Week 5

Saturday, 31 July 2021 19:23

BATCH UPLOAD

BACTERIAL ANALYSIS PIPELINE

1) PHENOTYPING
2) TYPING
3) PHYLOGENY

} POSSIBLE SERVICES

BATCH OF UNKNOWN SAMPLES

↓

BATCH UPLOAD

↓

ALL TYPING OF ALL SEQUENCES

- METADATA TEMPLATE
- FILL TEMPLATE
- UPLOAD TEMPLATE
- UPLOAD FILES
- GET RESULTS

BOLD → REQUIRED FIELD METADATA

IF UNASSEMBLED (RAW READS)
YOU HAVE TO FILL

SEQ_PLATFORM
SEQ_TYPE

OUTPUT

SAMPLE SPECIES MLST PLASMIDS

pMLSTs RESISTANCE GENES VIRULENCE GENES

CAN DOWNLOAD .CSV OR .XLS

YOU CAN KNOW WHICH TOOLS WERE USED
YOU CAN REANALIZE USING DIFFERENT
PARAMETERS & TRESHOLD

PHYLOGENETIC RELATEDNESS

CSI PHYLOGENY TOOL

PHYLOGENY DERIVED BY SNPs

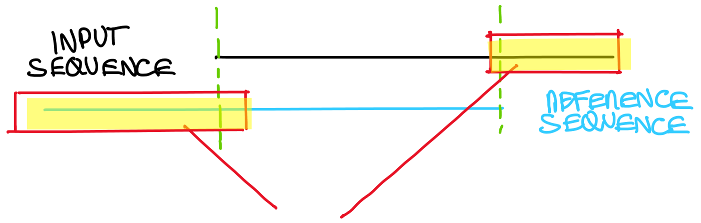
SINGLE
NUCLEOTIDE
POLYMORPHISM

ASSUMPTION SNPs ARE IID
(RANDOM & INDEPENDENT)

SNP CALLING

FINDING DIFFERENCE FROM
REFERENCE SEQUENCE

RAW READS → MAPPING SOFTWARE → REFERENCE SEQUENCE



IGNORING BITS THAT HAVE NO MATCH
BETWEEN THE 2 SEQUENCES

SNP FILTERING

MOBILE ELEMENTS DO NOT SHARE
THE SAME PHYLOGENY

BWA
SAMTOOLS
Z-SCORE →

$$\boxed{Z_{\text{SCORE}} = \frac{X-Y}{\sqrt{X+Y}}}$$

RAW READS ARE BETTER IF POSSIBLE
MUCH EASIER TO VALIDATE SNPs

DEPTH AT LEAST X READS TO COVER
EACH OF YOUR SNP POSITIONS

Z-SCORE USED TO SORT OUT
AMBIGUOUS SNP CALLS

IF $Z = 1.96$ P-VALUE $= 0.05$

IF $z = 3.26$ P-VALUE = 0.01

OUTPUT PHYLOGENY (NEWICK, PDF, PNG)

CAN OPEN .NEWICK W/ FIGTREE

SNP - MATRIX

PSEUDO - ALIGNMENT

QUALITY CONTROL

TIPS
- USE CLOSELY RELATED REFS
- CHECK % OF REF. GENOME COVERED BY ALL ISOLATES

EVERGREEN ONLINE

IDENTIFICATION OF FOODBORNE BACTERIAL OUTBREAKS

USE GENOMIC DATA TO TRACK DISEASES

GENOMIC EPIDEMIOLOGY

DISTANCE MATRIX (HAMMING)

CENTER FOR GENETIC EPIDEMIOLOGY

CLUSTER → PHYLOGENETIC TREES

MULTIPURPOSE DETECTION OF GENETIC MARKERS

My DB FINDER TOOL

QUICK ANALYSIS OF WGS DATA CAN BE USEFUL

DATABASES SOMETIMES DO NOT HAVE YOUR SPECIAL GENE OF INTEREST

GENERATING YOUR OWN DATA

→ USER PROVIDES DB
→ USER PROVIDES INPUT SEQUENCE

ALIGNMENT WITH INTERFACE

MAKING YOUR OWN DATABASE
ONLY DNA NOT PROTEIN SEQ
ONLY FASTA FORMAT

INPUT NUCLEOTIDES SEQ IN FASTA

SEPARATION HEADER USING SPACES

IDENTITY %ID TRESHOLD
MINIMUM LENGTH OF ALIGNMENT

FASTA HEADER

IDENTITY

CONTIG

QUERY/TEMPLATE LENGTH

POSITION IN CONTIG

YOU CAN SEE ALIGNMENT
MISSING BASE PAIR