Table1. Variables description

|  |  |  |
| --- | --- | --- |
|  | Variable | Description |
| 1 | "X" | index |
| 2 | "X.label" | Variable names |
| 3 | "FDA\_lab\_id" | FDA\_lab\_id |
| 4 | "HHS\_region" | - |
| 5 | "IFSAC\_category" | he Interagency Food Safety Analytics Collaboration (IFSAC) develops regulatory-focused schemes to help categorize isolate sourcing information. |
| 6 | "LibraryLayout" | Sequence Read Archive (SRA) library layout (PAIRED/SINGLE) |
| 7 | "PFGE\_PrimaryEnzyme\_pattern" | Pulsed-field gel electrophoresis (PFGE) primary enzyme pattern. PFGE is a DNA fingerprinting technique used to differentiate bacterial strains based on the pattern of DNA fragments that are created by digesting their complete genome with a restriction enzyme. |
| 8 | "PFGE\_SecondaryEnzyme\_pattern" | Pulsed-field gel electrophoresis (PFGE) secondary enzyme pattern. |
| 9 | "Platform" | Sequence Read Archive (SRA) sequencing platform. |
| 10 | "Run" | Sequence Read Archive (SRA) accession of the sequence that was used for the genome assembly. |
| 11 | "asm\_acc" | The accession number for the genome sequence from the Assembly database. Note that a transient state may occur where two isolates point to the same assembly when the submitter changes the taxonomic identifier for the biosample from one taxgroup to another. |
| 12 | "asm\_level" | Assembly level. The NCBI Assembly database, which includes pathogen isolates as well as eukaryotic organisms, represents genomes assembled to different levels. The Isolates Browser uses circle icons to represents the assembly levels, as follows:  - Complete Genome:   Complete genome assemblies, represented in the "Level" column as a completely filled black circle icon.  - Scaffold:   Assemblies that include scaffolds and contigs, represented in the "Level" column as a 1/2 filled circle icon.  - Contig:   Assemblies that include only contigs, represented in the "Level" column as a 1/4 filled circle icon. |
| 13 | "asm\_stats\_contig\_n50" | Assembly contig N50. This is a statistical measure that defines assembly quality. At least half of the bases in the assembly belong to contigs that have a length of N50 or longer. |
| 14 | "asm\_stats\_length\_bp" | Total length of the genome sequence assembly in number of base pairs (nucleotides). |
| 15 | "asm\_stats\_n\_contig" | Contigs. Number of contigs in the isolate's genome assembly. |
| 16 | "assembly\_method" | Assembly method |
| 17 | "attribute\_package" | Isolation type of an isolate: clinical OR environmental/other OR NULL. |
| 18 | "bioproject\_acc" | BioProject accession |
| 19 | "bioproject\_center" |  |
| 20 | "biosample\_acc\_x" |  |
| 21 | "isolate\_identifiers" | A list of alternative identifiers that the isolate may be known by. Ids are assembled from various fields in the BioSample record, including:  - auxiliary identifiers supplied with the Biosample  - sample\_name  - strain  - isolate (from BioSample)  - NARMS\_isolate\_number  - culture\_collection  - isolate\_name\_alias (split by delimiter) |
| 22 | "collected\_by" | Name of persons or institute who collected the sample |
| 23 | "collection\_date" | Date sample was collected, in the format the submitter supplied. |
| 24 | "epi\_type" | Isolation type of an isolate: clinical OR environmental/other OR NULL. |
| 25 | "fullasm\_id" |  |
| 26 | "geo\_loc\_name" | Geographical origin of the sample. The Location data field typically may have two parts: Country:Region. Country is a controlled vocabulary. Region is not controlled and can be anything (i.e., free text). |
| 27 | "host" | Host species |
| 28 | "host\_disease" | Host disease |
| 29 | "isolation\_source" | Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived |
| 30 | "lat\_lon" | The geographical coordinates (latitude and longitude) of the location where the sample was collected |
| 31 | "ontological\_term" | did not find |
| 32 | "outbreak" | The submitter designated name for an occurrence of more cases of disease than expected in a given area or among a specific group of people over a particular period of time. |
| 33 | "sample\_name" | sample\_name |
| 34 | "scientific\_name" | Scientific name (in NCBI Taxonomy) of the isolate from the submitter. |
| 35 | "serovar" | serovar |
| 36 | "source\_type" | The isolate source type. Possible values include Food, Animal, Environmental, Human, Animal feed. |
| 37 | "species\_taxid" | The NCBI Taxonomy identifier (TaxID) at the species level for this isolate. |
| 38 | "sra\_center" | The name of the center that submitted the data to the Sequence Read Archive (SRA). |
| 39 | "sra\_release\_date" | Sequence Read Archive (SRA) release date. |
| 40 | "strain" | Microbial strain name |
| 41 | "target\_acc" | Pathogen Detection accession of the isolate. The accession begins with the prefix "PDT," which stands for Pathogen Detection Target. This database is the primary resource issuing PDT accessions. |
| 42 | "target\_creation\_date" | That is the date on which the isolate was first seen by the Pathogen Detection system. The isolates are shown in reverse chronological order, with the newest ones appearing at the top. |
| 43 | "taxid" | The NCBI Taxonomy identifier (TaxID) for this isolate, which can have a classification that is narrower than species. |
| 44 | "wgs\_acc\_prefix" | The stable accession prefix that is assigned to a Whole Genome Shotgun (WGS) project. |
| 45 | "wgs\_master\_acc" | The Whole Genome Shotgun (WGS) accession for the master record. The WGS master record contains no sequence data, and instead lists all of the accession numbers for the individual sequence records that compose the genome assembly for the isolate. |
| 46 | "minsame" | Minimum SNP distance from this isolate to one of the same isolation type. For example, the minimum SNP distance from one clinical isolate to another clinical isolate, or from one environmental isolate to another environmental isolate. |
| 47 | "mindiff" | the minimum SNP distance from this isolate to one of a different isolation type. |
| 48 | "computed\_types" | "In-silico" typing results. Currently the results of executing SeqSero2 on Salmonella isolates (only) are presented in these subfields:  - serotype - The serovar computed from the reads (if available) or the assembly of the isolate.  - antigen\_formula - The antigenic formula computed from the reads (if available) or the assembly of the isolate.  Values for "Serotype" and "Antigen formula" in the Computed types field may not agree with the user submitted fields Serovar, TaxID, or Scientific name because those fields are reported by the submitter. The "computed\_types" field, on the other hand, is a computational prediction based on the sequence calculated as part of the Pathogen Detection Pipeline. |
| 49 | "number\_drugs\_resistant" | describe the column “**AST\_phenotypes**” |
| 50 | "number\_drugs\_intermediate" | describe the column “**AST\_phenotypes**” |
| 51 | "number\_drugs\_susceptible" | describe the column “**AST\_phenotypes**” |
| 52 | "number\_drugs\_tested" | describe the column “**AST\_phenotypes**” |
| 53 | "number\_amr\_genes" | describe the column “**AMR\_genotypes**” |
| 54 | "number\_core\_amr\_genes" | describe the column “**AMR\_genotypes**” |
| 55 | "AST\_phenotypes" | Antibiotic resistance phenotype, based on Antimicrobial Susceptibility Test (AST) results. (also referred to as antibiograms.)  For those BioSample records for which sequencing data is submitted, and which are also incorporated into the Pathogen resources, the Isolate Browser displays the antibiotic compounds from each antibiogram, binned into the **SIR (sensitive, intermediate, resistance)** calls as made by the submitter into a separate column: AST\_phenotypes.  where the values can be:  I (intermediate)  NS (nonsusceptible)  N, ND (not defined)  R (resistant)  S (susceptible, sensitive)  SSD (susceptible-dose dependent)  [Note] the format for this data field in the isolates browser is presented as a list of antibiotic compounds broken down by resistance call made by the data submitter. These are typically, done by using CLSI or EUCAST standards and those standards change over time OR the call is made by an automated instrument which may infer the cutoff. This may mean that data submitted using an earlier standard may have different resistance calls for the same antibiotic compound than data submitter using a later standard, and even for the same organism and same isolate, different tests may yield different results. Users can search this field by the antibiotic compound AND by the resistance call – the format is different than most other fields in this document. |
| 56 | "AMR\_genotypes" | Antimicrobial resistance (AMR) genes found in the isolate during analysis with AMRFinderPlus. This is a de-duplicated list, so multiple genes that share the same symbol will only be represented once. <NONE> indicates a lack of AMR genes identified by AMRFinderPlus, while an empty field means AMRFinderPlus results are not yet available. See the AMRFinderPlus analysis type, PD Ref Gene Catalog version, and AMRFinderPlus version fields for more information about the AMRFinderPlus analysis of this isolate. |
| 57 | "AMR\_genotypes\_core" | Core antimicrobial resistance (AMR) genes found in the isolate during analysis with AMRFinderPlus. The only differences between AMR genotypes core (AMR\_genotypes\_core) and AMR genotypes (AMR\_genotypes) column is that "plus" genes are not shown. This is a de-duplicated list, so multiple genes that share the same symbol will only be represented once. <NONE> indicates a lack of AMR genes identified by AMRFinderPlus, while an empty field means AMRFinderPlus results are not yet available. |
| 58 | "number\_stress\_genes" | related to the below |
| 59 | "stress\_genotypes" | Stress resistance genes found in the isolate during analysis with AMRFinderPlus. These can include metal, biocide, and heat resistance genes. This is a de-duplicated list, so multiple genes that share the same symbol will only be represented once. <NONE> indicates a lack of AMR genes identified by AMRFinderPlus, while an empty field means AMRFinderPlus results are not yet available. |
| 60 | "number\_virulence\_genes" | did not find (might be related to the below) |
| 61 | "virulence\_genotypes" | Virulence genes found in the isolate during analysis with AMRFinderPlus. This is a de-duplicated list, so multiple genes that share the same symbol will only be represented once. <NONE> indicates a lack of AMR genes identified by AMRFinderPlus, while an empty field means AMRFinderPlus results are not yet available. |
| 62 | "amrfinder\_version" | The version of the AMRFinderPlus software that was used to analyze a particular isolate.  - New isolates are analyzed using the latest version of AMRFinderPlus software.  - Older isolates may have been analyzed with earlier versions of AMRFinderPlus software.  There might be occasional updates to annotation on all isolates in special circumstances, such as the identification of a new genes (e.g., mobilized colistin resistance (mcr) genes).  This field will be empty if AMRFinderPlus results are not yet available. |
| 63 | "refgene\_db\_version" | PD Ref Gene Catalog version. The version of the Pathogen Detection Reference Gene Catalog that was used to analyze a particular isolate.  New isolates are analyzed using the latest version of the Pathogen Detection Reference Gene Catalog. Older isolates may have been analyzed with earlier versions of the Pathogen Detection Reference Gene Catalog. There might be occasional updates to annotation on all isolates in special circumstances, such as the identification of a new genes (e.g., mobilized colistin resistance (mcr) genes).  Because the "refgene\_db\_version" data field was added in **February 2020**, isolates that were analyzed prior to that time do not have a value in the corresponding "PD Ref Gene Catalog version" data column of the Isolates Browser display. |
| 64 | "amrfinder\_analysis\_type" | did not find this but found another “amrfinderplus\_analysis\_type”:  Indicates the data types that were used to analyze the isolate's genome sequences using AMRFinderPlus. Genome sequences are generally analyzed in two passes:  **NUCLEOTIDE**: this in an initial analysis that is done, using translated BLAST, immediately after a pathogen isolate genome is assembled. It identifies the proteins in the genome sequence.  **COMBINED**: this is a second, more sensitive analysis that runs AMRFinderPlus on both an isolate's nucleotide and protein sequences.  Protein BLAST, nucleotide BLAST, and HMMER are used to analyze the proteins. **The combined analysis can produce more sensitive results than the initial nucleotide analysis**. |
| 65 | "amrfinder\_applied" | - |
| 66 | "PDS\_acc" | snp cluster ID |
| 67 | "biosample\_acc\_y" | - |
| 68 | "gencoll\_acc" | - |