Supporting Information

Wang et al. 10.1073/pnas.1206614109

SI Materials and Methods

Sample Preparation. To amplify the selected human leukocyte antigen (HLA) genes, individual long-range PCR reactions were performed using 5-pmol phosphorylated primers, 100-μM dNTPs, and 2.5 units Crimson LongAmp Taq DNA polymerase (New England Biolabs, NEB) in a 25-μL reaction volume. The reaction included an initial denaturation at 94 °C for 2 min, followed by 40 cycles of 94 °C for 20 s, 63 °C for 45 s, and 68 °C for 5 min (for HLA-A, -B, and -C) or 7 min for HLA-DRB1. The quality and the molecular weight of each PCR was estimated (assessed) in a 0.8% agarose gel and the approximate amount of each product was estimated by the pixel intensity of the bands. From the amplicon of each gene, ~300 ng were pooled and purified using Agencourt AMPure XP beads (Beckman Coulter Genomics) following the manufacturer's instructions and subsequently ligated to form concatemers. For the ligation reaction, overhangs generated by Crimson Taq polymerase were removed by incubating the reaction with 3 units T4 polymerase (NEB), 2,000 units T4 DNA Ligase (NEB), and 1 mM dNTP's in 10× T4 DNA ligase buffer for 10 min at room temperature. This was followed by the addition of l μL 50% PEG and incubated at room temperature for 30 min. Then another 2,000 units of T4 DNA ligase (NEB) was added followed by an overnight incubation at 4 °C. After completion of the reaction, 1 μg of ligation product was randomly fragmented in a Covaris E210R DNA shearing instrument to generate 300- to 350-bp fragments. A total of 225 ng of fragmented DNA was end repaired using the Quick blunting kit (NEB) followed by addition of deoxyadenosines, using Klenow polymerase to facilitate addition of barcoded adapters using 5,000 units of Quick ligase (NEB). For multiplex processing, multiple samples were pooled together and purified using AMPure XP beads (Beckman Coulter). The samples were run on a Pippin Prep DNA size selection system (Sage Biosciences) to select 350- to 450-bp fragments. After elution of the sample, one-half of the eluate was enriched by 13 cycles of PCR using Phusion Hot Start high-fidelity polymerase (NEB). The enriched libraries were quantified and the quality checked by an Agilent 2100 Bioanalyzer (Agilent Technologies). The libraries were diluted to a 10-nM concentration using elution buffer, EB (Qiagen). Following denaturation with sodium hydroxide, the amplified libraries were sequenced at a final concentration of 3.5 pM on the Illumina GAIIx instrument using eight Illumina 36-cycle SBS sequencing kits (v5) to perform a paired-end, 2 × 150 bp, run. After sequencing, the resulting images were analyzed with the proprietary Illumina pipeline v1.3 software. Sequencing was done according to the manual from Illumina.

To verify discordant calls or potential novel alleles, products from an independent PCR amplification were used to confirm the results by Sanger sequencing using the Big Dye Terminator kit v3.1 (Life Technologies) and internal sequencing primers. Ten microliters of PCR products were digested with 1 unit shrimp alkaline phophatase and 1.0 unit of Exonuclease I (Affymetrix) at 37 °C for 15 min followed by a 20-min heat inactivation at 80 °C. The products were directly used in the sequencing reaction or cloned with a TOPO XL PCR Cloning kit with One Shot TOP10 Electrocomp *Escherichia coli* (Invitrogen) before sequencing on the 3730 instrument (Life Technologies).

Comparison of Allele Resolution and Combination Resolution When Different Regions Were Analyzed. Sequence-based typing (SBT) is considered the most comprehensive method for HLA typing. Due to technique difficulty and cost consideration, only the most polymorphic sites of HLA genes were analyzed by this method, which commonly uses exon 2 and 3 sequences for HLA class I analysis and exon 2 alone for HLA class II analysis. With more and more new alleles discovered in the past several years, the accumulated data show that besides those well-analyzed regions, other regions of HLA genes are polymorphic too. Because of this finding, International ImMunoGene Tics (IMGT)/HLA data have designated new names for each group of HLA alleles that have identical nucleotide sequences across exons encoding the peptide binding domains (exons 2 and 3 for HLA class I and exon 2 for HLA class II) with an uppercase "G," which follows the three-field allele designation of the lowest numbered allele in the group (http://www.ebi.ac.uk/imgt/hla/ambig.html).

To compare the allele resolution, which is defined as the percentage of alleles that can be resolved definitively when particular regions of a gene are analyzed, we counted the number of alleles that do not share the same sequence of the analyzed regions and calculated the percentage of those alleles over all alleles listed in the IMGT/HLA database, which was released on October 10, 2011. We applied the procedure if exons 1–7 (our method), or exons 2–4, or exons 2 and 3 (conventional SBT methods) are determined for HLA class I genes, or exons 2–5 (our method) or exon 2 (conventional SBT methods) for HLA-DRB1.

To compare the combination resolution, which is defined as the percentage of combinations of two heterozygous alleles that can be resolved definitively when particular regions of a gene are analyzed, we first enumerated the combined sequence pattern of the analyzed regions as if two heterozygous alleles were coamplified and determined by Sanger sequencing method and counted the number of combinations, each of which has a unique sequence pattern. We then calculated the percentage of those combinations of unique sequence pattern over all enumerated combinations. We applied the procedure if exons 1–7 (our method), or exons 2–4, or exons 2 and 3 (conventional SBT methods) are determined for HLA class I genes or exons 2–5 (our method) or exon 2 (conventional SBT methods) for HLA-DRB1.

For HLA-DRB1 genes, only 15 and 7% reference sequences cover regions in exons 3 and 4 in the IMGT/HLA database released on October 10, 2011. The procedure we used did not count difference in exons 3 and 4 if there is no sequence information. Therefore, the difference between different methods over HLA-DRB1 cannot be clearly illustrated.

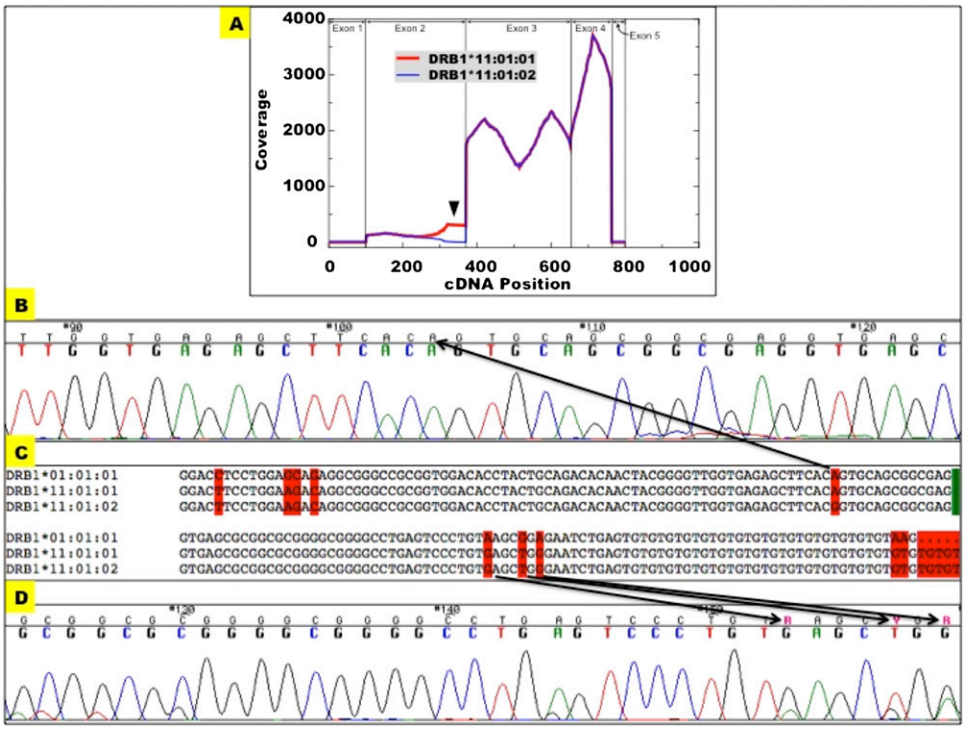

Fig. S1. Sanger sequencing validation of the HLA-DRB1 genotype of the cell-line FH11 (IHW09385). (*A*) Coverage plots for the reference allele HLA-DRB1*11:01:02 (blue) and the predicted allele HLA-DRB1*11:01:01 (red) where the black triangle points to the difference in the coverage plots of these two alleles. (*B*) Partial Sanger sequencing chromatogram of the amplification products in the exon 2 region of HLA-DRB1 locus. (*C*) Alignment of HLA-DRB1*01:01:01, HLA-DRB1*11:01:01, and HLA-DRB1*11:01:02 where the differences among the three alleles are highlighted in red and the intron–exon boundary is indicated in green. (*D*) Partial Sanger sequencing chromatogram of the amplification products in the intron 2 region of HLA-DRB1 locus. Arrows link positions that are different between the three references in the alignment and the corresponding positions in the chromatograms. The IMGT-HLA database reports that the HLA-DRB1 locus of FH11 is heterozygous for 01:01:01/11:01:02. Our Illumina data suggest that it should be heterozygous for 01:01:01/11:01:01. Chromatograms in *B* and *D* match the expected pattern of mixture of HLA-DRB1*01:01:01/11:01:01, instead of HLA-DRB1*01:01:01/11:01:02.

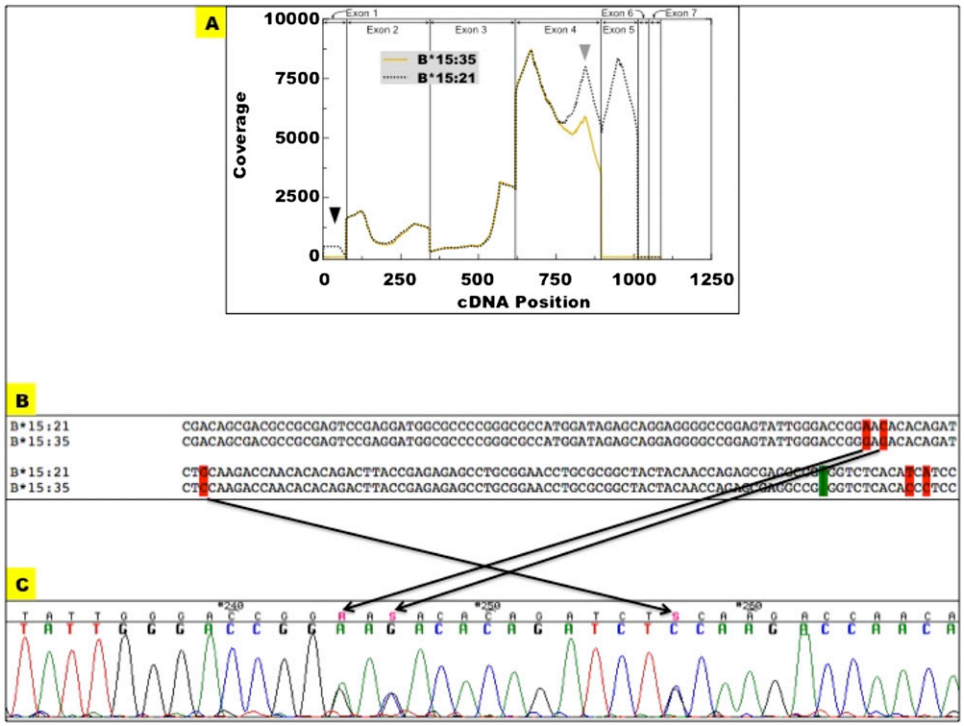

Fig. S2. Sanger sequencing validation of the genotype of HLA-B locus of the cell line FH34 (IHW09415). (*A*) Coverage plots for the reference allele HLA-B*15:35 (yellow line) and the predicted allele HLA-B*15:21 (black dashed line). Note there is no reference sequence for the HLA-B*15:35 allele in the exon 1 region, which is the reason for zero coverage in this region (highlighted by the black triangle). There is no reference sequence for the HLA-B*15:35 allele in exons 5–7 either. Although HLA-B*15:21 and HLA-B*15:35 are identical in exon 4, HLA-B*15:35 has lower coverage than HLA-B*15:21 (highlighted in gray triangle) due to removal of reads that did not pass the pair-end filter (see main text). (*B*) Alignment of HLA-B*15:35 and HLA-B*15:21 in the partial exon 2 and 3 regions where the differences among the three alleles are highlighted in red and the intron–exon boundary is indicated in green. (*C*) Partial Sanger sequencing chromatogram of the amplification products in the exon 2 region of HLA-B locus. Arrows point out the chromatogram pattern matching the expected pattern of mixture of HLA-B*15:21 and HLA-B*15:35. The reference alleles listed for HLA-B locus of FH34 is 15/15:21 and on the basis of our sequencing data, we are able to extend the resolution to 15:21/15:35.

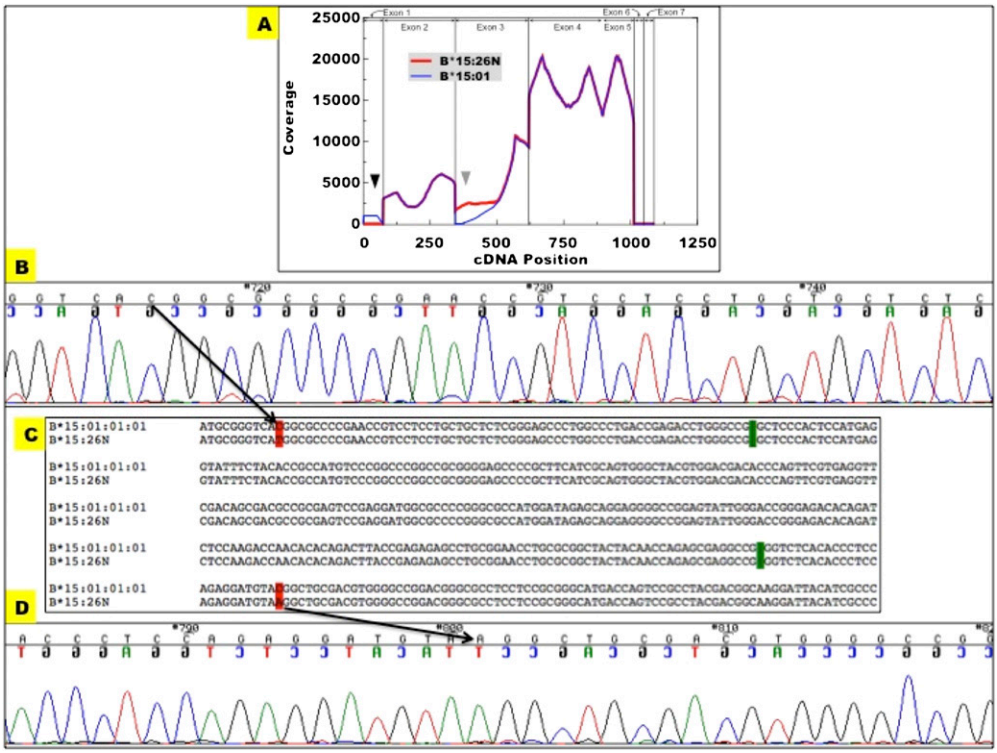

Fig. S3. Sanger sequencing validation of the HLA-B genotype of the cell line ISH3 (IHW09369). (*A*) Coverage plots for reference HLA-B*15:26N (red) and HLA-B*15:01 (blue). Reads align continuously onto exons 2–5, but not exon 1 of HLA-B*15:26N. There are reads aligning to exon 1 of HLA-B*15:01 (black triangle). (*B*) Partial Sanger sequencing chromatogram of the amplification products in the exon 1 region of HLA-B locus. The nucleotide in the 11th position of exon 1 is C as in HLA-B*15:01:01. (*C*) Alignment of HLA-B*15:01:01:01 and HLA-B*15:26N where the differences among the three alleles are highlighted in red and the intron–exon boundary is indicated in green. (*D*) Partial Sanger sequencing chromatogram of the amplification products in the exon 3 region of HLA-B locus. Arrows link positions that are different between the three references in the sequence alignment and the corresponding position in the chromatograms. The IHWG cell-line database reports that the HLA-B locus of ISH3 is homozygous for 15:26N. The chromatograms in *B* and *D* suggest that this is a previously undescribed allele with exon 1 sequence as that of HLA-B*15:01:01:01 and exons 2–5 sequence as that of HLA-B*15:26N.

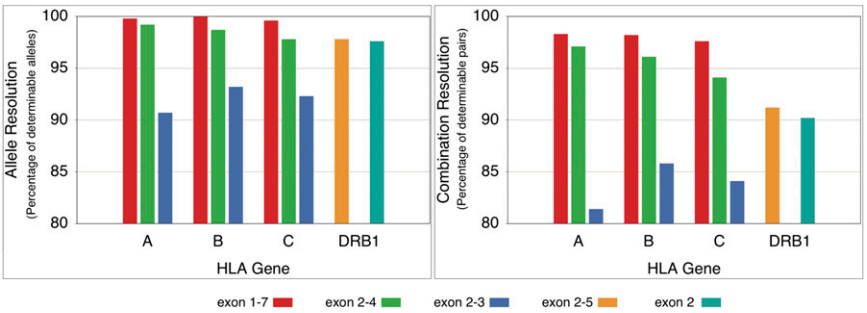

Fig. S4. Comparison of allele resolution (*Left*) and combination resolution (*Right*) if different regions of HLA genes were sequenced. Analysis was based on the IMGT/HLA reference sequence database released on October 10, 2011. The allele resolution is defined as the percentage of alleles that can be resolved definitively when particular regions of a gene are analyzed. The combination resolution is defined as the percentage of combinations of two heterozygous alleles that can be resolved definitively when particular regions of a gene are analyzed. Note that due to the lack of sequence information outside exon 2 for the HLA-DRB1 gene, where only 15% reference sequences cover exon 3 and 7% reference sequences cover the exon 4 region for the HLA-DRB1 gene, the difference between our method and conventional SBT methods over this gene can be estimated accurately.

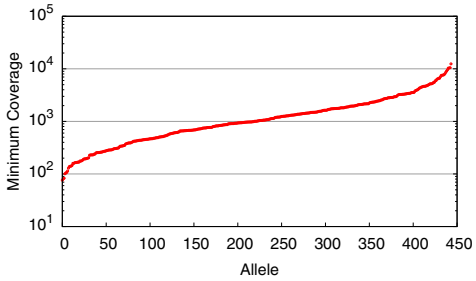

Fig. S5. Minimum coverage (sorted ascending) of all HLA alleles in 59 clinical samples. Only three alleles were typed with minimum coverage less than 100.

Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)