Guide to Supplementary Data and Results Files, and Online Resources

for the paper …

Language trees with sampled ancestors support a hybrid model for the origin of Indo-European languages

Heggarty et al. (2023) in *Science* — https://doi.org/10.1126/science.abg0818

Contents

1. Reproducibility

To facilitate reproducibility, this guide sets out how we have made available all files and software packages needed.

The pipeline for getting from the raw IE-CoR data tables to our phylogenetic results is set out in section 5.5 of the Supplementary Materials pdf text [Suppl_Materials.pdf], provided with the main paper. Below we also direct readers as appropriate to other specific sections of the Supplementary Materials text, referred to here as 'SM'.

This guide gives details of our supplementary **files**, and other resources, which are all available online.

1. The **full language data** in the IE-CoR database at https://iecor.clld.org, as well as the binary **encodings** of those data in nexus and .xml format for input to our phylogenetic analyses, and the code used to export from IE-CoR to the corresponding nexus file format.
2. The **software** packages used to perform our phylogenetic analyses: principally BEAST, as well as new code required for our tailored specific supplementary analyses.

3., 4. Our supplementary **data and results files**, e.g. .xml input files and .tree output files, from all our phylogenetic analysis runs. These are available within our full (and) supplementary files online, at:
https://share.eva.mpg.de/index.php/s/E4Am2bbBA3qLngC
All of these data and results files are also available within folder .zips published on Zenodo, under this title:
Supplementary Data and Results Files for Heggarty et al. (2023) in Science, https://doi.org/10.1126/science.abg0818

Here we use colour coding to distinguish URLs from [Folder_Names] and [File_Names] among our supplementary files.

2. Language Data

2.1 IE-CoR: The Indo-European Cognate Relationships Database

The full language data used in our analyses can be explored through our online IE-CoR database app at:
https://iecor.clld.org.

The raw data tables that underlie IE-CoR can be downloaded within the full CLDF dataset for IE-CoR 1.0, from:
https://zenodo.org/record/8089434.

2.2 Language Data Encoded in Binary Format: Nexus and .xml Files

So that they can be used by quantitative and phylogenetic analysis software, the language data in the IE-CoR raw data tables need to be converted into a dataset in a format that such software can process. In particular, most such analyses require the data to be **encoded** as binary data characters, i.e. with the presence/absence of each cognate set in each language taxon expressed in binary form: 0=absent, 1=present.

We therefore also make available the IE-CoR language data encoded as input to our phylogenetic analyses, in the form of a binary data matrix, in files of two formats , nexus and .xml.

In a first stage, we exported the IE-CoR data into a data file in the widely used **nexus format**. This was done using the export script [make_nexus.py], written by Hans-Jörg Bibiko and available within [05_Software], and at:

https://github.com/lexibank/iecor/blob/master/iecorcommands/make_nexus.py.

This script also exports the time calibrations for each non-modern language, also part of the IE-CoR data.

In a second stage, we prepared from the nexus files the corresponding **.xml file** for each analysis run, in the BEAST phylogenetic analysis software. The binary data matrix and time calibrations remain the same in the nexus and .xml file, although the latter includes the specifications of further parameters needed for each analysis run.

We make our nexus files available in the top-level folder [04_Nexus_Data_Files]; most analyses were based on the same main nexus file. The corresponding .xml file, with the further specification for each individual analysis run, is available within the sub-folder that corresponds to each phylogenetic analysis, as detailed in section 3 below.

2.3 Structure of the Binary Data Grid

In the nexus (and corresponding .xml) file for the **main analysis** (and all others except those detailed below), the language data matrix is a block of 161 rows by 4990 columns. The 161 rows correspond to the 161 language taxa in IE-CoR, while the 4990 columns correspond to cognate sets. The values in the matrix are in binary form, as follows:

- 0 = cognate set 'absent', or strictly: in this language (row), the *primary* lexeme for the precise IE-CoR target meaning does not belong to this cognate set.
- 1 = cognate set 'present', or strictly: in this language (row), the *primary* lexeme for the precise IE-CoR target meaning does belong to this cognate set.
- ? = missing data, i.e. there is no reliable enough record of what was the primary lexeme (and thus cognate set) for this target meaning in this language. (Generally only for some meanings in poorly attested ancient languages.)

The 4990 columns correspond to the 4820 cognate sets in IE-CoR that were included in this main data export (given the parameter settings for exclusions, see below), plus one extra column for each of the 170 IE-CoR reference meanings, for **ascertainment bias correction**. This column is in effect 'empty' (of 1s), in that it contains only 0s or (for languages that have no data for that meaning) ?s.

2.4 How to Read the Nexus Files

The nexus files contain additional user-friendly comment lines, as rich meta-data so that users can directly and immediately identify and refer, through the IE-CoR database website, to the actual language data (individual lexemes in each language) that correspond to any 1 in the matrix. We explain below how to read these comment lines.

- The nexus file begins with comment lines that list the inclusion or **exclusion parameters** for that export. In particular, these show whether and how cognate sets identified as parallel loans are included in that export (varied only for analysis SA9), along with other possible exclusion parameters such as ideophony or parallel derivations (set to true only for very few cognate sets, which were excluded in all analyses here). These settings determine how many cognate set columns are present in the data matrix in any individual exported nexus file, relative to the total number of 5013 cognate sets in IE-CoR. (This count is also affected by the ascertainment bias correction, which adds one 'empty' (no 1s) row/column per meaning, i.e. 170 extra rows/columns.)
- Next, the nexus file lists (from line 27 onwards) all cognate sets grouped by meaning, in alphabetical order of the IE-CoR reference meanings, from ANT to YESTERDAY. Each meaning block begins with an extra "_group" row, before any actual cognate sets for that meaning, for ascertainment bias correction.
- Just before the main data matrix, the file contains an additional comment line (5816 in the nexus file for the main analysis) to show which ranges of columns in the data matrix correspond to each of the 170 meanings in the IE-CoR reference set.
- Each meaning block is preceded by an 'empty' column for the ascertainment bias correction.

- The nexus file also contains the precise IE-CoR cognate set reference number that corresponds to each column in the data matrix. This is given in comment lines immediately before the main data matrix (lines 5820-5823 in the nexus file for the main analysis). Since this has to be given per single column, but IE-CoR has over 5000+ cognate sets, the IE-CoR number for each cognate set is shown arranged vertically in that column, digit by digit (up to four digits, as needed for numbers up to 5000+). So the very first data column, which corresponds to IE-CoR cognate set 5007, has the digits 5 0 0 7 in the successive comment lines 5820-5823. This can then be looked up online at https://iecor.clld.org/cognatesets/5007, and corresponds in this case to the most widely used cognate set across the Indo-European family for the first IE-CoR meaning ANT. Indeed within each meaning, cognate sets are sorted by how many of the 12 main branches of Indo-European they occur in, starting with whichever cognate set is found in the most branches. For ANT, the most widespread cognate set 5007 is defined as words that all originated in the Proto-Indo-European wordform **moru̯i-*. The IE-CoR webpage lists the individual lexemes that belong to that cognate set in each language, e.g. Ancient Greek *mýrmēx* (μύρμηξ), Old Icelandic *maurr*, Serbo-Croat *mrav*, and so on — those languages' words for ANT. Note that this ordering and the use of 12 standard reference clades is only to allow easier and more user-friendly exploration of the data block. Neither has any role or effect whatsoever on the actual phylogenetic analysis. For that analysis, no clade structure is built into the data-set itself, and the ordering of cognate sets within a meaning is irrelevant.
- The meaning names and cognate set numbers are repeated in comment lines also immediately *below* the main data matrix.

2.5 Variations in the Binary Data Matrix Across Analyses

All of our different model tests (M1 to M4), and many of our sensitivity analyses, use as their input exactly the same **language data matrix** as just described, based on the same, main nexus file. Some of the sensitivity analyses, however, test variations in the precise data-set used, and are therefore based on separate nexus files that can also be found within the top-level folder: [04_Nexus_Data_Files].

- Sensitivity analysis SA2 followed a different treatment of cognate sets of the 'parallel loan' type in IE-CoR, making for a slightly larger number of cognate sets (columns) in the binary data matrix — see SM §7.2.
- Sensitivity analysis SA5 excluded ten languages heavily affected by missing data, making for a slightly smaller number of languages (rows) in the binary data matrix — see SM §7.5.
- Sensitivity analysis SA8 uses a data matrix that differs enormously, because it did not use IE-CoR data at all, but for comparative purposes used instead the 'broad' subset (as defined by Chang et al. 2015) of the earlier IELex data-set — see SM §7.8 and Heggarty (2021). This file therefore also does *not* contain the same user-friendly meta-data as in the nexus files exported from IE-CoR.
- Sensitivity analysis SA9 does in fact start out from the same language data matrix as in most other analyses, although during the analysis run the tree likelihoods estimated for 57 of the 170 meanings (those with polymorphism at the Proto-Indo-European root) were excluded and did not go forward into the analysis — see below under 4.3, and SM §7.9
- Sensitivity analysis SA10, which used a recoding of cognate sets not in binary but in multistate terms. That is, each cognate set is taken in this analysis not as a binary character either present 1 or absent 0, as in all other analyses, but as one state of the corresponding IE-CoR meaning, each taken as a multistate data character (170 in total). This entails no binary data block in the .xml file, but a completely different data structure, converted into the multistate format in the .xml file specific to that analysis.

Further details on the precise differences between analyses are given in section 4.3 below.

3. Software for Bayesian Phylogenetic Analyses

- The main Bayesian phylogenetic analysis software used in this paper, **BEAST version 2.6.5**, is freely available at: https://www.beast2.org
- In addition, to implement the three ancestry constraint analyses within the **SA7** set of sensitivity analyses we used the BEAST2 **sampled-ancestors package**, available at: https://github.com/CompEvol/sampled-ancestors
- This includes in particular the **AncestryConstraint.java** code newly written by Denise Kühnert, available at: https://github.com/CompEvol/sampled-ancestors/blob/master/src/sa/evolution/tree/AncestryConstraint.java
- Sensitivity analysis **SA10** used a BEAST2 add-on package newly written by Benedict King to implement a multistate phylogenetic analysis model, the code for which is available at: https://github.com/king-ben/ConceptModels.

Copies of the last two files can also be found within the folder [05_Software].

4. Data and Results Files: Online Access and Folder Structure

4.1 Browsable Folder Structure — Or Pre-Packaged Zip Files for Download

The input data files and output results files from our phylogenetic analyses are freely available online, at:

https://share.eva.mpg.de/index.php/s/E4Am2bbBA3qLngC

All files can be explored within the folder structure there, and single, multiple or all files can be freely downloaded.

All files are also available within folder .zips published on Zenodo under this title:

Supplementary Data and Results Files for Heggarty et al. (2023) in Science, https://doi.org/10.1126/science.abg0818

The entire set of results and data files makes for a total of 12.9 GB, however, so for faster downloading we also provide a set of ready-made .zip files, available at the above link within the top-level folder [PrePackaged_Zip_Files].

The full set of all data and results files zips down to 4.9 GB, as:

- [IECoR_Suppl_Files_Full.zip] 4.9 GB
  All data and results files for all analyses, including the very large results files.

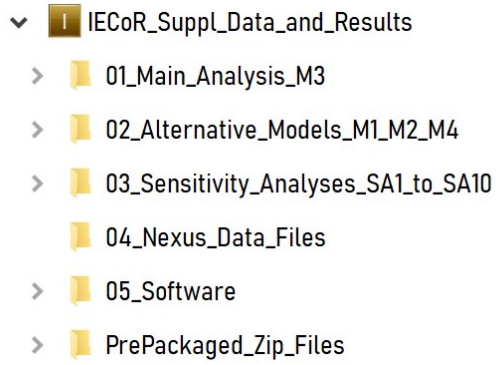
We expect, however, that many users will need only more limited subsets of the main results files. So we have also provided three slimmed down subsets, already gathered into individual .zip files for one-stop downloads.

- [IECoR_Suppl_Files_Full_Main_M3_Only.zip] 487.9 MB
  All data and results files for the main analysis only (M3), without the other models or the sensitivity analyses.
- [IECoR_Suppl_Files_Light.zip] 158.0 MB
  The core data and results files for all analyses, but without the very large results files (full .trees and .log files).
- [IECoR_Suppl_Files_Light_Main_M3_Only.zip] 16.3 MB
  The core data and results files for the main analysis only (M3), without the other models or the sensitivity analyses.

Both [IECoR_Suppl_Files_Full.zip] and [IECoR_Suppl_Files_Light.zip] unzip to the same folder structure as set out here.

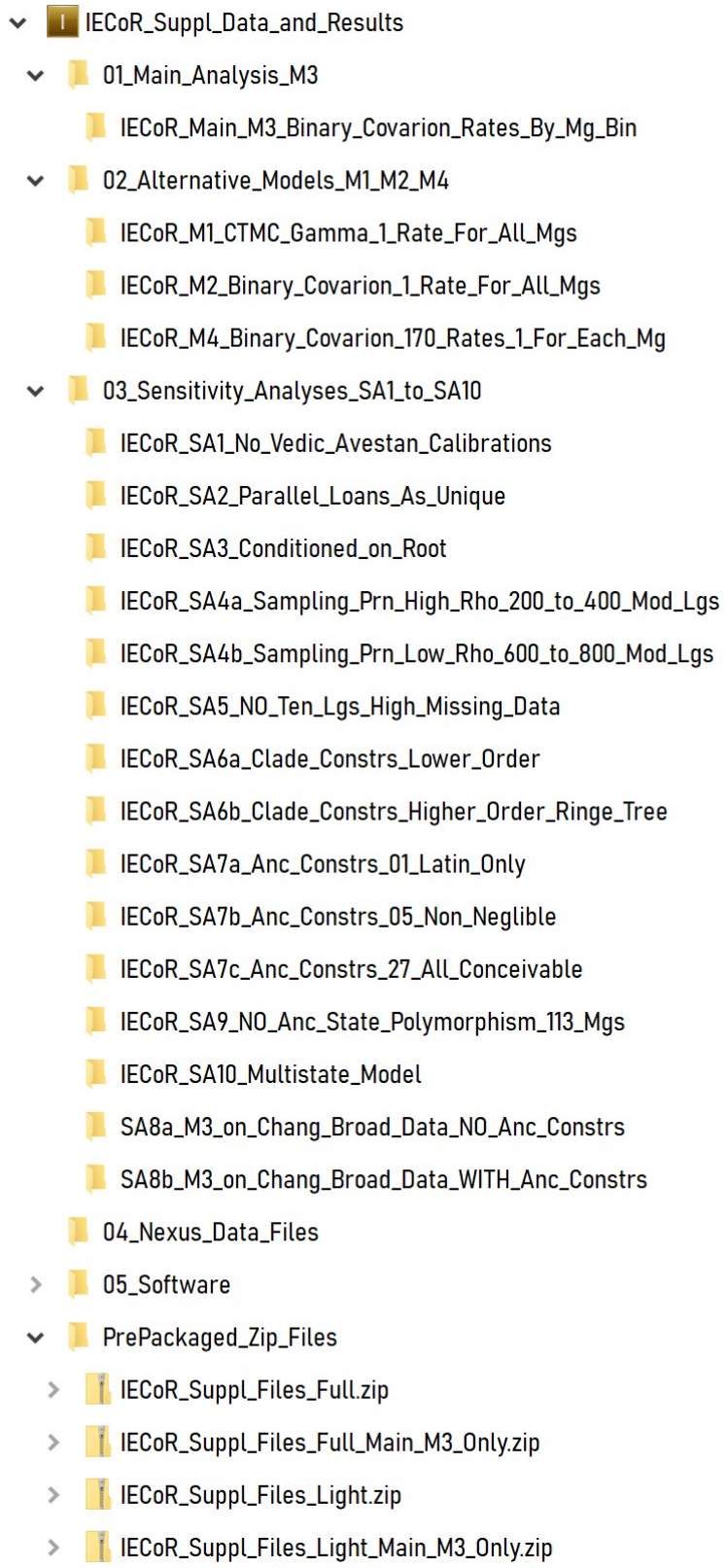
4.2 Folder Structure

The online IE-CoR supplementary data and results files are arranged in five top-level folders, plus the sixth folder [PrePackaged_Zip_Files], as shown below: top-level folders (left) and full expanded sub-folder structure (right).



- The contents of top-level folders 01-03 are explained in sections §5.1 to §5.3 below.
- Top-level folder 04 contains the language data as encoded in the nexus format export files (see §2.2 above).
- Top-level folder 05 contains the dedicated software newly coded for this research paper (see §2.2 and §3).

:



5. Data and Results Files: Details

5.1 Main Phylogenetic Analysis (Using Model M3)

We tested four different phylogenetic analysis models (M1 to M4), as explained in SM §5.

Of these four models, the best performing (see Table S4.1) was M3, using the binary covarion model and with meanings were grouped into one of eight bins, according to the number of different cognate sets found per meaning across all languages in IE-CoR. Rates of change in cognate status vary by bin (see Figure S4.1). As the best-performing model, M3 was therefore taken as our 'main analysis', on which we principally report in this paper.

The data and results files for this main analysis are within the top-level folder [01_Main_Analysis_M3], inside the single sub-folder [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin]. This contains firstly:

- The input **.xml** file, including the language data matrix, date calibrations, and the setup of the prior distributions: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin.xml]

With this input file, we performed three independent Beast analysis runs, starting from different random seeds (see Table G5.4 below), and where necessary (see §5.4) we resumed each run once, from another different seed.

When completed, we combined the log files and posterior distributions from each run (hence _combined in their filenames). The results files include:

- The **.log** file of the three combined phylogenetic analysis runs, including the posterior distribution of the main parameters: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin_combined.log]
- The **_mcc.tree** results summary file, which is the median Maximum Clade Credibility (MCC) summary tree of the posterior trees distribution, used to produce Figure S5.1: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin_combined-mcc.tree]

The full results also include two other very large files (190.9 MB and 568.6 MB),

- The **_per_meaning.log** file is an extension of the .log file, containing tree likelihoods per meaning, the sampled ages of the extinct languages, and the posterior distribution of further parameters: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin_combined_per_meaning.log]
- The **.trees** results file contains the full posterior distribution of all trees, from which the DensiTree (Figure 4 in the main article) and MCC summary tree (Figure S5.1 in the supplement) were calculated: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin_combined.trees]

Within the zipped files, the 'Light' set [IECoR_Suppl_Files_Light.zip] does not include these two very large files.

For this main analysis M3 only, the [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin] folder also contains:

- The **_PRIOR.log** file of the analysis run used to create the prior tree distribution: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin_combined_PRIOR.log]
- The **_PRIOR.trees** file contains the full prior distribution of all trees: [IECoR_Main_M3_Binary_Covarion_Rates_By_Mg_Bin_combined_PRIOR.trees]

Since the _PRIOR.trees file is so large (519 MB), it too is not included in the 'Light' zipped file [Suppl_Files_Light.zip].

If data and results files are desired *only* for this main M3 analysis, they can be downloaded as:

- [IECoR_Suppl_Files_Full_Main_M3_Only.zip] 476.5 MB
- [IECoR_Suppl_Files_Light_Main_M3_Only.zip] 9.5 MB

5.2 Alternative Models: M1, M2, M4

The files for each of the other three models (see SM §5 and Table S5.1) are within the folder [02_Alternative_Models_M1_M2_M4], in its three sub-folders:

- [IECoR_M1_CTMC_Gamma_1_Rate_For_All_Mgs]
- [IECoR_M2_Binary_Covarion_1_Rate_For_All_Mgs]
- [IECoR_M4_Binary_Covarion_170_Rates_1_For_Each_Mg]

Each of these sub-folders includes same files as in the Main Analysis with the M3 model, except that the _PRIOR.log file and _PRIOR.trees files are *not* provided for these three alternative model runs.

Within the zipped files, the 'Light' sets include only the input **.xml** file, the analysis run **.log** file, and the **_mcc.tree** results summary file. The 'Full' sets additionally include the much larger **_per_meaning.log** file and the full **.trees** results file.

As in the main analysis, for each of the models M1, M2 and M4 we also performed three independent Beast analysis runs, starting from different random seeds, and resuming each once from another random seed, before combining the logs and posterior distributions from each. The random seeds used are reported in Table G4.4 below.

5.3 Sensitivity Analyses

As detailed in SM §7-§7.8, as robustness tests we also ran a series of sensitivity analyses, SA1 to SA8. Table G4.3 below identifies these sensitivity analyses, gives the name of the corresponding sub-folder within which the data and results files are found, and indicates which section in the Supplementary Materials pdf file provides the full explanation of that analysis.

The input data and results files for these sensitivity analyses are to be found in the corresponding sub-folders within the top-level folder [03_Sensitivity_Analyses]. Each of these analyses uses the same language data, models and setup of the prior as for the main analysis M3, **except** for the aspect that was specifically varied for that sensitivity analysis. We now explain how the input .xml file for each sensitivity analysis *differs* from that used for the main analysis.

- SA1 differs from the other analyses in the calibrations section of the input .xml file, from which the **date calibrations** for two language taxa (Early Vedic and Younger Avestan) have effectively been removed.
  - Found in the sub-folder: [IECoR_SA1_No_Vedic_Avestan_Calibrations]

- SA2 differs in having a language data-set (i.e. the binary data matrix within the input .xml file) somewhat different to that used in most other analyses. The data matrix in SA2 differs only in the alternative handling of one type of loanwords (**parallel loan** sets), as explained in SM §3.8.3 and SM §7.2. This adds to the data matrix 789 more columns (5779 cognate sets, rather than 4990 in the main analysis), although each is a 'singleton' cognate set, such that in each of the 789 new columns, the binary value 1 (present) occurs in only a single language taxon (row).
  - Found in the sub-folder: [IECoR_SA2_Parallel_Loans_As_Unique]

- SA3 differs in the setup of the prior tree distribution within the input .xml file, in that the prior is **conditioned on the origin**, not the root.
  - Found in the sub-folder: [IECoR_SA3_Conditioned_on_Root]

	Sensitivity Analysis	Name of Sub-Folder	See…
SA1	With Vedic and Avestan tip calibrations removed	IECoR_SA1_No_Vedic_Avestan_Calibrations	SM §7.1
SA2	With parallel loans as unique cognate sets	IECoR_SA2_Parallel_Loans_As_Unique	SM §7.2
SA3	With prior conditioned on MRCA, not origin	IECoR_SA3_Conditioned_on_Root	SM §7.3
SA4a	Sampling probability assuming 200-400 language taxa	IECoR_SA4a_Sampling_Prn_High_Rho_200_to_400_Mod_Lgs	SM §7.4
SA4b	Sampling probability assuming 600-800 language taxa	IECoR_SA4b_Sampling_Prn_Low_Rho_600_to_800_Mod_Lgs	SM §7.4
SA5	Without ten languages with high missing data rates	IECoR_SA5_NO_Ten_Lgs_High_Missing_Data	SM §7.5
SA6a	With targeted lower-order **clade** constraints	IECoR_SA6a_Clade_Constrs_Lower_Order	SM §7.6
SA6b	With higher-order **clade** constraints to enforce Ringe & Anthony (2015) tree topology	IECoR_SA6b_Clade_Constrs_Higher_Order_Ringe_Tree	SM §7.6
SA7a	With **ancestry** constraint: Latin only.	IECoR_SA7a_Anc_Constrs_01_Latin_Only	SM §7.7
SA7b	**Ancestry** constraints: 5 with non-negligible support	IECoR_SA7b_Anc_Constrs_05_Non_Neglible	SM §7.7
SA7c	**Ancestry** constraints: all 27 remotely conceivable	IECoR_SA7c_Anc_Constrs_27_All_Conceivable	SM §7.7
SA8a	M3 on Chang et al. (2015) broad data-set (*not* IE-CoR) with no ancestry constraints	SA8a_M3_on_Chang_Broad_Data_NO_Anc_Constrs	SM §7.8
SA8b	M3 on Chang et al. (2015) broad data-set (*not* IE-CoR) with their ancestry constraints	SA8b_M3_on_Chang_Broad_Data_WITH_Anc_Constrs	SM §7.8
SA9	**Excluding** 57 meanings reconstructed in main analysis as having ancestral state polymorphism per meaning	IECoR_SA9_NO_Anc_State_Polymorphism_113_Mgs	SM §7.9
SA10	Using a **multistate** (not binary) model of cognate evolution, on the IE-CoR data in multistate format	IECoR_SA10_Multistate_Model	SM §7.10

Table G5.3 List of sensitivity analyses and corresponding folder names

- **SA4a** and **SA4b** differ also in the setup of the prior tree distribution within the input .xml file, in that they vary the **sampling proportion** by setting it higher and lower than in the main analysis.
  - Found in the sub-folder: [IECoR_SA4a_Sampling_Prn_High_Rho_200_to_400_Mod_Lgs]
  - Found in the sub-folder: [IECoR_SA4b_Sampling_Prn_Low_Rho_600_to_800_Mod_Lgs]

- **SA5** differs, like SA2, in having a language data-set (i.e. the binary data matrix within the input .xml file) that differs somewhat from that used in all other analyses. In SA5, ten languages heavily affected by missing data are removed, so this .xml file includes data for only 151 rather than the full 161 language taxa. Cognate sets present only in these ten excluded languages are thus also excluded. The data matrix thus contains ten fewer rows (161 rather than 151 language taxa) and 33 fewer columns (4967 rather than 4990 cognate sets). For full details, see SM §7.5.
  - Found in the sub-folder: [IECoR_SA5_NO_Ten_Lgs_High_Missing_Data]

- **SA6a** and **SA6b** differ from the main M3 analysis in that the input .xml file for each contains the specification of the particular set of **clade** constraints enforced in that analysis, see SM §7.6.
  - Found in the sub-folder: [IECoR_SA6a_Clade_Constrs_Lower_Order]
  - Found in the sub-folder: [IECoR_SA6b_Clade_Constrs_Higher_Order_Ringe_Tree]

- **SA7** differs in adding **ancestry** constraints to the analysis: three alternative sets, constraining 1, 5 and 27 ancient languages to be directly ancestral to later languages. SA7 thus differs in using additional BEAST code modified by a methodological innovation to implement a new analysis of such ancestry constraint specifications (see above under "Software for Bayesian Phylogenetic Analyses"). This allows an individual ancient or historical language to be constrained to be directly ancestral to a particular set of later, 'descendant' languages, as specified in the constraint specifications within the input .xml file specific to each analysis. For full details, see SM §7.7.
  - Found within the sub-folder: [IECoR_SA7a_Anc_Constrs_01_Latin_Only]
  - Found within the sub-folder: [IECoR_SA7b_Anc_Constrs_05_Non_Neglible]
  - Found within the sub-folder: [IECoR_SA7c_Anc_Constrs_27_All_Conceivable]

- **SA8a** and **SA8b** differ from all model analyses (M1 to M4) and all other sensitivity analyses (SA1 to SA7) in employing a completely different language data-set: not our own new IE-CoR data-set but the 'broad' subset of the IELex database, as defined and used in Chang et al. (2015). The corresponding binary data grid is included in the .xml file, which furthermore includes the specification of the very extensive set of 35 clade constraints used by Chang et al. (2015: Table 9), and also applied in this near-replication analysis. SA8a and SA8b differ from each other in that the SA8a analysis applied no ancestry *constraints* (although it did use our ancestry-*enabled* approach), whereas SA8b did apply the 8 ancestry constraints of Chang et al. (2015: Table 2). For full details, see SM §7.8.
  - Found within the sub-folder: [SA8a_M3_on_Chang_Broad_Data_NO_Anc_Constrs]
  - Found within the sub-folder: [SA8b_M3_on_Chang_Broad_Data_WITH_Anc_Constrs]

- **SA9** returns to the main IE-CoR database, but differs in that the analysis **excludes** all the data from any meanings for which ancestral state reconstructions (from the main M3 analysis) revealed that the model had returned more than one cognate set as present (i.e. **polymorphism**) in that meaning at the Proto-Indo-European root. Of the 170 IE-CoR reference meanings, 57 showed a posterior probability of root polymorphism above a threshold set conservatively at a probability of just 0.1. In SA9 these 57 meanings were excluded, and the SA9 analysis is based on only the remaining subset of 113 out of the total IE-CoR meanings (66.5%).

  The binary data matrix within the input .xml file in fact retains the full data-set as in most analyses. The model, however, in any case calculates a tree likelihood for each meaning, so in SA9 the tree likelihoods for the 57 meanings with root polymorphism were simply excluded, and did not go forward into the analysis (as specified in the sections of the xml code on "treeLikelihood."). So although the full cognate dataset appears at the start of the .xml file, only the cognate sets in 113 of the 170 meanings were actually used. For full details, see SM §7.9.
  - Found within the sub-folder: [IECoR_SA9_NO_Anc_State_Polymorphism_113_Mgs]

- **SA10** does use the main IE-CoR data-set, but differs from all model analyses (M1 to M4) and all other sensitivity analyses (SA1 to SA9) in taking the same data but expressing them in an alternative encoding format: not binary but **multistate**. This is because the analysis model used in SA10 likewise differs from all other analyses, in being a multistate not binary model. For full details, see SM §7.10.
  - Found within the sub-folder: [IECoR_SA10_Multistate_Model]

The sub-folder for each sensitivity analysis includes the same set of files as in the Main Analysis (see §5.1 above).

Within the zipped files, the 'Light' sets include only the input **.xml** file, the analysis run **.log** file, and the **_mcc.tree** results summary file. The 'Full' sets additionally include the much larger **_per_meaning.log** file and the full **.trees** results file.

5.4 Analysis Runs: Resumptions, Burn-Ins, and Random Seeds Used

For each model (M1 to M4) and each sensitivity analysis (SA1 to SA10) we performed **three** independent Beast analysis runs, i.e. three independent MCMC chains, starting from different random seeds, each for an initial 100 million steps. If, after 100 million steps, the estimated sample size (ESS) for all parameters had already reached the desired minimum of 200, then that run was taken as completed. If not, then the run was resumed for a further 100 million steps, starting from a different seed. The random seeds used, and any resumptions, are shown in Table G4.4 below.

The results for each analysis are the combination of the posterior distributions from each of these three analysis runs, including any resumptions.

Burn-in was set to 10,000,000 for every analysis run, except for the M1 analysis, and for resumptions of any analyses. For M1, two of the three analysis runs initially got stuck on sub-optimal peaks in tree-space, so run 2 required a burn-in of 70,000,000 while run 3 required a burn-in of 80,000,000. (See the [readme burnin seeds.txt] file in the M1 folder.) Where any run was resumed, for the log file after resumption the burn-in was set to 0.

For further detail on the analysis runs, see SM §5.4.

ID	Analysis	Runs	Seeds Used (two entries if analysis resumed)		
M3 Main	Binary Covarion with binned rates	3 (1 resumed)	1627120501827	1627120501531 1628767647568	1629105323605
M1	CTMC Gamma	3 (2 resumed)	1627120501882	1627120501844 1628766687383	1627120501955 1628766687382
M2	Binary Covarion with 1 rate for all meanings	3	1627120501712	1627120501776	1627120501492
M4	Binary Covarion with 170 rates (1 per meaning)	3 (3 resumed)	1627120501958 1628768079150	1627121593517 1628768079150	1630358876762 1630577992936
SA1	With Vedic and Avestan calibrations removed	3	1627331212541	1627331212312	1627331211998
SA2	With parallel loans as unique cognate sets	3	1627128486871	1627128486971	1627128486893
SA3	With prior conditioned on MRCA, not origin	3	1627331212257	1627331212480	1627331212508
SA4a	Sampling proportion assuming 200-400 taxa	3	1627331212440	1627331212404	1627331212553
SA4b	Sampling proportion assuming 600-800 taxa	3	1627331212497	1627331216661	1627331216662
SA5	Without ten languages with high missing data	3 (3 resumed)	1627128404288 1628770579152	1627128404394 1628770579155	1627128404412 1628770579152
SA6a	With targeted lower-order **clade** constraints	3	1629815182502	1630358876775	1629815182504
SA6b	With higher-order clade **constraints** (Ringe tree)	3	1627425698817	1627425698846	1627426007182
SA7a	With 1 **ancestry** constraint (Latin)	3	1627399912814	1627399912779	1627399912908
SA7b	With 5 **ancestry** constraints	3	1627400177550	1627399912852	1627399912766
SA7c	With 27 **ancestry** constraints	3	1629819865320	1629819865318	1630358876766
SA8a	On IELex 'broad' data-set, ancestry-enabled	3	1628609069855	1628609069604	1628609069674
SA8b	On IELex 'broad' data-set, ancestry-constraints	3	1628609069820	1628609069854	1628609069897
SA9	Without 57 meanings reconstructed in main M3 analysis with ancestral state **polymorphism**	3 (3 resumed)	1664208264134 1664362172715	1664457104149 1664962043157	1664457110018 1664962111399
SA10	Using a new **multistate** (not binary) model	3	1654095519292	1664455791828	1664456180594

Table G5.4: Random seeds used in all analysis runs