Haplotype frequency sets in the HAP-E and ATLAS matching algorithm of the Search & Match Service

Carefully calculated haplotype frequency sets are used in conjunction with probabilistic matching algorithms. They enable the calculation of match probabilities (shown as percentages) across donor phenotypes as well as individual loci when a search is performed in WMDA's Search & Match Service.

With the last release of our the Search & Match Service, the WMDA has maintained a similar haplotype frequency set configuration as was used for Optima tch (Search & Match v1). In most cases the ION of the donor/cord blood unit will determine which haplotype frequency set is used.

To provide a more up-to-date representation of the donor pools worldwide, the haplotype frequency set calculations have been performed using the latest data in the WMDA Donor and CBU database. Calculations were performed using the open source algorithm Haplo-o-Mat (https://github.com/DKMS/Hapl-o-Mat and https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5450239/) and applying it to high resolution donor typing results received from listing organisations to extrapolate the possible haplotypes of the region or organisation. Thus, an organisation or geographical region must meet a minimum threshold of both quantity and quality HLA typed records to build usable and valuable frequencies. For cord blood units, we will use the same sets that were determined for the donor populations.

The inclusion criteria aims to balance the high number of donors, the quality of the HLA types and the complexity of the haplotype frequency estimation.

5-Locus haplotype frequency estimation was performed on all organisations/populations with at least 10 000 donors typed for HLA-A, -B, -C, -DRB1 and -DQB1 where ambiguity was low enough to be useful. Organisations in the same country that recruit from within the same population are usually combined, e.g. AT-ABMDR (ION-2614, Austrian Bone Marrow Donors) and AT-GFL (ION-4961, Verein Geben für Leben) but **not** for US-NMDP (general US population) and US-GOL (predominantly Ashkenazi Jews).

A global frequency set was also calculated and will be used for all patients and donors that, for some reason, do not match a typical haplotype frequency set. For example, when a registry is new and has not yet been assigned to an appropriate existing haplotype frequency set or when the from this registry deviate substantially enough from any typical haplotype frequency set that it does not make sense to assign them to an existing one. The number of this frequency set is 999. Information about which haplotype frequency set is applied to a donor search is provided in the search results API endpoint.

Below, you can find information on how the haplotype frequency sets are assigned to the donors and cord blood units. The search service first determines if both registry ION and ethnicity of the donor or cord blood unit provided. Registry ION is then used to look up the country/region that the donor or cord blood unit belongs to. If yes, then it will check if an ethnicity specific set for this registry is assigned. If the donor, for example, is coming from USA and has ethnicity HISA, then we do not have a specific set available. However, there is a more broad Hispanic haplotype frequency set available (USA-HI) and will be used in this scenario.

If there is no ethnicity available for the donor or no ethnicity specific haplotype frequency set exists for that country, the system will check if a country specific haplotype frequency set is available or if the country is part of one of the regional haplotype frequency sets.

If no country specific haplotype frequency set is available, the Search & Match Service will use the global consensus haplotype frequency set.

Flowchart assigning a haplotype frequency (HF) set
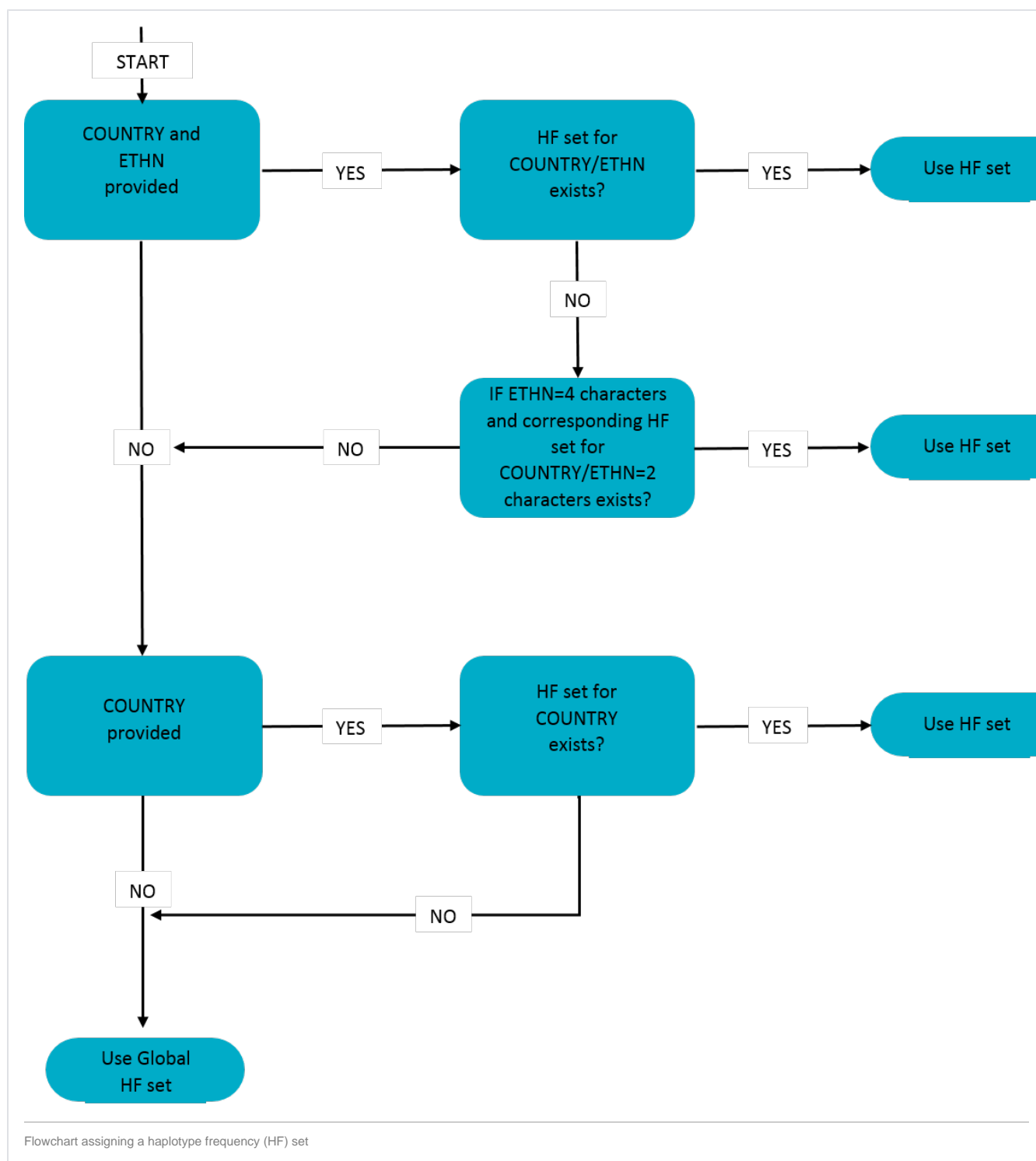
In the table below, you can find the haplotype frequency set numbers and which countries/organisations were included as well as the total sample size. In the last part from the table we included regional sets, some country/ethnicity specific sets and the global consensus set.

The regional sets are defined as follows:

East Asia (eas): CN (CN+CN1), HK, TW, JP (Set number: 32)

Eastern Europe (eeu):AT (A+A2), CZ (CS+CS2), CY (CY+CY2), GR2, PL (PL3+PL5+PL6), TR (TRAN+TRIS+TRKK), BG, HU, HR, LT, MK, RU (R2+R4), RO, RS, SK, SI (Set number: 33)

South America (sam): AR, BR, UY (Set number: 34)

				Input Data		Applied to:
HF set number (pop_id)	**Determination date**	**Sample size**	**ISO country code**	**ION**	**Search & Match Service_Registry_codes**	**IONs**
1	2022-05	226507	AR	5117	AR-INCUCAI	Same
2	2022-05	149750	AT	2614 / 4961	AT-ABMDR / AT-Verein	Same + 8162 - AT-Vita34 (CBB)
3	2022-05	41348	AU + NZ	7748 / 8261	AU-ABMDR / NZ-NZBMDR	Same
4	2022-05	36351	BE	4201	BE-MDPB	Same
5	2022-05	81859	BR	8766	BR-REDOME	Same
6	2022-05	68077	CA	5103 / 6912	CA-One Match / CA-HemaQuec	Same + 3066 - CA-VAR (CBB)
7	2022-05	114900	CH	9341	CH-SBSC	Same
8	2022-05	564150	CN + HK + TW	2197 / 6681 / 4070 / 3458	CN-CMDP / CN-Sunshine / HK-HKBMDR / TW-TzuChi	Same + 1212 - TW-SinoCell (CBB) 1714 - TW-Meribank (CBB) 3105 - HK-CBB (CBB) 5812 - TW-StemCyte (CBB) 6459 - TW-Bionet (CBB) 6692 - TW-Healthbanks (CBB) 9281 - HK-Mononuclear (CBB)
9	2022-05	74633	CY	4278 / 9751	CY-Paraskevaidio / CY-CBMDR	Same
10	2022-05	75810	CZ	4753 / 5440	CZ-CSCR / CZ-CNMDR	Same
11	2022-05	2627544	DE	5525 / 6939	DE-DKMS / DE-ZKRD	Same + DE-DUS (CBB) All ethnicities except for ASSW
12	2022-05	53849	DK	2015 / 7484	DK-DSCDW / DK-DSDE	Same
13	2022-05	81265	ES	7813	ES-REDMO	Same
14	2022-05	39567	FI	9738	FI-FSCR	Same
15	2022-05	90069	FR	1804	FR-FGM	Same
16	2022-05	636595	GB + IE	6354 / 1726 / 2731 / 9968 / 5590	GB-Anthony / GB-WBMDR / GB-BBMR / GB-DKMS / IE-IUBMR	Same
17	2022-05	102358	GR	4979	GR-HTO	Same
18	2022-05	651085	IL	5239 / 4987 / 4068	IL-Hadassah / IL-Ezer Miz. / IL-SHBB	Same

19	2022-05	530272	IN	2824	IN-GeneBand	Same
				4131	IN-MDR	
				4460	IN-ArjanVir	
				8196	IN-ArjanVir	
				8486	IN-Datri	
				9935	IN-DKMS-BMST	
20	2022-05	84477	IT	7450	IT-IBMDR	Same
21	2022-05	328557	NL	8139	NL-Matchis	Same
22	2022-05	16499	NO	7214	NO-NBMDR	Same
23	2022-05	1464544	PL	3918	PL-ALF	Same
				5391	PL-Poltranspl	
				7414	PL-DKMS	
24	2022-05	11029	PT	7358	PT-Cedace	Same
25	2022-05	97139	SA	1810	SA-KFSHRC	Same
				2107	SA-SSCDR	
26	2022-05	184283	SE	5285	SE-Tobias	Same
27	2022-05	78885	SG	3785	SG-BMDP	Same + 4291 - SG-SCBB (CBB)
28	2022-05	122133	TH	8362	TH-TSCDR	Same
29	2022-05	491941	TR	3893	TR-TRAN	Same
				5509	TR-TRIS	
				3503	TR-TURKOK	
30	2022-05	1723292	US	3553	US-NMDP	3553 with ethnicities: • None provided • MX • OT • UK and the following organisations (all ethnicities) 4857 - US-CSCC (CBB) 6579 - US-Cleveland (CBB) 6738 - ZA-DKMS 7470 - US-Lifebank 8118 - ZA-SABMR 8379 - US-StemCyte (CBB) 8691 - US-NCBP (CBB)
31	2022-05	97953	US-GOL	1033	US-GOL	Same
32	2022-05	564561	East Asia: CN+HK+TW+JP	*1212*		4364 - JP-JMDP 8405 - KR-KONOS All other IONs that were used as input have their own country/region specific set.
				1714		
				2197		
				3105		
				3458		
				4070		
				5812		
				6459		
				6681		
				6692		
				6933		

33	2022-05	2398213	Eastern Europe: AT+CZ+CY+GR+PL +TR+BG+HU+ HR+LT+MK+RU+RO +RS+SK+SI	1005 1372 1695 2073 2614 3503 3893 3918 4278 4307 4398 4565 4650 4753 4961 4979 5019 5391 5440 5509 5712 7197 7414 8162 8256 8714 9751 9778		1005 1372 4381 4398 4565 4650 5070 5712 7197 8256 8714 9778
34	2022-05	413438	South America: AR+BR+UY+CL+PY	1574 2547 5117 6517 8766	CL-DKMS PY-VKS AR-INCUCAI UY-SINDOME BR-REDOME	1574 - CL-DKMS 2547 - PY-VKS 4675 - CL-Vidacel (CBB) 6517 - UY-SINDOME Brazil and Argentina have their own country-specific sets.
35	2022-05	38530	ZA	6738 8118	ZA-DKMS ZA-SABMR	None yet. Under investigation.
36	2022-05	35895	IR	4993 6887	IR-INSCDN IR-ISCDP	None yet. Under investigation.
41	2022-05	11012	LU	3099	LU-LMDP	Same
100	2022-05	151,204	DE-ASSW	5525 6939	Subset of donors from these registries with ethnicity: • ASSW	Same
101	2022-05	656,591	US-AF*	3553	Subset of donors from this registry with ethnicities: • AF • AFNA • AFSS	Same +3034
102	2022-05	796,780	US-AS*	3553	Subset of donors from this registry with ethnicities: • ASNE • AS • ASCE • ASOC • ASSE • ASSO • ASSW	Same
103	2022-05	3,740,668	US-CA*	3553	Subset of donors from this registry with ethnicities: • CANA • CA • CAAU • CAER • CAEU	Same

| 104 | 2022-05 | 1,002,893 | US-HI* | 3553 | Subset of donors from this registry with ethnicities:
• HISA
• HI
• HICA | Same
+ 1671 |
| 999 | 2022-05 | 11,430,561 | | | All donors | All donors/ CBUs not specifically matching any other set. |

In order for WMDA to perform these calculations on behalf of member organisations, it is important for organisations to submit donor ethnicity data and high resolution typing when available. This greatly improves the haplotype frequency sets and allows for individual donor match grades to be more accurate.

NOTE: None of the Japanese and Korean donors have been typed for DQB1. This means that no Japanese or Korean donors were included in the calculations for set #32. Therefore, the haplotype frequency set applied to Japanese and Korean donors is likely to be biased against Japanese and Korean . We have kindly received a self-generated haplotype frequency set from the Japanese registry and are working on integrating it in the updated Search & Match Service. Whether to apply the provided Japanese haplotype frequency set or to apply set #32 is still under investigation.

NOTE: in some cases, a country's donors may be used in a national haplotype frequency set as well as a regional one. For example, Austrian donors are used in their own set (#2) as well as the Eastern Europe regional set (#33). When calculating match probabilities, the most specific set is always used. So for donors from Austrian Bone Marrow Donor Registry or Verein Geben für Leben set #2 is used, but for Slovakian donors there were not enough donors typed at high resolution to calculate their own set and so the regional set is assigned (#33).

There are also countries or donors with a specific ethnic background in a country for which there was no specific haplotype frequency set in Optimas, but for which there are now enough donors for their own haplotype frequency set and/or their haplotype. We are currently investigating whether it would provide benefit to host these separately. These include:

- IN-DATRI: Since DATRI is recruiting in a specific region of India, its haplotype frequencies may be significantly different from the haplotype frequencies found throughout the greater India (set #19).
- ZA: There are enough donors from South Africa typed at high resolution to merit its own set. Considering the unique heritage of South African donors, having frequency sets per ethnic group would be especially beneficial.

It is important to know that the haplotype frequency assigned to your organisation can be changed!

You can request that your donors be returned to the global consensus haplotype frequency set or to a regional set. Your request should include the reasons why you believe that another set is more applicable for your donors. Additionally, if your organisation has a frequency set available that you would like to see utilized for the donors of your organisation, we can implement it as well. In this case, your request should include at least a description about your population, the sample size, inclusion and exclusion criteria, calculation method, and the reasons why your haplotype frequency set would be better than the current set applied by WMDA. Both requests will be reviewed by the WMDA Bioinformatics & Innovation Working Group. Please send requests to support@wmda.info.

Date	Version	Description	Author
2017-10-31	1.0	Replacement BMDW global haplotype frequency set for more specific sets	JK
2018-01-24	1.1	Modification some sets; introduced during OptiMatch version 3.31.0	JK
2019-02-15	1.2	Replaced BMDW by WMDA / Search & Match Service; updated email address	JK
2022-08-15	1.3	Copied original page and updated with relevant info for HAP-E algorithm.	MM