



Ministerstwo
Zdrowia



NARODOWY
INSTYTUT
ZDROWIA
PUBLICZNEGO
PAŃSTWOWY INSTYTUT
BADAWCZY



Współfinansowane przez
Unię Europejską

WP8 – narzędzia bioinformatyczne w projekcie Plepiseq

Michał Łaźniewski, Michał Kadlof

„Towards the National Platform for Genomic Surveillance” PLEpiSeq is co-funded by the European Union and the Ministry of Health of the Republic of Poland.

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

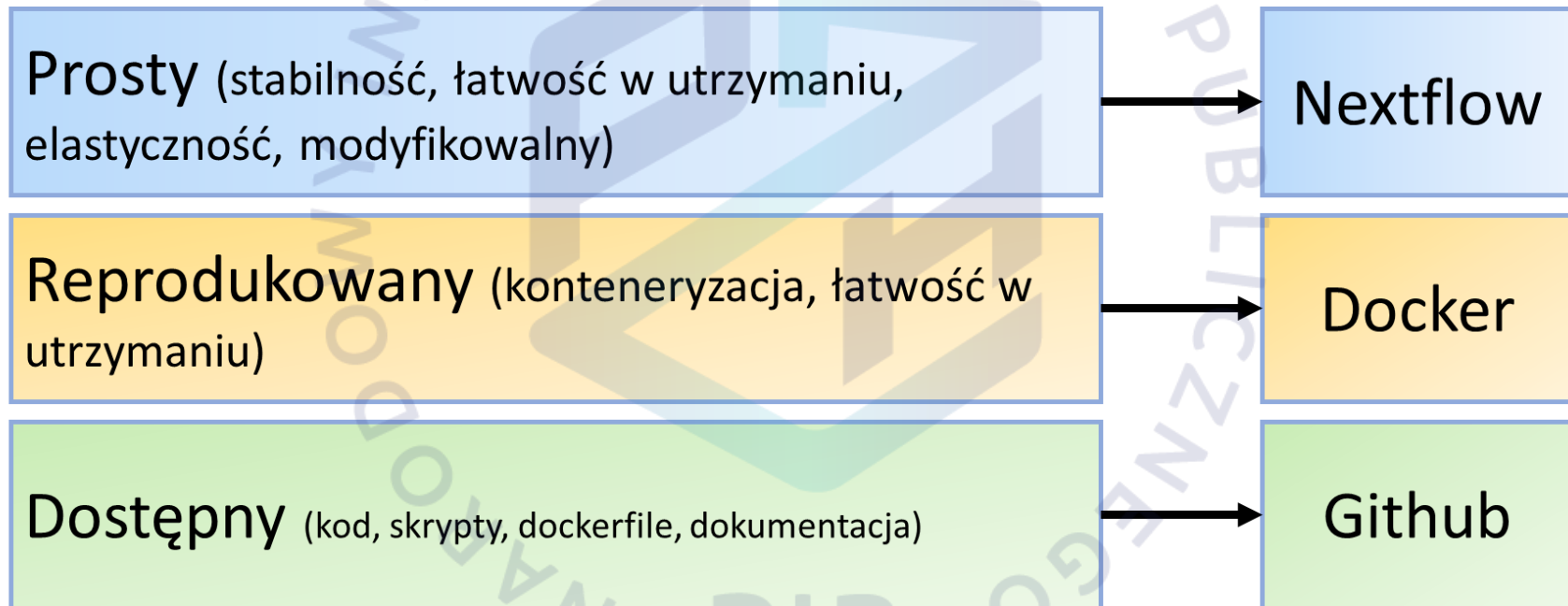
„Towards the National Platform for Genomic Surveillance” PLEpiSeq współfinansowany ze środków Unii Europejskiej oraz Ministerstwa Zdrowia Rzeczypospolitej Polskiej.

Wyrażone poglądy i opinie są jednak wyłącznie poglądami autora (autorów) i niekoniecznie odzwierciedlają poglądy Unii Europejskiej lub Europejskiej Agencji Wykonawczej ds. Zdrowia i Cyfryzacji. Unia Europejska ani organ przyznający pomoc nie mogą być za nie pociągnięte do odpowiedzialności.

Konspekt

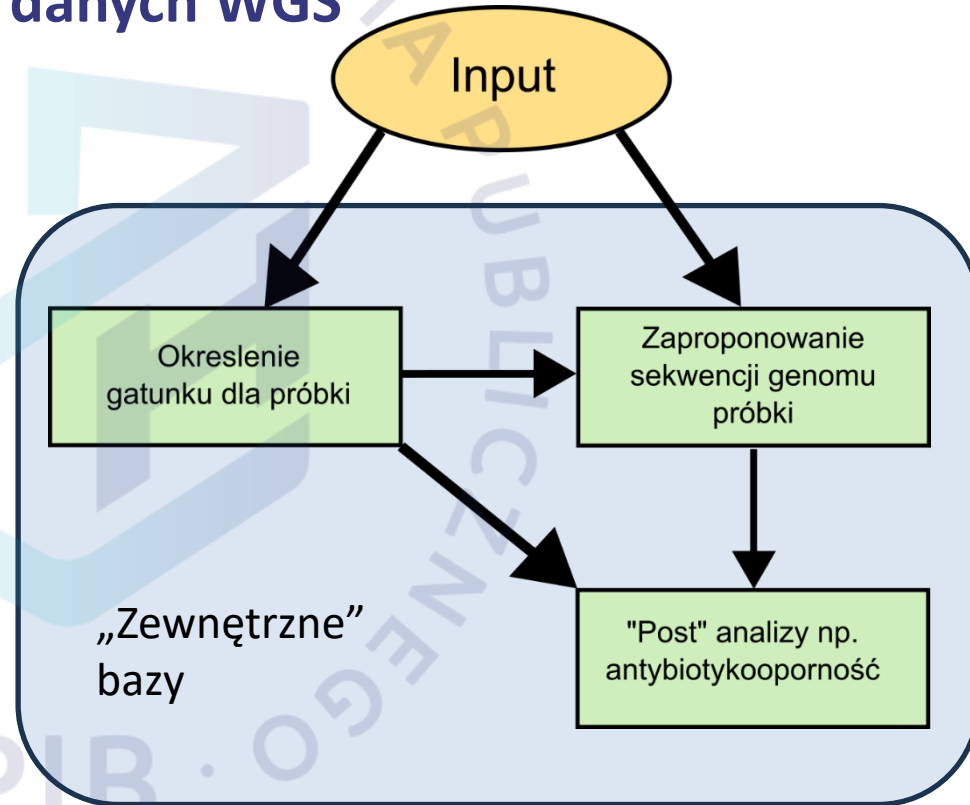
- Jak działają nasze programy - założenia?
- Jakie są dane wejściowe?
- Jakie dane generują programy ?
- Czy można „ufać” wynikom programów ?
- Czemu niektóre próbki mogą nie zostać przeanalizowane ?

Jak działają nasze programy? Założenia



Jak programy działają ? Analiza danych WGS

- **Input**
 - Plik(i) w formacie fastq.gz
 - Technologia sekwencjonowania
 - Informacja o primer'ach dla sekwencjonowania wirusów
 - Nazwa rodzajowa sekwencjonowanego organizmu
- **Określenie gatunku**
 - Weryfikacja jakości sekwencjonowania
 - Weryfikacja wskazanego organizmu
- **Zaproponowanie sekwencji genomu**
 - De novo w przypadku bakterii
 - Oparte o referencje w przypadku wirusów
- **„Post” analizy**
 - Wszystkie analizy jakie można przeprowadzić mając genom analizowanego organizmu

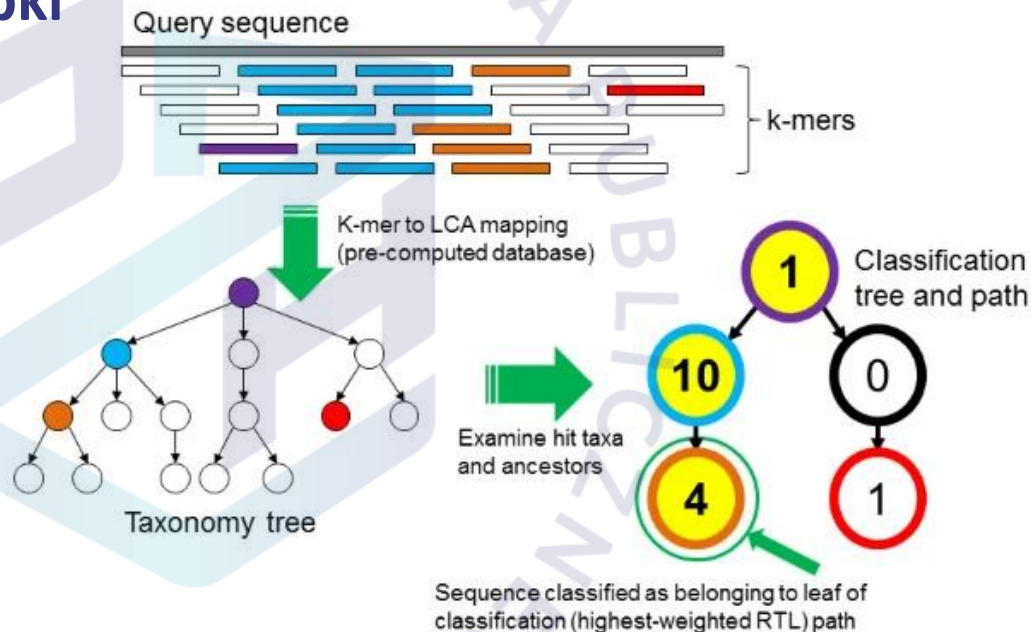


Input - Analiza danych WGS

- Plik(i) w formacie fastq.gz
- Technologia sekwencjonowania
 - Illumina – pair-end sequencing
 - Nanopore – single-end
- Nazwa organizmu
 - SARS-CoV-2
 - RSV (typ wykrywany automatycznie)
 - Influenza (typ i podtyp wykrywane automatycznie)
 - Salmonella, Escherichia, Campylobacter
- Informacja o primer'ach dla sekwencjonowania wirusów
 - Nieistotny w przypadku bakterii oraz wirusa grypy

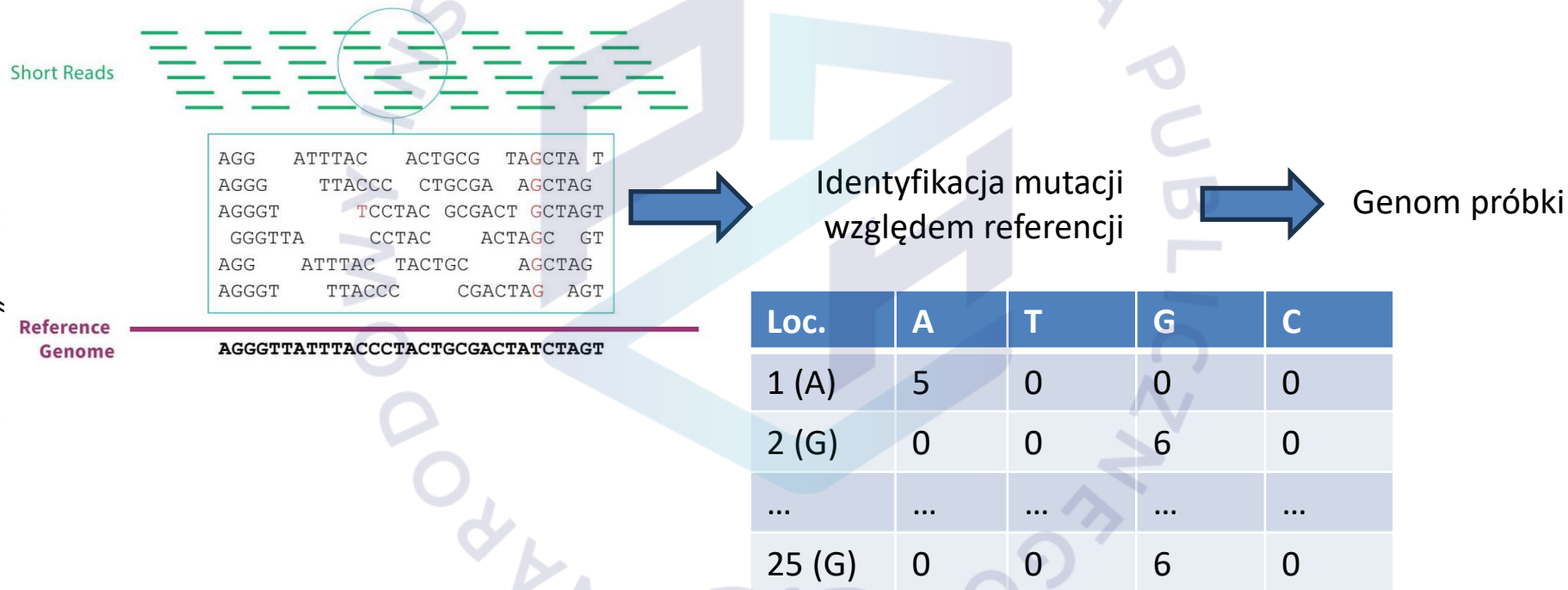
Określenie gatunku dla próbki

- Kraken2 -> próbki wirusowe
- Kraken2, Metaphlan4 i kmerfinder -> bakterie



<https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-3-r46/figures/1>

Proponowanie sekwencji genomu próbki



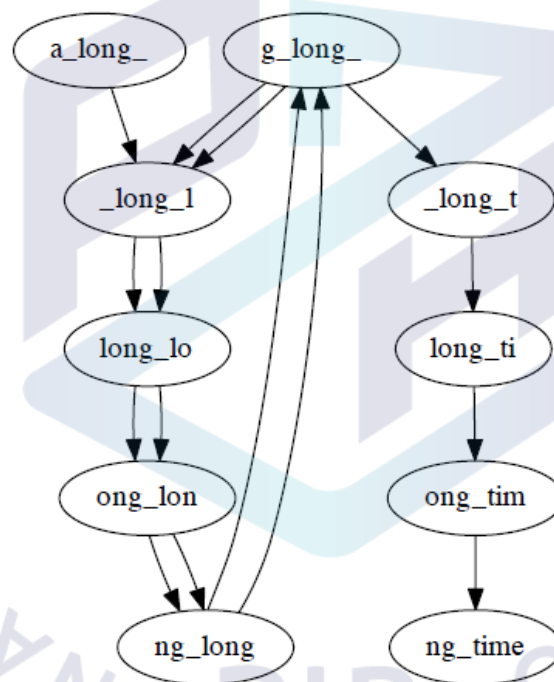
Sekwencjowanie

a_long_long_long
ng_long_l
g_long_time

6-mery

a_long_
_long_l
long_lo
ong_lon
...

Graf
de Bruijn'a



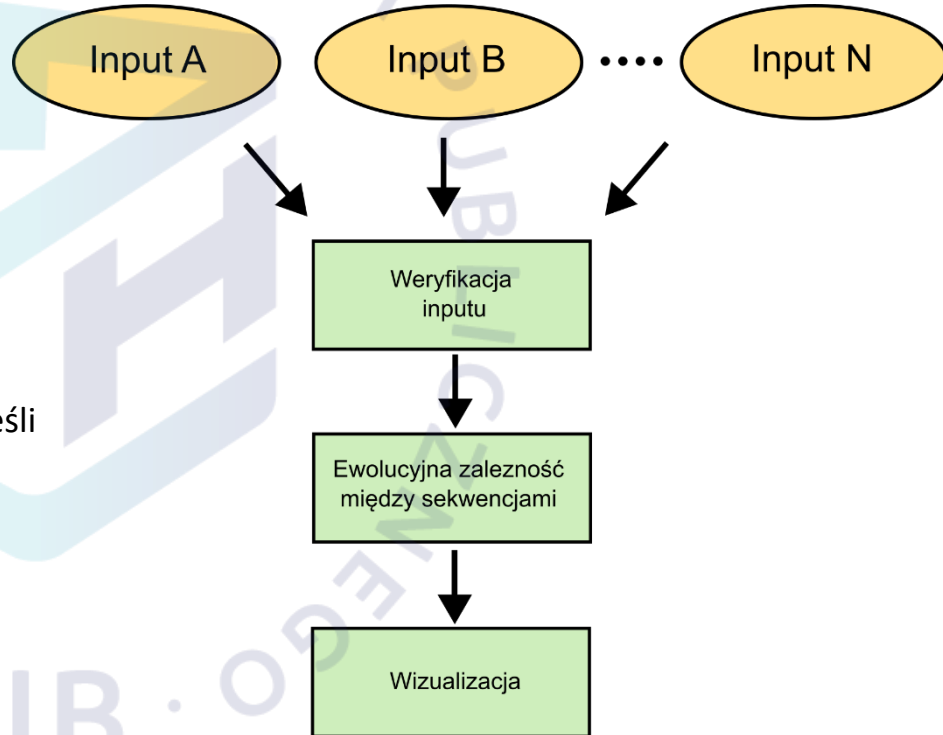
Ścieżka
Eulere'a

a_long_long_long_time

Kategoria danych	Opis
Genom	Próbki wirusowe ilość plików taka jak ilość segmentów, dla bakterii jeden plik z contig’ami
Obecność „zanieczyszczeń”	Informacja o najczęściej występujących rodzajach organizmów w próbce i ich procentowym udziale
Jakość sekwencjonowania	Podstawowe dane o sekwencjonowaniu (ilość odczytów, średnia jakość zasad, średnia długość itd. ...)
Jakość przewidywania genomu	Długość genomu, ilość pozycji o niskim pokryciu, L50, N50, ilość contig’ów
Klasyfikacja wirusów	Typ/podtyp, identyfikacja linii Pango i/lub Nextstrain
Klasyfikacja bakterii	Antygeny, klasyfikacja według schematów MLST i cgMLST
Antybiotykooporność	Oporność na antybiotyki i geny oporności stwierdzone w próbce
Plazmidy	Obecność plazmidów
Dane strukturalne	Struktury wybranych białek (np HA i NA dla grypy)
Dane unikalne dla danego organizmu	Dane uzyskiwane tylko dla jednego organizmu np. oporność na leki przeciwwirusowe dla próbek grypy

Jak programy działają ? Analiza filogenetyczna

- **Input**
 - Pliki w formacie fasta
 - Plik tekstowy z metadanymi
- **Weryfikacja inputu**
 - Czy analizie podlegają organizmy ?
 - Czy przesłane sekwencje spełniają kryteria jakościowe ?
- **Analiza filogenetyczna**
 - Czy analizowane organizmy są spokrewnione, jeśli tak to w jakim stopniu ?
 - Jak w czasie przebiegała ich ewolucja ?
- **Wizualizacja**
 - Przygotowanie wizualizacji -> Na tym etapie użytkownik ma styczność z programami

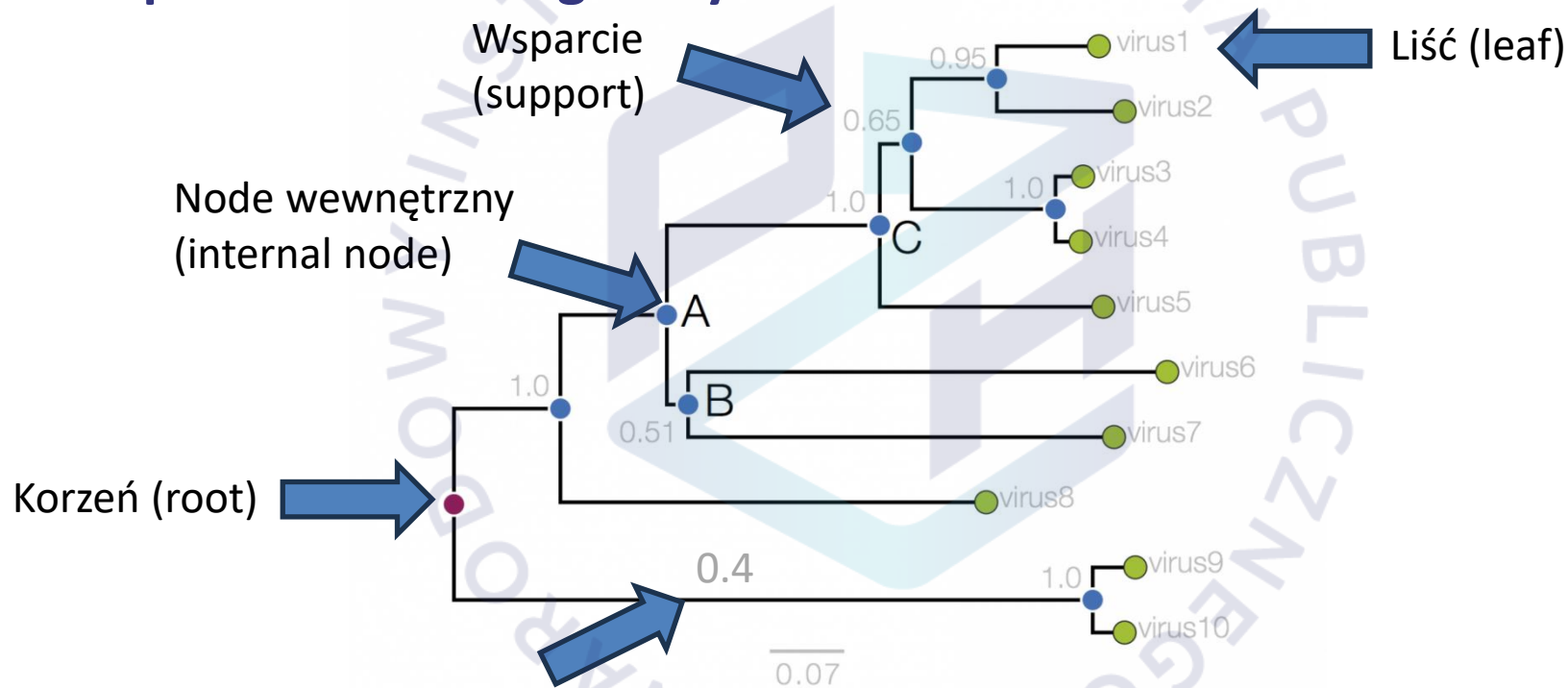


Input - Analiza filogenetyczna

- Pliki w formacie fasta z genomem każdej analizowanej próbki
- Metadane - plik csv, zawierający konkretne kolumny
 - (przykład kolumny dodatkowych informacji może być więcej/mniej)

Kolumna	Opis
strain	Identyfikator próbki – identyczny jak nazwa pliku w formacie dasta
date	Data powiązana z próbka w formacie YYYY-MM-DD
region	Kontynent
country	Kraj
division	Województwo/departament/stan
city	Miasto
Serovar	Serowar np. Enteritidis
MLST	ID bakterii według schematu 7-genomwego
cgMLST	ID bakterii według schematu cgMLST
HC5	Identyfikator klastra do którego należy bakteria przy poziomie odcięcia 5
HC10	Identyfikator klastra do którego należy bakteria przy poziomie odcięcia 10

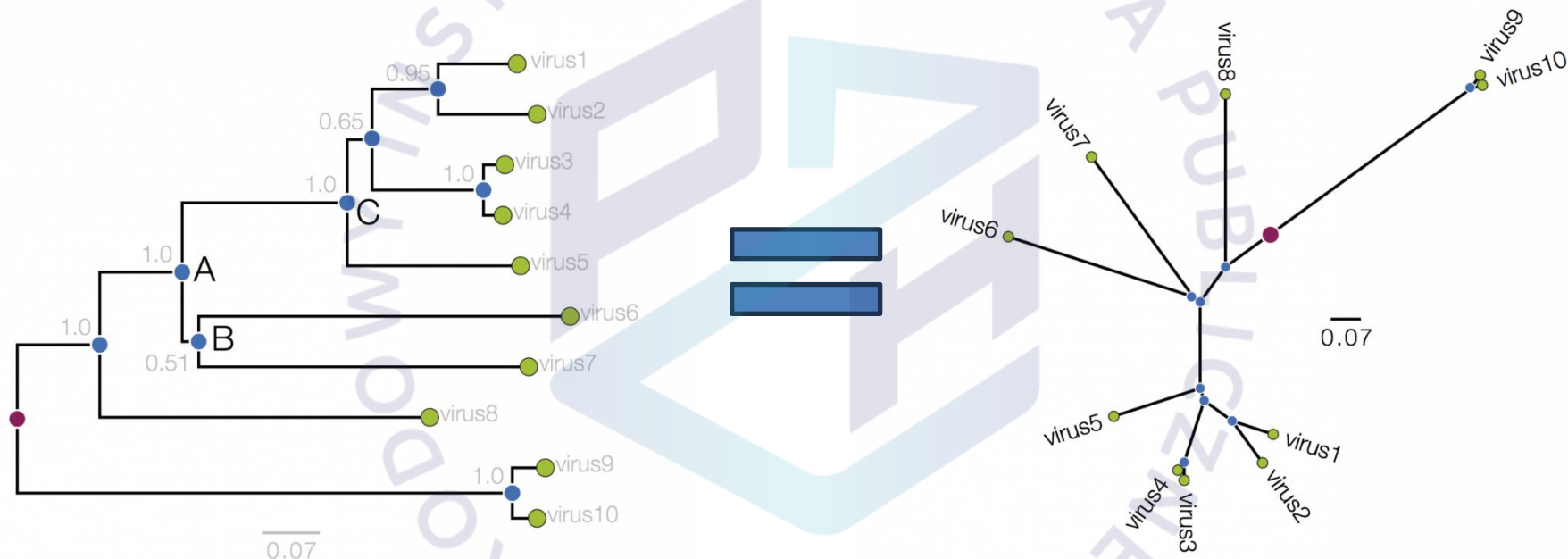
Output - Analiza filogenetyczna



Oczekiwana liczba zmian ewolucyjnych

<https://artic.network/guides/how-to-read-a-tree>

Output - Analiza filogenetyczna



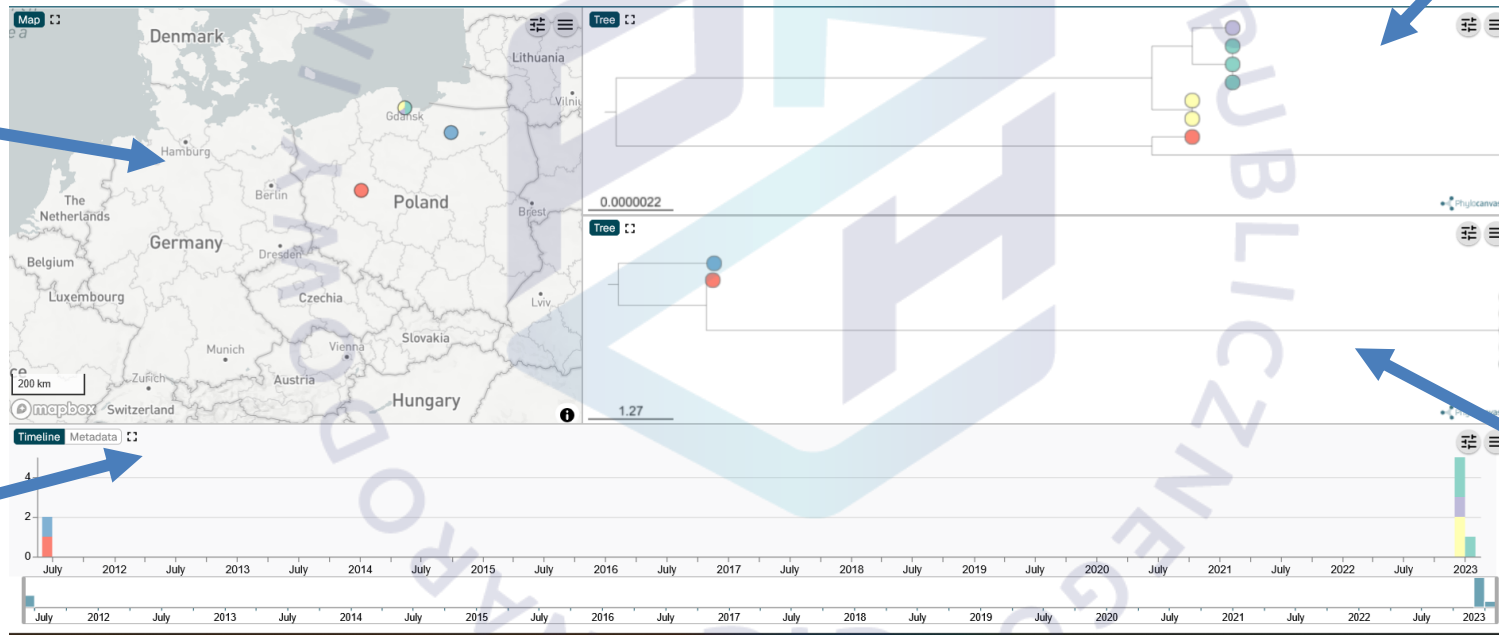
Drzewo ukorzenione

Drzewo nieukorzenione

Output - Analiza filogenetyczna

Jak ewolucyjnie
powiązane są próbki
(filogram)

Gdzie
występują
próbki



Kiedy
występują
próbki

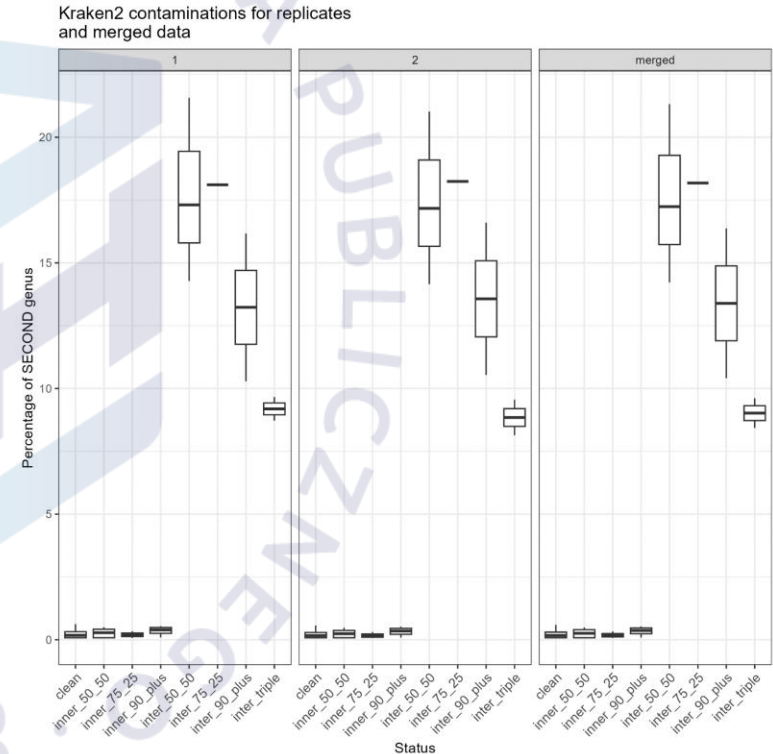
Chronogram
(długość
gałęzi
skalowana
czasem)

Ewaluacja programów

- Zdolność do poprawnej identyfikacji organizmu poddawanego sekwencjonowaniu
- Zdolność do poprawnego przewidywania sekwencji analizowanej próbki
- Zdolność do proponowania odtwarzalnych wyników w ramach tej samej platformy sekwencjonującej
- Zdolność do proponowania poprawnych fenotypów w ramach różnych platform sekwencjonujących
- Zdolność do poprawnej oceny pokrewieństwa próbek

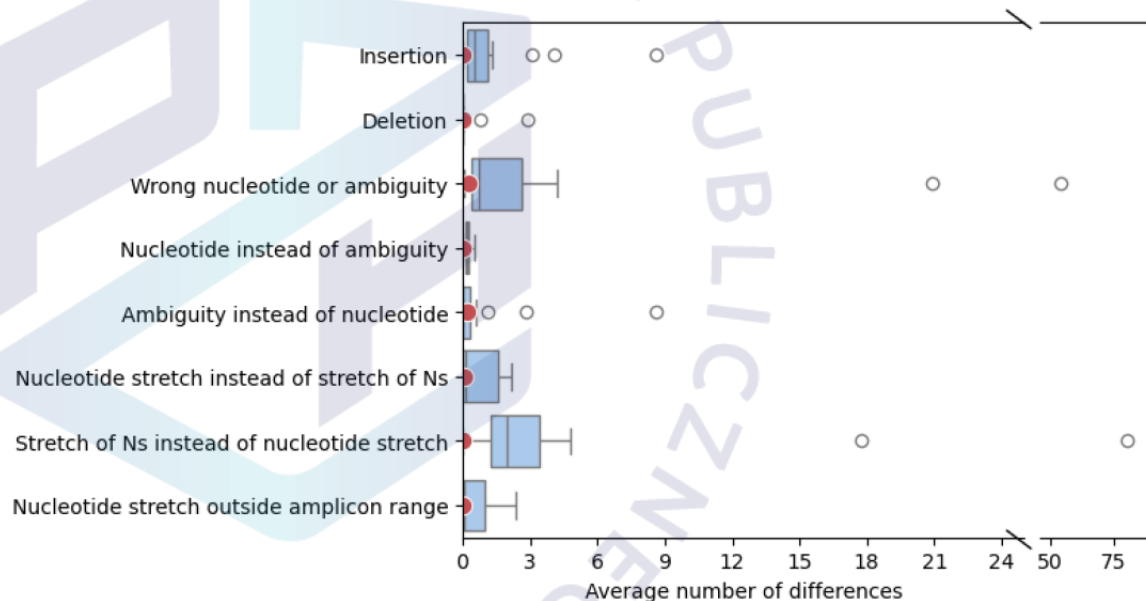
Zdolność do poprawnej identyfikacji organizmu w oparciu o dane WGS

- Dane
 - Wyniki sekwencjonowania 11 próbek w których materiałem była „czysta” kolonia danej bakterii
 - 5 Escherichia
 - 3 Campylobacter
 - 3 Salmonella
 - Wyniki sekwencjonowania 17 próbek w których intencjonalnie zmieszano próbki bakterii z różnych rodzajów (proporcje- 50:50, 75:25, 90:10)
- We wszystkich 11 próbkach „czystych” poprawnie zidentyfikowano właściwy rodzaj
- We wszystkich próbkach zanieczyszczonych zidentyfikowano znaczący udział odczytów z „niepożądanego” rodzaju



Zdolność do poprawnego przewidywania sekwencji analizowanej próbki

- European quality assesment (EQA) 2025
- 12 próbek (pliki fastq) z wirusem SARS-CoV-2
- Platform illumina
- Porównanie sekwencji ze złotym standardem w oparciu o różne kategorie błędów



Zdolność do proponowania odtwarzalnych wyników w ramach tej samej platformy sekwencjonującej

- DNA 12 próbek
 - 6 Escherichia
 - 6 Salmonella
- Sekwencjonowanie illumina
- 2 niezależne laboratoria

Field	Lab A	Lab B
Gatunek	Salmonella enterica	Salmonella enterica
Serotyp	Typhimurium	Typhimurium
Typ (7-genowy)	34	34
Typ (cgMLST)	local_241	local_241
Typ (cgMLST, publiczny)	337784	337784
Dystans do znanego profilu cgMLST	1	1
Liczba brakujących loci	2	2
Liczba zduplikowane loci	74	74
HierCC_L0	local_241	local_241
HierCC_L2	337784	337784
HierCC_L5	337784	337784
HierCC_L10	2	2
HierCC_L20	2	2
Oporność	Amikacin(oporny), Doxycycline(oporny), Minocycline(oporny), Tetracycline(oporny), Tobramycin(oporny)	Amikacin(oporny), Doxycycline(oporny), Minocycline(oporny), Tetracycline(oporny), Tobramycin(oporny)
Długość genomu	5,004,068	5,004,676
Ilość fragmentów	78	45
N50	225,224	319,194
L50	7	5
Średnie pokrycie	55.41	76.34
Liczba odczytów	1,059,384	1,443,215
Mediana jakości odczytów	37.0	35.0
Zanieczyszczenia	95.275% Salmonella 0.15% Escherichia	95.18% Salmonella 0.155% Escherichia

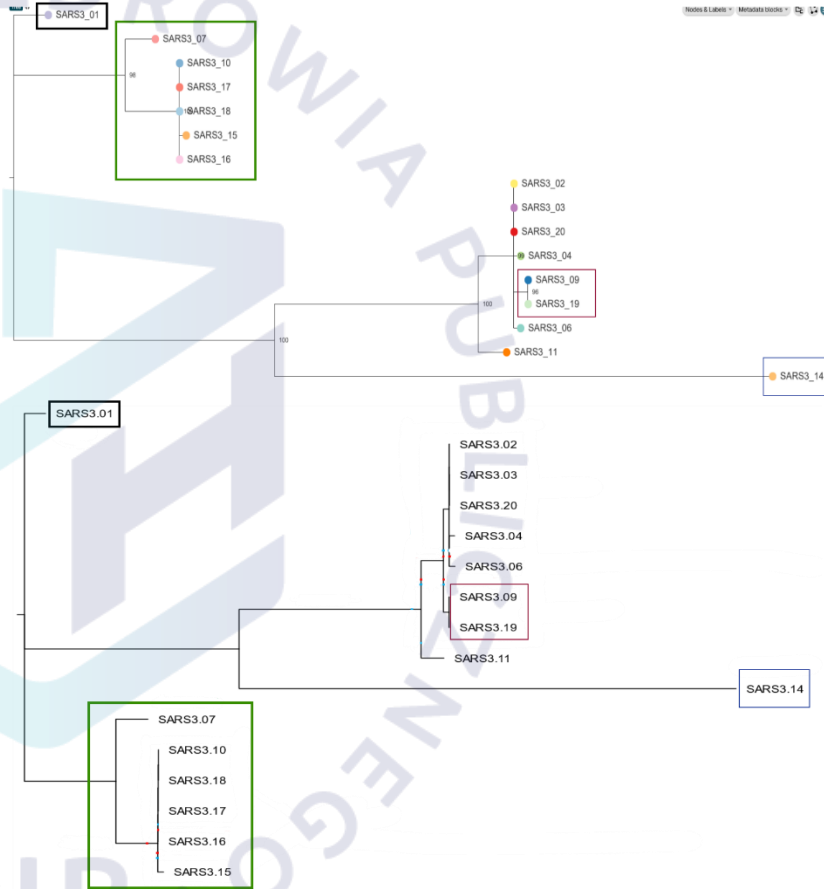
Zdolność do proponowania poprawnego fenotypu

- 6 próbek bakterii z rodzaju *Campylobacter*
- Sekwencjonowanie illumina i nanopore
- Antybiotykoodporność próbek określono z wykorzystaniem metody seryjnych rozcieńczeń
- Wyniki eksperymentalne porównano z wynikami programu do analizy danych WGS

	Illumina	Nanopore	Wet lab
cgMLST	local_41	Local_41	-
Closest external cgMLST	23 885 (197)	23 885 (200)	
AZI (Makrolid)	Tak (23S)	Tak (23S)	>64
CIP (fluorochinolon)	Tak (gyrA)	Tak (gyrA)	32
CLI (linkozamid)	Tak (23S)	Tak (23S)	8
ERY (Makrolid)	Tak (23S)	Tak (23S)	>64
FFN (Florfenicol)	-	-	2
TET (tetracykliny)	Tak (tet(O/32/O))	Tak (tet(O/32/O))	>64
GEN (aminoglikozyd)	Tak (aph(2'')-Ib)	Tak (aph(2'')-Ib)	>32
NAL (chinolon)	Tak (gyrA)	Tak (gyrA)	64

Zdolność do poprawnej oceny pokrewieństwa próbek

- European quality assessment (EQA) 2024
- SARS-CoV-2
- 20 sekwencji (przed oceną jakości)
- Drzewo filogenetyczne zachowuje kluczowe elementy topologii w stosunku do wzorca



EQA 2024

Czemu niektóre próbki mogą nie zostać przeanalizowane?

- Programy mają kilka zabezpieczeń w przypadku „słabego” sekwencjonowania np.
 - Plik fastq jest pusty, ma zły format, zawiera mniej niż 1000 odczytów
 - % odczytów związanych z oczekiwanym organizmem < 50%
 - Teoretyczne pokrycie wynosi < 20
 - Długość genomu składanego *de novo* wynosi < 75% oczekiwanej długości i powyżej 125 %
 - W schemacie MLST (7 genowym) > 2 zduplikowane geny

Kod i dokumentacja

- <https://github.com/mkadlof/plepiseq-wgs-pipeline>
- <https://github.com/mkadlof/plepiseq-phylogenetic-pipeline>
- https://github.com/michallaz/plepiseq_json

Dziękuję za uwagę
mlazniewski@pzh.gov.pl

„Towards the National Platform for Genomic Surveillance” PLEpiSeq is co-funded by the European Union and the Ministry of Health of the Republic of Poland.
Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

„Towards the National Platform for Genomic Surveillance” PLEpiSeq współfinansowany ze środków Unii Europejskiej oraz Ministerstwa Zdrowia Rzeczypospolitej Polskiej.
Wyrażone poglądy i opinie są jednak wyłącznie poglądami autora (autorów) i niekoniecznie odzwierciedlają poglądy Unii Europejskiej lub Europejskiej Agencji Wykonawczej ds. Zdrowia i Cyfryzacji. Unia Europejska ani organ przyznający pomoc nie mogą być za nie pociągnięte do odpowiedzialności.