Institute for Health Metrics and Evaluation

# Data Release Information Sheet

## *Data Summary*

Dataset name: Global Human Tuberculosis Molecular Epidemiology Systematic Review Dataset

Date of release: October 31, 2018

Summary:

Researchers from IHME and collaborating institutions conducted a study to map the global distribution of genotypes of bacterial strains that cause tuberculosis disease and examine whether any epidemiologically relevant clinical characteristics were associated with those genotypes. They performed a systematic review to create a comprehensive dataset of human TB molecular epidemiology studies that used representative sampling techniques. Data were extracted and synthesized from 206 studies that reported prevalence of bacterial genotypes (representing over 200,000 bacterial isolates collected over 27 years in 85 countries) and from 34 studies that reported clinical characteristics associated with those genotypes. This dataset contains the following: a screening sheet detailing all studies reviewed; raw genotype distribution data extracted in the systematic review; raw genetic clustering data extracted in the systematic review; and sheets containing MTBC genotype conversions for all genotyping methods included in this study.

Relevant publications and visualizations:

- Wiens K, Woyczynski L, Ledesma J, Ross J, Zenteno-Cuevas R, et al. Global variation in bacterial strains that cause tuberculosis disease: a systematic review and meta-analysis. *BMC Medicine*. 31 October 2018.

## Acknowledgements

Contributing organizations:

- Institute for Health Metrics and Evaluation (IHME)

Funders:

- Bill and Melinda Gates Foundation (BMGF)

Suggested Citation:

Institute for Health Metrics and Evaluation (IHME). Global Human Tuberculosis Molecular Epidemiology Systematic Review Dataset. Seattle, United States: Institute for Health Metrics and Evaluation (IHME), 2018.

## *File Inventory*

| File Name | Description | Date Produced |
|---|---|---|
| IHME_TB_SYST_REVIEW_DATA_FILE_2_Y2018M10D31.XLSX | **Additional file 2**: Literature screening sheet. Literature screening sheet including citation information for all literature included in the study. (XLSX 3.1 MB) | October 31, 2018 |
| IHME_TB_SYST_REVIEW_ DATA_FILE_3_Y2018M10D31.CSV | **Additional file 3**: Raw genotype distribution data. Raw genotype distribution data extracted in the systematic review. (CSV 1.9 MB) | October 31, 2018 |
| IHME_TB_SYST_REVIEW_ DATA_FILE_4_Y2018M10D31.CSV | **Additional file 4**: Raw genetic clustering data. Raw genetic clustering data extracted in the systematic review. (CSV 16 KB) | October 31, 2018 |
| IHME_TB_SYST_REVIEW_ DATA_FILE_5_Y2018M10D31.XLSX | **Additional file 5**: Genotype classification system. Sheets containing MTBC genotype conversions for all genotyping methods included in this study. (XLSX 147 KB) | October 31, 2018 |
| IHME_TB_SYST_REVIEW_ DATA_INFO_SHEET_Y2018M10D31.PDF | Data Release Information Sheet | October 31, 2018 |

## *Data Files Information*

### File 3: Raw Genotype Distribution Data

| Variable | Variable Label | Variable Definition |
|---|---|---|
| nid | NID | Unique identifier for the catalog record for the study publication and/or data in the Global Health Data Exchange (GHDx) |
| pmid_or_unique_identifier | PubMed ID or Unique Identifier | Unique study identifier based on either PubMed ID, Scopus EID, or other unique study name |
| country_name | Country Name | Name of the country in which the study was conducted |

| | | |
|---|---|---|
| country_code | Country Code | Country Code (ISO 3) for the country in which the study was conducted |
| year | Year | Time period of estimate. Represents midpoint between start and end year. Possible values: Years in the range 1994-2016 |
| page_num | Page Number | Page number from the study that contains the extracted data |
| table_num | Table Number | Table or Figure from the study that contains the extracted data |
| source_type | Source Type | Underlying mode of data collection. Possible values: Surveillance, Case notifications, Survey |
| location_name | Location Name | Location name corresponding to each location_id and ihme_loc_id |
| location_id | Location ID | Unique numeric ID used by IHME for each location |
| ihme_loc_id | IHME Location ID | Unique ID used by IHME for each location that combines country_code and location_id |
| smaller_site_unit | Smaller Site Unit | Whether or not study area is representative of the ihme_loc_id. 0 = Representative sample; 1 = representative of a smaller collection site (recorded in “site_memo”) |
| site_memo | Site Memo | Description of collection site when smaller_site_unit = 1 |
| admin_level | Administrative Level | Administrative level at which the data were collected at when smaller_site_unit = 1. Possible values: 1, 2, 3, 4, "precise place" |
| geography_name | Geography Name | Name of administrative unit or precise place at which data were collected when smaller_site_unit = 1 |
| year_start | Start Year | Start year of study. Possible values: Years in the range 1989 - 2016 |
| year_end | End Year | End year of study. Possible values: Years in the range 1995 - 2016 |
| cv_spoligotype | Spoligotyping Method | Genotyping done using spoligotyping. 1 = yes; 0 = no |
| cv_miru_vntr | MIRU-VNTR Method | Genotyping done using multilocus variable number of tandom repeats method (MIRU and/or VNTR). 1 = yes; 0 = no |
| cv_pcr | PCR Method | Genotyping done using PCR for large sequence polymorphisms (LSP). 1 = yes; 0 = no |
| cv_wgs | WGS Method | Genotyping done using whole genome sequencing (WGS). 1 = yes; 0 = no |
| spoligotype | Spoligotype Pattern | Spoligotype octal code in number format (15 digits between 0 and 7) |
| sit | Spoligotype International Type | Spoligotype international type (numeric) |

| | | |
|---|---|---|
| miru_vntr_lineage | MIRU-VNTR Lineage | Lineage name based on MVLA or MIRU-VNTR typing |
| sit_clade | Spoligotype Clade | Clade name based on spoligotyping |
| LSP_lineage_name | LSP Lineage Name | Lineage name based on PCR typing of large sequence polymorphisms |
| LSP_lineage_number | LSP Lineage Number | Lineage number based on PCR typing of large sequence polymorphisms |
| mtbc_lineage | MTBC Lineage | Mycobacterium tuberculosis complex lineage (to be used in meta-analysis). Possible values: Lineage_1, Lineage_2, Lineage_3, Lineage_4, Lineage_7, Lineage_Maf, Lineage_Animal, Lineage_Other |
| cases | Cases | Total number of each genotype identified in each study |
| collection_period | Collection Period | Measurement of sampling duration corresponding to collection_period_value. Possible values: months, years |
| collection_period_value | Collection Period Value | Value of sampling duration corresponding to collection_period (numeric) |
| sampling_method | Sampling Method | Sampling approach. Possible values: All cases, All new patients, Cluster, Multistage, Probability, Simple random |
| africa | Africa Region | Whether the study was conducted in Africa. 1 = yes; 0 = no |
| americas | Americas Region | Whether the study was conducted in North, Central, or South America. 1 = yes; 0 = no |
| europe | Europe Region | Whether the study was conducted in Europe. 1 = yes; 0 = no |
| east_asia | East Asia Region | Whether the study was conducted in East Asia. 1 = yes; 0 = no |
| west_asia | West Asia Region | Whether the study was conducted in West Asia. 1 = yes; 0 = no |
| oceania | Oceania Region | Whether the study was conducted in Oceania. 1 = yes; 0 = no |
| nationally_representative | Nationally Representative | Whether the study was nationally-representative. 1 = year; 0 = no |
| sampling_type | Sampling Type | Category of sampling method. Possible values: All cases, Survey |
| data_collection_method | Data Collection Method | Category of mode of data collection. Possible vlaues: Cohort, Cross-sectional |
| sample_size | Sample Size | Total number of cases with genotyping information by study, country, and year |
| estimated_national_tb_cases | Estimated National TB Cases | Total prevalent TB cases in the country and year in which the study was conducted using estimates from the Global Burden of Disease Study 2016 |

| | | |
|---|---|---|
| percent_of_tb_cases | Percent of TB Cases Represented | Percent of all estimated prevalent TB cases represented in each study |

## File 4: Raw Genetic Clustering Data

| Variable | Variable Label | Variable Definition |
|---|---|---|
| pmid_or_unique_identifier | PubMed ID or Unique Identifier | Unique study identifier based on either PubMed ID, Scopus EID, or other unique study name |
| country_code | Country Code | Country Code (ISO 3) for the country in which the study was conducted |
| region | Region | Region in which the study was conducted. Possible values: africa, america_europe, east_asia, oceania, west_asia |
| year_start | Start Year | Start year of study. Possible values: Years in the range 1998-2014 |
| year_end | End Year | End year of study. Possible values: Years in the range 2002-2015 |
| cv_spoligotype | Spoligotyping Method | Genotyping done using spoligotyping. 1 = yes; 0 = no |
| cv_miru_vntr | MIRU-VNTR Method | Genotyping done using multilocus variable number of tandom repeats method (MIRU and/or VNTR). 1 = yes; 0 = no |
| cv_pcr | PCR Method | Genotyping done using PCR for large sequence polymorphisms (LSP). 1 = yes; 0 = no |
| cv_wgs | WGS Method | Genotyping done using whole genome sequencing (WGS). 1 = yes; 0 = no |
| mtbc_lineage | MTBC Lineage | Mycobacterium tuberculosis complex lineage (to be used in meta-analysis). Possible values: Lineage_1, Lineage_2, Lineage_3, Lineage_4, Lineage_7, Lineage_Maf, Lineage_Animal, Lineage_Other |
| clustered_cases | Total Clustered Cases | Total number of clustered cases of each genotype identified in each study |
| total_cases | Total Cases | Total number of clustered and non-clustered cases of each genotype identified in each study |
| cv_prop_hiv | Proportion of Cases HIV-Infected | Proportion of the TB cases that were HIV-infected across all genotypes in the study sample |
| cv_prop_resist | Proportion of Cases Drug-Resistant | Proportion of the TB cases that were drug resistant across all genotypes in the study sample |
| cv_mean_age | Mean Age of Cases | Mean age of all TB cases in the study sample |
| cv_mean_age_SD | SD of Mean Age of Cases | Standard deviation of the mean age of all TB cases in the study sample |
| cv_prop_male | Proportion of Cases Male | Proportion of the TB cases that were male across all genotypes in the study sample |
| cv_prev_tb | Proportion of Cases with Previous TB | Proportion of the TB cases that had previously had TB across all genotypes in the study sample |

| | | |
|---|---|---|
| cv_eptb | Proportion of Cases with Extrapulmonary TB | Proportion of the TB cases that had any form of extrapulmonary TB across all genotypes in the study sample |
| cv_immigrant | Proportion of Cases Immigrants | Proportion of the TB cases that immigrants across all genotypes in the study sample |

## *Additional Information*

### Terms and Conditions

http://www.healthdata.org/about/terms-and-conditions

### Contact information

To request further information about this dataset, please contact IHME:

Institute for Health Metrics and Evaluation
2301 Fifth Ave., Suite 600
Seattle, WA 98121
USA
Telephone: +1-206-897-2800
Fax: +1-206-897-2899
Email: data@healthdata.org
www.healthdata.org

These files may be updated periodically, so we appreciate hearing feedback or additional information about how these data are being used.