



ORIGINAL ARTICLE

Genomic characterization of MICA gene using multiple next generation sequencing platforms: A validation study

Yizhou Zou¹  | Jamie L. Duke²  | Deborah Ferriola² | Qizhi Luo¹ | Jenna Wasserman² | Timothy L. Mosbrugger² | Weiguang Luo¹ | Liang Cai¹ | Kevin Zou¹ | Nikolaos Tairis² | Georgios Damianos² | Ioanna Pagkrati² | Debra Kukuruga³ | Yanping Huang² | Dimitri S. Monos^{2,4}

¹Department of Immunology, Central South University Xiangya School of Medicine, Changsha, Hunan, China

²Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania

³Department of Pathology, University of Maryland School of Medicine, Baltimore, Maryland

⁴Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania

Correspondence

Dimitri S. Monos, Department of Pathology and Laboratory Medicine, The Children's Hospital of Philadelphia, Abramson Research Bldg. 707A, 3615 Civic Center Blvd., Philadelphia, PA 19104.
Email: monosd@email.chop.edu

Present address

Yanping Huang, Department of Pathology, Anatomy and Cell Biology, Thomas Jefferson University, Philadelphia, Pennsylvania

Funding information

Children's Hospital of Philadelphia; Hunan Science and Technology Project Foundation, Grant/Award Number: 2018JJ2549; National Natural Science Foundation of China, Grant/Award Numbers: 81571562, 81873875

We have developed a protocol regarding the genomic characterization of the MICA gene by next generation sequencing (NGS). The amplicon includes the full length of the gene and is about 13 kb. A total of 156 samples were included in the study. Ninety-seven of these samples were previously characterized at MICA by legacy methods (Sanger or sequence specific oligonucleotide) and were used to evaluate the accuracy, precision, specificity, and sensitivity of the assay. An additional 59 DNA samples of unknown ethnicity volunteers from the United States were only genotyped by NGS. Samples were chosen to contain a diverse set of alleles. Our NGS approach included a first round of sequencing on the Illumina MiSeq platform and a second round of sequencing on the MinION platform by Oxford Nanopore Technology (ONT), on selected samples for the purpose of either characterizing new alleles or setting phase among multiple polymorphisms to resolve ambiguities or generate complete sequence for alleles that were only partially reported in the IMGT/HLA database. Complete consensus sequences were generated for every allele sequenced with ONT, extending from the 5' untranslated region (UTR) to the 3' UTR of the MICA gene. Thirty-two MICA sequences were submitted to the IMGT/HLA database including either new alleles or filling up the gaps (exonic, intronic and/or UTRs) of already reported alleles. Some of the challenges associated with the characterization of these samples are discussed.

KEYWORDS

genotyping, Illumina, MICA, NGS, Oxford Nanopore

Yizhou Zou and Jamie L. Duke contributed equally to this study.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. HLA: Immune Response Genetics published by John Wiley & Sons Ltd.

1 | INTRODUCTION

The major histocompatibility complex class I chain-related gene A (MICA) is located within the MHC and centromerically to HLA-B (43.5 kb apart).^{1,2} The MICA gene encodes a protein with a domain structure similar to the heavy chain of HLA class I molecules but does not associate with β 2-microglobulin and does not bind peptides like the HLA class I molecules. However, the molecule is quite polymorphic (more than 159 alleles corresponding to 92 different proteins and 4 null alleles, IMGT/HLA version 3.39.0) and functions as a stress-induced ligand for integral membrane protein receptor NKG2D. The NKG2D receptor is present on the cell surface of NK cells, $\gamma\delta$ T cells, and CD8⁺ T cells,³ whereby upon interaction with MICA these cells are activated and elicit their cytotoxic effect targeting epithelial tumor cells or other stressed cells (autoimmune diseases, DNA damage, ischemia-reperfusion injury, inflammation, and viral infections).⁴ MICA is mainly expressed intracellularly with just a small fraction appearing on the surface of some epithelial cells.⁵ Variations of the MICA gene are associated with a number of diseases related to NK activity, such as viral infections, cancer and allograft rejection. It remains debatable as to whether MICA polymorphisms are associated with graft vs host disease in hematopoietic cell transplantation.^{6,7}

Expanding on the importance of MICA in solid organ transplantation and specifically the role it may play in organ rejection, the possible mechanisms of its involvement can either be the direct recognition of MICA on allografts by NKG2D⁸ and/or the development of anti-MICA antibodies, when the donor and recipient are mismatched for MICA. An increased amount of evidence suggests that there is a link between anti-MICA antibodies and graft rejection.⁹ It therefore becomes relevant for the accurate assessment of alloreactivity to determine reliably the exact polymorphism of this gene.

Until recently the primary methods for MICA typing were PCR-SSP and PCR-SSOP, with the gold standard being Sanger sequencing of targeted subsegments of the MICA gene (exon 2 to exon 5 included).¹⁰ However due to the sequencing idiosyncrasy of this gene with the number of short tandem repeats (STRs) in exon 5, and difficulty in determining the phase of single nucleotide polymorphisms between different exons, its accurate and unambiguous characterization is quite challenging. The complexity presented by this gene, though, can be adequately addressed by using different next generation sequencing (NGS) technologies that when combined, can provide not only accurate but also thorough and unambiguous characterization of the MICA gene. In this study, we have validated our approach using a set of samples

that were typed for the MICA gene by using either SSOP or Sanger sequencing and then typed by using two NGS technologies: the MiSeq Illumina platform to provide the typing for most of the samples, when phasing of distant polymorphisms was attainable, supported by Oxford Nanopore Technology (ONT) when phasing and generation of complete and unambiguous consensus sequence of the entire MICA gene was challenging. The gene was amplified from 5' untranslated region (UTR) to the completion of the 3' UTR region. New alleles were described and the exonic and intronic sequences of already described alleles were completed.

2 | MATERIALS AND METHODS

2.1 | DNA samples

A total of 156 samples were amplified and sequenced for this study. Ninety-seven of these samples were previously characterized at MICA and were used to evaluate the accuracy, precision, specificity, and sensitivity of the assay. All genotype results produced by NGS for these 97 samples were compared to sequence specific oligonucleotide (SSO) probes or Sanger sequence-based typing (SBT). Samples were chosen in part to contain a diverse set of alleles. Of the 97 samples, 56 were received in our lab as DNA specimens of unknown tissue origin, 34 samples were extracted from whole blood (EDTA or ACD), 3 samples were extracted from crushed clots, and 4 samples were extracted from buccals swabs and cleaned and concentrated with Microcon-30 kDa centrifugal columns (MilliporeSigma, Burlington, Massachusetts). Of the 97 samples in the validation portion of the study, 46 of the specimens were from a South-Chinese Han population of volunteers, whereas the remaining 51 specimens were from individuals of various ethnicities within the US population, including Caucasian, African American, Hispanic, and Asian/Pacific Islanders. Even though, the ethnicity is known for some of the individuals, we do not have a full account of the ethnic background of all 51 individuals. An additional 59 DNA samples from US-based volunteers with unknown ethnicity were extracted from whole blood and amplified at MICA but were only genotyped by NGS. These were included in the evaluation of the amplification performance and quality metrics for sequencing.

2.2 | MICA genotyping by Sanger SBT

PCR amplification for MICA-SBT was performed for the 46 South-Chinese Han samples with generic primers to

generate the target DNA sequences covering from exon 2 to exon 5. The methods for MICA-SBT PCR including primers and cycling parameters were previously described.^{10,11} The fragments of the PCR products were visualized in 1.2% agarose gel under UV light and recorded by photograph. A total of 10 μ L amplified PCR products were incubated with one unit of shrimp alkaline phosphatase and two units of ExoI (USB Corp., Cleveland, Ohio). The cleaned DNA segments were sequenced with four primers on an ABI 3730 Sequencing System (PE Biosystems).¹¹ Four electronic sequencing files were used for MICA allele assignment on the software which was designed to assemble and align MICA sequences containing exons 2, 3, 4, and 5 with *MICA*001* as consensus. MICA alleles from the IMGT/HLA database v. 3.28 (April 2017, <http://www.ebi.uk/imgt/hla/align.html>) were used. For MICA short tandem repeat (STR) polymorphism, overlay of sequencing signals may occur due to potentially different number of (GCT)*n* repeats and deletion mutation in exon 5. These heterogeneity sequencing signals were used to identify the MICA-STR genotypes using an in-house designed computer algorithm. The resolution of MICA typing may be limited by not sequencing exons 1 and 6 of MICA. Ambiguous allele assignments can occur when two alleles are present and the composite sequence is identical for more than one combination (cis/trans ambiguities). The assigned MICA alleles for each samples were blinded to each MICA typing method. All homozygous alleles were confirmed by MICA PCR-SSP typing as previously described.¹²

2.3 | MICA genotyping by SSO

Fifty-one samples were also genotyped using the OneLambda LABType SSO kit for MICA (West Hills, California) following the manufacturer's protocol. MICA genotypes were assigned using the HLA Fusion software from OneLambda with the default settings and an IMGT/HLA database version that was within at least 1 year of the most current version at the time of typing.

2.4 | NGS-based sequencing and analysis of full-length MICA genes

Amplification of the MICA gene was performed with primers developed at the Children's Hospital of Philadelphia (CHOP) that delineate the entire length of MICA from the middle of the 5' UTR (CCG TGC TTA TGA AGT TGG AGC TG, GRCh38 chr6:31403325-31403347) through the entire gene and ends approximately 500 bases past the end of the 3' UTR (ACA GAC CTC TCT TTC

TCC CTG AAC C, GRCh38 chr6:31416289-31416313), creating a amplicon of approximately 13.0 Kb. Reagents from Qiagen LR PCR kits (Valencia, California) were combined with the MICA primers for amplification. Amplification conditions per 25 μ L reaction consisted of 1X Qiagen LR PCR buffer, 0.5 mM each dNTP, 0.4 μ M each primer, 2 U LR enzyme mix and approximately 150 ng of DNA. PCR was performed on a ThermoFisher Veriti thermal cycler (Waltham, Massachusetts) programmed for 3 minutes at 95°C for initial denaturation, followed by 35 cycles of 15 seconds at 95°C, 30 seconds at 62°C, and 13 minutes at 68°C. Cycling was followed by a final extension at 68°C for 10 minutes.

Library preparation for NGS was performed using the standard library preparation method and reagents from the Omixon V2 Holotype HLA kits (Budapest, Hungary). Briefly, MICA amplicons were quantitated with QuantiFluor dsDNA system (Promega, Madison, Wisconsin) and diluted to approximately 150 ng/ μ L, enzymatically fragmented, end repaired, ligated to indexed adaptors, pooled, and size selected on a Blue Pippin (Sage Science, Beverly, Massachusetts). The size selected MICA library was quantitated by qPCR with KAPA library quantification kits, (Wilmington, Massachusetts) diluted to 9 pM, and sequenced on an Illumina MiSeq (San Diego, California) using paired-end 2 X 150 V2 sequencing chemistry.

Data were demultiplexed and fastq files were generated by MiSeq Reporter on the platform. Fastq files were analyzed with Omixon Twin (versions 2.1.2 and 3.1.3) and GenDx NGSengine (Utrecht, Netherlands, version 2.13). The resulting genotype and quality metrics from both software programs using the MiSeq data were compared to produce a high confidence final genotype through an in-house CHOP developed software program.

2.5 | Oxford Nanopore sequencing

Twenty-two samples with alleles that were either not fully characterized in the IMGT/HLA database version 3.27, possessed ambiguities in the Illumina genotyping, or had novelty in either the exons or introns were also sequenced on a MinION (Oxford Nanopore Technologies, Oxford, UK) to obtain fully phased sequence and resolve ambiguities. These 22 samples were sequenced on multiple sequencing runs using an ONT 1D Native barcoding kit for genomic DNA with the manufacturer's protocol version NBE_9006_v103_revQ_21Dec2016 beginning at the end repair step. In short, 2 μ g of each amplicon were end repaired and dA-tailed with New England Biolabs (NEB) Ultra II End-prep enzyme and buffer (Ipswich, Massachusetts), ligated to ONT native barcodes with NEB Blunt/TA Ligase Master Mix, pooled equimolar, and

then ligated to ONT Barcode Adaptor Mix with NEB NEBNext 5X Quick Ligation Reaction Buffer and Quick T4 DNA Ligase. Between each step in the ONT protocol, reactions were cleaned with AMPure XP beads. The final adaptor ligated, barcoded pool was loaded on an ONT SpotON SQK-LSK108 flow cell on the MinION and amplicon strands were processed through the nanopores on the flow cell for 2 hours. Local basecalling was performed.

2.6 | Oxford nanopore sequence analysis and consensus generation

The general method to error correct Oxford Nanopore reads using Illumina data have been previously published.¹³ ONT reads were demultiplexed using Porechop version 0.2.3 and filtered to retain read lengths between 10 and 15 kb (R.Wick, Melbourne, VIC, Australia; <https://github.com/rrwick/Porechop>, last accessed May 1, 2018). Correction of the ONT reads with Illumina reads was performed with FMLRC version 0.1.2 using default settings to produce long hybrid reads that were used for subsequent analyses. MAFFT version 7.394 was used to generate multiple sequence alignment (MSA) on 100 randomly selected corrected ONT reads per sample. The adjust-direction argument was used to account for unstranded reads. The function `msaConsensusSequence` from the Bioconductor package `msa` version 1.10.0 was used to generate a consensus sequence from the MAFFT MSA. The resulting consensus was cleaned by removing gaps and replacing question marks with N and was reported in the sense orientation. A total of 1000 corrected ONT reads were aligned to the consensus sequence using `minimap2` version 2.10, and the alignments were left aligned using `GATK` function `LeftAlignIndels` version 3.3.0.

Base count pileups were created using `Pysam` version 0.14 (<https://pypi.org/project/pysam>, last accessed February 2018) and scanned for polymorphic positions, which were defined as positions with two nucleotides at a frequency of 0.30 or more and a deletion frequency less than 0.10. The observed bases at the polymorphic positions are stored for each read in the alignment. For each set of 10 polymorphic positions, the most commonly observed base combination is stored to create the most likely sequence for allele 1. The opposite base combination is created to represent the most likely sequence for allele 2. The Levenshtein distance similarity ratio is calculated between each read's polymorphic positions and the derived sequences of allele 1 and allele 2. If the ratio is greater than 75 in one of the comparisons, the read is assigned to the appropriate allele. If ratio is less

than 75 in both comparisons, it is removed from analysis. MAFFT was used to generate allele-specific MSA using 100 randomly selected allele-specific ONT reads. The function `msaConsensusSequence` was used to generate allele-specific consensus sequences from the MAFFT MSA. The allele-specific consensus sequences were compared to consensus sequences generated by `NGSEngine` using the original Illumina data and corrected ONT data. Any differences were manually inspected and resolved to derive the final MICA consensus sequence.

3 | RESULTS

Ninety-seven samples that were previously characterized by either SBT (exons 2-5 as one amplified region of 2.2 Kb) or OneLambda LABType SSOP (exons 2, 3, 4, and 5 individually) were selected for validating the protocol for the characterization of the MICA gene by NGS. An attempt was made to include as many different MICA alleles as possible. Two separate populations were used for this purpose, one sample set of 46 individuals from a South-China Han population and a second sample set of 51 individuals of various ethnicities from a US-based population. Furthermore, an additional set of 59 samples from US-based volunteers with unknown ethnicity was used to assess the quality of the overall MICA assay. These supplemental samples were not part of the validation study as no genotyping results from a second method were available for the MICA gene.

In brief, the entire MICA gene was amplified using long-range PCR, sequenced on the Illumina MiSeq platform using paired-end 2×150 chemistry. The resulting fastq files from Illumina were analyzed with two programs, the results combined, and a genotype determined based on the Illumina data. A small subset of samples were subsequently sequenced on the ONT MinION platform to address limitations of the Illumina platform, while in reverse the ONT sequencing reads were corrected using the Illumina fastq files, and using our pipeline produced a final consensus sequence for each allele.

3.1 | Amplifying the MICA gene

A single 13.0 kb amplicon was designed to include the entire MICA gene starting with a portion of the 5' UTR and extending 500 bases past the IMGT/HLA-defined 3' UTR (Figure 1). Given the length of the MICA amplicon (13 kb), long-range PCR is highly affected by the quality of the DNA extraction method. Amplification of the

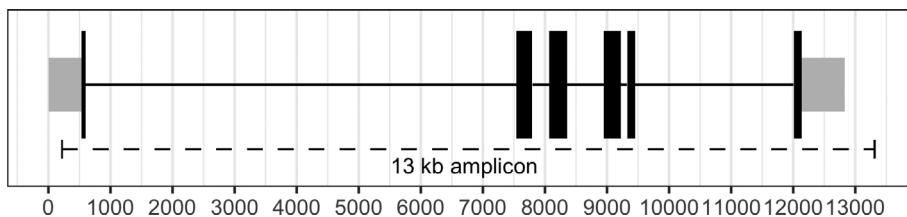


FIGURE 1 MICA gene. The exons are in the tall black bars, the untranslated regions are the shorter gray bars, while the introns are represented by lines. The amplicon used for next generation sequencing (NGS) is shown below with the dashed line

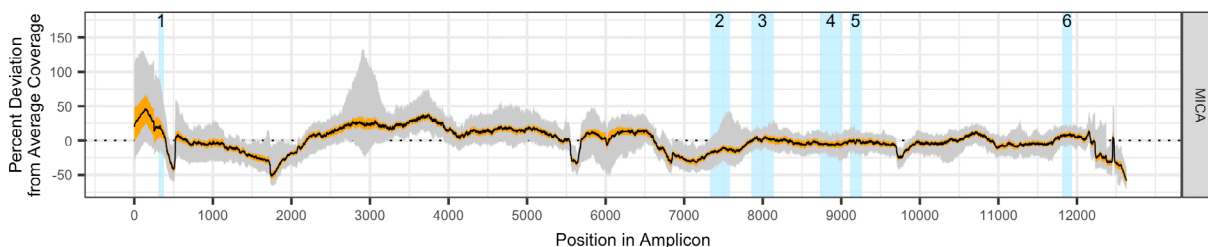


FIGURE 2 Uniformity of coverage across the MICA amplicon. The average depth of coverage was calculated for each sample and the depth across the amplicon was normalized to the mean to show the percent from which the depth at each position varies from the mean (represented as zero, dotted line). The average deviation per position (black line), the 25th to 75th quartile (orange), and the full range of deviation (gray) are shown for the samples with robust amplification (amplicon concentration greater or equal to 125 ng/ μ L). The background is highlighted in blue for the exons and labeled with the exon number

MICA gene was considered to be successful if the amplicon concentration was measured to be at least 50 ng/ μ L, and robust if the concentration was 125 ng/ μ L or greater. For the 97 validation samples, 18 of 97 samples failed to amplify on the first attempt (15 South-China Han, 3 US-based samples). All samples were able to produce a viable amplicon for sequencing when attempted a second time. In three instances, the amplicon concentration was below 50 ng/ μ L (35, 38, and 47 ng/ μ L), but a visible band at 13 kb was observed after gel electrophoresis (data not shown) and these samples were processed for sequencing. No failures were observed during the amplification of the 59 supplemental samples.

3.2 | Sequencing on the Illumina MiSeq platform

Samples were sequenced on a MiSeq using paired-end 2 x 150 sequencing chemistry on V2 flow cells. For the samples that had a robust amplification among all samples sequenced (130/156 samples), the coverage across the length of the 13 kb amplicon fluctuates, but is generally well represented. We can assess the uniformity of coverage for these samples by standardizing the depth of coverage at each position of the amplicon relative to the average depth of coverage within each sample, which allows for comparison across samples with varying depths of coverage due to differences in the amount of reads obtained from sequencing each sample. As can be seen in Figure 2, there is fluctuation in the depth of

coverage across the length of the 13 kb amplicon for robustly amplified samples, but generally the deviation from the mean is not substantial. There are four areas that are consistently lower than the mean in intron 1, where the coverage is between 25% and 50% lower than the average depth of coverage at the approximate positions of 500, 1750, 5600, and 6800 to 7200 bases away from the start of the amplicon. The remaining samples that did not have robust amplification at MICA (26/156 samples) indicated by low amplicon concentrations had considerably less uniformity of coverage showing a high depth of coverage at the ends of the amplicon and low depth of coverage through the majority of the gene, signifying that PCR started but was incomplete across the targeted region. The genotyping, though, of these less robustly amplified samples was not compromised because an adequate number of reads were generated during sequencing for each of these samples.

The Illumina fastq files for each sample were analyzed by two different software programs, Omixon Twin and GenDx NGSengine, and the genotype and quality metrics from both programs were compared, where a final genotype was determined based on the accumulation of data from both programs. The NGS method was able to accurately detect the presence of all alleles detected by the legacy typing methods (Sanger SBT or LABType SSOP) for the 97 validation samples (Table 1 and Supplemental Table S1). The sensitivity, specificity, precision and accuracy of the NGS method for MICA are all 100% when compared to the results of the legacy methods. Both legacy methods had difficulty to

unambiguously assign an allele call, whereby 55.7% (108/194) of all allele calls had at least one alternative allele possible at the first or second field and 30.9% (60/194) of all allele calls had at least one alternative allele with a different protein sequence possible. We excluded the ability to detect the *MICA*019* vs *MICA*019:01/MICA*019:02* as an ambiguity since the *MICA*019:02* allele was not described until July 2015 and may not have been described at the time of the legacy typing. By short-read NGS, 2.6% of the allele calls remain ambiguous at the allele level (5/194), and 0% of the alleles were ambiguous at the protein level, demonstrating ability of the NGS to resolve the ambiguity from the legacy methods by sequencing the entire gene in addition to being able to phase heterozygous positions both within and between exons. Four samples present us with five allele ambiguities in the validation set. Three allele ambiguities in three samples are caused by not being able to determine whether the *MICA*008:01* or the *MICA*008:04* allele is present, where the only exonic difference is found in exon 1. Using Illumina data, the heterozygous position in exon 1 is not phased to exon 2. The second allele present in these samples is one of the following: *MICA*008:02*¹ (sample 17-4303), *MICA*009:01* (sample 13-2530) and *MICA*049* (sample 17-3959), and none of these alleles are characterized in exon 1. Therefore, both 008:01 and 008:04 are possible given the unknown sequence of the second allele for the exon 1 polymorphic position. The fourth sample presents us with two ambiguous allele calls (*MICA*002:01* or *MICA*002:02* and *MICA*007:01* or *MICA*007:02*). In this case, exon 2 is not phased to exon 4 (a phasing gap of approximately 1100 bases) using Illumina data. This phase break causes a cis/trans ambiguity depending on the arrangement of the heterozygous positions in the two exons, whereby the two genotypes are possible: *MICA*002:01* + *MICA*007:01* or *MICA*002:02* + *MICA*007:02*. While there are other heterozygous positions present in exon 5 that are phased to exon 4, exon 5 is not characterized for both the *MICA*002:02* and *MICA*007:02* allele. This highlights the need for full characterization of genes when entered into the reference databases, as alternative allele combinations may be ruled out with additional sequence context.

Among the 97 validation samples when genotyped with IMGT/HLA version 3.39, only a single sample presented a novel protein sequence that was otherwise unknown. The sample was characterized by SSO to have a *MICA*043* allele in addition to a *MICA*002:020/055/086* allele. Upon sequencing with NGS, it was discovered that the *MICA*043* allele was present, and the second allele did not match any known allele. This second allele, as compared to the *MICA*002:01* allele, was observed to have identical nucleotide sequences in exons 1 to 5, but

exon 6 contained two novel differences. These two differences are found in codons 350, where GAT (encoding aspartic acid) changed to GCT (encoding alanine) in the novel allele, and in codon 360 where GCC (encoding alanine) changed to ACC (encoding threonine) in the novel allele. In addition, this novel allele is characterized by intronic sequences that are different than any other *MICA*002:01:XX* allele. The intronic sequences of this new allele are most similar to *MICA*002:01:03*, and when the two are compared they differ at 44 positions.

TABLE 2 Frequency of MICA alleles in Validation datasets based on Illumina-derived NGS typing

Allele	South-Chinese Han	US-based
<i>MICA*001</i>		1 (0.98%)
<i>MICA*002:01</i>	19 (20.65%)	14 (13.73%)
<i>MICA*002:01/02</i> ^a		1 (0.98%)
<i>MICA*004</i>		11 (10.78%)
<i>MICA*007:01</i>		5 (4.90%)
<i>MICA*007:01/02</i> ^a		1 (0.98%)
<i>MICA*007:07</i>	1 (1.09%)	
<i>MICA*008:01</i>	5 (5.43%)	13 (12.75%)
<i>MICA*008:02</i>		1 (0.98%)
<i>MICA*008:04</i>	17 (18.48%)	7 (6.86%)
<i>MICA*008:04/008:01</i> ^a	1 (1.09%)	3 (2.94%)
<i>MICA*009:01</i>	1 (1.09%)	8 (7.84%)
<i>MICA*009:02</i>		1 (0.98%)
<i>MICA*010:01</i>	16 (17.39%)	3 (2.94%)
<i>MICA*011</i>		5 (4.90%)
<i>MICA*012:01</i>	5 (5.43%)	3 (2.94%)
<i>MICA*015</i>		2 (1.96%)
<i>MICA*016</i>		3 (2.94%)
<i>MICA*017</i>	1 (1.09%)	2 (1.96%)
<i>MICA*018:01</i>		3 (2.94%)
<i>MICA*019:01</i>	13 (14.13%)	3 (2.94%)
<i>MICA*027</i>	4 (4.35%)	3 (2.94%)
<i>MICA*033</i>	1 (1.09%)	
<i>MICA*041</i>		1 (0.98%)
<i>MICA*043</i>		1 (0.98%)
<i>MICA*045</i>	6 (6.52%)	1 (0.98%)
<i>MICA*049</i>	2 (2.17%)	2 (1.96%)
<i>MICA*068</i>		3 (2.94%)
<i>MICA*NEW</i>		1 (0.98%)

Abbreviation: NGS, next generation sequencing.

^aAfter sequencing these alleles with ONT, the ambiguity has been resolved. See Supplemental Table S1 for the allelic resolution.

The sample size for the two validation populations are approximately equal, 46 individuals from a South-China Han population and 51 individuals from a US-based population. The frequency of the alleles, as determined by Illumina-based NGS, in each of the two validation populations are shown in Table 2. The 46 South-China Han population possess 13 different MICA alleles as described at the allele level (two fields), whereas the 51 individuals from the US-based population include a total of 24 different MICA alleles as described at the allele level (two fields), excluding unresolved ambiguities, for a total of 26 unique alleles at the allele level. Adding the 59 supplemental samples from the US-based population did not increase the allele diversity. All alleles present in the South-China Han population were also observed in the US-based population, with the exception of *MICA*007:07* and *MICA*033*. As of IMGT/HLA database version 3.39, there are a total of 159 MICA alleles, 126 alleles with different exonic nucleotide sequences (2 fields), and representing 92 proteins (one field) and 4 null alleles. Allele frequency data have been collected in 41 published populations available in the Allele Frequency Net Database (AFND; accessed October 28, 2019). A total of 62 allele entries (mix of alleles at first and second fields) have been published among those 41 populations, of which 37 entries have a frequency of at least 1% in any population, constituting 29 unique proteins (different first field), of which 19 are represented in this study. The alleles included in our validation study comprise more than 99% of the allele frequencies reported for US populations¹⁴⁻¹⁶ and Chinese populations¹⁷⁻²¹ in AFND. Overall, the large majority of the high frequency MICA alleles found worldwide are represented within this study and the NGS-based method is capable of detecting and properly genotyping these alleles.

MICA alleles are also classified by the STR that is present within exon 5. The STR is a repeat of the trinucleotide *GCT*, which encodes an alanine residue, leading to either 4, 5, 6, 9, 10, or 11 alanine residues within the transmembrane domain and is named A4, A5, A6, A9, A10, and A11, respectively. There is also a seventh representation, A5.1, whereby the third *GCT* repeat is instead *GGCT*, causing a frameshift ultimately leading to a premature stop codon. Accurately genotyping the STR can be challenging with other methods, particularly Sanger sequencing where the two MICA alleles are sequenced together. A benefit of the NGS method is that reads are haploid in nature, and allows for straightforward characterization of the different STR repeats that are present within each sample. The MICA alleles represented in this dataset contain 5 of the 7 STR alleles, with the exceptions of A10, present only in the *MICA*020* allele that is

specific to Brazil, Turkey and Thailand²²⁻²⁴ in AFND, and A11, which is only present in the *MICA*090* allele first described in the IMGT/HLA database version 3.39 and has not been associated with any specific population. The frequency of the exon 5 STR and the alleles that are associated with each STR is found in Table 3. Of the 5 STR alleles represented in this dataset, all are well represented within the US-based population (11.76%-26.47%), whereas 4 of the 5 alleles are well represented in the South-China Han population (13.04%-36.96%), with the exception being the A6 allele found at 3.26%.

Balanced representation of each allele is important for accurately detecting and genotyping each MICA allele present in a sample. If one allele is preferentially amplified, this will inhibit the ability of the software programs to accurately detect the allele that is less represented in the sample. The allelic balance was examined for all samples that were heterozygous within the exons including both the validation and supplemental datasets, leaving a total of 134 samples. The minor allele percentage was averaged for all heterozygous positions within each sample that were not insertions or deletions between the two alleles (Figure 3; Supplemental Table S1). The median minor allele percentage for all heterozygous samples was 47.2%. All but one sample (133/134) had a minor allele percentage greater than 30%. For the single sample below this 30% threshold, the minor allele was *MICA*004*, represented at 17.43% depth of coverage, and was in combination with *MICA*016*. The *MICA*004* allele was sequenced 21 other times in this dataset with no allelic preference observed (34.4%-57.0% depth of coverage, 50.1% average depth of coverage), with one of those cases also observed in combination with the *MICA*016* allele. In this particular instance, both MICA alleles in the sample were detected and correctly genotyped by both software programs, even with this level of preferential amplification.

3.3 | Sequencing on the ONT MinION platform

As has been previously mentioned, one of the major benefits of the NGS technology is the ability to phase heterozygous positions between the two alleles present within a single sample due to the haploid nature of the reads. The Illumina sequencing system used here was with paired-end sequencing, where a range of different sized DNA fragments broadly ranging between 250 and 750 bases length, are sequenced for each sample. The Illumina method is challenged when attempting to sequence larger fragments, however it is not uncommon within the MICA gene to require the phasing of heterozygous

TABLE 3 Frequency of the exon 5 STR and associated alleles in the validation dataset

STR Repeat	MICA Alleles	South-China Han		US-based	
A4	<i>MICA</i> *001, *007:01, *007:07, *012:01, *018:01, *043, *045	12	(13.04%)	15	(14.71%)
A5	<i>MICA</i> *010:01, *016, *019:01, *027, *033	34	(36.96%)	12	(11.76%)
A5.1	<i>MICA</i> *008:01, *008:02, *008:04	23	(25.00%)	24	(23.53%)
A6	<i>MICA</i> *004, *009:01, *009:02, *011, *049	3	(3.26%)	27	(26.47%)
A9	<i>MICA</i> *002:01, *015, *017, *041, *068, *NEW	20	(21.74%)	24	(23.53%)
A10		0	(0.00%)	0	(0.00%)
A11		0	(0.00%)	0	(0.00%)

Abbreviation: STR, sequence tandem repeat.

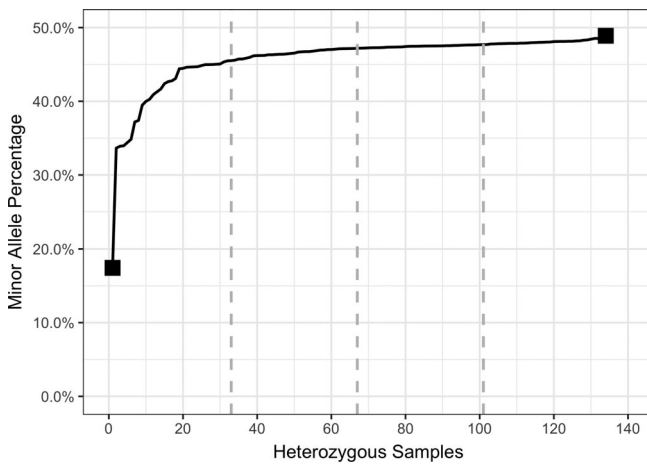


FIGURE 3 Allelic balance for heterozygous samples. The allele balance is shown for all samples with a heterozygous genotype at two fields for MICA, sorted by minor allele percentage ($n = 134$). The first and last data points are shown as squares, with the 25th, 50th, and 75th percentiles highlighted with a dashed line

positions further than 1000 bases apart. Nineteen samples were either homozygous or had only one heterozygous position, and therefore phasing was not required. One more sample had a pattern of heterozygosity that allowed phasing with Illumina data. The heterozygous positions of the remaining samples (136) were not phased. It is to be noted that the 136 heterozygous samples account for differences within both exonic and intronic regions. Considering that in most instances, full phasing of the amplicon is not necessary to unambiguously determine the alleles present at two-field resolution, in our validation dataset, there were only four samples that had an ambiguous typing by Illumina.

During the course of this work, 22 of the 156 samples (14.1%) were chosen to be sequenced on the MinION platform. These samples included loci that were unphased and/or included alleles that were either incompletely characterized or had novel sequence information. This platform is capable of sequencing the entire 13 Kb

MICA amplicon with a single, unfragmented read, allowing for phasing of distant heterozygous positions that remained unphased with the Illumina platform.

Consensus sequences were determined for all 22 samples (44 alleles). Of the 44 alleles that were sequenced with ONT, 32 alleles were determined to have novel sequence information, either in the completion of the unknown regions (exons, introns and UTRs) or were identified to have novel changes as compared to the current reference allele present in the IMGT/HLA database v 3.39 (Table 4). The sequences for these 32 alleles were submitted to IMGT/HLA database. The allele that contained a novel exon variant encoding a new MICA protein, which is similar to *MICA**002:01 with the exception of exon 6, was able to be completely phased, and assigned the name *MICA**110. Upon completing the consensus sequence with ONT for the *MICA**008 allele in sample 17-4303 and submitting the allele for official naming, we were presented with a new *MICA**008 allele in the second field, *MICA**008:13. The *MICA**008:02 allele was updated from IMGT/HLA version 3.39 by another lab to include exon 1 and exon 6 sequence content, of which the allele found in sample 17-4303 now differs when compared to the *MICA**008:02 in IMGT/HLA version 3.40 in exon 1, with a synonymous substitution in codon -17, whereby the *MICA**008:02 allele now contains TTT, and our allele *MICA**008:13 contains TTC, both encoding Phenylalanine. Completion of the exon sequence in addition to adding the intronic and UTR sequences was achieved for eight alleles: *MICA**007:07, *009:01, *011, *033, *041, *043, *049, and *068. Through ONT sequencing, complete allele characterization now includes intronic and UTR sequence for *MICA**045. For five samples that had ambiguous genotyping after Illumina NGS sequencing, four of these samples were sequenced with ONT and the ambiguity was resolved. For both samples 13-2530 and 17-4303, where *MICA**008:01 ambiguous with *MICA**008:04 with Illumina due to lack of phasing exon 1 to exons 2 to 5 and incomplete sequences for the

TABLE 4 Alleles sequenced on ONT MinION platform

Sample	Allele name	Reference allele	Contribution
17-4312	<i>MICA*110</i>	<i>MICA*002:01:01</i>	New MICA protein, all exons and introns characterized and partial UTR sequence
13-3835	<i>MICA*002:01:03</i>	<i>MICA*002:01:03</i>	partial UTR extensions
17-3779	<i>MICA*002:01:10</i>	<i>MICA*002:01:07</i>	Intronic differences and partial UTR extensions
17-6723	<i>MICA*002:01:07</i>	<i>MICA*002:01:07</i>	partial UTR extensions
12-0927	<i>MICA*004:01:07</i>	<i>MICA*004:01:03</i>	Intronic differences and partial UTR extensions
17-4305	<i>MICA*004:01:08</i>	<i>MICA*004:01:03</i>	Intronic differences and partial UTR extensions
17-4858	<i>MICA*004:01:03</i>	<i>MICA*004:01:03</i>	Partial UTR extensions
13-2303	<i>MICA*007:01:05</i>	<i>MICA*007:01:01</i>	Intronic differences and partial UTR extensions
15-7662	<i>MICA*007:01:05</i>	<i>MICA*007:01:01</i>	Same as 13-2303
17-3779	<i>MICA*007:01:06</i>	<i>MICA*007:01:01</i>	Intronic differences and partial UTR extensions
17-6715	<i>MICA*007:07</i>	<i>MICA*007:07</i>	Exon 6, introns and partial UTR
13-2303	NA	<i>MICA*008:01:01</i>	No contribution, matches reference
17-4305	NA	<i>MICA*008:01:02</i>	No contribution, matches reference
15-9488	<i>MICA*008:01:03</i>	<i>MICA*008:01:03</i>	Partial UTR extensions
15-9488	<i>MICA*008:01:03</i>	<i>MICA*008:01:03</i>	Same as 15-9488 (homozygous sample)
17-6717	<i>MICA*008:01:09</i>	<i>MICA*008:01:02</i>	Intronic and 3' UTR differences
17-6717	<i>MICA*008:01:09</i>	<i>MICA*008:01:02</i>	Same as 17-6717 (homozygous sample)
17-4306	<i>MICA*008:01:04</i>	<i>MICA*008:01:04</i>	Partial UTR extensions
17-4310	<i>MICA*008:01:08</i>	<i>MICA*008:01:02</i>	Intronic differences
17-6723	<i>MICA*008:01:09</i>	<i>MICA*008:01:02</i>	Same as 17-6717
17-4303	<i>MICA*008:13</i>	<i>MICA*008:02</i>	Exonic synonymous SNP in exon 1, intronic differences and partial UTR extensions
13-2530	<i>MICA*008:04:02</i>	<i>MICA*008:04:02</i>	Same as 17-6716
17-4303	<i>MICA*008:04:05</i>	<i>MICA*008:04:02</i>	Single intronic difference and partial UTR extensions
17-4310	NA	<i>MICA*008:04:01</i>	No contribution, matches reference
17-6716	<i>MICA*008:04:02</i>	<i>MICA*008:04:02</i>	Partial UTR extensions
13-2530	<i>MICA*009:01:01</i>	<i>MICA*009:01</i>	Exon 1, introns and partial UTR
17-6729	<i>MICA*009:01:06</i>	<i>MICA*009:01</i>	Exon 1, introns and partial UTR
12-0565	<i>MICA*009:02:01</i>	<i>MICA*009:02:01</i>	Partial UTR extension
17-6720	<i>MICA*010:01:05</i>	<i>MICA*010:01:05</i>	Partial UTR extensions
17-6722	<i>MICA*010:01:05</i>	<i>MICA*010:01:05</i>	Same as 17-6720
17-6735	<i>MICA*010:01:05</i>	<i>MICA*010:01:05</i>	Same as 17-6720
13-3835	<i>MICA*011:01:01</i>	<i>MICA*011</i>	Exon 1, introns and partial UTR
17-4306	<i>MICA*011:01:01</i>	<i>MICA*011</i>	Same as 13-3835
17-6720	<i>MICA*012:01:03</i>	<i>MICA*012:01:02</i>	Intronic difference and partial UTR extensions
17-6735	<i>MICA*012:01:02</i>	<i>MICA*012:01:02</i>	Partial UTR extensions
15-7662	<i>MICA*018:01:04</i>	<i>MICA*018:01</i>	Intronic differences
17-6729	<i>MICA*019:01:01</i>	<i>MICA*019:01:01</i>	Partial UTR extension
17-6722	<i>MICA*033</i>	<i>MICA*033</i>	Exons 1 and 6, introns and partial UTR
12-0565	<i>MICA*041</i>	<i>MICA*041</i>	Exons 1 and 6, introns and partial UTR
17-4312	<i>MICA*043</i>	<i>MICA*043</i>	Exons 1 and 6, introns and partial UTR
17-6715	<i>MICA*045:01:01</i>	<i>MICA*045</i>	Introns and partial UTR

(Continues)

TABLE 4 (Continued)

Sample	Allele name	Reference allele	Contribution
17-6716	<i>MICA*045:01:01</i>	<i>MICA*045</i>	Same as 17-6715
12-0927	<i>MICA*049:01:01</i>	<i>MICA*049</i>	Exon 1, introns and partial UTR
17-4858	<i>MICA*068:01:01</i>	<i>MICA*068</i>	Exon 1, introns and partial UTR

Abbreviations: MICA, major histocompatibility complex class I chain-related gene A; SNP, single nucleotide polymorphism; UTR, untranslated region.

other allele, in both cases, the *MICA*008:04* allele was present after sequencing with ONT. For samples 15-7662 and 17-3779, whereby each sample had a phase ambiguity between exons 2 and 4, ONT resolved the phasing to reveal the presence of *MICA*007:01* + *MICA*018:01* and *MICA*002:01* + *MICA*007:01*, respectively. We did not have any remaining material left to complete the ONT typing for the fifth sample, 17-3959, to resolve the *MICA*008:01/MICA*008:04* ambiguity. However, the second allele in this sample, *MICA*049*, which was previously not characterized in exon 1, was completed with another sample sequenced by ONT (12-0927), and with this additional sequence for exon 1 we are able to deduce that the ambiguity would be resolved and sample 17-3959 would now be typed as *MICA*008:04* + *MICA*049*.

Among the 22 samples sequenced with ONT, we have found that many of these samples contain new intronic variants as compared to either the known reference allele in IMGT or between alleles that were genotyped the same at 2-fields that were sequenced multiple times with ONT. For example, the *MICA*008:01* allele was sequenced a total of nine times with ONT: one completely matched *MICA*008:01:01*, one was a complete match for *MICA*008:01:02*, two sequences matched *MICA*008:01:03*, one sequence matched *MICA*008:01:04*, two sequences had new intronic variants (*MICA*008:01:08* and *MICA*008:01:09*), and the final two sequences both matched the newly described 008:01:09 allele. In the case of sequences that matched *MICA*008:01:03* and *MICA*008:01:04*, this study sequenced more of the 5' and 3' UTR sequences and have been submitted to IMGT to further characterize each of these alleles. The *MICA*007:01* allele was sequenced a total of three times with ONT. Two of the three sequences contained different intronic novelty as compared to the *MICA*007:01:01* allele in the IMGT/HLA Database version 3.39, now named *MICA*007:01:05* and *MICA*007:01:06*, and the third allele was identical the newly described *MICA*007:01:05* allele. This pattern of discovering intronic novelty continues for other MICA alleles that have typed the same at two fields (Table 4). The new sequences identified in this

limited dataset of 22 samples indicate that there may be substantial undiscovered intronic diversity within the locus. Additionally, since we did not interrogate the intronic sequences of the other 134 samples not sequenced with ONT further intronic novelty may exist that was not examined or reported.

An additional benefit of long-reads from the ONT platform is the ability to look at the specificity of the MICA amplification, and determine whether there was any coamplification of any of the highly similar MICB gene. We took all ONT reads and aligned then to both the MICA and MICB alleles in the IMGT/HLA database. We then examined the percent of reads that had at least 1000, 5000, or 10 000 bases aligned to any allele of either gene and compared the mapping rates. Overall, we find very little MICB contamination in the reads from the MICA amplification. Specifically, we find virtually no reads mapping to MICB at a length of 10 000 bases or more (average = 0.02%; 0.00%-0.13%). The rate minimally increases as the length required for alignment decreases. For an alignment length of at least 5000 bases, we find on average 0.11% of reads map to MICB (0.02%-0.25%), and for an alignment length of at least 1000 bases we find on average 0.31% of reads map to MICB (0.08%-1.25%). In the case requiring at least 1000 bases to map to MICB, we cannot exclude that this may be due to the strong homology between the MICA and MICB loci. Taken together, these results demonstrate that the amplification of the MICA gene is highly specific and that the MICA primers do not amplify any genomic elements of the MICB gene, in part or in full.

4 | DISCUSSION

NGS technology, thus far, has been established as the method of choice for the characterization of HLA polymorphisms. However, while the classical HLA genes with clinical relevance have been extensively studied and characterized, other HLA-like genes, that is, MICA gene, with less clinical utility, despite their potential interest have attracted less attention. NGS genotyping of MICA is available either by purchasing reagents for NGS from

GenDx (whole gene amplification) or as a service through DKMS (individual exons). However, no extensive studies regarding MICA characterization have been undertaken as there are no published reports in the literature to demonstrate their performance. This study is an effort to develop a comprehensive way to characterize the MICA gene, as this gene has established clinical utility in solid organ transplantation when non-HLA antibodies are suspected. Improving the characterization of the MICA gene may eventually optimize our means of characterizing anti-MICA antibodies and more accurately determine compatibility regarding MICA polymorphisms.

Our approach included 97 previously characterized samples for the MICA gene, utilizing legacy methods, and formed the basis of the validation study, supplemented by an additional 59 samples, that were not previously characterized for their MICA gene, to assess further quality metrics of the assay and attempt to expand the number of MICA alleles included in the study. The alleles included in this study represent greater than 99% of the alleles frequently found in the US- and Chinese-based populations, and are a good representation of the frequent alleles found in the world-wide populations. Of the first-field alleles with frequencies in AFND greater than 1% in any population worldwide, 10 were not characterized by this study: *MICA*005*, *MICA*006*, *MICA*013*, *MICA*020*, *MICA*023*, *MICA*026*, *MICA*030*, *MICA*046*, *MICA*048*, and *MICA*052*. Only three alleles were present at a frequency greater than 2%: *MICA*006* at 2.5% from an English population and 2.3% from a Turkish population, *MICA*048* at 11.8% from a Japanese population and *MICA*052* at 8.2% from the Thailand North East population. The reported allele frequencies beyond the first field are highly limited in the Allele Frequencies Net Database, and exist for only 10 alleles: *MICA*002:01*, *MICA*002:02*, *MICA*007:01*, *MICA*007:02*, *MICA*008:01*, *MICA*008:02*, *MICA*008:04*, *MICA*009:01*, *MICA*009:02*, and *MICA*012:01*. Eight of these alleles are represented in this study, with the *MICA*002:02* and *MICA*007:02* being excluded, but also not highly represented in any population (maximum frequency of 0.3% and 0.4% for each allele, respectively). Given that there are 92 unique proteins and a total of 126 alleles in the IMGT/HLA database (v.3.39), but only 39 alleles found in AFND with >1% frequency, it suggests that either the majority of the IMGT/HLA reported MICA alleles are rather infrequent or the community needs better characterization methods to distinguish highly similar alleles that currently exist as a single specificity in AFND as these specificities have been characterized many years ago. Overall, the large majority of the high frequency MICA alleles found worldwide are represented within this study and the full-length amplification

method is capable of detecting and properly genotyping all these alleles.

The selected PCR conditions, including primers, resulted in the successful amplification of MICA and its genotyping of all samples. Not unexpectedly a few samples with compromised quality of DNA needed to repeat the PCR before MICA was successfully amplified. Considering the length of the amplicon (close to 13 Kb) this is not surprising. Furthermore the PCR conditions secured the balanced amplification of the alleles present in virtually every sample (median minor allele percentage for all heterozygous samples was 47.2%, except one with an allele percentage of 17.4%). Uniformity of coverage was generally very good with some variation when amplifications of the MICA gene were not robust. Additionally, at specified positions coverage was 25% to 50% lower than the average, possibly due to lower complexity sequences, high GC content and/or high sequence homology with other HLA genes that result in removal of some of these reads. The strong performance of the assay overall resulted in excellent performance (100%) of all metrics including sensitivity, specificity, precision, and accuracy, when compared to the results of the legacy methods.

Having good representation of both alleles in a sample strongly improves the ability of the software programs to accurately genotype the sample, for both the MICA and HLA genes. Among the alleles tested in this study, only a single sample showed a slight allelic imbalance with the minor allele present at 17% depth of coverage; it was accurately detected nevertheless, by both programs. Through our experience, each of the two software programs used for genotyping in this study generally do not have a problem detecting alleles that are present at 20% depth of coverage. However, the accuracy for detecting alleles that are imbalanced with the minor allele at or below 10% is challenging, but is greatly improved when using our approach of employing two analysis software programs in combination.

Despite the key feature of NGS to sequence a single DNA molecule, enabling the ability to set phase at distant polymorphic positions and thereby eliminate ambiguities, the characterization of the long 13 kb amplicon included several cases whereby the Illumina platform with its known limitation to sequence fragments not exceeding 1Kb, resulted in a rather limited number of ambiguities (2.6% of allele calls), even though a large majority of these samples were not fully phased (136 out of 137 loci with multiple heterozygous positions). It is certain that the number of reported alleles in all HLA classical and non-classical loci will continue to increase, including the MICA locus, and therefore we expect that an increasing number of ambiguities will ensue if we do not manage to introduce appropriate methodologies that will sequence

long DNA fragments, set phase, and therefore eliminate ambiguities. Of course, the ability to phase a sample is highly dependent upon the allele combinations present, and any large regions with few heterozygous positions are problematic for the Illumina platform's short reads. Irrespective of the number of heterozygous positions within a sample, where some samples had as few as six heterozygous positions (*MICA*008:01* + *MICA*008:04*) and others with as many as hundreds, phasing is not possible due to the spacing of those heterozygous positions across the 13 kb amplicon. Nevertheless even unphased sequences provide unambiguous typing. However, it should be noted that unambiguous allele calls that derive from unphased sequences with the current IMGT/HLA database may in the future become ambiguous with the addition of new alleles into future releases of the database. In an effort to develop a system that takes advantage of the positive features of the different sequencing technologies, that is, Illumina's accuracy of base calling (0.297% error)²⁵ and ONT's sequencing of long fragments, we first sequenced the MICA amplicons on the Illumina platform and then selected the cases that presented existing ambiguities, lack of complete sequence, or new alleles to be characterized on the ONT platform.

Before submitting each ONT-derived consensus sequence for naming, each sequence was manually compared to the consensus derived from the Illumina reads. In these cases, we find that there are three main types of discrepancies that existed between the two consensus sequences: imbalanced positions within the ONT reads, correct determination of insertions and length of homopolymer sequences. With regard to imbalance in the ONT reads, there are some positions with a high level of noise (10%-20% wrong base calling) that the Illumina data is unable to correct. The true heterozygous positions are represented in the corrected ONT data at close to 50% depth of coverage, whereas the noise is significantly lower at 10% to 20% depth of coverage. The noise is manually corrected using the higher fidelity Illumina only data. Regarding insertions, both sequencing platforms detect them accurately. All insertions were manually checked for correct placement before submission.

Besides the overall known high error rate of the ONT platform (10%-13%, 11.78% in this study),^{26,27} accurate sequencing of homopolymers remain a major challenge for the ONT platform, especially as the length of the homopolymer increases. Two particular regions within the 3' UTR of the MICA gene with a long homopolymer of As or Ts were the biggest challenge in the generation of the consensus sequence. The number of repeated deoxynucleotides was between 9 and 15 for Ts and 15 to 25 for As. It is technically challenging to determine the length of each homopolymer with certainty due to

technical limitations of both the Illumina and ONT platforms. The longer the homopolymer the more likely it is to observe variation in the length of the homopolymer within a single sample and a single allele, a situation that is further complicated with a heterozygous sample. While both the polyA and polyT sites in the 3' UTR showed the same trend, it was only the polyA site that often has a length greater than 20 which proved exceedingly difficult to accurately determine the correct number of bases. In our submitted sequences to GenBank and IMGT we refrained from including the polyA in the submitted sequences and therefore the last 155 nucleotides of the reported MICA 3' UTR, not accounting for the A homopolymer, have not been included. It should also be mentioned that a smaller homopolymer of As in intron 1 did not pose the same challenge and we convincingly identified the correct length of this homopolymer as it was significantly smaller (10-12 nucleotides). This validation study has identified additional polymorphisms within the MICA gene and considering that our method amplifies the whole MICA gene, provides a comprehensive means for the complete characterization of this gene.

This work resulted in 32 unique sequences submitted to the IMGT/HLA database. A single allele had a novel protein sequence (*MICA*110*), a second allele had novel synonymous substitution in exon 1 (*MICA*008:13*), while 10 alleles included new polymorphisms within the intronic sequences. Additionally, 20 already existing alleles with incomplete exonic, intronic and/or UTR sequences were further characterized.

This study presents a robust protocol for the accurate and thorough characterization of the MICA gene; it contributed additional MICA sequences to the data basis of IMGT and enables a more accurate determination of MICA genomic sequences, and therefore indirectly anti-MICA antibodies, facilitating the interpretation of data analysis in a transplant setting.

ACKNOWLEDGMENTS

The authors would like to acknowledge the histocompatibility specialists at the Children's Hospital of Philadelphia for their work to obtain the MICA genotyping by SSO. The authors would also like to acknowledge grant support from the National Natural Science Foundation of China (grant numbers 81571562 and 81873875) and from the Hunan Science and Technology Project Foundation (2018JJ2549) to Y. Z. and institutional support from the Children's Hospital of Philadelphia to D. M.

CONFLICT OF INTEREST

D. M. is a consultant to, and owns options in Omixon. D. M., D. F., and J. L. D. receive royalties from Omixon. The other authors of this manuscript have no conflicts of interest to disclose.

AUTHOR CONTRIBUTIONS

Yizhou Zou collected samples and determined MICA genotypes by SBT from the South-Chinese Han population, participated in the Illumina sequencing and contributed to the writing and review of the manuscript. Jamie L. Duke participated in the experimental design of the study, data and statistical analysis of the Illumina and ONT sequencing data, interpretation of the results and contributed to the writing of the manuscript. Deborah Ferriola participated in the experimental design of the study, designed the primers and procedure for the amplification and sequencing of the full-length MICA gene, analysis of the Illumina and ONT sequencing data, and contributed to the writing of the manuscript. Jenna Wasserman conducted the Illumina sequencing experiments and reviewed the manuscript. Timothy L. Mosbrugger participated in the experimental design of the study, data analysis of the Illumina and ONT sequencing data, generation of the consensus sequences, and reviewed the manuscript. Qizhi Luo, Weiguang Luo, Liang Cai, and Kevin Zou all collected samples and determined MICA genotypes by SBT from the South-Chinese Han population and reviewed the manuscript. Nikolaos Tairis conducted the Illumina sequencing experiments and reviewed the manuscript. Georgios Damianos and Ioanna Pagkrati analyzed the ONT sequencing data and consensus sequences, submitted new alleles to Genbank and IMGT/HLA databases, and reviewed the manuscript. Debra Kukuruga collected samples and determined MICA genotypes by SSOP of unknown ethnicity from the US-based population and reviewed the manuscript. Yanping Huang participated in the experimental design of the study, analysis of the Illumina sequencing data and reviewed the manuscript. Dimitri S. Monos conceived and directed the study and contributed to the writing of the manuscript.

DATA AVAILABILITY STATEMENT

Data is available upon request from the authors.

ORCID

Yizhou Zou  <https://orcid.org/0000-0002-9261-1161>

Jamie L. Duke  <https://orcid.org/0000-0002-4768-0846>

ENDNOTE

¹ As part of this study, the new sequence of this allele was submitted to IMGT/HLA, and has now been named *MICA*008:13*.

REFERENCES

- Bahram S, Bresnahan M, Geraghty DE, Spies T. A second lineage of mammalian major histocompatibility complex class I genes. *Proc Natl Acad Sci U S A*. 1994;91(14):6259-6263.
- Leelayuwat C, Townend DC, Degli-Esposti MA, Abraham LJ, Dawkins RL. A new polymorphic and multicopy MHC gene family related to nonmammalian class I. *Immunogenetics*. 1994;40(5):339-351.
- Bauer S, Groh V, Wu J, et al. Activation of NK cells and T cells by NKG2D, a receptor for stress-inducible MICA. *Science*. 1999; 285(5428):727-729.
- Baranwal AK, Mehra NK. Major histocompatibility complex class I chain-related a (MICA) molecules: relevance in solid organ transplantation. *Front Immunol*. 2017;8:182.
- Choy M-K, Phipps ME. MICA polymorphism: biology and importance in immunity and disease. *Trends Mol Med*. 2010;16(3):97-106.
- Carapito R, Jung N, Kwemou M, et al. Matching for the non-conventional MHC-I MICA gene significantly reduces the incidence of acute and chronic GVHD. *Blood*. 2016;128(15):1979-1986.
- Askar M, Sobecks R, Wang T, et al. MHC class I chain-related gene A (MICA) donor-recipient mismatches and MICA-129 polymorphism in unrelated donor hematopoietic cell transplantations has no impact on outcomes in acute lymphoblastic leukemia, acute myeloid leukemia, or myelodysplastic syndrome: a Center for International Blood and Marrow Transplant Research Study. *Biol Blood Marrow Transplant*. 2017;23(3):436-444.
- Suárez-Alvarez B, López-Vázquez A, Baltar JM, Ortega F, López-Larrea C. Potential role of NKG2D and its ligands in organ transplantation: new target for immunointervention. *Am J Transplant*. 2009;9(2):251-257.
- Zou Y, Stastny P, Süsal C, Döhler B, Opelz G. Antibodies against MICA antigens and kidney-transplant rejection. *N Engl J Med*. 2007;357(13):1293-1300.
- Zou Y, Han M, Wang Z, Stastny P. MICA allele-level typing by sequence-based typing with computerized assignment of polymorphic sites and short tandem repeats within the transmembrane region. *Hum Immunol*. 2006;67(3):145-151.
- Zou Y, Stastny P. High resolution MICA genotyping by sequence-based typing (SBT). *Methods Mol Biol*. 2012;882: 183-195.
- Luo QZ, Lin L, Gong Z, et al. Positive association of major histocompatibility complex class I chain-related gene A polymorphism with leukemia susceptibility in the people of Han nationality of southern China. *Tissue Antigens*. 2011;78(3): 178-184.
- Duke JL, Mosbrugger TL, Ferriola D, et al. Resolving MiSeq-generated ambiguities in HLA-DPB1 typing by using the Oxford Nanopore Technology. *J Mol Diagn*. 2019;21(5):852-861.
- Petersdorf EW, Shuler KB, Longton GM, Spies T, Hansen JA. Population study of allelic diversity in the human MHC class I-related MIC-A gene. *Immunogenetics*. 1999;49(7-8):605-612.
- Zhang Y, Lazaro AM, Lavingia B, Stastny P. Typing for all known MICA alleles by group-specific PCR and SSOP. *Hum Immunol*. 2001;62(6):620-631.
- Zhang Y, Han M, Vorhaben R, Giang C, Lavingia B, Stastny P. Study of MICA alleles in 201 African Americans by multiplexed single nucleotide extension (MSNE) typing. *Hum Immunol*. 2003;64(1):130-136.
- Gong W, Fan L, Yang J, Xu L, Yao F. Analysis on polymorphism in exons 2,3 and 4 of the MICA gene in three different

- Chinese populations. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi*. 2002;19(4):336-339.
18. Zhu F, Zhao H, He Y, et al. Distribution of MICA diversity in the Chinese Han population by polymerase chain reaction sequence-based typing for exons 2-6. *Tissue Antigens*. 2009;73(4):358-363.
 19. Tian W, Cai JH, Wang F, Li LX. MICA polymorphism in a northern Chinese Han population: the identification of a new MICA allele, MICA*059. *Hum Immunol*. 2010;71(4):423-427.
 20. Chen E, Lin L, Chen CJ, Zhang XY, Luo QZ, Yu P. MIC gene polymorphism and haplotype diversity in Zhuang nationality of southern China. *Hum Immunol*. 2014;75(9):953-959.
 21. Wang YJ, Zhang NJ, Chen E, Chen CJ, Bu YH, Yu P. MICA/B genotyping of Tujias from Zhangjiajie, Hunan Province, China. *Hum Immunol*. 2016;77(4):340-341.
 22. Romphruk AV, Naruse TK, Romphruk A, et al. Diversity of MICA (PERB11.1) and HLA haplotypes in northeastern Thais. *Tissue Antigens*. 2001;58(2):83-89.
 23. Mizuki N, Meguro A, Tohnai I, Gül A, Ohno S, Mizuki N. Association of major histocompatibility complex class I chain-related gene A and HLA-B alleles with Behçet's disease in Turkey. *Jpn J Ophthalmol*. 2007;51(6):431-436.
 24. Oliveira LA, Ribas F, Bicalho MG, Tsuneto LT, Petzl-Erler ML. High frequencies of alleles MICA*020 and MICA*027 in Amerindians and evidence of positive selection on exon 3. *Genes Immun*. 2008;9(8):697-705.
 25. Duke JL, Lind C, Mackiewicz K, et al. Towards allele-level human leucocyte antigens genotyping—assessing two next-generation sequencing platforms: ion torrent personal genome machine and Illumina MiSeq. *Int J Immunogenet*. 2015;42(5):346-358.
 26. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol*. 2018;19(1):90.
 27. Tyler AD, Mataseje L, Urfano CJ, et al. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Sci Rep*. 2018;8(1):1-12.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Zou Y, Duke JL, Ferriola D, et al. Genomic characterization of MICA gene using multiple next generation sequencing platforms: A validation study. *HLA*. 2020;96:430–444. <https://doi.org/10.1111/tan.13998>