

Pipeline

The following is a recipe for arriving at the results shown in the paper, in addition to some analyses only shown in Supplementary Information and some further, optional analyses. Sequential execution, following the numbering, is essential. The R scripts referred to contain comments with additional information.

1. Dates

1.1. Data preparation: download Glottolog

The `glottolog_languoid.csv.zip` file is downloaded from <https://glottolog.org/meta/downloads> (here from version 5.1 of Glottolog, Hammarström et al. 2024).¹ Unzipped it will yield `languoid.csv`.

1.2. Data preparation: extract data from Glottolog

The script `generate_glottolog_file.R` is run on the relevant version of `languoid.csv` to produce `glottolog_classification_strings.txt`. This arranges languages according to their ISO 639-3 codes in the first column, with successive columns showing Glottolog ID, classification, latitude, and longitude.

1.3. Data preparation: extract data from ASJP

A file with the best attested ASJP doculects is produced. One doculect is selected per ISO 639-3 code. It should be the one that has most items on the 40-item list (in case of ties there is an arbitrary choice) and it should minimally have 26 items. Ancient and reconstructed languages are excluded (i.e., those whose population size label is given as -2) as well as languages classified as "Oth[er]" in WALS (creoles, pidgins, constructed languages). The reason why ancient languages are removed is that they are inappropriate for the distance-based dating method used. The interactive ASJP software at <https://github.com/Sokiwi/InteractiveASJP02> was used for carrying out this selection, but for convenience the resulting file is included here. It is called `listss20_pruned.txt`.

1.4. Data preparation: Update Glottolog classification for ASJP data

The files `glottolog_classification_strings.txt` (produced in Step 1.2) and `listss20_pruned.txt` (produced/provided in Step 1.3) are given as input to `update_glottolog_classification.R` so as to update the Glottolog classification in `listss20_pruned.txt`. The file produced is called `listss20_pruned_updated.txt`. If a doculect is classified as a ‘dialect’ or ‘family’ in Glottolog it will not receive a Glottolog classification since this would introduce an extra level of dating. There are 190 such cases. These will not be used.

1.5. Inspect ASJP data (optional)

In the ‘Editor’s Corner’ of <https://github.com/Sokiwi/InteractiveASJP02> a tab-delimited file can be produced from `listss20_pruned_updated.txt` for the purpose of inspection. It reveals that there are 5135 ISO 639-3 code languages (of which 190 are considered dialects in

¹ Data of Glottolog 5.1 is published under the following license: <https://creativecommons.org/licenses/by/4.0/>.

Glottolog and are not used). The earliest date of extinction of a language in the pruned dataset is 1740 (the year that PUMPOKOL is reported to have gone extinct).

1.6. Produce dates

The script `dates.R` is run on `listss20_pruned_updated.txt`. It requires the file `preamble.txt` and the program `asjp62c.exe` (Holman 2011) for calculating dates using the method of Holman et al. (2011). Currently it produces 3092 dates based on from 1 to 241,713 pairs of languages, with dates running from -88 BP to 11,493 BP.

2. BayesTraits homelands

All BayesTraits in- and output files are deposited at <https://zenodo.org/records/14453224>. The following describes how they are produced and parsed.

2.1. Prepare BayesTraits input files

This step is accomplished running `imputation_and_BT_input_file_preparation.R`, which has further detail, including instructions for downloading data and software. The general procedure for using Glottolog trees with branch lengths from ASJP for BayesTraits analyses was introduced in Wichmann (2023). The process is as follows. As input for BayesTraits, locations for languages and a phylogeny with distinctive branch lengths are needed. Locations are retrieved from Glottolog and trees are prepared by fitting normalized Levenshtein distances (LDN) obtained from ASJP data to Glottolog tree topologies. The Glottolog trees are pruned such that they only contain taxa at the ‘language’ level (not dialect, various subdialect or family levels). (This involves the use of the `keep_as_tip()` function of `glottoTrees` [Round 2021]; some other tree manipulation functions used throughout this and some other parts of the pipeline are from the `ape` package [Paradis and Schliep 2019]²). The fitting of LDNs to Glottolog trees is accomplished by first turning the patristic lengths (originally 1 unit per branch) in the Glottolog trees into ultrametric distances (using the `force.ultrametric()` function of `phytools` [Revell 2024]); then the matrices of ultrametric distances are used imputing missing LDN values through polynomial fitting (through the `lm()` and `predict()` functions of R [R Core Team 2024]); finally, distinctive branch lengths are added to the Glottolog trees by least-squares estimation (using the `nnls.tree()` function of the `phangorn` package [Schliep 2011]) based on the LDN matrices (some of which may have imputed values). The script is run with the latest Glottolog classification data (version 5.1, Hammarström et al. 2024). This has 423 trees, but this number is reduced by excluding non-genuine “families” (artificial languages, bookkeeping, mixed languages, pidgins, speech registers, unattested, unclassifiable, and sign languages), isolates, and other families with less than three members. 177 families remain. It is required that for a BT input file to be prepared a family should have 3 or more taxa represented in ASJP. This further reduces the set of families to 153. The trees with branch lengths does not introduce new nodes not in the Glottolog trees, so in this sense preserve the topologies of the latter. But some nodes are collapsed, probably because branch lengths are not supported by ASJP. For instance, in the Glottolog tree for Chibchan there is a subgroup called Core Chibchan consisting of all languages except Pech. In the tree made available for BayesTraits, however, the subgroups of

² Not all R packages used are specified and referenced explicitly here, only irreplaceable, non-generic ones serving crucial steps in the pipeline.

Core Chibchan are coordinate (i.e., polytomically united) with the Pech branch, the Core Chibchan node having disappeared.

2.2. Run BayesTraits

BayesTraits.R runs BayesTraits. It specifies the ‘Geo’ model and (as the only available option) MCMC. A folder BT_output will contain the output. The process of running BayesTraits is described in the manual (Meade and Pagel 2023: 59-60), which also explains the format of the output. BayesTraits will fail to run on star-shaped phylogenies, including any tree with just three taxa. Additionally it will cease to run and produce an error message for tree with branches that are ultrashort or negative. Currently such trees include those for Austronesian [aust1307], Kru [krua1234], Sahaptian [saha1239], Totonacan [toto1251], and Turkic [turk1311]. Later in the pipeline missing homeland data for such cases will be supplied by the Minimal Distance method. Since the output will vary for each run the output used in the paper is provided here as a file called

2.3. Parse the BayesTraits output

The script parse_BT.R parses the output and collects information about homelands in a file called homelands_BT.txt. This lists the best supported homeland coordinates for each internal node in the Glottolog trees (currently numbering 8812).

3. Minimal Distance homelands

3.1. Run homelands for the Minimal Distance method

The method was introduced in Wichmann and Rama (2021). Run md_fams_subgroups.R. Output is in md_homelands_glottolog_groups.txt.

4. Dates and homelands

4.1. Match clade IDs based on classifications with IDs based on member languages

Information for clades both in dates.txt and homelands_BT.txt needs to be combined. In dates.txt each clade is identified by its Glottolog classification string. In homelands_BT each clade is identified by a string with glottocodes for all member languages separated by the underscore character. The task is now to match the two kinds of IDs. The script clas_strings_to_ids.R extracts each unique classification string from glottolog_classification_strings.txt, which is the file where ISO-code languages are listed along with their glottocodes and Glottolog classifications. It then extracts each clade from those strings. For each clade, the membership of glottocode languages is found by searching for the classification string defining the clade in the beginning of the strings in the column for classification strings in the file glottolog_classification_strings.txt. The glottocodes found are put together separated by "_" in a style similar to how clades are identified in homelands_BT.txt. Matching Glottolog classification strings and clades IDs consisting of glottocodes strung together are output to the file clstrings_ids.txt.

4.2. Combine clade IDs for dated clades and clades with BayesTraits homelands

The information in `clstrings_ids.txt` is used to match clades in `dates.txt`, which are defined as classification strings, and clades in `homelands_BT.txt`, which are defined as sets of glottocodes pertaining to the clades. The matching of clades in the two files and output to `dates_homelands.txt` is done by `match_clades.R`. If a dated clade could not be matched to a clade for which a BayesTraits homeland is available the script will look up the homeland in `md_homelands_glottolog_groups.txt`, the file with Minimal Distance homelands and supply that instead. For an entire family and its subgroups BayesTraits homelands will not be available if the phylogeny is star-shaped or if it contains ultrashort or negative branch lengths (cf. 2.2). Moreover, for some families a Glottolog subgroup is not available for BayesTraits because the corresponding node collapsed with a higher node in the input tree (cf. 2.1). The output to `dates_homelands.txt` will include an indication of which method was used in each case.

5. Diffusion events

5.1. Put together basic information on diffusion events

Run `diffusion_events.R`. This reads `dates_homelands.txt` and combines the information for each pair of mother-daughter clade. Results are in `diffusion_events.txt`.

5.2. Add more information on diffusion events

Run `diffusion_events_annotator.R`. This reads `diffusion_events.R` and produces `diffusion_events_annotated.txt`, which adds information about estimated errors on homelands (not used in the paper), distance, speed, world area, major subsistence, bearing, number and names of main reconstructed biome traversed (“biome4”; Beyer 2020), percentage of trip pertaining to the main reconstructed biome, number and names of main modern biome traversed (“biome_98”), percentage of trips pertaining the main modern biome, reconstructed altitude (not used in the paper), rugosity, and net primary productivity (not used in the paper) (all the last three are from Beyer 2020). Since this file contains all the main results it is provided for convenience, obviating the need for running the pipeline up to this point.

More detail:

Step 5.2 crucially requires the following materials (more detail on required packages in the scripts):

- Installation of `pastclim` package along with the dataset from Beyer (2020) giving ecological information for the past 120k years, at intervals of 1/2 k years, and a resolution of 0.5 degrees in latitude and longitude.
- Presence of the file `families_areas_subsistence.txt` in the working directory
- Downloaded biome mapfiles.³³ The classification of terrestrial ecoregions is described in Olson and Dinerstein (1998) and Olson et al. (2001).

For the purpose of extracting the main (majority) biome traversed (modern or reconstructed) and the percentage of the trip pertaining to this main biome a line is drawn from the locations

³³ From: https://files.worldwildlife.org/wwfcmprod/files/Publication/file/6kcchn7e3u_official_teow.zip found at the website <https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world#:~:text=There%20are%20867%20terrestrial%20ecoregions,conserve%20biodiversity%20around%20the%20world.>

of the mother, M , to that of the daughter, D . The distance d in kilometer (as the crow flies) represented by the line is divided by 10 and rounded off to an integer i . The value of i is equal to the number, N , of evenly spaced intermediate points on the line between M and D . Included among the total set of points is M and D . Thus, in principle, there should minimally be two points; this should happens when $d < 5$. But the function `gcIntermediate()` from the package `geosphere` (Hijmans 2024) will introduce an intermediate point when $N = 0$. I.e. $N = 0$ is changed to $N = 1$. Thus, for $d < 5$ there will be $2 + 0 + 1 = 3$ points, for $5 \leq d < 15$ there will be $2 + 1 = 3$ points, for $15 \leq d < 25$ there will be $2 + 2 = 4$ points, and so on. In practice, the number of points varies between 3 and 1526.

6. Plots

6.1. Statistical plots

Figures 1-3 are produced by the function `as()` in the script `diffusion_events_analysis.R`. The parameter `what` defines what is to be plotted. Running relevant code inside the function may in some cases be necessary rather than running the function itself.

6.2. Plotting homelands

Running `allfams()` of the `HPD.R` script outputs pictures of homelands for families indicating the 95% HPD region and with coloring according to probability densities. `plot_HPD_multi()`, also of `HPD.R`, allows for multiple such picture to be combined and was used for Figure 4 (running code inside the function rather than the function itself may be necessary). In order to prepare data to be plotted `get_BT_data()` is used. This requires a numbered node in the BT output, which can be identified using the `get_node_id()` function of `get_node.R`.

7. Supplementary analyses

7.1. Diffusion rates for agriculturalists vs hunter-gatherers

When the `what` parameter is set to `subsistence` `diffusion_events_analysis.R` produces a plot of mean diffusion rates across time for agriculturalists vs hunter-gatherers, used for Figure S1.

7.2. EW/NS movement rations for agriculturalists vs. hunter.gatherers

When the `what` parameter is set to `bearings_subsistence` the script `diffusion_events_analysis.R` produces barplots of ratios of EW to NS movements for agriculturalists vs. hunter.gatherers, used for Figure S2.

7.3. Comparing traversal counts for reconstructed vs. modern biomes

The script `compare_biomes.R` produces a heatmap to compare results for biomes reconstructed to the times of different diffusion events with results for modern biomes, used for Figure S3.

8. Optional further analyses

8.1. Pictures showing BayesTraits results

parse_BT_geo_out.R produces pictures showing the dispersal of lineages in each family using BayesTraits results, with coloring indicating relative time depths.

References

- Beyer, Robert M., Mario Krapp, and Andrea Manica. 2020. High-resolution terrestrial climate, bioclimate and vegetation for the last 120,000 years. *Scientific Data* 7, 236. doi:10.1038/s41597-020-0552-1
- Güldemann, Tom and Harald Hammarström. 2020. Geographical axis effects in large-scale linguistic distributions. In Crevels, Mily and Pieter Muysken, *Language Dispersal, Diversification, and Contact: A global perspective*, 58–77. Oxford: Oxford University Press.
- Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Glottolog 5.1. Leipzig: Max Planck Institute for Evolutionary Anthropology. <https://doi.org/10.5281/zenodo.10804357> (Available online at <http://glottolog.org>, Accessed on 2024-10-29.)
- Holman, Eric W. 2011. Program for calculating ASJP dates (version 1.1). <https://asjp.cild.org/software>
- Holman, Eric W., Cecil H. Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, Oleg Belyaev, Matthias Urban, Robert Mailhammer, Johann-Mattis List, and Dmitry Egorov. 2011. Automated dating of the world's language families based on lexical similarity. *Current Anthropology* 52(6): 841–875.
- Meade, Andrew and Mark Pagel. 2023. BayesTraits V4.0.0. Manual updated October 2023. <https://www.evolution.reading.ac.uk/BayesTraitsV4.0.0/Files/BayesTraitsV4.0.0-Manual.pdf>.
- Olson David M. and Eric Dinerstein. 1998. The Global 200: A representation approach to conserving the Earth's most biologically valuable ecoregions. *Conservation Biology* 12: 502–515.
- Olson, David M., Eric Dinerstein, Eric D. Wikramanayake, Neil D. Burgess, George V. N. Powell, Emma C. Underwood, Jennifer A. D'Amico, Illanga Itoua, Holly E. Strand, John C. Morrison, Coulby L. Loucks, Thomas F. Allnutt, Taylor H. Ricketts, Yumiko Kura, John F. Lamoreux, Wesley W. Wettengel, Prashant Hedao, and Kenneth R. Kassem. 2001. Terrestrial ecoregions of the world: A new map of life on Earth. *Bioscience* 51(11): 933–938.
- Paradis E, Schliep K (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35, 526–528. doi:10.1093/bioinformatics/bty633 <<https://doi.org/10.1093/bioinformatics/bty633>>.
- R Core Team. 2024. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <<https://www.R-project.org/>>.
- Revell, Liam J. 2024. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 12, e16505.
- Round, Erich R. 2021. glottoTrees: phylogenetic trees in linguistics. R package version 0.1 URL <https://github.com/erichround/glottoTrees>

- Schliep Klaus Peter. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4) 592–593
- Wichmann, Søren. 2023. Tone and word length across languages. *Frontiers in Psychology* 14: 1128461.
- Wichmann, Søren and Taraka Rama. 2021. Testing methods of linguistic homeland detection using synthetic data. *Philosophical Transactions of the Royal Society B* 376: 20200202.