MINUTES

Haplotype Hackathon

Date | August 10, 2016 | Meeting called to order by Martin Maiers

In Attendance

Pradeep Bashyal
Hans-Peter Eberhard
Loren Gragert
Michael Halagan
Jan Hofmann
Steven Mack
Martin Maiers
Jurgen Sauter
Joel Schneider

Service Requirements

Meet the needs of:

- BMDW
- WMDA
- AFND

1. Standard input and output formats (genotypes and/or haplotypes)
    a. Genotype set = COHORT
    b. haplotypes & frequencies = HF
2. Ability to tolerate ambiguity (MAC, GL)
3. Ability to validate HLA (MAC service, GL service, ARS service)
4. Assign accession numbers (POP_ID, COHORT_ID, HF_ID)
5. Manage access control and associate licenses with datasets
    a. Creative Commons-Non-Commercial
6. Out of Scope
    a. HF Estimation Methods: treat as a black box
    b. Controlled vocabularies for population attributes (AFND: HLA Net, IDAWG)

Goal One

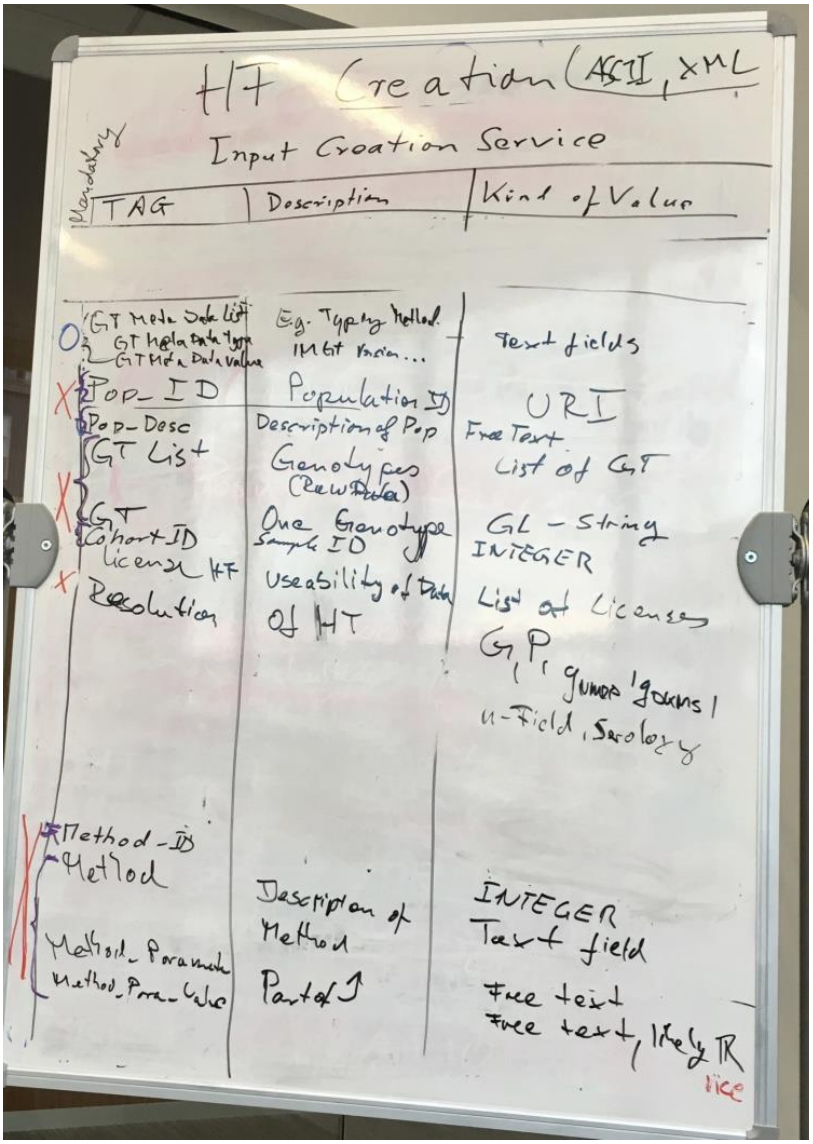
Establish the technical platform for assigning population-cohort-IDs and frequency-set-IDs following the pattern of gl-service, MAC service, and feature-service. To standardize the input and output file formats as well as required/optional parameters/variables.

Haplotype Frequency Creation Service

Mandatory	XML Tag	Values / Kind of Val.	Description
Yes	POP_ID	URI from POP DB	Population Identifiyer
Yes	GT_List	List of GTL	List of Genotypes
Yes	GTL	Record of GTL_Name, GTL_M_List	GTL_M_List is optional
Yes	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_Data	Free Text	
No	GTLP_M_Value	Free Text	
Yes	GT_Lic	License ID	License under which GT data is available
Yes	METHOD_LIST	List of METHOD	
Yes	METHOD	Record of METH_Type, METH_Value, METH_Comment, METH_Ref	
Yes	METH_Type	Free Text	
Yes	METH_Value	Free Text	
No	METH_Comment	Free Text	
No	METH_REF	Free Text	To be used to refer to an external source
No	HFCeS_COHORT_ID		As an alternative to GT List
No	HFCeS_METHOD_ID		As an alternative to Method List

Output

- An Inputfile to HFCuS
- HFCeS_METHOD_ID
- HFCeS_COHORT_ID



HT Creation (ASCII, XML

Input Creation Service

Mandatory	TAG	Description	Kind of Value
O	GT Meta Data List GT Meta Data Type GT Meta Data Value	Eg. Typing Method. IMGT Version...	Text fields
X	Pop_ID	Population ID	URI
	Pop_Desc	Description of Pop	Free Text
X	GT List	Genotypes (Raw Data)	List of GT
	GT	One Genotype	GL - String
	Cohort ID	Sample ID	INTEGER
	License HF	Useability of Data	List of Licenses
X	Resolution	Of HT	G, P, gumor 'gauns' u-Field, Serology
X	Method-ID		
	Method	Description of Method	INTEGER Text field
	Method_Parameter	Part of ↑	Free text
	Method_Para_Value		Free text, likely TR

tice

Haplotype Frequency 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 <= 1	
Yes	HT_Lic	License ID	License under which HF data is available
Yes	HT_Res	G, P, g_{NMDP}, g_{DKMS}, 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, NOT THE SAME IDs as used in the HFCeS
No	METHOD_ID		As an alternative to Method List, NOT THE SAME IDs as used in the HFCeS

Additional Stored Values:

- Timestamp
- Submiting UserID

Direct Output/Feedback:

- URI to dataset
- Method_ID
- Cohort_ID
- HF_ID

Internal Data Structure of HFCuS

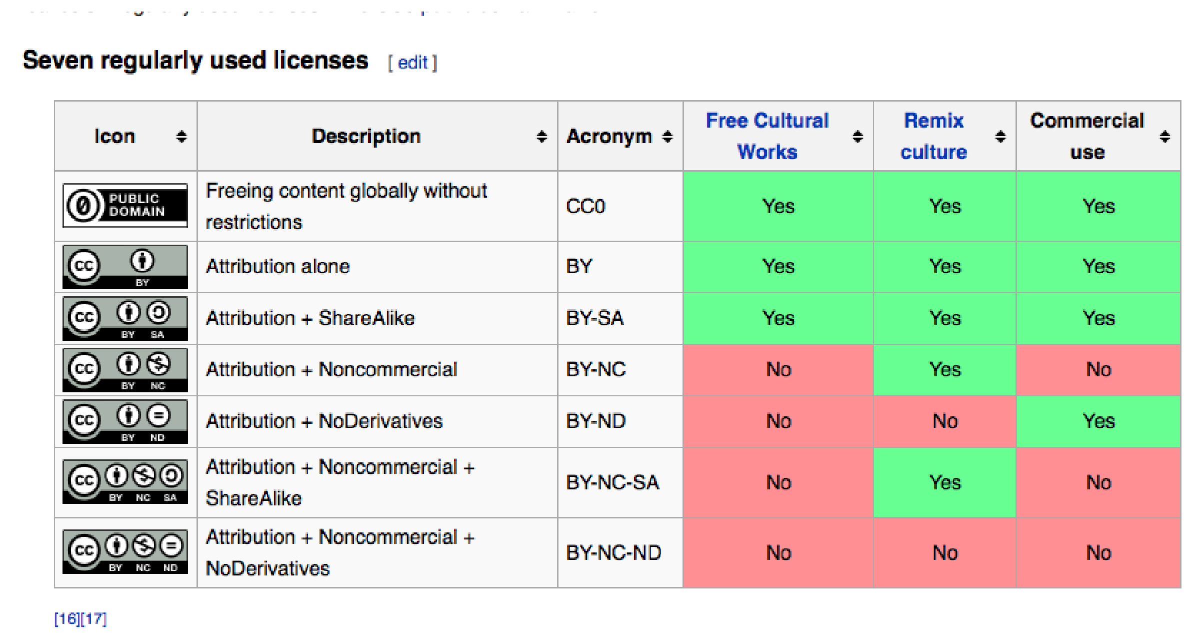
Basic Fields: The basic data structure of the HFCuS mimics the input file

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.

Madatory	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

Note: If the genotype list is available, GTs can be downloaded, HF resubmitted for the same Cohort_ID with a different (but 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
0 PUBLIC DOMAIN	Freeing content globally without restrictions	CC0	Yes	Yes	Yes
cc BY	Attribution alone	BY	Yes	Yes	Yes
cc BY SA	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes
cc BY NC	Attribution + Noncommercial	BY-NC	No	Yes	No
cc BY ND	Attribution + NoDerivatives	BY-ND	No	No	Yes
cc BY NC SA	Attribution + Noncommercial + ShareAlike	BY-NC-SA	No	Yes	No
cc BY NC ND	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	No	No	No

[16][17]

taken from: `https://en.wikipedia.org/wiki/Creative_Commons_license`

Note: User and Group Models are to be defined

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 Genotype 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 dsitribution
DIV_50	integer	Number of haplotypes needed (in descending order of frequency) to have the cumulative sum be > 0.5

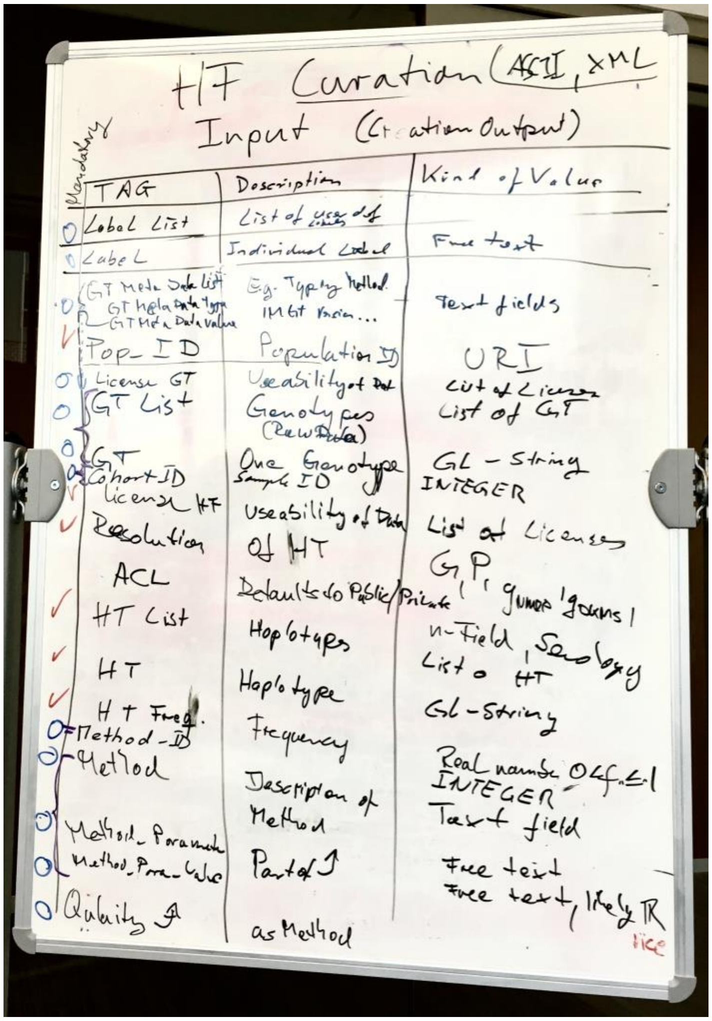
DIV_50_REL	Real, 0 <= x <= 1	Number of haplotypes needed (in descending order of frequency) to have the cumulative sum be > 0.5 divided by the number of HT
SAM_SIZE	integer	Number of GT
SAM_POP	integer	Size of Population (approx.)
DIV_PGD	Real, 0 <= x <= 1	Population genetics diversity (1-sum f_i ^2 N/(N-1))
DIV_HEAVY_TAIL	a	Martin knows that
RES_TRS_COUNT	Real, 0 <= x <= 1	Jan knows that
RES_TRS	Real, 0 <= x <= 1	Resolution score
RES_SHARE_AMBIG	Real, 0 <= x <= 1	Fraction of GT with a lower resolution than definied in the resolution tag
RES_MISS_LOCI	Real, 0 <= x <= 1	Fraction of GT with missing loci
DEV_HWE	real	Devition from HWE, method described in the method section
ERR_STD	Real, 0 <= x <= 1	Weighted average of standard error
ERR_SAMP_80_100	real	Laurent, Excoffier
SUM_FREQ_GAP	Real	Expected but unobserved, LD!
ERR_OFFSET	Real, 0 <= x <= 1	1-sum f_i
LD_MEASURE	real	Define in Method section
KFOLD_IMPUTE	real	% of imputable GT from HT
KFOLD_PRED_ACTUAL	Real	Divergence between prediction and actual
KFOLD_N	integer	Number of independent iterations

List of Labels

LABEL_CLASS	HH2016	
LABEL_TYPE	**VALUE**	**DESCRIPTION**
GT_REGISTRY	String	ION or other description of the entity hosting the GT
HT_ESTIMATION_ENT	String	ION or other description of the entity performed the HTF analysis

Open Issues for HFCuS

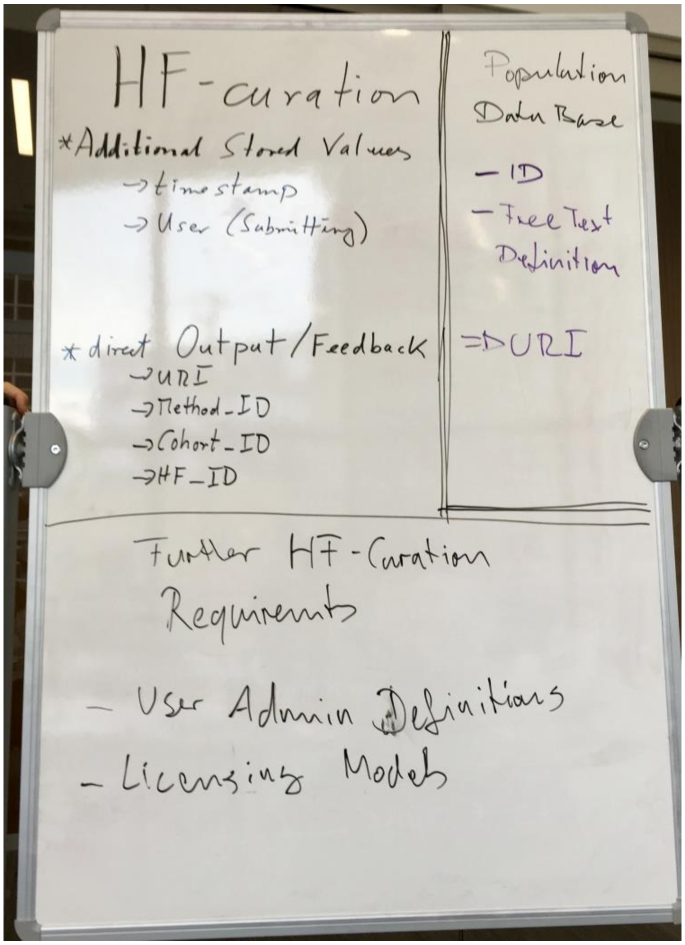
- User/Group Management, Schema, Specs
- ACL: Management and Specs
- Governance structure
- Implementation



HT Curation (ASCII, XML

Input (Creation Output)

Mandatory	TAG	Description	Kind of Value
O	Label List	List of user def labels	
O	Label	Individual Label	Free text
O	GT Meta Data list GT Meta Data Type GT Meta Data Value	Eg. Typing Method IMGT Version...	Text fields
✓	Pop_ID	Population ID	URI
O	License GT	Useability of Dat	List of Licenses
O	GT List	Genotypes (Raw Data)	List of GT
O	GT	One Genotype	GL - String
O	Cohort ID	Sample ID	INTEGER
✓	license HT	Useability of Data	List of Licenses
✓	Resolution	Of HT	G, P, ' ' 'gunos' 'gauns'
	ACL	Defaults to Public/Private	
✓	HT List	Haplotypes	n-Field, Serology
✓	HT	Haplotype	List of HT
✓	HT Freq.	Frequency	GL-String
O	Method-ID		Real number 0<f≤1
O	Method	Description of Method	INTEGER
O	Method_Parameter		Text field
O	Method_Para_Value	Part of ↑	Free text
O	Quality ↗	as Method	Free text, likely TR



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

Goal Two

To develop a strategy for standardizing/comparing methods (building on previous work). Establish methods for evaluating the quality of frequency sets, and computation methods for improving frequency sets.

Goal Two Notes

Quality Measurement Options:

- % of non-imputable donors in test set - Lower is better
- "If" between frequencies derived from 100% set and 80% training set - Lower is better
- Each haplotype estimate has a standard error. Weighted sum of errors over all haplotypes of (SE * HF) - Lower is better
- Use real patient typings and perform searches using matching algorithm on registry donors with their typing rolled back to lower resolution (A~B~DRB1 222) or pre-CT.
- Sum over the top 10 potentially-matched donors in search distance between match probabilities and truth.
- Could also simulate patients by drawing from haplotype frequencies.
- Ratio of alleles and haplotypes in 100% set versus 80% set - Lower is better
- HWE of population
- Overall number of donors used to generate frequencies
- Ratio of number of donors used to generate frequencies to number of donors to be imputed using those freqs (and fraction of excluded donors that can be explained with HFE)
- Typing quality of samples (typing resolution score, number of loci typed, typing resolution)

Advanced Quality Metrics (best to use code developed by Yoram Louzoun):

- Projecting how many alleles / haplotypes exist in general population versus how many are observed in cohort.
- SHF-based metric of sum of freqs of new haplotypes that need to be added to the distribution (alpha, beta).

QM Service Inputs:

- Imputation Output File
- Donor Typings (High Res)
- Quality Metric Name
- For patient-donor matching metrics, also need Patient typing (HR) and match probabilities instead of donor imputation output.

Procedure to Generate and Validate BMDW Haplotype Frequency Datasets

1. Population / Cohort Selection Step: Choose how to allocate donors to populations / cohorts. Randomly split this cohort of HLA typings into 80% training set / 20% test set.
2. Haplotype Frequency Estimation Step: - Calculate high resolution haplotype frequencies on training set. Presumably using EM.
3. Typing Simulation Step: Roll back high resolution data within test set back to low resolution and mask certain loci by simulation.
   a. All A~B~DRB1 222
   b. Other resolution levels: Capable of doing serology, 2-digit DNA, SSO, SBT and any combo of missing loci. Would some pops do better than others at 222, but worse at other levels?

4. Imputation / Matching Prediction Step: Impute back high resolution from rolled-back lower resolution test set.

5. Quality Measurement Step: Determine what percentage of donors are not imputable without fallback / SHF methods. Calculate predicted vs actual for match probabilities for patient-donor pairs.

6. Repeat steps #1-5 to get different splits and do K-fold cross-validation to get distribution for each quality metric.

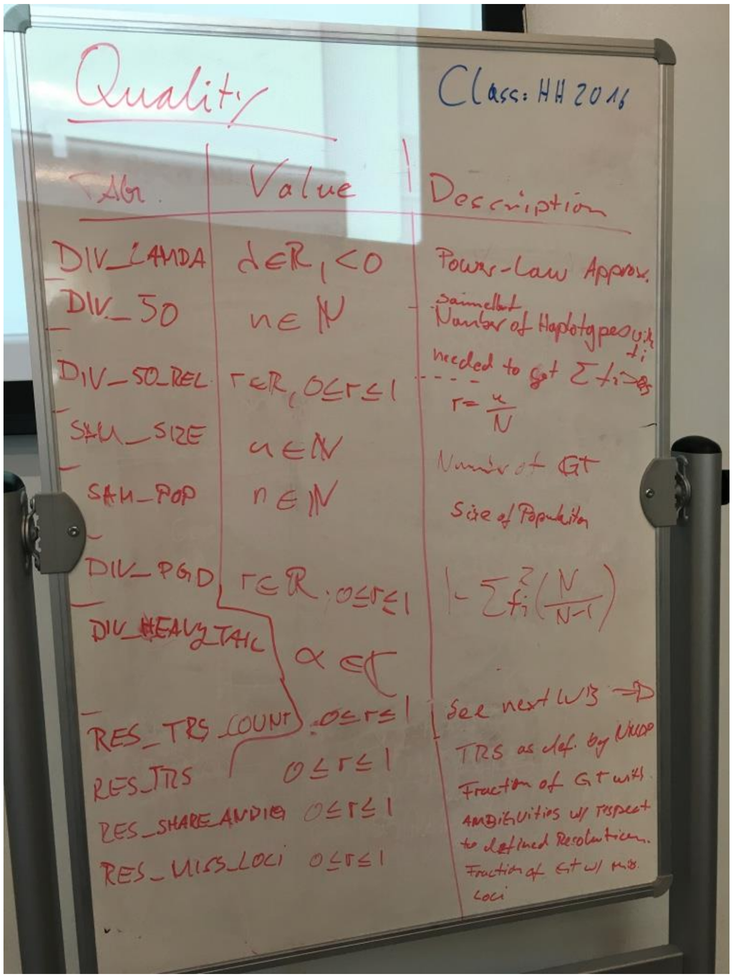
A diversity of options exist for each step. Goal will be to find the combination of options that work best. Operationally for the final frequency sets, we’ll use 100% of available data to generate frequencies.

Comparing Quality Metrics Between Populations:

- Differences in population diversity are accounted for - Results will show that more donors for diverse populations will be needed to reach same quality metrics

[*Above: Notes provided by Loren Gragert*]

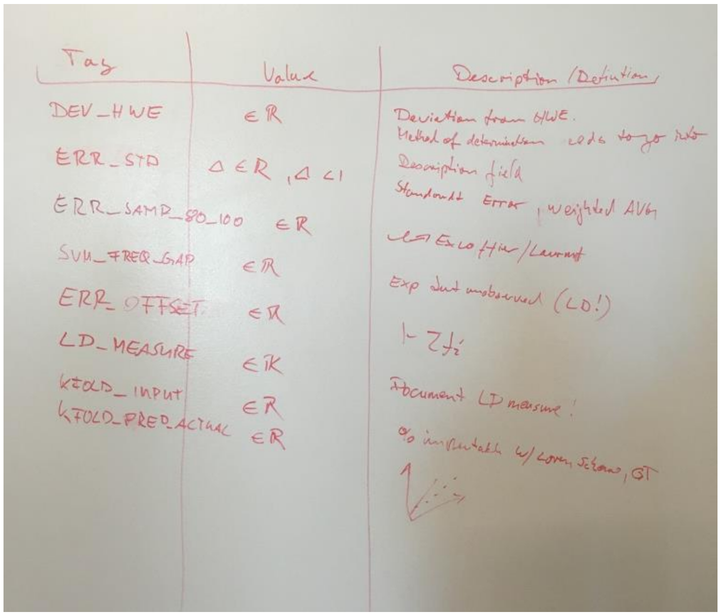
Note: Pictures below depict possible data structure for quality metric that can be persisted by the frequency curation service.



Quality

Class: HH 2016

Tag	Value	Description
DIV_LAMDA	$\lambda \in \mathbb{R}, < 0$	Power-Law Approx.
DIV_50	$n \in \mathbb{N}$	Saimellot Number of Haplotypes with f_i needed to get $\sum f_i \geq 50$
DIV_50_REL	$r \in \mathbb{R}, 0 \leq r \leq 1$	$r = \frac{n}{N}$
SAM_SIZE	$n \in \mathbb{N}$	Number of GT
SAM_POP	$n \in \mathbb{N}$	Size of Population
DIV_PGD	$r \in \mathbb{R}, 0 \leq r \leq 1$	$1 - \sum f_i^2 \left(\frac{N}{N-1}\right)$
DIV_HEAVY_TAIL	$\alpha \in \mathbb{C}$	
RES_TRS_COUNT	$0 \leq r \leq 1$	See next WB →
RES_TRS	$0 \leq r \leq 1$	TRS as def. by NMDP
RES_SHARE_AMBIG	$0 \leq r \leq 1$	Fraction of GT with AMBIGUITIES w/ respect to defined Resolution.
RES_MISS_LOCI	$0 \leq r \leq 1$	Fraction of GT w/ mis. Loci



Tag	Value	Description / Definition
DEV_HWE	$\in \mathbb{R}$	Deviation from HWE. Method of determination needs to go into Description field
ERR_STD	$\Delta \in \mathbb{R}, \Delta < 1$	Standard Error, weighted AVG
ERR_SAMP_80_100	$\in \mathbb{R}$	↳ Excel file/Laurent
SUM_FREQ_GAP	$\in \mathbb{R}$	Exp but unobserved (LD!)
ERR_OFFSET	$\in \mathbb{R}$	$1 - \sum f_i$
LD_MEASURE	$\in \mathbb{R}$	Document LD measure!
KFOLD_INPUT	$\in \mathbb{R}$	% imputable w/ common scheme, GT
KFOLD_PRED_ACTUAL	$\in \mathbb{R}$	

Possible Quality Related Services

Typing Simulation (TS) Service Inputs:

- Donor Typings (High Res HLA)
- Resolution Level

TS Service Outputs:

- Donor Typings (Low Res HLA)

Imputation (IMP) Service Inputs:

- Frequency File (multiple for multi-race?)
- Donor Typings
- Patient Typings (for cases where we perform actual searches and patient-donor match probabilities)

IMP Service Outputs:

- List of high resolution haplotype pairs and probabilities per donor

Goal Three

Identify populations where additional high-resolution typing is needed, set priorities and target cohort sizes algorithmically.

Goal Three Notes

Haplotype Frequency Estimation (HFE) of Bone Marrow Donors Worldwide (BMDW) 2.0 Data

- BMDW 2.0 data contains no ethnicity/population information
- Therefore, HFE can only be based on registry information
- HLA data quality (resolution, number of typed donors) with regard to HLA-A, -B, -C, -DRB1, -DQB1 (5 locus) Haplotype Frequency is heterogeneous

Results

- 32 haplotype frequency sets, including two region sets (South America (sam) and Eastern Europe (eeu) and a global BMDW consensus haplotype frequency.
- Assessed by the number of phemotypes explicable with a haplotype frequency set.
- Introduced two levels of substitution haplotype frequency

[*Above: Presented by Hans-Peter Eberhard*]

Prioritizing High-Resolution Typing:

- Decide on quality metric to sort populations HFE's (composite score of multiple quality metrics that has linear relationship with quality - values from 0 to 1)
- Weight quality metric by number of searching patients from population (cost-effectiveness scales linearly with this).
- More cost effective to recruit new donors and type them well, or to upgrade typing of existing donors? If we chose to upgrade typings, would we do it random or select "high value" donors (unique HLA typing, more likely to be selected, and/or more informative for defining HF distribution)?

How To Improve Frequency Datasets:

- Run all combinations of developed methods for cohort selection, haplotype frequency estimation, and imputation to search entire space. Lots of experiments. Develop new methods.
- Improve input data by upgrading typing or recruiting new donors.

Population / Cohort Selection Options:

- What are the rules for including a donor in one population rather than another?
- How are relationships between populations defined (hierarchical, multi-combinations)?
- What are the minimum requirements for quality of input HLA typing for a population? Does including low resolution typing make things worse? Number of donors that have certain loci typed and when imputed the most likely genotype is above a certain threshold. For ZKRD, they had minimum of 5,000 donors with 90% imputation threshold, and A, B, C, DRB1 DNA-based typing.
- How does the model for Registry Codes or BMDW Ethnic Codes map to Population IDs / Cohort ID used for generating frequencies or predictions?
- Which BMDW countries should be combined into one cohort (e.g China and Taiwan)? Which registries in the same country should be separately (USA1 NMDP versus USA4 Gift Of Life)?
- Registry IDs, ISO country code for combining Registries in same country together, BMDW Ethnic Groupings are all defined categories where the mapping are known. This is how we start.

Announcements

[Announcements]

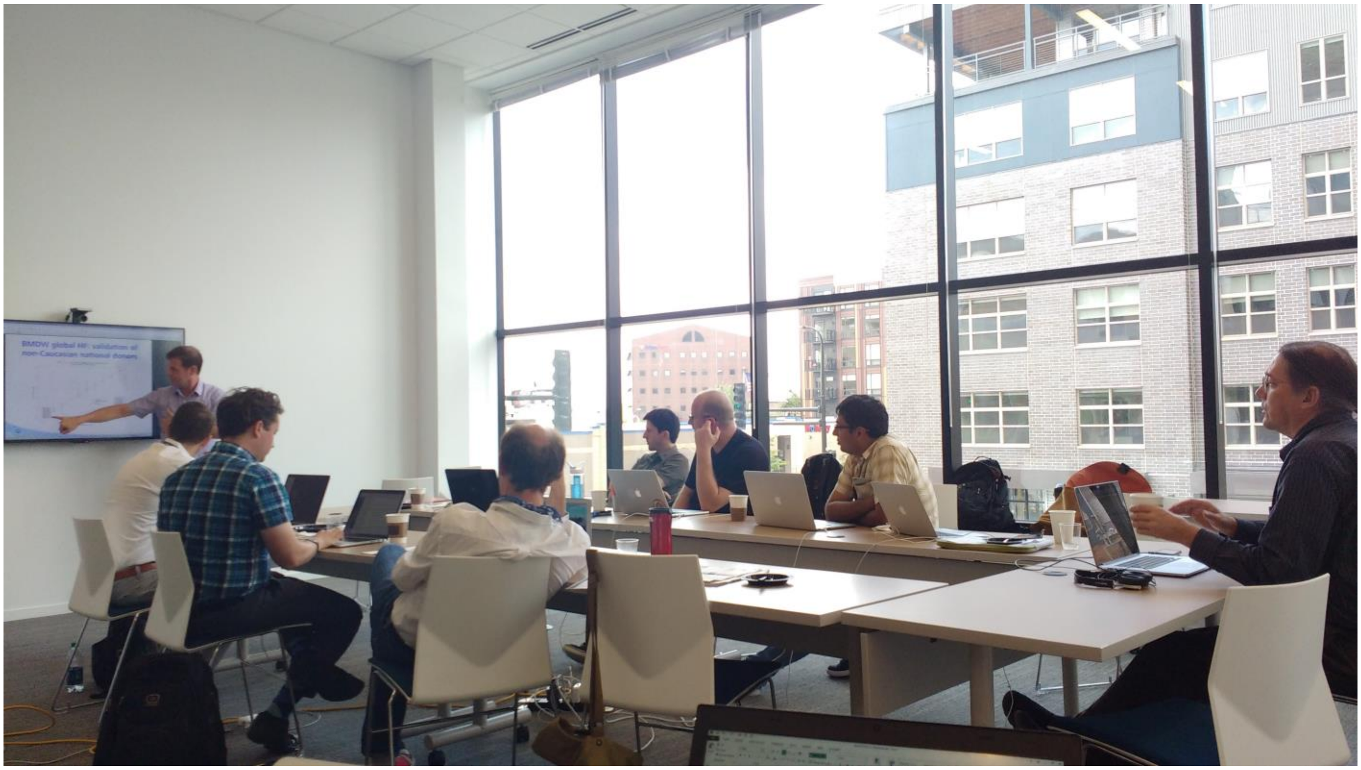
Next Meetings

First Haplotype Hackathon Follow-Up Teleconference:

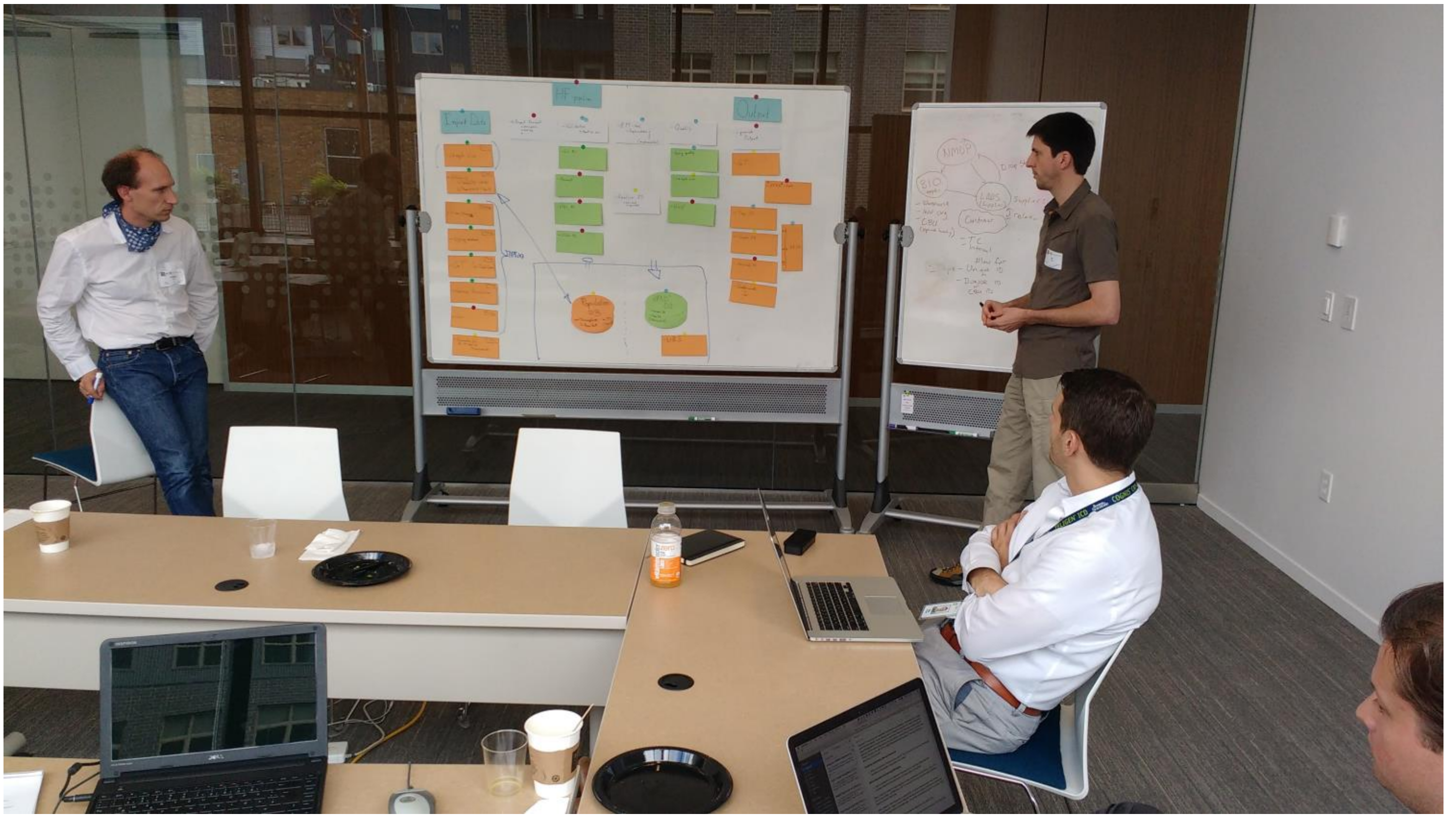
Second Haplotype Hackathon Follow-Up Teleconference:

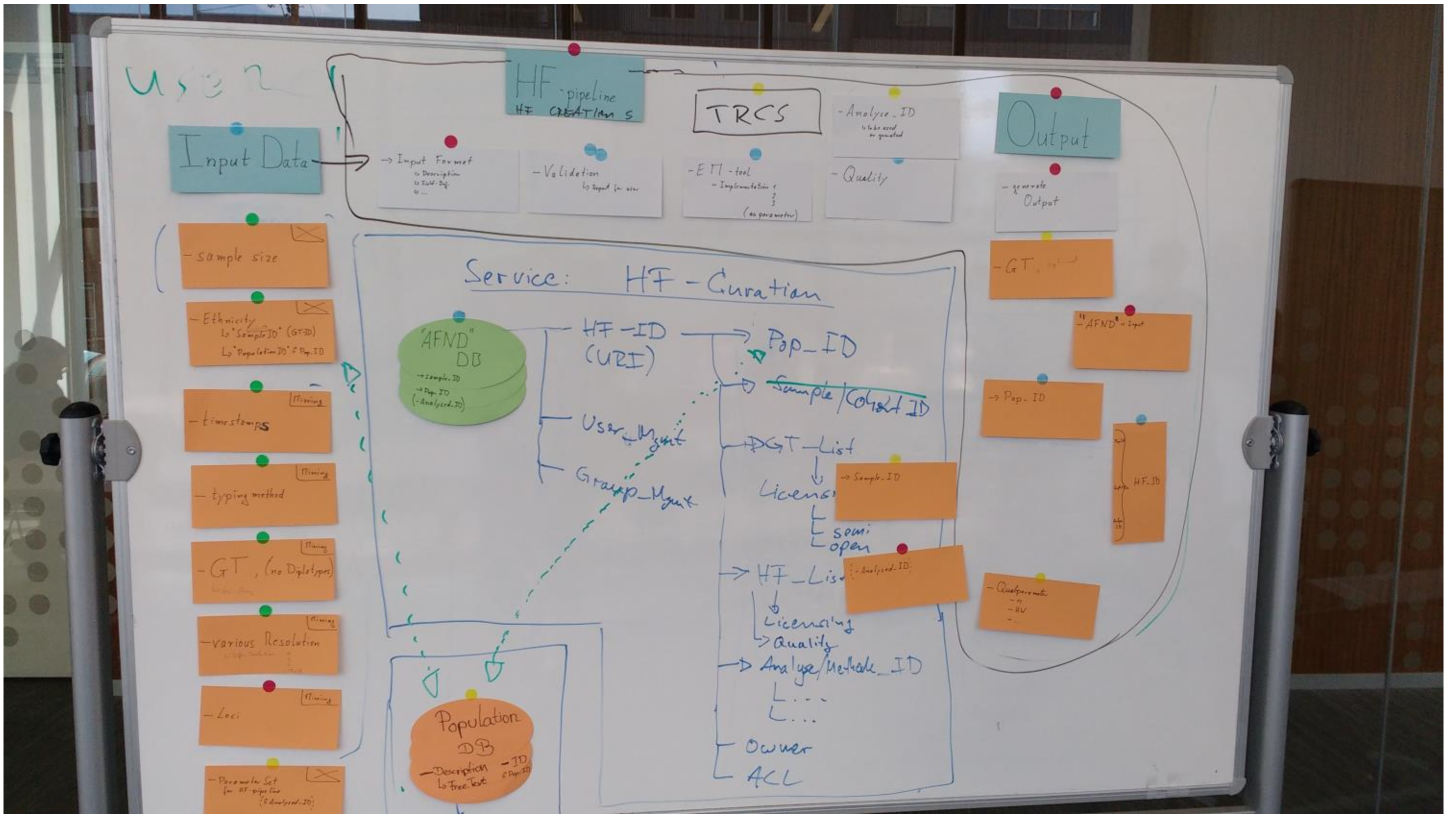
Photos

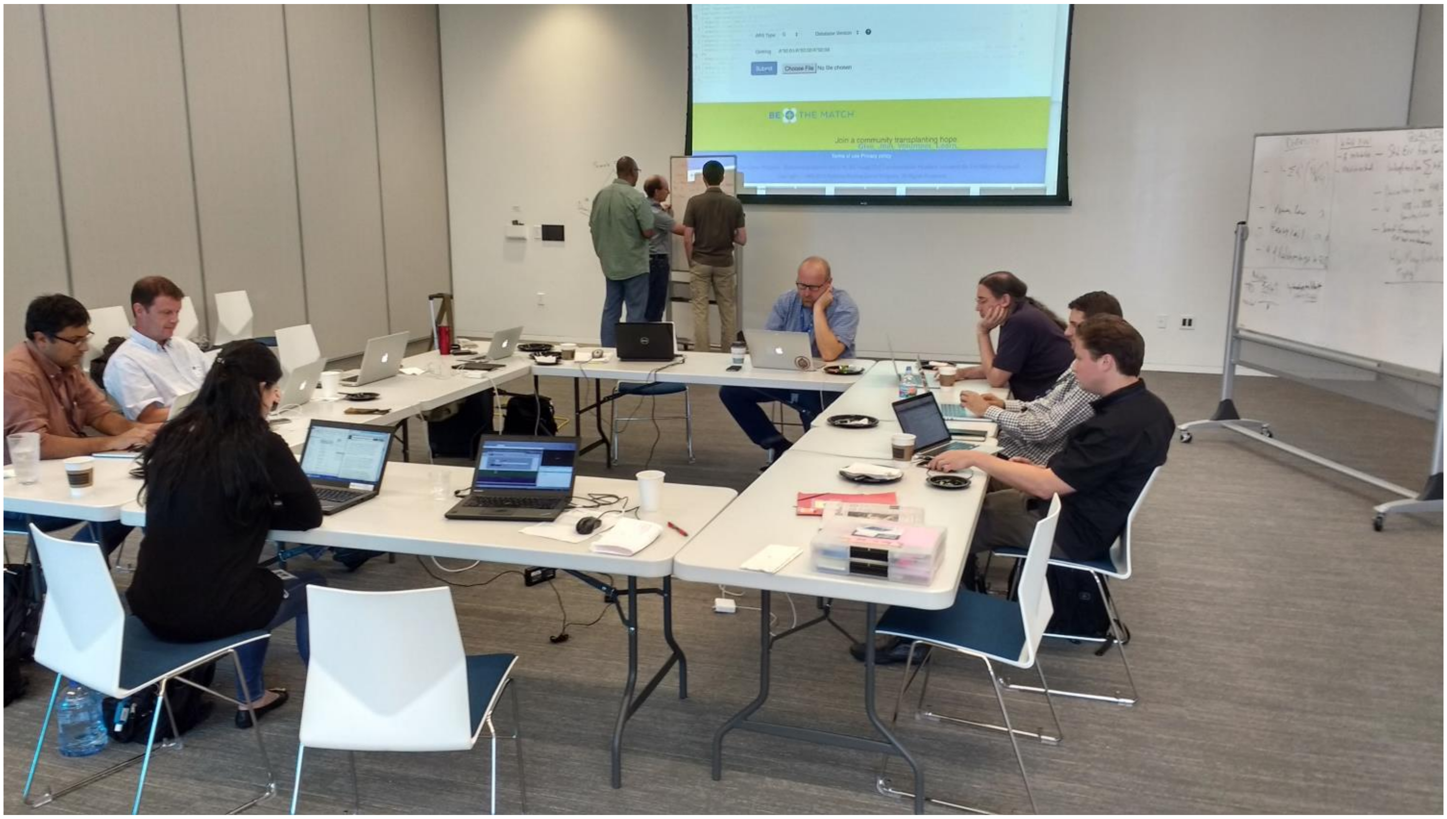
validation of
tional donors

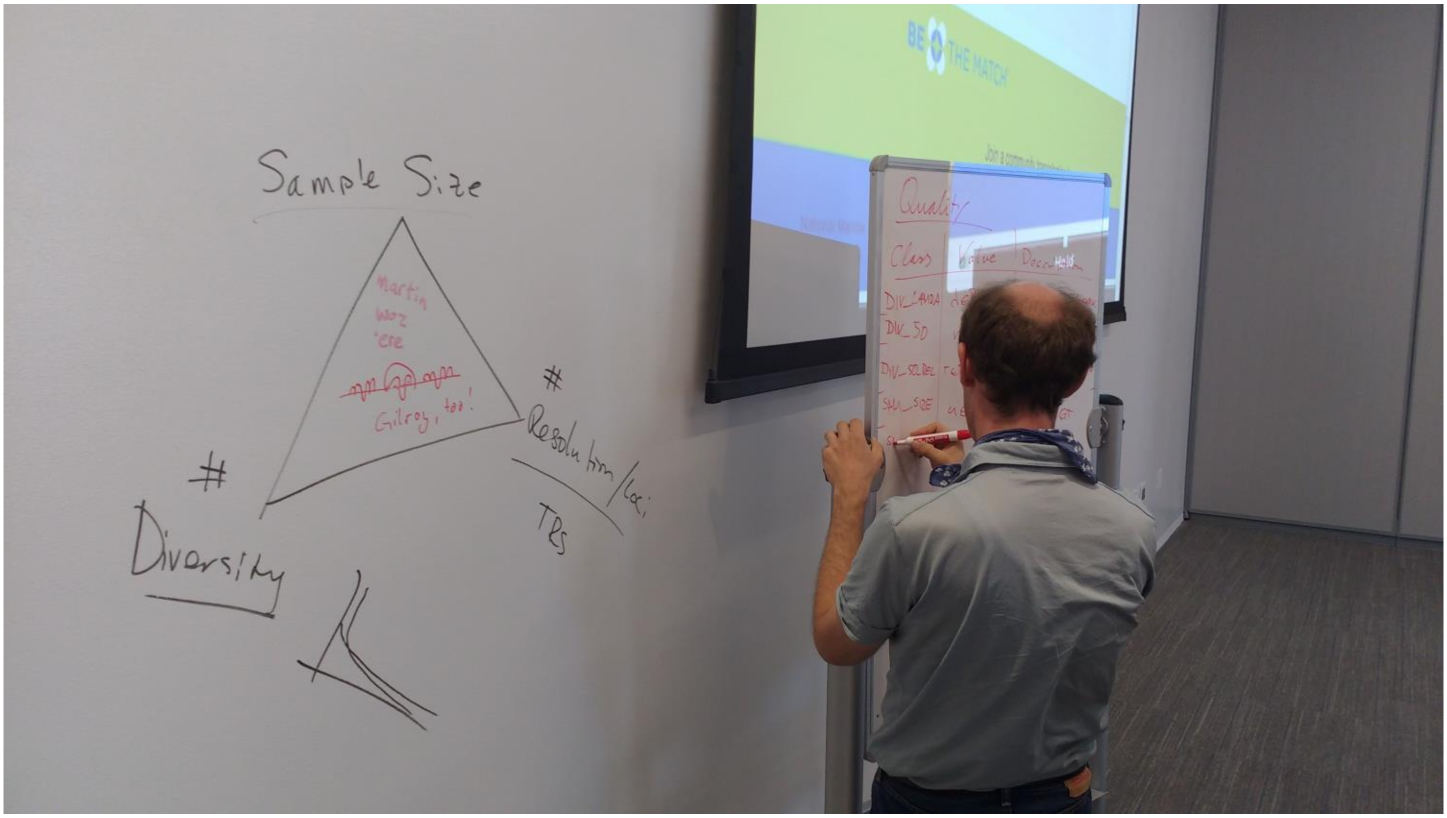

BMDW global HF: validation of
non-Caucasian national donors

HF pipeline
TRCS
Output
Input Data
Quality
GL
MAC
URI
Population DB
HF-Curation-Service
List of needed SERVICES
– GL
– MAC
– Population Service
– Typing resolution conversion service "TRS-Service"
– HF CURATION DB Service


HF pipeline
Input Data
Output
Population DB
NMDP


User
HF-pipeline
HF CREATIONS
TRCS
Input Data
-> Input Format
Description
Field-Def.
- Validation
Report for user
- ETI - tool
- Implementation 1
2
3
(as parameter)
- Analyse_ID
to be used or generated
- Quality
Output
- generate Output
- sample size
- Ethnicity
"SampleID" (GT-ID)
"PopulationID" & Pop.ID
Missing
- timestamps
- typing method
- GT, (no Diplotypes)
- various Resolution
- Loci
- Parameter Set for HF-pipeline (& Analysed_ID)
Service: HF-Curation
"AFND" DB
-> Sample_ID
-> Pop_ID
(-Analysed_ID)
HF-ID (URI)
User_Mgmt
Group_Mgmt
Pop_ID
~~Sample~~/Cohort ID
DGT-List
Licensing
semi
open
HF_List
Licensing
Quality
Analyse/Methode_ID
Owner
ACL
-> Sample_ID
- Analysed_ID
Population DB
- Description
Free Text
- ID
& Pop-ID
- GT
"AFND" -> Input
-> Pop_ID
HF_ID
- Qualparameter


BE THE MATCH
Join a community transplanting hope


Sample Size
Martin woz 'ere
Gilroy, too!
#
Resolution/Loci
TRs
#
Diversity
BE THE MATCH
Quality
Class
Value
Document
DIV_LANDA
DIV_SD
DIV_SZLBL
SMA_SIZE