Clustering and Alignment of Polymorphic Sequences for 

HLA-DRB1 Genotyping

Steven Ringquist1*, Gaia Bellone2, Ying Lu1, Kathryn Roeder2, Massimo Trucco1

1 Department of Pediatrics, Division of Immunogenetics, Children’s Hospital of Pittsburgh of UPMC, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America, 2 Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania, United States of America

Abstract

Located on Chromosome 6p21, classical human leukocyte antigen genes are highly polymorphic. HLA alleles associate with a variety of phenotypes, such as narcolepsy, autoimmunity, as well as immunologic response to infectious disease. Moreover, high resolution genotyping of these loci is critical to achieving long-term survival of allogeneic transplants. Development of methods to obtain high resolution analysis of HLA genotypes will lead to improved understanding of how select alleles contribute to human health and disease risk. Genomic DNAs were obtained from a cohort of n = 383 subjects recruited as part of an Ulcerative Colitis study and analyzed for HLA-DRB1. HLA genotypes were determined using sequence specific oligonucleotide probes and by next-generation sequencing using the Roche/454 GSFLX instrument. The Clustering and Alignment of Polymorphic Sequences (CAPSeq) software application was developed to analyze next-generation sequencing data. The application generates HLA sequence specific 6-digit genotype information from next-generation sequencing using MUMmer to align sequences and the R package diffusionMap to classify sequences into their respective allelic groups. The incorporation of Bootstrap Aggregating, Bagging to aid in sorting of sequences into allele classes resulted in improved genotyping accuracy. Using Bagging iterations equal to 60, the genotyping results obtained using CAPSeq when compared with sequence specific oligonucleotide probe characterized 4-digit genotypes exhibited high rates of concordance, matching at 759 out of 766 (99.1%) alleles.

Introduction

The classical human leukocyte antigen (HLA) loci -A, -B, and -C (class I) as well as -DRB1, -DQB1, and -DPB1 (class II) are expressed on the cell surface and function as antigen-presenting proteins. HLA loci were first identified during studies involving transplantation of tumor cells among different strains of mice, and later shown to be the principle genetic barrier to allogeneic transplantation of cells, tissues, and organs [1]. In humans, these genes are clustered within a roughly 4 Mb region on Chromosome 6p21 [2]. The genes encoding the HLA proteins are characterized by frequent polymorphisms, occurring as often as 1 out 8 nucleotides, resulting in greater than 7,800 characterized alleles and more than 5,700 protein variants [3]. Select alleles have been implicated in pharmacogenomics [4] and human diseases, associating with immune response to infectious agents as well as autoimmunity and narcolepsy [2].

The impact of HLA diversity on transplant genetics, the requirement to match HLA alleles between unrelated donor and recipient, is that there is currently an estimated 60% chance of finding a matched donor-recipient pair when genotyping for HLA-A, -B, and -DRB1 [5,6]. Timely matching of donor with recipient, therefore, depends upon ongoing recruitment of volunteer donors. Currently, screening of HLA genotypes from increasingly large cohorts is achieved using the sequence specific oligonucleotide (SSO) typing approach. In this methodology, genomic DNA is PCR amplified and the product detected by hybridization of allele specific probes [7]. When determining 4-digit resolution (e.g., HLA-DRB1*15:01 and denoting the protein sequence subtype) the SSO method is highly accurate and can be performed rapidly [8]. For example, the overall accuracy of SSO based genotyping of HLA-DRB1 is estimated at roughly 99% based upon blinded analysis of 1,652 reference samples [8], however, full resolution genotyping of alleles defining substitutions occurring among synonymous codons as well as within introns and untranslated regions are not routinely achieved.

In contrast, full resolution genotyping of HLA loci can be achieved by next-generation sequencing based protocols [9,10]. Since it was first developed pyrosequencing technology has advanced substantially allowing it to be used in ways that enable efficient preparation of thousands of individually prepared clones for analysis within a single sequencing run, scaling of the reaction to picoliter volumes, and increasing read lengths to several hundred nucleotides [11]. As the length of the polymorphic exons of HLA class I (exons 2 and 3) and class II loci (exon 2) are between 264 base pairs for HLA-DPB1 exon 2 and 276 base pairs for exon 3 of HLA-A, -B, and -C the method is capable of sequence based typing of HLA alleles. In fact, a number of research laboratories have already reported achieving high resolution typing of HLA class I and class II loci using genomic DNA isolated from human cell lines as well as blood samples [9,10,12,13] and from cDNA.
Materials and Methods

Ethics Statement

All subjects were recruited for genetic studies at the University of Pittsburgh under institutional review board approved protocols [18].

Materials

Residual DNA samples were selected from cohorts recruited as part of a study into the genetics of Ulcerative Colitis (UC). The UC cohort had been previously analyzed during a genome-wide scan of SNPs and the 4-digit HLA-DRB1 genotypes were inferred via the imputation of SNP’s approach and confirmed by SSO genotyping [18]. Individual DNA samples (n = 383), representing a subset of subjects recruited for the UC cohort, were chosen for next-generation sequencing in order to capture a wide range of alleles (Table 1). HLA-DRB1 alleles covered by the UC cohort accounted for 36 commonly occurring alleles, including 29 out of the 31 most frequently occurring alleles observed among subjects of European ancestry [19]. The cumulative frequency of allelic variance covered for -DRB1 was 99.5% among European populations and 84.3% among populations worldwide [19,20].

Amplicon Preparation

Forward barcoded oligonucleotide primer 5’-CCTATGGCGCTTCTGCCAGGCAATCAG(MID)CCGGATCTCCCTGGGTCCTCAGGACAC-3’ employing a multiplex identifier region (MID) and reverse primer 5’-CTATGGCGCTTCTGCCAGGCAATCAGGCTG-CACGTCGAGCTCTC-3’ were used to target HLA-DRB1 exon 2 for amplification, and were chimeras of titanium primers A and B along with HLA-DRB amplification primers [9]. DNA amplification was performed in 10 µl and used 40 ng genomic DNA, 400 nM oligonucleotide primers, 0.2 mM dNTP, and 1 unit FastStart High Fidelity Enzyme Blend (Roche Diagnostics). Thermal cycling was performed by incubation at 94°C for 3 minutes followed by 40 cycles at 94°C for 30 seconds, 61°C for 45 seconds, and 72°C for 1 minute. The reactions were then incubated at 72°C for 2 minutes.

Next-Generation Sequencing

Next-generation sequencing was performed using the Roche/454 titanium assay as described previously [18]. Briefly, emulsion of PCR reagents in microreactors was prepared by mixing beads, PCR reaction mix (1X amplification mix, Amplification Primers, 0.15 U/µl Platinum Taq [Invitrogen, Carlsbad, CA]), and emulsion oil and mixing vigorously using a Tissue Lyser (Qiagen, Valencia, CA). Emulsion was distributed into a PCR plate and template amplification was carried out in a thermocycler using the following cycling conditions: Hotstart activation for 4 minutes at 94°C, 40 cycles of 94°C for 30 seconds, 58°C for 1 minute, 68°C for 90 seconds followed by 13 cycles of 94°C for 30 seconds and 58°C for 6 minutes. Sequencing primers were added to the mixture of beads and annealing buffer and annealed to the template using the following thermocycler conditions: 65°C for 5 minutes, ramp to 30°C at 0.1°C/second, hold at 50°C for 1 minute, ramp to 40°C at 0.1°C/second, hold at 40°C for 1 minute, ramp to 15°C at 0.1°C/second, hold 15°C. Packing beads, sample beads and enzyme beads were applied to the picotiter plate as per manufacturer instructions. The sequencing reaction was performed in picotiter plates loaded onto the sequencer. The Roche/454 instrument used pyrosequencing chemistry and detected the incorporation of each nucleotide in real time.

CAPSeq Genotyping

HLA genotypes were determined from sequence data using the CAPSeq software application (Figure 1). The CAPSeq application uses next-generation sequences and their corresponding Q-Scores to enable alignment followed by classification of sequences into their corresponding allelic components. The consensus sequences are then compared with known HLA alleles [3] and the nearest match reported as the genotype. The application is written in R and PERL and can be run using the R command source (“CAPSeq.R”). The software is compatible with Linux and Mac operating systems.

Online Software and Sample Data

The CAPSeq software and HLA-DRB1 sample data are available at (http://wpicr.wpic.pitt.edu/WPICCompGen/) and consist of four R Scripts CAPSeq.R, CompleteTemp.R, diffuse.R, GenoFuns1.R; the six PERL Scripts CreateReadsQscore.pl, Edit_Alignment.pl, Edit_Numeric_SNP.pl, Format_Delta_Files.pl, Identity_Score.pl, Matrix_Check.pl and the HLA-DRB1 Sequence File Library_DRB_Release370_unique_exon2. Sample Data obtained from five subjects are available in the folder DRB_Example and are formatted for use with the CAPSeq application. All new data have been deposited in GenBank (SUB162897). The ReadMe file provides directions for installing and running the CAPSeq application.

Results

Next-generation Sequencing

Next-generation sequencing analysis of HLA loci -DRB1 exon 2 was performed using the Roche/454 GSFLX instrument and resulted in 371,657 sequences from the UC cohort of 383 subjects. Sequencing assays were performed so as to obtain data from both strands. Upon analysis of read lengths for those sufficient to enable direct analysis of exon 2, that is, without the need to assemble truncated DNA sequences there were 368,610 sequences remaining, accounting for 99% of the total. After data cleaning in which sequence barcodes were used to identify subjects [21] and sequence motifs found within HLA exons were used to identify the location of the -DRB1 exon start and finish boundaries there were 330,981 (89%) of the sequences remaining. Due to the complexity of the HLA-DRB sub-region in which amplicons in addition to those at the HLA-DRB1 locus were generated the data were subjected to additional cleaning steps designed to minimize the occurrence of sequences containing motifs found only in HLA-DRB pseudogenes (e.g., HLA-DRB6/7/8/9) and paralogous loci (i.e., HLA-DRB3/4/5). After data cleaning there were 130,513 (35%) sequences remaining corresponding to a mean and standard deviation of 340±132 sequences per subject.

HLA-DRB1 Genotyping

The Clustering and Alignment of Polymorphic Sequences (CAPSeq) software makes use of statistical analysis for calling each genotype. The CAPSeq application incorporates MUMmer [22] to align sequences and the R package diffusionMap [23,24] to classify the sequence data into allelic groups. The genotyping process is organized into three main steps (Figure 1). The first step consists in identifying the clusters of reads generated by the
different alleles. To do this, a pairwise weighted similarity score is computed as the difference between the alignment length and the number of mismatches among the aligned sequences. At this stage the information about the variable regions are incorporated via their overall sequence quality score, giving lower weight to the mismatches belonging to these regions (e.g., when Q-Scores are less than 30). Second, Diffusion Mapping [23,24] is used to generate a data transformation that can simplify cluster structures and protect from the presence of outliers (e.g., sequence reads containing infrequent variants generated during PCR amplification). The data representing the pairwise weighted similarity score are projected into diffusion space in which the Euclidean distance between two points is small if the points are highly connected in the original feature space and large otherwise. Then the K-means algorithm is applied and the reads are clustered into homogeneous groups. During this process the consensus sequence for each sequenced cluster is established. The genotype is determined by comparing the consensus sequence with that of known HLA allele sequences obtained from the HLA/IMGT database [3]. A third step is performed for ensuring classification precision. To this end Bootstrap Aggregating, Bagging, is used [25]. Bagging is a machine learning technique that generates multiple versions of the same

| HLA-DRB1 | European Freq (Rank) | Worldwide Freq (Rank) | UC Cohort Freq |
|----------|-----------------------|-----------------------|---------------|
| *01:01   | 0.09149 (4)           | 0.04123 (8)           | 0.07311       |
| *01:02   | 0.01703 (13)          | 0.01161 (26)          | 0.02872       |
| *01:03   | 0.00889 (19)          | 0.00329 (41)          | 0.02480       |
| *03:01   | 0.12916 (3)           | 0.06760 (3)           | 0.10183       |
| *04:01   | 0.09111 (5)           | 0.02896 (13)          | 0.07963       |
| *04:02   | 0.00972 (17)          | 0.00742 (32)          | 0.00392       |
| *04:03   | 0.00572 (23)          | 0.02659 (17)          | 0.00392       |
| *04:04   | 0.03634 (9)           | 0.01795 (19)          | 0.02219       |
| *04:05   | 0.00368 (25)          | 0.04776 (6)           | 0.00522       |
| *04:07   | 0.00947 (18)          | 0.01536 (23)          | 0.00783       |
| *04:08   | 0.00248 (26)          | 0.00188 (43)          | 0.00261       |
| *07:01   | 0.13767 (2)           | 0.06986 (2)           | 0.10574       |
| *08:01   | 0.02363 (12)          | 0.00875 (29)          | 0.03525       |
| *08:02   | 0.00025 (34)          | 0.02104 (17)          | 0.00131       |
| *08:03   | 0.00133 (29)          | 0.03864 (9)           | 0.00261       |
| *08:04   | 0.00006 (31)          | 0.00500 (36)          | 0.00392       |
| *08:06   | 0.00006 (41)          | 0.00101 (47)          | 0.00131       |
| *08:10   | 0.00000 (NA)          | 0.00003 (68)          | 0.00131       |
| *09:01   | 0.00820 (21)          | 0.05450 (5)           | 0.01175       |
| *10:01   | 0.00826 (20)          | 0.01284 (24)          | 0.01044       |
| *11:01   | 0.05654 (7)           | 0.05945 (4)           | 0.08747       |
| *11:02   | 0.00152 (28)          | 0.00604 (33)          | 0.00131       |
| *11:03   | 0.00483 (24)          | 0.00227 (42)          | 0.00522       |
| *11:04   | 0.03189 (10)          | 0.01780 (20)          | 0.04178       |
| *11:14   | 0.00000 (NA)          | 0.00003 (68)          | 0.00131       |
| *11:15   | 0.00000 (NA)          | 0.00000 (NA)          | 0.00131       |
| *11:29   | 0.00000 (NA)          | 0.00002 (69)          | 0.00131       |
| *12:01   | 0.01468 (14)          | 0.02712 (16)          | 0.03003       |
| *13:01   | 0.06283 (6)           | 0.03152 (12)          | 0.05875       |
| *13:02   | 0.04015 (8)           | 0.03746 (10)          | 0.03655       |
| *13:03   | 0.00991 (16)          | 0.00784 (30)          | 0.01436       |
| *13:27   | 0.00000 (NA)          | 0.00000 (NA)          | 0.00131       |
| *14:01   | 0.02459 (11)          | 0.03218 (11)          | 0.02742       |
| *15:01   | 0.14441 (1)           | 0.07864 (1)           | 0.13316       |
| *15:02   | 0.00775 (22)          | 0.04507 (7)           | 0.00653       |
| *16:01   | 0.01061 (15)          | 0.01656 (21)          | 0.02480       |

**Study Panel Total** 0.995 0.843

Allele frequencies and rankings are taken from Maiers et al. [19] for European -DRB1 and from Lancaster et al. [20] for worldwide frequencies. UC cohort frequencies are determined from CAPSeq genotyping results.

doi:10.1371/journal.pone.0059835.t001

PLOS ONE | www.plosone.org 3 March 2013 | Volume 8 | Issue 3 | e59835
Input Data
- Modified FASTQ Formatted Sequences
- Known HLA Allele Sequences

CAPSeq Application
Step 1. Calculate Weighted Pairwise Similarity Score
- Sample N sequences and corresponding Q-scores
- Align sequences using MUMmer software
- Determine weighted pairwise similarity score

Step 2. Identify Sequence Clusters
- Apply Diffusion Mapping data transformation
- Apply K-means to identify homogeneous sequence groups
- Align sequences within each group
- Determine consensus sequence for each group
- Determine HLA identity of each consensus

Step 3. Ensure Genotyping Precision
- Apply Bootstrap Aggregating by reiterations of Steps 1 and 2

Output Data
- HLA Genotype, Call Rate, Average Cluster Size, Average Identity Score

Table 1. The Clustering and Alignment of Polymorphic Sequences (CAPSeq) software application illustrated as a schematic. Input Data: Next-generation sequence data formatted as modified FASTQ files consisting of sequences and corresponding Q-scores along with an additional input data file containing known HLA allele sequences. CAPSeq Application: The analysis software can be broken down into 3 principle steps consisting of those developed to align sequences and use corresponding Q-scores to generate a weight-ed pairwise similarity score (step 1) that can be analyzed via diffusion mapping, followed by K-means clustering to enable the identification of homogeneous sequence groups (step 2) followed by Bootstrap Aggregating, Bagging, of multiple analyses of the data to ensure genotyping precision (step 3). Output Data: The HLA genotyping data is provided as a tab delimited text file containing the most likely allelic match between the CAPSeq generated consensus sequences and list of known HLA alleles.

doi:10.1371/journal.pone.0059835.g001

Figure 1. The Clustering and Alignment of Polymorphic Sequences (CAPSeq) software application illustrated as a schematic. Input Data: Next-generation sequence data formatted as modified FASTQ files consisting of sequences and corresponding Q-scores along with an additional input data file containing known HLA allele sequences. CAPSeq Application: The analysis software can be broken down into 3 principle steps consisting of those developed to align sequences and use corresponding Q-scores to generate a weighted pairwise similarity score (step 1) that can be analyzed via diffusion mapping, followed by K-means clustering to enable the identification of homogeneous sequence groups (step 2) followed by Bootstrap Aggregating, Bagging, of multiple analyses of the data to ensure genotyping precision (step 3). Output Data: The HLA genotyping data is provided as a tab delimited text file containing the most likely allelic match between the CAPSeq generated consensus sequences and list of known HLA alleles.

doi:10.1371/journal.pone.0059835.g001

Table 2. Comparison of SSO based 4-digit HLA-DRB1 genotyping with CAPSeq.

| Bagging Iterations | Concordant | Sensitivity Error | Specificity Error |
|--------------------|------------|-------------------|-------------------|
| 5                  | 733 (95.7%)| 31 (4.0%)         | 2 (0.3%)          |
| 10                 | 747 (97.5%)| 17 (2.2%)         | 2 (0.3%)          |
| 20                 | 752 (98.2%)| 11 (1.4%)         | 3 (0.4%)          |
| 40                 | 758 (99.0%)| 6 (0.8%)          | 2 (0.3%)          |
| 60                 | 759 (99.1%)| 5 (0.7%)          | 2 (0.3%)          |

doi:10.1371/journal.pone.0059835.t002

The genotyping results obtained using CAPSeq software are summarized in Table 2. Analysis was performed in quadruplicate at increasing Bagging iterations from 5 to 60 times. A greater number of iterations lead to improved identification of the HLA-DRB1 genotypes. For example, the rate of concordance between CAPSeq and SSO determined genotypes improved with increasing Bagging from 95.7% at Bagging 5 to 99.1% when Bagging was performed 60 times. The improvement resulted from a decreased number of sensitivity errors occurring when a genotype determined as heterozygous via the SSO approach was reported as homozygous during CAPSeq analysis of the next-generation sequencing data. For instance, there was a roughly 6-fold decrease in sensitivity errors when Bagging iterations were increased with 5 persistent sensitivity errors occurring at Bagging 60 compared to 31 sensitivity errors observed at Bagging 5 (Table 2).

Table 2. Comparison of SSO based 4-digit HLA-DRB1 genotyping with CAPSeq.

| Bagging Iterations | Concordant | Sensitivity Error | Specificity Error |
|--------------------|------------|-------------------|-------------------|
| 5                  | 733 (95.7%)| 31 (4.0%)         | 2 (0.3%)          |
| 10                 | 747 (97.5%)| 17 (2.2%)         | 2 (0.3%)          |
| 20                 | 752 (98.2%)| 11 (1.4%)         | 3 (0.4%)          |
| 40                 | 758 (99.0%)| 6 (0.8%)          | 2 (0.3%)          |
| 60                 | 759 (99.1%)| 5 (0.7%)          | 2 (0.3%)          |

doi:10.1371/journal.pone.0059835.t002

In contrast, persistent specificity errors were reported for 2 subjects (Table 2). In each instance inspection of individual sequences was consistent with the CAPSeq called genotype. For example, Subject 150 was identified as HLA-DRB1*03:01, *11:01 versus HLA-DRB1*03:01:01G, *11:01:01G while Subject 307 was genotyped as HLA-DRB1*11:04, *13:01 versus HLA-DRB1*11:01:01G, *13:01:01G by SSO and CAPSeq based methods, respectively (Supplementary Table S1). The CAPSeq identified genotype was strongly supported by the underlying sequence data accounting for 47% (236 out of 500) and 45% (157 out of 350) of the total number of sequences obtained from Subjects 150 and 307, and indicating the likely possibility that CAPSeq provided genotypes are, in fact, correct for these samples.

The overall quality of the genotyping results were evaluated by comparing the frequency of alleles observed when analyzing next-generation sequencing data against the expected frequency determined by the SSO genotyping method (Figure 3). For example, the frequency of genotypes determined using the SSO and CAPSeq methods exhibited strong correlation with a Pearson’s
correlation coefficient $r>0.999$ when the SSO genotypes and Bagging 60 results were compared (Figure 3). Moreover, as summarized in Table 1, the observed frequency of individual CAPSeq determined alleles were highly correlated with the expected frequencies of these $HLA-DRB1$ alleles observed for subjects of European ancestry and to a lesser extent with worldwide allele frequencies, exhibiting Pearson’s correlation coefficients $r>0.97$ and 0.75, consistent with the expected validity of the CAPSeq analysis approach.

Discussion

The $HLA$ loci constitute the most highly polymorphic genetic system in humans [2]. A stable, inherited polymorphism, the alleles, gives rise to alternative forms of the protein with 5,764 variants reported [3]. The molecular basis for the $HLA$ polymorphism resides in nucleotide sequence differences present in the coding regions of the HLA gene. Nearly all the $HLA$ loci have various alleles with $HLA-DRB1$ reported to have 1,260 distinct forms [3]. During recent years, substantial progress has been made in converting the protocols for molecular $HLA$ genotyping to next-generation sequencing technologies [28]. For example, the choice of the Roche/454 GSFLX instrument allows direct sequencing of polymorphic $HLA$ exons without the need to assemble partial reads. When compared with the SSO-based method, the next-generation sequencing approach to $HLA$ genotyping provides increased accuracy and sample throughput [28]. Moreover, a number of publications have reported achieving high resolution typing of $HLA$ loci using genomic DNA isolated from human cell lines, blood samples, as well as from cDNA created from isolations of mRNA [9,10,12,14].

The CAPSeq software application was developed as an open source solution for genotyping of $HLA$ alleles, offering a standalone approach to $HLA$ genotyping. The CAPSeq analysis method has been developed initially for genotyping $-DRB1$ alleles but has also been used with $-DPB1$ providing similar levels of accuracy (data not shown). In contrast, computational methods for analysis of next-generation sequencing data generated on the Roche/454 GSFLX instrument are commercially available [17,28]. In addition, use of the Genome Analysis Tool Kit (GATK) suite of open source software applications for $HLA$ typing has also been described [16] and is available as source code but unfortunately is no longer supported by the GATK help desk [29]. Recently, Illumina based next-generation sequencing of $HLA$ loci has also been reported with similar levels of accuracy [30]. Like CAPSeq, these various methods for analysis of next-generation sequences for $HLA$ genotypes have reported a roughly 99% concordance when compared with test samples [16,17], a threshold that may increase when subjects are examined directly via cloning and traditional Sanger based sequencing. The CAPSeq application uses the depth of sequencing provided by the next-generation based method to enable the R module diffusionMap in order to classify sequences based upon their similarity score. This feature results in the classification of sequences into their respective alleles. Along with MUMmer, used to align sequences, the application generates a consensus for each sequence class. The consensus sequence is used to identify known $HLA$ genotypes with sequence identity scores approaching 100%. The inclusion of Bagging improves the

![Figure 2. CAPSeq Bagging iterations result in improved genotyping sensitivity.](doi:10.1371/journal.pone.0059835.g002)
frequency of concordant calls in excess of 99% when compared with the results of SSO based typing. The application is freely available (http://wpic.wpirc.pitt.edu/WPICCompGen/) and consists of R and PERL scripts as well as sample data and allele reference files for HLA-DRB1.

Supporting Information

Table S1 HLA-DRB1 genotypes determined by SSO and CAPSeq analyses of next-generation sequencing data. The table lists the Subject Identifier (column 1); Classification whether the SSO and CAPSeq results were concordant or discordant (column 2); Description of the error type or genotype (column 3); the 4-digit and 6-digit SSO and CAPSeq derived genotypes (columns 4 through 7); along with the CAPSeq derived values for the frequency with which individual alleles were called during the Bagging steps, Call Rate (columns 8 and 9), the average number of times each allele was observed when sampling the sequence data, Average Cluster Size (columns 10 and 11); and how closely each CAPSeq derived consensus sequence matched the sequence of the closest known HLA-DRB1 allele, Average Identity Score (Columns 12 and 13).

Acknowledgments

We are grateful to the University of Pittsburgh Genomics and Proteomics Core Laboratories for assistance with next-generation sequencing.

Author Contributions

Conceived and designed the experiments: GB KR MT SR. Performed the experiments: GB YL. Analyzed the data: GB KR SR YL. Contributed reagents/materials/analysis tools: GB KR SR YL. Wrote the paper: KR MT SR.

References

1. Rossini AA, Greiner DL, Mordes JP (1999) Induction of immunologic tolerance for transplantation. Physiol Rev 79: 99–141.
2. Shina T, Hesomichi K, Itosu H, Kushi JK (2009) The HLA genomic loci map: expression, interaction, diversity and disease. J Hum Genet 54: 15–39.
3. Robinson J, Mistry K, McWilliam H, Lopez R, Parham P, et al. (2011) The IMGT/HLA database. Nucleic Acids Res 39: D1171–1176.
4. Recquermont L (2010) HLA: a pharmacogenomics success story. Pharmacogenomics 11: 277–281.
5. Karanes C, Nelson GO, Chirikukahitai P, Agura E, Ballen KK, et al. (2008) Twenty years of unrelated donor hematopoietic cell transplantation for adult recipients facilitated by the National Marrow Donor Program. Biol Blood Marrow Transplant 14: 8–15.
6. Ballen KK, Spitzer JR (2011) The great debate: haploidentical or cord blood transplant. Bone Marrow Transplant 46: 323–329.
7. Dunbar SA (2006) Applications of Luminex xMAP technology for rapid, high-throughput multiplexed nucleic acid detection. Clin Chim Acta 363: 71–82.
8. Ng J, Hurley CK, Baxter-Lowe LA, Chopek M, Coppo PA, et al. (1993) Large-scale oligonucleotide typing for HLA-DRB1/3/4 and HLA-DQB1 is highly accurate, specific, and reliable. Tissue Antigens 42: 473–479.
9. Bentley G, Higuchi R, Hoglund B, Goodridge D, Sayer D, et al. (2009) High-resolution, high-throughput HLA genotyping by next-generation sequencing. Tissue Antigens 74: 393–403.
10. Gabriel C, Danzer M, Hackl G, Kopal G, Huflaig P, et al. (2009) Rapid high-throughput human leukocyte antigen typing by massively parallel pyrosequencing for high-resolution allele identification. Hum Immunol 70: 960–964.
11. Rothberg JM and Leamon JH (2008) The development and impact of 454 sequencing. Nat Biotechnol 26: 1117–1124.
12. Lind C, Ferriola D, Mackiewicz K, Heron S, Rogers M, et al. (2010) Next-generation sequencing: the solution for high-resolution, unambiguous human leukocyte antigen typing. Hum Immunol 71: 1033–1042.
13. Shima T, Suzuki S, Ozaki Y, Taira H, Kakkawa E, et al. (2012) Super high resolution for single molecule-sequence-based typing of classical HLA loci at the 8-digit level using next generation sequencers. Tissue Antigens 80: 305–316.
14. Lank SM, Wiseman RW, Dudley DM, O’Connor DH (2010) A novel single cDNA amplicon pyrosequencing method for high-throughput, cost-effective sequence-based HLA class I genotyping. Hum Immunol 71: 1011–1017.
15. Lank SM, Golbach BA, Creager HM, Wiseman RW, Keskin DB, et al. (2012) Ultra-high resolution HLA genotyping and allele discovery by highly multiplexed cDNA amplicon pyrosequencing. BMC Genomics 13: 578.
16. Erlich RL, Jia X, Anderson S, Banke E, Guo X, et al. (2011) Next-generation sequencing for HLA typing of class I loci. BMC Genomics 18: 12–42.
17. Holcomb CL, Hoglund B, Anderson MW, Blake LA, Bohme I, et al. (2011) A multi-site study using high-resolution HLA genotyping by next generation sequencing. Tissue Antigens 77: 206–217.
18. Achkar JP, Klei L, de Bakker PIW, Bellone G, Rebert N, et al. (2012) Amino Acid Position 11 of HLA-DRβ1 is a Major Determinant of Chromosome 6p Association with Ulcerative Colitis. Genes and Immunity 13: 245–252.
19. Maier S, Gragert L, Kitz W (2007) High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol 68: 779–788.
20. Lancaster AK, Single RM, Solberg OD, Nelson MP, Thomson G (2007) PyPop update - a software pipeline for large-scale multifocus population genomics. Tissue Antigens 69: 192–197.
21. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat Methods 5(3): 235–237.
22. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.
23. Richards J, Freeman P, Lee A, Schafer C (2009) Exploiting low-dimensional structure in astronomical spectra. Astrophysical Journal 691: 32–42.
24. Lalun S and Lee AB (2006) Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Trans Pattern Anal Mach Intell 28: 1393–1403.
25. Breiman L (1996) Bagging predictors. Machine Learning 24: 123–140.
26. Voorter CE, Kik MC, van den Berg-Loonen EM (1998) High-resolution HLA typing for the DQβ1 gene by sequence-based typing. Tissue Antigens 51: 80–87.
27. van Dijk A, Melchers R, Tilanus M, Rozenmuller E (2007) HLA-DQβ1 sequencing-based typing updated. Tissue Antigens 69 Suppl 1: 64–65.
28. Erlich H (2012) HLA DNA typing: past, present, and future. Tissue Antigens 80: 1–11.
29. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297–1303.
30. Wang C, Krishnakumar S, Willehrny J, Bahrzadeh F, Stepanyan L, et al. (2012) High-throughput, high-fidelity HLA genotyping with deep sequencing. Proc Natl Acad Sci U S A. 109: 8676–8681.