Supplementary Table 1. Overview of key variables in dataset

Variable Name (*data_field_name*)	Description
Organism Group (*taxgroup_name)*	Organism group related to taxonomy for purposes of calculating SNP clusters. The organisms groups are manually constructed and may include sister species and outgroups.
Strain (*strain*)	Microbial strain name, if provided by the submitter. This field contains values exactly as they were entered by the data submitters.
Isolate identifiers (*isolate_identifiers*)	A list of alternative identifiers the isolate may be known by.
Serovar (*serovar*)	Combined field of sub-species, serotype, or serovar, if provided by the submitter.
Isolate (*target_acc*)	Pathogen Detection accession of the isolate. The accession begins with the prefix "PDT" (stands for Pathogen Detection Target).
Host (*host*)	Host species, if provided by the submitter.
Create date (*creation_date*)	The date on which this isolate was first seen by the Pathogen Detection System. Format: YYYY-MM_DD.
Location (*geo_loc_name*)	Geographical origin of the sample, if provided by the submitter. Matches the country qualifier of GenBank records. Location data field typically contains two parts: Country:Region - Country is controlled - Region is not controlled and can be anything (state abbreviation, province name, city name, zip code, etc.)
Lat/Lon (*lat_lon*)	The geographical coordinates (latitude and longitude) of the location where the sample was collected, if provided by the submitter.
Isolation source (*isolation_source*)	Describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived, if provided by the submitter.

Isolation type (*epi_type*)	Isolation type of an isolate: - Clinical OR environmental/other OR NULL The isolation type (epi_type) is used to calculate the SNP distance values Min-same and Min-diff.

SNP cluster (*erd_group*)	Pathogen SNP cluster accession.
	  - SNP cluster: group of isolates whose genome assemblies are closely related, depending on the clustering methodology used.
	  - ERD stands for Epidemiologically Related Distance.
	Each SNP cluster can be viewed as a phylogenetic distance tree in the SNP Tree Viewer.

Min-same (*minsame*)	Minimum SNP distance from this isolate to one of the same isolation type. - Ex: minimum SNP distance from one clinical isolate to another clinical isolate, or from one environmental isolate to another environmental isolate
Min-diff *(mindiff)*	Minimum SNP distance from this isolate to one of a different isolate type. - Ex: minimum SNP distance from a clinical isolate to an environmental isolate. Value will appear in this column only if the isolate has been found by Pathogen Detection Project data processing pipeline, belongs to a SNP cluster and another isolate in that cluster has a different "Isolation type" that is not NULL.
AMR genotypes (*AMR_genotypes*)	Antimicrobial resistance (AMR) genes found in the isolate during analysis (with AMRFinderPlus).
Stress genotype (*stress_genotypes)*	Stress resistance genes found in the isolate during analysis (with AMRFinderPlus).

Virulence genotypes (*virulence_genotypes*)	Virulence genes found in the isolate during analysis (with AMRFinderList). The genes that have been identified in an isolate's genome sequence are grouped into genotype categories - complete, partial, partial end of contig.