bifido

October 15, 2019

[1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%pylab inline
```

```
Populating the interactive namespace from numpy and matplotlib
```

[2]:
```
# file path
filename = '/home/joaocarlosgomesneto/Documents/bifido_test/transposed_report.
 ↪tsv'
data = pd.read_csv(filename, sep='\t', header=0, na_values='Nothing')
```

[3]:
```
data.head()
```

[3]:

	Assembly	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	\
0	10M9_contigs	171	57	
1	1B5_contigs	184	55	
2	1M8_contigs	147	47	
3	3B9_contigs	99	43	
4	3M6_contigs	101	47	

	# contigs (>= 5000 bp)	# contigs (>= 10000 bp)	# contigs (>= 25000 bp)	\
0	46	42	31	
1	36	31	28	
2	34	30	27	
3	32	26	22	
4	39	34	24	

	# contigs (>= 50000 bp)	Total length (>= 0 bp)	Total length (>= 1000 bp)	\
0	18	2504424	2478586	
1	21	2615370	2579241	
2	23	5499281	5472023	
3	18	2595177	2577668	
4	16	2664445	2647845	

	Total length (>= 5000 bp)	…	Total length (>= 50000 bp)	# contigs	\
0	2457637	…	1741752	60	
1	2539750	…	2190814	59	

```
2                    5447732  …                    5209401          50
3                    2552328  …                    2270283          50
4                    2627501  …                    2097783          54

   Largest contig  Total length  GC (%)     N50     N75  L50  L75  \
0          217393       2480747   65.25   80366   46690   11   21
1          271967       2582106   63.50   87410   64327    9   17
2          721821       5473997   41.88  248321  152293    7   13
3          226637       2582568   63.50  140347   72430    7   13
4          310499       2653208   60.46  123902   61066    7   14

   # N's per 100 kbp
0                0.0
1                0.0
2                0.0
3                0.0
4                0.0

[5 rows x 22 columns]
```

[4]:
```
data.describe()
```

[4]:
```
       # contigs (>= 0 bp)  # contigs (>= 1000 bp)  # contigs (>= 5000 bp)  \
count           21.000000              21.000000              21.000000
mean          1478.523810             254.142857              60.095238
std           5865.710636             774.199541              67.064823
min             73.000000              35.000000              29.000000
25%            101.000000              44.000000              32.000000
50%            162.000000              56.000000              46.000000
75%            184.000000              62.000000              49.000000
max          27057.000000            3567.000000             339.000000

       # contigs (>= 10000 bp)  # contigs (>= 25000 bp)  \
count                21.000000               21.000000
mean                 38.238095               25.761905
std                  11.999603                9.591167
min                  25.000000                1.000000
25%                  29.000000               23.000000
50%                  39.000000               28.000000
75%                  43.000000               32.000000
max                  67.000000               41.000000

       # contigs (>= 50000 bp)  Total length (>= 0 bp)  \
count                21.000000             2.100000e+01
mean                 17.095238             3.491058e+06
std                   6.212123             3.100314e+06
min                   0.000000             2.397236e+06
```

25%	17.000000	2.595177e+06
50%	18.000000	2.664445e+06
75%	20.000000	2.835210e+06
max	25.000000	1.673026e+07

	Total length (>= 1000 bp)	Total length (>= 5000 bp)	\
count	2.100000e+01	2.100000e+01	
mean	3.094200e+06	2.700611e+06	
std	1.519337e+06	7.473158e+05	
min	2.387056e+06	1.013016e+06	
25%	2.538521e+06	2.516356e+06	
50%	2.629553e+06	2.604373e+06	
75%	2.817420e+06	2.737212e+06	
max	9.100740e+06	5.447732e+06	

	Total length (>= 10000 bp)	…	Total length (>= 50000 bp)	\
count	2.100000e+01	…	2.100000e+01	
mean	2.549475e+06	…	2.033151e+06	
std	9.059729e+05	…	9.843005e+05	
min	3.933120e+05	…	0.000000e+00	
25%	2.465051e+06	…	1.813298e+06	
50%	2.515086e+06	…	2.097783e+06	
75%	2.647148e+06	…	2.244929e+06	
max	5.419392e+06	…	5.209401e+06	

	# contigs	Largest contig	Total length	GC (%)	N50	\
count	21.000000	21.000000	2.100000e+01	21.000000	21.000000	
mean	490.809524	279558.000000	3.257594e+06	61.080476	104164.333333	
std	1775.937404	147452.678227	2.165545e+06	5.162948	60880.986909	
min	40.000000	27726.000000	2.389826e+06	41.880000	2028.000000	
25%	50.000000	217393.000000	2.566729e+06	60.400000	76803.000000	
50%	60.000000	271967.000000	2.633698e+06	62.310000	89902.000000	
75%	69.000000	322519.000000	2.820415e+06	63.500000	123902.000000	
max	8192.000000	721821.000000	1.228660e+07	65.290000	248321.000000	

	N75	L50	L75	# N's per 100 kbp
count	21.000000	21.000000	21.000000	21.0
mean	57426.571429	86.047619	211.714286	0.0
std	31025.505386	314.103562	800.274274	0.0
min	974.000000	4.000000	9.000000	0.0
25%	46690.000000	7.000000	14.000000	0.0
50%	54289.000000	9.000000	19.000000	0.0
75%	68733.000000	11.000000	21.000000	0.0
max	152293.000000	1446.000000	3683.000000	0.0

[8 rows x 21 columns]

[5]:
```
data.loc[(data['# contigs'] == 0) | (data['# contigs'] >= 300) | (data['N50']␣
↪<= 25000)]
```

[5]:

	Assembly	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	\
8	5B1_contigs	1232	750	
20	Undetermined_contigs	27057	3567	

	# contigs (>= 5000 bp)	# contigs (>= 10000 bp)	# contigs (>= 25000 bp)	\
8	121	28	2	
20	339	67	1	

	# contigs (>= 50000 bp)	Total length (>= 0 bp)	\
8	0	2700537	
20	0	16730263	

	Total length (>= 1000 bp)	Total length (>= 5000 bp)	…	\
8	2456437	1013016	…	
20	9100740	2737212	…	

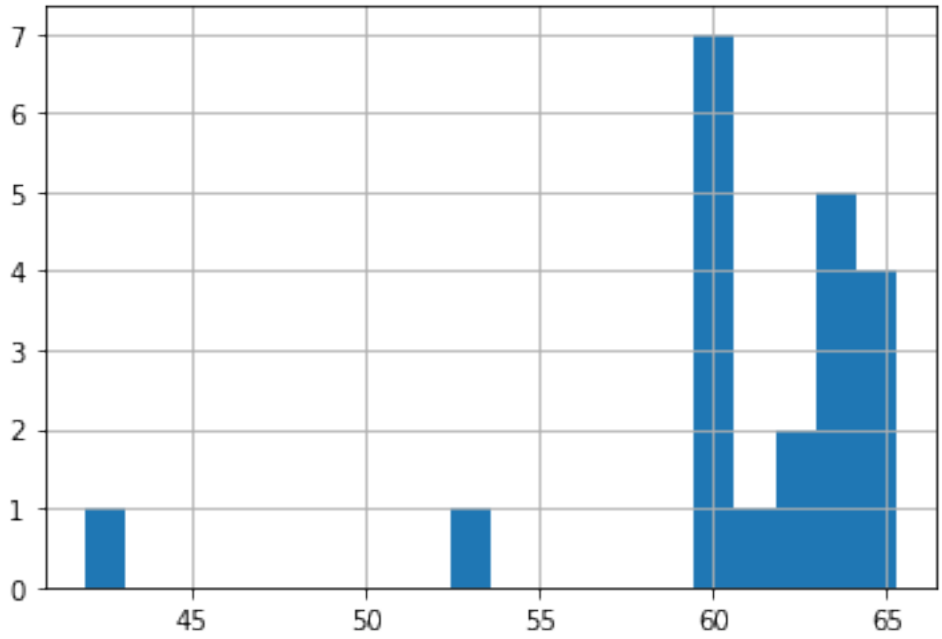
	Total length (>= 50000 bp)	# contigs	Largest contig	Total length	\
8	0	978	27726	2622976	
20	0	8192	28828	12286595	

	GC (%)	N50	N75	L50	L75	# N's per 100 kbp
8	60.33	3796	2065	190	423	0.0
20	61.59	2028	974	1446	3683	0.0

[2 rows x 22 columns]

[6]:
```
data['GC (%)'].hist(bins = 20)
```

[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fca321a2e10>
```



[ ]:

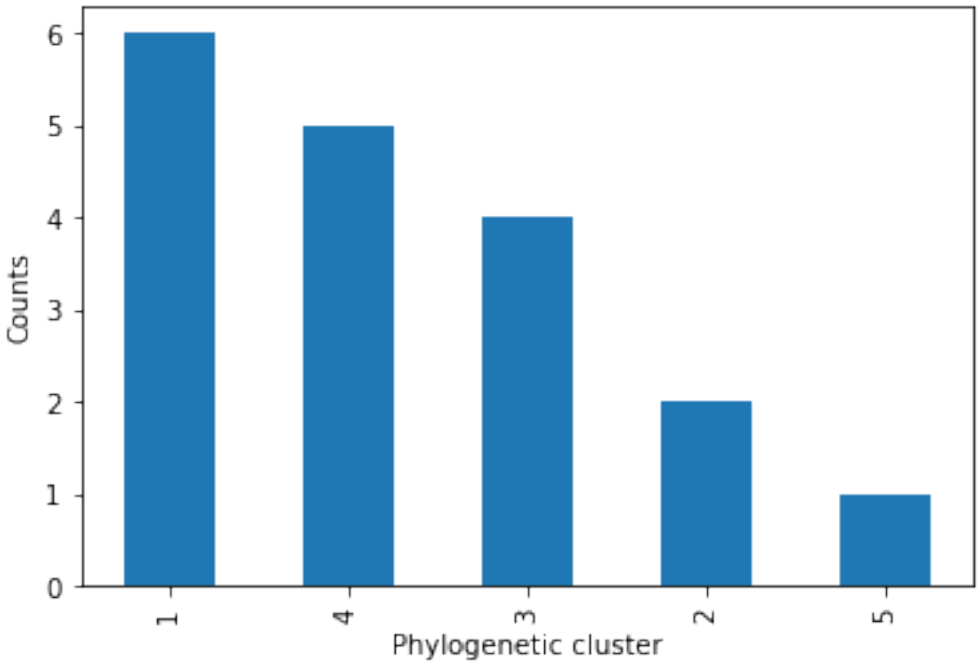
[7]:
```
# file1 path
file1 = '/home/joaocarlosgomesneto/Documents/bifido_reference_genomes/
 ↪species_table.csv'
data1 = pd.read_csv(file1, header=0, na_values='Nothing', index_col='id')
```

[8]:
```
data1.head()
```

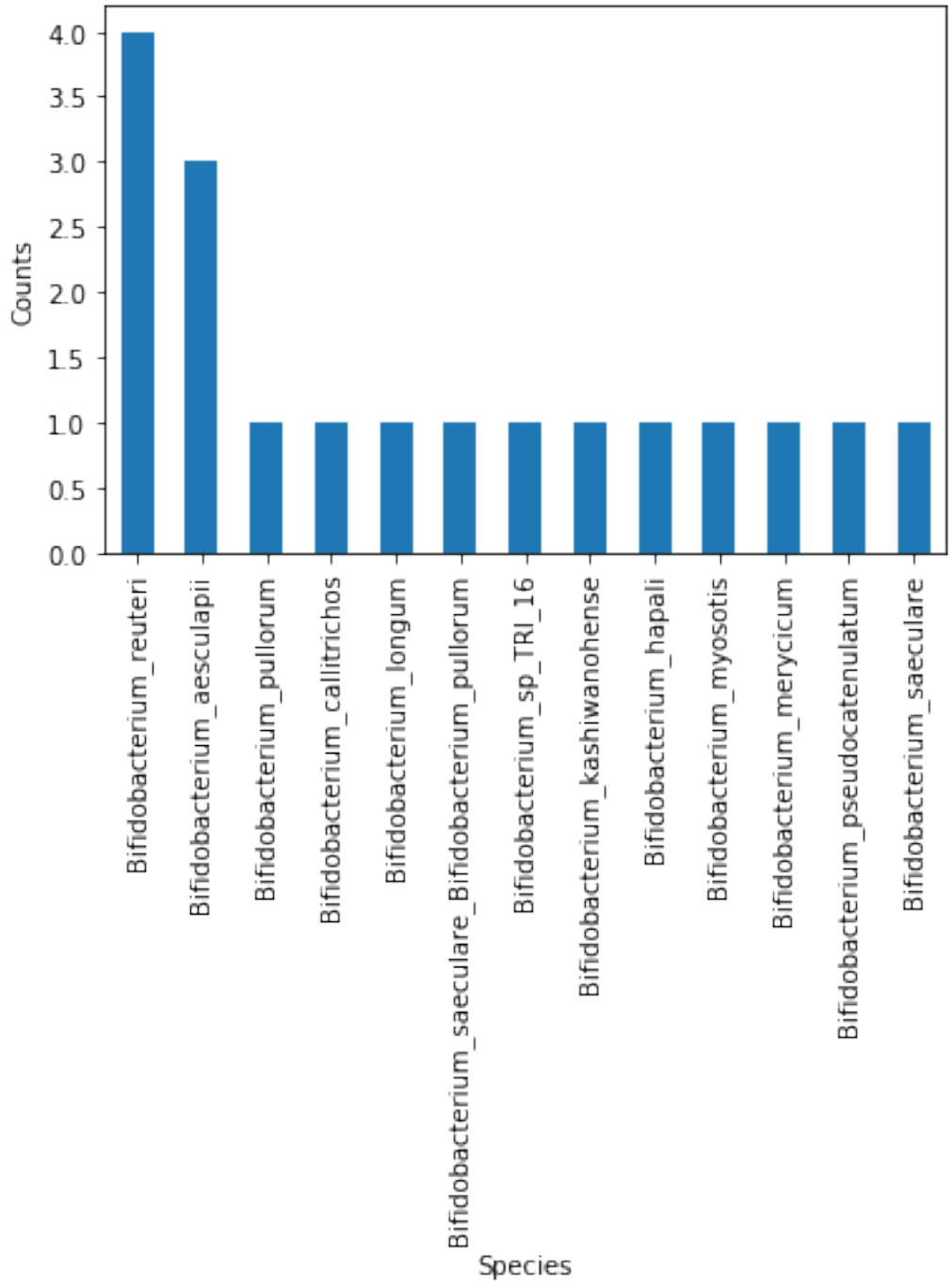
[8]:
```
                      species  core_genome_phylogenetic_cluster  \
id
8B3    Bifidobacterium_reuteri                                  1
3M6    Bifidobacterium_reuteri                                  1
5B2  Bifidobacterium_merycicum                                  1
5B8     Bifidobacterium_longum                                  1
4B7    Bifidobacterium_reuteri                                  1

     phylotype_within_phylogroups  phylogroup_gene  phylotype_gene
id
8B3                             1              NaN             NaN
3M6                             1              NaN             NaN
5B2                             1              NaN             NaN
5B8                             1              NaN             NaN
4B7                             2              NaN             NaN
```

[9]:
```
_ = data1['core_genome_phylogenetic_cluster'].value_counts().plot(kind = 'bar')
_ = plt.ylabel('Counts')
_ = plt.xlabel('Phylogenetic cluster')
plt.show()
```



[10]:
```
_ = data1['species'].value_counts().plot(kind = 'bar')
_ = plt.ylabel('Counts')
_ = plt.xlabel('Species')
plt.show()
```



[11]:
```
data1.isnull().sum()
```

[11]:
```
species                              0
core_genome_phylogenetic_cluster     0
phylotype_within_phylogroups         0
phylogroup_gene                     18
```

```
phylotype_gene                    18
dtype: int64
```

[12]:
```
# file2 path
file2 = '/home/joaocarlosgomesneto/Documents/bifido_reference_genomes/
 ↪pangenome_bifido.csv'
data2 = pd.read_csv(file2, header=0, na_values='Nothing')
```

[13]:
```
data2.head()
```

[13]:

	id	10M9	1B5	3B9	3M6	4B6	4B7	4M3	5B2	5B8	5B9	6M2	8B3	\
0	group_100	1	1	1	1	1	1	1	1	1	1	1	1	
1	pgm	1	1	1	1	1	1	1	1	1	1	1	1	
2	infB	1	1	1	1	1	1	1	1	1	1	1	1	
3	ffh	1	1	1	1	1	1	1	1	1	1	1	1	
4	adhE	1	1	1	1	1	1	1	1	1	1	1	1	

	8B4	8B6	8B9	8M5	9B2	9B6
0	1	1	1	1	1	1
1	1	1	1	1	1	1
2	1	1	1	1	1	1
3	1	1	1	1	1	1
4	1	1	1	1	1	1

[14]:
```
data3 = data2.transpose()
```

[15]:
```
data3.head()
```

[15]:

	0	1	2	3	4	5	6	7	8	9	\
id	group_100	pgm	infB	ffh	adhE	ywaC	rpsK	rpsZ	pdtaS	group_120	
10M9	1	1	1	1	1	1	1	1	1	1	
1B5	1	1	1	1	1	1	1	1	1	1	
3B9	1	1	1	1	1	1	1	1	1	1	
3M6	1	1	1	1	1	1	1	1	1	1	

	…	12197	12198	12199	12200	12201	\
id	…	group_9990	group_9991	group_9992	group_9993	group_9994	
10M9	…	0	0	0	0	0	
1B5	…	0	0	0	0	0	
3B9	…	0	0	0	0	0	
3M6	…	0	0	0	0	0	

	12202	12203	12204	12205	12206
id	group_9995	group_9996	group_9997	group_9998	group_9999
10M9	0	0	0	0	0
1B5	0	0	0	0	0
3B9	0	0	0	0	0

```
3M6            0           0           0           0           0

[5 rows x 12207 columns]
```

[16]:
```
data3.to_csv('/home/joaocarlosgomesneto/Documents/bifido_reference_genomes/
↪data3.csv', header=False, index=True)
```

[17]:
```
# file3 path
file3 = '/home/joaocarlosgomesneto/Documents/bifido_reference_genomes/data3.csv'
data4 = pd.read_csv(file3, header = 0, na_values='Nothing', index_col = 'id')
```

[18]:
```
data4.head()
```

[18]:
```
      group_100  pgm  infB  ffh  adhE  ywaC  rpsK  rpsZ  pdtaS  group_120  \
id
10M9          1    1     1    1     1     1     1     1      1          1
1B5           1    1     1    1     1     1     1     1      1          1
3B9           1    1     1    1     1     1     1     1      1          1
3M6           1    1     1    1     1     1     1     1      1          1
4B6           1    1     1    1     1     1     1     1      1          1

      …  group_9990  group_9991  group_9992  group_9993  group_9994  \
id    …
10M9  …           0           0           0           0           0
1B5   …           0           0           0           0           0
3B9   …           0           0           0           0           0
3M6   …           0           0           0           0           0
4B6   …           0           0           0           0           0

      group_9995  group_9996  group_9997  group_9998  group_9999
id
10M9           0           0           0           0           0
1B5            0           0           0           0           0
3B9            0           0           0           0           0
3M6            0           0           0           0           0
4B6            0           0           0           0           0

[5 rows x 12207 columns]
```

[19]:
```
data5 = pd.concat([data1, data4], axis = 1, sort = False, join = 'inner')
```

[20]:
```
data5.head()
```

[20]:
```
                      species  core_genome_phylogenetic_cluster  \
id
8B3    Bifidobacterium_reuteri                                 1
3M6    Bifidobacterium_reuteri                                 1
```

```
5B2  Bifidobacterium_merycicum                                  1
5B8     Bifidobacterium_longum                                  1
4B7    Bifidobacterium_reuteri                                  1

     phylotype_within_phylogroups  phylogroup_gene  phylotype_gene  group_100  \
id
8B3                             1              NaN             NaN          1
3M6                             1              NaN             NaN          1
5B2                             1              NaN             NaN          1
5B8                             1              NaN             NaN          1
4B7                             2              NaN             NaN          1

     pgm  infB  ffh  adhE  …  group_9990  group_9991  group_9992  \
id                        …
8B3    1     1    1     1  …           0           0           0
3M6    1     1    1     1  …           0           0           0
5B2    1     1    1     1  …           0           0           0
5B8    1     1    1     1  …           0           0           0
4B7    1     1    1     1  …           0           0           0

     group_9993  group_9994  group_9995  group_9996  group_9997  group_9998  \
id
8B3           0           0           0           0           0           0
3M6           0           0           0           0           0           0
5B2           0           0           0           0           0           0
5B8           0           0           0           0           0           0
4B7           0           0           0           0           0           0

     group_9999
id
8B3           0
3M6           0
5B2           0
5B8           0
4B7           0

[5 rows x 12212 columns]
```

[21]:
```
data5
```

[21]:
```
                                          species  \
id
8B3                          Bifidobacterium_reuteri
3M6                          Bifidobacterium_reuteri
5B2                       Bifidobacterium_merycicum
5B8                          Bifidobacterium_longum
4B7                          Bifidobacterium_reuteri
```

```
6M2                              Bifidobacterium_reuteri
9B2                             Bifidobacterium_myosotis
9B6                         Bifidobacterium_callitrichos
8B9                             Bifidobacterium_pullorum
10M9                          Bifidobacterium_aesculapii
8M5                           Bifidobacterium_aesculapii
8B6                           Bifidobacterium_aesculapii
4B6                    Bifidobacterium_pseudocatenulatum
4M3                       Bifidobacterium_kashiwanohense
3B9   Bifidobacterium_saeculare_Bifidobacterium_pull…
5B9                            Bifidobacterium_saeculare
1B5                            Bifidobacterium_sp_TRI_16
8B4                               Bifidobacterium_hapali

      core_genome_phylogenetic_cluster  phylotype_within_phylogroups  \
id
8B3                                  1                             1
3M6                                  1                             1
5B2                                  1                             1
5B8                                  1                             1
4B7                                  1                             2
6M2                                  1                             2
9B2                                  2                             1
9B6                                  2                             1
8B9                                  3                             1
10M9                                 3                             1
8M5                                  3                             1
8B6                                  3                             1
4B6                                  4                             1
4M3                                  4                             2
3B9                                  4                             3
5B9                                  4                             3
1B5                                  4                             3
8B4                                  5                             1

      phylogroup_gene  phylotype_gene  group_100  pgm  infB  ffh  adhE  … \
id                                                                      …
8B3               NaN             NaN          1    1     1    1     1  …
3M6               NaN             NaN          1    1     1    1     1  …
5B2               NaN             NaN          1    1     1    1     1  …
5B8               NaN             NaN          1    1     1    1     1  …
4B7               NaN             NaN          1    1     1    1     1  …
6M2               NaN             NaN          1    1     1    1     1  …
9B2               NaN             NaN          1    1     1    1     1  …
9B6               NaN             NaN          1    1     1    1     1  …
8B9               NaN             NaN          1    1     1    1     1  …
10M9              NaN             NaN          1    1     1    1     1  …
```

```
8M5               NaN             NaN          1    1     1    1     1  …
8B6               NaN             NaN          1    1     1    1     1  …
4B6               NaN             NaN          1    1     1    1     1  …
4M3               NaN             NaN          1    1     1    1     1  …
3B9               NaN             NaN          1    1     1    1     1  …
5B9               NaN             NaN          1    1     1    1     1  …
1B5               NaN             NaN          1    1     1    1     1  …
8B4               NaN             NaN          1    1     1    1     1  …

      group_9990  group_9991  group_9992  group_9993  group_9994  group_9995  \
id
8B3            0           0           0           0           0           0
3M6            0           0           0           0           0           0
5B2            0           0           0           0           0           0
5B8            0           0           0           0           0           0
4B7            0           0           0           0           0           0
6M2            0           0           0           0           0           0
9B2            0           0           0           0           0           0
9B6            0           0           0           0           0           0
8B9            0           0           0           0           0           0
10M9           0           0           0           0           0           0
8M5            0           0           0           0           0           0
8B6            0           0           0           0           0           0
4B6            0           0           0           0           0           0
4M3            0           0           0           0           0           0
3B9            0           0           0           0           0           0
5B9            0           0           0           0           0           0
1B5            0           0           0           0           0           0
8B4            1           1           1           1           1           1

      group_9996  group_9997  group_9998  group_9999
id
8B3            0           0           0           0
3M6            0           0           0           0
5B2            0           0           0           0
5B8            0           0           0           0
4B7            0           0           0           0
6M2            0           0           0           0
9B2            0           0           0           0
9B6            0           0           0           0
8B9            0           0           0           0
10M9           0           0           0           0
8M5            0           0           0           0
8B6            0           0           0           0
4B6            0           0           0           0
4M3            0           0           0           0
3B9            0           0           0           0
```

```
5B9            0           0           0           0
1B5            0           0           0           0
8B4            1           1           1           1

[18 rows x 12212 columns]
```

[22]:
```
# initial search for candidate genes using a binary search of total number

sums = data5.select_dtypes(pd.np.number).sum().rename('total')
data6 = data5.append(sums)
```

[23]:
```
data6
```

[23]:
```
                                                species  \
id
8B3                                 Bifidobacterium_reuteri
3M6                                 Bifidobacterium_reuteri
5B2                               Bifidobacterium_merycicum
5B8                                  Bifidobacterium_longum
4B7                                 Bifidobacterium_reuteri
6M2                                 Bifidobacterium_reuteri
9B2                               Bifidobacterium_myosotis
9B6                           Bifidobacterium_callitrichos
8B9                               Bifidobacterium_pullorum
10M9                            Bifidobacterium_aesculapii
8M5                             Bifidobacterium_aesculapii
8B6                             Bifidobacterium_aesculapii
4B6                     Bifidobacterium_pseudocatenulatum
4M3                        Bifidobacterium_kashiwanohense
3B9    Bifidobacterium_saeculare_Bifidobacterium_pull…
5B9                             Bifidobacterium_saeculare
1B5                             Bifidobacterium_sp_TRI_16
8B4                                Bifidobacterium_hapali
total                                                  NaN

       core_genome_phylogenetic_cluster  phylotype_within_phylogroups  \
id
8B3                                 1.0                           1.0
3M6                                 1.0                           1.0
5B2                                 1.0                           1.0
5B8                                 1.0                           1.0
4B7                                 1.0                           2.0
6M2                                 1.0                           2.0
9B2                                 2.0                           1.0
9B6                                 2.0                           1.0
8B9                                 3.0                           1.0
10M9                                3.0                           1.0
```

```
8M5                              3.0                              1.0
8B6                              3.0                              1.0
4B6                              4.0                              1.0
4M3                              4.0                              2.0
3B9                              4.0                              3.0
5B9                              4.0                              3.0
1B5                              4.0                              3.0
8B4                              5.0                              1.0
total                           47.0                             27.0

      phylogroup_gene  phylotype_gene  group_100   pgm  infB   ffh  adhE  \
id
8B3               NaN             NaN        1.0   1.0   1.0   1.0   1.0
3M6               NaN             NaN        1.0   1.0   1.0   1.0   1.0
5B2               NaN             NaN        1.0   1.0   1.0   1.0   1.0
5B8               NaN             NaN        1.0   1.0   1.0   1.0   1.0
4B7               NaN             NaN        1.0   1.0   1.0   1.0   1.0
6M2               NaN             NaN        1.0   1.0   1.0   1.0   1.0
9B2               NaN             NaN        1.0   1.0   1.0   1.0   1.0
9B6               NaN             NaN        1.0   1.0   1.0   1.0   1.0
8B9               NaN             NaN        1.0   1.0   1.0   1.0   1.0
10M9              NaN             NaN        1.0   1.0   1.0   1.0   1.0
8M5               NaN             NaN        1.0   1.0   1.0   1.0   1.0
8B6               NaN             NaN        1.0   1.0   1.0   1.0   1.0
4B6               NaN             NaN        1.0   1.0   1.0   1.0   1.0
4M3               NaN             NaN        1.0   1.0   1.0   1.0   1.0
3B9               NaN             NaN        1.0   1.0   1.0   1.0   1.0
5B9               NaN             NaN        1.0   1.0   1.0   1.0   1.0
1B5               NaN             NaN        1.0   1.0   1.0   1.0   1.0
8B4               NaN             NaN        1.0   1.0   1.0   1.0   1.0
total             0.0             0.0       18.0  18.0  18.0  18.0  18.0

      …  group_9990  group_9991  group_9992  group_9993  group_9994  \
id    …
8B3   …         0.0         0.0         0.0         0.0         0.0
3M6   …         0.0         0.0         0.0         0.0         0.0
5B2   …         0.0         0.0         0.0         0.0         0.0
5B8   …         0.0         0.0         0.0         0.0         0.0
4B7   …         0.0         0.0         0.0         0.0         0.0
6M2   …         0.0         0.0         0.0         0.0         0.0
9B2   …         0.0         0.0         0.0         0.0         0.0
9B6   …         0.0         0.0         0.0         0.0         0.0
8B9   …         0.0         0.0         0.0         0.0         0.0
10M9  …         0.0         0.0         0.0         0.0         0.0
8M5   …         0.0         0.0         0.0         0.0         0.0
8B6   …         0.0         0.0         0.0         0.0         0.0
4B6   …         0.0         0.0         0.0         0.0         0.0
```

```
4M3    …         0.0         0.0         0.0         0.0         0.0
3B9    …         0.0         0.0         0.0         0.0         0.0
5B9    …         0.0         0.0         0.0         0.0         0.0
1B5    …         0.0         0.0         0.0         0.0         0.0
8B4    …         1.0         1.0         1.0         1.0         1.0
total  …         1.0         1.0         1.0         1.0         1.0

       group_9995  group_9996  group_9997  group_9998  group_9999
id
8B3           0.0         0.0         0.0         0.0         0.0
3M6           0.0         0.0         0.0         0.0         0.0
5B2           0.0         0.0         0.0         0.0         0.0
5B8           0.0         0.0         0.0         0.0         0.0
4B7           0.0         0.0         0.0         0.0         0.0
6M2           0.0         0.0         0.0         0.0         0.0
9B2           0.0         0.0         0.0         0.0         0.0
9B6           0.0         0.0         0.0         0.0         0.0
8B9           0.0         0.0         0.0         0.0         0.0
10M9          0.0         0.0         0.0         0.0         0.0
8M5           0.0         0.0         0.0         0.0         0.0
8B6           0.0         0.0         0.0         0.0         0.0
4B6           0.0         0.0         0.0         0.0         0.0
4M3           0.0         0.0         0.0         0.0         0.0
3B9           0.0         0.0         0.0         0.0         0.0
5B9           0.0         0.0         0.0         0.0         0.0
1B5           0.0         0.0         0.0         0.0         0.0
8B4           1.0         1.0         1.0         1.0         1.0
total         1.0         1.0         1.0         1.0         1.0

[19 rows x 12212 columns]
```

[24]:
```
data7 = data5.loc[:, (data5.sum(axis=0) != 18)]
```

[25]:
```
data7
```

[25]:
```
                                          species  \
id
8B3                          Bifidobacterium_reuteri
3M6                          Bifidobacterium_reuteri
5B2                       Bifidobacterium_merycicum
5B8                          Bifidobacterium_longum
4B7                          Bifidobacterium_reuteri
6M2                          Bifidobacterium_reuteri
9B2                        Bifidobacterium_myosotis
9B6                    Bifidobacterium_callitrichos
8B9                        Bifidobacterium_pullorum
10M9                     Bifidobacterium_aesculapii
```

```
8M5                          Bifidobacterium_aesculapii
8B6                          Bifidobacterium_aesculapii
4B6                 Bifidobacterium_pseudocatenulatum
4M3                    Bifidobacterium_kashiwanohense
3B9   Bifidobacterium_saeculare_Bifidobacterium_pull…
5B9                          Bifidobacterium_saeculare
1B5                          Bifidobacterium_sp_TRI_16
8B4                             Bifidobacterium_hapali

      core_genome_phylogenetic_cluster  phylotype_within_phylogroups  \
id
8B3                                  1                             1
3M6                                  1                             1
5B2                                  1                             1
5B8                                  1                             1
4B7                                  1                             2
6M2                                  1                             2
9B2                                  2                             1
9B6                                  2                             1
8B9                                  3                             1
10M9                                 3                             1
8M5                                  3                             1
8B6                                  3                             1
4B6                                  4                             1
4M3                                  4                             2
3B9                                  4                             3
5B9                                  4                             3
1B5                                  4                             3
8B4                                  5                             1

      phylogroup_gene  phylotype_gene  group_1092  group_1108  map  coaD  \
id
8B3               NaN             NaN           1           1    1     1
3M6               NaN             NaN           1           1    1     1
5B2               NaN             NaN           1           1    1     1
5B8               NaN             NaN           1           1    1     1
4B7               NaN             NaN           1           1    1     1
6M2               NaN             NaN           1           1    1     1
9B2               NaN             NaN           1           1    1     1
9B6               NaN             NaN           1           1    1     1
8B9               NaN             NaN           1           1    1     1
10M9              NaN             NaN           1           1    1     1
8M5               NaN             NaN           1           1    1     1
8B6               NaN             NaN           1           1    1     1
4B6               NaN             NaN           1           1    1     1
4M3               NaN             NaN           1           1    0     0
3B9               NaN             NaN           1           1    1     1
```

```
5B9              NaN             NaN           1           1    1     1
1B5              NaN             NaN           1           1    1     1
8B4              NaN             NaN           0           0    1     1

      aroP  …  group_9990  group_9991  group_9992  group_9993  group_9994  \
id          …
8B3      1  …           0           0           0           0           0
3M6      1  …           0           0           0           0           0
5B2      1  …           0           0           0           0           0
5B8      1  …           0           0           0           0           0
4B7      1  …           0           0           0           0           0
6M2      1  …           0           0           0           0           0
9B2      1  …           0           0           0           0           0
9B6      1  …           0           0           0           0           0
8B9      1  …           0           0           0           0           0
10M9     1  …           0           0           0           0           0
8M5      1  …           0           0           0           0           0
8B6      1  …           0           0           0           0           0
4B6      1  …           0           0           0           0           0
4M3      1  …           0           0           0           0           0
3B9      1  …           0           0           0           0           0
5B9      1  …           0           0           0           0           0
1B5      1  …           0           0           0           0           0
8B4      0  …           1           1           1           1           1

      group_9995  group_9996  group_9997  group_9998  group_9999
id
8B3            0           0           0           0           0
3M6            0           0           0           0           0
5B2            0           0           0           0           0
5B8            0           0           0           0           0
4B7            0           0           0           0           0
6M2            0           0           0           0           0
9B2            0           0           0           0           0
9B6            0           0           0           0           0
8B9            0           0           0           0           0
10M9           0           0           0           0           0
8M5            0           0           0           0           0
8B6            0           0           0           0           0
4B6            0           0           0           0           0
4M3            0           0           0           0           0
3B9            0           0           0           0           0
5B9            0           0           0           0           0
1B5            0           0           0           0           0
8B4            1           1           1           1           1

[18 rows x 11919 columns]
```

[26]:
```
data5.shape #entire data
```

[26]:
```
(18, 12212)
```

[27]:
```
data7.shape #entire data - core genes
```

[27]:
```
(18, 11919)
```

[28]:
```
12209-11916 #calculation for the number of core genes
```

[28]:
```
293
```

[100]:
```
# filter out phylogroup 1 using data7 which does not include the core genes
phy1a = data7[data7.core_genome_phylogenetic_cluster == 1] #filtering out␣
 ↪phylogroup 1
phy1b = phy1a.loc[:, (phy1a.sum(axis=0) == 6)] #filtering out genes present in␣
 ↪all isolates
```

[101]:
```
phy1_genes = ['lsrA', 'mdtH', 'mdtK']
phy1c = pd.DataFrame(phy1b, columns = phy1_genes)
```

[102]:
```
phy1c # list of candidate genes for phylogroup 1
```

[102]:

id	lsrA	mdtH	mdtK
8B3	1	1	1
3M6	1	1	1
5B2	1	1	1
5B8	1	1	1
4B7	1	1	1
6M2	1	1	1

[103]:
```
# check if phylogroup 1 genes (lsrA) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.lsrA)
```

[103]:

lsrA	0	1
core_genome_phylogenetic_cluster		
1	0	6
2	2	0
3	4	0
4	5	0
5	1	0

[104]:
```
# check if phylogroup 1 genes (mdtH) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.mdtH)
```

[104]:
```
mdtH                                0  1
core_genome_phylogenetic_cluster
1                                   0  6
2                                   2  0
3                                   4  0
4                                   5  0
5                                   1  0
```

[105]:
```
# check if phylogroup 1 genes (mdtK) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.mdtK)
```

[105]:
```
mdtK                                0  1
core_genome_phylogenetic_cluster
1                                   0  6
2                                   2  0
3                                   4  0
4                                   5  0
5                                   1  0
```

[61]:
```
pd.crosstab(phy1a.phylotype_within_phylogroups, phy1a.aprN) # checking␣
↪candidate gene for phylogroup 1 & phylotype 1
```

[61]:
```
aprN                          0  1
phylotype_within_phylogroups
1                             0  4
2                             2  0
```

[62]:
```
pd.crosstab(phy1a.phylotype_within_phylogroups, phy1a.acm_2) # checking␣
↪candidate gene for phylogroup 1 & phylotype 1
```

[62]:
```
acm_2                         0  1
phylotype_within_phylogroups
1                             0  4
2                             2  0
```

[65]:
```
p1p1genes = ['aprN', 'acm_2'] # checking candidate genes for phylotype1 within␣
↪phylogroup1
new1 = pd.DataFrame(phy1a, columns = p1p1genes)
new1
```

[65]:
```
      aprN  acm_2
id
8B3      1      1
3M6      1      1
5B2      1      1
5B8      1      1
4B7      0      0
```

```
6M2     0      0
```

[106]:
```
p1p1genes = ['aprN', 'acm_2'] # checking candidate genes for phylotype1 across␣
 ↪the entire database
new1 = pd.DataFrame(data7, columns = p1p1genes)
new1
```

[106]:

id	aprN	acm_2
8B3	1	1
3M6	1	1
5B2	1	1
5B8	1	1
4B7	0	0
6M2	0	0
9B2	0	0
9B6	0	0
8B9	0	0
10M9	0	0
8M5	0	0
8B6	0	0
4B6	0	0
4M3	0	0
3B9	0	0
5B9	0	0
1B5	0	0
8B4	0	0

[107]:
```
p1p2genes = ['luxC', 'group_2734'] # checking candidate genes for phylotype2␣
 ↪within phylogroup1
new2 = pd.DataFrame(phy1a, columns = p1p2genes)
new2
```

[107]:

id	luxC	group_2734
8B3	0	0
3M6	0	0
5B2	0	0
5B8	0	0
4B7	1	1
6M2	1	1

[108]:
```
p1p2genes = ['luxC', 'group_2734'] # checking candidate genes for phylotype2␣
 ↪across the entire database
new2 = pd.DataFrame(data7, columns = p1p2genes)
new2
```

[108]:
```
       luxC  group_2734
id
8B3       0           0
3M6       0           0
5B2       0           0
5B8       0           0
4B7       1           1
6M2       1           1
9B2       0           0
9B6       0           0
8B9       0           0
10M9      0           0
8M5       0           0
8B6       0           0
4B6       0           0
4M3       0           0
3B9       0           0
5B9       0           0
1B5       0           0
8B4       0           0
```

[109]:
```
# filter out phylogroup 2 using data7 which does not include the core genes
phy2a = data7[data7.core_genome_phylogenetic_cluster == 2] #filtering out␣
 ↪phylogroup 2
phy2b = phy2a.loc[:, (phy2a.sum(axis=0) == 2)] #filtering out genes present in␣
 ↪all isolates
```

[110]:
```
phy2b
```

[110]:
```
     phylotype_within_phylogroups  group_1092  group_1108  map  coaD  aroP  \
id
9B2                             1           1           1    1     1     1
9B6                             1           1           1    1     1     1

     csd  group_1746  rplO  folC  …  aml_2  group_3206  group_3207  \
id                                 …
9B2    1           1     1     1  …      1           1           1
9B6    1           1     1     1  …      1           1           1

     group_3208  ybiT_2  group_3210  group_3211  group_489  group_500  nagB_2
id
9B2           1       1           1           1          1          1       1
9B6           1       1           1           1          1          1       1

[2 rows x 1592 columns]
```

[111]:
```
phy2_genes = ['cas3', 'pbpE']
phy2c = pd.DataFrame(phy2b, columns = phy2_genes)
```

[112]:
```
phy2c
```

[112]:
```
     cas3  pbpE
id
9B2     1     1
9B6     1     1
```

[113]:
```
# check if phylogroup 2 genes (cas3) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.cas3)
```

[113]:
```
cas3                              0  1
core_genome_phylogenetic_cluster
1                                 6  0
2                                 0  2
3                                 4  0
4                                 5  0
5                                 1  0
```

[144]:
```
# check if phylogroup 2 genes (pbpE) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.pbpE)
```

[144]:
```
pbpE                              0  1
core_genome_phylogenetic_cluster
1                                 6  0
2                                 0  2
3                                 4  0
4                                 5  0
5                                 1  0
```

[ ]:
```

```

[ ]:
```

```

[115]:
```
# filter out phylogroup 3 using data7 which does not include the core genes
phy3a = data7[data7.core_genome_phylogenetic_cluster == 3] #filtering out␣
 ↪phylogroup 3
phy3b = phy3a.loc[:, (phy3a.sum(axis=0) == 4)] #filtering out genes present in␣
 ↪all isolates
```

[116]:
```
phy3b
```

[116]:
```
      phylotype_within_phylogroups  group_1092  group_1108  map  coaD  aroP  \
id
8B9                              1           1           1    1     1     1
```

```
10M9                              1           1           1    1     1     1
8M5                               1           1           1    1     1     1
8B6                               1           1           1    1     1     1

      csd  group_1746  rplO  folC  …  group_435  group_548  group_549  \
id                                   …
8B9     1           1     1     1  …          1          1          1
10M9    1           1     1     1  …          1          1          1
8M5     1           1     1     1  …          1          1          1
8B6     1           1     1     1  …          1          1          1

      group_578  ykoT_1  group_610  group_62  group_757  ulaF_2  group_857
id
8B9           1       1          1         1          1       1          1
10M9          1       1          1         1          1       1          1
8M5           1       1          1         1          1       1          1
8B6           1       1          1         1          1       1          1

[4 rows x 1586 columns]
```

[117]:
```
phy3_genes = ['nikQ', 'lipM', 'bceA', 'bag']
phy3c = pd.DataFrame(phy3b, columns = phy3_genes)
```

[118]:
```
phy3c
```

[118]:
```
      nikQ  lipM  bceA  bag
id
8B9      1     1     1    1
10M9     1     1     1    1
8M5      1     1     1    1
8B6      1     1     1    1
```

[119]:
```
# check if phylogroup 3 genes (nikQ) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.nikQ)
```

[119]:
```
nikQ                              0  1
core_genome_phylogenetic_cluster
1                                 6  0
2                                 2  0
3                                 0  4
4                                 5  0
5                                 1  0
```

[120]:
```
# check if phylogroup 3 genes (lipM) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.lipM)
```

[120]:
```
lipM                                0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   0  4
4                                   5  0
5                                   1  0
```

[121]:
```
# check if phylogroup 3 genes (bceA) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.bceA)
```

[121]:
```
bceA                                0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   0  4
4                                   5  0
5                                   1  0
```

[122]:
```
# check if phylogroup 3 genes (bag) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.bag)
```

[122]:
```
bag                                 0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   0  4
4                                   5  0
5                                   1  0
```

[ ]:

[123]:
```
# filter out phylogroup 4 using data7 which does not include the core genes
phy4a = data7[data7.core_genome_phylogenetic_cluster == 4] #filtering out␣
 ↪phylogroup 4
phy4b = phy4a.loc[:, (phy4a.sum(axis=0) == 4)] #filtering out genes present in␣
 ↪all isolates
```

[124]:
```
phy4b
```

[124]:

id	map	coaD	csd	glmS	rng	regX3_2	ybaK	amt	nrdE1	group_842	…	\
											…	
4B6	1	1	1	1	1	1	0	1	1	1	…	
4M3	0	0	0	0	0	0	1	0	0	0	…	
3B9	1	1	1	1	1	1	1	1	1	1	…	
5B9	1	1	1	1	1	1	1	1	1	1	…	
1B5	1	1	1	1	1	1	1	1	1	1	…	

```
     group_2097  murA1  uvrY_2  group_2103  group_2105  araQ_16  group_2107  \
id
4B6           0      0       1           0           0        1           1
4M3           1      1       0           1           1        0           0
3B9           1      1       1           1           1        1           1
5B9           1      1       1           1           1        1           1
1B5           1      1       1           1           1        1           1

     group_2109  group_31  accA1
id
4B6           0         0      0
4M3           1         1      1
3B9           1         1      1
5B9           1         1      1
1B5           1         1      1

[5 rows x 162 columns]
```

[133]:
```
phy4_genes = ['tetD', 'murA1', 'amt']
phy4c = pd.DataFrame(phy4b, columns = phy4_genes)
phy4c
```

[133]:
```
     tetD  murA1  amt
id
4B6     0      0    1
4M3     1      1    0
3B9     1      1    1
5B9     1      1    1
1B5     1      1    1
```

[132]:
```
phy4b.loc[:, (phy4b.sum(axis=0) == 4)]
```

[132]:
```
     map  coaD  csd  glmS  rng  regX3_2  ybaK  amt  nrdE1  group_842  …  \
id                                                                      …
4B6    1     1    1     1    1        1     0    1      1          1  …
4M3    0     0    0     0    0        0     1    0      0          0  …
3B9    1     1    1     1    1        1     1    1      1          1  …
5B9    1     1    1     1    1        1     1    1      1          1  …
1B5    1     1    1     1    1        1     1    1      1          1  …

     group_2097  murA1  uvrY_2  group_2103  group_2105  araQ_16  group_2107  \
id
4B6           0      0       1           0           0        1           1
4M3           1      1       0           1           1        0           0
3B9           1      1       1           1           1        1           1
5B9           1      1       1           1           1        1           1
```

```
1B5           1      1       1           1           1        1           1

     group_2109  group_31  accA1
id
4B6           0         0      0
4M3           1         1      1
3B9           1         1      1
5B9           1         1      1
1B5           1         1      1

[5 rows x 162 columns]
```

[127]:
```
# check if phylogroup 4 genes (tetD) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.tetD)
```

[127]:
```
tetD                                0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   4  0
4                                   1  4
5                                   1  0
```

[128]:
```
# check if phylogroup 4 genes (amt) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.murA1)
```

[128]:
```
murA1                               0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   4  0
4                                   1  4
5                                   1  0
```

[146]:
```
p4p1genes = ['chuR'] # checking candidate genes for phylotype1 across the␣
 ↪entire database
new1 = pd.DataFrame(data7, columns = p4p1genes)
new1
```

[146]:
```
      chuR
id
8B3      0
3M6      0
5B2      0
5B8      0
4B7      0
6M2      0
```

```
9B2      0
9B6      0
8B9      0
10M9     0
8M5      0
8B6      0
4B6      1
4M3      0
3B9      0
5B9      0
1B5      0
8B4      0
```

[135]:
```
p4p2genes = ['adaB', 'afuB_1', 'agaA_1'] # checking candidate genes for␣
 ↪phylotype2 across the entire database
new2 = pd.DataFrame(data7, columns = p4p2genes)
new2
```

[135]:

id	adaB	afuB_1	agaA_1
8B3	0	0	0
3M6	0	0	0
5B2	0	0	0
5B8	0	0	0
4B7	0	0	0
6M2	0	0	0
9B2	0	0	0
9B6	0	0	0
8B9	0	0	0
10M9	0	0	0
8M5	0	0	0
8B6	0	0	0
4B6	0	0	0
4M3	1	1	1
3B9	0	0	0
5B9	0	0	0
1B5	0	0	0
8B4	0	0	0

[136]:
```
p4p3genes = ['aadK', 'acoC', 'agaSK_1'] # checking candidate genes for␣
 ↪phylotype1 across the entire database
new3 = pd.DataFrame(data7, columns = p4p3genes)
new3
```

[136]:

id	aadK	acoC	agaSK_1
8B3	0	0	0

```
3M6      0     0        0
5B2      0     0        0
5B8      0     0        0
4B7      0     0        0
6M2      0     0        0
9B2      0     0        0
9B6      0     0        0
8B9      0     0        0
10M9     0     0        0
8M5      0     0        0
8B6      0     0        0
4B6      0     0        0
4M3      0     0        0
3B9      1     1        1
5B9      1     1        1
1B5      1     1        1
8B4      0     0        0
```

[137]:
```
# filter out phylogroup 5 using data7 which does not include the core genes
phy5a = data7[data7.core_genome_phylogenetic_cluster == 5] #filtering out␣
 ↪phylogroup 5
phy5b = phy5a.loc[:, (phy5a.sum(axis=0) == 1)] #filtering out genes present in␣
 ↪all isolates
```

[138]:
```
phy5b
```

[138]:
```
     phylotype_within_phylogroups  map  coaD  csd  glmS  rng  regX3_2  ybaK  \
id
8B4                             1    1     1    1     1    1        1     1

     amt  nrdE1  …  group_9990  group_9991  group_9992  group_9993  \
id                …
8B4    1      1  …           1           1           1           1

     group_9994  group_9995  group_9996  group_9997  group_9998  group_9999
id
8B4           1           1           1           1           1           1

[1 rows x 2083 columns]
```

[139]:
```
phy5_genes = ['cse1_2', 'btuD']
phy5c = pd.DataFrame(phy5b, columns = phy5_genes)
phy5c
```

[139]:
```
     cse1_2  btuD
id
8B4       1     1
```

[142]:
```
# check if phylogroup 5 genes (cse1_2) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.cse1_2)
```

[142]:
```
cse1_2                              0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   4  0
4                                   5  0
5                                   0  1
```

[143]:
```
# check if phylogroup 5 genes (btuD) are not shared with other phylogroups
pd.crosstab(data7.core_genome_phylogenetic_cluster, data7.btuD)
```

[143]:
```
btuD                                0  1
core_genome_phylogenetic_cluster
1                                   6  0
2                                   2  0
3                                   4  0
4                                   5  0
5                                   0  1
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]: