation from Whole-Genome Sequence Data. *PLoS Genet* 10(1).

### Presentations

Witherspoon DJ, Tackney JC, Parsons TJ, and Jorde LB.  
2014 *Relationship detection with high-density SNP genotypes obtained from sub-nanogram amounts of fragmented DNA.*  
IN: 64<sup>th</sup> Annual Meeting of the American Society of Human Genetics. San Diego, CA. October 18-22

Tackney JC, Witherspoon DJ, and Jorde LB.  
2013 *Successful Illumina Infinium Beadchip high-density genotyping from fragmented and low concentration samples.*  
IN: 63<sup>rd</sup> Annual Meeting of the American Society of Human Genetics. Boston, MA. October 22-26.

Witherspoon DJ, Tackney J, Parsons TJ, and Jorde LB.  
2013 *High-confidence detection of distant relationships using high-density SNP genotypes obtained from small amounts of degraded DNA*  
IN: 25<sup>th</sup> International Congress of the International Society for Forensic Genetics. Melbourne, Australia. September 2-7.

### Other

Detailed, bench-ready protocols have been prepared and delivered to scientists at the International Commission on Missing Persons (ICMP). At the American Academy of Forensic Sciences (AAFS) meeting (Feb 2015), Dr. Witherspoon discussed this work and its potential with industry, government, law enforcement, nonprofit and academic researchers and institutional representatives, all of whom are developing or are interested in using high-throughput genotyping capabilities for forensic DNAs.