

# HFCuS

HaplotypeFrequency  
Curation Service

## Input file

ASCII File, XML

## Structure

Mandatory	XML Tag	Values / Kind of Val.	Description
Yes	POP_ID	URI from POP DB	Population Identifier
Yes	HT_List	List of HT	
Yes	HTL	Pair of HTL_Name, HTL_Freq	
Yes	HTL_Name	GL String	
Yes	HTL_Freq	$0 < f \leq 1$	
Yes	HT_Lic	License ID	License under which HF data is available
Yes	HT_Res	G, P, g <sub>NMDP</sub> , g <sub>DKMS</sub> , n-Field, Serology	
No	GT_List	List of GTL	List of Genotypes
No	GTL	Record of GTL_Name, GTL_M_List	
No	GTL_Name	GL String	Raw Data
No	GTL_M_List	List of GTLP_Meth	Additional Information on GT, Typing
No	GTLP_Meth	Pair of GTLP_M_Data, GTLP_M_Value	
No	GTLP_M_TYPE	String	Free Text / Predefined Tag
No	GTLP_M_VALUE	Free	
No	GT_Lic	License ID	License under which GT data is available
No	METHOD_LIST	List of METHOD	
No	METHOD	Record of METH_Type, METH_Value, METH_CLASS	
No	METH_Type	String	Free Text
No	METH_Value	String	Free Text
No	METH_CLASS	String	From predefined list / Text
No	QUALITY_LIST	List of QUALITY	

No	QUALITY	Record of QUAL_TYPE, QUAL_VALUE, QUAL_CLASS	
No	QUAL_TYPE	String	Free Text
No	QUAL_VALUE	String	Free Text
No	QUAL_CLASS	String	Predefined list / Free Text
No	LABEL_LIST	List of LABEL	
No	LABEL	Pair of LABEL_TYPE, LABEL_VALUE	
No	LABEL_TYPE	Free Text	Predefined List and Free Text
No	LABEL_VALUE	Free Text	
No	LABEL_CLASS	Free Text	Predefined List and Free Text
No	ACL	ToBeDefined <Defaults to public/private> ??	Access Control List
No	COHORT_ID		As an alternative to GT List, <b>NOT THE SAME IDs as used in the HFCeS</b>
No	METHOD_ID		As an alternative to Method List, <b>NOT THE SAME IDs as used in the HFCeS</b>

#### Additional Stored Values

- Timestamp
- Submitting UserID

#### Direct Output/Feedback

- URI to dataset
- Method\_ID
- Cohort\_ID
- HF\_ID

## Internal Data Structure of HFCuS

### Basic Fields

The basic data structure of the HFCuS mimics the input file

### Additional Fields

To allow for curation of submitted data sets, comments can be used. They are an independent submission to the HFCuS but refer to an existing HF data set. The field COM\_REF\_SPEC can be used to specifically address a comment to a certain piece of data in the original set.








Mandatory	Data_Field	Content	Description
	COMMENT	Record of COM_TIME, COM_USER, COM_REF_HF, COM_LIST	Mandatory fields: COM_TIME, COM_USER, COM_REF_HF, COM_LIST
	COM_TIME	Timestamp	Time of addition of the comment
	COM_USER	UserID	The user of HFCuS adding the comment
	COM_REF	Pair of COM_REF_TARGET, COM_REF_ID	The HF set the comments refer to
	COM_REF_TARGET	Free Text/COHORT_ID, METHOD_ID, HF_ID, POP_ID, COMMENT	
	COM_REF_ID	ID	Appropriate Of the above target
	COM_LIST	List of COM_REC	
	COM_REC	Record of COM_TEXT, COM_REF_SPEC	
	COM_TEXT	Free Text	The comment
	COM_REF_SPEC	Free Text	Some hints what the comment is referring to

Also, if the genotype list is available, GTs can be downloaded, HF resubmitted for the same Cohort\_ID with a different (better!) methodology.

## License Models

People submitting to HFCuS shall choose one of the following options for licensing:

### Seven regularly used licenses [\[ edit \]](#)

Icon ⇅	Description ⇅	Acronym ⇅	Free Cultural Works ⇅	Remix culture ⇅	Commercial use ⇅
	Freeing content globally without restrictions	CC0	Yes	Yes	Yes
	Attribution alone	BY	Yes	Yes	Yes
	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes
	Attribution + Noncommercial	BY-NC	No	Yes	No
	Attribution + NoDerivatives	BY-ND	No	No	Yes
	Attribution + Noncommercial + ShareAlike	BY-NC-SA	No	Yes	No
	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	No	No	No

[\[16\]](#)[\[17\]](#)

taken from: [https://en.wikipedia.org/wiki/Creative\\_Commons\\_license](https://en.wikipedia.org/wiki/Creative_Commons_license)

## User and Group Models

TBdone, TBdefined

## List of Method Tags

METH_CLASS		HH2016
METH_TYPE	VALUE	DESCRIPTION
EM_ALGORTIHM	String	The EM Algorithm used
EM_VERSION	String	Version of the EM
EM_ALG_REF	String	A reference to the algorithm used
MAC_SERVICE	String	The MultiAlleleCodeService used
MAC_SER_REF	String	A reference to the MAC Service
MAC_VERSION	String	Version of the MAC Service
ARS_SERVICE	String	The service used to translate typing resolutions
ARS_SERV_REF	String	A reference to the ARS service used
ARS_VERSION	String	Version of the ARS Service
HWE_METHOD	String	The Method used for HWE deviation estimation
HWE_REF	String	A reference to the HWE deviation estimation method
LD_METHOD	String	The LD estimation Method
LD_METHOD_REF	String	A reference to the LD estimation method
EM_PARAM_...		
ARS_PARAM_...		
HWE_PARAM_..		

## List of Geotype-Method Tags

METH_CLASS		HH2016
GTLP_M_TYPE	GTLP_M_VALUE	DESCRIPTION
TYPING_METHOD	SSO, SSP, Serology, SangerSequencing, NGS, Free Text	The Typing Method used
TYPING_REF	String	A reference to the typing method
TYPING_DATE	Date	Date of typing
TYPING_IMGT_VER	String	Version on IMGT(/HLA) used to type the sample
MIRING_REF	Reference	A reference to a MIRING compliant set of details to the typing of the sample

## List of Quality Tags

QUAL_CLASS HH2016		
QUAL_TYPE	VALUE	DESCRIPTION
DIV_LAMBDA	real, < 0	Exponent of Power Law fit to HTF distribution (This is called alpha in Slater et al. Power Laws for Heavy-Tailed Distributions?)
DIV_50	integer	Number of haplotypes needed (in descending order of frequency) to have the cumulative sum be > 0.5 (Sample size sensitive!)
DIV_50_REL	Real, $0 \leq x \leq 1$	Number of haplotypes needed (in descending order of frequency) to have the cumulative sum be > 0.5 divided by the number of HT
SAM_SIZE	integer	Number of GT
SAM_POP	integer	Size of Population (approx.)
DIV_PGD	Real, $0 \leq x \leq 1$	Population genetics diversity ( $1 - \sum f_i^2 N / (N-1)$ )
DIV_HEAVY_TAIL	Real, $0 \leq x \leq 1$	a is an independence parameter of the Bayesian SHF model that describes how allele frequency products correlate with haplotype frequencies (also correlates with the fraction of nonzero categories) – From Yoram SHF MS
RES_TRS_COUNT	Real, $0 \leq x \leq 1$	Jan knows that – (Average number of possible genotypes per individual?)
RES_TRS	Real, $0 \leq x \leq 1$	Typing Resolution Score – Average sum of square of genotype probabilities (imputation using population-specific HF estimate, could also do uniform HF global)

<b>RES_SHARE_AMBIG</b>	Real, $0 \leq x \leq 1$	Fraction of GT with a lower resolution than defined in the resolution tag
<b>RES_MISS_LOCI</b>	Real, $0 \leq x \leq 1$	Fraction of GT with missing loci (separate qual_type per locus?)
<b>DEV_HWE</b>	Real	Deviation from HWE (using HWE with ambiguity method)
<b>ERR_STD</b>	Real, $0 \leq x \leq 1$	Weighted average of standard errors across all haplotypes
<b>ERR_SAMP_80_100</b>	Real	Laurent, Excoffier “If” between frequencies derived from 100% set and 80% training set
<b>SUM_FREQ_GAP</b>	Real	Sum of haplotype frequencies for unobserved haplotypes that are expected in population by SHF model
<b>ERR_OFFSET</b>	Real, $0 \leq x \leq 1$	$1 - \sum f_i$ (Difference between predicted full HF distribution using SHF versus actual including test set?)
<b>LD_MEASURE</b>	Real	Define in Method section – (Where is LD measured for quality?)
<b>KFOLD_IMPUTE</b>	Real, $0 \leq x \leq 1$	% of imputable GT in 20% test set from HT generated in 80% training set
<b>KFOLD_PRED_ACTUAL</b>	Real, $0 \leq x \leq 1$	Divergence between predicted and actual with Log Loss function (for test set predictions on simulated lower-resolution typings)
<b>KFOLD_N</b>	integer	Number of independent training-test folds (k)

## List of Labels

LABEL_CLASS HH2016		
LABEL_TYPE	VALUE	DESCRIPTION

<b>GT_REGISTRY</b>	String	ION or other description of the entity hosting the GT
<b>HT_ESTIMATION_ENT</b>	String	ION or other description of the entity performed the HTF analysis

## Open Issues for HFCuS

- User / Group Management, Schema, Specs
- ACL: Management and Specs
- Governance structure
- Implementation



H/F Curation (ASCII, XML)			
Input (Creation Output)			
Mandatory	TAG	Description	Kind of Value
○	Label List	List of user def. labels	
○	Label	Individual Label	Free text
○	GT Meta Data List	E.g. Typing Method, IMGT version...	Text fields
○	GT Meta Data Type		
○	GT Meta Data Value		
✓	Pop-ID	Population ID	URI
○	License GT	Usability of data	List of Licenses
○	GT List	Genotypes (Raw Data)	List of GT
○	GT	One Genotype	GL - String
○	Cohort ID	Sample ID	INTEGER
✓	License HT	Usability of data	List of Licenses
✓	Resolution	Of HT	GP, gnum, gnum1
✓	ACL	Defaults to Public/Private	n-Field, Serology
✓	HT List	Haplotypes	List of HT
✓	HT	Haplotype	GL - String
✓	HT Freq.	Frequency	Real number 0.4f. ≤ 1
○	Method-ID	Description of Method	INTEGER
○	Method		Text field
○	Method-Parameter	Part of	Free text
○	Method-Parameter		Free text, likely TR
○	Quality A	as Method	

# HF-curation

## \*Additional Stored Values

- timestamp
- User (Submitting)

## \*direct Output/Feedback

- URI
- Method-ID
- Cohort-ID
- HF-ID

## Population Data Base

- ID
- Free Text  
Definition

⇒ URI

## Further HF-Curation Requirements

- User Admin Definitions
- Licensing Models

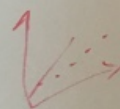


# Quality

Class: HH 2016

Abb.	Value	Description
DIV_LANDA	$\lambda \in \mathbb{R}, \lambda < 0$	Power-Law Approx.
DIV_50	$n \in \mathbb{N}$	Number of Haplotypes <sup>sampled</sup> needed to get $\sum f_i \geq 0.5$
DIV_50_REL	$r \in \mathbb{R}, 0 \leq r \leq 1$	$r = \frac{u}{N}$
SAH_SIZE	$u \in \mathbb{N}$	Number of GT
SAH_POP	$n \in \mathbb{N}$	Size of Population
DIV_PG_D	$r \in \mathbb{R}, 0 \leq r \leq 1$	$1 - \sum f_i^2 \left( \frac{N}{N-1} \right)$
DIV_HEAVY_TAIL	$\alpha \in \mathbb{R}$	
RES_TRS_COUNT	$0 \leq r \leq 1$	See next WS $\rightarrow$
RES_TRS	$0 \leq r \leq 1$	TRS as def. by UNDP
RES_SHARE_ANDID	$0 \leq r \leq 1$	Fraction of GT with ANDID variants w/ respect to defined Resolution.
RES_MISS_LOCI	$0 \leq r \leq 1$	Fraction of GT w/ miss. loci

Tag	Value	Description / Definition
DEV_HWE	$\in \mathbb{R}$	Deviation from HWE.
ERR_STD	$\Delta \in \mathbb{R}, \Delta < 1$	Method of determination leads to go into Description field
ERR_SAMP_80_100	$\in \mathbb{R}$	Standard Error, weighted AV61
SVM_FREQ_GAP	$\in \mathbb{R}$	Exp. Ex. 10. 11. 12. / Laurent
ERR_OFFSET	$\in \mathbb{R}$	Exp. just unobserved (LD!)
LD_MEASURE	$\in \mathbb{K}$	$1 - 2 \frac{1}{2}$
KFOLD_INPUT	$\in \mathbb{R}$	Document LD measure!
KFOLD_PRED_ACTUAL	$\in \mathbb{R}$	% imputable w/ Lorenz, Sjöberg, GT



## DIVERSITY

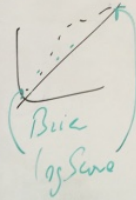
- $1 - \sum f_i^2 \left( \frac{N}{N-1} \right)$
- Power Law  $\lambda$
- Heavy Tail  $\alpha$
- # of Haplotypes to get to 50%

Resolution  
TRS  $\frac{\sum_i (q_i^2)}{N}$   
"circular"

Avg Genotype possibility #  
Uniform HFE global

## k-fold X Val

- % imputable
- Pred vs actual



## QUALITY

✓ Std Err for Each Haplotype  
Weighted Sum  $\sum_i h f_i \cdot \text{std}_i$

✓ Deviation from HWE  
✓ if 100% vs 80% Limit  
Sampling Error Exceeded

✓ Smol frequency "gap"  
Exp but unobserved

⌋ How Many Prospective  
Typing **ALL!!!**