Czech Technical University in Prague

Faculty of Electrical Engineering

Doctoral Thesis

April 2013

David Steiner

Czech Technical University in Prague

Faculty of Electrical Engineering

Department of Cybernetics

PROBABILISTIC MATCHING IN SEARCH FOR UNRELATED HEMATOPOIETIC STEM CELL DONORS

Doctoral Thesis

David Steiner

Prague, April 2013

Ph.D. Programme: Electrical Engineering and Information Technology

Branch of study: Artificial Intelligence and Biocybernetics

Supervisor: ***Doc. Ing. Lenka Lhotská, Csc.***

Supervisor-Specialist: ***Ing. Kamil Matoušek, Ph.D.***

Acknowledgements

Firstly, I would like to thank to my supervisor *Doc. Ing. Lenka Lhotská, Csc.* and my supervisor-specialist *Ing. Kamil Matoušek, Ph.D.* for their advices during my work on this thesis. I really appreciate good conditions and support I had in the research group BioDat.

This work would not be possible without cooperation with several stem cell donor registries around the world that have provided anonymous data for test purposes, advices and/or feedback: the Czech Stem Cells Registry (*Mgr. Marie Kuříková, MUDr. Lenka Záhlavová*), the Czech National Marrow Donor Registry (*MUDr. Pavel Jindra*), the Slovak National Bone Marrow Donor Registry (*MUDr. Mária Kušíková, Denisa Stemnická*), the Finnish Bone Marrow Donor Registry (*Matti Korhonen, MD, PhD, Anne Arvola*), the ALF Marrow Donor Registry in Poland (*Leszek Kauc*, *Monika Sankowska*), the Tobias Registry in Sweden (*Mats Bengtsson*, *Bert Svensson, Rosa Hellström*), the South African Bone Marrow Registry (*prof. Ernette Du Toit*, *Veronica Borrill*, *Terry Schlaphoff*), the Ezer Mizion Bone Marrow Donor Registry (*Nira Shriki*), the Cyprus Bone Marrow Donor Registry (*Dr. Paul Costeas*, *Anita Koumouli*), the Hungarian Bone Marrow Donor Registry (*Dr. Rajczy Katalin*), the British Bone Marrow Registry (*Dr. Ann Green*), the Marrow Donor Program Belgium (*Hildegarde Broos*), the Netcord Foundation (*Dr. Etienne Baudoux*), the Nigerian Bone Marrow Donor Registry (*Seun Adebiyi*) and others.

Many thanks also to my colleagues with whom I consulted my work or from whom I got inspiration: *Carlheinz R. Müller MD PhD, Werner Bochtler PhD* and *Hans-Peter Eberhard PhD* from the Zentrales Knochenmarkspender-Register Deutschland (Germany), *Machteld Oudshoorn* and *Henk van der Zanden, MSc* from the Europdonor Foundation (Netherlands), *Martin Maiers, BSc PhD, Abeer Madbouly, Loren Gragert* and *Neil Smeby* from the National Marrow Donor Program (USA), *Pierre-Antoine Gourraud* from the University of California San Francisco (USA), *Steven Marsh, BSc PhD ARCS* from the Anthony Nolan Trust (Great Britain), *Univ. Prof. Dr. Agathe Rosenmayr* from the Austrian Bone Marrow Donors (Austria), *Amar Baouz* from the France Greffe de Moelle (France), *Colette Raffoux* from the IRGHET International Research Group on Hematopoietic stem cells Transplantation (France), *Julia Pingel*, *Jan Hofmann* and *Alexander Schmidt* from the DKMS (Germany), *MUDr. Michal Pročke, Karel Peyerl, Lucie La Mar* and other co-workers (Czech republic).

And last but not least, I would like to thank my *wife Anna*, for her support, tolerance and great conditions during my work on the thesis.

This research has been supported by the research program No. MSM 6840770012 "Transdisciplinary Research in the Area of Biomedical Engineering II" of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of the Czech Republic.

Abstract

The most important factor in the successful outcome of the hematopoietic stem cell transplantation is that a patient and a donor are matched for the Human Leukocyte Antigens (HLA). Mismatching within HLA alleles (antigens) between a recipient and a donor increases the incidence and severity of an alloreactive immune response. Because of financial and technological limits, HLA data of donors are not complete, so we have to deal with fuzzy information. Therefore selection of the potentially best donor is not an easy task. Information and communication technologies play a key role in the donor search process in international registries of volunteer donors.

This work focuses on the development of a modern search algorithm, one of the major challenges for donor registry computer systems. Our algorithm uses probabilistic matching that predicts, for each donor, the probability to be HLA allele identical to the patient.

To achieve this goal, we have estimated HLA haplotype frequencies (population genetics models) for several populations, studied properties and reliability of these models, run simulations and validated the overall system.

Abstrakt

Úspěch transplantace krvetvorných buněk je nejvíce závislý na HLA genetické shodě mezi pacientem a dárcem. Případné neshody HLA alel (nebo antigenů) zvyšují riziko a závažnost selhání transplantace. Kvůli finančním a technologickým omezením, registry dárců nemají kompletní HLA data o všech dárcích, takže nemáme k dispozici přesné informace. Díky tomu není lehké vybrat nejvhodnějšího dárce. Informační a komunikační technologie hrají důležitou roli při hledání celosvětově nejlepšího dárce.

Tato práce se zaměřuje na vývoj moderního vyhledávacího algoritmu, což je klíčová funkce počítačového systému registrů dárců krvetvorných buněk. Náš algoritmus používá pravděpodobnostní přístup, který pro každého dárce spočítá pravděpodobnost, s jakou tento dárce bude HLA shodný s pacientem.

Abychom dosáhli tohoto cíle, tak jsme spočítali HLA haplotypové frekvence několika populací a vytvořili tak populační modely. Dále jsme studovali vlastnosti těchto modelů, jejich spolehlivost, provedli jsme simulace a nakonec jsme validovali celý systém.

Content

Lists

List of abbreviations

AF	Allele Frequencies
AB	HLA-A and HLA-B loci; “AB donor” is an individual with HLA typing results on (at least) these two loci, but without typing on HLA-DRB1 locus
ABDR	HLA-A, HLA-B and HLA-DRB1 loci; “ABDR donor” is an individual with typing on at least these three loci
ABCDRDQ	HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DQB1 loci on at least these three loci
AUD	Adult Unrelated Donor
BMDW	Bone Marrow Donors Worldwide
CB	Cord Blood
CBU	Cord Blood Unit
CS	BMDW abbreviation of the Czech Stem Cells Registry
CS2	BMDW abbreviation of the Czech National Marrow Donor Registry
CSCB	BMDW abbreviation of the Cord Blood Bank Czech Republic
CSCR	Czech Stem Cells Registry
CT	Confirmatory Typing
DKMS	Deutsche Knochenmarkspenderdatei
DNA	Desoxyribonucleic Acid
DR	Donor Registries
EM	Expectation-Maximization
H	BMDW abbreviation of the Hungarian Bone Marrow Donor Registry
HF	Haplotype Frequencies
HFE	Haplotype Frequeny Estimation/Estimates
HLA	Human Leukocyte Antigens
HR	High resolution
HWE	Hardy-Weinberg-Equilibrium
IR	Intermediate resolution
IT	Information Technology
LD	Linkage Disequilibrium
LR	Low resolution
MHC	Major Histocompatibility Complex
ML	Maximum-Likelihood
NMDP	National Marrow Donor Program
SK	BMDW abbreviation of the Slovak National Bone Marrow Donor Registry
SKCB	BMDW abbreviation of the Eurocord Slovakia (SRPKB)
WMDA	World Marrow Donor Association
ZKRD	Zentrales Knochenmarkspender-Register Deutschland

List of mathematical symbols

c_j	The number of genotypes leading to the *j*-th phenotype, see (1)
$\tilde{c}_j$	The number of all possible genotypes that could lead to phenotype *j*, considering all ambiquities
s_j	The number of heterozygous loci in the *j*-th phenotype, see (1)
h_i	The haplotype *i*, see (26)
l_i	Locus *i*
a_i	Allele or antigen *i*
p_i	Frequency of the haplotype *i*
P_j	The probability of the *j*-the phenotype, see **(2)**
$h_k h_l$	Two haplotype forming the genotype, see (3)
$L(\ )$	The likelihood of the haplotype frequencies, see (6)
I_F	Similarity index, see (22)
$T()$	Haplotype type, see (27)
$\lfloor x \rfloor$	floor function; largest integer not greater than *x*
lg()	Common logarithm; $\log_{10}(\ )$, mathematical symbol by ISO 31-11 standard
$R(\ )$	Ambiguity rank of the dataset, see (25)
U	Logarithmic Score Function, see (34)

List of figures

List of tables

1. Introduction

Hematopoietic stem cell transplantation (HSCT) [1] (commonly referred to as bone marrow transplantation) is a medical procedure in the field of hematology and oncology. HSCT is the treatment of choice for people with hematopoietic malignancies (e.g. leukemia), bone marrow failure and certain types of cancer (e.g. lymphoma) which result in a compromised immune system. The principle is that intravenous infusion of stem cells collected from donor bone marrow, peripheral blood or umbilical cord blood is used to replace the hematopoietic functions of a patient with these conditions. The most important factor in the successful outcome of HSCT is that the patient and donor are matched for the Human Leukocyte Antigens (HLA) [2]. Mismatching within HLA alleles (antigens) between a recipient and a donor increases the incidence and severity of an alloreactive immune response when transplanting hematopoietic stem cells. The level of the matching required varies with the source of stem cells used for HSCT.

In most cases (in Europe) patients have no suitable HLA matched donor within their family, so physicians must activate a 'donor search process' by interacting with national and international donor registries who will search their databases for adult unrelated donors or cord blood units (CBU) [3].

Information and communication technologies play a key role in the donor search process in donor registries both nationally and internationally. One of the major challenges for the donor registry computer systems is the development of a reliable search algorithm [4]. Our previous work [5] had focused on design and implementation of combinatorial approach. In principle, the algorithm compares patient with donors by counting all known and visible HLA mismatches. Implementations of such algorithms are commonly used, including the Bone Marrow Donors Worldwide computer system (BMDW) [6]. In 2011, the Information Technology (IT) working group of the World Marrow Donor Association (WMDA) has issued recommendations [7] that summarize current knowledge about implementation of this approach.

Nowadays, there are more than 20 million stem cell donors and cord blood units available worldwide [6]. Due to character of HLA system, history of HLA typing techniques and limitation of resources, we do not have full information about HLA types of these donors. Search coordinators often see very long lists of partly HLA matching, partly HLA typed donors and they have to guess which donors should be selected for further HLA typing or testing. Limitation of resources (time and money) and risk of detours makes their choice tricky. An 'expert system' that would better lead the coordinator is needed to make faster and more accurate decisions.

1.1 Goals of the work

This work implements the **probabilistic matching** method that can predict donor data even if they are invisible or fuzzy at the moment. The main motivation [8] of the probabilistic matching is to help search coordinators to:

- identify easy, difficult and (almost) futile searches
- predict the level of patient-donor matching realistically achievable
- speed up the donor search by choosing the most promising candidates and avoiding detours
- make ultra-urgent searches feasible in spite of ambiguous or missing HLA data

In our previous work, we had used the **combinatorial matching** method that observes visible donor data and analyzes them, especially HLA mismatches.

Currently, probabilistic matching systems are used in daily operations only by the biggest registries in the World. The Zentrales Knochenmarkspender-Register Deutschland (ZKRD) has pioneered this innovative technology and developed the OptiMatch® system [9] and the National Marrow Donor Program (NMDP) uses HapLogic™ system [10]. These registries have invested huge efforts into the development of the systems but their internals are not published. However, even if they publish them or provide them to smaller registries for free, it is not clear if others can use them and approximate local population by German or American models and what would be the reliability of such predictions.

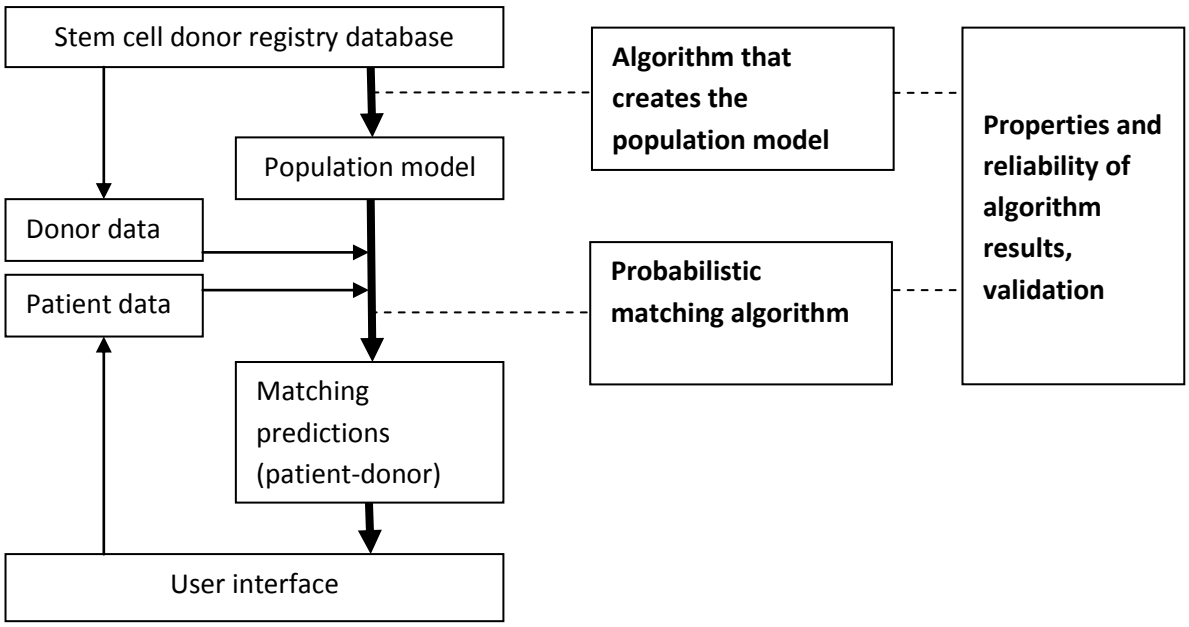


Figure 1: Probabilistic matching system - structure of the work

We will systematically implement new probabilistic matching system, see Figure 1. In order to do it, we have to answer these questions and satisfy the underlying goals:

- **How can we design and implement algorithm that creates population model?** The population model will be represented by HLA haplotype frequencies estimates (HFE) and we will focus on the problem of estimation of HLA gene and haplotype frequencies of a human population. For this purpose, we want to use datasets of registries of unrelated hematopoietic stem cells donors that are the biggest available databases of HLA data for most of the populations in the world. These databases have been built and maintained for more than 10 years. HLA typing of d onors were determined by different typing techniques and a lot of data is missing. The combination with the complexity of HLA system and the size of these databases (up to millions of individuals) brings the problem to another level. Therefore the estimation of gene and haplotype frequencies in such conditions is a real challenge.

- **What are the properties and reliability of the model (HFE) in general?** Since we will approximate the local population by its stem cell donor registry datasets of different sizes and structures, we have to understand quality of the result. We need to study the dependency on the size of the population, genetic properties of the population, size of the sample (registry) and resolution of the donor typing. We are also limited by computational time. In practice, we have to deal with all these factors together.
- **How can we design and implement the probabilistic matching algorithm?** We are looking for a solution even to countries for which it is not possible to create own model. The algorithm must be able to handle all types of cases, patient-donor pairs, even if the patient or the donor does not fit to our model (e.g. other ethnic). It has to be fast enough and give reliable results.
- **How can we validate the whole system? Can we apply it for all registries and populations?** The whole system must be validated before its use. In some countries we can use historical data for validation, but in most countries, we don't have enough data. Therefore we need to find novel method for validation, using simulation.

HFEs are useful not only to support search for unrelated donors, but could be used in other applications, we will present some of them.

1.2 Structure of the work

The work is organized as follows: chapter 2 gives introduction to the HLA system, unrelated haematopoietic stem cell transplantation, stem cell donor registries and selection of unrelated stem cell donors. Chapter 3 focuses on the overview of computer algorithms in the search for unrelated stem cell donors.

Chapter 4 is the overview of possible methods of HFE with focus on maximum likelihood function and its solution by the iterative Expectation-Maximalization (EM) algorithm. A method that can verify reliability of estimates is presented.

Main part of the work starts by Chapter 5 that discusses the implementation of the HFE algorithm and its usage on datasets of stem cell donor registries – challenges, pitfalls and possible solutions. Chapter 6 gives new methods for testing of reliability of the HFE algorithm with stem cell donor registry datasets. Chapter 7 presents real results, using methods of chapters 5 and 6.

Chapter 8 presents some applications of HFE, but we focus on the prediction of the HLA match in the chapter 9. Top-down design of the algorithm is described. We compare our approach with other implementations in the world (ZKRD, NMDP).

Chapter 10 describes methods of validation of the HLA matching prediction algorithm and our results.

Chapter 11 shows application of the algorithms and tools in daily operation of stem cell donor registries.

Chapters 12 and 13 conclude the work.

2. HLA and haematopoietic stem cell transplantation

This chapter gives introduction to the HLA system, unrelated haematopoietic stem cell transplantation, stem cell donor registries and selection of unrelated stem cell donors.

2.1 Basic terms

In the following text, we will use the terminology with the following meaning:

- **Locus** – gene; HLA locus, e.g. DRB1
- **Antigen** - one of the alternative versions of a gene at a given location (locus) along a chromosome; substances that are recognized by the immune system and induce an immune reaction.
- **Allele** - one of the alternative versions of a gene at a given location (locus) along a chromosome; an individual inherits two alleles for each gene, one from each parent. If the two alleles are the same, the individual is **homozygous** for that gene. If the alleles are different, the individual is **heterozygous**. [9]
- **Haplotype** – set of specific loci with antigen/allele designations. From each parent, a haplotype is inherited as unit [10].
- **Genotype** – particular combination of two multi-locus haplotypes [10].
- **Phenotype** – multi-locus genotype whose haplotype phase is unknown a priori [10].
- **Linkage disequilibrium** – association of alleles at two or more loci, combinations of alleles in a population that is more or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies [11].

2.2 HLA system

Human leukocyte antigen (HLA) genes are located on the short arm of chromosome 6. HLA genes are extremely polymorphic and play critical role in immune recognition and response. Each individual has two sets of genes; consequently, the combination of HLA markers of each individual is rare or almost unique in various populations.

Polymorphism is beneficial for population studies, because it allows determination of genetic affinities among different populations. Haplotype studies are also important in complex research of genetic diseases, when we want to know association of diseases or risks with specific haplotypes.

2.2.1 Human Leukocyte Antigen

The major histocompatibility complex (MHC) [12] is a large genomic region or gene family found in most vertebrates. It is the most gene-dense region of the mammalian genome and plays an important role in the immune system, autoimmunity, and reproductive success. MHC genes are some of **the most genetically variable** coding genes in mammals. The proteins encoded by the MHC are expressed on the surface of cells in all jawed vertebrates, and display fragments of molecules from invading microbes or dysfunctional cells (e.g. tumor cells) to a particular type of white blood cell called a T cell that has the capacity to kill or co-ordinate the killing of the microbe, infected cell or malfunctioning cell.

The best-known genes in the MHC region are the subset that encodes cell-surface antigen-presenting proteins [12]. In humans, these genes are referred to as human leukocyte antigen (HLA) genes.

The most intensely-studied HLA genes (also called loci, sg. locus) are the nine so-called classical MHC genes: HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, and HLA-DRB1. In humans, the MHC is divided into three regions: Class I, II, and III. The A, B, and C genes belong to MHC class I, whereas the six D genes belong to class II.

Besides being scrutinized by immunologists for its pivotal role in the immune system, the MHC has also attracted the attention of many evolutionary biologists, due to the **high levels of allelic diversity** found within many of its genes. Indeed, much theory has been devoted to explaining why this particular region of the genome harbors so much diversity, especially in light of its immunological importance.

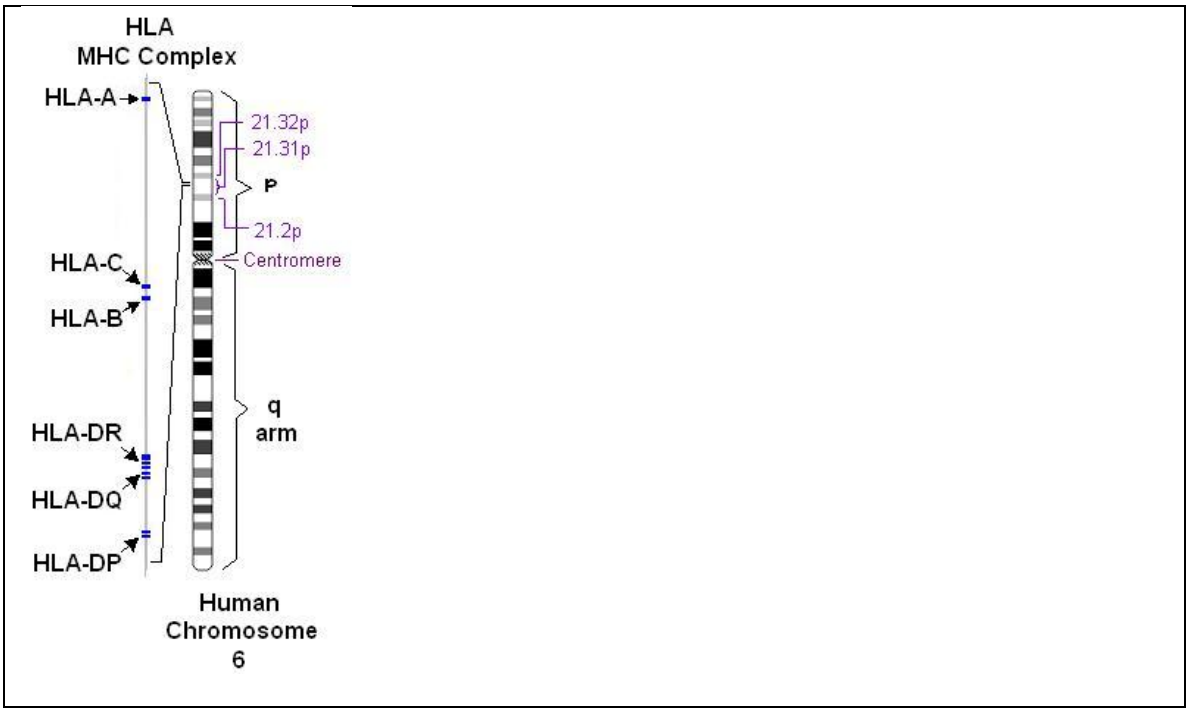


Figure 2: HLA complex on human chromosome 6 [13]

2.2.2 Nomenclature of HLA System

Nomenclature of HLA system is under responsibility of the WHO Nomenclature Committee [14]. Stem cell donor registries also follow the WMDA standards and recommendations [15]. Each HLA allele name has unique two, three or four field names. Fields are separated by colon (":"). The length of the allele designation depends on the sequence of the allele and that of its nearest relative. All alleles receive at least a two field name, three and four field names are only assigned when necessary.

The first field (number) describes the type, which often corresponds to the serological antigen carried by an allotype. The second field (number) is used to list the subtypes, numbers being assigned in the order in which DNA sequences have been determined. Alleles whose numbers differ in the first two fields must differ in one or more nucleotide substitutions that change the amino acid sequence of the encoded protein [16].

Full code	Abbreviation (unofficial)	Description
HLA-A	A	A means locus name (HLA gene)
HLA-A2	A2	Serological antigen A2. Result of a serology typing method.
HLA-A23	A23(9) or A23	A23 is a **split** serological antigen of **broad** serological antigen A9. Result of serology typing method.
HLA-A*02:XX	A*02:XX or A*02	Group of alleles (subtypes of antigen A2). **Low resolution (LR)** result of a DNA typing method.
HLA-A*02:01	A*02:01	Allele A*02:01. An example of the **high resolution (HR)** results of a DNA typing method.
HLA-A*02:01/02:02	A*02:01/02:02 or A*02:AB	Group of two alleles A*02:01 and A*02:02. A*02:AB is the **NMDP multiple allele code** that represents the group. An example of **intermediate resolution (IR)** results of DNA typing methods.

Table 1: Nomenclature of HLA System

In this work, we will use both official and abbreviated nomenclature.

2.2.3 Resolution of the HLA typing

Based on the quality of HLA typing we can get HLA typing results of five different levels:

- Broad serology antigen
- Split serology antigen
- Low resolution (LR) DNA typing
- Intermediate resolution (IR) DNA typing
- High resolution (HR) DNA typing.

Broad and split serology antigen results are based on serology typing methods, while others are based on molecular biology typing methods.

Low resolution means the identification for the first two digits of the HLA nomenclature, i.e. all alleles with the same first field. **Intermediate resolution** means selection of at least two allele codes, all belonging to the same serological antigen groups. **High resolution** typically means one allele designation with two or more fields (at least four digits). In some countries (e.g. Germany), multiple allele codes are still considered as high resolution, if all the alleles covered are identical over exons 2 and 3 for HLA class I or over all of exon 2 for HLA class II.

Nomenclature of locus names differs, if we speak about serology typing results or molecular biology (DNA) typing results.

Serology	HLA-A	HLA-B	HLA-C	HLA-DR	HLA-DQ
DNA	HLA-A*	HLA-B*	HLA-C*	HLA-DRB1	HLA-DQB1

Table 2: Nomenclature of locus names of different typing methods

In general, relation between typing results of different levels of typing of one individual is quite complex (see Table 3):

- Broad serology antigen always represents a group of split serology antigens, so their relation is 1:n. E.g. broad serology antigen A9 represents group {A23, A24}.
- LR DNA code represents a group of HR resolution DNA codes and every HR DNA code belongs exactly to one LR DNA code. I.e. A*01:XX represents group {A*01:01, A*01:01:01, A*01:01:02, …, A*01:02, A*01:03, …, A*01:20, …}.
- Other relations are more complicated (m:n). IR DNA codes (also called NMDP codes or multiple-allele-codes) represent a group of HR DNA codes. I.e. A*01:AAXP represents group {A*01:02, A*01:08, A*01:14}. But a HR DNA code can belong to many IR DNA codes. I.e. A*01:01 belongs to A*01:AB, A*01:AC, A*01:AAJ, etc.
- HR DNA typing result DRB1*11:16 can have corresponding split serology antigen DR11(5) or DR13(6) [17], but DR11(5) have many corresponding DNA typing results (DRB1*11:01, DRB1*11:02, …, DRB1*11:16, …, DRB1*11:60).
- Other relations (m:n) are derived from previous facts.

Resolution	**Split**	**LR**	**IR**	**HR**
Broad	1:n	m:n	m:n	m:n
Split		m:n	m:n	m:n
LR			m:n	1:n
IR				m:n
HR				

Table 3: Relation between levels of HLA typing (m:n – many to many, 1:n – one to many)

2.2.4 Examples of HLA typing results

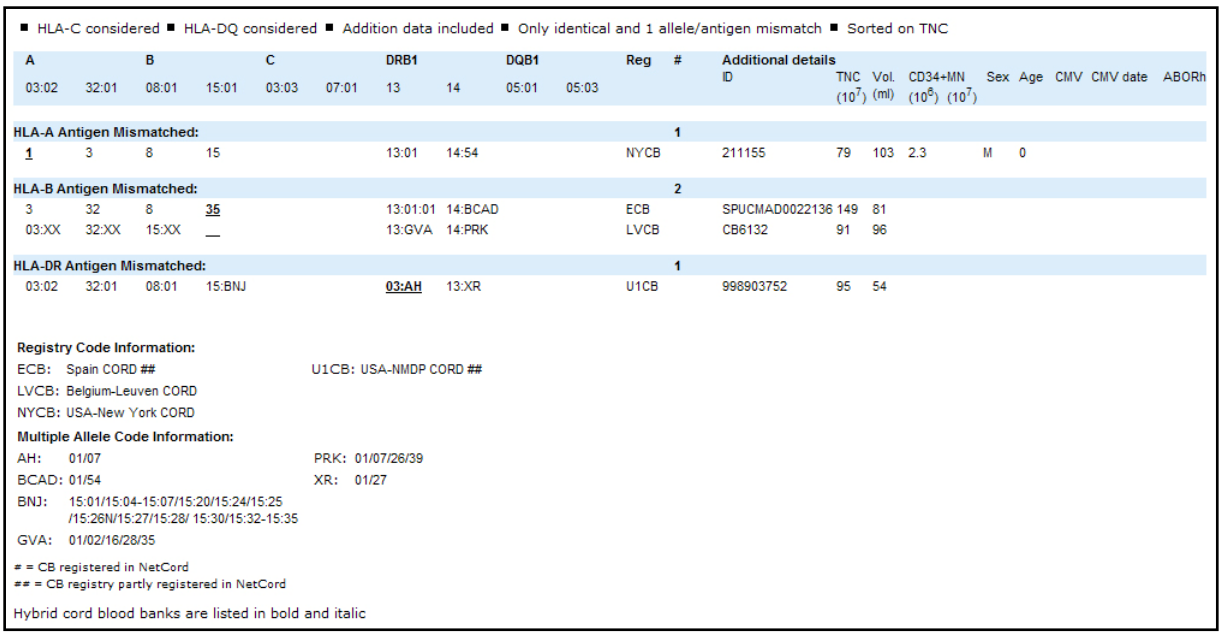
Individual A	**Individual B**	**Individual C**
HLA-A*01:01, 26:01	HLA-A*03:01,32:BYJT	HLA-A1,9
HLA-B*38:01, 57:01	HLA-B*35:01,38:01	HLA-B17,40
HLA-C*06:02, 12:03	HLA-C*04:BRXU,12:03	
HLA-DRB1*04:02,15:01	HLA-DRB1*01:01,13:03	
HLA-DQB1*03:02,06:02	HLA-DQB1*03:01,05:01	
(tested by molecular biology typing methods)	(tested by molecular biology typing methods)	(tested by serology typing methods)

Table 4: Examples of HLA typing results

2.3 Unrelated donor selection process

Search for unrelated stem cell donors typically follows these steps [18] [19]:

1. Patient HLA typing is determined. At least HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DQB1 loci are tested. Sometimes also HLA-DRB3/4/5. Patient should be typed at intermediate or high resolution.
2. Search coordinator runs the search algorithm in national and international registries.



■ HLA-C considered ■ HLA-DQ considered ■ Addition data included ■ Only identical and 1 allele/antigen mismatch ■ Sorted on TNC

A		B		C		DRB1		DQB1		Reg	#	Additional details ID	TNC (10^7)	Vol. (ml)	CD34+MN (10^6)	(10^7)	Sex	Age	CMV	CMV date	ABORh
03:02	32:01	08:01	15:01	03:03	07:01	13	14	05:01	05:03												
HLA-A Antigen Mismatched:											1										
1	3	8	15			13:01	14:54			NYCB		211155	79	103	2.3		M	0			
HLA-B Antigen Mismatched:											2										
3	32	8	35			13:01:01	14:BCAD			ECB		SPUCMAD0022136	149	81							
03:XX	32:XX	15:XX	__			13:GVA	14:PRK			LVCB		CB6132	91	96							
HLA-DR Antigen Mismatched:											1										
03:02	32:01	08:01	15:BNJ			03:AH	13:XR			U1CB		998903752	95	54							

Registry Code Information:

ECB: Spain CORD ##
U1CB: USA-NMDP CORD ##
LVCB: Belgium-Leuven CORD
NYCB: USA-New York CORD

Multiple Allele Code Information:

AH: 01/07
PRK: 01/07/26/39
BCAD: 01/54
XR: 01/27
BNJ: 15:01/15:04-15:07/15:20/15:24/15:25 /15:26N/15:27/15:28/ 15:30/15:32-15:35
GVA: 01/02/16/28/35

= CB registered in NetCord
= CB registry partly registered in NetCord

Hybrid cord blood banks are listed in bold and italic

Figure 3: Example of donor search result [6]

3. List of potential donors (see Figure 3) typically contains a lot of gaps (missing HLA typing results) or HLA ambiguities. Based on transplant protocol, consultation with transplant centres and local experience (or expert system predictions!), the search coordinator can select several (3-10+) potential donors for additional typing. These tests could be done by local or remote laboratories. Number depends on frequency of patient's alleles & haplotypes (if rare, more donors are selected), clinical urgency (more urgent case requires simultaneous testing of several potential donors) and may be also limited also by patient's financial situation (i.e. limited funding by healthcare insurance company) – requested services have to be paid by the applicant (hundreds of Euros).
4. Some potential donors will be unavailable, so missing results will never be obtained. Contacting the donor, logistics of the blood sample and execution of the requested tests will take several weeks.
5. Requested donor HLA typing results could show mismatch with the patient, so next rounds of additional typing procedure may be initiated. Common patient HLA types can usually find donor on first match run, less common may require a more sophisticated search using HLA expert help to prioritize donors/cords. Unfortunately, some searches are finished without finding a match. Then, other solution has to be found – physician has to change transplant protocol (e.g. mismatched donor or cord blood unit) or select non-transplant treatment.

6. If a suitable donor is found, the transplant centre – donor centre handshaking process is started (formal work-up requests, donor is examined, etc.), that may end up with the transplant operation.

3. Computer algorithms in the search for unrelated stem cell donors

This chapter gives an overview of computer algorithms in the search for unrelated stem cell donors.

3.1 Search algorithm

The purpose of the donor search algorithm is to find and present a selected list of potential donors and/or CBUs, in which the most likely an optimal stem cell source for the patient are sorted to the top of the list [7]. Selection and sorting criteria are based on HLA compatibility and may also take into consideration secondary preference criteria, such as CMV antibody status, gender and age.

Basic requirements for the search system used by stem cell donor registries are:

- **Deterministic** behavior that ensures the same results with the same input. This means, the algorithm has to reproduce exact decisions at every step.
- **Clear ranking order** results.
- **Exhaustive** - all donors available for transplant in the source database should be included in the search algorithm. Exceptions must be clearly indicated to the end-user. For example some algorithms exclude donors that are typed only at HLA-A and HLA-B.
- **Scalable** - the system should be able to handle databases of varying size and type.
- **Fast** – search algorithms are also used in user-interactive systems, so the results should be received in seconds.
- **Configurable** – search coordinator must be able to define patient-donor HLA match criteria and secondary preference criteria (CMV status, gender, age).
- **Consistently matched** [20] - The data presented should be uniformly matched as a set for a given instance of a patient search. Different primary algorithms or matching criteria shall not be used within a single patient search.

The search algorithm is usually implemented as the key component of the stem cell donor registry software system. It has several inputs and a single output. The following input data are essential:

- Patient's data: HLA type (minimum HLA-A, HLA-B and HLA-DRB1 typing).
- Patient's match criteria (position and number of allowable mismatches)
- Database of adult unrelated and cord blood units (CBUs) (optional)
- HLA nomenclature code-lists
- Allele and haplotype frequencies (optional, depending on type of the algorithm)

The algorithm itself usually follows these steps:

a. **Pre-processing**: fast pre-selection of donors based on predetermined internal indices
b. **Processing:** comparison of every (pre-selected) donor with the patient, calculation of match grades, matching probabilities and filtering
c. **Post-processing:** linking corresponding donor/CBU details.

The search output, which returns a sorted list of potential donors and CBUs can be presented either in the user interface, on a printed report or transmitted to other systems (EMDIS). The presentation

output may be calculated within in the search engine software. e.g. it is common practice to highlight patient-donor HLA mismatches. As well as match grade and matching probability this may require additional data extraction from internal information calculated during the execution of the algorithm.

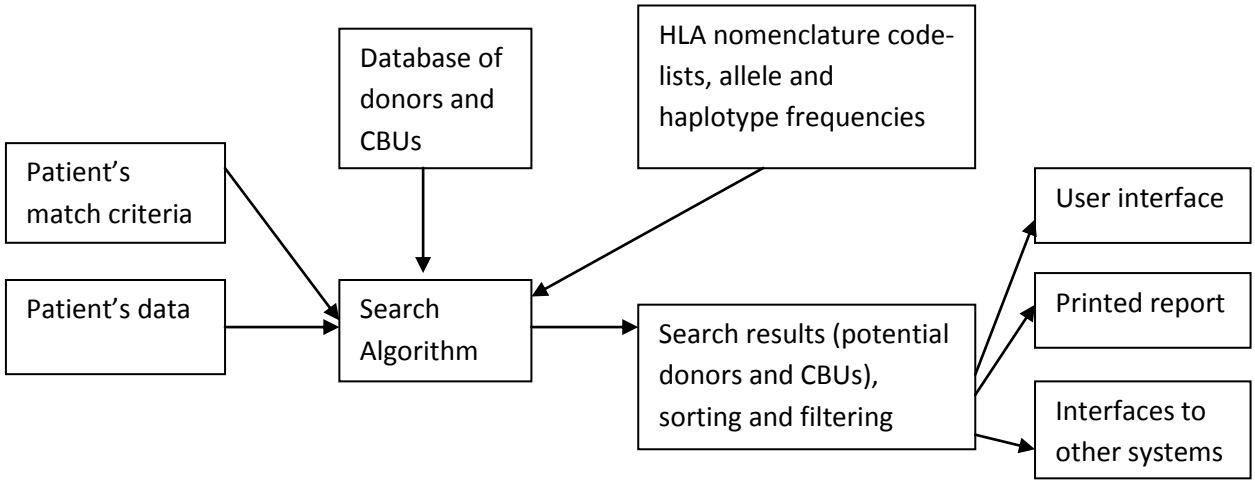


Figure 4: Basic concept of the donor search algorithm

3.1.1 Patient's data

Patient's HLA typing data must correspond to the valid HLA nomenclature and WMDA guidelines [15] and should be typed at the highest possible resolution, i.e. least intermediate resolution. According to some algorithms may return unexpected search results, if low resolution HLA typing data is provided.

Example: B*35:76 has no mapping to 'Unambiguous Serology' [16], but is mapped to 'Possible Serology' B35 and B22. B22 is the broad HLA code with splits B54, B55 and B56. Therefore a patient carrying B*35:XX is a potential match with a donor carrying B*56:XX. Such a result is likely to be confusing for healthcare professionals. This problem would not appear if the patient was typed at higher resolution (the B*35:76 allele is excluded). An alternative solution would be to apply an exceptions or filter by application of additional criteria, e.g. matching probabilities with threshold (it is very unlikely the B*35:XX will become B*35:76).

3.1.2 Patient's match criteria

Some algorithms have hard-coded or fixed match criteria, but more sophisticated search algorithms allow users to define matching preferences for each individual search. EMDIS Matching Preferences [21] define these criteria:

- Counting method for mismatches: count graft-versus-host (GvH) mismatches only or host-versus-graft (HvG) mismatches only
- Maximum number of antigen/allele mismatches for adult donors
- Maximum number of antigen/allele mismatches for CBUs
- Maximum number of antigen/allele mismatches at loci A/A*, B/B*, C/C*, DR/DRB1*, DQ/DQB1*
- Maximum age of the donor, gender matching, CMV matching

3.1.3 Database of donors and cord blood units (CBUs)

Database of unrelated stem cell donors and CBUs should correspond to these requirements [20]:

- **Current** - The data used by the algorithm should be up to date.
- **Detailed** - The data presented should contain all relevant fields to the determination of match. The set of data elements should be consistent amongst the registry community.
- **Integrated** - The data presented should be considered as a set and should be available to the matching party as a part of a singular search event.
- **Recognizable** - The data presented should uniquely reference individual sources using the identifier that is directly associated with the donor/CBU or would appear on any biological samples associated with the product.
- **Comprehensive** - The data presented should represent a consolidated view of the inventory. Uniform depth of access to all donors is needed.

Good implementation of the donor database is essential for acceptable performance of the search algorithm. Not all database structures of HLA applications are suitable as the data source for the algorithm.

Many small to middle size registries are co-located in a single centre with the HLA typing laboratory and there is a need for data integration of these two departments. It may seem the registry system stores and manages the HLA typing results in the same way as the HLA laboratory information management system (LIMS), and some registries have implemented such data storage. It is a mistake to use these in search algorithms. The main differences between registry database and HLA LIMS database are:

- The registry system needs fast access to the most current and comprehensive HLA typing results, which does not always mean the last test typing. This may be combination of multiple tests performed in the past by multiple typing techniques. The registry system always needs access to the full set of all loci that should be stored at one place, while the HLA lab system order includes only requested tests and loci, so HLA typing results of an individual may be spread in multiple typing orders.
- When the HLA lab supervisor approves the order results, it cannot be changed in the lab system. However, the registry system has to keep historical HLA typing results up-to-date according to the latest HLA nomenclature, so it needs to update them (deleted and renamed alleles, new HLA nomenclature).

Database of donors/CBUs can simply be organized in a single relational database table. Even this may be problematic. A logical database approach is to organize HLA code-lists in separated tables (multiple-allele-codes, alleles, antigens and their relations) and define master-detail relationship between donor data and HLA codes. These systems have been implemented in some registries. The storage of donor record is using only primary keys of HLA codes (as foreign keys). The disadvantage of the master-detail storage is that retrieval of donor's HLA typing is inefficient. Often the solution for data retrieval in such a structure is cumbersome, because the database system has to join data (database natural join) from tens of tables .The advantage is easy manipulation with the properties

of HLA codes or even the renaming of HLA allele codes. But such operations are much less common, compared to data retrieval.

3.1.4 HLA nomenclature code-lists

In all cases, the algorithm has to recognize the description of HLA typing codes (e.g. multiple-allele-codes) and relations between HLA codes, especially DNA to serology mapping. Some algorithms even use antigen recognition site matching, amino acid sequences or nucleotide sequences. It is recommended that code-lists and code attributes are downloaded from specialist reference web sites [16] and [22].

Donors have been typed by various different typing techniques and many of them are registered with HLA serological assignments. The database of donors could be pre-processed, so all interpretations and mapping of HLA codes could be saved in advance, but generally, the patient's HLA type is known only at the time of the search, so HLA nomenclature code-lists are needed. Of some concern is that a minority of patients are still typed only by serologic typing techniques! This means that search algorithms must be capable of using these in the search process.

3.2 Pre-processing

Several variants of search algorithms are being used by stem cell donor registries. Selection of the algorithm is influenced by available resources, size of the donor database, availability of haplotype frequencies of the supported population(s), etc. We will discuss commonly used search algorithms.

I. Simple pre-selection

The goal of the algorithm is to find potential donors for one patient. The phenotype of the patient is compared with all donors phenotypes in the donor registry database that are 'available' for transplantation purposes (simple pre-selection).

```
For every donor D in the database
      Count Match Grade (patient P – donor D)
      If the Match Grade is acceptable, store
            data of donor D in the list of
            potential donors of patient P
```

This kind of algorithm can be used only for small to middle sized registries. Implementation enhancements can help to improve this situation. For example, increasing current capacities of server memories allows caching of all donors in the random access memory (RAM) of the server. The advantage of this algorithm is mainly in its simplicity and simple validation process. It also has very straightforward implementation of distributed or parallel computing. The drawback is the speed and memory limitation, especially where donor database is growing

This algorithm could be extended to multiple patient searches [5] that might be useful, for example, for EMDIS repeat searches [21], when search results from several thousands of donors have to be generated and compared with previous results. Again, the list of all patients could be cached in the server memory with one additional loop.

```
For every donor D in the database
      For every patient P in the database
            Count Match Grade (patient P – donor D)
            If the Match Grade is acceptable, store
                  data of donor D in the list of
                  potential donors of patient P
```

II. Search determinants

Databases from Registries and cord blood banks store the HLA types in many formats depending whether typing was by serology or by DNA-based methods. Registries must take these different assignments to create a match algorithm to search for a patient. This comparison is usually facilitated by the conversion of phenotypes to "search determinants" prior to development matching algorithms.

The phenotype of the patient/donor is mapped to 'Search Determinants' (SD) [23] [24]. The SD is a data record, based on serological antigens, corresponding to the original HLA phenotype. For example, it might be a group of six HLA serologic-based assignments – three pairs for HLA-A, HLA-B and HLA-DRB1 loci. An individual can have multiple SDs. SDs are used as an index to select the set of matching phenotypes. Then, more precise match grades are counted and the list of donors is filtered.

The main application of SDs is the speeding up of the match process by using SDs as keys values in conjunction with a database and a matching algorithm [25]. The main disadvantage is the need for regular checks and updates of SDs of all donors in the database, due to changes of donor data, HLA nomenclature updates and changes in the "DNA to serology" mapping. There are particular problems where there is no serological equivalent for a DNA allele.

III. DNA matching only

The National Marrow Donor Program (NMDP) in the United States has developed an algorithm [26] that does not use SDs for the initial matching step as this is done by directly comparing patient DNA type to donor DNA type. The algorithm is able to account for all serologic typing possibilities with the use of a special table called the "Serology to DNA Allele Table".

3.3 Processing

The key element of the processing step of the algorithm is the 'match grade function' that can compare data (HLA, ethnic group) of two individuals (usually patient and donor) and return their match grade and/or matching probabilities. The threshold function then filters out donors that do not match patient's match criteria.

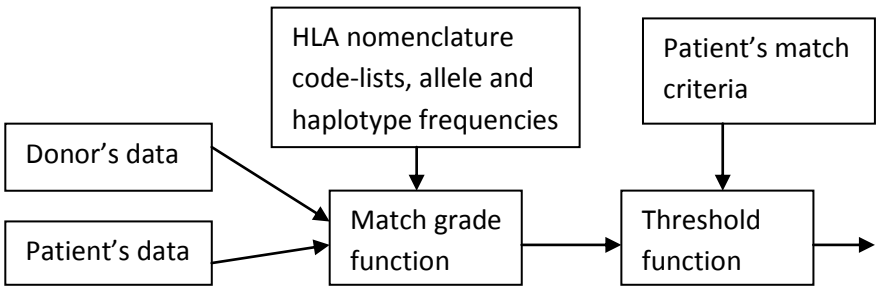


Figure 5: Match grade function

Original versions of matching algorithms compared HLA typing only at HLA-A and HLA-B loci. DNA typing was not performed. Later generations added other loci, especially HLA-DRB1, but also HLA-C and HLA-DQB1. Today, some algorithms even use HLA-DRB3/4/5, HLA-DPB1 and other loci.

Earlier versions of matching algorithms also used only serological assignments; DNA typing either did not exist or was not taken into account. Later versions have converted DNA typing results into serological assignments or vice versa, so the algorithm has a uniform typing technique view on all donors. Current search algorithms use DNA typing results as much as possible and switch to serology comparisons only if DNA typing is not provided or if they want to refine DNA to serology mapping.

The Information Technology (IT) Working group of the World Marrow Donor Association (WMDA) has issued two key resources that describe the correct handling of HLA data and key patient-donor matching procedures:

- Framework for the implementation of HLA matching programs in hematopoietic stem cell donor registries and cord blood banks [7]. This article gives a bottom-up approach to the design of search algorithms: comparison of individual HLA codes, then HLA single-locus phenotypes and eventually HLA multi-locus phenotypes.
- Guidelines for use of HLA nomenclature and its validation in the data exchange among hematopoietic stem cell donor registries and cord blood banks [15]

A common mistake in the design of search algorithm is the violation of the rule 2.1 of the guidelines [15]: “Laboratories must assign DNA nomenclature to results obtained using DNA-based methods and serologic nomenclature to results obtained using antibody reagents.”. Some computer systems need to permanently store serology derived results of DNA codes, usually because of simple DNA-serology matching. However, the mapping should be done automatically by the system and not by the user. Derived serology values must be clearly distinguished from real serology results obtained using antibody reagents. Where mapping has changed, the registry system has to know if stored serologic results should be updated or not. Moreover, some alleles are mapped to multiple serology equivalents and the system has to take this into account.

In addition to match grade, some information can be calculated. In these, the probability of HLA matching at the allele level based on local population haplotype frequencies in the underlying population can be calculated. Such prediction algorithm system has been developed and validated by the NMDP (HapLogic™ II) [27].

The latest, state-of-the-art versions of search algorithms (OptiMatch®, HapLogic™ III) use these probability calculations to determine the rank order of HLA matches as the main searching and sorting criteria.

3.4 Post-processing

At this stage, the system retrieves corresponding donor details of all selected donors that will be displayed in the search results. If the matching probabilities are not used as the main sorting criteria, the search system can apply them at this stage (ProMatch [28], Hap-E [29] and EasyMatch [30]).

3.5 Validation of the search algorithm

All implementations of the search algorithms need to be validated before being used. The WMDA Information Technology Working Group provides validation sets of patients and donors that are used for matching trials and comparison of results with expected outcomes [31] [7]. Algorithms that do not use simple pre-selection approach, but use more complex pre-selection, have to be validated for completeness. It is important not to miss any relevant donors in the pre-selection [7].

Validation of the processing phase, especially the match grade function, can be done by running several automated unit tests, addressing all kinds of matches and mismatches, exceptions and rare cases. Interfaces to software source code classes, modules or libraries are tested with a variety of input arguments to validate that the results that are returned are as expected [32].

4. Haplotype Frequencies Estimation

This chapter gives an overview of possible methods of HFE with focus on maximum likelihood function and its solution by the iterative Expectation-Maximalization (EM) algorithm. A method that can verify reliability of the estimates is presented.

4.1 Number of genotypes

The number of genotypes (c_j) leading to the *j*-th phenotype is a function of the number of heterozygous loci s_j:

(1)

$$c_j = \begin{cases} 2^{s_j - 1} & if \ s_j > 0 \\ 1 & if \ s_j = 0 \end{cases}$$

Example 1

Assume the following phenotype of an individual ($s_j = 3$): A1,2 B7,8 DR1,4

Then all possible genotypes are ($c_j = 2^2 = 4$):

A1 B7 DR1 A2 B8 DR4	A1 B7 DR4 A2 B8 DR1	A1 B8 DR1 A2 B7 DR4	A1 B8 DR4 A2 B7 DR1

□

Only one of these c_j genotypes is the proper one.

4.2 Problem formulation

Typing techniques allow the survey of many polymorphic loci, but do not allow distinguishing gametic phase of haplotypes. For heterozygous diploids the direct sequencing of the PCR (polymerase chain reaction) product results in the amplification of both alleles and does not allow resolving the haplotypes when the diploid individual is heterozygous at more than one locus.

The data set consists of individuals (sample of a population) and their unphased HLA typing results at one or more loci.

The goal is to find the best estimates of the haplotype frequencies in the population using only limited information included in the phenotype (unphased genotype) sample data.

4.3 Methods

The main methods of solution of the problem are:

1. Family studies – adding some additional information.
2. Remove heterozygous individuals – ignoring the problem.

3. Parsimony method – counting phase known individuals
4. Two by two tables – solution only for two loci
5. Bayesian methods
6. Maximum likelihood approach

4.3.1 Family studies

Multi-loci haplotypes can be usually determined by additional genealogical study of the individual. [35]

Family members of many individuals could not be reachable for tests. Therefore the family studies of all individuals in the data set are not possible. Moreover, to avoid redundant information and possible bias, some members of the families must be excluded from the data set, so the costs would be extremely high. This approach is not scalable for large data sets.

4.3.2 Remove heterozygous individuals

The easiest possibility would be to remove all heterozygous individual from the sample and keep only homozygous ones. Then calculate haplotype frequencies by direct counting.

This approach is problematic, because it might lead to a bias.

4.3.3 Parsimony method

Clark's algorithm [33] and its variation [34] start to examine complete homozygotes and single-locus heterozygotes and creates list of haplotypes that must be present unambiguously in the sample. If such individual does not exist, then the algorithm cannot start. Then other individuals are screened for a possible occurrence of previously recognized haplotypes. For each positive identification, the complementary haplotype is added to the list of the recognized haplotypes, and so forth. Problems of the approach are:

(a) homozygous individuals are not always present in stem cell donor registry databases or there can be only few of them;
(b) the final result depends on the order of individuals in the sample as shown in [35].
(c) in the end there could remain unresolved individuals.

4.3.4 Two by two tables

The estimation method [34] [36] counts the phenotype frequencies of each antigen in the sample for both (two) loci and uses these to calculate the linkage disequilibrium of each haplotype consisting of two alleles *i* and *j* as follows:

$$D_{ij} = \sqrt{\frac{d_{ij}}{n}} - \sqrt{\frac{b_{ij} + d_{ij}}{n} + \frac{c_{ij} + d_{ij}}{n}}$$

in which *a*, *b*, *c* and *d* are the phenotype frequencies of the +/+, +/-, -/+ and -/- combinations of the allele in each haplotype and *n* is the sum of *a*, *b*, *c* and *d*. The haplotype frequency of allele *i* from the first locus and allele *j* from the second locus is then:

$$p_{ij} = D_{ij} + g_i \times g_j$$

where g_i and g_j are gene frequencies of allele *i*, resp. *j*.

This method is computationally simple, but unfortunately it works only for two loci and gives worse results than maximum likelihood approach [34].

4.3.5 Bayesian methods

The PHASE algorithm [37] threats haplotype configuration for each unresolved individual as an unobserved random quantity and aims to evaluate their conditional distribution, given a sample of unphased data. Goal of the Bayesian framework is to approximate posterior distribution of haplotype configurations $f(\mathbf{G}|\mathbf{P})$, where

$\mathbf{G} = (\mathbf{G}_1, \ldots, \mathbf{G}_n)$ denote to unknown corresponding haplotype pairs (genotypes), *n* is number of individual in the sample and $\mathbf{P} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$ are known unphased phenotypes. The method implements Markov chain Monte Carlo (MCMC) methods (Gibbs sampling) to sample from $f(\mathbf{G}|\mathbf{P})$. It starts with random configuration $\mathbf{G}^{(0)}$, repeatedly selects unresolved individuals at random and samples from their possible haplotype configurations, assuming all other individuals to be correctly resolved. Repeating this process enough times results in an appropriate sample from $f(\mathbf{G}|\mathbf{P})$. In other words, it constructs Markov chain $\mathbf{G}^{(0)}$, $\mathbf{G}^{(1)}$, $\mathbf{G}^{(2)}$,… with stationary distribution $f(\mathbf{G}|\mathbf{P})$ on the space of possible haplotype reconstructions.

The output of the PHASE algorithm is haplotype frequency estimation and reconstruction of haplotypes of each individual in the sample.

We have cooperated with Mr. Urban on his Master's thesis [38] regarding the problem of haplotype frequency prediction and haplotype resolution using statistical methods in general, and specifically in the context of HLA data. Mr. Urban has proposed a new Bayesian approach that uses the available prior knowledge to solve this task. The algorithms has been compared with our approach (EM algorithm) and even though it gave worse results in terms of accuracy, its robustness in speed when faced with large datasets with missing or ambiguous information in principle allows for processing of register data on a massive scale.

4.3.6 Maximum likelihood approach

Under of the assumption of the Hardy-Weinberg equilibrium [39] and random mating, the probability P_j of the *j*-the phenotype is given by the sum of the probabilities of each of the possible c_j genotypes:

(2)

$$P_j = \sum_{i=1}^{c_j} P(genotype\ i) = \sum_{i=1}^{c_j} P(h_k h_l)$$

where $P(h_k h_l)$ is the probability that the *i*-th genotype is composed of haplotypes *k* and *l*:

(3)

$$P(h_k h_l) = \begin{cases} p_k^2 & if\ \ k = l \\ 2 p_k p_l & if\ \ k \neq l \end{cases}$$

and p_i denotes the frequency of the *i*-th haplotype h_i in the population.

The probability of a sample of *n* individuals, conditioned by phenotype frequencies $P_1, P_2, \ldots, P_m$ is given by the multinomial expression

(4) $$P(sample | P_1, P_2, \ldots, P_m) = \frac{n!}{n_1! n_2! \cdots n_m!} \times P_1^{n_1} \times P_2^{n_2} \times \cdots \times P_m^{n_m}$$

where *m* denotes the total number of phenotypes and n_j is the number of individuals carrying the *j*-the phenotype observed in the sample:

(5) $$\sum_{j=1}^{m} n_j = n$$

Substituting equation (2) into equation (4), we obtain the probability of the sample as a function of the unknown the haplotype frequencies. Therefore, the likelihood of the haplotype frequencies given phenotypic counts is:

(6) $$L(p_1, p_2, \ldots, p_h) = \frac{n!}{n_1! n_2! \cdots n_m!} \times \prod_{j=1}^{m} \left(\sum_{i=1}^{c_j} P(h_{ik} h_{il}) \right)^{n_j}$$

(7) $$\sum_{i=1}^{h} p_i = 1$$

4.4 Solutions of maximum likelihood function

Possible methods of solution of the maximum likelihood function are:

1. Analytic solution [42]
2. Genetic algorithms (own attempt)
3. EM algorithm [43]

4.4.1 Analytic solution

We can logarithmize the equation (6) and get:

(8) $$\log L(p_1, p_2, \ldots, p_h) = a_1 + \sum_{j=1}^{m} n_j \log P_j$$

where a_1 is a constant incorporating the multinomial coefficient.

The maximum likelihood estimates of haplotype frequencies could be, in principle, found analytically or numerically by solving a set of equations resulting from the $h-1$ partial derivatives equated to 0:

(9) $$\frac{\partial \log L}{\partial p_t} = \sum_{j=1}^{m} \frac{n_j}{P_j} \frac{\partial P_j}{\partial p_t}, \quad t = 1,2,\ldots,h-1$$

However the nonlinearity of (9) and a large number of equations when practical data are analyzed (tens of thousands for real data) make this approach prohibitive. Moreover the *h* is often unknown a priori.

Numerical methods must be used to solve these equations and find the maximum. Many numerical methods are sensitive to rounding errors and they are usually not able to prove that a particular solution is the global maximum. Procedures based on analytical solution are limited to a few loci and polymorphism.

4.4.2 Genetic algorithms

Maximum likelihood approach is an optimization problem, so we can consider genetic algorithms (GA) to solve it. Fitness function is very straightforward, because it is the Maximum likelihood function.

But we are in troubles with the definition of GA-chromosome. It should store the result of the algorithm, which is the list of haplotypes and their frequencies. Maximal length of the list is $n \times 2^m$, where m is number of heterozygous loci in the sample ($m = \max s_j$). Every item of this list (haplotype frequency estimation) is a real number (the frequency) and *m* loci with allele designations defining the haplotype. Each HLA locus can have approx. up to 1000 different alleles, so we can encode them to 10 bits. If we encode a real number to 32 bits (frequencies could be very small numbers) we get the size of the GA-chromosome to max. $n \times 2^m \times (m \times 10 + 32)$. For real data ($n = 10^6$ and $m = 3$) we get the GA-chromosome bigger than 0.5 MB which is not feasible for GA. GA could solve only small instances of the problem and is not applicable in our situation.

4.4.3 EM algorithm

One of the most widely used methods of haplotype reconstruction is Expectation-Maximalization (EM) algorithm, which estimates haplotype frequencies iteratively. Since we have used this approach as a basis of our solution, we will describe this algorithm in the following chapter.

4.5 Expectation-Maximalization (EM) algorithm

Association of haplotype structures and sample of unphased genotypes can be expressed by likelihood function (see also section 4.3.6). The relation (6) is complicated and cannot be maximized by standard techniques, as has been discussed before.

The Expectation Maximalization (EM) algorithm was formalized by Dempster A.P. et al. in 1977 [40]. Dempster has proven the monotone behaviour of the likelihood and derived the convergence of the algorithm. Its application to the problem of haplotype reconstruction was formulated in 1995 by several authors [10] [35] [41]. Since then the method and its properties were further analyzed by several studies [42] [43] [44]. They have shown it can be used for wide variety of population and data-set scenarios.

4.5.1 Algorithm description

The EM algorithm is an interactive method of computing sets of haplotype frequencies $p_1, p_2, \ldots, p_h$ starting with arbitrary initial values $p_1^{(0)}, p_2^{(0)}, \ldots, p_h^{(0)}$. These initial values are used to estimate genotype frequencies $\tilde{P}(h_k h_l)$ as if they were the unknown true frequencies (the expectation step). These expected genotype frequencies are standardized and used, in turn, to estimate haplotype frequencies $\hat{p}$ at the next iteration (the maximization step), and so on, until convergence is reached.

4.5.2 Initial conditions

There are several possibilities of initializing the haplotype frequencies $p_1^{(0)}, p_2^{(0)}, \ldots, p_h^{(0)}$ with respect to equation (7). They can be summarized as follows:

- (IC1) All haplotypes are equally likely

(10) $$p_t^{(0)} = \frac{1}{n_h}, \quad t = 1,2,\ldots,n_h .$$

- (IC2) All possible genotypes of each phenotype are equally likely

(11) $$P_j(h_k h_l)^{(0)} = \frac{1}{c_j}, \quad j = 1,2,\ldots,m .$$

- (IC3) Initial haplotype frequencies are chosen at random.
- (IC4) All initial haplotype frequencies are equal to the product of the corresponding single-locus allele/antigen frequencies (complete linkage equilibrium).
- (IC5) The input data influence the initial haplotype frequencies.

4.5.3 The expectation step

- Estimation of genotype frequencies, given haplotype frequencies:

(12) $$\tilde{P}(h_k h_l)^{(g)} = \begin{cases} p_k^{(g)2} & if\ k = l \\ 2 p_k^{(g)} p_l^{(g)} & if\ k \neq l \end{cases}$$

4.5.4 The maximization step

- Estimation of phenotype frequencies, given genotype frequencies

(13) $$P_j^{(g)} = \sum_{i=1}^{c_j} \tilde{P}(genotype\ i)^{(g)} = \sum_{i=1}^{c_j} \tilde{P}(h_k h_l)^{(g)}$$

- Standardization of genotype frequencies

(14) $$P(h_k h_l)^{(g)} = \frac{n_j}{n} \frac{\tilde{P}(h_k h_l)^{(g)}}{P_j^{(g)}}$$

- A genotype has one or two specific haplotypes, so genotype frequencies can be used to estimate haplotype frequencies by direct counting of all occurrences of a haplotype within all sample genotypes.

(15) $$\hat{p}_t^{(g+1)} = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{c_j} \delta_{it} P_j (h_k h_l)^{(g)}$$

where δ_{it} is an indicator variable equal to the number of times haplotype *t* is present in the genotype *i*:

(16) $$\delta_{it} = \begin{cases} 0 & if\ (t \neq k) \wedge (t \neq l) \\ 1 & if\ ((t \neq k) \wedge (t = l)) \vee ((t = k) \wedge (t \neq l)) \\ 2 & if\ t = k = l \end{cases}$$

4.5.5 The stopping criterion

The stopping (convergence) criterion can be defined as:

- (SC1) the relative difference between the consecutive ML function values is less than an arbitrary parameter $\varepsilon > 0$.
- (SC2) the absolute value of difference between the consecutive ML function values is less than an arbitrary parameter $\varepsilon > 0$ [43].
- (SC3) when the changes in haplotype frequency in consecutive iterations are less than an arbitrary parameter $\varepsilon > 0$:

(17) $$\left| \hat{p}_t^{(g+1)} - \hat{p}_t^{(g)} \right| \leq \varepsilon \quad t = 1,2,\ldots,h\,.$$

4.6 Properties of EM algorithm

Sample size

As expected, the algorithm performs better for larger samples sizes, i.e. give better estimates, as shown in [10].

Multiple local maxima

EM algorithm climbs the multidimensional likelihood surface, but there is no guarantee that the surface is convex, i.e. there is no proof for uniqueness of a likelihood function maximum, so the likelihood surface may have multiple local maxima [43].

To ensure finding global maximum likelihood, the EM algorithm should be started from several initial conditions [10].

Deviation from HWE

Departure from HWE may be a substantial source of error, because the algorithm relies on HWE in its expectation step. However, deviation from HWE will not result in a significant differentiation in the haplotype frequency estimation [45]. Also linkage disequilibrium does not impact highly on the common haplotype frequencies [42].

Convergence speed

Most studies confirm high convergence speed of EM algorithm, e.g. in less 20 iterations by [43] or in less than 50 iterations by [42].

Other properties that could be studies are: shape of log-likelihood graph, sensitivity to stopping criteria, LD and departures from HWE and sensitivity to different initial conditions.

4.7 Reliability of haplotype frequency estimation

There is no single measure of performance of EM algorithm, because there are many possible uses of it and the choice of a measure depends on the intended purpose [10]. Anyway some properties could be observed.

4.7.1 Haplotypes with low frequency

When we run haplotype frequency estimation algorithm, we might get list of tens of thousands of haplotypes, but some of them could have very low frequency (e.g. $p_i < 10^{-500}$). The question is if these low frequencies are reliable or not.

We use similar approach as [46], which estimates the minimal registry size in order to calculate reliable haplotype frequencies. In our case, we have fixed size of the sample (registry) and we want to know the reliability of haplotype estimates.

Reliable estimation of the frequency of a haplotype should be supported by at least one individual in the sample carrying the haplotype. If the frequency of *i*-th haplotype is p_i and the sample size is *n*, then the probability that the individual hasn't the *i*-th haplotype is $\overline{P}_i = (1-p_i)^2$, because the individual has two haplotypes. The probability *Q* that at least one individual with *i*-th haplotype is found in *n* individuals is:

(18) $$Q = 1-\overline{P}_i^{\,n} = 1-(1-p_i)^{2n}$$

If we want to reach certain probability *Q*, we can fix it as constant and we get

$$\ln(1-Q) = 2n\ln(1-p_i)$$

(19) $$p_i = 1-e^{\frac{\ln(1-Q)}{2n}}$$

Table 5 shows examples of minimal reliable p_i values for different *n* and *Q* values.

	Q		
N	0.95	0.99	0.999
10^2	1.487×10^{-2}	2.276×10^{-2}	3.395×10^{-2}
10^3	1.498×10^{-3}	2.300×10^{-3}	3.448×10^{-3}
10^4	1.498×10^{-4}	2.302×10^{-4}	3.453×10^{-4}
10^5	1.498×10^{-5}	2.302×10^{-5}	3.454×10^{-5}

Table 5: Minimal reliable value of haplotype frequency estimation.

On the other hand, if a haplotype exists in the sample, then at least one individual has to carry it. Since number of haplotypes in the sample is 2*n*, the minimal frequency of any haplotype must be

(20) $$p_i = \frac{1}{2n}$$

Combining these two approaches, we get

(21) $$Q = 1-\overline{P}_i^{\,n} = 1-\left(1-\frac{1}{2n}\right)^{2n} \approx 1-\frac{1}{e}$$

The calculated value of Q is 0.63 for all values of *n* mentioned in Table 5.

4.7.2 Lab-based verification of the EM algorithm

Verification of the algorithm can be done by this scheme [42]:

1. Generate a model of "true" population, including "true" haplotype frequencies **T**.
2. Do the sampling process, i.e. select or generate individuals according to the population model. As a result, we have phase-known sample and sample haplotype frequencies **S**.
3. Hide the phase information in the sample, i.e. convert genotypes to phenotypes.
4. Estimate haplotype frequencies **E**.

If we compare estimated haplotype frequencies **E** with "true" population haplotype frequencies **T**, we get the assessment of the validity of the final haplotype frequency.

If we compare **S** and **T**, we get the sampling error. As confirmed in [42], the accuracy of the frequency estimation depends on the proper sampling procedure.

4.7.3 Distance from true frequencies

To examine how close estimated frequencies **E** are to "true" frequencies **T**, we can use the similarity index I_F [10]:

$$I_F = \sum_{i=1}^{h} \min(\hat{p}_i, p_{0i})$$

where $\hat{p}_i$ are the estimated frequencies, p_{0i} are the true simulated frequencies and *h* is the number of unique haplotypes in the union of both sets (estimated and true). It varies between zero, when the sets of "true" and estimated haplotypes with non-zero frequency have empty intersection and one, when true and estimated frequencies are identical. This index gives more weight to the high-frequency haplotypes.

$$2\sum_{i=1}^{h} \min(\hat{p}_i, p_{0i}) + \sum_{i=1}^{h} |\hat{p}_i - p_{0i}| = \sum_{i=1}^{h} \hat{p}_i + \sum_{i=1}^{h} p_{0i} = 2$$

holds, so we can express similarity index in other form:

(22) $$I_F = \sum_{i=1}^{h} \min(\hat{p}_i, p_{0i}) = 1 - \frac{1}{2}\sum_{i=1}^{h} |\hat{p}_i - p_{0i}|$$

Other possibilities of comparison of **T**, **S** and **E**, include Goodness of fit, Pearson's r and Spearman's coefficient tests.

5. Design and implementation of HFE algorithm for stem cell donor registry datasets

This chapter discusses our own design and implementation of the HFE algorithm and its usage on datasets of stem cell donor registries – challenges, pitfalls and possible solutions.

5.1 HLA data from stem cell donor registries

Databases of stem cell donor registries are unique and very valuable sources for population genetic studies. The most of the HLA typing results were obtained in accredited HLA laboratories with high quality control standards, which is very important. These data are not "dead", but they are daily used and continuously updated by stem cell donor registries staff in order to find unrelated donors for stem cell transplantation.

On the other hand, HLA haplotype estimation from a sample of a stem cell donor registry is demanding because of the following reasons:

- Missing data.
- Registry data contain HLA results that have been done by different typing techniques, so it contains different typing resolution (see chapter 2.2.3).
- HLA system is extremely polymorphic and people still find a lot of new alleles, see Table 6 and Table 7.
- There are quite a lot of HLA loci for which it would be useful to estimate haplotype frequencies: A, B, C, DRB1, DRB3, DRB4, DRB5, DQA1, DQB1, DPA1 and DPB1. Reliable and unbiased data of DRB3, DRB4, DRB5, DQA1, DPA1 and DPB1 are rare and insufficient for haplotype frequency calculation, therefore for practical reasons, we will consider only A, B, C, DRB1 and DQB1. Consequently, haplotypes could have up to 5 loci.

	Number of possible values				
Resolution	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1
Missing data	1	1	1	1	1
Serology broad	11	32	8	10	4
Serology split	28	61	10	21	9
DNA low resolution	21	36	14	13	5
DNA interm. Resolution	$>10^4$	$>10^5$	$>10^4$	$>10^5$	$>10^4$
DNA high resolution	853	1249	361	659	99

Table 6: Number of possible values (antigens/alleles) in the HLA system (August 2009) [47]

	Number of possible values				
Resolution	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1
Missing data	1	1	1	1	1
Serology broad	11	32	8	10	4
Serology split	28	61	10	21	9
DNA low resolution	21	36	14	13	5
DNA interm. Resolution	$>10^5$	$>10^5$	$>10^5$	$>10^5$	$>10^4$
DNA high resolution	2188	2862	1746	1285	193

Table 7: Number of possible values (antigens/alleles) in the HLA system (January 2013) [47]

5.2 Input and output typing resolution

Our goal is to design and develop a general method that takes as the input a population sample data, a stem cell donor registry database, and calculates haplotype frequencies that cover user-defined set of loci and each locus is calculated at user-requested resolution.

When we start to "play" with different typing resolution, we must keep in mind that all haplotypes entering the EM calculation and appearing in the result set must be disjoint.

Example configuration	A/A*	B/B*	C/C*	DR/DRB1*	DQ/DQB1*
#1				Low res.	
#2		High res.	High res.		
#3	Serol. broad	Serol. broad		Serol. broad	
#4	Low res.	Low res.		Low res.	
#5	High res.	High res.	High res.	High res.	High res.

Table 8: Examples of configuration of HLA haplotype frequency estimation

Table 8 shows examples of desired settings. This variability of configuration is quite challenging. Let us breakdown all possible combinations of input-output relations at any locus, see Table 9.

	Output data - Required resolution of HLA haplotypes				
Input data	Serology Broad	Serology Split	DNA low res.	DNA interm. res.	DNA high res.
Missing data	{01}	{02}	{03}	{04}	{05}
Serology Broad	{11}	{12}	{13}	{14}	{15}
Serology Split	{21}	{22}	{23}	{24}	{25}
DNA Low res.	{31}	{32}	{33}	{34}	{35}
DNA interm. res.	{41}	{42}	{43}	{44}	{45}
DNA high res.	{51}	{52}	{53}	{54}	{55}

Table 9: Input and output HLA typing resolutions.

Most of HLA studies work with uniform input level of typing resolution of all individuals. In order to have such uniform dataset, they:

- Exclude volunteers with different typing resolution (e.g. donors without HLA-DR typing) or
- Collapse serology split level antigens to broad level (e.g. A23 to A9).

We can use datasets with multiple level of typing resolution, because it is not necessary to require the level of typing resolution to be statistically independent on the HLA type [48].

In fact, the situation about input data is more complicated, because an individual can have two HLA codes of different resolution at one locus and we must have solution that can deal with it. We could get results like DRB1*01:XX, 07:01 (mix of low and high resolution). Nevertheless we will expect the input HLA typing complains WMDA guidelines for use of HLA nomenclature [15] that is true for databases of stem cell donor registries. Therefore we do not have to deal with mix of serology and DNA typing results at one locus (e.g. DR1, DRB1*04:XX).

Table 9 defines 30 different situations that could happen at a locus:

- Cases {*X*4} make no sense, neither for practical purposes nor for extreme diversity of intermediate resolution HLA codes.
- **(EQ)** Cases {*XY*}, where *X* = *Y*, are the easiest ones, because we do not have to convert input and output HLA codes.
- **(LO)** Cases {*XY*}, where *X* < *Y*, mean conversion of codes from lower to higher resolution. In other words, it is expectation of higher resolution typing, given a lower resolution typing. Special cases {*0Y*}, i.e. first row of the table, handle missing data.
- **(HI)** Cases {*XY*}, where *X* > *Y*, mean conversion of codes from higher to lower resolution. In other words, it is degradation of HLA typing results to lower resolution.

Cases (HI) are also important. The most of studies performing HLA haplotype frequency estimation on serology broad/split level just ignore DNA typing results of individuals in the sample. But this information should not be ignored, because it can improve the serology typing results of an individual. This approach is also in harmony with findings of the study [48].

5.3 Missing data

We consider a phenotype to present a **missing value** when no antigens/alleles are reported at a particular locus. We assume that the presence of missing values is independent on hidden values and other reported values.

Example: The typing result of an individual could be A1,2 B7,8, so just A and B loci are HLA typed. Locus DR is not typed, therefore contain missing values.

There are several methods how to handle missing values in population data:

(MI-1) **Ignoring individuals** with incomplete information (EH software). This approach introduces sampling error and overestimates common haplotypes.

(MI-2) Treating a missing antigen/allele as **any other antigen/allele** (ARLEQUIN software). This approach generates unreal haplotypes.

(MI-3) Consider missing value as **any allele**. The best approach, but computationally demanding. It means to generalize definition of c_j in equation (1), so now the $\tilde{c}_j$ is number of all possible genotypes that could lead to phenotype j. Then sums in equations (13) and (15)

iterates through all pseudo-haplotypes, i.e. haplotypes compatible with the given phenotype.

(MI-4) Consider missing value as **any allele that is already found** associated with the observed alleles at the other loci in the dataset where considered to substitute missing values [49]. This approach is an optimization of the previous one. The idea behind is based on the fact that EM algorithm in the previous approach will gradually withdraw those haplotypes that are not directly observed in the sample (in complete phenotypes). This method therefore provides the same result as the previous one.

The study [49] shows that the MI-4 method is better than MI-1 and MI-2, especially when the study is focused on rare haplotypes.

(MI-5) An enhanced approach of MI-4 (Henk van der Zanden, personal communication, 2008):

- Transform input dataset with missing values to new one, without missing values
- Missing values are guessed according to analysis of phenotypes without missing values.
- One phenotype with missing values is substituted by more phenotypes without missing values and the original number of individuals of this phenotype is proportionally divided between new phenotypes.

Problem of this approach is there could be missing values which cannot be substituted. Advantage of this approach is it simplifies the computation. On the other hand it tries to do some work in advance that should be done by the EM algorithm. Its influence on the accuracy of haplotype frequency estimation should be tested, but we think it will not provide better estimates, maybe the same ones.

(MI-6) An enhanced approach of MI-1 [50]: Calculate full (3-locus) haplotype frequencies ignoring individuals with incomplete information (like MI-1). Then correct these haplotype frequencies by adjusting them according to the ratio of the resulting (2-locus) marginal frequencies and the direct estimate from the full registry.

5.4 Lower to higher typing resolution

HLA typing techniques often give results as ‘ambiguities’, which means the result is not perfectly determined (high resolution), but some of the known alleles could be discarded. Such result could be a list of possible alleles or multiple allele code. In fact the missing value according to approach MI-3 (resp. MI-4) is also a kind of multiple allele code that represents all existing alleles (resp. all observed allelic combinations in the sample). The study [49] suggests ambiguities “could easily be handled using the same statistics as those presented for missing values”, but “this theoretically simple process becomes complicated to implement”.

5.4.1 Mapping serology broad to split values

These cases refer to the cell {12} in the Table 9. The study [51] is the first one that maps broad antigens to all possible split antigens in order to generate all possible genotypes to be considered. It

is similar to the approach MI-3 and shows the complexity of HLA system even for serology haplotypes.

Example 2

A9 is mapped to split group {A23, A24}. An individual with phenotype A2, A9; B8, B35 could have one of these (split) genotypes:

- A2 – B8 / A23 – B35
- A2 – B8 / A24 – B35
- A2 – B8 / A23 – B35
- A2 – B35 / A24 – B8

□

The worst case for calculation of HLA-A, HLA-B, HLA-DR haplotypes is the phenotype A10,19; B15,22; DR5,6 (six broad antigens), resulting in 3456 different possible genotypes.

5.4.2 Overlapping mapping of multiple allele codes

But the situation with ambiguities is more complex than with missing values. Both MI-3 and MI-4 map a missing value to exclusive set of alleles.

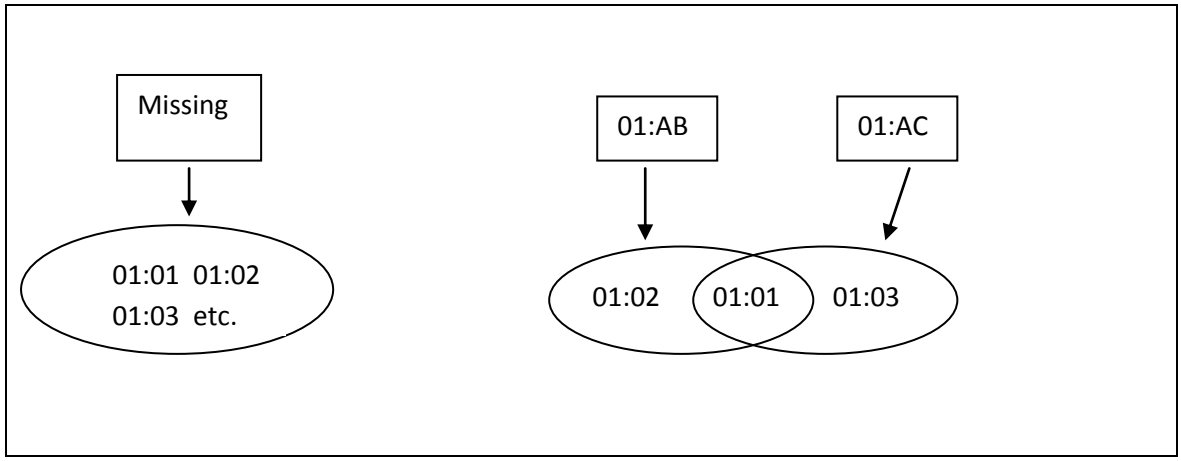


Figure 6: Comparison of missing value and other ambiguities.

On the other hand, multiple allele codes are mapped to sets of alleles that have nonempty intersection, see Figure 6. Other problem is multiple allele codes can contain only few special alleles which are not observable in the registry database as high resolution result. This leads to the conclusion that MI-3 and MI-4 do not give the same result for ambiguities, as shown in the following experiment.

Experiment 1

Data set:

- One individual with A*01:AB (=01:01/01:02)
- One individual with A*01:AC (=01:01/01:03)
- One individual with A*01:AG (=01:01/01:06)
- One individual with A*01:02
- One individual with A*01:03
- As we can see A*01:01 and A*01:06 are not directly represented in the dataset, so MI-4 would not work for A*01:AG.

Required HLA haplotypes:

- A* (high resolution)

Results: After 16 iterations of the EM algorithm with MI-3 strategy, the A*01:01 is the most frequent allele (0.447), followed by A*01:02 (0.276) and A*01:03 (0.276) and A*01:06 (<0.001).

□

Experiment 2

Data set:

- 10 individuals with A*01:AB (=01:01/01:02)
- 1 individual with A*01:AC (=01:01/01:03)
- 1 individual with A*01:02
- 1 individual with A*01:03
- As we can see A*01:01 is not directly represented in the dataset, so MI-4 would ignore it.

Required HLA haplotypes:

- A* (high resolution)

Results: After 19 iterations of the EM algorithm with MI-3 strategy, the A*01:02 is the most frequent allele (0.498), followed by A*01:01 (0.410) and A*01:03 (0.090).

□

As conclusion, for ambiguities we should use similar strategy as MI-3, take into account all possible alleles.

5.4.3 Overlapping serology to DNA mapping

Serology to DNA mapping is very practical, but its impact on EM algorithm hasn't been previously studied. S. GE Marsh publishes mapping of HLA alleles to antigens [52], so in order to get HLA serology to alleles mapping, we should calculate the reverse index. Other mappings, such as serology to low resolution DNA, can be obtained from the previous one. But it also raises some problems.

Example 3 (HLA nomenclature as of January 2013 [52])

A*02:65 could be A31, therefore A31 should be mapped to A*02:65, A*31:01…A*31:71, A*33:09. Reducing this list to low resolution we get A31 mapping to A*02, A*31, A*33. But:

- EM algorithm would prefer A*02 mapping of A31, because A*02 is more common than A*31 and A*33.
- It seems like A31 could be potentially A*02:01, which is not true.

□

Example 4

Broad A28 has splits A68 and A69. A*02:55 could be A28 or A2 (assumed). Therefore A28 could be A*02, A*68 or A*69. In context of A28, the A*02 group contains just one allele (A*02:55), the A*68 group contains at least 40 alleles and A*69 contains just one allele (A*69:01). So it is very likely the A28 will be A*68. In order to observe how the EM will deal with a phenotype containing A28 in the context of real data, we have tried the following experiment.

□

Experiment 3

Data set:

- The Cord Blood Bank Czech Republic, November 2008, n = 2825
- Additional individual with the phenotype P_A: A11,28 B*35:XX, DRB1*01:XX.

Required HLA haplotypes:

- A*-B*-DRB1* (low resolution - low res. - low res.)

Results: Table 10 shows distribution of possible genotypes of the phenotype during EM iterations and behavior of maximization step (14).

	$\frac{\widetilde{P}\left(h_k h_l\right)^{(g)}}{P_j^{(g)}}$		
Iteration	GENOTYPE 1 h_k: A*02,B*35,DRB1*01 h_l: A*11,B*35,DRB1*01	GENOTYPE 2 h_k: A*68,B*35,DRB1*01 h_l: A*11,B*35,DRB1*01	GENOTYPE 3 h_k: A*69,B*35,DRB1*01 h_l: A*11,B*35,DRB1*01
1	79,642%	19,055%	<0,001%
2	80,609%	19,348%	<0,001%
3	81,737%	18,261%	<0,001%
4	83,170%	16,830%	<0,001%
5	84,432%	15,568%	<0,001%
6	85,379%	14,621%	<0,001%
7	86,044%	13,956%	<0,001%
8	86,496%	13,504%	<0,001%
9	86,806%	13,194%	<0,001%
10	87,023%	12,977%	<0,001%
…	…	…	…

Table 10: Distribution of possible genotypes of the phenotype during EM iterations in the experiment

This experiment shows the test phenotype "helps" more genotype 1 than the more accurate genotype 2. And the EM tends to prioritize the genotype 1 during its iterations. This behavior will lead to the overestimation of the haplotype A*02,B*35,DRB1*01 and underestimation of the haplotype A*68,B*35,DRB1*01.

□

This problem comes from two facts:

- HLA-A antigens are mapped to set of HLA-A* alleles that are overlapping.
- The maximization step does not reflect relations between HLA alleles of different typing resolution. Therefore all feasible genotypes of a phenotype are handled in the same way.

This problem comes from the equations (12) - (14) in combination with MI-3 approach, because they do not reflect HLA nomenclature and handle all mapping values in the same way. Therefore we propose to change the equation (12) to the extended form:

(23) $$\tilde{P}(h_k h_l)^{(g)} = \begin{cases} p_k^{(g)2} \times P\left(genotype\ h_k h_l \middle| phenotype\ j\right) & if\ k = l \\ 2\, p_k^{(g)} p_l^{(g)} \times P\left(genotype\ h_k h_l \middle| phenotype\ j\right) & if\ k \neq l \end{cases}$$

Unfortunately we do not know these conditional probabilities. But if we assign

(24) $$P\left(genotype\ h_k h_l \middle| phenotype\ j\right) = \frac{1}{c_j}$$

the EM algorithm will behave in the same way as original approach, because it does not affect the equation (14). If the value of $P\left(genotype\ h_k h_l \middle| phenotype\ j\right)$ is higher than $\frac{1}{c_j}$, the genotype $h_k h_l$ is "promoted" over other possible genotypes of phenotype *j*. If it is lower, the genotype $h_k h_l$ is suppressed. It does not affect convergent properties of the EM algorithm.

If we know "true" haplotype frequencies, we could easily calculate $P\left(genotype\ h_k h_l \middle| phenotype\ j\right)$. This leads to the following algorithm:

1. Assign $P\left(genotype\ h_k h_l \middle| phenotype\ j\right) = \frac{1}{c_j}$.
2. Run EM algorithm, using equation (21).
3. Calculate new $P\left(genotype\ h_k h_l \middle| phenotype\ j\right)$.
4. Repeat steps 2 and 3 until $P\left(genotype\ h_k h_l \middle| phenotype\ j\right)$ is "stable" – e.g. until maximal relative change of any $P\left(genotype\ h_k h_l \middle| phenotype\ j\right)$ is lower than ε.

This approach would be very computationally demanding, since it runs the EM algorithm several times.

Other possibility is to approximate $P(genotype\ h_k h_l | phenotype\ j)$ by HLA nomenclature relations.

Example 5

A31 is mapped to {A*02:65, A*31:01…A*31:71}. The size of the set is 72 alleles (two fields only). Therefore the probability the A31 will be A*02:65 is 1/72. Consequently the probability the A31 will be A*02 is also 1/72. Genotypes containing A*02 are suppressed among all genotypes of phenotype with A31.

□

Experiment 4

Data set:

- The Cord Blood Bank Czech Republic, November 2008, *n* = 2825
- Additional individual with the phenotype P_A: A11,28 B*35:XX, DRB1*01:XX.

Required HLA haplotypes:

- A*-B*-DRB1* (low resolution - low res. - low res.)

Results: Table 11 shows distribution of possible genotypes of the phenotype after run of the EM algorithm and behavior of adjusted maximization step using equation (21).

	$\frac{\tilde{P}(h_k h_l)^{(g)}}{P_j^{(g)}}$		
Iteration	GENOTYPE 1 h_k: A*02,B*35,DRB1*01 h_l: A*11,B*35,DRB1*01	GENOTYPE 2 h_k: A*68,B*35,DRB1*01 h_l: A*11,B*35,DRB1*01	GENOTYPE 3 h_k: A*69,B*35,DRB1*01 h_l: A*11,B*35,DRB1*01
38	14,69%	85,31%	<0,01%

Table 11: Distribution of possible genotypes of the phenotype after run of the EM algorithm in the experiment

This experiment shows the correction by equation (21) managed to prioritize genotype 2 over the genotype 1.

□

5.5 Higher to lower typing resolution

Mapping of higher resolution to lower resolution is quite straightforward. Split serology antigen can be easily mapped to broad. Allele codes could be mapped to serology code(s) by [52]. Other mapping could be obtained by combination of these two. Therefore we can always get set of lower typing resolution codes that are assigned to higher typing resolution code.

Example 6

A*01AB is mapped to A*01 (intermediate to low resolution)
A*01AB is mapped to A1 (intermediate to split/broad resolution)
□

5.6 Data preprocessing

For practical implementation of the algorithm, data preprocessing steps are necessary. Challenges and problems of the input database are described in the chapter 5.1

5.6.1 Checking of input data

As the first step, the preprocessor should check input data for errors and bring them to the consistent form [51].

5.6.2 Grouping of phenotypes

Summarization in equation (15) runs over all phenotypes. In highly polymorphic system, it is more efficient to sum over individuals, because there are fewer individuals sampled than potential phenotypes. It is also very useful to group all the same phenotypes in the sample into one record and count number of occurrences n_j of such phenotype. This is especially useful for individual with missing data (e.g. HLA-AB typed donors).

5.6.3 Feasible genotypes and haplotypes

The probabilities appearing in equations (12)-(16) are indexed by both haplotype and genotype numbers. Given the observed phenotypes, we can generate and index list of all feasible genotypes and haplotypes as proposed in [43].

The indexing of haplotypes is natural since a haplotype could be shared by many genotypes and phenotypes. However we have found the indexing of genotypes does not substantially increase the performance of the EM algorithm on typical HLA samples, because there is almost no redundancy.

Experiment 5

Data set: The Cord Blood Bank Czech Republic, November 2008, *n* = 2825

Required HLA haplotypes: A*-B*-DRB1* (low resolution - low res. - low res.)

Results:

- 3887 possible haplotypes
- 20198 feasible genotypes
- 20153 unique genotypes

□

During initialization phase of the EM algorithm all possible genotypes derivable from an input phenotype should be generated. This includes finding mapping of all input HLA codes (including missing values) to list of output typing resolution codes. Generating such list is time consuming procedure (e.g. HLA antigen to list of alleles mapping) and the list occupies a lot of memory space. Therefore we have found useful to cache these lists and reuse them. This is especially useful with the mapping of missing values to output typing resolution codes.

5.7 Computational problems

The EM algorithm can theoretically handle an arbitrary number of polymorphic loci and arbitrary level of polymorphism. But in practice it is limited by the number of possible genotypes that could be handled by computers.

Number of possible genotypes is influenced by:

- Number of polymorphic loci – exponential relation, according to equation (1)
- Sample size
- Homozygosis – degree of homozygosis of individuals, number of heterozygous individuals
- Missing data or typing of individuals at different resolution than required
- Degree of polymorphism at observed loci

Addressing these issues is the main challenge of the EM algorithm implementation.

5.8 Our implementation

Our object-oriented implementation of the EM algorithm was built by 64bit version of the Embarcadero Delphi XE2 compiler. HLACORE library [53], kindly provided by ZKRD, was used as the low level library for handling HLA data according to the HLA nomenclature [15].

5.8.1 Universal configuration

We have implemented uniform solution of input-output typing resolution options, see chapter 5.2.

The software covers all desired input-output configurations, see Table 12. Since serology typing is declining and less accurate, the serology as output resolution is not our point of interest. It is better to map serology data to DNA than vice-versa.

	Output data - Required resolution of HLA haplotypes				
Input data	Serology Broad	Serology Split	DNA low res.	DNA interm. res.	DNA high res.
Missing data	Not needed.	Done.	Done.	Does not make sense.	Done.
Serology Broad	Mapping is not needed.	Done.	Done.	Does not make sense.	Done.
Serology Split	Not needed.	Mapping is not needed.	Done.	Does not make sense.	Done.
DNA Low res.	Not needed.	Not needed.	Mapping is not needed.	Does not make sense.	Done.
DNA interm. res.	Not needed.	Not needed.	Done.	Does not make sense.	Done.
DNA high res.	Not needed.	Not needed.	Done.	Does not make sense.	Mapping is not needed.

Table 12: Input and output HLA typing resolutions.

5.8.2 Data preprocessing

The program implements data preprocessing ideas described in this work, including:

- During initialization phase, conditional probabilities $P(genotype\ h_k h_l | phenotype\ j)$ are calculated, see chapter 5.4.3. This is used mainly for low resolution output.
- Memory sharing of haplotypes and genotypes, caching of input-output resolution HLA code mappings, see chapter 5.6.3

In order to limit the computational complexity, the user can set limit - maximum acceptable number of genotypes per donor ($\tilde{c}_j$), for example 10^6. This will exclude donors with the poorest information about background haplotypes. This approach has to be used carefully as discussed in [48] and [54].

5.8.3 Haplotype data structure and indices

One of the key issues in the design of HFE algorithm is development of efficient data structure that keeps lists of all relevant haplotypes. Fast access to these haplotypes is essential for good performance of the HFE algorithm. With the data structure, we perform two critical operations: adding new haplotypes (INSERT) and searching for specific haplotype without knowledge of the haplotype index (SEARCH). These two operations are frequently called even in the initialization phase of the EM algorithm, when the final number of all haplotypes is not known. Then the Expectation step of the EM algorithm needs to quickly access specific haplotype with known index (GET) and the Maximization step loops though all haplotypes (LOOP) and updates their frequencies.

In general, a haplotype is a vector of HLA allele/antigen codes, see (26) in chapter 9.2. These HLA allele/antigen codes are alpha-numerical strings that can be up to 12 characters long (”01:01:01:01N”).

Jan Hofmann [55] uses rooted three structures that store antigens/alleles in nodes of the tree. First locus is stored in nodes with distance one to the root, second locus in the next level, etc. (see Figure 7).

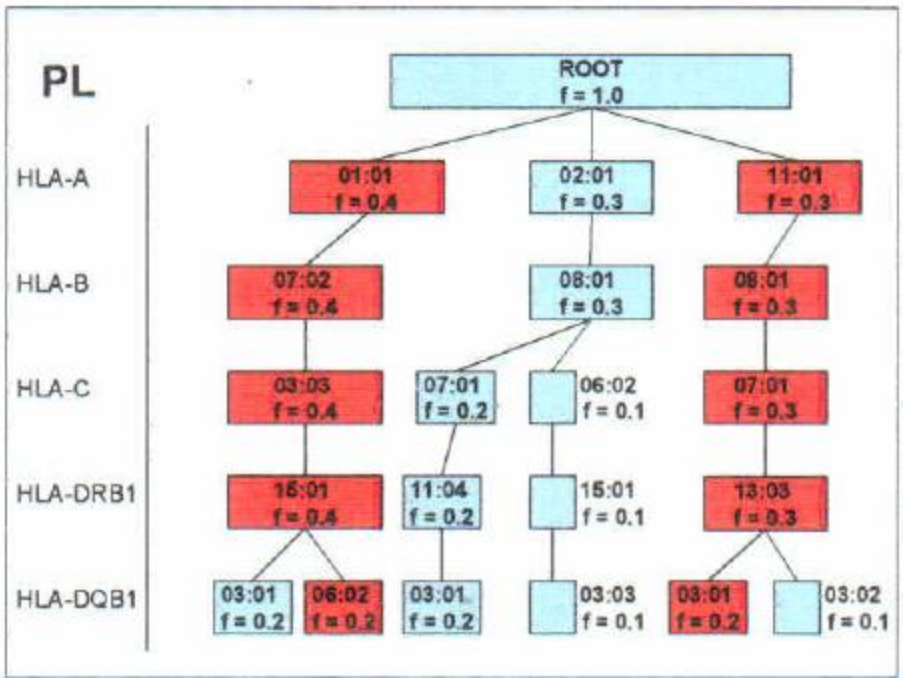


Figure 7: Haplotype data structure as a tree [55]

Data structure of an individual node needs to hold up to thousands of edges to the next level. So we still need to address the issue of fast indexing of HLA antigens/alleles in the node and fast INSERT operation. The GET operation is now more complicated. So we have rather focused on linear data structures.

Easiest possibility is to index all relevant haplotypes by consequence integers and then store them in a list, array or matrix. This is quite easy implementation, requires just O(1) for the INSERT and GET operations and LOOP is also easy. But the SEARCH operation requires O(N), which is not acceptable.

We can sort the list of haplotypes by their vector of HLA codes (e.g. alphabetical order). By this approach, the SEARCH operation has the complexity O(log(N)), but the INSERT operation has increased to O(N log(N)), which might be problematic. We have implemented this approach, but we do lazy sorting, e.g. the sorting is not done after every INSERT operation, but after every 100 INSERT operations. This decreases 100 times number of calls of the slow sort operation, but increases the SEARCH operation by a constant, maximally 100, because these unsorted haplotypes have to be checked if SEARCH operation fails on the sorted lists. Constant 100 has been chosen experimentally.

We have found out the k constant in the SEARCH operation $O(\log(N)) = k \times \log(N)$ is too high, because comparison of two haplotypes requires comparison of several HLA antigen/allele codes, i.e. several string operations. Therefore we have encoded haplotype into single integer and reduced the haplotype comparison operation by single processor cycle. Encoding is done in this way:

- All existing HLA allele and antigen codes at a locus are alphabetically sorted. For example, for locus A/A*, we get sequence: “01:01”, “01:02”, etc.
- We assign them integers, starting from 0. So “01:01” gets 0, “01:02” gets 1, etc.
- Since there are less than 3000 known alleles at a single locus, all HLA codes at a single locus could be encoded by 12 bit integer.
- Haplotype index is created by concatenation of these HLA code integers. For 5 loci haplotype, we get 60 bit integer. Current processors can handle 64 bit integers in single operation.

5.8.4 Allele list reduction

Exponential growth of HLA nomenclature allele list in recent years complicates the EM algorithm and dramatically increases the computational complexity. However, most of these new and rare alleles will never be observed in the sample. Therefore it is good idea to reduce considered HLA alleles. This could be done by:

1. Applying additional knowledge of the sample population or ethnic group and usage of known allele list estimated in the past on similar population or ethnic group (e.g. Caucasian). This could be for example list of “Common and Well-Documented HLA Alleles” (CWD) [56].
2. Several runs of the EM algorithm on the sample. We can calculating allele frequencies first, then filter less likely ones (e.g. with probability lower than $p_i = \frac{1}{2n}$).
3. The greedy algorithm that begins with a set of reference alleles defined for particular population and adds additional alleles in order by which allele allows the most new donor typings to be interpreted. Reinterpretation is done at each cycle and the allele list grows until all donors have valid genotype lists. This algorithm has been implemented by NMDP [57].

In our implementation, we use the second option, because it is more universal. In case the data preprocessing phase finds an HLA code that cannot be interpreted by reduced allele list, it takes the first compatible allele outside the filtered range, i.e. it find the most likely allele with the probability bellow p_i that interprets the problematic HLA code.

5.8.5 Partial haplotype list reduction

Similarly, partial haplotypes (see chapter 9.4) could be pre-estimated and the algorithm can reduce haplotype list by filtering those haplotypes that do not match to any of the pre-selected partial haplotypes (probability bellow p_i).

Thanks to strong linkage disequilibrium (see chapter 2.1) we have used this method for B-C and DRB1-DQB1 haplotypes.

5.8.6 Haplotype list reduction

In extreme case, we can run the EM algorithm with already known list of output haplotypes. The EM algorithm ‘just’ estimates their probabilities.

5.8.7 Genotype list reduction

If output haplotypes are known before the EM algorithm starts and even their probabilities are known (at least approximately), we can filter less likely genotypes (see chapter 5.6.3) with low probability.

5.8.8 User interface

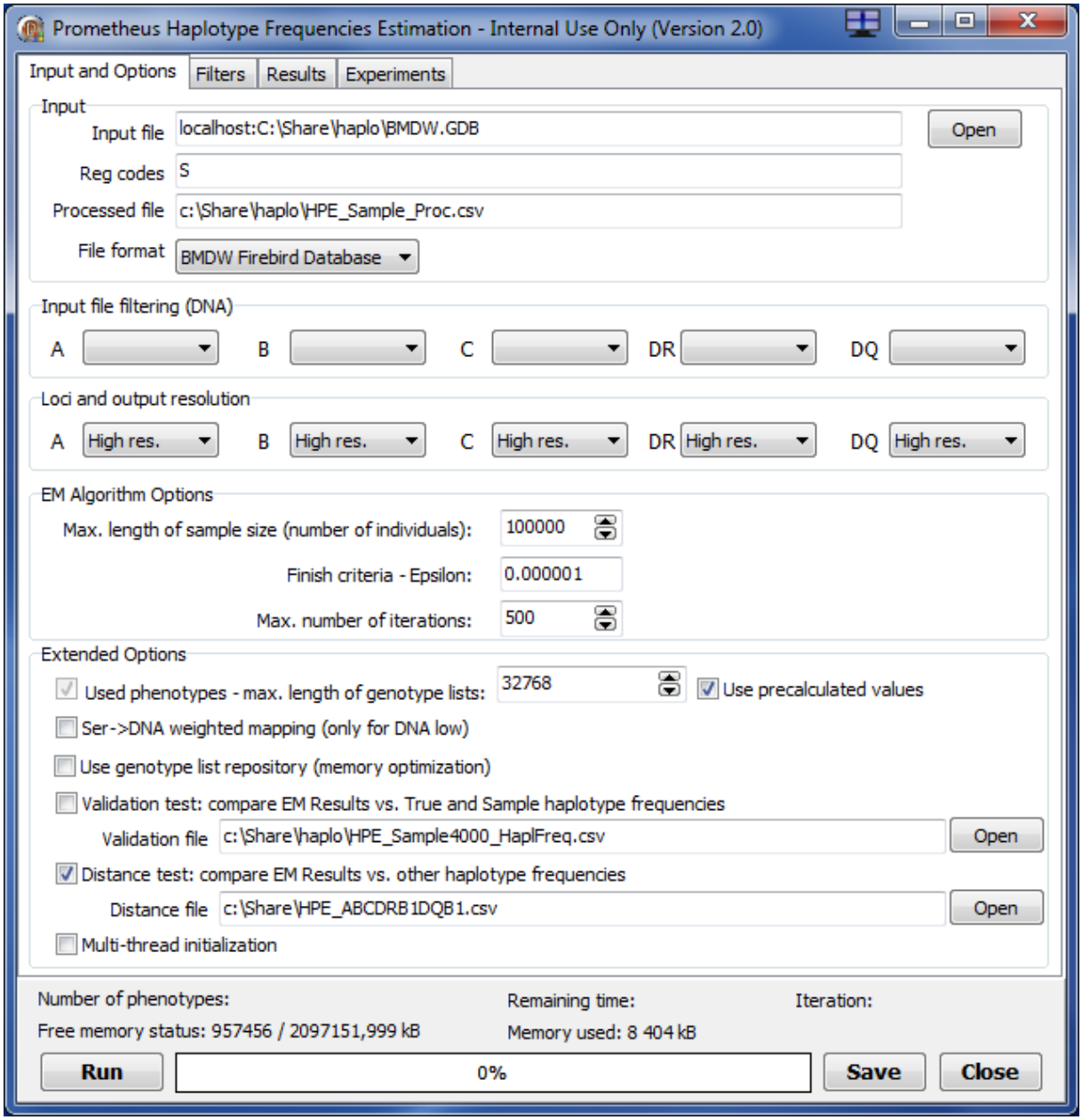


Figure 8: User interface of our HFE implementation

The algorithm is run via user interface implemented under Windows OS. Figure 8 shows screenshot of the window with most important settings:

- **Input**: input file with the sample, registry ID selection, file format of the input file (CSV, BMDW file format or relational database)
- **Input file filtering**: by default, all input phenotypes are accepted, but user can filter out phenotypes that do not meet minimum desired level of typing (e.g. low resolution). For example, this can be used to filter out donors without DRB1 typing.
- Loci and output resolution: selection of loci and requested resolution of output haplotypes. Resolution can be set individually at each locus (see also chapter 5.2).

- **EM algorithm options**: maximum number of individuals in the sample, finish criteria (see chapter 4.5.5), maximum number of iterations
- **Other options**:
    - Maximum length of genotype list $\tilde{c}_j$ (see chapter 5.8.2).
    - Serology to DNA weighted mapping (see chapter 5.4.3)
    - Genotype list repository (see chapter 5.8.7)
    - Validation tests and distance calculations between result and reference frequencies
    - Multi-thread initialization: possibility to use parallel computing during the data preprocessing phase (see chapter 5.6.3)
- **Optional filters**:
    - Allele list reduction (see chapter 5.8.4)
    - Partial haplotype list reduction (see chapter 5.8.5)
    - Haplotype list reduction (see chapter 5.8.6)

5.8.9 Hardware

We have run experiments on a PC with Windows 7 Professional SP1 64bit, Intel Core i3-2120 CPU @ 3.30 GHz, 16 GB RAM.

5.9 Other studies and implementations of the HFE algorithms

5.9.1 Small samples

Computer programs described in most papers work with quite small instances:

- [58] (EH): max. 30 alleles per loci
- [35] (HAPLO): up to 114 haplotypes, 114 observed phenotypes, and 500 genotypes.
- [42]: 8-14 biallelic markers per gene in 300 individuals
- [10]: 2-8 highly polymorphic loci with 20 possible alleles. They have considered samples where the total number of possible haplotypes did not exceed 16384.
- [41]: 619 individuals, three loci HLA-A, HLA-B, HLA-C, serological testing

One of the first analyses of stem cell donor registries [59] calculated ABDR haplotype frequencies of registries in the 22nd edition of the Bone Marrow Donors Worldwide (1997):

- HLA-A, HLA-B and HLA-DR
- broad antigens have been preferentially used instead of their splits
- some registries were excluded from the analysis because of various problems (e.g. deviation from HWE).
- Maximal size of a registry dataset was about 50,000 individuals.

5.9.2 State-of-the-art HLA studies

HLA system is much more complex, see Table 6. The biggest state-of-the-art HLA studies are performed in Germany and the United States, which have the biggest databases of bone marrow donor registries:

- **2003 - German Blood Donors** [51]: three loci HLA-A, HLA-B and HLA-DRB1; conversion of broad to split antigens, 13,000 individuals, about 10,000 haplotypes; a single individual with the typing result A10,19; B15,22; DR5,6 (six broad antigens), has 3456 possible genotypes.

- **2005 - German registry ZKRD** [60] [61]: three loci HLA-A, HLA-B and HLA-DRB1; about 1 million donors, 412,494 of these individuals were typed for HLA-DRB1 at low or intermediate resolution and another 90,673 at high resolution level. HLA-A and B were analyzed using serological nomenclature without associated antigens. For high res. frequencies donors only typed for A and B were excluded due to algorithmic limitations. Low resolution data were then used to correct a possible selection bias in the restricted data set. Computation took 2 resp. 9 days.

- **2006 – ZKRD** (presented at the WMDA conference 2006): HLA-A, HLA-B and HLA-DRB1 high resolution haplotype frequencies estimations; 120,000 individuals; 10^7 haplotypes to consider; up to 5×10^8 diplotypes per phenotype to consider; description matrix (specifying which pairs of haplotypes are to be considered for a given phenotype) has 10^{19} elements, 10^{10} of them are positive

- **2007 – NMDP** [62]: three loci HLA-A (max. 21 antigens), HLA-B (max. 42 antigens) and HLA-DRB1 (max. 250 alleles); 3.5 million individuals; 21 x 42 x 250 = 220,500 total haplotypes; a single individual with the typing result A10,19; B15,22 (DRB1 not tested), has more than 9.5 million possible genotypes. 5.5 hours running on a cluster of five Sun Fire V100 servers (2 GB RAM).

- **2007 – NMDP** [62]: five loci HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DQB1; high resolution; comparison of US ethnic groups; up to 6,500 individuals in one ethnic group. Because of limitation of the EM algorithm at greater than three loci with registry data, four- and five-locus haplotype frequencies were estimated using initial EM runs on the two tightly linked locus clusters (C-B and DRB1-DQB1) followed by a second three-locus EM run that considered the tightly linked clusters as a single locus.

- **2008 - ZKRD** (Carlheinz Muller, personal communication): five loci HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DQB1; tens of thousands individuals; high resolution; computed on server with 64 GB RAM; program runs more than ten hours.

- **2010 – ZKRD** [63] [64]: five loci HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DQB1; hundreds of thousands of individuals; high resolution.

- **2011 – DKMS** [65]: 20 thousand Polish stem cell donors, four loci: HLA-A, HLA-B, HLA-C and HLA-DRB1.

- **2012 – NMDP** (Loren Gragert, presented at the 16th IHIWS conference in Liverpool):
    - five loci HLA-A, HLA-B, HLA-C, HLA-DRB1 and HLA-DQB1
    - NMDP can run EM algorithm on BMDW database, for every registry and every country
    - only DNA based typing is considered, but donors without C and DQB1 typing are still included

- Experience: genotypic ambiguity of BMDW HLA typing is too high for conventional EM to be practical. Two main strategies were implemented to reduce ambiguities, reducing ambiguity: Allele list reduction by greedy algorithm and Blocks / Imputation.

5.10 Comparison of our implementation with others

For comparison between the main HFE implementations, including our algorithm, see the Appendix D. The table shows applications of HLA HFE algorithms of research groups that cooperate in the Registry Diversity Subcommittee of the World Marrow Donor Association (WMDA) Information Technology Working Group. It gives overview of technology (platforms, programming languages), limitations of the algorithms (maximum number of loci, maximum number of phenotypes, accepted input), initial and terminating conditions, internal methods (mapping of alleles, handling of ambiguities), running time on common tasks and practicalities (output format).

6. Reliability of HFE algorithm on registry datasets

This chapter describes own research results of the reliability of HFE algorithm on real registry datasets. The reliability of HFE depends on typing ambiguities of registry donors, computational complexity and used heuristics, population size, sample size and population homogeneity. We will study these parameters independently in controlled data environment and finally, we will combine them together, like in real registry dataset.

6.1 Typing ambiguities and computational complexity

Key factors that influence the reliability of HFE are the structure of the registry and ambiguity of HLA typing results of donors in the sample. This also influences computational complexity of the HFE algorithm, especially values $\tilde{c}_j$.

Previous studies have also pointed out this important aspect. ZKRD has visualized structure of the registry [63] by three-dimensional graph. Every field represents different combination of missing/low-resolution/intermediate-resolution/high-resolution typing at five loci (A*, B*, C*, DRB1*, DQB1*). The horizontal axe shows the first class loci and the vertical axe shows the second class loci. The more dark blue, the relative number of donors is higher.

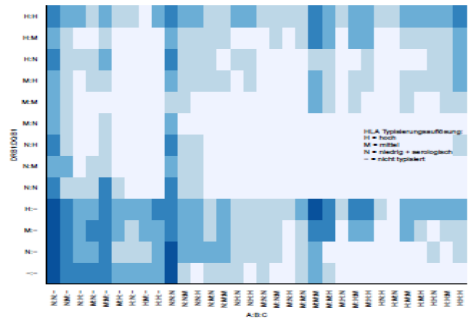


Figure 9: Visualization of the HLA typing ambiguities in ZKRD [63]

We need different visualization that would better represent computational complexity and value $\tilde{c}_j$ - number of genotypes per donor. Computational complexity is one of the main obstacles when someone tries to calculate HFE. Following example demonstrates the problem.

Example 7

- Output: A*-B*-C*-DRB1*-DQB1* high resolution haplotypes
- HLA nomenclature: April 2012
- An individual with HLA type A*01:01, B*08:01, C*07:01, DRB1*03:01, DQB1*02:01 is high resolution typed, homozygous, so there is just one possible genotype, $\tilde{c}_j = 1$
- An individual carrying HLA type A2, B7,62 was typed by serology techniques, so there are many possible genotypes, $\tilde{c}_j \approx 6 \times 10^{26}$. CSCR registry has more than 20 individuals with this HLA type.

□

This example shows $\tilde{c}_j$ can grow to more than 25 digits. In the same way, we have analyzed all donors in a registry and visualized number of genotypes per donor $\lfloor \lg(\tilde{c}_j) \rfloor$ vs. number of donors carrying such level of ambiguity $D(\lfloor \lg(\tilde{c}_j) \rfloor)$. E.g. the first donors in the Example 7 has $\lfloor \lg(\tilde{c}_j) \rfloor = 0$ and the second has $\lfloor \lg(\tilde{c}_j) \rfloor = 26$.

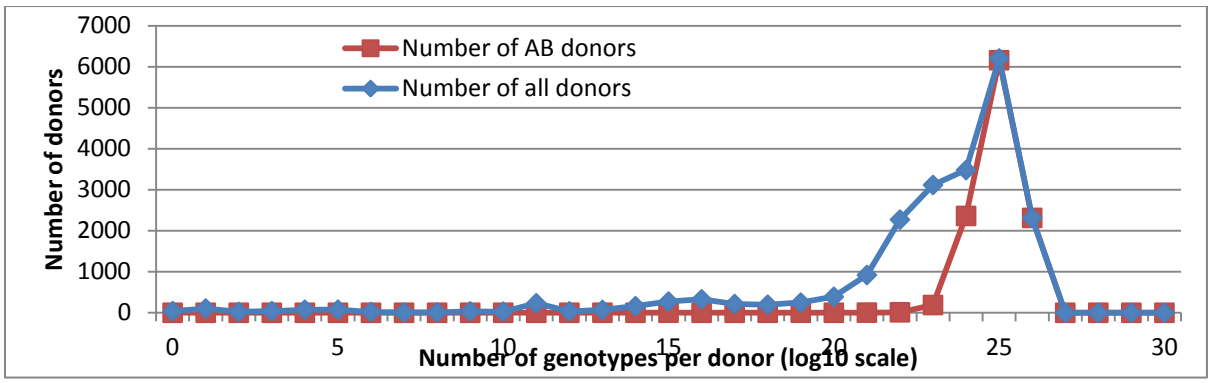


Figure 10: Visualization of the HLA typing ambiguities and computational complexity in CSCR, May 2012

The graph shows huge number of donors with $\tilde{c}_j \approx 10^{25}$. Most of these donors are only serologically AB typed.

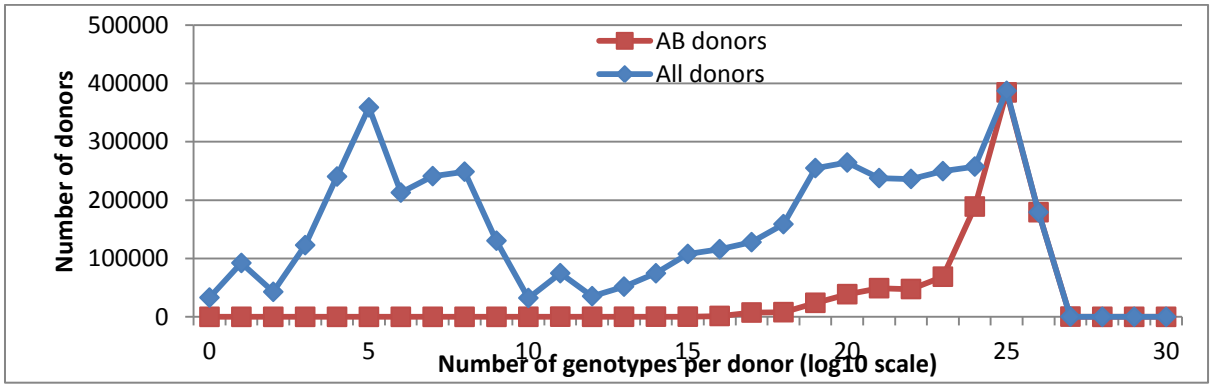


Figure 11: Visualization of the HLA typing ambiguities and computational complexity in ZKRD, May 2012

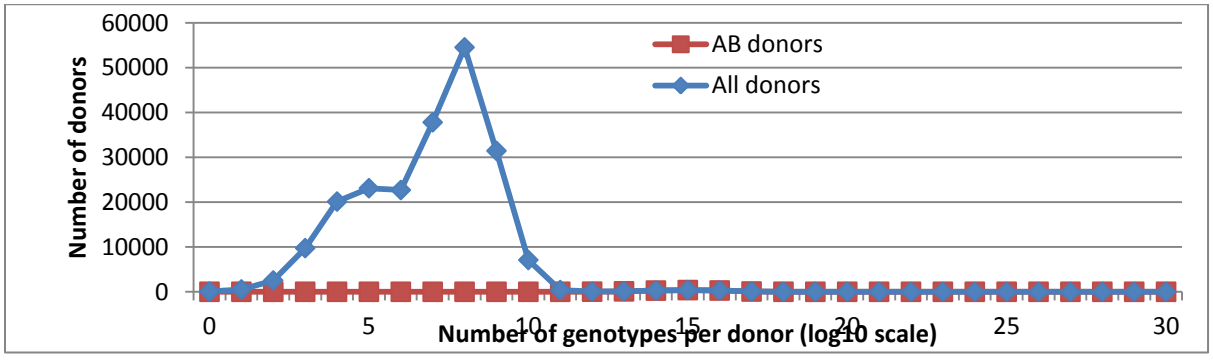


Figure 12: Visualization of the HLA typing ambiguities and computational complexity in DKMS Polska, May 2012 [67]

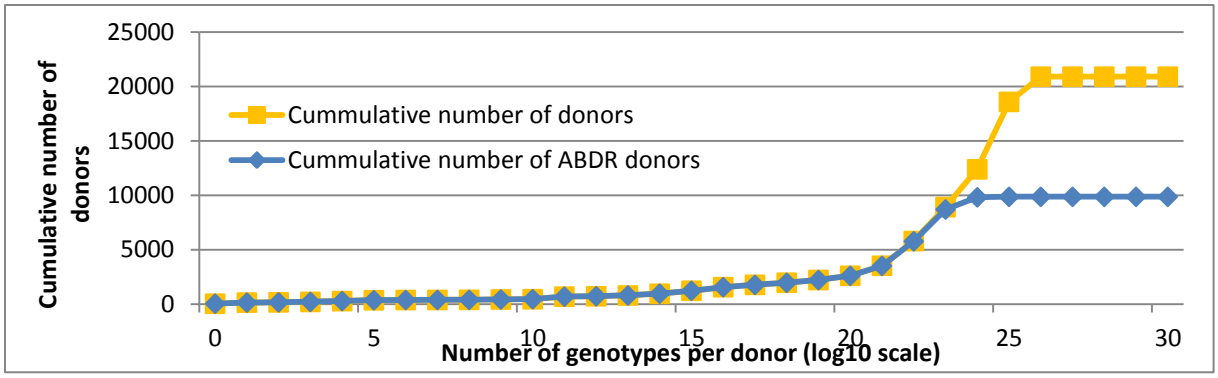


Figure 13: Visualization of the HLA typing ambiguities and computational complexity in CSCR, May 2012

The graph shows the most of the donors have $\tilde{c}_j \geq 10^{10}$. There are only relatively few donors with reasonable number of genotypes.

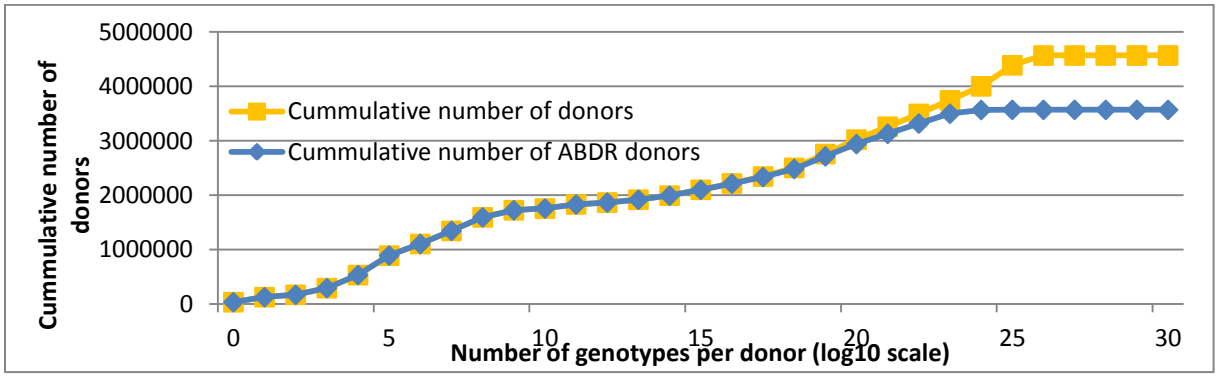


Figure 14: Visualization of the HLA typing ambiguities and computational complexity in ZKRD, May 2012

The graph shows different the ZKRD registry has much more donors that are better typed than CSCR. There are more than 500 000 donors with $\tilde{c}_j \leq 10^5$.

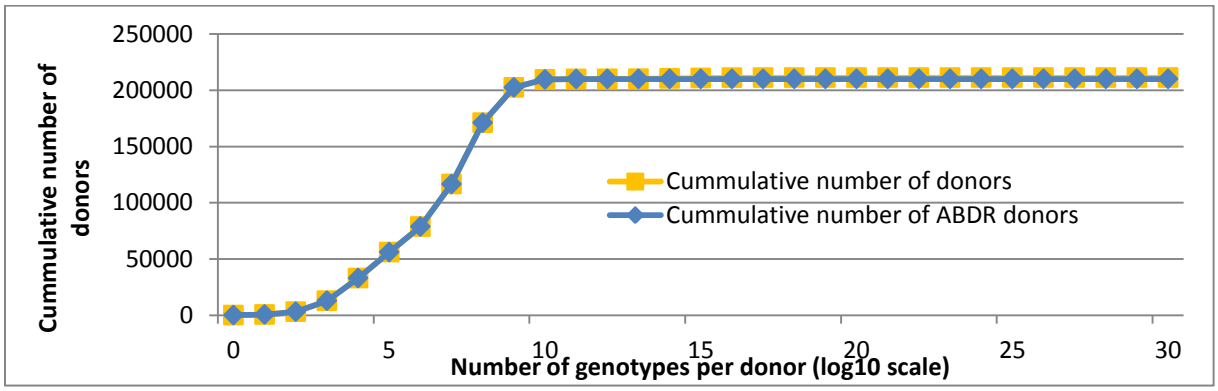


Figure 15: Visualization of the HLA typing ambiguities and computational complexity in DKMS Polska, May 2012

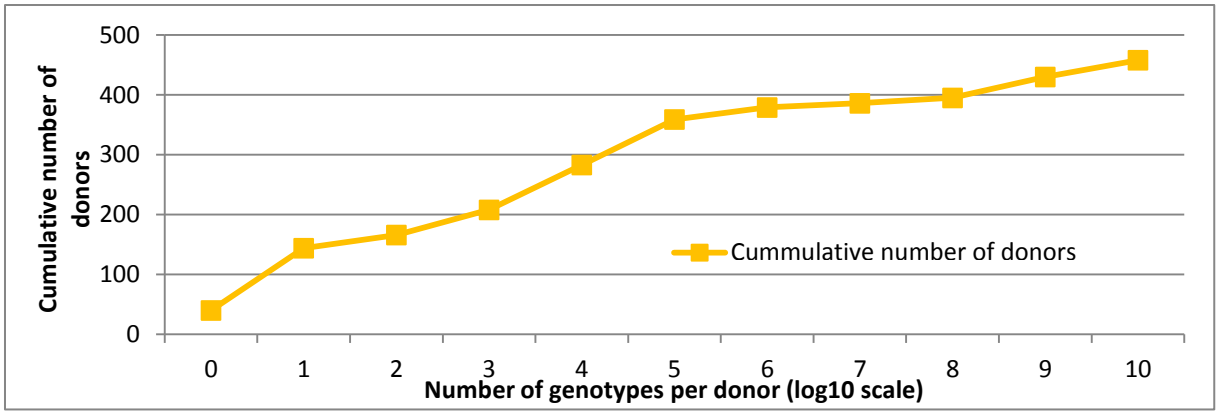


Figure 16: Visualization of the HLA typing ambiguities and computational complexity in CSCR, May 2012 (extract from previous graph)

The graph shows in detail all donors with reasonable level of HLA ambiguities. Only few hundred donors are relevant for HFE algorithm (5 loci, high resolution).

Ambiguity rank of the dataset

If we calculate the mean of $\lg(\tilde{c}_j)$, we get interesting **ambiguity rank** of the whole dataset.

(25)

$$R = \frac{\sum_{j=1}^{n} \lg(\tilde{c}_j)}{n}$$

Following table shows comparison of the datasets.

Dataset / registry	*R*
CSCR, May 2012	23.3
ZKRD, May 2012	15.3
DKMS Polska, May 2012	7.3

Table 13: Ambiguity rank of selected registries

6.2 Typing ambiguities

Previous graphs show extremely big computational complexity of the HFE problem on real registry data. Donors with high level of ambiguity (> 10^{10}) do not bring a lot of specific information about **two** underlying haplotypes, because these haplotypes are "hidden" in the set of all compatible genotypes (up to 10^{27}). HFE benefit of such donors is very poor, but they bring extreme increase in the computational expenses.

Since 2008, we participate in the Registry Diversity Subcommittee of the Information Technology Working Group of the World Marrow Donor Association (WMDA). The group, lead by Martin Maiers (USA), Steven GE Marsh (UK) and Carlheinz Muller (Germany) is a great platform for discussion, research and development of HFE methods. [56]

We have compared different programs for HLA haplotype frequency estimation in a controlled data environment. Simulated data set of the same sample size (100 000 individuals) contained the same donors, but with different proportions of typing ambiguities.

The work, summarized below, was presented at the 15th IHIWS conference [66]. Our HFE implementation has number 1.

COMPARING DIFFERENT PROGRAMS FOR HLA HAPLOTYPE FREQUENCY ESTIMATION IN A CONTROLLED DATA ENVIRONMENT

Hans-Peter Eberhard[1], Marie-Lorraine Balère[2], Pierre-Antoine Gouraud[3], Loren Gragert[4], Hazael Maldonado-Torres[5], David Steiner[6], Henk van der Zanden[7], Martin Maiers[4], Steven GE Marsh[5], Carlheinz R. Mueller[1]

[1] ZKRD Zentrales Knochenmarkspender-Register Deutschland, Ulm, Germany, [2] France Greffe de Moelle, Agence de Biomedécine, Paris, France, [3] Unité INSERM 558, Faculté de médecine, F-31073 Toulouse, France, [4] National Marrow Donor Program, Minneapolis, USA, [5] Anthony Nolan Research Institute, London, UK, [6] Czech Technical University, Prague, Czech Republic, [7] Europdonor Foundation and Dept. of Immuno-Hematology and Bloodtransfusion, Leiden University Medical Centre, Leiden, Netherlands

Introduction

Haplotype frequencies estimation (HFE) is an indispensable tool for many applications in biomedicine including the HLA domain. Therefore, the estimated haplotype frequencies (HF) must be reliable, valid and the many variables influencing their accuracy should be known qualitatively and quantitatively.
The two studies presented here are working steps towards validated tools for population analysis specifically designed to address issues of datasets from large donor registries.

Materials and Methods

We have compared the results from six different implementations of the Expectation Maximisation (EM) algorithm when applied to two different tasks with known ideal results. In each task, the algorithms were challenged with 100 simulated datasets of 100,000 individuals randomly created using a set of 3938 haplotypes with defined frequencies modeled after the French population. The simulation process allows the possibility for rare haplotypes to be missing by sampling effects. The average number of haplotypes in a simulated data set is 3730, ranging from 3698 to 3759.

For task 1 samples were provided with complete HLA-A-B-DR typing in order to measure the estimation error introduced by the EM and to separate it from the sampling error.

For task 2 data sets with an increasing rate of missing HLA-DR typings were generated. The HLA-DR types in these sets were revealed in different proportions after undergoing simulated patient driven typing or random selection. In the data set designation below the first number indicates the percentage of random DR typing and the second the percentage of patient driven DR typing in the simulated data set.

Here, each data set was used for two estimations, the first including only donors with *complete* typing and the second including *all* donors allowing the EM to handle missing values (abbreviated with *c* and *a* below). For cross-validation, a data set with complete DR typing was analysed in the same way as in task 1.

For the error quantifications we used the established identity coefficient I transformed into a distance measure D = 1 – I with

$$I = \sum_i \min(f_i, g_i) = 1 - \frac{1}{2} \times \sum_i |f_i - g_i|$$

$$D = \frac{1}{2} \times \sum_i |f_i - g_i|$$

where the range of i is the union of all haplotypes in the vectors f and g. For the exploratory data analysis of task 2 we have used the individual frequency differences between estimation and sample for each data set plotted in the order of their frequency.

Results

Task 1

As expected from an earlier study for the 14th IHIWS, task 1 did not reveal serious discrepancies in the results of the different EM implementations but helped to remove some subtle errors and allowed for further calibration of the individual tools. The average distance between the frequencies of the haplotypes actually used in the 100 samples and the theoretical values is 0.037 (min: 0.035, max: 0.038).

Results (continued)

The average distance between the estimated frequencies and the real frequencies in the samples is 0.035 (0.034 - 0.037), and the average total estimation error i.e. the distance of the estimated from the true theoretical frequencies is 0.05 (0.035 - 0.054). So the relation between sampling and estimation error is not additive (figure 1).

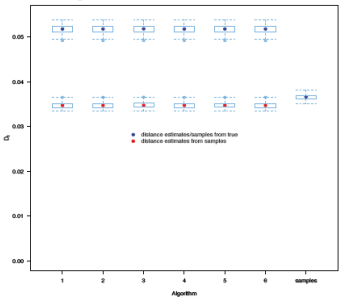


Figure 1: Distances of estimates/samples from theoretical (true) frequencies and distances of estimates from samples for task 1.

The log likelihood (LLH) of the frequencies obtained by the six implementations also did not differ substantially. The average difference between the maximum and minimum of the LLH is 0.59 which reflects a distance between the estimates of about 10^{-4}. The biggest fraction of the LLH differences can be accounted to differences of the stopping criteria, output precision and starting values used.

Task 2

Task 2 revealed the problems missing typing data can create for implementations of the EM although the ability of dealing with incomplete information is one of its major features. Only four algorithms fully completed task 2, one algorithm reported only results for about 10 of the 100 simulations in the missing data situation. The cross-validation results were consistent with task 1 for all algorithms.

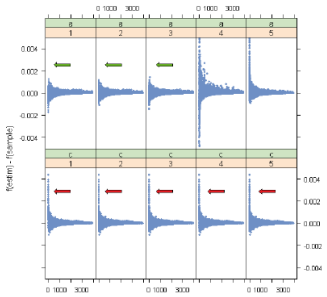


Figure 2: Differences of estimates from the sample frequencies for the 100 simulations with 20% patient driven DR-typing. Haplotypes ordered by sample frequency.

Results (continued)

Figure 2 depicts the difference between the estimated and true frequencies from the biased (bottom; c) and full (top; a) data sets in the most complicated situation: no random typing and only 20% DR patient driven typing. It demonstrates that algorithms 1-3 efficiently make use of the full data set to compensate the striking bias shown by the large discrepancies on the left (frequent haplotypes) of all five bottom plots (see the five red and three green arrows). The analysis of the other DR ratios showed the same picture but less blur especially for the frequent haplotypes.

The overall D comparison in figure 3 summarises the findings so far and additionally illustrates the effect of patient driven typing. Apart from an outlier of algorithm 3 for simulations 1 to 69 of task 0-60c, it becomes clear, that the trials with a random part yield better results for any given DR ratio. Algorithm 4 is censored in figure 3 since the spreading of the differences seen in 4a of figure 2 finds its continuation in the distance analysis and would show far outside the plotting area.

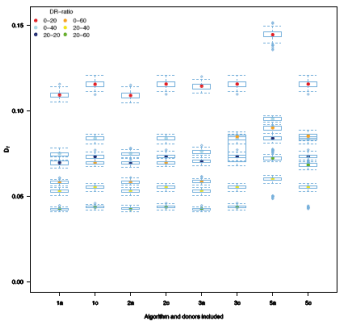


Figure 3: Distances of estimates from samples for 100 simulations and all 6 DR-typing ratios tested.

Discussion

Our analysis allows us to measure the quality and validity of HFs and provides insight into intrinsic properties of EM. This poster describes the current interim results of this 15th IHIWS component while the work is still continuing. The experience of the group from the last workshop shows that by refining the individual programs a broad consensus on the best practice will be achieved. In particular, we are working on general recommendations for the implementation and application of the EM for HFE in HLA and related systems.

A further analysis with random DR typing only would be interesting with regard to bias and estimation error due to total DR rate. The results of these studies will be the basis to investigate more complex situations such as deviation from Hardy-Weinberg equilibrium, population substructures and molecular HLA data with varying resolution.

Correspondence

Carlheinz R. Müller, MD, PhD
Phone: +49-731-1507-00
Fax: +49-731-1507-51
E-Mail: Carlheinz.Mueller@zkrd.de

6.3 Population and sample size

Let's now focus on sample size and its influence on the reliability of HFE. For this purpose, we have done the following experiment (see also chapter 4.7.2):

- Generate population of N individuals (genotypes). Calculate "**Population** HF".
- Simulate the registry by sampling the population. Take random subsets of 500, 1000, 2000, 4000, etc. individuals. Calculate "**Sample** HF".
- Convert genotypes to phenotypes (hide phasing information). Estimate HFE, using the sample by **EM** algorithm.
- Compare distance **(22)** between HFE of the EM algorithm, "Sample HF" and "Population HF".

Results of these experiments are shown in the following graphs.

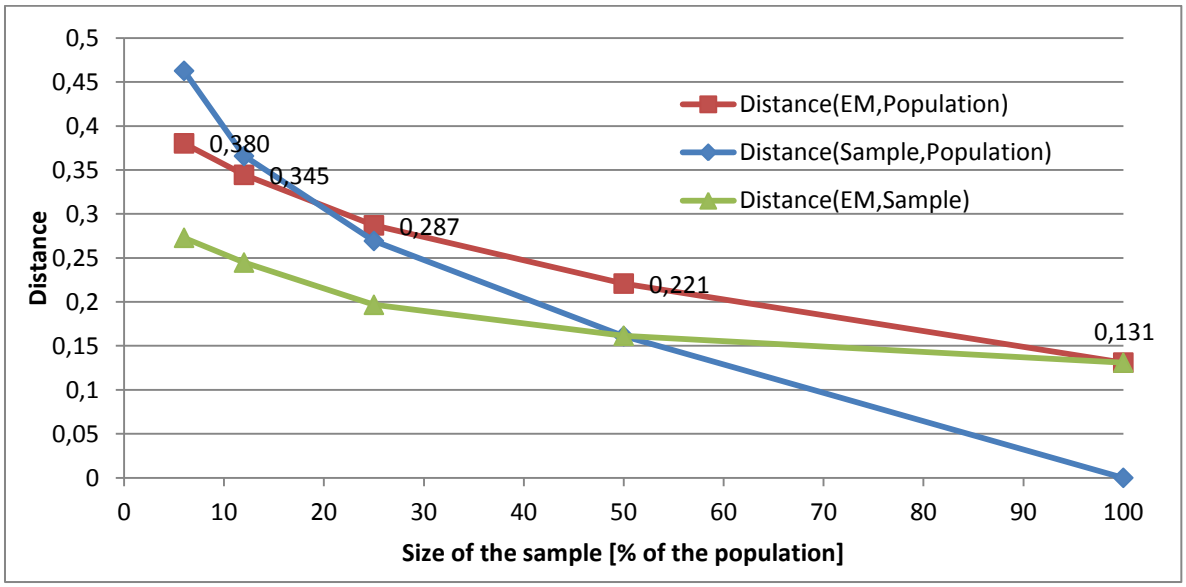


Figure 17: Sample size and reliability of HFE: Artificial population of 8 000 individuals based on [HPE-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

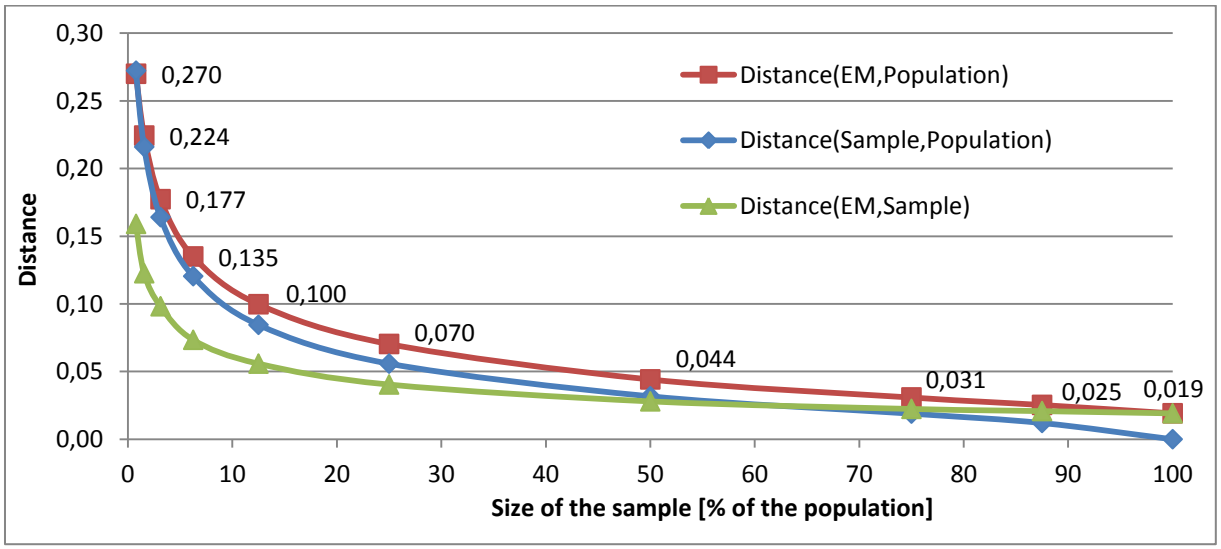


Figure 18: Sample size and reliability of HFE: Artificial population of 512 000 individuals based on [HPE-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

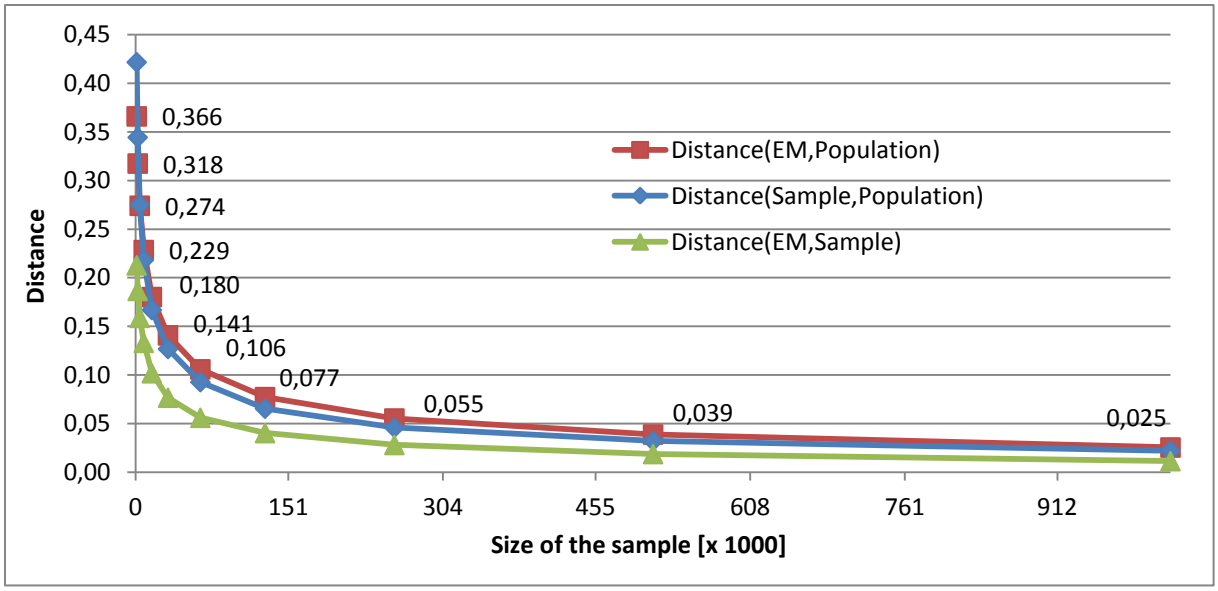


Figure 19: Sample size and reliability of HFE: Artificial population of 10 000 000 individuals based on [HPE-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

This graph simulates population of 10 million individuals, similar size like the population of the Czech Republic and other Central European countries. The experiment gives us very good understanding of the sampling error of the small to middle size stem cell donor registry. The sampling error of all donors recruited in the Czech Republic (less than 100 thousand donors) is more than 0.1.

We can also compare HFE of the EM algorithm (on the sample) and the Sample HF.

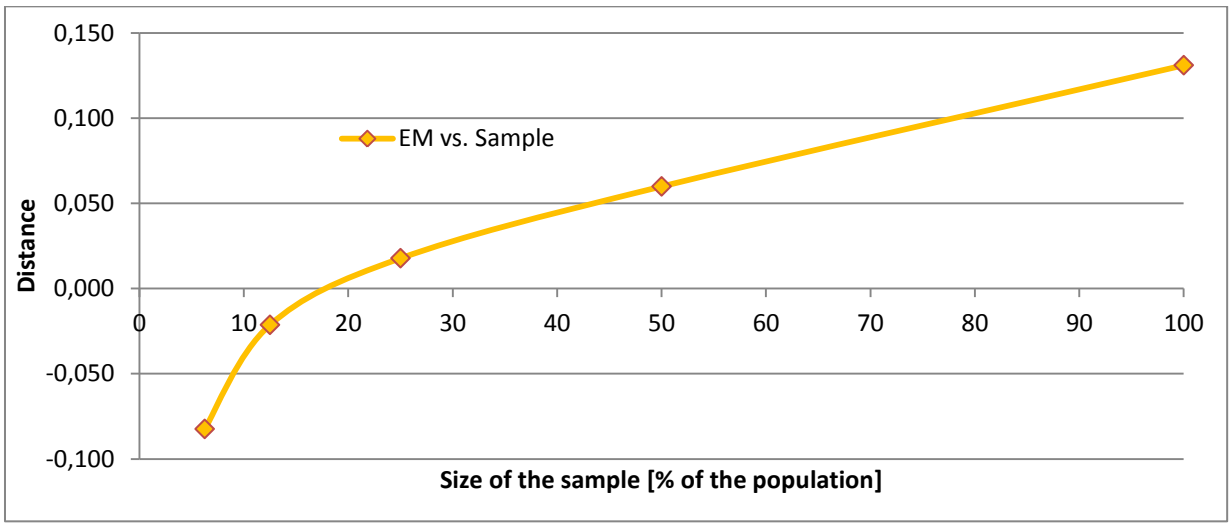


Figure 20: Comparison of HFE and the sample HF: Artificial population of 8 000 individuals based on [HPE-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

In general, the sample itself outperforms the EM algorithm, especially with growing sample size. When the sample size reaches 100% of the size of the population, there is no sampling error, because the sample contains the whole population. But the beginning of the curve might bring unexpected (and unreliable) results. With small sample sizes (up to 17% of the population), the EM algorithm may outperform the sample itself. This paradox could be observed mainly in small populations. If we increase the size of the population, we get the following result.

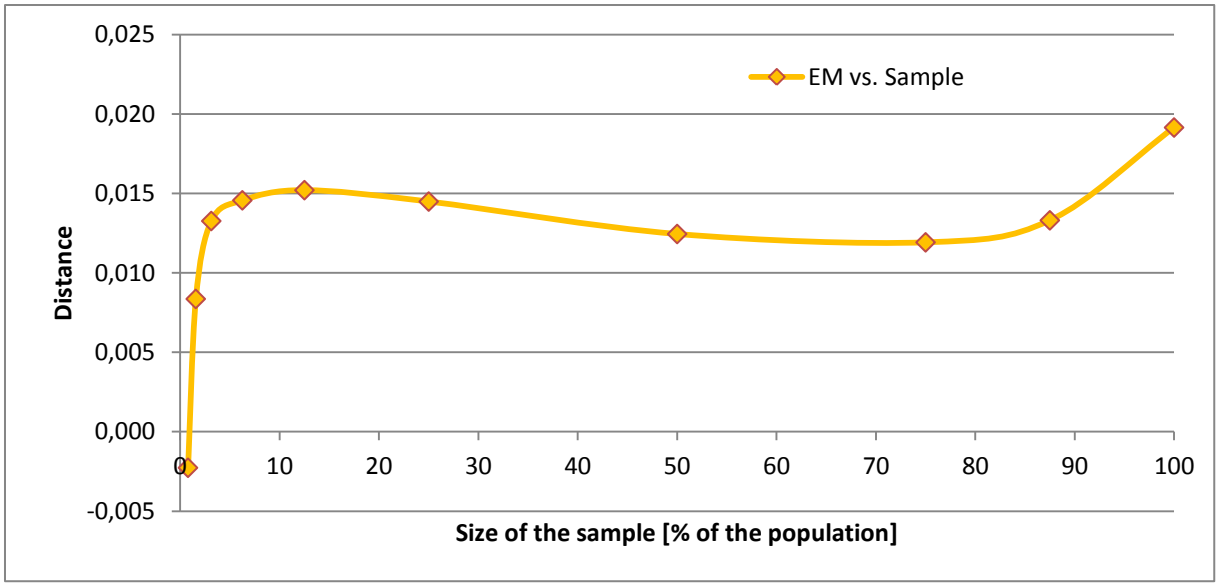


Figure 21: Comparison of HFE and the sample HF: Artificial population of 512 000 individuals based on [HPE-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

For small sample sizes, the EM algorithm may be still slightly better than the sample itself, but only until the sample size reaches about 1% of the population. For the population of the size of the Czech Republic, this drops to 0.05% of the population size.

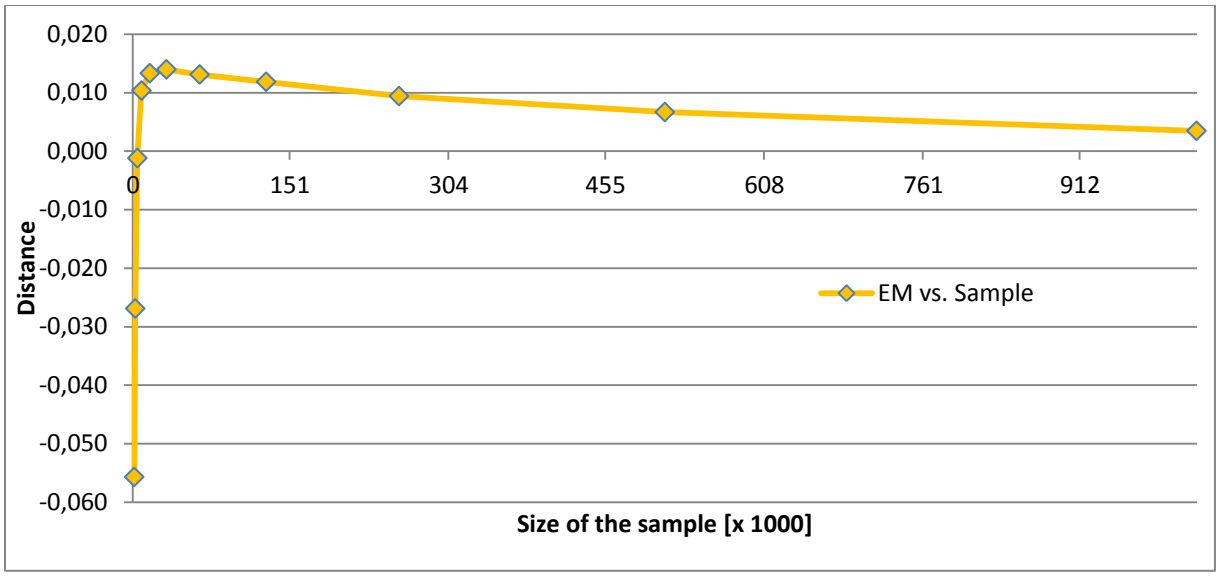


Figure 22: Comparison of HFE and the sample HF: Artificial population of 10 000 000 individuals based on [HPE-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

This behavior on very small sample sizes looks strange, but could be explained. Let's imagine extreme case, a sample of single heterozygous individual. In fact, there are two haplotypes and estimations of population frequencies of these two haplotypes are 0.5. Since true haplotype frequencies are close to 0, the distance of these estimates from true frequencies is almost 1. The EM algorithm does not know these two correct haplotypes, so in case of 5 loci typing, it will consider 16 haplotypes with frequencies 0.0625. Only two of them are correct and there is a high chance their

true frequency is less than 0.0625, so the algorithm overestimates frequencies of these two haplotypes. But there is a quite good chance at least one of remaining 14 haplotypes exists in the population. Then EM algorithm finds a haplotype that does not exist in the sample, but exists in the population and the overall HFE is better estimates of the sample itself.

We have discussed this topic with Carlheinz Muller which results in two additional comments:

- Observations for small sample size depends on ignoring confidence intervals which are extremely wide in such cases. Small sample sizes have big sampling error and therefore observations related to such samples are not reliable.
- *"The major drawback of EM is that it incorrectly works on a continuous instead of a discrete number space. In a sample, all allele or haplotype counts must be integers and the maximum should only be sought within such an integer valued domain. … anything depending on seriously ignoring this constraint refers to artifacts or useless or unreliable numbers produced by this algorithm. This refers in particular to the accuracy of estimates and the low-frequency estimates (low = "count in the sample < 3")."*

6.4 Population homogeneity

All experiments in the previous chapter were done using artificial population based on [HPE-2010]. But other populations, represented by other HF sets, are more homogenous (see Appendix A).

To test this influence, we have generated several artificial populations using different datasets in the Appendix A. We have found out the HFE depends on the population homogeneity – higher homogeneity of the population results in better HFE. The following graph shows the extreme case of artificial population based on [FI-2010]. HFEs are 2-10x better than those shown on the Figure 18.

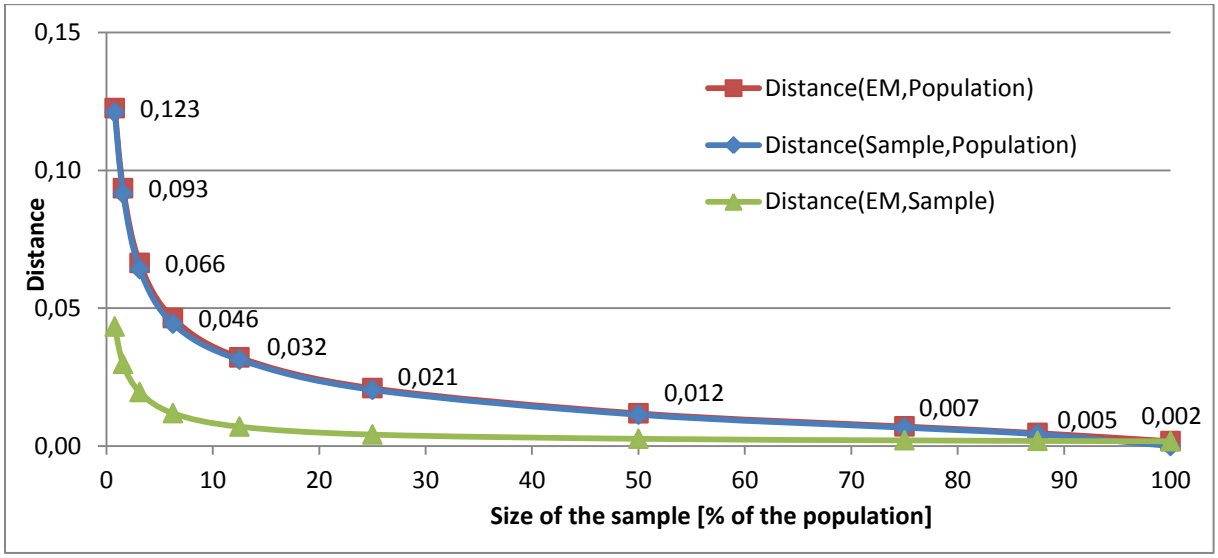


Figure 23: Sample size and reliability of HFE: Artificial population of 512 000 individuals based on [FI-2010], five loci high resolution typing (A-B-C-DRB1-DQB1).

We expect the Czech population is slightly more homogeneous than the German population (see Appendix A). This is probably caused by smaller population and country size. It means simulations of Czech HFEs could be done using [HFE-2010] and our conclusions are the same or slightly worse than the reality (e.g. Figure 19). This means we are on the save side.

6.5 Computational complexity

As discussed earlier, donors with high level of ambiguity (> 10^{10}) do not bring a lot of specific information about two underlying haplotypes, but bring extreme computational complexity. Can we exclude them? What is the influence on the HFE?

In order to simulate this dependency, we have selected all German phenotypes [BMDW-201205] that are at least intermediate resolution typed at loci A*, B*, C*, DRB1* and DQB1*. There were 380567 of such records. We have sorted them by growing $\tilde{c}_j$. Then, a subset of N first records was selected, HFE was performed and results were compared to [HFE-2010].

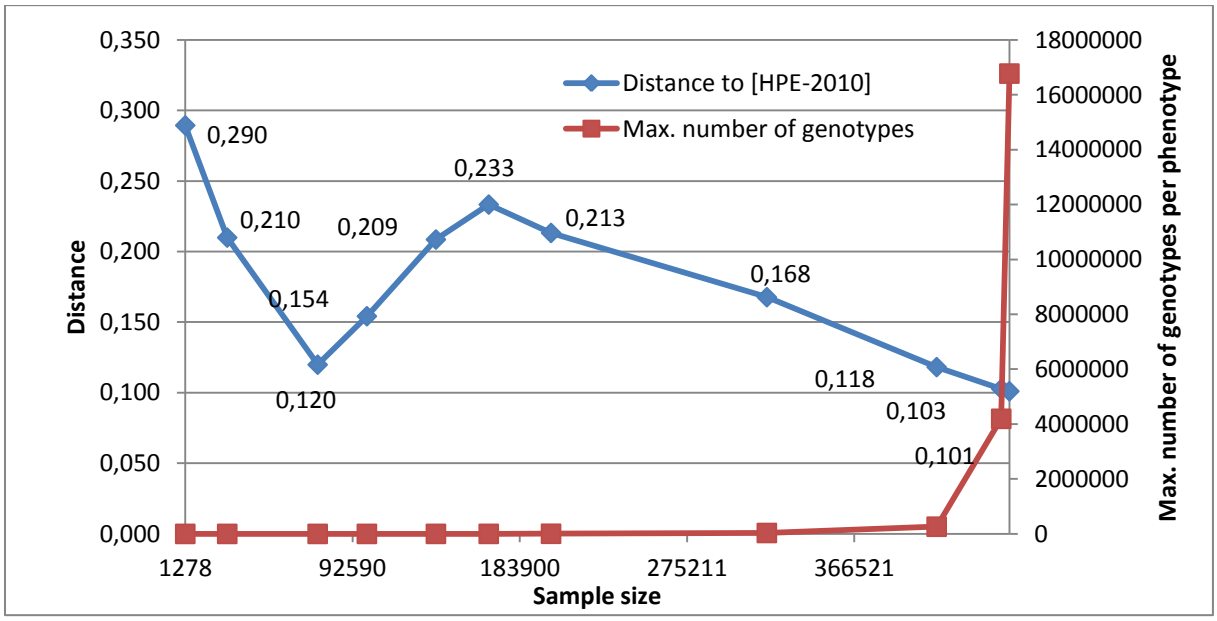


Figure 24: Growing sample size, computational complexity vs. reliability of HFE. Used data: the ZKRD registry (May 2012), at least intermediate resolution typing (A-B-C-DRB1-DQB1), 5 loci high resolution HFE, reference haplotype frequencies [HPE-2010].

The Figure 24 shows results of the simulation on ZKRD registry data. First estimate uses just about 1 300 donors who are high resolution typed and homogeneous ($\tilde{c}_j = 1$). The third estimate uses all high resolution typed donors ($\tilde{c}_j \leq 16$, about 90 000 individuals) and the estimate is very good. Mixture of high and intermediate resolution typed donors increase the distance from [HPE-2010], but with growing sample size, the distance gets closer and closer to [HPE-2010]. However, computational costs (time and memory) grow exponentially, so at final stage we managed to include 451 190 donors with algorithm running time 7,5 hours (PC with Windows 7 Professional SP1 64bit, Intel Core i3-2120 CPU @ 3.30 GHz, 16 GB RAM) and the distance to [HPE-2010] was just 0.1.

6.6 Simulation of real dataset

We have seen the reliability of HFE algorithm depends on several factors, such as typing ambiguities of registry donors, computational complexity (and limitations of hardware), population size, sample size and population homogeneity. But real live registry dataset has combination of all

these factors. What is the reliability of HFE for a real population? Especially, what is the reliability of HFE for the Czech population?

In order to simulate the reliability of HFE on a registry dataset, we need to have similar data in a controlled data environment.

Therefore, we have designed and run this complex simulation:

1. **Population homogeneity** (see chapter 6.4): Take appropriate high resolution HF, with similar homogeneity as real population. These are "background haplotype frequencies".
2. **Population size** (see chapter 6.3): Generate the artificial population - create individuals according to the population model (HFE). As a result, we have phase-known population and its "true haplotype frequencies". Size of the artificial population will be the same as the real population.
3. **Sample size** (see chapter 6.3): Simulate the recruitment process - do the sampling of the artificial population. Sample size will be the same as the registry dataset. We get "sample haplotype frequencies". Hide the phase information in the sample, i.e. convert genotypes to phenotypes. Every real donor has corresponding artificial donor in the simulated dataset (donor pair).
4. **Typing ambiguities** (see chapters 6.1 and chapter 6.2): For every real donor, analyze the typing ambiguity. Simulate the HLA typing of the corresponding artificial donor to the similar level of typing ambiguity as the real donor. We get simulated dataset.
5. **Computational complexity** (see chapter 6.5): Estimate haplotype frequencies on the simulated dataset ("estimated haplotype frequencies") using the same techniques, algorithms and heuristics as on the real registry dataset.
6. **Reliability of HFE**: Count the distance (22) between "estimated haplotype frequencies" and "true haplotype frequencies". This is also approximation of the reliability of HFE of the real registry dataset. If "estimated haplotype frequencies" do not contain all loci as "true haplotype frequencies" or some of these loci are not estimated at high resolution level, we need to convert "true haplotype frequencies" to the same resolution as "estimated haplotype frequencies", before the distance can be counted.

The first step is difficult, because we need to take some HFs with similar homogeneity as real population. But we may not know precisely the homogeneity of the real population. As discussed in the chapter 6.4, it is better to take HFs of a population with lower homogeneity than the real population.

But the trickiest is the fourth step that has to be done very carefully. Artificial donor virtual HLA typing process must maximally correspond to real donor HLA typing techniques. But the artificial donor is different individual (from different population) than the real donor, which complicates this step.

We can do virtual intermediate resolution typing by applying commercial SSOP typing kits and their characteristics. This technique has been implemented by NMDP (not published). A problem could be selection of the vendor, since we don't know by what typing technique (serology, SSP, SSO, SBT) and what typing kit the real donor was typed.

We have implemented different approach that will be demonstrated by the following example.

6.6.1 Example: Simulation of the CBB Czech Republic

In this example, we will simulate HFE of the Czech population, using the real dataset of the Cord Blood Bank Czech Republic. Simulation steps are:

1. We take German population and [HPE-2010] as background haplotype frequencies. Germany is neighbor country, has the biggest registry in Europe and both populations are Caucasian. We expect the homogeneity of the German population is lower than the homogeneity of the Czech population, because Germany is about 8x bigger country. This is also confirmed by HFE (see the Appendix A).
2. The Czech population has about 10 million people (May 2012), generate artificial population of 10 million individuals.
3. The CBB has less than 4000 CBUs (May 2012). Simulate recruitment process of 4000 individuals.
4. (A) Replace artificial (German) donor by reference donor phenotype in the (German) registry.
    i. Select all donors in the reference (German) registry [BMDW-201205] with no HLA mismatch [7] (HLA-A, -B, -C, -DRB1, -DQB1) against the artificial donor.
    ii. In the set of these donors, find a donor with the most similar typing ambiguity as the real donor (CBU) in the simulated dataset (CBB Czech Republic) - take the one with the smallest absolute distance of $\widetilde{c}_j$ between reference (German) donor and real CBU. This reference donor has our simulated HLA typing of the artificial donor.
5. Estimate HF of the simulated dataset.
6. Count distance between “estimated haplotype frequencies” of the simulated dataset and “true haplotype frequencies” of the artificial population.

By this approach we get following key properties of the simulated dataset:

1. Similar population homogeneity, maybe little bit more pessimistic than the reality.
2. Same population size.
3. Same sample size.
4. Similar typing ambiguities ($\widetilde{c}_j$), based on real HLA typing techniques.
5. Similar computational complexity, see Figure 25.
6. Similar reliability of HFE, see Figure 26.

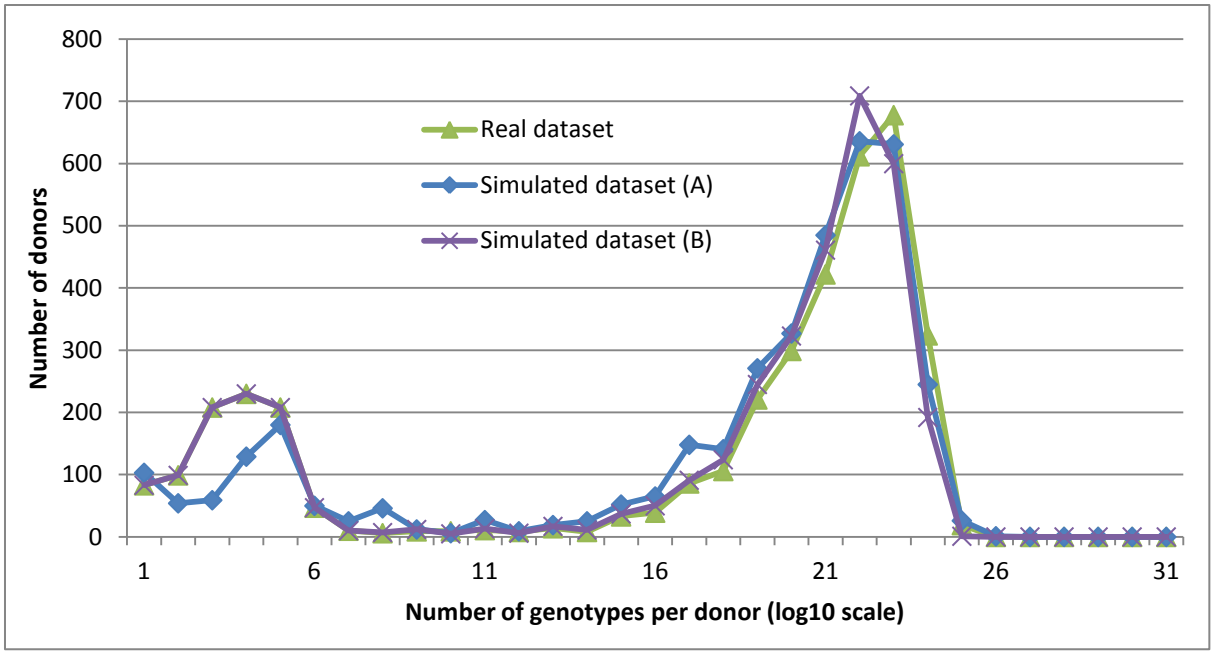


Figure 25: Simulation of the real registry (Cord Blood Bank of the Czech Republic) by artificial population (based on German HFE) and virtual recruitment and virtual donor typing. Used data: the ZKRD registry (May 2012), [HPE-2010], CBB Czech Republic (May 2012). 5 loci high resolution genotypes (A-B-C-DRB1-DQB1).

The sampling error of 4000 individuals in our artificial 10 million population is 0.275. The HFE algorithm is limited mainly by computational complexity, so not all donors could be considered in the estimation. The Figure 26 shows dependency between number of donors considered by HFE algorithm and reliability of haplotype frequency estimates.

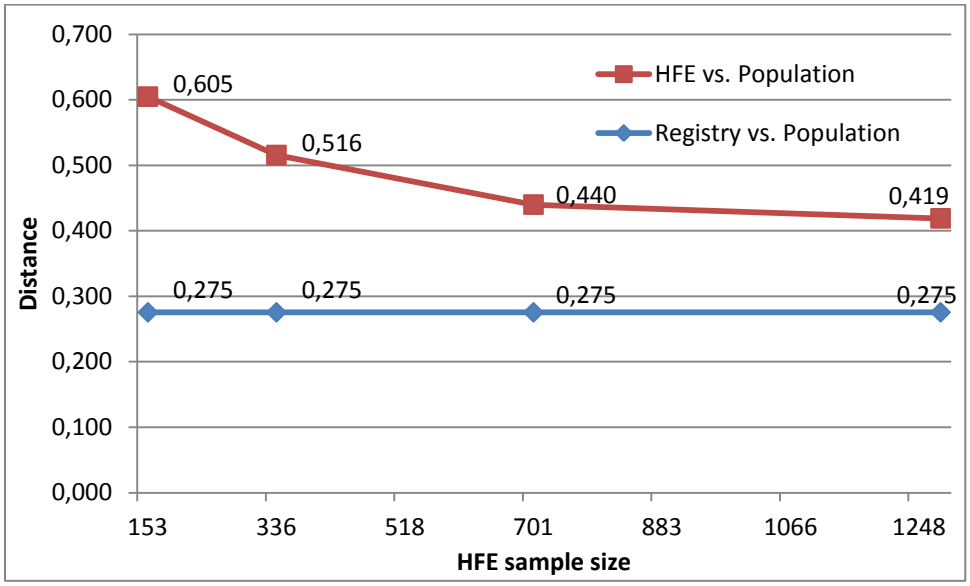


Figure 26: Simulation of reliability of HFE of the Cord Blood Bank of the Czech Republic (May 2012).

The implementation (A) of the step 4 in this example does very realistic virtual HLA typing of the artificial donor and it keeps typing relations between loci as it was done by real HLA typing techniques – for example using ABDR SSP typing kits.

However, this approach has also some drawbacks. It may lead to replacement of low resolution ABDR typed donor by intermediate resolution AB typed donor with the same $\widetilde{C}_j$. We can improve the virtual HLA typing by searching only ABDR typed donors, if the donor was ABDR typed and other similar improvements, but it would be difficult to cover all possibilities and exceptions.

Other option would be to "type" loci individually, which is also common practice in the HLA laboratories that use typing kits focused only on one locus. We can also do virtual HLA typing at each locus independently. This means we need several real donors to simulate HLA typing of one artificial donor.

The alternative implementation (B) of the step 4:

4. (B) For every locus (HLA-A, -B, -C, -DRB1, -DQB1), simulate the HLA typing process by replacing artificial donor typing by reference donor type at the loci.
    i. Select all donors in the reference (German) registry [BMDW-201205] with no HLA mismatch [7] **at the locus** with the artificial donor.
    ii. In the set of these donors, find a donor with the most similar typing ambiguity **at the locus** as the real donor (CBU) in the simulated dataset (CBB Czech Republic) - take the one with the smallest absolute distance of $\widetilde{C}^L{}_j$ between reference (German) donor typing at the locus and real CBU typing at the locus. This reference donor has our simulated HLA typing of the artificial donor at the locus.

Both approaches (A) and (B) are extremely computationally demanding and such simulation takes several days. We have to:

- Analyze and calculate length of the genotype lists for all donors in both the real and the reference dataset. In our case, it means more than 4 million donors for approach (A) and more than 20 million loci for approach (B).
- Run the search for all donors of the simulated dataset in the reference dataset. For the registry like in our example (4000 donors only), it means to run 4000 donor searches in the file of 4 million donors. For approach (B) it is even 20 000 donor searches in the reference dataset. These results must be sorted by decreasing smallest absolute distance of genotype list length, which is also not trivial procedure.

Simulation by approach (B) is more demanding, but gets better results, especially for better typed donors (see Figure 25). This is as expected – it might be difficult to find well typed reference donor, HLA compatible with the artificial donor. However, if we search by individual loci, it is more likely we will find a well typed reference donor, matching with the artificial donor at selected locus.

7. Results of HFE on registry datasets

In this chapter we will present HFE of several populations, mainly in the Central Europe. Given a stem cell donor registry dataset, the goal is to estimate the "best possible" haplotype frequencies for the registry population. The "best" means:

- Maximum number of loci, highest possible typing resolution. Gold standard is the estimation of 5 loci high resolution haplotype frequencies (A-B-C-DRB1-DQB1).
- Maximum reliability of estimates, so they represent the whole population.

These two criteria go against each other – if we estimate higher resolution haplotype frequencies with more loci, the reliability will be lower than haplotype frequencies with lower resolution or with less loci.

7.1 Hungary

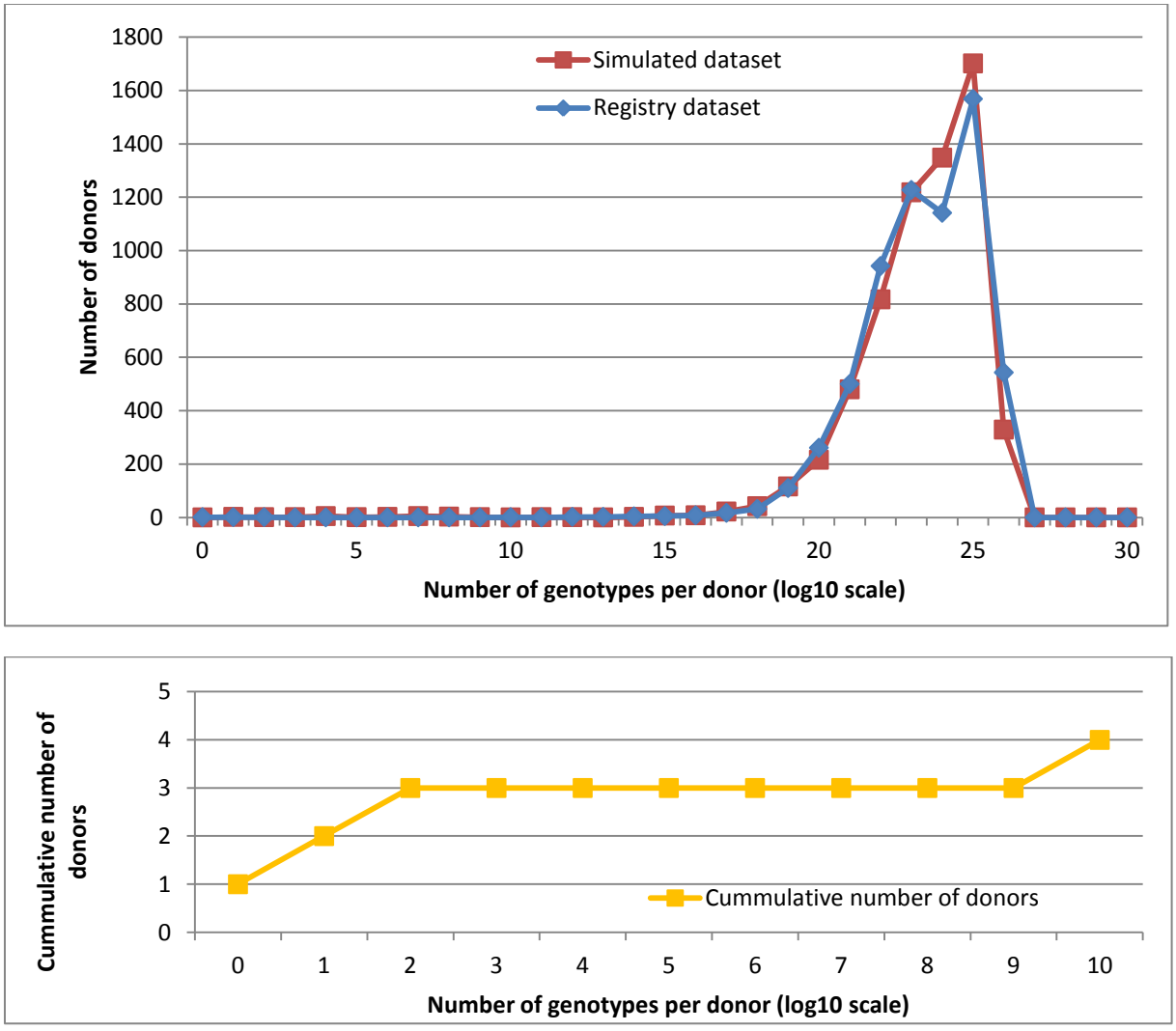


Figure 27: Visualization of the HLA typing ambiguities and computational complexity in the Hungarian registry: 5 loci high resolution haplotype frequencies (A-B-C-DRB1-DQB1), May 2012.

The Hungarian registry (May 2012) has 6366 active potential stem cell donors in the registry. There are almost no donors with small number of typing ambiguities, e.g. only 2 high resolution A-B-

C-DRB1-DQB1 typed donors (see Figure 27). Even estimation of HLA-A allele frequencies is not reliable, because only three high resolution typed HLA-A alleles can be found in the dataset (A*01:01, A*02:01 and A*03:01).

This means we cannot estimate high resolution allele and haplotype frequencies. But we can try to estimate low resolution A-B-C-DRB1-DQB1 frequencies.

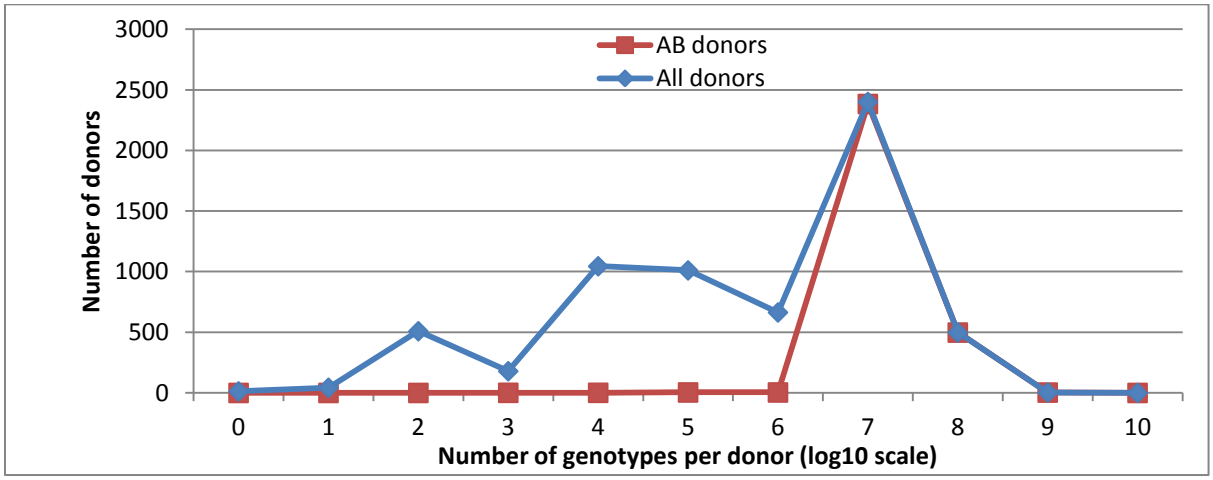


Figure 28: Visualization of the HLA typing ambiguities and computational complexity in the Hungarian registry, 5 loci low resolution haplotype frequencies (A-B-C-DRB1-DQB1), May 2012.

This is computationally feasible, but not reliable, since there are only 28 donors typed at all five loci by DNA typing techniques. HFE of the simulated dataset have distance 0.452 from the true frequencies (estimation of the reliability of 5 loci low resolution haplotype frequencies).

So finally, we can estimate low resolution ABDR (A-B-DRB1) haplotype frequencies. The registry has 3471 ABDR typed donors (54.5%), the rest is AB typed only.

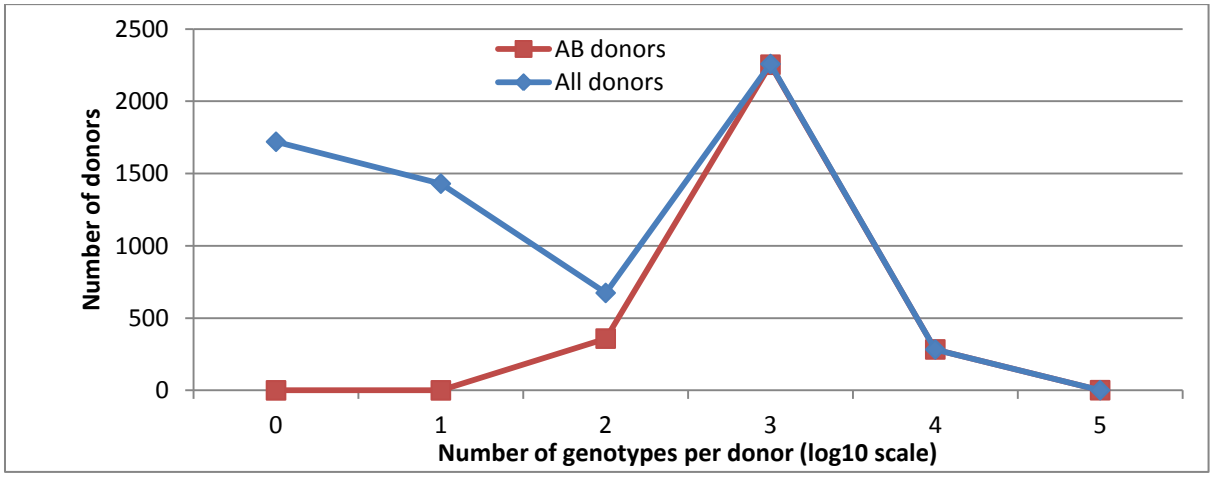


Figure 29: Visualization of the HLA typing ambiguities and computational complexity in the Hungarian registry, 3 loci low resolution haplotype frequencies (A-B-DRB1), May 2012.

Table bellow shows results of the HFE algorithm, considering all donors, including those AB typed only.

Rank	A*	B*	DRB1*	Frequency
1	01:XX	08:XX	03:XX	0.056816
2	02:XX	18:XX	11:XX	0.0157
3	02:XX	44:XX	04:XX	0.014903
4	02:XX	13:XX	07:XX	0.012188
5	02:XX	44:XX	16:XX	0.011933
6	02:XX	27:XX	16:XX	0.011915
7	02:XX	15:XX	04:XX	0.010172
8	03:XX	07:XX	15:XX	0.009535
9	03:XX	35:XX	01:XX	0,008859
10	02:XX	08:XX	03:XX	0,008491

Table 14: Most frequent ABDR low resolution haplotype frequencies of the Hungarian registry (May 2012).

The simulated datasets has average distance 0.13 from the population, which is also estimation of the registry sampling error for ABDR low resolution haplotype frequencies. HFE of the simulated datasets have avg. distance 0.324 from true frequencies – this is also rough estimation of the reliability of ABDR low resolution haplotype frequencies.

7.2 Slovakia

There are two registries in the Slovak Republic – one for adult donors (SK) and one public cord blood bank (SKCB). The adult donor registry has 3144 donors (May 2012) and the CBB has 1734 units (May 2012). Together, we have 4878 individuals and almost all of them are ABDR typed.

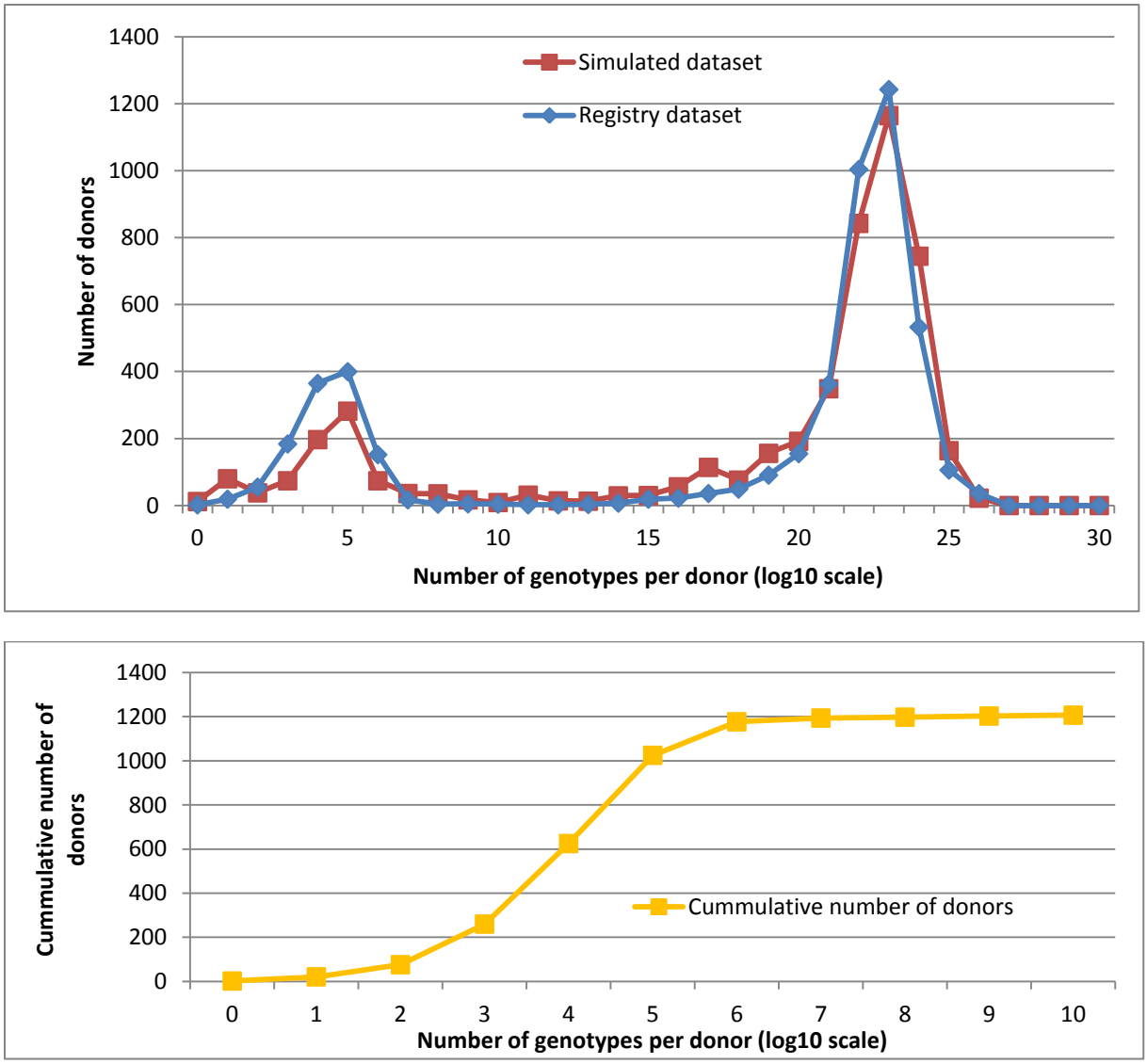


Figure 30: Visualization of the HLA typing ambiguities and computational complexity in the Slovak registries (SK, SKCB), 5 loci high resolution haplotype frequencies (A-B-C-DRB1-DQB1), May 2012.

As we can see from the graph, there are about 1200 very well typed donors. This number is already comparable with African American, Hispanic and Asian ethnic groups used in the HFE of the American study [62].

The simulated datasets have average distance 0.27 from the population, which is also estimation of the registry sampling error for A-B-C-DRB1-DQB1 high resolution haplotype frequencies. HFE of the simulated datasets have average distance 0.444 from true frequencies.

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	01:01	08:01	07:01	03:01	02:01	0,043228
2	03:01	07:02	07:02	15:01	06:02	0,025822
3	25:01	18:01	12:03	15:01	06:02	0,012463
4	02:01	07:02	07:02	15:01	06:02	0,012246
5	02:01	38:01	12:03	13:01	06:03	0,01157
6	02:01	44:02	07:04	16:01	05:02	0,010398
7	02:01	15:01	03:04	04:01	03:02	0,008965
8	02:01	44:02	05:01	04:01	03:01	0,007578
9	02:01	13:02	06:02	07:01	02:01	0,006094
10	02:01	13:02	06:02	07:01	02:02	0,006094

Table 15: Most frequent ABCDRDQ high resolution haplotype frequencies of the Slovak population (May 2012).

Computational complexity of the estimation of low resolution ABCDRDQ haplotype frequencies is shown on the Figure 31 and results are provided in the Table 16.

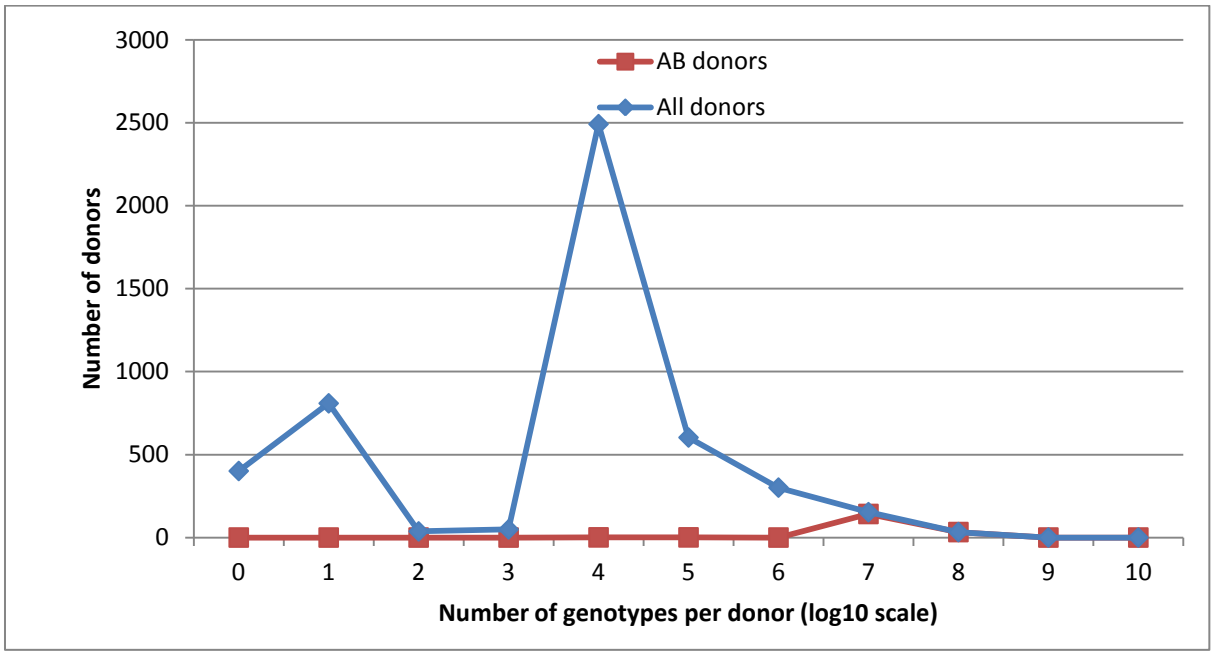
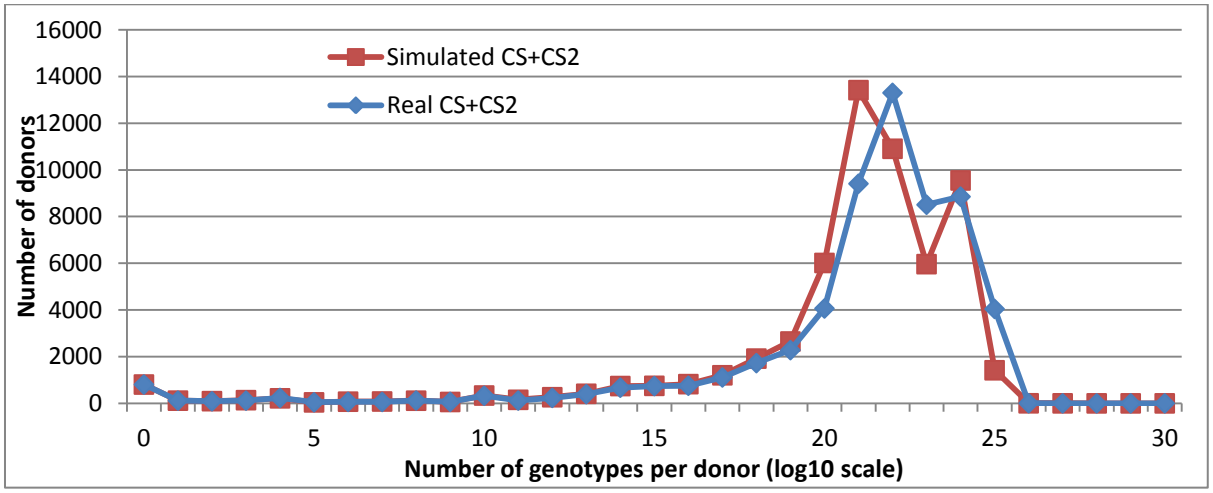


Figure 31: Visualization of the HLA typing ambiguities and computational complexity in the Slovak registries (SK, SKCB), 5 loci low resolution haplotype frequencies (A-B-C-DRB1-DQB1), May 2012.

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	01:XX	08:XX	07:XX	03:XX	02:XX	0,062804
2	03:XX	07:XX	07:XX	15:XX	06:XX	0,027063
3	02:XX	18:XX	07:XX	11:XX	03:XX	0,017698
4	02:XX	07:XX	07:XX	15:XX	06:XX	0,015893
5	02:XX	44:XX	05:XX	04:XX	03:XX	0,014746
6	02:XX	15:XX	03:XX	04:XX	03:XX	0,012922
7	02:XX	13:XX	06:XX	07:XX	02:XX	0,012229
8	02:XX	38:XX	12:XX	13:XX	06:XX	0,011592
9	23:XX	44:XX	04:XX	07:XX	02:XX	0,011306
10	25:XX	18:XX	12:XX	15:XX	06:XX	0,010862

Table 16: Most frequent ABCDRDQ low resolution haplotype frequencies of the Slovak population (May 2012).

7.3 Czech Republic



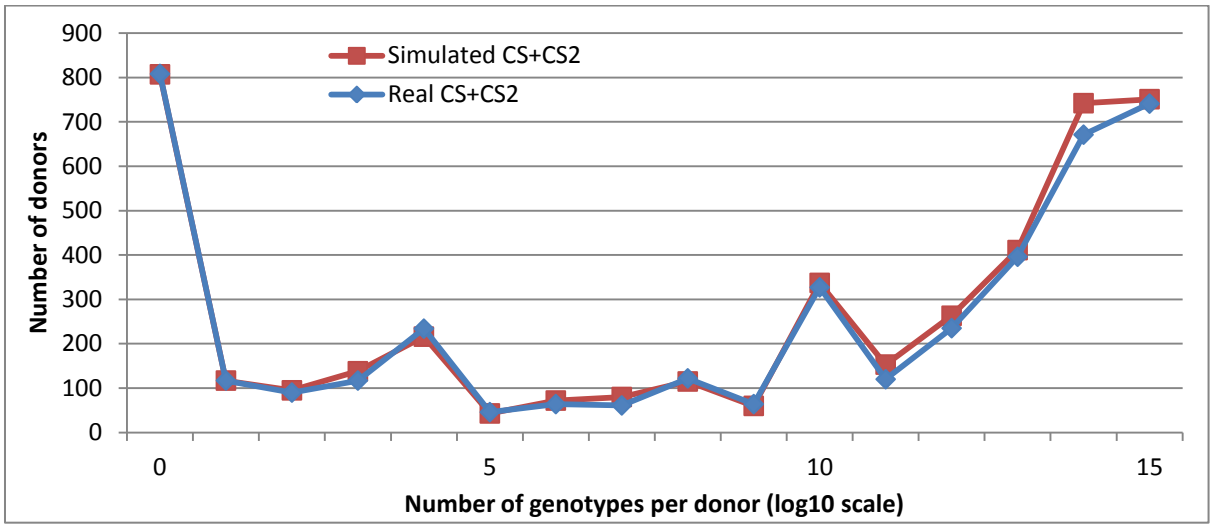


Figure 32: Visualization of the HLA typing ambiguities and computational complexity in the Czech registries (CS, CS2), 5 loci high resolution haplotype frequencies (A-B-C-DRB1-DQB1), May 2012.

There are two adult registries (Czech Stem Cell Registry and Czech National Marrow Donors Registry) and one public cord blood bank in the Czech Republic – together, they have 62 084 individuals (May 2012). We have already shown example of simulation of the CBB (see Figure 25).

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	01:01	08:01	07:01	03:01	02:01	0,074842
2	03:01	07:02	07:02	15:01	06:02	0,048162
3	02:01	13:02	06:02	07:01	02:02	0,022213
4	02:01	07:02	07:02	15:01	06:02	0,019257
5	01:01	57:01	06:02	07:01	03:03	0,014887
6	23:01	44:03	04:01	07:01	02:02	0,014544
7	03:01	35:01	04:01	01:01	05:01	0,01417
8	25:01	18:01	12:03	15:01	06:02	0,011151
9	02:01	44:02	05:01	04:01	03:01	0,010263
10	30:01	13:02	06:02	07:01	02:02	0,009327

Table 17: Most frequent ABCDRDQ high resolution haplotype frequencies of the Czech population (May 2012).

Average distance of high resolution HFEs of the simulated datasets to the true frequencies is 0.355.

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	01:XX	08:XX	07:XX	03:XX	02:XX	0,064548
2	03:XX	07:XX	07:XX	15:XX	06:XX	0,040355
3	02:XX	13:XX	06:XX	07:XX	02:XX	0,019092
4	02:XX	44:XX	05:XX	04:XX	03:XX	0,017204
5	02:XX	07:XX	07:XX	15:XX	06:XX	0,017202
6	23:XX	44:XX	04:XX	07:XX	02:XX	0,012991
7	02:XX	18:XX	07:XX	11:XX	03:XX	0,012938
8	03:XX	35:XX	04:XX	01:XX	05:XX	0,012271
9	02:XX	15:XX	03:XX	04:XX	03:XX	0,011751
10	01:XX	57:XX	06:XX	07:XX	03:XX	0,010672

Table 18: Most frequent ABCDRDQ low resolution haplotype frequencies of the Czech population (May 2012).

Average distance of low resolution HFEs of the simulated datasets to the true frequencies is 0.262.

The following results will be presented without simulated estimation of distance to the true frequencies. It is not clear whether simulation can be used for populations that are far from reference Caucasian population (north Europe, Cyprus, Africa, etc.). Finish population is much more homogeneous than reference German population, but there could be also other hidden problems (e.g. linkage disequilibrium).

7.4 Finland

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	03:01	35:01	04:01	01:01	05:01	0,115949
2	01:01	08:01	07:01	03:01	02:01	0,066685
3	03:01	07:02	07:02	15:01	06:02	0,043546
4	02:01	27:05	02:02	08:01	04:02	0,028982
5	02:01	07:02	07:02	15:01	06:02	0,028114
6	02:01	15:01	03:04	04:01	03:02	0,025415
7	03:01	07:02	07:02	13:01	06:03	0,023425
8	02:01	15:01	04:01	08:01	04:02	0,021335
9	02:01	15:01	03:03	13:01	06:03	0,020245
10	02:01	13:02	06:02	07:01	02:02	0,017724

Table 19: Most frequent ABCDRDQ high resolution haplotype frequencies of the Finnish population (May 2012, 980 donors used, FI and FICB datasets).

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	03:XX	35:XX	04:XX	01:XX	05:XX	0,096065
2	01:XX	08:XX	07:XX	03:XX	02:XX	0,051767
3	03:XX	07:XX	07:XX	15:XX	06:XX	0,036495
4	02:XX	15:XX	03:XX	04:XX	03:XX	0,027472
5	02:XX	07:XX	07:XX	15:XX	06:XX	0,026341
6	03:XX	07:XX	07:XX	13:XX	06:XX	0,02582
7	02:XX	27:XX	02:XX	08:XX	04:XX	0,023214
8	02:XX	15:XX	03:XX	13:XX	06:XX	0,021598
9	02:XX	13:XX	06:XX	07:XX	02:XX	0,020913
10	02:XX	15:XX	04:XX	08:XX	04:XX	0,016104

Table 20: Most frequent ABCDRDQ low resolution haplotype frequencies of the Finnish population (May 2012, 3356 donors used, FI and FICB datasets).

7.5 Sweden

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	01:01	08:01	07:01	03:01	02:01	0,053935
2	02:01	07:02	07:02	15:01	06:02	0,033879
3	03:01	35:01	04:01	01:01	05:01	0,026681
4	02:01	15:01	03:04	04:01	03:02	0,026362
5	02:01	40:01	03:04	13:02	06:04	0,021377
6	02:01	44:02	05:01	04:01	03:01	0,018612
7	03:01	07:02	07:02	15:01	06:02	0,01501
8	02:01	15:01	03:03	04:01	03:02	0,010864
9	02:01	40:01	03:04	01:01	05:01	0,009709
10	02:01	27:05	02:02	01:01	05:01	0,009526

Table 21: Most frequent ABCDRDQ high resolution haplotype frequencies of the Swedish population (May 2012, 812 donors used, S and SCB datasets).

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	01:XX	08:XX	07:XX	03:XX	02:XX	0,045838
2	02:XX	44:XX	05:XX	04:XX	03:XX	0,04313
3	02:XX	15:XX	03:XX	04:XX	03:XX	0,030709
4	02:XX	40:XX	03:XX	13:XX	06:XX	0,022565
5	02:XX	07:XX	07:XX	15:XX	06:XX	0,017791
6	03:XX	07:XX	07:XX	15:XX	06:XX	0,017476
7	03:XX	35:XX	04:XX	01:XX	05:XX	0,017084
8	29:XX	44:XX	16:XX	07:XX	02:XX	0,011913
9	02:XX	40:XX	03:XX	04:XX	03:XX	0,009044
10	02:XX	08:XX	07:XX	03:XX	02:XX	0,008687

Table 22: Most frequent ABCDRDQ low resolution haplotype frequencies of the Swedish population (May 2012, 3296 donors used, S and SCB datasets).

7.6 Cyprus

The Cyprus Bone Marrow Donor Registry and Cord Blood Bank register more than 120 thousand individuals. It is one of the biggest registries in Europe.

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	24:XX	35:XX	04:XX	11:XX	03:XX	0,031285
2	32:XX	35:XX	04:XX	11:XX	03:XX	0,017396
3	33:XX	14:XX	08:XX	01:XX	05:XX	0,015306
4	02:XX	35:XX	04:XX	14:XX	05:XX	0,013654
5	24:XX	18:XX	07:XX	11:XX	03:XX	0,012446
6	02:XX	44:XX	02:XX	16:XX	05:XX	0,01128
7	11:XX	35:XX	04:XX	11:XX	03:XX	0,011086
8	02:XX	51:XX	14:XX	04:XX	03:XX	0,010685
9	24:XX	35:XX	04:XX	16:XX	05:XX	0,010259
10	02:XX	35:XX	04:XX	11:XX	03:XX	0,009933

Table 23: Most frequent ABCDRDQ low resolution haplotype frequencies of the Greek Cypriot adult population (October 2012).

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	33:XX	14:XX	08:XX	01:XX	05:XX	0,02768
2	24:XX	35:XX	04:XX	11:XX	03:XX	0,024564
3	32:XX	35:XX	04:XX	11:XX	03:XX	0,015889
4	03:XX	35:XX	04:XX	11:XX	03:XX	0,010468
5	11:XX	35:XX	04:XX	11:XX	03:XX	0,010428
6	24:XX	18:XX	07:XX	11:XX	03:XX	0,010153
7	24:XX	35:XX	04:XX	16:XX	05:XX	0,010093
8	32:XX	40:XX	02:XX	16:XX	05:XX	0,009607
9	01:XX	08:XX	07:XX	03:XX	02:XX	0,009271
10	02:XX	39:XX	12:XX	16:XX	05:XX	0,009111

Table 24: Most frequent ABCDRDQ low resolution haplotype frequencies of the Greek Cypriot young population (Cord Blood Bank, October 2012).

These results have been used by the Cyprus Bone Marrow Donor Registry to study genetic changes of the Greek Cypriot population. The study has shown lower homogeneity of the young Cyprus populations thanks to mixture with other nations (immigrants, mixed couples).

7.7 South Africa

The South African Bone Marrow Donor Registry (SABMR) has more than 64 thousand donors. It is the biggest registry in Africa. In fact, there are only two registries in Africa, so the SABMR is very unique for the different ethnic groups in the register. We have been asked by medical director of the SABMR to focus on the black population.

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	02:XX	58:XX	06:XX	11:XX	03:XX	0,013385
2	02:XX	58:XX	06:XX	11:XX	06:XX	0,013385
3	29:XX	44:XX	07:XX	11:XX	06:XX	0,011765
4	02:XX	58:XX	03:XX	13:XX	06:XX	0,010348
5	30:XX	08:XX	07:XX	03:XX	04:XX	0,009766
6	02:XX	58:XX	07:XX	07:XX	02:XX	0,009374
7	68:XX	15:XX	03:XX	11:XX	06:XX	0,008583
8	30:XX	18:XX	07:XX	11:XX	06:XX	0,00853
9	02:XX	44:XX	16:XX	13:XX	06:XX	0,008019
10	02:XX	08:XX	07:XX	03:XX	02:XX	0,007241

Table 25: Most frequent ABCDRDQ low resolution haplotype frequencies of the Black population in South Africa, based on 582 individuals (SABMR, October 2012).

Rank	A*	B*	C*	DRB1*	DQB1*	Frequency
1	33:XX	07:XX		03:XX		0,042531
2	33:XX	58:XX		13:XX		0,019308
3	66:XX	58:XX		13:XX		0,018958
4	02:XX	44:XX		13:XX		0,018881
5	33:XX	44:XX		11:XX		0,01813
6	02:XX	58:XX		11:XX		0,017367
7	24:XX	07:XX		15:XX		0,015563
8	02:XX	15:XX		03:XX		0,014922
9	33:XX	15:XX		11:XX		0,0142
10	02:XX	58:XX		07:XX		0,013871

Table 26: Most frequent ABDR low resolution haplotype frequencies of the Black population in South Africa, based on 2592 individuals (SABMR, October 2012).

7.8 Nigeria

This work [67] proves the need of setting up the new registry in Nigeria, by comparing Nigerian HLA haplotype frequencies with Afro-American frequencies.

HLA Study in Nigeria

David Steiner [1], Terry Schlaphoff [2], Veronica Borrill [2], Professor Ernette du Toit [2], Colette Raffoux [3]

[1] Czech Technical University, Prague, Czech Republic
[2] South African Bone Marrow Registry, Cape Town, South Africa
[3] IRGHET International Research Group on Hematopoietic stem cells Transplantation, Paris, France

INTRODUCTION

The South African Bone Marrow Registry (SABMR) has number of donors from other African countries on its database as patients come from all parts of Africa to seek medical treatment in South Africa.
In December 2009, we were approached by DKMS Americas regarding a patient in the USA, originally from Nigeria, who was not able to finding a matching unrelated donor anywhere in the worlds. DKMS Americas agreed to assist the patient by holding a donor drive in Nigeria. However, they needed to find a host registry for these persons & after discussion with our Medical Director, it was agreed that we would help. The reason for approaching the SABMR is that we are the only functional registry in Africa. The donors completed SABMR application forms.

MATERIALS AND METHODS

The Nigerian file contains 274 healthy individuals that were typed by SBT on A*, B*, C* and DRB1* in an EFI accredited laboratory. Resolution of typing was at least intermediate (multiple-allele-codes or high resolution codes). Even if the file might not fully represent Nigerian population, it is the biggest and best typed sample of Nigerian HLA data.
Haplotypes were estimated by maximum likelihood approach and Expectation-Maximalization (EM) algorithm. Multiple-allele-codes were expanded to all possible alleles according to current HLA nomenclature. Results were compared to Afro-American (AFA) population (Maiers M, Gragert L, Klitz W.: High-resolution HLA alleles and haplotypes in the United States population, Hum Immunol. 2007 Sep;68(9):779-88. Epub 2007 May 24.).

Nigeria C-B haplotypes					NMDP	
C*	B*		NIG_rank	freq	AFA_rank	freq
04:10	53:01		1	10,40%	NA	
04:06	53:01		2	8,23%	NA	
06:02	58:02		3	4,93%	6	3,96%
03:04	15:10		4	4,85%	11	2,37%
02:10	15:28		5	4,53%	NA	
04:06	44:03		6	2,41%	NA	
17:03	42:01		7	2,38%	4	5,32%
16:01	52:01		8	2,01%	17	1,23%
04:04	53:01		9	1,91%	NA	
07:21	57:03		10	1,88%	NA	

Table 1: Top Nigerian C*-B* haplotypes

Nigeria A-B haplotypes					NMDP		
A*	B*		NIG_rank	freq	AFA_rank	freq	detail
36:01	53:01		1	7,70%	5	0,01535	
02:02	53:01		2	3,57%	16	0,01042	
68:02	15:10		3	2,06%	12	0,01202	
33:03	15:16		4	1,83%	32	0,00549	
30:01	42:01		5	1,82%	1	0,02963	30:01/30:24
30:24	42:01		6	1,82%	1	0,02963	30:01/30:24
68:02	53:01		7	1,67%	8	0,01351	
34:02	44:03		8	1,52%	22	0,00782	
23:19Q	53:01		9	1,28%	NA		
30:02	15:10		10	1,26%	87	0,00290	

Table 2: Top Nigerian A*-B* haplotypes

Nigerian A-B-C-DRB1 low resolution haplotypes						
A*	B*	C*	DRB1*	NIG_rank	NIG_freq	AFA_freq
36:XX	53:XX	04:XX	11:XX	1	3,32%	0,78%
02:XX	53:XX	04:XX	13:XX	2	2,24%	0,35%
30:XX	42:XX	17:XX	03:XX	3	2,03%	1,89%
36:XX	53:XX	04:XX	15:XX	4	1,75%	0,02%
02:XX	35:XX	16:XX	13:XX	5	1,55%	0,06%
34:XX	44:XX	04:XX	15:XX	6	1,44%	0,56%
36:XX	53:XX	04:XX	03:XX	7	1,39%	0,16%
74:XX	15:XX	02:XX	15:XX	8	1,28%	0,38%
68:XX	15:XX	03:XX	03:XX	9	1,25%	0,86%
02:XX	53:XX	04:XX	03:XX	10	1,24%	0,27%

Table 3: Top Nigerian A*-B*-DRB1* haplotypes

RESULTS

Differences were observed mainly in B-C high resolution haplotypes. 6 of top 10 most common B-C haplotypes are not present in AFA population (1. B*53:01-C*04:10, 2. B*53:01-C*04:06, 5. B*15:28-C*02:10, 6. B*44:03-C*04:06, 9. B*53:01-C*04:04 and 10. B*57:03-C*07:21). On the other hand 3 of top 3 most common AFA B-C haplotypes were not observed in the file (1. B*53:01-C*04:01g, 2. B*15:03g-C*02:02 and 3. B*07:02g-C*07:02).

All most common A-B haplotypes are known also in AFA, but with different frequencies and ranks: 1. A*36:01-B*53:01 (rank in AFA is 5), 2. A*02:01-B*53:01 (AFA: 16), 3. A*68:02-B*15:10 (AFA: 12). 3 of top 3 most common AFA A-B haplotypes were not observed in the file (2. A*03:01g-B*07:02g and 3. A*02:01g-45:01g).

Most frequent A-B-C-DRB1 haplotypes in the Nigerian file are: 1. A*36-B*53-C*04-DRB1*11 (NIG: 3,32%, AFA: 0,78%), 2. A*02-B*53-C*04-DRB1*13 (NIG: 2,24%, AFA: 0,35%), 3. A*30-B*42-C*17-DRB1*03 (NIG: 2,03%, AFA: 1,89%), 4. A*36-B*53-C*04-DRB1*15 (NIG: 1,75%, AFA: 0,02%) and 5. A*02-B*35-C*16-DRB1*13 (NIG: 1,55%, AFA: 0,06%).

CONCLUSION

These results show the Nigerian population is different than Afro-American population in United States and the need of further development of unrelated stem cell registries in Africa.

CONTACT

http://www.sabmr.co.za

8. Usage of haplotype frequency estimations

This chapter presents some applications of HFE.

8.1 Examples of applications

HLA haplotype frequency estimates could be used to:

1. To plan development of a stem cell donor registry, especially its size and effectively in finding an unrelated stem cell donor for a new random patient [51] [68] [69] [70] [71].
2. To select donors that are HLA-A and HLA-B typed only for prospective HLA-DRB1 typing by their HLA-AB-phenotype, so that after a defined number of typings performed the expected "population coverage" of the registry is maximized [60].
3. Selective recruitment of stem cell donors [72].
4. To analyze and compare HLA genetic relations and properties of different populations [73] [54] [74].
5. To calculate the probability of HLA high resolution match between a particular donor and patient. Based on this, we can construct new generation of the search algorithm that ranks donors according to their probability of HLA high resolution match with the patient. Such state-of-the-art approach is used in Germany (Optimatch®) and in the United States (HapLogic℠).
6. To interfere HLA haplotype information for a specific donor, for who we cannot perform family study [75].
7. To calculate the probability of finding a suitable related or unrelated stem cell donor [76] [36].

This work focuses on the point 5 (and partly also 6 and 7) that is further elaborated in the following chapters 9, 0 and 11. However, in the next paragraphs, we will mention some of our results related to previous points.

8.2 Phylogenetic threes and population maps

We have cooperated with students of the Czech Technical University on their bachelor and diploma works. They have used our data as input of their applications. J. Těhník has implemented program that can analyse database of a registry and projects trends [77], see Figure 33.

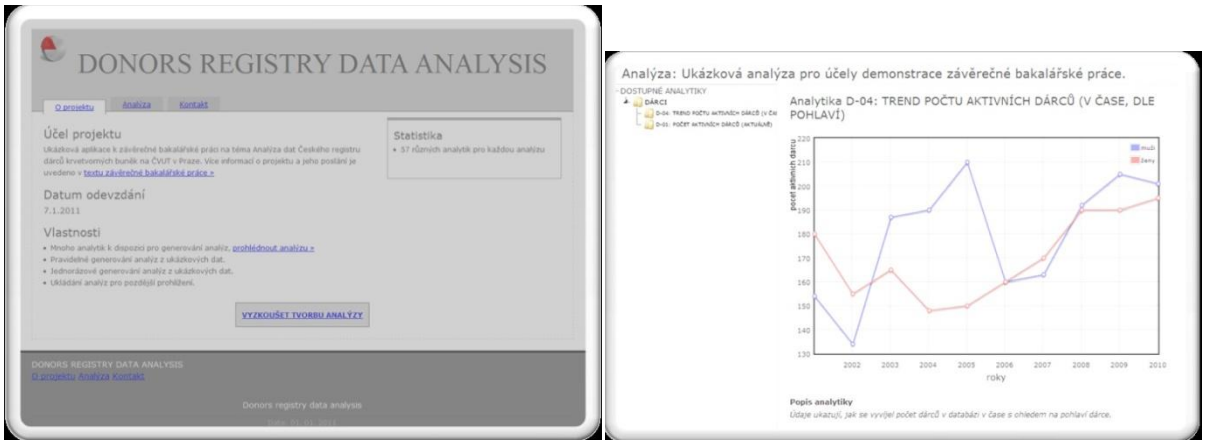


Figure 33: Bachelor work [77] – analysis of database of a stem cell donor registry.

L. Kábrt has developed a web application that visualizes HLA data and their location on Google maps [73]. For location of the donor, postal codes have been used. For example, the map of Finland shows different frequencies of HLA allele groups in regions with Swedish speaking population (see Figure 34). In case of the Czech republic, we did not find significant regional differences. Similar study has been done in the UK and Germany [70].

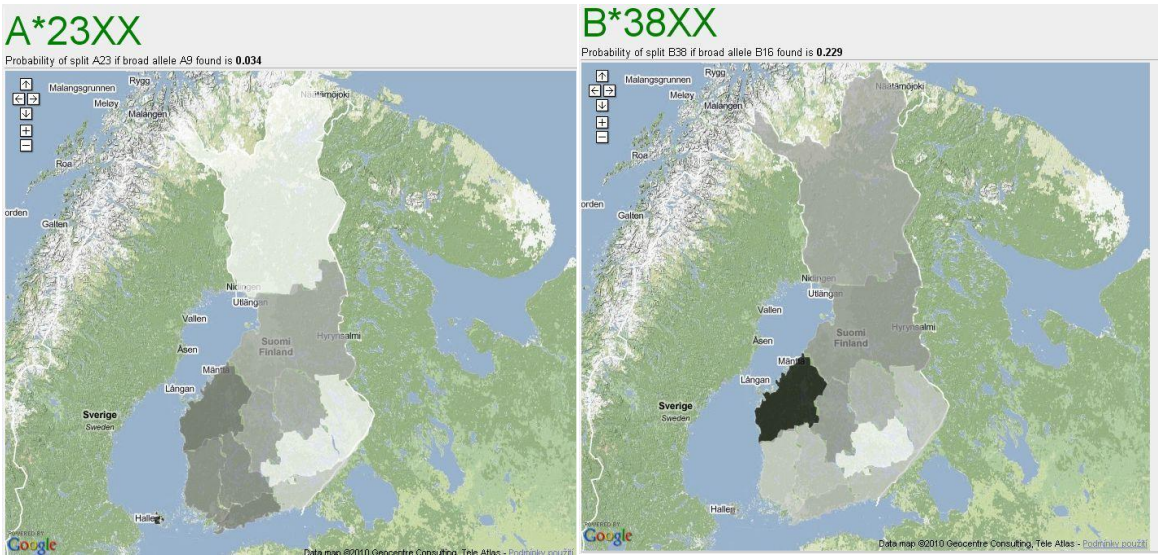


Figure 34: Diploma work [73] – analysis of database of a stem cell donor registry.

8.3 HLA Explorer

We have developed an internet application HLA Explorer (www.hlaexplorer.net) [78] that implements user-friendly interface for browsing HLA haplotype frequencies estimations. Goal of the project was to develop system that helps physicians (transplant centers) and coordinators (stem cells donor registries) to examine Linkage Disequilibrium of HLA system in order to assist to find suitable unrelated stem cells donor.

The application has more than 200 registered users worldwide.

WWW.HLAEXPLORER.NET — FREE INTERNET APPLICATION FOR EXPLORING DEPENDENCIES IN HLA SYSTEM

David Steiner[1], Jakub Brož[1], Libuše Kupková[2], Agathe Rosenmayr[3], Derek Middleton[4], Lenka Lhotská[1]

[1] Czech Technical University in Prague/Gerstner Laboratory, Technická 2, Prague, Czech Republic

[2] Czech Stem Cells Registry, Prague, Czech Republic

[3] Austrian Bone Marrow Donor Registry, Wien, Austria

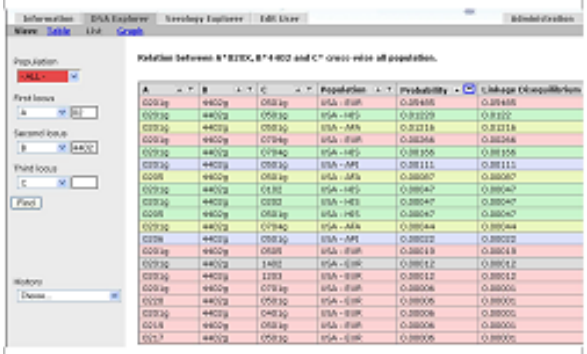
[4] Northern Ireland Histocompatibility & Immunogenetics Laboratory, Belfast, UK

Introduction

The goal of the project HLA Explorer (www.hlaexplorer.net) is to develop a new internet based information system that will help coordinators (stem cells donor registries) and physicians (transplant centers) to examine Linkage Disequilibrium (LD) of HLA system in order to assist to find suitable unrelated stem cells donor.

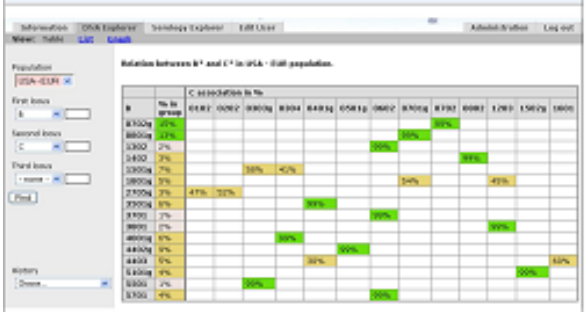
DNA Explorer - List View

- DNA high resolution relations of HLA alleles in the list that you can sort according to any column (selected loci, population, probability and Linkage Disequilibrium).



- Example shows comparison of high resolution A*-B*-C* frequencies between different American populations. Filter (on the left) is set to A*02-B*4402-C* and no specific population is selected ("ALL").

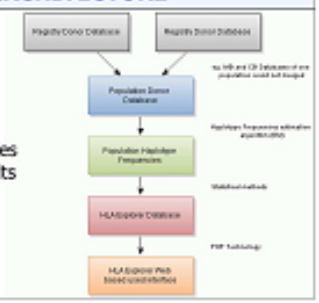
DNA Explorer - Table View



- DNA high resolution relations of HLA alleles in the table.
- Example shows 2 locus relation B*-C* in USA-EUR population. Filter is set to "USA-EUR" and B*-C*. Yellow color stresses high LD and green color very high LD.

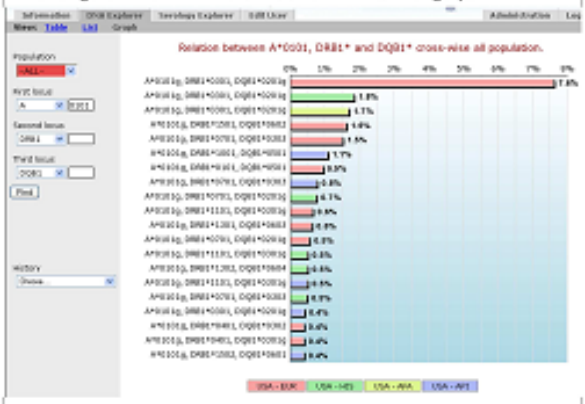
System Architecture

- Data from Registry Donor Databases are merged into Population Donor Database. (e.g. CS, CSCB and CS2 to "Czech population")
- We count haplotype frequencies by EM algorithm and store results in special database.
- Web based application shows results in user-friendly way.



DNA Explorer - Graph View

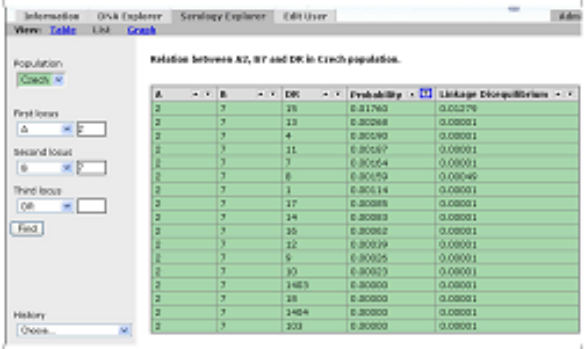
- DNA high resolution relations of HLA alleles in the graph



- Example shows comparison of 3 locus high resolution haplotype frequencies between different American populations. Filter is set to A*0101-DRB1*-DQB1* and no specific population is selected.

Serology Explorer - List View

- Low resolution relations of HLA alleles in the list



- Example shows haplotype frequencies in the Czech population. Filter is set to "Czech" and A2-B7-DR.

Used Data

- **NMDP**: Maiers, M., Gragert, L., Klitz, W. High resolution HLA alleles and haplotypes in the US population. Human Immunology (2007) 68, 779-788.
- **www.allelefrequencies.net** - Derek Middleton
- **Donor Registries**: CS, CSCB, CS2, SK, A (BMDW registry codes)
- *We would like to cooperate with other registries and with BMDW.*

Conclusions

The chance of any patient to obtain fully matched unrelated stem cells donors is dependent on possibility to match with more than 12 million donors in BMDW. But, up to now many HLA typing results of unrelated donors are incomplete. In case two or more donors - who are potentially HLA identical - are available, coordinators and physicians can use "HLA Explorer" in order to choose donors for complementary genomic typing. The project "HLA Explorer" provides extensive information for haplotype frequencies.

Using of modern technologies such as HLA Explorer increase the chance to find a suitable donor in the shortest term possible.

8.4 Phenotype analysis

Another interesting usage of haplotype frequencies are applications that analyze given phenotype and resolve them into possible genotypes. This can be done for multiple populations and we get multiple results. If the ethnical or family background of the individual is unknown, the comparison of results may help to associate the patient with an ethnic group and or focus attention on rare combinations of patient's alleles. Such information help to refine the donor search strategy before starting the search process.

The publically available tool with such analysis has been developed by NMDP and is available at www.haplostats.org. Another example is the French EasyMatch [30].

The phenotype analysis tool is an internal component of the predictive matching (see next chapter).

9. Prediction of HLA Match

In this chapter we will design new computational method for matching predictions. Top-down design of the algorithm is described. We will also compare our approach with other implementations in the world (ZKRD, NMDP).

9.1 Criteria for the new matching prediction algorithm

A. **Usability**: We need to compare predictions for all donors. The method must be able to handle all cases, patient-donor pairs.
B. **Time**: It would be desirable if the method is used in the interactive user interface. We need to use the method for at least hundreds of patient-donor pairs. Therefore the method must be able to give result quickly, in fraction of one second.
C. **Correctness**: If the method gives results, it must be reliable. We need to understand how reliable is the method. Therefore the method must be validated.

9.2 Definitions

For the purpose of this chapter, we will mathematically define terms haplotype and phenotype that have been already defined in chapter 2.1.

Haplotype h_i is a set of pairs. Each pair is composed of DNA/serology locus designation l_j and allele/antigen code a_j according to the HLA nomenclature (see chapter 2.2.2). Empty value is considered as special valid code. Let *s* be number of loci in the haplotype. We can explicitly write number of loci as upper index of the haplotype.

(26) $$h_i = h_i^s = \{(l_1, a_1), (l_2, a_2), ..., (l_s, a_s)\}$$

l_j must be different from each other.

Let's call the corresponding set of loci as **haplotype type**.

(27) $$T(h_i) = T(h_i^s) = \{l_1, l_2, ...l_s\}$$

So the condition on distinct l_j can be written as

$$\left|T(h_i^s)\right| = s$$

Where $|X|$ is the **size** of the set *X* (number of elements).

Phenotype *Phen* is a set of pairs. Each pair is composed of DNA/serology locus designation l_j and a set of two allele/multiple-allele/antigen codes $a_{j,1}$ and $a_{j,2}$

$$Phen = \{(l_1, \{a_{1,1}, a_{1,2}\}), (l_1, \{a_{2,1}, a_{2,2}\}), ..., (l_s, \{a_{s,1}, a_{s,2}\})\}$$

We say two phenotypes $Phen_X$ and $Phen_Y$ are the **same**, if $Phen_X = Phen_Y$. Otherwise they are **different**.

We say phenotype *Phen* **matches** the haplotype h_i (and vice versa) if all elements of the haplotype h_i "match" with corresponding elements (the same locus) in the phenotype *Phen*. It means it "matches" at least one of two alleles/antigens present at the same locus on the phenotype *Phen*. Our "matching" is described by [5] [7] and basically it means no **mismatch** is observed. If the locus is not present in the phenotype *Phen*, it is considered as "match", i.e. no mismatch is observed.

This predicate (*Phen* matches h_i) is written as

$$M(Phen, h_i)$$

Example 8

$$h_1 = h_1^5 = \{(A, "26:01"), (B, "38:01"), (C, "12:03"), (DRB1, "04:02"), (DQB1, "03:02")\}$$

$$h_2 = h_2^5 = \{(A, "01:01"), (B, "57:01"), (C, "06:02"), (DRB1, "15:01"), (DQB1, "06:02")\}$$

$$h_3 = h_3^5 = \{(A, "01:01"), (B, "38:01"), (C, "06:02"), (DRB1, "04:02"), (DQB1, "03:02")\}$$

$$\left|T(h_1^5)\right| = |T(\{A, B, C, DRB1, DQB1\})| = 5.$$

The individual A (chapter 2.2.4) could be represented as

$$Phen_A = \{(A, \{"01:XX", "26:XX"\}), (B, \{"38", "57"\}), (C, \{"06:02", "12:03"\}),$$
$$(DRB1, \{"04:02", "15:01"\}), (DQB1, \{"03:02", "06:02"\})\}$$

Locus HLA-A has been low resolution typed, locus HLA-B has been typed by serology technique and other loci are high resolution typed.

Then predicates

$M(Phen_A, h_1)$, $M(Phen_A, h_2)$ and $M(Phen_A, h_3)$

are all True.

□

9.3 Matching prediction method

Given an **original phenotype** *Phen* of an individual and *s*-loci haplotype frequencies h_i of its population, the algorithm selects all *s*-loci (high-resolution) haplotypes that match with the phenotype *Phen*.

Then the algorithm loops all matching haplotypes and tries to combine them together into pairs forming **predicted diplotype** (h_i h_j), still matching the original phenotype. h_i and h_j must be complementary, i.e. predicted diplotype form the **predicted phenotype** $Phen_{A,k}$ that also matches the phenotype *Phen*.

Example 9

Following the Example 8: Let's consider these three haplotypes as the only matching haplotypes with the phenotype $Phen_A$. They can from three different diplotypes (h_1 h_2), (h_1 h_3) and (h_2 h_3), but only $Phen_{A,1} = (h_1\ h_2)$ matches the phenotype $Phen_A$.

$$Phen_{A,1} = (h_1\ h_2) = \{(A,\{"01:01","26:01"\}),(B,\{"38:01","57:01"\}),(C,\{"06:02","12:03"\}),(DRB1,\{"04:02","15:01"\}),(DQB1,\{"03:02","06:02"\})\}$$

□

Let's focus on probabilities P_j as defined by equations **(2)** and (3) on page 37.

Let's say we have got *m* possible predicted phenotypes. We normalize their probabilities P_j by

$$P_j^* = \frac{P_j}{\sum_{i=1}^{m} P_i} \quad \textbf{(28)}$$

Therefore

$$\sum_{j=1}^{m} P_j^* = 1$$

holds.

P_j^* are conditional probabilities expressing the event the given phenotype is in fact predicted phenotype *j*.

We use this algorithm twice, to analyze both patient and donor phenotype. Let P_j^D are normalized probabilities of predicted phenotypes $Phen_j^D$ of the donor, m^D is their count, P_j^P are normalized probabilities of predicted phenotypes $Phen_j^P$ of the patient and m^P is their count.

Finally, we find all predicted phenotypes that are common for the donor and the patient, and multiply their conditional probabilities. We get equation for the matching prediction

$$mp = \sum_{j=1}^{m^D} \sum_{k=1}^{m^P} P_j^D P_k^P \delta_{jk} \quad \textbf{(29)}$$

Where $\delta_{jk} = 0$, if phenotypes *j* and *k* are different and $\delta_{jk} = 1$, if phenotypes *j* and *k* are the same, i.e. no mismatch (see chapter 9.2). Similarly, the following equation calculates the probability of one mismatch

$$mp^{MM1} = \sum_{j=1}^{m^D} \sum_{k=1}^{m^P} P_j^D P_k^P \delta_{jk}^{MM1} \quad \textbf{(30)}$$

Where $\delta_{jk}^{MM1} = 1$, if phenotypes *j* and *k* have exactly one mismatch (see also [7]) and $\delta_{jk}^{MM1} = 0$ otherwise.

The probability of a match at specific locus is estimated by

$$mp^L = \sum_{j=1}^{m^D} \sum_{k=1}^{m^P} P_j^D P_k^P \delta_{jk}^L \quad \textbf{(31)}$$

Where *L* is the locus designation and $\delta_{jk}^L = 1$, if phenotypes *j* and *k* are the same at locus *L* and $\delta_{jk}^L = 0$ otherwise.

9.4 Phenotypes cannot be explained

Previously described method does not meet criterion A (see chapter 9.1), because it can fail if the patient or donor phenotype cannot be "**explained**". For patient, it means $m^P = 0$ and the patient's set of predicted phenotypes $\{Phen_j^P\}$ is empty. This can happen if:

- There are no matching *s*-loci (high-resolution) haplotypes (let's call them **full haplotypes**).
- There are such full haplotypes, but they cannot form matching predicted diplotypes.

In such case, our method tries to find **matching partial haplotypes**, i.e. haplotypes with less than *s* loci that match the original phenotype.

Partial haplotype h_i^r is non-empty subset of any original (high-resolution) haplotype h_i^s.

$$h_i^r \subset h_i^s, \; h_i^r \neq h_i^s, \; h_i^r \neq \{\}$$

$L(h_i^r)$ is **partial haplotype type**.

$L(h_i^r) \subset L(h_i^s)$, $L(h_i^r) \neq L(h_i^s)$, $L(h_i^r) \neq \{\}$

Matching partial haplotype is the partial haplotype matching the phenotype *Phen*. Obviously, if the (high-resolution) haplotype matches the phenotype *Phen*, then all derived partial haplotypes are matching the phenotype *Phen*. But it is not always true vice versa, i.e. mismatched (high-resolution) haplotype may include matching partial haplotypes.

Let PH^r be the set of partial haplotypes with *r* loci (0<*r*<*s*)

$$PH^r = \bigcup_i \left\{ h_i^t \middle| h_i^t \subset h_i^s \wedge \left|h_i^t\right| = r \right\} = \bigcup_i \left\{ h_i^t \middle| h_i^t \subset h_i^s \wedge \left|L(h_i^t)\right| = r \right\}$$

Conditions $h_i^r \neq h_i^s$ and $h_i^r \neq \{\}$ are forced by the condition 0<*r*<*s*.

Matching partial haplotypes *MPH* are

$$MPH = \bigcup_{r=1}^{s-1} MPH^r = \bigcup_{r=1}^{s-1} \left\{ h_i^r \middle| h_i^r \in PH^r \wedge M(h_i^r, Phen) \right\}$$

The method combines these partial haplotypes of different types together, forming **artificial haplotypes** that cover all *s* loci. In case of two partial haplotypes we get

$$H_2^s = \left\{ h_i^t \cup h_j^v \middle| 0 < t < s \wedge 0 < v < s \wedge \left|L(h_i^t) \cup L(h_j^v)\right| = s \wedge \left|L(h_i^t) \cap L(h_j^v)\right| = \left|h_i^t \cap h_j^v\right| \right\}$$

The condition $\left|L(h_i^t) \cup L(h_j^v)\right| = s$ could be also written as $\left|L(h_i^t \cup h_j^v)\right| = s$, which means the artificial haplotype covers all *s* loci.

The condition $\left|L(h_i^t) \cap L(h_j^v)\right| = \left|h_i^t \cap h_j^v\right|$ means that for all loci that appear in both h_i^t and h_j^v, also corresponding alleles/antigens must be the same at all sharing loci of both haplotypes.

But in general, even more than two partial haplotypes can form one artificial haplotype. For triplets we get

$$H_3^s = \left\{ \begin{array}{l} h_i^t \cup h_j^v \cup h_k^w \middle| 0 < t < s \wedge 0 < v < s \wedge 0 < w < s \wedge \left|L(h_i^t) \cup L(h_j^v) \cup L(h_k^w)\right| = s \wedge \\ \left|L(h_i^t) \cap L(h_j^v)\right| = \left|h_i^t \cap h_j^v\right| \wedge \left|L(h_i^t) \cap L(h_k^w)\right| = \left|h_i^t \cap h_k^w\right| \wedge \left|L(h_j^v) \cap L(h_k^w)\right| = \left|h_j^v \cap h_k^w\right| \end{array} \right\}$$

… and so on.

All artificial haplotypes are

$$H^s = \bigcup_{i=2}^{s} H_i^s$$

Full haplotypes could be perceived as H_1^s

Haplotype frequency of the partial haplotype is calculated as the sum of all full haplotypes that are supersets of the partial haplotype.

(32) $p_i^t = \sum_{\{j;h_j \supset h_i\}} p_j^s$

Allele frequencies are special case of partial haplotype frequencies, where $t = 1$.

Example 10

Following the Example 8: If the full haplotype type is {A, B, C, DRB1, DQB1}, $s = 5$, then an example of **partial haplotype type** is {A, B, C, DRB1}, which covers 4 loci and the locus DQB1 is omitted. An example of partial haplotype is

$$h_1^4 = \{(A,"26:01"),(B,"38:01"),(C,"12:03"),(DRB1,"04:02")\}$$

□

How to find these partial haplotypes? Number of partial haplotype types corresponds to all subsets of the full haplotype type, except empty set, which grows by exponential function $2^s - 1$. In case of five loci, we get $2^s - 1 = 2^5 - 1 = 31$, so there are 31 partial haplotype lists. For each partial haplotype type, we need to search for matching partial haplotypes. Then these lists are combined together. There are up to $(2^s - 1)^2$ possible pairs of partial haplotype types we need to check. Each check combines two lists, so its complexity is $O(n^2)$. There are $(2^s - 1)^3$ triplets, etc., so the total complexity of the calculation is extreme

$$\sum_{i=2}^{s} (2^s - 1)^i O(n^i)$$

These are maximum numbers, not all of them make sense to combine, for example {A*} is already included in {A*, B*} and the combination does not make sense. I.e. combining such two haplotype types we cannot create artificial haplotypes that cover all 5 loci. But still the number of combinations is too high for efficient computing.

Therefore we check only selected partial haplotype types. We have also implemented heuristics that first checks bigger partial haplotype types and then, if not successful, others. The algorithms uses this order of partial haplotype types:

- **{A, B, C, DRB1, DQB1}** … full haplotypes
- {A, B, DRB1, DQB1} … locus C excluded
- {A, B, C, DRB1} … locus DQB1 excluded
- **{A, B, DRB1}** … typical 3 loci matching (loci C and DQB1 excluded). First versions of HapLogic™ and OptiMatch® used these three loci for predictive matching (see chapter 11.3). BMDW [6] also uses these three loci for basic matching.
- {A, B, C} … first class loci
- {B, DRB1, DQB1} … second class loci and the closest first class locus, see Figure 2.
- {A} … individual locus

- {B} … individual locus
- {C} … individual locus
- {DRB1} … individual locus
- {DQB1} … individual locus

If the partial haplotype type can be used for explanation of patient (partial) phenotype, we select it. We continue to the next partial haplotype type, until we cover all loci by the loci in all selected partial haplotype types. Since single locus haplotype types are in the end of the list ({A}, {B}, {C}, {DRB1} and {DQB1}), we will always find solution that cover all loci. This means, in the worst case, allele frequencies will be used.

Example 11

Partial haplotype types {A, B, DRB1}, {B, DRB1, DQB1} and {C} together cover all five loci, i.e. haplotype type {A, B, C, DRB1, DQB1}.

□

Note: Theoretically, it could happen even the (phenotype) typing result at a locus cannot be explained by allele frequencies. This means we are trying to estimate probabilities of alleles that have never been observed in the underlying population, so allele frequencies are zero or almost zero. This can happen if the individual does not belong to the model population.

After this procedure, we get list of partial haplotype types and corresponding lists of matching partial haplotypes. Now, we need to form artificial haplotypes and estimate their frequencies.

Artificial haplotypes are formed by combination of partial haplotypes from all lists of matching partial haplotypes. We take only those combinations that match, i.e. if there is non-empty intersection of two partial haplotype types and corresponding partial haplotypes must share the same alleles at all loci in the intersection. Haplotype frequency of the artificial haplotype is estimated as haplotype frequency of the first partial haplotype (forming the artificial haplotype) multiplied by normalized multiplication of all other partial haplotypes forming the artificial haplotype.

In case of two partial haplotypes h_i^t and h_j^v forming an artificial haplotype $h_a^s = h_i^t \cup h_j^v$ we define artificial haplotype frequency as

$$p_a^s = p_{i,j}^s = p_i^t \frac{p_j^v}{\sum p_x^v} \quad \textbf{(33)}$$

where:

- p_x^v are frequencies of h_x^v (one of them is also h_j^v)
- h_x^v have the same partial haplotype type as h_j^v
- $M\left(Phen, h_x^v\right)$

This means h_x^v are all possible extensions of h_i^t, that belong to single partial haplotype type and still form matching artificial haplotype with h_i^t.

This definition of partial haplotype frequency is consistent with the property $p_i^t = \sum_j p_{i,j}^s$ (see equation (32)).

If these partial haplotypes do not share any locus (intersection is empty set), then $\sum p_x^v = 1$

and these are two independent **fragments** (without common loci) also form new haplotype. Haplotype frequency is calculated as multiplication of frequencies of forming partial haplotypes (fragments).

In case three partial haplotypes h_i^t, h_j^v and h_k^w forme an artificial haplotype $h_a^s = h_i^t \cup h_j^v \cup h_k^w$ we define artificial haplotype frequency as

$$p_a^s = p_{i,j,k}^s = p_i^t \frac{p_j^v}{\sum p_x^v} \frac{p_k^w}{\sum p_y^w}$$

where:

- p_x^v are frequencies of h_x^v (one of them is also h_j^v)
- p_y^w are frequencies of h_y^w (one of them is also h_k^w)
- h_x^v have the same partial haplotype type as h_j^v
- h_y^w have the same partial haplotype type as h_y^w
- $M\left(Phen, h_x^v\right)$
- $M\left(Phen, h_y^w\right)$

Similarly for four and more partial haplotypes.

In extreme case, only allele frequencies are used (partial haplotype types {A}, {B}, {C}, {DRB1} and {DQB1}) and the haplotype frequency is calculated as multiplication of allele

Example 12

Let haplotypes h_1, h_2, h_3 and h_4 are the only haplotypes in our haplotype list.

$$h_1 = h_1^5 = \{(A,"26:01"),(B,"38:01"),(C,"12:03"),(DRB1,"04:02"),(DQB1,"03:02")\}$$

$$h_2 = h_2^5 = \{(A,"01:01"),(B,"57:01"),(C,"06:02"),(DRB1,"15:01"),(DQB1,"06:02")\}$$

$$h_3 = h_3^5 = \{(A,"26:01"),(B,"38:01"),(C,"06:02"),(DRB1,"04:02"),(DQB1,"03:02")\}$$

$$h_4 = h_4^5 = \{(A,"26:01"),(B,"38:01"),(C,"06:02"),(DRB1,"15:01"),(DQB1,"03:02")\}$$

Let their frequencies be $p_1 = 0.1$, $p_2 = 0.2$, $p_3 = 0.3$ and $p_4 = 0.4$.

The frequency of partial haplotype

$$h_{11}^3 = \{(B,"38:01"),(DRB1,"04:02"),(DQB1,"03:02")\}$$

is $p_{11} = p_1 + p_3 = 0.4$

The second partial haplotype of the same type {B, DRB1, DQB1} is

$$h_{12}^3 = \{(B,"57:01"),(DRB1,"15:01"),(DQB1,"06:02")\}$$

with frequency $p_{12} = p_2 = 0.2$

And the third one is

$$h_{13}^3 = \{(B,"38:01"),(DRB1,"15:01"),(DQB1,"03:02")\}$$

with frequency $p_{13} = p_4 = 0.4$. There is no other partial haplotype type and therefore $p_{11} + p_{12} + p_{13} = 1$

Similarly, partial haplotype

$h_{21}^3 = \{(A,"26:01"),(B,"38:01"),(C,"12:03")\}$ has frequency $p_{21} = p_1 = 0.1$

This partial haplotype h_{21}^3 can be extended by h_{11}^3 or h_{13}^3 to form the full haplotype.

h_{21}^3 and h_{13}^3 form new artificial haplotype

$$h_{a,1}^5 = \{(A,"26:01"),(B,"38:01"),(C,"12:03"),(DRB1,"15:01"),(DQB1,"03:02")\}$$

with frequency $p_{a,1} = p_{21}\frac{p_{13}}{p_{11}+p_{13}} = 0.1\frac{0.4}{0.4+0.4} = 0.05$

The new artificial haplotype $p_{a,1}$ may help to explain the input haplotype.

h_{21}^3 and h_{11}^3 form again h_1 , but with different frequency

$$p_{a,2} = p_{21}\frac{p_{11}}{p_{11}+p_{13}} = 0.1\frac{0.4}{0.4+0.4} = 0.05$$

Summary of the example: All full haplotypes starting with partial haplotype h_{21}^3 (only h_1 in our example) were replaced by all possible extensions of h_{21}^3 (two options). This has added new artificial haplotype(s). Frequencies of newly formed haplotypes were reshuffled, but total frequency of all of them is the same as original haplotypes.

□

9.5 Validation of the concept of artificial haplotypes

We form artificial haplotypes are formed only in case normal haplotypes fail to resolve (explain) the input phenotype. But it might be useful in more difficult cases.

In order to validate the concept of artificial haplotypes, we have run the following simulation:

1. Select all high resolution A*-B*-C*-DRB1*-DQB1* phenotypes from the dataset [BMDW-2011].
2. Try to explain these phenotypes by standard full haplotypes.
3. Select phenotypes that cannot be explained by full haplotypes, but can be explained by artificial haplotypes.
4. Decrease high resolution to low resolution for all five loci. Estimate probability of low resolution phenotype to become the high resolution phenotype, using these three methods:
    - Artificial haplotypes, combined by partial haplotypes that overlap. For example types { A*, B*, C*} and {B*, DRB1*, DQB1*}
    - Artificial haplotypes, combined by partial haplotypes that do not overlap. For example types { A*, B*, C*} and { DRB1*, DQB1*}
    - Artificial haplotypes, combined by single locus partial haplotypes (types { A*}, { B*}, { C*}, { DRB1*} and { DQB1*}).
5. Calculate average *U* value (see (34)) for all these three approaches.

In the database [BMDW-2011], we have found 595 haplotypes that cannot be explained by full haplotypes, but can be explained by artificial haplotypes. We have also used [PROM-CT]. As HFE, we have used [ZKRD-2008] and [HPE-2010]. The more haplotypes we have in the HFE, the lower number of validation cases for this exercise we find.

Table 27 displays results of the simulation. It shows the concept of artificial phenotypes has better results than other two concepts.

Dataset	Haplotype frequencies	Number of validation cases	Artificial haplotypes, combined by partial haplotypes that overlap	Artificial haplotypes, combined by partial haplotypes that do not overlap	Artificial haplotypes, combined by single locus partial haplotypes
[BMDW-2011]	[ZKRD-2008]	595	3.8707379473	4.7190691871	5.9372123576
[PROM-CT]	[ZKRD-2008]	206	0.1497170941	0.3691394013	0.5350892222
[PROM-CT]	[HPE-2010]	68	0.0499954565	0.1668737778	0.2416876335

Table 27: Validation of the concept of artificial haplotypes, table shows U values

9.6 Situation in the world

9.6.1 OptiMatch®

OptiMatch® matching prediction method is roughly described in [63]. The system calculates the matching prediction in the same way as our method, i.e. our equation (29) and OptiMatch® equation on the Figure 35 are similar.

$$mp = \frac{\sum_{k \in (P^S \cap P^P)} p_k^S \cdot p_k^P}{\sum_{k \in P^S} p_k^S \cdot \sum_{k \in P^P} p_k^P}$$

mit

p_k^S Phänotypfrequenzen der Spenderpopulation
p_k^P Phänotypfrequenzen der Patientenpopulation.
P^P, P^S Mengen der möglichen Phänotypen von Patient und Spender

Figure 35: Matching prediction method equation of OptiMatch® [63]

However, other aspects of OptiMatch® matching prediction methods are not published, e.g. how to handle patients and donors with phenotypes that cannot be explained (see chapter 9.4).

You will find more information about the OptiMatch® system in chapter 11.3.1

9.6.2 HapLogic™

As far as we know, HapLogic™ prediction methods have not been published.

You will find more information about the HapLogic™ system in chapter 11.3.2

9.6.3 Others

The Hap-E system [79] uses probably similar prediction method as OptiMatch®. Mathematical description, internals and handling of problematic cases has not been published.

EasyMatch [30] focuses on a priori analyses of patient's phenotype, rather than patient-donor matching predictions.

10. Validation of Matching Predictions

This chapter describes methods of validation of the HLA matching prediction algorithm, including new simulation framework and provides our results.

10.1 Methods

The quality of prognostic matching algorithm and the population model used (allele and haplotype frequencies) have to be validated as well. This is usually done by retrospective or prospective studies.

Hans-Peter Eberhard has used the Logarithmic Score Function [63].

(34) $$U(x,q) = \frac{1}{l}\sum_{i=1}^{l}\begin{cases} \log(q_i) & for\ x_i = 1\,(Match) \\ \log(1-q_i) & for\ x_i = 0\,(Mismatch) \end{cases}$$

where *l* is the number of matching predictions and q_i are matching predictions. In case of $q_i = 0$, the value $q_i = 10^{-4}$ is taken instead.

More typical option is to use all VTs performed by the registry that meet specific criteria. These criteria are:

- Patient has been typed at high resolution
- Donor was not typed at high resolution before the typing request, but has been high resolution typed at the time of typing request (or later).
- No discrepancy between a priori and final HLA type.

Table 28: Criteria for validation typing request

The review process retrospectively calculates the matching prognosis and compares the predicted and observed percentage of allele matches.

10.2 Validation using verification typings

Validation of matching predictions was carried out similarly to Optimatch/Haplogic. We have taken all **verification typing requests** (VTs, formerly known as confirmatory typing requests, CTs) performed by the registry. This was not easy task, because most of the registries recorded such data only in paper form. In last four years, we have helped to connect at least 10 stem cell donor registries in Europe, Asia and Africa to the EMDIS network (see Appendix C). Thanks to this effort, these registries have started to record all international and national VTs in electronic form. This has been one of the key building blocks of this work. As VTs we have used EMDIS "Sample request" messages (SMP_REQ) [21]. We have collected more than 5000 VTs (Czech Stem Cells Registry, Slovak BMDR, Polish ALF Registry, Swedish Tobias Registry, Finnish BMDR, South African BMR and Ezer Mizion BMDR).

From these VTs, we have selected only those that met these requirements:

- patient has been typed at high resolution level (HLA-A, -B, -C, -DRB1, -DQB1)
- high resolution (HLA-A, -B, -C, -DRB1, -DQB1) data for loci examined as a VT result (or later)
- no discrepancy between a priori and final HLA type

Table 29: Criteria for validation VTs

About one third of VTs satisfy the criteria and that could be used for validation.

We have faced two problems:

- unlike ZKRD and NMDP, other registries do not have enough donors that could be used for estimation of 5 locus high resolution haplotype frequencies. Haplotype frequencies could be calculated, but their confidence is questionable.
- smaller registries also do not have enough VTs that could be used for validation of the prediction algorithm. ZKRD used 9843 CTs in 2008 [8] and 22255 CTs in 2010 [63]. These numbers are not achievable by smaller registries.

In order to overcome these problems, we have approximated the local population to the German (ZKRD) population, i.e. we have used our estimation of German haplotype frequencies [D-1205]. We have also joined VTs from multiple registries using Prometheus software. As result, we have collected 1406 VTs for validation. Unlike ZKRD or NMDP that have enough VTs only for their donors, our VTs represent a mix of Caucasian donors from different countries.

Then we have calculated (retrospectively) the matching prognosis and compared the predicted and observed percentage of 10/10 (resp. 9/10) allele matches at 10% or 20% prediction intervals.

Patient	**Donor typing before VT**	**Probabilities**	**Donor typing after VT**
(German patient)	(Finnish donor)	P(10/10)=**0.943** P(9/10)=0.057	(10/10 allele match)
A*02:01,03:01	A2,3	P(A)=0.999	A*02:01,03:01
B*15:01,44:02	B62,44	P(B)=0.943	B*15:01,44:02
C*03:03,05:01	C*03:03,05:01	P(C)=1.000	C*03:03,05:01
DRB1*07:01,11:01	DRB1*07:01,11:01	P(DR)=1.000	DRB1*07:01,11:01
DQB1*02:02,03:01	DQ2,3	P(DQ)=0.999	DQB1*02:02,03:01
(German patient)	(Finnish donor)	P(10/10)=0.049 P(9/10)=0.017	(7/10 allele match)
A*01:01,24:02	A*01:XX,24:XX;	P(A)=0.998	A*01:01,24:02
B*08:01,15:17	B*08:CCWB,15:XX;	P(B)=**0.0668**	B*08:01,**15:01**
C*07:01,07:01		P(C)=**0.0661**	C*07:01,**04:01**
DRB1*07:01	DRB1*07:XX	P(DR)=0.999	DRB1*07:01
DQB1*02:02,02:02		P(DQ)=**0.342**	DQB1*02:02,**03:03**

Table 30: Examples of the VTs. In the first case, the VT has proven, the donor has the same typing as the patient (prediction for the 10/10 allele match was 94.3%). In the second case, the VT has shown, the donor has multiple mismatches at B*, C* and DQB1* (low predictions at these tree loci).

PROMETHEUS PROBABILITY MATCHING
COMMUNITY TECHNOLOGY PREVIEW

David Steiner[1,2], Matti Korhonen[3], Marie Kuříková[4], Mária Kušíková[5], Monika Sankowska[6], Bert Svensson[7], Emette Du Toit[8], Nira Shriki[9], Karel Peyerl[2], Colette Raffoux[10]

[1] Czech Technical Univeristy in Prague, [2] Steiner, Ltd., [3] Finnish Bone Marrow Donor Registry, [4] Czech Stem Cells Registry, [5] Slovak National Bone Marrow Donor Registry, [6] ALF Marrow Donor Registry, [7] The Tobias Registry, [8] South African Bone Marrow Registry, [9] Ezer Mizion Bone Marrow Donor Registry, [10] IRGHET International Research Group on Hematopoietic stem cells Transplantation, Paris FR

INTRODUCTION

- Prometheus is a software solution for stem cell donor registries. The system is currently in use in 20 countries. The smallest registry has 600, while the largest one has 600 000 donors. Prometheus serves as a national software system for more than 1000 patients a year.
- The system uses donor search algorithm that fully supports the EMDIS matching preferences (allele/antigen mismatch specification, age, gender, CMV, etc.) and also participates in the WMDA Matching Validation project (IT Working Group).
- However, a new generation of matching algorithms has been developed by ZKRD (Optimatch®) [1,2] and NMDP (Haplogic™) [3].
- These algorithms are currently based on 5 locus high resolution haplotype frequencies (Optimatch Version 2 from June 2008 and Haplogic III from December 2011).
- The goal of this project is to independently develop and integrate these technologies into the Prometheus software. Such technology must be validated before it can be reliably used by search coordinators.

METHOD

- Estimation of 5 locus high resolution haplotype frequencies from a donor registry database.
- A matching program calculating, for each donor, the probability of being allele identical to the patient.
- Validation (carried out similarly to Optimatch/Haplogic): all CTs performed by the registry, high resolution data for loci examined as a CT result, no discrepancy between a priori and final HLA type. Calculate (retrospectively) the matching prognosis and compare the predicted and observed percentage of allele matches.

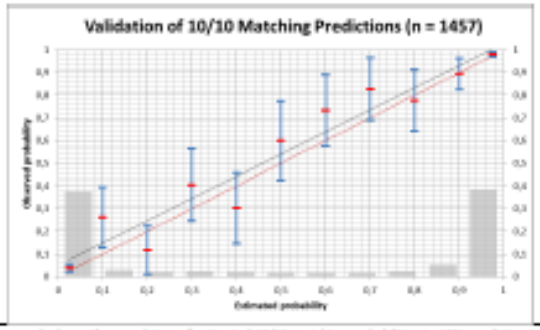
PROBLEMS

- We are facing two problems: unlike ZKRD and NMDP, other registries do not have enough donors that could be used for estimation of 5 locus high resolution haplotype frequencies. Haplotype frequencies could be calculated, but their confidence is questionable.
- Smaller registries also do not have enough CTs that could be used for validation of the prediction algorithm. ZKRD used 9843 CTs in 2008 and 22255 CTs in 2010. These numbers are not achievable by smaller registries.
- In order to overcome these problems, we have approximated the local population to the German (ZKRD) population. We have also joined CTs from multiple registries using Prometheus software.

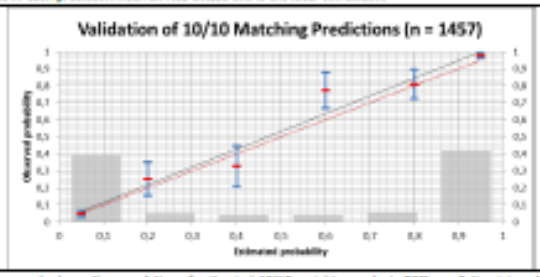
RESULTS

- In this project, we have collected more than 5000 EMDIS Sample Requests (CTs) from the Czech Stem Cells Registry, Slovak BMDR, Polish ALF Registry, Swedish Tobias Registry, Finnish BMDR, South African BMR and Ezer Mizion BMDR. 1457 of them meet our criteria and could be used for validation. We will continue to gather more CTs also from other registries who are willing to cooperate.
- We compared the predicted and observed percentage of 10/10 (resp. 9/10) matches at 10% prediction intervals. The correlation was r = 0.96 (resp. r = 0.98).

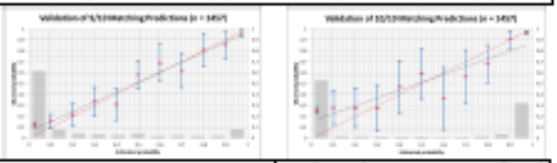
[1] Hans-Peter Eberhard: Validation of the Predictions of Optimatch®, IDRC Bern 2008
[2] Hans-Peter Eberhard: Schätzung von hochaufgelösten HLA-Haplotypfrequenzen deutscher Blutstammzellspender und ihre Anwendung bei der Patientenversorgung, 2010
[3] Jason Dehn: Haplogic III, NMDP Council Meeting 2011



The graph shows the correlation of estimated 10/10 matching probabilities in 10% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population. Blue bars show 95% confidence intervals of estimated probabilities. Since we have less CTs than the ZKRD, confidence intervals are bigger. Grey bars show relative number of CTs in each prediction interval. Red dotted line is the ideal correlation.



The graph shows the correlation of estimated 10/10 matching prob. in 20% prediction intervals and corresponding observed prob. The population model is approximated by German population.



The graph shows the correlation of estimated 9/10 matching probabilities in 10% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population. Also in this model, in order to increase the absolute number of cases within the prediction intervals (and shrink the confidence interval), you would need to decrease the number of prediction intervals by increasing the size of these intervals.

We also used European American (NMDP) population as an approximation of local populations. The results were less reliable (r=0.91) than when using the German (ZKRD) population, but very similar when decreasing the precision to 20% prediction intervals (r=0.97). The graph shows the correlation of estimated 10/10 matching probabilities in 10% prediction intervals and corresponding observed probabilities.

CONCLUSION

- The probability matching algorithm can use both German (ZKRD) and European American (NMDP-EUR) populations as an approximation for other Caucasian populations. The results are satisfactory.
- The study is limited by small number of local CTs for validation. For this reason, our validation uses 10% prediction intervals instead of the 5% intervals used by ZKRD and NMDP. Importantly the algorithm can identify donors that are more likely or less likely to be a 10/10 match.
- Haplotype frequencies are the basis for modern methods for unrelated donor searching. Retrospective analysis of several CTs has shown the prediction algorithm may speed up the donor search process.

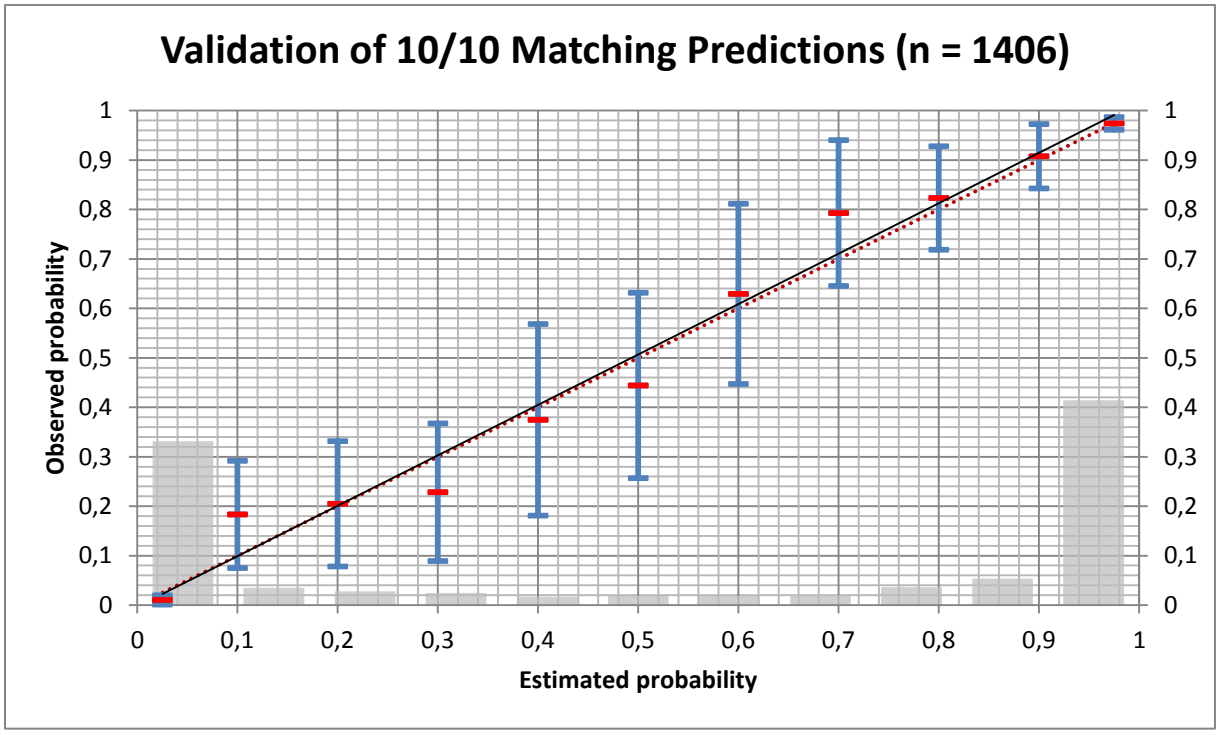
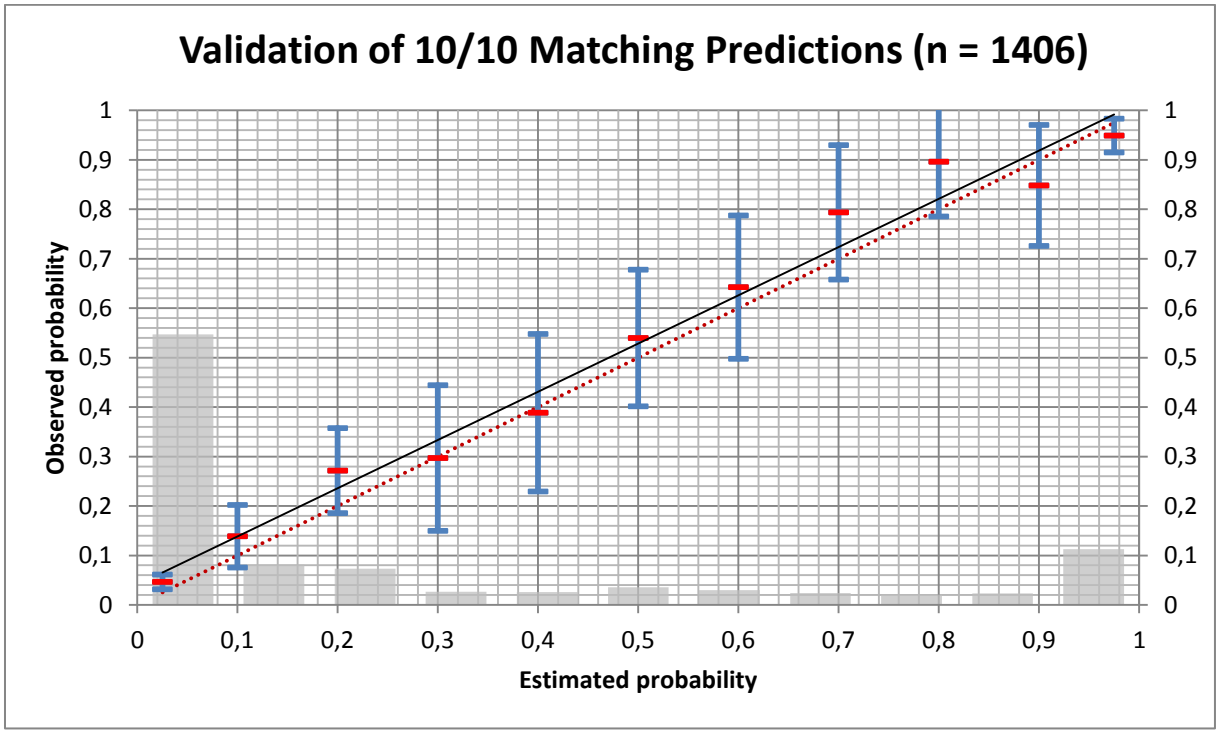


Figure 36: The graph shows the correlation of estimated 10/10 matching probabilities in 10% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. Blue bars show 95% confidence intervals of estimated probabilities. Since we have less VTs than the ZKRD, confidence intervals are bigger. Grey bars show relative number of VTs in each prediction interval. Red dotted line is the ideal correlation. The correlation is r = 0.99.



The graph shows the correlation of estimated 9/10 matching probabilities in 10% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. Also in this model, in order to increase the absolute number of cases

within the prediction intervals (and shrink the confidence interval), you would need to decrease the number of prediction intervals by increasing the size of these intervals. The correlation is r = 0.99.

Unfortunately, for validation of individual locus predictions, we don't have enough validation cases that would sufficiently fill in all 10% prediction intervals, so we have to do the validation in 20% prediction intervals.

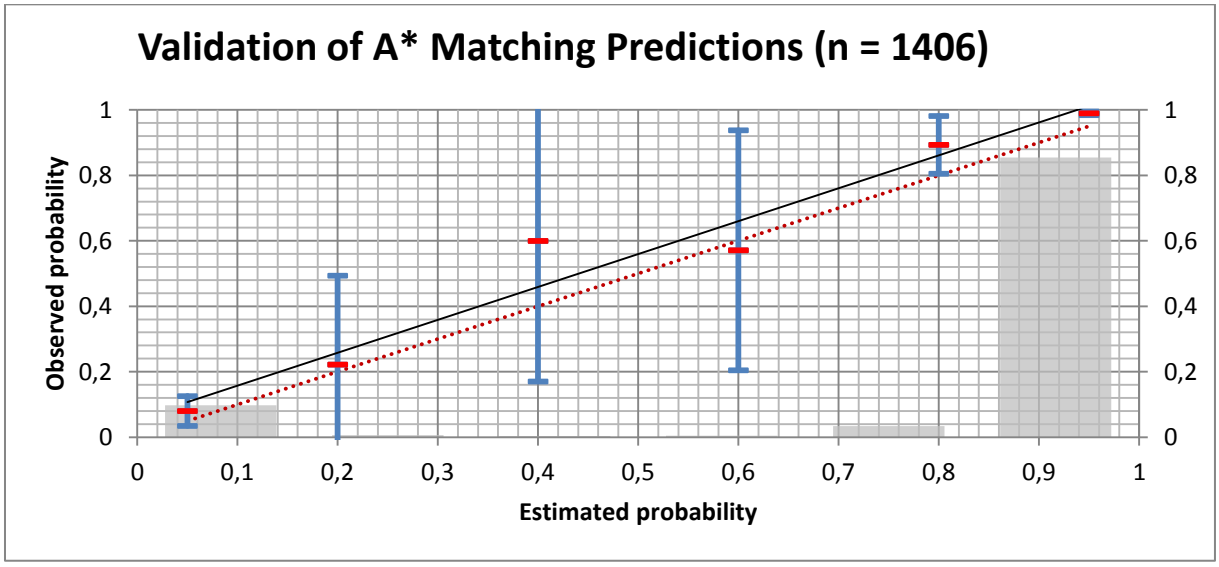


Figure 37: The graph shows the correlation of estimated A* matching probabilities in 20% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. The correlation is r = 0.98.

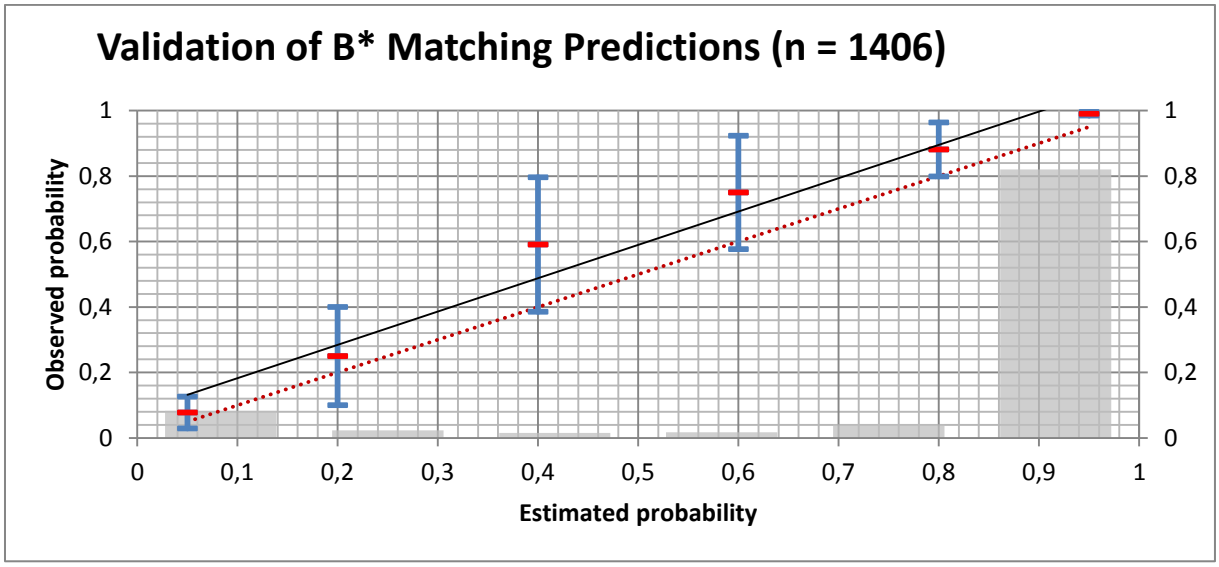


Figure 38: The graph shows the correlation of estimated B* matching probabilities in 20% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. The correlation is r = 0.98.

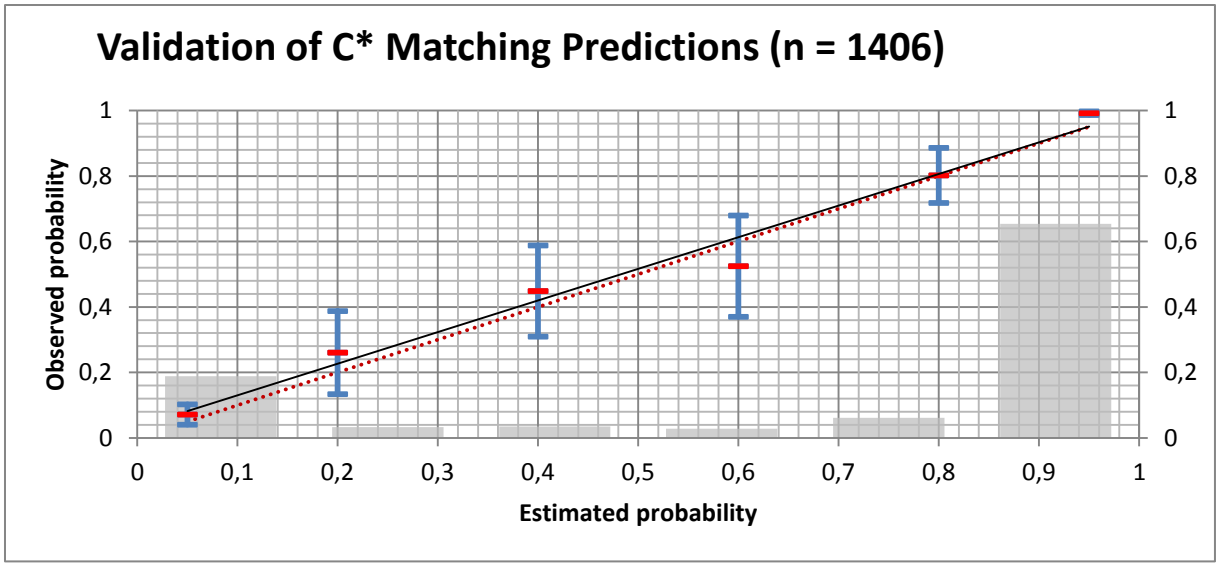


Figure 39: The graph shows the correlation of estimated C* matching probabilities in 20% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. The correlation is r = 0.997.

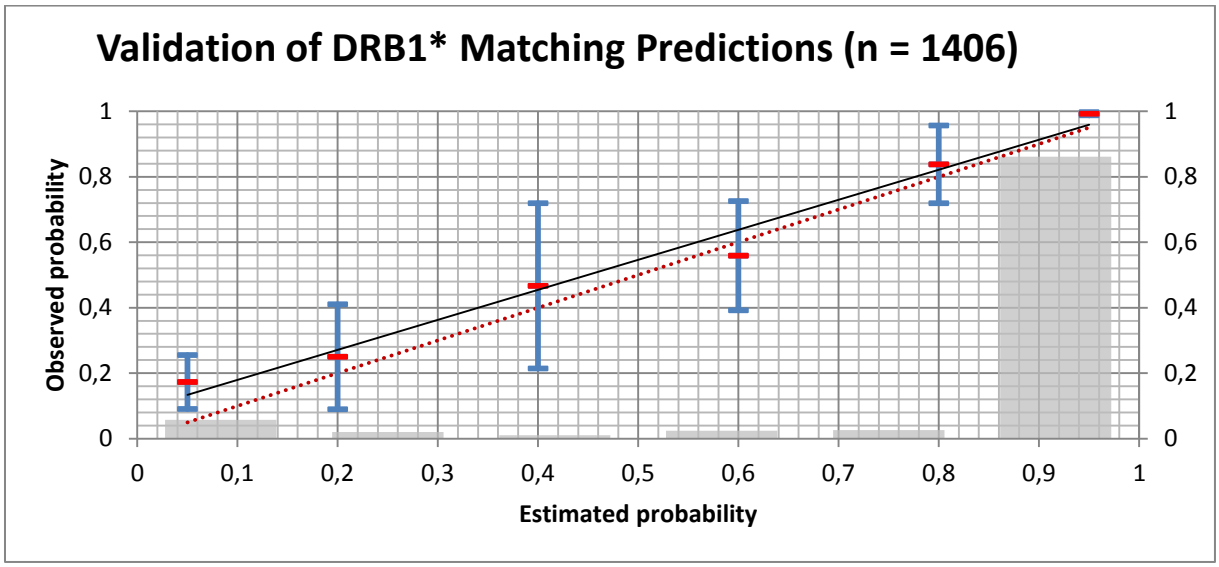


Figure 40: The graph shows the correlation of estimated DRB1* matching probabilities in 20% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. The correlation is r = 0.99.

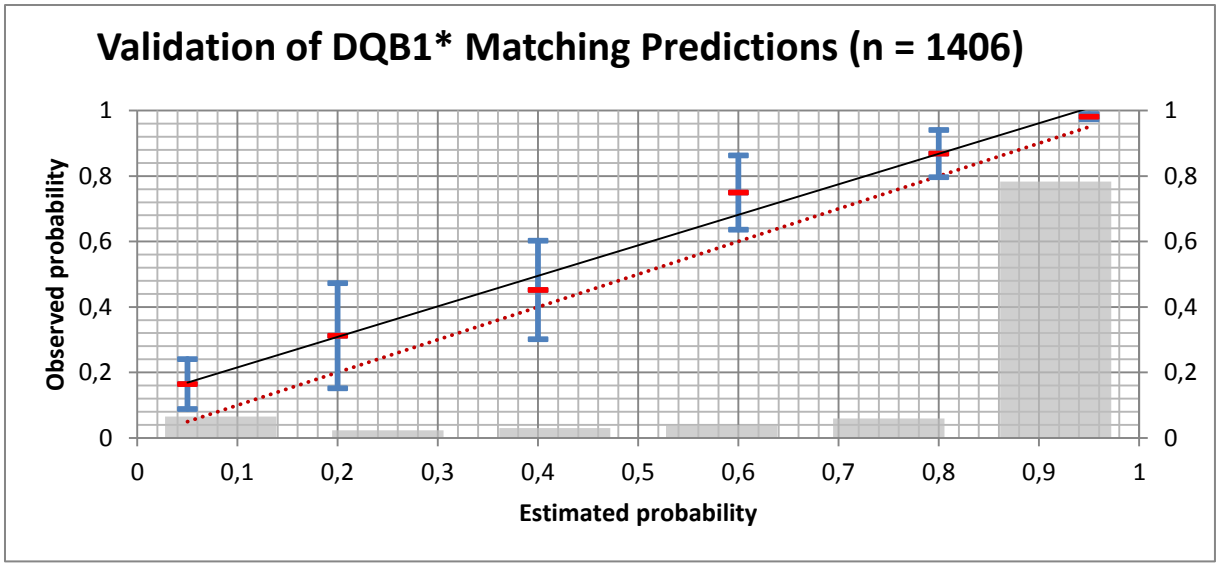


Figure 41: The graph shows the correlation of estimated A* matching probabilities in 20% prediction intervals and corresponding observed probabilities. The population model is approximated by the German population [D-1205]. The correlation is r = 0.99.

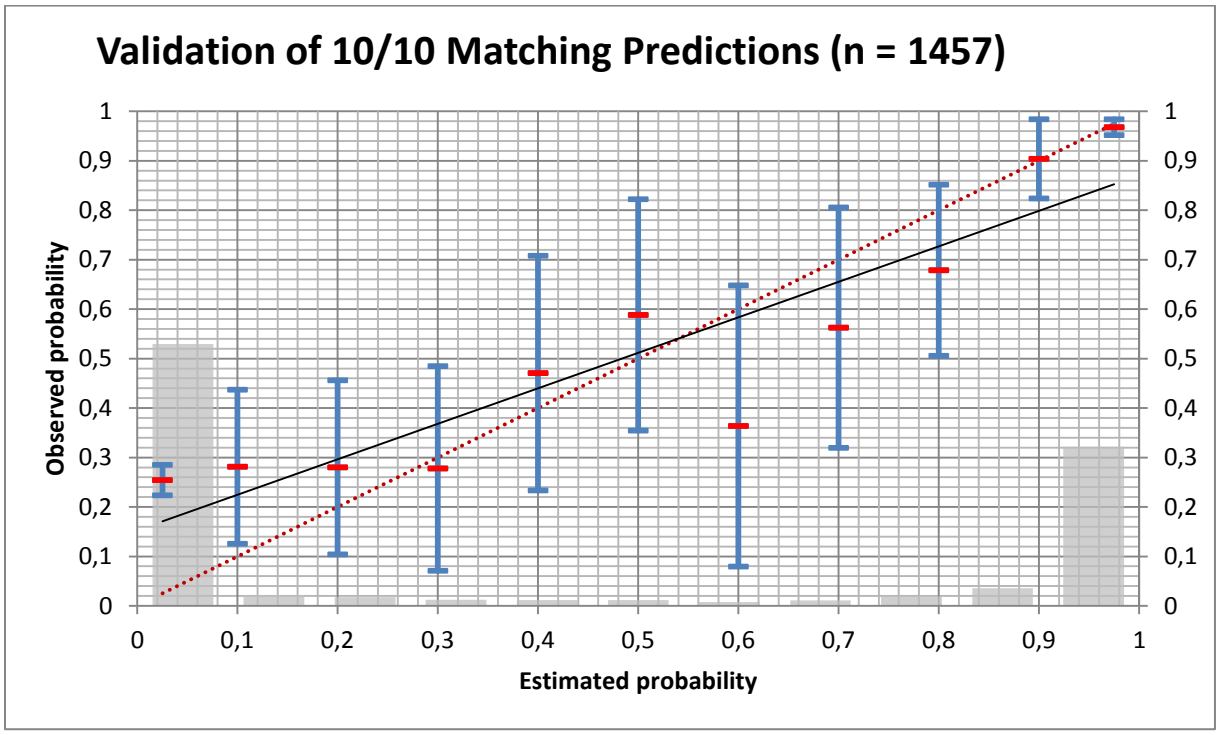


Figure 42: We also used European American (NMDP) population [62] as an approximation of local populations. The results were less reliable (r=0.91) than when using the German (ZKRD) population, but very similar when decreasing the precision to 20% prediction intervals (r=0.97). The graph shows the correlation of estimated 10/10 matching probabilities in 10% prediction intervals and corresponding observed probabilities.

Our interpretation of these results and conclusions are:

- The matching prediction tool works well (the algorithm and its implementation).

- The probability matching algorithm can use both German (ZKRD) and European American (NMDP-EUR) populations as an approximation for other Caucasian populations. The results are satisfactory.
- The study is limited by small number of local VTs for validation. For this reason, our validation uses 10% prediction intervals instead of the 5% intervals used by ZKRD and NMDP. Importantly the algorithm can identify donors that are more likely or less likely to be a 10/10 match.

If we want to distinguish usage of our predictive matching tool (ProMatch) for the registry in general vs. prediction for local donors only, we have to go further. In order to prove it works for local donors, we would need to have enough VTs for local donors that we don't have. For example when we did analysis in February 2012, there were just 20 useful EMDIS VTs in the Finnish registry database. This means we are not able to confirm the ProMatch (with German haplotype frequencies) gives reliable estimates for the Finish donors. We can only confirm it works for the mixed Caucasian population.

Intuitively, we expect the ProMatch with German haplotype frequencies will better work for populations that are closer to Germans, i.e. there is probably correlation between "genetic distance of the population of a small registry to Germans" with "reliability of ProMatch predictions". But again, we do not have enough data to prove this hypothesis.

10.3 Validation using simulated dataset

We do not have enough VTs (patient-donor-sample pairs) that would allow us to decrease the prediction intervals. For about 2000 VTs we can use only 10% prediction intervals. If we want to use 5% prediction intervals (like NMDP or ZKRD), we have to have much more VTs (at least 4000).

To overcome this problem and extensively validate the algorithm implementation, we can create simulated VTs. We have designed and implemented this method to create simulated VTs that meet our criteria (see Table 29):

a) Take the simulated dataset of Czech adult donors (see chapter 7.3). For all of them we know both simulated HLA lab typing and background high resolution typing of the artificial donor. Almost all AB typed donors have probability lower than 1% and these donors are very rarely requested for VT. In order to make it more realistic, we have excluded these donors. Donor with probability lower 1% will still form quite big group.
b) We need some patients records. We can simulate them as well, but this way all patient phenotypes would be based on our haplotype frequencies. In real world, some patients cannot be “explained” by reference haplotypes. So we will use different approach. Let’s take all patients in the CSCR registry that were registered in year 2010 and 2012 (real patient cases). We will consider only high resolution typed patients (about 50 thousand patients). We get high resolution typed patients from different ethnic groups.
c) For every donor in the set a), try to find a matching patient in the set b). Match means there is no mismatch at HLA-A, -B, -C, -DRB1 and –DQB1, i.e. patient and donor are potential match.

d) One donor in the set a) can match with multiple patients in the set b) and vice versa. But in order to keep maximum diversity of VTs and avoid bias, we will use each patient record and each donor record only once. This means, the donor-patient pair is exclusive.

These triples (simulated donor HLA typing + artificial donor typing + real patient typing) are our simulated VTs. This way, we have generated about 8000 VTs that meet our criteria!

Now we have quite a big database of VTs and we can run several validation procedures, using different haplotype frequencies.

10.3.1 German haplotype frequencies

The artificial donors were created using ZKRD reference dataset [HPE-2010]. This means these artificial donors have similar genetic background as real Germans. Our first key validation is based on our haplotype frequencies [HPE-2010]. Since we have enough VTs, we can 5% prediction intervals.

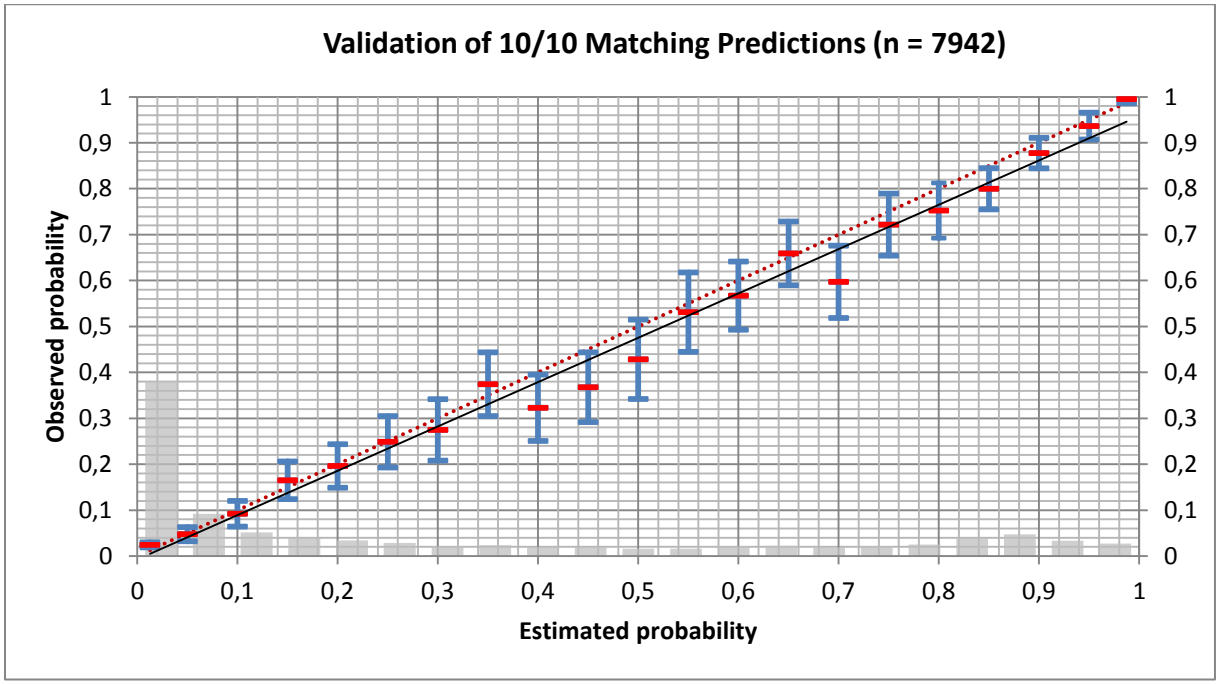


Figure 43: Validation of 10/10 matching predictions using simulated VTs and dataset [HPE-2010]. U_H = 0,3497, R=0.994

Figure 43 shows excellent results. **This validates our algorithm** design and implementation. All other pieces in the validation process are fixed: haplotype frequencies are ideal (true frequencies [HPE-2010]) and VTs are very realistic.

Even if we have excluded AB typed donors, we will still find the majority of estimated probabilities bellow 20%. However, all 5% prediction intervals have at least 126 cases. That is sufficient amount to calculate the average in all intervals.

Now, under the presumption the algorithm is validated, we can focus on validation of our German haplotype frequency estimates [D-1205].

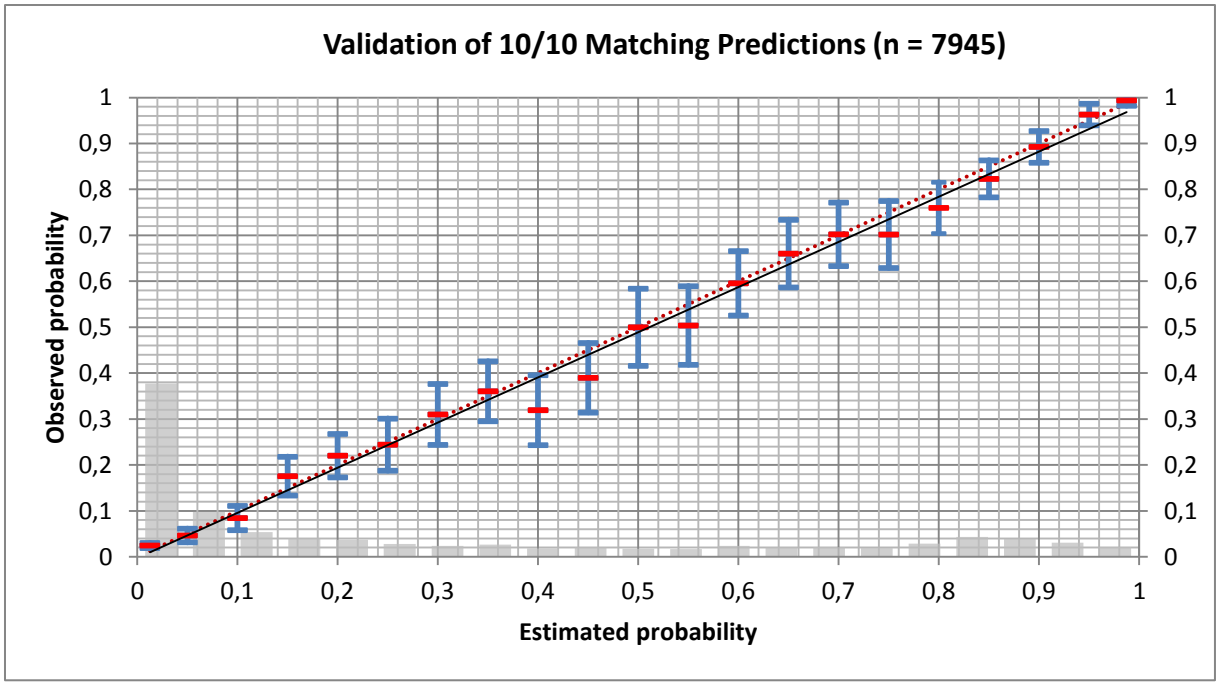


Figure 44: Validation of 10/10 matching predictions using simulated VTs and dataset [D-1205]. U_H = 0,3500798930, R=0.995

Again, Figure 44 shows excellent results, almost identical to [HPE-2010]. **This validates** the dataset [D-1205], i.e. the dataset has similar quality as the reference [HPE-2010].

10.3.2 NMDP-EUR haplotype frequencies

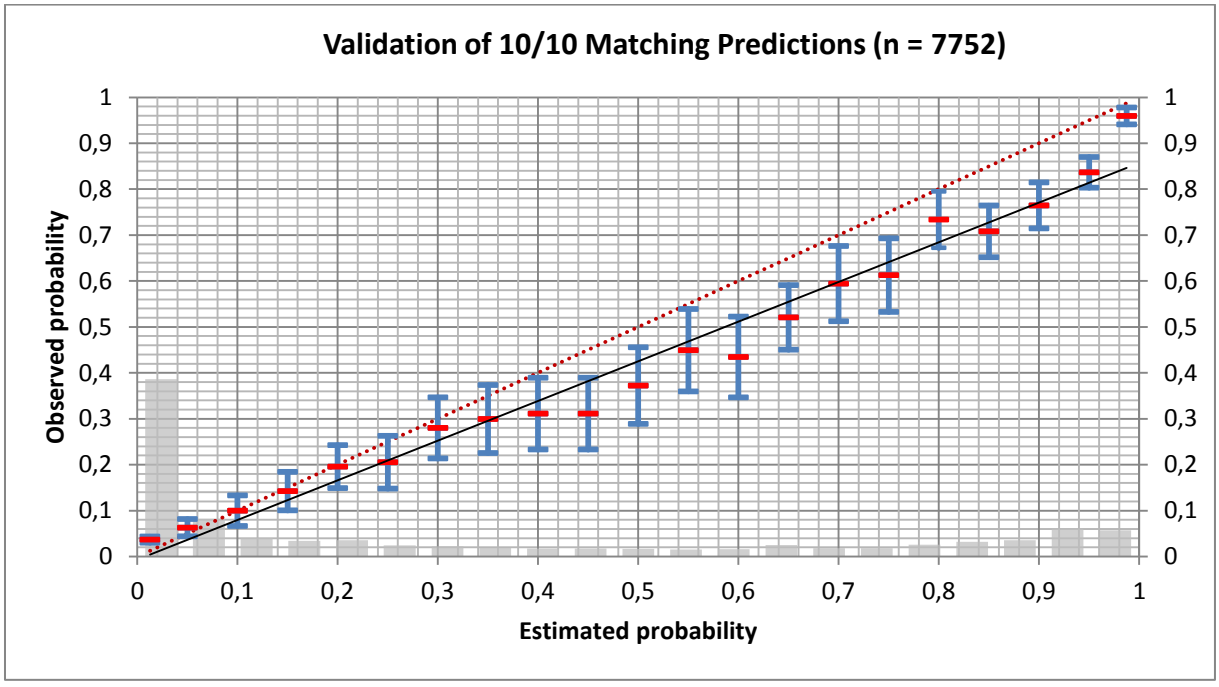


Figure 45: Validation of 10/10 matching predictions using simulated VTs and dataset [NMDP-EUR-2007], U_H = 0,4126, R=0.987

Results are slightly worsej than using German population HFE, the system underestimates the observed probabilities. Interestingly, also HapLogic III. underestimates the probabilities as well (see Figure 50).

10.3.3 Frequencies estimated from the simulated dataset

In this experiment we will estimate haplotype frequencies directly from the simulated Czech dataset and then, use them for the validation.

For HFE, we have used only 2188 best typed donors. This is still comparable to US study [62]. The result HFE dataset include only 2253 haplotypes that is much less than [HPE-2010] and [D-1205].

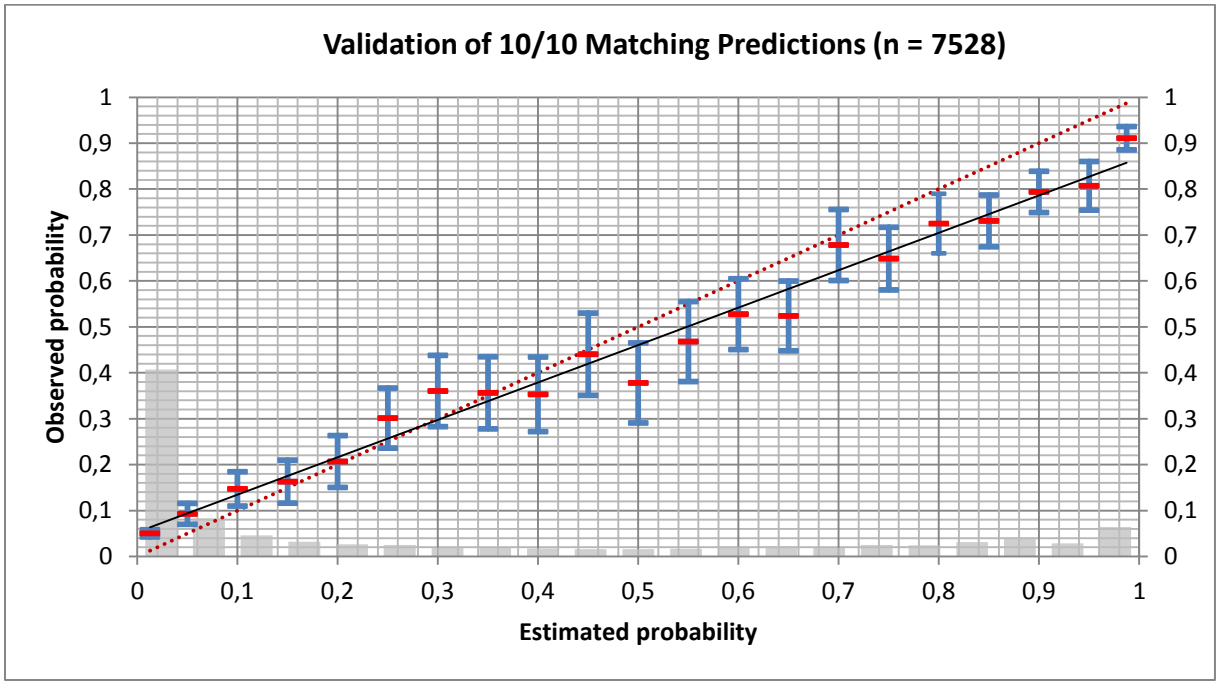


Figure 46: Validation of 10/10 matching predictions using simulated VTs and HFE from the simulated dataset (U_H = 0,4735, R = 0,9896).

Results are worse than German population HFE, but still satisfactory. This has important consequence for the Czech population HFE: given a real data of the Czech registry database (CS+CS2), we can estimate haplotype frequencies of the Czech population (see chapter 7.3). These frequencies can be used for the matching prediction algorithm and the algorithm is able to deliver satisfactory matching predictions for donors in the Czech registry database. This validation overcomes the problem of the lack of VTs that we do not have in the Czech registry database.

	Real world	Simulation
True population frequencies	Unknown	[HPE-2010]
Registry database	Czech registry database CS+CS2 (58 295 donors)	Simulated Czech registry database (58 295 donors)
HFE Algorithm	ProMatch HFE (see chapter 5.8)	
Haplotype frequencies	Czech population HFE (1237 haplotypes with frequency >= 10^{-4})	Simulated Czech population HFE (1340 haplotypes with frequency >= 10^{-4})
Prediction Algorithm	ProMatch (see chapter 9)	
Validation dataset	Hundreds of real VTs (insufficient number)	Thousands of simulated VTs (sufficient number)
Validation result	Unknown (not enough data)	Pass

Table 31: Validation of the Czech registry (population) matching prediction algorithm using simulated dataset and simulated VTs.

These results are promising, especially for registries (populations) that cannot be approximated by other population. However, its use for populations with true population frequencies that differ a lot from [HPE-2010] is questionable.

For the Czech population itself, it does not solve the question which HFEs are better for the matching prediction of the Czech donors – limited Czech haplotype frequencies [CZ-2012] or comprehensive German haplotype frequencies [D-1205]? We are not able to answer this question, mainly thanks to insufficient number of real VTs for Czech donors.

10.4 Situation in the world

10.4.1 OptiMatch®

The validation of OptiMatch® (see also chapter 11.3.1) has been done in 2008 with 9843 CTs that satisfy these conditions [8]:

- No high resolution data for the locus / loci examined at the time of request
- High res data for the locus / loci examined obtained as a CT result (or even later)
- No discrepancy between the a priory and final HLA type

For this file of CTs, the ZKRD has calculated (retrospectively) the matching prognosis of OptiMatch® and compared the predicted and observed percentage of allele matches in 5% prediction intervals. Results are shown on the Figure 47.

As we have mentioned before, we have adopted the same method in chapter 10.2

Another published validation of OptiMatch® has been done in 2010 with 22255 CTs. Results are shown on the Figure 48. These results are excellent and there is no doubt OptiMatch® is very accurate in the HLA predictions of German donors.

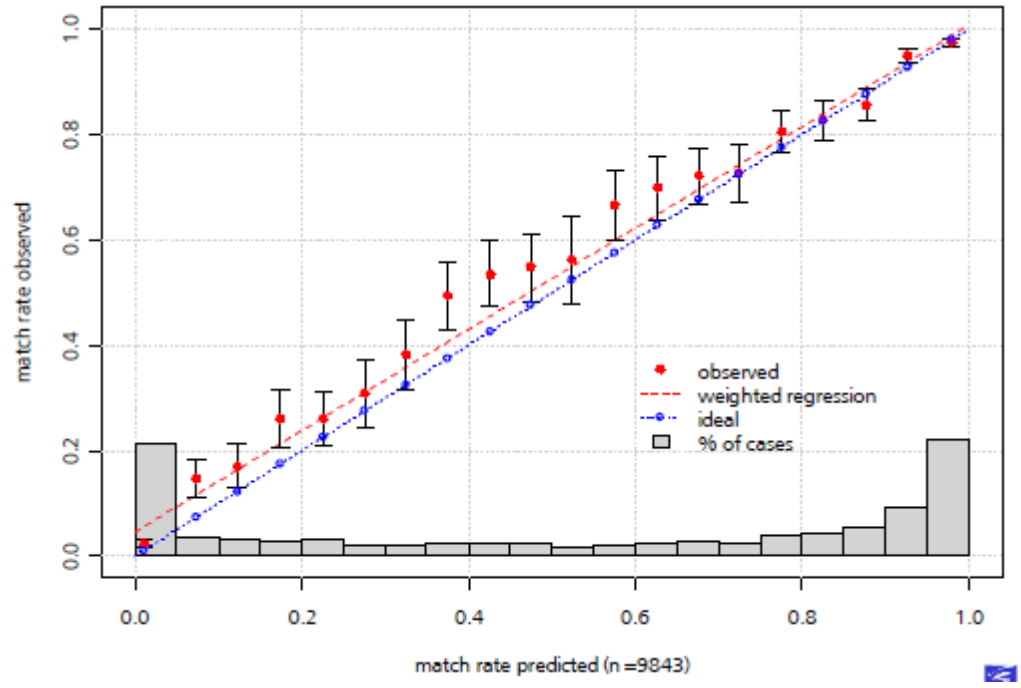


Figure 47: Validation of 10/10 matching predictions of the OptiMatch® system in 2008 using 9843 CTs [8]

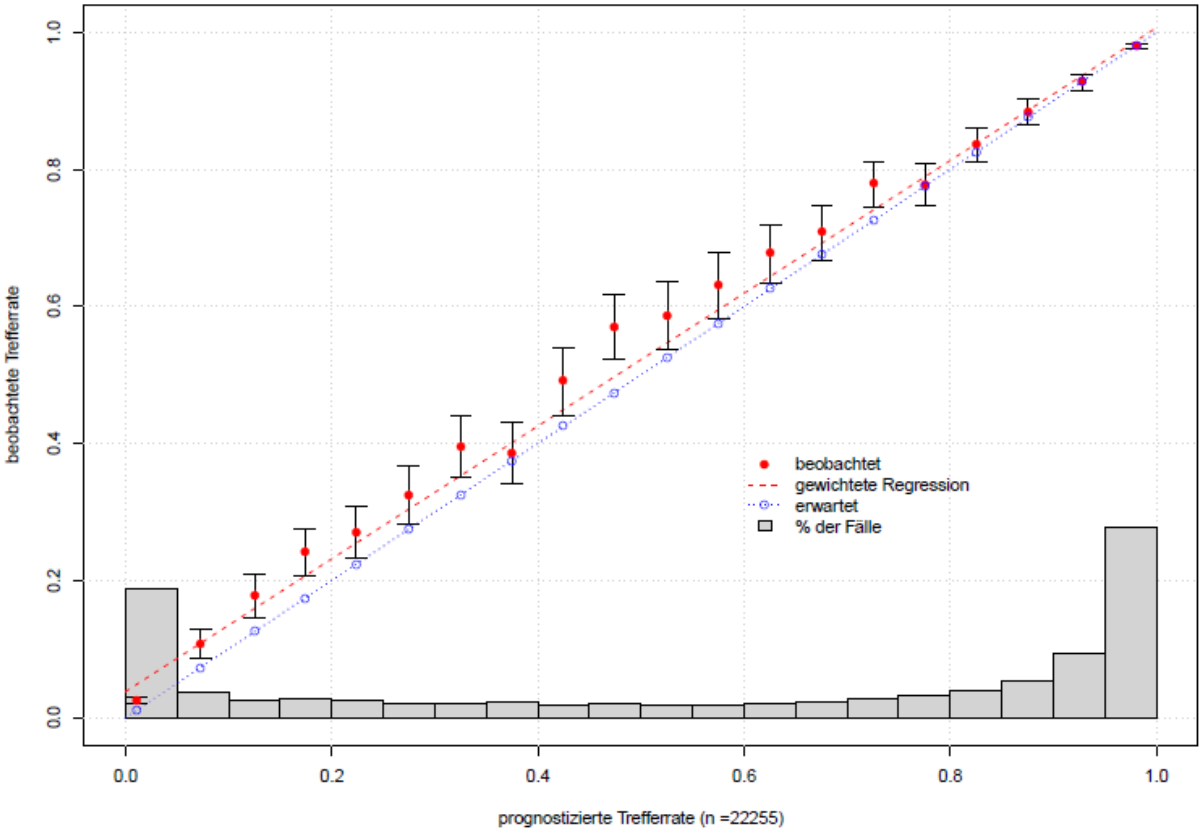


Figure 48: Validation of 10/10 matching predictions of the OptiMatch® system in 2010 using 22255 CTs [63]

10.4.2 Haplogic™

Validation methods of Haplogic™ (see also chapter 11.3.2) are probably similar as those used by OptiMatch® however details have not been published. Haplogic™ II. results are shown on the Figure 49 and Haplogic™ results on the Figure 50. Haplogic™ takes into account the ethnic group of the donor, so it has to use several sets of HFE, which is very interesting feature of the system. But it is not clear if NMDP does single validation using CTs from all ethnic groups or if it does validations per ethnic group and what are the numbers of CTs. HapLogic™ III currently uses 21 ethnic groups.

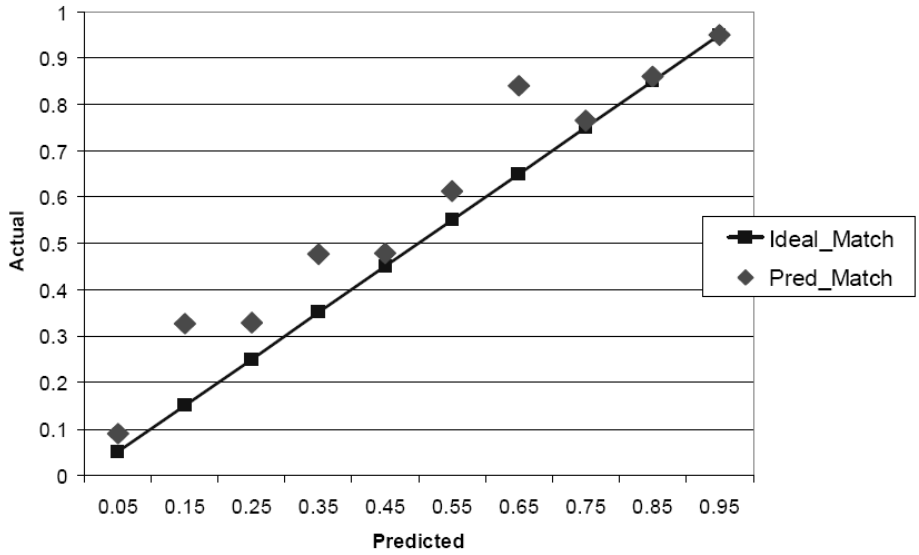


Figure 49: Validation of 6/6 matching predictions of the HapLogic II system [graph provided by NMDP]

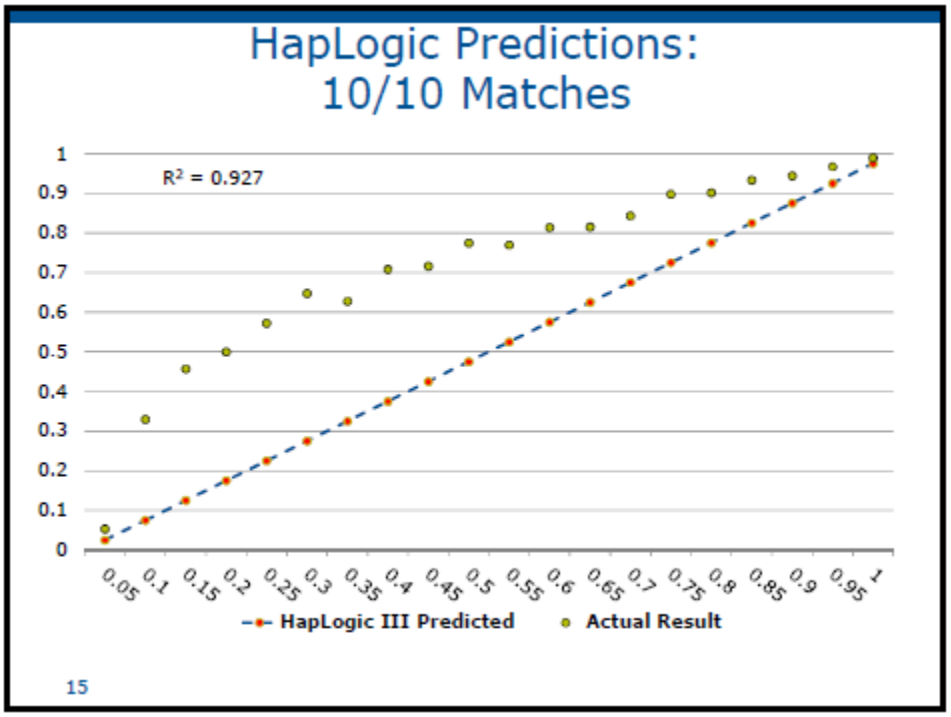


Figure 50: Validation of 10/10 matching predictions of the HapLogic™ III system [80]

11. Implementation of matching prediction methods

This chapter presents applications of the algorithms and tools in daily operation of stem cell donor registries.

11.1 ProMatch system

Our implementation of the matching prediction method is called **ProMatch** (**Pro**babilistic **Match**ing). This functionality has been integrated with the Prometheus system [28] – software for stem cell donor registries used in more than 20 countries, mainly in Europe. This was the key step towards practical usage of these methods in registry operations.

11.2 User interface

Donor search results in Prometheus software are presented in the table. User can switch between deterministic matching ("Best First by Match Grade") and the new probabilistic matching ("Best First by Probability"), see Figure 51. This feature is not common in other systems (OptiMatch® and HapLogic™).

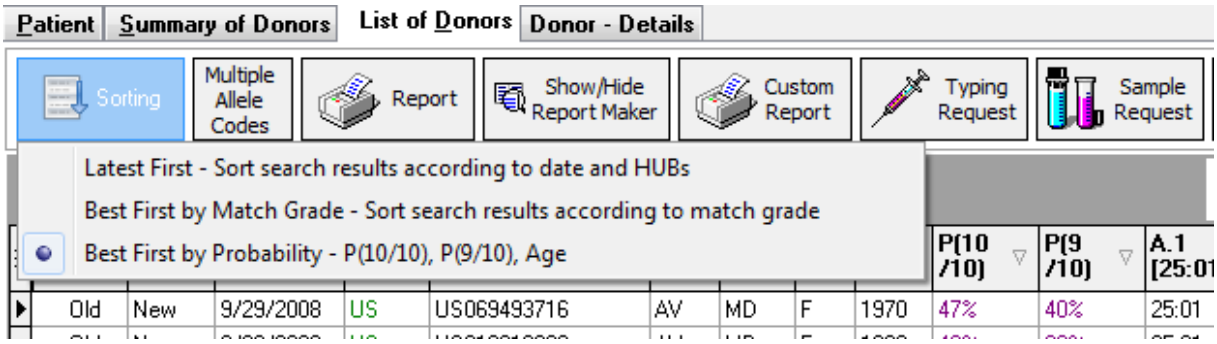


Figure 51: ProMatch – sorting options of the donor search results: Time, Deterministic matching and Probability matching.

Potential donors are listed in the table. The system displays:

- The probability of 10/10 HLA-A, -B, -C, -DRB1 and –DQB1 allele match, calculated by (29), column "P(10/10)".
- The probability of 9/10 HLA-A, -B, -C, -DRB1 and –DQB1 allele match, calculated by (30), column "P(9/10)".
- Probabilities of HLA-A, -B, -C, -DRB1 and –DQB1 allele match at individual loci, calculated by (31), columns "P(A)", "P(B)", "P(C)", "P(DRB1)" and "P(DQB1)".

Sorting "Best First by Probability" means donors are sorted by "P(10/10)", then by "P(9/10)", see Figure 52 and Figure 53.

Display method corrects probabilities by deterministic matching observations:

- Presented probabilities are rounded and displayed in per cents (0 – 100%).
- If patient and donor are not mismatched at specific locus (potential match) and displayed rounded value would be 0%, it is corrected to 1%.
- If patient and donor typing do not have the same high resolution allele codes (potential match) and the displayed rounded value would be 100%, it is corrected to 99%.

	CZ47549P Test Patient 2						TC CZBRN			A 25:01 29:01	B 18:01	C 07:01 12:03	DRB1 11:04 15:01		DQB1 03:01 06:02	
P(10/10)	P(9/10)	A.1 [25:01]	A.2 [29:01]	P(A)	B.1 [18:01]	B.2	P(B)	C.1 [07:01]	C.2 [12:03]	P(C)	DR.1 [11:04]	DR.2 [15:01]	P(DRB1)	DQ.1 [03:01]	DQ.2 [06:02]	P(DQB1)
47%	40%	25:01	29:ANTJ	49%	18:AERV	18:AESD	99%			87%	11:ARAP	15:VYF	99%			96%
43%	36%	25:01	29:ANTJ	47%	18:AERV	18:AESD	99%	07:ANNG	12:NP	99%	11:KBB	15:STF	80%			92%
39%	33%	25:01	29:ATDP	42%	18:AC	18:XX	99%			83%	11:AD	15:AB	84%			96%
37%	32%	25	29	39%	18		99%			82%	11:XX	15:XX	79%	03:XX	06:XX	97%
37%	32%	25	29	39%	18		99%			82%	11:XX	15:XX	79%	03:XX	06:XX	97%
37%	31%	25	29	39%	18		99%			82%	11	15	79%			96%
37%	31%	25	29	39%	18		99%	7		82%	11	15	79%			96%
36%	31%	25	29	40%	18		99%			83%	11:ARX	15:ADE	77%			92%
36%	31%	25	29	40%	18		99%			83%	11:ARX	15:ADE	77%			92%
35%	31%	26:AZRB	29:XX	39%	18:XX		99%			81%	11:AU	15:AZRM	80%			92%
7%	6%	25:XX	29:XX	22%	18:XX		99%			43%			15%			19%
7%	6%	25	29	22%	18		99%	7		43%			15%			19%
7%	6%	25	29	22%	18		99%			43%			15%			19%
7%	6%	25:01	29:ANTJ	22%	18:AHUA	18:XX	99%			44%			15%			19%
7%	6%	25:01	29:ANTJ	22%	18:AHUA	18:XX	99%			44%			15%			19%
7%	6%	25	29	22%	18		99%			43%			15%			19%
7%	6%	25	29	22%	18		99%	7		43%			15%			19%
7%	6%	25	29	22%	18		99%	6		43%			15%			19%
7%	6%	25	29	22%	18		99%			43%			15%			19%
7%	6%	25	29	22%	18		99%	7		43%			15%			19%

Figure 52: ProMatch – example of donor search results (probability matching). The main sorting criteria is the probability of 10/10 HLA-A, -B, -C, -DRB1 and –DQB1 match, see column P(10/10).

P(10/10)	P(9/10)	A.1 [25:01]	A.2 [29:01]	P(A)	B.1 [18:01]	B.2	P(B)	C.1 [07:01]	C.2 [12:03]	P(C)	DR.1 [11:04]	DR.2 [15:01]	P(DRB1)	DQ.1 [03:01]	DQ.2 [06:02]	P(DQB1)
7%	6%	25	29	22%	18		99%			43%			15%			19%
1%	1%	25	19	1%	18		99%			7%			1%			3%
0%	97%	25	2	0%	18:CHAY		99%	07:XX	12:AUCW	99%	11:UBR	15:BKET	99%	03:CFTD	06:WG	99%
0%	84%	25:XX	02:XX	0%	18:XX		98%			89%	11:04	15:01	100%			96%
0%	68%	25:XX	02:XX	0%	18:XX		99%	07:XX	12:XX	98%	11:XX	15:XX	73%			96%
0%	44%	25	32	0%	18		95%			64%	11:XX	15:XX	74%	03:XX	06:XX	91%
0%	41%	25	32	0%	18		95%			63%	11	2	70%			85%
0%	35%	25	1	0%	18		98%			55%	11:XX	15:XX	77%			96%
0%	35%	25:XX	24:XX	0%	18:XX		99%			59%	11:XX	15:XX	71%			96%
0%	26%	25	3	0%	18		98%			59%	11	15	52%			95%
0%	26%	25	11	0%	18		58%			81%	11:XX	15:XX	77%			93%
0%	26%	25	3	0%	18		98%			59%	11	15	52%			95%
0%	8%	25	33	0%	18		97%			33%			28%			31%
0%	8%	25	33	0%	18		97%			33%			28%			31%
0%	8%	25	33	0%	18		97%			33%			28%			31%
0%	8%	25	33	0%	18		97%			33%			28%			31%
0%	7%	25	31	0%	18		99%			62%			11%			21%
0%	7%	25	32	0%	18		97%			52%			12%			17%
0%	7%	25:01	29:02	0%	18:RRG	18:RRG	99%			34%			10%			16%
0%	7%	25	31	0%	18		99%			62%			11%			21%
0%	7%	25	32	0%	18		97%			52%			12%			17%

Figure 53: ProMatch – example of donor search results (probability matching). The second sorting criteria is the probability of 9/10 HLA-A, -B, -C, -DRB1 and –DQB1 match, see column P(9/10).

11.3 Situation in the world

Until 2011, only two HLA matching prediction systems were available in Germany and the United States. They have been implemented by two biggest registries in the world – ZKRD and NMDP

– that have invested a lot of resources in R&D . Names of these systems are registered and protected: OptiMatch® and Haplogic™.

Except these two systems and our work, some activities are being done by the German donor centre DKMS. Their Hap-E system is used only internally [79].

11.3.1 OptiMatch®

OptiMatch® [81] [8] is a matching program calculating, for each donor, the probability to be allele identical to the patient. The program is developed and used by the German registry ZKRD.

First version (since October 2006) was based on 3 locus high resolution haplotype frequencies had sorting of potential donors according to the probability of 6 of 6 allele match probability (HLA-A, -B and -DRB1) and secondary sorting on HLA-C and HLA-DQB1 matching probabilities, age and gender. The current version (since June 2008) is based on 5 locus high resolution haplotype frequencies (HLA-A,-B,-C,-DRB1 and –DQB1).

OptiMatch® is able to do serology to DNA mapping, so predictions are calculated also for serology typed donors. Current version's primary matching can be based on the probability of matching 6 of 6, 8 of 8 (including C or DQB1) or 10 of 10 (including both) alleles, and then the probability of 1 or, finally, 2 allele mismatches.

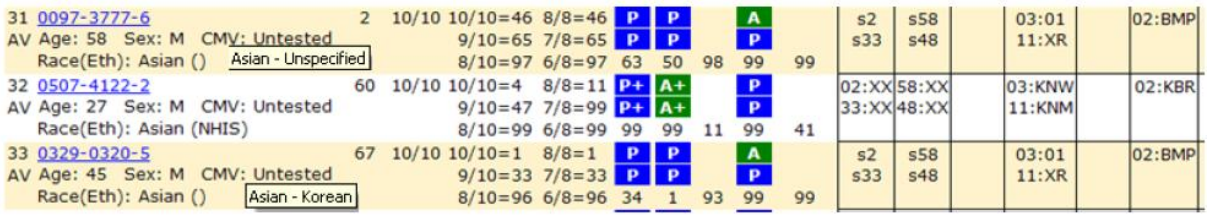
User-friendly web based user interface shows a list of potential donors with 7 probabilities: A* match, B* match, C* match, DRB1* match, DQB1* match and overall probabilities of 10/10 match and 9/10 match.

11.3.2 Haplogic™

HapLogic™ I. [27] was developed and used by NMDP registry since 2006. It works in similar way like OptiMatch®. It calculates the likelihood of allele-level matching based on calculated HLA haplotype frequencies within major American racial and ethnic populations. HapLogic™ I. predicted high-resolution matching at HLA-A, -B and -DRB1 (6 of 6 allele match, 5 of 6 allele match and 2-allele match at each of the three loci) [82] [83] [84].

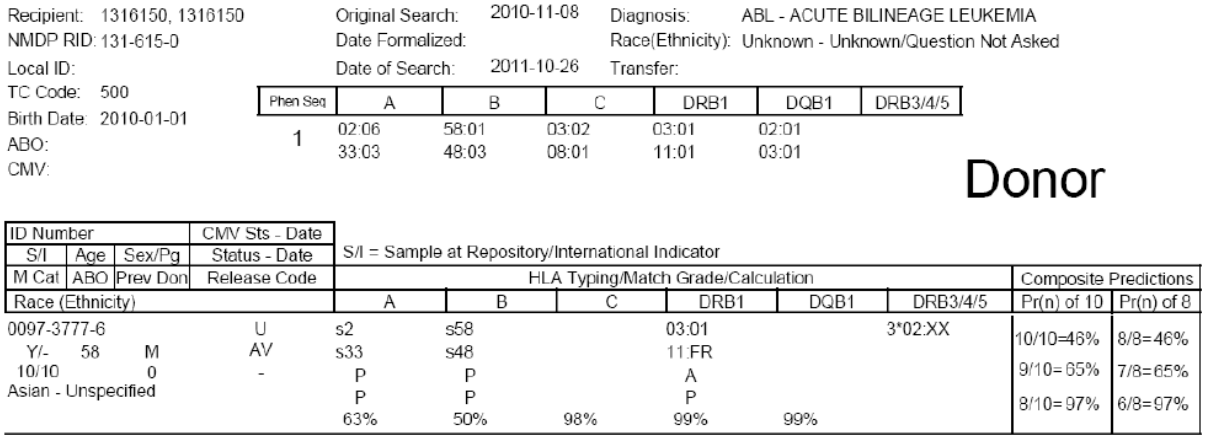
HapLogic II. (2008) is able to incorporate HLA-C and HLA-DQ matching (2-allele match). The latest version III, introduced in November 2011, sorts donors based on probability of matching 10 alleles, using 5 locus high resolution haplotypes (like OptiMatch®). HapLogic also uses 5 broad and 21 detailed race/ethnic groups.

The web based user interface shows a list of potential donors with several probabilities: A* match, B* match, C* match, DRB1* match, DQB1* match and overall probabilities of 10/10 match, 9/10 match, 8/10 match, 8/8 match, 7/8 match, 6/8 match and for cord blood units also 6/6 match, 5/6 match and 4/6 match. Screenshot of the user interface is shown on the Figure 54 and example of the printed report on the Figure 55.



31	0097-3777-6	2	10/10	10/10=46	8/8=46	P	P		A		s2	s58		03:01		02:BMP
AV	Age: 58 Sex: M CMV: Untested			9/10=65	7/8=65	P	P		P		s33	s48		11:XR		
	Race(Eth): Asian () Asian - Unspecified			8/10=97	6/8=97	63	50	98	99	99						
32	0507-4122-2	60	10/10	10/10=4	8/8=11	P+	A+		P		02:XX	58:XX		03:KNW		02:KBR
AV	Age: 27 Sex: M CMV: Untested			9/10=47	7/8=99	P+	A+		P		33:XX	48:XX		11:KNM		
	Race(Eth): Asian (NHIS)			8/10=99	6/8=99	99	99	11	99	41						
33	0329-0320-5	67	10/10	10/10=1	8/8=1	P	P		A		s2	s58		03:01		02:BMP
AV	Age: 45 Sex: M CMV: Untested			9/10=33	7/8=33	P	P		P		s33	s48		11:XR		
	Race(Eth): Asian () Asian - Korean			8/10=96	6/8=96	34	1	93	99	99						

Figure 54: Screenshot of Haplogic™ III [80]



Recipient: 1316150, 1316150
NMDP RID: 131-615-0
Local ID:
TC Code: 500
Birth Date: 2010-01-01
ABO:
CMV:

Original Search: 2010-11-08
Date Formalized:
Date of Search: 2011-10-26

Diagnosis: ABL - ACUTE BILINEAGE LEUKEMIA
Race(Ethnicity): Unknown - Unknown/Question Not Asked
Transfer:

Phen Seq	A	B	C	DRB1	DQB1	DRB3/4/5
1	02:06 33:03	58:01 48:03	03:02 08:01	03:01 11:01	02:01 03:01	

Donor

ID Number			CMV Sts - Date
S/I	Age	Sex/Pg	Status - Date
M Cat	ABO	Prev Don	Release Code
Race (Ethnicity)			

S/I = Sample at Repository/International Indicator

HLA Typing/Match Grade/Calculation						Composite Predictions	
A	B	C	DRB1	DQB1	DRB3/4/5	Pr(n) of 10	Pr(n) of 8
s2	s58		03:01		3*02:XX	10/10=46%	8/8=46%
s33	s48		11:FR			9/10=65%	7/8=65%
P	P		A			8/10=97%	6/8=97%
P	P		P				
63%	50%	98%	99%	99%			

0097-3777-6 U
Y/- 58 M AV
10/10 0 -
Asian - Unspecified

Figure 55: Printed report of Haplogic™ III [80]

12. Contribution of the work

Main contributions of this work are:

- **Design and implementation of Haplotype Frequencies Estimation algorithm and further exploration and extension of underlying methods**
  - We have given an overview of different methods for HFE (chapter 4.3).
  - We have designed and implemented powerful algorithm (based on EM algorithm) and tool for HFE that uses real HLA data of stem cell donor registries. Several tricks that decreases computational costs, i.e. time and memory were included (chapter 5).
  - We have used $\tilde{c}_j$ … a method that transforms qualitative parameters of the HLA typing results of an individual to the quantitative attributes (chapter 5.8.2).
  - We have done research of reliability of HFE algorithm on registry datasets. New framework that can simulate real stem cell donor registry and estimates reliability of HFE (chapter 6) was presented.
- **Probability Matching algorithm and its validation**
  - We have designed and implemented the algorithm for the prediction of HLA match by top-down design (chapter 9).
  - We have introduced new concept of partial haplotypes (chapter 9.4).
  - We have validated the HLA match prediction algorithm using both real and simulated datasets (chapter 10).
- **Real data and deployment of the software into routine operations**
  - We have estimated most accurate HLA haplotype frequencies for several populations. HFE of some populations have never been published (Hungary, Slovakia, Nigeria, etc.). These haplotypes have several applications, not only in the medicine (chapter 7).
  - The most importantly, the work has practical benefits for the patients. Results of the work (the software) have been deployed in several countries and it is used in daily operations of several stem cell donor registries around the world (chapter 11).
  - Main benefits are: it helps search coordinators to identify easy, difficult and (almost) futile donor searches, to predict the level of patient-donor matching realistically achievable, speed up the donor search by choosing the most promising candidates and avoiding detours and make ultra-urgent searches feasible in spite of ambiguous or missing HLA data [8]. The speed at which a suitable donor is identified can significantly impact patient survival [2].

13. Conclusion and future work

A reliable and efficient search algorithm is the key component of the unrelated stem cell donor registry computer system. In our previous work [5] we have implemented combinatorial search algorithm that compares patient with donors by counting all known and visible HLA mismatches. In this work we have designed and implemented a new probabilistic matching method. The production software system combines both methods together, the first one for rough pre-selection and the second one for fine grading and sorting.

In the first part of the work, we have given the overview of search algorithms, their design and implementation aspects (chapter 33.1). A top-down design approach that first lists algorithm requirements, specifies input and output parameters and then goes deeper into details, was selected. The importance of validation prior to the deployment of a new matching algorithm has been emphasized (chapter 3.5).

In the introduction, we have posed these questions that represent underlying goals (chapter 1.1):

How can we design and implement algorithm that creates population model?

Haplotype frequencies are the basis for modern methods for unrelated donor searching. However, the problem of estimation of HLA gene and haplotype frequencies of a human population is very difficult (chapter 5.1). We have mathematically formulated the problem (chapter 4.2). Then we provided an overview of all methods that could be used for its solution (chapter 4.3). Different methods were discussed, especially its possible usage for databases of stem cell donor registries (chapter 4.4). Bayesian methods are also promising and worth further investigation (chapter 4.3.5). But currently we think the maximum likelihood approach with the Expectation-Maximization algorithm is the best approach in our situation (chapter 4.5). Properties of the algorithm (chapter 4.6) and reliability of results were discussed (chapter 4.7). We have shown the complexity of HLA system and databases of stem cell donor registries and reasons for its computational difficulties (chapter 5.1).

We have proposed a framework of arbitrary HLA typing resolution as user-specified input and output of the EM algorithm (chapter 5.2). It is generalization of all previous efforts of dealing with data of multiple typing resolutions. Several methods of handling missing values were discussed and compared (chapter 5.3). We have presented some examples and results of experiments that show these methods cannot be easily applied for serology to DNA mapping. We have proposed a modification of the EM algorithm that solves the problem (chapter 5.4).

The EM algorithm in our context is very computationally demanding (chapter 5.7). In our implementation (chapter 5.8), we have used several optimizations that speed up the process and save computer memory.

We have presented the situation in the world and overview of the state-of-the-art HLA haplotype frequencies estimation programs (chapter 5.9). Our implementation was compared with these programs in the international workshop project that tested behavior of EM algorithm in controlled data environment and within the scope of this exercise it provided similar results as algorithms of other international research groups (chapter 5.10).

What are the properties and reliability of the model (HFE) in general?

We have approximated local populations by its stem cell donor registry datasets of different sizes and structures. In order to better understand the quality of the result model, we have studied different properties of the EM algorithm in the controlled data environment. We have inspected quality dependencies on typing ambiquities (chapter 6.1 and chapter 6.2), population size (chapter 6.3), sample size (chapter 6.3), population homogeneity (chapter 6.4) and restriction of computational complexity (chapter 6.5). The final simulation of real stem cell donor registry dataset

has combined all these aspects together and provided approximation of the distance of HFE and true population frequencies (chapter 6.6).

We have applied our methods and estimated HLA-A*-B*-C*-DRB1*-DQB1* haplotype frequencies for Czech, Slovak, Hungarian, Finnish, Swedish, Cypriote, South African and Nigerian populations on the best possible resolution (chapter 7). Such precise estimates of these populations have never been published. Our results have been already used in different analyses of stem cell donor registries in these countries.

But possible usage of the data exceeds the field of stem cell transplantation. We have presented some examples of other applications (chapter 8).

How can we design and implement the probabilistic matching algorithm?

We have defined criteria for the matching prediction algorithm (chapter 9.1) and then designed the new computational method (chapter 9.2 and chapter 9.3). A lot of intention has been dedicated to special cases, where standard method fails and patient or donor phenotypes cannot be resolved (chapter 9.4). We have proposed a system of so called artificial haplotypes and their usage in matching predictions. This proposal has been validated on real data (chapter 9.5).

How can we validate the whole system? Can we apply it for all registries and populations?

The search algorithm cannot be deployed, if it is not validated. The crucial element of validation is the availability of sufficient amount of data (validation cases). Five years ago, most of registries in our interest had all these data only in paper form. Since then, we have implemented and deployed automated software systems (implementation of EMDIS) in more than 15 countries that support daily operations of these registries. One of the outputs of these efforts was the database of validation cases in electronic format that was used in this work for validation of the matching prediction algorithm and HFE. We have collected more than 1400 validation cases, but still it was not enough for detail validation (chapter 10.2).

We have done also another validation, using simulated datasets (chapter 10.3). By this method we have validated both our probabilistic algorithm and HFE [D-1205], the approximation of European Caucasian population model.

This work was not only academic research. Designed algorithms and methods have been implemented (chapter 11) and deployed in several countries in Europe and help search coordinators of stem cell donor registries in daily work to find the best match for patients in need. First registry that adopted these algorithms was the Czech Stem Cell Registry in Prague. Nowadays, match lists for all Czech patients are ranked and can be sorted by matching probabilities. This helps to identify difficult searches, predict realistically achievable results and speed up the donor search.

Deployment of the system in several other countries is on the way, for example in Finland, Sweden, Switzerland, Slovakia, Belgium, England, Ireland, etc. We are in touch with all of them.

There are also some countries that are interested as well, but we don’t have reliable solution yet. These are populations that do not belong to European Caucasian group, such as South Africa,

Argentina, Saudi Arabia, etc., but also some minorities in Europe, such as gypsies. The problem is we cannot approximate them by Caucasians and we don't have enough data for estimation of their own high resolution haplotype frequencies. We also don't have enough validation data to verify any kind of proposed solution. Overcoming of these problems will be our future work.

Bibliography

[1] Wikipedia, "Hematopoietic stem cell transplantation," [Online]. Available: http://en.wikipedia.org/wiki/Hematopoietic_stem_cell_transplantation.

[2] S. Lee, J. Klein, M. Haagenson, L. A. Baxter-Lowe, D. L. Confer, M. Eapen, M. Fernandez-Vina, N. Flomenberg, M. Horowitz, C. K. Hurley, H. Noreen, M. Oudshoorn, E. Petersdorf and others, "High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation," *Blood,* vol. 110, no. 13, pp. 4576-4583, 2007.

[3] L. Foeken, A. Green, C. Hurley, E. Marry, T. Wiegand, M. Oudshoorn and D. R. W. G. World Marrow Donor Association (WMDA), "Monitoring the international use of unrelated donors for transplantation: the WMDA annual reports," *Bone Marrow Transplantation,* vol. 45, p. 811–818, 2010.

[4] D. Steiner, "Computer Algorithms in the Search for Unrelated Stem Cell Donors," *Bone Marrow Research,* Vols. 2012, doi:10.1155/2012/175419, p. Article ID 175419, 2012.

[5] D. Steiner, "Search for Unrelated Bone Marrow Donors," Diploma Thesis, FEL ČVUT, 2007, 2007.

[6] BMDW, "BMDW Database," [Online]. Available: http://www.bmdw.org/index.php?id=downloads&no_cache=1. [Accessed 24 02 2011].

[7] W. Bochtler, M. Maiers, J. Bakker, M. Oudshoorn, S. Marsh, D. Baier, C. Hurley, C. Muller and I. T. W. G. World Marrow Donor Association, "World Marrow Donor Association framework for the implementation of HLA matching programs in hematopoietic stem cell donor registries and cord blood banks," *Bone Marrow Transplantation,* vol. 46, p. 338–343, 2011.

[8] H. Eberhard, "Validation of the Predictions of OptiMatch®," in *IDRC*, Bern, 2008.

[9] N. H. G. R. Institute, "Glossary of Gentic Terms - Allele," [Online]. Available: http://www.genome.gov/glossary/index.cfm?id=4. [Accessed 20 09 2012].

[10] L. Excoffier and M. Slatkin, "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population," *Mol Biol Evol,* vol. 12, p. 921–927, 1995.

[11] Wikipedia, "Linkage disequilibrium," [Online]. Available: http://en.wikipedia.org/wiki/Linkage_disequilibrium. [Accessed 14 02 2013].

[12] Wikipedia, "Major histocompatibility complex," [Online]. Available: http://en.wikipedia.org/wiki/Major_histocompatibility_complex. [Accessed 01 05 2007].

[13] Wikipedia, "HLA complex," [Online]. Available: http://upload.wikimedia.org/wikipedia/en/4/4f/HLA_complex1.JPG. [Accessed 01 05 2007].

[14] S. G. E. Marsh, E. D. Albert, W. F. Bodmer, R. E. Bontrop, B. Dupont, H. A. Erlich, M. Fernandez-Vina and others, "Nomenclature for factors of the HLA system, 2010," *Tissue Antigens,* vol. 75, p. 291–455, 2010.

[15] W. Bochtler, M. Maiers, M. Oudshoorn, S. Marsh, C. Raffoux, C. Mueller and C. Hurley, "World Marrow Donor Association guidelines for use of HLA nomenclature and its validation in the data exchange among hematopoetic stem cell donor registries and cord blood banks," *Bone Marrow Transplantation,* vol. 39, p. 737–741, 2007.

[16] S. G. Marsh, "Nomenclature of HLA alleles," [Online]. Available: http://hla.alleles.org/wmda/index.html. [Accessed 27 08 2009].

[17] BMDW, "DNA Reference Tables," [Online]. Available: http://www.bmdw.org/index.php?id=downloads. [Accessed 01 05 2007].

[18] C. Hurley, "The NMDP Matching Guidelines: What's New," in *NMDP Council Meeting*, Minneapolis, 2011.

[19] R. Bray, C. Hurley, N. Kamani, A. Woolfrey, C. Müller, S. Spellman, M. Setterholm and D. Confer, "National marrow donor program HLA matching guidelines for unrelated adult donor hematopoietic cell transplants," *Biol Blood Marrow Transplant,* vol. 14, p. 45–53, 2008.

[20] M. Prestegaard, "Unrelated Hematopoietic Stem Cell Donor Search and Facilitation Information Systems Principles," 2012.

[21] A. Timm and H. Eberhard, "The semantics of EMDIS messages Version 1.28," 2011.

[22] NMDP, "HLA Resources - Allele Code Lists," [Online]. Available: http://bioinformatics.nmdp.org/HLA/Allele_Codes/Allele_Code_Lists/Allele_Code_Lists.aspx. [Accessed 23 07 2012].

[23] C. K. Hurley, M. Maiers, S. G. E. Marsh and M. Oudshoorn, "Overview of registries, HLA typing and diversity, and search algorithms," *Tissue Antigens,* vol. 69, p. Suppl. 1 (3–5), 2007.

[24] C. Hurley and e. al, "Hematopoietic stem cell donor registry strategies for assigning search determinants and matching relationships," *Bone Marrow Transplantation,* vol. 33, p. 443–450, 2004.

[25] ZKRD, "IMGT/HLA Database - Search Strategies - ZKRD," [Online]. Available: http://www.ebi.ac.uk/imgt/hla/searchdet_zkrd.html. [Accessed 23 07 2012].

[26] NMDP, "IMGT/HLA Database - Search Strategies - NMDP," [Online]. Available: http://www.ebi.ac.uk/imgt/hla/searchdet_nmdp.html. [Accessed 23 07 2012].

[27] C. Malmberg, "Search Strategy and HapLogic: Case Studies," NMDP, [Online]. Available: http://marrow.org/News/Events/Council_Meeting/HapLogic_III_Search_Strategy_With_Answ

ers.aspx. [Accessed 23 07 2012].

[28] D. Steiner, M. Korhonen, M. Kurikova, M. Kusikova, M. Sankowska, B. Svensson, E. Du Toit, N. Shriki, K. Peyerl and C. Raffoux, "PROMETHEUS PROBABILITY MATCHING: COMMUNITY TECHNOLOGY PREVIEW," in *9th International Donor Registry Conference (IDRC)*, Sydney, 2012.

[29] J. Pingel, J. Hofmann, D. Baier, U. Solloch, A. Grathwohl, U. Ehninger, A. Schmidt and G. Ehninger, "Hap-E Search: Haplotype-Enhanced Search," in *9th International Donor Registry Conference (IDRC)*, Sydney, 2012.

[30] P. Gourraud, M. Balère, A. Dormoy, P. Loiseau, E. Marry and F. Garnier, "COMPUTER ASSISTED SEARCH FOR UNRELATED DONORS USING THE EASYMATCH - TOOL AT FRANCE GREFFE DE MOELLE," in *16th International HLA and Immunogenetics Workshop and Joint Conference*, Liverpool, 2012.

[31] M. Maiers, J. Bakker, W. Bochtler, H. Eberhard, S. Marsh, C. Muller, H. Rist and I. T. W. G. World Marrow Donor Association, "Information technology and the role of WMDA in promoting standards for international exchange of hematopoietic stem cell donors and products," *Bone Marrow Transplantation,* vol. 45, p. 839–842, 2010.

[32] Wikipedia, "Test automation," [Online]. Available: http://en.wikipedia.org/wiki/Test_automation. [Accessed 23 07 2012].

[33] A. Clark, "Inference of haplotypes from PCR-amplified samples of diploid populations," *Mol. Biol. Evol.,* vol. 7, no. 2, p. 111–122, 1990.

[34] R. F. Schipper, J. D'Amaro, P. de Lange, G. M. Th. Schreuder, J. J. van Rood and M. Oudshoorn, "Validation of Haplotype Frequency Estimation Methods," *Human Immunology,* vol. 59, pp. 518-523, 1998.

[35] M. E. Hawley and K. K. Kidd, "HAPLO: A Program Using the EM Algorithm to Estimate the Frequencies of Multi-site Haplotypes," *The Journal of Heredity,* vol. 86, no. 5, pp. 409-411, 1995.

[36] R. Schipper, J. D'Amaro and M. Oudshoorn, "The probability of finding a suitable related donor for bone marrow transplantation in extended families," *Blood,* vol. 87, pp. 800-804, 1996.

[37] M. Stephens, N. J. Smith and P. Donnelly, "A new statistical method for haplotype reconstruction from population data," *American Journal of Human Genetics,* vol. 68, pp. 978-989, 2001.

[38] H. Urban, "Bayesian haplotype frequency prediction," Diploma Thesis, FEL ČVUT, Prague, 2011.

[39] Wikipedia, "Hardy-Weinberg principle," [Online]. Available: http://en.wikipedia.org/wiki/Hardy-Weinberg_principle. [Accessed 26 08 2009].

[40] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc.,* vol. 39, no. 1, p. 1–38, 1977.

[41] J. C. Long, R. C. Williams and M. Urbanek, "An E-M algorithm and testing strategy for multiple-locus haplotypes," *Am. J. Hum. Genet.,* vol. 56, no. 3, p. 799–810, 1995.

[42] D. Fallin and N. Schork, "Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization algorithm for unphased diploid genotype data," *Am. J. Hum. Genet.,* vol. 67, no. 4, p. 947–959, 2000 .

[43] J. Polanska, "The EM Algorithm and its implementation for the estimation of frequencies of SNP-haplotypes," *Int. J. Appl. Math. Comput. Sci.,* vol. 13, no. 3, pp. 419-429, 2003.

[44] K. Rohde and R. Fuerst, "Haplotyping and Estimation of Haplotype Frequencies for Closely Linked Biallelic Multilocus Genetic Phenotypes Including Nuclear Family Information," *Human mutation,* vol. 17, pp. 289-295, 2001.

[45] T. Niu, Z. Qin, X. Xu and J. Liu, "Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms," *American Journal of Human Genetics,* vol. 70, pp. 157-169, 2002.

[46] R. Schipper, M. Oudshoorn, J. D'Amaro, H. Zanden, P. Lange, J. Bakker, J. Bakker and J. Rood, "Validation of large data sets, an essential prerequisite for data analysis: An analytical survey of the Bone Marrow Donors Worldwide," *Tissue Antigens,* vol. 47, pp. 169-178, 1996.

[47] A. N. R. Institute, "HLA Nomenclature - statistics," [Online]. Available: http://hla.alleles.org/nomenclature/stats.html. [Accessed 26 08 2009].

[48] C. Kollman, M. Maiers, L. Gragert, C. Muller, M. Setterholm, M. Oudshoorn and C. Hurley, "Estimation of HLA-A, -B, -DRB1 Haplotype Frequencies Using Mixed Resolution Data from a National Registry with Selective Retyping of Volunteers," *Human Immunology ,* vol. 68, pp. 950-958, 2007.

[49] P. Gourraud, E. Génin and A. Cambon-Thomsen, "Handling missing values in population data: consequences for maximum likelihood estimation of haplotype frequencies," *European Journal of Human Genetics,* vol. 12, pp. 805-812, 2004.

[50] U. Feldmann and C. Muller, "Methods to correct the HLA selection bias in haplotype frequency estimations using donor registry data," *Human Immunology,* vol. 64, p. S134, 2003.

[51] C. Muller, G. Ehninger and S. Goldmann, "Gene and Haplotype Frequencies for the Loci HLA-A, HLA-B and HLA-DR Based on Over 13,000 German Blood Donors," *Human Immunology,* vol. 64, p. 137–151, 2003.

[52] A. N. R. Institute, "HLA Nomenclature in WMDA format," [Online]. Available: http://hla.alleles.org/wmda/index.html. [Accessed 26 08 2009].

[53] W. Bochtler, M. Beth, S. Marsh and C. Muller, "HLACORE – A comprehensive low level HLA library for programmers," *Human Immunology,* vol. 66, p. 49, 2005.

[54] M. Maiers, L. Gragert, A. Madbouly, D. Steiner, S. Marsh, P. Gourraud, M. Oudshoorn, H. van der Zanden, A. Schmidt, J. Pingel, J. Hofmann, C. Müller and H. Eberhard, "16(th) IHIW: Global analysis of registry HLA haplotypes from 20 Million individuals: Report from the IHIW Registry Diversity Group.," *Int J Immunogenet.,* p. doi: 10.1111/iji.12031, 2012.

[55] J. Hofmann, J. Pingel, U. Ehninger, D. Baier, A. Grathwohl, A. Stahr, A. H. Schmidt and G. Ehninger, "How to plant a haplotype tree: speeding up haplotype frequency-based searches, P33," *Tissue Antigens,* vol. 77, p. 409, 2011.

[56] P. Cano, W. Klitz, S. J. Mack, M. Maiers, S. Marsh, H. Noreen, E. Reed, D. Senitzer, M. Setterholm, A. Smith and M. Fernández-Viña, "Common and Well-Documented HLA Alleles: Report of the Ad-Hoc Committee of the American Society for Histocompatiblity and Immunogenetics," *Human Immunology,* vol. 68, p. 392–417, 2007.

[57] L. Gragert and M. Maiers, "A greedy algorithm for generating abridged," in *Abstracts for the 20th Annual BSHI Conference, Leeds, UK*, 2009.

[58] J. H. Zhao, D. Curtis and P. C. Sham, "Model-Free Analysis and ermutation Tests for Allelic Associations," *Hum Hered,* vol. 50, p. 133–139, 2000.

[59] R. F. Schipper, J. D'Amaro, J. T. Bakker, J. Bakker, J. J. van Rood and M. Oudshoorn, "HLA Gene and Haplotype Frequencies in Bone Marrow Donors Worldwide Registries," *Human Immunology,* vol. 52, pp. 54-71, 1997.

[60] C. Muller, "Optimized strategies for prospective DRB1-typing for unrelated stem cell donor registries," *Human Immunology,* vol. 64, p. S134, 2003.

[61] C. Muller, U. Feldmann, H. Eberhard, D. Baier and A. Schmidt, "HLA-A-B-DRB1*-haplotype frequencies of the German population," *Human Immunology,* vol. 66, p. 51, 2005.

[62] M. Maiers, L. Gragert and W. Klitz, "High-resolution HLA alleles and haplotypes in the United States population," *Human Immunology,* vol. 68, pp. 779-788, 2007.

[63] H.-P. Eberhard, "Schätzung von hochaufgelösten HLA-Haplotypfrequenzen deutscher Blutstammzellspender und ihre Anwendung bei der Patientenversorgung," Institut für Transfusionsmedizin, Universität Ulm, 2010.

[64] H. Eberhard, U. Feldmann, W. Bochtler, D. Baier, C. Rutt, A. Schmidt and C. Müller, "Estimating unbiased haplotype frequencies from stem cell donor samples typed at heterogeneous resolutions: a practical study based on over 1 million German donors.," *Tissue Antigens,* vol. 76, pp. 352-361, 2010.

[65] A. Schmidt, U. Solloch, J. Pingel, D. Baier, I. Böhme, K. Dubicka, S. Schumacher, C. Rutt, A. Skotnicki, J. Wachowiak and G. Ehninger, "High-resolution human leukocyte antigen allele and

haplotype frequencies of the Polish population based on 20,653 stem cell donors," *Human immunology,* vol. 72, pp. 558-865, 2011.

[66] H. P. Eberhard, M. L. Balere, P. A. Gouraud, L. Gragert, H. Maldonado-Torres, D. Steiner, H. van der Zanden, M. Maiers, S. Marsh and C. Mueller, "Comparing different programs for HLA haplotype frequency estimation in a controlled data environment," *Tissue Antigens,* vol. 72, no. 3, pp. 263-264, 2008.

[67] D. Steiner, T. Schlaphoff, V. Borrill, E. Du Toit and C. Raffoux, "HLA study in Nigeria," *Tissue Antigens,* vol. 77, pp. 399-400, 2011.

[68] J. Čejka, "Analýza dat Českého registru dárců krvetvorných buněk," Bachelor Thesis, FEL ČVUT, Prague, 2011.

[69] A. Schmidt, D. Baier, U. Solloch, A. Stahr, N. Cereb, R. Wassmuth, G. Ehninger and C. Rutt, "Estimation of high-resolution HLA-A, -B, -C, -DRB1 allele and haplotype frequencies based on 8862 German stem cell donors and implications for strategic donor registry planning.," *Human immunology,* vol. 70, no. 11, pp. 895-902, 2009.

[70] A. Schmidt, U. Solloch, D. Baier, A. Stahr, R. Wassmuth, G. Ehninger and C. Rutt, "Regional differences in HLA antigen and haplotype frequency distributions in Germany and their relevance to the optimization of hematopoietic stem cell donor recruitment.," *Tissue Antigens,* vol. 76, no. 5, pp. 362-379, 2010.

[71] J. Pingel, U. Solloch, J. Hofmann, V. Lange, G. Ehninger and A. Schmidt, "High-resolution HLA haplotype frequencies of stem cell donors in Germany with foreign parentage: How can they be used to improve unrelated donor searches?," *Hum Immunol.,* pp. pii: S0198-8859(12)00607-6. doi: 10.1016/j.humimm.2012.10.029, 2012.

[72] A. H. Schmidt, D. Stahr, D. Baier, S. Schumacher, G. Ehninger and C. Rutt, "Selective recruitment of stem cell donors with rare human leukocyte antigen phenotypes," *Bone Marrow Transplantation,* vol. 40, pp. 823-830, 2007.

[73] L. Kábrt, "HLA genetická příbuznost Čechů s ostatními národy," Diploma Thesis, FEL ČVUT, Prague, 2009.

[74] S. Mack, B. Tu, A. Lazaro, R. Yang, A. Lancaster, K. Cao, J. Ng and C. Hurley, "HLA-A, -B, -C, and -DRB1 allele and haplotype frequencies distinguish Eastern European Americans from the general European American population," *Tissue Antigens,* vol. 73, no. 1, pp. 17-32, 2009.

[75] P. A. Gourraud, P. Lamiraux, N. El-Kadhi, C. Raffoux and A. Cambon-Thomsen, "Inferred HLA Haplotype Information for Donors From Hematopoietic Stem Cells Donor Registries," *Human Immunology,* vol. 66, pp. 563-570, 2005.

[76] P. G. Beatty, K. M. Boucher, M. Mori and E. L. Milford, "Probability of Finding HLA-mismatched Related or Unrelated Marrow or Cord Blood Donors," *Human Immunology,* vol. 61, pp. 834-

840, 2000.

[77] J. Těhník, "Geographical Analysis of Databases of Stem Cell Donor Registries," Diploma Thesis, FEL ČVUT, Prague, 2011.

[78] J. Brož, "HLA Explorer," Diploma Thesis, FEL ČVUT, 2008, 2008.

[79] J. Hofmann, U. V. Solloch, J. Pingel, D. M. Baier, A. Grathwohl, U. Ehninger, S. F. Winter, A. H. Schmidt and G. Ehninger, "Hap-E Search: A Haplotype-Enhanced Search Algorithm for Directed Quality Programs in Donor Centers," in *9th International Donor Registry Conference (IDRC)*, Sydney, 2012.

[80] J. Dehn, "HapLogic III," NMDP, [Online]. Available: http://marrow.org/News/Events/Council_Meeting/2011_Presentations/A3_B3_Putting_HapLogic_to_Work_for_You.aspx.

[81] W. Bochtler, H. Eberhard, M. Beth and C. Mueller, "Optimatch® – Optimized Selection of Allele Matched Unrelated Donors from a Large Database," *Biology of Blood and Marrow Transplantation,* vol. 14, p. 52, 2008.

[82] D. Confer and P. Robinett, "The US National Marrow Donor Program role in unrelated donor hematopoietic cell transplantation," *Bone Marrow Transplantation,* vol. 42, pp. S3-S5, 2008.

[83] C. Hurley, J. Wagner, M. Setterholm and D. Confer, "Advances in HLA: Practical Implications for Selecting Adult Donors and Cord Blood Units," *Biology of Blood and Marrow Transplantation,* vol. 12, pp. 28-33, 2006.

[84] M. Maiers and M. Setterholm, "Haplogic: The NMDP registry match algorithm," *Human Immunology,* vol. 67, p. S127, 2006.

[85] A. Green and D. Steiner, "Information Technology (IT) and Information Management in an unrelated stem cell donor registry," in *Donor Registries Working Group (DRWG) - Registry Handbook*, Leiden, WMDA, 2013.

[86] C. Muller, W. Hontscha, S. Cleaver, L. Canham, A. Baouz, C. Raffoux and editors, "EMDIS European Marrow Donor Information System. A model for the communication between heterogeneous databases.," *RIAO 1994 Conference Proceedings. Paris: C.I.D.-C.A.S.I.S,* p. 132 – 134, 1994.

[87] A. Baouz and C. Raffoux, "EMDIS: European Marrow Donor Information System," *Comput Methods Programs Biomed,* vol. 45, pp. 45-6, 1994.

[88] D. Steiner, "European Marrow Donor Information System: Concept and Praxis," *Transplantation Proceedings,* vol. 42, no. 8, pp. 3255-3257, 2010.

[89] C. Muller, "Computer applications in the search for unrelated stem cell donors," *Transplant*

Immunology, vol. 10, p. 227–240, 2002.

Appendix A: Used datasets

ID	Description	Number of haplotypes	Haplotype Rank of the median haplotype
[FI-2011]	HLA Haplotype Frequencies of the Finnish registry, calculated in 2011 by David Steiner	>= 10^{-4}: 442 >= 10^{-5}: 3093	**20**
[CZ-2011]	HLA Haplotype Frequencies of the Czech population, calculated in 2011 by David Steiner	>= 10^{-4}: 1236 >= 10^{-5}: 3746	**93**
[CZ-2012]	HLA Haplotype Frequencies of the Czech population, calculated in 2012 by David Steiner	>= 10^{-4}: 1476	**96**
[NMDP-EUR-2007]	HLA Haplotype Frequencies of the NMDP registry [62], Caucasian population, calculated in 2007	3380	**102**
[ZKRD-2008]	HLA Haplotype Frequencies of the ZKRD registry, calculated in 2008	7686	**154**
[HPE-2010]	HLA Haplotype Frequencies of the ZKRD registry [63], calculated in 2010	24449	**158**
[D-1205]	HLA Haplotype Frequencies of the ZKRD registry (May 2012), calculated by David Steiner	33102	**216**

Table 32: HFE datasets and their identification used in the experiments of the work

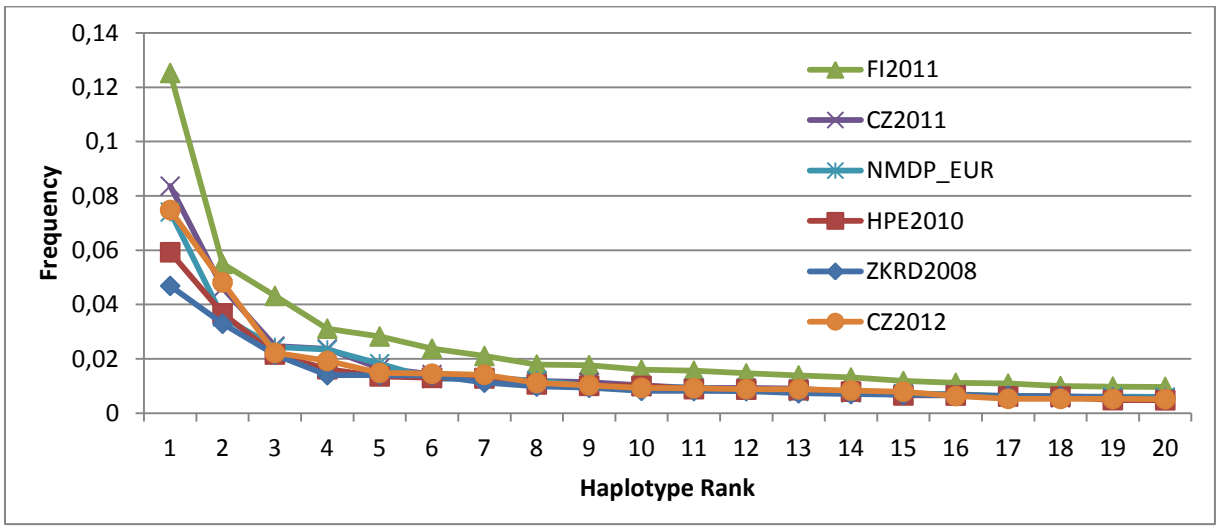


Figure 56: HFE datasets used in the experiments of the work, frequencies of top 20 haplotypes

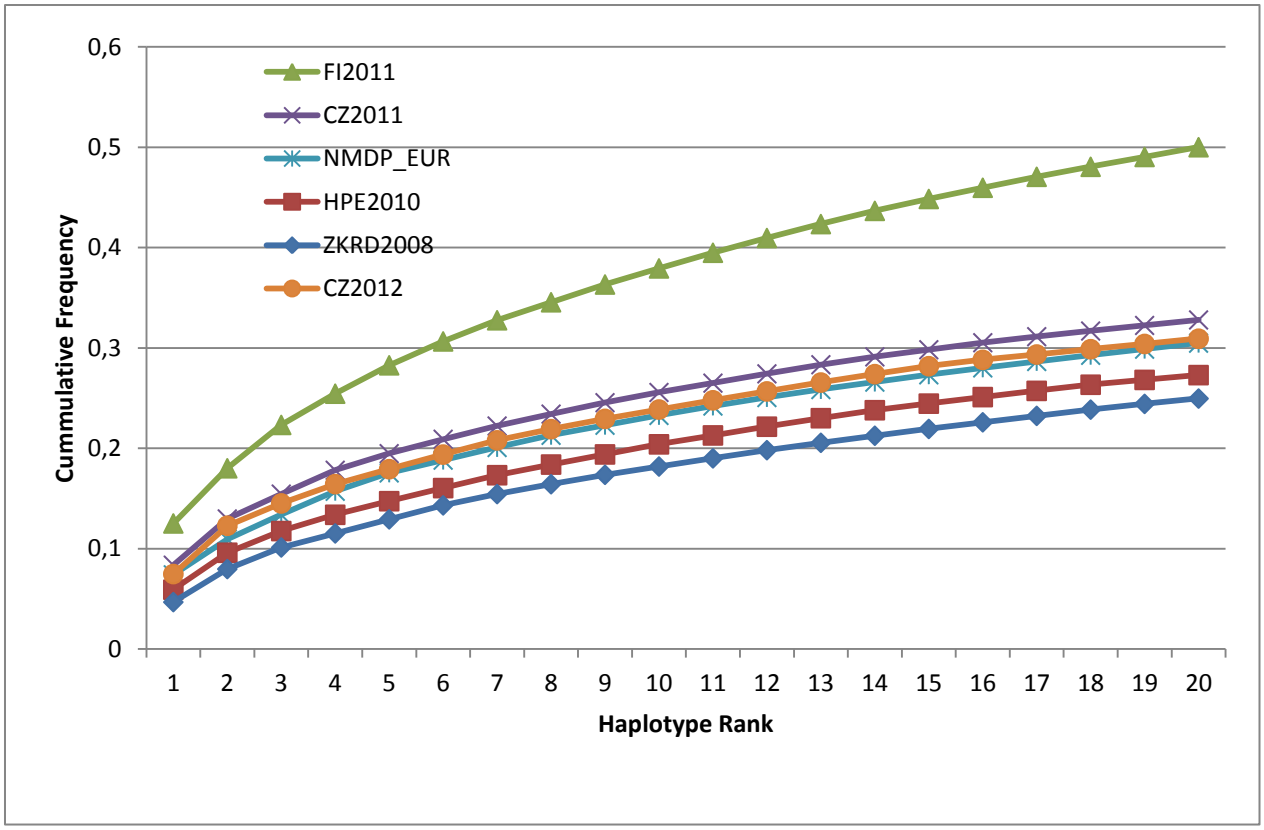


Figure 57: HFE datasets used in the experiments of the work, cumulative frequency of top 20 haplotypes

Figure 57 demonstrates heterogeneity of datasets. This corresponds with the statistics "Haplotype Rank of the median haplotype" (see Table 32). Due to small registry size and sampling error, we expect Czech population is more heterogeneous than we can currently see.

Other referred datasets

[BMDW-2011] … BMDW Database [6], February 2011

[BMDW-201205] … BMDW Database [6], March 2012

[PROM-CT] … EMDIS Verification Typing requests and their results from selected registries running Prometheus software (see Chapter 10.2)

Appendix B: Stem cell donor registry software specification

Text has been taken from [85].

WMDA standards require that 'all patient and donor communications and records must be stored to ensure confidentiality and to allow for traceability of the donors and steps of the donation process' (WMDA 5.01.2).

This section describes architecture, data and functional requirements of the registry IT system.

It is essential that the registry analyses the following:

- what are the key modules and functions of the system
- what information and how it should be stored on the database
- what are the business processes of the registry and how should they be supported by the system
- who are the end users of the system, what are their roles in the system
- what are the interactions of the system with the outside world, what interfaces should be built

The **architecture** of the system follows the stem cell donor registry organisation. There are several aspects:

- **Situation**: The registry might be completely independent, located in the administrative building or be to part of a hospital, blood transfusion institute or other medical organisation. If the registry belongs to the bigger medical organisation, it has to follow specified rules and usually has to be well integrated. Very often, small registries are organisationally connected with the HLA laboratory, which necessitates the interface between systems of these departments.
- **Donor centres**: The registry may be the national HUB that does not recruit donors directly, but cooperates with the network donor centres and cord blood banks. In some countries, the registry does not have access to donor contact details, so the donors may not be contacted directly. In this case this 'master record' of the donor is in the donor centre and the registry only has a copy. Other settings, typical for small registries, are based on integrated registry with the donor centre. Donor recruitment is organised by the registry itself or a network of partner organisations that, after recruitment, transfers all donor data to the registry database, so the master record of the donor is managed by the registry itself.
- **Access** to the registry database is usually restricted registry staff. Partner institutions must contact the registry in order to access the database or changes will be visible after next off-line upload of partner institution data (e.g. cord blood bank). The alternative option is to build an on-line interface or allow partner institutions to access the database directly, for example, donor centres and cord blood banks can manage their donor, CBU records and transactions directly. The registry may look like a single institution for the international community (EMDIS, international registries), but is actually a network of donor, transplant and search coordination centres that are spread across the whole country.

The list of key functional requirements that a registry may consider to include, when considering new or improving existing registry system:

- **Donor database** is the key module. Donor record must include:

    - **Donor identification**: a unique, invariant registry ID is the primary reference, but a data set can also include social security number, donor centre ID, recruitment ID, cord blood bank ID, ID of the mother of the CBU, ISBT128 donation code, stored sample ID, stored DNA ID, EMDIS ID, among other fields.
    - **HLA data**: separate fields for serology and DNA typing results [15], typing laboratory, date of typing, primary typing data, NIMA, etc. The registry should consider of how HLA data are imported into the database as this may be from either an internal or an external source. Reference to the white paper [15] has to be made regarding standardisation of nomenclature and data formatting.
    - **Demographics**: name, title, gender, date of birth, ethnic group, insurance company, etc.
    - **Relationships**: family or personal relations to other donors or patients, used for family reports of the patient
    - **Recruitment**: donor centre, date of recruitment, recruitment method (website, patient-draft, blood donor, etc.), blood donor flag, platelets donor flag, etc.
    - **Donor status**: reservation of the donor, temporary or permanent withdrawal, reason of withdrawal (age, medical, personal, etc.)
    - **Contact details**: permanent, temporarily and work address, email, phones, social media networks, communication language, preferred contact, history of communication with the donor, etc.
    - **Medical questionnaire**: weight, height, blood group, kell, haemoglobin, number of pregnancies, number of blood transfusions, donor consent to different types of donations, diseases in the past, etc.
    - **Infectious disease markers**: CMV status, Toxoplasmosis, EBV status, HIV status, HIV p24 antigen, antibodies to HIV, hepatitis B and C status and antibodies, Lues status, ALT status, etc. with dates of tests and laboratories that performed tests.
    - **Products**: information about the stored donor samples or cord blood unit product, its position in the freezer, etc.
    - **Cord blood unit data**: volume of CBU, nucleated cells, CD34+ cells, mononucleated cells, white blood cells, processing methods, fractions, mother tests, etc.
    - **Harvests**: date and place of harvest, date and place of transplant, patient ID, source of stem cells (bone marrow, PBSC, DLI, cord blood, other)
    - **Audit**: who and when has inserted or modified the donor record, search-able history of changes of the donor record (who, when, old data, new data).
- **Patient database** functions include:
    - Need of the record for both national and international patients
    - **Patient identification**: unique, invariant registry ID, but can also include social security number, transplant centre ID, hospital record ID, EMDIS ID, physician, etc.
    - **HLA data**: separate fields for serology and DNA typing results [15], typing laboratory, date of typing, primary typing data, etc.
    - **Demographics**: name, title, gender, date of birth, ethnic group, insurance company, etc.
    - **Relationships**: family or personal relations to donors, used for family reports of the patient
    - **Patient status**: donor search status, transplant status, closure of the case (date, reason)
    - **Medical information**: diagnosis, disease phase, weight, blood group, CMV status, etc.
    - **Transplants**: date and place of harvest, date and place of transplant, donor ID, source of stem cells (bone marrow, PBSC, DLI, cord blood, other), etc.
    - **Audit**: who and when has inserted or modified the patient record, search-able history of changes of the patient record (who, when, old data, new data).
- Both donor and patient database must be searchable by different attributes.

- **Quality control**: the system should control quality of data according to registry policies. There should be no expired reservations of donors, no over aged donors that are 'available for transplant purposes' on the searches, no donors missing critical data (e.g. date of birth, gender), HLA data should be always valid according to the latest HLA nomenclature (renamed or deleted alleles should be corrected), etc.
- Regular update of reference tables of **HLA nomenclature** [16] and multiple-allele-codes [22].
- **Reports**: customizable reports of donor and patient details, export to PDF files, letters and emails to donors by user-defined templates.
- **WMDA annual report**: Many registries do not systematically collect data for the WMDA annual report; leading to time spent searching paper records/excel spreadsheets when preparing the WMDA questionnaire. There is a huge advantage to building in the functionality to generate this data automatically at the start of the project. This also increases the reliability of data reported to WMDA.
- **Donor searches**: The donor search algorithm is the key and probably most difficult element of the stem cell donor registry software. It should follow WMDA recommendations and guidelines. For more information about the search algorithm see the section 'Search Algorithm'.
- **Management of requests**: the system must allow users to create and track different national and international requests for donors. This includes typing requests, VT sample requests, IDM requests, donor reservation requests and workup requests. Traceability of requests means clear information about the status of the request (result, inability to do the service, cancellation, denial) and related events (acknowledgement by the partner, contact of the donor, reminders, invoice).
- The system should support the **work-flow management** of requests for different scenarios (e.g. unsuccessful CT collection, cancelled workup). Each step in the search process (e.g. patient registration and any request, result or update) shall be documented with all relevant attributes and a time stamp (WMDA 5.04.3). Management of requests includes both:
    - National requests - national patient and national donor
    - International requests - national patient and international donor or vice versa; electronic on-line requests (EMDIS or web interface) and fax requests (outside EMDIS)
- **Financial module** can be integrated into the request management work-flow. Closed requests are usually invoiced to the requesting institution. Integration with external economical software system requires synchronisation of services (invoice items) and clients (invoice recipients).
- **Transplant records, donor and patient follow-up records** with automated reminders of incomplete or missing records.
- **Document management system**: possibility to store and maintain different kinds of electronic documents, linked to donor, patient, search and other types of records.
- **International interfaces**: the registry should be well integrated to the international community, mainly due to efficient donor searches:
    - **BMDW**: regular export of donor and CBU database to Bone Marrow Donors Worldwide (www.bmdw.org)
    - **EMDIS, EMDIScord**: on-line peer-to-peer network of stem cell donor registries (www.emdis.net). You will find more information about the EMDIS system bellow in this chapter.
    - **NMDP**: some international registries are listed as donor centres in the NMDP network, so they regularly export data to NMDP database (www.nmdp.org).
    - **Netcord**: member cord blood banks of this organisation regularly export data to the central database (www.netcord.org/).
    - **HLA:** regular import of the current HLA nomenclature (http://hla.alleles.org/wmda/index.html, NMDP allele code nomenclature)

- **National interfaces**: the registry serves as the national HUB that connects different institutions and individuals within the country. Following on-line interfaces might be useful:
    - **HLA laboratory**: registry sends electronic typing requests for its donors and patients to the laboratory and HLA typing results are returned to the registry. The registry can also access information about donor samples stored in the HLA laboratory freezers, so registry coordinators know if they can use this stored DNA sample for the additional HLA typing.
    - **Donor centres**: donor centres and cord blood banks in the registry network may have their own IT systems that should be interfaced to the registry system.
    - **Harvesting centre**: once the patient-donor pair is identified, the registry may send donor record to the harvesting centre system and get back details about the stem cell product.
    - **Search units:** search units in the registry network may have their own IT systems that should be interfaced to the registry system.
    - **Transplant centres**: transplant centres and hospitals need to communicate with the registry. An on-line solution instead of fax / paper / phone solution is desirable.
    - **Donors**: On-line web portal helps to keep the contact with donors. Such portal can include contact details change form, on-line forum, news from the registry, reimbursement form, etc. Some registries also use social media networks such as Facebook or Twitter.
    - **Sponsors**: On-line web portal for registry sponsors may increase their motivation. The system can manage sponsor accounts and show statistics how many donors were recruited for the sponsorship, how many of them were requested for VTs, workups, etc.

TIP: It may seem that a registry system stores and manages the HLA typing results in the same format as the HLA laboratory information management system (LIMS), and some registries have implemented such data storage.

It is a mistake to use these in search algorithms. The main differences between registry database and HLA LIMS database are:

- The registry system needs fast access to the most current and comprehensive HLA typing results, which does not always mean the last test typing. This may be combination of multiple tests performed in the past by multiple typing techniques. The registry system always needs access to the full set of all loci that should be stored at one place, while the HLA lab system order includes only requested tests and loci, so HLA typing results of an individual may be spread in multiple typing orders.
- When the HLA lab supervisor approves the order results, it cannot be changed in the lab system. However, the registry system has to keep historical HLA typing results up-to-date according to the latest HLA nomenclature, so it needs to update them (deleted and renamed alleles, new HLA nomenclature).

Appendix C: Inter-Registry Communication System (EMDIS)

Text has been taken and adapted from [85].

Reliable communications and data transfer of donor and patient records between all partners in this huge network is one of the most important success factors in stem cell transplantation.

The internet gives us great opportunities in registry to registry connections, including the software support of the whole process - from the preliminary search request to transplantation.

EMDIS (European Marrow Donor Information System) [31] [86] [87] [88] is an open computer network for data exchange among different unrelated hematopoietic stem cell donor registries. Today, it covers around 75% of all potential unrelated stem cell donors and cord blood units registered in BMDW (www.bmdw.org) and became the de-facto standard communication system for unrelated HSCT registries worldwide. The EMDIS community provides documentation, status information, software tools, support and a project management platform [31] (www.emdis.net).

C.1 Technical background

The decrypted content of an EMDIS message is a text in special format, called the Flexible Message Language (FML). EMDIS emails are not read by humans, but computer systems that parse the FML text into elemental attributes and data fields that are further processed.

On the basis of this technical background about 30 message types are defined, including preliminary requests and patient updates, search results, typing requests and results, sample requests, notification of sample arrival date and sample testing results, IDM (Infectious disease markers) requests and results, donor reservation requests and results, workup requests and results, etc.

The most advanced feature in EMDIS is the donor search process. When a national registry initiates an international donor search for a specific patient, its data is broadcasted to other EMDIS registries. Every recipient (i.e. computer system) makes a donor search in the local database using its own algorithm and technology and replies with a set of potential donors. Then the requesting registry composes these partial results into one global EMDIS search result. In praxis, these results are received within several hours.

After this procedure, the patient is in the “Preliminary status” and no further action is taken. But the local registry can change this status to “Active” by broadcasting the Patient status change message to other registries. The preliminary search result could be outdated after a few days. If the patient is in the Active status, every remote registry runs a regular repeat search process for this patient and checks if the search result has changed. The differential update is sent back to the patient’s registry. It could contain new and better donors than previously reported or other changes in the current search result.

Finally, when the patient case is closed, the national registry broadcasts the Patient status change message with the new status “Stopped” and the repeat search process of this patient is ended.

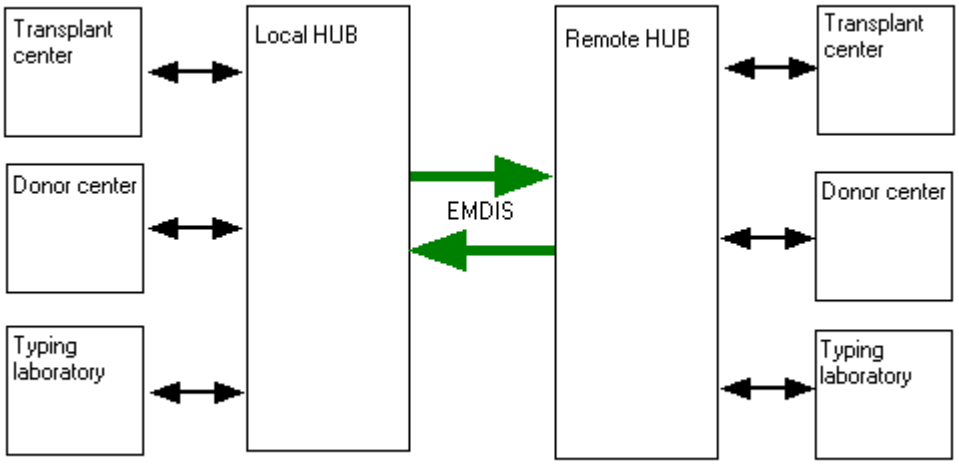


Figure 58: EMDIS communication. HUB is a national stem cell donor registry.

C.2 Software Implementation

The basic components of the EMDIS software include:

1. An email system to send and receive messages
2. Software based on ECS (EMDIS Communication System) rules to control the sending and receipt of messages
3. Software to encrypt and decrypt messages
4. Software to validate the EMDIS FML message (the FML parser). FML = Flexible Mesage Language.
5. Functions to interpret process and respond to messages – EMDIS message processor.
6. Search engine to run preliminary and repeat searches
7. User interface to create and manage EMDIS messages

The first four components exist outside of the registry software and are currently available free of charge. The four form a package called ESTER (**ECS** message **T**ransfer between **E**MDIS **R**egistries) (http://www.steinersw.eu/en/ester.html), also commonly known as middleware,. ESTER uses the FML parser developed by ZKRD. ESTER runs under the Windows operating system.

A platform independent implementation of the first three components is called PerlECS, which was developed by NMDP.

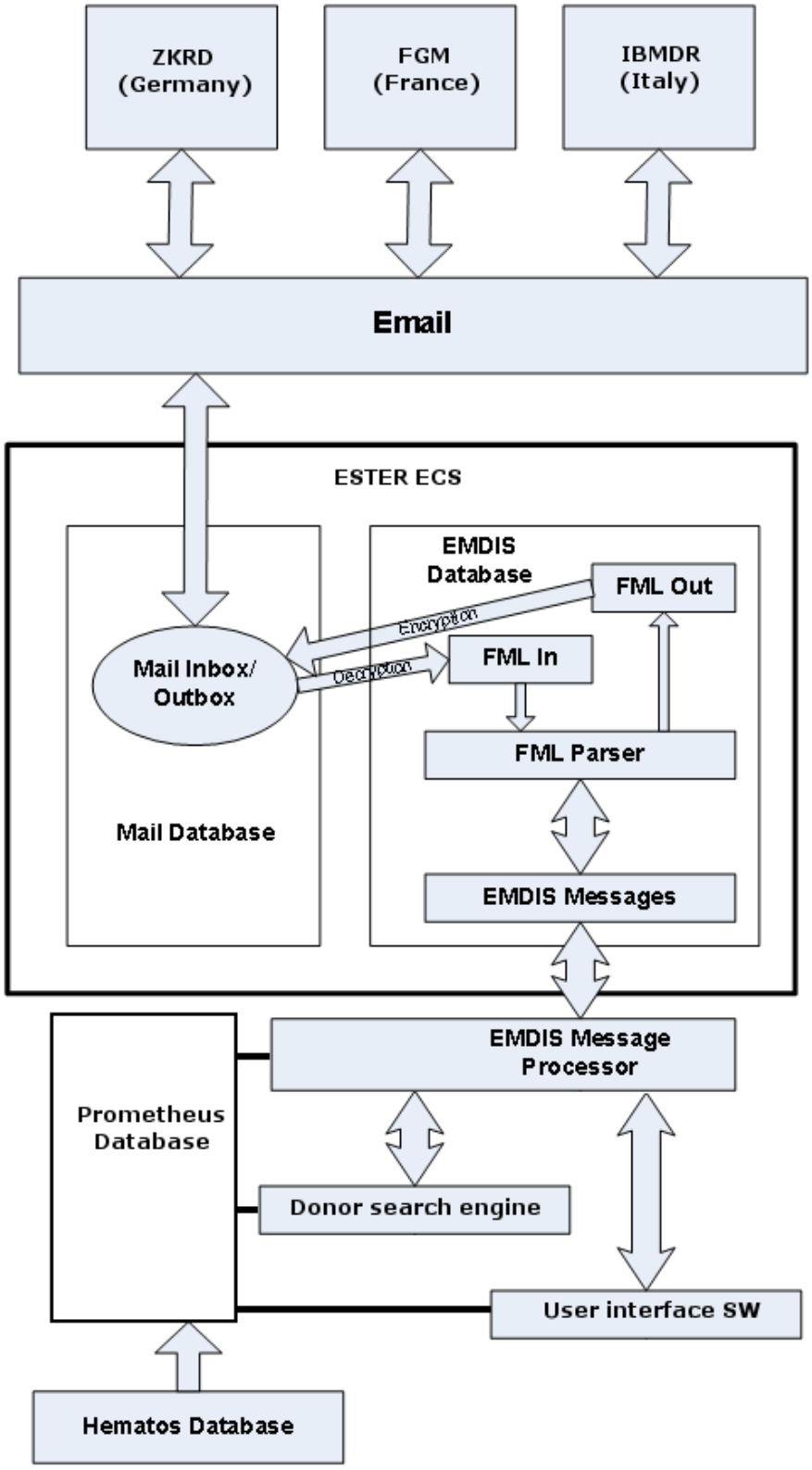


Figure 59: EMDIS Implementation of the British Bone Marrow Donor Registry.

The fifth, sixth and seventh components, the EMDIS message processor, search engine and user interface, are the most complex ones. They are available as separate piece of software, known as Prometheus, required linking ESTER to a copy of the local registry database.

EMDIS implementation can vary from one registry to another. Typically, a registry receives a search or sample request from its own national or regional transplant centre by e-mail or fax. These are then passed via EMDIS to all of the active EMDIS nodes. Responses to these requests are sent back from the external EMDIS nodes and then relayed by some other means to the originating transplant centre. This is patient-related EMDIS messaging.

If the local system implements the original idea of a 'single virtual international registry', it must maintain the same status of the patient in all EMDIS registries. And this could include the national registry itself. Then there is no difference between a local and a remote search, it is only an EMDIS search. The advantage is that the local system also notifies changes of search result as it does for foreign patients.

The registry can also receive and respond to search or sample requests from other registries directly via EMDIS. This is donor-related EMDIS messaging.

Not all registries have chosen, or are able, to respond to all of the available EMDIS messages and some registries process donor-related messages only.

C.3 EMDIS Governance

Bidirectional messaging between registries follows highly structured protocols and standard nomenclature agreed and controlled by the EMDIS community.

The EMDIS organizational structure and rules are described by EMDIS House Rules and reflect the procedures of a working party with a high level of user involvement and a focus on practical issues.

EMDIS User Group coordinates the advancement of EMDIS to achieve the goals of the network; sanctions and approves new EMDIS Users; validates and prioritises User needs; liaises with the Technical Group over specifications, time-tables and feasibility of requirements.

EMDIS Technical Group protects the integrity of the EMDIS system, technology and infrastructure; defines technical requirements for the participation in EMDIS, defines interfacing rules and prepares the necessary documentation; reviews proposals for new developments emanating from the User Group; prepares specifications and timetables for implementation by national development teams; liaises with the User Group and the national development teams of the member registries.

These groups meets regularly to discuss requests for change and to oversee the implementation of new versions of EMDIS. General maintenance, training and operational issues are also supported by the WMDA IT Working Group.

EMDIS membership is open to unrelated donor registries that actively use the EMDIS system (EMDIS hubs). Membership application has to be submitted to the chair of the EMDIS User Group for review and be approved by the EMDIS User Group.

Appendix D: Comparison of HFE programs

	CZ (Czech republic, this work)	ZKRD (Germany)	DKMS (Germany)	FGM (France)	ANT (UK)	Europdonor (Netherlands)	NMDP (USA)
Program name	Prometheus HFE	OptiHapfreq	Haplomat	Estihaplo	Cactus	Haplo3v5.exe	NA
Language(s)	Embarcadero Delphi 2007	Perl and C	Perl	C++	Perl and C	Visual Basic 6.0	perl
Platform	Windows	Linux/MAC	Windows/Linux	Windows/Linux/MAC	Linux/MAC	Windows	Windows/Linux/MAC
Max # loci/alleles	5/None	None	6/none	None	None	3/2000	none
# limit of phenotypes/individual for serology/low res	Theoretically no limit, but practically, 4.1 million individuals.	No	Yes, not used for high resolution haplotype inference	No	No	No	Only if estimating high res haplotype freqs
# limit of phenotypes/individual for high res	same	No	Yes, hardware limitations at large numbers (>1 million)	No	No	No	No
Accepted Input	Serology, Nomenclature v2, v3, NMDP allele codes	Serology, Nomenclature v2, v3, NMDP allele codes, genotypes lists	Serology, Nomenclature v3, NMDP allele codes, genotypes lists	Serology, Nomenclature v2, v3, NMDP allele codes	Nomenclature v2	Serology, Nomenclature v2, v3	Serology, Nomenclature v2, v3, NMDP allele codes, genotypes lists
Alleles abbreviated to 2 fields	yes	Optional	Yes	No	Optional	No	Optional
Alleles mapped to p-	optional	Optional	Yes	No	no	no	Optional

like groups							
Ambiguities	DNA, missing loci	Serologic, DNA, missing loci, null antigen/allele	DNA, missing loci, null antigen/allele (only in antigen level setting for DRB1)	Serologic, DNA, missing loci, null antigen/allele	Serologic, DNA, missing loci	None	Serologic, DNA, missing loci, null antigen/allele
Method to handle ambiguities	All possible genotypes are considered.	Consider all possibilities	If "p-identical" over exon 2/3 including nulls then merged to "g"-nomenclature. [69] If SBT typing ambiguities, then include all possible combinations (According to IMGT/HLA Release number in use). If intermediate resolution typing results, then include all possible combinations. Missing data: Include all possible combinations.	Consider all possible diplotypes combinations	Expanding diplotypes generated by phenotypes with missing/ambiguous typing and let EM process them	Remove from records	Consider all possibilities
EM with HWE	yes	Yes	Yes	Yes	Yes	No	Yes
Starting values	Equal, user-defined, at	Equal, user-defind, at random, based	user-defined, simulated	user-defined, at random, simulated	Equal, at random, based on allele-	Equal	Equal, user-defined, at random, based on

	random, based on allele-frequencies	on allele-frequencies	distribution	distribution	frequencies		allele-frequencies
Terminating criteria	Based on likelihood change, # iterations	Based on frequency change	Based on frequency change	Based on frequency change	Based on likelihood change, # iterations	Based on frequency change, # iterations	Based on frequency change
Terminating threshold	Specified by user: 0.00001	Specified by user: $1E^{-6}$ likelihood changed between iterations	Fixed. value used in task 1: Sum(Abs(diff(f_n-f_{n+1})))<1e-5 value used in task 2 Max(Abs(diff(f_n-f_{n+1})))<1e-6	Specified by user: $\sum\left\|\delta^{t}-\delta^{t+1}\right\|$	Specified by user: $1E^{-6}$ likelihood changed between iterations	Specified by user: 2048 iterations for tasks 1 and 2	Specified by user: $1E^{-6}$ likelihood changed between iterations
Handling low frequency haplotypes	low frequency haplotypes are excluded, threshold = 1/(2xSample_size)	Iterative tail truncation as long as LLH increases	No special handling	No skimming – output is optional	No special handling	Mathematically no problem, values of $4.3E^{-310}$ might be observed	Haplotypes with count < 0.01 are not reported
Key features	Output loci and resolution could be customized: serology broad, serology split, DNA low res, DNA high res.		High resolution haplotype frequency inference from intermediate to high resolution HLA typing results. Antigen resolution typing results are used exclusively for antigen resolution haplotype		Developed as a perl module. Main script is coded using module	EM implemented by iterating the probabilities of discretion of phenotypes	- 2-locus + 3-locus LD - Standard error - Allele frequency and 2-locus haplotype tables - HW exact test - Wn statistic

			inference only				
Output format	Floating point	Fixed point (10 digits)	Floating point	Fixed point : 6 digit	Floating point	Fixed point (14 digits), floating point	Floating point
Task 1 running time/CPU/memory	Standard PC, Windows XP, 1 processor, 4GB RAM. less than one hour		10.5h CPU 4 memory 7.9 GB	<24h	6m 45s/2.4 GHz/140MB	Typical input 8 million A, B, DR low res BMDW individuals in 12 hours running time on PC	< 1 day / 2 Intel Xeon X5690 6-core 3.47 GHz / 100 GB
Task 2 running time/CPU/memory	Same/15 hours	40 hrs	4 days CPU 12 memory 192 GB	<24h	0-20: 2h 45m/3 GHZ/850 MB 20-60: 18m 40s/3 GHZ/470 MB		Same

Table 33: Characteristics of the seven HFE computer programs.