Opis formatu plików dla genomowej oceny wartości hodowlanej

Dokument opisujący wymagane formaty plików dla wymian danych z Instytutem Zootechniki - Państwowym Instytutem Badawczym.

Wymagane typy i formaty plików

Rodzaje wymaganych typów i formatów plików:

1. Plik z genotypami,
2. Plik z informacją rodowodową (pedigree osobników).

Wymagania dla laboratoriów normy ISO17025

W przypadku, gdy usługa weryfikacji rodzicielskiej jest obowiązkowa w procesach weryfikacji danych, należy pamiętać, że od 2022 r. akredytacja ICAR dla laboratoriów przeprowadzających genotypowanie SNP zgodnie z normą ISO17025 będzie obowiązkowym wymogiem minimalnym zgodnie z wytycznymi: https://www.icar.org/index.php/certifications/certification-and-accreditation-of-dna-genetic-laboratories/guidelines-for-str-and-snp-based-parentage-testing-in-cattle/

Kompresja danych

W celu przyspieszenia transferu danych oraz redukcji przestrzeni magazynowania danych, wysyłane dane winny być poddane kompresji **xz**. W celu kompresji IZ PIB proponuje wykorzystanie darmowego oprogramowania do kompresji i dekompresji plików w wymaganym formacie dla środowiska Windows http://www.7-zip.org/ oraz dla środowiska Linux https://tukaani.org/xz/. Jeśli kompresja plików nie jest możliwa w formacie xz należy wykorzystać kompresję gzip http://www.gzip.org/.

1. Plik z genotypami

Dane z laboratoriów genotypowania winny być dostarczone w formacie *Illumina* FinalReport. Dane winny zawierać poszczególne kolumny **oddzielone tabulatorem**:

1. **Nazwa markera SNP (zgodnie z nazewnictwem *Illuminy*) - SNP Name,**
2. **Polski numer osobnika (14 znakowy) - Sample ID,**
3. **Allele1 (w formacie A/B) - Allele1 - AB,**
4. **Allele2 (w formacie A/B) - Allele2 - AB,**
5. **Allele1 (w formacie TOP) - Allele1 - TOP,**
6. **Allele2 (w formacie TOP) - Allele2 - TOP,**
7. (opcjonalnie) - wynik statystyki GC - GC Score,
8. (opcjonalnie) - wynik statystyki GT - GT Score,

Przykład:

```
[Header]
```

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

Dodatkowe informacje:

- brak danych powinien być oznaczony jako znak „-” zgodnie z domyślnymi ustawieniami oprogramowania Illumina GenomeStudio lub Beeline,
- minimalny call rate dla przesyłanych danych powinien wynosić 0.95 na osobnika,
- dla każdego rodzaju mikromacierzy proszę o utworzenie osobnego pliku,
- zwierzęta wysyłane na konkretną ocenę, genotypowane przy użyciu jednej mikromacierzy proszę umieszczać w jednym pliku.
- istnieje możliwość wysłania dodatkowych kolumn w pliku z genotypami zgodnie z zasadami generowania ich w programie Illumina GenomeStudio lub Beeline,
- wszystkie odstępstwa od w/w formatu prosimy zgłaszać przed transferem plików.

Nazewnictwo pliku z genotypami

Plik z genotypami winien podlegać następującym kryteriom nazewnictwa:
YYYYMMDD_RodzajMikromacierzy_Final_Report.txt.xz

Gdzie:

- YYYYMMDD - data przygotowania pliku odzwierciedlająca rok (YYYY), miesiąc (MM) oraz dzień (DD) przygotowania pliku,
- RodzajMikromacierzy - rodzaj mikromacierzy na której genotypowano osobniki w pliku. W chwili obecnej możliwe wartości to: *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, EuroGMDv3, EuroGMDv4*
- Final_Report - stała jednostka odpowiadająca rodzajowi przygotowanego pliku w programie Illumina GenomeStudio lub Beeline,
- rozszerzenie – txt.xz odpowiadające plikowi testowemu skompresowaniu w formacie xz.

*kolorem szarym oznaczono mikromacierze nie używane w Polsce.

Przykład:

20140310_EuroG10Kv3_Final_Report.txt

- oznacza dane przygotowane dnia 10 marca 2014 roku na mikromacierzy EuroG10Kv3

Przykład:

20170829_BovineSNP50v3_Final_Report.txt

- oznacza dane przygotowane dnia 29 sierpnia 2017 roku na mikromacierzy Illumina BovineSNP50 w wersji v3

2. Plik z informacją rodowodową (pedigree osobników)

Istnieje możliwość przesłania pliku z informacją rodowodową w dwóch formatach:

2.1. Zgodnie ze standardem Interbull - pliki w formacie file200 (**Tabela 1**).

Tabela 1. Format rodowodowe zwierząt w formacie file200.

Pozycja pierwszego znaku w linii rekordu	Opis pola	Typ danych i liczba znaków	Przykład
1	Typ wpisu	znakowy (character) 3	200
Międzynarodowy numer zwierzęcia (International ID of ANIMAL)			
5	Rasa zwierzęcia[1]	znakowy (character) 3	HOL
8	Kraj pierwszej rejestracji zwierzęcia[2]	znakowy (character) 3	CAN
11	Płeć zwierzęcia[3]	znakowy (character) 1	M
12	Numer zwierzęcia[4]	znakowy (character) 12	000000A12345
Międzynarodowy numer ojca zwierzęcia (International ID of Sire of ANIMAL)			
25	Rasa ojca zwierzęcia	znakowy (character) 3	HOL
28	Kraj pierwszej rejestracji ojca zwierzęcia[2]	znakowy (character) 3	CAN
31	Płeć	znakowy (character) 1	M
32	Numer ojca zwierzęcia[4]	znakowy (character) 12	556912367589
Międzynarodowy numer ojca zwierzęcia (International ID of Dam of ANIMAL)			
45	Rasa matki zwierzęcia	znakowy (character) 3	HOL
48	Kraj pierwszej rejestracji matki zwierzęcia[2]	znakowy (character) 3	CAN
51	Płeć	znakowy (character) 1	F
52	Numer matki zwierzęcia[4]	znakowy (character) 12	123569874521
65	Data urodzenia zwierzęcia (YYYYMMDD)	Liczba całkowita (integer) 8	19870215
74	Status zwierzęcia[5]	Liczba całkowita (integer) 2	10
77	Data urodzenia pierwszej córki AI (YYYYMMDD)	Liczba całkowita (integer) 8	19890314
86	Imię zwierzęcia	znakowy (character) 30	Cantarello
Narodowy numer zwierzęcia			
117	Rasa zwierzęcia[1]	znakowy (character) 3	HOL
120	Kraj pierwszej rejestracji zwierzęcia[2]	znakowy (character) 3	CAN

123	Płeć zwierzęcia[3]	znakowy (character) 1	M
124	Numer zwierzęcia[4]	znakowy (character) 12	000000A12345
137	Kraj wysyłający informację	znakowy (character) 3	CAN

*1 HOL= rasa holsztyńsko-fryzyjska (HF) odmiana czarno i czerwono biała; BSW= rasa Brunatna szwajcarska (Brown Swiss); JER= rasa Jersey; SIM= rasy Simental; RED= rasy czerwone bydła.
*2 Kraj, w którym pierwszy raz zarejestrowano zwierzę, wpisany zgodnie oznaczeniem krajów według tablicy ISO alpha-3 (dla przykładu Polska to POL, Niemcy to DEU itd.)
*3 Płeć zwierzęcia M – buhaj, samiec, F – jałówka, krowa, samica
*4 Unikalny, 12 znakowy numer zwierzęcia, wszystkie krótsze numery (poniżej 12 znaków) winny być dopełnione zerami od początku numeru (od lewej strony)
*5 Status buhaja: 00 – nieznany, stosowany również dla krów, 10 – buhaj hodowlany, 20 – inny buhaj

- Dane powinny być dostarczone jako plik tekstowy (*.txt)
- Brak numeru ojca lub matki powinien być oznaczony UUUUUUUUUUUUUUUUUUU (19xU)
- Brak w dacie urodzenia (YYYYMMDD) osobnika powinien być oznaczony jako 00000000 (8x0)
- Brak w dacie urodzenia pierwszej córki AI (YYYYMMDD) osobnika powinien być oznaczony jako 00000000 (8x0)
- Brak w statusie buhaja powinien być oznaczony jako 00 (2x0)
- Całkowita linia rekordu powinna posiadać wraz ze spacjami 140 znaki

Przykład:

```
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
…
```

*w przykładzie spacje oznaczono jako _

INFORMACJA RODOWODOWA W FORMACIE FILE200 WINNA BYĆ PRZYGOTOWANA ZGODNIE Z ZASADĄ JAK NAJEGŁĘBSZYCH RODOWODÓW ZARÓWNO PO STRONIE OJCOWSKIEJ JAK I MATECZNEJ.

2.2. Format danych identyfikacyjnych zwierząt (**Tabela 2**).

Tabela 2. Format danych identyfikacyjnych zwierząt przesyłanych do badania na mikromacierzach genotypowych

Kolumny	Znaki	Format	Uwagi
Numer próby	20		
Chip[1]	14		BovineSNP50

[1] BovineSNP50 dla osobników nie wymagających imputacji; EuroG10K dla osobników wymagających imputacji

			EuroG10K
Rodzaj próby	8		krew nasienie wycinek inny
Data pobrania	8	YYYY-MM-DD	
Numer zwierzęcia w Polsce	14		
Nazwa zwierzęcia	30		
Rasa/odmiana	2		
Data urodzenia	8	YYYY-MM-DD	
Płeć	1		M - buhaj F - krowa
Embriotransfer	2		ET
Numer ojca w Polsce	14		
Nazwa ojca	30		
Numer matki w Polsce	14		
Nazwa matki	30		
Numer ojca matki	14		
Nazwa ojca matki	30		
Operator genotypu[1]	16		WCHiRZ Poznań MCHiRZ Łowicz MCB Krasne SHiUZ Bydgoszcz PFHBiPM Warszawa

*1 lub inny podmiot do tego upoważniony.

PLIK WINIEN BYĆ ZAPISANY W FORMACIE CSV (ang. comma-separated values, wartości rozdzielone przecinkiem) Z WARTOŚCIAMI ODDZIELONYMI PRZECINKAMI https://pl.wikipedia.org/wiki/CSV_(format_pliku).

Nazewnictwo pliku z pedigree

Plik z informacją rodowodową winien podlegać następującym kryteriom nazewnictwa:
YYYYMMDD_pedigree_file200.txt.xz
YYYYMMDD_pedigree_dane.csv.xz

Gdzie:

- YYYYMMDD - data przygotowania pliku odzwierciedlająca rok (YYYY), miesiąc (MM) oraz dzień (DD) przygotowania pliku, data powinna odpowiadać dacie pliku z genotypami,
- pedigree - stała jednostka odpowiadająca rodzajowi przygotowanego pliku,
- file200.txt.xz - stała jednostka odpowiadająca rodzajowi przygotowanego pliku rodowodowemu zwierząt w formacie file200 – file200, skompresowane w formacie xz.
- dane.csv.xz - stała jednostka odpowiadająca rodzajowi przygotowanego danych identyfikacyjnych zwierząt – dane, skompresowane w formacie xz.

Przykład:

20140310_pedigree_file200.txt.xz

- oznacza skompresowane dane (xz) rodowodowe w file200 przygotowane dnia 10 marca 2014 roku.

Przykład:

20170829_pedigree_dane.csv.xz

- oznacza skompresowane dane (xz) identyfikacyjne zwierząt w pliku csv przygotowane dnia 29 sierpnia 2017.

Transfer plików z genotypami na serwer ftp

Zgodnie z indywidualnie przesłanymi informacjami o sposobie transferu danych przygotowane pliki przed umieszczeniem na serwerze ftp winny być zweryfikowane oraz skompresowane.

Transfer plików z informacją rodowodową na serwer ftp oraz email

Zgodnie z indywidualnie przesłanymi informacjami o sposobnie transferu danych przygotowane pliki przed umieszczeniem na serwerze ftp winny być zweryfikowane oraz skompresowane.

Dane z informacją rodowodową zarówno w formacie file200 jak i/lub w formacie danych identyfikacyjnych zwierząt winny być wysłane na adres dr inż. Kacper Żukowski kacper.zukowski@iz.edu.pl oraz dr inż. Monika Skarwecka monika.skarwecka@iz.edu.pl

Informacje kontaktowe

Instytut Zootechniki Państwowy Instytut Badawczy
ul. Krakowska 1
32-083 Balice

Osoba kontaktową w sprawie przyjmowania genotypów jest dr inż. Kacper Żukowski e-mail: kacper.zukowski@iz.edu.pl.
Osoba kontaktową w sprawie przyjmowania informacji rodowodowej jest dr inż. Monika Skarwecka e-mail: monika.skarwecka@iz.edu.pl.

Historia formatu

2023-10-30	-	rewizja pliku, dodanie mikromacierzy
2021-05-06	-	dodano format TOP na równi z AB, zmiana @
2020-12-21	-	rewizja pliku, dodanie mikromacierzy
2020-06-07	-	dodatnie wymogów odnośnie normy ISO17025
2019-05-30	-	rewizja pliku, dodanie mikromacierzy

2018-01-15	-	dodano wersje anglojęzyczną dokumentu
2017-10-09	-	dodatnie opisu formatu file200
2017-09-29	-	usystematyzowanie formatu plików