gatk_50_files_metadata_fields

50 xml files from vcf from:

-13 experiments (13 submissions)

-10 studies

-46 biosamples

–FILES (50)

- RUN

- [ ] PRIMARY_ID e.g "SRR10903401" (SRR)
- [ ] total_spots e.g "476632" "676694" (50)
- [ ] total_bases e.g "143565674" "203832203" (50)
- [ ] Statistics
  - [x] ~~read_counts (39) e.g "476632" "676694" "28282964" "5246584" ()~~
  - [ ] 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"
- [x] ~~EXPERIMENT_REF~~
  - [x] ~~accession (EXP PRIMARY ID) "SRX7571571"~~
- [x] ~~Member~~
  - [x] ~~PRIMARY ID e.g "SRS6007144"~~
  - [x] ~~EXTERNAL ID "SAMN13872787"~~
- [x] ~~attributes~~
  - [x] ~~accession (same as PRIMARY ID) e.g "SRR10903401" (SRR)~~
  - [x] ~~alias e.g "wuhan2_1.fq.gz", "nCov1.bam" (run filename)~~
  - [x] ~~is_public/ cluster_name ("true", "public")~~
- [x] ~~total_spots e.g "476632" "676694" (same as read counts)~~
- [x] ~~nreads ("2", "variable")~~
- [x] ~~nspots (same as read counts, same as total_spots)~~

–EXPERIMENT (13)

- SAMPLE

- [x] ~~SAMPLE_ATTRIBUTES~~
  - [x] ~~sample type ("genomic RNA")~~

- [x] ~~Extraction Method ("QIAamp Viral RNA kit using carrier tRNA")~~

- EXPERIMENT

- [ ] IDENTIFIERS
  - [ ] PRIMARY_ID eg. "SRA1041081" (13 experiments)
- [ ] TITLE (string e.g "Total RNA sequencing of BALF (human reads removed)")
- [x] ~~STUDY_REF~~
- [ ] DESIGN
  - [x] ~~DESIGN_DESCRIPTION (string) e.g "Low input total RNA library"~~
  - [x] ~~SAMPLE_DESCRIPTOR (ID) e.g "SRS6007144" (= SAMPLE PRIMARY ID = sample accession)~~
  - [ ] LIBRARY_DESCRIPTOR
    - [x] ~~LIBRARY_NAME e.g Wuhan2 not homogenous useful values, no commercial library names or anything~~
    - [ ] 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")

–SUBMISSION (13)

- SUBMISSION

- [x] ~~IDENTIFIER~~
  - [x] ~~PRIMARY_ID e.g "SRA1027290" (13) (same info as EXPERIMENT ID/ accession)~~
- [x] ~~ACCESSION~~
  - [ ] ACCESSION e.g "SRA1027290" (13) (same info as EXPERIMENT ID and SUBMISSION ID)

- ORGANIZATION

- [x] ~~Name e.g "Wuhan University"~~
- [ ] Address
  - [ ] Country ("China", "USA", "Colombia", "United States of America", "Australia") * Map to GAZ ontology
  - [ ] Sub ("Hubei" , "WI" , "Risaralda", "wa" , "WA" , "California", "Vic" ,"UT") * Map to GAZ ontology

- [ ] City ("Wuhan", "Shanghai", "Beijing" , "Madison" , "Pereira", "seattle", "Seattle" , "San Diego", "Melbourne" , "Salt Lake City") * Map to GAZ ontology
- [x] ~~Contact~~
  - [x] ~~Adress~~
  - [x] ~~Name~~

--STUDY (10)

- STUDY
- [ ] IDENTIFIERS
  - [ ] PRIMARY_ID (=study accession) e.g "SRP242226"(all are SRP
  - [ ] EXTERNAL_ID (=study alias) e.g "PRJNA601736" (all are PRJ)
- [x] ~~STUDY_TYPE e.g "Other" (all 10 studies are of type "Other")~~
- [x] ~~CENTER_PROJECT_NAME (Not all have this, the ones have this info "Wuhan seafood market pneumonia virus", "Severe acute respiratory syndrome coronavirus 2")~~

-(--STUDY --INFO) Or maybe to HOST/INDIVIDUAL Endpoint?

- [ ] STUDY_TITLE e.g "Metatranscriptomics of two pneumonia cases"
- [ ] STUDY_ABSTRACT e.g "Discovery and characterization of a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak" (at the end)

- SAMPLE

- [ ] SAMPLE_ATTRIBUTES
  - [ ] link_addit_analys (article)("https://wwwnc.cdc.gov/eid/article/26/6/20-0516_article")

--BIOSAMPLE (45)

- SAMPLE

- [ ] IDENTIFIERS
  - [ ] PRIMARY_ID (=sample accession) e.g "SRS6007144" (all SRS)
  - [ ] EXTERNAL_ID e. g "SAMN13872787" (all SAMN)
  - [x] ~~alias (from papers maybe?) e.g "WHU02" , "UT-00005" , "nCov6", "Human-BALF"~~
- [ ] 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_LINKS:SAMPLE_LINK:XREF_LINK
  - [x] ~~DB (NULL, "bioproject")~~
  - [x] ~~ID e.g "601736" (what does this mean?)~~
  - [x] ~~LABEL (=study alias) e.g "PRJNA601736" (all PRJ)~~
- [ ] SAMPLE_ATTRIBUTES
  - [x] ~~BioSampleModel e.g Human~~
  - [x] ~~isolate e.g P02 (34/50 xml)~~
  - [x] ~~biomaterial_provider e.g "State Key Laboratory of Virology" (only 1 value)~~
  - [x] ~~ref_biomaterial ("BEI Resources catalog NR-52281 (lot 70033135)")~~
  - [ ] … (list all 31 in file SAMPLE_ATTRIBUTES and sample.attributes.pdf)
  - [ ] 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")
  - [x] ~~host_disease_stage ("Acute")~~

– VIRUS

- SAMPLE

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