**Model the current and future risk of locally acquiring Chagas with Habitat Suitability Distribution of Chagas Disease vectors triatomine in North America with presence-only data**

**Introduction (edited)**

The Chagas Disease, traditionally considered as a disease prevalent in Latin America, is estimated to infect approximately 6-7 million people worldwide.1 The Chagas Disease has two phases, acute and chronic, with symptoms ranging from asymptomatic in the acute phase to cardiac or gastrointestinal disease during the chronic phase.2 The infection persists for a lifetime without effective treatment.2 Regarding treatment, the decades-old effective drugs benznidazole and nifurtimox produce side effects which causes one in five patients to abandon treatment,3 while vaccines are still in development4.5

The Chagas disease is caused by the pathogen *Trypanosoma cruzi*, which is majorly transmitted by the infectious waste of triatomines contacting with human’s membrane or wound, or food contaminated by the infectious waste been consumed by human.1 Chagas can also be transmitted through contaminated blood transfusion and organ transplantations, ingestion of contaminated food, and postnatal infections. Other than humans, animals can also be infected. For example, non-human primates would be infected by eating the triatomines.6 Under climate change, the habitat of the triatomines might extend into traditionally unoccupied regions in North America, increasing the risk of unprepared populations locally acquiring Chagas.3

[~~In the United States, with no large-scale screening of Chagas Disease,~~~~7~~ ~~the infected population is only an estimation number, and the estimated number of locally acquiring cases is outdated, making it hard to estimate the infectious risk. The estimated 288,000 - 347,000 infected patients are calculated from combining immigrants' population and the infection rate in their countries of origin~~~~8~~~~.~~~~7~~ ~~The 10,000 autochthonous Chagas infection cases~~~~87~~ ~~is calculated from an estimation in 2012 that 1 in 354,000 blood donors in United States locally acquired Chagas~~~~9~~~~.~~]

(to write the above paragraph nicely)

In the United States, with no large-scale screening of Chagas Disease,7 currently, the best estimation of 288,000 - 347,000 infected patients and 10,000 autochthonous Chagas infection cases7,8 was calculated from outdated survey data. Alternative measures of the current and future risks of locally acquiring the Chagas Disease are required for us to better evaluate the risk.

This study seeks to provide an alternative measure by studying the Habitat Suitability Distribution (HSD) of the Chagas Disease vector triatomines. In all the triatomine species observed in North America, we selected and modeled 14 species from our database. All the 14 species have been reported to carry *T. cruzi* and transmitted Chagas1011121314.15

We created a most complete database for triatomines in North America and conduct HSD modeling for both historical (1991-2020) and CMIP6 future projected (2071-2100) scenarios with two machine learning models: MaxEnt and downsample Random Forest, which have good performance for modeling species distribution with sparse presence-only data. The triatomine database and HDS maps are presented to support healthcare providers and public health researchers in their decision making.

**Material and Method**

Study Design (edited)

We first constructed a North America triatomine database, including observation data till July 2023. For each species, we use observation data points from this dataset combining with pseudo-absence data points to sample historical environmental predictors and train MaxEnt and downsample Random Forest. The environmental predictors include bioclimatic and land cover data for historical period (1991-2020) and 2071-2100 under 4 CMIP6 scenarios: SSP1 RCP2.6, SSP2 RCP4.5, SSP3 RCP7.0 and SSP 5 RCP8.5 (CO2 emission increases as we go along these 4 scenarios). Then, the two models produce HSD maps for historical and 2071-2100 CMIP6 scenarios with these environmental predictors. We analyzed the model evaluation metrics and the differences of the HSD maps. Figure 1 is a flowchart presenting the modeling process.

Triatomine Database (edited)

We compiled the North America Triatominae Database by combing multiple published datasets: ‘New American dataset’ (19,600 records),16 the Disease Vector Database from Moffett et al. (567 records),17 Texas triatomine database 2012-2015 (382 records),18 North Mexico Triatominae dataset (388 records),19 and iNaturalist research-standard citizen science observations (2901 records).20,21  Additionally, we incorporated an unpublished dataset from Alejandro Martínez-Ibarra (220 records) and 875 newly gathered records of triatomine observations published in January 2022 to July 2023. These recent records were collected by our team from databases Semantic Scholar, Science Direct and DataTri.

The raw data of all species was cleaned and filtered before feeding into models. The cleaning process removed points which are duplicated, wrongly/incompletely georeferenced, or not having a clear capture time are removed (scripts available in Appendix 1). The filtering process removed points which are closer than 5 km with the R spThin library. The North America Triatominae Database and filtered datasets are available on Texas Data Repository22.

Environmental Predictors (edited)

The 39 environmental predictors (Table 1) consists of bioclimatic and land cover variables, the bioclimatic variables are from the AdaptWest bioclimatic dataset,23,2425 and the land cover variables are from the Global 7-land-types LULC projection dataset under SSPs- RCPs dataset,26.27 Both datasets are of 1-km resolution and have historical and future projected data under 4 CMIP6 scenarios. For the land cover historical predictor, we use their 2015 land cover data for historical period since the average land cover for 1991-2020 has not been found and most of our triatomine observations were between 2010-2020.

Machine Learning Models (edited)

We use the MaxEnt from R dismo package and downsample RF from randomForest R package to train and predict habitat suitability. These two have been found to be top performance models for presence-only modeling in a previous study.28 Pilot study has been done (see Appendix 3) to show that MaxEnt performs the best with pseudo-absence data selecting randomly outside a 5-km buffer and no environmental predictor preprocessing (eliminate correlated variables and perform PCA) with methods inspired by Konowalik, Kamil, and Agata Nosol.29 For each species, we train MaxEnt and downsample RF with actual observation points and 10,000 pseudo-absence points. Different from traditional RF, down- sample RF randomly selects the pseudo-absence points to match the number of observation points at tree level. Aligning with Valavi, Roozbeh, et al.,28 we implemented MaxEnt with dismo::maxent (needs maxent.jar) R package, and downsample RF using the randomForest R package by modifying code from Valavi, Roozbeh, et al.30

**Results (edited)**

Database (edited)

We created a database for all the triatomine observation data, which is publicly available at Texas Data Repository. The raw dataset contains 24,933 observations. The cleaned Americas dataset for the Americas contains 20,277 observations and cleaned North America contains 12,578 observations. From all the species with observations in the cleaned North America dataset, we further cleaned them to remove observation points that are closer than 5 km with spThin R package (Figure 2), and selected the 14 species with more than 39 records (the number of environmental predictors). We proceed to HSD modeling with these 14 species: *T. sanguisuga*, *T. dimidiata*, *T. gerstaeckeri*, *T. protracta*, *T. longipennis*, *T. rubida*, *M. pallidipennis*, *T. barberi*, *T. mexicana*, *P. lecticularia*, *T. recurva*, *T. mazzottii*, *P. hirsute* and *T. indictiva*.

Model Performance (edited)

The performance of MaxEnt and RF tends to agree on their AUC and TSS scores as the number of training samples increases, while this trend is not observed in MAE and Bias (Figure 3) [plot difference plots]. Spearman correlation analysis shows that the mean and standard deviation of AUC, TSS, and Bias of the two models have similar correlation pattern with number of observation points (positive correlation) and average nearest neighbor distance of observation points (negative correlation), indicating in general the models perform better with a larger number of observation points at a greater distance apart from each other, while the MAE mean has very different correlation values (Appendix 4).(? Do we want to leave all this section to the next paper)

It is agreed by 5 variable importance measurements of MaxEnt and RF that, DD18 (Degree-days above 18 °C) is the most important variables for *T. sanguisuga*; for *T. dimidiata* it is TD (Difference between MCMT (Mean temperature of the coldest month) and MWMT (Mean temperature of the warmest month), as a measure of continentality), for *T. gerstaeckeri* it is PAS (Precipitation as snow), and for *T. protracta* it is CMD (Hargreave’s climatic moisture index). The importance measurement spider charts for all 12 species can be found in Appendix 1.

Habitat Suitability Distribution (edited)

The HSDs for the four species having the most observation points are shown below, and the HSDs for the other 10 species can be found in Appendix 1. In historical HSD maps, predictions from both models cover regions located by observation points, with RF giving a higher suitability values for those regions. For 2071-2100 predictions, HSDs tend to spread out more in higher emission scenarios for both models (Figure 4, for the four species with the most observation points after filtering), with predictions from RF generally more aggressive, consistently displaying a higher suitability value and broader spatial coverage. Among all, *T. sanguisuga, T. protracta, P. lecticularia, P. hirsute* and *T. indictiva* have the most significant suitable habitat expansion to the north in 2071-2100.

**Discussion (edited)**

We created the most complete open-sourced North America triatomine database combining well-established databases, newly published datasets, and citizen science dataset. The cleaned data could be used for conducting further triatomine ecological niches studies on regional scale, and the filtered version could be directly used for model development.

MaxEnt and downsample RF both have good agreements on evaluation metric results for most of the nine species with training samples no fewer than 143; therefore, we would regard habitat suitability maps for these nine species as reliable: *T. sanguisuga*, *T. dimidiata*, *T. gerstaeckeri*, *T. protracta*, *T. longipennis*, *T. rubida*, *M. pallidipennis*, *T. barberi*, *T. mexicana.*

We applied a new way to evaluate differences between prediction maps on regional scale, for which we calculated the percentage difference of pixel-wise summation of two prediction raster maps. This percentage differences converges to a low value at the right end where models have more training samples(Figure 5).

RF models have a consistent negative Bias score (Figure 3), which indicates they could be used as an upper bound. At the same time, the observation points are well covered by high suitability regions predicted by RF, while the ‘outlier’ points are sometimes missed by those from MaxEnt models (Figure 4 and Appendix 1).

Our study has three key achievements. First, we generated the most comprehensive North America triatomine database. Second, we computed the most reliable HSDs for 14 most abundant species. Third, we conducted a quantitative assessment of model performance differences using real presence-only observations.

The expanding suitable habitat for triatomines signals a growing risk of locally acquired Chagas diseases. Early diagnosis during asymptomatic period is essential for effective treatments and potentially prevent congenital transmission and entering chronic phase. However, no large-scale screening exists in the United States for detecting infections during this critical period.7 Currently, Chagas Disease has only been reported in Texas, Arizona, Arkansas, Tennessee, Mississippi, Louisiana, Utah, plus Los Angeles, California; the District of Columbia; and Boston, Massachusetts, metropolitan areas3,7,31 Given an increasing locally acquiring risk, integrating habitat suitability modeling with screening and public education programs will be essential for preparing at-risk populations for the challenges ahead.

**Tables**

Table 1 (edited)

|  |  |
| --- | --- |
| Species | Number of Observations in Cleaned North America Dataset |
| *Triatoma gerstaeckeri* | 4464 |
| *Triatoma sanguisuga* | 2217 |
| *Triatoma dimidiata* | 1679 |
| *Triatoma protracta* | 864 |
| *Triatoma rubida* | 492 |
| *Triatoma longipennis* | 482 |
| *Meccus pallidipennis* | 465 |
| *Triatoma barberi* | 380 |
| *Triatoma mexicana* | 358 |
| *Paratriatoma lecticularia* | 275 |
| *Triatoma indictiva* | 204 |
| *Triatoma mazzottii* | 158 |
| *Triatoma recurva* | 112 |
| *Dipetalogaster maxima* | 96 |
| *Triatoma phyllosoma* | 71 |
| *Paratriatoma hirsuta* | 64 |

**Figures**

Figure 1

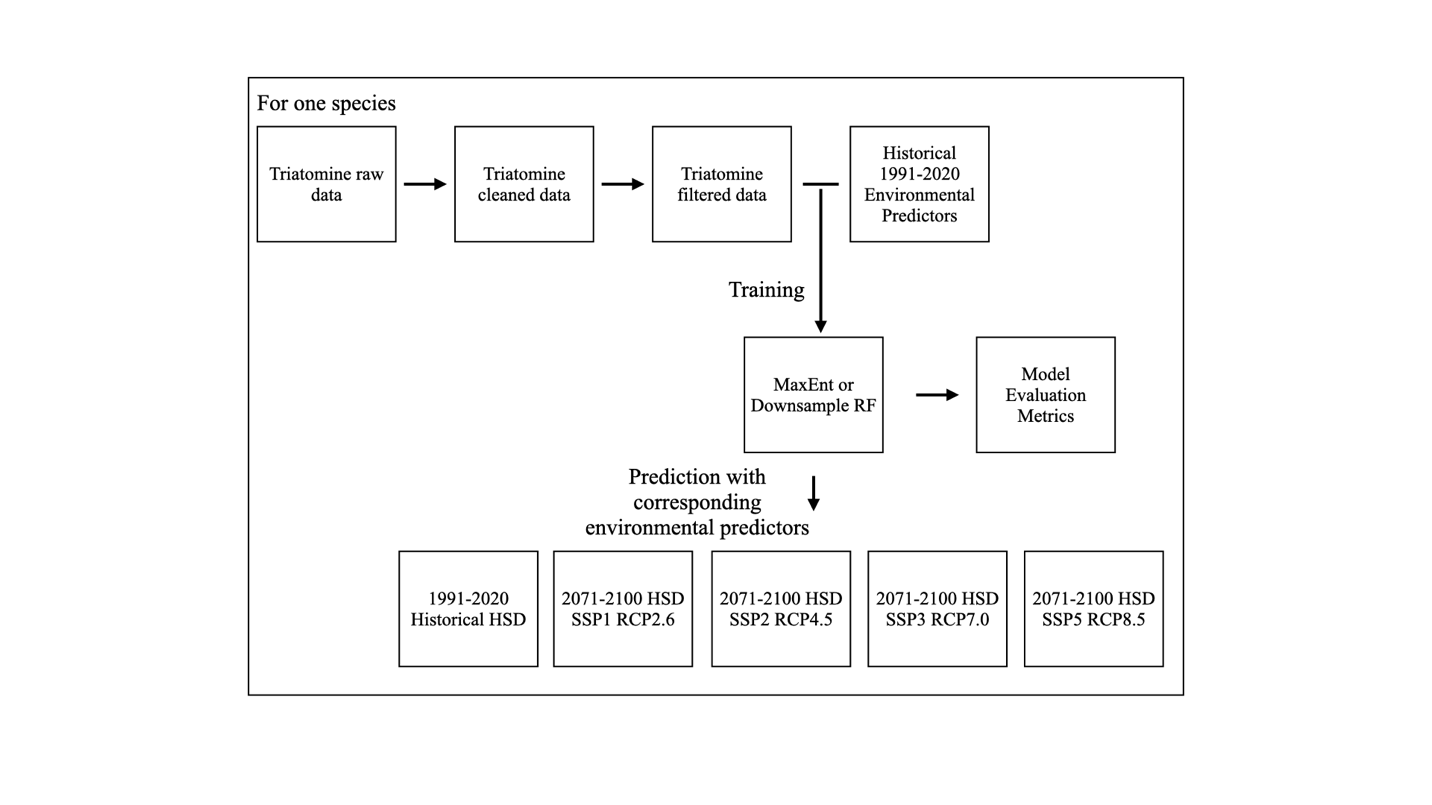


Figure 2 (edited): The geospatial distribution of filtered observation points for the 16 species: A) *T. sanguisuga*, B) *T. dimidiata*, C) *T. gerstaeckeri*, D) *T. protracta*, E) *T. longipennis*, F) *T. rubida*, G) *M. pallidipennis*, H) *T. barberi*, I) *T. mexicana*, J) *P. lecticularia*, K) *T. recurva*, L) *T. mazzottii*, M) *P. hirsuta*, N) *T. indictiva*, O), *D. maxima* P) *T. phyllosoma*. The number of filtered observation points (red) and the number of cleaned observation points (black) are noted at the bottom right of each subplot. The subplots are arranged in the ascending order of the number of filtered observation points.

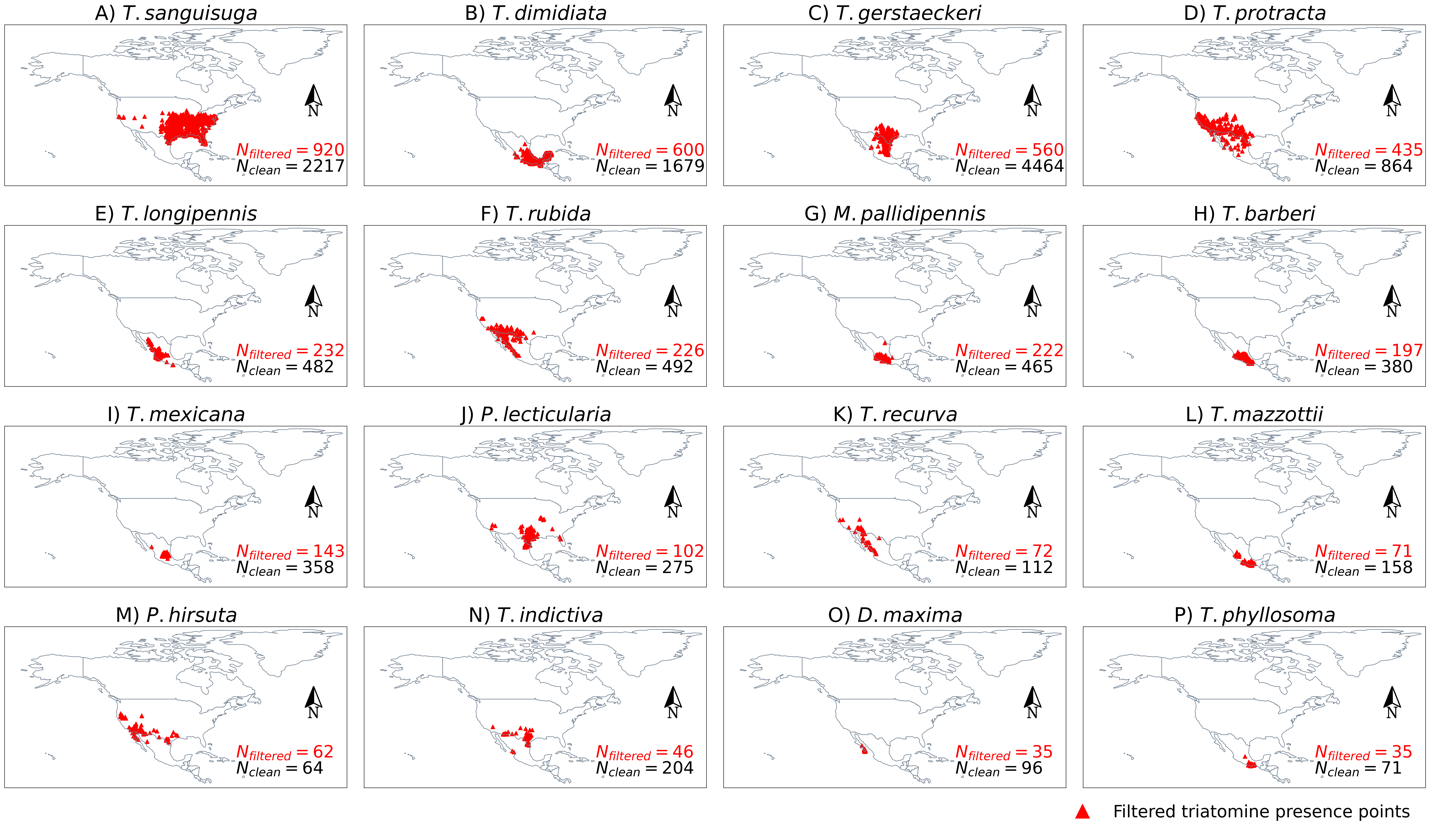


Figure 3 (edited)

A graph of different sizes and shapes

AI-generated content may be incorrect.

Figure 4a (edited)

A screenshot of a map

AI-generated content may be incorrect.

Figure 4b (edited)

A screenshot of a computer screen

AI-generated content may be incorrect.

Figure 4c (edited)

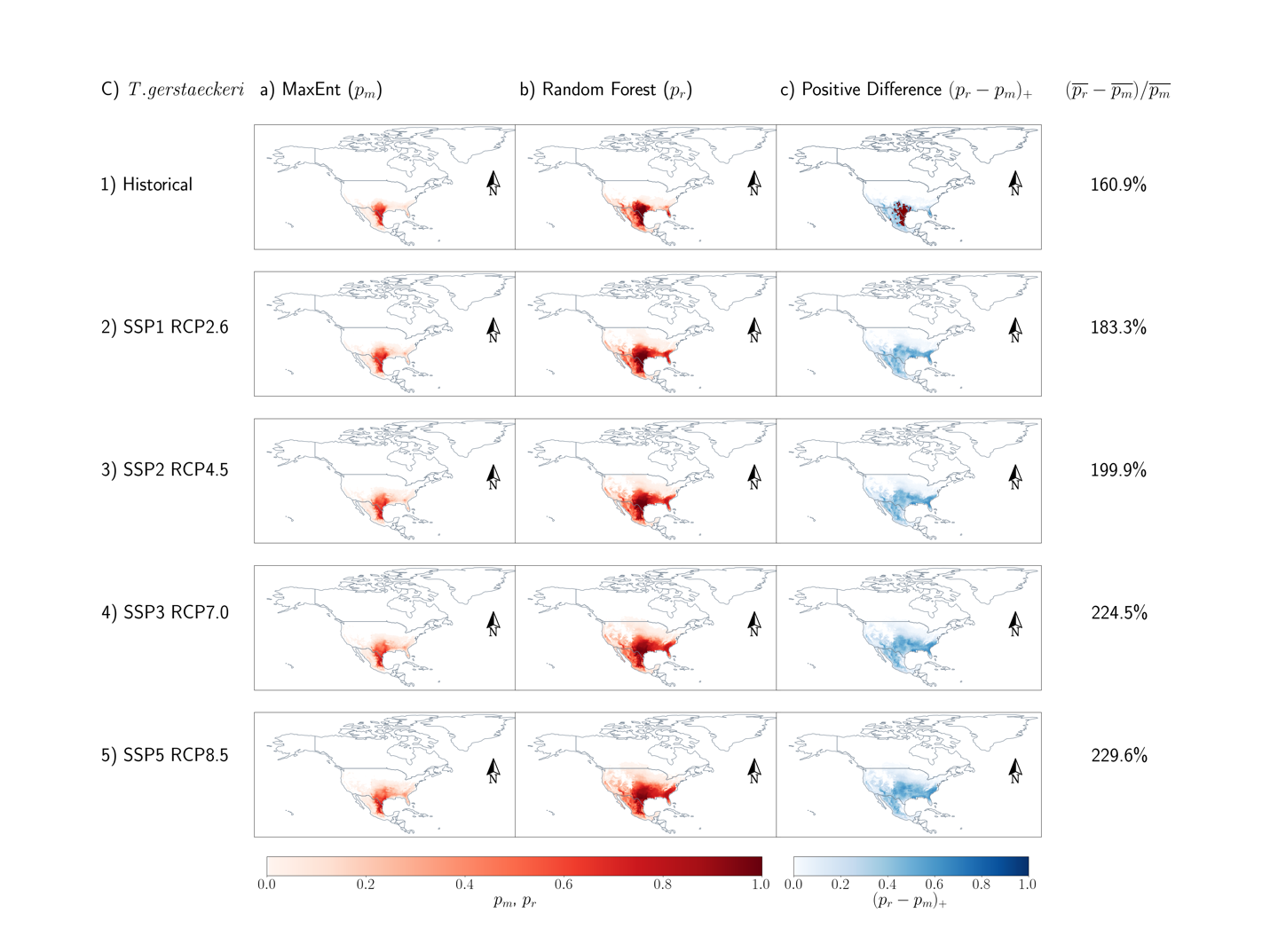


Figure 4d (edited)

A screenshot of a map

AI-generated content may be incorrect.

Figure 5 (edited)

A graph of number of filtereded points

AI-generated content may be incorrect.

**References**

1. WHO W. Chagas disease (American trypanosomiasis). Epidemiology. 2020:Date ac- cessed: July 16, 2024.
2. Bern C, Messenger LA, Whitman JD, Maguire JH. Chagas disease in the United States: a public health approach. Clinical microbiology reviews. 2019;33(1):10–1128.
3. Higuita NIA, Beatty NL, Forsyth C, et al. Chagas disease in the United States: a call for increased investment and collaborative research. The Lancet Regional Health–Americas. 2024;34.
4. Dumonteil E, Herrera C, Marx PA. Safety and preservation of cardiac function follow- ing therapeutic vaccination against Trypanosoma cruzi in rhesus macaques. Journal of Microbiology, Immunology and Infection. 2023;56(2):400–407.
5. Jones KM, Poveda C, Versteeg L, Bottazzi ME, Hotez PJ. Preclinical advances and the immunophysiology of a new therapeutic Chagas disease vaccine. Expert Review of Vaccines. 2022;21(9):1185–1203.
6. Hodo CL, Wilkerson GK, Birkner EC, Gray SB, Hamer SA. Trypanosoma cruzi transmis- sion among captive nonhuman primates, wildlife, and vectors. EcoHealth. 2018;15:426– 436.
7. Irish A, Whitman JD, Clark EH, Marcus R, Bern C. Updated estimates and mapping for prevalence of Chagas disease among adults, United States. Emerging infectious diseases. 2022;28(7):1313.
8. Manne-Goehler J, Umeh CA, Montgomery SP, Wirtz VJ. Estimating the burden of Cha- gas disease in the United States. PLoS neglected tropical diseases. 2016;10(11):e0005033.
9. Cantey PT, Stramer SL, Townsend RL, et al. The United States Trypanosoma cruzi In- fection Study: evidence for vector-borne transmission of the parasite that causes Chagas disease among United States blood donors. Transfusion. 2012;52(9):1922–1930.
10. Bern C, Kjos S, Yabsley MJ, Montgomery SP. Trypanosoma cruzi and Chagas’ disease in the United States. Clinical microbiology reviews. 2011;24(4):655–681.
11. Waleckx E, Suarez J, Richards B, Dorn PL. Triatoma sanguisuga blood meals and poten- tial for Chagas disease, Louisiana, USA. Emerging infectious diseases. 2014;20(12):2141.
12. Reisenman CE, Lawrence G, Guerenstein PG, Gregory T, Dotson E, Hildebrand JG. Infection of kissing bugs with Trypanosoma cruzi, Tucson, Arizona, USA. Emerging infectious diseases. 2010;16(3):400.
13. Salazar Schettino PM, Rosales Pin ̃a JS, Rojas Wastavino G, Cabrera Bravo M, Vences Blanco M, Lo ́pez Ca ́rdenas J. Triatoma mexicana (Hemiptera: Reduviidae) in Guanajuato, Mexico: house infestation and seasonal variation. Memorias do Instituto Oswaldo Cruz. 2007;102:803–807.
14. Ramsey J, Ordon ̃ez R, Cruz-Celis A, et al. Distribution of domestic Triatominae and stratification of Chagas disease transmission in Oaxaca, Mexico. Medical and Veterinary Entomology. 2000;14(1):19–30.
15. Magall ́on Gast ́elum E, Magdaleno Pen ̃aloza NC, Katthain Duchateau G, Trujillo Con- treras F, Lozano Kasten FJ, Hern ́andez Guti ́errez RJ. Distribucio ́n de los vectores de la enfermedad de Chagas (Hemiptera: Reduviidae: Triatominae), en el estado de Jalisco, M ́exico.. Centro de Investigaciones Regionales” Dr. Hideyo Noguchi”. Universidad .... 1998.
16. Ceccarelli S, Balsalobre A, Cano ME, et al. Datos de ocurrencia de triatominos amer- icanos del Laboratorio de Triatominos del CEPAVE (CONICET-UNLP). Version 1.6. Centro de Estudios Parasitol ́ogicos y de Vectores (CEPAVE).. 2022.
17. Moffett A, Strutz S, Guda N, et al. A global public database of disease vector and reservoir distributions. PLoS Neglected Tropical Diseases. 2009;3(3):e378.
18. Arroyo TF. Texas 2012-2015. 2024.
19. Rebollar-T ́ellez EA. Triatomine North Mexico Dataset. 2024.
20. GBIF.Org User . Occurrence Download. 2023.
21. contributors . iNaturalist Research-grade Observations. 2022.
22. Huang L, Sa ́nchez A, Rebollar-T ́ellez EA, et al. Triatomine Database North and South Americas 2023 version. 2024.
23. Project A. Gridded current and projected climate data for North America at 1km reso- lution, generated using the ClimateNA v7.30 software (T. Wang et al., 2022). 2022.
24. Wang T, Hamann A, Spittlehouse D, Carroll C. Locally downscaled and spatially cus- tomizable climate data for historical and future periods for North America. PloS one. 2016;11(6):e0156720.
25. Mahony CR, Wang T, Hamann A, Cannon AJ. A global climate model ensemble for downscaled monthly climate normals over North America. International Journal of Cli- matology. 2022;42(11):5871–5891.
26. Chen G, Li X, Liu X. Global land projection based on plant functional types with a 1-km resolution under socio-climatic scenarios. Scientific Data. 2022;9(1):125.
27. Chen G, Li X, Liu X. Future global land datasets with a 1-km resolution based on the SSP-RCP scenarios. Zenodo. 2021.
28. Valavi R, Guillera-Arroita G, Lahoz-Monfort JJ, Elith J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological monographs. 2022;92(1):e01486.
29. Konowalik K, Nosol A. Evaluation metrics and validation of presence-only species dis- tribution models based on distributional maps with varying coverage. Scientific Reports. 2021;11(1):1482.
30. Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. Modelling species presence-only data with random forests. Ecography. 2021;44(12):1731–1742.
31. Bennett C. Chagas disease surveillance activities—seven states, 2017. MMWR. Morbidity and Mortality Weekly Report. 2018;67.