Human Immunology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

journal homepage: www.elsevier.com/locate/humimm

# An integrated genotyping approach for HLA and other complex genetic systems

Wyatt C. Nelson [a,b,1], Chul-Woo Pyo [a,b,1], David Vogan [a,b], Ruihan Wang [a,b], Yoon-Soo Pyon [a], Carly Hennessey [a], Anajane Smith [b], Shalini Pereira [b], Akiko Ishitani [b,c], Daniel E. Geraghty [a,b,c,*]

[a] Fred Hutchinson Cancer Research Center, Seattle, WA 98109, United States
[b] Scisco Genetics Inc., Seattle, WA 98115, United States
[c] Scisco Genetics Inc., Shinga-cho, Kashihara, Nara 634-0006, Japan

ARTICLE INFO

*Article history:*
Received 9 February 2015
Revised 22 April 2015
Accepted 2 May 2015
Available online xxxx

*Keywords:*
Single molecule sequencing
HLA genotyping
KIR haplotyping
High resolution
Multiplex PCR
System integration
Genotyping software

ABSTRACT

Clinical immunogenetics laboratories performing routine sequencing of human leukocyte antigen (HLA) genes in support of hematopoietic cell transplantation are motivated to upgrade to next-generation sequencing (NGS) technology by its potential for cost savings as well as testing accuracy and flexibility. While NGS machines are available and simple to operate, there are few systems available that provide comprehensive sample preparation and data analysis workflows to complete the process. We report on the development and testing of the Integrated Genotyping System (IGS), which has been designed to specifically address the challenges associated with the adoption of NGS in clinical laboratories. To validate the system for a variety of sample DNA sources, we have tested 336 DNA specimens from whole blood, dried blood spots, buccal swabs, and lymphoblastoid cell lines. HLA class I and class II genotypes were derived from amplicon sequencing of HLA-A, -B, -C for exons 1–7 and HLA-DPA1, -DPB1, -DQA1, -DQB1, -DRB1, -DRB3, -DRB4, -DRB5 for exons 1–4. Additionally, to demonstrate the extensibility of the IGS to other genetic loci, KIR haplotyping of 93 samples was carried out in parallel with HLA typing using a workflow based on the HLA system. These results are discussed with respect to their applications in the clinical setting and consequent potential for advancing precision medicine.



## 1. Introduction

For over 30 years, clinical immunogenetics laboratories (CILs) have provided human leukocyte antigen (HLA) typing for use in the clinical setting. Over 190 HLA typing laboratories are currently accredited by the American Society for Histocompatibility and Immunogenetics (ASHI) to carry out HLA typing for clinical applications, including the thousands of allogeneic hematopoietic cell transplants (HCT) performed each year around the world [1,2]. In the absence of an HLA-matched sibling donor, HCT programs mandate stringent criteria for HLA matching of patients and unrelated donors [3–5], and the success of unrelated donor HCT has been amply demonstrated for adults and children with hematologic malignancies [6–10]. CILs that support unrelated donor HCT programs require high resolution HLA typing technology.

While capillary electrophoresis (CE)-based dye-terminator Sanger sequencing is the current gold standard for high resolution HLA typing in clinical laboratories [11,12], the technology is limited in throughput, precision, and dynamic range when compared with NGS systems [13]. *Throughput:* In the Sanger workflow locus-specific PCR amplification is followed by a dye-terminator sequencing reaction, which generates a collection of fluorescently-labeled DNA fragments for each target. Since each collection of fragments is analyzed in a separate capillary tube, throughput of the Sanger CE sequencer can be limited to approximately 48 targets per hour per machine. With current NGS DNA library preparation techniques, a conservative throughput estimate for an NGS machine, such as the MiSeq (Illumina, San Diego, CA), is $10^4$ of 600 bp targets per hour [13]. *Precision*: Sanger CE sequencing detects signals from both alleles of a heterozygous target, often generating diploid ambiguities that require extended testing or statistical inference for resolution. In contrast, NGS methods read

* Corresponding author at: Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, D4-100, Seattle, WA 98109-1024, United States.
*E-mail address:* geraghty@fhcrc.org (D.E. Geraghty).

[1] These authors each contributed substantially to this work and are responsible for this report as co-first authors.

http://dx.doi.org/10.1016/j.humimm.2015.05.001
0198-8859/

single molecules of each allele separately, yielding phased data in a single pass. *Dynamic Range*: In Sanger sequencing, the maximum signal-to-noise ratio is close to 10, meaning that in a sequencing reaction consisting of two alleles wherein one is represented by 10 times as many dye-terminated fragments as the other (perhaps as a result of amplification bias), only the former allele will be detected and the minor allele will be lost in noise. The configuration of NGS makes it possible to simultaneously detect signals that vary in intensity over a much broader range because the signal intensity is proportional to the number of clusters, which is routinely $>10^7$ for the MiSeq device. NGS is thus more tolerant of biases that would otherwise lead to allele drop-off (e.g., preferential amplification, variations in DNA template quality or quantity, etc.) as well as being extensible to quantitative assays such as chimerism testing [14–16].

Recently, a range of NGS systems including reagents and software for HLA typing have been reported [17–19]. Long-range PCR with shotgun sequencing approaches [20–24] are well suited for research applications that require whole gene sequences. Alternatively, amplicon approaches [25–28] have been designed for applications that do not necessarily require complete gene sequences. These systems can be adapted to either the MiSeq or the Ion Torrent (Life Technologies, Carlsbad, CA) platforms. Both systems yield phased data, but the clonal amplification and sequence detection mechanisms of each platform are different. The MiSeq uses on-chip microfluidic cluster generation for clonal amplification with reversible dye-terminator chemistry and optical detection [29]. The Ion Torrent uses an off-chip bead-based clonal amplification by emulsion PCR or isothermal amplification, with microarray pH-sensing for sequence detection [30–32]. Although both amplicon and long-range PCR methods have been demonstrated using the Roche 454 system [33–35], the 454 platform will apparently be discontinued by the manufacturer in 2016.

Here we describe an amplicon-based NGS approach designed for a clinical laboratory setting. The Integrated Genotyping System (IGS) provides a complete package from assay setup through data analysis and delivery of results. This study reports high-resolution HLA class I and class II typing of samples from multiple DNA sources including whole blood, blood spots, buccal swabs, and cell lines. To demonstrate IGS extensibility to other immune complex genotyping, KIR gene content and copy number determination was performed in parallel with HLA typing using an essentially similar workflow. These results were considered and discussed with respect to design constraints and improvement opportunities specific to the clinical laboratory setting.

## 2. Materials and methods

### 2.1. DNA specimens

190 DNAs purified from whole blood, 42 dried blood spots, 53 buccal swabs, and 51 lymphoblastoid cell line (LCL) specimens were tested in parallel. Whole blood specimens were obtained as purified DNA from the UCLA Immunogenetics Center (Los Angeles, CA). DNA was extracted from 6 mm dried blood spots, from healthy donors, using a scalable protocol [36]. Buccal swabs from healthy donors were extracted using a ZR-96 Quick-gDNA™ kit (Zymo Research, Irvine CA). DNA was extracted from LCLs, supplied by the International Histocompatibility Working Group (IHWG, Fred Hutchinson Cancer Research Center, Seattle, WA), using an isopropanol precipitation method [26].

The IHWG and UCLA panels were chosen for this study in large part because there are high-resolution genotyping records for most or all of the HLA class I and class II loci analyzed by the IGS. The inclusion of well-characterized IHWG specimens known to be homozygous was intended to test the IGS for false positive allele assignments across multiple HLA loci. While specimens were not specifically chosen to represent infrequent or rare HLA alleles, the UCLA panel, comprising approximately 33% Hispanic, 20% Caucasian, 20% Black, and 15% Asian (12% with no ethnicity reported), was intended to provide a diversity of HLA genotypes.

### 2.2. Targeted amplicon DNA library preparation

All HLA class I and class II genotyping and KIR analysis followed the manufacturer's specifications (Scisco Genetics Inc., Seattle WA). HLA Version 3.0 IGS kits (Scisco Genetics Inc., Seattle WA) were used for all HLA typing for this report. Briefly, the HLA and KIR procedures consisted of amplicon PCR, amplicon pooling, enzymatic cleanup, barcoding, barcode pooling, magnetic bead-based size selection, and quantification. Complete exons are sequenced for HLA genotyping with the exceptions of exon 3 of HLA-DQA1 and exon 2 of -DRB1 and -DRB345, as noted in Table 1.

#### 2.2.1. Amplicon PCR, pooling, and purification

The IGS primer sets amplify exons as shown in Table 1. The HLA class I Version 3.0 includes 9 amplicon PCR reactions generating 42 targets for HLA-A, B, C (2 alleles × 3 loci × 7 exons). The HLA class II Version 3.0 includes 8 amplicon PCR reactions generating 48 targets for HLA-DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, -DPB1 (2 alleles × 6 loci × 4 exons), depending on the variable presence of HLA-DRB3, -DRB4 and -DRB5 loci. Amplicon PCR setup time was approximately 1 h for 96 samples and thermocycling time is 2.5 h. Subsequent to amplicon PCR, reaction products for each sample are pooled using volumes appropriate to maintaining a balance of targets. A mix of exonuclease I (Epicentre, Madison, WI) and alkaline phosphatase (Roche, Basel, Switzerland) is added to the pooled products, followed by incubation according to the manufacturer's specifications. The KIR haplotyping kit Version 1.0 was used according to the enclosed protocol (Scisco Genetics Inc., Seattle WA).

#### 2.2.2. Barcoding samples

To enable barcoding, DNA linkers containing adaptor sequences are attached to the purified amplicon targets and serve as primer binding sites in a dual-indexing barcode PCR. In this process, all amplicon targets of each individual sample are barcoded with the same index sequences so that every sample has a different barcode, essentially as previously described [37].

#### 2.2.3. Final purification and MiSeq Prep

After barcoding, reaction products from different samples are pooled together. Each pool undergoes a size selection using Agencourt AMPure XP beads (Beckman Coulter, Brea, CA) with a 0.7:1 beads to DNA ratio, followed by column purification using a QIAQuick PCR Purification kit (Qiagen, Venlo, Limburg, Netherlands). The purified pools are quantified using a Quant-iT™ PicoGreen dsDNA Assay kit (Life Technologies, Carlsbad, CA) with Scisco Genetics Quantification Standards (Scisco Genetics Inc., Seattle WA) and pooled into one complete library that is ready for dilution and denaturation according to the Illumina MiSeq loading protocol. For this study, MiSeq v3 kits were used with a loading concentration of 6.5 pM, resulting in a cluster density of approximately 1000 k/mm$^2$.

### 2.3. Illumina MiSeq sequencing

The IGS utilizes Illumina MiSeq paired-end sequencing-by-synthesis (SBS). In this scheme, there are two long reads of each sequencing cluster, 'R1' reads in the 5′–3′ direction, and 'R2' in the 3′–5′ direction. R1 and R2 are coupled together as

Please cite this article in press as: Nelson WC et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.05.001

**Table 1**
IGS kits.

| Type | Version | Exons covered | Notes | # Amplicon PCR Rxns | # Barcode PCR Rxns |
|---|---|---|---|---|---|
| HLA-CI | 3.0 | 1–2–3, 4–5, 6–7 | "X–Y" means exons are phased by sequencing | 9 | 3 |
| HLA-CII | 3.0 | 1, 2, 3, 4 | Segments not covered:<br>DQA1 exon 3: 10 bp (3' end)<br>DRB1345 exon 2: 2 bp (5'end) and 6 bp (3'end) | 8 | 1 |
| KIR | 1.0 | 3, 4, 5, 9 | | 4 | 1 |

a 'read pair'. An additional two short index reads record the barcode sequences. In this way, a read pair can be categorized by sample based on its barcode. Currently, in MiSeq v3 paired-end sequencing chemistry, each read covers up to 300 bp of a target, so that 600 bp can be captured by each pair.

We emphasize that the target generation and library preparation processes described above produce not only target segments, but also non-target segments such as primer-dimers, chimeric PCR products [38], and pseudogene sequences. Data corresponding to non-target sequences, typically <20% of total read pairs, are identified and excluded using computational methods incorporated into the GeMS-Analyzer described below.

A paired-end sequencing run using MiSeq v3 chemistry generates approximately 25 million read pairs. Prior to genotyping analysis, MiSeq control software filters and discards 10–20% of the read pairs due to poor imaging quality and about 5% are set aside into special FASTQ files as unidentified reads with no recorded barcode sequences. Thus, the total number of analyzable reads is approximately 20 million pairs. Theoretically, this number of reads should translate into a run capacity of several thousand samples, but a routine capacity is approximately 400 samples. Reduction in capacity is due to biases causing an imperfect balance of targets within the library, as well as the presence of non-target PCR products.

## 3. Results

### 3.1. Workflows

The workflows for the HLA class I and class II Version 3.0 genotyping kits are outlined with each discrete step in the process indicated by a colored block for each reaction per sample, accompanied by average timelines required for execution of the corresponding steps (Fig. 1). In the class I workflow, there are nine amplicon PCR reactions, which together generate targets for HLA-A/B/C exons 1 through 7. After initial amplifications, the products are pooled into three reactions per sample. After barcoding PCR, the reactions are further consolidated into one tube in preparation for the Illumina MiSeq loading procedure. For the HLA class II system, each of 4 exons for HLA-DRB1, -DRB3, -DRB4, -DRB5, -DQA1, -DQB1, -DPA1, -DPB1 are targeted in a total of 8 reaction mixes. Amplicons were designed to include complete exons in all cases except where primer binding sites are excluded from analysis (Table 1).

Overlapping amplicons and database intersection were used for coverage and phasing of exons in HLA class I (Fig. 2). For the kit version used in this study (Table 1), exons 1 through 7 are sequenced with phasing between a subset of amplicons provided by alignment of overlapping positions that distinguish alleles. In order to establish phase between exons 3 and 4 and between exons 5 and 6, a database intersection is implemented. This computational method, described below, calculates phase by comparing the set of phased subsections of the genes with known genotypes in the IMGT/HLA database (http://www.ebi.ac.uk/ipd/imgt/hla).

### 3.2. Bioinformatics data analysis

The IGS data analysis workflow, illustrated in Fig. 3, uses an automated typing application, GeMS-Analyzer, which exists within the Amazon AWS framework that is also home to Illumina's BaseSpace. These applications are linked so that HLA data analysis can be initiated via a BaseSpace account and no local computer processors are required. When desired, GeMS-Analyzer is able to enlist a multiplicity of virtual servers for parallelized computations in order to reduce analysis times, e.g., 20 servers can process about 1 sample/min of HLA class I and class II data. When initiated by the IGS user, GeMS-Analyzer accesses FASTQ files from BaseSpace, analyzes them, and returns a typing results file to BaseSpace for user download. This results file is then opened using the GeMS-HLA desktop application for the IGS data viewing, editing, and reporting.

#### 3.2.1. GeMS-Analyzer

Read pairs are packaged by the MiSeq control software into FASTQ files, two per sample. Specifically, for each sample, all R1 are grouped into one FASTQ and the corresponding R2 are grouped into a second FASTQ. Read identification numbers are retained within each FASTQ so that pairs can be analyzed together. The first step in GeMS-Analyzer processing is to dispose of read pairs containing adaptor-dimers and chimeric PCR product sequences, accounting for roughly 10–30% of the total read pairs per sample. Adaptor-dimers and chimeras are identified using string searches for forbidden sequences. In the second phase of analysis, reads are subdivided into clusters according to loci and exons. This clustering process uses string searches for primer sequences and other motifs in order to allocate read pairs. The final phase of analysis involves independently comparing all read pairs in a given cluster with a database of known alleles. For each cluster, the allele database is composed of target exon sequences as well as non-target pseudogenes from the IMGT database (http://www.ebi.ac.uk/ipd/imgt/hla/) [39]. Upon each read pair comparison, the known allele is given a matching score based on how well it is represented by the read pair. The read comparison algorithm is based on the Smith-Waterman [40] approach for pairwise alignments. After all read pairs in a cluster have been processed, the allele database is ranked by matching score. Using the top-ranked alleles, logical operations are used to check these candidate sequences for chimeric PCR products, SNPs, allele drop-off, and novel alleles before assigning a typing result.

#### 3.2.2. GeMS-HLA

Typing reports generated by GeMS-Analyzer were reviewed and edited using GeMS-HLA, an application for manual evaluation and reporting of sequencing data. GeMS-Analyzer attaches warning flags to typing results in order to alert the user of results that need verification due to the presence of mismatched positions, low coverage, PCR recombination products, and allele bias. For such cases, GeMS-HLA allows a user to interrogate the flagged results and edit the report as necessary.

Amplicons that do not overlap cannot be phased by sequencing and GeMS uses database intersection logic to infer phased

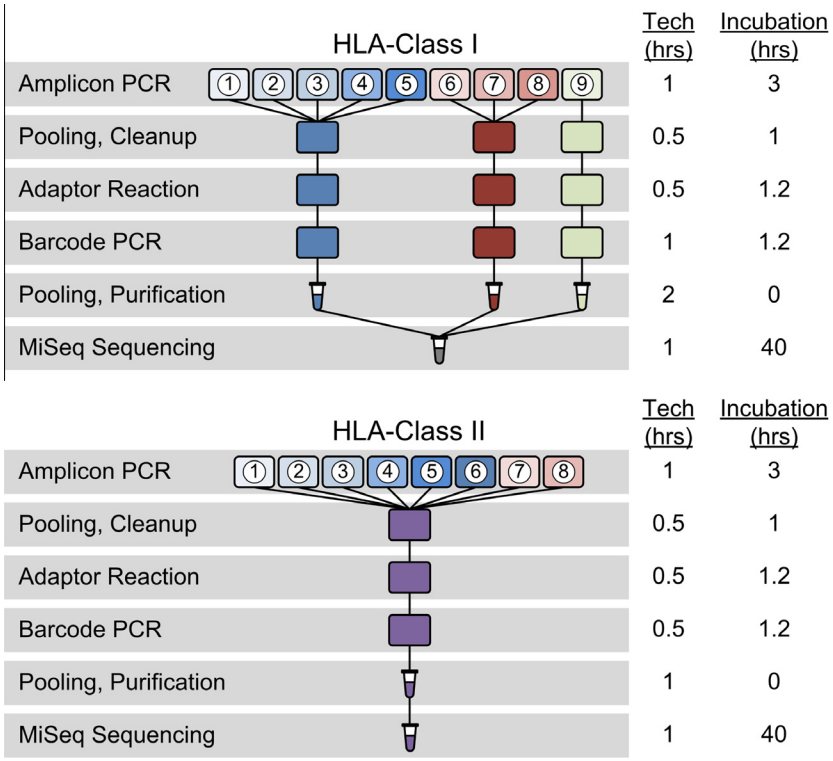


**Fig. 1.** IGS workflow schematic for HLA-class I and II Version 3 kits. Colored blocks represent the number of reactions per sample at each step. Technician and incubation times are listed to the right of each step. In steps following barcode PCR, all samples are pooled together into tubes for further processing and MiSeq preparation. At the MiSeq Prep step, libraries from class I, class II, and KIR may be pooled together for parallel sequencing.

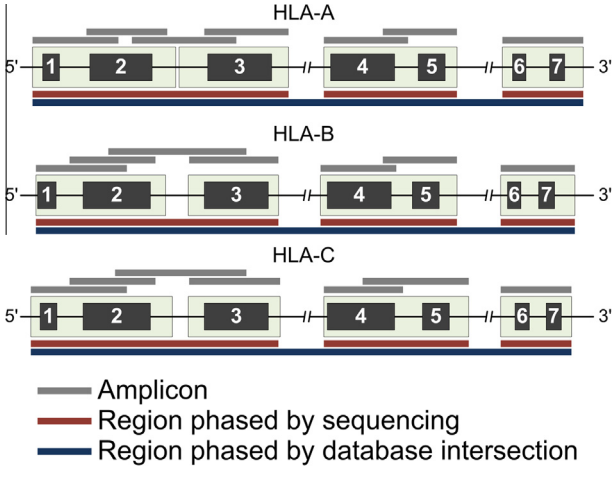


**Fig. 2.** Gene coverage diagram corresponding to IGS HLA-class I Version 3. The gray bars represent target amplicons, of which there are 7 per locus. Regions covered by amplicons are phased by sequencing when overlapping regions contain one or more positions that distinguish alleles. Genotypes are phased across regions not covered by amplicons, i.e., introns 3 and 5, by database intersection logic within GeMS software.

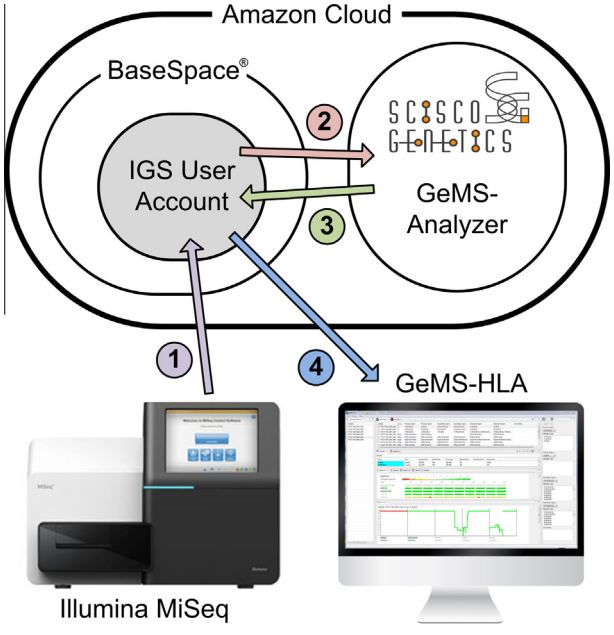


**Fig. 3.** Illustration of the cloud-based GeMS data pipeline for analyzing MiSeq reads and delivering IGS sequencing results. (1) Upon MiSeq sequencing, data is streamed to the IGS user's BaseSpace account. (2) From the BaseSpace account, the user launches GeMS-Analyzer. (3) GeMS-Analyzer generates a typing results file and delivers it to BaseSpace, where is it available for download. (4) Once downloaded, the results file is opened using the GeMS-HLA desktop application. GeMS-HLA is then used for data validation and report generation.

genotypes. In a heterozygous specimen, alleles A1 and B1 are two alleles of exon 1, and A2 and B2 are two alleles of exon 2 generating two possible pairwise combinations: A1A2 + B1B2 and A1B2 + B1A2. If the available HLA database contains both of these combinations, the result is diploid ambiguous. Alternatively, if the database contains only one of these pairwise combinations, the known combination is the most likely and ranked above the unknown combination. GeMS-Analyzer has been designed to also consider coverage levels, recombinant detection, and sequencing quality in allele ranking. Importantly, an allele is equally likely to be detected whether or not it exists in the database.

Please cite this article in press as: Nelson WC et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.05.001

### 3.3. IGS genotyping study

To demonstrate the system, 336 DNA samples were genotyped using the IGS HLA class I and II Version 3.0 systems described in Table 1. In order to test the versatility of the system, DNAs from sources commonly encountered in CILs were tested including blood, cell lines, buccal swabs, and blood spots. The samples were processed in two separate MiSeq runs with each run using Illumina v3 chemistry for 250-cycle paired-end sequencing and producing 20 million read pairs. In the first run, the sequencing library consisted of 95 whole blood and 51 LCL samples. The second library consisted of 95 whole blood, 42 blood spot, and 53 buccal swab samples. The second run also contained a library prepared using the IGS Version 1.0 KIR haplotyping kit.

For each sample type and HLA locus, the total number of genotypes is equal to twice the sample number, excepting the HLA-DRB3, -DRB4, -DRB5 loci, which are present in linkage with specific HLA-DRB1 alleles. The analysis process began with automated GeMS-Analyzer analysis followed by manual GeMS-HLA data evaluation performed independently by two operators, who reviewed flagged results. Manual edits of flagged results involved: (i) coverage below 30 read pairs, (ii) allele biases, which were flagged when the coverage of one allele was larger than 4-fold that of the other allele, and (iii) detection of a novel allele sequence with high coverage at mismatched positions among the top-ranked allele candidates.

The two edited GeMS-HLA reports were cross-checked for consistency and the number of manual data edits was recorded (Table 2). The finalized report was compared with reference panel typing comprised of Sanger sequencing data from the IHWG and the UCLA Immunogenetics Center. IGS results were checked for concordance with reference types provided that they have coverage >20× over exons 2–4 in HLA class I and exons 2–3 in HLA class II. This coverage criteria is consistent with other reports showing that 20× coverage is a sufficient minimum threshold for typing accuracy [20,30] Genotypes missing data from those polymorphic exons were deemed to have no data. Table S1 summarizes coverage data corresponding to the sequencing results in Table 2.

### 3.4. IGS typing accuracy

A comparison of all specimens having pre-existing fields 1:2 or higher HLA typing data indicated 100% concordance with IGS data, excepting three alleles listed in Table 3. For subject IHW09367, we reported a homozygous DQB1*03:03:02, whereas the IHWG Cell and DNA Bank record (http://www.ihwg.org) reports a heterozygous DQB1*03:03:02 + 05:02:01. The same discrepancy was observed in a previous report where alternate analysis was performed [41] that confirmed the typing reported here. For subject IHW09377, the IHWG record lists DRB4*01:01:02, and we report DRB4*01:01:01. In this case, it appears IHWG result was a typographical error because there is no record of DRB4*01:01:02 in the IMGT/HLA database, past or present. Finally, for subject IHW09385 we report DRB1*11:01:01 and the IHWG record contains a DRB1*11:01:02. Wang et al. [20] described the same discrepancy in their study, and confirmed that the correct result is DRB1*11:01:01 using Sanger sequencing as well as an alternative NGS method.

No false positive results were observed in these samples, i.e., no incorrect results that failed to be flagged by GeMS-Analyzer. In some instances, the typing results retained diploid ambiguities that could not be resolved upon sequencing all target exons, e.g., the pairs DPB1*04:01:01/02 + 04:02:01/02 and DPB1*105:01 + 126:01 are not distinguished due to the inability to phase between the two alternative combinations of exons 2 and 3. There were a limited number of diploid ambiguities encountered in this dataset as summarized in Table S2. Two similar diploid ambiguities observed in DQB1 involved alleles with unknown exon sequences. Specifically, DQB1*06:02:01 + 06:04:01 cannot be distinguished from DQB1*06:39 + 06:84 based on the current database information. In the same way, DQB1*06:02:01 + 06:09:01 and DQB1*06:84 + 06:88 are indistinguishable.

Due to the extensive homology across the HLA genes, we observed various pseudogenes, co-amplified with the target sequences. A pseudogene sequence was detected in DQA1 exon 4 that prevented the IGS system from distinguishing DQA*01:04:01:01/02 and DQA1*01:05:02. Additionally, we observed several apparent pseudogene sequences that match HLA alleles present in the IMGT/HLA database. For example, the exon 2 sequence of the DQB1*06:146 allele submitted to IMGT in April 2014 was detected at a low level (approximately 20% coverage level compared with the target) in all DQB1*06:02:01-positive samples. This study and other work in our laboratory have ruled out contamination or other sources of this sequence, suggesting that this allele sequence may be derived from a pseudogene or other gene locus (e.g., -DQB2). Also in this study, 21 samples were found to have a -DPA1 exon 3 variant sequence that matched DPA1*02:02:02 except for a C instead of T at position 447. Beyond diploid ambiguities and pseudogene interference, other allelic ambiguities remained due to insufficient read coverage at exons, e.g., low read pairs covering B exon 1 prevented resolution of B*08:01:01 and B*08:01:20 in subject IHW01154.

Four of the samples included in this study have a DQB1*04:01 allele according to external legacy typing results. Our results revealed a discrepancy with the expected IMGT/HLA database record in the exon 1 sequence for these alleles. Specifically, the IGS-sequenced exon 1 for each of the four DQB1*04:01:01 alleles examined included an A at position 79 of exon 1 (identical to DQB1*04:02:01) rather than the G at this position reported in the latest IMGT/HLA database. Read coverage data detected by GeMS-Analyzer are listed in Table 4. These samples were heterozygous in DQB1, including alleles DQB1*03:01, DQB1*02:01, and DQB1*05:01, each of which has a distinguishing exon 1 sequence. No other alleles were detected over exons 1–4 in these samples. The exon 1 G reported for DQB1*04:01:01 is a unique sequence among all DQB1 exon 1 sequences and, as such, somewhat suspect. A reexamination of the original sample (LKT3/IHW9107) may be required to determine if a sequencing error was reported or if the four specimens in this study have novel -DQB1 alleles with exon 1 sequences matching DQB1*04:02:01, and exons 2–4 matching DQB1*04:01:01. The latter scenario was observed for subject IHW09369, for which the IHWG database reports a homozygous HLA-B*15:26N. Our results indicated that the exon 1 sequence for this cell line matches B*15:01:01:01 instead, as shown in Table 4. Wang et al. [20] also reported that this cell line possesses a novel HLA-B allele in which exon 1 matches B*15:01:01:01 and exons 2–7 match B*15:26N.

### 3.5. DNA sources

The sample sets used for these analyses had a broad range of DNA concentrations over the four sources tested, simulating specimens that might be encountered in any genotyping laboratory. HLA analysis showed that typing accuracy and exon coverage (Tables 2 and S1) were largely independent of sample source and DNA quantity. The number of read pairs per sample plotted over a broad range of template DNA concentrations (0.6–96.3 ng/μl) demonstrated that the balance of data per sample in the IGS sequencing library was not sensitive to genomic DNA concentration (Fig. 4a), suggesting that the amplicon PCR reaches saturation and has a normalizing effect.

Please cite this article in press as: Nelson WC et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.05.001

**Table 2**
Results of IGS HLA genotyping study.

| Sample source | Samples # | Conc. range (ng/μl) | | Results | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Types # | GeMS-HLA edits # (%) | No data # (%) | Concordance # (%) | Resolution[a] | | |
| | | | | | | | | CI ex 2–3 CII ex 2 | CI ex 2–4 CII ex 2–3 | CI ex 1–7 CII ex 1–4 |
| Whole blood | 190 | 5.9–96.3 | *A* | 190 | 13 (2.3) | 1 | 189 (100) | 189 (100) | 189 (100) | 182 (96.3) |
| | | | *B* | 190 | | 1 | 189 (100) | 189 (100) | 186 (98.4) | 185 (97.9) |
| | | | *C* | 190 | | 1 | 189 (100) | 189 (100) | 188 (99.5) | 178 (94.2) |
| | | | DPA1 | 190 | 8 (0.7) | 0 | 190 (100) | 190 (100) | 190 (100) | 190 (100) |
| | | | DPB1 | 190 | | 1 | 189 (100) | 189 (100) | 189 (100) | 181 (95.8) |
| | | | DQA1 | 190 | | 0 | 190 (100) | 190 (100) | 190 (100) | 190 (100) |
| | | | DQB1 | 190 | | 1 | 189 (100) | 189 (100) | 189 (100) | 186 (98.4) |
| | | | DRB1 | 190 | | 0 | 190 (100) | 190 (100) | 190 (100) | 190 (100) |
| | | | DRB345 | 183 | | 0 | 183 (100) | 183 (100) | 183 (100) | 172 (94.0) |
| Blood spot | 42 | n/a | *A* | 42 | 1 (0.7) | 0 | 42 (100) | 42 (100) | 42 (100) | 40 (95.2) |
| | | | *B* | 42 | | 0 | 42 (100) | 42 (100) | 41 (97.6) | 41 (97.6) |
| | | | *C* | 42 | | 0 | 42 (100) | 42 (100) | 42 (100) | 41 (97.6) |
| | | | DPA1 | 42 | 3 (1.2) | 0 | 42 (100) | 42 (100) | 42 (100) | 42 (100) |
| | | | DPB1 | 42 | | 0 | 42 (100) | 42 (100) | 42 (100) | 42 (100) |
| | | | DQA1 | 42 | | 0 | 42 (100) | 42 (100) | 42 (100) | 42 (100) |
| | | | DQB1 | 42 | | 0 | 42 (100) | 42 (100) | 42 (100) | 42 (100) |
| | | | DRB1 | 42 | | 0 | 42 (100) | 42 (100) | 42 (100) | 42 (100) |
| | | | DRB345 | 42 | | 0 | 42 (100) | 42 (100) | 42 (100) | 38 (90.5) |
| Buccal swab | 53 | 0.6–40.0 | *A* | 53 | 4 (2.5) | 0 | 53 (100) | 53 (100) | 53 (100) | 47 (88.7) |
| | | | *B* | 53 | | 0 | 53 (100) | 53 (100) | 53 (100) | 50 (94.3) |
| | | | *C* | 53 | | 0 | 53 (100) | 53 (100) | 53 (100) | 47 (88.7) |
| | | | DPA1 | 53 | 3 (0.9) | 0 | 53 (100) | 53 (100) | 53 (100) | 53 (100) |
| | | | DPB1 | 53 | | 0 | 53 (100) | 53 (100) | 53 (100) | 51 (96.2) |
| | | | DQA1 | 53 | | 0 | 53 (100) | 53 (100) | 53 (100) | 53 (100) |
| | | | DQB1 | 53 | | 0 | 53 (100) | 53 (100) | 53 (100) | 53 (100) |
| | | | DRB1 | 53 | | 0 | 53 (100) | 53 (100) | 53 (100) | 53 (100) |
| | | | DRB345 | 52 | | 0 | 52 (100) | 52 (100) | 52 (100) | 49 (92.4) |
| LCL | 51 | 1.1–22.1 | *A* | 51 | 4 (2.7) | 0 | 51 (100) | 51 (100) | 51 (100) | 48 (94.1) |
| | | | *B* | 51 | | 0 | 51 (100) | 51 (100) | 51 (100) | 50 (98.0) |
| | | | *C* | 51 | | 0 | 51 (100) | 51 (100) | 51 (100) | 46 (90.2) |
| | | | DPA1 | 51 | 7 (2.4) | 0 | 51 (100) | 51 (100) | 51 (100) | 50 (98.0) |
| | | | DPB1 | 51 | | 0 | 51 (100) | 51 (100) | 51 (100) | 50 (98.0) |
| | | | DQA1 | 51 | | 0 | 51 (100) | 51 (100) | 51 (100) | 50 (98.0) |
| | | | DQB1 | 51 | | 0 | 50 (98.0) | 50 (100) | 50 (100) | 49 (98.0) |
| | | | DRB1 | 51 | | 0 | 50 (98.0) | 50 (100) | 50 (100) | 49 (98.0) |
| | | | DRB345 | 49 | | 0 | 48 (97.9) | 48 (100) | 48 (100) | 47 (97.9) |

[a] Instances of diploid and allele ambiguities are listed in Table S2.

**Table 3**
HLA study non-concordant results.

| Cell line | IMGT record | This study | Comments |
|---|---|---|---|
| IHW09367 | DQB1*03:03:02 + DQB1*05:02:01 | DQB1*03:03:02 + DQB1*03:03:02 | Resolved by external typing:<br>Wittig et al. [41] report homozygous DQB1*03:03:02, obtained by an alternative NGS method. Homozygous DQB1*03:03:02 is also reported in the protocol of a commercial SSP-based HLA-DQ Typing kit (GeneBox, Cantanhede, Portugal) |
| IHW09377 | DRB3*02:02:01 + DRB4*01:01:02 | DRB3*02:02:01 + DRB4*01:01:01 | Prior result is a typo:<br>DRB4*01:01:02 is not found in the IMGT database |
| IHW09385 | DRB1*01:01 + DRB1*11:01:02 | DRB1*01:01:01 + DRB1*11:01:01 | Previously resolved by Wang et al. [21]:<br>DRB*11:01:01 and 11:01:02 are distinguished by A versus G at position 357 in exon 2. This study detected no evidence of a G at that position, and Wang et al. obtained the same result by Sanger and an alternative NGS method |

### 3.6. Allele bias, drop-off, and other exceptions

To examine allele bias across all exons and loci, heterozygous genotypes were plotted as relative coverage, i.e., the ratio of allele 1 to allele 2, where allele 1 has fewer reads (Fig. 4b and c). With this definition a perfectly balanced PCR has a relative coverage of 1. The data indicated that a significant majority of loci exhibited relative coverage of greater than 0.5, and in the worst cases was near 0.2. As a result, GeMS-Analyzer is designed to alert the user to data with relative coverage less than 0.25. Allele imbalance is also detected in GeMS-Analyzer by checking for consistency across all exons using database intersection logic. For example, if allele drop-off at one exon leads to a homozygous result where a heterozygous result would be consistent with genotypes in the IMGT/HLA database, the result is flagged for user evaluation in GeMS-HLA. In a single exception due to the complete divergence of intron 2 sequences for the B*73:01 and B*73:02 alleles, GeMS-Analyzer uses a special case analysis algorithm to assemble a complete exon 2 in addition to relying on data from exons 1, 3–7 and partial data from exon 2 assembled via the general case analysis algorithms. Among the 336 samples analyzed, 3 samples with B*73:01 were correctly genotyped.

Please cite this article in press as: Nelson WC et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.05.001

**Table 4**
Discrepant exon 1 IMGT database record.

| Sample | Legacy result | GeMS candidate allele | Read pair coverage | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Exon 1 | Exon 2 | Exon 3 | Exon 4 | Exon 5 | Exon 6 | Exon 7 |
| PL1–28 | DQB1*04:01 | DQB1*04:01:01 | 0[a] | 283 | 484 | 142 | – | – | – |
| | | DQB1*04:02:01 | 103 | 0 | 484 | 142 | – | – | – |
| PL2–30 | DQB1*04:01 | DQB1*04:01:01 | 0[a] | 144 | 215 | 45 | – | – | – |
| | | DQB1*04:02:01 | 40 | 0 | 215 | 45 | – | – | – |
| PL2–52 | DQB1*04:01 | DQB1*04:01:01 | 0[a] | 143 | 243 | 74 | – | – | – |
| | | DQB1*04:02:01 | 31 | 0 | 243 | 74 | – | – | – |
| PL3–70 (IHW09370) | DQB1*04:01 | DQB1*04:01:01 | 0[a] | 46 | 434 | 93 | – | – | – |
| | | DQB1*04:02:01 | 45 | 0 | 434 | 93 | – | – | – |
| PL3–69 (IHW09369) | B*15:26N | B*15:26N | 0[a] | 817 | 771 | 253 | 154 | 1101 | 590 |
| | | B*15:01:01:01 | 91 | 0 | 0 | 253 | 154 | 1101 | 590 |

[a] Coverage = 0 due discrepancy with IMGT database record for exon 1 sequence.

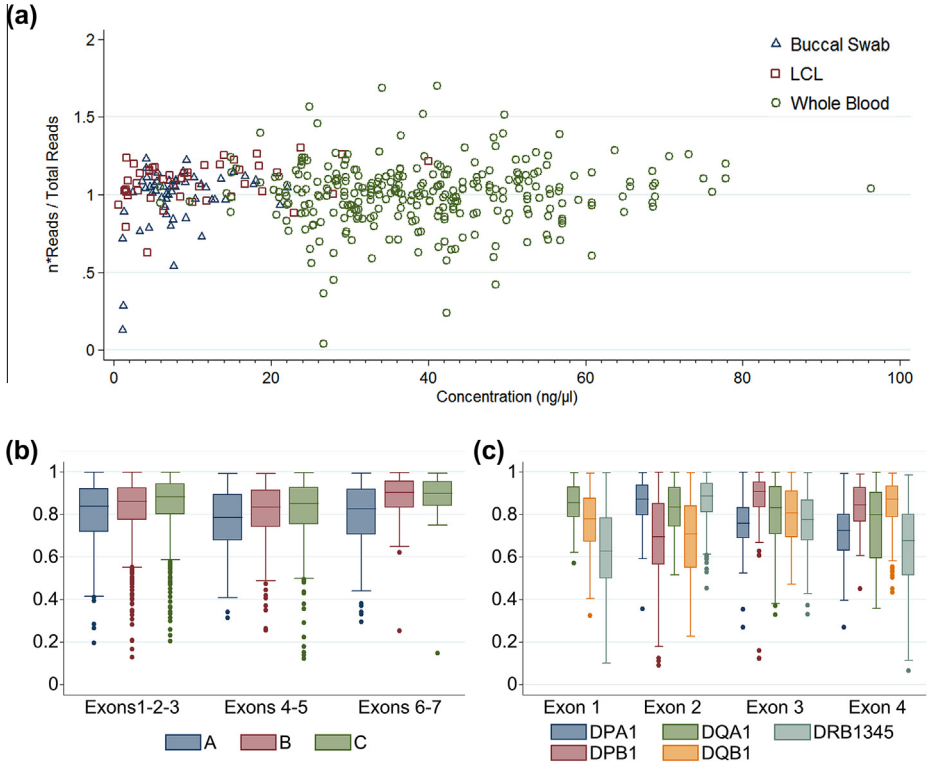


**Fig. 4.** (a) Normalized read pair counts for IGS libraries from different DNA sources versus amplicon PCR template concentration. *N* = number of samples in each library. For an ideal balance of reads per sample, all normalized read pair counts would be equal to 1. (b) Class I relative allele coverage calculated for all heterozygous samples (total allele pairs for each box plot: exons 1–2–3/4–5/6–7: *A* = 732/506/244, *B* = 781/362/124, *C* = 690/475/115) as the ratio of the read pairs for allele 1 to allele 2, where the numerator allele has lower coverage. Boxes represent median and interquartile ranges, whiskers show range, and outliers are plotted separately as dots. (c) Class II relative allele coverage calculated for all heterozygous samples (total allele pairs for each box plot: exon 1/2/3/4: DPA1 = 0/144/134/117, DBP1 = 0/250/176/93, DQA1 = 248/267/265/270, DQB1 = 212/275/271/236, DRB = 300/469/430/147); plotted in the same manner as (b).

### 3.7. Intron data and null alleles

A selection of specimens to validate the detection of null alleles including A*68:11N, C*04:09N, and DRB5*01:08N were examined in this study. Cell lines IHW09418 and IHW09419 were correctly typed as A*68:11N, which is distinguished from A*68:01 by positions in exon 1, i.e., a single base deletion at position 46 and a T at position 47. The cell line IHW09501 was correctly typed as C*04:09N based on the single base deletion at position 1095 in exon 7. This report also included the detection in 4 specimens having the 19 base exon 3 deletion that characterizes the DRB5*01:08N allele.

As described above (Fig. 2), the HLA class I system covers intron sequences that can be utilized for null allele detection in GeMS-Analyzer. For example, B*15:01:01:02N is distinguished from the wild type by a 10-base deletion at position 175 of intron 1. Similarly, A*24:02:01:02L is detected by a substitution at position 708 in intron 2. One LCL specimen (IHWG09504) was typed in this study to demonstrate the accurate detection of A*24:02:01:02L. In total, sequences identifying all 8-digit HLA-A and HLA-B alleles reported in the IMGT v.3.19 with the N, L, or S designation, excepting HLA-A*02:01:01:02L, can be detected by the IGS and reported by GeMS-Analyzer (Table S3).

### 3.8. IGS KIR haplotyping

One of the advantages of the amplicon-based NGS approach for genotyping is its extensibility to other genetic loci without

significant changes in the workflow. To demonstrate this, the KIR haplotyping protocol previously described [42] using sequencing with ABI dye terminators was adapted to the IGS system, querying KIR exons 3, 4, 5 and 9 for gene content. Whereas the dye terminator protocol depended on fluorescent peak heights of variant positions marking specific KIR loci, the IGS protocol utilizes direct counting of relative read numbers from KIR loci-specific amplicons amplified with the same primer pairs. Those locus-specific sequences measured in the dye terminator protocol were queried using this NGS approach, and additional locus-specific sequences from exons 5 and 9 were incorporated into the analysis algorithm to provide further confirmation.

The commercially available KIR IGS kit v1.0, which determines haplotype content in the manner described above, was used on 81 blood and 12 LCL samples. Positive controls included 24 samples that had been previously KIR typed for gene content or sequenced in their entirety [42,43]. The workflow employed essentially the same steps carried out for HLA class I or class II (Fig. 1), with 4 amplicon PCR reactions followed by consolidation into one reaction per sample for barcoding. Samples were pooled and run together with HLA class I and class II libraries on the same MiSeq run. The current version of KIR analysis software measures gene content through string searches and includes partial allelic data. A KIR IGS version enabling complete allele determination using the GeMS-Analyzer/GeMS-HLA framework is under development.

For 93 samples, data were recovered for all targeted exons, all samples yielded complete haplotype data, and all detected patterns matched previously identified KIR haplotypes (Table S4). A representative variety of KIR haplotypes with frequencies similar to those found in previous studies [42] were detected (Table 5). All 24 positive control samples yielded haplotypes concordant with previous studies. The samples examined are commonly available samples (i.e., from UCLA panels, IHWG panels) and therefore these data may find utility as a comparative dataset in other studies. The capacity to include KIR haplotyping with HLA in parallel demonstrated the capability for expansion of the IGS to other complex genetic systems in an entirely horizontal fashion, capitalizing on the previously developed HLA components including the protocol, workflow, and analysis frameworks.

## 4. Discussion

As interest in NGS for genotyping HLA and other genetic systems continues to grow, there is little doubt that the technology will soon move aggressively into clinical immunogenetics laboratories supporting HCTs. This study demonstrated the use and capabilities of an integrated genotyping system using amplicon-based NGS approaches for HLA and KIR genotyping, and has defined challenges to be met and improvements to be made that can accelerate this transition. The approach demonstrated the ability to generate high-resolution genotypes from a variety of DNA sources common to research and clinical genotyping applications. The protocols, reagents, and software applications are commercially available and are now being used in CILs and for research in support of vaccine trials [44,45].

In designing the IGS laboratory workflow, amplicon PCR was chosen as the target generation mechanism for a number of reasons. First, it is generally applicable to any locus, with the design of new genotyping assays constrained primarily by the existence of adequate genomic data resources. As genome sequence data accumulates rapidly [46,47], we expect that this limitation, to the extent that it exists, will be short-lived, even for the more challenging genomic regions that are equivalent in complexity to the MHC and KIR. The data analysis approach and software tools applied are also extensible if genomic data resources are available by the use of rule sets for assignment of allele types. This extensibility enables the IGS to support and expand into new areas for genotyping with relative ease using the established framework for genotyping the HLA system, itself among the most complex and polymorphic genetic systems.

A second and technical advantage of amplicon-based NGS for CILs in particular is the simplicity of application. The basic approach includes amplicon PCR and incubation steps that are already in common use in CILs, and avoids steps that involve variable interpretations and adjustments in protocol, such as random fragmentation of larger DNA products or gel extraction for size selection. This is a notable advantage given the needs of CILs to provide results in a time-constrained setting, with the associated low tolerance for sample repeats. A third advantage is the high tolerance for DNA sources and concentrations with as little as 20 ng of total DNA sufficient for the complete six-digit HLA class I and class II eleven locus typing reported here (Fig. 4). This is an essential consideration in particular given variability in both quality and quantity of DNA derived from common sample sources available to clinical laboratories. Quality of DNA is a consideration for DNA from blood spots, and to a lesser extent from buccal swabs, which may provide DNA below of the size range suitable for long range PCR.

In addition to their application in the CIL setting, amplicon-based approaches are applicable in medium and high throughput settings as well. The bulk of each workflow involves setting up and then transferring and aggregating PCR reactions and products. Given the availability of pipetting robotics that can perform these tasks, several satisfactory alternatives for equipping a laboratory with a high throughput workflow system can be envisioned. As a pointed example, the entirety of the HLA and KIR genotyping reported in this study – *without* any supporting robotics – was prepared for two MiSeq runs by one technician within 2 weeks (average times required for each step are indicated in Fig. 1).

An advantage of long-range (LR) PCR shotgun approaches over amplicon-PCR is the completeness of the DNA sequences obtained. While LR-PCR provides intron sequences, it is not at all clear that such data will have any impact in the HCT or other settings beyond the identification of new – and at this point rare – intron-encoded null or otherwise functionally altered alleles. Indeed, the most effective application of the LR approach may be in generating genomic sequences that supplement databases to further enhance amplicon-based approaches. Given that the majority of the variation in HLA lies in exons 2 and 3 of HLA class I and exon 2 of the HLA class II genes, intron-encoded polymorphism is relatively lower in extent [48] and even more so in functional consequence [49]. At the same time, the amplicon approach can effectively identify intron-encoded null alleles (Table S3) by capturing introns 1, 2, 4, and 6 in HLA class I (Fig. 2), and could be easily extended to include all HLA class I intron data if that proved useful. Another proposed advantage of the LR approach is the ability to phase data more completely and thus to lower the number of diploid ambiguities reported. However, this is apparently of minimal impact given the small number of diploid ambiguities remaining when sequencing HLA class I exons 1–7 and HLA class II exons 1–4 using the amplicon approach. In fact, of the diploid ambiguities that do remain, most of these persist with shotgun approaches when alleles share common sequences extending longer than the average sequence extent generated by the sequencing technology (e.g., the distance between exons 2 and 3 is ~2500 bp for HLA-DQB1 and ~4000 bp for HLA-DPB1). The only technology capable of resolving such ambiguities may be the single molecule real time sequencing approach from Pacific Biosciences [50], to the extent that this method can be routinely applied to HLA.

Please cite this article in press as: Nelson WC et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.05.001

**Table 5**
Results of IGS KIR haplotyping study.

| Sample source | Samples # | Conc. range (ng/μl) | Results | |
|---|---|---|---|---|
| | | | Haplotypes detected[a] | # (Frequency) |
| Whole blood | 81 | 6.7–13.3 | cA01\|tA01 + cA01\|tA01 | 19 (0.23) |
| | | | cA01\|tA01 + cA01\|tB01 | 10 (0.12) |
| | | | cA01\|tA01 + cA01\|tB01 − del7 | 1 (0.01) |
| | | | cA01\|tA01 + cB01\|tA01 | 9 (0.11) |
| | | | cA01\|tA01 + cB01\|tA01 − hybd1 | 1 (0.01) |
| | | | cA01\|tA01 + cB01\|tA01 − hybd2 | 1 (0.01) |
| | | | cA01\|tA01 + cB01\|tA01 − ins4 | 1 (0.01) |
| | | | cA01\|tA01 + cB01\|tB01/cA01\|tB01 + cB01\|tA01 | 5 (0.06) |
| | | | cA01\|tA01 + cB02\|tA01 | 6 (0.07) |
| | | | cA01\|tA01 + cB02\|tB01/cA01\|tB01 + cB02\|tA01 | 9 (0.11) |
| | | | cA01\|tA01 + cB02\|tB01 − del6 | 4 (0.05) |
| | | | cA01\|tA01 − del5 + cB01\|tA01 − del10 | 1 (0.01) |
| | | | cA01\|tA01 − ins4 + cA01\|tB01 | 1 (0.01) |
| | | | cA01\|tB01 + cA01\|tB01 | 2 (0.02) |
| | | | cB01\|tA01 + cB01\|tA01 | 1 (0.01) |
| | | | cB01\|tB01 + cB02\|tA01/cB01\|tA01 + cB02\|tB01 | 1 (0.01) |
| | | | cB02\|tA01 + cB01\|tA01 | 2 (0.02) |
| | | | cB02\|tA01 + cB01\|tA01 − hybd1 | 1 (0.01) |
| | | | cB02\|tA01 + cB02\|tA01 | 1 (0.01) |
| | | | cB02\|tA01 + cB02\|tB01 | 1 (0.01) |
| LCL | 12 | 2.4–16.2 | cA01\|tA01 + cB01\|tA01 | 2 (0.17) |
| | | | cA01\|tA01 + cB01\|tA01 − hybd2 | 1 (0.08) |
| | | | cA01\|tA01 + cB01\|tB01/cA01\|tB01 + cB01\|tA01 | 2 (0.17) |
| | | | cA01\|tA01 + cB02\|tA01 | 3 (0.25) |
| | | | cA01\|tA01 + cB02\|tB01/cA01\|tB01 + cB02\|tA01 | 3 (0.25) |
| | | | cA01\|tB01 − del8 + cB02\|tA01 | 1 (0.08) |

[a] KIR haplotype nomenclature codes from [43] available in Table S4.

### 4.1. Challenges to improve the system

In an amplicon-based approach, as with other NGS systems, the primary goal of the laboratory preparation protocol is to produce a well-balanced sequencing library, ideally one in which every target is equally represented. The key challenge is maximizing the number of targets per reaction while minimizing preferential amplification. Given 20 million reads and completely balanced amplifications of all targets, the theoretical number of typings per MiSeq run is over 10,000 samples. In practice, 400 samples per run is routinely achievable, with the differential being largely due to imbalances between different amplicons, and the need to capture the lowest frequency product among them. This requires that the average read number is closer to 500 copies lowering the potential typeable number to ~1000 samples, with the remainder of the reduction due to primer dimer and chimeric reads that are eliminated by the software. While there is substantial room for increasing the sample count per run by better balancing of the amplicon reaction products, 400 samples per run fares very well against other approaches and is about 5-fold the routine throughput of typical CILs.

### 4.2. Improvements in software

The main area for development and optimization within the data analysis mechanism of the IGS is to reduce the number of flagged results per run, thereby reducing the amount of time spent by an operator evaluating data in GeMS-HLA. Ideally, there would only be flags to indicated failed typings due to the absence of a valid read pair population. Given the complexity and diversity of HLA libraries however, in this study and routinely, many flags (approximately 0.5 flags per sample) result because the primary focus has been on designing GeMS-Analyzer to be fault-tolerant – meaning that low coverage, high allele bias, novel alleles, and interference from non-target read pairs leads to flagged results beyond any false positives. Ultimately, this fault-tolerant design can be optimized through improvements in the operational matching statistics that allow quantitative measures of confidence to be established for each genotype. This iterative optimization process involves cycles of building more effective logic and testing it on large and diverse datasets. Additionally, the implementation of an independent parallel analysis algorithm within GeMS-Analyzer for crosschecking the current algorithm can provide additional confidence in establishing well-founded statistical probabilities.

Another key component for advancing GeMS-Analyzer is to incorporate new data into further building and refining the allele databases. Already, the number of new allele sequences added to the IMGT/HLA and other reference databases is rapidly expanding, primarily due to the application of NGS to build large HLA typing repository data [25]. With this expansion, however, care must be taken in ensuring that these submissions are accurate and thoroughly characterized by independent methods. In our experience, co-amplified pseudogene sequences or other derivative sequence targets are the main source of interference in HLA typing by NGS, and a fault-tolerant design is required to prevent misinterpreting a pseudogene or chimeric PCR product as the target allele. Going forward, the current software architecture of GeMS-Analyzer allows for the capture and accumulation of frequency data on all alleles genotyped, which in turn will contribute to probability estimates and establishment of more powerful databases in support of genotyping. While the IMGT/HLA database is well established and largely complete beyond the addition of novel rare alleles, this approach can have a significant impact on the rapid establishment of new allele databases in support of other gene systems, such as KIR and others of potential immunological consequence [51].

### 4.3. Clinical use of the IGS

The IGS was developed to enable low- to mid-throughput clinical laboratories to utilize NGS and drive progress in precision medicine [52,53]. This transition will build upon a long history of

using genetic data in HCT clinical practice. In the short term, CILs will have the opportunity to adopt NGS for HLA by using the IGS and similar systems. Longer term opportunities presented by NGS include broadening the scope of precision medicine in CILs well beyond HLA. The laboratory and computational components of the IGS were designed to facilitate this process, enabling clinics to perform new genetic tests without having to install a new workflow, which by itself can be prohibitive in a smaller CIL setting. An NGS device such as the Illumina MiSeq can serve as a genetic testing platform that can generate a variety of results, in some ways analogous to a computer operating system on which diverse software programs can run. Furthering the analogy, the IGS is designed to be an expandable suite of programs, each providing a different genetic test result. For example, when coupled with the IGS a MiSeq run can be used to simultaneously obtain HLA and KIR data as this study demonstrated. This consolidation of genetic testing platforms is applicable to the typing of virtually any genetic locus including a variety of assays such as chimerism and minimal residual disease [54,55].

Through the use of NGS, CILs may play a central role in not only developing more effective donor–recipient matching criteria based on HLA and other immune loci, but also extending precision medicine beyond HCTs into treatments for other diseases. While the IGS is innovative in the CIL environment, the CILs themselves have been innovating in clinical medicine for decades. HLA typing advancements and corresponding improvements in HCT outcomes would simply not have arrived at the present state without the innovations brought to fruition by the CILs supporting HCTs in collaboration with clinical scientists. These partnerships have proven to be extremely effective in advancing the state of HLA typing and chimerism testing and revealing new outcome correlations. By providing these laboratories and their respective clinical environments with the ability to reduce costs and at the same time expand the repertoire of genetic testing, additional improvements to clinical outcomes are likely to come. The extensibility of amplicon-based protocols facilitates the ready addition of new gene testing components. The overall efficiency and ease of use of the general approach can be maintained, allowing CILs to participate in new genotyping applications without disrupting routine operations. Given this capability, physicians will be more likely to include novel tests in their clinical research protocols in order to meet emerging needs, and thereby accelerate the development of precision medicine.

## Conflict of interest

Daniel E. Geraghty is founder and owner of Scisco Genetics Inc. (Seattle, WA). Wyatt Nelson, Chul-Woo Pyo, David Vogan, and Ruihan Wang are employed by Scisco Genetics Inc. Shalini Pereira and Anajane G. Smith are consultants of Scisco Genetics Inc. The remaining authors declare no financial conflict of interest.

## Acknowledgments

We are especially grateful for the support provided by the entire staff, past and present, of the FHCRC laboratory that have contributed to HLA typing over the past 9 years. This work was supported in part by the National Institutes of Health Grant No. R42RR018669.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.humimm.2015.05.001.

## References

[1] Fernandez-Vina MA, Wang T, Lee SJ, Haagenson M, Aljurf M, Askar M, et al. Identification of a permissible HLA mismatch in hematopoietic stem cell transplantation. Blood 2014;123:1270.
[2] Petersdorf EW, Malkki M, Hsu K, Bardy P, Cesbron A, Dickinson A, et al. 16th IHIW: international histocompatibility working group in hematopoietic cell transplantation. Int. J. Immunogenet. 2013;40:2.
[3] Furst D, Solgi G, Schrezenmeier H, Mytilineos J. The frequency of DRB1*1454 in South German Caucasians. Tissue Antigens 2010;76:57.
[4] Lee SJ, Klein J, Haagenson M, Baxter-Lowe LA, Confer DL, Eapen M, et al. High-resolution donor–recipient HLA matching contributes to the success of unrelated donor marrow transplantation. Blood 2007;110:4576.
[5] Spellman SR, Eapen M, Logan BR, Mueller C, Rubinstein P, Setterholm MI, et al. A perspective on the selection of unrelated donors and cord blood units for transplantation. Blood 2012;120:259.
[6] Dehn J, Arora M, Spellman S, Setterholm M, Horowitz M, Confer D, et al. Unrelated donor hematopoietic cell transplantation: factors associated with a better HLA match. Biol. Blood Marrow Transplant. 2008;14:1334.
[7] MacMillan ML, Davies SM, Nelson GO, Chitphakdithai P, Confer DL, King RJ, et al. Twenty years of unrelated donor bone marrow transplantation for pediatric acute leukemia facilitated by the National Marrow Donor Program. Biol. Blood Marrow Transplant. 2008;14:16.
[8] Sasazuki T, Juji T, Morishima Y, Kinukawa N, Kashiwabara H, Inoko H, et al. Effect of matching of class I HLA alleles on clinical outcome after transplantation of hematopoietic stem cells from an unrelated donor. Japan Marrow Donor Program. N. Engl. J. Med. 1998;339:1177.
[9] Hansen JA, Gooley TA, Martin PJ, Appelbaum F, Chauncey TR, Clift RA, et al. Bone marrow transplants from unrelated donors for patients with chronic myeloid leukemia. N. Engl. J. Med. 1998;338:962.
[10] Venepalli N, Rezvani K, Mielke S, Savani BN. Role of allo-SCT for CML in 2010. Bone Marrow Transplant. 2010;45:1579.
[11] Latham K, Little AM, Madrigal JA. An overview of HLA typing for hematopoietic stem cell transplantation. Methods Mol. Biol. 2014;1109:73.
[12] Voorter CE, Palusci F, Tilanus MG. Sequence-based typing of HLA: an improved group-specific full-length gene sequencing approach. Methods Mol. Biol. 2014;1109:101.
[13] Grumbt B, Eck SH, Hinrichsen T, Hirv K. Diagnostic applications of next generation sequencing in immunogenetics and molecular oncology. Transfus. Med. Hemother. 2013;40:196.
[14] McCann SR, Crampe M, Molloy K, Lawler M. Hemopoietic chimerism following stem cell transplantation. Transfus. Apher. Sci. 2005;32:55.
[15] Aydin F, Oguz F, Onal AE, Bilgen H, Oguz R, Anak S, et al. FISH detection of chimerism in pediatric hematopoietic stem cell transplantation. Int. J. Lab. Hematol. 2007;29:208.
[16] Kwon M, Martinez-Laperche C, Infante M, Carretero F, Balsalobre P, Serrano D, et al. Evaluation of minimal residual disease by real-time quantitative PCR of Wilms' tumor 1 expression in patients with acute myelogenous leukemia after allogeneic stem cell transplantation: correlation with flow cytometry and chimerism. Biol. Blood Marrow Transplant. 2012;18:1235.
[17] Erlich H. HLA DNA typing: past, present, and future. Tissue Antigens 2012;80:1.
[18] Danzer M, Niklas N, Stabentheiner S, Hofer K, Proll J, Stuckler C, et al. Rapid, scalable and highly automated HLA genotyping using next-generation sequencing: a transition from research to diagnostics. BMC Genomics 2013;14:221.
[19] Niklas N, Proll J, Danzer M, Stabentheiner S, Hofer K, Gabriel C. Routine performance and errors of 454 HLA exon sequencing in diagnostics. BMC Bioinformatics 2013;14:176.
[20] Wang C, Krishnakumar S, Wilhelmy J, Babrzadeh F, Stepanyan L, Su LF, et al. High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc. Natl. Acad. Sci. USA 2012;109:8676.
[21] Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, Slavich L, et al. Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum. Immunol. 2010;71:1033.
[22] Ehrenberg PK, Geretz A, Baldwin KM, Apps R, Polonis VR, Robb ML, et al. High-throughput multiplex HLA genotyping by next-generation sequencing using multi-locus individual tagging. BMC Genomics 2014;15:864.
[23] Lind C, Ferriola D, Mackiewicz K, Papazoglou A, Sasson A, Monos D, et al. Filling the gaps – the generation of full genomic sequences for 15 common and well-documented HLA class I alleles using next-generation sequencing technology A multi-site study using high-resolution HLA genotyping by next generation sequencing. Hum. Immunol. 2013;74:325.
[24] Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics 2013;14:355.
[25] Lange V, Bohme I, Hofmann J, Lang K, Sauter J, Schone B, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics 2014;15:63.
[26] Smith AG, Pyo CW, Nelson W, Gow E, Wang R, Shen S, et al. Next generation sequencing to determine HLA class II genotypes in a cohort of hematopoietic cell transplant patients and donors. Hum. Immunol. 2014;75:1040.
[27] Moonsamy PV, Williams T, Bonella P, Holcomb CL, Hoglund BN, Hillman G, et al. High throughput HLA genotyping using 454 sequencing and the Fluidigm

Access Array System for simplified amplicon library preparation. Tissue Antigens 2013;81:141.
[28] Chang CX, Tan AT, Or MY, Toh KY, Lim PY, Chia AS, et al. Conditional ligands for Asian HLA variants facilitate the definition of CD8+ T-cell responses in acute and chronic viral diseases. Eur. J. Immunol. 2013;43:1109.
[29] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53.
[30] Gabriel C, Furst D, Fae I, Wenda S, Zollikofer C, Mytilineos J, et al. HLA typing by next-generation sequencing –- getting closer to reality. Tissue Antigens 2014;83:65.
[31] De Santis D, Dinauer D, Duke J, Erlich HA, Holcomb CL, Lind C, et al. 16(th) IHIW : review of HLA typing by NGS. Int. J. Immunogenet. 2013;40:72.
[32] Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature 2011;475:348.
[33] Holcomb CL, Hoglund B, Anderson MW, Blake LA, Bohme I, Egholm M, et al. A multi-site study using high-resolution HLA genotyping by next generation sequencing. Tissue Antigens 2011;77:206.
[34] Erlich RL, Jia X, Anderson S, Banks E, Gao X, Carrington M, et al. Next-generation sequencing for HLA typing of class I loci. BMC Genomics 2011;12:42.
[35] Trachtenberg EA, Holcomb CL. Next-generation HLA sequencing using the 454 GS FLX system. Methods Mol. Biol. 2013;1034:197.
[36] Saavedra-Matiz CA, Isabelle JT, Biski CK, Duva SJ, Sweeney ML, Parker AL, et al. Cost-effective and scalable DNA extraction method from dried blood spots. Clin. Chem. 2013;59:1045.
[37] Adey A, Morrison HG, Asan Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 2010;11:R119.
[38] Holcomb CL, Rastrou M, Williams TC, Goodridge D, Lazaro AM, Tilanus M, et al. Next-generation sequencing can reveal in vitro-generated PCR crossover products: some artifactual sequences correspond to HLA alleles in the IMGT/HLA database. Tissue Antigens 2014;83:32.
[39] Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SG. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. 2015;43:D423.
[40] Smith TF, Waterman MS. Identification of common molecular subsequences. J. Mol. Biol. 1981;147:195.
[41] Wittig M, Anmarkrud JA, Kassens JC, Koch S, Forster M, Ellinghaus E, et al. Development of a high-resolution NGS-based HLA-typing and analysis pipeline. Nucleic Acids Res. 2015.
[42] Pyo CW, Wang R, Vu Q, Cereb N, Yang SY, Duh FM, et al. Recombinant structures expand and contract inter and intragenic diversification at the KIR locus. BMC Genomics 2013;14:89.
[43] Pyo CW, Guethlein LA, Vu Q, Wang R, Abi-Rached L, Norman PJ, et al. Different patterns of evolution in the centromeric and telomeric regions of group A and B haplotypes of the human killer cell Ig-like receptor locus. PLoS One 2010;5:e15115.
[44] Dommaraju K, Kijak G, Carlson JM, Larsen BB, Tovanabutra S, Geraghty DE, et al. CD8 and CD4 epitope predictions in RV144: no strong evidence of a T-cell driven sieve effect in HIV-1 breakthrough sequences from trial participants. PLoS One 2014;9:e111334.
[45] Gartland AJ, Li S, McNevin J, Tomaras GD, Gottardo R, Janes H, et al. Analysis of HLA A*02 association with vaccine efficacy in the RV144 HIV-1 vaccine trial. J. Virol. 2014;88:8242.
[46] Siva N. UK gears up to decode 100000 genomes from NHS patients. Lancet 2014.
[47] Torjesen I. Genomes of 100,000 people will be sequenced to create an open access research resource. BMJ 2013;347:f6690.
[48] Cox ST, McWhinnie AJ, Robinson J, Marsh SG, Parham P, Madrigal JA, et al. Cloning and sequencing full-length HLA-B and -C genes. Tissue Antigens 2003;61:20.
[49] Pohl M, Bortfeldt RH, Grutzmann K, Schuster S. Alternative splicing of mutually exclusive exons – a review. Biosystems 2013;114:31.
[50] Chen CY. DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present. Front. Microbiol. 2014;5:305.
[51] Li SS, Gilbert PB, Tomaras GD, Kijak G, Ferrari G, Thomas R, et al. FCGR2C polymorphisms associate with HIV-1 vaccine protection in RV144 trial. J. Clin. Invest. 2014;124:3879.
[52] UK catapults precision medicine, Nat. Biotechnol. 33 (2015) 119.
[53] Collins FS, Varmus H. A new initiative on precision medicine. N. Engl. J. Med. 2015.
[54] Pagani IS, Spinelli O, Mattarucchi E, Pirrone C, Pigni D, Amelotti E, et al. Genomic quantitative real-time PCR proves residual disease positivity in more than 30% samples with negative mRNA-based qRT-PCR in chronic myeloid leukemia. Oncoscience 2014;1:510.
[55] Rieneck K, Bak M, Jonson L, Clausen FB, Krog GR, Tommerup N, et al. Next-generation sequencing: proof of concept for antenatal prediction of the fetal Kell blood group phenotype from cell-free fetal DNA in maternal plasma. Transfusion 2013;53:2892.

Please cite this article in press as: Nelson WC et al. An integrated genotyping approach for HLA and other complex genetic systems. Hum Immunol (2015), http://dx.doi.org/10.1016/j.humimm.2015.05.001