Bioinformatics Tools for Improving Matching for Hematopoietic Stem Cell Transplantation

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Michael S. Halagan

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Dr. Abeer Madbouly

August 2017

Acknowledgements

This work would not have been possible without the approval and support of the Ezer Mizion Bone Marrow Donor Registry. Sigal Manor and Jerry Stein were instrumental in setting up this partnership. I am grateful for the passion and determination they added to our collaboration.

This work would also not have been possible without the support of the National Marrow Donor Program. I am thankful for the support of my supervisors Yung-Tsi Bolon, Caleb Kennedy, and Martin Maiers.

I would like to thank my advisor Abeer Madbouly for motivating me to work on my thesis. She has been a source of encouragement and advice for as long as I have known her. She also played a critical role in the collaboration with Ezer Mizion, for which I am very grateful.

I would also like to thank Claudia Neuhauser for all of the support and advice she has given me throughout my graduate career. I have greatly appreciated how responsive and helpful she has been with all of the questions I have had along the way.

Abstract

Hematopoietic Stem Cell Transplantation (HSCT) is a curative therapy for multiple malignant and non-malignant blood disorders. Multiple opportunities exist for facilitating and improving the accuracy of matching potential donors with patients in need of HSCT. The global donor pool does not adequately represent many regions of the world; therefore, donor searches would benefit from the haplotype analysis and modeling of underserved populations. Utilizing sequence data in matching algorithms also has potential to improve HSCT for patients in need. We developed the Gene Feature Enumeration (GFE) ecosystem to supplement the current HLA nomenclature by retaining all sequence data, hence enhancing matching precision. To improve the global donor pool, we performed a haplotype frequency analysis and registry modeling on the Ezer Mizion registry in Israel. Combining all these bioinformatics tools provides invaluable resources for unrelated donor registries to help serve HSCT patients worldwide.

Table of Contents

List of Tables 5

List of Figures 6

INTRODUCTION 1

Chapter 1: The Gene Feature Enumeration Ecosystem 3
- **SYNOPSIS** 3
- **INTRODUCTION** 4
- **METHODS** 5
  - Feature Service 5
  - Annotation Pipeline 6
  - GFE Service 7
  - GFE DB 8
  - ACT Service 13
  - Validation 14
- **RESULTS** 15
  - GFE DB 15
  - ACT Service 17
- **DISCUSSION** 18

Chapter 2: High-resolution HLA A~B~DRB1 haplotype frequencies from the Ezer Mizion Bone Marrow Donor Registry in Israel 22
- **SYNOPSIS** 22
- **INTRODUCTION** 23
- **METHODS** 24
  - Study population 24
  - HLA genotyping 25
  - HLA allele and haplotypes frequency analysis 26
  - Clustering analysis 26
- **RESULTS** 27
  - HLA allele frequency 27
  - Frequent HLA-A, HLA-B and HLA-DRB1 haplotypes 28
  - Classification of populations and haplotypes 28
- **DISCUSSION** 31

Chapter 3: Impact of Ethnicity on Donor Match Rates in the Ezer Mizion Bone Marrow Donor Registry 34
- **SYNOPSIS** 34
- **INTRODUCTION** 35
- **METHODS** 38
  - Study Population 38
  - Modeling and Definitions 39
  - Availability of Donors 39
  - Statistical Analysis 40
  - Marginal Benefit Analysis 41
- **RESULTS** 43
  - Adult Donor Match Rates 43

Donor Registry Growth 46
Marginal Benefit Analysis with Be the Match Donor Registry 47
DISCUSSION 49
BIBLIOGRAPHY 53
APPENDIX A 56

List of Tables

- **Table 1.** Number of HLA sequence persisted in GFE DB (Page 15)
- **Table 2.** Number of observed GFE that match the expected (Page 15)
- **Table 3.** List of the 19 populations analyzed (Page 23)
- **Table 4.** Adult donor availability (Page 38)
- **Table 5.** 6/6 and >5/6 adult donor match rates and marginal benefit (Page 43)
- **Table 6.** 6/6 adult donor match rates within and outside populations (Page 46)
- **Table 7.** Number of exact accessions matches by feature (Appendix A: Page 54)
- **Table 8.** Number of exacts matches when the observed sequences are off by three base pairs (Appendix A: Page 55)

List of Figures

- **Figure 1.** The GFE Ecosystem (Page 5)
- **Figure 2.** GFB DB Schema (Page 7)
- **Figure 3.** CLUTO Clustering of 19 Ezer Mizion Populations (Page 28)
- **Figure 4.** 6/6 and 5/6 Stack Match Rates (Page 41)
- **Figure 5.** Adult Donor 6/6 Project Match Rates (Page 44)

INTRODUCTION

In 1979 the first unrelated allogeneic hematopoietic stem cell transplantation (HSCT) was performed[1]. Since then advancements in the field of human leukocyte antigen (HLA) have helped establish HSCT as a curative therapy for a myriad of malignant and non-malignant blood-related diseases. Clinical outcomes of HSCT have gradually improved with our understanding of the significance of HLA matching between donor and recipient. The best outcomes for HSCT are associated with high-resolution HLA matching between donors and recipients[2]. Matching at HLA-A, -B, -C, -DRB1, and –DQB1 is the current gold standard when searching for an unrelated donor. However, recent studies have highlighted the potential importance of donor age, CMV status, KIR and other HLA loci on HSCT outcomes[2,3]. These studies will continue to improve with the now routine use of next generation sequencing (NGS) and growing sample sizes[4]. Growing sample sizes and NGS typing will also improve HSCT donor matching algorithms by allowing for the reduction of phase and allelic ambiguity. Matching algorithms evaluate phase and allelic ambiguity for scoring the suitability of potential donors. If a matching algorithm cannot utilize haplotype frequencies that reflect the patient's ethnicity, then the best-suited donor may not be found due to poorly resolved phase ambiguity. For this reason, population-specific haplotype frequencies are critical when matching for HSCT. Even with well-described haplotype frequencies, ambiguities still arise at the allelic level that can impact the outcomes of HSCT. Before a transplant, high-resolution HLA typing is routinely done on the recipient and any prospective donors

to resolve these ambiguities. However, allelic ambiguities still exist that are not considered during the matching process. Matching is currently done at the second field (protein level) of the HLA nomenclature; therefore, ambiguities to the third and fourth field still exist between the donor and recipient[5]. Matching to the fourth field would utilize more detailed linkage information that would allow for more precise match predictions. With current NGS typing technologies, we could go beyond that and determine the best match between donor and recipient at the sequence level. Sequence level analysis of patients HLA data could also be extremely informative for HLA disease association studies[6]. Unfortunately, these types of studies cannot be done with the current HLA nomenclature, because it does not retain all sequence variations.[7] The current HLA nomenclature will need to be extended to resolve allelic ambiguity any further. Phase ambiguity present in matching algorithms could be reduced with the development of haplotype frequencies and donor modeling for poorly represented regions of the world. Our aim was to improve the current capabilities of resolving phase and allelic ambiguity for HSCT. We developed the gene feature enumeration (GFE) ecosystem as a way of extending the current nomenclature, and therefore as a way to further resolve any allelic ambiguity. We developed haplotype frequencies and performed modeling for the Ezer Mizion Bone Marrow Donor Registry in Israel. These results will better define the HLA haplotype distributions of the Middle East, and will, therefore, help resolve phase ambiguity for patients with Middle Eastern ancestry. Together, the GFE Ecosystem and Ezer Mizion analysis provide invaluable resources for unrelated donor registries to help better serve HSCT patients worldwide.

Chapter 1: The Gene Feature Enumeration Ecosystem

SYNOPSIS

A publicly available service for assigning HLA allele names to consensus sequences would be a valuable community resource. It would allow researchers the ability to Our aim was to use existing technologies to develop a publicly available, graph-based HLA allele calling service. Gene feature enumeration (GFE) notation has been developed as a way of describing sequence variation outside of current HLA nomenclature. We generated GFE notation for every sequence in the IPD-IMGT/HLA Database, and loaded the results into a neo4j graph database (neo4j.b12x.org), which includes HLA, GFE and sequence feature nodes. We developed an allele-calling tool (ACT) that leverages this graph to return an HLA allele call and GFE notation from submitted consensus sequence. ACT uses the GFE service (gfe.b12x.org) to convert the consensus sequences to GFE notation, and then searches the graph database for HLA alleles that share the most features with the generated GFE. A RESTful service interface makes ACT easy to use and allows for cross-platform compatibility (act.b12x.org). We used 30,000 Be the Match® donors typed at high resolution and with consensus sequences available for HLA-A, B, and C to test ACT. We compared the HLA calls ACT produced to the allele names reported by the typing labs. Comparisons were made with and without mapping alleles to their G group equivalent. When mapping alleles to their respective G group equivalent, the reported HLA typing for all 30,000 donors matched the ACT typing to the 2nd field with no ambiguity. Without mapping alleles to their G groups, the number of exact matches to the 4th field at HLA-A, B and C was 17,142 (57%), 16,540 (55%) and

10,024 (33%), respectively. Using G groups, ambiguity at the 4th field was reduced substantially; the number of exact matches increased to 29,914 (99.7%), 29,733 (99.1%) and 29,985 (99.9%) for HLA-A, B, and C, respectively. In 75 instances, ACT called alleles at a higher resolution than reported by the typing lab. HLA class I allele names can be accurately called from consensus sequences using our GFE-based ACT. This service will allow anyone to easily convert HLA class I consensus sequences into the corresponding GFE notation and HLA allele name. Class II HLA and KIR are already represented in the graph, and we plan to extend ACT functions to these loci.

INTRODUCTION

Innovations in human leucocyte antigen (HLA) genotyping technologies have caused a dramatic expansion in the number of known HLA alleles several times over the last decade[7]. HLA is the most polymorphic region of the human genome, making the accurate typing of HLA genes complicated[2]. Only recently have sequencing companies developed typing platforms and kits that are capable of sequencing HLA with high accuracy. These improvements in HLA sequencing technologies create challenges for the current HLA nomenclature[8]. Polymorphisms present in features of a gene are not always captured accurately with the current nomenclature[5]. Potentially useful sequence data is lost because of the divide between sequencing technology and nomenclature. Bridging this divide could extend the precision of disease association studies and potentially improve outcomes for hematopoietic stem cell transplantation (HSCT). The gene feature enumeration (GFE) notation was established to supplement the current HLA nomenclature by retaining all the lost sequence information[9]. The feature service

(feature.nmdp-bioinformatics.org) was developed as a way of storing and retrieving the accession numbers in a GFE notation. Programmatically generating GFE notation with consensus sequences became possible after the creation of the GFE service (gfe.b12x.org). These services made GFE accessible to the community and created an active discussion around the future potential of such technologies. From these discussions, the concept of a GFE based, publicly available, open sourced system for extending the IPD-IMGT/HLA database and HLA nomenclature emerged. Establishing a link between the GFE notation and the current IPD-IMGT/HLA database was an essential part of the GFE Ecosystem. Our aim was to complete the GFE Ecosystem by developing a publicly available GFE database (GFE DB) and a graph-based HLA allele-calling service.

METHODS

Feature Service

To make the GFE notation useful there needed to be a database for storing and retrieving the accession numbers. The feature service provides this by offering a RESTful interface for generating and retrieving the accession numbers (Figure 1A). A user can POST a Sequence Ontology (OS) term (ex. Exon, Intron, etc.), the term rank, the gene name and the sequence and receive the corresponding accession number. Users can also POST a SO term, rank, gene name and accession number and receive the corresponding sequence. The accession numbers are created by incrementing at each newly observed sequence for the provided gene name, OS term and rank. Therefore, the accession number for the first sequence submitted for each gene, SO term, and rank will always be one. To make sure

the accession numbers are consistent throughout the community there is only one feature service publicly available. The feature service is intended as a storage and retrieval service only and has no real computational capabilities. Joining the feature accession numbers associated with a gene sequence creates a GFE notation, but the feature service does not directly generate GFE. The accession numbers are generated by first extracting the gene features from the sequence.

Annotation Pipeline

The annotation pipeline offers a way of extracting the features from a given sequence using Clustal Omega (Figure 1B)[10]. The annotation pipeline consumes sequence data from FASTA files and produces text files containing the sequences associated with each feature in the sequence. We generated reference alignment files for each HLA and KIR locus that allowed for the quick identification of the coordinates of gene features in a sequence. A phylogenetic analysis was done to determine the best sequences to use in the reference alignments for each locus. Comparing the results to the sequence annotations on IPD-IMGT/HLA validated the annotation pipeline. The annotations for class I loci (HLA-A, B, and C) are very accurate due to highly conserved regions but are less accurate for class II loci (HLA-DRB1, DQA1, DQB1, DPA1 and DPB1). The annotation pipeline is publicly available on Github (github.com/nmdp-bioinformatics/HSA), but is not available as a service or easily built locally.

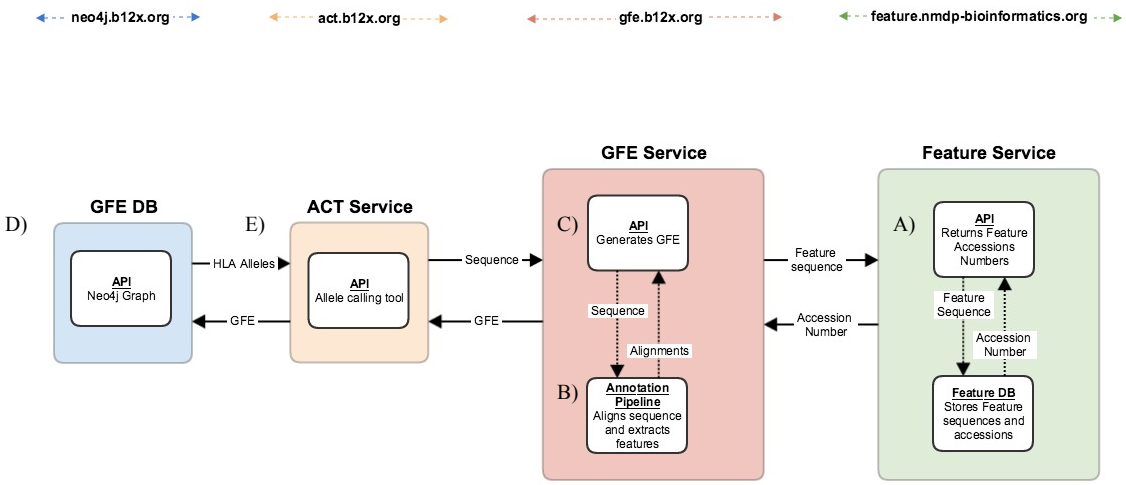

Figure 1. The GFE Ecosystem. **A)** The feature service accepts sequence features and returns the corresponding accession number. **B)** The annotation pipeline is used to extract sequences of features for submitting to the feature service. **C)** The GFE Service uses the annotation pipeline to align sequences and then submits them to the feature service. The corresponding accession numbers are combined to create a GFE notation. **D)** The GFE DB contains all of the sequence data in IPD-IMGT/HLA and also persists the corresponding GFE notation. **E)** The ACT Service creates GFE notation from consensus sequences using the GFE service and then searches the GFE DB for the corresponding HLA alleles.

GFE Service

At the fourth Immunogenetics Data Standards Hackathon (DaSH) in Vienna, Austria we developed the GFE service so that anyone could generate GFE notation from sequence data (Figure 1C). This service has played an important role in the adoption GFE notation by allowing anyone to produce GFE notation from a RESTful API easily. Before the GFE service was available all GFE notation was created manually and could not be done for large numbers of sequences. The GFE service was developed with the same RESTful principals as the feature service, allowing anyone on any platform to utilize it. Users can post HLA and KIR sequence data to the service and receive the GFE notation in return. Within the GFE Service the annotation pipeline is being run on the submitted sequences to extract the features. The features are passed to the feature service to attain the

corresponding accession numbers. The returned accession numbers are then placed in their genomic order and joined to create the GFE notation. Users can get back the full sequence associated with a GFE by submitting the GFE to the 'seq' API. Like the feature service, it is also publicly available on Github (github.com/nmdp-bioinformatics/service-gfe-submission) and Dockerhub (hub.docker.com/r/nmdpbioinformatics/service-gfe-submission), making it exceptionally easy to deploy locally.

GFE DB

The sequences in IPD-IMGT/HLA are connected in many different ways, and IPD-IMGT/HLA does not persist these connections. These connections, such as shared exons, could be represented in a graph database and efficiently queried. This type of functionality would be a tremendous improvement from what the IPD-IMGT/HLA currently offers. Establishing a link between the GFE notation and the HLA allele names on IPD-IMGT/HLA was a crucial reason for building the GFE DB.

We first established a database schema that would capture as many connections as possible while still retaining the original information.

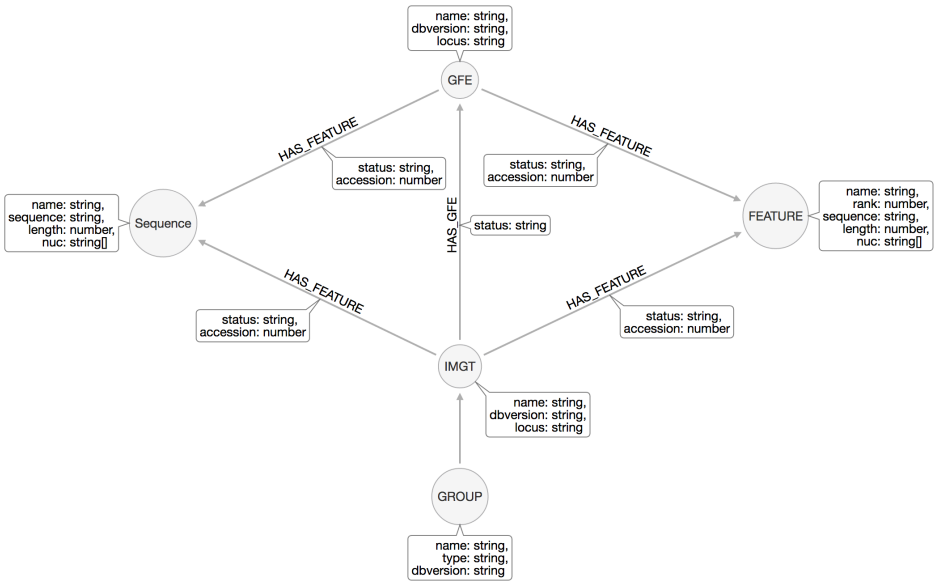

Figure 2. The database schema used for modeling the GFE DB.

We developed the schema for the GFE DB (Figure 2) by first creating the nodes we wanted represented in the graph: IMGT, GFE, FEATURE, SEQUENCE, and GROUP. The IMGT nodes represent the HLA allele names from IPD and do not have any sequence information contained in the node properties. The IPD-IMGT/HLA database version, allele name, and locus are simply the dbversion, name, and locus properties in the node. The IMGT nodes are connected to GFE, FEATURE and SEQUENCE nodes. Like the IMGT nodes, the GFE nodes only have dbversion, name and locus as properties and are also connected with the FEATURE and SEQUENCE nodes. Both the IMGT and GFE nodes are connected to the FEATURE and SEQUENCE nodes by the HAS_FEATURE relationship. This relationship has two critical properties: status and accession. The accession numbers returned by the feature service are stored as the accession property in the relationship. The status indicates whether the feature was

generated from taking the exact coordinates from IPD-IMGT/HLA (Expected) or if the feature was generated using the GFE service (Observed).

For each HLA sequence in IPD-IMGT/HLA an observed and expected relationship exists. The HAS_GFE relationship also has the status property but does not contain any information on relating to the sequence. The SEQUENCE nodes contain the full sequence taken from IPD-IMGT/HLA and therefore only have expected HAS_FEATURE relationships. The FEATURE nodes have the same properties as the SEQUENCE nodes, which are: name, sequence, rank, length, and nuc. In the SEQUENCE node, the name and rank are "Sequence" and 0. However, in the FEATURE nodes they represent the name of the feature (exon, intron, etc.) and the rank of that feature. The "nuc" property is an array of the sequence, which allows for querying specific positions in a sequence.

Broader queries can be made using the GROUP nodes, which are the only nodes without any inward pointing edges. The GROUP nodes represent different groupings of the HLA alleles, such as G_GROUP and P_GROUPS, which are defined on IPD-IMGT/HLA. HLA alleles that have the same nucleotide sequence in the antigen recognition site are considered to be in the same G_GROUP. Similarly, alleles that have the same protein sequence in the antigen recognition site are in the same P_GROUP. The group nodes contain the type of node being represented, the name of the group (ex. HLA-A*01:01:01G) and the IMGT/HLA database version associated with that group. Each group node points to the alleles that are found within that group.

With the schema established, we worked on producing the data needed for populating the graph. We produced the raw data files needed for populating the graph in three steps: generating the expected, generating the observed and finally formatting the results into bulk load files. We used the HLA.xml file for the IPD-IMGT/HLA version 3.26.0 to create both the expected and observed results. Every HLA allele in IPD-IMGT/HLA is represented by an element within the XML file. Within the allele element is a sequence element, which contains the full sequence associated with the allele. The sequence element also contains tags for each specific feature present in the sequence and the corresponding coordinates for those features within the sequence. We produced the expected sequence features by using these coordinates and the full sequence. For every allele in IMGT/HLA, we extracted the sequence features and then passed them to the feature service to get the expected accession numbers. We used these expected accession numbers to create the expected GFE notation for each HLA allele in IMGT/HLA. The produced expected file contained the HLA allele name, the GFE notation, the full sequence, and a list of the features with the corresponding sequence and accession number for each HLA allele. The observed file was formatted the same way, however, instead of generating the GFE notation and features from the known coordinates we used the GFE service.

The GFE service can convert a consensus sequence into GFE notation in about 5 to 10 seconds, depending on the locus and the length of the sequence. When dealing with a handful of sequences the runtime is not an issue, but it can be a substantial barrier when dealing with thousands of sequences. To expedite the processing of all the IPD-

IMGT/HLA sequences we stood up several Amazon Web Service (AWS) instances and used docker to run the GFE service locally on each machine. We broke up the HLA.xml file containing the IPD-IMGT/HLA database into three equal parts and allocated one part per machine. Each sequence was run through the GFE service in parallel using a nextflow script. Nextflow allows for the GFE client scripts to be run in parallel and efficiently utilizes the available CPU[11]. Up to 40 sequences ran in parallel on each machine, allowing for roughly 120 sequences to be run in parallel across all machines.

Building a neo4j graph database is straightforward when using the neo4j-import tool. Only four arguments are needed for building a neo4j graph database with the import tool: id-type, into, nodes, and relationships. The “--into” argument specifies the location of the graph database and needs to match the “dbms.directories.data” variable in the neo4j.conf file. The “--id-type” argument will either be INTEGER or CHARACTER depending on the type of IDs used for the nodes and relationships. The “--nodes" and “--relationships" arguments specify CSV files that contain the nodes and the relationships in the graph. For the GFE DB, we created two node files and three relationship files. We generated these files from the expected and observed results using a Perl script. Running the import command with these files populates a Neo4j graph database and starts a server with a web interface available on port 7474. The code for building the graph can be found on Github (github.com/nmdp-bioinformatics/gfe-db) and a docker image containing the graph is available on Dockerhub (hub.docker.com/r/nmdpbioinformatics/gfe-db).

ACT Service

Designing the API for the ACT Service was done using the Swagger editor (swagger.io). We designed five APIs for utilizing the GFE DB: hla, gfe, ars, sequence, and act. We defined the parameters and responses for each API call in the YAML swagger specification. We generated a Python Flask server using the Swagger specification file and the Swagger code generation tool. The generated server code was modified to import python modules that contained the main functionality for each API. The function of each API was made possible through cypher queries that search the GFE DB.

The "/gfe" API allows users to find every GFE notation associated with a particular HLA allele. Similarly, the "/hla" API allows users to find every HLA allele associated with a particular GFE notation. The ARS group associated with a given GFE notation or HLA allele can be determined using the "/ars" API. Getting the sequence associated with a GFE notation, HLA allele or feature for an HLA allele can be done using the "/sequence" API. The main API exposed by the ACT service is the "/act" API, which also offers the most complex functionality. The ACT API generates GFE notation from consensus sequences and makes HLA allele calls to the fourth field. Users provide an HLA locus and consensus sequence and receive an array of HLA alleles associated with that consensus sequence as well as the GFE notation.

The functionality behind the "/act" API is broken up into five parts, each using separate cypher queries for getting data from the GFE DB. First, the service checks whether the provided sequence exists in the GFE DB. If the sequence exists, then the service simply returns the HLA and GFE associated with that sequence in the graph. If

the sequence does not exist in the graph, then it is converted to GFE notation by making a call to the GFE service. The service extracts the accession numbers associated with exons two and three from the generated GFE. The typing of these exons is required for every HLA class I sequence submitted to IPD-IMGT/HLA. Classification into the "G" groups is also done at these exons, making them informative for producing HLA allele calls. The accession numbers associated with these exons are then used to search the graph for all GFE that share them. The service returns a list of GFE notations that share the same exons two and three as the previously created GFE. From this list, the GFE notations that share the most accession numbers with the generated GFE are taken. The remaining GFE will have the most features identical in sequence to the posted consensus sequence. An HLA allele call is made by finding the alleles associated with these returned GFE. The service will return a JSON object with a "gfe" element that contains the generated GFE, a version element that has the ACT version and an "hla" element that contains an array of HLA alleles. If the service fails to make an allele can then the "hla" element will be empty.

Validation

Extensive validation is essential for proving the utility of any new tool.

Comparing expected to observed results was the validation approach used for both the GFE DB and the ACT service. We compared the number of alleles successfully loaded in the GFE DB to the expected number of alleles at each locus. We also compared the raw data from IPD-IMGT/HLA to nodes with expected relationships.

We compared the accession numbers of the expected relationships to the observed for each feature type and locus. Similarly, we compared the expected GFE notation to the observed at each locus. To identify how often the expected and observed sequences differed by three base pairs, we wrote cypher queries to compare the beginning and ends of sequences.

The accuracy of the ACT service was evaluated using 30,000 Be the Match® donors typed at HLA-A, B, and C with consensus sequences available. We stood up three AWS c4.large instances and used Docker to deploy both the GFE and ACT services locally on each machine. A python tool that utilizes the ACT service client was run in parallel on each machine using a nextflow script. We compared the allele calls produced by the ACT service to the allele calls reported by Histogenetics for each donor. We made comparisons with at the second, third and fourth field with and without alleles reduced to their ARS equivalent.

RESULTS

GFE DB

In five hours all of the IPD-IMGT/HLA sequences finished being processed by the GFE service. The number of HLA alleles persisted in the GFE DB matches the number of HLA alleles available on IPD-IMGT/HLA. The nodes with expected relationships correctly represent all of the sequence data available on IPD-IMGT/HLA. Table 1 shows how many alleles successfully ran through the GFE service for each locus. As expected, a significant portion of sequences for class II could not be processed. However, every sequence for class I was successfully processed by the GFE service. HLA-DPB1

performed the worst, with only 26 sequences being successfully processed by the GFE service. We expected this due to variability among HLA-DPB1 sequences that the reference alignment file could not appropriately capture. The number of times the observed GFE matches the expected for each sequence in IPD-IMGT/HLA is depicted in Table 2. HLA-A performed the best with only 10.5% of the observed GFE matching the expected. None of the observed class II GFE matched the expected. Slight variations in the sequences would cause the expected to mismatch with the observed; therefore these numbers were not expected to be high. When comparing the observed to the expected at the level of the features, the accuracy was much higher (Appendix A: Table 7). On average the expected sequence matched the observed 71% of the time at the exons for class I and 99% of the time at the introns. The expected 5' UTR sequence matched the observed only 10% of the time for class I, which impacted the majority of the mismatches between the expected and observed GFE. Some of the mismatches at the feature level are explained by three base pairs being removed or added at the beginning or end of an exon. In these cases, the annotation pipeline incorrectly identifies the exon boundaries and nucleotides are added to or removed from an adjacent exon. These are observed over 200 times at exon three for each class I locus and is most commonly observed at HLA-B (Appendix A: Table 8).

Locus	Persisted Sequences	Sequences in IPD-IMGT/HLA
HLA-A	3657	3657
HLA-B	4459	4459
HLA-C	3296	3296
HLA-DPB1	26	716
HLA-DQB1	116	978
HLA-DRB1	95	1972

Table 1. The number of HLA sequences that are persisted in the GFE DB.

Locus	Matched GFE	Percent Match
HLA-A	381	10.46%
HLA-C	96	2.16%
HLA-B	319	9.70%
HLA-DRB1	0	0.00%
HLA-DQB1	0	0.00%
HLA-DPB1	0	0.00%

Table 2. The number of times the expected GFE notation matched the observed.

ACT Service

Using nextflow and three AWS instances we ran all 30,000 donors through the ACT service in 10 hours. When mapping alleles to their G group equivalent, the reported HLA typing for all 30,000 donors matched the ACT typing to the 2nd field with no ambiguity. Without mapping alleles to their G groups, the number of exact matches to the 4th field at HLA-A, B, and C was 17,142 (57%), 16,540 (55%) and 10,024 (33%), respectively. Using G groups, ambiguity at the 4th field was reduced substantially; the number of exact matches increased to 29,914 (99.7%), 29,733 (99.1%) and 29,985 (99.9%) for HLA-A, B, and C, respectively. In 89 instances, ACT called alleles at a higher resolution than reported by the typing lab.

DISCUSSION

We have shown that the GFE notation offers solutions to many of the current issues facing the HLA nomenclature; therefore, it would be a useful addition to the next version of the nomenclature. Researchers interested in studying sequence variation not captured by the current naming convention could utilize the GFE Ecosystem for generating and analyzing GFE notation. The GFE DB offers researchers the ability to query the sequence data in IPD-IMGT/HLA in ways that are not currently possible. Sequence data can now be easily retrieved for alleles or features by simple queries. With the IPD-IMGT/HLA database, these operations would have to be done manually and could be time-consuming tasks. More flexibility is also offered with the GFE DB interface, which allows users to interact with and visualize the data. For developers, the Docker image of the GFE DB offers the ability to quickly stand up a version locally. Being able to run the GFE DB locally makes it more likely to be adopted by the broader immunogenetics community. The GFE DB also offers valuable insight into the current limitations of the annotation pipeline.

The annotation pipeline is not currently capable of reliably processing class II consensus sequences. This limitation was known before the creation of the GFB DB, but it was not clear which alleles would fail. The next iteration of the annotation pipeline will include more reference alignments that better represent the sequences that failed to be processed. Adding a step to the annotation pipeline to determine what reference alignment to use could address this issue as well. Partial class II and KIR sequences may still fail to be processed with this approach; therefore, we will need to apply different

methods for annotating sequences. One option is to use a Hidden Markov Model (HMM) for identifying the gene features in a sequence. Numerous research groups have validated the accuracy of using HMMs to identify gene features[12]. Taking this approach would be useful for partial sequence data where the exons are separated by large unreported introns. Alignment algorithms often fail to align these sequences because of the large gaps between the exons. These gaps would not effect the prediction of the gene features when using a HMM. We have already started the development of a HMM that is specifically built for HLA and will be included in the next release of the annotation pipeline.

The coordinates of the gene features returned by the annotation pipeline are slightly off for the vast majority of the sequences in IPD-IMGT/HLA (Table 2). While this inaccuracy highlighted issues with the annotation pipeline, it did not affect the accuracy of the allele-calling tool. These slight differences actually help improve the accuracy of the ACT, by providing the service with more variation to compare with. A typing lab can produce numerous unique sequences that all represent the same HLA allele, and each would represent a different GFE notation. Every variation captured in the GFE DB for an allele makes the ACT service more accurate at calling that specific allele. Therefore, without both the expected and observed GFE results persisted, the ACT service wouldn't have been as accurate. In next version of the ACT service and GFE DB, there will be an option for users to persist submitted results to the database. Being able to store sequences that identify with a specific allele would improve the accuracy of the ACT and allow researchers the ability to curate their data.

The technologies adopted by the GFE Ecosystem will allow for seamless integration into existing data pipelines. The RESTful framework separates platform and language giving developers the freedom to program clients in any language they desire. With Docker, developers can package entire web applications into isolated containers that can be run on any machine. Anyone who has Docker installed can build the entire GFE Ecosystem locally in minutes by pulling from Dockerhub with one docker-compose command. Unlike Dockerhub, Github contains everything that is needed for anyone to contribute to the development process of the GFE Ecosystem. Researchers in the field have already shown interest by providing feedback and by following the repositories on Github. Continued community engagement in the development process will ensure the adoption and sustainability.

The most active engagement has been around the ACT service and how it can be utilized in current pipelines. HLA class I allele names can be accurately called from consensus sequences using our GFE-based ACT. Having no ambiguity at the second field after reducing to the ARS equivalents is a significant feat for an allele-calling tool and could be a useful tool for research groups. None of the publicly available tools for making HLA allele calls work well with consensus sequences and do not utilize the same technologies as the GFE Ecosystem. The ACT service is the only available allele-calling tool that implements a RESTful framework, which allows it to be utilized with simple HTTP requests. No publicly available tool exists for making Class II and KIR allele calls from consensus sequences. Extending the capabilities of the ACT service to these loci will make it substantially more useful. The next release of the ACT service will include

the ability to make class II and KIR allele calls using the updated version of the annotation pipeline.

Taken together, all of these services make up the GFE Ecosystem: an open sourced, community driven system for extending the IPD-IMGT/HLA database and nomenclature. The GFE Ecosystem offers a solution to the problems facing the current nomenclature and will allow everyone to partake in the process of creating HLA allele names. Sequence data will no longer be lost in allele names and waiting for IPD-IMGT/HLA to provide a name for a new sequence will no longer be an issue. Most importantly, this system opens up the development process and tools to the whole community. Hospitals and research labs paying for tools whose functions are done by the GFE Ecosystem could save money and gain the advantages of using GFE. Unforeseen research opportunities could also be found from utilizing the data exposed by the GFE DB. All of these benefits demonstrate potential for the GFE Ecosystem to have a dramatic impact on the field of immunogenetics. The current system will inevitably adopt some change, with or without GFE. With the GFE Ecosystem the immunogenetics community will be better prepared for impending flood of sequencing data. This will allow for cutting edge research to be done and life saving discoveries to be made.

Chapter 2: High-resolution HLA A~B~DRB1 haplotype frequencies from the Ezer Mizion Bone Marrow Donor Registry in Israel

Sigal Manor[a‡], Michael Halagan[b‡], Nira Shriki[a], Isaac Yaniv[a,c,d], Bracha Zisser[a], Martin Maiers[b], Abeer Madbouly[b], Jerry Stein[c,d]

[a] Ezer Mizion Bone Marrow Donor Registry, Petach Tikva, Israel
[b] Bioinformatics Research, National Marrow Donor Program, Minneapolis, MN, USA
[c] Bone Marrow Transplant Unit, Schneider Children's Medical Center of Israel, Petach Tikva, Israel
[d] Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[‡] These authors contributed equally to this work

I performed every analysis and generated the figures, tables and supplementary tables for this study. I wrote the methods and the first part of the results section. I also contributed to the revising of the introduction and discussion sections.

DOI: https://doi.org/10.1016/j.humimm.2016.09.004

SYNOPSIS

We have investigated HLA population alleles and haplotype frequencies for the ethnicities that comprise the contemporary population of Israel, using a large data set from the Ezer Mizion Bone Barrow Donor Registry. We genotyped 275,699 individuals at the HLA-A, -B and -DRB1 loci using HLA genotyping methods. HLA A~B~DRB1 haplotype frequencies were estimated from 19 sub-ethnic Jewish populations and other non-Jewish minorities using the maximum likelihood model, which accommodates typing ambiguities. We present overall and sub-ethnicity specific HLA diversity results of the registry, which will help guide a data-driven strategy for future registry expansion.

INTRODUCTION

Israel is home to the entire genetic spectrum of the Jewish Diaspora as well as for non-Jewish minorities, leading to marked ethnic diversity in a country of only eight million individuals. Before the founding of the modern state of Israel (1948), the Jewish Diaspora consisted of separate Jewish communities in Europe, North Africa, and Asia. Ancient communities of Jewish exiles formed in Iran and Iraq from the 7th to the 5th centuries BCE. From the 1st century BCE onward, Jewish communities spread westward through the north and south coasts of the Mediterranean basin, throughout the Levant and inland into the European continent. Subsequently, Jewish populations expanded in central and eastern Europe, along the North African coast and in the southern Arabian peninsula. Jews migrated eastwards from the Persian Empire and reached as far as India and China. Toward the late 19th century through the 20th century, there was significant immigration of Jews from Europe to North and South America. Admixtures of Jewish migrants with indigenous host populations led to increasing genetic diversity between individual Diaspora communities, while cultural and religious forces maintained coherence of the Jewish people[13–15]. During the last 100 years, increasing immigration to Israel, together with the growth of Muslim and Druze populations within the boundaries of the State have resulted in a panoply of subethnicities that make up the patchwork of modern Israeli society. As second and third generation Israelis start marrying outside of their ancestral subethnicities, an additional layer of allelic diversity has been introduced into the HLA landscape in Israel, leading to intergenerational immunogenetic differences within the Israeli population[16]. In the last decades, the successful use of highly matched unrelated

volunteer donors for hematopoietic stem cell transplantation (HSCT) has stimulated the development of many national volunteer unrelated donor stem cell registries. This paper outlines analyses of HLA population haplotype frequencies in a large stem cell donor registry to characterize the genetic population profile of the contemporary Israeli population. In addition to the large extent of the population size analyzed, the high-resolution HLA profile presented in this study improves on data obtained in previous studies[14,17–20]. This analysis is the first phase of a project that seeks to guide strategic planning for donor recruitment and expansion of the Ezer Mizion Bone Marrow Donor Registry (EM BMDR) in Israel.

METHODS

Study population

We examined all 754,135 unrelated volunteer donors registered in EM BMDR from its inception through June 2014 to gauge HLA haplotype diversity within the Israeli population. All subjects provided informed consent for registration at recruitment and self-reported their parents' country of origin, permitting us to assign subethnicities. Each sub-ethnic population included only individuals who self-reported the same sub-ethnic population for both parents; multi- or mixed-ethnicity donors were excluded from this analysis. The study was approved by the Ethics Committee of Rabin Medical Center, and was conducted in accordance with the 2014 Ministry of Health (Israel) Guideline for Clinical Trials in Human Subjects.

Table 3. A list of all 19 analyzed Ezer Mizion populations and sample counts.

Ezer Mizion populations	Sample counts
Arab	12,300
Argentina[a]	4,307
Ashkenazi	4,625
Bukhara	2,317
Druze	5,914
Ethiopia	5,928
Georgia	4,471
Iran	8,153
Iraq	13,270
Israel	69,716
Kavkaz	2,840
Libya	3,739
Morocco	36,718
Poland	13,871
SEE[b]	11,179
Tunisia	9,070
USA[a]	6,058
USSR[c]	45,681
Yemen	15,542
Total	**275,699**

[a] Argentina and USA population are derived from emigrants of European Jews
[b] SEE include Romania, Bulgaria, Moldova, Greece, Yugoslavia, Albania, Serbia, Transylvania and Cyprus
[c] USSR include Russia, Ukraine, Belarus, Lithuania, Latvia and East Europe

HLA genotyping

HLA typing analyses evolved during the building of the EM BMDR roster, starting with serologic determinations and evolving to DNA- based testing at low, intermediate, and allele resolution (SSO-, SSP-, or SBT based). Only data genotyped by DNA-methods was used in this analysis. The initial dataset included 67 sub-ethnicities, with an average sample size of 5200 (2–79,066) per population. However, only 19 populations, containing a total of 275,699 donors, were large enough for analysis, based on a minimum sample size per of >200 high-resolution typed samples at HLA-A, -B and -

DRB1 loci per sub-ethnicity (Table 3). It should be noted that Argentina and USA population are derived from emigrants of European Jews; SEE sub-ethnic population include Jews from Romania, Bulgaria, Moldova, Greece, Yugoslavia, Albania, Serbia, Transylvania and Cyprus; USSR sub-ethnic population include Jews from Russia, Ukraine, Belarus, Lithuania, Latvia and East Europe.

HLA allele and haplotypes frequency analysis

Three-locus haplotype frequencies (A~B~DRB1) were estimated for each of the 19 populations, resolving phase and allelic ambiguity using the expectation–maximization (EM) algorithm[21,22]. The applied EM algorithm was designed to handle mixed resolution data[23,24]. Allele frequencies were calculated by summing across the haplotype frequencies. Deviations from Hardy– Weinberg equilibrium (HWE) were assessed at the allele-family level (first nomenclature field) using a chi-squared test as implemented in the software PyPop[25].

Clustering analysis

HLA haplotype frequency visualizations on the study populations were created using the CLUTO software[26] by clustering the top 100 haplotypes (rows) in the 19 studied populations, and separating them into eight clusters of haplotypes based on haplotype similarity across populations (columns). It should be noted that since CLUTO examines only the top 100 haplotypes in each sub-ethnicity, this analysis may highlight the dominant themes of a population's composition as compared with other tools such as principal components analysis or neighbor-joining (NJ).

RESULTS

HLA allele frequency

HLA alleles frequency data are presented for 19 ethnic groups containing a total of 275,699 subjects from the EM BMDR. A summary of the 20 most common HLA-A, HLA-B and HLA-DRB1 alleles and their respective estimated frequencies are presented in Supplementary Table S2, S3 and S4. Top 100 alleles frequencies of HLA-A, HLA-B and HLA–DRB1 loci are provided in Supplementary Table S1. The most frequent alleles for the HLA-A locus were A*01:01, A*02:01 and A*03:02. The most frequent alleles for the locus HLA-B were B*35:08, B*49:01 and B*38:01. The most frequent alleles for the locus HLA-DRB1 were DRB1*11:04, DRB1*07:01 and DRB1*13:02. HLA-DRB1*11:04 was observed with a frequency greater than 15% in all populations except Yemen and Ethiopia. Locus level deviations from HWE were detected at HLA-A, -B, and -DRB1 in 11 of the 19 analyzed populations. The most significant HWE deviation was detected at HLA-B among those donors who self-designated themselves as ‘‘Israeli” between the homozygous expected and observed counts (5326 expected, 5922 observed, $p = 4\ 10^{16}$). This result was expected since the ‘‘Israel” group comprises of donors who reported that their parents were born in Israel, however their specific ethnic origin was not indicated. Thus, the ‘‘Israel” group is likely to be of mixed ethnicity with parents who do not necessarily share the same ethnic origin. Data results of HWE analysis are provided in Supplementary Table S5 and S6.

Frequent HLA-A, HLA-B and HLA-DRB1 haplotypes

Supplementary Table S1 illustrates the top 50 ranked A~B~DRB1 haplotypes for all 19 analyzed populations. The number of haplotypes with a minimum frequency of 10^{4} ranged from 692 to 1579 depending on sub-ethnicity groups. Several haplotypes were shared among sub-ethnic populations while others remained private. The haplotype A*26:01~B*38:01~DRB1*04:02 was shared among Ashkenazi populations with a frequency of 2.3–6.7%. The haplotypes A*33:01~B*14:02~DRB1*01:02 and A*30:01~B*13:02- ~DRB1*07:01, common in the US Middle Eastern/North African coast population, were common in most sub-ethnicities with frequencies 0.02–5.3%. Also, A*02:05~B*50:01~DRB1*07:01, common to most Arab populations, was among the top 100 haplotypes in the Ezer Mizion data with frequencies ranging from 0.02% to 1.29%.

Classification of populations and haplotypes

Figure 3 shows the clustering of the Top 100 haplotypes (rows) in the 19 studied sub-ethnicities, which could be resolved into eight haplotype clusters based on similarities across populations (columns). Several inferences can be drawn from the generated clusters. As expected, the largest cluster of haplotypes (cluster 4) includes the Ashkenazi populations. These populations cluster more tightly together than the non-Ashkenazi groups. Other clusters correlate with geographic proximity, such as Iran and Iraq (cluster 3), Bukhara and Iraq (cluster 2, although Bukhara also contained some private haplotypes) and Libya and Tunisia (cluster 5). Other clustering trends were noted. As

expected, populations related by immigration patterns form clusters such as Morocco and Tunisia, and Ashkenaz, Poland and USSR. In general, HLA haplotypes of most Ashkenazi populations cluster together while those of non-Ashkenazi populations are more divergent.

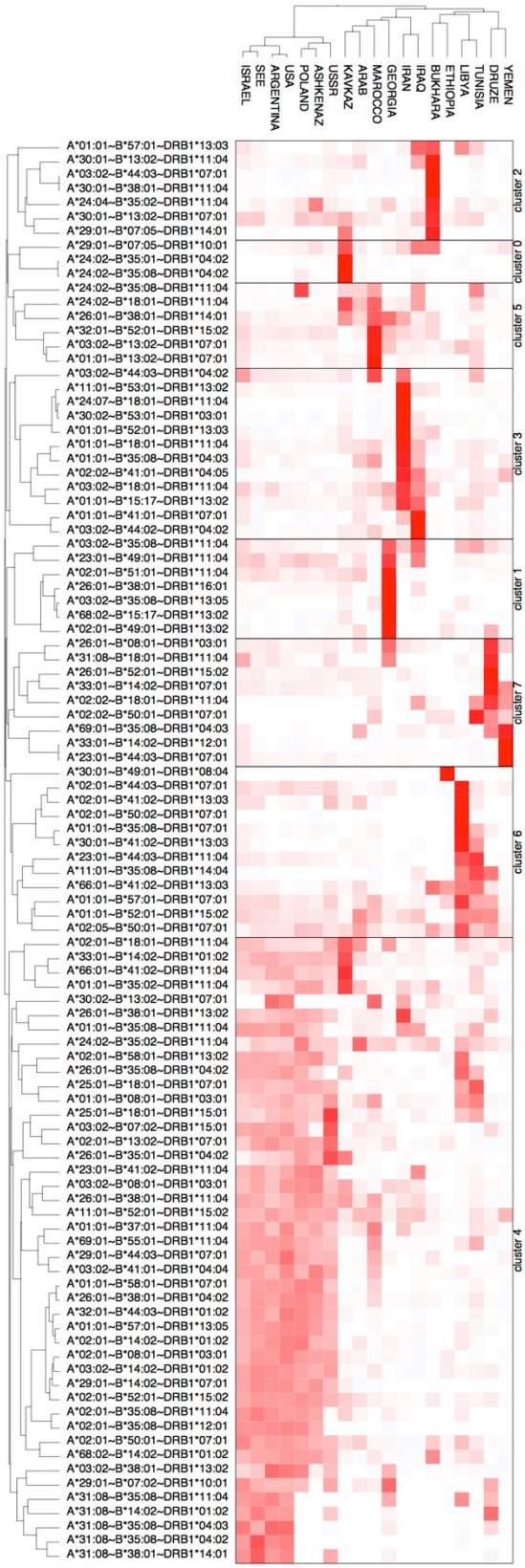

Figure 3. Clustering of the top 100 haplotypes on 19 sub-ethnic populations defined by country of origin.

DISCUSSION

We described allele and haplotype frequency distribution based on data obtained from 275,699 hematopoietic stem cell donors, representing 19 different ethnic groups contained in the Ezer Mizion Bone Marrow Donor Registry using a maximum likelihood model that can resolve genotyping ambiguities. The large sample size of this study population validates the findings of previously published haplotype frequencies that relied on a smaller sample size with a less stringent HLA-typing resolution[14,17–20]. Addition- ally, we present information regarding unique sub-ethnicities that have not previously been reported (Bukhara) or have been reported with limited number of subjects (Ethiopia and Druze)[19]. The sub-ethnic populations of the donors included in this study were stratified according to the donors' self-reporting of their parents' origin. It should be mentioned that the extent to which self- reporting ethnicity or geographic ancestry will correspond to genetic ancestry is likely to vary as previously discussed by Hollenbach et al.[27]. This fact is specifically observed in one of our large, but undistinguished, self-reported population that is composed of donors who listed themselves as ''Israel". This heterogeneous group likely consists of many mixed sub-ethnicity donors for whom sub-ethnic assignations cannot be made.

The cross-sectional nature of our registry permits a detailed view of the contemporary Israeli genetic profile. We compared our results to a previously published study by the Hadassah registry (HD) in Israel[14] which in some ways were similar to our data and in other divergent, most likely due to differing sample sizes, resolution level and

recruitment strategies. In order to compare both registries we rolled up the EM BMDR data to the 2-digit level, which could introduce a source of discrepancy between the results. As expected, we have observed that populations of geographic proximity in both studies share common haplotype. Examples are (to name a few) Iran, Libya and Israeli in the data sets from both registries, Algeria and Morocco in the data set of HD vs. Morocco in EM BMDR data set, Argentina in the HD data set vs. Argentina and Uruguay in EM BMDR, Germany and Poland in the HD study vs. Poland in EM BMDR donors. Some haplotypes emerged as common across different geographic areas, for example, the haplotype A*26:01~B*38:01~DRB1*04:02 and A*24:02~B*35:02~DRB1*11:04 are common in most Ashkenazi populations while the haplotype A*02:05~B*50:01~DRB1*07:01 and A*02:01~B*50:01~DRB1*07:01 were shared in Middle Eastern populations in both studies. Some Jewish sub-ethnicities in both the Hadassah and EM BMDR registries share some common haplotypes such as: A*26~B*38~DRB1*04, A*24~B*35~DRB1*04 and A*24~B*35~DRB1*11. It also should be noted that some allele families appear to be shared among most populations (e.g. A*02, 24 and 26; B*35, 38 and 50 and DRB1*04, 07 and 11).

Sub-ethnic and genetic heterogeneity within the Israeli population necessitates a population specific unrelated bone marrow donor registry that reflects the commonalities and diversities of the overall population. Enhanced representation of both common and uncommon alleles increases the likelihood of HLA-matching between Israeli donors and Israeli patients (including Jewish patients in the Diaspora) of all sub-ethnicities. The

results or our analysis have implications on cross-population matching and can help in donor searches and population-based recruitment strategies. Gragert et al. published a recent study regarding the likelihood of finding a suitably matched adult donor or cord-blood unit in the NMDP registry that included projections that account for future registry growth[28]. The development of accurate models for prediction of optimal registry size and expansion require the determination of high-resolution HLA haplotype frequencies within the target population that includes cross-sectional coverage of sub-ethnicities within said population. The haplotype frequencies of the major sub-ethnic groups in the EM BMDR presented in this current study are part of a strategic effort to guide recruitment goals and expansion of the Ezer Mizion volunteer adult unrelated donor registry.

Chapter 3: Impact of Ethnicity on Donor Match Rates in the Ezer Mizion Bone Marrow Donor Registry

Michael Halagan[‡a], Sigal Manor[‡b], Nira Shriki[b], Isaac Yaniv[b,c,d], Bracha Zisser[b], Abeer Madbouly[a], Martin Maiers[a], Jerry Stein[c,d]

[a] Bioinformatics Research, National Marrow Donor Program, Minneapolis, MN, USA
[b] Ezer Mizion Bone Marrow Donor Registry, Petah-Tikva, Israel
[c] Bone Marrow Transplant Unit, Schneider Children's Medical Center of Israel, Petah-Tikva, Israel
[d] Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, Israel

[‡] These authors contributed equally to this work

I performed every analysis and generated the figures, tables and supplementary tables for this study. I wrote the methods and results sections. I also contributed significantly to the revising of the introduction and discussion sections.

DOI: http://dx.doi.org/10.1016/j.bbmt.2017.04.005

SYNOPSIS

Human Leukocyte Antigen (HLA) haplotype frequencies in a volunteer bone marrow donor registry should reflect the frequencies of potential transplant recipients served by that registry, a challenge in a country with diverse sub ethnicities of immigrants from Eastern and Western cultures, such as Israel. We measured the likelihood of finding suitable donors for hypothetical patients drawn from defined sub-ethnicities in the Ezer Mizion Bone Marrow Donor Registry (EM BMDR) both from donors within and outside the registry now and during the coming decade. On average, bioinformatics modeling predicts that, given current donor recruitment trends, 6/6 high-resolution HLA match rates for Israelis, which currently stand at 40-55% for most sub-ethnicities, will rise by up to 1% per year over the next decade. Sub- ethnicities with historically lower rates of

inter-ethnic admixture are less likely to find matches outside of their designated group but will also benefit from expansion of the registry, while ethnically-directed drives will enhance match rates for currently under- represented sub-ethnicities. Donor searches for the same cohort using a large extramural registry was of only slight benefit for most of the 19 EM BMDR sub- ethnicities evaluated, confirming that local donor registries that reflect the ethnic diversity of the community being served are best equipped to serve the needs of their respective communities. Contemporary trends of increasing multi-ethnic admixture in Israel may impact the effect of ethnic profiling in assessing future match rates for EM BMDR.

INTRODUCTION

Hematopoietic Stem Cell Transplantation (HSCT) can be life saving for patients with lethal hematologic malignancies and for patients with an ever-expanding list of non-malignant hematologic and immunologic disorders[29,30]. As family size in many western countries shrinks, patients in need of HSCT must frequently turn to burgeoning bone marrow donor registries to seek a matched unrelated donor (MUD). The growth of donor registries in countries around the world, their online accessibility, and the improving outcome of unrelated HSCT have made this procedure a reality for patients in need[31,32]. The success of unrelated-donor HSCT increases commensurate to the degree of Human Leukocyte Antigens (HLA) matching between the donor and the recipient[33–36]. As ethnicity affects HLA allele and haplotype frequency and thereby influences the outcome of a donor search, a national registry's donor pool should reflect the ethnicities of that

nation's population[27].

The Ezer Mizion Bone Marrow Donor Registry (EM BMDR), established in 1998, is the largest Jewish registry in the world with a roster of more than 800,000 volunteer adult donors. From its inception and through Dec 2015 the EM BMDR provided 950 Hematopoietic Product Cells (HPC) for Israeli patients and 1176 HPCs for patients from outside of Israel. Establishing a comprehensive donor pool for the Israeli population is an immense challenge. Contemporary Jews comprise an aggregate of ethnoreligious communities in Israel and in the Jewish Diaspora. Genetic divergence within the greater Jewish population was caused by admixture with indigenous host populations on a backbone of Mediterranean ancestry, while cultural and religious forces maintained coherence of the Jewish people[13,14,37,38]. Israel is a home to the entire genetic spectrum of the Jewish Diaspora as well as to large minorities comprised of non-Jewish ethnic groups, leading to substantial ethnic diversity in a country of only 8 million individuals. As second-generation Israelis start marrying outside of their ancestral sub-ethnicities, an additional layer of diversity has been introduced into the HLA landscape in Israel, leading to immunogenetic inter- generational differences within the Israeli population[16].

Since 2005, the EM BMDR has enrolled stem cell volunteer donors primarily at the central induction center of the Israel Defense Forces where military conscripts, at the age of 18, are offered enrolment as part of their enlistment process. All Israeli teenagers (male and female) of Jewish, Druze, Bedouin or Circassian descent are required to

register at this center on or around their 18th birthday. EM BMDR's recruitment strategy has increased the number of young, healthy donors in the registry (37% of registered donors are now between 18 and 25 years of age) and has enhanced its HLA diversity to reflect the representation of nearly all the sub- ethnicities in the Israeli population (according to the Israeli Census report 2014, table 2.8). EM BMDR also conducts ethnically focused donor drives within Jewish, Arab and Druze communities in efforts to enhance the ethnic representation of specifically targeted groups. Approximately 10% of the Israeli adult population is registered in the EM BMDR making it the registry with the highest number of HLA-A, -B, -DR stem cell donors per 10,000 inhabitants in the world[39]. Despite the size of this donor pool, many Israeli patients in need of a MUD stem cell transplant do not find a suitably matched local donor[32]. According to the 2014 WMDA annual report, 72% of the MUD products required for transplants in Israel were procured from Israeli BMDR's, of which 68% were provided by EM BMDR[39].

Haplotype frequency estimation using phenotypic population data can permit, among other things, an estimate of the size of a theoretical donor pool that will meet the needs of a specific patient population[40–42]. In this paper, the likelihood of finding donors at different matching stringencies was computed for hypothetical patients from each subethnicity, in an effort to show the ability of the current registry (size and sub-ethnic representation) to provide stem cell donors for the Israeli population. We also project the effects of donor registry growth on the likelihood of successful donor searches within the registry.

METHODS

Study Population

The initial data set included all 754,135 adult volunteer donors registered at EM BMDR from its inception through June 2014. All subjects provided informed consent for registration at recruitment and provided self-reported information regarding the country of origin of each of their parents. Donors were asked to write in the country of parental origin and did not choose from a registry-generated list, so as not to limit their answers. The study was approved by the Rabin Medical Center Ethics Committee. Study subjects were restricted to those individuals who reported the same sub-ethnicity for both parents. Multi- or mixed-ethnicity donors were excluded from this analysis. The USA, Argentina, and Ashkenazi (donors who did not specify their parents' geographical origin but indicated that they are of Ashkenazi ancestry) sub-ethnic populations are predominantly emigrants from Eastern and Western Europe. The SEE (Southeast Europe) sub-ethnic population includes Jews from Romania, Bulgaria, Moldova, Greece, Yugoslavia, Albania, Serbia, Transylvania, and Cyprus. The USSR sub-ethnic population includes Jews from Russia, Ukraine, Belarus, Lithuania, Latvia and Eastern Europe. Donors who listed "Israel" as their sub-ethnicity reflect a diverse group whose parents do not necessarily share the same ethnic origin; these donors did not indicate their parents' ethnicity but their parents' country of birth. For the purpose of this study, all EM BMDR donors who met the above criteria were used as potential patients seeking an unrelated stem cell donor.

Modeling and Definitions

Match rate projection tools were developed by the National Marrow Donor Program (NMDP)/Be The Match® Bioinformatics Research Department and were previously applied to the Be The Match (USA) registry[43]. For the purpose of this study, HLA-matching models were based on donor–recipient three-locus (HLA-A, -B, and - DRB1) high-resolution haplotype frequencies of the EM BMDR sub-ethnicities[28]. Matching at all three loci is termed a 6/6 HLA match. A ≥5/6 allele match includes matches of all six alleles of the donor-recipient pairs or allows for a single mismatched allele ("5/6 or better"). In addition, we analyzed the probability of identifying donors from “within” and "outside" of the patient’s sub ethnic group accounting for current donor availability rates. The match rate for searches outside the donor's ethnic group is calculated by subtracting the "within" match rate from the "cumulative" match rate using the entire registry, it therefore represents the matches that could not be found from first searching within the given population.

Availability of Donors

Many factors affect the availability of registered volunteer donors[44]. Availability rates of potentially HLA-matched donors are given for the three stages of the MUD search process: Confirmatory Typing (CT), donor validation, and medical clearance. At the CT stage, a blood sample is obtained from identified potentially matching volunteers to confirm HLA typing. Preliminary serological testing for infectious agents is also performed. Donor unavailability at this stage may result from failure to locate the donor,

donor health-related issues, or scheduling conflicts. At the donor validation stage, the initially inferred HLA typing at preliminary search is compared with the results of CT; discrepancies might invalidate the donor. Lastly, MUDs receive detailed information about the donation process and their medical eligibility is determined; at this stage, a donor may decline to continue or may be deemed medically unfit to donate stem cells. Donor availability rates were similar among the various sub-ethnic populations in the EM BMDR registry, and the cumulative availability factor was calculated by multiplying the percentages of availability at the three respective stages, treating each stage as provisional for each subsequent event (Table 4). Match rates were adjusted for availability in our model by multiplying the number of donors in each analyzed population by this cumulative availability factor[43].

Table 4. Adult-Donor Availability in 2015 requested from EM BMDR

†Confirmatory Typing Available (%)	‡ Typing not Discrepant (%)	§ Workup Available (%)	Available Overall (%)
80	99.4	94	75

† Shown are data for donors who can be contacted and who have a DNA sample collected for confirmatory HLA typing.
‡ Shown are data for donors whose confirmatory HLA typing was consistent with HLA typing performed at recruitment.
§ Shown are data for donors who were cleared as healthy by means of a medical examination and who agreed at this stage to proceed toward donation.

Statistical Analysis

We included 275,699 donors of the initial dataset from 67 sub-ethnicities who had been genotyped by DNA methods at HLA-A, -B and -DRB1 loci, with an average sample size of 5,200 per sub-ethnicity [2 – 69,717]. Due to progressive changes in HLA genotyping technology, the registry data contains donors at varying levels of resolution. Only sub-ethnicities with a minimum of 200 high-resolution (first two nomenclature fields) typed

samples at HLA-A, -B and -DRB1 were included in this study; nineteen sub-ethnic populations met this selection criterion[44]. We used the Expectation Maximization (EM) algorithm designed to handle mixed resolution data and to resolve both allelic and phase ambiguity[21–23,45], in order to estimate HLA A~B~DRB1 haplotype frequencies for each population[28]. The haplotype frequencies and effective donor registry sizes for each population were put into a matching model[46–48] and deviations from Hardy–Weinberg equilibrium (HWE) were 7 assessed at the allele-family level[28]. We used the model to calculate the population- specific HLA match rates for the given registry size and match definitions. Match rates were defined as the likelihood at which each individual from a given sub-ethnic population would find an allele matched adult donor by searching the same sub- ethnic population, other sub-ethnic populations, or the entire EM BMDR donor list. Adult donor match rates were modeled for each of the sub-ethnic populations over a range of potential registry sizes that predicted growth proportional to the initially reported sub-ethnic population size. A projected annual registry growth rate of 6% for the next 10 years was selected for this analysis based on annual growth rates of the registry during the previous decade. We assume a proportional expansion of each sub-ethnicity from its current representation based on this growth rate and did not account for future ethnically-driven donor recruitment drives.

Marginal Benefit Analysis

We modeled the allele-level 6/6 and ≥5/6 adult donor match rates for patients utilizing the donor pools of both EM BMDR and Be the Match registries, in order to assess the

effect conferred by the existence of EM BMDR on finding donors for potential patients drawn from the study populations enumerated above. We used previously published A~B~DRB1 haplotype frequencies for 21 race groups for modeling the Be the Match registry. The number of donors in each race group ranged from 1,469 for Alaskan Native to 2,899,081 for European. A donor availability factor could not be assessed for some ethnicities due to the low number of transplants performed for minority patients in Be the Match registry. We accounted for donor availability in the models by multiplying the number of donors in each of the 21 populations in Be the Match registry by the donor availability of their corresponding broad race group. We calculated match rates attained from searching 8 only the Be the Match registry for all individuals from each of the 19 sub-ethnic populations analyzed from the EM BMDR at the 6/6 and ≥5/6 level, and termed this "marginal benefit"[24].

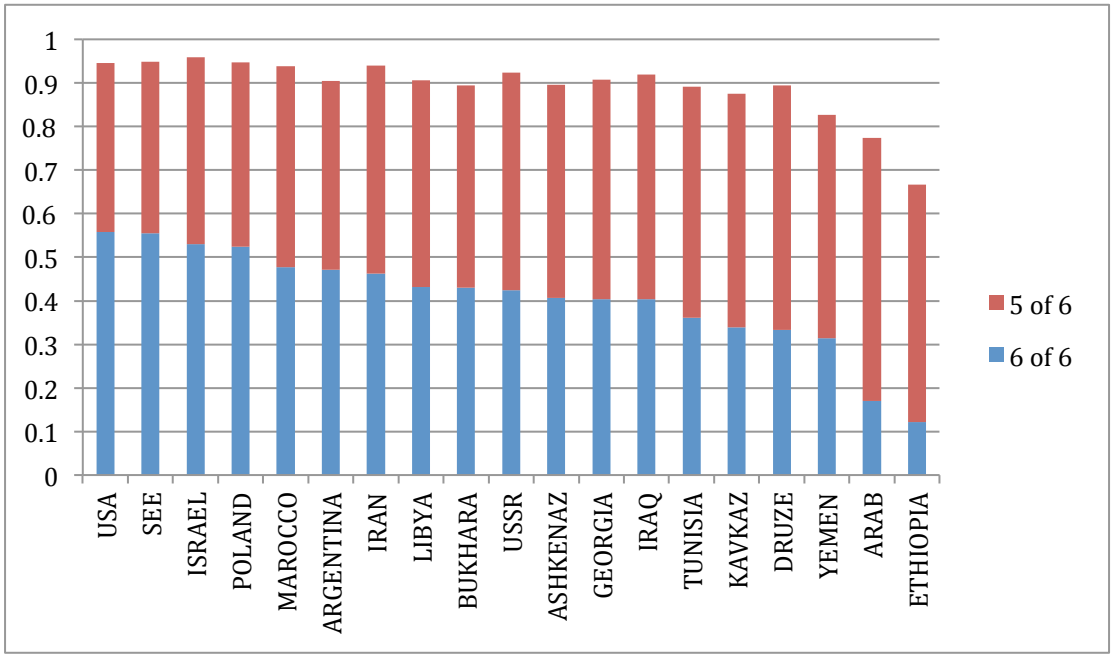

Figure 4. 6/6 and 5/6 stacked match rates considering 75% overall donor's availability. For each population, the shown match rates describe the rate at which patients from this population would find an allele-matched donor by searching the entire Ezer Mizion registry.

RESULTS

Adult Donor Match Rates

Figure 4 shows the 5/6 and 6/6 overall match rates for the sub-ethnic populations analyzed in the EM BMDR considering a 75% cumulative donor availability rate. Most patients will have a 6/6 or ≥5/6 HLA MUD available from within the registry. For the majority of sub-ethnic populations, 6/6 match rates are 40-55%, with the exception of the Tunisia (36%) Kavkaz (33%), Druze (33%) Yemeni (31%), Arab (17%) and Ethiopian (12%) populations. When allowing for a single HLA-allele mismatched donor, potential transplant candidates from all sub-ethnic populations have a match rate ≥ 80% with the exception of Arab (77%) and Ethiopian (66%) sub- ethnic populations. Although 6/6 and

≥5/6 loci HLA-match rates are greatest within most sub-ethnic population (with the exception of Argentina, Ashkenaz, SEE and USA – relatively small and very heterogenous groups)(Table 6), exploiting potential donors from other ethnic groups may enhance the chances of finding suitably matched donors (Supplementary Table S7). Individuals belonging to sub-ethnic populations with high levels of genetic admixture, (Ashkenazi and European [Argentina, Poland, USA], have 21-31% 6/6 matched rate with donors identified outside their ethnic group. By contrast, potential patients from sub-ethnicities with lower levels of genetic admixture and unique allele representations[44], rarely find extramural 6/6 matched rate donors (Bukhara 7.6%; Druze 7.7%; Yemen 2.3%; Ethiopia 0.8%). It should be noted that patients belong to sub-ethnicities with large sample size (Israel, Morocco, USSR) also present low rate of extramural matching since they have higher chance to find matched donors from within their own population.

Table 5. 6/6 and ≥5/6 adult donor match rates and marginal benefit for EM BMDR patients utilizing the donor pools of both EM BMDR and Be the Match®

Ethnicity	Cumulative 6/6	Marginal Benefit 6/6	Cumulative ≥5/6	Marginal Benefit ≥5/6
ARAB	0.394	0.225	0.945	0.171
ARGENTINA[a]	0.542	0.071	0.954	0.050
ASHKENAZ	0.499	0.091	0.948	0.053
BUKHARA	0.487	0.056	0.937	0.042
DRUZE	0.389	0.056	0.950	0.057
ETHIOPIA	0.169	0.046	0.822	0.155
GEORGIA	0.469	0.065	0.946	0.038
IRAN	0.544	0.082	0.978	0.039
IRAQ	0.504	0.100	0.965	0.045
ISRAEL	0.599	0.069	0.983	0.024
KAVKAZ	0.479	0.140	0.954	0.079
LIBYA	0.502	0.070	0.954	0.048
MAROCCO	0.560	0.084	0.975	0.036
POLAND	0.654	0.130	0.987	0.039
SEE[b]	0.629	0.074	0.977	0.029
TUNISIA	0.460	0.099	0.948	0.056
USA	0.643	0.085	0.977	0.030
USSR[c]	0.568	0.143	0.975	0.052
YEMEN	0.343	0.029	0.878	0.051

[a] Argentina and USA population are derived from emigrants of European Jews
[b] SEE include Romania, Bulgaria, Moldova, Greece, Yugoslavia, Albania, Serbia, Transylvania and Cyprus
[c] USSR include Russia, Ukraine, Belarus, Lithuania, Latvia and East Europe

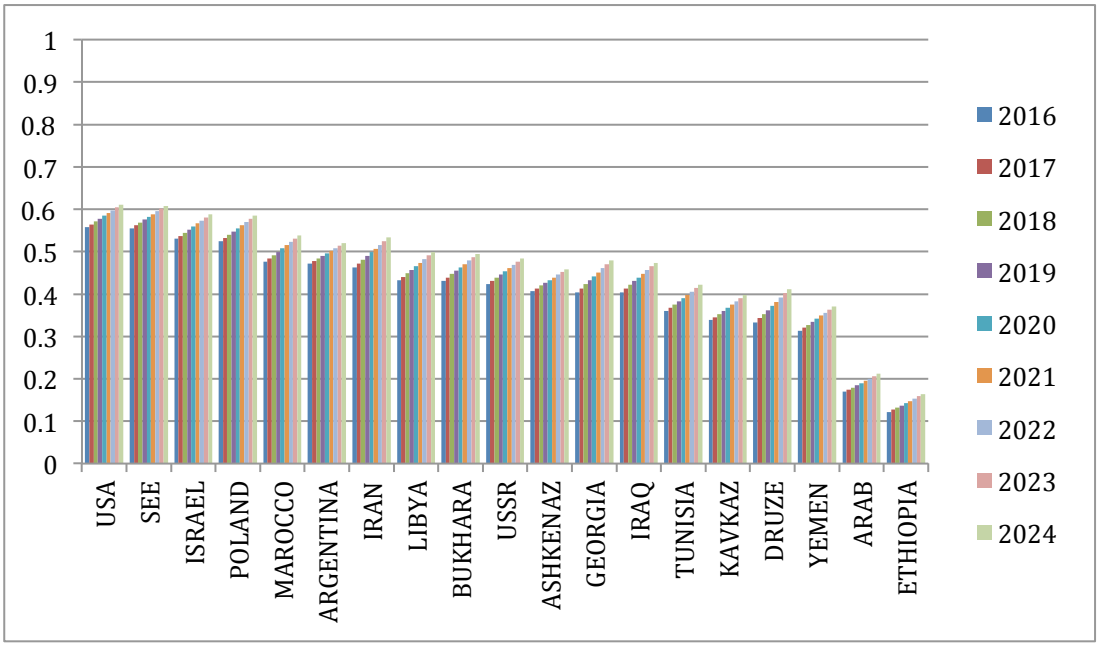

Figure 5. Adult donor 6/6 projected match rates (based on current donor availability) for each sub-ethnic population were modulated for 2014 through 2024. Calculations were based on anticipated growth of 6% cumulatively each year.

Donor Registry Growth

We projected expansion of the registry roster from 2016 to 2026. Projected match rates were calculated based on anticipated growth of 6% cumulatively each year and 75% donor availability. We forecast that aggregate 6/6 (Figure 5) match rates will improve by 0.5 to 1% per population per year through 2026. We compared changes in match rates for specific sub-ethnicities and found that Druze (9.8%) and Georgia (9.3%) populations experienced the most rapid growth in this metric. As expected, populations who already have high match rates using the current registry donor roster and who have high levels of genetic admixture, (Ashkenazi and European) will reap lower benefits from projected registry growth (6-7.5%). We also modeled 6/6 and ≥5/6 HLA-allele donor match rates given a doubling the total registry roster from 750,000 to 1,500,000 donors (Figure S8 in supplementary material). “Six of 6” allele aggregate match rates rose from 52% to 62%

for the entire registry cohort. The increase of donor-recipient matches by sub-ethnicity ranged from 6.8% – 12.3%, with highest levels among the Druze community. The aggregate increase in match rate for ≥5/6 donor-recipient pairs with doubling of the registry size was a more modest 2.5% (from 94.2% to 96.7%), with Ethiopian, Arab and Yemeni match rates rising 7.9%, 5.3% and 4%; all other groups showed only nominal increases.

Marginal Benefit Analysis with Be the Match Donor Registry

Allele-level 6/6 match rates for patients searching both the Ezer Mizion and Be the Match registries ranged from 16.9% to 65.4% for Ethiopia and Poland sub- ethnicities, respectively (Table 6). Cumulative 6/6 match rates in Table 5 represent the sum of the cumulative 6/6 match rates from Table 6 (using the EM BMDR pool alone) plus the marginal benefit added by searching the Be The Match pool of donors. For most potential patients searching both registries, the likelihood of finding a 6/6 matched donor exceeds 46%; exceptions include Ethiopian, Druze, Yemeni and Arab individuals. On average, potential patients from the Ethiopia, Druze, Yemen, and Arab sub-ethnic populations are 22% less likely to find a 6/6 matched donor in both registries than individuals from other populations. When allowing for a single HLA-allele mismatch ("5/6 or better"), most sub-ethnic populations have match rates exceeding 94%, with the exception of the Yemen and Ethiopia sub-ethnic populations.

Table 6. Probability of identifying adult donors from within and outside of the patient's sub ethnic group incorporating consideration of donor availability.

Ethnicity	Cumulative 6/6	Within Population 6/6	Outside Population 6/6	Cumulative ≥5/6	Within Population ≥5/6	Outside Population ≥5/6
ARAB	0.170	0.069	0.101	0.774	0.543	0.231
ARGENTINA[a]	0.471	0.163	0.309	0.904	0.594	0.310
ASHKENAZ	0.407	0.164	0.243	0.895	0.595	0.301
BUKHARA	0.431	0.355	0.076	0.895	0.751	0.144
DRUZE	0.333	0.256	0.077	0.894	0.762	0.132
ETHIOPIA	0.122	0.114	0.008	0.667	0.565	0.101
GEORGIA	0.404	0.314	0.091	0.908	0.768	0.141
IRAN	0.463	0.381	0.082	0.940	0.834	0.105
IRAQ	0.404	0.318	0.086	0.919	0.820	0.100
ISRAEL	0.530	0.446	0.084	0.959	0.918	0.041
KAVKAZ	0.338	0.217	0.122	0.875	0.610	0.265
LIBYA	0.432	0.280	0.152	0.907	0.720	0.187
MAROCCO	0.477	0.420	0.057	0.939	0.884	0.055
POLAND	0.525	0.309	0.216	0.948	0.797	0.151
SEE[b]	0.555	0.275	0.281	0.948	0.766	0.182
TUNISIA	0.361	0.205	0.155	0.892	0.716	0.176
USA	0.558	0.239	0.319	0.947	0.708	0.239
USSR[c]	0.424	0.327	0.097	0.923	0.852	0.072
YEMEN	0.314	0.291	0.023	0.827	0.753	0.074

[a] Argentina and USA population are derived from emigrants of European Jews
[b] SEE include Romania, Bulgaria, Moldova, Greece, Yugoslavia, Albania, Serbia, Transylvania and Cyprus
[c] USSR include Russia, Ukraine, Belarus, Lithuania, Latvia and East Europe

The average marginal benefit achieved for potential patients from EM BMDR by searching the Be the Match donor pool for a 6/6 matched donor was 9%, but varied by sub-ethnicity. Specifically, marginal benefits of this extramural search ranged from 3% for Yemeni recipients to 22% for Arab patients. The likelihood of finding a 6/6 matched donor for potential EM BMDR recipients only in the Be the Match® registry is < 10% for most patients, except for Poland (13%), Kavkaz (14%), USSR (14%), and Arab (22%) sub-ethnic populations. Comparative figure for equivalent match rates in EM BMDR are 52%, 33%, 42%, and 17% for these same populations, respectively. The Arab population was the only population more likely to find a 6/6 matched donor in the Be the

Match registry than in EM BMDR. The likelihood of finding a ≥5/6 matched donor only in the Be the Match registry for EM BMDR potential patients was less than 8%, except for the Ethiopia (15%) and Arab (17%) sub-ethnic populations.

DISCUSSION

Analysis of a target population's HLA profile is integral to strategic planning for the establishment and expansion of stem cell donor registries that will provide an optimal representation of the population that they are meant to serve. We used HLA A~B~DRB1 haplotype frequencies from 19 sub-ethnic populations of adult volunteer donors from within our registry to estimate match rates for hypothetical domestic patients given the current size of the EM BMDR. We assessed the effects that non- directed short-term growth of the registry (expansion of the donor pool without ethnicity-directed donor drives) would have on matching frequencies for Israelis patients of these sub-ethnicities who seek a stem cell donor. Using a population- based genetic model[43] with selection criteria that predict the likelihood of finding 6/6 or ≥5/6 HLA-matched donor and accounting for our current rate of donor availability, we charted the effect that enlarging the registry would have on HLA-match rates over the next decade. Each donor whose parents were of the same sub-ethnicity was used as a potential stem cell transplant recipient for the purpose of this analysis.

Our results show that 40-55% of potential transplant recipients from most of the sub- ethnic populations studied will find a 6/6 HLA allele-matched donor within the EM BMDR registry. Searches performed at lower stringency (≥5/6 matching stringency),

resulted in match rates of ≥ 80% for patients from most of the sub-ethnic populations. Some sub-ethnic populations continue to have lower match rates within EM BMDR because of lower representation in the registry (Arab and Druze donors) or due to distinct HLA allele frequencies, likely the result of limited admixture with the greater Jewish Diaspora (Ethiopian, Kavkazi, and Yemen). Given projected donor enrollment rates, we anticipate adding approximately 500,000 donors to EM BMDR by 2026. We predict an aggregate improvement in 6/6 match rates of 0.5% to 1% per population per year during this period (Figure 5). Although directed donor recruitment will alter this dynamic for currently under-represented groups, all other population in our registry will benefit from ongoing recruitment efforts. More importantly, our model does not account for future shifts in the immigration patterns or for multi-ethnic admixtures in Israeli society over the coming decade. Assuming that the discrete sub-ethnicities on whom we have reported maintain their unique HLA haplotype frequencies, we forecast improved match rates for some currently under-represented sub-ethnicities that are moderately higher than the effects of registry growth on the EM BMDR population as a whole. A registry's mandate is to expeditiously identify the best available donor[49]. Our findings are consistent with those reported by Gragert et al.,[43] and point to the futility of delaying transplant for a patient who does not find a donor in the hope that one will be identified in the near future.

The EM BMDR is predominantly supported by charitable and personal contributions, and donor recruitment is a costly process. As the stewards of philanthropic funds, we must allocate the resources of the registry responsibly. As such, we assessed

the need for registry expansion by looking at the match rates from extra-mural donors, taken in this case from the Be the Match registry. Results of this analysis showed that existence of the EM BMDR adds substantial benefit for Israelis who would require an unrelated stem cell transplant. Of interest, the Arab population in our country gained most from searching the extramural registry. We project that directed donor drives in this community will be required to improve local match rates for Arab patients. Our model calculated match rates using 6 HLA alleles typed at high resolution. Searches at higher levels of stringency, such as inclusion of typing at HLA-C, will likely result in reduced match rates[50]. Additionally, our models assume HWE and were solely conducted using donors who listed both of their parents as belonging to the same sub-ethnicity. Recent data suggests that contemporary Israel is undergoing profound ethnic changes[15]; East is meeting West and they are bearing children. As such, models assuming HWE may not be valid. The results of sub-ethnic admixture in contemporary Israeli society are already apparent in the lower numbers of fully HLA matched unrelated donors actually found for children as compared with adult transplant recipients in Israel[16]. Exploring haplotype frequencies and match rates for the growing multi-ethnic population base in the EM BMDR will set the foundation for elaborating strategies for recruitment and expansion of the registry, and will highlight the value added to the international community through the contribution of multi- ethnic donors in an increasingly globalized community.

This study utilized donor self-reporting of sub-ethnicity, which has been shown to be less than completely reliable[27]. Additionally, a large group of donors was excluded

from analysis due to absence of data regarding parental ethnicity or due to multi- ethnic lineages. As noted above, changes in reproductive patterns in contemporary Israel will likely change the sub-ethnic landscape in the coming decades. As sub- ethnic admixture will likely result in changing representation of HLA-allele frequencies in our populations, we have begun to collect data regarding grandparental sub-ethnicities in an effort to guide future analyses. Our data-driven approach will help plan the expansion and recruitment policies of the EM BMDR and aid Israeli and non-Israeli patients worldwide in their search for a stem cell donor.

BIBLIOGRAPHY

1. Hansen, J. A. *et al.* Transplantation of Marrow from an Unrelated Donor to a Patient with Acute Leukemia. *N. Engl. J. Med.* **303,** 565–567 (1980).
2. Shaw, B. E., Arguello, R., Garcia-Sepulveda, C. A. & Madrigal, J. A. The impact of HLA genotyping on survival following unrelated donor haematopoietic stem cell transplantation. *Br. J. Haematol.* **150,** 251–258 (2010).
3. Tiercy, J.-M. How to select the best available related or unrelated donor of hematopoietic stem cells? *Haematologica* **101,** 680–7 (2016).
4. Cereb, N., Kim, H. R., Ryu, J. & Yang, S. Y. Advances in DNA sequencing technologies for high resolution HLA typing. *Hum. Immunol.* **76,** 923–927 (2015).
5. Robinson, J. *et al.* The IPD and IMGT/HLA database: Allele variant databases. *Nucleic Acids Res.* **43,** D423–D431 (2015).
6. Loiseau, P. *et al.* HLA Association with Hematopoietic Stem Cell Transplantation Outcome: The Number of Mismatches at HLA-A, -B, -C, -DRB1, or -DQB1 Is Strongly Associated with Overall Survival. *Biol. Blood Marrow Transplant.* **13,** 965–974 (2007).
7. Soormally, A. R., Hayhurst, J. D. & Marsh, S. G. E. The IPD-IMGT/HLA Database – New developments in reporting HLA variation. *Hum. Immunol.* **77,** 233–237 (2016).
8. Marsh, S. G. E. *et al.* Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* **75,** 291–455 (2010).
9. Mack, S. J. A gene feature enumeration approach for describing HLA allele polymorphism. *Hum. Immunol.* **76,** 975–981 (2015).
10. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539–539 (2014).
11. Cruz, S. Nextflow enables reproducible computational workflows. (2017). doi:10.1145/2443416.2443417
12. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* **10,** 402–15 (2009).
13. Malamat, A. & Ben-Sasson, H. H. *A History of the Jewish people*. (Harvard University Press, 1976).
14. Klitz, W. *et al.* Genetic differentiation of Jewish populations. *Tissue Antigens* **76,** 442–458 (2010).
15. Mahler, R. A History of Modern Jewry. *Schocken* (1971).
16. Israeli, M. *et al.* Age-dependent HLA profiles of the Israeli population: impact on hematopoietic cell donor recruitment and availability. *Immunogenetics* **66,** 525–533 (2014).
17. Bonn?-Tamir, B. *et al.* HLA Polymorphism in Israel: 9. An Overall Comparative Analysis. *Tissue Antigens* **11,** 235–250 (1978).
18. Brautbar, C., Battat, S., Sherman, L., Benhamu, R., Cohen, O. HLA antigens in

Israeli Ashkenazi and non-Ashkenazi Jews. *HLA Asia-Oceania, Proc. Third Asia Ocean. Histocompat. Work. Conf.* 324–327
19. Amar, A. *et al.* Molecular analysis of HLA class II polymorphisms among different ethnic groups in Israel. *Hum. Immunol.* **60,** 723–730 (1999).
20. Roitberg-Tambur, A. *et al.* HLA polymorphism in Moroccan Jewry. *Hum. Immunol.* **40,** 61–67 (1994).
21. Excoffier, L. & Slatkin, M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12,** 921–927 (1995).
22. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. (1976).
23. Kollman, C. *et al.* Estimation of HLA-A, -B, -DRB1 haplotype frequencies using mixed resolution data from a National Registry with selective retyping of volunteers. *Hum. Immunol.* **68,** 950–8 (2007).
24. Gragert, L., Madbouly, A., Freeman, J. & Maiers, M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.* **74,** 1313–20 (2013).
25. Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P. & Thomson, G. PyPop update--a software pipeline for large-scale multilocus population genomics. *Tissue Antigens* **69 Suppl 1,** 192–7 (2007).
26. Karypis, G. CLUTO - A Clustering Toolkit. (2002).
27. Hollenbach, J. A. *et al.* Race, Ethnicity and Ancestry in Unrelated Transplant Matching for the National Marrow Donor Program: A Comparison of Multiple Forms of Self-Identification with Genetics. *PLoS One* **10,** e0135960 (2015).
28. Gragert, L. *et al.* HLA Match Likelihoods for Hematopoietic Stem-Cell Grafts in the U.S. Registry. *N. Engl. J. Med.* **371,** 339–348 (2014).
29. Appelbaum, F. R. Hematopoietic-Cell Transplantation at 50. *N. Engl. J. Med.* **357,** 1472–1475 (2007).
30. Copelan, E. A. Hematopoietic Stem-Cell Transplantation. *N. Engl. J. Med.* **354,** 1813–1826 (2006).
31. Gratwohl, A. *et al.* Hematopoietic Stem Cell Transplantation<subtitle>A Global Perspective</subtitle> *JAMA* **303,** 1617 (2010).
32. Baldomero, H. *et al.* The EBMT activity survey 2009: trends over the past 5 years. *Bone Marrow Transplant.* **46,** 485–501 (2011).
33. Lee, S. J. *et al.* High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood* **110,** (2007).
34. Fürst, D. *et al.* High-resolution HLA matching in hematopoietic stem cell transplantation: a retrospective collaborative analysis. *Blood* **122,** 3220–9 (2013).
35. Horan, J. *et al.* Evaluation of HLA matching in unrelated hematopoietic stem cell transplantation for nonmalignant disorders. *Blood* **120,** 2918–2924 (2012).
36. Dehn, J. *et al.* 8/8 and 10/10 high-resolution match rate for the be the match unrelated donor registry. *Biol. Blood Marrow Transplant.* **21,** 137–41 (2015).
37. Kwon, O. J. *et al.* Immunogenetics of HLA class II in Israeli Ashkenazi Jewish, Israeli non-Ashkenazi Jewish, and in Israeli Arab IDDM patients. *Hum. Immunol.*

62, 85–91 (2001).
38. Behar, D. M. *et al.* The genome-wide structure of the Jewish people. *Nature* **466,** 238–242 (2010).
39. WMDA Annual Report.
40. Müller, C. R., Ehninger, G. & Goldmann, S. F. Gene and haplotype frequencies for the loci hLA-A, hLA-B, and hLA-DR based on over 13,000 german blood donors. *Hum. Immunol.* **64,** 137–51 (2003).
41. Kollman, C. *et al.* Assessment of optimal size and composition of the U.S. National Registry of hematopoietic stem cell donors. *Transplantation* **78,** 89–95 (2004).
42. Sonnenberg, F., Eckman, M. & Pauker, S. Bone marrow donor registries: the relation between registry size and probability of finding complete and partial matches [see comments]. *Blood* **74,** (1989).
43. Manor, S. *et al.* High-resolution HLA A~B~DRB1 haplotype frequencies from the Ezer Mizion Bone Marrow Donor Registry in Israel. *Hum. Immunol.* **77,** 1114–1119 (2016).
44. Confer, D. L. The National Marrow Donor Program. Meeting the needs of the medically underserved. *Cancer* **91,** 274–8 (2001).
45. Gragert, L., Madbouly, A., Freeman, J. & Maiers, M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.* **74,** 1313–1320 (2013).
46. Beatty, P. G., Boucher, K. M., Mori, M. & Milford, E. L. Probability of finding HLA-mismatched related or unrelated marrow or cord blood donors. *Hum. Immunol.* **61,** 834–40 (2000).
47. Mori, M., Graves, M., Milford, E. L. & Beatty, P. G. Computer program to predict likelihood of finding and HLA-matched donor: methodology, validation, and application. *Biol. Blood Marrow Transplant.* **2,** 134–44 (1996).
48. Dehn, J. *et al.* HapLogic: A Predictive Human Leukocyte Antigen?Matching Algorithm to Enhance Rapid Identification of the Optimal Unrelated Hematopoietic Stem Cell Sources for Transplantation. *Biol. Blood Marrow Transplant.* **22,** 2038–2046 (2016).
49. Spellman, S. R. *et al.* A perspective on the selection of unrelated donors and cord blood units for transplantation. *Blood* **120,** 259–265 (2012).
50. Eberhard, H.-P. & M??ller, C. R. The Impact of HLA-C Matching on Donor Identification Rates in a European-Caucasian Population. *Front. Immunol.* **5,** 501 (2014).

APPENDIX A

Locus	Feature	Rank	Number Matched	Percent Matched
HLA-A	EXON	1	383	50.86%
HLA-A	EXON	2	2667	73.19%
HLA-A	EXON	3	1495	41.03%
HLA-A	EXON	4	1348	97.89%
HLA-A	EXON	5	383	52.04%
HLA-A	EXON	6	383	76.14%
HLA-A	EXON	7	383	76.14%
HLA-A	EXON	8	383	77.53%
HLA-A	FIVE_PRIME_UTR	1	88	26.83%
HLA-A	INTRON	1	383	100.00%
HLA-A	INTRON	2	383	100.00%
HLA-A	INTRON	3	383	100.00%
HLA-A	INTRON	4	383	100.00%
HLA-A	INTRON	5	383	100.00%
HLA-A	INTRON	6	383	100.00%
HLA-A	INTRON	7	383	100.00%
HLA-A	THREE_PRIME_UTR	8	353	100.00%
HLA-B	EXON	1	416	44.30%
HLA-B	EXON	2	3175	71.28%
HLA-B	EXON	3	1730	38.84%
HLA-B	EXON	4	1602	99.63%
HLA-B	EXON	5	416	49.52%
HLA-B	EXON	6	416	77.18%
HLA-B	EXON	7	416	78.64%
HLA-B	FIVE_PRIME_UTR	1	6	1.78%
HLA-B	INTRON	1	416	100.00%
HLA-B	INTRON	2	416	100.00%
HLA-B	INTRON	3	416	100.00%
HLA-B	INTRON	4	415	99.76%
HLA-B	INTRON	5	416	100.00%
HLA-B	INTRON	6	416	100.00%
HLA-C	EXON	1	590	71.78%
HLA-C	EXON	2	3267	99.30%
HLA-C	EXON	3	651	19.79%
HLA-C	EXON	4	1171	96.94%

HLA-C	EXON	5	590	72.66%
HLA-C	EXON	6	590	74.87%
HLA-C	EXON	7	590	84.17%
HLA-C	EXON	8	590	91.90%
HLA-C	FIVE_PRIME_UTR	1	13	2.37%
HLA-C	INTRON	1	590	100.00%
HLA-C	INTRON	2	590	100.00%
HLA-C	INTRON	3	590	100.00%
HLA-C	INTRON	4	590	100.00%
HLA-C	INTRON	5	590	100.00%
HLA-C	INTRON	6	590	100.00%
HLA-C	INTRON	7	590	100.00%
HLA-C	THREE_PRIME_UTR	8	556	100.00%

Table 7. The number of times the expected accession number for a given feature and rank matches the observed.

Locus	Feature	Rank	Count
HLA-A	EXON	2	137
HLA-A	EXON	3	222
HLA-A	EXON	4	5
HLA-A	EXON	5	14
HLA-B	EXON	2	163
HLA-B	EXON	3	285
HLA-B	EXON	5	2
HLA-C	EXON	3	270
HLA-C	EXON	4	3

Table 8. The number of times the expected and observed sequences for a given feature and rank are off by three base pairs at the beginning or end of the sequence.