Supporting Information

Pagel et al. 10.1073/pnas.1218726110

SI Text

Eurasiatic Language Superfamily. Dating back to Alfredo Trombetti's 1905 (1) monograph, several authors have proposed a Eurasiatic language superfamily uniting a core group of the Indo-European, Altaic, Uralic, Eskimo-Aleut, and Chukchi-Kamchatkan language families (2–8). Different authors vary in the inclusion or not of other language families. Greenberg's (6) proposal has probably received the most attention and he includes Nivkh and Etruscan but not Dravidian and Kartvelian. Here we follow the Languages of the World Etymological database (LWED, see below) and include these latter two families. Ethnologue (9) and Ruhlen (10) provide descriptions of the geographical extent of these families, as summarized below.

Altaic is a proposed language family that today comprises 64 living languages spoken widely across northern and Central Asia, and including Turkic, Mongolian, Tungusic, and Japonic languages. The family is named for the Altai mountains of Western Mongolia, where it might have originated or was at least once thought to be centered. The Orkhon inscriptions date back to the eighth century AD (10). Altaic is a controversial family with proponents noting many similarities among its languages, thought to indicate common descent. Opponents of Altaic suggest these similarities arise from widespread adoptions among speakers of languages living in close proximity.

Chukchi-Kamchatkan (also Chukotko-Kamchatkan and Chukchee-Kamchatkan) contains five languages whose speakers live predominantly in northeastern Siberia.

Dravidian comprises 73 languages found in parts of India, Pakistan, and Afghanistan. The precise origins of the Dravidian language family are unclear, but they have been epigraphically attested since the sixth century BC, and it is widely believed that Dravidian speakers must have been spread through India before the arrival of the Indo-European speakers (10, 11).

For Eskimo, linguists have historically used the name Eskimo-Aleut to refer to the languages of the indigenous peoples of far north-eastern Russia, parts of Alaska, and Greenland. However, Eskimo is now considered an outdated term politically and socially and the LWED does not include the Aleut languages, so we will refer to this group as Inuit-Yupik to denote the languages the LWED includes.

Indo-European is the fourth largest language family in the World (after Austronesian, Niger-Congo, and trans-New Guinea), with 430 living languages. Recent evidence suggests it arose around 8,000–9,000 y ago (12) and then spread throughout Europe and into present day Iran, Afghanistan, Pakistan, and India with the advent of farming (13).

Kartvelian comprises only five extant languages. Today, speakers of Kartvelian languages live in the country of Georgia and some parts of southern Russia and Turkey.

Uralic has 36 languages. Most Uralic speakers live in northern Europe (with the exception of Hungary) and northern Asia, extending from Scandinavia across the Ural mountains into Asia (10). Ruhlen (10) describes three hypotheses for the original homeland of Uralic people: a region including the Oka River south of Moscow and central Poland, the Volga and Kama Rivers, or western and northwestern Siberia. See also ref. 14 for a discussion of Uralic.

Languages of the World Etymological Database. The LWED is part of the Tower of Babel project founded by the late Sergei Starostin and his team of researchers, and including contributions from Anna Dybo, Vladimir Dybo, Alexander Militarev, Oleg Mudrak, Sergei Nikolayev, Ilia Peiros, George Starostin, Olga Stolbova, John Bengtson, Merritt Ruhlen, William Wang, George Van Driem, R. Rutgers, and J. Tolsma (http://starling.rinet.ru/cgi-bin/main.cgi). The LWED is also affiliated with the Evolution of Human Languages project at the Santa Fe Institute (http://ehl.santafe.edu/main.html). The etymological database contains reconstructed forms (proto-forms or proto-words) and proposed cognacy relations for 41 language families spanned by five major long-range reconstructed macrofamilies: Macro-Khoisan, Austric, Sino-Caucasian, Afroasiatic, and Nostratic. The LWED includes the seven Eurasiatic language families we study within Nostratic.

The LWED is unique in providing such a range of reconstructions and is in substantial agreement with others' proposed proto-words for the Indo-European and Uralic language families. Many proto–Indo-European (PIE) proposals, including those in the LWED, take the widely used Pokorny Dictionary (15) as their starting point, and the LWED's proto-Uralic (PU) reconstructions have an 80% agreement with Janhunen's reconstructions (16). We used the LWED cognacy judgements for the Chukchi-Kamachatkan family to derive a phylogenetic tree for those languages (see *Phylogenetic Inference*, below), and found a tree that fits with expectations for that family (9, 10).

Reconstructed Proto-Words. We recorded the reconstructed proto-words as proposed in the LWED for each of the 200 meanings in the Swadesh fundamental vocabulary for the seven language families. We excluded 12 meanings from the list of 200 for which the LWED provided reconstructions for only one or at most two language families. These words are: "and" (conjunction), "at" (preposition), "because" (conjunction), "here" (adverb), "how" (adverb), "if" (conjunction), "in" (preposition), "some" (adjective), "there" (adverb), "when" (adverb), "where" (adverb), and "with" (preposition).

Often, more than one proto-word is reconstructed for a meaning, reflecting the uncertainty as to the true ancestral word. In deciding which proto-words to include or exclude for a given meaning, we sought to adhere to the precise meaning. For example, for the item "hand" we excluded all modified versions of that meaning (for example, "left-hand," "take into hands," "palm of hand") but did allow plural forms (i.e., "hands"). For adjectives, such as "dry," we also accepted "to be dry." We also included words with additional meanings alongside the one being explored (that is, polysemous words), such as, in proto–Inuit-Yupik (PIY) the form *anəʁ, meaning both "spark" and "fire" was included under the meaning "fire."

The meanings "to cut" and "to burn" have 26 and 21 reconstructed forms in PIE, respectively. This variety probably arises from their vague or general nature. For example, "to burn" can be used in the sense of cooking, as in "to sear" and "to boil," but also in the sense of the sun burning, as in "to glitter," "to shine," and "to scorch." The word can be used in terms of mood, as in "to be angry" and "to grieve," as having to do with temperature rising, such as in "to heat" and "to dry," or with a change of state, as in "to turn black" and "to turn to ashes."

Recording all of the proto-words that met the criteria outlined above, we identified 3,804 proto-words among the 200 meanings in seven language families. The mean number of proto-words per meaning is 2.89 ± 2.81 (SD), but this number is strongly influenced by a few outliers with large numbers of reconstructions, such as "to cut" or "to burn." Thus, the median number of proto-words per meaning is two, and the modal number is one (Fig. S1).

Cognate Sets and Cognate Class Sizes. We used the LWED to obtain information as to whether reconstructed words from different language families are putative cognates or not. Two proto-words are cognate if they are judged to derive from a common ancestral word-meaning pairing. Where two or more proto-words with the same meaning are cognate across language families, they are coded as forming a "cognate set." To ensure reliability, we adopted a conservative coding when constructing cognate sets, accepting as cognate only those proposed proto-words that preserved exact meaning. Thus, for example, the proto-Dravidian (PD) form *er-Vc-, meaning "wild dog," and the proto-Kartvelian (PK) form *xwir-, meaning "male dog," are both rejected as reconstructions for the meaning "dog" because they are too narrow in their semantic definition. Furthermore, we required a two-way correspondence in the meanings and cognacy judgements. For example, for the numeral "two," we excluded the PU form *to-ńće, meaning "second" and the PK form tq̇ub-, meaning "twins," even though the LWED judges these forms to be cognate to the PIE form *duwo and the Proto-Altaic form *t̯i̯ubu, both of which mean "two," and both of which are indicated in the LWED as being cognate to the other.

We scored each meaning for its cognate class size, which is defined as the number of language families that are judged cognate for a particular word. For example, for the meaning "we" (first person plural), the Indo-European proto-word *we- is listed as cognate with the Kartvelian proto-word *čwen-, but not with any other proto-word, giving a cognate class size of two. For 108 or 8% of the possible (200 − 12) × 7 = 1,316 possible combinations of meanings and language families, no proto-word is reconstructed in the LWED. Kartvelian and Chukchi-Kamchatkan are missing 38 (19%) and 29 (15%) items, respectively, possibly because these language families now each comprise just five languages and so often no clear picture emerges of the proto-word. All these missing data were treated as noncognate, when constructing cognate class sizes, which is conservative because the missing data do not necessarily point to language unrelatedness, they simply tell us that we do not have enough information either way. In cases where none of the reconstructed proto-words was cognate with any other reconstructed forms for a particular meaning, the size of the cognate class is one (i.e., a class consisting of one language). When more than one proto-word is reconstructed for a meaning, it is possible to derive more than one cognate class size. For all such proto-words, we used the largest observed cognate class size.

Table S1 reports the largest cognate class size in the Eurasiatic language families for each meaning in the Swadesh vocabulary list. Multiple rows for a given meaning indicate that there are multiple ways of achieving the maximal cognate class size. At the phylogenetic inference step (see below), the data in Table S1 were reduced to 200 lines by calculating for every pair of languages whether there was any evidence that their words for a particular meaning were cognate.

Phylogenetic Inference. We estimated a posterior distribution of phylogenetic trees from a Markov chain Monte Carlo procedure (17) applied to the pairs of distances between languages on phylogenetic trees. The Markov chain proposes a new tree and branch-lengths each iteration of the chain, and then evaluates the likelihood of the distances that tree implies. We estimate the likelihood of a distance between a pair of languages i and j by evaluating $L_{ij} = \prod_{k=1}^{m} \sum_{i=1}^{4} \gamma_i P_{k0}, \times \prod_{k=m+1}^{n} \sum_{i=1}^{4} \gamma_i P_{k1}$ for a given t or unknown time, where $P_{k0} = (1 - e^{-r_k t})$ and $P_{k1} = (e^{-r_k t})$, m corresponds to words in the Swadesh list that we scored as not cognate between the two language families, $n - (m + 1)$ counts the words scored as cognate, r_k is the rate of change for the k^{th} word in units of lexical replacement per unit time, as estimated in the Indo-European languages (rates taken from ref. 18), and γ_i is the usual γ-rate heterogeneity (19), summed over four rate categories. The value of n is nominally 188 but can vary because some words were missing in some language families. To adjust for this, we normalized each L_{ij} by the number of words on which it was based. A given tree implies finding L_{ij} for all 21 pairs of language families, yielding an overall likelihood that is their product. This procedure thereby counts some portions of the tree more than once, so to check that this was not a source of bias we ran the procedure on a set of seven Indo-European languages. The tree matches that of Gray and Atkinson (12) and the estimated time-depths of its nodes correlate 0.99 with their tree. Running the same procedure on a set of 87 Indo-European languages yields a tree virtually indistinguishable from those reported in refs. 12 and 18 for the same 87 languages.

To estimate the Eurasiatic superfamily tree, we ran many independent Markov chains billions of iterations each. This process allowed us to derive large posterior densities of trees ($n > 40{,}000$) sampled at widely spaced intervals to ensure that successive trees in the chains were uncorrelated. The same consensus tree (Fig. 4*A*) emerged from five such independent runs. The consensus topology was also the most frequently occurring topology in the posterior sample, and 9 of the top 10 most frequent topologies placed PD and PK outside the remaining language families. The consensus topology accounted for 7.2% of all trees in the posterior sample against a random expectation of 0.11% (there are 945 possible topologies for seven taxa), yielding a Bayes Factor of 68, indicating strong support (20). The first nine topologies together accounted for 29% of all trees in the posterior sample against an expectation of 1.06%; this yields a Bayes Factor of 27.4 also indicating strong support (20).

The support at internal nodes of this tree is low, and so to gauge the consensus tree's posterior values we counted the number of times each node is present in the 945 possible topologies for these seven language families. We then compared these to the observed support values obtained from the five independent Markov chains. These values are shown in Table S2 and there was virtually no run-to-run variation in any of the posterior support values (less than 1% in all cases). The PK-PD and proto-Altaic (PA), proto–Chukchi-Kamchatkan (PCK), (PIY) groupings both have Bayes Factors exceeding five, indicating positive support (20). The low value for PCK-PIY is influenced by the exceedingly long branch to PCK, which allows it to move around without substantially altering the likelihood. Thus, deleting it increases the PA-PIY posterior to 61.

An alternative procedure for finding the random posterior distribution randomizes the cognacy data among all of the pairs of language families, but takes account of the fact that some families had missing data. This procedure returns "star" phylogenies with the "random" node supports. Importantly, this process also removes the link we observe in the actual data between a word's rate of lexical replacement and its probability of being retained as cognate, further evidence that the observed data depart from random.

Dating the Eurasiatic Tree. Our likelihood function (described above) directly estimates the lengths of the branches of the phylogenetic tree and, by implication times separating all pairs of languages. To establish a timing for the root or proto-Eurasiatic language at the base of this tree we first needed to root the tree using midpoint rooting along the branch leading to PD as an outgroup (see main text). Unlike most trees, however, this tree's tips do not represent contemporary languages but rather proto-languages that existed at varying times in the past, which means the tree "floats" somewhere along a time line potentially beginning near the present but more probably further back in time.

To establish where the tree sits along this time line means we have to have independent evidence of the age of at least two of the proto-languages. Gray and Atkinson (12) have previously estimated PIE to be ~8,700 ± 544 (SD) y old based on a set of 87 contemporary languages and 14 dated calibration points throughout the tree corresponding to historically attested texts or divergence events. The 8,700 y age has since been verified using a second dataset of 24 primarily ancient languages (21).

In addition to this estimate, we sought to provide a date for PCK because it extends furthest toward the present in our tree. To estimate this date we studied patterns of cognacy among the words of the Swadesh list for the five recognized languages of the Chukchi-Kamchatkan family—Chukchi, Koryak, Palan, Alutor, and Itelmen—using the LWED. Cognacy judgements for 166 words from the Swadesh list are available in the LWED. We inferred the phylogenetic tree of the five PCK languages using the likelihood function, as described previously. This process returned a well-supported tree describing the two recognized subgroups of this family, one containing Itelman and the other containing Chukchi and Koryak and then Palan and Alutor as sister groups. The estimated root age of these five languages is 692 ± 67 (SD) y. This is not necessarily an estimate of the age of PCK per se. For example, Fortescue (22) suggests there was a flourishing neolithic culture on the Kamchatkan peninsula around 2000 BC and that the PCK language speakers entered this region at or soon after this time. Rather, the 692 figure estimates the age of the construct that represents PCK on our tree, that PCK being based on the five languages. All this date needs to do is provide a way to calibrate the rate of evolution on our tree for the set of words we studied. As these words are held constant throughout the tree (and are the same set of words that were used to date PIE), this construct (the reconstructed PCK ancestor) is the correct one to use.

We then instituted a phylogenetic inference procedure that takes into account our uncertainty about the PIE and PCK dates. At each iteration of our Bayesian inference step, we sample a date independently and at random from PIE and PCK distributions normally distributed around their respective dates and with the SDs given above. We then repeated this procedure billions of times collecting a posterior sample of trees, each one calibrated according to the pair of random dates for PIE and PCK, and then dated at the root. The inferred age of the tree using PD as the outgroup is 14.45 ± 1.75 (SD) kya. The 95% confidence interval (CI) based on the Bayesian posterior distribution of root ages is 11.72–18.38 kya; using PK as the outgroup, it is 15.61 ± 2.29 kya, 95% CI is 11.72–20.40 (the agreement between the two lower 95% CIs is coincidental).

Rates of Borrowing. Haspelmath and Tadmor (23) describe a large-scale study of adoption or borrowing of words from a study of 1,000–2,000 meanings in 41 languages taken from a worldwide sample of 26 language families. Language families include Indo-European, Uralic, Altaic, Afro-Asiatic, Sino-Tibetan, Mayan, Austronesian, and several creoles and pidgins. The authors have compiled the results of this study in the World Loanword Database (WOLD) (23).

Each meaning in WOLD receives a borrowing score within each of the 41 languages on a five-point scale range, where 0.0 indicates "no evidence of borrowing," 0.25 denotes "very little chance of borrowing," 0.50 indicates "perhaps borrowed," 0.75 denotes "probably borrowed," and 1.0 is used for words that are "clearly borrowed." Wherever a word is suspected of being borrowed, WOLD documents the language the word was most likely borrowed or adopted from. Once the borrowing scores for each meaning are recorded, they are averaged over the 41 languages to an overall borrowing score for that meaning.

We used WOLD to assess borrowing scores for the meanings in the Swadesh Fundamental Vocabulary list. Overall, the 200 meanings have a mean borrowing score of 0.12 ± 0.07 (SD), ranging from 0.0 to a maximum of 0.46 (median = 0.1, mode = 0.1). The 23 high-frequency of use meanings included in Table 1 have a mean borrowing score of 0.07 ± 0.05 (SD), ranging from 0.0 to a maximum of 0.19 (median = 0.05, mode = 0.04).

The very low borrowing scores and high-frequency of use of Swadesh word list items makes it unlikely that the patterns of cognacy we have identified arise from adoptions of words by one set of speakers from another. For adoptions systematically to affect our results, lexical items would have to have been exchanged so frequently among the many extant languages of two or more language families as to cause them to be reconstructed as the proto-words of these families.

It might be speculated that adoptions occurred very early in the evolution of these language families, thousands of years ago before there were large numbers of different languages. These adoptions might then have been retained for thousands of years into the present. This scenario cannot be ruled out but seems unlikely, and for the same reasons as given for current adoptions: the Swadesh word list items have low adoption scores and there is no reason to postulate that the words behaved differently in the past. The structure of the topology we derive in Fig. 4*A* supports these arguments by placing language families that are geographical neighbors in distinct regions of the tree. For example, the Altaic language family includes modern day Turkish, which is surrounded by Indo-European languages, and yet PA is placed distantly to PIE. Similarly PD and especially PK are distant to PIE and PA, despite their likely central Asian origins.

Checks for Bias in Cognacy Judgements. Nine of the words in Table 1 are closed-class words of simple phonology ("thou," "I," "not," "that," "we," "who," "this," "what," "ye"), whose short length might mean that resemblances between proto-words are more likely to arise by chance. We think it is unlikely this affects our data because all 12 meanings that we excluded from our analyses (main text) because the LWED linguists could not derive proto-words for them are closed-class words of this type, showing that linguists are well aware of this potential source of bias. Removing the nine closed-class words from Table 1 does not change any of our conclusions.

Perhaps there is an expectation among linguists that frequently used words will be conserved, and this produces a bias toward identifying them as cognate between families. Our data do not support this notion. Even though the majority of frequently occurring words are conserved, there are some relatively high-frequency words (e.g., "to say," "day," and "to know," along with the number words) with cognate class sizes of two or less, and some infrequently used words are conserved (e.g., "bark," "ashes," and "worm"). In addition, the LWED proposes many more possible proto-words for the less-frequently used words (reflecting their greater variability within and among languages), and so just by chance one expects more cognate links to be found among them: but we find the opposite.

Frequently used words might have clearer or more easily defined meanings ("more straightforward glossing"): contrast "two" of something with what it means to "burn" something. If so, this could mean that infrequently used words would not be as likely to be recognized as cognate across language families, giving the trend we find but for the wrong reasons. We think this theory unlikely because linguists look for evidence of just this sort of "glossing" problem, often proposing a large number of possible proto-words for a given meaning. This situation occurs in the LWED predominantly for the infrequently used forms because they are the ones with higher rates of lexical replacement and thus they are more likely to have synonyms or near synonyms within languages and vary more among the attested languages. The large number of proposed proto-words for these meanings then explore a wide semantic space.

The LWED then checks all of the proposed proto-forms for a given meaning in a language family with all those proposed for another language family (such as Indo-European and Uralic).

This process makes it far more likely that at least one match will be found just by chance. Thus, far from biasing the data toward finding fewer cognates, infrequently used words are likely to produce a bias in the opposite direction. However, as predicted of the infrequently used words, they actually form fewer, not more cognate relationships.

Might frequently used words draw on a smaller range of distinct sounds than infrequently used words, and so be more likely to appear—just by chance—to be cognate across language families? We compared the phonemic diversity of the 10 most frequently used words from the Swadesh list with the phonemic diversity of the 10 least-used words separately for English, German, and French by counting the number of distinct phonemes among these top and bottom 10 words.

In each case we found that frequently used words are more diverse (more different from each other) per phoneme than the infrequently used ones. We attribute this finding to an evolutionary pressure in languages for frequently used words to be more different from one another in order that they will be easily distinguished in everyday speech. This means that by chance alone one is less likely to find a phoneme match at any given position of two reconstructed proto-words from different language families when they are derived from frequently used words, but the actual data show more matches.

1. Trombetti A (1905) *L'unità d'Origine del Linguaggio* (Libreria Treves, Bologna, Italy).
2. Bomhard A (1998) in *Nostratic: Sifting the Evidence*, eds Salmons J, Joseph B (John Benjamins, Amsterdam, The Netherlands), pp 17–50.
3. Bomhard A, Kerns J (1994) *The Nostratic Macrofamily. A Study in Distant Linguistic Relationship* (Mouton de Gruyter, Amsterdam, The Netherlands).
4. Illich-Svitych V (1989) in *Reconstructing Languages and Cultures*, ed Shevoroshkin V (Universitatsverlag Brockmeyer, Bochum), pp 125–176.
5. Dolgopolsky A (1986) in *Typology, Relationship and Time*, eds Shevoroshkin V, Markey T (Karoma, Ann Arbor, MI), pp 27–50.
6. Greenberg J (1998) in *Nostratic: Sifting the Evidence*, eds Salmons J, Joseph B (John Benjamins, Amsterdam, The Netherlands), pp 51–60.
7. Greenberg J (2000) *The Eurasiatic Language Family: Indo-European and Its Closest Relations. Volume I: Grammar* (Stanford Univ Press, Stanford, CA).
8. Ruhlen M (1994) *On the Origin of Languages: Studies in Linguistic Taxonomy* (Stanford Univ Press, Stanford, CA).
9. Lewis M (2009) *Ethnologue: Languages of the World* (SIL Institute, Dallas, TX).
10. Ruhlen M (1987) *A Guide to the World's Languages, Volume 1: Classification* (Stanford Univ Press, Stanford, CA).
11. Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature* 461(7263):489–494.
12. Gray RD, Atkinson QD (2003) Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* 426(6965):435–439.
13. Renfrew C (1987) *Archaeology and Language: The Puzzle of Indo-European Origins* (Pimlico, London).
14. Janhunen J (2009) Proto-Uralic–what, where, and when? *Suomalais-Ugrilaisen Seuran Toimituksia* 258:57–78.
15. Pokorny J (1959) *Indogermanisches Etymologisches Wörterbuch* (Francke, Bern).
16. Janhunen J (2007) The primary laryngeal in Uralic and beyond. *Mém de la Soc Finno-Ougrienne* 253:203–227.
17. Gilks WR, Richardson S, Spiegelhalter DJ (1996) in *Markov Chain Monte Carlo in Practice*, eds Gilks WR, Richardson S, Spiegelhalter DJ (Chapman and Hall, London), pp 1–19.
18. Pagel M, Atkinson QD, Meade A (2007) Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449(7163):717–720.
19. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J Mol Evol* 39(3):306–314.
20. Raftery AE (1996) in *Markov Chain Monte Carlo in Practice*, eds Gilks WR, Richardson S, Spiegelhalter DJ (Chapman and Hall, London), pp 163–188.
21. Atkinson QD, Nicholls G, Welch D, Gray RD (2005) From words to dates: Water into wine, mathemagic or phylogenetic inference? *Trans. Phil. Soc* 103(2):193–219.
22. Fortescue M (2003) Diachronic typology and the genealogical unity of Chukotko-Kamchatkan. *Ling Typology* 7:51–88.
23. Haspelmath M, Tadmor U (2009) *Loanwords in the World's Languages: A Comparative Handbook* (Mouton de Gruyter, Amsterdam, The Netherlands).

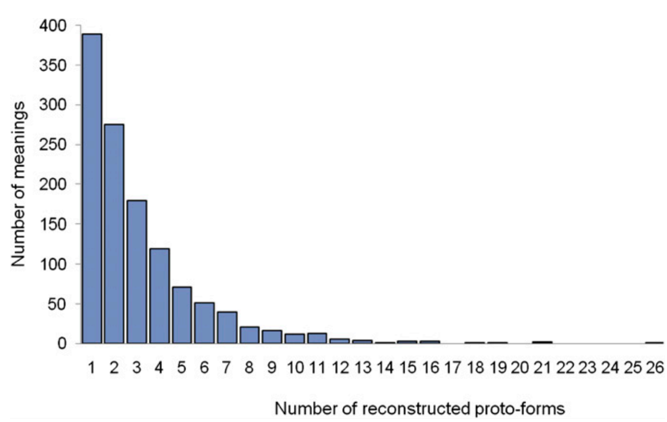

Fig. S1. The distribution of proto-word reconstructions per meaning for the 188 vocabulary items (meanings) in the 200-word Swadesh list (12 meanings were excluded for lack of data). Total number of proto-words is 3,804. Mean number per meaning is 2.89 ± 2.81 (SD) (median = 2, mode = 1).

Other Supporting Information Files

Table S1 (DOCX)
Table S2 (DOCX)