Received: 21 September 2021 | Revised: 8 January 2022 | Accepted: 21 January 2022

DOI: 10.1111/tan.14562

ORIGINAL ARTICLE

HLA Immune Response Genetics WILEY

Personalized HLA typing leads to the discovery of novel HLA alleles and tumor-specific HLA variants

Irantzu Anzar[1] | Angelina Sverchkova[1] | Pubudu Samarakoon[1] | Espen Basmo Ellingsen[2] | Gustav Gaudernack[2] | Richard Stratford[1] | Trevor Clancy[1]

[1]NEC OncoImmunity AS, Oslo Cancer Cluster, Oslo, Norway

[2]Ultimovacs ASA, Oslo Cancer Cluster, Oslo, Norway

Correspondence
Trevor Clancy, NEC OncoImmunity AS, Oslo Cancer Cluster, Ullernchausseen 64/66, 0379 Oslo, Norway.
Email: trevor@oncoimmunity.com

Funding information
Norges Forskningsråd

Accurate and full-length typing of the HLA region is important in many clinical and research settings. With the advent of next generation sequencing (NGS), several HLA typing algorithms have been developed, including many that are applicable to whole exome sequencing (WES). However, most of these solutions operate by providing the closest-matched HLA allele among the known alleles in IPD-IMGT/HLA Database. These database-matching approaches have demonstrated very high performance when typing well characterized HLA alleles. However, as they rely on the completeness of the HLA database, they are not optimal for detecting novel or less well characterized alleles. Furthermore, the database-matching approaches are also not adequate in the context of cancer, where a comprehensive characterization of somatic HLA variation and expression patterns of a tumor's HLA locus may guide therapy and clinical outcome, because of the pivotal role HLA alleles play in tumor antigen recognition and immune escape. Here, we describe a personalized HLA typing approach applied to WES data that leverages the strengths of database-matching approaches while simultaneously allowing for the discovery of novel HLA alleles and tumor-specific HLA variants, through the systematic integration of germline and somatic variant calling. We applied this approach on WES from 10 metastatic melanoma patients and validated the HLA typing results using HLA targeted NGS sequencing from patients where at least one HLA germline candidate was detected on Class I HLA. Targeted NGS sequencing confirmed 100% performance for the 1st and 2nd fields. In total, five out of the six detected HLA germline variants were because of Class I ambiguities at the third or fourth fields, and their detection recovered the correct HLA allele genotype. The sixth germline variant let to the formal discovery of a novel Class I allele. Finally, we demonstrated a substantially improved somatic variant detection accuracy in HLA alleles with a 91% of success rate in simulated experiments. The approach described here may allow the field to genotype more accurately using WES data, leading to the discovery of novel HLA alleles and help characterize the relationship between somatic variation in the HLA region and immunosurveillance.

KEYWORDS

HLA typing, HLA variant calling, novel HLA allele discovery, tumor-specific HLA variants, whole exome sequencing (WES)

1 | INTRODUCTION

Full-length typing of the HLA (also known as high-resolution HLA genotyping) is a continuous challenge, as it is one of the most complex and polymorphic regions in the human genome.[1–3] The classical Class I HLA proteins bind in a complex with peptides that may be presented on the cell surface. Once presented at the cell surface, these complexes may then be recognized by effector T cells of the adaptive immune system. Class I HLA proteins present peptides on the surface of all human cells, and consequently the identification of the precise HLA genotype has implications in organ transplantation, with crucial implications in hematopoietic stem cell transplantation, where donors and recipients need to be HLA matched.[4–6] Precise knowledge of HLA genotypes of individuals is also important in disease association studies, where HLA allelic variants have strong genetic associations to many common human diseases.[7] Additionally, variations in HLA alleles have been frequently linked to disease susceptibility in many studies,[4,8] in addition to drug sensitivity[9] and susceptibility to adverse drug responses.[10] In cancer, it has been demonstrated that specific HLA genotypes,[11] and in particular diversity in the HLA genotype of a patient, can predict response to immune checkpoint inhibitors (ICIs).[11–13]

The clinical importance of precise HLA typing is very well established, and next generation sequencing (NGS) data has recently been adopted by many diagnostic laboratories as the preferred data source to perform reliable HLA typing.[14,15] The main outcome of HLA typing is the assignment of a unique HLA name, referred as an HLA allele, that constitutes up to four fields of resolution separated by colons, (e.g., *HLA-A*02:01:01:01*). The four fields of this HLA nomenclature represent: (1) allelic group, (2) protein group, (3) synonymous DNA changes within the protein coding regions, and (4) variants in non-coding regions.

NGS-based HLA typing methods can currently be divided into two categories related to the type of input data: HLA-targeted sequencing (e.g., PCR-based target amplification) with high sequence depth, and standard NGS (e.g., whole exome sequencing [WES], whole genome sequencing [WGS], and RNA-sequencing [RNA-Seq]) with moderate sequence depth. Targeted HLA sequencing is the most information-dense type of NGS, and therefore most often used to discover novel HLA alleles and resolve HLA typing ambiguities in the exons that encode the peptide binding cleft (typed with all other protein coding exons at the first and second fields of resolution). The emphasis on the binding cleft exons is because of the critical importance of determining the donor HLA-peptide complex presentation for tissue compatibility in transplantations. Although targeted HLA sequencing may be subject to PCR amplification errors[16] in a small fraction of samples,[17] it is arguably considered the gold standard for HLA typing in clinical applications.[14] Targeted HLA sequencing, to date, has mostly been used to resolve HLA genotypes and ambiguities in the peptide binding cleft exons. Consequently, most of the described HLA alleles have incomplete sequences with enriched coverage for the binding cleft exons and only a minority of the alleles come with complete and full-length HLA sequences.[18,19] This lack of full-length HLA sequences is not optimal, as identification of the complete HLA sequence has important clinical and research applications. Full-length HLA sequence typing is useful, for example, to generate ancestry-based analyses[20] and has been shown to be critically important for identifying causal variants in HLA-based disease association studies.[21] The importance of full-length HLA sequence typing has also been shown to help optimize donor selection, improve clinical outcome, and result in fewer transplant complications, as clearly demonstrated in hematopoietic cell transplantations (HCT).[22–26] Furthermore, full-length HLA typing may provide novel insights into the transcript expression regulation of HLA genes, including epigenetic mechanisms leading to improved understanding of complex immune diseases.[27,28]

WES captures most exons in the coding regions of the HLA region, not only the binding cleft exons. Although deep intronic variants are difficult to sequence with WES, most exome capture kits extend beyond the defined exon boundaries and also sequence some intronic regions at lower depth, leaving the possibility of complete or full-length HLA typing and characterization of many intronic HLA variants. Compared with WES; WGS[29] and targeted NGS sequencing is considered to be more laborious,[14] or particularly expensive in the case of WGS.[29] As an estimated 85% of Mendelian inherited disease causing mutations and many disease associated single nucleotide polymorphisms (SNPs) are located in the exome,[30] WES has therefore become a very popular alternative in many clinical and research settings.[29] Given the widespread accessibility, low cost, high speed, interpretability, and broad abundance of WES data; there is arguably a necessity to develop methods that allow full-length HLA typing and novel allele discovery where sequencing has been performed using WES only.

This advent of widely available NGS data has resulted in an increased number of computational NGS-based HLA typing solutions, many of which can be applied to WES.[16,18,19,31–33] However, the majority of these tools perform HLA typing by identifying the closest-matched HLA allele through sequence alignment of WES NGS reads against the reference sequences in the IPD-IMGT/HLA Database.[1] Unfortunately, because of the current incomplete nature of the HLA databases (an average of 10% Class I HLA alleles have their full-length sequences available in old versions of IPD-IMGT/HLA Database and 55% in the latest to date 3.41.2[19]); previous WES database-matching based solutions do not reliably perform full-length HLA typing. Importantly, although some of these computational methods[19] do provide the functionality to output full-length HLA genotypes using WES data; until quite recently[34–36] there was credible benchmarking data available only for the protein coding sequence. De novo assembly methods do not have the limitation of directly relying on the IPD-IMGT/HLA Database, and therefore have the potential of identifying novel HLA alleles.[33,37] However, these tools are computationally expensive, and their accuracy is dependent on deep coverage of the HLA region and the necessity of using long reads for correct phasing.

In cancer, typing of the HLA locus has critical importance, as it facilitates recognition and subsequent killing of tumor cells by the adaptive immune system. Many of the cancer immunotherapies developed in the recent years are HLA-dependent immunotherapies.[38] Although these cancer immunotherapies improved clinical outcome greatly, only a fraction of patients currently respond to treatment.[39,40] This may partially be because of the lack of computational tools allowing a comprehensive profiling of the HLA status in a tumor.[41–43] Currently a limited number of methods exist that allow the detection of somatic mutations in HLA.[44,45] Unfortunately, these current strategies are unable to discover novel alleles, have a limited variant calling approach, do not assess the expression of HLA in a tumor, and finally, use an outdated version of the IPD-IMGT/HLA Database.

Here, we describe an HLA typing solution that attempts to rectify all these shortcomings in somatic HLA profiling in cancer. The approach is also based on alignments to known HLA sequences at the IPD-IMGT/HLA Database, but simultaneously enables the discovery of novel germline and somatic HLA alleles by leveraging the systematic integration of variant calling. We applied this personalized HLA typing method to WES data and demonstrated its ability to identify novel HLA alleles and rectify HLA ambiguities, particularly at the third and fourth fields of resolution. We validated the approach using targeted HLA sequencing from the normal blood of 10 metastatic melanoma patients and confirmed the prediction of a novel Class I HLA allele not yet characterized in the reference library. Furthermore, as somatic HLA mutations in cancer have an association with tumor-immune escape[44,46–48]; with the personalized HLA genotypes in hand, we demonstrated the performance and utility of our approach to identify tumor-specific variants in HLA in the metastatic melanoma tumors.

2 | MATERIALS AND METHODS

2.1 | HLA database closest-matched typing from WES data

An overview of the "NeoOncoHLA" workflow for personalized HLA typing and tumor-specific HLA variant calling is illustrated in Figure 1. The first step involved the assignment of the closest-matched HLA allele from the IPD-IMGT/HLA Database (see step 1 in Figure 1), which then further served as a reference for the personalized HLA variant detection in subsequent steps. The enormous complexity of the HLA region makes conventional mapping approaches to the reference genome result in inaccurate HLA typing. This complexity is resolved by aligning NGS reads either to an HLA reference sequence library or by applying *de novo* assembly methods. The method outlined in Figure 1 is based on the former and performed HLA typing by aligning reads to the IPD-IMGT/HLA Database of known HLA sequences,[2] using a previously published HLA database closest-matched approach, "OncoHLA."[19] Once WES reads were aligned to all known HLA alleles, the HLA allele was then determined by the using an integer linear programing (ILP) algorithm which uses prior probabilities of the allelic ethnic frequencies.[19] The output includes the closest-matched HLA allele from the HLA database and the associated HLA sequence, up to four fields of resolution for each allele (see step 1 in Figure 1).

2.2 | Integration of germline variant calling to achieve personalized HLA typing from WES data

NeoOncoHLA, which incorporated variant calling, then processed the reads from the WES data and aligned them

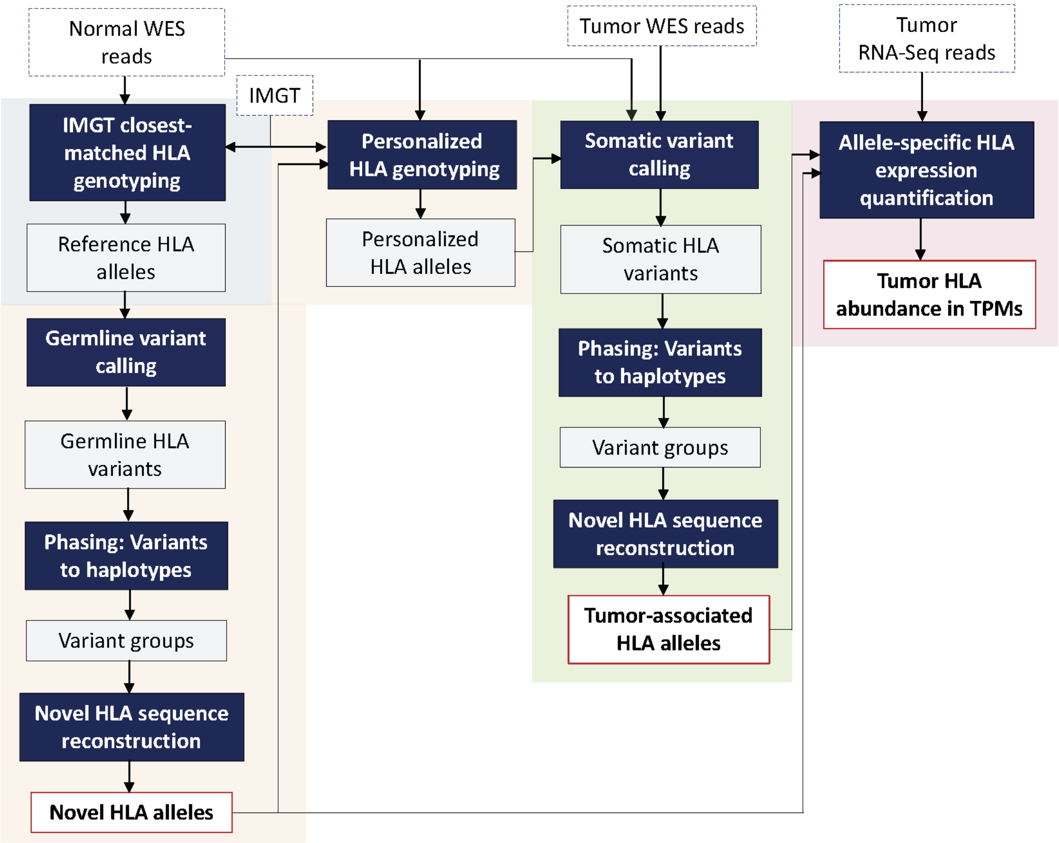
Normal WES reads
Tumor WES reads
Tumor RNA-Seq reads
IMGT
IMGT closest-matched HLA genotyping
Reference HLA alleles
Personalized HLA genotyping
Personalized HLA alleles
Somatic variant calling
Somatic HLA variants
Phasing: Variants to haplotypes
Variant groups
Novel HLA sequence reconstruction
Tumor-associated HLA alleles
Allele-specific HLA expression quantification
Tumor HLA abundance in TPMs
Germline variant calling
Germline HLA variants
Phasing: Variants to haplotypes
Variant groups
Novel HLA sequence reconstruction
Novel HLA alleles

FIGURE 1 NeoOncoHLA workflow. This figure illustrates the main steps executed during NeoOncoHLA framework, where NGS reads are used to carry a comprehensive patient-tailored HLA typing, including the characterization of somatic HLA variants in the tumor and expression quantification of the resultant MHC molecules. Its modular architecture consists in four main steps, each one highlighted in a different color: (1) IPD-IMGT/HLA Database closest-matched HLA-typing (blue), (2) personalized HLA profiling enhanced by germline variant calling (orange), (3) robust screening of tumor HLA status through comprehensive somatic variant characterization (green), (4) allele-specific expression quantification of inferred HLA molecules (pink). Text boxes colored with dark blue represent major steps of NeoOncoHLA pipeline, those in white with blue dashed edge are input files, those in gray stand for intermediate output files and those in white with red edge correspond to final output files of the pipeline. MHC, major histocompatibility complex; NGS, next generation sequencing

to the closest-matched HLA sequences produced from the previous HLA typing step[19] (see step 1 in Figure 1). GSNAP[49] (version 2020-05-30) was used for alignment, because of its high accuracy in mapping NGS reads to highly polymorphic regions.[50] The resulting alignment files were processed following standard practices (including sorting and duplicate marking of the reads) and provided as input into two state-of-the-art germline variant calling tools, GATK-HaplotypeCaller[51] (v4.0.6) and Strelka2[52] (v2.0.10; see step 2 in Figure 1). It was reasoned that each germline variant was potentially related to a mistyped or a novel HLA allele. To reduce false positives, only those variants detected by both germline variant callers, having a read depth >10, and a variant allele frequency (VAF) of >0.30 were considered for further analyses.

2.3 | Integration of ensemble somatic variant calling for the identification of tumor-specific HLA variants

We extended the functionality of an ensemble somatic variant calling pipeline,[53] that incorporates six different state-of-the art somatic variant calling tools, to identify somatic mutations in the HLA region. To achieve this, we aligned matched tumor-normal WES reads in the HLA region against the previous typed personalized HLA sequences (known or potentially novel; see Figure 1). because of the high complexity of detecting somatic variants in HLA alleles, we fine-tuned the ensemble variant calling pipeline to consider only high-quality candidate variants reported by at least three out of six variant calling tools.[52,54–58] Each tool has its own algorithm and intrinsic set of rules to distinguish a variant from background noise and therefore, combining them reduces the false positive detection rate. Some additional filters were also applied to discard potential false positive calls, including minimum read depth of 10 for both tumor and normal data at a variant position, and a minimum number of alternative (mutant) reads of three in the tumor and zero in the normal data. The Ensembl variant effect predictor (VEP)[59] toolkit was then used to evaluate the impact of the detected variants on the resulting gene products.

2.4 | Reconstruction of personalized HLA alleles and tumor-specific HLA alleles with correct phased haplotypes

Identifying the correct genomic phase (or haplotype) of the HLA variants was crucial for the subsequent accurate reconstruction of fully phased candidate variant HLA sequences. WhatsHap[60] (v0.17) was used to determine the phase relationship between heterozygous variants along two target HLA alleles. Once the phasing was conducted, Haplosaurus,[61] a method embedded into the Ensembl VEP[59] (v95), was then used to evaluate the functional impact of the detected variants in the HLA allele sequences. We developed customized features that extends the Haplosaurus functionality to annotate and fully reconstruct candidate variant HLA gene, transcript, and protein sequences. Once the variant HLA sequences were successfully reconstructed, a compulsory additional round of HLA typing was conducted, providing the reconstructed germline variant HLA alleles in addition to those available in the IPD-IMGT/HLA Database (see Figure 1, step 2). In this round of HLA typing, NGS reads were aligned not only to the IPD-IMGT/HLA Database, but also to the candidate personalized HLA variant sequences, and the closest-matched allele was assigned accordingly.

The reconstruction of tumor-specific HLA allele sequences was performed in a similar manner as that guided by germline variants, described above. However, in contrast to the germline workflow, all the generated tumor-specific HLA alleles were retained as valid potentials, since tumors may violate the diploid background assumption (because of somatic copy number alterations [CNAs] affecting the ploidy of HLA genes, or tumor heterogeneity; see Figure 1, step 3).

2.5 | Allele-specific expression quantification

We relied on the precise mapping of RNA-Seq reads to the personalized HLA sequences from the patient to obtain reliable expression levels for the patient's HLA alleles. For that purpose, Kallisto[62] (v0.43.1) was used for transcript isoform-level expression quantification. However, we extended Kallisto's functionality whereby previously inferred HLA genotypes were used as an index to assign RNA-Seq reads back to their corresponding HLA sequences (see step 4 in Figure 1). The output included patient-specific HLA allelic abundance measurements, reported as transcripts per million mapped reads (TPM). In addition, this step served to attempt to deconvolute the correct allele from the expression of the corresponding isoforms of the inferred HLA alleles when phasing is not complete.

2.6 | In silico spike-in of germline variants to simulate novel HLA allele discovery

The ability of the personalized HLA typing step to infer novel or uncharacterized HLA genotypes was first evaluated on simulated novel HLA alleles. This simulation framework is summarized here and in Figure 2. To preserve the sequencing error profiles and complexity of biological data, and thus, keep the simulation as faithful as possible to reality, we used BamSurgeon[63] to spike germline variations into the WES data. Three different Class I alleles belonging to three different normal WES patients were selected randomly to apply the simulations (*HLA-A*68:01:02:01*, *HLA-B*51:01:01:01* and *HLA-C*03:04:01:01*). The WES with spiked-in variant reads were then used as input to evaluate the capability our approach to predict novel HLA alleles. In total 1800 independent simulation experiments containing 3200 variants overall were carried out including, SNPs, insertions, and deletions. A wide range of effects in the resultant protein were simulated, including missense, synonymous, in-frame, frameshift, stop gain and stop loss variants. The variants were spiked in both individually and in phased co-occurrences for the purpose of modeling more dissimilar alleles. The simulation framework first verified whether the spiked in variants were detected by the germline variant callers, if not, the experiment was labeled as a miscall. If the variant was called correctly, the variant HLA allele sequence was required to be correctly reconstructed and chosen as the best-matching HLA allele over its reference HLA counterpart. If the reference allele was outputted in this process, the experiment was labeled as an HLA mistype. These simulations were performed for a further 36 HLA-A, -B and -C alleles using 8450 spiked-in single variants to demonstrate the robustness of novel HLA allele discovery across a broader spectrum of HLA alleles.

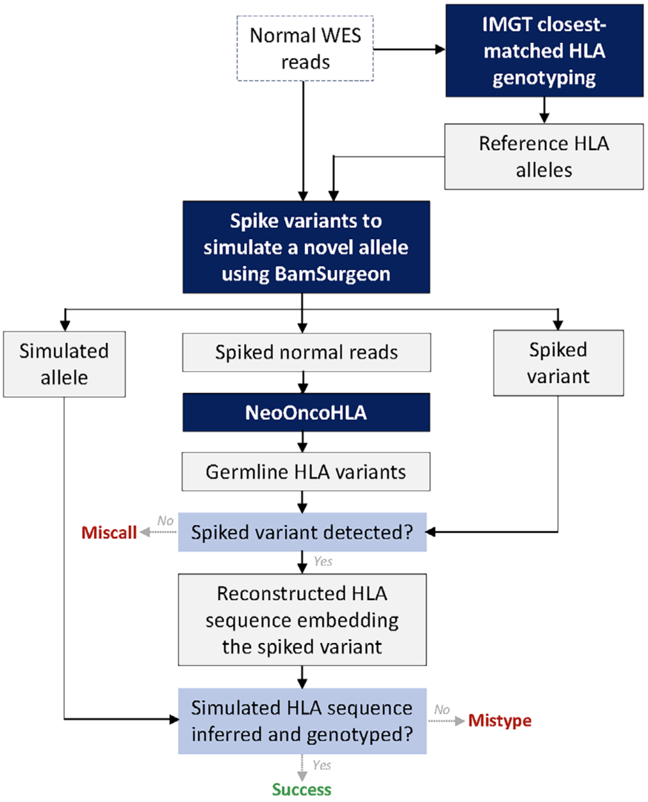
Normal WES reads
IMGT closest-matched HLA genotyping
Reference HLA alleles
Spike variants to simulate a novel allele using BamSurgeon
Simulated allele
Spiked normal reads
Spiked variant
NeoOncoHLA
Germline HLA variants
Miscall
No
Spiked variant detected?
Yes
Reconstructed HLA sequence embedding the spiked variant
Simulated HLA sequence inferred and genotyped?
No
Mistype
Yes
Success

FIGURE 2 In silico novel allele experiment design. After typing IPD-IMGT/HLA Database closest-matched reference HLA alleles using real data, three different Class I alleles (*HLA-A*68:01:02:01*, *HLA-B*51:01:01:01* and *HLA-C*03:04:01:01*) belonging to three different patients were chosen randomly to carry the simulations. A comprehensive set of different germline variant types were spiked-in to simulate novel HLA allele discovery, including SNVs, insertions, deletions. NeoOncoHLA's performance to type novel alleles was assessed attending to its ability to detect the spiked variant and reconstruct the novel HLA allele sequence. If the spiked variant is missed by the variant callers, the experiment was labeled as "miscall," while if the variant is detected but the allele is not correctly inferred afterwards, it was tagged as "mistype." Text boxes colored with dark blue represent major steps of the experiment, those in light blue are checkpoints and those in gray stand for output files

2.7 | Simulation of somatic variants in HLA alleles

Following a similar workflow as in the novel HLA allele simulation experiments, the performance detecting somatic variants on HLA alleles was tested on simulated somatic variants on real data. In this case, the somatic variant was spiked in the tumor WES reads, while the normal WES reads were kept unaltered. In total, 740 simulation experiments were performed, covering the same three HLA alleles as before. The simulations included single nucleotide variants (SNVs), and small insertion and deletions with a VAF ranging from 0.01 to 0.5, allowing for the simulation of heterogeneous tumor subclones or sample contamination. The simulation results were benchmarked against POLYSOLVER (v4),[44] a state-of-the art tumor-specific HLA profiling tool (see Table S1 for the command used). The performance of each tool was then assessed by its ability to detect each simulated somatic variant.

2.8 | WES and HLA targeted sequencing on normal PBMCs and WES on matched metastatic melanoma patients

To assess the performance of the proposed solution on clinical samples, we applied the HLA typing pipeline on 10 WES samples from the normal peripheral blood mononuclear cell (PBMC) of 10 metastatic melanoma donors. Additionally, WES was performed on 14 metastatic melanoma samples (matched to the 10 normal PBMCs from the same metastatic melanoma cohort[64]). All the research and ethics approval and permits together with the written informed consents from all the participants were obtained prior to sample collection. Exome enrichment of the samples was performed using the Agilent AllExome v5 kit, according to the vendor's protocol. The sequencing was carried out by the Illumina HiSeq4000 system using paired-end mode with 151 bp per read and producing 50 million reads per sample, on average.

To validate the accuracy of the results, five samples that presented at least one germline variant in at least one HLA allele were subject to targeted HLA sequencing using NGSgo-MX11-3 HLA-targeted amplification kit and analyzed with NGSengine (v2.20) (GenDx, Utrecht, Netherlands). The NGSgo-MX11-3 kit comprises amplification primers for 11 loci (including HLA-A, -B, -C, -DRB1, -DQB1, -DPB1, DRB3/4/5, DQA1, and DPA), multiplexed in three tubes, resulting in HLA locus-specific amplicons that were then used for HLA typing. To eliminate the possibility of analyses performed on different versions of the IPD-IMGT/HLA Database, both the WES and the HLA targeted sequencing analyses was performed using identical versions of the IPD-IMGT/HLA Database (v.3.41.2).

2.9 | Orthogonal validation of somatic HLA variants using RNA-Seq

We next attempted to validate the somatic HLA variants using WES data from 10 metastatic melanoma patients, by also conducting an orthogonal validation using the available RNA-Seq data for the same tumor samples. Because of the nature of RNA-Seq data, only those variants located on exons regions could be subject of this validation. A variant discovered in the WES data was considered confirmed when at least one read harboring the alternate allele was found in the RNA-Seq data. A variant remained unconfirmed because of a false positive WES call, low quality of RNA-Seq data or expression down-regulation, where the allele's expression was switched off or lowly expressed. Confirming the presence of the detected variant at RNA level (when feasible), significantly reduced the probability of the variant being called erroneously because of a sequencing error.

3 | RESULTS

3.1 | Simulating the discovery of novel HLA alleles from germline WES data

We first evaluated the capability of the personalized HLA typing approach to capture novel HLA alleles from WES data on simulated HLA variants. The HLA alleles *HLA-A*68:01:02:01*, *HLA-B*51:01:01:01* and *HLA-C*03:04:01:01* were used to perform 1800 independent simulations by applying different combinations of germline variations to assess the performance of our approach (see Section 2, Figure 2). We simulated a total of 3200 HLA variants randomly distributed across the length of the selected alleles. Table 1 summarizes the results, where experiments are

TABLE 1 Novel allele simulation results by spiked germline HLA variants

Allele	Simulated variant combination	Total number of simulations	Successful experiments	Success rate (%)	Miscall	Mistyped
A	SNV	100	100	100.00	0	0
	SNV x2	100	79	79.00	7	14
	SNV x3	100	76	76.00	18	6
	Deletion	100	98	98.00	0	2
	Deletion x2	100	62	62.00	37	1
	Insertion	100	100	100.00	0	0
	Insertion x2	100	62	62.00	37	1
	SNV + deletion	100	99	99.00	1	0
	SNV + insertion	100	100	100.00	0	0
B	SNV	50	47	94.00	0	3
	SNV x2	50	46	92.00	1	3
	SNV x3	50	36	72.00	14	0
	Deletion	50	49	98.00	0	1
	Deletion x2	50	28	56.00	22	0
	Insertion	50	50	100.00	0	0
	Insertion x2	50	31	62.00	17	2
	SNV + deletion	50	49	98.00	0	1
	SNV + insertion	50	50	100.00	0	0
C	SNV	50	46	92.00	0	4
	SNV x2	50	40	80.00	2	8
	SNV x3	50	29	58.00	10	11
	Deletion	50	39	78.00	0	11
	Deletion x2	50	23	46.00	25	2
	Insertion	50	42	84.00	0	8
	Insertion x2	50	29	58.00	5	16
	SNV + deletion	50	41	82.00	0	9
	SNV + insertion	50	43	86.00	1	6

Note: The table shows a comprehensive overview of the results of 1800 simulation experiments grouped attending to reference allele and spiked variant type combination. Alleles: A (*HLA-A*68:01:02:01*), B (*HLA-B*51:01:01:01*) and C (*HLA-C*03:04:01:01*).

classified according to the mutated allele, variant type, and number of co-occurring simulated variants. Overall, we detected simulated variants and correctly inferred the novel HLA allele at a success rate of 83% across all the experiments. Considering only those experiments in which single mutations were spiked in into the alleles, a 97%, 93% and 96% success rate was observed for SNVs, deletions and insertions respectively (see Table S2). As expected, a reduction in the success rate was observed when simulating co-occurring phased variants in the same allele (see Table 1). This reduction was particularly notable when simulating co-occurring indels, because of the challenge of mapping reads with multiple indel related mismatches. With respect to specific HLA genes, an 86%, 86% and 74% success rate was observed for A, B and C alleles, respectively (see Table S2 for detailed overview). To demonstrate the approach's ability to identify novel alleles across a wider spectrum of HLA alleles, we extended these experiments with 8450 spiked-in single variants further simulations across a diverse range of an additional 36 HLA-A, -B and -C alleles to achieve an overall success rate of 97% (see Table S2). Overall, the simulations summarized in Table 1 and Table S2 indicate the potential of the proposed approach to accurately identify germline variants affecting HLA alleles, leading to novel HLA sequences from WES-based NGS data.

3.2 | Validation of WES-based HLA typing at protein coding sequence level

We then ran NeoOncoHLA on WES data from the PBMCs of 10 donors. For validation purposes, five samples from the 10 donors, where at least one germline variation in an HLA Class I allele was predicted, were also subject to targeted HLA sequencing from the GenDx NGSgo-MX11-3 kit (see Section 2). We compared the HLA typing results obtained here with NeoOncoHLA, OncoHLA (a previously published typing pipeline that does not incorporate variant calling[19]) and the high-resolution targeted HLA sequencing (GenDx). The validation results are described in Table 2. There was a 100% overlap in the HLA typing between NeoOncoHLA and OncoHLA using WES data and HLA targeted sequencing at protein coding sequence level (i.e., at the first and second field of resolution). This 100% performance overlap with HLA targeted sequencing was a validation of both NeoOncoHLA and OncoHLA, for the first and second field of resolution.[19]

OncoHLA, the HLA typing from WES data without variant calling integration,[19] had a reduced performance at the third and fourth fields, with 86.7% and 70% for the third and fourth fields respectively (see Table 2). This reduced performance was as expected, as for any WES-based HLA typing solution, because of the lower coverage of HLA sequences in the IPD-IMGT/HLA Database for all HLA exons and non-coding sequences in addition to the moderate read depth of WES compared with targeted NGS. The performance for our WES-based HLA typing solutions however improved significantly when using deep or targeted HLA NGS data as input (see Table 2 and Table S3).

3.3 | Integration of germline variant calling enhances personalized HLA typing and enables novel HLA discovery from WES data

Personalized HLA typing, through the integration of variant calling (see Figure 1), significantly improved the performance at the third and fourth fields of resolution using WES data (see Table 2). It was demonstrated that the performance was enhanced through the integration of variant calling, raising the accuracy to 96.7% from 86.7%, and to 86.7% from 70% for the third and fourth fields, respectively. The complete results for all five patients that had at least one germline mutated HLA allele is available in Table S3. In Table 3, the HLA typing results with a discrepancy using WES data versus targeted HLA sequencing data is also depicted. In all the six cases of variant HLA alleles among the normal PBMC samples of five patients, where there was at least one germline variant detected, NeoOncoHLA was able to

TABLE 2 Overlap percentage on the validation of HLA typing of Class I alleles by our previous tool OncoHLA using WES data versus OncoHLA with targeted HLA NGS sequencing data versus NeoOncoHLA using WES data. NeoOncoHLA has incorporated variant calling, whereas OncoHLA does not

Resolution	OncoHLA with WES data	OncoHLA with targeted NGS data	NeoOncoHLA with WES data
1 field	100	100	100
2 field	100	100	100
3 field	86.67	93.33	96.67
4 field	70	86.67	86.67

TABLE 3 Validating potential mistype correction and discovery of novel HLA genotypes using personalized germline variant calling by NeoOncoHLA

Sample_ID	OncoHLA HLA typing using WES data	GenDx HLA typing using targeted HLA sequencing data	Discrepancy with targeted HLA sequencing	HLA germline variants detected by NeoOncoHLA	Outcome of NeoOncoHLA with personalized germline variant calling
UV1-0001	*HLA-B*40:01:01*	*HLA-B*40:01:02:01*	3rd field	Yes	Mistyping fixed
UV1-0002	*HLA-A*24:02:32*	*HLA-A*24:02:01:01*	3rd field	No	No germline variants detected
UV1-0002	*HLA-B*35:01:01:01*	*HLA-B*35:01:01:02*	4th field	Yes	Mistyping fixed
UV1-0002	*HLA-C*04:01:01:01*	*HLA-C*04:01:01:13*	4th field	No	No germline variants detected
UV1-0002	*HLA-C*04:01:115*	*HLA-C*04:01:01:05*	3th field	Yes	Mistyping fixed
UV1-0006	*HLA-B*40:01:01*	*HLA-B*40:01:02:01*	3rd field	Yes	Mistyping fixed
UV1-0006	*HLA-C*03:04:01:01*	*HLA-C*03:04:01:02*	4th field	No	No germline variants detected
UV1-0011	*HLA-B*44:02:01:03*	*HLA-B*44:02:01:03*	Equal	Yes	Potential new allele characterization
UV1-0013	*HLA-B*52:01:01:01*	*HLA-B*52:01:01:02*	4th field	Yes	Mistyping fixed
UV1-0013	*HLA-C*12:02:02:02*	*HLA-C*12:02:02:01*	4th field	No	No germline variants detected

Note: NeoOncoHLA has incorporated variant calling, whereas OncoHLA does not, enabling the rescue of the correct HLA allele when mistyping occurs and importantly, the detection of novel alleles.

correct the HLA mistypes. In one of these cases, the analysis led to the discovery of a novel Class I HLA allele confirmed by both WES, targeted NGS sequencing, and officially assigned the name *HLA-B*44:02:01:52* (see Table 3), and consequently improved the typing as shown in Table 2. Figure 3 illustrates the germline variant found in the 3′UTR of the closest-matched reference *HLA-B*44:02:01:03* allele, responsible of the novel *HLA-B*44:02:01:52* allele. This Class I novel variant was also subject to confirmation at the transcriptional level from the tumor sample matched to the same patient (see Section 2: “Allele-specific expression quantification”).

3.4 | Evaluation of ensemble somatic variant calling for enhanced detection of somatic variants in HLA alleles

The ability of NeoOncoHLA to detect somatic HLA variants was benchmarked against POLYSOLVER. In total, 740 simulation experiments were conducted, covering three HLA alleles (one HLA-A, one HLA-B and one HLA-C, see Section 2); including SNVs, small insertions and deletions; with a VAF ranging from 0.01 to 0.5, to mimic tumor heterogeneity. The performance of each tool was assessed by its ability to detect the spiked somatic variant. Overall, 677 (91%) and 556 (75%) of the 740 simulated somatic variants were detected by NeoOncoHLA and POLYSOLVER, respectively (see Section 2 for simulation description, and Table S4 for detailed results). A summary of the benchmarking comparison is illustrated in Figure 4. NeoOncoHLA outperformed POLYSOLVER across all the alleles and variant types. In addition, NeoOncoHLA had an improved performance across all the various simulated VAF values (see Table S4).

3.5 | Somatic variant calling in Class I HLA alleles using personalized germline HLA alleles as reference

We next applied NeoOncoHLA on WES data from the 14 metastatic melanoma samples from 10 patients. To capture somatic HLA variants with improved fidelity, we called the variants by using the personalized HLA sequences derived from the matched normal PBMCs.

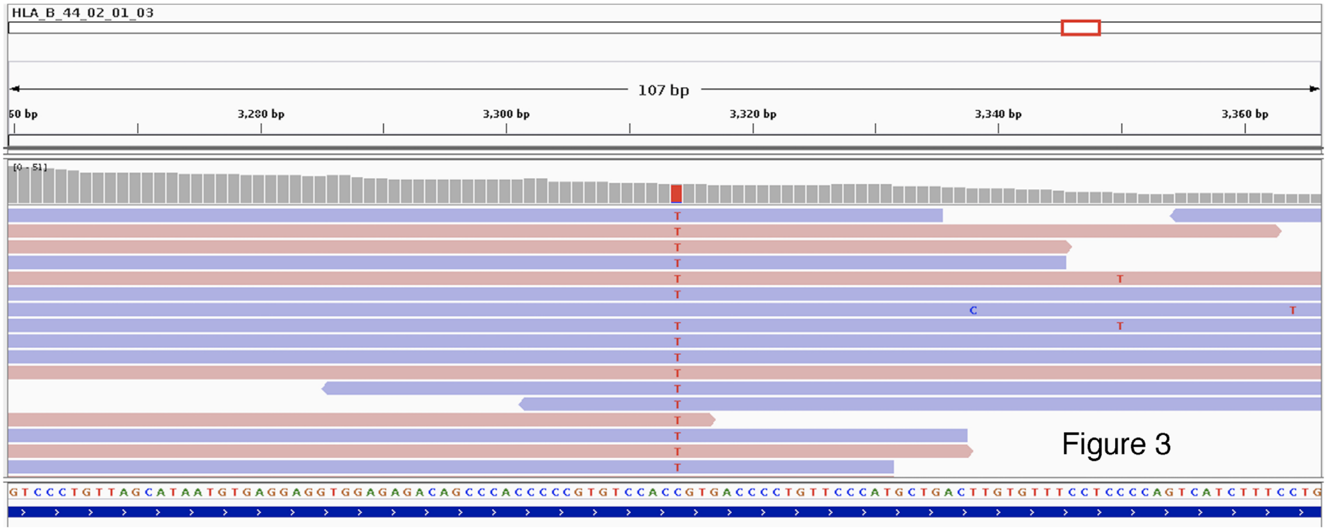

FIGURE 3 Integrative Genomics Viewer (IGV) visualization of the germline variant conforming the detected novel allele. SNP affecting the 3′UTR region of reference *HLA-B*44:02:01:03* allele, leading to the characterization of the novel *HLA-B*44:02:01:52* allele

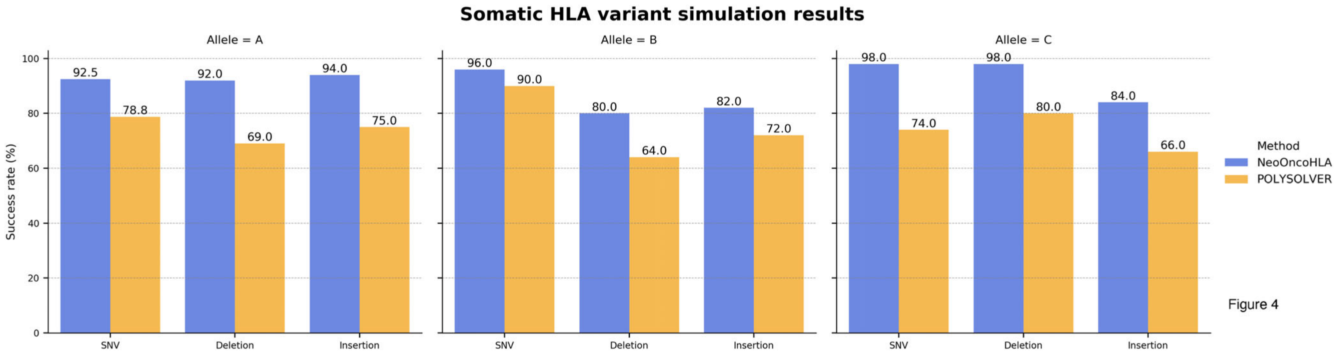

FIGURE 4 NeoOncoHLA VS POLYSOLVER performance benchmark in somatic HLA variant simulation experiments. The figure is divided into three histograms, one per Class I allele under evaluation. The *y*-axis shows the success rate of each tool to detect the spiked somatic variant type specified in the *x*-axis

In total, 15 somatic mutations were detected in classical Class I alleles across the 14 melanoma samples, 7 (47%) of those were found in HLA-A, five (33%) in HLA-B and the remaining three (20%) in HLA-C alleles (see table 1 in Table S5). The functional consequences of the somatic variants were inferred from their predicted effect on the resultant protein using VEP. In total, five out of 15 (33%) variants were situated in protein-coding regions, while 10 (66%) were in non-coding genomic regions. The predicted functional effects of the 15 detected somatic variants sorted from low to high impact were distributed as follows: five intron variants, two 5′UTR variants, three 3′UTR variants, one synonymous variant and four missense variants (see table 2 in Table S5). One of the four missense variants occurred, interestingly, in the binding cleft of *HLA-A*02:01:01:01*.

The mean VAF across the detected somatic changes was moderately low (0.13), indicating a broad genomic heterogeneity and the presence of sub-clonal mutations private to subpopulations of cancer cells. However, one of the somatic HLA variants located at *HLA-A*03:01:01:01* of sample UV1-0009-T01 showed a high VAF of 0.707. This variant was annotated as missense, changing the second amino acid encoded by the allele from an alanine to valine. The variant was also confirmed in tumor RNA-Seq data, and the resultant novel somatic allele expressed with a high abundance (TPM of 914.75).

Orthogonal validation using RNA-Seq data was conducted to confirm the five somatic HLA variants on protein-coding regions (see Section 2). A variant discovered in the WES data was considered confirmed when at least one read harboring the alternate allele was found in the RNA-Seq data. We observed three of the five exonic variants confirmed in RNA-Seq data. The expression abundance of the total of four mutant HLA alleles sequences was calculated using Kallisto (see Section 2). As expected, the mutant HLA alleles harboring those unconfirmed RNA-seq variants and with very low VAF (0.01–0.02) were not expressed. The remaining two alleles had a TPM of 7.23 and 914.76 (the latter being the variant with high VAF of 0.707 in *HLA-A*03:01:01:01* of sample UV1-0009-T01, mentioned above).

POLYSOLVER applied to the same 14 metastatic melanoma samples detected 28 somatic variants compared with the 15 from NeoOncoHLA. Only one variant was called by both tools that being the high VAF (0.707) SNV, mentioned above. One source of the large difference of note was that NeoOncoHLA uses all the alleles available at IPD-IMGT/HLA Database to align NGS reads, including classical Class I, classical Class II and non-classical alleles, allowing the NGS reads to map back and align to their true origin, whereas POLYSOLVER uses classical Class I alleles only. POLYSOLVER reported the same variant twice, once per typed HLA allele (see table 3 in Table S5). This was addressed in NeoOncoHLA by being more stringent with the filters (see Section 2) where multiple variant calling tools must detect a somatic variant compared with a single tool in POLYSOLVER, improving the mapping of the NGS reads by allowing them to map against Class II and non-classical HLA alleles, and also by including allelic variant phasing steps taking advantage of both WES and RNA-Seq data.

4 | DISCUSSION

We have reported herein an NGS-based HLA typing approach that relies on alignments to known HLA alleles in the IPD-IMGT/HLA Database, while simultaneously also enabling the discovery of novel HLA alleles, and tumor-specific HLA variants. In summary, the typing method described here maximizes the value of HLA database-matching method and is also capable of discovering novel and tumor-specific HLA alleles through the systematic integration of variant calling applied to WES data. The vast majority of NGS-based HLA typing tools that use the IPD-IMGT/HLA Database are limited to typing only known HLA alleles. Hence, the accuracy of HLA typing from database-matching based methods relies highly on the completeness of the IPD-IMGT/HLA Database. This limitation, as previously mentioned, is particularly problematic for individuals harboring uncharacterized HLA alleles and can also cause considerable challenges in the characterization of tumor-specific HLA variants.

De novo assembly-based NGS algorithms[33,37,65,66] can achieve novel HLA discovery by building a consensus sequence from reads without relying on a reference library. However, these algorithms are computationally expensive (particularly when deeper sequencing is required) and require high coverage and longer reads for accurate phased HLA typing. Furthermore, the outputted sequences of de novo assembly methods still need to be aligned to their closest match in the IPD-IMGT/HLA Database to fully characterize the HLA allelic variants, consequently resolving ambiguities and discovering full-length novel HLA alleles remains challenging even for de novo assembly methods. In addition to the de novo assembly efforts, there are other methods designed with a similar motivation as this study, namely, to detect HLA allelic variants that are not characterized in the IPD-IMGT/HLA Database. The ALPHLARD is one such study that used a probabilistic model to infer new HLA alleles.[67] However, although demonstrating good performance for HLA typing, no novel candidate HLA alleles was identified in that study.[67]

A methodology capable of accurately typing the HLA region and identify novel alleles, using standard WES data alone, could help improve the completeness of HLA libraries and therefore enhance the HLA typing accuracy in clinical or research applications. In the approach described in this study, we also relied on the IPD-IMGT/HLA Database, however novel HLA allele discovery, and tumor-specific HLA variants, was made possible through the systematic integration of variant calling tools. Strong validation for HLA typing was demonstrated at the first and second fields of resolution using targeted HLA sequencing data from the blood of the five donors that had at least one candidate germline variant in their HLA alleles. This validation performance was consistent with much of the recent literature on the performance of NGS-based HLA typing. The reduction of the validation observed at the third and fourth fields of resolution was improved when applying the germline variant calling strategy, resulting in the recovery of the correct allele, or inferring a novel allele not described in IPD-IMGT/HLA Database. This was particularly promising given that standard WES data was used to perform the HLA typing, and therefore had very sparse coverage of reads in non-coding regions (introns and UTR's), where ambiguities lie at the fourth field of resolution. Interestingly, non-synonymous HLA variants were never detected in exons that encode the peptide-binding cleft in any of our analyses (i.e., in exons 2 and 3 for Class I); reflecting the comprehensive coverage of these exons in the IPD-IMGT/HLA Database for HLA allelic variants that bestow different antigen presentation patterns. The ability of the approach described to discover new alleles was demonstrated by the detection of a novel HLA-B allele. The novel allele has been assigned the name *B*44:02:01:52* and officially cataloged by the WHO Nomenclature Committee for Factors of the HLA System. Finally, an improved toolkit to interrogate HLA variation in the tumors, to help understand the interplay between cancer progression and the adaptive immune system,[41,68] was demonstrated in simulated experiments and in metastatic melanoma WES samples. Improved tools for the identification of HLA variants from WES in tumors is important, as large-scale WES studies have previously revealed that somatic variation in HLA Class I genes[69–72] and their expression are associated to immunosurveillance and clinical outcome.[44,72]

The approach described here is capable of effectively identifying novel Class I HLA alleles and tumor-specific HLA variants, through the systematic integration of variant calling. However, the approach is limited to alleles whose sequences share a relatively high degree of similarity to those of already known alleles in the HLA reference databases. Hence, this approach has been mostly tested on classical Class I alleles, the most extensive collection of full-length allele sequences available in IPD-IMGT/HLA Database. We hope to benchmark this approach on Class II and non-classical alleles as full-length HLA gene sequences become increasingly submitted for these groups into the IPD-IMGT/HLA Database. For the identification of highly dissimilar novel HLA alleles (as those presenting structural variants such as large insertions or deletions) from NGS data, a de novo alignment approach may be more optimal.[45] Additionally, the personalized germline HLA variant calling steps in this study were restricted to the consensus of three state-of-the-art variant callers. As numerous NGS-based germline variant callers have been developed in recent years,[73] with solutions to variant detection in complex genome regions like HLA continuously emerging[74,75]; it may be beneficial to investigate ensemble approaches to variant calling,[53] taking input from numerous germline variant callers in order to optimize the accuracy of germline variant detection in the HLA region.

A well-known source of ambiguous HLA typing results is characterized by the difficulties to reliably infer the phase relationship between variants along two HLA alleles and consistently reconstruct the fully phased HLA haplotypes.[16] This limitation can be attributed to the very nature of short reads in WES data. To correctly phase alleles with short NGS reads, they must adequately cover the variant region; making it particularly difficult to reliably span the distance between variants located at both ends of the HLA gene. Incorrect phasing is an arduous challenge in the HLA genotype field, currently, and constitutes a major source of spurious HLA typing results.[16] When ambiguous phasing occurs, our method employs RNA-Seq to only select those reconstructed HLA alleles with verification of RNA expression. Long read sequencing could also be used to address the challenge of haplotype phasing of HLA alleles harboring long intronic regions, however, this was out of the scope of this current study.

In summary, the HLA typing approach described herein showed good performance for full-length HLA typing from WES data, when validated using targeted HLA sequencing and demonstrated an ability to detect novel HLA alleles and tumor-specific HLA variants. With a large amount of WES-based NGS data being continuously accumulated in many clinical and research studies worldwide, this approach may lead to the discovery of more novel HLA alleles and help fill some of the gaps in the IPD-IMGT/HLA Database. Furthermore, using a personalized germline reference HLA genotype to perform somatic variant calling, allows tumor-specific HLA variants to be identified with increased fidelity, and help to characterize HLA associated tumor-immune escape.

CONFLICT OF INTEREST DECLARATION

IA, AZ, PS, RS and TC are employees of NEC OncoImmunity, a subsidiary of NEC Corporation. EBE and GG are employees of Ultimovacs ASA.

AUTHOR CONTRIBUTIONS

Irantzu Anzar and Trevor Clancy designed the study and drafted the manuscript. Irantzu Anzar, Angelina Sverchkova and Pubudu Samarakoon performed software development, and bioinformatics data analysis and collection. All authors performed data interpretation and reviewed the manuscript. All authors approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in IMGT at http://www.imgt.org/. These data were derived from the following resources available in the public domain: - EGAS00001005253 (EGA accession number), https://ega-archive.org/- HLA variant simulation code, https://github.com/OncoImmunity/hla_simulator

ORCID

Trevor Clancy https://orcid.org/0000-0001-9896-0613

REFERENCES

1. Robinson J, Malik A, Parham P, Bodmer JG, Marsh SG. IMGT/HLA database—a sequence database for the human major histocompatibility complex. *Tissue Antigens*. 2000;55:280-287.

2. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res*. 2015;43:D423-D431.
3. Horton R, Wilming L, Rand V, et al. Gene map of the extended human MHC. *Nat Rev Genet*. 2004;5:889-899.
4. Morishima Y, Sasazuki T, Inoko H, et al. The clinical significance of human leukocyte antigen (HLA) allele compatibility in patients receiving a marrow transplant from serologically HLA-A, HLA-B, and HLA-DR matched unrelated donors. *Blood*. 2002;99:4200-4206.
5. Bray RA, Hurley CK, Kamani NR, et al. National marrow donor program HLA matching guidelines for unrelated adult donor hematopoietic cell transplants. *Biol Blood Marrow Transplant*. 2008;14:45-53.
6. Lee SJ, Klein J, Haagenson M, et al. High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood*. 2007;110:4576-4583.
7. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*. 2008;40:695-701.
8. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. *Annu Rev Genomics Hum Genet*. 2013;14:301-323.
9. Lee MT, Mahasirimongkol S, Zhang Y, et al. Clinical application of pharmacogenomics: the example of HLA-based drug-induced toxicity. *Public Health Genomics*. 2014;17:248-255.
10. Fan WL, Shiao MS, Hui RC, et al. HLA association with drug-induced adverse reactions. *J Immunol Res*. 2017;2017:3186328.
11. Richard C, Fumet JD, Chevrier S, et al. Exome analysis reveals genomic markers associated with better efficacy of Nivolumab in lung cancer patients. *Clin Cancer Res*. 2019;25:957-966.
12. Ivanova M, Shivarov V. HLA genotyping meets response to immune checkpoint inhibitors prediction: a story just started. *Int J Immunogenet*. 2021;48:193-200.
13. Chowell D, Krishna C, Pierini F, et al. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat Med*. 2019;25:1715-1720.
14. Profaizer T, Pole A, Monds C, Delgado JC, Lazar-Molnar E. Clinical utility of next generation sequencing based HLA typing for disease association and pharmacogenetic testing. *Hum Immunol*. 2020;81:354-360.
15. Bravo-Egana V, Sanders H, Chitnis N. New challenges, new opportunities: next generation sequencing and its place in the advancement of HLA typing. *Hum Immunol*. 2021;82:478-487.
16. Klasberg S, Surendranath V, Lange V, Schofl G. Bioinformatics strategies, challenges, and opportunities for next generation sequencing-based HLA genotyping. *Transfus Med Hemother*. 2019;46:312-325.
17. Schofl G, Lang K, Quenzel P, et al. 2.7 million samples genotyped for HLA by next generation sequencing: lessons learned. *BMC Genomics*. 2017;18:161.
18. Orenbuch R, Filip I, Comito D, Shaman J, Pe'er I, Rabadan R. arcasHLA: high-resolution HLA typing from RNAseq. *Bioinformatics*. 2020;36:33-40.
19. Sverchkova A, Anzar I, Stratford R, Clancy T. Improved HLA typing of Class I and Class II alleles from next-generation sequencing data. *HLA*. 2019;94:504-513.
20. Rotimi CN, Jorde LB. Ancestry and disease in the age of genomic medicine. *N Engl J Med*. 2010;363:1551-1558.
21. Zhou F, Cao H, Zuo X, et al. Deep sequencing of the MHC region in the Chinese population contributes to studies of complex disease. *Nat Genet*. 2016;48:740-746.
22. Flomenberg N, Baxter-Lowe LA, Confer D, et al. Impact of HLA class I and class II high-resolution matching on outcomes of unrelated donor bone marrow transplantation: HLA-C mismatching is associated with a strong adverse effect on transplantation outcome. *Blood*. 2004;104:1923-1930.
23. Mayor NP, Hayhurst JD, Turner TR, et al. Recipients receiving better HLA-matched hematopoietic cell transplantation grafts, uncovered by a novel HLA typing method, have superior survival: a retrospective study. *Biol Blood Marrow Transplant*. 2019;25:443-450.
24. Vazirabad I, Chhabra S, Nytes J, et al. Direct HLA genetic comparisons identify highly matched unrelated donor-recipient pairs with improved transplantation outcome. *Biol Blood Marrow Transplant*. 2019;25:921-931.
25. Agarwal RK, Kumari A, Sedai A, Parmar L, Dhanya R, Faulkner L. The case for high resolution extended 6-loci HLA typing for identifying related donors in the Indian subcontinent. *Biol Blood Marrow Transplant*. 2017;23:1592-1596.
26. Buhler S, Baldomero H, Ferrari-Lacraz S, et al. High-resolution HLA phased haplotype frequencies to predict the success of unrelated donor searches and clinical outcome following hematopoietic stem cell transplantation. *Bone Marrow Transplant*. 2019;54:1701-1709.
27. Hosomichi K, Shiina T, Tajima A, Inoue I. The impact of next-generation sequencing technologies on HLA research. *J Hum Genet*. 2015;60:665-673.
28. Carapito R, Radosavljevic M, Bahram S. Next-generation sequencing of the HLA locus: methods and impacts on HLA typing, population genetics and disease association studies. *Hum Immunol*. 2016;77:1016-1023.
29. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. *Brief Funct Genomics*. 2016;15:374-384.
30. Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. *J Hum Genet*. 2014;59:5-15.
31. Wittig M, Schmöhl M, Koch S, et al. HLAssign 2.0: an advanced graphical user Interface for the analysis of short and long read human leukocyte antigen-typing data. *bioRxiv*. 2005;2020(2020):2025.
32. Nordin J, Ameur A, Lindblad-Toh K, Gyllensten U, Meadows JRS. SweHLA: the high confidence HLA typing bio-resource drawn from 1000 Swedish genomes. *Eur J Hum Genet*. 2020;28:627-635.
33. Lee H, Kingsford C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol*. 2018;19:16.
34. Creary LE, Guerra SG, Chong W, et al. Next-generation HLA typing of 382 international histocompatibility working group reference B-lymphoblastoid cell lines: report from the 17th international HLA and immunogenetics workshop. *Hum Immunol*. 2019;80:449-460.

35. Osoegawa K, Vayntrub TA, Wenda S, et al. Quality control project of NGS HLA genotyping for the 17th international HLA and Immunogenetics workshop. *Hum Immunol*. 2019;80:228-236.
36. Turner TR, Hayhurst JD, Hayward DR, et al. Single molecule real-time DNA sequencing of HLA genes at ultra-high resolution from 126 international HLA and Immunogenetics workshop cell lines. *HLA*. 2018;91:88-101.
37. Lee H, Kingsford C. Accurate assembly and typing of HLA using a graph-guided assembler Kourami. *Methods Mol Biol*. 2018;1802:235-247.
38. Wang C, Xiong C, Hsu Y-C, Wang X, Chen L. Human leukocyte antigen (HLA) and cancer immunotherapy: HLA-dependent and -independent adoptive immunotherapies. *Ann Blood*. 2020;5(14).
39. Rosenberg SA. Raising the bar: the curative potential of human cancer immunotherapy. *Sci Transl Med*. 2012;4(127):127ps8–127ps8.
40. Sambi M, Bagheri L, Szewczuk MR. Current challenges in cancer immunotherapy: multimodal approaches to improve efficacy and patient response rates. *J Oncol*. 2019;2019:4508794.
41. Jhunjhunwala S, Hammer C, Delamarre L. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. *Nat Rev Cancer*. 2021;21:298-312.
42. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol*. 2020;20:651-668.
43. Algarra I, Garrido F, Garcia-Lora AM. MHC heterogeneity and response of metastases to immunotherapy. *Cancer Metastasis Rev*. 2021;40:501-517.
44. Shukla SA, Rooney MS, Rajasagi M, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol*. 2015;33:1152-1158.
45. Hayashi S, Moriyama T, Yamaguchi R, et al. ALPHLARD-NT: Bayesian method for human leukocyte antigen genotyping and mutation calling through simultaneous analysis of Normal and tumor whole-genome sequence data. *J Comput Biol*. 2019;26:923-937.
46. Balas A, Planelles D, Goterris R, Rodriguez-Cebria M, Vicario JL. Somatic mutation in the two HLA-B genes of a patient with acute myelogenous leukemia. *HLA*. 2019;94:360-364.
47. Montesion M, Murugesan K, Jin DX, et al. Somatic HLA class I loss is a widespread mechanism of immune evasion which refines the use of tumor mutational burden as a biomarker of checkpoint inhibitor response. *Cancer Discov*. 2021;11:282-292.
48. Garcia-Lora A, Algarra I, Garrido F. MHC class I antigens, immune surveillance, and tumor immune escape. *J Cell Physiol*. 2003;195:346-355.
49. Wu TD, Reeder J, Lawrence M, Becker G, Brauer MJ. GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality. *Methods Mol Biol*. 2016;1418:283-334.
50. Tian S, Yan H, Neuhauser C, Slager SL. An analytical workflow for accurate variant discovery in highly divergent regions. *BMC Genomics*. 2016;17:703.
51. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. 2017.;
52. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;15:591-594.
53. Anzar I, Sverchkova A, Stratford R, Clancy T. NeoMutate: an ensemble machine learning framework for the prediction of somatic mutations in cancer. *BMC Med Genomics*. 2019;12:63.
54. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31:213-219.
55. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res*. 2012;22:568-576.
56. Lai Z, Markovets A, Ahdesmaki M, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016;44:e108.
57. Larson DE, Harris CC, Chen K, et al. SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;28:311-317.
58. Erik Garrison GM. Haplotype-based variant detection from short-read sequencing. *arXiv*. 2012.;
59. McLaren W, Gil L, Hunt SE, et al. The Ensembl variant effect predictor. *Genome Biol*. 2016;17:122.
60. Patterson M, Marschall T, Pisanti N, et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *J Comput Biol*. 2015;22:498-509.
61. Spooner W, McLaren W, Slidel T, et al. Haplosaurus computes protein haplotypes for use in precision drug design. *Nat Commun*. 2018;9:4128.
62. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34:525-527.
63. Ewing AD, Houlahan KE, Hu Y, et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods*. 2015;12:623-630.
64. Aamdal E, Inderberg EM, Ellingsen EB, et al. Combining a universal telomerase based cancer vaccine with Ipilimumab in patients with metastatic melanoma - five-year follow up of a phase I/IIa trial. *Front Immunol*. 2021;12:663865.
65. Huang Y, Yang J, Ying D, et al. HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med*. 2015;7:25.
66. Warren RL, Choe G, Freeman DJ, et al. Derivation of HLA types from shotgun sequence datasets. *Genome Med*. 2012;4:95.
67. Hayashi S, Yamaguchi R, Mizuno S, et al. ALPHLARD: a Bayesian method for analyzing HLA genes from whole genome sequence data. *BMC Genomics*. 2018;19:790.
68. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. *Nat Immunol*. 2002;3:991-998.
69. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*. 2014;505:495-501.
70. Ozcan M, Janikovits J, von Knebel DM, Kloor M. Complex pattern of immune evasion in MSI colorectal cancer. *Onco Targets Ther*. 2018;7:e1445453.
71. Stransky N, Egloff AM, Tward AD, et al. The mutational landscape of head and neck squamous cell carcinoma. *Science*. 2011;333:1157-1160.

72. Castro A, Ozturk K, Pyke RM, Xian S, Zanetti M, Carter H. Elevated neoantigen levels in tumors with somatic mutations in the HLA-A, HLA-B, HLA-C and B2M genes. *BMC Med Genomics*. 2019;12:107.
73. Koboldt DC. Best practices for variant calling in clinical sequencing. *Genome Med*. 2020;12:91.
74. Cooke DP, Wedge DC, Lunter G. A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol*. 2021;39:885-892.
75. Poplin R, Chang PC, Alexander D, et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol*. 2018;36:983-987.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Anzar I, Sverchkova A, Samarakoon P, et al. Personalized HLA typing leads to the discovery of novel HLA alleles and tumor-specific HLA variants. *HLA*. 2022;99(4): 313-327. doi:10.1111/tan.14562