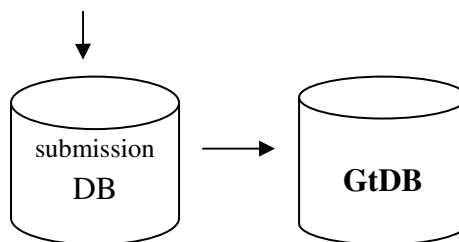


GtDB DATA FORMATS

- 1. Data Formats2
 - 1.1 Genotype Data.....2
 - 1.2 Pedigree Data3
 - 1.3 Map Data.....4
- 2. File format.....5
- 3. File transfer5
- 4. Codes for change and deletion reasons6

Genotyping centers:

Laboratory	Country code	Lab code	Genotypes	Instrument	Data Handling
FGC	246	03	Microsatellite	MegaBACE1000	SQL*LIMS
FGC	246	03	SNPs	Sequenom 7K	SMdb
KTL	246	02	Microsatellite	ABI	SMdb
KTL	246	02	SNPs		BioDataBase
UPP-RUDBECK	752	03	Microsatellite	ABI	MS Excel / Genotyper
UPP-MOLMED	752	02	SNPs	Orchid	SQLServer
UK	826	99	Microsatellite	ABI	
Australia	036	99	Microsatellite	ABI	



1. Data Formats

1.1 Genotype Data

New Results and Updates

Column	Description	e.g.
GENOTYPE ID	id number that identifies the result row in the database or unique running number in file (if no unique id available)	LIMS:301320659 12
PROJECT ID	local code for subproject under GenomEUtwin or N/A if not available	NIKO N/A
EUIDNUM	EUID number	208100361715
CSAMPLEID	unique local sample identifier as used in the lab	NIKO.123456
MARKER ID	marker identifier of polymorphic site. DS or RS-number	DS12345
ASSAY ID	PCR assay as used in lab dot marker set name and version if modified For one marker there can be different pcr assays with different PCR primers	DS12345.LMS2.2
ALLELE 1,2	microsat: binned allele result snp: genotype result in nucleotide (A, C, T, G and I, D)	15 A
ALLELE SIZE 1,2	allele size in bp (only in microsatellites)	215.32
INSTRUMENT	used instrument	MegaBACE1000
QUALITY DESCRIPTION	qc field what the instrument gives out for the measurement	2.8
METHOD	'snp' or 'str'	str
RUN DATE	run date YYYYMMDD - Year, month, day	20030115
TIME STAMP	time stamp when data is inserted or modified in database. YYYYMMDD:hhmmss	20030129:150211
MENDEL ERROR	mender error 0 = no mendel error, 1 = mendel error, 9 = unknown	9
[REPLACE] ¹⁾	GENOTYPE ID of previous result	
[CHANGE REASON CODE] ¹⁾	reason code used to update or reject the data (see. ch 4)	2

¹⁾ for update only

Rejection

For rejection only the GENOTYPE ID and reason code are needed.

Genotype ID

I) ID if genotypes are maintained in local database and they have unique identifiers

LocalNameOfDatabase:internalUniqueIdentifier e.g. LIMS:301320659

e.g. SEU:15943842

II) ID if data is stored only in flat files and do not have unique/maintained identifier (not preferred).

row_number e.g. 512

1.2 Pedigree Data

Can be used as in *Instructions to send samples and information for genotyping Data Form2* or following:

Column	Description	e.g.
EUIDNUM	EUid number	208100361715
GENDER	sex of person 1= male, 2= female	2
FAMILY ID ¹⁾	family identification	PED.1355
FID ^{1) 2)}	id of father	208100361712
MID ^{1) 2)}	id of mother	208100361713
ZYG ³⁾	zygosity 1 = MZ, 2 = DZ, 9 = unknown	2
BIRTHDATE ⁴⁾	date of birth (YYYYMMDD)	19211012

¹⁾ Only for family data

²⁾ Use EUid numbers, if parents have any phenotypic data available

³⁾ Only for twins

⁴⁾ If the Date of Birth information is incomplete, you can use one of the following ways

1) As in *Data Format and Variable Standard for GenomEUtwin's Phenotype Database prototype*

2) Two separate fields.(Birth month, Birth year)

1.3 Map Data

Map and marker data is coordinated centrally in GenomEUtwin project.

Microsatellite

Column	Description	e.g.
MAP ID	map name	LMS2
CHROMOSOME	chromosome	1
MARKER ID	DS-number of marker	DS12345
PHYSICAL LOCATION	marker location in bp (start of forward primer)	75 560
SEQUENCY ASSEMBLY	origin and version	Ensembl 19.34a.1

SNP (*incomplete*)

Column	Description	e.g.
MAP ID	map name	
CHROMOSOME	chromosome	1
MARKER ID	marker name	
MARKER NAME	marker name used in lab	
RS NUMBER	RS number of marker	
PHYSICAL LOCATION		
SEQUENCY ASSEMBLY	origin and version	Ensembl 19.34a.1

- possible alleles for the marker
- the way marker is read, Kaisa: forward (is this part of assay data?)

2. File format

All data columns are case sensitive strings.

Data is prepared in ASCII comma-delimited format using semicolon (;) as delimiter with variable names in a first row. If a value contains text with spaces or semicolons(;), then all text is surrounded by double quotes (“”).

3. File transfer

To simplify tracing of data transfers between GenomEUtwin centres it is recommended to name data files as follows:

GT_country_centre_typeoperation_YYYYMMDD_NN

GT	identification of GenomEUtwin project
Country	value of COUNTRY variable, as it appears in data file
Centre	value of CENTRE variable, as it appears in data file
Type	G = genotype result data, P = pedigree data, M = map data
Operation	insert (I), update (U) or delete (D)
YYYYMMDD	year, month and date when file was created
NN	sequential number of transfers

e.g. GT_246_03_GI_20030828_01 *new genotype results*

e.g. GT_246_03_GU_20030828_01 *updated genotype results*

e.g. GT_246_03_GD_20030828_01 *genotype rejections*

e.g. GT_246_03_PI_20030828_01 *new pedigree information*

4. Codes for change and deletion reasons

Code	Description
1	BINNING ADJUSTMENT
2	CORRECTION BASED ON NEW DATA
3	DOUPLE RECOMBINATION
4	ERROR IN LABORATORY WORKING PROCEDURES
5	GENOTYPING ERROR
6	NON-MENDELIAN INHERITANCE
7	RESULTS NOT CORRECT IN PREVIOUS FILE
8	TYPING ERROR
9	UNRELIABLE RESULT
10	WRONG PLATE CONTROL VALUES USED
11	OTHER ERROR

Report new needed change reasons to Juha Muilu (juha.muilu@helsinki.fi).