A data format description for genomic breeding evaluation

This document describes the required file format for exchanging data with the National Research Institute of Animal Production.

Type and format of files required for data exchange

Type of files required for data exchange:

1. The file is containing genotype information.
2. The file is containing pedigree information.

Lab ISO17025 requirements

In case that the Parentage verification service is obligatory in data verification processes, please be aware that from 2022 onwards for ICAR accreditation to be granted to laboratories carrying out SNP genotyping ISO17025 accreditation will be a mandatory minimum requirement as per guidelines. https://www.icar.org/index.php/certifications/certification-and-accreditation-of-dna-genetic-laboratories/guidelines-for-str-and-snp-based-parentage-testing-in-cattle/

Data compression

To speed up data transfer, as well as to save storage space, the genotype file should be compressed with xz format. Compression of xz file can be performed with software available at http://www.7-zip.org/ for Windows and at http://tukaani.org/xz/ for Linux. If using xz is not possible, gzip compression should be performed.

1. Genotype file

Data from laboratories should be delivered in *Illumina* FinalReport format. Each column in the data set should be tab-delimited:

1. **SNP-name (according to *Illumina* nomenclature) - SNP Name,**
2. **Short number of individual (14 characters) or International ID of ANIMAL (Interbull ID – 19 characters) - Sample ID,**
3. **Allele1 (in A/B format) - Allele1 - AB,**
4. **Allele2 (in A/B format) - Allele2 - AB,**
5. (optional) – GC statistic result - GC Score,
6. (optional) – GT statistic result - GT Score,

Example:

```
[Header]
GSGT Version     1.9.4
Processing Date  33/01/2020 1:01 PM
Content          BovineSNP50_v3_A1.bpm
```

Genomika Polska

```
Num SNPs    53218
Total SNPs  53218
Num Samples 1526
Total Samples     1526
[data]
SNP Name    Sample ID   Allele1 - AB      Allele2 - AB      GC Score    GT
Score
ARS-BFGL-BAC-10172      PL000123456789    A     B     0.9140      0.8767
ARS-BFGL-BAC-1020 PL000123456789    A     B     0.9288      0.8919
ARS-BFGL-BAC-10245      PL000123456789    B     B     0.7227      0.7447
ARS-BFGL-BAC-10345      PL000123456789    -     -     0     0
…
```


Remarks:

- unknown genotypes of a SNPs should be marked as „-" according to default settings of Illumina GenomeStudio or Beeline,
- data exchanged should come from analysis with a call rate of at least 0.95 per individual
- for each type of separate microarray, file should be created
- animals genotyped using one kind of microarray should be saved into one file
- there is a possibility to replace data in A/B format with a TOP format
- the additional column can be sent in a file with genotypes data according to principles of file generation in Illumina GenomeStudio or Beeline,
- please report every exception from the remarks mentioned above before the data transfer!

The naming of the files:

The exchanged file should be named according to the following nomenclature criteria: YYYYMMDD_ NameOfChip_Final_Report.txt.xz

Where:

- YYYYMMDD – file preparation date: year (YYYY), month (MM) and day (DD)
- NameOfChip – the type of microarray used for genotyping. Possible values are: *BovineSNP50v1, BovineSNP50v2, BovineSNP50v3, EuroG10Kv3, EuroG10Kv4, EuroG10Kv5, EuroG10Kv6, EuroG10Kv7, EuroG10Kv8, EuroG10Kv8b, GeneSeekHDv1, GeneSeekHDv1, GeneSeekHDv3, HD, LDv1, LDv1.1, LDv2, ZoetisMD2, GeneSeekGGPv2, GeneSeekGGPv3, GeneSeekGGPv4, ZoetisMD2, Neogen GGP100Kv1, EuroGMDv1, EuroGMDv2*
- Final_Report – permanent unit corresponding to the type of prepared file whether in Illumina GenomeStudio or Beeline
- extension – csv.xz representing xz compressed text file.

*Microarrays not used in Poland are marked with grey colour.

Example:

20140310_EuroG10Kv3_Final_Report.txt

- indicates data prepared on 10th March 2014 with EuroG10Kv3 microarray

Example:

20170829_BovineSNP50v3_Final_Report.txt

- indicates data prepared on 29th August 2017 with Illumina BovineSNP50 version v3

2. Pedigree information

Pedigree information can be sent in two formats:

2.1. According to standard Interbull – in format file200 (**Table 1**).

Table 1. Pedigree file format in file200 type.

Position of the first character in the line	Description	Data type and number of characters	Example
1	Type of file	character 3	200
International ID of ANIMAL			
5	Animal breed[1]	character 3	HOL
8	Animal first registration country[2]	character 3	CAN
11	Sex[3]	character 1	M
12	Animal ID[4]	character 12	000000A12345
International ID of Sire of ANIMAL			
25	Sire breed	character 3	HOL
28	Country of first registration of sire[2]	character 3	CAN
31	Sex	character 1	M
32	Sire ID[4]	character 12	556912367589
International ID of Dam of ANIMAL			
45	Dam breed	character 3	HOL
48	Country of first registration of dam[2]	character 3	CAN
51	Sex	character 1	F
52	Dam ID[4]	character 12	123569874521
65	Animal birthdate (YYYYMMDD)	integer 8	19870215
74	Status of animal[5]	integer 2	10
77	Birthdate of first AI daughters (YYYYMMDD)	integer 8	19890314
86	Name of the animal	character 30	Cantarello
The national ID of the animal			
117	Animal breed[1]	character 3	HOL
120	Country of first registration of the animal[2]	character 3	CAN
123	Sex[3]	character 1	M
124	Animal ID [4]	character 12	000000A12345
137	Country of data exchange	character 3	CAN

*1 HOL= Holstein Friesian breed (HF) for black and white and red and white type; BSW= Brown Swiss; JER= Jersey; SIM= Simental; RED= national red breeds.

*2 Country, where the animal was first registered, described according to ISO alpha-3 country designation table (for example POL for Poland, DEU for Germany etc.)

*3 Sex M –sire, bull, F – cow, dam, heifer
*4 Unique, 12-digit number of animal, all shorter names (below 12 characters) should be completed with „zero" starting from the beginning of ID (from the left side)
*5 Bull status: 00 – unknown, also used for cows, 10 – AI bull, 20 – other bull

- Data should be provided as a text file (*.txt)
- Lack of sire or dam number (ID) should be marked as UUUUUUUUUUUUUUUUUUU (19xU)
- Lack of birth date (YYYYMMDD) of the animal should be marked as 00000000 (8x0)
- Lack of the birth date of first daughter AI (YYYYMMDD) should be marked as 00000000 (8x0)
- Gaps in the bull status should be marked as 00 (2x0)
- Total record line should contain 140 characters including spaces

Example:

```
200_HOLPOLM005247600495_HOLCANM000007816429_UUUUUUUUUUUUUUUUUUU_20120321_00
_00000000_ORINOKO_____________________HOLPOLM005247600495_POL
200_HOLPOLM005250595474_UUUUUUUUUUUUUUUUUUU_HOLPOLF005154309641_20120308_00
_00000000_____________________________HOLPOLM005250595474_POL
200_HOLPOLM005255079702_HOLUSAM000062169176_HOLPOLF005147728718_20120805_00
_00000000_MOROZ_______________________HOLPOLM005255079702_POL
…
```

*spaces marked as _

INFORMATION OF PEDIGREE IN FILE200 FORMAT SHOULD BE PREPARED ACCORDING TO RULE AS DEEP AS POSSIBLE, for both sire and dam side.

2.2. The data format of animals identification (**Table 2**).

Table 2. The data format of animals sent to microarray genotype analysis.

Columns	number of characters	Format	Remarks
Sample number	20		
Chip[1]	14		BovineSNP50 EuroG10K
Sample type	8		blood semen ear tissue other
Date of sampling	8	YYYY-MM-DD	
Number of the animal (Short name)	14		
Name of the animal	30		
Breed/type	2		
Birthdate	8	YYYY-MM-DD	
Sex	1		M - bull F - cow

[1] BovineSNP50 for the animals that do not require imputation; EuroG10K for the animals that require imputation

Embryotransfer	2		ET
Number of sire (Short number)	14		
Name of sire	30		
Number of dam (Short number)	14		
Name of dam	30		
Number of sire dam (Short number)	14		
Name of sire dam (Short number)	30		
Genotype operator[1]	16		WCHiRZ Poznań MCHiRZ Łowicz MCB Krasne SHiUZ Bydgoszcz PFHBiPM Warszawa

*1 or other institution authorized for that purpose.

FILE SHOULD BE SAVED IN CSV (comma-separated values) https://pl.wikipedia.org/wiki/CSV_(file format).

The naming of file with pedigree

File with pedigree information should be subjected to the following naming criteria:
YYYYMMDD_pedigree_file200.txt.xz
YYYYMMDD_pedigree_dane.csv.xz

Where:

- YYYYMMDD – data of file preparation containing year (YYYY), month (MM) and day (DD) of file preparation, the date should correspond to the date of file with genotypes
- pedigree – constant unit corresponding to the type of prepared file
- file200.txt.xz – constant unit corresponding to the type of prepared pedigree file of animals in file200 – file200 format, compressed in xz format
- dane.csv.xz – constant unit corresponding to the type of animals identification data – data compressed with xz format.

Example:

20140310_pedigree_file200.txt.xz

- indicates compressed pedigree data (xz) in file200 prepared on 10th March 2014.

Example:

20170829_pedigree_dane.csv.xz

- indicates compressed animal identification data (xz) in csv file prepared on 29th August 2017.

Genotype file transfer to ftp server

According to individually sent information about the way of data transfer, prepared files should be verified and compressed before sending to ftp server.

Pedigree file transfer to ftp server and email

According to individually sent information about the way of data transfer, prepared files should be verified and compressed before sending to ftp server.

Data with pedigree information both in file200 format and/or animal identification file format should be sent to Kacper Żukowski kacper.zukowski@izoo.krakow.pl and Monika Skarwecka monika.skarwecka@izoo.krakow.pl.

Contact information

National Research Institute of Animal Production
1, Krakowska Street
PL32083 Balice
Poland

Contact person regarding genotype data transfer is Kacper Żukowski email: kacper.zukowski@izoo.krakow.pl.
Contact person regarding pedigree information transfer is Monika Skarwecka email: monika.skarwecka@izoo.krakow.pl.

Format history

2019-12-21	-	format revision, chips added
2020-06-07	-	add ISO requirements
2019-05-30	-	format revision, chips added
2018-01-15	-	added English version of document
2017-10-09	-	addition of file200 to format description
2017-09-29	-	systematization of file format