High-resolution characterization of allelic and haplotypic HLA frequency distribution in a Spanish population using high-throughput next-generation sequencing


⁎ Corresponding author at: Histocompatibility, Immunogenetics & Disease Profiling Laboratory (HIDPL), Department of Pathology, Stanford University, School of Medicine, 3373 Hillview Avenue, Palo Alto, CA 94304, USA.
E-mail address: gmonter2@stanford.edu (G. Montero-Martín).

ARTICLE INFO

Keywords:
Human Leukocyte Antigen (HLA)
Next-Generation Sequencing (NGS)
Population study
Spain

ABSTRACT

Next-generation sequencing (NGS) at the HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1 and -DRB3/4/5 loci was performed on 282 healthy unrelated individuals from different major regions of Spain. High-resolution HLA genotypes defined by full sequencing of class I loci and extended coverage of class II loci were obtained to determine allele frequencies and also to estimate extended haplotype frequencies. HLA alleles were typed at the highest resolution level (4-field level, 4FL); with exception of a minor deviation in HLA-DPA1, no statistically significant deviations from expected Hardy Weinberg Equilibrium (HWE) proportions were observed for all other HLA loci. This study provides new 4FL-allele and haplotype frequencies estimated for the first time in the Spanish population. Furthermore, our results describe extended haplotypes (including the less frequently typed HLA-DPA1 and HLA-DQA1 loci) and show distinctive haplotype associations found at 4FL-allele definition in this Spanish population study. The distinctive allelic and haplotypic diversity found at the 4FL reveals the high level of heterozygosity and specific haplotypic associations displayed that were not apparent at 2-field level (2FL). Overall, these results may contribute as a useful reference source for future population studies, for HLA-disease association studies as a healthy control group dataset and for improving donor recruitment strategies of bone marrow registries. HLA genotyping data of this Spanish population cohort was also included in the 17th International Histocompatibility and Immunogenetics Workshop (IHIW) as part of the study of HLA diversity in unrelated worldwide populations using NGS.

https://doi.org/10.1016/j.humimm.2019.02.005
Received 10 July 2018; Received in revised form 7 February 2019; Accepted 9 February 2019
Available online 11 February 2019
0198-8859/ © 2019 American Society for Histocompatibility and Immunogenetics. Published by Elsevier Inc. All rights reserved.
1. Introduction

Mainland Spain is located on the Iberian Peninsula in Southwestern Europe and it is also very close to North Africa being just separated by the Strait of Gibraltar. The Spanish territory also includes the Balearic Islands in the Mediterranean Sea, the Canary Islands off the North African Atlantic coast and two cities, Ceuta and Melilla, located on the northern coast of Africa. As a consequence of its unique geographic location, Spain shows an extensive cultural diversity within its population (e.g. Spanish is the main language spoken but Catalan in the East, Galician in the Northwest and Euskera in the Western Pyrenees are also spoken languages). The Spanish population diversity resulting from migrations through its history is well documented (e.g. Christian Visigoths, North African Muslims and Sephardic Jews coexisted in the Iberian Peninsula for several centuries) [1]; recent migrations have further increased diversity (e.g. migrants who represent approximately 13% of the Spanish census, who are mainly coming from countries of Eastern Europe, South America and Northern Africa) [2].

In the clinical histocompatibility setting, assessment of HLA allelic and haplotypic diversity of each population is important, especially for hematopoietic stem cell transplantation (HSCT) [3] with unrelated donors. Previous studies have described allelic and haplotypic HLA frequency distribution in Spanish population [4–11]. However, most of these previous studies were only based on lower-resolution HLA typing data (allele resolution level at the 1-field or at the 2-field) generated by traditional HLA molecular typing techniques most commonly used in routine practice (sequence-specific primer (SSP) or sequence-specific oligonucleotide (SSO) probe technologies and sequence-based typing (SBT)). The majority of these earlier studies analyzed only sets of individuals coming from specific regions of Spain. At the same time, HLA typing was performed for certain loci but not for all major HLA loci. As a result, most of these previous studies did not define complete extended HLA haplotypes. Recent works have placed emphasis in the importance of elucidating HLA diversity for all major HLA genes and at a higher allele resolution level for all worldwide populations [12]; this is an unmet need for the Spanish population [4], that may improve current donor search criteria and therapeutic strategies in the HSCT field [3]. Application of next-generation sequencing (NGS) for high-resolution molecular HLA typing has enabled to obtain full-length and/or extended sequences and genotypes of all major HLA genes. This is based on the clonal sequencing nature and the increased read length, throughput, accuracy and resolution that NGS offers for describing the high polymorphism presented by HLA genes [13]. Therefore, NGS-based HLA typing methods permit to overcome many of the limitations of legacy techniques (SSP, SSO and SBT) and also facilitate detection of novel alleles [14]. At the same time, implementation of high-throughput platforms in this NGS technology allows cost-effective and large-scale population genetics studies [12].

The aim of the present study was to describe allelic and haplotypic frequency distribution by typing all major HLA class I and class II genes (HLA- A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1 and -DRB3/4/5) at high-resolution (allele resolution level at the 4-field) and obtaining genotypes by using high-throughput NGS for a representative sample of the Spanish population including a cohort of 282 healthy unrelated individuals from different major regions of the country.

2. Materials and methods

2.1. Sample collection and testing methods

This population study includes 282 healthy unrelated individuals randomly selected from Spain in collaboration with the Spanish Working Group in Histocompatibility and Transplant Immunology (GETHIT) of the Spanish Society for Immunology (SEI). Collection of all genomic DNA samples consisted of 11 participant clinical laboratories that are situated in 10 different locations in Spain (Santander, Salamanca, Madrid (which included 2 different participant clinical laboratories), Barcelona, Valencia, Murcia, Córdoba, Sevilla, Málaga and Gran Canaria) which provided a set of 25–26 samples per institution (Fig. 1). This HLA Spanish population study was approved by the
Institutional Review Board (IRB) of the 17th International Histocompatibility and Immunogenetics Workshop (IHIIW) as well as the respective local research and ethics committee of each Spanish participant institution and it was carried out in accordance with the principles of the Declaration of Helsinki. Samples were tested in parallel: i) All 282 samples were genotyped by using a commercial NGS-based HLA genotyping method [15] at the Stanford Histocompatibility, Immunogenetics and Disease Profiling Laboratory (HIDPL); ii) at the same time, the 11 Spanish participant clinical laboratories performed HLA typing tests (with a variable range of allele resolution level and number of HLA genes tested) of their respective sets of 25–26 samples by using other HLA molecular typing techniques (either using an in-house NGS platform or commercial/in-house SSO or SBT technologies).

2.2. NGS-HLA sequencing and genotyping performed at Stanford HIDPL

All samples were genotyped for HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1 and -DRB3/4/5 loci using the MIA FORA NGS FLEX HLA Typing 11 Kit 96 Tests (Immucor, Inc. Norcross, GA, USA), following manufacturer’s semi-automated protocol and as described previously [15]. Briefly, paired-end sequence reads were generated by using this aforementioned NGS-based HLA genotyping method, which specifically amplifies these 11 HLA genes with extensive coverage of the HLA genomic region by long-range polymerase chain reaction (PCR). All fragments of the final prepared DNA library were sequenced using sequencing by synthesis (SBS) chemistry in a massively parallel fashion on the Illumina® NGS sequencing platform (Illumina, Inc. San Diego, CA, USA). For assignment of HLA genotypes, NGS paired-end reads were analyzed using the MIA FORA FLEX version 3.0 software (Immucor, Inc. Norcross, GA, USA) and according to IPD-IMGT/HLA database version 3.25.0. All HLA genotyping calls automatically assigned by the software were manually reviewed (e.g. for evaluation of ambiguities at the 4-field [see Section 2.3]) and confirmed by the user.

2.3. Standardization of ambiguities at the 4-field

Some HLA assignments resulted ambiguous when trying to distinguish alleles at the 4-field allele resolution level (intronic and untranslated (UTR) sequence level). In these particular cases, called allele candidates present differences only in length of either homopolymer sequences or short tandem repeats (STRs); these were not sequenced candidates present differences only in length of either homopolymer translated (UTR) sequence level). In these particular cases, called allele candidates present differences only in length of either homopolymer sequences or short tandem repeats (STRs); these were not sequenced.

At the 4-field allele resolution level, no overall deviations from expected Hardy-Weinberg equilibrium (HWE) proportions (based on exact test of Guo and Thompson), the Ewens-Watterson homozygosity (EWH) test of neutrality (tested by Slatkin’s implementation of the Monte-Carlo approximation of the Ewens-Watterson exact test, using a two-tailed test (p < 0.05) of the null hypothesis of neutrality) and all pairwise linkage disequilibrium (LD) estimates [17]. Hapl-o-Mat version 1.1 software was used to estimate extended haplotype frequencies from this current Spanish genotypic data using a maximum likelihood estimation via an expectation-maximization (EM) algorithm [18]. Finally, in order to compare allele frequencies (at the HLA-A, -B, -C, -DQB1 and -DRB1 loci) between the 3 different geographical Spanish regions established for this study (Northern-Central, Eastern and Southern Spain) as well as between the 10 Spanish locations studied in the present work, a population dendogram was constructed using POPTREEW (web version of POPTREE software) [19]. A total of 1000 dendrogram replicates based on the matrices of Ne genetic distances (DA) [20] were generated using the neighbor-joining (NJ) method [21].

3. Results

3.1. Evaluation of concordance of HLA typing results obtained from this study

HLA genotyping results obtained for all 282 samples by using this commercial NGS-based method [15] at Stanford HIDPL are 100% concordant with those available HLA typing results (e.g. HLA-DPA1 locus was not tested locally) obtained by using other HLA molecular typing techniques (either using an in-house NGS platform or commercial/in-house SSO or SBT technologies) respectively at the 11 local participant clinical laboratories from Spain (Supplementary Table 1). Therefore, we confirmed that all samples were tested correctly by all the participating laboratories without any sample-switching error, allele dropout (for the HLA loci tested respectively) and neither contamination.

3.2. Evaluation of deviations from expected Hardy-Weinberg equilibrium proportions

At the 4-field allele resolution level, no overall deviations from expected HWE proportions are observed in any of the HLA loci analyzed with the exception of a minor but significant departure at the HLA-DPA1 locus (p-value = 0.0104) (Supplementary Table 2). To further investigate this HLA-DPA1 departure, collapsed 2-field and 3-field HLA genotyping datasets of this same Spanish population cohort (n = 282) were evaluated (data not shown) and no HWE deviation was observed at any of the HLA loci. Furthermore, estimated homozygosity (Waterson’s homozygosity F statistic (F)) in HLA-DPA1 locus at the 4-field allele resolution level shows a much lower value (F = 0.164) in comparison to collapsed 2-field (F = 0.649) and 3-field (F = 0.635) HLA genotyping datasets. Altogether, this can be interpreted as estimated deviations from HWE may not be corrected properly for multiple comparisons including low number counts of alleles or genotypes when they are present at the 4-field allele resolution level. Thus, this observed deviation may be explained by the fact that HLA alleles presenting low frequencies (e.g. HLA-DPA1 alleles) would not be considered properly when evaluating HWE proportions and their contribution to HWE deviation would be being estimated higher than it should be at this 4-field allele resolution level. Overall, the HLA dataset of this present study was considered valid for proceeding with the rest of statistical analyses.

3.3. HLA allelic frequencies in this Spanish population cohort

The frequency distribution of HLA-A, -B, -C, -DPA1, -DPB1, -DQA1, -DQB1, -DRB1, and -DRB3/4/5 alleles at 4-field level of resolution are summarized in Supplementary Table 3. 36 HLA-A, 53 HLA-B, 40
3.4. Identification of two new HLA alleles in this Spanish population cohort

Two novel HLA alleles were identified during this Spanish population study using this aforementioned NGS-based HLA genotyping method [15] (Supplementary Fig. 1). One individual (17th IHII sample ID no. H00035f6, from Barcelona, Spain) presents a single base mismatch with HLA-B*38:20 allele reference sequence in exon 3 (codon 99), which leads to a synonymous substitution (Tyr (TAC) to Tyr (TAT)) (Supplementary Fig. 1a-c). Complete HLA genotyping result of this individual including the novel allele is:

1. HLA-A*29:02:01:01, HLA-A*25:01:01:01, HLA-C*12:03:01:01, HLA-C*03:03:01:01, HLA-B*38:20:02:01, HLA-B*15:01:01:01;
2. HLA-DRB1*03:01:01:01, HLA-DRB3*02:02:01:02, HLA-DRB1*07:01:01:01, HLA-DRB1*13:01:01:01:01, HLA-DQA1*02:01:01:01, HLA-DQB1*06:03:01:01, HLA-DPA1*02:01:01, HLA-DPA1*01:03:01:01, HLA-DPB1*19:01:01, HLA-DPB1*02:01:02.

Also in another different subject (17th IHII sample ID no. H00036d1, from Málaga, Spain), a single base mismatch with HLA-DRB3*02:02:01:01 allele reference sequence is detected in exon 3 (codon 166), which leads in this case to a non-synonymous substitution and, therefore, to an amino acid change (Arg (CGG) to Gln (CAG)) (Supplementary Fig. 1d-1f). Complete HLA genotyping result of this other subject including the novel allele is:

1. HLA-A*11:01:01:01, HLA-A*11:01:01:01; HLA-C*05:01:01:02, HLA-C*15:02:01:01, HLA-B*44:02:01:01, HLA-B*51:01:01:01; HLA-DRB5*01:01:01, HLA-DRB3*02:02:01:01, HLA-DRB1*15:01:01:01, HLA-DRB1*03:01:01:01; HLA-DQA1*05:01:01:01, HLA-DQB1*06:02:01:01, HLA-DQB1*02:01:02, HLA-DPA1*01:03:01:01, HLA-DPA1*01:03:01:01; HLA-DPB1*04:01:01:01, HLA-DPB1*02:01:02.

To confirm these findings, sequence-based typing (SBT) was performed using respective SBT/excellerator kits (GenDx, Utrecht, The Netherlands) on a 3130xl Genetic Analyzer (Applied Biosystems, Foster City, CA, USA) and SBTengine HLA typing software version 3.14.0.2783 (GenDx, Utrecht, The Netherlands) at the corresponding local Spanish clinical laboratories of origin.

Reported sequences of both identified exon variants were submitted to GenBank and to the IPD-IMGT/HLA Database. These two new alleles have been officially assigned by the WHO HLA Nomenclature Committee for Factors of the HL system [22]. In the case of the new HLA-B*38:20 allele, the official name given is HLA-B*38:20:02 (GenBank accession no. MG76848) and IPD-IMGT/HLA submission no. HWS10051845. Regarding the new HLA-DRB3*02:02:01:01 allele, the official name given is HLA-DRB3*02:71 (GenBank accession no. MG922498 and IPD-IMGT/HLA submission no. HWS10051607).

3.5. Ewens-Watterson homozygosity test of neutrality

EWH test of neutrality was used for analysis of selective processes based on HLA allelic diversity at the 4-field allele resolution level of this Spanish population cohort. All HLA loci analyzed showed levels of observed homozygosity (F_o) that are below the expected homozygosity under neutrality (F_e) with the exception of HLA-DBP1 locus (Table 1). Furthermore, HLA-B, -DQA1 and -DQB1 are the only loci that show statistically significant deviation from neutrality and, therefore, are consistent with a more pronounced balancing selection (F_ad < 0). As previously described across human populations [23,24], we also observed for this Spanish population cohort (in spite of presenting a relatively small sample size) an overall direction towards balancing selection for most of the classical HLA class I and II loci with the striking exception of HLA-DP genes. These latter (especially HLA-DRB1 locus, based on our results at the 4-field allele resolution level) seem to be more under directional selection, in which only a set of few alleles become selected (e.g. HLA-DPB1*04:01:01:01). These interpretations however need to be confirmed on a larger Spanish cohort, considering also the diverse nature of the regional subpopulations included in this study.

3.6. 2-locus haplotype linkage disequilibrium analysis

Estimated 2-locus haplotype frequencies and measure of overall LD (Hedrick’s D’ statistic) of pairs of neighboring genetic HLA loci (B ~ C, DPA1 ~ DBP1, DQA1 ~ DQB1 and DQB1 ~ DRB1) at the 4-field allele resolution level are shown in Supplementary Table 4. Interestingly, it can be observed unique 2-locus haplotype associations in non-coding regions at the 4-field allele resolution level that are not apparent at the 2-field level. For instance, alleles of the HLA-B*35 allele group show very distinctive associations with HLA-C alleles at the 4-field level. On one hand, at the 2-field level, HLA-B*35:01, HLA-B*35:02, HLA-B*35:03 and HLA-B*35:08 alleles show a strong and common association with HLA-C*04:01 allele. Nevertheless, at the 4-field level we observed that in the case of the intron variant HLA-B*35:01:01:01 it displays a specific association with HLA-C*04:01:01:01. Whereas, B*35:01:01:02 intron variant presents associations with not only HLA-C*04:01:01:01 allele but also with HLA-C*04:01:01:05 and HLA-C*04:01:01:06 alleles. In the case of HLA-B*35:02:01:01, we observed it seems to display a specific association with HLA-C*04:01:01:06 in Spanish population. As for HLA-B*35:03:01, it presents association with HLA-C*04:01:01:01. Finally, HLA-B*35:08:01 shows association with
HLA-A∼B play a role in determining this measurement. Associations between HLA types as previously reported [25,26], LD patterns of show weaker LD associations than any of the other pairwise comparisons. HLA-B suggests that differences in diversity between HLA-A∼B, DRB1∼DRB5/4/3 the contiguous and/or physically close HLA loci pairs including {HLA-A, -B, -C, -DQA1, -DQB1, -DRB1} and {DRB3/4/5} at the 4-field resolution level (and according to IPD-IMGT/HLA database version 3.25.0) in this Spanish population study (n = 282 subjects).

<table>
<thead>
<tr>
<th>Locus Pair HLA-</th>
<th>D'</th>
<th>Wn</th>
</tr>
</thead>
<tbody>
<tr>
<td>B∼C</td>
<td>0.93630</td>
<td>0.77226</td>
</tr>
<tr>
<td>A∼C</td>
<td>0.59490</td>
<td>0.43926</td>
</tr>
<tr>
<td>A∼B</td>
<td>0.64988</td>
<td>0.45530</td>
</tr>
<tr>
<td>DPB1∼DQB1</td>
<td>0.82896</td>
<td>0.71883</td>
</tr>
<tr>
<td>DQA1∼DQB1</td>
<td>0.97854</td>
<td>0.78901</td>
</tr>
<tr>
<td>DQA1∼DRB1</td>
<td>0.98990</td>
<td>0.86147</td>
</tr>
<tr>
<td>DQB1∼DRB1</td>
<td>0.97446</td>
<td>0.80953</td>
</tr>
<tr>
<td>DPB1∼B</td>
<td>0.47923</td>
<td>0.37854</td>
</tr>
<tr>
<td>DPB1∼DQB1</td>
<td>0.43446</td>
<td>0.38512</td>
</tr>
<tr>
<td>B∼DRB1</td>
<td>0.70620</td>
<td>0.46705</td>
</tr>
<tr>
<td>B∼DQA1</td>
<td>0.66583</td>
<td>0.49954</td>
</tr>
<tr>
<td>B∼DQB1</td>
<td>0.65365</td>
<td>0.49127</td>
</tr>
<tr>
<td>DRB1∼DRB3</td>
<td>0.96724</td>
<td>0.86072</td>
</tr>
<tr>
<td>DRB1∼DQB1</td>
<td>0.97158</td>
<td>0.69772</td>
</tr>
<tr>
<td>DRB1∼DRB5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

HLA-C*04:01:01:01:06. Furthermore, we also found distinct haplotypic associations at the intronic level in several other HLA class I and II loci pairs (e.g. HLA-DQA1*05:01:01 intron variants, HLA-B*18:01:01 intron variants, HLA-C*05:01:01 intron variants or HLA-C*06:02:01 intron variants). In contrast, HLA loci pairs as B*07:02:01~C*07:02:01:01, DQA1*01:01:01:02~DQB1*05:01:01:03 and DQB1*02:02:01:01~DRB1*07:01:01:01 are some examples of 4-field highly conserved associations found in this Spanish population cohort.

To evaluate the overall linkage disequilibrium we considered (Table 2) two different locus-pair-level measures. The D' (normalized Heddrick's D statistic) parameter, expressed as the normalization of the product of allele frequencies at each locus, weights the LD contribution of specific allele pairs [17]. Whereas the second parameter, Wn (Cramer's V statistic), calculates also a normalization in this case of the chi-square statistic for deviations between observed and expected haplotype frequencies [17]. The strongest associations are observed for the contiguous and/or physically close HLA loci pairs including DRB1~DRB5/4/3, DRB1~DQA1, DQA1~DQB1 and DRB1~DQB1 followed by B~C. HLA-DPA1~DPB1 pair appears associated with less strength. Interestingly, in spite of HLA-A~C pair being physically closer than HLA-A~B the strength of the LD between the latter is higher, suggesting that differences in diversity between HLA-B and -C loci may play a role in determining this measurement. Associations between HLA-A~B and HLA-B~DRB1 appear in similar ranges. HLA-DP loci show weaker LD associations than any of the other pairwise comparisons. As previously reported [25,26], LD patterns of HLA-DP loci seem to be driven primarily in a different manner compared to the other HLA loci (e.g. relatively higher rate of recombination and combined DPA1~DPB1 amino acid epitope have been suggested to contribute on this distinctive selection).

### 3.8. Estimation of extended HLA haplotype frequencies

Maximum likelihood estimation via an expectation-maximization algorithm is a statistical method commonly used for HLA haplotype inference and estimation of haplotype frequencies in unrelated individuals from a population-specific genotype data as in the present study. Moreover, this statistical method serves as an alternate approach when it is not possible to rely on family segregation studies [18]. Inferred extended HLA haplotypes (encapsulating 6-locus, 7-locus and 9-locus respectively) were evaluated for the estimation of haplotype frequencies in this Spanish population cohort: HLA~A∼C∼B~DRB3/4/5∼DRB1~DQA1~DQB1 (Supplementary Table 5); HLA~A∼C∼B~DRB3/4/5∼DRB1~DQA1~DQB1~DPA1~DPB1 (Supplementary Table 6); and HLA~A∼C∼B~DRB3/4/5∼DRB1~DQA1~DQB1 (Supplementary Table 7).

Similarly to what we found in 2-locus haplotypes, it can be observed very distinctive extended haplotype associations in non-coding regions at the 4-field level that are not apparent at lower allele resolution level (2-field or 3-field) results that are obtained when using legacy methodologies (e.g. SSP or SSO) with important limitations in sequence coverage and phasing in comparison to NGS-based typing [12,13].

### 3.9. Different HLA allele distributions found between Spanish regions

Finally, we examined the disparity/similarity of allelic distributions within this Spanish population cohort based on the results (at the 4-field allele resolution level) of the current study. In this sense, we carried out a comparison of HLA distributions (based on allele frequencies found at HLA-A, -B, -C, -DQB1 and -DRB1 loci) between the 3 different geographical Spanish regions established (Northern-Central, Eastern and Southern Spain) as well as between the 10 Spanish locations studied in the present work (Supplementary Table 8).

Despite of limitations in the sample size shown by these different Spanish population sub-groups in the present study. At the HLA allele level, it can be observed that most frequent alleles at a national level (considering entire Spanish population, termed as “ESP”, n = 282) are fairly evenly distributed and well represented among the different Spanish regions and locations evaluated here, with some minor exceptions (specifically found at the different 10 Spanish locations presenting a limited and small sample size comparatively) that need to be further analyze by future larger-scale population studies (see Supplementary Table 8.b)). Taking into account genetic distances evaluated here (see Supplementary Table 8.c and 8.d)), the present entire Spanish population cohort shows a Mediterranean genetic substrate that seems to be represented more predominantly by Eastern and Central regions/locations situated within the Central Plateau as previously described [4–11]. Whereas the most Northern and Southern regions/locations (which are mountainous areas that are more isolated geographically unlike this Central Castilian Plateau region in mainland Spain; or even being very unique island areas such as Canary Islands) diverge from this aforementioned Mediterranean HLA distribution as reported in previous works [4–11]. For instance, although we considered Barcelona location as part of the Eastern region of Spain for this study, we clearly observed how this Catalan location seems to be more related to other Northern locations than to Mediterranean sites such as Valencia or Murcia. Interestingly, Salamanca location population group (situated very close to the frontier that separates Spain from Portugal) describes a pronounced distinctive HLA distribution in comparison to other Northern-Central locations in Spain as previously described and it also exemplifies the extensive HLA diversity of the Iberian Peninsula [4,8]. Furthermore, the striking divergence observed in Malaga and Gran Canaria locations (see Supplementary Table 8.d) may be explained by the reported historic genetic contribution from North African Berber populations [1,5].

We also attempted to do this regional study at the extended HLA haplotype level (data not shown). However, due to these limited small sample sizes found at the different Spanish regions and locations it was not possible to estimate accurately haplotype frequencies via an expectation-maximization algorithm [18] to evaluate haplotype sharing between regions/locations.

Overall, in spite of presenting a relatively small sample size, the present Spanish population study has allowed us to see the great potential of NGS-based HLA population studies in order to identify 4-field HLA allele signatures at a regional level as a consequence of both differential regional historic events and the characteristic regional
orography that favors more isolation of certain local populations. Nonetheless, future studies of larger population sample size at a wider geographic scale will be needed to assess more accurately the HLA diversity in Spanish population in order to confirm these observations and findings of our study as well as to reveal other unknown but significant polymorphism within the HLA system.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.

4. Discussion

In the present study, we characterized HLA allelic sequences of 11 major HLA genes with extensive coverage and phased-alleles with minimum heterozygous ambiguity per locus at the 4-field for a representative Spanish population cohort (n = 282) by applying this novel high-throughput NGS-based HLA typing method [15]. We also examined allelic and haplotypic HLA frequency distributions at the 4-field allele resolution level in the Spanish population for the first time.

At the HLA allele level, regarding HLA class I loci we observed that HLA-B locus presents the highest allele diversity in comparison to HLA-A and -C loci in relation to the number of alleles (k) found in this population. Nevertheless, the 4-field allele resolution level has allowed us to see a significant diversity at the nucleotide level for HLA-A population. Nevertheless, the 4-field alleleresolution level has allowed HLA-B field alleleresolution level in the Spanish population for the first time.
DRB1*07:01:01:01 ~ DQA1*02:01:01:01 ~ DQB1*02:02:01:01; or HLA-A*01:01:01:01 ~ C*06:02:01:01 ~ B*57:01:01 ~ DRB4*01:03:01:02N ~ DRB1*07:01:01:01 ~ DQA1*02:01:01:01 ~ DQB1*03:02:02:01; tentative haplotypes that contain rare alleles such as: HLA-A*02:01:01:01 ~ C*12:166 ~ B*52:01:01:02 ~ DRB5*01:02:02 ~ DRB1*15:02:01:01 ~ DQA1*03:01:01:01 ~ DQB1*06:01:01; or HLA-A*02:02:01:01 ~ C*12:03:01:01 ~ B*15:220 ~ DRB4*01:03:01:01 ~ DRB1*07:01:01:01 ~ DQA1*02:01:01:01 ~ DQB1*02:02:01:01). Nonetheless, future studies of larger population sample size at a wider geographic scale will be needed to obtain a more accurate determination of rare alleles and respective allele-carrying haplotypes [43].

In addition to other genetic markers (e.g. Y-chromosome and mitochondrial DNA), assessment of HLA haplotypes diversity within the worldwide populations and its geographical variation also contributes in the analysis of tracking migrations of modern populations as well as in anthropological studies [44,45]. Interestingly, we observed certain unique haplotypes that reflect both significant historic Sephardic Jewish and Arab genetic contributions (e.g. common haplotype: HLA-A*24:02:01:01 ~ C*04:01:01:06 ~ B*35:02:02 ~ DRB3*02:02:01:02 ~ DRB1*11:04:01 ~ DQA1*05:05:01:01 ~ DQB1*03:01:01:02 ~ DPA1*01:03:04:01 ~ DPB1*04:01:01:01) [1,3,34,36] as well as a gene flow of relatively itinerant ethnic groups (e.g. common Spanish Gypsy haplotype: HLA-A*01:01:01:01 ~ C*15:02:01:02 ~ B*40:06:01:02 ~ DRB3*02:02:01:01 ~ DRB1*14:04:01 ~ DQA1*04:02:02 ~ DQB1*05:03:01:02 ~ DPA1*01:03:01:02 ~ DPB1*02:02:02:02) [46] in the present Spanish population cohort. In addition to this observed genetic substrate from various ethnic groups, it is well documented the relevant cultural heritage as well as the institutional recognition of all ethnic communities in Spain [47–49].

5. Conclusion

Results of the present Spanish population study show that NGS reveals distinctive allelic distributions and haplotypic associations. Moreover, HLA 4-field level data of worldwide populations may contribute to revise and to update CWD alleles list, knowledge of disease-associated alleles and/or haplotypes as well as current donor-recipient matching algorithms. Furthermore, our study shows that allelic and haplotypic HLA frequencies of Spanish population present a relative homogenous distribution that fits HWE proportions at all loci with the exception of HLA-DPA1 locus. At the same time, the distinctive HLA diversity found at both the allele and the haplotype levels is in concordance with the well-documented migrations presenting episodes of gene flow that have occurred through the course of history in Spain [1]. In this sense, application of NGS technology has allowed us to obtain a first glance of the HLA diversity at the 4-field level in the Spanish population. At the same time, future larger and wider geographic scale NGS studies will provide a more accurate description of this vast genetic diversity.

Overall, these results may contribute as a useful reference source for future population studies as well as a healthy control group dataset for evaluating HLA-disease associations. Furthermore, this HLA data may provide helpful information for improving donor recruitment strategies of bone marrow registers.

Conflict of interest

The authors have declared no conflicting interests.

Acknowledgments

We are grateful to the reviewers of this manuscript for their helpful suggestions. We also thank all participant members of the Spanish Working Group in Histocompatibility and Transplant Immunology (GETHTI) of the Spanish Society for Immunology (SEI) for their significant contribution and collaboration to this study. This study was supported by Stanford Blood Center (as the official organizer and sponsor of the 17th International Histocompatibility and Immunogenetics Workshop-2017 (17th IHWW) and by United States National Institutes of Health (NIH) grant U19NS095774.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhimmun.2019.02.005.

References


Human Immunology 80 (2019) 429–436
G. Montero-Martín, et al.


