Application of high-throughput, high-resolution and cost-effective next generation sequencing-based large-scale HLA typing in donor registry

M. Zhou1,†, D. Gao2,†, X. Chai1, J. Liu2, Z. Lan1, Q. Liu2, F. Yang1, Y. Guo1, J. Fang1, L. Yang1, D. Du2, L. Chen1, X. Yang1, M. Zhang1, H. Zeng1, J. Lu2, H. Chen1, X. Zhang1, S. Wu1, Y. Han2, J. Tan1, Z. Cheng1, C. Huang1 & W. Wang1,†

1 BGI Diagnosis Co., Ltd, Shenzhen, China
2 China Marrow Donor Program, Beijing, China

Key words
Allelic level; cost effective; genotyping; high throughput; human leukocyte antigen; next generation sequencing

Correspondence
Wei Wang
BGI Diagnosis Co., Ltd
Shenzhen
China
Tel: +86 755 25273780
Fax: +86 755 25273620
e-mail: wangw@genomics.cn

Received 22 January 2014; revised 26 August 2014; accepted 15 October 2014
doi: 10.1111/tan.12477

Abstract

Next generation sequencing (NGS)-based human leukocyte antigen (HLA) typing was used for ultra large-scale genotyping of registry donors for the China Marrow Donor Program (CMDP). More than 79,000 samples were subjected to HLA genotyping at 4-digit allelic level without ambiguities for HLA-A, -B, -C, DRB1 and DQB1 loci, with throughput up to 2068 samples per lane in a HiSeq flow cell (eight lanes per run), and cost reduced by 95% compared with that of Sanger-based typing. Two percent of randomly selected samples were quality control (QC) tested at 4-digit allelic level by the CMDP QC laboratory, yielding a concordance of 99.72%. These results demonstrate that NGS is a cost effective and valuable tool for HLA typing of registry donors.

Introduction

The genes that encode the human leukocyte antigen (HLA) are the most polymorphic genetic system found in the human genome. The classical HLA molecules are known to play an important role in solid organ and hematopoietic stem cell transplantation (HSCT), autoimmune disease and drug hypersensitivity reactions etc. In order to select the best-matched unrelated donors for recipients who are undergoing HSCT, high-resolution HLA typing is required (1, 2). However, for the majority of the donor registries, HLA typing is performed at the intermediate resolution level characterized by National Marrow Donor Program (NMDP) code or at an acceptable high-resolution level whose alleles encode identical nucleotide sequence in the antigen recognition site domain. This practice inevitably identifies many potentially matched unrelated donors for each recipient. Additional typing at high resolution of four or more digits is subsequently performed to select a best-matched donor, which is often a costly and time-consuming process.

Currently, the Sanger sequencing-based HLA typing (SSBT) that targets only the most polymorphic regions i.e. the antigen recognition sites (ARS), usually exons 2 and 3 for class I genes and exon 2 for class II genes, produces a large number of unresolved ambiguous combinations. Additional time and cost are spent to resolve these ambiguities. Furthermore, extensive training for data analysis and editing is required for laboratory staff to become proficient in allele assignment.

Next generation sequencing (NGS) technology, which is characterized by massively parallel clonal sequencing followed by, automatic data analysis, gives bright prospect for achieving cost effective and allelic level genotyping. The methodologies of NGS-based genotyping have been reported by several groups (3–12). By using the combination of polymerase chain reaction (PCR) on Fluidigm Access Array microfluid chip and MiSeq sequencing and targeting exons 2 and 3 only for all six HLA loci, Lange et al. (13) reported their achievement of high-resolution typing of 12,000 donor samples per week. However, no report has been published on the practical application of NGS in truly large-scale typing such as those routinely processed by HLA registries.

Herewith, we document our experience with high-throughput typing for China Marrow Donor Programs (CMDP’s) registry donors using a HiSeq platform-based short-read multiplex sequencing method (11). We mainly target the HLA-A, -B, -C,
-DRB1 and HLA-DQB1 genes, the routinely tested HLA loci for both the registry donors and clinical HSCT, and provide 4-digit allelic level typing results to meet CMDP’s typing requirements. Before applying the test in practice, a panel of 470 previously typed samples was used to validate its performance. The panel included the common and well documented (CWD) alleles, rare alleles and unannotated alleles, which cover most of the allele groups across all five loci. Concordance of alleles unambiguously assigned at 4-digit allelic level in the first pass by NGS was 100% with the previous genotypes determined by SSBT. Following the validation, 79721 CMDP donor samples were typed by NGS between February 2011 and December 2012. For this high volume HLA typing project, we improved our throughput to 2068 samples per lane in a HiSeq flow cell (eight lanes per run). Taking the turnaround time into consideration, SSBT was used to repeat testing for those loci without allele assignment initially by NGS. High-resolution SSP was used to resolve class II ambiguities. For quality control (QC) purposes, CMDP’s own QC laboratory will randomly select and perform 4-digit allelic level SBT typing on 2% of samples shipped to each contract laboratory on an annually basis. On an average, the concordance between CMDP’s QC results with our combined results of NGS plus SSBT and SSP was 99.72%, further demonstrating that NGS can be a cost effective and valuable tool in practical application.

**Materials and methods**

**Samples**

The validation panel included DNA from 389 healthy Chinese volunteers whose genotype was previously determined by SSBT and 81 DNA samples from the UCLA International DNA Exchange Proficiency Test (PT) Program. The panel covered CWD alleles, rare alleles and unannotated (novel) alleles. Detailed genotype information of the panel is listed in Table S1.

The blood specimens from the 79,721 donors were collected from the Chinese volunteers by CMDP. Written consent forms were signed by all volunteers. DNA was extracted using a Thermo Scientific KingFisher Flex (Thermo Fisher Scientific, Waltham, MA) instrument according to the manufacturer’s instructions.

**Methods**

**Amplification**

Exons 1–7 were amplified in four PCR reactions for each class I locus, DRB1 exon 2 was in one PCR, and exons 2 and 3 for DQB1 locus were amplified in a multiplex PCR reaction. In total, 14 PCR reactions were required for each sample (Figure 1). In our assay, 94 samples and 2 negative controls were placed onto a template tray. Identical target regions for all samples in the template tray were amplified together in a PCR tray. In essence, each PCR tray was used to amplify one set of locus-specific target regions. Therefore, 14 trays of PCR reactions were required for a template tray to complete the typing of five loci of 94 samples. To distinguish each sample and amplicon in the subsequent data analysis, multiplex identifiers (MIDs) were synthesized together with both forward and reverse primers. Thus, 96 sets of MIDs were used for a template tray. Each 15 μl PCR reaction contained 1×KAPA2G Robust HotStart ReadyMix (Kapa Biosystems, Boston, MA), 1 μM of each primer (Life Technologies, Grand Island, NY) and 50 ng genomic DNA. The Bravo Liquid Handling Workstation (Agilent Technologies, Santa Clara, CA) was used to aliquot all reagents. The thermal cycling conditions are as follows: 95°C for 3 min, 95°C − 15 s, Annealing temperature for 15 s, 72°C − 5 s; 33 cycles, 72°C − 5 min. Annealing temperature depends on the primer pair used, and it ranges from 55°C to 66°C. After amplification, 2 negative controls and 16 PCR products (well c1–c8 and F5–12) from each PCR tray were verified by a 2% agarose gel. Whenever amplification failures were detected in three or more wells from a PCR tray, re-amplification was performed for the entire tray; if two wells of failure were detected, electrophoresis for the entire tray was performed to verify the quality of amplification. In this process, once failures from more than nine wells were detected in the entire tray, re-amplification for the entire tray was performed; if only one well failed, test will be continued. However, all failed samples will be repeated in the next experiment.

**Fragmentation, library construction and sequencing**

Fourteen plates of PCR products amplified from a template tray that contains 94 samples were pooled as a pre-library. The volume of each plate of amplicons in pooling is calculated from a formula: 

\[ C1 \times V1/L1 = Cn \times Vn/Ln \]

where C: concentration of amplicon; V: pooling volume; L: length of amplicon; 1 or n: amplicon ID. The concentrations for each plate’s amplicons were determined by NanoDrop 8000 UV-Vis Spectrophotometer (Thermo Fisher Scientific). The data of pooling volume for each plate of amplicons were used to guide the subsequent pooling in the large-scale typing. The pooled products were purified by QIAquick PCR Purification Kit (QIAGEN, Hilden, Germany) and diluted to 50 ng/μl with elution buffer (EB). One hundred microliter of this dilution was fragmented using the Covaris S2 system (Covaris Inc., Woburn, MA). In this process, amplicons were sheared into DNA fragments with different length. Because MIDs were only added into the 5’ end of forward and reverse primers, both of the DNA fragments with MIDs and without MIDs would be generated. However, the DNA fragments without MIDs were filtered out in subsequent data analysis. All of the resulting fragments were collected and purified with QIAquick PCR Purification Kit.

The amplification-free Illumina sequencing-library preparation method was adopted for the library construction, including
end-blunting and phosphorylation of DNA fragments by T4 DNA polymerase (Enzymatics Inc., Beverly, MA), Klenow Fragment (Enzymatics Inc.), T4 Polynucleotide Kinase (Enzymatics Inc.) and A-Tailing using Klenow Fragment (3′→5′exo-) (Enzymatics Inc.).

The Illumina barcoded amplification-free sequencing adaptors were ligated to the A-tailed product using T4 DNA ligase (Enzymatics Inc.) at 20°C in the Thermomixer comfort (Eppendorf Research, Hauppauge, NY) for 30′.

As described above, the MIDs and primer sequences were used as tags to trace each sample and amplicon, respectively. The adaptor was utilized as an ID to trace each library. Because 94 samples were prepared as one library, the 470 samples from the panel test were allocated randomly into five groups and prepared as five libraries, which were subsequently ligated to five sets of adapters, respectively. For the CMDP test, 22 libraries, which were mixed as a unit for sequencing, were ligated to 22 sets of adapters, respectively.

After adaptor ligation, the libraries were purified using 1.2× the library volume of AMPure XP Beads (Beckman Coulter Inc., Indianapolis, IN) and eluted with 30 μl of EB. The purified libraries were mixed equimolarly which was determined by quantitative PCR (qPCR) (StepOnePlus, Life Technologies, Foster, CA). For the panel test, the library mixture contains five libraries that were prepared from 470 samples. For the CMDP test, the library mixture contains 22 libraries that were prepared from 2068 samples. Electrophoresis was performed for the library mixture using Low Range ultra gel (2% w/v, Bio-Rad Laboratories Inc., Hercules, CA). Adapter-linked DNA fragments between 400 and 700bp were excised from the library mixture (Figure 2) and purified with Qiaquick Gel Extraction kit (QIAGEN). The length and quantity of the

![Figure 1](image-url)
purified 400–700 bp DNA fragments were verified by both DNA Bioanalyzer 2100 (Agilent Technologies) and qPCR StepOne, and then sequenced with Illumina HiSeq2000 Illumina Inc., San Diego, CA using a pair-end (PE) 150 bp read protocol. Given that the purified 400–700 bp fragments from 22 libraries (2068 samples) were tiled and sequenced in one lane, eight lanes were included in a flow cell. Therefore, 16544 samples can be tested in one run.

**Data analysis and allele assignment**

The variant calling and phasing, single nucleotide polymorphism (SNP) haploid assembling and alignment were utilized for data analysis as described by Cao et al. (11). First, the reads in each lane were allocated into the corresponding libraries according to their respective adapter information. In the subsequent data filtering step, clean reads, with an average quality score no less than 15 and uncalled based less than two, were retained and grouped into individual samples for specific amplicons according to MIDs and genomic primer sequence. During this process, reads without ID information or mis-matched in pair ends were designated as invalid and filtered out. Subsequently, the grouped valid reads were mapped to a reference sequence to make variants calling and phasing (14–16). These variants were used to assemble SNP haploids. By aligning the SNP haploids to allele sequences in the IMGT-HLA database, each allele was called and assigned to a specific locus. The computing procedures for data analysis and genotyping are designed by BGI Co., and it is run through the Bioinformatics Linux Cluster Operation System. The allele database utilized for analysis is updated in May and December annually. The versions of database that we used were from February 2011 to December 2012 including 3.4.0, 3.6.0, 3.8.0 and 3.10.
For donor registry typing, repeat typing could be performed by NGS for those loci without allele assignment. However, for quick turnaround time in data reporting, SSBT was used for retyping, and high-resolution SSP was utilized to either resolve class II ambiguities or determine the correct allele combinations.

**Results**

**Results of validation**

The sizes of the amplicons from the 14 PCR reactions ranged approximately from 250 to 1000 bp. During the process of verification of PCR products, multiple bands, weak bands or no bands that may have resulted from nonspecific amplification, poor amplification and amplification failure were observed from gel electrophoresis.

In the panel test, approximately 0.66 Gb raw data, on an average, were obtained for each library. After filtering out low quality and invalid reads, 0.64 Gb (96.15%) of clean data and 0.46 Gb (71.34%) of valid data were obtained, respectively, and the latter was used for subsequent analysis.

The amount of valid PE reads for each amplicon ranged from 480 to 3000. The extremely low or high outliers, indicating that the amount of valid reads were either extremely redundant or insufficient, were detected for some amplicons. In the data analysis process, the extreme low outliers would lead to failure of variant calling, while the extreme high outliers would result in false homozygosity calling of a variant, especially for HLA-DRB1. The latter is likely because of the false homozygosity calling resulted from imbalanced ratio of reads between the two heterozygous alleles as the number of reads highly increased for one allele but extremely low for the second allele. It may also be because of the formation of PCR artifacts (crossover products) as described by Holcomb et al. (17). The trend for the distribution of the valid reads for each amplicon (Figure 3) can be used as an indicator for pooling condition of the PCR products.

Read depth is one of the factors that affect accurate genotype calling. In our panel test, read depth, on an average, for exon region was 550x. Approximately 98% of the exonic regions had depths above 100x. The minimal depth that still generates an accurate allele call in the panel test was 20x (Figure 4). There were dips in read depth shown in Figure 4. The dips mainly fall into the central part of the amplicons whose length were up to 1000 bp. Additionally, only 150 bp from both ends of the DNA fragments were read in the PE 150 bp sequencing protocol. Therefore, data in the central part of the amplicons are much less than that in the ends, especially for the long PCR products.

According to the resolution level, the final results of the test panel can be divided into three groups: samples with unambiguous genotype assignment, samples with ambiguous genotype assignment and samples with no genotype assignment (Table 1). The concordance of unambiguous genotypes assigned at 4-digit allelic level with those previously determined for all five loci was 100%. In order to evaluate the efficiency of HLA typing strategies, the term of ‘initial success rate’ was introduced. The initial success rates for the HLA-A, -B, -C, DRB1 and DQB1 loci in the panel test were 93.4%, 97.2%, 96.6%, 40.2% and 96.6%, respectively (Table 1).

In reviewing the ambiguous genotype group, we observed a false homozygous assignment phenomenon. The DRB1 locus of the sample typed by SSBT was DRB1*09:01/13:02, while DRB1*13:02/13:128, 13:02/13:128 were assigned by NGS. To check this phenomenon, we reviewed all of the DRB1 homozygotes and samples containing a DRB1*09:01 allele, only one sample appeared to show allele dropout.

In the no genotype assignment group, around 5.8% was caused by amplification or sequencing failure, 65.1% resulted from failure of genotype calling owing to nonspecific amplification, 18.4% resulted from insufficient reads and 10.7% resulted from the new alleles. New alleles were also classified into this
Figure 4 The depth distribution of valid data. The graph depicts depth distributions of valid data for each amplicon. The X-axis indicates the amplicon; the Y-axis indicates the depth. Each black line represents a sample. The red line shows the mean depth for each locus. The green bar in X-axis represents the exon region in each amplicon. For HLA-B exon 3, the left side bar is narrower and lower than the right side bar, which is because of high GC content in the 5′-most sequences.

Reproducibility is one of the most important performance parameters to assess a new test method or technology. In order to assess the reproducibility of the method, 94 samples were repeat tested. The percent of samples that were assigned with the same genotype in both runs was 100% for B and DQB1 loci. It was 96.81% for A locus, 95.74% for C locus and 90.43% for DRB1 locus (Table 2). The percent of samples that failed to obtain genotype assignments in either of the two runs were 3.19%, 4.26% and 9.57% for A, C and DRB1 loci, respectively. Approximately 6.25% of unsuccessful genotype assignments in both runs were resulted from amplification failure, 68.75% resulted from nonspecific amplification and 25.00% resulted from insufficient data caused by poor amplification.
Results of large-scale genotyping

From February 2011 to December 2012, 79,721 CMDP samples were subjected to NGS HLA typing on the HiSeq2000 platform. Approximately 0.67 Gb raw data on an average was obtained for each library, about 0.63 Gb (94.03%) clean data and 0.47 Gb (74.88%) valid data on average were retained after filtering out low quality reads and invalid reads. Read depth, on an average, for the exonic regions was 500×, and 96% of the exonic regions had depth above 100×. The minimal depth which still could generate an accurate allele call was 20×. Most of the alleles can be assigned based on our exonic data. Only several combinations, such as B*15:01,15:02 and B*15:15,15:25, of which the intronic sequences were taken into consideration for the analysis. When 4-digit allelic level results were not assigned, and several possible combinations were given, manual analysis is required to obtain an unambiguous genotype. Unambiguous results for about 15% of DRB1 locus that mainly centered on eight sets of specific combinations, and 1% of the class I loci can be obtained this way. It takes a technician about 40 s to edit and obtain a genotype. Manual analysis for DQB1 is not common. When no alleles were assigned to a locus, computing procedures can flag the loci and remind the technician to perform retyping. The mean value of the initial success rates (including results obtained through manual analysis) for HLA-A, -B, -C, DRB1 and DQB1 were 97.1%, 98.6%, 97.4%, 42.3% and 98.0%, respectively. A 100% initial success rate was obtained for B locus in about 15% of the total libraries. On an average, 2.3% of total loci had no allele assignment in the first pass by NGS and required retyping by SSBT. About 48.5% of DRB1 samples typed required high-resolution SSP testing to determine the correct allele combinations. For QC purposes, CMDP’s own QC laboratory will randomly select and perform 4-digit allelic level SBT typing on 2% of samples sent to contracted laboratory on an annual basis. The concordance between CMDP’s QC results with our combined corresponding results of NGS plus SSBT and SSP was 99.72%. None of the errors were attributable to typing methodology.

Discussion

Conventional SSBT typing strategies create numerous combina-
tional ambiguities. It is both costly and labor intensive to resolve these ambiguities. In our large-scale NGS HLA typing strategies, typing results with unambiguous genotype assignments were obtained on an average of 97.7% for HLA-A, -B, -C and DQB1 alleles. Although on an average, only 42.3% of the DRB1 alleles were assigned at the allelic level. This successful rate is significantly lower than for the other loci. We believe that the low success rate is not caused by typing technique or platform, but by targeting exon 2 only in our typing schema. When reviewing the ambiguous DRB1 results, we found that most of the ambiguities (Table S2) were located in exon 3 and 5’ end of the exon 2 initiation codon except for DRB1*04:06/04:49 where the SNP differences reside in exon 4. High-resolution SSP can distinguish DRB1*04:06 from DRB1*04:49. For the other ambiguities, our unpublished data from the newly designed primer pairs that cover part of DRB1-exon 1, exons 2 and 3 were able to resolve most of the ambiguities. Three hundred seventy-four DNAs with known genotype were used to validate these new primer pairs, and genotypes at 4-digit allelic level were assigned initially for 354 subjects at 4-digit allelic level were assigned initially for 354 subjects (94.6%). The concordance between the genotypes assigned and the previous typing results was 100%. The increased success rate of DRB1 assignment was also observed by Danzer et al. (4) when both exons 2 and 3 were incorporated into the typing strategy.

In order to avoid the false homozygous assignments detected from the panel test, we use the generic SSP to verify the DRB1 homozygotes. It was observed that the DRB1*09:01 allele dropout appeared to be in part of samples, when DRB1*09:01 was in combination with one of the following DRB1*04:08/11/12/13/14/16 allele group, however such an allele dropout was not observed when it paired with one of the following DRB1*01/03/07/09/10/15 allele group. Similar phenomenon was also found during the development of AmpliTtype kit Cetus Co., Emeryville, CA (18, 19). As Walsh et al. described (20), an incorrect or ambiguous genetic typing

Table 1 NGS genotyping results of panel study

<table>
<thead>
<tr>
<th>Locus</th>
<th>Number of samples with unambiguous genotype assignment</th>
<th>Number of samples with ambiguous genotype assignment</th>
<th>Number of samples with no genotype assignment</th>
<th>Concordance of unambiguous assigned genotype with Sanger result (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>439 (93.4%)</td>
<td>457 (97.2%)</td>
<td>454 (96.6%)</td>
<td>189 (40.2%)</td>
</tr>
<tr>
<td></td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
<td>252 (53.6%)</td>
</tr>
<tr>
<td></td>
<td>31 (6.6%)</td>
<td>13 (2.8%)</td>
<td>16 (3.4%)</td>
<td>29 (6.2%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locus</th>
<th>Number of samples with identical genotype in both runs</th>
<th>Number of samples failed to assign genotypes in either one of the two runs</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>91 (96.81%)</td>
<td>3 (3.19%)</td>
</tr>
<tr>
<td>B</td>
<td>94 (100%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>C</td>
<td>90 (95.74%)</td>
<td>4 (4.26%)</td>
</tr>
<tr>
<td>DRB1</td>
<td>85 (90.43%)</td>
<td>9 (9.57%)</td>
</tr>
<tr>
<td>DQB1</td>
<td>94 (100%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

Table 2 Results of allele assignment in the reproducibility test
could be generated because of preferential PCR amplification of one allele relative to the other in a heterozygous sample. Several mechanisms could potentially lead to preferential PCR amplifications, e.g. significant different GC% between a pair of alleles; the PCR products in different length amplified from different alleles; less efficient priming of DNA synthesis of one allele vs the other caused by mismatches between the primer and the specific allelic template, etc. In our typing strategies, several primer pairs were used to amplify DRB1 exon 2, thus competition may have occurred between the primer pairs corresponding to the specific alleles. In large-scale HLA typing, we found that homozygotes occurred in 4% of the samples typed. Generic SSP was used to verify the homozygosity status, and about 1.2% of the total time was spent for this verification task.

Sequencing run time and throughput vary with different platforms. For HiSeq2000 platform, 15 days are required to run a flow cell. Within the flow cell, more than 16,544 samples are able to be tiled and sequenced. Based on an 8-h workday, and on a per sample basis, the entire process from DNA extraction to allele assignment takes about 26 min when HiSeq2000 is used. For the HiSeq 2500 platform, one run containing 6000 samples can be achieved within 24 h (unpublished data) and approximately 15 min to process one sample. Both platforms are suitable for various high volume tests. To accommodate a need for low volume or urgent sample testing, a 3-day turnaround time can be achieved with the smaller MiSeq platform. From our preliminary data, typing for 188 samples can be completed in 3 days on MiSeq with paired-end and 150 bp read lengths. The workflow includes 2.5 h for DNA extraction, 4 h for PCR amplification, 17 h for library preparation, 27 h for sequencing and 4 h for data analysis.

Typing cost is another important factors of concern. In the report by Holcomb et al., 10 HLA genes (17 exons/sample) for 20 samples were typed in one single 454 sequencing run with the reagent costs approximately equivalent to SSBT. In the study by Cao et al., the cost of typing five HLA genes (24 exons/sample) for 96 samples were reduced by 80%, and the throughput were increased by 10 (24 exons/sample) for 96 samples were reduced by 80%, and the study by Cao et al., the cost of typing five HLA genes for 20 samples were typed in one single 454 sequencing run in comparison with SSBT. In our test, 2068 samples were processed in a HiSeq2000 lane. Although high-resolution SSP was needed to resolve ambiguities for 48.5% of DRB1 samples typed and generic SSP was needed to verify 4% of DRB1 homozygotes detected, the cost of SSP tests was much reduced by using in-house prepared reagents. The SSP tests account for about 8% of total cost, and approximately 4.5% of the total time was spent to perform the SSP tests. Sequencing time per HiSeq2000 run is longer than SSBT, however, it cuts costs by 95% and has improved the throughput by 210× in comparison with SSBT. When the workflow is adopted by automation coupled with a labor-saving automatic data analysis pipeline, costs may even be reduced more substantially.

Aerosol contamination is one of the most problematic issues for DNA-based typing. Because more PCR reactions are performed and pooled in NGS workflow, high concentrated nucleotide fragments are more likely accumulated, which greatly increases the possibility of aerosol contamination. Therefore, effective measures to detect cross-contamination and maintain a contamination-free environment must be taken into consideration for laboratories that intend to implement NGS for HLA typing. Facilities, containing laminar flow hoods, may be a good option. To prevent other frequent mishaps such as sample switches during large-scale typing and errors in tracing sample IDs, establishing an effective matching algorithm between tracing tags and samples in data analysis as well as accurate record keeping throughout the entire operation play a vital role in ensuring an error-free NGS typing laboratory.

Approximately 1100 new HLA alleles were identified through NGS typing in our laboratory from February 2011 to December 2012. Assuming that all of these new alleles were named, to identify and assign new nomenclature to these alleles are difficult and expensive tasks with conventional SSBT. For NGS HLA typing, automated data analysis is available. When supercomputers, such as bioinformatics linux clusters, are adopted in data analysis, genotype calling becomes extremely cost effective. In our laboratory, allele assignments for 2068 samples can be achieved in 1.5 working days, which represents a 19-fold increase in efficiency over SSBT (typically 75 samples per working day). The combination of NGS HLA typing and data analysis by supercomputer also makes it cost effective for registry donors or cord blood banks to update the information on samples that were previously identified as novel alleles. These new alleles will receive WHO nomenclatures much sooner from the newly issued IMGT database, which in turn will help both the donor registry and cord blood banks to update their own databases.

Donor registries or cord blood banks with allelic level HLA typing data can significantly accelerate the search for optimally compatible donors. In the past, the cost and time involved have made it prohibitive to build up such cord blood banks or registries. However, the application of high-accuracy, high-throughput and cost-effective NGS HLA typing for registry donors or cord blood units may further this goal. We believe that to adopt NGS HLA typing will completely transform donor workup strategies from searching for multiple potentially matched donors to select the best-matched donor(s) in one single step.

Acknowledgments
We are grateful to Susan H. Hsu from Histocompatibility/Molecular Genetics department at American Red Cross Blood Services, Penn-Jersey Region, Philadelphia, PA and Birgitte Simen from 454 Life Sciences/Roche for careful review of this manuscript. We thank Jian Li from BGI-Shenzhen/Department of Biology, University of Copenhagen, for the suggestion of the paper and the members of BGI-Shenzhen genome sequencing team for help with sequencing.
Conflict of interest

The authors have declared no conflicting interests.

References


Supporting Information

The following supporting information is available for this article:

Table S1. DNA sample information of the panel.

Table S2. The ambiguities of HLA-DRB1 locus generated from NGS HLA typing.