World Academy of Science, Engineering and Technology
International Journal of Bioengineering and Life Sciences
Vol:9, No:3, 2015

Accurate HLA Typing at High-Digit Resolution from NGS Data

Yazhi Huang, Jing Yang, Dingge Ying, Yan Zhang, Vorasuk Shotelersuk, Nattiya Hirankarn, Pak Chung Sham, Yu Lung Lau, Wanling Yang

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

***Abstract*—**Human leukocyte antigen (HLA) typing from next generation sequencing (NGS) data has the potential for applications in clinical laboratories and population genetic studies. Here we introduce a novel technique for HLA typing from NGS data based on read-mapping using a comprehensive reference panel containing all known HLA alleles and de novo assembly of the gene-specific short reads. An accurate HLA typing at high-digit resolution was achieved when it was tested on publicly available NGS data, outperforming other newly-developed tools such as HLAminer and PHLAT.

***Keywords*—**Human leukocyte antigens, next generation sequencing, whole exome sequencing, HLA typing.

I. Introduction

THE human leukocyte antigens (HLA) include a large number of genes crucial to immune system function. They play important roles in immune responses to infection, transplant rejection, pathogenesis of autoimmune diseases, adverse drug reaction, and cancer development. Thus, HLA typing is very important for both clinical laboratories and biomedical research. Since having NGS data for large number of healthy individuals is rapidly becoming a reality, potential benefit of HLA screening from this type of data is multi-fold.

However, HLA typing has always been challenging due to the complexity of this group of genes, including existence of large number of alleles for most HLA genes, major sequence difference between these alleles, sequence similarity among the paralogous HLA genes, and long range linkage disequilibrium (LD) in this region [1], [2]. For example, for HLA-DRB1 gene alone, over a thousand alleles have been reported in human populations according to IMGT/HLA database (IMGT/HLA 2012 Release 3.10.0) [3]. In addition, many HLA-DRB5 alleles have great sequence similarity to those of DRB1, adding more difficulties for accurately calling HLA alleles from sequencing data [4].

HLA typing has been done via various technologies such as serological, cellular, and molecular assays [5]. Sequencing-based methods have been rapidly gaining popularity due to technology advancement, especially in research settings. With the development of NGS, large amount of sequencing data are becoming widely available. Although most of them were not generated for this purpose, they still provide valuable resources for HLA typing. NGS data might be useful in many aspects such as preliminary screening for potential organ donors, for individuals that are potentially susceptible to adverse drug responses, risk prediction for complex diseases, and population genetic studies [6], [7]. They also provide much more comprehensive information on this region than any other traditional methods of HLA typing, potentially useful in sorting out the complex structure of the genetic variants in this region.

However, due to the complexity of the HLA loci, the large amount of NGS data has not been made informative on HLA genotypes. Many efforts have been made on HLA typing by mining NGS data, including the alignment-based method that relies on counting the number of short reads aligned to each specific allele [8], the assembly- and scoring-based method that takes into account of good quality contigs and their scores for each candidate HLA allele [9]. These methods capitalize on the increasing accessibility and affordability of NGS sequencing and have greatly reduced the time and cost required to make an HLA call comparing to traditional standard PCR-based solutions. Unfortunately, all these methods are only capable of achieving low-digit resolution and perform poorly at higher-digit resolution, which is required for clinical applications.

In this study, we introduce a novel approach for accurate HLA typing at high-digit resolution based on a strategy of comparing sequence reads to a comprehensive reference panel containing all the known HLA alleles for high efficiency mapping, followed by assembly of the mapped reads to contigs, stepwise matching and designation of the contigs to HLA alleles and decision on HLA allele calling. Testing of the method on a set of public and internal whole exome sequencing (WES) data demonstrated that this new method is capable of reporting HLA alleles at a high-digit resolution with great accuracy. These preliminary results highlight the potential applications of this method for HLA calling from NGS data, which may have significant implications in many important clinical aspects. A toolkit (HLAreporter) was developed to facilitate the use of this method for HLA typing from NGS data, which can be downloaded from http://paed.hku.hk/genome/.

Yazhi Huang is with the Department of Paediatrics and Adolescent Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, 21 Sassoon Road, Hong Kong (corresponding author to provide phone: 00852-62850139; e-mail: bill0@ connect.hku.hk).

Jing Yang, Dingge Ying, Yan Zhang, Pak Chung Sham, Yu Lung Lau, and Wanling Yang are with The University of Hong Kong, Hong Kong.

Vorasuk Shotelersuk and Nattiya Hirankarn are with the Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand.

TABLE I
EXOME SEQUENCING DATA WITH 601 AND 110 KNOWN HLA ALLELES FOR HAPMAP AND 1000 GENOMES SAMPLES RESPECTIVELY

Sample	*Run ID*	*Alternat. Run* [a]	*HLA-A* [b]	*HLA-B*	*HLA-C*	*HLA-DRB1*	*HLA-DQB1*	*HLA-DQA1*
NA18502	SRR764722	SRR764723	2424 7401	1403 5801	0701 0802	1301 0701	0501 0201	0102 0201
NA18505	SRR716648	SRR716649	2601 7401	1503 5301	0202 0701	1503 0301	0602 0201	0102 0501
NA18507	SRR764745	SRR764746	2301 3001	1503 4201	0202 1701	1302 0804	0605 0301	0102 0501
NA18956	SRR766028	-	0201 0201	1501 4002	0303 0304	1502 0901	0601 0303	0103 0301
NA18978	SRR716650	-	0301 3101	4002 4402	0501 1402	1301 0403	0603 0302	0103 0301
NA18994	SRR716431	-	2402 3101	0702 4002	0304 0702	0101 0802	0501 0302	0101 0401
NA18992	SRR716428	-	2402 3303	4403 5201	1202 1403	1302 1502	0604 0601	0102 0103
NA18997	SRR702078	-	2402 2603	1501 3901	0303 0702	0406 -	0302 0303	0301 0301
NA19005	SRR715906	-	0201 3303	5801 6701	0302 0702	1501 1302	0602 0605	0102 0102
NA18508	SRR716637	SRR716638	3303 6802	4201 5301	0401 1701	1303 0302	0201 0402	0201 0401
NA19099	SRR748771	SRR748772	2301 3601	0702 0702	0702 1505	1503 0901	0502 0201	0102 0301
NA18995	SRR764775	-	2402 3303	1507 4403	0303 1403	1302 1101	0604 0301	0102 0501
NA19222	SRR748214	SRR177280	3001 3303	4202 4501	1601 1701	1303 0403	0201 0302	0201 0301
NA18965	SRR764771	SRR764772	2601 3101	4006 5601	0401 0801	0901 0901	0303 0303	0301 0301
NA19137	SRR792560	SRR792542	0201 3001	0702 4501	0702 1601	1503 1102	0602 0301	0102 0501
NA19239	SRR792159	SRR792097	0201 6802	3501 5201	0401 1601	1301 1201	0501 0301	0103 0501
NA19240	SRR792767	SRR792091	3001 6802	3501 5703	0401 1801	1602 1201	0502 0301	0102 0501
NA19238	SRR792121	SRR792165	3001 3601	5301 5703	0401 1801	1602 1101	0502 0602	0102 0102
NA18971	SRR077447	SRR078842	0206 2402	4002 4002	0304 0702	1401 0901	0503 0303	0101 0301
NA18968	SRR077480	SRR081231	1101 3101	4001 5101	0702 1402	15xx 1403	0602 0301	0102 0501
NA18975	SRR078849	SRR081225	0201 0206	0702 1501	0401 0702	0101 0406	0501 0302	0101 0301
NA18981	SRR077477	SRR077751	0201 3101	5101 5101	1402 1402	0802 1501	0402 0301	0401 0501
NA19152	SRR071135	SRR071167	0301 2601	1510 5601	0102 0804	1104 1401	0502 0503	0102 0101
NA19131	SRR070494	SRR070783	6602 6802	4201 5301	0401 1701	0302 0804	0402 0301	0401 0501
NA12144	SRR766058	-	0301 0201	3501 4402	0401 0704	0101 0407	0501 0301	0101 0301
NA07000	SRR766039	-	0201 6801	4402 4001	0304 0704	0301 1101	0201 0301	0501 0501
NA06985	SRR709972	-	0301 0201	0702 5701	0702 0602	1501 1501	0602 0602	0102 0102
NA10851	SRR766044	-	2402 0101	4001 0801	0304 0701	0404 0701	0302 0303	0301 0201
NA07357	SRR764689	SRR764690	2402 0101	3906 0801	0702 0701	0404 0301	0302 0201	0301 0501
NA12044	SRR766060	-	0201 0101	0702 0702	0702 0702	1301 1501	0603 0602	0103 0102
NA12043	SRR716423	SRR716424	0206 2601	3501 3801	0401 1203	0101 0404	0501 0302	0101 0301
NA11881	SRR766021	-	2601 0301	0702 0702	0702 0702	1501 1501	0602 0602	0102 0102
NA11829	SRR710128	-	0201 0201	4402 1501	0501 0304	0401 0401	0301 0302	0301 0301
NA11832	SRR766003	-	3201 0201	4002 2705	0202 0704	1301 1501	0603 0602	0103 0102
NA11830	SRR766026	-	0201 0201	1402 1401	0802 0802	1303 07xx	0301 0201	0501 0201
NA11831	SRR709975	-	0101 0301	0801 0702	0701 0702	0301 1501	0201 0602	0501 0102
NA11992	SRR701474	-	0201 0101	3501 0801	0401 0701	0101 0301	0501 0201	0101 0501
NA11995	SRR766010	-	0101 0101	5701 0801	0602 0701	1301 1501	0603 0602	0102 0103
NA11994	SRR701475	-	0101 1101	5101 0702	1502 0702	0402 0404	0302 0302	0301 0301
NA12234	SRR716435	-	3201 1101	4002 4403	1502 1601	0701 1404	0201 0503	0201 0101
NA12156	SRR764691	-	0101 1101	5101 5001	1502 0602	0407 0701	0301 0201	0301 0201
NA12154	SRR702067	-	0101 3101	0801 4001	0701 0304	0301 0404	0201 0302	0501 0301
NA12155	SRR702068	-	0101 0201	0801 4402	0701 0501	0301 0401	0201 0301	0501 0301
NA12005	SRR718067	-	2902 0201	0702 2705	0702 0102	1501 1501	0602 0602	0102 0102
NA12006	SRR716422	-	2501 1101	1801 1501	1203 0303	1501 0404	0302 0602	0102 0301
NA12750	SRR794547	SRR794550	3101 0201	1501 0702	0303 0702	1501 0404	0602 0302	0102 0301
NA12814	SRR715914	-	0301 0201	0702 0702	0702 0702	- -	0501 0602	0101 0102
NA12813	SRR718077	SRR718078	2902 2402	4403 5701	1601 0602	- -	0303 0201	0201 0201
NA12812	SRR715913	-	0201 0201	4402 4402	0501 0501	- -	0603 0301	0103 0301
NA12815	SRR716646	-	0101 2402	0801 5501	0701 0702	- -	0603 0301	0103 0301
NA12873	SRR702070	-	2402 0301	3905 0702	0702 0702	- -	0602 0602	0102 0102
Sample	*Run ID*	*Platform*	*Instrument*	*HLA-A*	*HLA-B*	*HLA-C*	*HLA-DRB1*	*HLA-DQB1*
HG01756	SRR359102	V2(SureSelect)	GAIIx	*30:02; *66:01	*18:01; *41:02	*05:01; *17:01	*30:01	*02:01
HG01757	SRR359103	V2(SureSelect)	GAIIx	*01:01; *02:01	*18:01; *57:01	*07:01	*03:01; *07:01	*02:01; *03:03
NA20313	SRR359108	V2(SureSelect)	GAIIx	*03:01; *68:02	*35:01; *53:01	*04:01	*04:05; *08:04	*03:01; *03:02
HG01872	SRR359298	V2(SureSelect)	GAIIx	*11:02; *24:07	*27:04; *39:05	*08:01; *12:02	*08:03; *12:02	*03:01; *06:01
HG01873	SRR359295	V2(SureSelect)	GAIIx	*02:03; *03:01	*35:03; *55:02	*04:01; *12:03	*08:02; *14:05	*04:02; *05:03
HG01886	SRR360655	V2(SureSelect)	GAIIx	*30:02; *74:01	*15:03; *57:03	*02:10; *07:01	*11:01; *13:02	*05:02; *06:09
HG01953	SRR360288	V2(SureSelect)	GAIIx	*02:01; *02:11	*15:04; *35:05	*01:02; *04:01	*04:11; *09:01	*03:02; *03:03
HG01968	SRR360391	V2(SureSelect)	GAIIx	*02:01; *68:01	*07:02; *40:02	*03:04; *07:02	*01:03; *09:01	*03:03; *05:01
HG02014	SRR360148	V2(SureSelect)	GAIIx	*02:01; *36:01	*35:01; *40:01	*03:04; *04:01	*01:01; *15:01	*05:01; *06:02
HG02057	SRR359301	V2(SureSelect)	GAIIx	*02:03; *31:01	*13:01; *48:01	*03:03; *03:04	*11:01; *13:12	*03:01
NA20313	SRR359098	V2(SureSelect)	GAIIx	*03:01; *68:02	*35:01; *53:01	*04:01	*04:05; *08:04	*03:01; *03:02

[a]: Samples listed in this column were integrated into their corresponding run (column "Run ID") in an attempt to enhance the data quality for HLA typing. They were released in the same year with their corresponding run (column "Run ID"), with the same read length. [b]: The HLA types of these genes can be found at [1] , [12] and [13].

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

II. Materials and Methods

A. Classification of Short Reads to HLA Genes through Mapping Using a Comprehensive Reference Panel (CRP)

In order to achieve an accurate typing, the first essential step is to accurately classify the short sequencing reads into specific HLA genes. Many of the short sequencing reads from HLA genes are not mapped properly or labelled as unmapped for most NGS data processing procedures due to great sequence differences among HLA alleles of the same gene and sequence similarity between paralogous genes. Recognizing this, we designed a comprehensive reference panel (CRP) for collecting reads corresponding to HLA genes. Allele differences were fully accounted for during mapping by adopting all the known HLA alleles in IMGT/HLA as references, which ensured a complete capture of the HLA reads for further analysis [3], [10].

Mapping was performed using Burrows-Wheeler Aligner (BWA) 0.6.1 [11] using default parameters, using all the raw reads from fastq file for NGS data against the reference sequences in CRP. We mainly considered a gene's polymorphic exons for HLA typing in this work (exon2, 3, 4 for Class-I genes HLA-A, HLA-B, HLA-C and exon2 and 3 for the Class-II genes). In order to capture reads containing partial intron sequences, a comprehensive panel of references was designed by appending 50 base pairs of intron sequences extracted from IMGT/HLA to both ends of a reference exon. For the alleles without intron sequence data in IMGT/HLA, corresponding intron sequences from other alleles of the gene were used to include intron sequences at the junctions. As a result, the panel is capable of capturing short reads partly falling out of the targeted exons by as many as 50 bp, for both WES data and whole genome sequencing (WGS) data. In addition to the targeted genes that we aimed to type (i.e. HLA-A, -B, -C, -DRB1, -DQB1, -DQA1, -DPB1, -DRB3, 4, 5), we also included all known allelic sequences of a set of minor HLA genes (i.e. HLA-E, -F, -G, -H, -J, -K, -L, -V, -P, -DMA, -DMB, -DOA, -DOB, -DPA1, -DRA) in the CRP panel. These genes serve as "mapping competitors" to ensure accurate mapping of the sequencing reads. After classification, short reads mapped to a specific gene were grouped together and collected for further analysis (Fig. 1).

We excluded those ambiguous reads where a short read captured by one gene on the mapping panel could also be perfectly mapped to another gene. Reads with equal mapping score towards multiple genes but with imperfect sequence alignment were retained for further analysis since this level of similarity is expected among different HLA genes. The short reads mapped to a particular HLA gene were assembled respectively using de novo assembly. An assembler called TASR [9] was used here for the de novo assembly (for detailed description of TASR algorithm see Warren, et al., 2012 [9]). During this process, only reads with 100% match in the overlapped region were assembled. On average, 30% of short reads with mismatches would be excluded based on the NGS data used in this study, eliminating potential effects of sequencing error on assembly.

B. Design of Reference Database on HLA Alleles for Matching Assembled Contigs

We designed two reference databases on HLA alleles for matching the assembled contigs to the corresponding HLA alleles, one with sequences for major polymorphic exons (exon 2 and 3 for Class-I genes and exon2 for Class-II genes) and one with additional sequences for minor polymorphic exons (exon 4 for Class-I genes and exon 3 for the Class-II genes).The first reference database was designed as the major database (mDB) that contains sequence information of all known alleles on the major polymorphic exons from IMGT/HLA and this database will be queried first for each assembled contig (exon 2 and 3 for Class-I genes and exon2 for Class-II genes). When multiple HLA alleles have identical nucleotide sequences across the major exons, an upper case 'G' will be appended to their names as suffix based on HLA nomenclature (http://hla.alleles.org/alleles/g_groups.html) and they will be further examined. Unlike CRP panel that was designed for mapping of the sequencing reads that may contain intron sequences, the reference database here does not contain intron extensions, and no "competitor sequences" are included.

The second reference database was designated as the additional database (aDB) that recorded the minor polymorphic exon sequences for all the 'G' group members. When a specific allele could not be designated via mDB, the contigs corresponding to minor exon sequences will be examined by aligning them to the candidate alleles in aDB. We separated the minor exons from the major ones since a number of alleles in IMGT/HLA do not have information on the minor exons. Meanwhile, this also enhances the efficiency of the analysis on the assembled contigs.

C. A Stepwise HLA Typing Process

The assembled contigs are then matched to the sequences in the two databases, mDB and aDB, sequentially. To guarantee accuracy, a stringent standard was adopted during the contig-HLA allele matching process, where only a perfect match to the candidate alleles on the exonic regions is considered. Contigs supported by an average read coverage depth less than five folds were also excluded since contigs with lower depth are less reliable. All contigs with different lengths supported by an average read depth of five folds or above were considered in cases that not enough long contigs could be generated by the assembler for certain HLA alleles. By analyzing all the assembled contigs using a scoring system based on the length of the contig in the exonic region and the coverage depth, the candidate HLA alleles that match those assembled contigs are assigned (for the scoring system and the assignment algorithm see [9]). Briefly, the score of a contig is the product of the contig size (bp), the average coverage depth of the contig, and percentage of the contig's exonic sequence. Accordingly, the score of an allele supported by multiple contigs (e.g. exon 2 and 3 for Class-I genes) is the sum of the contig scores. Each candidate allele is measured by the corresponding contig scores and is sorted according to the score value in a descending order. Based on the sorted scores, HLA alleles are reported whenever a uniquely matched contig that

has not already been assigned is detected. The overall flow of this technique is presented in Fig. 1, where classification of reads to a specific gene using multiple reference-based mapping is shown in stage 1, 2, and assembly, contig-HLA assignment are shown in stage 3, 4, and 5.

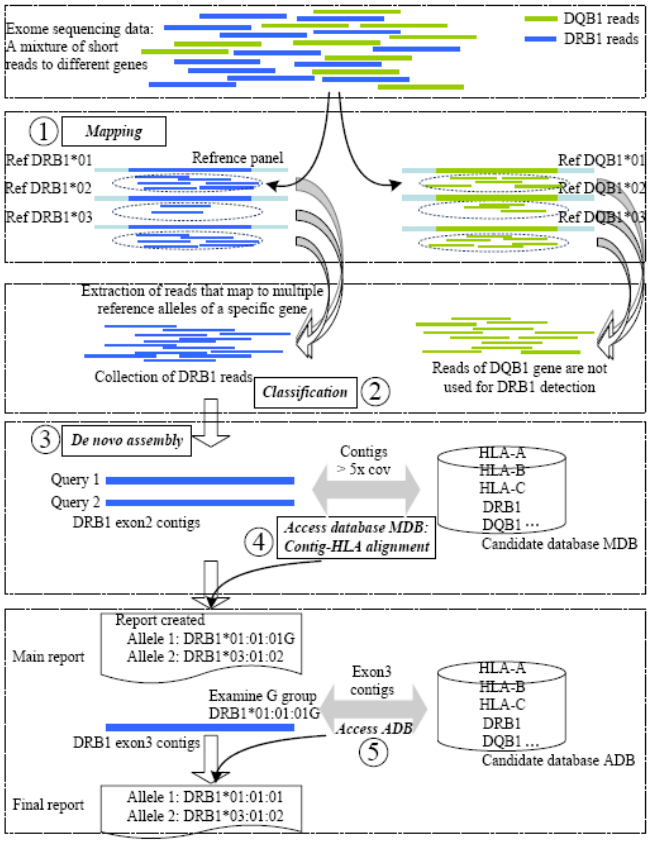

Fig. 1 HLAreporter detection flow with HLA-DRB1 gene as an example

Since we only allowed for perfect matches between assembled contigs and the HLA alleles in the database, any assembled contigs with mismatches were not considered at this step. For the purpose of novel allele detection, HLAreporter would document all the assembled contigs for the main exons, and report these contigs and their quality to the users, so that further analysis of them could reveal novel alleles not existing in the database.

D.Application of HLAreporter on WES Data

Data from 82 samples with a total of 791 verified HLA alleles using other experimental methods were selected to test the performance of HLAreporter, including 62 publicly available samples [1], [12]-[14] and 20 internal samples from Thai population. WES data in fastq format were downloaded from the 1000 Genomes Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/data/) and the HapMap Project (http://hapmap.ncbi.nlm.nih.gov/). Of the 62 public samples, the 11 samples from 1000 Genomes Project were generated using Illumina GAIIx and allelic HLA typing was performed using SeCore HLA Sequencing Reagents. The 51 HapMap samples were released in 2013 by Baylor College of Medicine (BCM) and Washington University Genome Sequencing Center (WUGSC). We selected these samples because they have publicly available calls on HLA Class-I and Class-II genes and relatively longer read length and higher coverage depth, a feature of the more recently released samples. The details of the WES data including 62 publicly available samples with 711 verified HLA types can be found in Table I.

Using NGS data for HLA typing, there could be phasing ambiguities [15]. For those HLA genes with phasing issues, alleles with higher frequency in the public database were assigned over those with lower population frequencies. For example, allele pairs (A*02:03:01G; A*31:01:02G) and (A*02:152; A*unknown-allele) could both explain the observed genotypes for an individual due to phasing ambiguity. Since A*02:152 has a much lower frequency than A*02:03:01G and there is also an unknown allele with 0 reported population frequency in IMGT/HLA, allele pair (A*02:03:01G; A*31:01:02G) would be assigned here.

III. Results

A. Mapping Efficiency

In this method, a comprehensive reference panel CRP was used for short read collection for HLA genes. To study the mapping efficiency for this group of genes, we compared the number of reads captured for HLA-DRB1 gene between using a traditional single reference-based mapping method and the CRP panel-based mapping developed in this study, using BWA as the mapping tool in both cases.

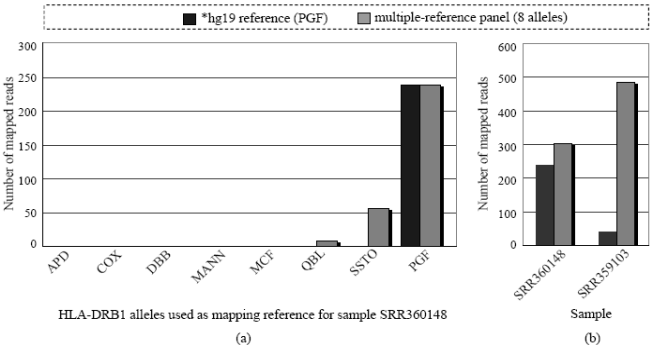

Fig. 2 Mapping efficiency for HLA-DRB1 genes. (a) Difference in the number of reads captured on exon2 region for sample SRR360148. (b) The total number of reads captured on exon2 region

Fig. 2 shows the difference in mapping efficiency between a single reference and multiple reference-based mapping. Sample SRR360148 is with allele DRB1*01:01 and DRB1*15:01. Sample SRR359103 is with allele DRB1*03:01 and DRB1*07:01.As can be seen in Fig. 2 (a), comparing to using a single reference hg19 only during mapping, sample SRR360148, whose DRB1 alleles are *01:01 and *15:01, was more effectively mapped by a reference panel with multiple alleles corresponding to the eight haplotypes in hg38, namely APD, COX, DDB, MANN, MCF, QBL, SSTO, and PGF (the allele for hg19). The short reads matching the two DRB1 alleles

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

*01:01 and *15:01 were better captured by the multiple reference panel, even though allele *01:01 was still underrepresented by the eight haplotype panel and far fewer short reads were captured compared to those corresponding to allele PGF (*15:01). Clearly, using a single allele hg19 *15:01:01:01 as the reference can only capture those short reads similar to itself, which would result in a loss of short reads belonging to allele *01:01 and an incorrect DRB1 designation.

Fig. 2 (b) summarizes the total number of reads captured for samples SRR360148 and SRR359103 using hg19 *15:01:01:01 as reference only versus a multiple allele based reference panel. Using a single allele as the mapping reference, as done by most WES processing tools, would lose quite a number of short reads and likely result in incorrect HLA typing. For SRR360148, the number of short reads that a single reference can capture only amounts for about 75% of that captured by the multiple allele reference panel. The differences were much greater for sample SRR359103 who has allele DRB1*03:01 and *07:01 for this gene, where majority of sequence reads were not captured using reference hg19 *15:01:01:01. This resulted in an HLA detection failure for this sample when we used those reads for HLA typing, emphasizing the deficiency of traditional mapping approaches for HLA detection.

TABLE II
HLA PREDICTIONS OF CLASS-I AND CLASS-II GENES

	HLAminer	*HLAreporter*	*HLAminer*	*HLAreporter*	*HLAminer*	*HLAreporter*
SRR	*HLA-A*	*HLA-A*	*HLA-B*	*HLA-B*	*HLA-C*	*HLA-C*
359102	*30:01; *30:02	*30:02:01G	*15:83; *18:01; *18:26	*18:01:01G	*05:01 *17:01	*05:01:01G
	*30:04; *66:01	*66:01:01G	*41:01; *45:01; *50:01	*41:02:01	--	*17:01:01G
359103	*01:01; *01:03; *02:01	*01:01:01G	*18:01; *18:03; *57:01	*18:01:01G	*07:01	*07:01:01G
	*02:03; *11:02; *68:08	*02:01:01G	--	*57:01:01G	--	--
359108	--	*03:01:01G	--	*35:01:01G	*04:01	*04:01:01G
	--	*68:02:01G [P]	--	*53:01:01	--	--
359298	*11:01; *11:02; *11:50	*11:02:01G	*27:04; *27:25; *39:34	*27:04:01G	*08:01; *08:21; *12:02	*08:01:01G
	*24:02; *24:07; *24:20 [a]	*24:07 [P]	*40:02; *40:06	*39:05:01	*12:03	*12:02:01G [P]
359295	*02:03; *03:01	*02:03:01G	*35:01; *35:03; *37:01	*35:03:01G	*01:02; *04:01; *04:03	*04:01:01G
	--	*03:01:01G [P]	*55:02; *55:48; *56:01	*55:02:01G [P]	*12:03; *15:02; *15:16 [a]	*12:03:01G
360655	*30:01; *30:02; *30:04	*30:02:01G	*15:03; *57:01; *57:06	*15:03:01G	*02:02; *02:11; *07:01	*02:10
	*32:01; *74:01; *74:11	*74:01:01G	*57:11	*57:03:01 [P]	--	*07:01:01G [P]
360288	*02:01	*02:01:01G	*15:01; *15:07; *15:32	*15:04	*01:02; *04:01; *04:03	*01:02:01G
	--	*02:11:01G	*35:14; *58:01	*35:05:01	*04:06	*04:01:01G
360391	*02:01; *02:48	*02:01:01G	*07:02; *40:02; *40:06	*07:02:01G	*03:03; *03:04; *07:02	*03:04:01G
	*68:01	*68:01:02G [P]	--	*40:02:01G [P]	--	*07:02:01G [P]
360148	*01:01; *02:01	*02:01:01G	*07:02; *35:01; *35:41	*35:01:01G	*03:02; *03:04; *04:01	*03:04:01G
	*36:01	*36:01	*40:01; *40:79; *53:01 [a]	*40:01:01G [P]	*04:03; *15:02; *15:17 [a]	*04:01:01G
359301	*02:03; *11:02; *31:01	*02:03:01G	*13:01; *48:01	*13:01:01G	*03:03; *03:04	*03:03:01G
	*32:01; *74:01; *74:11	*31:01:02G [P]	--	*48:01:01G	--	*03:04:01G
359098	--	*03:01:01G	--	*35:01:01G	*04:01	*04:01:01G
	--	*68:02:01G [P]	--	*53:01:01	--	--
	HLAminer	*HLAreporter*		*HLAminer*	*HLAreporter*	
SRR	*HLA-DRB1*	*DRB1(exon2)*	*DRB1 (exon2&3)* [b]	*HLA-DQB1*	*DQB1(exon2)*	*DQB1 (exon2&3)* [b]
359102	*03:01; *07:01	*03:01:01G	*03:01:01	*02:01	*02:01:01G	*02:01:01
	--	--	--	--	--	--
359103	*03:01; *07:01	*03:01:01G	*03:01:01	*02:01; *03:03	*02:01:01G	*02:01:01
	--	*07:01:01G	*07:01:01	--	*03:03:02G	*03:03:02
359108	--	*04:05:01	*04:05:01	--	*03:01:01G	*03:01:04
	--	*08:04:01	*08:04:01	--	*03:02:01G	*03:02:01
359298	*04:03; *08:03; *12:01	*08:03:02	*08:03:02	*03:01; *06:01	*03:01:01G	*03:01:01
	*14:54 [a]	*12:02:01	*12:02:01	--	*06:01:01G	*06:01:01
359295	*04:03; *07:01; *08:03	*08:02:01	*08:02:01	*03:02; *03:03	*04:02:01	*04:02:01
	*14:05	*14:05:01	*14:05:01	*03:05; *05:03	*05:03:01G	*05:03:01
360655	*07:01; *08:03; *11:01	*11:01:02	*11:01:02	*05:01; *05:03	*05:02:01G	*05:02:01
	*13:02	*13:02:01	*13:02:01	*06:09	*06:09:01	*06:09:01
360288	*04:03; *07:01	*04:11:01	*04:11:01	*03:02	*03:02:01G	*03:02:01
	--	*09:01:02	*09:01:02	--	*03:03:02G	*03:03:02
360391	*01:01; *07:01	*01:03	*01:03	*03:03; *05:01	*03:03:02G	*03:03:02
	--	*09:01:02	*09:01:02	--	*05:01:01G	*05:01:01
360148	*01:01; *01:02; *07:01	*01:01:01G	*01:01:01	*05:01; *06:02	*06:02:01G	*06:02:01
	*15:01	*15:01:01G	*15:01:01	--	*05:01:01G	*05:01:01
359301	*07:01; *08:03; *11:01	*11:01:01G	*11:01:01	*03:01	*03:01:01G	*03:01:01
	--	*13:12:01	*13:12:01	--	--	--
359098	--	*04:05:01	*04:05:01	--	*03:01:01G	*03:01:04
	--	*08:04:01	*08:04:01	--	*03:02:01G	*03:02:01

[a]Additional ambiguity at 4-digit resolution was not shown; [b]: All alleles with identical exon2 & 3 sequences will be reported. For example, after examining exon2 & 3 of allele DQB1*03:03:02G, HLAreporter will report alleles 03:03:02:01/03:03:02:02/03:03:02:03. Since the last 2 digits out of 8 digit-based HLA nomenclature are determined by intronic sequences, we only presented the first 6 digits *03:03:02 in the table after examining the minor exon; [P]: Phase was reported.

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

B. Predictions of HLA Class-II Genes

Predictions of HLA Class-I and Class-II genes for the 11 publicly available 1000 Genomes samples with adequate NGS data are presented in Table II. Results on 51 additional HapMap samples are shown in Table III (Data not shown for 20 internal Thai samples). Predictions made by HLAminer [9] using de novo assembly are also shown as a comparison. The columns "HLA-DQB1(exon2)" and "HLA-DQB1 (exon2&3)" represent predictions by examining exon2 only and predictions after further examining exon3, respectively.

TABLE III
HLA PREDICTIONS OF CLASS-I AND -II GENES FOR 51 HAPMAP SAMPLES

SRR[hmp]	*HLA-B[b]*	*HLA-DRB1*	*HLA-DQB1*	*HLA-DQA1*
764722	14:03 58:01:01G	**13:01:01G 07:01:01G**	05:01:01G 02:01:01G	**01:02:01G 02:01**
716648		**15:03:01G 03:01:01G**	**06:02:01G 02:01:01G**	**01:02:01G 05:01:01G**
764745		**13:02:01 08:04:01**	*06:09:01 03:01:01G[a]*	*01:02:01G 05:01:01G*
766028		*15:02:01 09:01:02*	**06:01:01G 03:03:02G**	**01:03:01G 03:01:01G**
716650		13:01:01G 04:03:01	**06:03:01G 03:02:01G**	**01:03:01G 03:01:01G**
716431		**01:01:01G 08:02:01**	*03:02:01G 03:02:01G[e]*	01:01:01G04:01:01G
716428		**13:02:01 15:02:01**	06:04:01G 06:01:01G	**01:02:01G 01:03:01G**
702078		**04:06:01G 09:01:02**		
715906		**15:01:01G 13:02:01**	**06:02:01G 06:09:01**	**01:02:01G 01:03:01G**
716637		**13:03:01 03:02:01**	02:01:01G 04:02:01	**02:01 04:01:01G**
748771	**07:02:01G 07:02:01G**	**15:03:01G 09:01:02**	**05:02:01G 02:01:01G**	**01:02:01G 03:01:01G**
764775		*11:01:01G 13:23:02[l]*	*06:04:01G 03:01:01G*	*01:02:01G 05:01:01G*
748214	42:02 45:01:01G [p]	13:03:0104:03:01		*02:01 03:01:01G*
764771		**09:01:02 09:01:02**		*03:01:01G 03:01:01G*
792560	**07:02:01G 45:01:01G [p]**	**15:03:01G 11:02:01**	**06:02:01G 03:01:01G**	*01:02:01G 05:01:01G*
792159		13:01:01G 12:01:01G	**05:01:01G 03:01:01G**	
792767		*16:02:01 12:01:01G*	*05:02:01G 03:01:01G*	
792121	53:01:03 57:03:01	**16:02:01 11:01:02**	**05:02:01G 06:02:01G**	
077447				**01:01:01G 03:01:01G**
077480		**15:01:01G 14:03:01**	**06:02:01G 03:01:01G**	01:02:01G 05:01:01G
078849		**01:01:01G 04:06:01G**	05:01:01G 03:02:01G	**01:01:01G 03:01:01G**
077477		08:02:01 15:01:01G [a]	*04:02:01 03:01:01G*	**04:01:01G 05:01:01G**
071135		11:04:02 14:01:01G	**05:02:01G 05:03:01G**	01:02:01G 01:01:01G
070494		03:02:01 08:04:01	*04:02:01 03:01:01G[a]*	**04:01:01G 05:01:01G**
766058	35:01:01G 44:02:01G [p]	**01:01:01G 04:07:01G[p]**	**05:01:01G 03:01:01G**	**01:01:01G 03:01:01G**
766039	44:02:01G [a] 40:01:01G [a]	03:01:01G 11:01:01G	*02:01:01G 03:01:01G*	*05:01:01G 05:01:01G*
709972	07:02:01G 57:01:01G [a]	**15:01:01G 15:01:01G**	**06:02:01G 06:02:01G**	**01:02:01G 01:02:01G**
766044	**40:01:01G 08:01:01G[p]**	**04:04:01 07:01:01G**	**03:02:01G 03:03:02G**	
764689		04:04:01 03:01:01G	*02:01:01G 03:02:01G[a]*	
766060	**07:02:01G 07:02:01G**	**13:01:01G 15:01:01G**	**06:03:01G 06:02:01G**	**01:03:01G 01:02:01G**
716423		**01:01:01G 04:04:01[p]**	**05:01:01G03:02:01G**	01:01:01G 03:01:01G
766021	**07:02:01G 07:02:01G**	**15:01:01G 15:01:01G**	**06:02:01G 06:02:01G**	*01:02:01G 01:02:01G*
710128	15:01:01G 44:34:02 [l]	04:01:01 04:01:01	*03:01:01G 03:02:01G*	03:01:01G 03:01:01G
766003	40:02:01G 27:05:02G [p]	**13:01:01G 15:01:01G**	**06:03:01G 06:02:01G**	**01:03:01G 01:02:01G**
766026	14:02:01 14:02:01 [l]	13:03:01 07:01:01G	*03:01:01G 02:01:01G*	*02:01 05:01:01G*
709975		**03:01:01G 15:01:01G**	*02:01:01G 06:02:01G*	05:01:01G 01:02:01G
701474		01:01:01G 03:01:01G	05:01:01G 02:01:01G	01:01:01G 05:01:01G
766010		**13:01:01G 15:01:01G**	*06:03:01G 06:02:01G*	01:02:01G 01:03:01G
701475	**51:01:01G 07:02:01G**	**04:02:0104:04:01**	*03:02:01G 03:02:01G*	03:01:01G 03:01:01G
716435		*07:01:01G 07:01:01G[e]*		02:01 01:01:01G
764691		04:07:01G 07:01:01G	*03:01:01G 02:01:01G*	
702067	**08:01:01G 40:01:01G[p]**	**03:01:01G04:04:01**	*02:01:01G 03:02:01G*	05:01:01G 03:01:01G
702068	08:01:01G 44:02:01G [a]	*04:01:01 04:01:01[e]*	*02:01:01G 03:01:01G[a]*	05:01:01G 03:01:01G
718067	07:02:01G 27:05:02G	**15:01:01G 15:01:01G**	**06:02:01G 06:02:01G**	**01:02:01G 01:02:01G**
716422		**15:01:01G 04:04:01**	**03:02:01G 06:02:01G**	01:02:01G 03:01:01G
794547		**15:01:01G 04:04:01**	06:02:01G 03:02:01G	
715914		- -	*06:02:01G 05:01:01G[a]*	01:01:01G 01:02:01G
718077		- -	**03:03:02G 02:01:01G**	**02:01 02:01**
715913		- -	*06:03:01G 03:01:01G[a]*	**01:03:01G03:01:01G**
716646		- -	*06:03:01G 03:01:01G[a]*	01:03:01G 03:01:01G
702070		- -	06:02:01G 06:02:01G	**01:02:01G 01:02:01G**

[a]: This prediction is ambiguous; [b]: For the sake of space limit, only HLA-B gene is shown as an example for class I genes; [p]: Phase was reported in this sample; [l]: This allele only reached low digit resolution (2-digit); [e]: This allele was not correctly typed at low digit resolution (2-digit); [hmp]: These samples are HapMap samples, whose names are prefixed with "SRR". For all the HapMap samples, BOLD font (e.g. 01:01:01) represents this gene reaches a data quality of ">10x" = 100% and ">20x" >= 98%; NORMAL font (e.g. 01:01:01) represents this gene only reaches a data quality of ">10x" = 100% and ">20x" >= 90%; SLASH font (e.g. 01:01:01) represents this gene only reaches a data quality of ">10x" >= 95%; Blank entry means this gene failed the pre-set quality threshold.

As we can see from Table II, for HLA Class-II genes, our results showed complete consistency with the reported alleles at the 4-digit resolution (The 110 known HLA types are presented in Table I). In contrast, HLAminer mistyped

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

heterozygosity to homozygosity in the case of SRR360288, failed to achieve the 4-digit resolution in the case of SRR360655, and reported an incorrect result even at 2-digit resolution in the case of SRR359295, just to give a few examples. It is observed that for some Class-II alleles, examining exon2 sequences alone would be sufficient (e.g. for DQB1*06:09:01, polymorphism is not currently found in any other exons or introns except exon2). While for the alleles with identical exon2 sequences but differences in other exons, further examination of exon3 region would be necessary.

C. Typing Results on HLA Class-I Genes

While accurate predictions without ambiguity were achieved for HLA Class-II genes in all the samples checked, HLA typing of Class-I alleles appeared to be more affected by phase ambiguity. Generally, phase becomes an issue when the size of the non-polymorphic gap between any two alleles is greater than the read length, since different combinations could result in different alleles. We would report this phase ambiguity to users so that further measures can be taken accordingly (Data not shown). We adopted the same measures as in HLAminer by [9], i.e. sensitivity, specificity, and ambiguity as assessment metrics. To summarize, for HLAminer, sensitivity, specificity, and ambiguity on this group of genes were 85%, 88%, and 56%, respectively. Of 66 Class-I alleles tested, only 25 alleles (38%) were accurately reported by HLAminer without ambiguity. In contrast, although with phase issues on ambiguous haplotypes, all predictions made by HLAreporter were completely consistent with the reported HLA alleles at the 4-digit resolution.

Notably, running on our Unix server, HLAminer requires a month for processing a single sample. On the other hand, HLAreporter only requires less than three hours for processing a comparable sample using the same system.

D. Prediction Accuracy

Thus the reliability of this tool was verified by the 110 HLA alleles from 1000 Genomes samples, whose typing was determined through standard PCR-based method. In addition, the 51 HapMap samples and 20 internal samples from Thai population were also tested. The HLA predictions were completely consistent with the experiment-based typing results for all the samples that passed our quality test. Table IV presented the statistics for the HapMap samples with 601 known HLA alleles. With 20 fold coverage depth for 98% of the sequences in the exonic region, we achieved 100% typing accuracy at 4-digit resolution for all HLA genes (Table IV, row "10x=100% & 20x>=98%"). When the quality standard was reduced to 90% of the exonic regions with coverage of 20 folds (Table IV row "10x=100% & 20x>=90%"), we still had a desirable performance at 4-digit resolution, particularly for Class-II genes (Accuracy > 99%). Although with certain ambiguity at 4-digit resolution, 2-digit resolution remained at 100%. HLA-DQA1 apparently is the most tolerant gene to low data quality, where it still had 100% accuracy at 4-digit resolution even when the quality standard was reduced to 95% of the regions with only 10 fold coverage depth (Table IV). This is probably due to the least polymorphic nature of this HLA gene. HLA-DRB1 and -DQB1 failed to achieve 100% accuracy at 2-digit resolution under this lower coverage depth, where three genes were mistyped as homozygous since one of the two alleles were missed for each case (Table III).

To summarize, in total, data from 82 samples with 791 known HLA types were tested using HLAreporter. With a 20-fold coverage quality standard for most targeted exonic sequences (i.e. Table IV "10x=100% & 20x>=98%"), all 288 alleles (36% of all the tested alleles) that passed the quality threshold were correctly typed at 4-digit resolution. Using a more lenient quality standard for the less polymorphic Class-II genes (90% of the regions with coverage of 20 folds), of all the 370 alleles (47% of all the tested alleles), only one HLA-DRB1 allele was ambiguous at 4-digit resolution, while the rest 369 alleles were correctly typed. Based on these results, for HLAreporter, calls made based on coverage lower than 20 folds in more than 2% of the exonic sequences will be accompanied with a warning sign for the user to check the quality of the call. Since next generation sequencing is becoming more and more accurate and with higher coverage depth, our algorithm is advantageous with its call accuracy despite its high demand on data coverage and quality.

IV. Discussion

We have shown that our approach is an efficient technique for HLA typing from whole exome sequencing data. The reliability of this tool is verified by testing 791 known HLA Class-I and Class-II alleles from 82 samples, whose HLA alleles were experimentally confirmed. To achieve reliable predictions, a stringent assembly was conducted to form contigs (zero mismatch tolerance), followed by a stringent HLA assignment process to assign alleles (zero mismatch tolerance on exonic regions), processes that would ensure accuracy for HLA calls.

Since the proposed technique relies on de novo assembly, read length is critical for typing accuracy. A shorter read length would worsen the phase issue. Generally, phase becomes an issue when the size of the non-polymorphic gap between any two alleles is greater than the read length (100 bp in our tested 1000 Genomes data), since different combination might result in different alleles. In addition, phase also becomes an issue when sequences between two exons could have different combinations, which has been reported in IMGT/HLA. For example, for HLA alleles C*08:21, C*08:01:01G, C*08:16:01, C*12:02:01G, and C*12:49, there is a 110 bp gap within exon2 and an inter-exon gap between exon2 and exon3, with different combinations specifying different alleles. Apparently this problem cannot be solved by the current sequencing technology using short reads and paired-end information of reads can only partially solve the problem. While Class-II genes seem to have little phase problem, Class-I gene typing is significantly affected by this issue.

TABLE IV
STATISTICS OF HLA TYPING RESULTS FROM 51 HAPMAP SAMPLES

Quality Standard QS [a]	*HLA gene*	*# Total Genes*	*# Pass QS*	*Pass QS (%)*	*4-digit (%)* [b]	*2-digit (%)* [b]
10x=100% & 20x>=98%	HLA-A	51	3	6%	100%	100%
	HLA-B	51	7	14%	100%	100%
	HLA-C	51	4	8%	100%	100%
	HLA-DRB1	46	28	61%	100%	100%
	HLA-DQB1	51	20	39%	100%	100%
	HLA-DQA1	51	20	39%	100%	100%
10x=100% & 20x>=90%	HLA-A	51	18	35%	81%	100%
	HLA-B	51	18	35%	83%	100%
	HLA-C	51	11	22%	100%	100%
	HLA-DRB1	46	40	87%	99%	100%
	HLA-DQB1	51	27	53%	100%	100%
	HLA-DQA1	51	35	69%	100%	100%
10x>=95%	HLA-DRB1	46	45	98%	96%	98%
	HLA-DQB1	51	46	90%	91%	99%
	HLA-DQA1	51	43	84%	100%	100%

[a]: "10x" represents the percentage of locations with coverage depth greater than 10 folds on the targeted exon. Accordingly, "10x>=95%" means the pre-defined percentage (i.e. "10x") is 95% or above. ("20x" has the similar definition); [b]: The 4-digit (2-digit) percentage is equal to the number of HLA calls at 4-digit (2-digit) resolution without ambiguity divided by the total number of alleles.

To achieve a high accuracy, data with good coverage on HLA genes are also essential. Depth test on each exon of the targeted gene is necessary to ensure accurate typing. If sequencing reads were poorly captured during the enrichment process on exonic fragments, it would be unlikely to properly detect HLA alleles. While there is no golden standard, 30-fold depth on every location of the targeted exons is recommended for adequate coverage and HLA calls (Fig. 3). Balanced capture of the different alleles is also important, a process that should be considered during the design of the probes for enriching exonic genomic fragments. In the 82 real samples we tested, it was showed that 20-fold depth on every targeted location could reach 100% accuracy at 4-digit resolution. Lower coverage depth would increase the risk of either failing to make a call or missing one of the two alleles, calling homozygous on a heterozygous genotype.

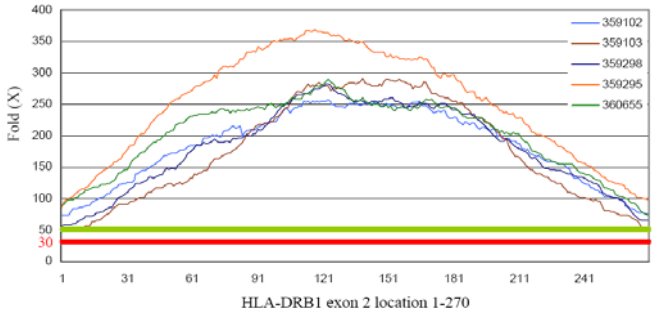

Fig. 3 Depth test for sequencing reads of five 1000 Genomes samples captured by our comprehensive reference panel

The proposed technique focused on the main exon sequences for HLA typing by classification of sequencing reads using a comprehensive reference-based mapping strategy. We have shown that traditional mapping approach using a single sequence as reference is incapable of dealing with the great allelic differences of HLA genes, with a large number of sequencing reads being missed. The comprehensive mapping panel guarantees a full information retrieval from the source data. Classification of sequencing reads to a specific gene first helps avoid the difficulty of de novo assembly using all the sequencing reads. HLAreporter also documents the assembled contigs without 100% match to the candidate alleles in the designated database, so it provides a chance of novel allele detection, making full use of the advantages that the NGS technology can bring.

During the revision of this article, [16] and [14] developed two algorithms for HLA typing from NGS data, respectively. The two up-to-date algorithms are both based on an alignment strategy without performing contig assembly, which aim at typing HLA from different NGS data with distinct read length, region coverage, and coverage depth. Bai's algorithm PHLAT reported 93% accuracy for WES data at 4-digit resolution. We found that a fraction of publicly available WES data they used overlapped with our dataset. Therefore, we checked these overlapped alleles and compared the performance of their method with ours. We found that PHLAT mistyped five alleles out of the 100 alleles, three of which were mistyped at 2-digit resolution. Clearly, HLAreporter outperformed PHLAT with 100% accuracy at 4-digit resolution. Notably, these 100 alleles are covered with high quality reads and high depth (Fig. 3).

Meanwhile, [16]'s algorithm focused on Class-I genes and reported 94% accuracy for WES data at 4-digit resolution. We tried to replicate their experiments using their WES data but unfortunately their data failed our data quality test (Data not shown). This alignment-based algorithm still provides predictions even when the coverage depth is as low as three folds [16]. These data with low coverage depth on HLA region might only be suitable for alignment-based approach instead of the de novo assembly-based approach such as HLAreporter. We simulated 60 samples based on data used in Major et al.'s report by generating 10 folds of error-free short reads for each HLA allele per sample, with an insert size of 250 bp and a read length identical with the original data (30 WES samples and 30 WGS samples), and we correctly predicted all HLA alleles from these samples (Data not shown). The simulation suggested that the proposed method also applies to WGS data. For RNAseq data, however, given that the CRP panel was designed to collect certain short reads with intronic sequences,

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf

it might not properly capture the short reads covering exon/exon junctions during mapping, thus some modification is needed to apply HLAreporter for RNAseq data.

In the current method we proposed, only the main exons of HLA genes are examined. These exons determine the amino acid residues of the peptide binding groove that is important for antigen presentation. Yet, there could be a small number of HLA alleles that share the identical sequences on the main polymorphic exons while exhibiting polymorphism on other exons, such as between Class-I gene C*01:02:01 and C*01:02:11 alleles and between Class-II gene DRB1*12:01:01 and DRB1*12:10 alleles. Relatively, these sequences are less important to the binding specificity of the encoded protein [17]. To reach an even higher resolution, further analysis on additional exons would be needed for these alleles.

With efficient HLA calling, analysis such as HLA allele distribution profiles, haplotype prediction, and disease/drug response association studies could be carried out using available NGS data. In summary, the method introduced here is timely and may help us make full use of NGS data and to better connect the alleles in this region with diseases, drug responses, and transplant rejections.

V. Conclusion

This study presented a novel technique for HLA typing from whole exome sequencing data or other NGS data, capable of accurate typing of HLA alleles at high-digit resolution. Accurate HLA typing from NGS data holds much promise for applications in clinical laboratories and biomedical research. Preliminary analysis on both public and local datasets indicates a great potential for broad application of this method.

Acknowledgment

WY and YLL thank support from Research Grant Council of the Hong Kong Government (GRF 17125114, HKU783813M, HKU 784611M, and HKU 770411M) and S K Yee Medical Foundation general award. YZH is partially supported by Centre for Genomic Sciences of Faculty of Medicine, University of Hong Kong and Hong Kong RGC AoE program on nasopharyngeal cancer for the University of Hong Kong.

References

[1] de Bakker PI, McVean G, Sabeti PC, et al., "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC," *Nat Genet*, vol. 38, 2006, pp. 1166–1172.

[2] de Bakker PI, Raychaudhuri S., "Interrogating the major histocompatibility complex with high-throughput genomics," *Hum Mol Genet*, vol. 21, 2012, pp. 29–36.

[3] IMGT/HLA (International immunogenetics project/Human major histocompatibility complex). London: Royal Free Hospital, Anthony Nolan Research Institute, HLA Informatics Group, 1998. http://www.ebi.ac.uk/ipd/imgt/hla/.

[4] Bentley G, Higuchi R, Hoglund B, et al. High-resolution, high-throughput HLA genotyping by next-generation sequencing. *Tissue Antigens*, vol. 74, 2009, pp. 393–403.

[5] Lind C, Ferriola D, Mackiewicz K, et al., "Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing," *Hum Immunol*, vol. 10, 2010, pp. 1033–1042.

[6] Noble JA, Martin A, Valdes AM, et al., "Type 1 diabetes risk for HLA-DR3 haplotypes depends on genotypic context: Association of DPB1 and HLA class I loci among DR3 and DR4 matched Italian patients and controls," *Hum Immunol*, vol. 69, 2008, pp. 291–300.

[7] Solberg OD, Mack SJ, Lancastera AK, et al, "Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies," *Hum Immunol*, vol. 69, 2008, pp. 443–464.

[8] Boegel S, Lower M, Schafer M, et al, "HLA typing from RNA-Seq sequence reads," *Genome Med*, vol. 4, 2012, pp. 102–113.

[9] Warren RL, Choe G, Freeman D, et al, "Derivation of HLA types from shotgun sequence datasets," *Genome Med*, vol. 4, 2012, pp. 95–102.

[10] Gonzalez-Galarza FF, Christmas S, Middleton D, et al., "Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations," *Nucleic Acids Res*, vol. 39, 2011, pp. 913-919.

[11] Li H, Durbin R, "Fast and accurate long-read alignment with Burrows-Wheeler Transform," *Bioinformatics*, vol. 26, 2010, pp. 589–595.

[12] Inflammgen (The laboratory in genetics and genomic medicine of inflammation). http://www.inflammgen.org/.

[13] Erlich RL, Jia X, Anderson S, et al., "Next-generation sequencing for HLA typing of class I loci," *BMC Genomics*, vol. 12, 2011, pp. 42-54.

[14] Bai Y, Ni M, Cooper B, et al., "Inference of high resolution HLA types using genome-wide RNA or DNA sequencing reads," *BMC Genomics*, vol. 15, 2014, pp. 325-340.

[15] Stephens M, Smith NJ, Donnelly P, "A new statistical method for haplotype reconstruction from population data," *Am J Hum Genet*, vol. 68, 2001, pp. 978–989.

[16] Major E, Rigo K, Hague T, et al., "HLA typing from 1000 genomes whole genome and whole exome illumina data," *PLoS ONE*, vol. 8, 2013, pp. 11–19.

[17] RefSeq (NCBI reference sequence database). Bethesda: National Library of Medicine, National Center for Biotechnology Information, 2002. http://www.ncbi.nlm.nih.gov/refseq/.

Open Science Index, Bioengineering and Life Sciences Vol:9, No:3, 2015 publications.waset.org/10000791.pdf