Journal of Clinical & Cellular Immunology

Gowda et al., J Clin Cell Immunol 2016, 7:2
http://dx.doi.org/10.4172/2155-9899.1000399

Research Article Open Access

Comparative Analyses of Low, Medium and High-Resolution HLA Typing Technologies for Human Populations

Malali Gowda[1,2*], Sheetal Ambardar[1], Nutan Dighe[3], Ashwini Manjunath[1], Chandana Shankaralingu[1], Pradeep Hirannaiah[1], John Harting[4], Swati Ranade[4], Latha Jagannathan[3*] and Sudhir Krishna[5*]

[1]Next Generation Genomics Laboratory, Centre for Cellular and Molecular Platform, National Centre for Biological Sciences, TIFR Bangalore, India

[2]Genomics Discovery Program, TransDisciplinary University, Foundation for Revitalization of Local Health Traditions, Bangalore, India

[3]Bangalore Medical Services Trust, Bangalore, India

[4]Pacific Biosciences, California, USA

[5]National Centre for Biological Sciences, TIFR, Bangalore, India

[*]**Corresponding authors:** Malali Gowda, Next Generation Genomics Laboratory, Centre for Cellular and Molecular Platform, National Centre for Biological Sciences, TIFR Bangalore, India, E-mail: malalig@frlht.org; malalig@ccamp.res.in

Latha Jagannathan, Bangalore Medical Services Trust, Bangalore, India, E-mail: latha@bmstindia.org

Sudhir Krishna, National Centres for Biological Sciences, TIFR, Bangalore, India, E-mail: skrishna@ncbs.res.in

Received date: December 09, 2015; **Accepted date:** March 01, 2016; **Published date:** March 09, 2016

Abstract

Human Leukocyte Antigen (HLA) encoding genes are part of the major histocompatibility complex (MHC) on human chromosome 6. This region is one of the most polymorphic regions in the human genome. Prior knowledge of HLA allelic polymorphisms is clinically important for matching donor and recipient during organ/tissue transplantation. HLA allelic information is also useful in predicting immune responses to various infectious diseases, genetic disorders and autoimmune conditions. India harbors over a billion people and its population is untapped for HLA allelic diversity. In this study, we explored and compared three HLA typing methods for South Indian population, using Sequence-Specific Primers (SSP), NGS (Roche/454) and single- molecule sequencing (PacBio RS II) platforms. Over 1020 DNA samples were typed at low resolution using SSP method to determine the major HLA alleles within the South Indian population. These studies were followed up with medium resolution HLA typing of 80 samples based on exonic sequences on the Roche/454 sequencing system and high-resolution (6-8 digit) typing of 8 samples for HLA alleles of class I genes (HLA-A, B and C) and class II genes (HLA-DRB1 and DQB1) using PacBio RS II platform. The long reads delivered by SMRT technology, covered the full-length class I and class II genes/alleles in contiguous reads including untranslated regions, exons and introns, which provided phased SNP information. We have identified three novel alleles from PacBio data that were verified by Roche 454 sequencing. This is the first case study of HLA typing using second and third generation NGS technologies for an Indian population. The PacBio platform is a promising platform for large-scale HLA typing for establishing an HLA database for the untapped ethnic populations of India.

Keywords: HLA typing; Haplotypes; Sequence-specific primer; Pyrosequencing; Single molecule real-time sequencing

Introduction

The major histocompatibility complex (MHC) region on the short arm of chromosome 6, is one of the most complex regions (4 Mb) in the human genome with extreme levels of polymorphism (>10,000 alleles) [1,2]. The HLA region comprises over 200 genes, but six HLA genes (class I-A, B, C and class II-DR, DP, DQ) that are crucial for self and non-self recognition. Class I proteins expresses on the surface of all nucleated cells in the human body, but class II proteins can be found on the antigen-presenting phagocytes such as dendritic cells, mononuclear phagocytes [3,4]. The MHC class I complex consists of three peptide domains - α1, α2, and α3 and a non-MHC small petptide, β2-microglobulin. The central region of the α1/α2 heterodimer in the HLA complex forms antigen binding groove to presents the antigen(s) to CD8 T lymphocytes, thereby recognize the self and non-self fate of human cells [3,4]. Whereas class II complex consists of two chains, α and β peptides. The class II antigen presenting peptide groove is formed through the interaction of heterodimer of α1 and β1 peptides, which presents the antigen to $CD4^+$ lymphocytes [3,4].

A complex pattern of polymorphism is observed in antigen presenting regions of HLA genes including exon 2 and 3 of class I and exon 2 of class II. Currently over 10,000 HLA class I and II alleles are available at IMGT(the international ImMunoGeneTics) database (http://www.ebi.ac.uk/imgt/hla/). The HLA proteins play a pivotal role in the immune response and are implicated in numerous human pathological conditions including autoimmune disease, infectious diseases, cancer and drug reactions [3,4,5]. Clinically, HLA gene sequence information is widely used in organ transplantion to identify matching donor and recipient HLA alleles. Highly similar alleles improve the organ transplant outcome and reduce the risk of rejection [6].

Many HLA typing laboratories across the globe have adopted SSO (specific oligonucleotides), SSP (sequence specific primers) and Sanger sequencing methods. However, SSO or SSP can only detect known alleles with high accuracy, while Sanger sequencing is unable to

identify phased heterozygous SNPs and it is also expensive and laborious. Recent next generation sequencing (NGS) technologies have increased the HLA typing speed (large multiplexed samples (96 to 384), resolution (6 to 11 genes), typing confidence (high depth of sequence coverage (100 to 500x) and reduced cost per sample (Rs. 10,000 to 5,000) [7,8]. Various NGS platforms including 454, Illumina and Ion Torrent have been explored for HLA typing [7,9,10]. 454 sequencing was one of the first NGS platforms tested for HLA typing. It was considered advantageous due to availability of large-numbers of barcodes and relatively longer reads (upto 1 Kb) [11]. As the HLA genes (A, B, C, DRB1 and DQB1) are more than 5 Kb, the 454 or other short read sequencing was not able to resolve haplotypes of donors. To overcome the phasing limitations across distant SNPs especially those found in class II genes [12], third generation sequencing technology has been utilised where full-length HLA genesare amplified using long range PCR and sequenced using single molecule real-time (SMRT) PacBio sequencing [13]. Pac-Bio sequencing of HLA genes provides high-resolution (6 to 8 digit) HLA allelic information for multiple HLA genes (6 genes) with phased SNPs [13].

The Indian population exhibits a wide variety of ethnic, cultural, geographic and linguistic diversity. Its 1 billion people are categorized into 3824 castes, 25,000 sub-castes and 461 tribes [14]. Thus, India is generally considered a living laboratory for human genetic and genomic studies [15]. Southern India is one of the oldest geophysical regions of the world and home to one of the most primitive Dravidian communities, which is untapped for diversity and immunological studies [16]. Recent studies have revealed over 30 % novel SNPs which were unique to Indian tribal population [17].

Current study was aimed for understanding of HLA alleles in Indian population. Therefore, we typed low resolution HLA alleles for 1020 samples from South Indian population using SSP method. Subsequently we sequenced 80 and 8 samples using second (454/Roch GSF LX+) and third generation NGS (PacBio RS II) platforms, respectively. Our work demonstrated that single molecule sequencing by Pacific Biosciences is an ideal platform to catalogue full-length HLA alleles for diverse populations with few known reference sequences in the IMGT database. Our study is the first of its kind from an Indian population, which will improve the accuracy of HLA match for transplantation and also predict the severity of individual towards diseases and durgs.

Methods

Sample collection

A total of 1020 human genomic DNA samples were studied from a cohort of the renal and bone marrow transplant patients and donors who were typed at from BMST (Bangalore Medical Service Trust) based on requests from the transplant physician and informed consent of patient and donor. BMST had obtained ethical clearance and patient consent information to collect blood from Indian population as per Indian Council Medical Research policy. Genomic DNA was extracted from 4-5 ml of blood from patients and donors using Qiagen kit (Cat. No. 51304) using an automated DNA extraction system, Qiacube (QIAcube 230v/9001293/QIAgen/Germany).

Sequence-specific primer based HLA typing

Genomic DNA of selected clinical samples was checked for the quality and quantity prior to HLA typing. Micro SSP DNA typing kit of One-Lambda (ThermoFisher, USA) was used to type the HLA genes. The 96-wells coated pre-optimized primers for Exon 2 and 3 of HLA class I and exon 2 of HLA-DRB1 of class II and internal control (human β-globin gene) was used for SSP analysis. The amount of each primer was adjusted for optimal amplification using 100ng of human sample DNA. The specific formulated dNTP-buffer mix (Thermofisher, USA) along with recombinant Taq polymerase resulted in amplification of specific exons of the genes. The amplified DNA product is analyzed on 2.5% Agarose gel and interpreted based on presence or absence of a specific amplified DNA fragment. The SSP analysis was supported by One-Lambda software to classify HLA alleles at 2-digits for low to intermediate resolution.

454 library preparation and sequencing

Out of 1020 samples, 80 DNA samples were selected and further checked for quality on 0.7% agarose gel and quantity by Qubit fluorometery. The DNA sample was diluted to 5 ng/µl and used for amplification of HLA alleles. The Roche GS GType primer (HLA) medium resolution kit (Cat. No. 05872529001) which target primers for 8 exons (exon 2 and 3 for HLA class I and exon 2 for class II) was used for HLA alleles amplification (Figure 1). A set of 10 DNA samples were amplified using barcoded primers, amplicons were pooled and purified using AMPureXP beads. The amplicon library was quantified using QuantiT Pico Green ds-DNA assay Kit (Cat. No. P11496) on Qubit Fluorometer. The emulsion PCR was carried out followed by Pyrosequencing [9].

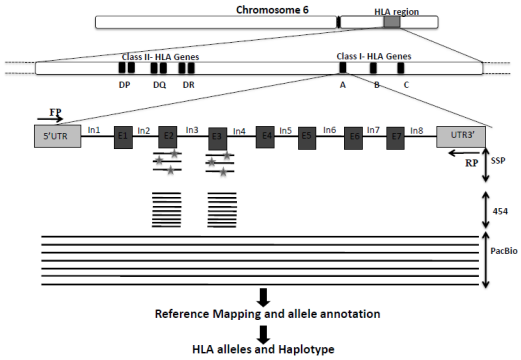

Figure 1: HLA typing strategy of HLA - class I and II genes using SSP and 454, and PacBio methods. The FP and RP represents the forward and reverse primers of PacBio sequencing and star represents the allele specific primer sites used for SSP method and E1 - E7 represents exon from 1 to 7 and In1 - In8 represents intron from 1 to 8.

PacBio library preparation and sequencing

Eight of the 80 samples characterized by 454 sequencing were used to amplify full-length targeted class I HLA-A, B, C, and Exon 2 to 4 of HLA-DRB1 and DQB1 genes using GenDx HLA primers (Cat. No. 2841102) and high-fidelity long range polymerase (Cat. No. 206402) (Figure 1). All amplicons of class I and II genes were purified using AMPurePB beads and quantified using a Qubit Fluorimeter. The size of amplicons was checked on agarose gel and Caliper LabChip GX (Perkin Elmer, USA). The PacBio SMRT library was prepared by pooling PCR amplicons of all HLA genes with equimolar

concentration for each sample. The PacBio library was prepared using the SMRT template preparation kit (Cat. No. 100-259-100) as per manufacturer's protocol. The libraries were validated by Agilent Bio-analyzer using 12K Chip (Cat. No. 5067-1508) and quantified using a Qubit Fluorimeter. All eight DNA libraries were pooled together and mixed with Magnetic beads and loaded onto SMRT Cells for sequencing.

Data analysis and allele calling

454 Pyrosequencing data was demultiplexed based on index sequences and individual sample sequencing reads were analysed using Connexio Genomics ATF software (http://www.conexio-genomics.com). HLA alleles were assigned by aligning the 454 sequencing reads to IMGT-HLA reference database (http://www.imgt.org/). PacBio data was de-multiplexed based on adapter sequence followed by trimming of primer sequences. PacBio sequencing reads were further processed through long SMRT Analysis software version 2.3 to obtain the consensus sequences, and allele calling was done by comparing with nearest IMGT HLA alleles.

Validation of novel alleles

HLA full gene sequences were compared with the IMGT reference database to determine SNP/InDel polymorphisms at introns/exons as well as UTRs of the HLA genes. The variations in exonic sequence of novel HLA alleles were identified using high quality of PacBio reads (≥ QV34 demonstrated). and de novo consensus sequences with high depth of novel allele 1 with 255X (NCBI ID - 1873316), allele 2 with 264X (NCBI ID- 1873302) and allele 3 with 67X (NCBI ID - 1873312). The contiguous full-length single reads were further validated by 454 technologies.

Results

In the present study, we characterized the BMST registry samples using three different HLA typing technologies including SSP/SSOP, 454 Roche (second generation sequencing) and PacBio sequencing (third-generation sequencing).

Major HLA alleles typing using SSP method

In the present study, 1020 human samples of Indian origin were selected from the BMST registry of solid organ transplant patients and donors for SSP analysis. This cohort consists of samples from various regions (state of origin), religion, language and caste. About 75% of BMST registry samples consist of South Indian population (Figure 2). These samples were typed by SSP (2- digit) HLA typing method (Supplemental Table 1).

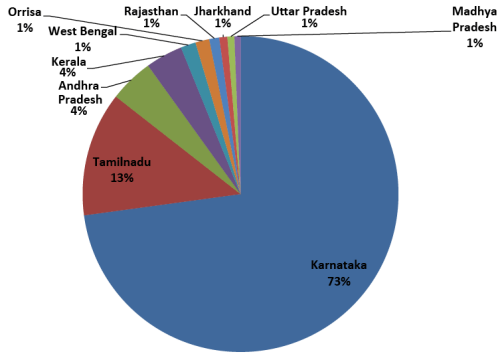

Figure 2: Population diversity of South Indian HLA registry.

SSP analysis revealed the predominance of HLA-A*02, A*24; HLAB* 35, B*40 and HLA-DRB1*15; DRB1*04 alleles in the southern population. Whereas HLA- A*02, A*24; HLA-B*35, B*40 and HLA-DRB1*14; DRB1*15 alleles were the most frequently occurring alleles in North Indian population. Nearly 70% of SSP based HLA typed samples were from Karnataka where the dominant alleles were HLA-A*02, A*24; HLA-B*35, B*40, and HLA-DRB1*15; DRB1*07. We identified three major castes including Vokkaliga (24.5%), Lingayat (16.9%), and Brahmin (12.4%) having different HLA allelic frequency (Table 1) in the BMST's South Indian registry.

High-throughput HLA typing using 454 sequencing

Based on SSP analysis, we selected most commonly occuring HLA allelic samples for 454 analyses. A cohort of 80 renal transplant patient and donor samples from BMST registry were selected for 454 Pyrosequencing. The antigen presenting regions of exon 2 and 3 for HLA-A, B and C of class I and exon 2 of class II for HLA-DRB1 and DQB1 genes were amplified using the medium resolution HLA typing kit. The analysis of multiple HLA loci for multiple samples in a single 454 run was facilitated by the incorporation of barcode identifier tags in the PCR fusion primers. Approximately 2000 reads per sample were obtained with the median read length of 320 nt (Figure 3a).

The 454 data provided over 100x coverage for each of HLA loci. About 85% of alleles had shown 100% similarity to their respective alleles in reference database. We identified HLA-A*24:02, A*0101; B*40:06, B*35:01; C*07:01, C*04:01; DRB1*15:01, DRB1*07:01; DQB1*06:01and DQB1*03:01 alleles, which are the dominating alleles in South Indian population using 454 sequencing technology (Table 2). The common alleles identified from SSP/SSOP were compared with 454 data (Supplemental Table 2).

Caste	Sample Size	HLA A_1	No	HLA A_2	No	HLA B_1	No	HLA B_2	No	HLA DRB1_1	No	HLA DRB1_2	No
Vokkaliga	114	A2	56	A24	26	B7	32	B61	22	DRB1 4	24	DRB1 15	36
Lingayat	79	A1/A2	20	A24	13	B35	23	B51	14	DRB1 7	18	DRB1 15	35
Brahmin	58	A1	14	A33	23	B35	15	B58	13	DRB1 7	14	DRB1 15	16

Table 1: Frequency of HLA alleles from SSP method for three major castes of the Karnataka.

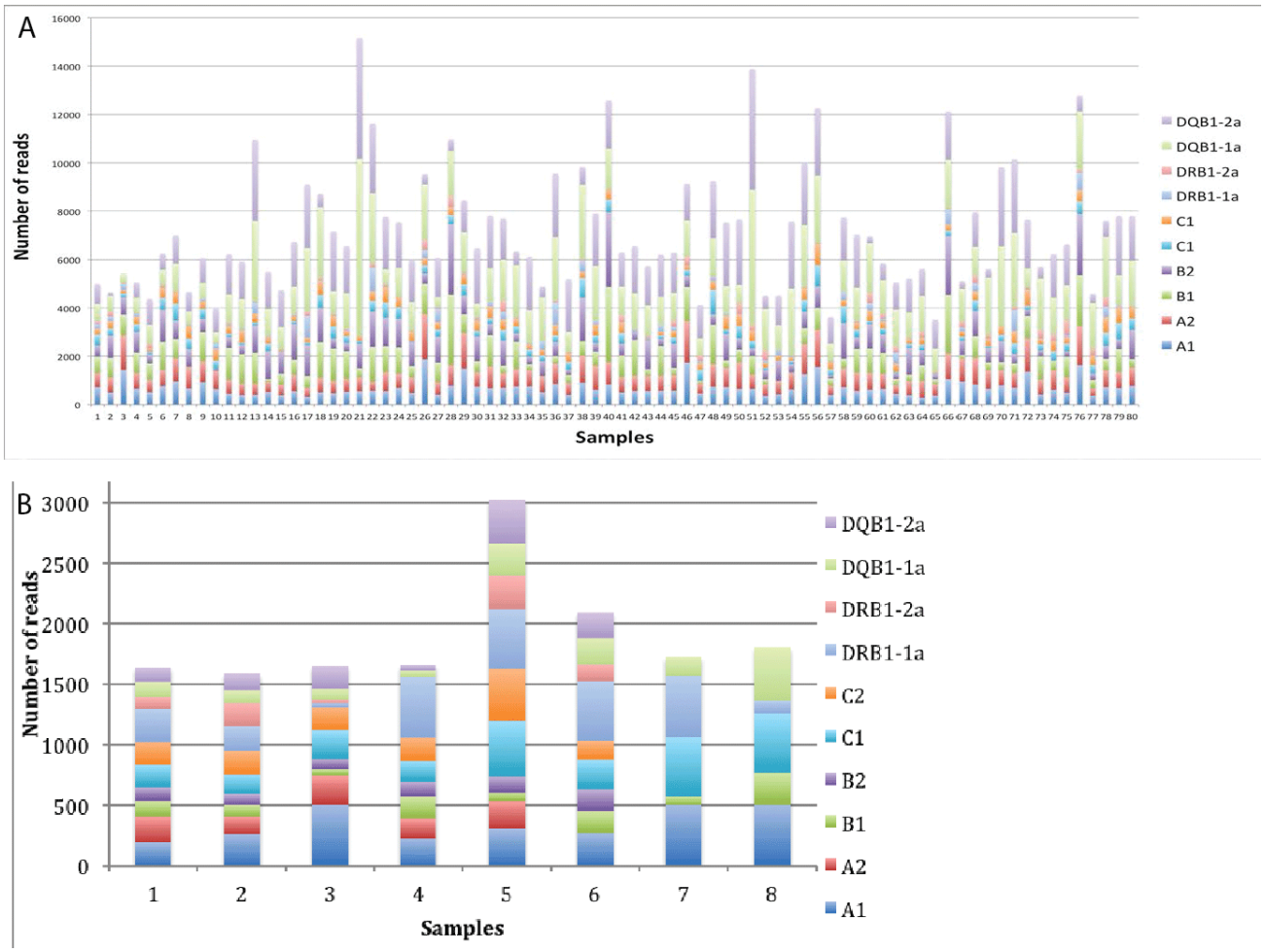

Figure 3: Read depth of coverage per individual by 454 **(A)** and PacBio sequencing **(B)**.

	4-digit Alleles	Number of Alleles
HLA A Allele	A*2402	38
	A*2407	2
	A*0101	26
	A*0103	1
HLA-B Allele	B*4006	16
	B*4001	3
	B*3501	10
	B*3503	9
	B*3559	1
HLA-C Alleles	C*04:01	18
	C*04:03	2
	C*07:01	21
	C*07:02	17
	C*07:04	4
	C*07:26	2
HLA-DRB1 Alleles	DRB1*07:01	34
	DRB1*15:01	18
	DRB1*15:02	20
	DRB1*15:06	2
	DRB1*15:03	1
HLA-DQB1 Alleles	DQB1*03:01	19
	DQB1*03:02	13
	DQB1*03:03	14
	DQB1*06:01	31
	DQB1*06:01	1

Table 2: Frequency of dominant HLA alleles from 454 sequencing.

High-resolution haplotyping of HLA genes using PacBio sequencing

We have selected 8 samples from 454 data-set and sequenced full-length HLA genes for class I (A, B and C) and class II (DQB1 and DRB1) using GenDx HLA specific primers and PacBio sequencing. The GenDx primers designed from 5' and 3' UTRs to obtain full-length gene for HLA class I genes and from exon 2 to 4 for HLA class II genes. The sequencing data was demultiplexed which yielded 2700 reads per sample with the median read length of 2780 nt (Figure 3b). These reads were further analyzed by SMRT Analysis software to generate long consensus sequences. The data coverage for each HLA loci ranged from 36-500X and over 90% of the reads with more than 100X coverage (Table 3). The consensus sequences were assigned to nearest HLA alleles using IMGT database (http://www.imgt.org/). We obtained high resolution data for eight samples with haplotype and phasing information (Table 3). The two homozygous samples of siblings (Sample 7 and 8) had the same paternal and maternal alleles (Table 3).

Samples	Genes	Depth of Allele 1	Closest HLA genotype Allele 1	Depth of Allele 2	Closest HLA genotype Allele 2
S1	HLA-A	183	A*01:01:01	223	A*68:01:02:01
(Heterozygous)	HLA-B	127	B*37:01:01	110	B*27:05:02
	HLA-C	185	C*02:02:02	189	C*06:02:01:01
	HLA-DRB1	271	DRB1*10:01:01	103	DRB1*13:01:01
	HLA-DQB1	123	DQB1*05:01:01	117	DQB1*06:03:01
S2	HLA-A	255	A*02:11:01	152	A*30:01:01
(Heterozygous)	HLA-B	92	B*07:02:01	87	B*50:01:01
	HLA-C	165	C*03:04:01:02	189	C*06:02:01:02
	HLA-DRB1	205	DRB1*07:01:01	194	DRB1*15:06:01
	HLA-DQB1	105	DQB1*02:02:01	139	DQB1*05:02:01
S3	HLA-A	500	A*26:01:01	242	A*30:01:01
(Heterozygous)	HLA-B	50	B*50:01:01	92	B*53:01:01
	HLA-C	227	C*04:01:01	192	C*06:02:01:02
	HLA-DRB1	36	DRB1*15:06:01	24	DRB1*15:03:01
	HLA-DQB1	101	DQB1*05:02:01	179	DQB1*06:02:01
S4	HLA-A	224	A*01:01:01	164	A*03:01:01:01
(Heterozygous)	HLA-B	179	B*37:01:01	125	B*40:06:01:01
	HLA-C	167	C*03:04:01:02	198	C*06:02:01:01
	HLA-DRB1	500	DRB1*10:01:01	51	low quality reads
	HLA-DQB1	45	DQB1*04:02:01	224	DQB1*05:01:01
S5	HLA-A	296	A*11:01:01:01	234	A*24:02:01:01
(Heterozygous)	HLA-B	70	B*35:01:01	132	B*56:01:01
	HLA-C	457	C*04:01:01:01	425	C*07:04:01
	HLA-DRB1	500	DRB1*14:04	279	DRB1*15:02:01
	HLA-DQB1	267	DQB1*05:03:01	362	DQB1*06:01:01
S6	HLA-A	264	A*02:11:01	-	No reads
(Heterozygous)	HLA-B	184	B*07:02:01	184	B*18:01:01:02
	HLA-C	240	C*07:01:01	151	C*07:02:01:03
	HLA-DRB1	500	DRB1*03:01:01	134	DRB1*04:04:01
	HLA-DQB1	222	DQB1*03:02:01	201	DQB1*02:01:01
S7	HLA-A	500	A*01:01:01:01		
(Homozygous)	HLA-B	66	B*37:01:01		
	HLA-C	500	C*06:02:01:01		
	HLA-DRB1	500	DRB1*07:01:01:01		
	HLA-DQB1	152	DQB1*02:02:01		
S8	HLA-A	500	A*01:01:01:01		
(Homozygous)	HLA-B	256	B*37:01:01		
	HLA-C	500	C*06:02:01:01		
	HLA-DRB1	104	DRB1*07:01:01		
	HLA-DQB1	441	DQB1*02:02:01		

Table 3: High resolution HLA typing using PacBio sequencing.

The PacBio data from sample 4 for HLA-DRB1 and sample 6 for HLA-A was not used to assign HLA alleles due to low quality of reads. We found total 68 HLA alleles from 8 samples and the consenses PacBio data is shown in Table 4. The comparison of SSP, 454 and PacBio data is shown in Table 4 and Supplemental Table 3.

Novel HLA alleles for Indian population

There were 46 different HLA alleles obtained from 8 samples based on IMGT database. There were 19 HLA alleles that were not similar (100%) to any nearest alleles in the IMGT database. These were further classified into 7 alleles for HLA-DQB1, 6 alleles for HLA-DRB1, 5 alleles for HLA-A and one allele for HLA B gene (Table 5a). This indicates that HLA DQB and DRB genes showed maximum polymorphism in Indian population. All the alleles had mismatch in intron regions except for the three alleles, which had functional variations in exon 2 and 3 regions. These variations were further confirmed by the 454 sequencing data. These changes in the peptide

binding domain of HLA protein sequence may affect structure, configuration and function.

Sample 1				Sample 2			
HLA Alleles	**SSP/ SSOP**	**454/ Roche**	**PacBio**	**HLA Alleles**	**SSP/ SSOP**	**454/ Roche**	**PacBio**
A 1	A*01	A*01:01:01:01	A*01:01:01:01	A1	A*11	A*11:01:01:01	A*11:01:01:01
		A*01:01:01:02N				A*11:01:03	
A 2	A*68	A*68:01:02	A*68:01:02:01	A 2	A*24	A*24:02:01:01	A*24:02:01:01
		A*68:11N				A*24:02:01:02L	
B 1	B*37	B*37:01:01	B*37:01:01	B 1	B*35	B*35:01:01:01	B*35:01:01:01
		B*37:03N				B*35:01:01:02	
B 2	B*27	B *27:05:02	B*27:05:02	B 2	B*56	B*56:01:01	B*56:01:01
		B *27:05:04				B*56:26	
C 1	NA	C*02:02:02	C*02:02:02	C 1	NA	C*04:01:01:01	C*04:01:01:01
		C*02:29				C*04:01:01:02	
C 2	NA	C*06:02:01:01	C*06:02:01:01	C 2	NA	C*07:04:01:01	C*07:04:01
		C*06:02:01:02				C*07:11	
DRB1	DRB1*10	DRB1*10:01:01	DRB1*10:01:01	DRB1	DRB1*14	DRB1*14:04	DRB1*14:04
DRB1	DRB1*13	DRB1*13:01:01	DRB1*13:01:01	DRB1	DRB1*15	DRB1*15:02:01	DRB1*15:02:01
		DRB1*13:01:08				DRB1*15:19	
DQB1	NA	DQB1*05:01:01	DQB1*05:01:01	DQB1	NA	DQB1*05:03:01	DQB1*05:03:01
						DQB1*05:03:03	
DQB1	NA	DQB1*06:03:01	DQB1*06:03:01	DQB1	NA	DQB1*06:01:01	DQB1*06:01:01
		DQB1*06:14:01				DQB1*06:01:03	

Table 4: Comparative HLA alleles analyses from SSP/SSOP, 454 sequencing and PacBio sequencing.

Sample	Nearest neighbour in IMGT database	InDel/SNP	Type	Position (nt)	Location
Sample 1	A*01:01:01	SNP	G>C	172	Intron
	DRB1*13:01:01	INS	GT	495	Intron
	DRB1*10:01:01	DEL	GA	476-477	Intron
	DQB1*06:03:01	SNP	T>G	25	Intron
	DQB1*05:01:01:(03)	SNP	T>G	25	Intron
		SNP	A>G	965	Intron
Sample 2	A*02:01:01	SNP	T>C	432	Exon 2
		SNP	G>C	434	Exon 2
Sample 3	DRB1*15:03:01	SNP	T>C	426	Exon 2
		SNP	C>T	487	Exon 2
	DQB1*06:02:01	SNP	T>G	25	Intron
Sample 4	A*01:01:01	SNP	G>C	172	Intron
	B*40:06:01:01	SNP	C>G	1047	Intron
	DRB1*10:01:01	SNP	T>A	525	Intron
	DQB1*05:01:01	SNP	T>G	25	Intron
		SNP	A>G	963	Intron
Sample 5	A*11:01:01:01	SNP	A>G	151	Intron
	DRB1*15:02:01	DEL	GTGT	680	Intron
	DQB1*05:03:01	SNP	T>G	25	Intron
	DQB1*06:01:01	SNP	T>G	25	Intron
Sample 6	A*02:11:01	INS	A	741	Exon 3
		DEL	T	2771	Exon 7

	DRB1*03:01:01(01)	INS	GT	512-513	Intron
		SNP	A>C	1212	Intron
	DQB1*03:02:01	SNP	T>G	25	Intron

Table 5a: Novel HLA alleles from PacBio sequencing of Indian population.

Sample No	HLA Allele	Exonic mutation and location (nt)	Change in amino acid	NCBI Accession ID
Sample 2	A*02:01:01	Exon 2 (T>C) gDNA (432 nt)	No change	1873316
		Exon 2 (G>C) gDNA (434nt)	No change	
Sample 3	DRB1*15:03:01	Exon 2 (T>C) gDNA (426 nt)	Tyrosine to Histidine	1873312
Sample 6	A*02:11:01	Exon 3 (Ins A) cDNA (741 nt)	No change	1873302

Table 5b: Amino acid changes in HLA proteins lead to novel HLA allele.

Discussion

The HLA complex represents the most polymorphic component of human genome located on chromosome 6q21.3 with several important immune functions. HLA alleles information play a major role in solid organ and haematopoietic stem cell transplantation, autoimmune disease (e.g. diabetes, rheumatoid arthritis, coeliac), infectious diseases (HIV, Dengue, Hepatitis C), allergy, cancer, vaccine development and population structure [3,4,18]. The HLA loci sequencing have inherent limitations due to extreme polymorphism, ambiguities in haplotype phasing, unknown reference standards, technical problems like shorter read lengths etc [3,4]. The analysis of whole exome sequences of over 7000 pairs of tumor and healthy tissues (http://www.1000genomes.org) has revealed large-scale mutations in HLA genes, which is proposed to be immune evasion by altering HLA protein function as a contributory mechanism in cancer [19].

In the current study, SSP based HLA typing was carried out for South Indian registry of 1020 samples. This analyses revealed the most frequent alleles in South Indian population including HLA-A*02 and A*24; HLA- B*40 and B*35, and HLA-DRB1*15 and DRB1*04. Balakrishnan et al [20] have previously reported that HLA-DRB1*15 allele was the most frequent in Kani tribe (45.19%) and less frequent in Narikkuravars (1.02%) from Tamil Nadu and Kerela, respectively. Additionally, HLA-DRB1*10 and DRB1*07 were found to be most common alleles in South Indian populations, whereas Caucasians DRB1*01 allele was also reported in the Namboothiris of Kerala, Narikkuravars of Tamil Nadu and Maratha of Maharastra [14,20]. Mehra et al [21] reported that HLA-A*02 was the most frequently occuring HLA allele in North Indian populations. The HLA-A*3303 was frequent in Mongoloid population of North Indians. Our study also identified HLA-A*33 in the BMST registry, which was found as the common allele in Brahamins of Karnataka. We speculate that some of Indian sub-populations might have admixtured alleles from non-Indian tribal populations due to continental movement, inter-tribal marriage and human migration [22]. The low resolution HLA typing methods have proven ineffective for HLA typing for bone marrow transplant which remains low resolution, labor-intensive, time-consuming and expensive.

The NGS provides an alternative opportunity to overcome SSP/SSOP and SBT based HLA typing. To test this hypothesis, we used second generation sequencing NGS platforms where more samples were multiplexed in a single sequencing run to obtain long sequences (~ 500 nt). Moreover 454 sequencing provided high depth (>200×) for each gene and generated medium resolution for HLA allele for exon 2 and 3 for class I and exon 2 for class II genes. The accuracy of 454 sequencing demonstrated the high degree (over 95%) concordance with low resolution SSP/SSOP data. There were 72 samples from 454 data mapped accuratly (100%) with SSP alleles. About 1% (8 out of 800 alleles) of HLA alleles dropped out from this study, which could be novel alleles or experimental errors during PCR amplification or sequencing, which need to verified in future studies. The limitation we noticed in 454 typing is that the short reads mapped to more than 2 alleles per gene in most of our samples (90%). Also 454 data resulted in the ambiguity in haplotype phasing because we typed only few exons and missed the information from other exons and introns [9,11]. Among the second-generation NGS platforms, 454 carry more errors in especially around homo-polymorphic sequences [4].

Currently, available HLA database is limited to exonic regions of clinical importance without phase information [4]. Full-length HLA gene sequences with high resolution are required to establish the Indian HLA database of ethnic Indian population. The PacBio sequencing is a single molecule sequencing platform which provides long, full-length and haplotype sequence information for HLA genes/alleles [13]. We identified three novel HLA alleles from PacBio data. Example two SNPs found in the exon 2 of HLA-A gene at 432 nt (T>C) and 434 nt (G>C) in Sample 2. Insertion of 'A' in exon 3 of HLA-A gene at 741 nt position was found in Sample 6. However SNPs in Sample 2 and 6 did not change the amino acid sequence in the HLA protein (Table 5a and b). SNP in exon 2 of HLA class II DRB1 at 426 nt (T>C) position in Sample 3 has changed the amino acids in the protein sequence (Tyrosine to Histidine) (Tables 4a and 4b). The presence of SNPs and InDels has been observed from 454 and PacBio HLA datasets, thus confirming the presence of novel HLA alleles (Table 5b). Interestingly, sample 7 and 8 are homozygous twins who inherited the same alleles from both parents (Table 4). Homozygosity is more frequent when there is a history of consanguinity, which is common in South Indian communities [16]. This can predispose to several genetic disorders and diseases, leading to drastic population decline.

Finally affordable cost is an important factor in the clinical relevance [4,7]. The low resolution by SSP typing was lower [Rs. 3000 ($46) /sample] as compared to medium resolution by 454 [Rs. 7000 ($107)/sample] and high resolution by PacBio typing [Rs. 20,000 ($308)/sample]. We recommend PacBio typing for clinical setting. Although cost was higher during our pilot study, but this will drastically reduce to Rs. 5000 ($75) per sample due to newly launched sequencer, Sequel and 96-barcodes by Pacific Bioscience. The reduced cost is reasonable for Indian clinical setup. However, there will be further scope to reduce cost to less than Rs. 1000 (10 to 15$) per sample. It would be interesting to try other long sequencing technologies for HLA typing [23].

Conclusion

This is the first study of high-resolution HLA typing of Indian population. We characterized 1020 individuals of South Indian population using SSP based typing, which identified most common and dominant alleles such as HLA-A*01 and A*24; HLA-B*40 and B*35, and HLA-DRB1*15 and DRB1*07. We demonstrated the large scale multiplexing of human samples using second generation (454) and third generation (PacBio) NGS platforms to sequence of exon 2 and 3 of class I (HLA-A, B and C) and exon 2 of class II (HLA-DRB1 and DQB1). SSP and 454 can only capture HLA protein information from the selected exonic regions where as PacBio provides complete HLA information including proteins (all exons) and non-coding (introns and UTRs) regions. PacBio sequencing generated long haploid reads for accurate phasing information, which is crucial for transplation and haplotype studies to detect novel alleles from Indian populations. According to our study, PacBio based HLA typing is an ideal technology platform or large-scale typing of HLA registry and other studies. We believe this data will be useful establishing future HLA typing methods for solid organ as well as Bone Marrow transplant. Furthermore HLA plays important role in disease association and drug interactions. Assays developed in this publication will not only lay the path for building a reference database for Indian population, but will also lead to establishment of standard assays that can be used to systematically study the Indian populations. As Indian healthcare is slowly moving from traditional to personalized medical health care, these and many more allele types can be further studied in great details on large-sample size, to identify their role and establish a significant path forward for Indian personalized medicine initiative.

Acknowledgements

We thank Mohammad Zahid from Shiva Scientific/GenDx for providing PacBio HLA primers. We also thank Shanmugasundaram Jayabalan, Roche India for helping in 454 library preparation. Thanks to Anil Singh from Institute of Himalayan Bioresource Technology, Palampur and also thank Paras Yadav, Institute of Life Science, Delhi for their help to access PacBio sequencer. We appreciate Centre for Cellular and Molecular Platforms and Department of Biotechnology, Government of India for financial support for this project.

Conflict of Interest

The Authors declare no conflict of interest.

References

1. De Bakker PI, McVean G, Sabeti PC, Miretti MM, Green T, et al. (2006) A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat Genet 38: 1166-1172.
2. Horton R, Gibson R, Coggill P, Miretti M, Allcock RJ, et al. (2008) Variation analysis and gene annotation of eight MHC haplotypes: the MHC Haplotype Project. Immunogenetics 60: 1-18.
3. Shiina T, Hosomichi K, Inoko H, Kulski JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54: 15-39.
4. De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, et al. (2013) IHIW : review of HLA typing by NGS (16th edn). Int J Immunogenet 40: 72-76.
5. Carrington M, O'Brien SJ (2003) The influence of HLA genotype on AIDS. Annu Rev Med 54: 535-551.
6. Morishima Y, Sasazuki T, Inoko H, Juji T, Akaza T, et al. (2002) 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 99: 4200-4206.
7. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, et al. (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71: 1033-1042.
8. Wittig M, Anmarkrud JA, Kässens JC, Koch S, Forster M, et al. (2015) Development of a high-resolution NGS-based HLA-typing and analysis pipeline. Nucleic Acids Res 43: e70.
9. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, et al. (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens 74: 393-403.
10. Erlich RL, Jia X, Anderson S, Banks E, Gao X, et al. (2011) Next-generation sequencing for HLA typing of class I loci. BMC Genomics 12: 42.
11. Moonsamy PV, Williams T, Bonella P, Holcomb CL, Höglund BN, et al. (2013) High throughput HLA genotyping using 454 sequencing and the Fluidigm Access Arrayâ„¢ System for simplified amplicon library preparation. Tissue Antigens 81: 141-149.
12. Nelson WC, Pyo CW, Vogan D, Wang R, Pyon YS, et al. (2015) An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol 76: 928-938.
13. Mayor NP, Robinson J, McWhinnie AJ, Ranade S, Eng K, et al. (2015) HLA Typing for the Next Generation. PLoS One 10: e0127153.
14. Shankarkumar U (2010) Complexities and similarities of HLA antigen distribution in Asian subcontinent. Indian J Hum Genet 16: 108-110.
15. Dobzhansky T (1973) Is genetic diversity compatible with human equality?. Soc Biol 20: 280-288.
16. Thomas R, Banerjee M (2005) HLA-A allele frequency and haplotype distribution in the dravidian tribal communities of south India. Indian J Hum Genet 11: 140-144.
17. Xing J, Watkins WS, Hu Y, Huff CD, Sabo A, et al. (2010) Genetic diversity in India and the inference of Eurasian population expansion. Genome Biol 11: R113.
18. Jagannathan L, Chaturvedi M, Satish B, Satish KS, Desai A, et al. (2011) HLA-B57 and gender influence the occurrence of tuberculosis in HIV infected people of south India. Clin Dev Immunol 2011: 549023.
19. Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, et al. (2015) Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. Nat Biotechnol 33: 1152-1158. Balakrishnan K, Rathika C, Kamaraj R, Subashini R, Saravanan MP, et al. (2012) Gradients in distribution of HLA-DRB1 alleles in castes and tribes of South India. Int J Hum Genet 12: 45-55.
20. Balakrishnan K, Rathika C, Kamaraj R, Subashini R, Saravanan MP, et al. (2012) Gradients in distribution of HLA-DRB1 alleles in castes and tribes of South India. Int J Hum Genet 12: 45-55.
21. Mehra NK (2010) Defining genetic architecture of the populations in the Indian subcontinent: Impact of human leukocyte antigen diversity studies. Indian Journal of Human Genetics 16: 105-107.
22. Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, et al. (2003) Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res 13: 2277-2290.
23. Szalay T, Golovchenko JA (2015) De novo sequencing and variant calling with nanopores using PoreSeq. Nat Biotechnol 33: 1087-1091.