EXTRACTION OF STUDY ACCESION NUMBERS OF DATA USED IN GMHI MANUSCRIPT

```
#Studies and samples used to construct a meta-dataset composed of
4,347 human stool metagenomes.
import pandas as pd

df =
pd.read_excel("/home/aleksandra/Pulpit/BIOINFORMATYKA/mikrobiom/gmhi/
gmhi1.xlsx")
df_unique = df["Unnamed: 3"].unique()
print(df_unique)
gmhi_test = df_unique.tolist()
```

```
['Study Accession' nan 'PRJEB21528' 'PRJEB7774' 'PRJEB6070'
'PRJEB12449'
 'PRJEB10878' 'PRJEB1220' 'PRJNA385949' 'PRJNA389280' 'PRJEB15371'
 'PRJNA278393' 'PRJNA422434' 'PRJEB1786' 'PRJEB6337' 'PRJEB12123'
 'PRJEB19090' 'PRJNA305507' 'PRJNA268964' 'PRJEB1690' 'PRJNA319574'
 'PRJEB11532' 'PRJEB6997' 'PRJNA299502' 'PRJDB3601' 'PRJEB8094'
 'PRJEB13870' 'PRJEB6456' 'PRJEB4336' 'PRJNA177201' 'PRJNA373879'
 'PRJNA328899' 'PRJNA48479' 'PRJNA290729' 'PRJEB12947']
```

```
#Studies and samples used to construct the independent validation set
composed of 679 human stool metagenomes
import pandas as pd

df =
pd.read_excel("/home/aleksandra/Pulpit/BIOINFORMATYKA/mikrobiom/gmhi/
gmhi4.xlsx")
df_unique = df["Unnamed: 3"].unique()
print(df_unique)
gmhi_validation = df_unique.tolist()
```

```
['Study Accession' nan 'PRJNA321058' 'PRJEB17784' 'PRJEB6337'
 'PRJNA397112' 'PRJNA373901' 'PRJNA375935' 'PRJNA447983' 'PRJDB4176'
 'PRJEB27928']
```

EXTRACTION OF STUDY ACCESSION NUMBERS OF DATA USED IN GMWI2 MANUSCRIPT

```
#Supplementary Data 2. Stool shotgun metagenome samples (n = 8069)
from 54 independently published studies ranging across healthy and 11
non-healthy phenotypes.

import pandas as pd

df =
```

```
pd.read_excel("/home/aleksandra/Pulpit/BIOINFORMATYKA/mikrobiom/gmwi2/
gmwi2.xlsx")
df_unique = df["Unnamed: 1"].unique()
print(df_unique)
gmwi2_test = df_unique.tolist()
```

```
['BioProjectID' nan 'PRJNA384246' 'PRJNA665061' 'PRJEB39223'
'PRJEB6456'
 'PRJEB17632' 'PRJNA688274' 'PRJNA588805' 'PRJNA340216' 'PRJNA421881'
 'PRJNA397112' 'PRJEB7774' 'PRJNA400072' 'PRJEB12124' 'PRJNA598446'
 'PRJEB15371' 'PRJNA48479' 'PRJNA275349' 'PRJNA690543' 'PRJEB21528'
 'PRJEB1786' 'PRJEB33013' 'PRJEB4336' 'PRJNA328899' 'PRJNA398089'
 'PRJEB27005' 'PRJNA373901' 'PRJNA354235' 'PRJEB1220' 'PRJNA268964'
 'PRJNA485056' 'PRJNA504891' 'PRJNA530971' 'PRJNA422434' 'PRJEB6337'
 'PRJNA672125' 'PRJNA395744' 'PRJNA319574' 'PRJNA389280' 'PRJNA392180'
 'PRJNA693850' 'PRJNA529124' 'PRJNA529400' 'PRJDB4176' 'PRJNA447983'
 'PRJEB28543' 'PRJEB12449' 'PRJNA375935' 'PRJNA429990' 'PRJEB27928'
 'PRJEB9576' 'PRJNA429097' 'PRJNA763023' 'PRJNA475246' 'PRJEB10878'
 'PRJEB11532' 'PRJEB6070' 'PRJEB6997' 'PRJNA602729' 'PRJNA602731'
 'PRJNA638404' 'PRJNA638403' 'PRJNA602732' 'PRJNA638405']
```

```
#Supplementary Data 7. Human stool metagenome samples (n = 1140) in
the external validation dataset.
import pandas as pd

df =
pd.read_excel("/home/aleksandra/Pulpit/BIOINFORMATYKA/mikrobiom/gmwi2/
gmwi7.xlsx")
df_unique = df["Unnamed: 1"].unique()
print(df_unique)
gmwi2_validation = df_unique.tolist()
```

```
['BioProjectID' nan 'PRJNA834801' 'PRJEB49206' 'PRJNA890008'
'PRJEB28545'
 'PRJEB24557' 'PRJNA832909']
```

EXTRACTION OF STUDY ACCESSION NUMBERS OF DATA USED IN THE hiPCA MANUSCRIPT

```
#"For the discovery cohort and validation cohort, we used GMHI data"
import pandas as pd
import numpy as np

df =
pd.read_excel('/home/aleksandra/Pulpit/BIOINFORMATYKA/mikrobiom/hiPCA/
hipca1.xlsx', header=None, sheet_name = 'Discovery dataset')
row_idx = df[df[0] == 'Study Accession'].index[0]
study_accessions = df.loc[row_idx, 1:].values
unique_accessions = np.unique(study_accessions)
```

```
print(unique_accessions)
hipca_validation = unique_accessions.tolist()
```

```
['PRJDB3601' 'PRJEB10878' 'PRJEB11532' 'PRJEB12123' 'PRJEB1220'
 'PRJEB12449' 'PRJEB12947' 'PRJEB13870' 'PRJEB15371' 'PRJEB1690'
 'PRJEB1786' 'PRJEB19090' 'PRJEB21528' 'PRJEB4336' 'PRJEB6070'
'PRJEB6337'
 'PRJEB6456' 'PRJEB6997' 'PRJEB7774' 'PRJEB8094' 'PRJNA177201'
 'PRJNA268964' 'PRJNA278393' 'PRJNA290729' 'PRJNA299502' 'PRJNA305507'
 'PRJNA319574' 'PRJNA328899' 'PRJNA373879' 'PRJNA385949' 'PRJNA389280'
 'PRJNA422434' 'PRJNA48479']
```

```
#For the test cohort, all sequencing data for this analysis can be
obtained from the European Nucleotide Archive (ENA) databases, and the
project numbers are PRJEB27005, PRJEB29127, PRJNA449784, PRJNA504891,
PRJNA529124, PRJNA529400, and PRJNA531203
import pandas as pd
import numpy as np

df =
pd.read_excel('/home/aleksandra/Pulpit/BIOINFORMATYKA/mikrobiom/hiPCA/
hipca1.xlsx', header=None, sheet_name = 'Test dataset')
row_idx = df[df[0] == 'study_accession'].index[0]
study_accessions = df.loc[row_idx, 1:].values
unique_accessions = np.unique(study_accessions)
print(unique_accessions)
hipca_test = unique_accessions.tolist()
```

```
['PRJEB27005' 'PRJEB29127\t' 'PRJNA449784' 'PRJNA504891' 'PRJNA529124'
 'PRJNA529400' 'PRJNA531203']
```

SUMMARY

```
#Preparation of csv file consisting of all accession number of data
used in GMWI2, GMHI and hiPCA and summarizing number of repeats
import pandas as pd

def pad_list(lst, length):
    return lst + [None] * (length - len(lst))

max_len = max(len(gmhi_test), len(gmhi_validation), len(gmwi2_test),
len(gmwi2_validation), len(hipca_test), len(hipca_validation))

data = {
    'gmhi_test': pad_list(gmhi_test, max_len),
    'gmhi_validation': pad_list(gmhi_validation, max_len),
    'gmwi2_test': pad_list(gmwi2_test, max_len),
    'gmwi2_validation': pad_list(gmwi2_validation, max_len),
    'hipca_test': pad_list(hipca_test, max_len),
    'hipca_validation': pad_list(hipca_validation, max_len)
```

```
}
df = pd.DataFrame(data)

test_cols = [col for col in df.columns if col.endswith('test')]
validation_cols = [col for col in df.columns if
col.endswith('validation')]

test_series = pd.concat([df[col] for col in test_cols],
ignore_index=True).dropna()
validation_series = pd.concat([df[col] for col in validation_cols],
ignore_index=True).dropna()

def summarize_duplicates(series, group_cols):

    counts = series.value_counts()
    duplicates = counts[counts > 1]
    if duplicates.empty:
        print("No repeats")
        return
    print("Accession number | Number of repeates | Indices")
    for accession in duplicates.index:
        cols_with_accession = [col for col in group_cols if accession
in df[col].values]
        print(f"{accession} | {counts[accession]} |
{cols_with_accession}")

print("\nRepats of studies in test groups:")
summarize_duplicates(test_series, test_cols)

print("\nRepats of studies in validation groups:")
summarize_duplicates(validation_series, validation_cols)

all_series = pd.concat([df[col] for col in df.columns],
ignore_index=True).dropna()
all_counts = all_series.value_counts()
all_duplicates = all_counts[all_counts > 1]

print("\nRepats of studies in all groups:")
if all_duplicates.empty:
    print("No repats")
else:
    print("Accession number | Number of repeats | Indices")
    for accession in all_duplicates.index:
        cols_with_accession = [col for col in df.columns if accession
in df[col].values]
        print(f"{accession} | {all_counts[accession]} |
{cols_with_accession}")


Repats of studies in test groups:
```

```
Accession number | Number of repeates | Indices
PRJNA529124 | 2 | ['gmwi2_test', 'hipca_test']
PRJNA529400 | 2 | ['gmwi2_test', 'hipca_test']
PRJNA389280 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB15371 | 2 | ['gmhi_test', 'gmwi2_test']
PRJNA422434 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB1786 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB6337 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB27005 | 2 | ['gmwi2_test', 'hipca_test']
PRJNA504891 | 2 | ['gmwi2_test', 'hipca_test']
PRJNA268964 | 2 | ['gmhi_test', 'gmwi2_test']
PRJNA319574 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB11532 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB6997 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB6456 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB4336 | 2 | ['gmhi_test', 'gmwi2_test']
PRJNA328899 | 2 | ['gmhi_test', 'gmwi2_test']
PRJNA48479 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB21528 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB7774 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB6070 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB12449 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB10878 | 2 | ['gmhi_test', 'gmwi2_test']
PRJEB1220 | 2 | ['gmhi_test', 'gmwi2_test']

Repats of studies in validation groups:
Accession number | Number of repeates | Indices
PRJEB6337 | 2 | ['gmhi_validation', 'hipca_validation']

Repats of studies in all groups:
Accession number | Number of repeats | Indices
PRJEB6337 | 4 | ['gmhi_test', 'gmhi_validation', 'gmwi2_test',
'hipca_validation']
PRJEB15371 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB1786 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA422434 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA389280 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA268964 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB4336 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA48479 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA328899 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB6456 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB6997 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB11532 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA319574 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB1220 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB10878 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB21528 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB7774 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
```

```
PRJEB6070 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJEB12449 | 3 | ['gmhi_test', 'gmwi2_test', 'hipca_validation']
PRJNA278393 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA529124 | 2 | ['gmwi2_test', 'hipca_test']
PRJEB1690 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA305507 | 2 | ['gmhi_test', 'hipca_validation']
PRJEB19090 | 2 | ['gmhi_test', 'hipca_validation']
PRJEB12123 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA529400 | 2 | ['gmwi2_test', 'hipca_test']
PRJEB8094 | 2 | ['gmhi_test', 'hipca_validation']
PRJEB13870 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA299502 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA397112 | 2 | ['gmhi_validation', 'gmwi2_test']
PRJEB12947 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA290729 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA373879 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA177201 | 2 | ['gmhi_test', 'hipca_validation']
PRJDB3601 | 2 | ['gmhi_test', 'hipca_validation']
PRJNA373901 | 2 | ['gmhi_validation', 'gmwi2_test']
PRJNA375935 | 2 | ['gmhi_validation', 'gmwi2_test']
BioProjectID | 2 | ['gmwi2_test', 'gmwi2_validation']
PRJEB27928 | 2 | ['gmhi_validation', 'gmwi2_test']
PRJDB4176 | 2 | ['gmhi_validation', 'gmwi2_test']
PRJNA504891 | 2 | ['gmwi2_test', 'hipca_test']
PRJEB27005 | 2 | ['gmwi2_test', 'hipca_test']
PRJNA447983 | 2 | ['gmhi_validation', 'gmwi2_test']
Study Accession | 2 | ['gmhi_test', 'gmhi_validation']
PRJNA385949 | 2 | ['gmhi_test', 'hipca_validation']
```