metadata fields from 50 illumina files for virus beacon schema v1

–FILE (metadata from fastq not vcf)

- RUN

□ PRIMARY_ID e.g "SRR10903401" (SRR)

□ total_spots e.g "476632" "676694" (50)

□ total_bases e.g ”143565674" "203832203" (50)

□ Statistics

  □ read_ave e.g "150.55" (39)

  □ read_stdev e.g "0.74”, "53.60”, "12.44" (39)

□ total_bases e.g "143565674”, "203832203" "8031043214" "8325534"

□ size e.g "72426963" "104687344" "2743427127" "2158046"

–EXPERIMENT

- SAMPLE
- EXPERIMENT

□ IDENTIFIERS

  □ PRIMARY_ID eg. "SRA1041081" (13 experiments)

□ TITLE (string e.g ”Total RNA sequencing of BALF (human reads removed)")

□ DESIGN

  □ LIBRARY_DESCRIPTOR

    □ LIBRARY_STRATEGY (“RNA-Seq”, "WGS”, "AMPLICON”, "Targeted-Capture”)

    □ LIBRARY_SOURCE (“METATRANSCRIPTOMIC”, "METAGENOMIC”, "GENOMIC" , "VIRAL RNA”)

    □ LIBRARY_SELECTION ("RANDOM”, "RT-PCR”, "RANDOM PCR”, "unspecified”, "PCR”, "cDNA”)

    □ LIBRARY_LAYOUT ("PAIRED" "SINGLE”)

□ PLATFORM

ILLUMINA INSTRUMENT MODEL (“Illumina MiSeq”, “Illumina MiniSeq” , ”Illumina HiSeq 2500” ,”NextSeq 500" , ”NextSeq 550”, "Illumina iSeq 100")

–BIOSAMPLE

- SAMPLE

□ IDENTIFIERS

  □ PRIMARY_ID (=sample accession) e.g "SRS6007144" (all SRS)

    - □ EXTERNAL_ID e. g "SAMN13872787" (all SAMN)
- □ SAMPLE_NAME
    - □ TAXON_ID e.g "9606" ("9606”, ”433733”, “2697049")
    - □ SCIENTIFIC_NAME e.g "Homo sapiens" *(Is this an error? they are citing host but shouldn’t it be the sequence source, i.e the virus)
      ("Homo sapiens”, "human lung metagenome”, “Severe acute respiratory syndrome coronavirus 2”, “Wuhan seafood market pneumonia virus")
- □ SAMPLE_ATTRIBUTES
    - □ tissue/ isolation_source (biosample type) e.g "Bronchoalveolar lavage fluid”, "oropharyngeal swab”, "passage" > Map to UBERON ontology
    - □ culture_collection/ Laboratory Host ("FDA:FDAARGOS_983”, "Vero E6 cells (CRL-1586)") > Map to CL ontology
    - □ passage_history e.g "Original (not passaged)"
    - □ collection_date (different formats) e.g "02-Jan-2020”, "2020-02-14" , "2020”, "2020-03"

–HOST/INDIVIDUAL

- SAMPLE
- □ SAMPLE_ATTRIBUTES
    - □ host (Species) e.g "Homo sapiens"
    - □ age/host_age e.g “21”
    - □ sex/host_sex “female”, “male”
    - □ geo_loc_name/ country/ at_lon (different formats) e.g "USA:WI:Madison”/ "USA: CA, San Diego County”/ “30.52 N 114.31 E" > harmonise, map to GAZ ontology
    - □ host_disease ("nCoV pneumonia”, "COVID-19" , "severe acute respiratory syndrome")
    - □ host_disease_outcome ("Survived")
- SAMPLE (INFO?)
- □ SAMPLE_ATTRIBUTES
    - □ link_addit_analys (article)(“https://wwwnc.cdc.gov/eid/article/26/6/20-0516_article”)

– VIRUS

- SAMPLE
- □ SAMPLE_ATTRIBITES
    - □ strain (22) e.g "2019-nCoV/USA-WI1/2020"