



# PHYLOGENETIC PIPELINE

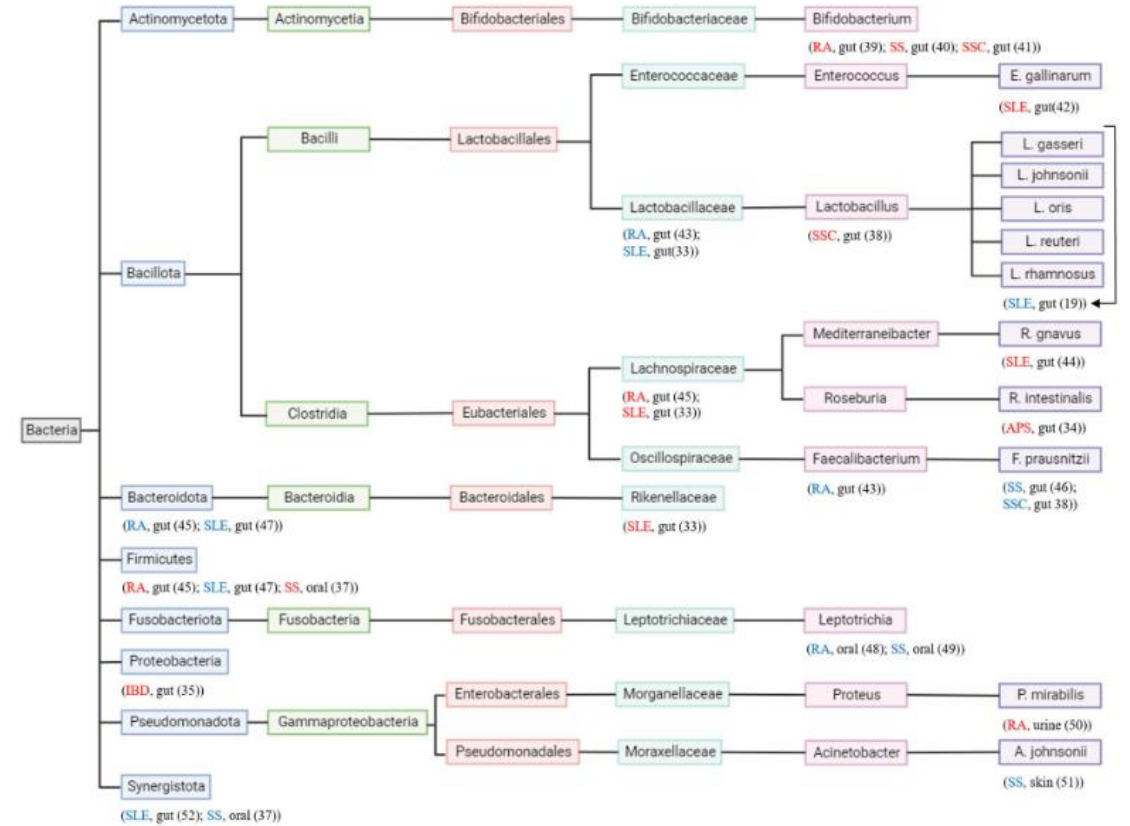
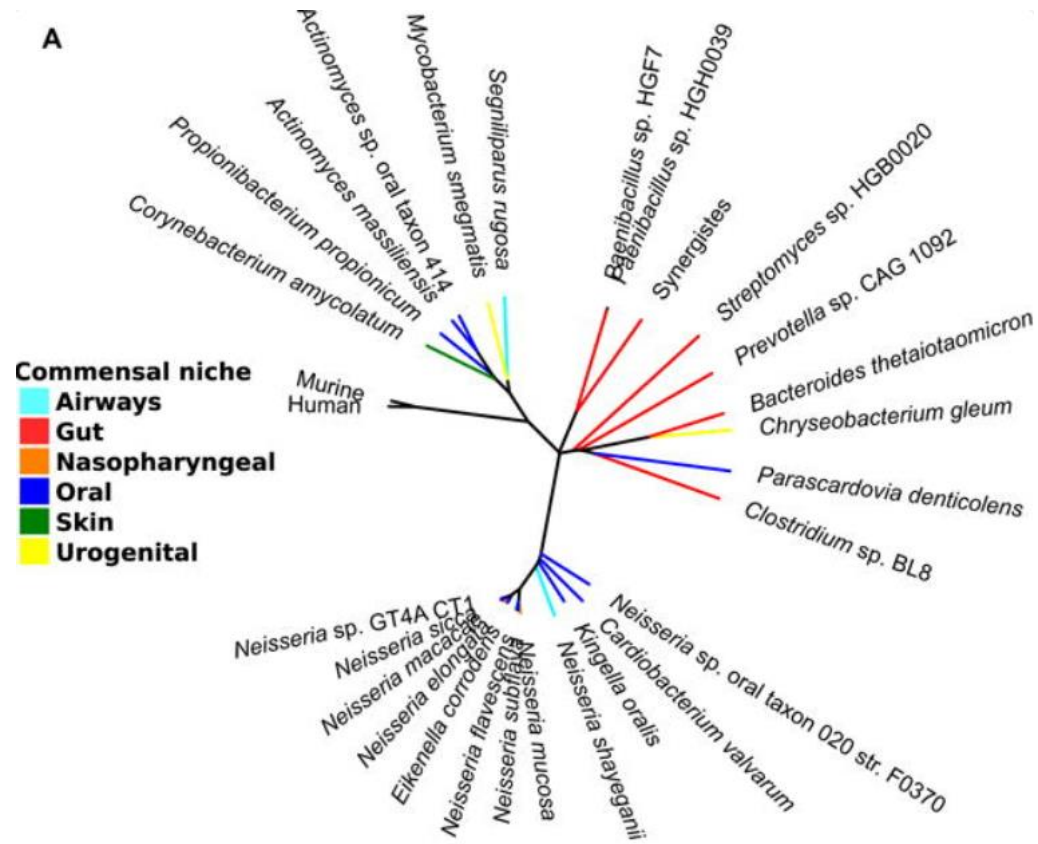
Jakub Giezgala

# HARMONOGRAM

1. Motywacja
2. Pipeline
3. Wyniki
4. Refleksje

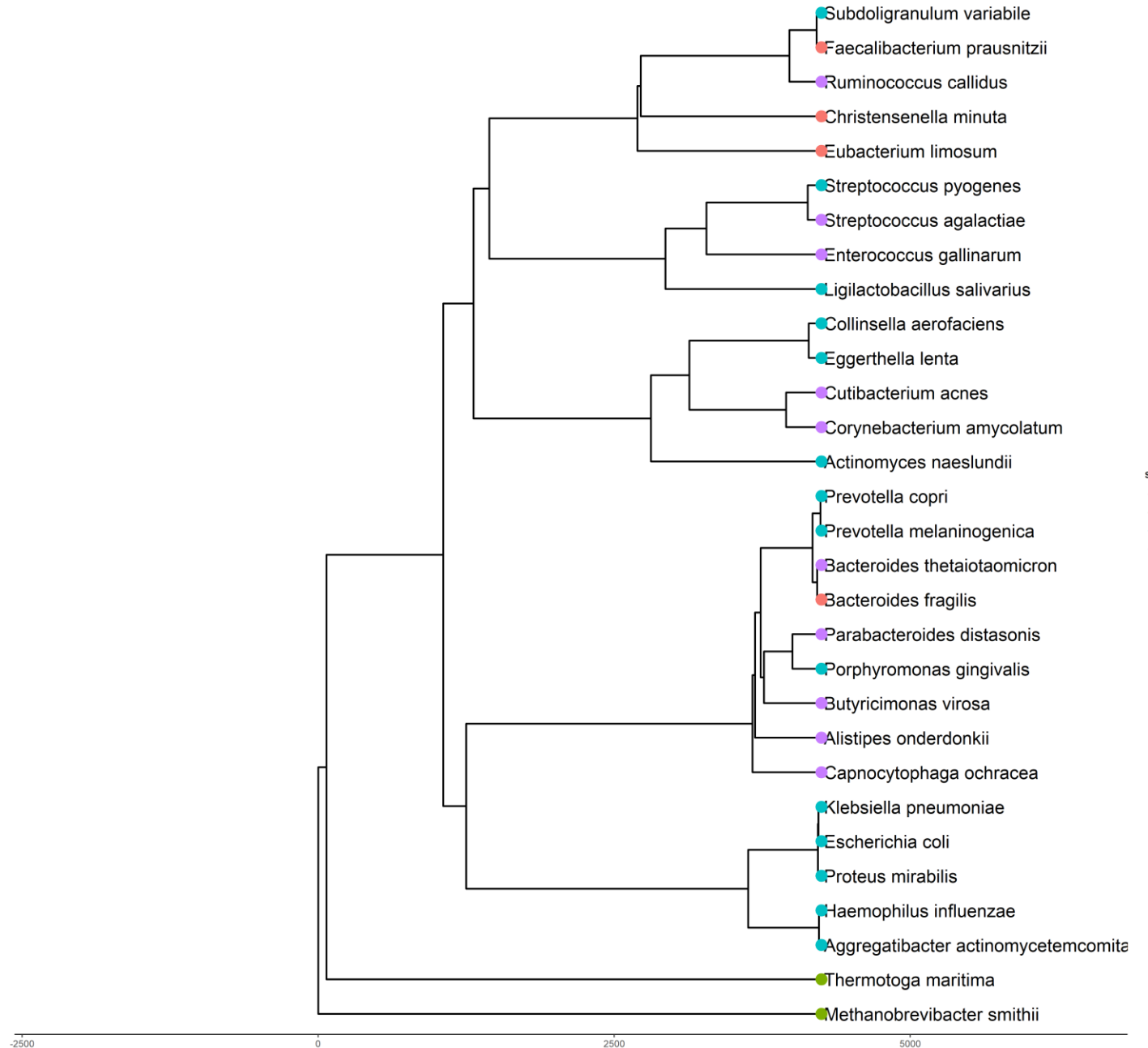


# WPROWADZENIE





# WYBÓR ORGANIZMÓW



# PIPELINE

Pobranie sekwencji z bazy danych

Klastrowanie (+ Korekcja)

Multiple Sequence Alignment

Inferencja drzew metodą NJ/ML (ewentualnie ze wsparciem)

Obliczenia drzewa konsensusowego i superdrzewa (ewentualnie dla paralogów)

Do pobrania  
wszystkich,  
wybranych,  
zwalidowanych  
proteomów  
wykorzystuję NCB  
I Datasets CLI.

## POBRANIE SEKWENCJI

```
1 id,species,strain,accession,bioproject,type,source
2 G01,Methanobrevibacter_smithii,OF4,GCF_000016525.1,PRJNA224116,Outgroup,Control
3 G02,Bacteroides_thetaiotaomicron,VPI_5482,GCF_000011065.1,PRJNA224116,Bacteroidetes,SLE
4 G03,Bacteroides_fragilis,NCTC_9343,GCF_000025985.1,PRJNA224116,Bacteroidetes,Commensal
5 G04,Prevotella_copri,FDAARGOS_1573,GCF_020735445.1,PRJNA224116,Bacteroidetes,RA
6 G05,Prevotella_melaninogenica,ATCC_25845,GCF_000144405.1,PRJNA31383,Bacteroidetes,RA
7 G06,Parabacteroides_distasonis,ATCC_8503,GCF_000012845.1,PRJNA13485,Bacteroidetes,SLE
8 G07,Butyrivibrio_virosa,DSM_23226,GCF_025148635.1,PRJNA746600,Bacteroidetes,SLE
9 G08,Porphyrimonas_gingivalis,ATCC_33277,GCF_000010505.1,PRJDA19051,Bacteroidetes,RA
10 G09,Capnocytophaga_ochracea,DSM_7271,GCF_000023285.1,PRJNA29403,Bacteroidetes,SLE
11 G10,Alistipes_nderdonkii,3BBH6,GCF_006542645.1,PRJDB8150,Bacteroidetes,SLE
12 G11,Cutibacterium_acnes,NBRC_107605,GCF_006739385.1,PRJDB8042,Actinobacteria,SLE
13 G12,Corynebacterium_amycolatum,FDAARGOS_1189,GCF_016889425.1,PRJNA231221,Actinobacteria,SLE
14 G13,Eggerthella_lenta,APC055-529-1D,GCF_021378605.1,PRJNA787074,Actinobacteria,RA
15 G14,Collinsella_aerofaciens,JCM_10188,GCF_010509075.1,PRJNA603246,Actinobacteria,RA
16 G15,Actinomyces_naeslundii,FDAARGOS_1037,GCF_016127855.1,PRJNA231221,Actinobacteria,RA
17 G16,Enterococcus_gallinarum,EG81,GCF_021496385.1,PRJNA622390,Firmicutes,SLE
18 G17,Ruminococcus_callidus,ATCC_27760,GCF_049562855.1,PRJNA1247524,Firmicutes,SLE
19 G18,Subdoligranulum_variabile,DSM_15176,GCF_025152575.1,PRJNA746600,Firmicutes,RA
20 G19,Eubacterium_limosum,ATCC_8486,GCF_000807675.2,PRJNA270275,Firmicutes,Commensal
21 G20,Faecalibacterium_prausnitzii,M21/2,GCF_000154385.1,PRJNA18203,Firmicutes,Commensal
22 G21,Ligilactobacillus_salivarius,B4311,GCF_035231985.1,PRJNA932943,Firmicutes,RA
23 G22,Streptococcus_pyogenes,NCTC12064,GCF_900475035.1,PRJEB6403,Firmicutes,RA
24 G23,Streptococcus_agalactiae,NGBS128,GCF_001552035.1,PRJNA293561,Firmicutes,SLE
25 G24,Christensenella_minuta,DSM_22607,GCF_003628755.1,PRJNA374621,Firmicutes,Commensal
26 G25,Escherichia_coli,K-12_substr.MG1655,GCF_000005845.2,PRJNA225,Proteobacteria,RA
27 G26,Proteus_mirabilis,HI4320,GCF_000069965.1,PRJNA12624,Proteobacteria,RA
28 G27,Klebsiella_pneumoniae,HS11286,GCF_000240185.1,PRJNA78789,Proteobacteria,RA
29 G28,Aggregatibacter_actinomycetemcomitans,PA7,GCF_000017205.1,PRJNA16720,Proteobacteria,RA
30 G29,Haemophilus_influenzae,FDAARGOS_1560,GCF_020736045.1,PRJNA231221,Proteobacteria,RA
31 G30,Thermotoga_maritima,JGI,GCF_000230655.2,PRJNA63051,Outgroup,Control
```

# KLASTROWANIE

MMseqs2 (Many-  
against-Many  
sequence  
searching)

Tryb: easy-  
linclust **--min-  
seq-id 0.25 -c  
0.6 --cov-mode  
0**

Wszystkich  
klastrów: 21474,  
ale 12958 to  
singletony

187 do  
Konsensusu (z  
Paralogami 199)

# MAFFT & trimAl



MAFFT : --AUTO



TRIMAI: --QUIET (ML)



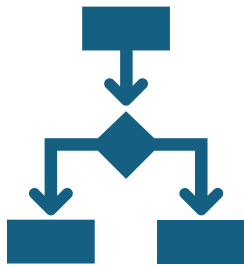
# INFERENCJA DRZEW

- FastTree: -fastest (NNI, S-H Test)
- Po przefiltrowaniu  $> 0.85$  zostaje jedynie 1784 drzewa





# DRZEWO GENOMÓW



Konsensus większościowy (--  
minsup 0.50) obliczony za  
pomocą IQ-TREE



Superdrzewo obliczone za  
pomocą programu Fasturec (-Y)



Ukorzenie drzewa:  
*Methanobrevibacter smithii*  
(Archea)

# WNIOSKI & WIZUALIZACJA

