Full speed ahead: Improving performance of Hap-E Search, a probabilistic algorithm based on haplotype frequencies

Christine Gnahm, Alexander H. Schmidt, Jan Hofmann

DKMS German Bone Marrow Donor Center, Kressbach 1, 72072 Tübingen, Germany

Introduction

We present a new search kernel for Hap-E Search, the DKMS search algorithm developed in 2012. For all potential 10/10 and 9/10 donors to a patient, it provides 10/10 and 9/10 and accordingly 9/10 and 8/10 match probabilities based on HLA-A, -B, -C, -DRB1 and -DQB1 haplotype frequencies. The new approach meets the challenges of searching more than seven million registered donors with different typing profiles in terms of typed loci and typing resolution. All donors down to a minimal resolution of serologically typed loci HLA-A and –B are included in the search.

Features

- WMDA conform mismatch counting and treatment of null-alleles (Bochtler et al. 2011,BMT,46(3),338)
  - alleles with expression-level suffices N, S and C are treated as absent, i.e. the other HLA typing result of the locus is treated as homozygous
  - HLA-mismatches of homozygous loci for both patient and donor are counted as one mismatch
- Matching probabilities for HLA-typings which cannot be represented by the used haplotype data are approximated by assuming minimal frequencies for the „unknown“ genotypes

Methods

Preparation of donor data

The algorithm gains performance by using donor information calculated beforehand:

a. Map all donors with the same HLA-typing and ethnicity to one representative.

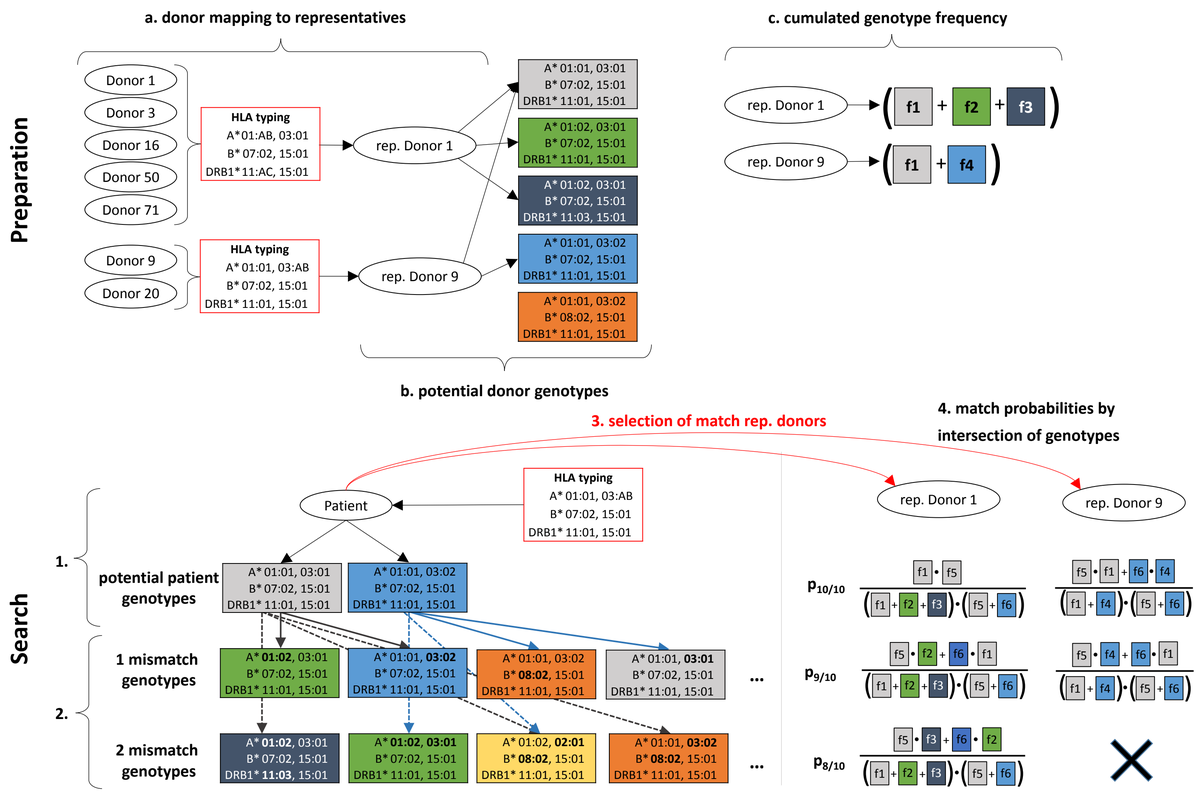
b. Determine potential donor genotypes according to HLA-typing and haplotype data.

c. Calculate cumulated frequency of all potential donor genotypes.

The information is constantly updated to the current state of the donor database.

Search process

1. Determine potential patient genotypes based on HLA-typing.
2. Determine genotypes having 1 or 2 mismatches to the patient genotypes.
3. Select potential 10/10 and 9/10 match representative donors based on HLA-typing and antigen recognition domains. This step is independent of used haplotype data in order to select also donors whose HLA typing cannot be represented by these haplotypes.
4. Calculate 10/10, 9/10 and 8/10 match probabilities for all donor-patient pairs selected in step 3 by intersecting the patient and donor genotype information.
5. Expand results for representative donors to the full donor set.



Results – Performance

Preparation of donor data

1. Existing donor base
   - Data preparation for all donors has to be performed once before launch of the new search kernel.
   - Preparation takes 1-2 days and can be further accelerated by parallelization.
2. Maintain donor information up-to-date
   - Preparation is performed for each new donor / HLA-typing.
   - Less then 9 minutes / day are required (assuming 1,000,000 new donors per year).

Preparation of donor genotypes				Relative preparation duration	
number of haplotypes	12,407		34,071	34,071/ 12,407	12,407
number of donors	5 million	1 million	1 million	1 million	5 million/1 million
preparation for existing donor database	28 h 53 min	6 h 39 min	15 h 35 min	2.3	4.3
preparation update for new donors	0.19 s	0.18 s	0.19 s	1.06	1.06
average duration per day to keep up-to-date *	8 min 41 s	8 min 13 s	8 min 41 s	1.06	1.06

* assuming 1,000,000 new donors per year (on average 2,740 per day)

Search process

a. Duration of complete search
   - ≥ 10 fold speed-up of the search process (compared to former search kernel).
   - Higher performance gain for expensive searches (> 1000 potential matches).
   - Dependence of runtime on the number of used haplotypes and searched donors is favorable.
     - ➔ Large donor sets and improved haplotype data can be used.

b. Duration of parts of the search
   - Step 1 & 2 (patient genotype preparation): independent of the number of donors
   - Step 3 (donor selection): independent of the number of haplotypes
   - Step 4 (calculation of match preparation): dependent on the underlying datasets, the number of match donors selected in step 3 and patient HLA data

Search duration in seconds Average search duration for 15 arbitrary patients with 5 loci high-resolution HLA typing. Each search was performed for two different numbers of haplotypes and donors.				Relative search duration Average relative search duration for 15 patients comparing different numbers of haplotypes and donors.	
number of haplotypes	12,407		34,071	34,071/ 12,407	12,407
number of donors	5 million	1 million	1 million	1 million	5 million/1 million
Complete search	29.2 ± 12.2	14.4 ± 4.5	22.0 ± 9.6	1.5 ± 0.4	2.0 ± 0.5
patient and MM genotypes (step 1 & 2)	1.4 ± 0.6	1.4 ± 0.6	2.2 ± 1.2	1.7 ± 0.8	1.1 ± 0.5
donor selection (step 3)	13.7 ± 5.7	9.2 ± 3.5	10.3 ± 5.4	1.1 ± 0.5	1.5 ± 0.4
calculation of probabilities (step 4)	13.7 ± 9.7	3.6 ± 2.6	9.3 ± 8.2	2.5 ± 1.1	3.9 ± 1.2

Validation

- Evaluation with the WMDA matching validation task 3 (Bochtler et al. 2016,HLA,87(6),439)
- Match probabilites for all patient-donor pairs except one where identical to the consensus data
- Deviation of one 9/10 probability by 1% (87% instead of 88%) due to rounding (P000656-D009637)
- Average duration per search: 6.8 s (1,000 patients in 10,000 donors)

Conclusion

With the new approach, performance of Hap-E Search 2.0 is tuned to allow for a more efficient workflow within DKMS quality programs, as well as to improve user friendliness of our external service Donor Navigator®.

DKMS gGmbH
Kressbach 1
72072 Tübingen, Germany
www.dkms.de

Author contact:
gnahm@dkms.de