BMC Genomics

METHODOLOGY ARTICLE Open Access

Phase-defined complete sequencing of the HLA genes by next-generation sequencing

Kazuyoshi Hosomichi[1], Timothy A Jinam[1], Shigeki Mitsunaga[2], Hirofumi Nakaoka[1] and Ituro Inoue[1*]

Abstract

Background: The human leukocyte antigen (HLA) region, the 3.8-Mb segment of the human genome at 6p21, has been associated with more than 100 different diseases, mostly autoimmune diseases. Due to the complex nature of HLA genes, there are difficulties in elucidating complete HLA gene sequences especially HLA gene haplotype structures by the conventional sequencing method. We propose a novel, accurate, and cost-effective method for generating phase-defined complete sequencing of HLA genes by using indexed multiplex next generation sequencing.

Results: A total of 33 HLA homozygous samples, 11 HLA heterozygous samples, and 3 parents-child families were subjected to phase-defined HLA gene sequencing. We applied long-range PCR to amplify six HLA genes (*HLA-A, -C, -B, DRB1, -DQB1*, and *–DPB1*) followed by transposase-based library construction and multiplex sequencing with the MiSeq sequencer. Paired-end reads (2 × 250 bp) derived from the sequencer were aligned to the six HLA gene segments of UCSC hg19 allowing at most 80 bases mismatch. For HLA homozygous samples, the six amplicons of an individual were pooled and simultaneously sequenced and mapped as an individual-tagging method. The paired-end reads were aligned to corresponding genes of UCSC hg19 and unambiguous, continuous sequences were obtained. For HLA heterozygous samples, each amplicon was separately sequenced and mapped as a gene-tagging method. After alignments, we detected informative paired-end reads harboring SNVs on both forward and reverse reads that are used to separate two chromosomes and to generate two phase-defined sequences in an individual. Consequently, we were able to determine the phase-defined HLA gene sequences from promoter to 3′-UTR and assign up to 8-digit HLA allele numbers, regardless of whether the alleles are rare or novel. Parent–child trio-based sequencing validated our sequencing and phasing methods.

Conclusions: Our protocol generated phased-defined sequences of the entire HLA genes, resulting in high resolution HLA typing and new allele detection.

Keywords: HLA, Next generation sequencer

Background

The human leukocyte antigen (HLA) region on the chromosome 6p21 comprising six classical polymorphic HLA genes and at least 132 protein coding genes plays important roles in regulation of immune system as well as fundamental molecular and cellular processes [1]. The completion of a continuous 3.6 Mb of HLA genomic sequence together with annotation of 224 genes, was first reported by The MHC Sequencing Consortium in 1999 [2]. In addition, the MHC Haplotype Project conducted by the Sanger Institute provided genomic sequences and gene annotation of eight different HLA haplotypes, which were registered in the UCSC hg19 and NCBI GRCh37 reference assembly [3-5]. This 3.6 Mb segment occupies only 0.13% of the human genome but is associated with more than 100 different diseases, mostly autoimmune diseases such as type I diabetes, rheumatoid arthritis, psoriasis, and atopic asthma. Recently, HLA genes attracted special attentions, because specific alleles of HLA genes are strongly associated with drug hypersensitivity induced by specific drugs. For example, strong associations between carbamazepine-induced Stevens-Johnson syndrome (SJS) or toxic epidermal necrolysis (TEN) and *HLA-B*15:02* [6,7],

* Correspondence: itinoue@nig.ac.jp
[1]Division of Human Genetics, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan
Full list of author information is available at the end of the article

abacavir-induced liver injury and *HLA-B*57:01* [8-11], and allopurinol-induced SJS or TEN and *HLA-B*58:01* [12] have been reported in various populations. For better understanding of disease causality and drug hypersensitivity, phase-defined complete HLA gene sequencing is required. Furthermore, complete HLA gene sequences are essential to minimize risk of graft versus host disease in hematopoietic transplantation because unknown determinants could be located around HLA genes.

Two methods of HLA genotyping, sequence specific oligonucleotide hybridization (SSO) and capillary sequencing with chain-termination reaction (Sanger sequencing or SBT), have been commonly applied in the past ten years. SSO requires the preparation of specific oligonucleotides corresponding to various genotypes in advance and potential difficulties may arise when new alleles are present. SBT or Sanger sequencing simultaneously sequences two chromosomes, thereby, phasing of the highly polymorphic HLA genes is very difficult *per se*. The common practice of SBT involves sequencing exons 2 and 3 of HLA Class I genes and exon 2 of HLA Class II genes. However, in some cases, different alleles share similar sequences across the sequenced region, leading to ambiguity in allele determination. Moreover, allele determination is generally based on sequence alignment to the IMGT/HLA database where there is an inherent limitation.

Rapid progress of sequencing technologies, so called next generation sequencing (NGS), resulted in revolutionary changes in medical genomics by providing massive sequencing data of human samples. Indeed, the 1,000 genomes project already reported novel variants including both rare and common types from population-scale sequencing [13]. Despite these progresses, complete sequence of HLA region could not be provided by the whole genome analysis because of the extraordinarily polymorphic and complex nature of the HLA region. Therefore, specific analytical procedures should be developed for completion of HLA sequencing and HLA haplotype determination. NGS technologies have potential advantages over Sanger method in sequencing HLA genes, i.e., sequence of single chromosome can be obtained at high throughput. Thus far, several high-throughput HLA typing methods using NGS have been developed [14-17]. One of those involved HLA class I typing by utilizing the 454 GS FLX Titanium sequencing platform with barcoding and multiplexing protocol, resulting in a 4-digit (fields 1 and 2 of HLA allele nomenclature) resolution with high accuracy in *HLA-A* (95.9%), *HLA-B* (99.4%), and *HLA-C* (94.4%) [16]. Recently, more comprehensive analyses of *HLA* typing using the Illumina platform were reported to demonstrate accurate genotyping via high coverage and extensive sequencing of the first seven exons of class I genes (*HLA-A, -B,* and *C*) and exons 2–5 of class II gene (*HLA-DRB1*) [17]. Also, cDNA amplicons of HLA genes were extensively sequenced [18,19] and these exon-centric analyses are successful in determining genotypes after consulting with the IMGT/HLA database to detect the closest HLA gene sequence. However, non-coding regions that may have impact on gene regulation [20,21], or mRNA splicing [22-24] are ignored. Most recently, 8-digit sequencing of HLA-genes is partially achieved using a combination of long-range PCR and Roche GS Junior sequencer and/or IonPGM sequencer [25]. In their study, the closest HLA gene sequence from the IMGT/HLA database was selected as the reference sequence for alignment and phasing, and subsequently they could construct consensus sequence to call HLA alleles. However, the phasing of single nucleotide variants (SNVs) separated at distances longer than the sequence reads, are dependent on the reference sequence because single read sequences of approximately 500 bp from GS Jr and 260 bp from IonPGM could not clarify phase ambiguities of those SNVs. In addition, if a target sequence is not registered in the database, it is not feasible to obtain complete sequences.

In the current study, we completely sequenced long-range PCR amplicons encompassing entire regions of each of the following HLA genes (*HLA-A, -C, -B, -DRB1, -DQB1,* and *-DPB1*). PCR amplicons were subjected to transposase-based library construction and multiplex sequencing with the MiSeq sequencer. Paired-end reads of 2 × 250 bp enables us to demonstrate phase-defined allele determination (also defined as HLA gene haplotype) for 33 HLA homozygous samples, 11 HLA heterozygous samples, and 3 parents-child families.

Results

PCR amplification of the HLA genes and library preparation

Genomic DNAs from 33 HLA homozygous cell lines, 11 heterozygous individuals, and 3 parents-child families were PCR amplified and subjected to HLA genes sequencing. We applied long-range PCR to amplify six HLA genes (*HLA-A, -C, -B, DRB1, -DQB1,* and *–DPB1*) that are known to be highly polymorphic. PCR primers were designed to anneal where known polymorphic sites were not observed according to the dbSNP build 135 database, and to amplify regions spanning the promoter to 3′UTR of the HLA genes (Additional file 1: Table S1). As shown in Additional file 2: Figure S1, specific amplification products of each gene were obtained; the PCR amplicon sizes of *HLA-A*, *-C*, *-B*, *-DRB1*, *-DQB1,* and *-DPB1* were 3,398 bp, 4,296 bp, 4,440 bp, 11,899 bp, 7,118 bp and 13,605 bp, respectively. Generally allelic imbalance and allele drop-out as a result of PCR is manifested by skewed allelic call in next generation