NATIONAL MARROW DONOR PROGRAM®

BE THE MATCH®

Accessing the HaploStats Application
May 12, 2014

Accessing HaploStats

HaploStats (http://www.haplostats.org) is a web application provided by the National Marrow Donor Program (NMDP) Bioinformatics group as a tool for the analysis of HLA typing, and to facilitate access to HLA haplotype frequency information relative to various global, country and ethnically specific populations. HLA frequencies made available through HaploStats are used to generate match probabilities in NMDP search reports. Input to the HaploStats application consists of the HLA dataset (frequencies), populations, the HLA loci that the user wishes to retrieve, and an HLA typing the user wishes to analyze. HaploStats also features typing ambiguity score (TAS) and the analysis of unphased genotypes for the input typing for each population. TAS is a measure that evaluates the amount of ambiguity or uncertainty in a typing. Low scores indicate highly ambiguous typing and high scores (up to 1) show the typing has very little ambiguity. TAS can be measured at the level of phased or unphased genotypes and it informs us of how ambiguous the typing is, when imputed to that level. It is a modified version of entropy measure previously published in *V Paunić, L Gragert et al. "Measuring ambiguity in HLA typing methods", PloS One 7 (8), e43585, 2012* (link). The Unphased Genotypes (HLA type) section contains information about the unphased genotypes imputed from the input HLA typing, and as such is more relevant to donor/patient matching operations. It essentially contains aggregated information from the Phased Genotypes section so that the pairs of haplotypes are converted into unphased genotypes, and all the information, such as frequency and likelihood, are aggregated over the haplotype pairs.

Methods Used to Estimate Haplotype Frequencies

Initial frequency tables made available in the legacy version of HaploStats (which is still accessible by clicking on "Please click here to access Classic HaploStats Application" in the current application) have been compiled from the NMDP registry database using the methods outlined in *Maiers, M.,et al. "High resolution HLA alleles and haplotypes in the US population", Human Immunology (2007) 68, 779-*

788 (link). With time, improvements to these methods have been made and HLA frequencies have been created for additional populations. When appropriate, these frequencies are added to the database and made available to the public through the HaploStats online application. Results from the newest publication by *Loren Gragert, et al. "Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry" Human Immunology Volume 74, Issue 10, October 2013, Pages 1313–1320* are available through HaploStats and in frequency tables (link). Information on each of the Datasets and Population statistics is given in the Appendix.

How to Use HaploStats

Before requesting HLA frequency information from HaploStats, you must select the "HLA Dataset", "Population" and "Haplotype Loci" combination that you want to look at (see Figure I). The "HLA Dataset" selected determines which "Populations" and "Haplotype Loci" are available for selection. "Populations" are not available for selection if they have grayed out checkboxes. Only the "Haplotype Loci" available for any given "HLA Dataset" are included on the respective pull down list.

HaploStats can also be accessed directly with URL parameters. For example typing or pasting the following link:
http://haplostats.org/haplostats?dataset=NMDP%20full%202011&populations=AFA,API,CAU,HIS,NAM&loci=A~C~B~DRB1~DQB1&a1=01:01&a2=01:02&b1=08:01&b2=08:01&drb11=03:01&drb12=15:01
into a browsers address line will return the HaploStats input screen fully loaded. Click the <SUBMIT QUERY> button and the <A*01:01, 01:02~ B*08:01, 08:01~DRB1*03:01, 15:01> HLA typing identified in the URL parameter will be submitted for the "NMDP full 2011"[1] HLA dataset (frequencies), the AFA, API, CAU, HIS and NAM Populations and the "A~C~B~DRB1~DQB1" Haplotype Loci combination indicated in the URL parameters.

Filling in the input screen

HLA information can be entered into the HaploStats application for any combination of HLA-A, B, C, DRB1/3/4/5 or DQB1 loci. The HLA can be entered as (see Table 1. for examples):

- Version 2 or version 3 nomenclature
- Any level of allele nomenclature detail (for example B*15:01:01:01)
- Serology
- NMDP allele codes
- Low resolution codes such as A*01:XX (2-digit DNA)

For homozygous loci, it is only necessary to fill in one of the "HLA Type" fields.
Note: some alleles will return a different label than was entered if they are part of a "g" group. "g" groups are sets of alleles that have the same amino acid sequences expressed across the domain of the antigen recognition site (ARS) in HLA proteins. A list of "g" group sets of alleles can be accessed at https://bioinformatics.bethematchclinical.org/WorkArea/DownloadAsset.aspx?id=10534 as an Excel file download (upon clicking this link an excel file is downloaded to your computer and you may find it in the "downloads" folder). You can find the "g" group sets of alleles in Appendix A of the 2011

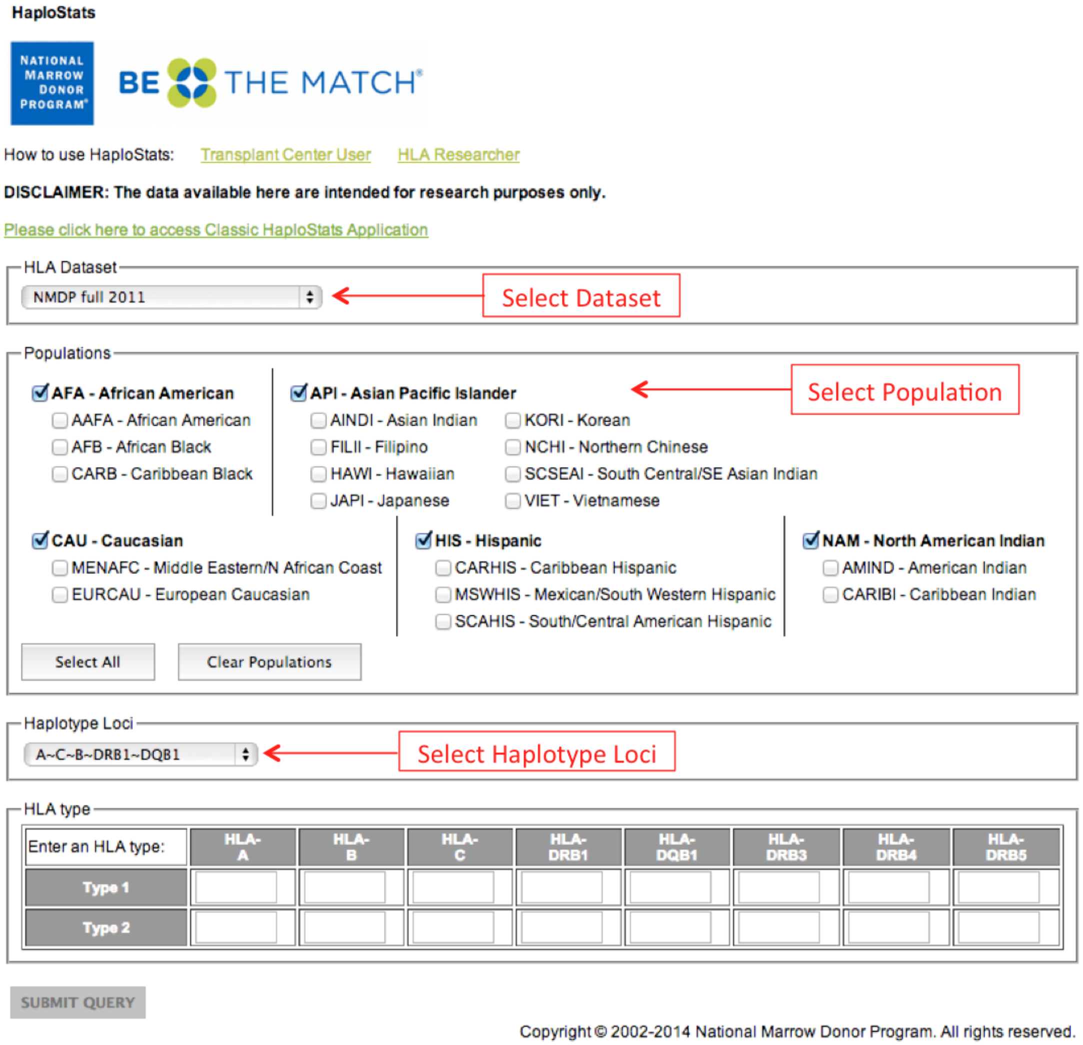
[1] Note that in the URL, space characters are percent-encoded with "%20", for proper usage (also known as URL encoding). So, the percent-encoded representation of dataset name "NMDP full 2011" is "NMDP%20full%202011". However, all modern browsers will also handle space characters in the URL without the percent encoding.

frequency manuscript as "Supplementary data 1" (link). The first allele listed in the group is the allele that will be displayed in the corresponding output of the HaploStats program. So for example if A*01:103 is entered on the HLA typing input screen of the HaploStats program, the associated output information reported for this allele will be labeled A*01:01. A*01:01 is the first allele in the "g" group which includes A*01:103.

This list can potentially change on a quarterly basis as new HLA alleles are discovered. Null alleles are not currently recognized by the HaploStats program. Null allele input is treated as a blank resulting in a homozygous representation of the companion allele at the locus where the null allele was entered. Haplotype frequency tables, which include null alleles in "g" groups and those, which stand by themselves, are available at http://frequency.nmdp.org/NMDPFrequencies2011/. For an up-to-date list of "g" groups with null alleles go to:
https://bioinformatics.bethematchclinical.org/WorkArea/DownloadAsset.aspx?id=10534
or refer to the publication by *Loren Gragert, et al. "Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry" Human Immunology Volume 74, Issue 10, October 2013, Pages 1313–1320* (link).

Once all HLA information has been entered in the appropriate fields for the desired loci, click the <SUBMIT QUERY> button to retrieve haplotype information.

Figure I: HaploStats Input Screen



HaploStats

How to use HaploStats: Transplant Center User HLA Researcher

DISCLAIMER: The data available here are intended for research purposes only.

Please click here to access Classic HaploStats Application

HLA Dataset

NMDP full 2011

Select Dataset

Populations

☑ **AFA - African American**
□ AAFA - African American
□ AFB - African Black
□ CARB - Caribbean Black

☑ **API - Asian Pacific Islander**
□ AINDI - Asian Indian
□ FILII - Filipino
□ HAWI - Hawaiian
□ JAPI - Japanese
□ KORI - Korean
□ NCHI - Northern Chinese
□ SCSEAI - South Central/SE Asian Indian
□ VIET - Vietnamese

Select Population

☑ **CAU - Caucasian**
□ MENAFC - Middle Eastern/N African Coast
□ EURCAU - European Caucasian

☑ **HIS - Hispanic**
□ CARHIS - Caribbean Hispanic
□ MSWHIS - Mexican/South Western Hispanic
□ SCAHIS - South/Central American Hispanic

☑ **NAM - North American Indian**
□ AMIND - American Indian
□ CARIBI - Caribbean Indian

Haplotype Loci

A~C~B~DRB1~DQB1

Select Haplotype Loci

HLA type

Enter an HLA type:	HLA-A	HLA-B	HLA-C	HLA-DRB1	HLA-DQB1	HLA-DRB3	HLA-DRB4	HLA-DRB5
Type 1								
Type 2								

Copyright © 2002-2014 National Marrow Donor Program. All rights reserved.

Table 1. Acceptable Formats for Entering Allele Information into HaploStats

Format	HLA-B Allele Code Examples	Alleles Included in the HaploStats Search
Version 2 Nomenclature		
4 Digit Allele Code	0801	B*08:01
Serology	21	B*49:01-49:05, 50:01, 50:02, 50:04
NMDP Code	15CDF	B*15:01, 15:26N, 15:28, 15:32, 15:33
2 Digit "Wild Card" Allele Code	13XX	B*13:01-13:04, 13:06, 13:07N, 13:08Q, 13:09-13:23, 13:25-13:48, 13:49N, 13:50-13:55, 13:56N, 13:57-13:62, 13:63N, 13:64-13:71
Version 3 Nomenclature		
4 Digit Allele Code	08:01	B*08:01
Serology	21	B*49:01-49:05, 50:01, 50:02, 50:04
NMDP Code	15:CDF	B*15:01, 15:26N, 15:28, 15:32, 15:33
2 Digit "Wild Card" Allele Code	13:XX	B*13:01-13:04, 13:06, 13:07N, 13:08Q, 13:09-13:23, 13:25-13:48, 13:49N, 13:50-13:55, 13:56N, 13:57-13:62, 13:63N, 13:64-13:71

HaploStats output

The main output screen shows the original HLA input as well as three output sections:

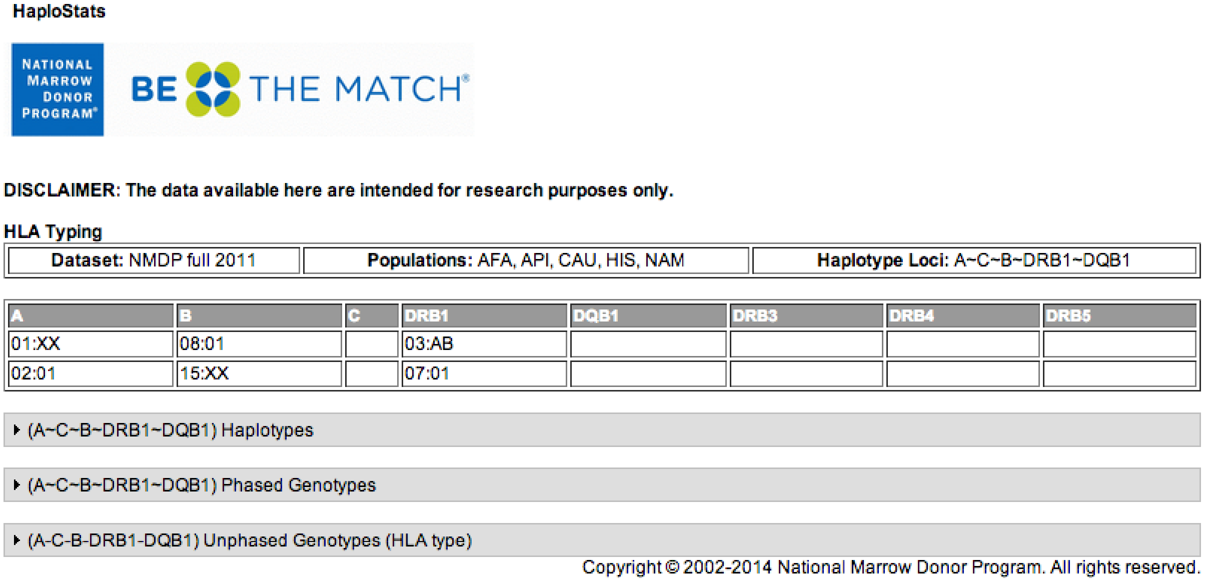
1. ***Haplotypes***
2. ***Phased Genotypes***
3. ***Unphased Genotypes (HLA type)***

Each of these output sections can be expanded or collapsed by clicking on the pointer to the left of the section labels.

1. Haplotypes

When expanded, the ***Haplotypes*** section shows the list of all haplotypes (from permutations of the HLA inputs on the initial HaploStats screen) where at least one of the selected “Populations”/“Haplotype Loci” combinations from the reference “HLA Dataset” has an observed frequency (Figure III). The frequency and rank of each haplotype is shown in the labeled column for each population. The “Collapse” button will collapse this section back.

Figure II: Primary output displayed for the query submitted in Figure I



HaploStats

DISCLAIMER: The data available here are intended for research purposes only.

HLA Typing

Dataset: NMDP full 2011	**Populations:** AFA, API, CAU, HIS, NAM	**Haplotype Loci:** A~C~B~DRB1~DQB1

A	B	C	DRB1	DQB1	DRB3	DRB4	DRB5
01:XX	08:01		03:AB				
02:01	15:XX		07:01				

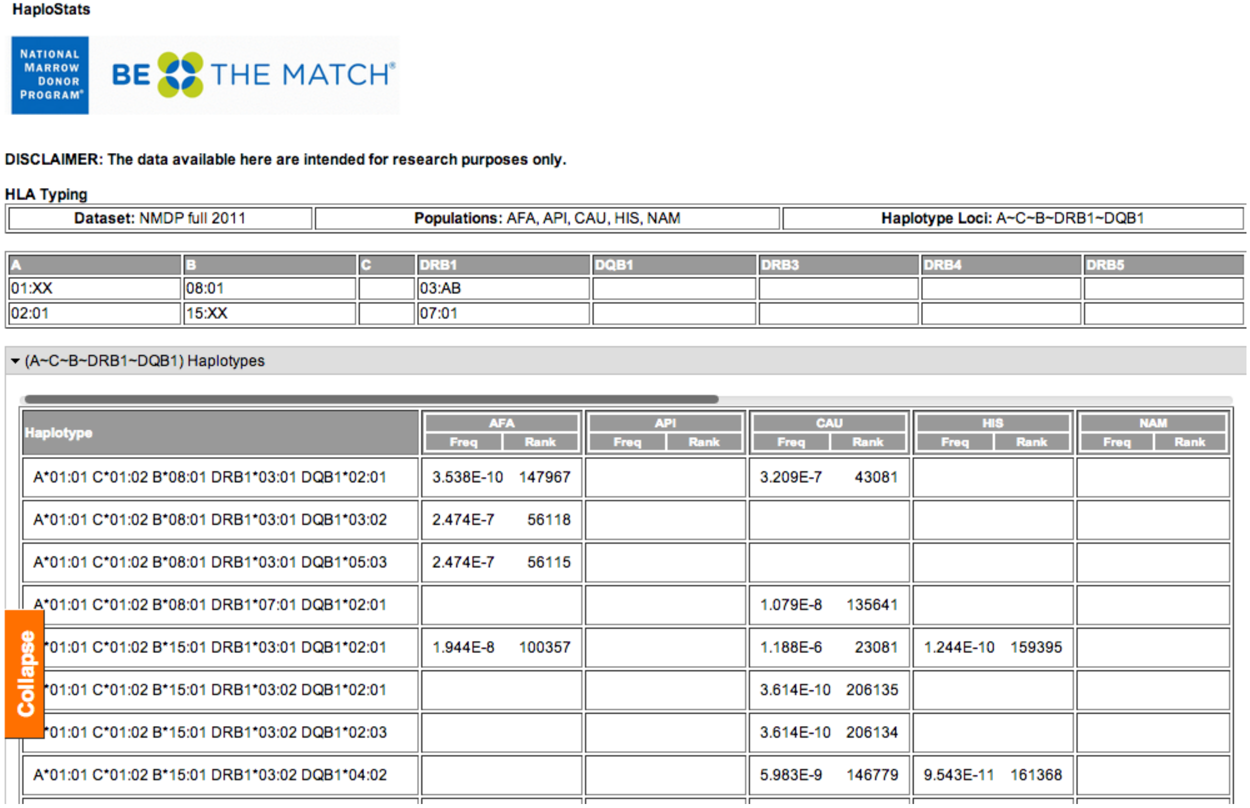
▸ (A~C~B~DRB1~DQB1) Haplotypes

▸ (A~C~B~DRB1~DQB1) Phased Genotypes

▸ (A-C-B-DRB1-DQB1) Unphased Genotypes (HLA type)

Copyright © 2002-2014 National Marrow Donor Program. All rights reserved.

Figure III: Expanded display for the *Haplotypes* output section



HaploStats

DISCLAIMER: The data available here are intended for research purposes only.

HLA Typing

Dataset: NMDP full 2011	**Populations:** AFA, API, CAU, HIS, NAM	**Haplotype Loci:** A~C~B~DRB1~DQB1

A	B	C	DRB1	DQB1	DRB3	DRB4	DRB5
01:XX	08:01		03:AB				
02:01	15:XX		07:01				

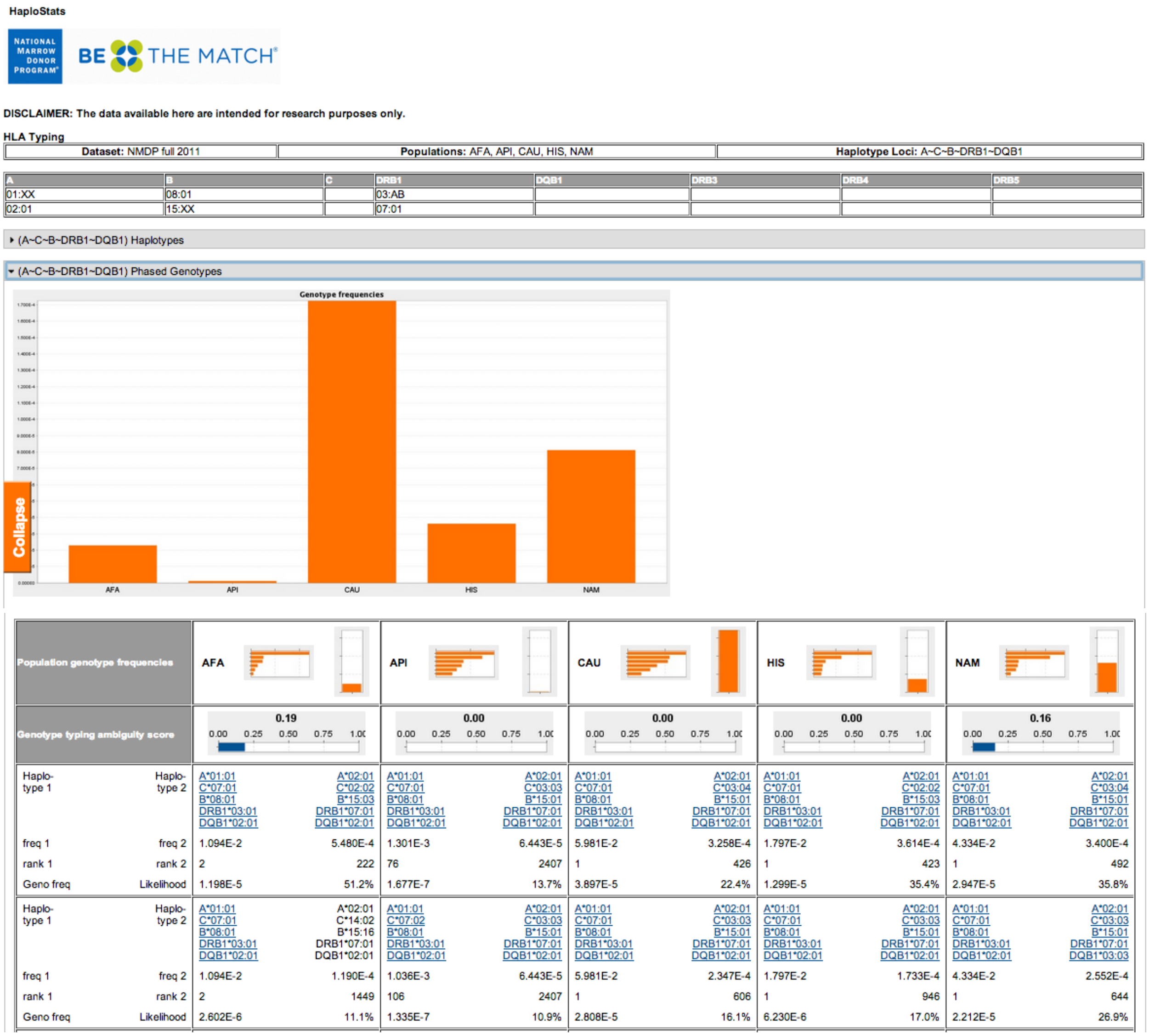
▾ (A~C~B~DRB1~DQB1) Haplotypes

Haplotype	AFA		API		CAU		HIS		NAM	
	Freq	Rank	Freq	Rank	Freq	Rank	Freq	Rank	Freq	Rank
A*01:01 C*01:02 B*08:01 DRB1*03:01 DQB1*02:01	3.538E-10	147967			3.209E-7	43081				
A*01:01 C*01:02 B*08:01 DRB1*03:01 DQB1*03:02	2.474E-7	56118								
A*01:01 C*01:02 B*08:01 DRB1*03:01 DQB1*05:03	2.474E-7	56115								
A*01:01 C*01:02 B*08:01 DRB1*07:01 DQB1*02:01					1.079E-8	135641				
*01:01 C*01:02 B*15:01 DRB1*03:01 DQB1*02:01	1.944E-8	100357			1.188E-6	23081	1.244E-10	159395		
*01:01 C*01:02 B*15:01 DRB1*03:02 DQB1*02:01					3.614E-10	206135				
*01:01 C*01:02 B*15:01 DRB1*03:02 DQB1*02:03					3.614E-10	206134				
A*01:01 C*01:02 B*15:01 DRB1*03:02 DQB1*04:02					5.983E-9	146779	9.543E-11	161368		

2. Phased Genotypes

When expanded the ***Phased Genotypes*** section displays several features (see Figure IV below).

Figure IV: Expanded display for the *Haplotype Pair Frequency* output section



HaploStats

DISCLAIMER: The data available here are intended for research purposes only.

HLA Typing

Dataset: NMDP full 2011	**Populations:** AFA, API, CAU, HIS, NAM	**Haplotype Loci:** A~C~B~DRB1~DQB1

A	B	C	DRB1	DQB1	DRB3	DRB4	DRB5
01:XX	08:01		03:AB				
02:01	15:XX		07:01				

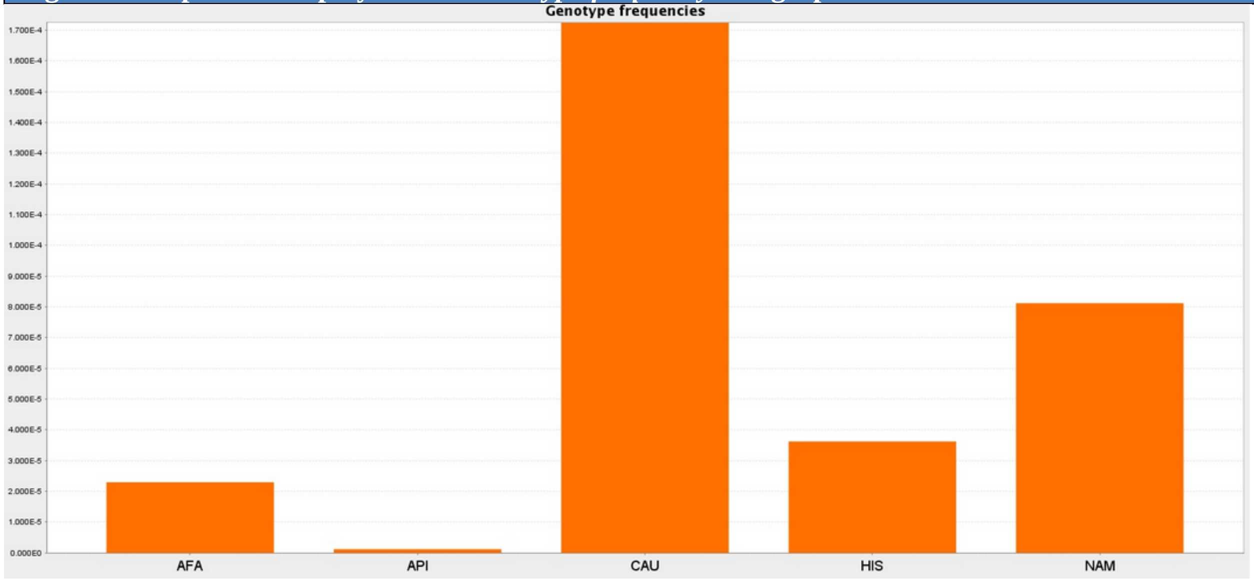
▸ (A~C~B~DRB1~DQB1) Haplotypes

▾ (A~C~B~DRB1~DQB1) Phased Genotypes

Population genotype frequencies		AFA		API		CAU		HIS		NAM	
Genotype typing ambiguity score		0.19		0.00		0.00		0.00		0.16	
Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*02:02 B*15:03 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:04 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*02:02 B*15:03 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:04 B*15:01 DRB1*07:01 DQB1*02:01
freq 1	freq 2	1.094E-2	5.480E-4	1.301E-3	6.443E-5	5.981E-2	3.258E-4	1.797E-2	3.614E-4	4.334E-2	3.400E-4
rank 1	rank 2	2	222	76	2407	1	426	1	423	1	492
Geno freq	Likelihood	1.198E-5	51.2%	1.677E-7	13.7%	3.897E-5	22.4%	1.299E-5	35.4%	2.947E-5	35.8%
Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*14:02 B*15:16 DRB1*07:01 DQB1*02:01	A*01:01 C*07:02 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*03:03
freq 1	freq 2	1.094E-2	1.190E-4	1.036E-3	6.443E-5	5.981E-2	2.347E-4	1.797E-2	1.733E-4	4.334E-2	2.552E-4
rank 1	rank 2	2	1449	106	2407	1	606	1	946	1	644
Geno freq	Likelihood	2.602E-6	11.1%	1.335E-7	10.9%	2.808E-5	16.1%	6.230E-6	17.0%	2.212E-5	26.9%

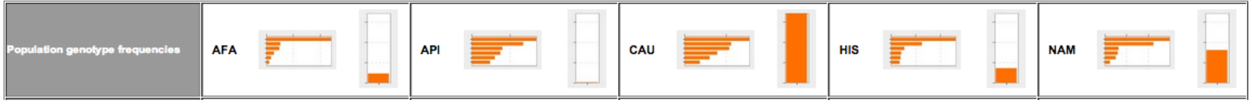
- At the top is a bar graph (orange bars) labeled *Genotype frequencies*, which indicates the relative (sum) frequency of individual genotype frequencies listed under each displayed population. The population with the highest sum frequency over all genotypes is used as the reference population when determining the scale for the overall bar graph display. Clicking on this area brings up a separate scrollable display for the bar graph (see Figure V below).

Figure V: Expanded display for the *Genotype frequency* bar graph



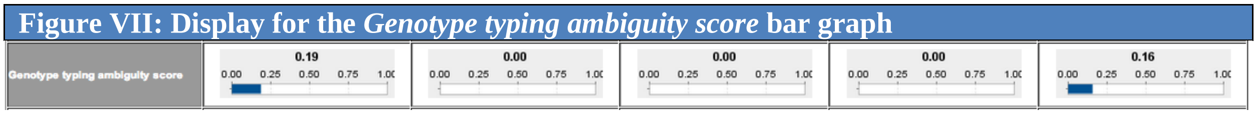
- Under the bar graph of relative population frequencies is a section labeled *Population genotype frequencies* (Figure VI). There are horizontal bar graphs displaying relative frequencies for the top 6 genotypes found within each of the populations requested. The vertical bar to the right of the horizontal bar graph represents the same information found in the larger *Genotype frequency* bar graph at the top of the ***Phased Genotypes*** section display.

Figure VI: Display for the *Genotype frequency* bar graph



- Under the *Population genotype frequencies* area is a set of horizontal blue bars labeled *Genotype typing ambiguity score* (Figure VII). Typing ambiguity score measure the content of ambiguity or uncertainty in the typing for a given population. It is a modified version of *entropy* measure previously published in *V Paunić, L Gragert et al. Measuring ambiguity in HLA typing methods. PloS one 7 (8), e43585, 2012* (link). For details on how the typing ambiguity score is computed please refer to the Appendix.

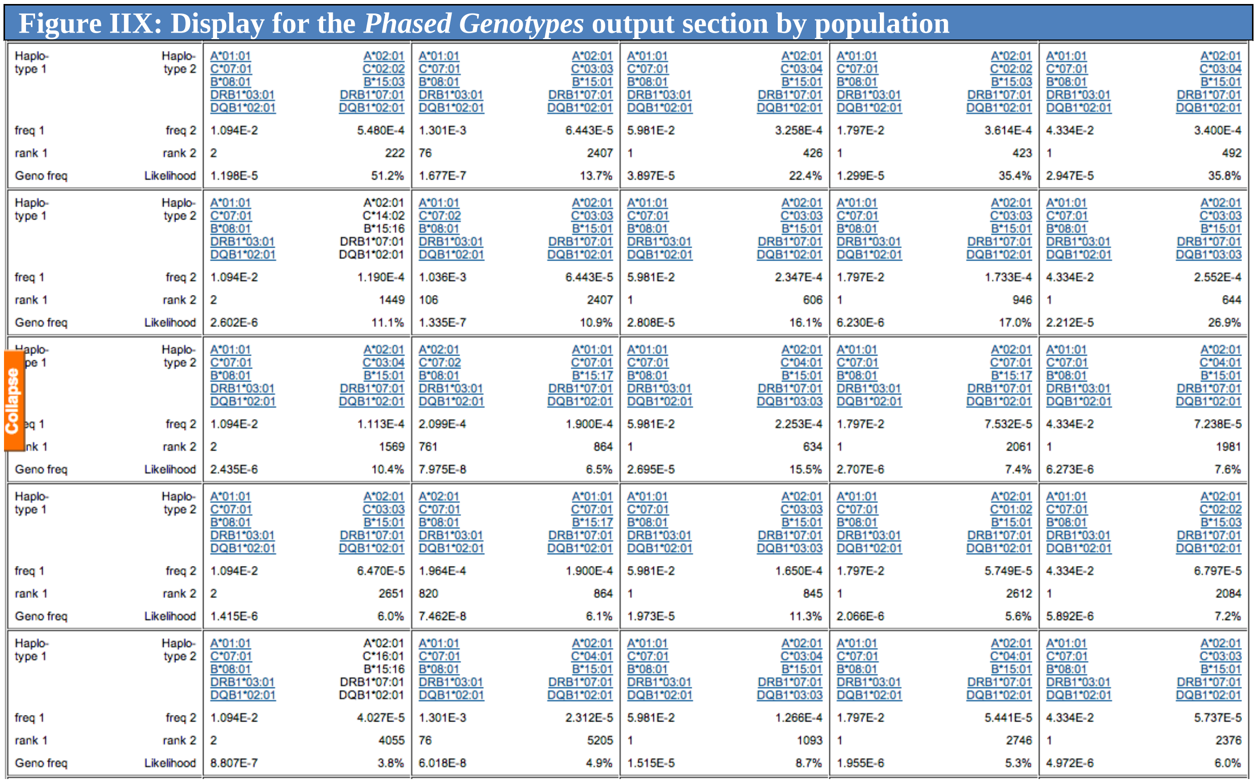
Figure VII: Display for the *Genotype typing ambiguity score* bar graph



- Under the row of bar graphs for the Genotype typing ambiguity score, the list of haplotype pairs is displayed (Figure IIX). There is a column for each population originally selected for inclusion in the results. The population frequency, and rank of each haplotype in the set of haplotype pairs is reported along with the frequency, rank and likelihood of the genotype they comprise being observed in the reference population.

- For low-resolution codes, all possible related combinations, where there is frequency data available, will be displayed in the results (Note: haplotype and genotype frequencies are displayed in Scientific Notation with 3 decimal places).
- When the number of loci for which HLA has been input is less than the number of loci associated with the selected “Haplotype Loci” for the “HLA Dataset” of interest, all possible combinations of haplotype pairs are reported that are an extension of the original HLA input.
- Haplotype and genotype frequency and rank information is broken out by broad and detailed race/ethnicity groups. Ranks are based on the relative position of their frequencies within the reference population. Haplotype pairs are sorted within each population from highest to lowest frequency. Therefore, haplotype pairs do not line up across columns but are sorted independently within population columns.
- **Note that all allele codes entered at a resolution higher than the protein level of nomenclature will have haplotype frequencies rolled up to protein level.** For example, the estimated frequency for haplotypes containing the allele A*02:05:49 will be represented by the frequency of the observed and related haplotypes containing A*02:05.
- The frequencies reported in HaploStats are combined across antigen recognition site (ARS) equivalent alleles. So, for example, A*01:06 and A*01:126 allele observations have been combined in the estimation of haplotype frequencies. Only the first allele in the Alpha/Numeric sort of the nomenclature of the ambiguous set will be listed in the output. So in the example above, if A*01:126 is the input in the initial HLA input screen, HaploStats will display A*01:06 in all of the output. The user must be aware of this constraint. As an aid, a list of ARS equivalent alleles and which allele will be displayed in HaploStats output can be found at https://bioinformatics.bethematchclinical.org/HLA-Resources/Allele-Codes/ under the link “G Group Allele Code List (XLSX)”.

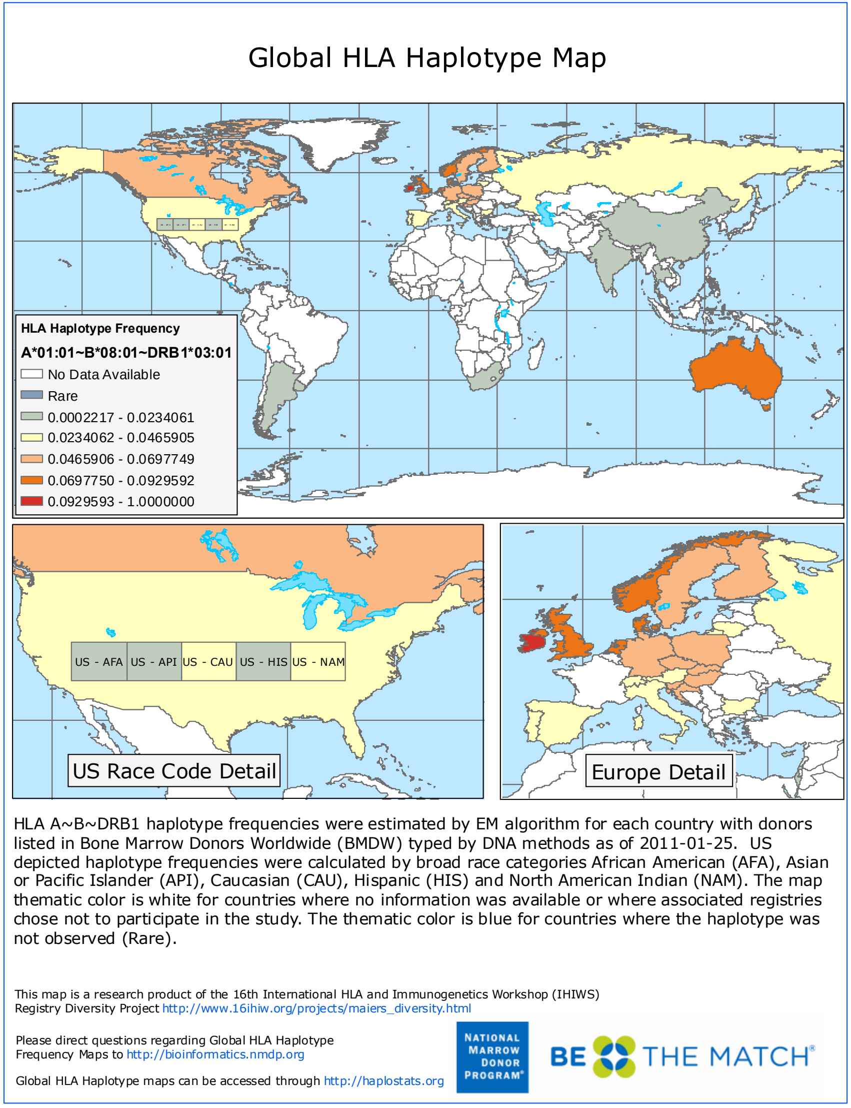
Figure IIX: Display for the *Phased Genotypes* output section by population



Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*02:02 B*15:03 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:04 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*02:02 B*15:03 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:04 B*15:01 DRB1*07:01 DQB1*02:01
freq 1	freq 2	1.094E-2	5.480E-4	1.301E-3	6.443E-5	5.981E-2	3.258E-4	1.797E-2	3.614E-4	4.334E-2	3.400E-4
rank 1	rank 2	2	222	76	2407	1	426	1	423	1	492
Geno freq	Likelihood	1.198E-5	51.2%	1.677E-7	13.7%	3.897E-5	22.4%	1.299E-5	35.4%	2.947E-5	35.8%
Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*14:02 B*15:16 DRB1*07:01 DQB1*02:01	A*01:01 C*07:02 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*03:03
freq 1	freq 2	1.094E-2	1.190E-4	1.036E-3	6.443E-5	5.981E-2	2.347E-4	1.797E-2	1.733E-4	4.334E-2	2.552E-4
rank 1	rank 2	2	1449	106	2407	1	606	1	946	1	644
Geno freq	Likelihood	2.602E-6	11.1%	1.335E-7	10.9%	2.808E-5	16.1%	6.230E-6	17.0%	2.212E-5	26.9%
Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:04 B*15:01 DRB1*07:01 DQB1*02:01	A*02:01 C*07:02 B*08:01 DRB1*03:01 DQB1*02:01	A*01:01 C*07:01 B*15:17 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*04:01 B*15:01 DRB1*07:01 DQB1*03:03	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*07:01 B*15:17 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*04:01 B*15:01 DRB1*07:01 DQB1*02:01
freq 1	freq 2	1.094E-2	1.113E-4	2.099E-4	1.900E-4	5.981E-2	2.253E-4	1.797E-2	7.532E-5	4.334E-2	7.238E-5
rank 1	rank 2	2	1569	761	864	1	634	1	2061	1	1981
Geno freq	Likelihood	2.435E-6	10.4%	7.975E-8	6.5%	2.695E-5	15.5%	2.707E-6	7.4%	6.273E-6	7.6%
Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01	A*02:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*01:01 C*07:01 B*15:17 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*03:03	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*01:02 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*02:02 B*15:03 DRB1*07:01 DQB1*02:01
freq 1	freq 2	1.094E-2	6.470E-5	1.964E-4	1.900E-4	5.981E-2	1.650E-4	1.797E-2	5.749E-5	4.334E-2	6.797E-5
rank 1	rank 2	2	2651	820	864	1	845	1	2612	1	2084
Geno freq	Likelihood	1.415E-6	6.0%	7.462E-8	6.1%	1.973E-5	11.3%	2.066E-6	5.6%	5.892E-6	7.2%
Haplo- type 1	Haplo- type 2	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*16:01 B*15:16 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*04:01 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:04 B*15:01 DRB1*07:01 DQB1*03:03	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*04:01 B*15:01 DRB1*07:01 DQB1*02:01	A*01:01 C*07:01 B*08:01 DRB1*03:01 DQB1*02:01	A*02:01 C*03:03 B*15:01 DRB1*07:01 DQB1*02:01
freq 1	freq 2	1.094E-2	4.027E-5	1.301E-3	2.312E-5	5.981E-2	1.266E-4	1.797E-2	5.441E-5	4.334E-2	5.737E-5
rank 1	rank 2	2	4055	76	5205	1	1093	1	2746	1	2376
Geno freq	Likelihood	8.807E-7	3.8%	6.018E-8	4.9%	1.515E-5	8.7%	1.955E-6	5.3%	4.972E-6	6.0%

- As of April 2011, global HLA Haplotype Maps are available for any haplotype highlighted and underscored in the *Phased Genotypes* output section. The highlighting and underscore indicate an active link to a .pdf file which displays a global frequency map of the haplotype (by country). These maps were developed in collaboration with consenting registries listing donors in Bone Marrow Donors Worldwide (BMDW) (http://www.bmdw.org) and are a product of the 16th International HLA and Immunogenetics Workshop (IHIWS) Registry Diversity Project (http://www.16ihiw.org/projects/maiers_diversity.html). For information on the methodologies used to create the frequencies and maps, see *Gragert, L., Williams E., Madbouly A., Maiers M., Oudshoorn M (2011) High-resolution HLA A-B-DRB1 haplotype frequencies for BMDW registries. Tissue Antigens 77(5): 409-410.* Not all A-B-DRB1 haplotypes reported in HaploStats will have maps available. Only the top 40,000 A-B-DRB1 haplotypes, according to BMDW registries contributing to the individual country frequency calculations, had maps created. See figure IX for an example of a Global HLA Haplotype Map.

Figure IX: Global HLA Haplotype Map



Global HLA Haplotype Map

HLA A~B~DRB1 haplotype frequencies were estimated by EM algorithm for each country with donors listed in Bone Marrow Donors Worldwide (BMDW) typed by DNA methods as of 2011-01-25. US depicted haplotype frequencies were calculated by broad race categories African American (AFA), Asian or Pacific Islander (API), Caucasian (CAU), Hispanic (HIS) and North American Indian (NAM). The map thematic color is white for countries where no information was available or where associated registries chose not to participate in the study. The thematic color is blue for countries where the haplotype was not observed (Rare).

This map is a research product of the 16th International HLA and Immunogenetics Workshop (IHIWS) Registry Diversity Project http://www.16ihiw.org/projects/maiers_diversity.html

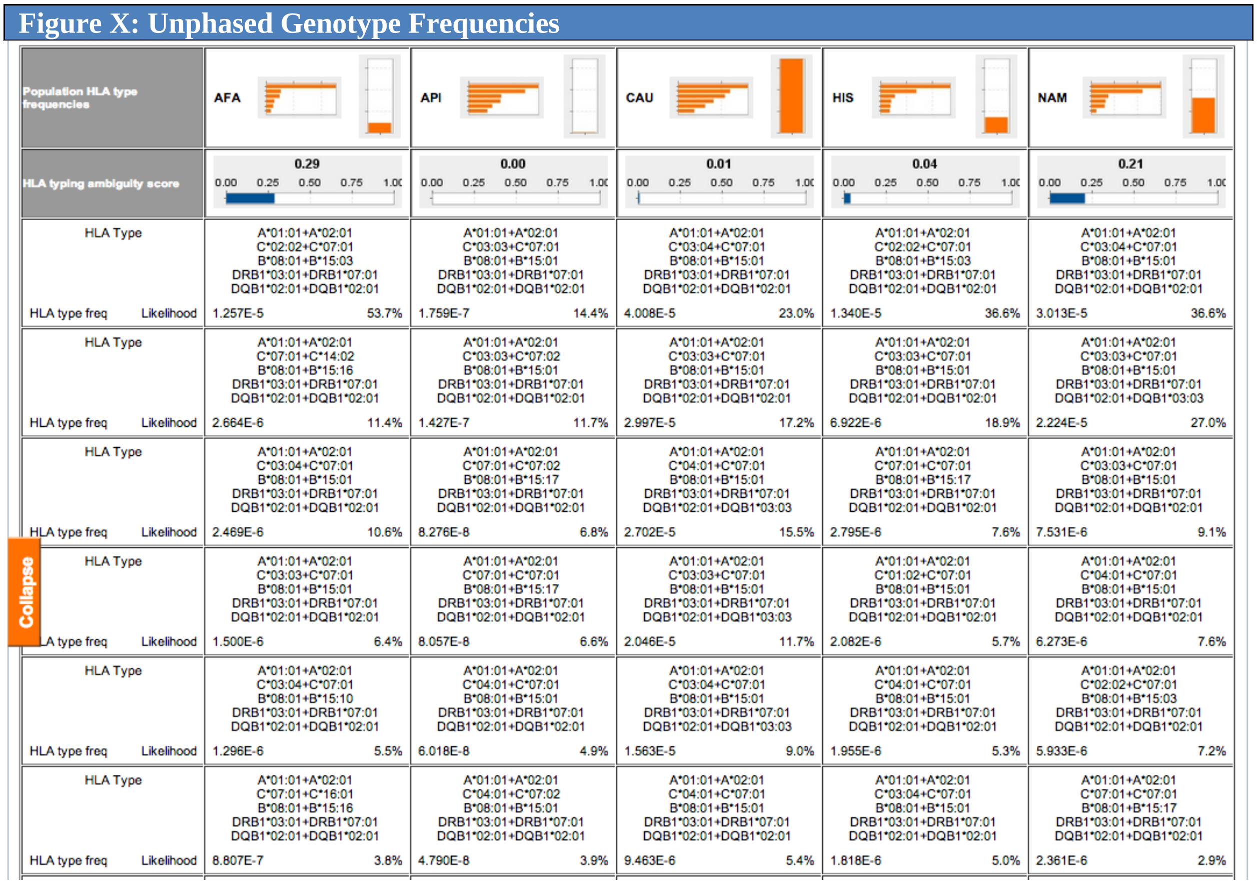
Please direct questions regarding Global HLA Haplotype Frequency Maps to http://bioinformatics.nmdp.org

Global HLA Haplotype maps can be accessed through http://haplostats.org

3. Unphased Genotypes (HLA type)

This section contains information about the unphased genotypes imputed from the HLA typing entered by a user (see Figure X below). It is similar in layout to the Phased Genotypes section as it essentially contains aggregated information from this section. The difference between these two sections is that the *Unphased Genotypes* section does not include any haplotype or phase information.

Figure X: Unphased Genotype Frequencies



Population HLA type frequencies	AFA	API	CAU	HIS	NAM
HLA typing ambiguity score	0.29	0.00	0.01	0.04	0.21
HLA Type	A*01:01+A*02:01 C*02:02+C*07:01 B*08:01+B*15:03 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:04+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*02:02+C*07:01 B*08:01+B*15:03 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:04+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01
HLA type freq Likelihood	1.257E-5 53.7%	1.759E-7 14.4%	4.008E-5 23.0%	1.340E-5 36.6%	3.013E-5 36.6%
HLA Type	A*01:01+A*02:01 C*07:01+C*14:02 B*08:01+B*15:16 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:02 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*03:03
HLA type freq Likelihood	2.664E-6 11.4%	1.427E-7 11.7%	2.997E-5 17.2%	6.922E-6 18.9%	2.224E-5 27.0%
HLA Type	A*01:01+A*02:01 C*03:04+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*07:01+C*07:02 B*08:01+B*15:17 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*04:01+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*03:03	A*01:01+A*02:01 C*07:01+C*07:01 B*08:01+B*15:17 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01
HLA type freq Likelihood	2.469E-6 10.6%	8.276E-8 6.8%	2.702E-5 15.5%	2.795E-6 7.6%	7.531E-6 9.1%
HLA Type	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*07:01+C*07:01 B*08:01+B*15:17 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:03+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*03:03	A*01:01+A*02:01 C*01:02+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*04:01+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01
LA type freq Likelihood	1.500E-6 6.4%	8.057E-8 6.6%	2.046E-5 11.7%	2.082E-6 5.7%	6.273E-6 7.6%
HLA Type	A*01:01+A*02:01 C*03:04+C*07:01 B*08:01+B*15:10 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*04:01+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:04+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*03:03	A*01:01+A*02:01 C*04:01+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*02:02+C*07:01 B*08:01+B*15:03 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01
HLA type freq Likelihood	1.296E-6 5.5%	6.018E-8 4.9%	1.563E-5 9.0%	1.955E-6 5.3%	5.933E-6 7.2%
HLA Type	A*01:01+A*02:01 C*07:01+C*16:01 B*08:01+B*15:16 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*04:01+C*07:02 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*04:01+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*03:04+C*07:01 B*08:01+B*15:01 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01	A*01:01+A*02:01 C*07:01+C*07:01 B*08:01+B*15:17 DRB1*03:01+DRB1*07:01 DQB1*02:01+DQB1*02:01
HLA type freq Likelihood	8.807E-7 3.8%	4.790E-8 3.9%	9.463E-6 5.4%	1.818E-6 5.0%	2.361E-6 2.9%

All genotypes in the ***Phased Genotypes*** section are converted into unphased genotypes, and all the information, such as frequency and likelihood, are aggregated over all phased genotypes that could be expanded from the given unphased genotype. For example, given the following list of potential genotypes in the *Phased Genotypes* section, (returned from a given input typing):

Table 2. Example (phased) genotypes with frequencies and likelihoods

Genotype	Genotype frequency	Likelihood
A*26:01~B*08:01~DRB1*03:01, A*02:01~B*51:01~DRB1*12:01	3.192e-05	0.83
A*02:01~B*08:01~DRB1*03:01, A*26:01~B*51:01~DRB1*12:01	3.192e-05	0.11
A*02:02~B*08:01~DRB1*03:01, A*26:01~B*51:01~DRB1*12:01	2.307e-06	0.06

The *Unphased Genotypes* section will contain the following unphased genotypes and corresponding frequencies and likelihoods:

Table 3. Example unphased genotypes with frequencies and likelihoods

Unphased genotype	Unphased genotype frequency	Likelihood
A*26:01, A*02:01 B*08:01, B*51:01 DRB1*03:01, DRB1*12:01	6.384e-05 (3.192e-05+3.192e-05)	0.94 (0.83+0.11)
A*26:01, A*02:02 B*08:01, B*51:01 DRB1*03:01, DRB1*12:01	2.307e-06	0.06

For comments and questions about HaploStats, please send correspondence to:

haplostats@nmdp.org

Appendix

Data Sets Currently Available in HaploStats

There are currently 2 HLA Datasets available to select from. The “NMDP high res 2007” and “NMDP full 2011” HLA Datasets. Information in the “NMDP high res 2007” HLA Dataset is an expansion of the original published haplotype dataset outlined in the 2007 Maiers, M., Gragert, L., and Klitz, W. publication. The North American Indian race/ethnicity group has been added to this HLA Dataset. Statistical characteristics of this dataset and the populations that can be selected are presented in Table 4.

Table 4.
Summary of HaploStats A-B-DRB1 haplotype sampling characteristics among five US census population categories

Race Code	Total sample size (2N)	Number of haplotypes with n>=1	% of samples in the top 100 haplotypes	% of haplotypes with n>1
AFA	9,178	2,780	31	98.5
API	7,348	1,977	39	99.0
CAU	54,874	4,522	46	99.2
HIS	9,960	2,858	35	98.6
NAM	3,174	1,063	45	98.0
Overall	84,534	13,200	39	98.7

Information in the “NMDP full 2011” HLA Dataset is from the original published haplotype dataset outlined in the 2013 publication by Loren Gragert, Abeer Madbouly, John Freeman and Martin Maiers “Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry” Human Immunology Volume 74, Issue 10, October 2013, Pages 1313–1320 and are available through HaploStats and in frequency tables at link. Statistical characteristics of this dataset and the populations that can be selected are presented in Table 5. This HLA Dataset is the same data referenced to create NMDP search reports. The broad race groups from the 2007 publication are also available to select for the “NMDP full 2011” HLA Dataset; see list and statistical parameters in Table 6.

Table 5. Size of Haplotype Frequency Generating Samples - detailed race

Race Code	Detailed Race/Ethnic Description	Broad	Count	Typed C	Typed DQB1	Typed DRB3/4/5
AAFA	African American	AFA	416581	99946	16178	134076
AFB	African	AFA	28557	6975	1488	8516
AINDI	South Asian Indian	API	185391	29635	8409	44484
AISC	American Indian - South or Central Am.	NAM	5926	1255	228	894
ALANAM	Alaska Native or Aleut	NAM	1376	288	100	347
AMIND	North American Indian	NAM	35791	7006	2398	13821

CARB	Caribbean Black	AFA	33328	10012	1856	9115
CARHIS	Caribbean Hispanic	HIS	115374	21286	4420	31097
CARIBI	Caribbean Indian	NAM	14339	5631	937	1372
EURWRC	European Caucasian	CAU	1242890	395676	81106	212472
HAWI	Hawaiian or other Pacific Islander	API	11499	3110	505	3355
JAPI	Japanese	API	24582	3552	852	7886
KORI	Korean	API	77584	11656	2107	25082
MENAFC	Middle Eastern or N. Coast of Africa	CAU	70890	22337	4415	17609
MSWHIS	Mexican or Chicano	HIS	261235	50875	12721	85021
NCHI	Chinese	API	99672	16621	3753	23569
SCAHIS	Hispanic – South or Central American	HIS	146714	31446	5764	29331
SCAMB	Black - South or Central American	AFA	4889	927	203	1677
SCSEAI	Southeast Asian	API	27978	5579	1321	3946
VIET	Vietnamese	API	43540	10511	1032	2446

Table 6. Size of haplotype frequency generating samples - broad race

Broad Race Code	Race/Ethnic Description	Count	Typed C	Typed DQB1	Typed DRB3/4/5
AFA	African American	505 250	123 871	21 408	156 764
API	Asian or Pacific Islander	568 597	104 027	21 814	142 755
CAU	Caucasian	3 912 440	1 808 061	502 117	1 596 577
DEC	Declined	23 072	17 526	805	3 241
HIS	Hispanic	712 764	166 192	31 700	163 539
MLT	Multiple Race	390 676	139 805	26 268	83 066
NAM	Native American	46 148	9 533	2 977	15 469
OTH	Other	22 241	7 173	536	667
UNK	Unknown	367 838	321 915	318 348	219 988

Typing Ambiguity Score

Typing ambiguity score (TAS) is calculated as follows:

$$TAS = 1 - \min(1, H)$$
$$H = -\sum p * log_{10}(p)$$

Where *p* is a vector of normalized genotype frequencies (likelihoods), and H is entropy. Note that vector *p* must sum to 1, otherwise the equation will return meaningless results. Since entropy goes from large positive value to zero, zero being the best case, typing ambiguity score is simply 1-entropy in order to bring the range to zero to one, where one is the best case. Min (1, H) implies that, the maximum entropy that we allow in our application is 1. Anything above that gets rounded to 1 to produce typing ambiguity score of 0.

Example

We will use 3-locus example for simplicity. Given ambiguous typing at all loci:
A*02:KRAW, A*26:KPVR
B*08:MCWE, B*51:MDCD
DRB1*03:JZMU, DRB1*12:JUFV

All possible genotypes, their frequencies and likelihoods returned by imputation are:

Table 7. Example genotypes along with their frequencies and likelihoods

Genotype	Genotype frequency	Likelihood
A*26:01~B*08:01~DRB1*03:01, A*02:01~B*51:01~DRB1*12:01	3.192e-05	0.83
A*02:01~B*08:01~DRB1*03:01, A*26:01~B*51:01~DRB1*12:01	3.192e-05	0.11
A*02:02~B*08:01~DRB1*03:01, A*26:01~B*51:01~DRB1*12:01	2.307e-06	0.06

To compute the typing ambiguity score with respect to the phased genotypes, we use the following vector p = (0.83, 0.11, 0.06) in the above equation for TAS.

To compute the typing ambiguity score with respect to unphased genotypes, we obtain aggregated likelihoods by summing over likelihoods of the participating phased genotypes, as shown in the following table.

Table 8. Example Unphased genotypes along with their frequencies and likelihoods

Unphased genotype	Unphased frequency	Likelihood
A*26:01, A*02:01 B*08:01, B*51:01 DRB1*03:01, DRB1*12:01	6.384e-05 (3.192e-05+3.192e-05)	0.94 (0.83+0.11)
A*26:01, A*02:02 B*08:01, B*51:01 DRB1*03:01, DRB1*12:01	2.307e-06	0.06

To obtain TAS, we plug in p = (0.94, 0.06) into the TAS equation.