virus beacon schema v1

NOTE: (Metadata fields should be extracted from XML files. See metadata fields from illumine files to feed virus beacon schema v1)

–VARIANT BASIC (basic beacon variant schema)

□ ref_assembly:
□ start_nucleotide
□ end_nucleotide
□ ref
□ alt

–VARIANT ANNOTATION

*(Metadata for Variant Annotation is not on XML, has to come from VCF and virus annotation file)

□ variant_id (optional)
□ region: 5UTR,ORF1ab, S, ORF3a, Intergenic, E,M, ORF6, ORF7a, ORF8, N, ORF10, 3UTR
□ variant_type: missense variant .. (SO variant type ontology)

–VARIANT IN SAMPLE

*(Metadata for Variant in Sample, except Info, comes from VCF not XML)

□ variant_id (ours, also global if it exists)
□ biosample_id: e.g "SRS6007144"
□ host_id:
□ variant_file_id: (run id) e.g "SRR10903401"
□ variant_frequency_dataset (dataset):
□ variant_frequency_accross (all data available):

NOTE> variant_frequency can be calculated also using filters such as country, etc, and displayed upon query by variant?

□ INFO
  - □ study_info:
    - □ study_id: (study accession): e.g "SRP242226"
    - □ study_ref: (article PUMED ID)
  - □ experiment_info
    - □ exp_id (experiment accession): e.g "SRX7571571"

  - □ exp_title: e.g "Total RNA sequencing of BALF (human reads removed)"
  - □ exp_lib_strategy: ("RNA-Seq", "WGS", "AMPLICON", "Targeted-Capture")
  - □ exp_lib_source: ("METATRANSCRIPTOMIC", "METAGENOMIC", "GENOMIC" , "VIRAL RNA")
  - □ exp_lib_selection: ("RANDOM", "RT-PCR", "RANDOM PCR", "unspecified", "PCR", "cDNA")
  - □ exp_lib_layout: ("PAIRED" "SINGLE")
  - □ exp_platform: ("Illumina MiSeq", "Illumina MiniSeq" , "Illumina HiSeq 2500" ,"NextSeq 500" , "NextSeq 550", "Illumina iSeq 100")

–BIOSAMPLE

- □ biosample_id: e.g "SRS6007144"
- □ biosample_alt_id: e.g "SAMN13872787"
- □ biosample_type: e.g "Bronchoalveolar lavage fluid", "oropharyngeal swab", "passage" > Map to UBERON ontology
- □ culture_cell: e.g: "Vero E6 cells (CRL-1586)" > Map to CL ontology (NULL or none if not culture)
- □ culture_passage_history e.g "Original (not passaged)" (NULL or none if not culture)
- □ collection_date (different formats) e.g "02-Jan-2020", "2020-02-14" , "2020", "2020-03" (homogenize)
- □ study_ref (article PUMED ID)

–HOST/INDIVIDUAL

- □ host_taxon_id e.g "9606" ("Homo sapiens")
- □ host_age: e.g "21" (age in default schema)
- □ host_sex: "female", "male" (sex in default schema)
- □ geo_origin: (different formats) e.g "USA:WI:Madison"/ "USA: CA, San Diego County"/ "30.52 N 114.31 E" > harmonise, map to GAZ ontology (geographic origin in default schema)
- □ disease ("nCoV pneumonia", "COVID-19" , "severe acute respiratory syndrome")
- □ comorbidities (diseases in default schema)
- □ disease_course: e.g "mild" (harmonized maybe from disease)
- □ disease_outcome: e.g "resolution/discharge" (harmonized from "Survived")
- □ info
  - □ study_ref (article PUMED ID)

– VIRUS

□ taxon_id: e.g ”433733”

□ taxon_name: e.g “Severe acute respiratory syndrome coronavirus 2”

□ strain_id:

□ strain_name: e.g "2019-nCoV/USA-WI1/2020"