Master Thesis Titled:

“Bioinformatic analysis of intra and inter population HLA diversity in Crete for the creation of a regional public cord blood bank”

MSc Candidate: Stylianakis Emmanouil

Student Code: 1120040

Heraklion, July 2020
Acknowledgements

I would like to thank the Principal Researcher Dr. George Potamias for accepting me in his lab and for giving me the opportunity to make my first steps on research. I would also like to give special thanks to the Postdoctoral Researcher Dr. Alexandros Kanterakis for giving me the opportunity to get involved in such an interesting topic. I also owe to Dr. Alexandros Kanterakis and to the Postdoctoral Researcher Dr. Helen Latsoudis many thanks for their excellent collaboration and for their patience and willingness to guide me in my first years in research. I would also like to really thank the Professor of Haematology, Helen Papadaki as well as the Professor Emeritus of Basic and Clinical Immunology, Anastasios Germenis for trusting me and letting me join their research team and for broadening my scientific way of thinking. It would be an omission not to thank the Research Associate Dr. Pavlos Pavlidis, Dr. Julia Pingel and Dr. Jurgen Sauter for their advice throughout the study.
Abstract

In hematopoietic stem cell transplantation (HSCT), histocompatibility between a potential marrow donor and the recipient depends on the highly polymorphic human leukocyte antigen (HLA) system. Five of the most crucial HLA genes for histocompatibility that reside in this region are A, B, C, DRB1 and DQB1. Both the genotypes and the formed haplotypes of these genes have been a central topic of research in this area. Allogeneic HSCT is a curative therapy for hematologic, genetic diseases and immunodeficiencies. The source of hematopoietic stem cells (HSC) may be matched sibling donors (MSD), matched unrelated donors (MUD) and umbilical cord blood (UCB) units from public UCB banks. The use of haploidentical HSCT has extended the availability of allo-HSCT and has limited the use of UCB. However, the existence of public UCB banks increases the chance of finding an ideal MUD in countries with a relatively small population and large genetic diversity. This study was initiated by the regional public UCB bank of Crete, Greece operating in the University Hospital of Heraklion. The main aim of this study is to assess the need of a regional public UCB bank in Crete by creating an HLA profile of the Cretan population, by comparing it afterwards with one of the most representative European unrelated donor registries (UDR), i.e. the German DKMS and finally, by comparing the HLA genetic diversity between the 4 Prefectures of Crete. These assessments can provide valuable insights regarding the potential need for the enrichment of the national/international UDR with uncommon/rare HLA haplotypes through UCB donor recruitment. The comparative analyses of the estimated HLA allelic and haplotypic frequencies between the Cretan population and different cohorts from DKMS resulted in significant differences in the distribution of specific alleles and haplotypes, as well as in a low number of common haplotypes and a relatively large number of unique Cretan alleles with 9 possible novel alleles. Additionally, the 4 Prefectures of Crete do not seem to have similar immunogenetic profiles, probably due to significant geographical or cultural isolation. These findings suggest that the Cretan population is underrepresented in the common DKMS UDR whilst exhibiting high genetic diversity. Therefore, the regional public UCB bank could play an important role for the enrichment of the national and international UDR improving the possibility for a patient to find a MUD when needed.
Περίληψη

Στη μεταμόσχευση αιμοποιητικών βλαστικών κυττάρων (HSCT), η ιστοσυμβατότητα μεταξύ ενός πιθανού δότη μυελού και του λήπτη εξαρτάται από το εξαιρετικά πολυμορφικό σύστημα HLA. Πέντε από τα πιο κρίσιμα γονίδια HLA για την ιστοσυμβατότητα που βρίσκονται σε αυτήν την περιοχή είναι τα A, B, C, DRB1 και DQB1. Τόσο οι γονότυποι όσο και οι απλότυποι που σχηματίζονται από αυτά τα γονίδια υπήρξαν κεντρικό θέμα έρευνας σε αυτόν τον τομέα. Η αλλογενή HSCT (allo-HSCT) είναι μια θεραπευτική επιλογή για αιματολογικές και γενετικές ασθένειες, καθώς και για ανοσοανεπάρκειες. Η πηγή των αιμοποιητικών βλαστικών κυττάρων (HSC) μπορεί να είναι από συμβατούς αμφιθαλείς αδελφούς δότες (MSD), από συμβατούς μη σχετιζόμενους συγγενικά δότες (MUD) ή από μονάδες αίματος ομφαλίου λόφου (UCB) από αντίστοιχες δημόσιες τράπεζες. Η αναζήτηση αντισυμβατικών σε μεταμοσχεύσεις αιμοποιητικών βλαστικών κυττάρων έχει αυξήσει τη διαθεσιμότητα της αλλογενούς HSCT και έχει περιορίσει τη χρήση των μονάδων αίματος ομφαλίου λώρου. Ωστόσο, η ύπαρξη δημόσιων τραπεζών UCB αυξάνει την πιθανότητα εύρεσης ενός συμβατού, μη σχετιζόμενο συγγενικά, δότη σε χώρες με σχετικά μικρό πληθυσμό και μεγάλη γενετική ποικιλομορφία. Η μελέτη αυτή ξεκίνησε από την Περιφερειακή Δημόσια Τράπεζα UCB της Κρήτης, που λειτουργεί στο Πανεπιστημιακό Γενικό Νοσοκομείο Ηρακλείου (ΠΑΓΝΗ). Ο κύριος σκοπός αυτής της μελέτης ήταν να εκτιμήσει την ανάγκη μιας Περιφερειακής Δημόσιας Τράπεζας UCB στην Κρήτη, δημιουργώντας ένα προφίλ HLA του Κρητικού πληθυσμού συγκρίνοντας το, στη συνέχεια, με ένα από τα πιο αντιπροσωπευτικά Ευρωπαϊκά Μητρώα μη σχετιζόμενων δοτών (UDR), όπως το αντίστοιχο μητρώο της Γερμανίας (DKMS) και τέλος, συγκρίνοντας τη γενετική ποικιλομορφία HLA μεταξύ των Νομών της Κρήτης. Αυτές οι αξιολογήσεις μπορούν να παρέχουν πολύτιμες πληροφορίες σχετικές με την πιθανή ανάγκη εμπλουτισμού του εθνικο/διεθνούς UDR μέσω της πρόσληψης δοτών UCB. Οι συγκριτικές αναλύσεις των εκτιμώμενων συχνοτήτων των HLA αλληλομόρφων και απλοτύπων μεταξύ του Κρητικού πληθυσμού και των διαφορετικών εθνικοτήτων από το DKMS είχαν ως αποτέλεσμα την εύρεση σημαντικών διαφορών στην κατανομή συγκεκριμένων αλληλομόρφων και απλοτύπων. Τανόχθισε επίσης ενδιαφέροντα ιδιαίτερα αριθμούς κουνών απλοτύπων και σχετικά μεγάλος αριθμός μοναδικών Κρητικών αλληλομόρφων με 9 πιθανά καινούργια/αχαρακτήριστα αλληλόμορφα. Επιπλέον, οι τέσσερες νομοί της Κρήτης δεν είχαν πολλά κοινά αλληλομόρφα, πιθανότατα λόγω σημαντικής γεωγραφικής ή πολιτιστικής απομόνωσης. Τα ευρήματα αυτά υποδηλώνουν ότι ο πληθυσμός της Κρήτης δεν εκπροσωπείται απαραίτητα στο DKMS, ενώ παρουσιάζει υψηλή γενετική ποικιλομορφία. Επομένως, η περιφερειακή δημόσια τράπεζα UCB θα μπορούσε να διαδραματίσει σημαντικό ρόλο για τον εμπλουτισμό της εθνικής και διεθνούς UDR βελτιώνοντας τη δυνατότητα για έναν ασθενή να βρει MUD όταν χρειαστεί.
# Table of Contents

**Acknowledgements** | 2  
**Abstract** | 3  
**Περίληψη** | 4  
**Introduction** | 7  
- Umbilical Cord Blood (UCB) banks | 7  
- Genetic region of the Human Leukocyte Antigen (HLA) | 7  
- Nomenclature of HLA Alleles | 9  
- HLA matching in transplantation - Finding a MUD - Importance of well organized biobanks | 10  
- Genetic/Disease background of Crete | 11  
- State of the art and innovation | 11  
- The basic bioinformatics tools utilized | 12  
- Other bioinformatics resources and tools for immunogenetic studies | 12  
**Methods** | 14  
- HLA typing in the Cretan population | 14  
- European sample | 14  
- Data Source and Processing | 14  
- Cretan sample for analysis and novel HLA alleles | 15  
- ‘g’ grouping | 15  
- Computing haplotype frequencies with Hapl-o-Mat v 1.1 | 15  
- Haplotype and allele frequency estimation and comparison | 16  
- Principal Component Analysis (PCA) | 16  
- Hardy-Weinberg Equilibrium (HWE) | 17  
- Linkage Disequilibrium (LD) | 17  
- Performance assessment and comparison of tools on artificial data | 18  
- Arlequin v 3.5.2.2 VS Genepop v 4.7 for Linkage Disequilibrium (LD) | 18  
**Results** | 20  
- Contrasting the Cretan HLA profile with the DKMS cohorts | 20  
- Statistic comparison of the frequencies of the common alleles | 20  
- Unique Cretan HLA alleles | 23  
- PCA on Cretans and DKMS cohorts | 23  
- Haplotype comparison between DKMS registries and the Cretan population | 24  
- **HLA profile with 6 genes of 1146 Cretans** | 26  
- Allele Frequencies | 26  
- Haplotype Frequencies | 28  
- Hardy-Weinberg Equilibrium (HWE) | 28  
- Linkage Disequilibrium (LD) | 30  
- **HLA profile of the 4 Cretan Prefectures** | 30  
- Allele Frequencies | 30  
- Haplotype frequencies | 35  
- Hardy-Weinberg Equilibrium (HWE) | 36  
- Linkage Disequilibrium (LD) | 37  
**Discussion** | 38  
**Availability** | 40  
**Achievements** | 40  
**References** | 42
Introduction

Umbilical Cord Blood (UCB) banks

Allogeneic haematopoietic stem cell transplantation (allo-HSCT) is a curative therapy for hematologic, genetic diseases and immunodeficiencies (Little and Storb 2002). In HSCT, histocompatibility between recipient and marrow donor depends on the human leukocyte antigen (HLA) system, a multigenic region on chromosome 6 with high allelic diversity and significant linkage disequilibrium (LD) (Figure 1). The haematopoietic stem cell (HSC) source may be HLA matched sibling donors (MSD), matched unrelated donors (MUD) and umbilical cord blood (UCB) units from public UCB banks. The recruitment and registration of volunteer donors in national registries, as well as establishment of public UCB banks, is particularly important for the probability of finding an HLA-compatible transplant for any patient in need. It is well known that HLA allele and haplotype frequencies differ significantly between populations (Carapito, Radosavljevic, and Bahram 2016); (Gandhi et al. 2017); (Lee et al. 2007); (Gragert et al. 2014; Tiercy 2016). Thus, the effectiveness of registries in providing MUD for patients in need depends on several variables, including the size of the registry or UCB bank and the genetic diversity of the target population.

Genetic region of the Human Leukocyte Antigen (HLA)

The HLA system, also known as the human major histocompatibility complex (MHC), spans a 4Mb genetic region (~0.13% of the human genome) on the short arm of chromosome 6 (6p21) encoding the classical transplantation HLA genes and at least 132 protein coding genes (Figure 1). The HLA complex is divided into three genetic regions containing genes of different classes (I, II, and III). The class I and II regions include genes encoding the HLA class I (HLA-A, -B, -C) and II (HLA-DR, -DQ, -DP) histocompatibility antigens, while the class III region contains genes coding for molecules involved in immune function that are not targets for allorecognition.

HLA is the most polymorphic region in the human genome. As of today the total number of the HLA alleles is estimated to be 26,889 with HLA class I alleles being 19,587 and HLA class II alleles being 7,302 as referenced in the Nomenclature of the IPD-HLA Database (Robinson 2001). It has been proved that they play a significant role in some fundamental molecular and cellular processes with most crucial the regulation of the immune system. This small genetic region has been associated with more than 100 different common diseases and autoimmune disorders in humans, such as diabetes, asthma, psoriasis and rheumatoid arthritis (Shiina et al. 2009).

The MHC genomic region is one of the most gene-dense and best-defined regions within the human genome, and the undefined sequences contribute to only a low percentage of the MHC region (Shiina et al. 2009). It is composed mainly of genes, retrotanposons, transposons, regulatory elements, pseudogenes and a few remaining undefined sequences. Many of the MHC gene products are receptors, ligands, signaling factors, interacting proteins and transcription regulators involved in the interactions with NK cells and cytokines as part of the innate immune responses and in the inflammatory response, antigen processing and presentation as part of the adaptive immune response.
Figure 1. Gene map of the human leukocyte antigen (HLA) region. SOURCE: (Shiina et al. 2009)
Nomenclature of HLA Alleles

Each HLA allele name has a unique number corresponding to up to four sets of digits (also referred as 8-digit coding) separated by colons (see Figure 2). The type of the sequencing of the allele defines the length of the allele designation. The lower error rate the sequencing has, the better characterization of the HLA allele will be (HLA typing), as its name will contain more digits. The allele names of the HLA genes may have 2, 4, 6 or 8 digits.

The digits before the first colon describe the allele group (type), which often corresponds to the serological antigen carried by an allotype. The second set of digits are used to list the specific HLA protein (subtypes). Alleles whose numbers differ in these two sets of digits must differ in one or more nucleotide substitutions that change the amino acid sequence of the encoded protein. Alleles that differ only by synonymous nucleotide substitutions (also called silent or non-coding substitutions) within the coding sequence are distinguished by the use of the third set of digits. Alleles that only differ by sequence polymorphisms in the introns, or in the 5' or 3' untranslated regions that flank the exons and introns, are distinguished by the use of the fourth set of digits (an example is shown on Figure 2).

Apart from the unique allele number, there are some more optional suffixes that may be added to an allele to indicate its expression status. Alleles that have been shown not to be expressed - 'Null' alleles - have been given the suffix 'N'. Alleles that have been shown to be alternatively expressed may have the suffix 'L' ('Low' cell surface expression when compared to normal levels), 'S' (the protein is expressed as a soluble, 'Secreted' molecule but is not present on the cell surface), 'C' (the protein is present in the 'Cytoplasm' and not on the cell surface), 'A' ('Aberrant' expression where there is some doubt as to whether a protein is actually expressed) or 'Q' (the expression of an allele is 'Questionable'). More information about the Nomenclature of the HLA alleles can be found in the IPD-IMGT/HLA Database (Robinson 2001) in the section HLA Nomenclature (How an HLA allele is named).

![Figure 2. Example of an HLA Nomenclature from IPD-IMGT/HLA Database. SOURCE: http://hla.alleles.org/nomenclature/naming.html](http://hla.alleles.org/nomenclature/naming.html)
HLA matching in transplantation - Finding a MUD - Importance of well organized biobanks

The optimal high-resolution DNA matching for HLA class I loci A, B, C, and class II loci DRB1 and DQB1 in MUD is a 10/10 match in the recipient’s and donor’s haplotypes, whereas a 8/8 match is considered the minimum level of matching associated with the highest survival after the transplantation (Lee et al. 2007). Even though over 37 million donors and almost 800,000 UCB units are now registered in the international database of WMDA, many patients will still not have a fully matched donor because of the extremely great diversity of HLA alleles and haplotypes. The probability of identifying a MSD depends only on the number of siblings (i.e. 25% for patients with one sibling, 44% for those with two, 58% for those with three, 68% for those with four) (Gragert et al. 2014; Tiercy 2016), whereas the probability of identifying a highly MUD depends on the frequency of the patient’s HLA haplotypes. When no MSD is available, an estimate of the probability of finding a fully MUD, based on the frequency of the patient’s haplotypes, will help the transplant center in taking a decision on whether to search for an unrelated donor or look for an alternative source of HSC (haploidentical donor or UCB unit from a public bank). The gold standard for MUD is to look for a 10/10 or 8/8 matched HSC or UCB transplants, respectively (NetCord-FACT International Cord Blood Standard, 7th Edition).

The significant HLA allelic diversity combined with the strong LD, which drives a non-random combination of the respective alleles that is highly influenced by the population characteristics of the patient’s country of origin (i.e. founder’s effects, population admixture, etc.), still remains a significant challenge influencing turn-around time requirements of the HSCT field. In practice, the mixing and ethnic diversity of a population increases the frequency of novel haplotypes and thus complicates the search for a compatible donor. Beyond doubt, the introduction of NGS to the standard HLA typing methodology of population samples from different geographic locations, unraveled the complexity of this genetic region (i.e. influenced by allelic heterogeneity, balancing selection, genetic drift) (Alcaide 2010; Sanchez-Mazas, Lemaître, and Currat 2012), revealed the importance of demographic history in the evolution of HLA variants and also highlighted the importance of diversity over quantity in UCB banking. A wealth of data (Lee et al. 2007; Gragert et al. 2014; Tiercy 2016; Alcaide 2010; Sanchez-Mazas, Lemaître, and Currat 2012) confirms that the search for potential matched donors or the associations of HLA alleles and haplotypes with specific phenotypes may be insufficient or even misleading if one does not take into effect the genetic makeup of the representative population. For instance, the likelihood of finding a 8/8 MUD in the National Marrow Donor program “Be The Match” (https://bethematch.org) for patients of White European descent is calculated at 75% for 2017 and drops to 18% and 34%-37% for populations where diversity is inherently higher (e.g. Africans) and admixture dominates (e.g. Latin Americans), respectively (Bravo-Acevedo et al. 2019; Arrieta-Bolaños, Oliveira, and Barquera 2020).

Building representative population reference datasets will significantly improve variant interpretation globally while at the same time will accelerate the discovery of HLA alleles associated with rare phenotypes, particularly in populations characterized by unique properties as a result of a geographical or a cultural isolation. Undoubtedly, some existing registries and UCB banks are neither efficient nor able to provide suitable donors. Under this prism, mapping the regional HLA haplotype spectrum may improve donor availability, particularly in countries of relatively small population size and significant genetic diversity (i.e.}
bottlenecks/founding effects, population admixture) ((Lee et al. 2007) ; (Gragert et al. 2014; Tiercy 2016) ; (Alcaide 2010) ; (Sanchez-Mazas, Lemaître, and Currat 2012) ; (Bravo-Acevedo et al. 2019) ; (Arrieta-Bolaños, Oliveira, and Barquera 2020)).

Genetic/Disease background of Crete

The island of Crete has been characterized as an elegant example of the opportunities that translational biology offers through the discovery of rare alleles of specific neurological (Dedoussis et al. 2005) (Tzagournissakis et al. 1995) and cardiovascular or metabolic traits (Panoutsopoulou et al. 2014) (Tachmazidou et al. 2013). For instance, an unusual form of Huntington’s disease has been detected (Tzagournissakis et al. 1995) in the prefecture of Heraklion which has distinctly different characteristics compared to other cases worldwide (i.e. a 15-20yrs later median age at onset and a stably transmitted (CAG)n repeat expansion). The mountainous geography and the wealth of historical events are thought to influence the genetic history and cultural characteristics of the Cretan population. For instance, occupation of Crete by Arabs, Venetians and Ottomans has resulted in multiple revolutions against conquerors in specific regions (e.g. Mylopotamos, Sfakia) and the subsequent clustering of the respective inhabitants, represented in Principal Component Analysis (PCA) as distinct subpopulations with an east-to-west gradient in gene frequencies (Drineas et al. 2019). In support of the hypothesis that unique properties may underlie the genetic makeup of the Cretan population is the recent establishment of the inhabitants from the mountainous Mylopotamos villages (HELIC-MANOLIS) as genetic isolates (Panoutsopoulou et al. 2014). Moreover, the distinctive characteristics of the aforementioned cohort (i.e. common founder, population admixture effects) resulted in the detection of a rare cardioprotective variant in APOC3 (Tachmazidou et al. 2013).

State of the art and innovation

The knowledge about HLA allele and haplotype frequencies, as well as presence of rare alleles within a population, is very useful for an effective HSCT programme particularly for countries with anticipated HLA diversities because of geographic specific position and various influences throughout history. Therefore, databases containing population-and donor-specific HLA allele and haplotype frequencies are of particular importance for the development of strategies for MUD search and recruitment. The combination of NGS for increased sequence lengths and advanced bioinformatics tools has emerged as the gold standard for high resolution HLA typing and haplotype estimation. In this project we aimed to map the immunogenetic HLA profile of the general population of Cretan origin residing in the island of Crete both in total and assigned to one of the 4 defined Cretan Prefectures; to compare the HLA genetic diversity of the Cretan population with one of the most representative European unrelated donor registries (UDR) i.e. the German DKMS; to substantiate the need of the regional public UCB bank of Crete for UCB donor recruitment for enrichment of the national/international UDR with uncommon/rare HLA haplotypes. The bioinformatics tools that were used throughout the analysis were Hapl-o-Mat v 1.1, Arlequin v 3.5.2.2, and the web version of Genepop 4.7. All the analysis was implemented on a Linux environment with the above tools and with Python 3.6.9. The significance of the present study is featured by the fact that such an approach has never been performed at the regional or national level in Greece.
The basic bioinformatics tools utilized

Throughout the study we came across to a variety of tools that we could use for our analysis, but we chose Hapl-o-Mat v 1.1 and Arlequin v 3.5.2.2 (see Arlequin v 3.5.2.2 VS Genepop v 4.7 for Linkage Disequilibrium (LD) section) since they met better our expectations and they were commonly used in other similar analysis in the bibliography. For instance, Arlequin as of today has 11,770 citations on Google Scholar for the version 3.5 and for an older version (3.0) has 16,189 citations.

**Hapl-o-Mat v 1.1.** This tool calculates the HLA haplotype frequencies using an expectation-maximization (EM) algorithm from population data including an arbitrary number of loci. It can also process different HLA typing resolutions within a given population sample and handle the ambiguities recorded via multiple allele codes (MAC) or genotype list strings (GLS). Another benefit of this computational program is that it gets updated manually by incorporating the latest IPD-IMGT/HLA ambiguous allele combinations files, which enables the efficient handling of the latest allele definitions and/or ambiguities (Sauter, Schäfer, and Schmidt 2018). Last commit from Hapl-o-Mat was in 2018 which indicates that it is no longer maintained (abandonware). Installing Hapl-o-Mat v 1.1 was a challenge since it did not support later versions of the Gnu C Compiler. After debugging this issue and posting this solution in Github, we had 2 users acknowledging our effort. This means that a very prominent bioinformatics tool in HLA studies would not have been useful to at least two people had we not posted our correction ([https://github.com/DKMS/Hapl-o-Mat/issues/5](https://github.com/DKMS/Hapl-o-Mat/issues/5)).

**Arlequin v 3.5.2.2.** Arlequin v 3.5.2.2 is an integrated software for population genetics data analysis and uses a Markov Chain (MC) sampling scheme for its calculations (such as Linkage Disequilibrium estimation). A MC is a stochastic model describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event (Excoffier and Lischer 2010).

**Genepop v 4.7.** Genepop v 4.7 is a population genetics software package (Raymond and Rousset 1995) that can also run on the web with some limitations on the number of populations tested and the number of the genetic loci. These limitations did not affect our analysis.

Other bioinformatics resources and tools for immunogenetic studies

**HLA.net** is a European network of the HLA diversity for histocompatibility, clinical transplantation, epidemiology and population genetics (Nunes 2015).

**Basic statistics** is an online tool on HLA.net that can analyze the allele and haplotype frequencies, Hardy-Weinberg Equilibrium (HWE) and Linkage Disequilibrium (LD) and requires as input a valid UNIFORMAT (version 3) file with the HLA loci data with .unif extension. The input file is uploaded and then, the results are either sent on the user’s email or saved on a browser. The result files are deleted after 48 hours. When we used this online tool in March 2020, we could not obtain the results and it seemed that it was no longer maintained since we could not ask any question on their site, probably due to the COVID-19 outbreak.
Nevertheless, a recent activity on the Overview section of the online tools, where users can post questions, indicates that the service is not totally abandoned.

Similar analyses can also be performed by PyPop, an environment for Population Genomics in the programming language Python (Lancaster et al. 2007). PyPop runs in two ways: (i) batch mode, where someone can supply all the command line options that the program needs and (ii) interactive mode, where the program will prompt the user to directly type the input it needs. Interactive is the suggested mode when a single population is analyzed.

Additionally, there are some packages in the Comprehensive R Archive Network (CRAN) that can perform these analyses, such as haplo.stats for haplotype frequency estimation and HardyWeinberg (Tierney 2012).

An interesting study has been conducted regarding the tool comparison for common analyses of immunogenetic population data (Mack et al. 2012). Many tools were tested on different analyses and each tool had a score about its performance on each analysis. In this study we were interested in the haplotype frequency estimation, the HWE and the LD. PyPop had the highest score in all of the 3 analyses. Regarding haplotype estimation and LD, CLUTO (Karypis 2002) and PyPop had the same efficiency although CLUTO cannot estimate the HWE. The second best tool for these analyses was Arlequin, followed by GenePop, which had the same scoring with Arlequin on HWE and on LD, but it cannot perform haplotype estimation. Tied with Arlequin on the haplotype estimation was the R package haplo.stats. EstiHaplo was tied with PyPop on the estimation of the haplotypes but it cannot perform any other analysis. It seems that this tool is abandoned since its Home Page cannot be found, as well as any reference regarding its creation. It would be an omission not to state that PLINK (Purcell et al. 2007), which is a popular toolset for whole-genome association and population-based linkage analyses (as of today it has 20,205 citations), is not recommended for immunogenetic analysis although it performs greatly with other kinds of data. Although it seems that PyPop would have been a better choice than Arlequin, we chose the latter because it was widely used in similar studies, its latest version was released 3 years later than PyPop’s and has way more citations. Hapl-o-Mat was not included in the Mack et al. 2012 comparison, because it was released 5 years later. As a future work, it would be really interesting to see how well it performs on haplotype estimation compared to all other tools.
Methods

HLA typing in the Cretan population

Peripheral blood samples were obtained from 1,204 unrelated individuals of Cretan origin living in various regions of Crete with self-assessed parentage and grand parentage at recruitment. Only donors with at least 3 out of the 4 grandparents being born in Crete were included in the study. Genomic DNA was extracted (iPrep Purification Instrument, Thermo Fisher Scientific) and HLA genotyping was performed by Next Generation Sequencing (NGS) using the AllType NGS kit (One Lambda). Template preparation and sequencing were performed in the Ion Chef Instrument and the S5 Sequencing System (Thermo Fisher Scientific), respectively. Allele assignment was performed by the Type Stream Visual software (One Lambda) using the IMGT ver. 3.35.00. A 4-digit genotypic HLA profile was generated for 6 HLA genes: A, B, C, DRB1, DQB1 and DPB1.

European sample

For the inter population comparison of the HLA profiles, we chose one of the most representative European unrelated donor registries (UDR) i.e. the German DKMS (Pingel et al. 2013). The 4-digit profiles of the HLA loci -A, -B, -C, -DRB1 for 13 available populations of DKMS were obtained from the Allele Frequency Net Database, where allelic and haplotypic frequencies of many populations worldwide based on the genomic region of the HLA are stored (Gonzalez-Galarza et al. 2020). Additionally, because we were also interested in comparing the Cretan HLA profile with the DKMS populations as a whole, we created a pool of all the available DKMS cohorts together and we adjusted the frequencies in their alleles and haplotypes according to the sample size of each cohort contributing to the pool. This pool consists of Austrians, Bosnians, Chinese, Croatians, French, Greeks, Italians, Netherlands, Portuguese, Romanians, Spaniards, Turkish and UK citizens and represents a sample of possible donors in Europe. The sample size of the pool was 21,314 individuals.

Data Source and Processing

The Hematological Department of the University General Hospital of Heraklion (PAGNI), has undertaken the task of the analysis of histocompatibility antigens of the Cretan population under the supervision of Prof. Helen Papadaki, Head of the Public Umbilical Cord Blood Bank of Crete. This initiative was funded from the 7th Heath Region of Crete (ΥΠΕ). These data are not publicly available due to active legislation on private data protection. Data was registered in a Microsoft Excel (.xlsx) file. Each line represented an individual whose information was stated in the columns. This information were the date and place of birth, the origins of each individual’s parents and grandparents, the medical history on a few diseases and the HLA genotypes. This file was parsed and processed with ‘pandas’ library (Bernard 2016) of Python 3.6.9 in the form of DataFrame (see Availability section):

```python
import pandas as pd
pd.read_excel(excel_file)
```
Cretan sample for analysis and novel HLA alleles

The initial samples were 1,204 unrelated Cretans. First-degree relatives (e.g. mother and child) were not included in the analyses in order to avoid type I errors (i.e. false positives) resulting from the rejection of a true Null hypothesis due to over-estimated frequencies of rare alleles and haplotypes, which would render the sample not representative of the real frequencies in the general population. Exclusion of 37 samples with typing ambiguities (i.e. not fully genotyped on the HLA-A, B, C and DRB1) and 9 samples with high resolution typing, but, yet, unassigned World Health Organization (WHO) name for new alleles that also required Sanger verification, finally resulted in 1,158 individuals available for inter-population analyses. Following Sanger verification the 9 novel HLA alleles will be submitted on the IPD-IMGT/HLA Database (Robinson 2001) if some certain criteria of the sequencing are met. For the mapping of the immunogenetic HLA profile of the potential donors of Cretan origin residing in the island of Crete, we excluded 12 more individuals who were not fully genotyped on the HLA-DQB1 and HLA-DPB1 genes. These 1,146 Cretans were assigned to the corresponding Prefecture according to their origins (at least 3 out of the 4 grandparents should have been from this Prefecture) and formed the sample that was used for the intra-population comparison of the HLA profiles in Crete as follows.

‘g’ grouping

The 4-digit HLA profile of the DKMS cohorts obtained from the Allele Frequency Net Database had a special categorization of the alleles which is called ‘g’ grouping and is used for the report of ambiguous HLA allele typings. This is not a grouping as defined by the WHO and it differs from G and P groups (uppercase on both), which are the main groupings for the HLA alleles. HLA alleles that have identical nucleotide sequences across the exons encoding the peptide binding domains (exon 2 and 3 for HLA class I and exon 2 only for HLA class II alleles) belong to the same ‘G’ group and HLA alleles that have nucleotide sequences that encode the same protein sequence for the peptide binding domains (exon 2 and 3 for HLA class I and exon 2 only for HLA class II alleles) belong to the same ‘P’ group. The main difference to ‘P’ groups is that ‘g’ groups also include null-alleles while ‘P’ Groups only include expressed alleles and the main difference to ‘G’ groups is that ‘g’ groups allow for synonymous mutations (resulting in the same protein) while ‘G’ groups require DNA sequence identity in relevant exons for the antigen recognition domain. For this reason, in order to have comparable HLA alleles and haplotypes, the Cretan HLA profile was categorized to ‘g’ grouping with the help of the Hapl-o-Mat software running on a Linux environment. This tool is provided by DKMS (Sauter, Schäfer, and Schmidt 2018).

Computing haplotype frequencies with Hapl-o-Mat v 1.1

We estimated the frequency of the Cretan haplotypes with 4 digits consisting of the HLA-A, -B, -C and -DRB1 with the Hapl-o-Mat software while grouping the alleles into ‘g’ groups so that there can be a comparison with the DKMS cohorts. As we mentioned above, the calculation of the haplotype frequencies for the comparison with the sample residing in Europe required ‘g’ grouping (desired typing resolution parameter on Hapl-o-Mat: ‘g’) while the estimation of the haplotype frequencies for the creation of the Cretan HLA profile did not
require it, so we did not perform this categorization (desired typing resolution parameter on Hapl-o-Mat: ‘4d’, which leaves the haplotypes with just the 4 digits) and we included the HLA-DQB1 and HLA-DPB1 genes on the haplotypes. Our input form for the Cretan population was in the form of MAC and the parameters from the EM algorithm were Epsilon: 1e-06 and Seed: 1000. Here, Epsilon was the value for the stopping criterion, so the algorithm continues as long as the maximal change between consecutive haplotype frequency estimations is larger than the assigned value and Seed sets the seed of the used pseudo random number generator. For the output haplotypes we set the cutoff frequency to 1e-06 and we set the RENORMALIZE_HAPLOTYPEFREQUENCIES parameter to ‘True’ so that the estimated haplotype frequencies would be normalized to sum up to 1. Finally, we set the parameters DO_AMBIGUITYFILTER and EXPAND_LINES_AMBIGUITYFILTER to ‘False’ since our genotype data had no ambiguities and all the rest parameters were set to default settings. Had we used another input format, other than MAC, we would have access to some more parameters related to this format. Since we used the MAC input format, we had access to the parameters at the ‘parametersMAC’ file and in order for Hapl-o-Mat to be executed, this file along with the executable file ‘haplomat’ must be in the same directory. An example of how this tool runs on the command line is “./haplomat MAC > Log.txt”. On the Log.txt file the details of the current execution are saved.

Haplotype and allele frequency estimation and comparison

We first estimated the frequencies of the Cretan haplotypes as discussed above. We then compared the frequencies of the common haplotypes between Cretan and the DKMS cohorts one by one and as a whole with Fisher’s exact test having significance level \( a = 0.05 \). We should note here that the haplotypes of the Italians and of Turks had an additional HLA gene on their haplotype (HLA-DQB1) compared to the other 11 DKMS populations. Because the comparison with the DKMS pool should have been with 4 genes in the haplotypes (A, B, C and DRB1) we adjusted the frequencies of the 5-locus haplotypes to 4-loci by dropping the 5th locus and then adding the frequencies of the identical 4-loci haplotypes left. For the inter population allele frequency comparison between Cretan and the DKMS registries we calculated the Cretan ‘g’ grouped allele frequencies from the frequencies of the Cretan haplotypes and compared the common alleles which were not rare in each population (frequency greater than 1/2n, where n is the sample size of the given population) in their frequencies with contingency chi-square test having significance level \( a = 0.05 \). For the creation of the HLA profile with the 6 HLA genes where ‘g’ grouping was not needed, we estimated the allele and genotypic frequencies directly from the genotypes. All the above steps were implemented in Python 3.6.9.

Principal Component Analysis (PCA)

In order to contrast the genetic distances between the HLA profiles of the DKMS cohorts (excluding the Chinese) and the corresponding Cretan profiles as a whole and by Prefecture, we proceeded on multidimensional scaling for all the common and non-rare alleles of HLA-A, HLA-B, HLA-C and HLA-DRB1 genes (frequency > 1/2n). So, the input was a DxN dimensional matrix, where D was the number of the populations examined in each case and N was the number of their common and non-rare alleles. For this we employed a Principal
Component Analysis (PCA) approach. PCA is a multivariate statistical technique using sophisticated underlying mathematical principles to transform a number of possibly correlated variables into a smaller number of variables called principal components. PCA is used to remove the least beneficial features (dimensionality reduction) so that we can have a smaller data set, but without losing too much predictive power (Jolliffe and Cadima 2016). The PCA analysis was performed in Python 3.6.9.

Hardy-Weinberg Equilibrium (HWE)

While creating the HLA profile of the Cretan population, it was important to test whether the HLA alleles on all 6 loci were in Hardy-Weinberg Equilibrium (HWE) as a measure of genetic diversity. This equilibrium indicates that the allele and genotype frequencies in a population remain constant from generation to generation in the absence of other evolutionary influences, like genetic drift and natural selection (Wigginton, Cutler, and Abecasis 2005). We calculated HWE in the whole Cretan population and in its subpopulations at the level of Prefectures using Arlequin v 3.5.2.2 software on a Linux environment. Holms-Bonferroni correction was used to account for multiple testing considering the number of typed HLA loci (6 loci). Significant differences (p-value < 0.05/6) between Observed Heterozygosity (OH) and Expected Heterozygosity (EH) indicated a deviation from HWE.

Linkage Disequilibrium (LD)

Another measure of HLA genetic diversity in our population that we also considered was the estimation of Linkage Disequilibrium (LD) between the alleles of the 6 HLA studied loci. LD is the pairwise non-random association of alleles at different loci in a given population and is influenced by evolutionary factors like mutation rate and population structure. The level of LD between two genetic loci A and B can be quantified by the coefficient of linkage disequilibrium $D_{AB}$. The effect size statistic $D'$ ($D/D_{max}$ - normalized) is used to express the overall degree of deviation in the studied population with values falling between 0 (equilibrium) and 1 (linkage) (Reich et al. 2001). The calculations, including the statistical significance for the whole Cretan population as well its sub-groups (according to the 4 Prefectures), were estimated with Arlequin v.3.5.2.2 software on a Linux environment (see below). Holms-Bonferroni correction was used to account for multiple testing considering the number of all the possible pairs of the 6 HLA loci (C(6,2)=15). Statistically significant deviations of this measure (p-value < 0.05/15) may indicate the existence of events in the past (e.g. founder effect, genetic drift) that affected the distribution of EH and subsequently the genetic makeup of the respective population.

The settings for Arlequin v 3.5.2.2. This tool runs differently when the operating system is Windows or iOS or Linux. In our case, where we had Linux, Arlequin needed 2 files for execution. The first file (.ars) has the parameters of all the analyses Arlequin can implement and the user must change the task number parameter in order to set the wanted analysis and change its parameters if necessary. The second file (.arp) is the input file with the population data. One can modify the Profile section of the .arp file according to the type of the population data. The input files .ars were downloaded from Arlequin’s Download Page (Example files_win.zip) and the .arp files were created in Python 3.6.9. We used the default parameters.
of the downloaded .ars files for the HWE and LD analyses for HLA genotypic data with unknown gametic phase. An example of how Arlequin v 3.5.2.2 is executed on a 64bit computer with Linux on the command line is: arlecore3522_64bit example.ars example.arp. The outputs of this command are several files with the execution details and a .res directory, where the results (.xml file) and the details of the analysis are stored (rest of the files).

Performance assessment and comparison of tools on artificial data

Due to the complexity of the genomic region of HLA, the calculations for HWE and LD do not follow the standard steps. As stated in the Introduction, HLA is the most polymorphic region in the human genome and thus, every HLA loci is multi-allelic. In order to find the most suitable software resource for these analyses that is: specific on the HLA region - because some popular tools like PLINK (Purcell et al. 2007) cannot handle HLA genotypes, we tested their performances on appropriately artificially generated genotypic HLA data. The tested tools were Arlequin v 3.5.2.2 and the web version of Genepop v 4.7.

As far as HWE is concerned, the 2 tools did not give significantly different p-values. But we were facing an issue on LD because the p-values were not always comparable. So, in order to find out and to prove which of these tools had a better efficacy on the LD estimation, we created artificial HLA genotypes of 3 loci (A, B and C) for 100 samples with Python’s 3.6.9 ‘random’ library (see Availability section). Locus A had 5 alleles, locus B had 6 and locus C had 7. We assumed that probability p = 1.0 for 2 loci meant that there is a perfect LD between them, while p = 0.0 for 2 loci meant that there is no LD between them. The genotypes of the loci pair A-C had p = 0.8, so we created them to be linked, on the contrary to the genotypes of the loci pairs A-B and B-C, which had random linkage. Therefore, we expected the two tools to find strong linkage between the loci pair A-C, and random linkage between the other 2, with the one that exhibits the best overall performance to be used in the analysis. The advantages for Arlequin v 3.5.2.2 were that it is the most commonly used tool for this type of analysis in the bibliography and more parameters can be customized by the user, in contrast to Genepop, where only 4 parameters could be tuned by the user, including the dememorization number (was set to 10.000 in our experiments), the number of batches (set to 100), the number of iterations per batch (set to 10.000) and the statistics test (we used both log likelihood ratio statistic and probability test and their results were very similar).

Arlequin v 3.5.2.2 VS Genepop v 4.7 for Linkage Disequilibrium (LD)

It is also worth of mentioning that even if with Arlequin v 3.5.2.2 one may estimate haplotype frequencies from HLA genotypic data, we chose Hapl-o-Mat v 1.1 because, unlike Arlequin v 3.5.2.2, it is able to perform ‘g’ grouping, and in addition, it is created and provided by DKMS, so our methodology is tuned and aligned and is directly comparable with respective DKMS results. The results from the comparison of Genepop v 4.7 and Arlequin v 3.5.2.2, according to the task of assessing their ability to better estimate the LD, are illustrated in Figure 3 (with log10 transformed p-values). It was proved that Arlequin v 3.5.2.2 met better our expectations because it’s p-values were more correct on the random linkage pairs (B-C) than
Genepop v 4.7. For some reason that we could not understand, the p-values from Genepop v 4.7 were always really low on the LD analyses.

Figure 3. The transformed LD p-values on every pair of the artificial data that we created prove that Arlequin v 3.5.2.2 was the best choice for our analysis.
Results

Contrasting the Cretan HLA profile with the DKMS cohorts

Statistic comparison of the frequencies of the common alleles

HLA typing of the retained 1,158 Cretan individuals resulted in the detection and annotation of 35, 68, 30 and 34 ‘g’ grouped alleles for HLA-A, -B, -C and -DRB1 loci, respectively. Subsequent comparisons showed that the majority of the detected alleles for each locus were common between the cohorts studied (33 were in common on HLA-A, 62 on HLA-B, 27 on HLA-C and 34 on HLA-DRB1). In order to compare the HLA profiles of the Cretan population with the DKMS pool, the first step is the statistical comparison of the frequencies of the common alleles.

The alleles are presented by descending order of the Cretan frequencies (red bars, Figure 4, Figure 5, Figure 6, Figure 7) and are filtered by having at least one occurrence in every population and with contingency chi-square p-value < 0.05 (or k > 1.3, where k is -log10(p-value) as shown). This is a common method to represent different allele frequencies between two or more populations (Soto-Nava et al. 2018). In addition, estimated allelic differences are accompanied by 95% confidence intervals (CI), as well as proportion difference confidence intervals (PDCI).

The most significant differences were observed for HLA alleles A*24:02g (k=12.06), B*35:02g (k=18.8), C*04:01g (k=32.55) and DRB1*11:04g (k=62.06). All of these alleles were observed at a higher frequency in the Cretan sample compared to the DKMS cohorts. Interestingly, the lowest p-value from all 4 of them was observed for HLA-DRB1, which is the most important HLA gene in transplantations (Petersdorf et al. 1995). Most of the common alleles of Crete and the DKMS pool had significant differences in their frequencies which is a first indication of the uniqueness of the HLA profile of Crete.
Figure 4. Statistic comparison of the frequencies of the common alleles of the HLA-A gene between Crete and the DKMS pool that have frequency greater than 1/2n and contingency chi-square p-value < 0.05. Confidence Intervals (CI) are colored black and the Proportion Difference Confidence Intervals (PDCI) brown.

Figure 5. Statistic comparison of the frequencies of the common alleles of the HLA-B gene between Crete and the DKMS pool that have frequency greater than 1/2n and contingency chi-square p-value < 0.05. Confidence Intervals (CI) are colored black and the Proportion Difference Confidence Intervals (PDCI) brown.
Figure 6. Statistic comparison of the frequencies of the common alleles of the HLA-C gene between Crete and the DKMS pool that have frequency greater than 1/2n and contingency chi-square p-value < 0.05. Confidence Intervals (CI) are colored black and the Proportion Difference Confidence Intervals (PDCI) brown.

Figure 7. Statistic comparison of the frequencies of the common alleles of the HLA-DRB1 gene between Crete and the DKMS pool that have frequency greater than 1/2n and contingency chi-square p-value < 0.05. Confidence Intervals (CI) are colored black and the Proportion Difference Confidence Intervals (PDCI) brown.
Unique Cretan HLA alleles

Besides the detected common alleles, **11 alleles of the class I HLA genes that were found in the Cretan cohort were not detected in the DKMS registries.** Apart from A*02:90 which was observed twice on two different individuals, all the remaining 10 alleles were observed only once in the Cretan sample (see Table 1). Additionally, in the Cretan sample we observed 9 alleles that could not be assigned with the IMGT ver. 3.35.00 and were assigned as A*03:NEW, A*24:NEW, B*07:NEW, B*13:NEW, B*51:NEW, C*04:NEW, DRB1*01:NEW, DRB1*04:NEW and DRB1*11:NEW. In order to be identified as novel alleles, the steps that are described on the IPD-IMGT/HLA Database need yet to be followed. The observed allelic diversity of the studied HLA loci suggests that the resultant haplotypes could diversify in a similar manner.

<table>
<thead>
<tr>
<th>Allele</th>
<th>Allele Frequency</th>
<th>Allele</th>
<th>Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*02:90</td>
<td>0.00086 (2 counts)</td>
<td>B*53:16</td>
<td>0.00043 (1 count)</td>
</tr>
<tr>
<td>A*23:53</td>
<td>0.00043 (1 count)</td>
<td>B*58:65</td>
<td>0.00043 (1 count)</td>
</tr>
<tr>
<td>B*07:18</td>
<td>0.00043 (1 count)</td>
<td>C*01:09</td>
<td>0.00043 (1 count)</td>
</tr>
<tr>
<td>B*15:37</td>
<td>0.00043 (1 count)</td>
<td>C*01:103</td>
<td>0.00043 (1 count)</td>
</tr>
<tr>
<td>B*35:158</td>
<td>0.00043 (1 count)</td>
<td>C*06:03</td>
<td>0.00043 (1 count)</td>
</tr>
<tr>
<td>B*51:24</td>
<td>0.00043 (1 count)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 1. 11 HLA alleles that were observed in the Cretan sample and are absent from the DKMS pool.**

PCA on Cretans and DKMS cohorts

The differentiation between HLA alleles frequencies was used to assess the genetic variation among the Cretan cohort and the DKMS populations world-wide using PCA multidimensional scaling. The results of PCA, which illustrate the genetic distances of the HLA profiles of all the datasets (except the one with donors of Chinese parentage), are shown in Figure 8. In this PCA plot we can observe some clusters that mirror the geographic locations of the corresponding countries. Crete, which is the red dot on the right side of the plot, is far from every population and is closer to the Mediterranean countries. Interestingly, Crete is closer to Turkey than it is to Greece which can be explained by the fact that the island was occupied by the Ottomans for over than two centuries and thus, there was interference in the DNA of the two populations. These results may also indicate that sometimes Crete may find a MUD in Turkey easier than in the rest of Greece. Likewise, Turkey’s HLA profile may be closer to Crete’s for some of its regions, probably in Asia Minor, and a MUD may be easier found in Crete rather than in the rest of Greece. However, the above observations are just speculative, as the specific place of origin is not available for any of the studied DKMS populations.
Figure 8. PCA plot of the common alleles of the 4 examined HLA genes for each population, apart from the Chinese, with allele frequency greater than 1/2n.

Haplotype comparison between DKMS registries and the Cretan population

A clearer image of the differentiation of two populations' HLA profiles can be obtained when comparing their haplotype frequencies. The DKMS pool, as already stated, represents the possible donors that reside in Europe. It consists of 21,314 individuals and was calculated to have 10,094 haplotypes for 4 HLA genes (A, B, C and DRB1), with their frequencies to sum up to 0.975. This can be explained due to the low coverage that may have caused ambiguities on the HLA typing. Likewise, our Cretan sample consists of 1,158 individuals and was calculated to have 993 haplotypes, with their frequencies to sum up to 1. The common haplotypes were 564 (Figure 9). Out of them, 199 had statistically significant differences in their frequencies (Fisher's exact test p-value < 0.05). The range of these p-values were from 1.46e-16 to 0.0479.

Figure 9. Haplotype number comparison between Crete and the DKMS pool.
If we merge the Cretan sample with the DKMS pool, we may observe the uniqueness of each population based on its unique number of haplotypes (Table 2). Table 2 also presents the number of common, as well as the significantly different, haplotypes between Crete and every DKMS population. The high percentage (43.2%) of the Cretan haplotypes that are unique in this merged pool compared to the other populations is a significant factor of differentiation.

Table 2. Inter population haplotypic comparison. The third column of the table shows the number of unique haplotypes that every population has if we merge the DKMS pool with the Cretan cohort. The fourth column shows how many haplotypes are in common with the Crete’s and out of them, on the fifth column, how many have statistically significant differences in their frequencies (Fisher’s exact test p-value < 0.05).

<table>
<thead>
<tr>
<th>POPULATION SIZE</th>
<th># HAPLOTYPES</th>
<th># UNIQUE HAPLOTYPES</th>
<th># COMMON WITH CRETE’S</th>
<th># SIGNIFICANT (p&lt;0.05)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRE (1158)</td>
<td>993</td>
<td>429 (43.2%)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AUS (1698)</td>
<td>1304</td>
<td>501 (38.4%)</td>
<td>169</td>
<td>26</td>
</tr>
<tr>
<td>BOS (1028)</td>
<td>797</td>
<td>223 (27.9%)</td>
<td>147</td>
<td>24</td>
</tr>
<tr>
<td>CHI (1282)</td>
<td>1119</td>
<td>871 (77.8%)</td>
<td>53</td>
<td>10</td>
</tr>
<tr>
<td>CRO (2057)</td>
<td>1341</td>
<td>444 (33.1%)</td>
<td>243</td>
<td>40</td>
</tr>
<tr>
<td>FRA (1406)</td>
<td>1207</td>
<td>506 (41.9%)</td>
<td>169</td>
<td>25</td>
</tr>
<tr>
<td>GRE (1894)</td>
<td>1546</td>
<td>614 (39.7%)</td>
<td>269</td>
<td>28</td>
</tr>
<tr>
<td>ITA (1159)</td>
<td>1064</td>
<td>443 (41.6%)</td>
<td>163</td>
<td>12</td>
</tr>
<tr>
<td>NET (1374)</td>
<td>998</td>
<td>344 (34.4%)</td>
<td>153</td>
<td>28</td>
</tr>
<tr>
<td>POR (1176)</td>
<td>1049</td>
<td>431 (41%)</td>
<td>155</td>
<td>22</td>
</tr>
<tr>
<td>ROM (1234)</td>
<td>1063</td>
<td>350 (32.9%)</td>
<td>185</td>
<td>21</td>
</tr>
<tr>
<td>SPA (1107)</td>
<td>1015</td>
<td>450 (44.3%)</td>
<td>146</td>
<td>27</td>
</tr>
<tr>
<td>TUR (4856)</td>
<td>3255</td>
<td>1829 (56.2%)</td>
<td>347</td>
<td>72</td>
</tr>
<tr>
<td>UK (1043)</td>
<td>806</td>
<td>286 (35.4%)</td>
<td>101</td>
<td>20</td>
</tr>
</tbody>
</table>

The most common Cretan haplotype was A*24:02g~B*35:02g~C*04:01g~DRB1*11:04g with frequency 1.8%. This haplotype was ranked 14th in the DKMS pool with frequency ~0.5% (p=2.22e-10). Interestingly, the typical European haplotype A*01:01g~B*08:01g~C*07:01g~DRB1*03:01g was on rank 1 (3.5%) both for the pool of DKMS countries, as well as for each European country (including the Greek DKMS cohort) separately, ranging between 2.4% and 8.03%. However, this haplotype was ranked 3rd on the Cretan cohort with frequency 1.4% (p=1.28e-9). As may be observed in Figure 10, there is a statistical comparison of the
frequencies of the 10 most common haplotypes in Crete. As the Fisher’s exact test p-values (p) indicate, 6 out of these 10 haplotypes have significant differences in their frequencies between the 2 groups studied. The top haplotypes in Crete are represented in Europe in a way that, alongside with the different rankings, the differentiation of the two HLA profiles is indicated.

Figure 10. Haplotype frequency comparison of the 10 most common haplotypes in Crete with the frequencies in the DKMS pool (Fisher’s exact test p-values (p) are shown).

**HLA profile with 6 genes of 1146 Cretans**

### Allele Frequencies

Besides estimating the differences in the distribution of HLA alleles and haplotypes between Crete and other countries world-wide, we set to delineate the HLA allelic and haplotypic profile of Cretans considering not only the 6 important HLA loci in HSCT, but also the genetic diversity that different Cretan population clusters might exhibit. The latter was achieved following stratification of the total Cretan sample by grandparents’ age place of origin and classifying available samples into 4 major categories based on the 4 Prefectures of Crete (i.e., Chania, Rethymno, Heraklion, Lasithi).

Moving onto the characterization of the Cretan HLA profile with 6 genes (A, B, C, DRB1, DQB1 and DPB1) we first present the 10 most common alleles of Class I (Table 3) and Class II (Table 4) HLA genes. ‘g’ grouping was not needed and thus, not performed. The ‘N’ columns show the estimated counts of the alleles in the Cretan population (1146 fully genotyped Cretans on these 6 genes).
### Table 3.

10 most common alleles of Class I HLA genes of 1146 Cretans. ‘N’ columns show the estimated counts of the alleles in the Cretan population and ‘AF’ columns show the estimated frequency of the alleles.

<table>
<thead>
<tr>
<th>HLA-A</th>
<th>N</th>
<th>AF</th>
<th>HLA-B</th>
<th>N</th>
<th>AF</th>
<th>HLA-C</th>
<th>N</th>
<th>AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>A*02:01</td>
<td>429</td>
<td>0.187010</td>
<td>B*35:01</td>
<td>225</td>
<td>0.098082</td>
<td>C*04:01</td>
<td>547</td>
<td>0.236448</td>
</tr>
<tr>
<td>A*24:02</td>
<td>374</td>
<td>0.163034</td>
<td>B*51:01</td>
<td>211</td>
<td>0.091979</td>
<td>C*07:01</td>
<td>291</td>
<td>0.126853</td>
</tr>
<tr>
<td>A*01:01</td>
<td>234</td>
<td>0.102005</td>
<td>B*18:01</td>
<td>189</td>
<td>0.082389</td>
<td>C*06:02</td>
<td>205</td>
<td>0.089364</td>
</tr>
<tr>
<td>A*11:01</td>
<td>178</td>
<td>0.077594</td>
<td>B*35:03</td>
<td>119</td>
<td>0.051874</td>
<td>C*12:03</td>
<td>193</td>
<td>0.084133</td>
</tr>
<tr>
<td>A*03:01</td>
<td>163</td>
<td>0.071055</td>
<td>B*13:02</td>
<td>109</td>
<td>0.047515</td>
<td>C*02:02</td>
<td>154</td>
<td>0.067132</td>
</tr>
<tr>
<td>A*32:01</td>
<td>153</td>
<td>0.066696</td>
<td>B*49:01</td>
<td>106</td>
<td>0.046207</td>
<td>C*07:02</td>
<td>116</td>
<td>0.050567</td>
</tr>
<tr>
<td>A*26:01</td>
<td>126</td>
<td>0.054926</td>
<td>B*35:02</td>
<td>98</td>
<td>0.042720</td>
<td>C*15:02</td>
<td>90</td>
<td>0.039233</td>
</tr>
<tr>
<td>A*23:01</td>
<td>89</td>
<td>0.038797</td>
<td>B*44:02</td>
<td>88</td>
<td>0.038361</td>
<td>C*08:02</td>
<td>83</td>
<td>0.036181</td>
</tr>
<tr>
<td>A*31:01</td>
<td>65</td>
<td>0.028335</td>
<td>B*14:02</td>
<td>76</td>
<td>0.033130</td>
<td>C*12:02</td>
<td>72</td>
<td>0.031386</td>
</tr>
<tr>
<td>A*30:01</td>
<td>63</td>
<td>0.027463</td>
<td>B*52:01</td>
<td>75</td>
<td>0.032694</td>
<td>C*03:03</td>
<td>64</td>
<td>0.027899</td>
</tr>
</tbody>
</table>

### Table 4.

10 most common alleles of Class II HLA genes of 1146 Cretans. ‘N’ columns show the estimated counts of the alleles in the Cretan population and ‘AF’ columns show the estimated frequency of the alleles.

<table>
<thead>
<tr>
<th>HLA-DRB1</th>
<th>N</th>
<th>AF</th>
<th>HLA-DQB1</th>
<th>N</th>
<th>AF</th>
<th>HLA-DPB1</th>
<th>N</th>
<th>AF</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRB1*11:04</td>
<td>375</td>
<td>0.163470</td>
<td>DQB1*03:01</td>
<td>730</td>
<td>0.318221</td>
<td>DPB1*04:01</td>
<td>767</td>
<td>0.334350</td>
</tr>
<tr>
<td>DRB1*11:01</td>
<td>230</td>
<td>0.100262</td>
<td>DQB1*05:02</td>
<td>287</td>
<td>0.125109</td>
<td>DPB1*02:01</td>
<td>467</td>
<td>0.203575</td>
</tr>
<tr>
<td>DRB1*16:01</td>
<td>215</td>
<td>0.093723</td>
<td>DQB1*05:01</td>
<td>265</td>
<td>0.115519</td>
<td>DPB1*04:02</td>
<td>381</td>
<td>0.166085</td>
</tr>
<tr>
<td>DRB1*07:01</td>
<td>213</td>
<td>0.092851</td>
<td>DQB1*02:02</td>
<td>224</td>
<td>0.097646</td>
<td>DPB1*10:01</td>
<td>105</td>
<td>0.045772</td>
</tr>
<tr>
<td>DRB1*03:01</td>
<td>137</td>
<td>0.059721</td>
<td>DQB1*05:03</td>
<td>175</td>
<td>0.076286</td>
<td>DPB1*10:04</td>
<td>85</td>
<td>0.037053</td>
</tr>
<tr>
<td>DRB1*14:54</td>
<td>126</td>
<td>0.054926</td>
<td>DQB1*03:02</td>
<td>141</td>
<td>0.061465</td>
<td>DPB1*03:01</td>
<td>83</td>
<td>0.036181</td>
</tr>
<tr>
<td>DRB1*01:01</td>
<td>100</td>
<td>0.043592</td>
<td>DQB1*02:01</td>
<td>139</td>
<td>0.060593</td>
<td>DPB1*13:01</td>
<td>81</td>
<td>0.035310</td>
</tr>
<tr>
<td>DRB1*13:01</td>
<td>83</td>
<td>0.036181</td>
<td>DQB1*06:03</td>
<td>100</td>
<td>0.043592</td>
<td>DPB1*17:01</td>
<td>66</td>
<td>0.028771</td>
</tr>
<tr>
<td>DRB1*13:02</td>
<td>81</td>
<td>0.035310</td>
<td>DQB1*06:04</td>
<td>55</td>
<td>0.023976</td>
<td>DPB1*14:01</td>
<td>63</td>
<td>0.027463</td>
</tr>
<tr>
<td>DRB1*15:01</td>
<td>80</td>
<td>0.034874</td>
<td>DQB1*06:01</td>
<td>48</td>
<td>0.020924</td>
<td>DPB1*01:01</td>
<td>51</td>
<td>0.022232</td>
</tr>
</tbody>
</table>
Haplotype Frequencies

The most crucial aspect in the characterization of a population-specific HLA profile is the distribution of respective haplotypes. Table 5 presents the 10 most common haplotypes composed of 6 HLA loci estimated using Hapl-o-Mat v 1.1. Following the same pattern as before, ‘g’ grouping was not needed and thus, not performed. The ‘N’ columns show the number of individuals carrying each detected haplotype. Due to the rounding, different haplotype frequencies (‘HF’) may correspond to the same absolute number of individuals.

Table 5. 10 most common haplotypes with 6 genes of 1146 Cretans. The ‘N’ column shows the estimated counts of the haplotypes in the Cretan population and the ‘HF’ column shows the estimated frequency of the haplotype.

<table>
<thead>
<tr>
<th>Haplotypes</th>
<th>N</th>
<th>HF</th>
</tr>
</thead>
<tbody>
<tr>
<td>A<em>33:01-B</em>14:02-C<em>08:02-DPB1</em>04:01-DQB1<em>05:01-DRB1</em>01:02</td>
<td>32</td>
<td>0.0139</td>
</tr>
<tr>
<td>A<em>24:02-B</em>35:02-C<em>04:01-DPB1</em>04:01-DQB1<em>03:01-DRB1</em>11:04</td>
<td>22</td>
<td>0.0095</td>
</tr>
<tr>
<td>A<em>01:01-B</em>08:01-C<em>07:01-DPB1</em>01:01-DQB1<em>02:01-DRB1</em>03:01</td>
<td>19</td>
<td>0.0083</td>
</tr>
<tr>
<td>A<em>02:01-B</em>18:01-C<em>07:01-DPB1</em>04:02-DQB1<em>03:01-DRB1</em>11:04</td>
<td>17</td>
<td>0.0076</td>
</tr>
<tr>
<td>A<em>02:01-B</em>13:02-C<em>06:02-DPB1</em>17:01-DQB1<em>02:02-DRB1</em>07:01</td>
<td>16</td>
<td>0.0068</td>
</tr>
<tr>
<td>A<em>31:01-B</em>35:03-C<em>04:01-DPB1</em>10:01-DQB1<em>05:02-DRB1</em>16:01</td>
<td>15</td>
<td>0.0065</td>
</tr>
<tr>
<td>A<em>24:02-B</em>35:02-C<em>04:01-DPB1</em>04:02-DQB1<em>03:01-DRB1</em>11:04</td>
<td>15</td>
<td>0.0064</td>
</tr>
<tr>
<td>A<em>01:01-B</em>57:03-C<em>07:01-DPB1</em>04:01-DQB1<em>02:01-DRB1</em>03:01</td>
<td>14</td>
<td>0.0061</td>
</tr>
<tr>
<td>A<em>23:01-B</em>49:01-C<em>07:01-DPB1</em>04:01-DQB1<em>03:01-DRB1</em>11:01</td>
<td>12</td>
<td>0.0052</td>
</tr>
<tr>
<td>A<em>02:01-B</em>13:02-C<em>06:02-DPB1</em>04:01-DQB1<em>02:02-DRB1</em>07:01</td>
<td>11</td>
<td>0.0049</td>
</tr>
</tbody>
</table>

Hardy-Weinberg Equilibrium (HWE)

In order to unravel the significance of genetic diversity measures (e.g. HWE, LD) in the estimated allelic frequencies we set out to calculate observed and expected heterozygosity (OH and EH) at a locus-by-locus level using a Markov Chain sampling scheme with 1x10^6 steps implemented in Arlequin v 3.5.2.2. Moreover, all estimations were made under the assumption of the population under study being in HWE. The lowest heterozygosity in the Cretan sample for each HLA locus studied was observed for HLA-DPB1 (81.5%), also demonstrating the only significant deviation between OH and EH measures under HWE (Table 6). We further examined the homozygosity percentage of HLA-DPB1 and we found that more than half of the individuals (139 out of the 212) who were homozygous on HLA-DPB1 were also homozygous on another HLA gene. Specifically, we found that 29 Cretans out of 1146 (2.53%) were both homozygous on HLA-A and on HLA-DPB1, 18 (1.57%) on HLA-B and on HLA-DPB1, 23 (2.01%) on HLA-C and on HLA-DPB1, 26 (2.27%) on HLA-DRB1 and on HLA-DPB1 and finally, 43 (3.75%) on HLA-DQB1 and on HLA-DPB1.
Table 6. Observed and Expected Heterozygosity at each HLA locus studied with accompanied significance p values in the 1146 Cretans.

<table>
<thead>
<tr>
<th>Locus (number of alleles)</th>
<th>Observed Heterozygosity</th>
<th>Expected Heterozygosity</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLA-A (37)</td>
<td>0.90227</td>
<td>0.90433</td>
<td>0.61161</td>
</tr>
<tr>
<td>HLA-B (69)</td>
<td>0.94415</td>
<td>0.95537</td>
<td>0.18490</td>
</tr>
<tr>
<td>HLA-C (35)</td>
<td>0.89180</td>
<td>0.89733</td>
<td>0.28870</td>
</tr>
<tr>
<td>HLA-DRB1 (35)</td>
<td>0.91798</td>
<td>0.92790</td>
<td>0.27166</td>
</tr>
<tr>
<td>HLA-DQB1 (18)</td>
<td>0.82984</td>
<td>0.84348</td>
<td>0.31523</td>
</tr>
<tr>
<td>HLA-DPB1 (38)</td>
<td>0.81501</td>
<td>0.81074</td>
<td>0.00196</td>
</tr>
</tbody>
</table>

Calculated significance values following comparison of OH and EH at each HLA locus under the assumption of HWE are graphically represented in Figure 11. The p-values are log transformed for better visualization and the threshold has been adjusted with Holms-Bonferroni correction. As we can see, only alleles at locus HLA-DPB1 are not in HWE which may be related to the fact that the OH is higher than the EH (see Table 6). It seems that some evolutionary factors are affecting this genetic locus.

Figure 11. HWE p-values on every HLA locus studied on 1146 Cretans.
Linkage Disequilibrium (LD)

Arlequin v 3.5.2.2 was also used to calculate delta (D<sub>ij</sub>) and standardized delta (D<sub>ij</sub>') values to measure LD for pairwise nonrandom associations of alleles at different HLA loci and their statistical significance. Results are presented in Figure 12 (with the p-values for every pair of HLA locus being transformed and adjusted as done above). All HLA gene pairs seem to be in strong LD (p<10<sup>-4</sup>), except from these containing the gene HLA-DPB1. The weak LD estimated for HLA loci pairs encompassing DPB1 may be due to the hotspot recombination region mapped between HLA-DP and -DQ loci (Cullen et al. 2002). Even though 4 intense hotspot recombination regions have been identified (DPB1 to RING3, DQB3 to DQB1, BAT2 to LTA, and telomeric to HLA-F), rates of recombination and disruption of recombination events may vary significantly among individuals resulting in a profound effect on haplotypic blocks distribution and gene typing of particular traits. Additionally, it does not contradict the fact that this locus is not in HWE and that it has the highest homozygosity percentage probably because some genetic and evolutionary factors are taking place on this genetic locus. The HLA pair DQB1-DPB1 seems more to be in linkage compared to the other DPB1 pairs probably because these two loci are closer to each other on the 6th chromosome.

**Figure 12.** LD p-values on every pair of the examined HLA loci studied on 1146 Cretans.

**HLA profile of the 4 Cretan Prefectures**

**Allele Frequencies**

The 1,146 Cretans were grouped in each Prefecture according to their origins as explained in the Methods. 312 individuals were from Chania, 229 from Rethymno, 412 from Heraklion and 193 from Lasithi. Table 7 shows not only the total number of samples per Prefecture and stratified by the 6 HLA loci, but also the number of alleles shared between all 4 Prefectures at each locus.
Table 7. Number of HLA alleles per Prefecture.

<table>
<thead>
<tr>
<th></th>
<th>HLA-A</th>
<th>HLA-B</th>
<th>HLA-C</th>
<th>HLA-DRB1</th>
<th>HLA-DQB1</th>
<th>HLA-DPB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chania  (n=312)</td>
<td>28</td>
<td>47</td>
<td>31</td>
<td>28</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>Rethymno (n=229)</td>
<td>28</td>
<td>48</td>
<td>28</td>
<td>30</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>Heraklion (n=412)</td>
<td>34</td>
<td>54</td>
<td>27</td>
<td>34</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>Lasithi (n=193)</td>
<td>29</td>
<td>49</td>
<td>25</td>
<td>28</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>total (n=1146)</td>
<td>37</td>
<td>69</td>
<td>35</td>
<td>35</td>
<td>18</td>
<td>38</td>
</tr>
<tr>
<td>Common</td>
<td>23</td>
<td>37</td>
<td>24</td>
<td>23</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Subsequent comparisons regarding the distribution of alleles for every HLA gene studied across the 4 Prefectures (‘g’ grouping was not performed) are shown in Figure 13, Figure 14, Figure 15, Figure 16, Figure 17, Figure 18. The alleles are sorted by descending order according to the frequency that they have in Crete as a whole (black line) which can be treated as a reference frequency. Alleles on HLA-C (Figure 15), HLA-DQB1 (Figure 17) and HLA-DPB1 (Figure 18) have similar frequencies across the 4 Prefectures on the contrary to the rest of the HLA genes where a higher differentiation is clear. For example, on HLA-A (Figure 13) the allele A*26:01 is one of the alleles that have significantly higher frequency on Lasithi compared to the other 3 Prefectures.

![Figure 13. Allelic distribution of HLA-A gene across the 4 Prefectures of Crete.](image-url)
**Figure 14.** Allelic distribution of HLA-B gene across the 4 Prefectures of Crete.

**Figure 15.** Allelic distribution of HLA-C gene across the 4 Prefectures of Crete.
Figure 16. Allelic distribution of HLA-DRB1 gene across the 4 Prefectures of Crete.

Figure 17. Allelic distribution of HLA-DQB1 gene across the 4 Prefectures of Crete.
Figure 18. Allelic distribution of HLA-DPB1 gene across the 4 Prefectures of Crete.

Every allele that was observed only in one Prefecture along with its numeric count (‘#’ columns) is presented on Table 8. The allele C*07:06 was observed 6 times only in Chania and had the most counts of every unique allele. Heraklion was the only Prefecture that had unique alleles in all 6 HLA genes, probably because of its biggest sample size.

Table 8. Unique alleles across the 4 Prefectures of Crete.

<table>
<thead>
<tr>
<th>Prefecture (2 alleles)</th>
<th>Total (1146 x 2 alleles)</th>
<th>HLA-A</th>
<th>HLA-B</th>
<th>HLA-C</th>
<th>HLA-DRB1</th>
<th>HLA-DQB1</th>
<th>HLA-DPB1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chania (312 x 2 alleles)</td>
<td>A*24:02 L</td>
<td>B*27:09</td>
<td>2</td>
<td>C*07:06</td>
<td>6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*35:15</td>
<td>1</td>
<td>C*01:09</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C*07:04</td>
<td>1</td>
<td>C*06:03</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*58:65</td>
<td>1</td>
<td>C*14:03</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rethymno (229 x 2 alleles)</td>
<td>-</td>
<td>B*13:01</td>
<td>2</td>
<td>C*04:03</td>
<td>2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*07:18</td>
<td>1</td>
<td>C*01:103</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*15:24</td>
<td>1</td>
<td>C*06:06</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*53:16</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Heraklion (412 x 2 alleles)</td>
<td>A*33:05</td>
<td>B*39:05</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A*02:90</td>
<td>B*39:31</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A*02:06</td>
<td>B*18:18</td>
<td>1</td>
<td>C*02:29</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A*34:02</td>
<td>B*07:10</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*15:37</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lasithi (193 x 2 alleles)</td>
<td>A*23:53</td>
<td>B*37:04</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*15:10</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*15:16</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*64:29</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B*51:24</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
We wanted to see how close these groups of Cretans were to the DKMS registries. So, we performed once again PCA multidimensional scaling in order to get an idea about the **genetic distances of the 4 Prefectures with respect to all the available DKMS cohorts**. We performed ‘g’ grouping on the alleles of the genes HLA-A, HLA-B, HLA-C and HLA-DRB1 because our sample did not have the other 2 genes genotyped. As we can see from Figure 19, the 4 Prefectures are closer to the Mediterranean countries (just like Crete as a total from Figure 8) with the Prefecture of Chania being closer to the Greek cohort of the DKMS and away from the other 3 Prefectures.

![PCA on Cretan Prefectures and DKMS, COMMON alleles, FILTERED](image)

**Figure 19.** PCA plot of the common alleles of the 4 examined HLA genes for each population, apart from the Chinese, with allele frequency greater than 1/2n.

Haplotype frequencies

Table 9 shows the total, as well as the unique, number of haplotypes composed of 6 HLA genes in every Prefecture. The high number of unique haplotypes per Prefecture and the fact that only 4 are in common in all Prefectures can be explained to an extent due to the fact that the haplotype has 6 genes. When we decreased the number of the genes in the haplotype to 5 and 4 we found again a high number of unique haplotypes and the common were only 8 in both cases (Tables not shown). All these suggest that the 4 Prefectures of Crete are differentiated probably because of a geographical or a cultural isolation and finding possible donors for transplantation from one Prefecture to another might be more difficult than expected. The 5 most common haplotypes with 6 genes along with their respective frequencies (‘HF’) in every Prefecture are presented in Figure 20.
Table 9. Number of HLA haplotypes with 6 genes per Prefecture.

<table>
<thead>
<tr>
<th>Prefecture</th>
<th>Number of TOTAL haplotypes</th>
<th>Number of UNIQUE haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chania</td>
<td>408</td>
<td>354</td>
</tr>
<tr>
<td>Rethymno</td>
<td>300</td>
<td>245</td>
</tr>
<tr>
<td>Heraklion</td>
<td>536</td>
<td>458</td>
</tr>
<tr>
<td>Lasithi</td>
<td>287</td>
<td>249</td>
</tr>
<tr>
<td>common</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 20. Top 5 haplotypes with 6 genes per Prefecture along with their frequencies (‘HF’).

Hardy-Weinberg Equilibrium (HWE)

Using the Arlequin software (v 3.5.2.2) we estimated the HWE p-values on every HLA locus in all Prefectures. Surprisingly, all genetic loci in all cases seem to be in HWE (p-value lower than the Holms-Bonferroni adjusted threshold) (Figure 21). As we presented in Figure 11, where we examined the HWE in the total Cretan population, HLA-DPB1 was not in HWE, but when we stratified the population by its Prefectures then it was in HWE for all of them. These demographic traits may have occurred due to possible genetic events (bottleneck, genetic drift, natural selection process etc) or are due to the sample size of the subgroups which may be small for this kind of analysis.
Linkage Disequilibrium (LD)

We also estimated the LD levels for all possible HLA pairs in every Prefecture with Arlequin v 3.5.2.2 and we found out that most of the pairs are not in LD (Figure 22). There was no pattern on the pairs that were strongly linked, so no safe induction can be made. These unpredictable results unravel the complexity of HLA region, which may be influenced by the small sample size used or the demographic history of the populations under study affected by genetic diversity measures (e.g. genetic drift, bottleneck effects) that may underlie the balancing selection and/or mutation rate of HLA alleles. In order to get a clearer image of the LD in these subpopulations and be more confident about whether the results are due to luck or not, we must either increase the sample size or perform some statistical analysis, like False Discovery Rate (FDR) validation or power analysis.
Discussion

Crete is an island with remarkable geographic landscape and rich history which have given cultural characteristics and thus unique genetic makeup in the Cretan population. Prompt by the significance of genetic diversity on the evolution and allelic heterogeneity of HLA, we set out to construct the HLA immunogenetic profile of 1,204 Cretan donors representatively selected from different regions of Crete with self-assessed parentage and grand parentage at recruitment. Moreover, we were interested in the detection of potential differences in the frequencies and distribution of HLA alleles and haplotypes between Crete and other well-characterized populations. For this purpose, we investigated the HLA diversity of the island and compared it with other known populations stored in one of the most representative European unrelated donor registries (UDR), the German DKMS.

Our main objective was to investigate whether the HLA immunogenetic profile of Crete is adequately represented in one of the best known European UDR registry for allogeneic stem cell transplantation. We found that a little less than half of the Cretan haplotypes were common with the haplotypes of DKMS. The distribution of common alleles was significantly different and Crete had a relatively high number of unique alleles compared to the DKMS. Moreover, 9 samples with high resolution typing, but, yet, unassigned WHO names for new alleles, are further processed by Sanger sequencing in order to verify the existence of 9 potential novel alleles in the Cretan cohort. The highest HLA haplotype frequency in Crete was recorded for A*24:02~B*35:02~C*04:01~DRB1*11:04 (1.8%, Figure 10), which was on rank 14 (0.5%) in the combined DKMS sample. Interestingly, the typical European haplotype A*01:01~B*08:01~C*07:01~DRB1*03:01, which was on rank 1 for all European countries (including the Greek DKMS cohort) ranging between 2.4% and 8.03%, was on rank 3 in Crete (1.4%). Subsequent multidimensional scaling for HLA-A, -B, -C and -DRB1 high-resolution allele frequencies showed two distinctive geographic clusters consisted of South-Eastern (Bosnia, Croatia, Romania) and North-Western (United Kingdom, Austria, Netherlands, France) European countries, while a third cluster was composed of countries from the Mediterranean basin (Spain, Portugal, Italy, Turkey). It is intriguing that the Greek minority of DKMS was placed between the South-Eastern European and the Mediterranean basin clusters, while Crete was closer to the latter (Figure 8).

In order to explore the validity of our observations we estimated the significance of confounding population structure measures (e.g. HWE, LD) on potential nonrandom associations of alleles at different HLA loci. Subsequent analyses revealed that the Cretan population was in HWE at all loci studied, except for DPB1. Moreover, our results verified that all pairs of HLA loci are in LD, except for those including HLA-DPB1. The weak LD estimated for HLA loci pairs encompassing DPB1 may be due to the hotspot recombination region mapped between HLA-DP and -DQ loci (Cullen et al. 2002). Even though 4 intense hotspot recombination regions have been identified (DPB1 to RING3, DQB3 to DQB1, BAT2 to LTA, and telomeric to HLA-F), rates of recombination and disruption of recombination events may vary significantly among individuals resulting in a profound effect on haplotypic blocks distribution and gene typing of particular traits.

Intrigued by the comparisons between the Cretan cohort and the 13 populations of DKMS we asked whether the specific analyses might have been influenced by potential genetic diversity that is known to characterize specific population clusters in Crete. Thus, following
data stratification by grandparentage place of origin we constructed the HLA profile of the 4 Cretan prefectures. Results showed that the 4 Prefectures of Crete do not seem to have similar immunogenetic profiles, probably due to geographical or cultural isolation. These findings suggest that Crete constitutes an underrepresented population in the common DKMS UDR with high genetic diversity. Our data also suggest that building representative population reference datasets will significantly improve the variant interpretation globally, while at the same time will accelerate the discovery of HLA alleles associated with rare phenotypes, particularly in populations characterized by unique properties as a result of geographical or cultural isolation. Therefore, the regional public UCB bank could play an important role for the enrichment of the national and international UDR improving the possibility for a patient to find a MUD when needed.

Throughout the study we realized that there was no Greek immunogenetic profile available for comparison with high-resolution and up-to-date methods for accurate HLA typing, other than the Greek minority of the DKMS. Additionally, we realized that the hospitals and other health related organizations that store this information in Greece do not collect detailed place of origin data of the donors’ samples. Our approach on HLA analyses has never been performed at the regional or national level in Greece, and with this study we demonstrated the benefits of such an initiative. We believe that it is of great importance for Greece and for the rest of the world to know that the more information of the stored HLA samples a hospital or organization has, the easier it will be for patients to find a MUD when needed without losing precious time which can cost their lives.

Besides the detailed donor registry and the accurate/high-resolution typing of the A, B, C, DRB1 and DQB1 HLA genes, the proposed analyses unravel the importance of haplotype phasing in HLA studies. For instance, pairwise parent-child sequencing will define the phase of the respective haplotypes in all, crucial for HSCT, HLA loci. In this way, the real haplotype frequencies of the sample can be calculated avoiding a considerable error or false discovery rate resulting from the maximum likelihood frequencies estimated using algorithms implemented in various bioinformatic resources. We should note here that in case of population analysis, the parent’s/child’s profile should not be counted in the sample and should be used for the haplotype frequency calculation only, because some alleles may have higher frequencies in the sample which would not represent the frequency of the general population. It may seem like a more expensive procedure at first glance, but the current methods actually cost more because of some additional lab experiments for reassuring the histocompatibility of the possible donor and recipient which won’t be needed if the above information is collected from the start.

After the completion of the current study, we are envisaging some possible future projects/plans. First of all, we are expecting approximately 800 more Cretan samples with the final sample size reaching the range of ~2000. All the above analyses (HLA profile creation and intra/inter population comparison), which has been automated to a high extent (see Availability section), will be redone using the final sample. Additionally, our aim is to expand the proposed intra-population HLA analyses at the level of Municipalities in Crete in order to find populations that may need more representation in the national UCB biobanks. For example, preliminary population structure analyses within our cohort following stratification by place of origin, i.e. Mylopotamos (n=112) versus rest of Crete (n=1,046), revealed a
considerable deviation from HWE for specific HLA loci and noticeable differences in HLA locus pairwise LD, both in Mylopotamos. The observed differences in population structure measures (e.g. HWE, LD) could be due to the small sample size of the cohort residing in the Mylopotamos area or could result from the unique genetic makeup of this specific Cretan population. Due to the complexity of the HLA region, the rare HLA haplotypes are just as important in HSCT as the most common haplotypes. However, the smaller the sample size, the less chance to detect the rare haplotypes will be. For this reason, our main research efforts in future studies will focus on the estimation of the adequate sample size that will be used to safely delineate the HLA immunogenetic profile of the enriched Cretan population cohort.

Moving on, we are actively looking for collaborations with other hospitals/institutes that may have similar HLA data stored, so that we will compare our Cretan cohort with representative samples from the general population of Greece. This task was impossible in the present study due to the unavailability of such samples. This will allow us to better understand the HLA diversity in Crete and to unravel the immunogenetic profile of the Greek population. Other projects that we have in mind are the creation of a model that will estimate the minimum sample size that a biobank should have according to its haplotype diversity and the creation of a bioinformatics tool that will calculate the haplotype frequencies from genotype HLA data alongside with their confidence intervals and p-values, because as far as we know none of the available tools do so.

As a conclusion, in this study, we had the privilege to take part in the analysis of samples that were collected and typed with meticulous care and quality. We also had the chance to examine both the potentials and the weaknesses of local UCB banks. This gave us the rare opportunity to form a long-term future work, as presented above, that we hope to bring to reality. With these projects and the current one, we further hope to inform and persuade the community that in MUD registries, it is not only the sample size that should be taken into consideration, but also the diversity of the HLA haplotypes. We are also making a case that when MUD registries and public UCB banks are expanded, special care should be taken so as to include donors from regions with high HLA diversities.

Availability

The complete code of this analysis will be available at a public repository under the MIT License after the publication of this work in a Hematology Journal. HLA calls and individual origins data are not available due to active legislation on private data protection. Allele and haplotype frequencies will be made available in the Allele Frequency Net Database after the completion of the analysis with the final sample of ~2000 Cretans.

Achievements

This work was accepted and presented at the 25th European Hematology Association (EHA) Annual Congress as an e-poster and was given a travel grant award. The abstract and the e-poster are presented below. After the completion of the analysis with the final sample, the work will be submitted for publication in a Hematology Journal.
ABSTRACT AT EHA

THE IMPORTANCE OF THE REGIONAL PUBLIC CORD BLOOD BANK OF CRETE FOR ALLOGENEIC HEMATOPOIETIC CELL TRANSPLANTATION ACCESS OF THE CRETAN POPULATION

Authors: Emmanuel Stylianakis, MSc thesis, July 2020

ABSTRACT

Introduction

The importance of the regional public cord blood bank of Crete for allogeneic hematopoietic cell transplantation (HCT) has increased in recent years. The regional public cord blood bank of Crete is a key component in providing a source of hematopoietic stem cells for HCT, which has been increasingly recognized as an effective treatment for various hematologic malignancies and autoimmune diseases.

Methods

The regional public cord blood bank of Crete has been active since 2012 and has collected cord blood stem cells from single donors. The bank is registered with the European Cord Blood Bank Network (ECBB) and is a member of the NCBB network. The bank currently has a research license and is approved by the Hellenic Council for Research.

Results

Since its establishment, the regional public cord blood bank of Crete has collected cord blood stem cells from 300 donors. The bank has successfully processed and cryopreserved cord blood stem cells for transplantation. The bank has also established a coordination and management plan to ensure the quality and safety of cord blood stem cells.

Conclusions

The regional public cord blood bank of Crete is an important resource for allogeneic HCT in Crete. The bank has successfully processed and cryopreserved cord blood stem cells for transplantation. The regional public cord blood bank of Crete is an important resource for allogeneic HCT in Crete. The bank has successfully processed and cryopreserved cord blood stem cells for transplantation.

REFERENCES


CONTACT INFORMATION

Email: stylianakis@med.uoi.gr
Phone: +30 694 123 4567
Address: Regional Public Cord Blood Bank of Crete, University of Crete, Heraklion, Greece

E-HA
EUROPEAN HEMATOPOIETIC ASSOCIATION

e-POSTER AT EHA

THE IMPORTANCE OF THE REGIONAL PUBLIC CORD BLOOD BANK OF CRETE FOR ALLOGENEIC HEMATOPOIETIC CELL TRANSPLANTATION ACCESS OF THE CRETAN POPULATION


INTRODUCTION

The importance of the regional public cord blood bank of Crete for allogeneic hematopoietic cell transplantation (HCT) has increased in recent years. The regional public cord blood bank of Crete is a key component in providing a source of hematopoietic stem cells for HCT, which has been increasingly recognized as an effective treatment for various hematologic malignancies and autoimmune diseases.

METHODS

Peripheral blood samples were collected from 300 unrelated individuals of Crete origin. The samples were stored in liquid nitrogen at -196°C. The samples were then sent to the regional public cord blood bank of Crete for storage and processing.

RESULTS

Since its establishment, the regional public cord blood bank of Crete has collected cord blood stem cells from 300 donors. The bank has successfully processed and cryopreserved cord blood stem cells for transplantation. The bank has also established a coordination and management plan to ensure the quality and safety of cord blood stem cells.

CONCLUSIONS

The regional public cord blood bank of Crete is an important resource for allogeneic HCT in Crete. The bank has successfully processed and cryopreserved cord blood stem cells for transplantation. The regional public cord blood bank of Crete is an important resource for allogeneic HCT in Crete. The bank has successfully processed and cryopreserved cord blood stem cells for transplantation.
References


