Revisiting Unchanged Cognates as a Criterion in Linguistic Subgrouping

Sudheer Kolachina and Taraka Rama

ICHL 2011

Preliminaries

- One of the main aims of Historical linguistics
  - Classification of languages into *language families*
- *Subgrouping*
  - internal classification of languages within a language family
  - Branching structure of the *family tree*
  - How daughter languages within a single family are related to one another?

Preliminaries

- Subgrouping
  - The only generally accepted criterion for subgrouping is *shared innovation*.
- *Shared innovation*
  - a linguistic change which shows a departure for some trait of the proto-language and is shared by a subset of the daughter languages (Campbell, 2004)

Preliminaries

- *Not all shared innovations are useful for establishing subgroups*
  - Naturalness of change
  - Very natural changes => Parallel development
- *Shared retentions*
  - unchanged inheritance in daughter languages from the proto-language regardless of whether the daughter languages belong to the same subgroup or not
  - Of no value to subgrouping

Preliminaries

- A closer look at the nature of (sound) Change
- How is sound change implemented ?
  - Neogrammarian answer: regular sound change, analogy and borrowing
  - Regularity hypothesis: Sound change is regular and affect all items qualified for change at once
  - Cases of irregular change: result of analogy or dialect borrowing
- Other possible answers? Yes...

Preliminaries

- *Lexical diffusion* hypothesis (Wang 1969)
  - Gradualness of (sound) change
  - Sound change affects the sound in certain words and then gradually diffuses to other words in the lexicon
  - When change diffuses across the lexicon to reach all words, it becomes a regular change
- Controversial in Historical linguistics

Preliminaries

- Shared innovations alone as a criterion for subgrouping
  - Implicit assumption: Neogrammarian regularity of change
- But, what if we take the lexical diffusionist perspective?
  - Do there exist sources of information about subgrouping other than shared innovations?
- Above question addressed in previous work

Previous Work

- *'Unchanged Cognates as a Criterion in Linguistic Subgrouping'* , Bh. Krishnamurti, Lincoln Moses, Douglas G. Danforth, Language, Vol. 59, No. 3 (Sep., 1983), pp. 541-568
- *'Areal and Lexical Diffusion of Sound Change: Evidence from Dravidian'* , Bh. Khrishnamurti, Language, Vol. 54, No. 1 (Mar., 1978), pp. 1-20

Previous Work

- Krishnamurti et al. (1983)
  - Within the framework of lexical diffusion, can unchanged cognates serve as a source of information about linguistic subrelations?
  - One of the early works to incorporate 'computational thinking' into Historical Linguistics
  - Tree-scoring based on a set of postulates that sound similar to Maximum parsimony
  - Excerpts from the abstract (Source: Krishnamurti et al., 1983)

Krishnamurti et al. (1983)

- **If a sound change has lexically diffused without completing its course, one finds that among the lexical items qualified for the change, some have already changed (c), others have remained unchanged (u), and still others show variant forms (u|c). When such a change has affected a group of genetically related languages, the consequent comparative pattern u-ulc-c can be used to set up subrelations among languages**. In this paper, we draw on data from six languages belonging to the South-Central subfamily of Dravidian, with reference to an atypical sound change called 'apical displacement'. There are 63 etymologies which qualify for the study. A total of 945 possible binary-labeled trees fall into six types for the six languages under study. In terms of our postulates, **that tree is the best which scores the lowest m, i.e. the minimum number of independent instances of change needed to account for the u-c-o (o = no cognate) pattern of a given entry**. Each of the 63 entries has been applied to the possible 945 trees, and the trees have been scored for the value m by computer. **The one tree which scored the lowest (71 points) is identical with the traditionally established tree for these languages**. This paper shows that: (a) one shared innovation is sufficient to give genetic subrelations among languages,within the framework the theory of lexical diffusion; (b) **unchanged cognates are as important as changed cognates in giving differential scores for possible trees**; and (c) the notion of shared innovation can be further refined within the theory of lexical diffusion.

Previous Work

- Krishnamurti (1978)
  - An earlier work on which Krishnamurti et al. (1983) build
  - Provides quantitative evidence in support of areal and lexical diffusion from Dravidian
  - Same sound change: 'Apical displacement'
  - Gradual lexical spread of this change can be observed from the percentage of changed items out of the total items qualified for change

Krishnamurti (1978)

- Excerpts from the abstract (Source: Krishnamurti, 1978)
  - "**...of the items which fulfill the structural conditions of the change, 72% are covered by it in Kui, about 63% in Kuvi, Pengo, and Manda; but only about 20% in Gondi and Konda**"
  - "**A chronological layering of lexical items is established in terms of particular combinations of languages which share the cognates-with-change**"
- Dataset containing numbers of shared cognates-with-change (Table 8)
- U-statistical hierarchical clustering (D' Andrade, 1978) applied to this dataset and results discussed in Krishnamurti et al. 1983

Our work

- Our work in this presentation
  - A critique of these previous works in the light of recent advances in computational historical linguistics
  - Application of well-known methods for phylogenetic inference to the datasets used in these previous works
  - Main focus on Krishnamurti et al. 's (1983) claim about usefulness of unchanged cognates for inferring subgrouping relations

Our work

- Two different kinds of datasets used the previous works
  - Data about numbers of shared cognates-with-change from the 1978 paper (dataset 1)
  - Data about changed/ unchanged status of 63 etymologies from the 1983 paper (dataset 2)
- Two different kinds of inference methods applied to these two datasets (D' Andrade's clustering versus Krishnamurti's MP-like postulates)
- Results of both methods claimed to be in agreement with the standard tree
- Therefore, subrelations inferred from numbers of shared innovations are also recoverable from the changed / unchanged cognate lexical diffusion data and hence, the importance of unchanged cognates as a criterion for subgrouping

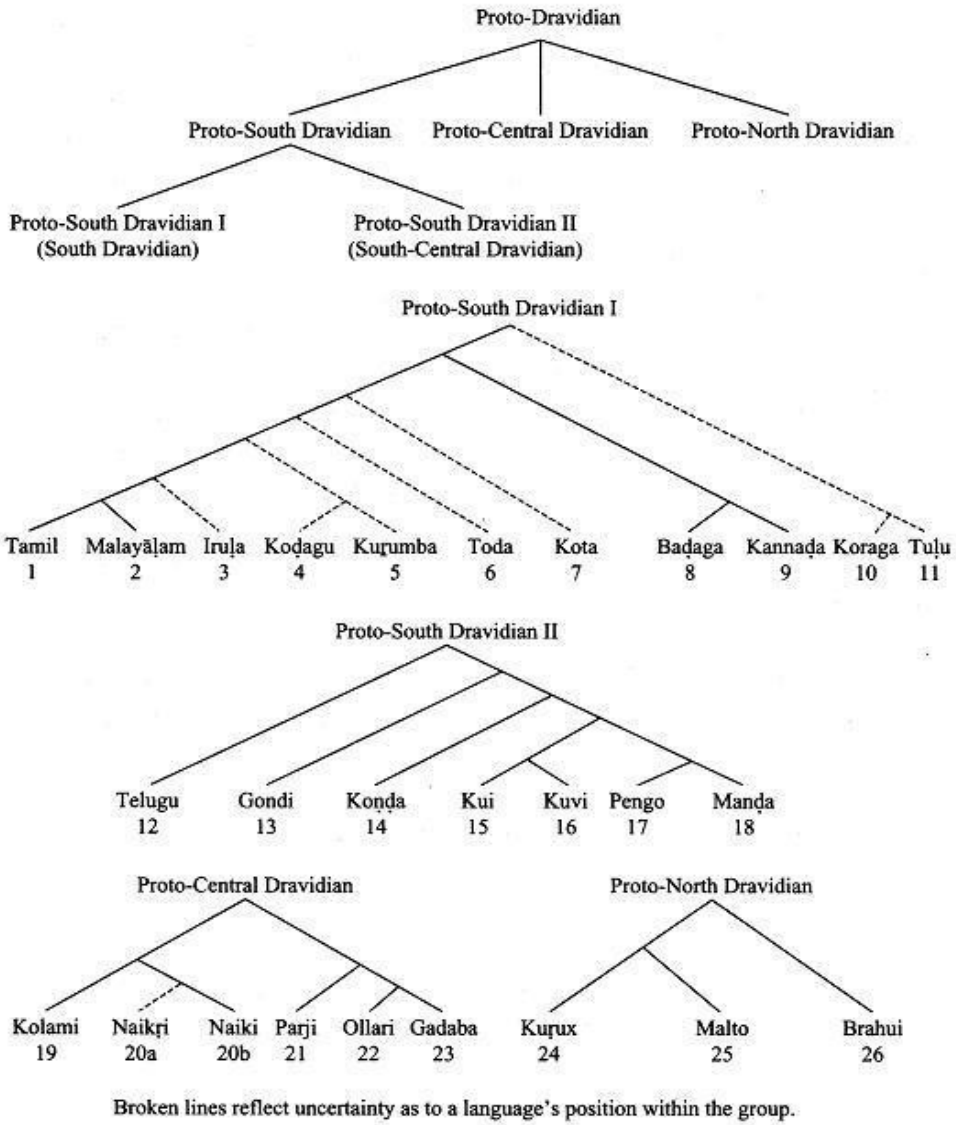
Our work

- Problem with this conclusion
  - The method applied are fundamentally different. D' Andrade's clustering algorithm is a distance-based method while the MP-like postulates resemble character-based methods
- Our proposal
  - Test these claims by applying same methods to both the datasets
  - Step 1: Transform the character-like lexical diffusion data (dataset 2) into a distance matrix
  - Step 2: Apply different well-known distance-based methods (Fitch-Margoliash, Minimum Evolution, UPGMA, NJ) to both the datasets and check if there is agreement
- Before we go into the details of our experiments, some trivia about Dravidian languages

Dravidian Languages

- 26 languages spoken by over 200 million people in South Asia making it the world's fifth largest language family (Krishnamurti, 2003)
- Most of them geographically located in the southern and the central parts of the India with a few scattered pockets in Northern India (Kurux, Malto) and Nepal (Kurux) and a lone population in Pakistan (Brahui)
- Latest family tree (Source: Krishnamurti, 2003)

Dravidian Language Family Tree



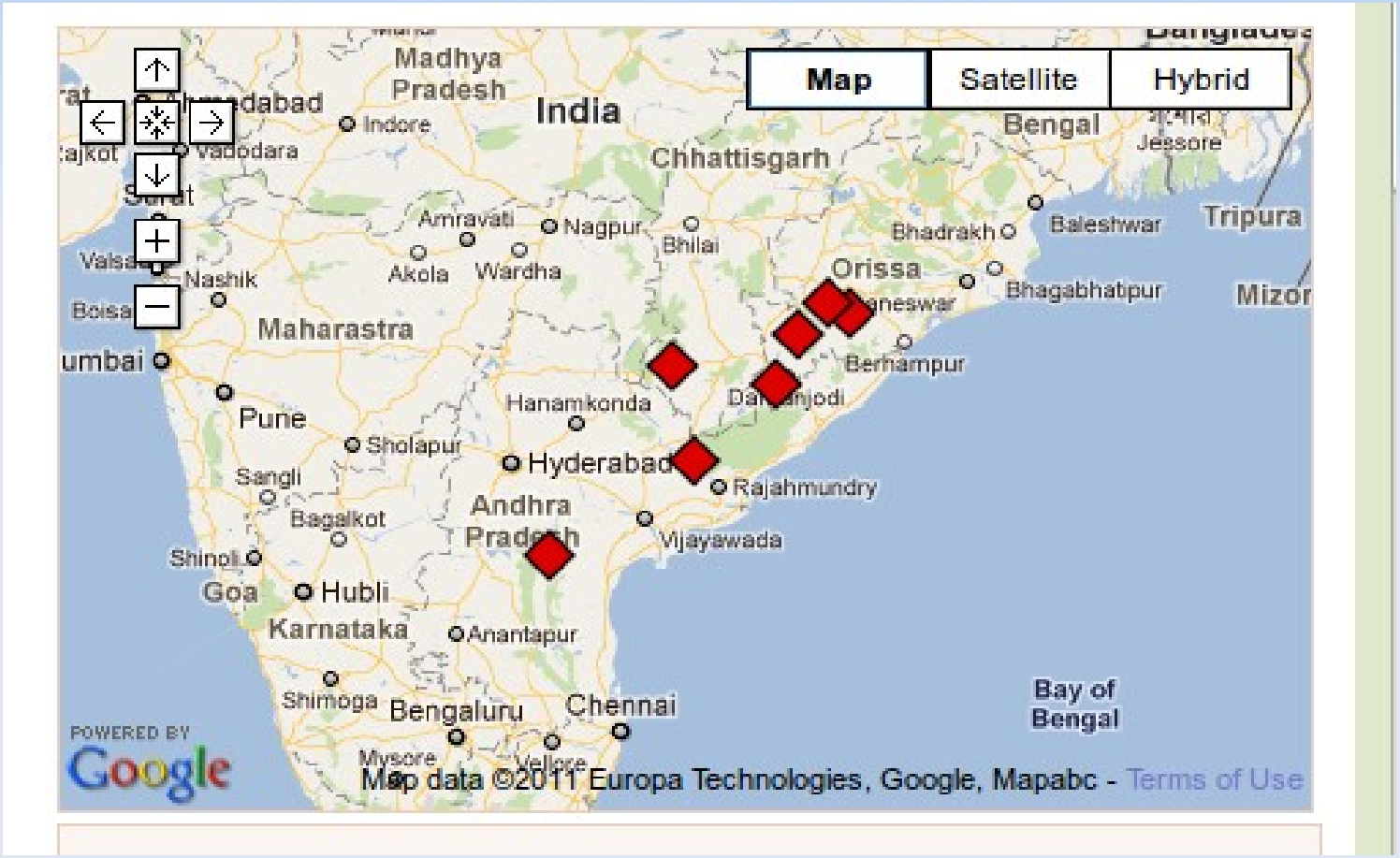
Broken lines reflect uncertainty as to a language's position within the group.

South Dravidian II

- South Dravidian II subfamily
  - South-Central Dravidian in an earlier classification
- Telugu, Gondi, Konda, Kui, Kuvi, Pengo and Manda
- Telugu, the lone literary language, excluded from this study due to the relative certainty of its position within the subgroup

SD II: Geographical Distribution

- WALS Interactive Reference Tool



Datasets

- Two datasets for six South Dravidian II (formerly South Central Dravidian) languages
- Dataset 1: Matrix containing pairwise number of shared cognates with change (Source: Krishnamurti, 1978)
  - Change – 'apical displacement'
  - 'Shared innovation' dataset
  - Information only about shared innovations
  - Number of common shared innovations between two languages– measure of their 'proximity'

Shared Cognates-with-change dataset

- Dataset 1

Gondi					
Konda	16				
Kui	18	18			
Kuvi	22	20	88		
Pengo	11	19	48	49	
Manda	10	9	40	42	57

Dataset I

- Each entry in the matrix represents a *proximity* value between two languages, which is, inverse of distance
- What about application of D' Andrade's clustering algorithm to this dataset ?
- Can distance-based methods be applied directly to this dataset?
- How would results vary if we transform the proximity values into distances?
- How to transform this data into distance? (s.t. distance is a value between 0 and 1)

Distance transformation: Dataset I

- Number of items qualified for apical displacement in each language
  - Gondi – 211, Konda – 178, Kui – 169, Kuvi – 137, Pengo – 97 and Manda – 110
- Normalize the pairwise value in the matrix
  - By the average (A.M.) number of items qualified for change (denominator = (n1 + n2)/2)
  - By the minimum of the number of items qualified for change (denominator = min(n1,n2))
- Other possible normalizations: use numbers of *items with change* rather than *qualified for change* ??

Dataset II

- Dataset II
  - Data about changed (c), unchanged (u) status of 63 etyma qualified for apical displacement from the same six languages (Source: Krishnamurti et al. (1983))
- Lexical diffusion dataset
- Information about both shared innovations (c) and retentions (u) (and non-occurrence (o))
- Information about the u-o-c distribution of apical displacement in these six genetically related languages

Dataset II

- Krishnamurti's MP-like postulates infer subgrouping relations from this distribution
- Lexical diffusion data resembles character-based data
  - Innovation (c) coded as 1
  - Retention (u) coded as 0
  - non-occurrence (o) coded as ?
- Transforming character-based data into distance-based data
  - Discussed in previous work on Linguistic phylogeny (Nakhleh et al., 2005)

Character to distance transformation: Dataset II

- Distance between two languages estimated as Hamming distance between character sequences
  - Hamming Distance: the number of sites at which two sequences differ
- Ambiguous states ignored
  - ? treated as ambiguous state
- Distance normalized by length of Hamming sequence

Distance Datasets: Summary

- Derivatives from Dataset I
  - Datasets I, IA, IB
  - I – raw numbers of pairwise shared cognates-with-change
  - IA and IB – normalized values
  - All three contain only information about shared innovations
- Derivatives from Dataset II
  - Dataset IIA – lexical diffusion u-o-c data converted to distance matrix
  - Information about both shared innovations and retentions

Aim of our experiment

- To verify if subrelations inferred from datasets I, IA and IB match with those inferred from IIA by the same phylogenetic inference methods
- If Yes,
  - Subgrouping information is recoverable from distances based on distribution of change
  - Unchanged cognates => useful information for subgrouping
  - Caution: Distribution of changed and unchanged cognates (Not unchanged cognates alone!!)

Distance-based methods

- Distance-based phylogenetic inference methods considered in our study
  - Fitch-Margoliash
  - Minimum Evolution
  - Neighbor Joining
  - UPGMA

What do they do?

- Fitch-Margoliash
  - Tries to find the tree with least squares branch length
- Minimum Evolution
  - Fits the tree's branch lengths using Fitch-Margoliash criterion
  - Searches for a tree topology by minimizing the branch lengths

Distance-based methods

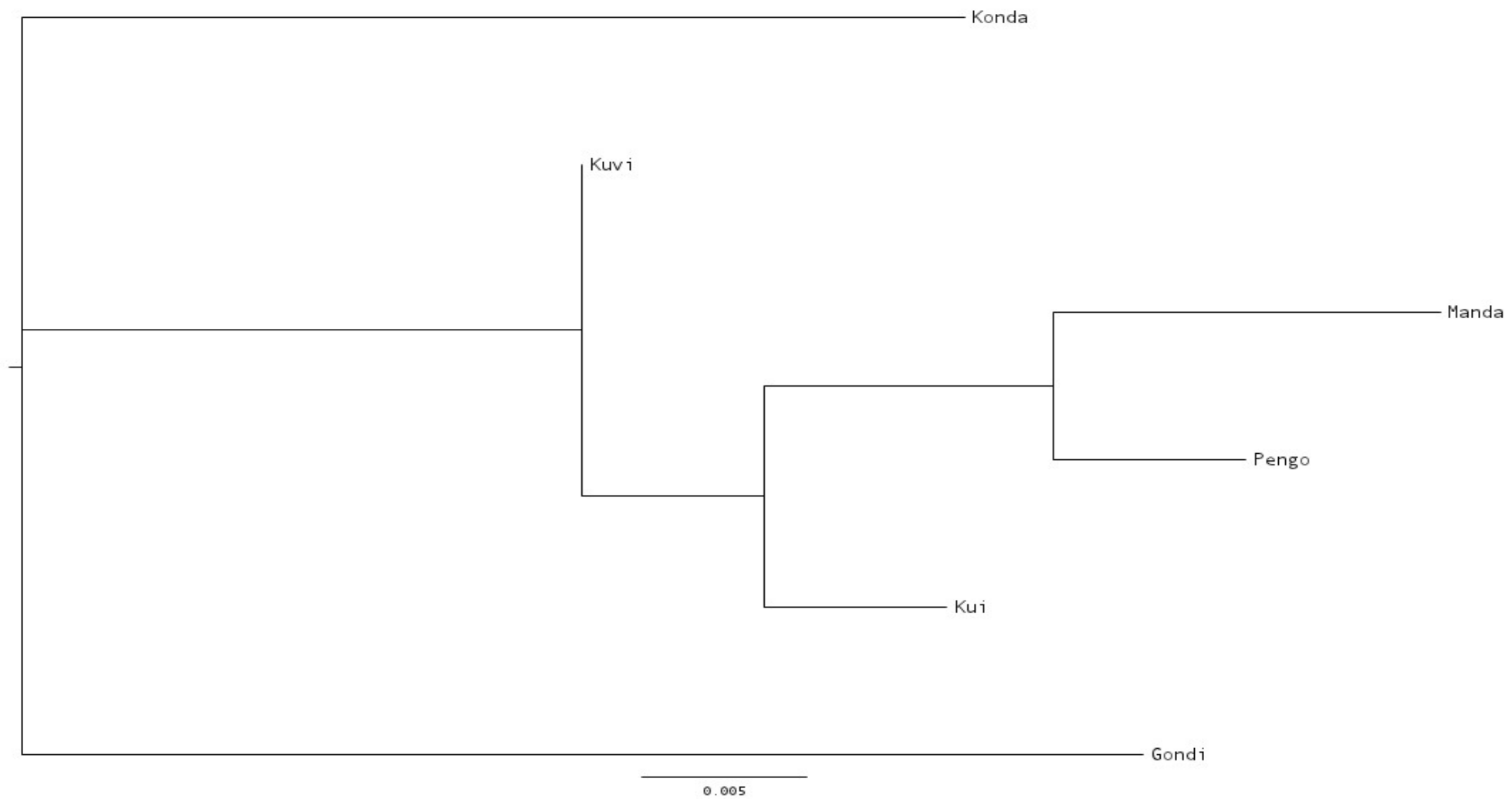
- UPGMA
  - A hierarchical algorithm and assumes clock-like evolution
  - Usually performs the worst (Nakhleh et al, 2005)
- Neighbor Joining
  - A greedy algorithm
  - Tries to minimize an estimate of the total branch length of the tree at each step

Experiments

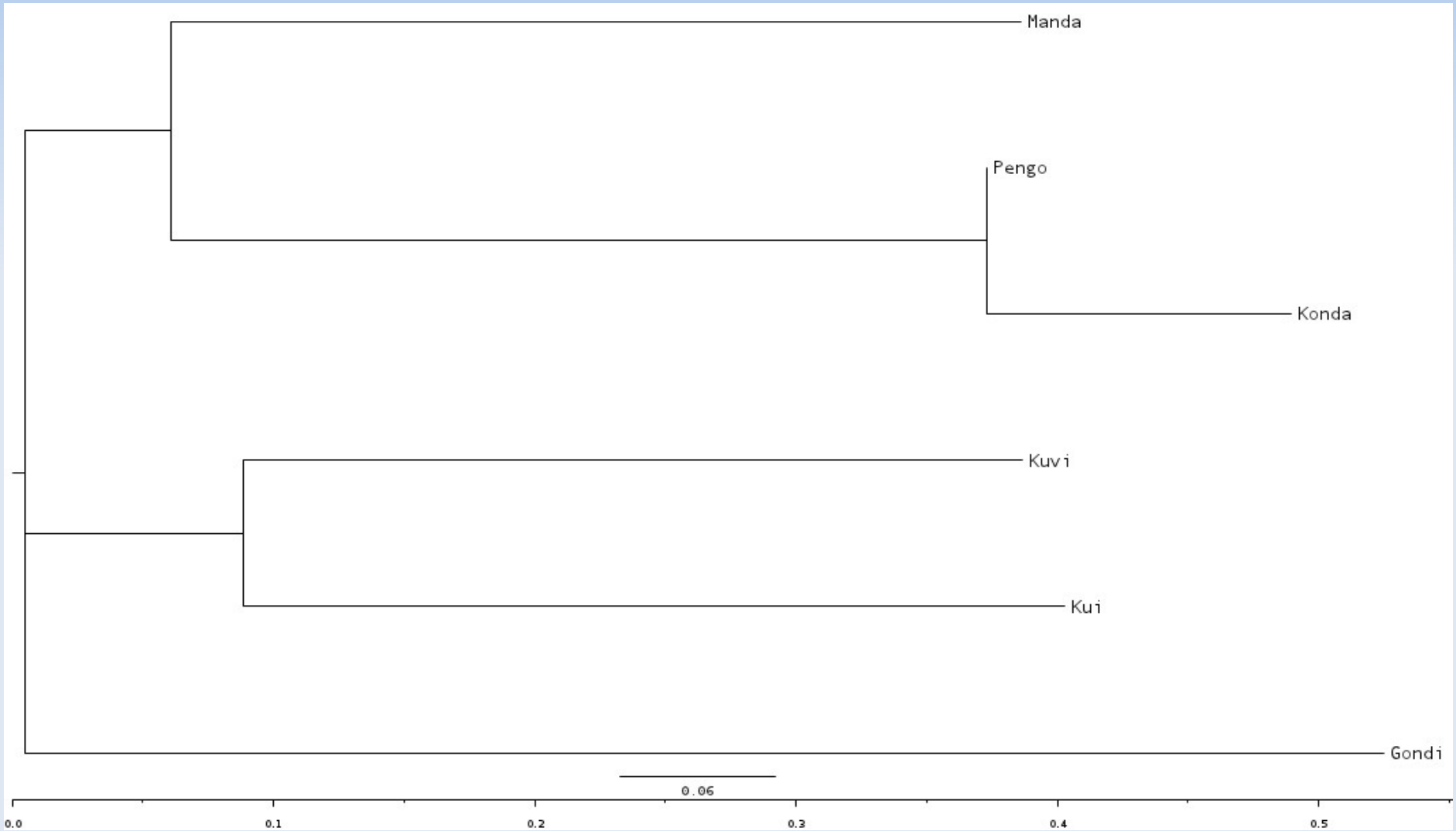
- Applied each of these methods to 4 datasets
  - I, IA and IB – datasets derived from the number of shared cognates-with-changes (innovations)
  - II A – distance matrix derived from the lexical diffusion data containing u-o-c distribution of apical displacement
- Implementations of all methods in PHYLIP (Felsenstein, 2003)

Results on dataset I derivatives (Shared Innovation datasets)

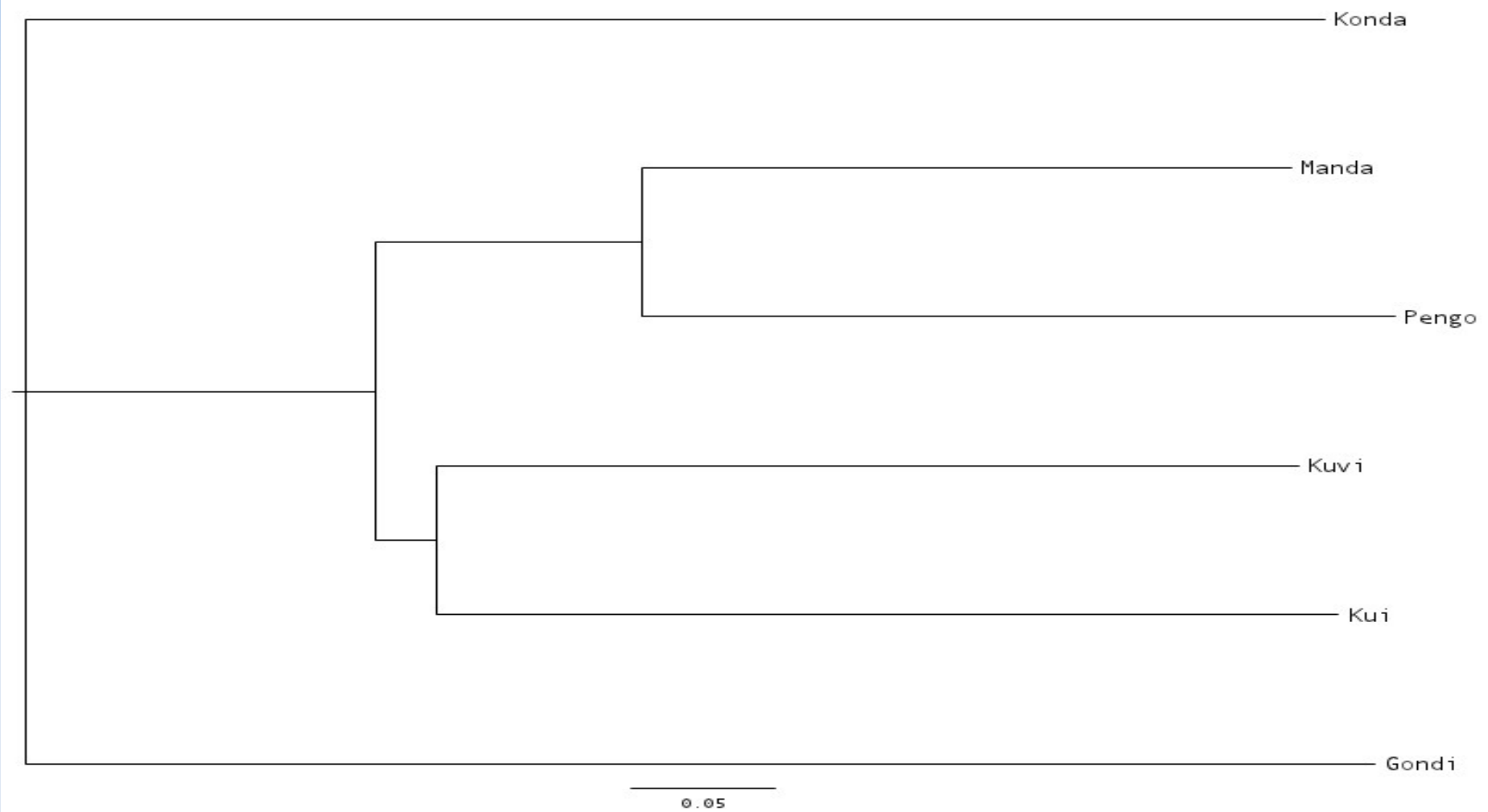
Fitsch-Margoliash (FM) 1



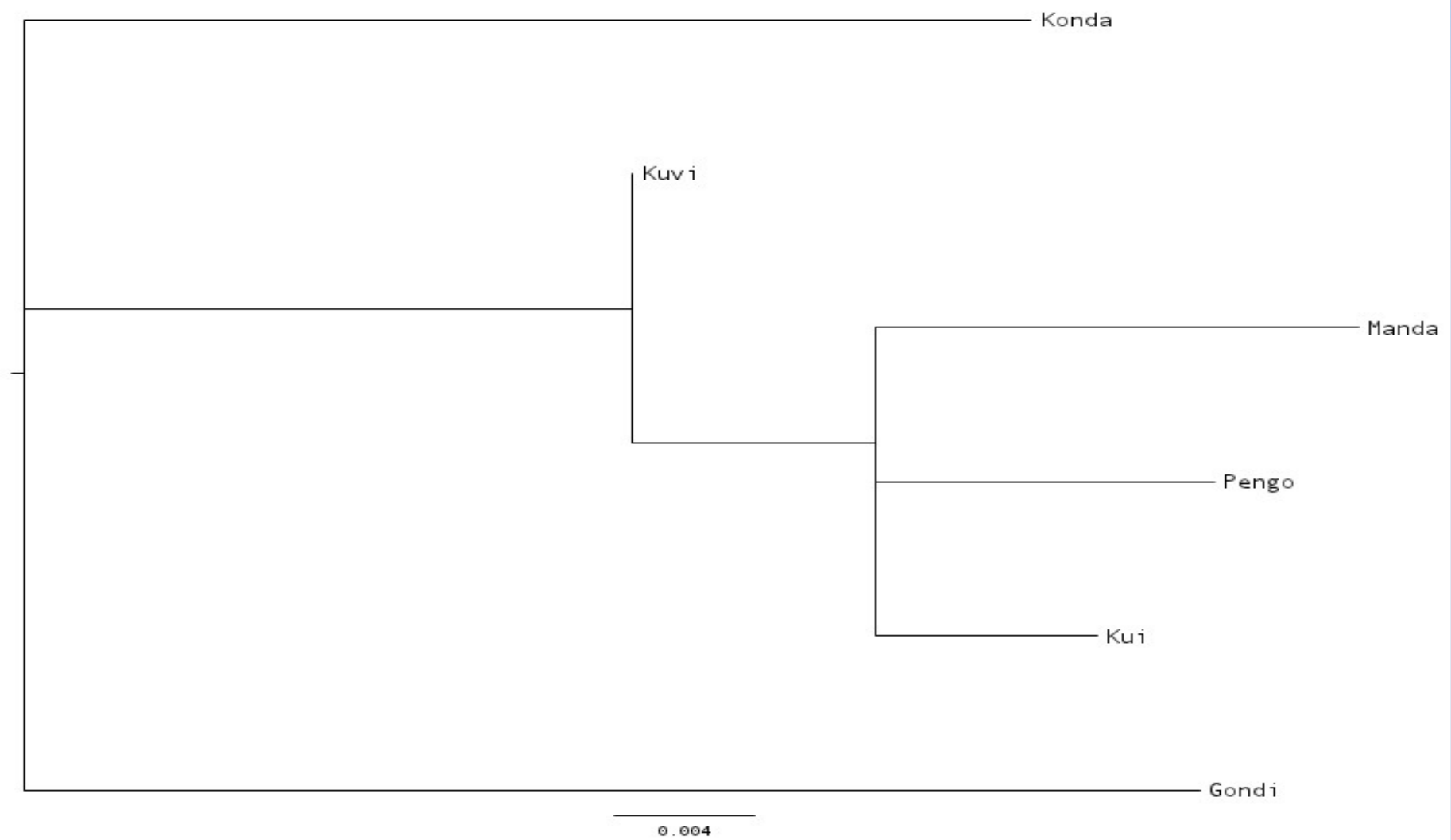
FM 1A



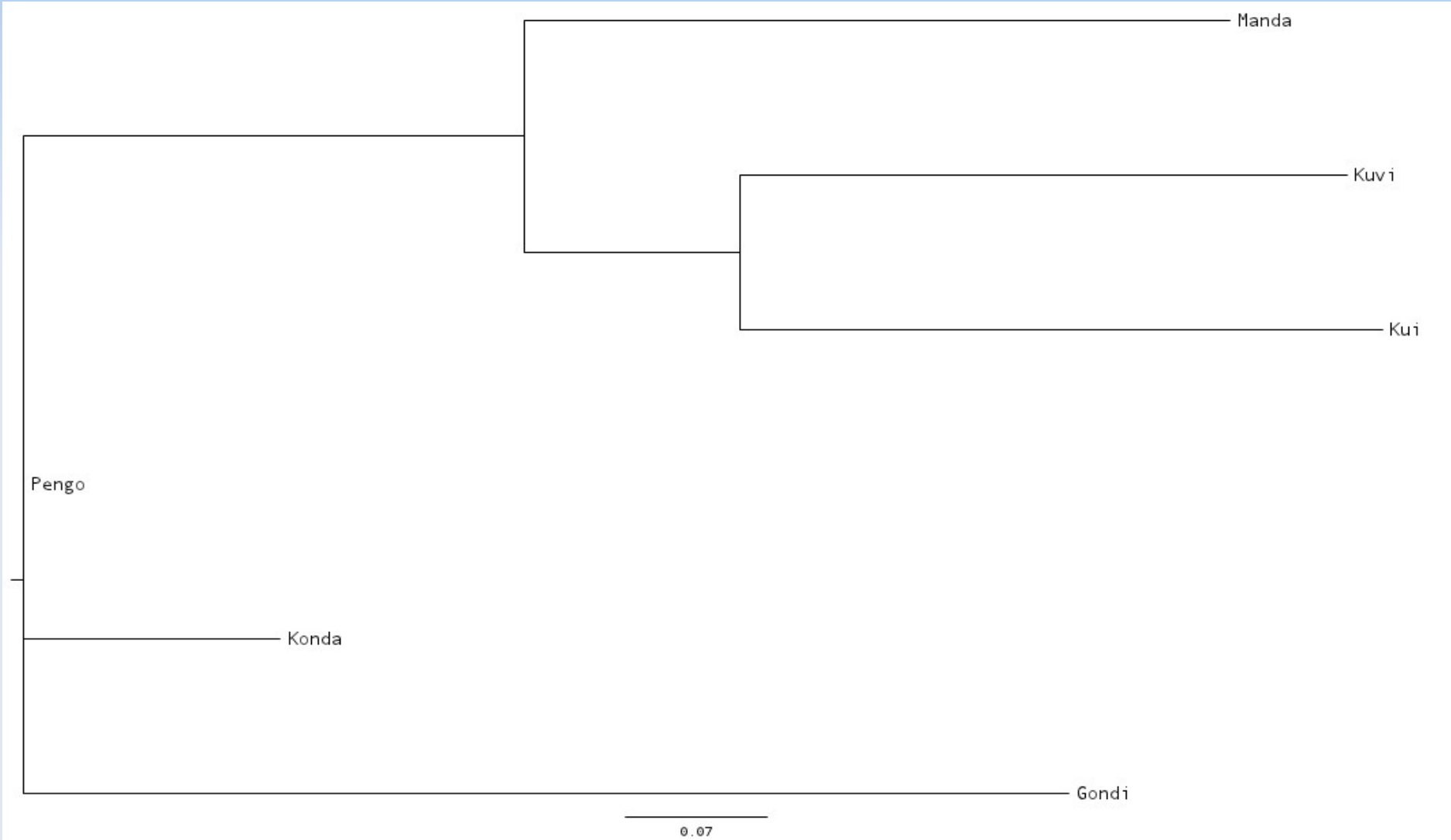
FM 1B



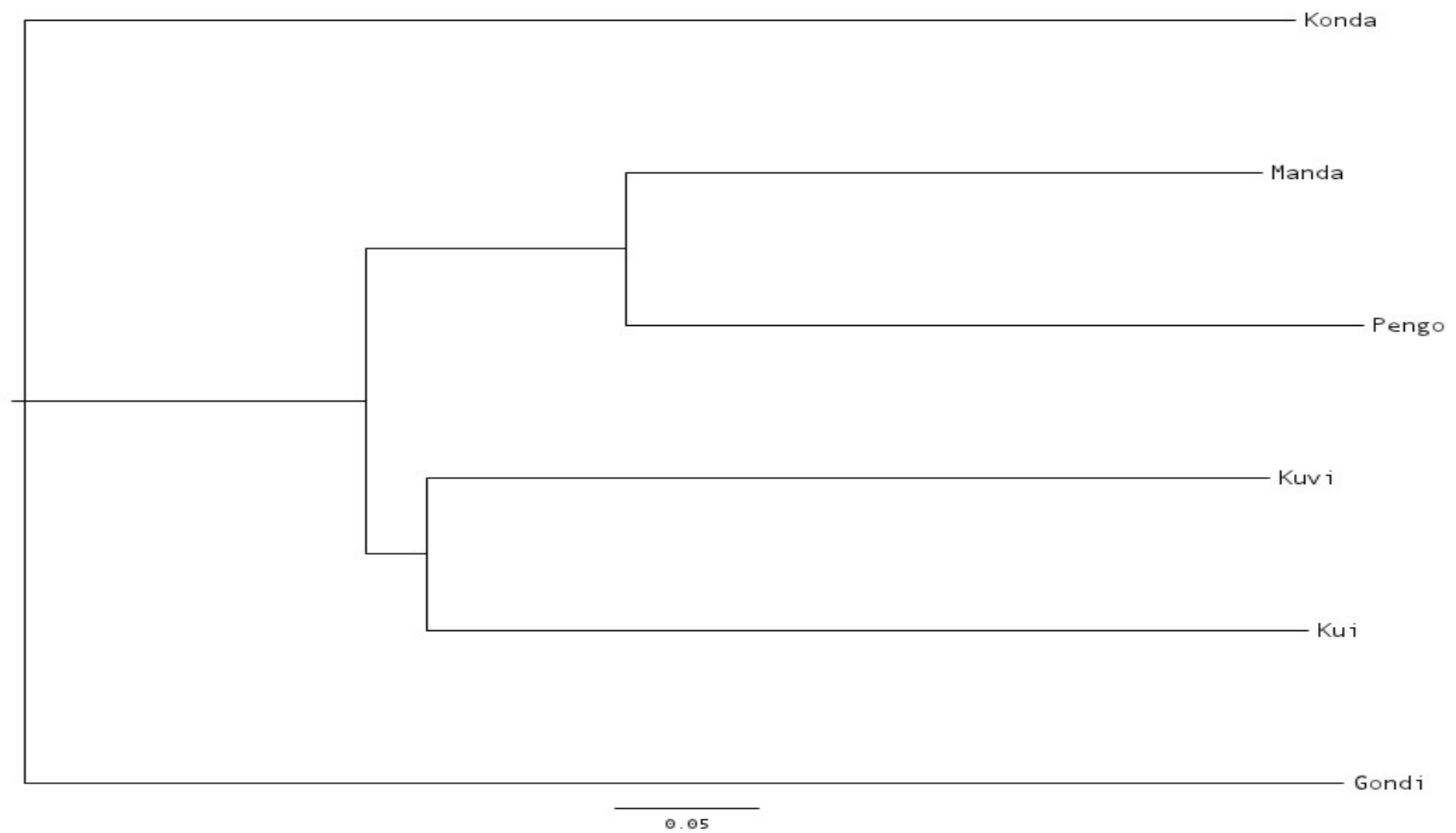
Minimum Evolution (ME) I



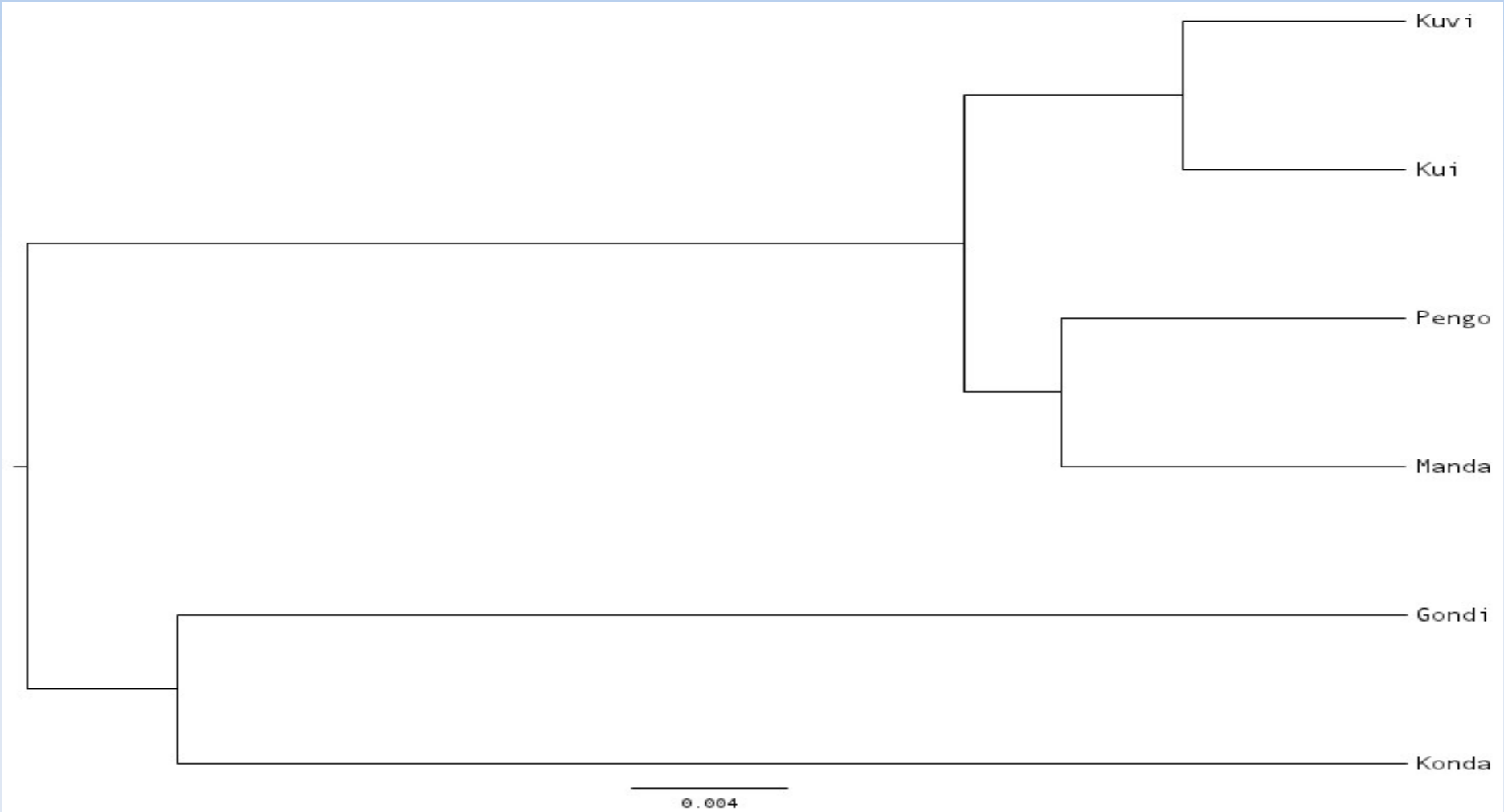
ME IA



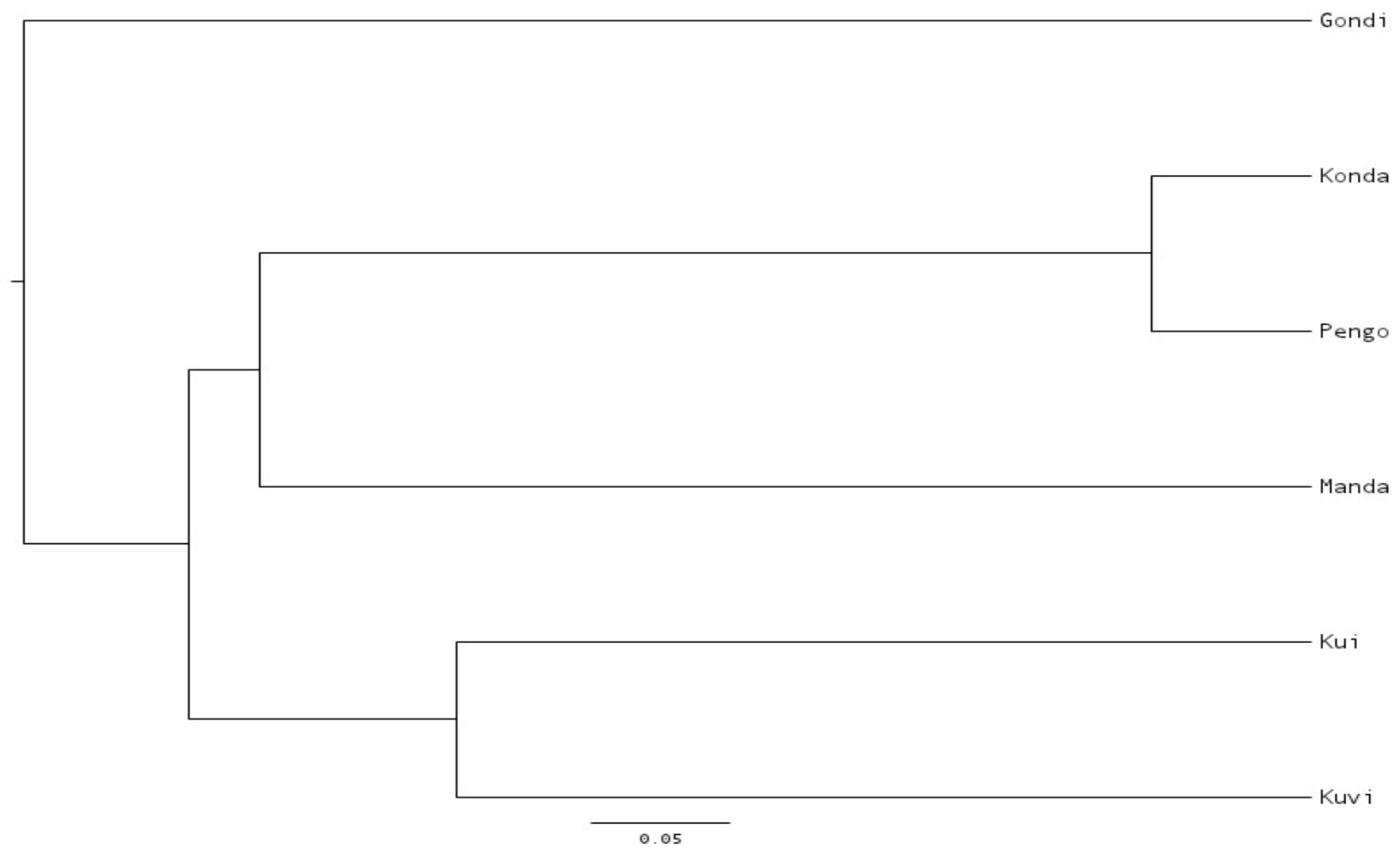
ME IB



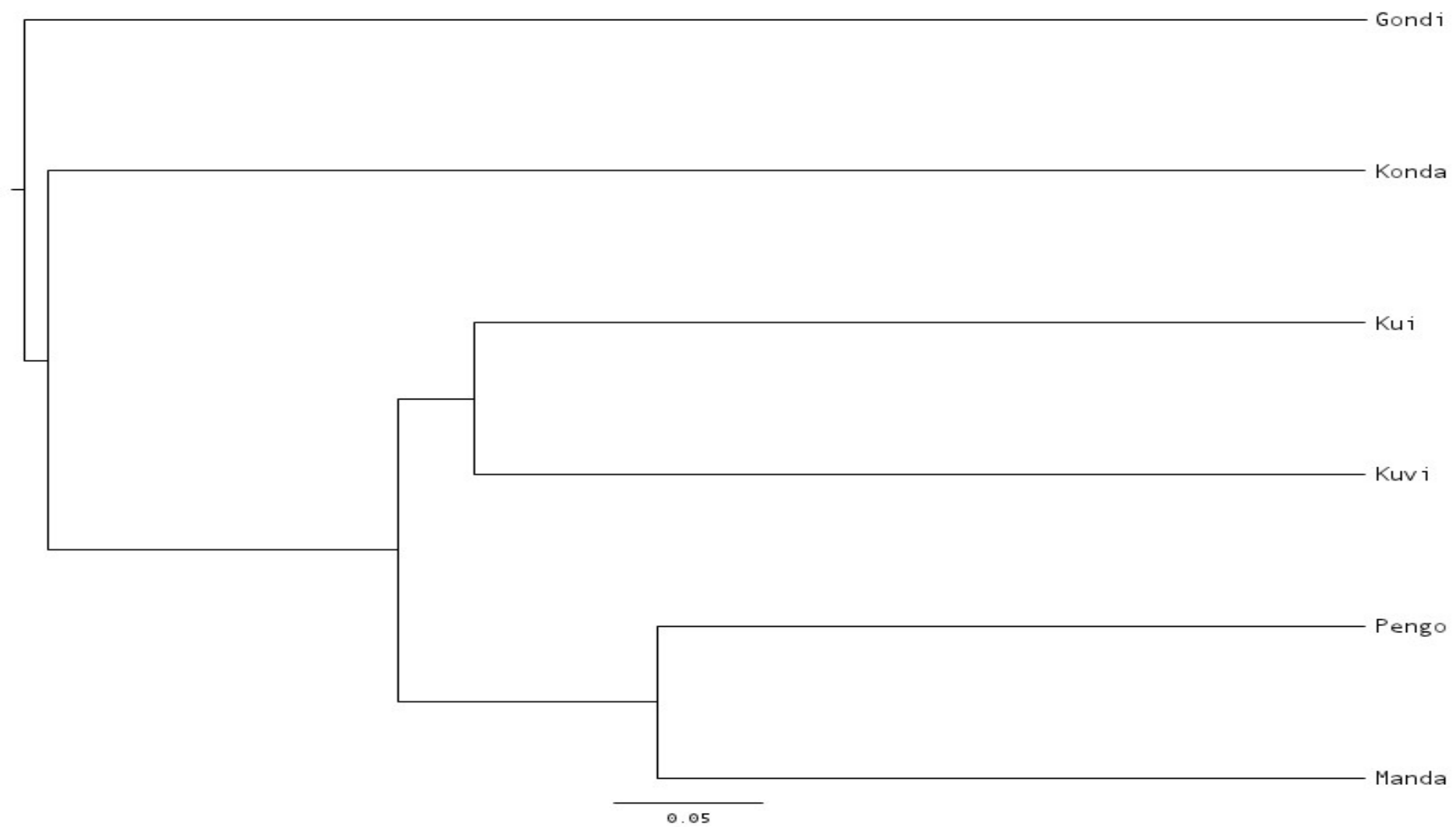
UPGMA I



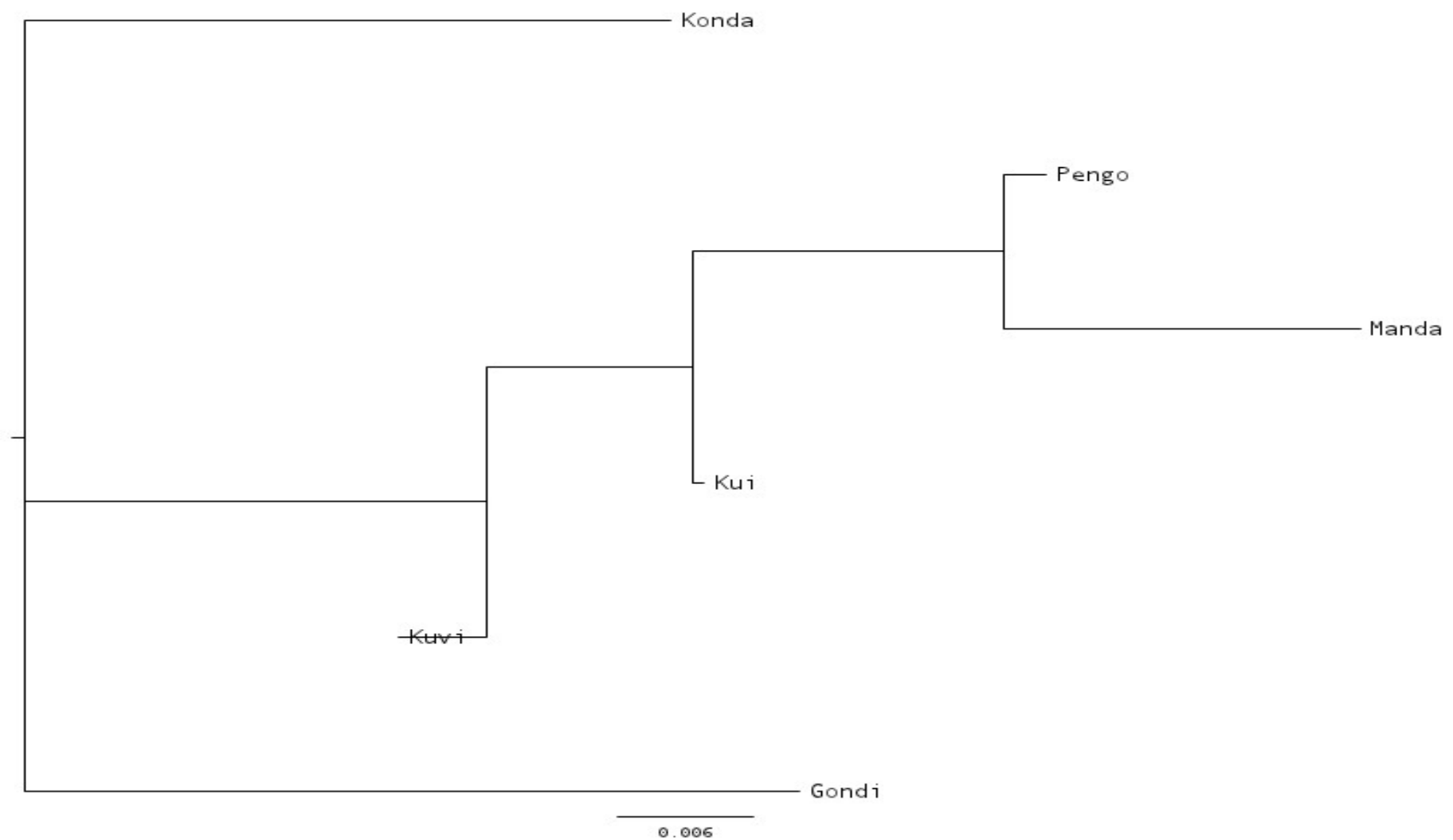
UPGMA IA



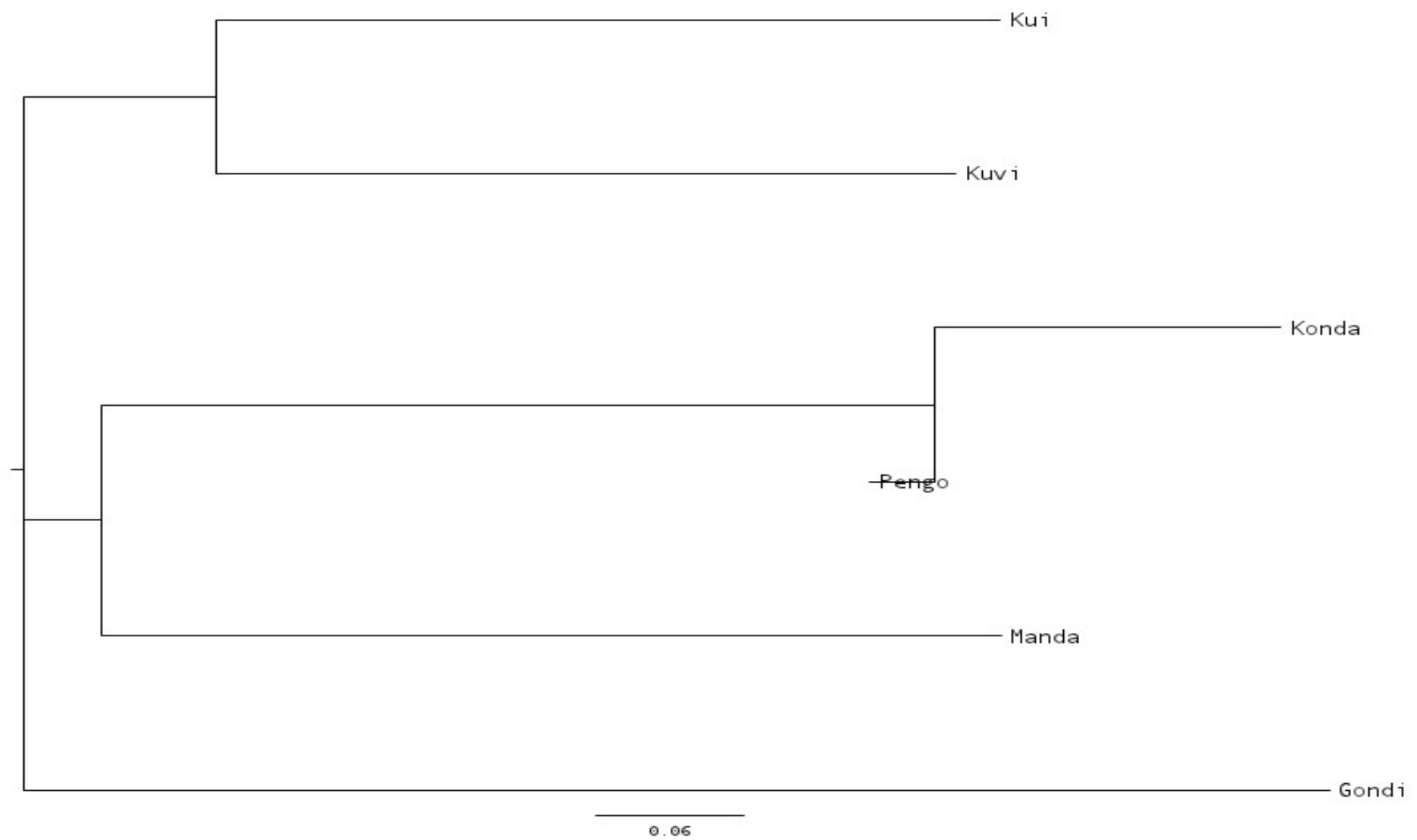
UPGMA IB



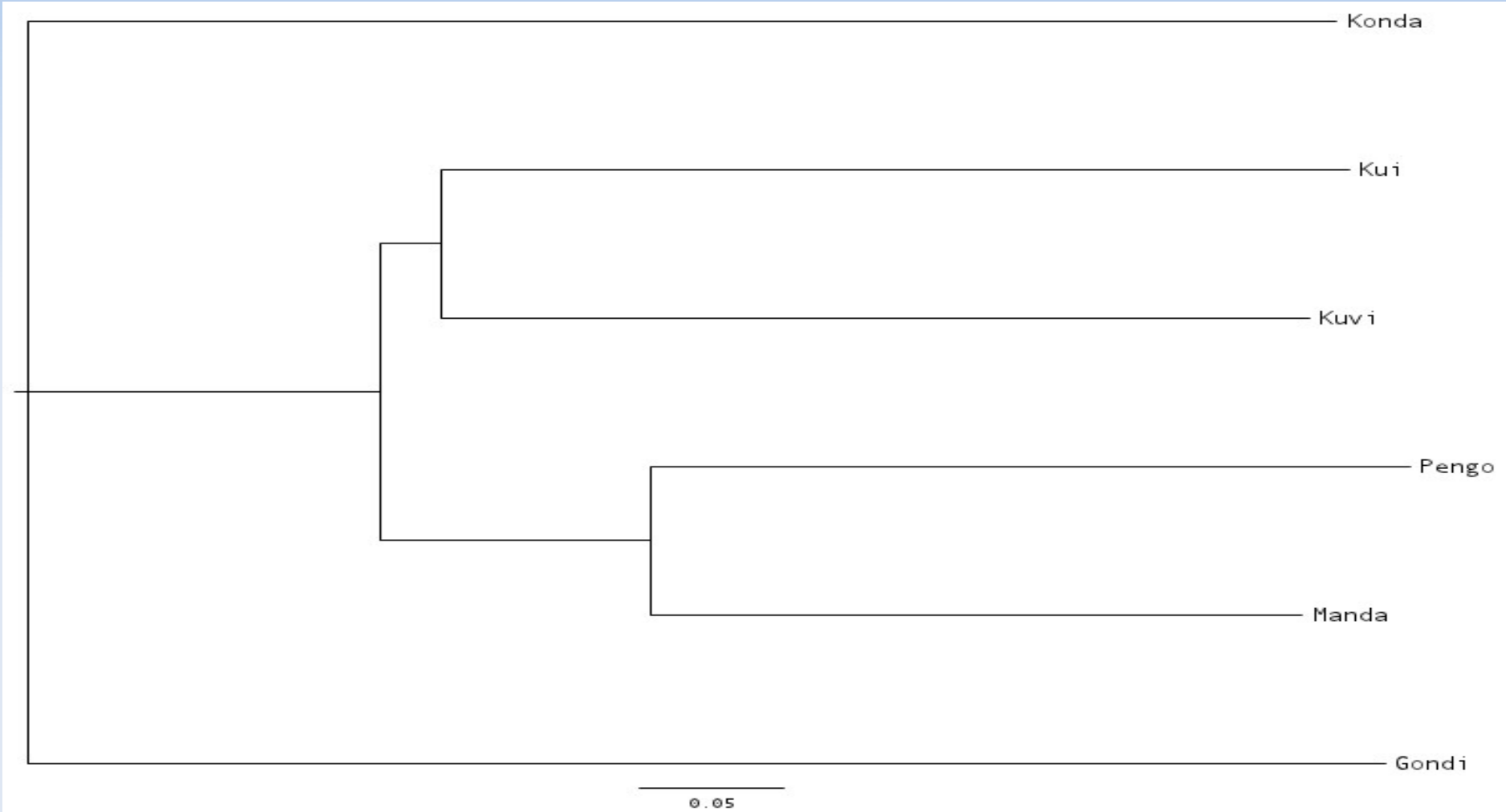
Neighbor Joining (NJ) I



NJ IA

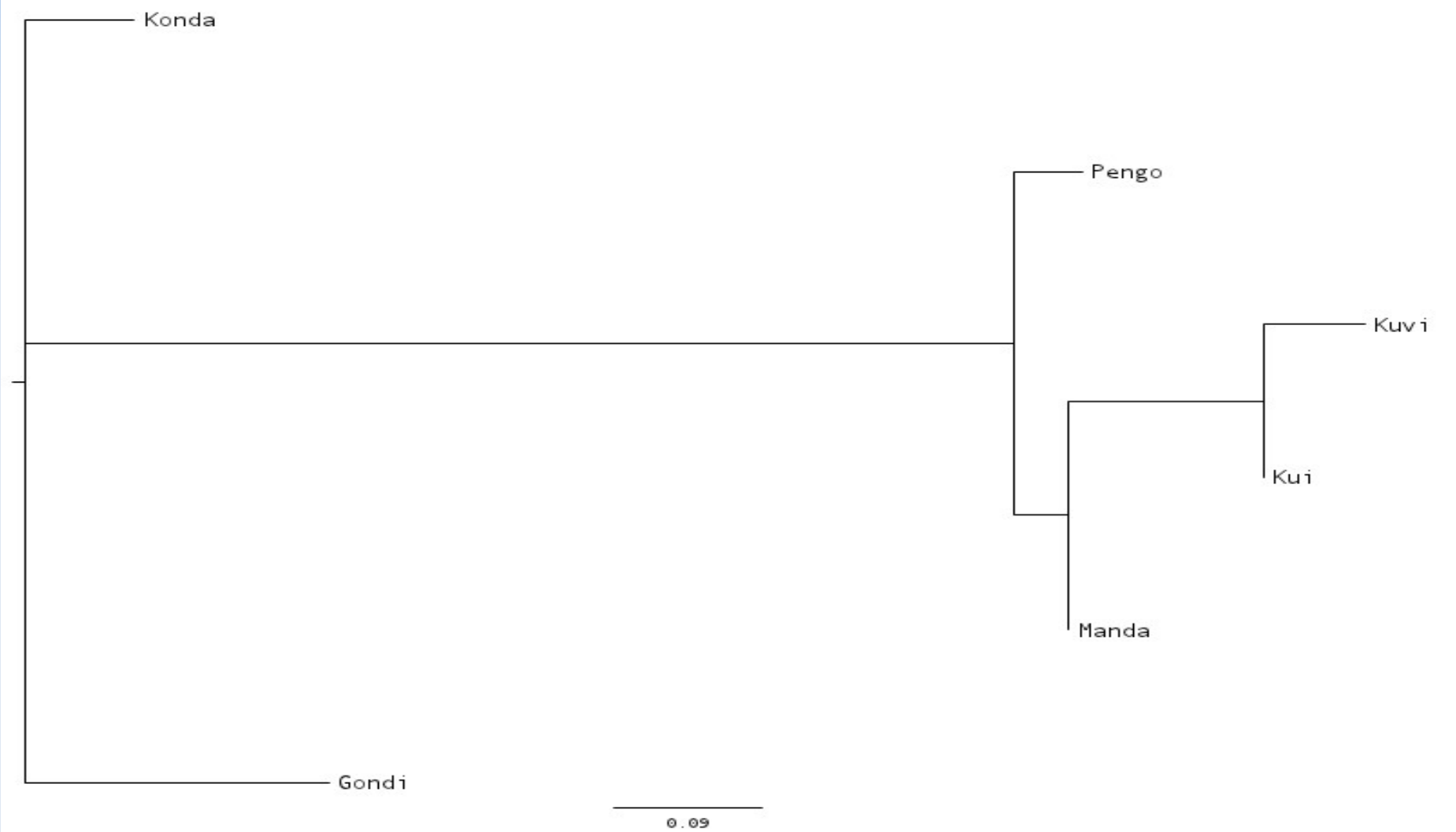


NJ 1B

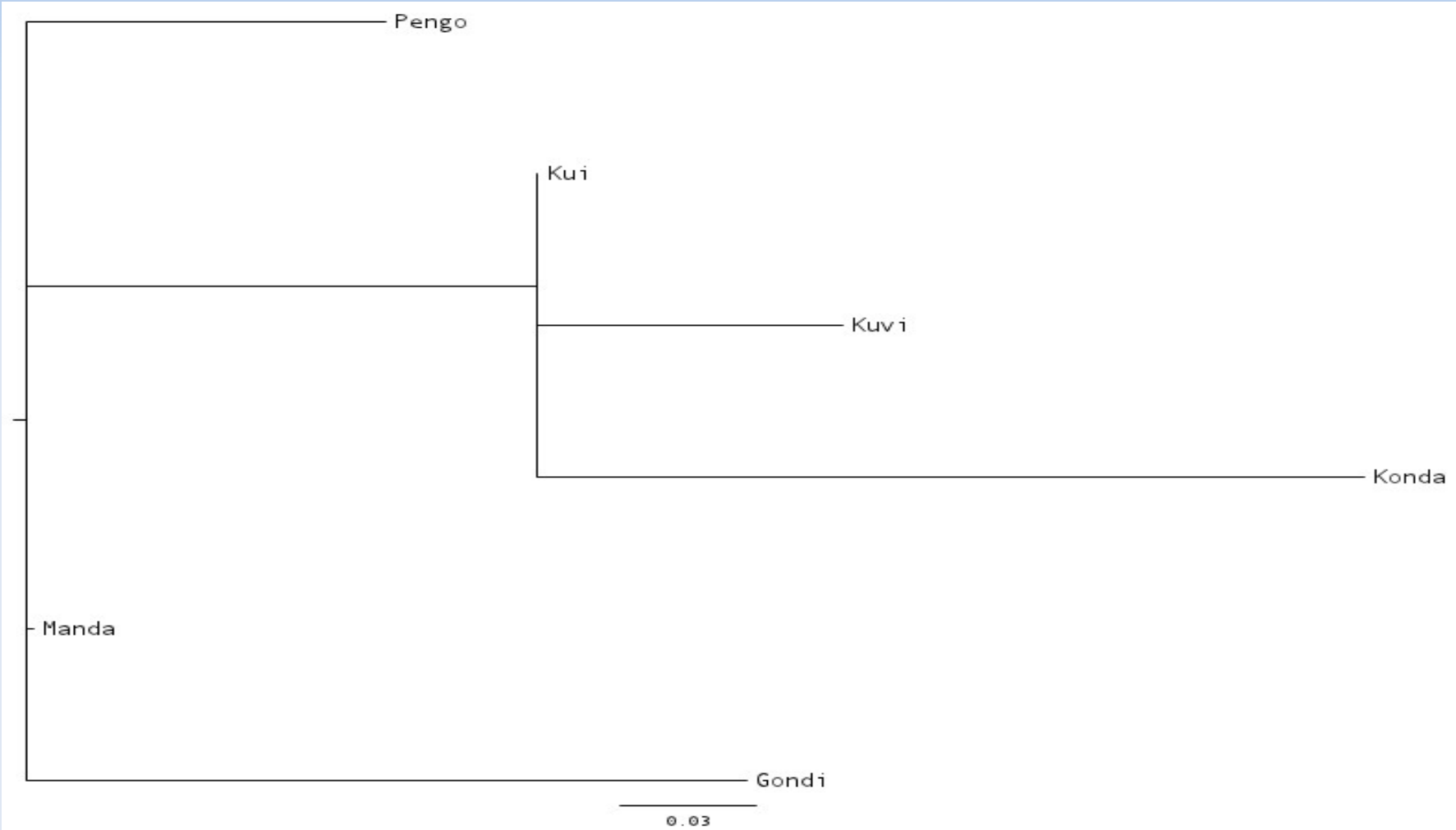


Results on dataset II A (Lexical Diffusion containing information about shared retentions)

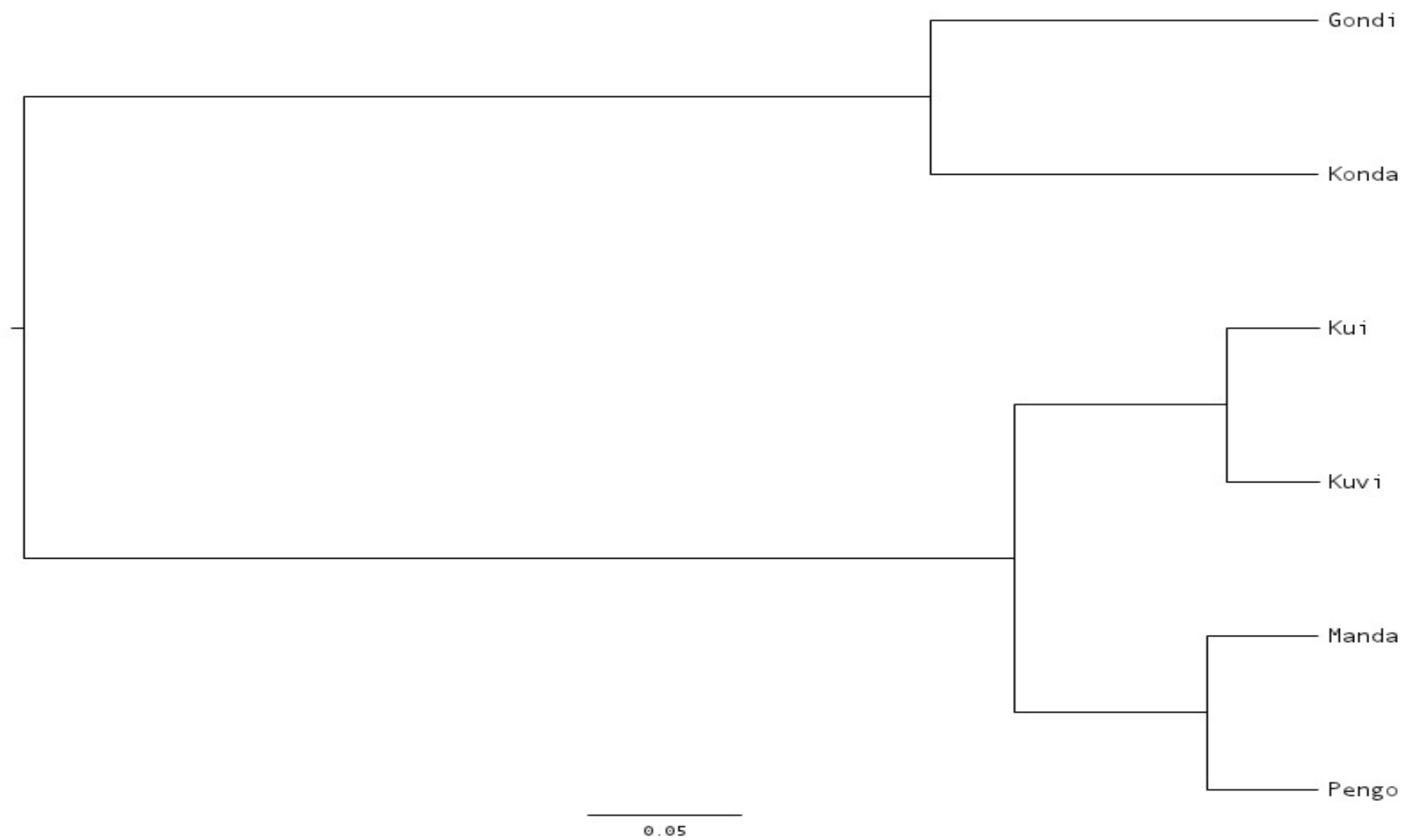
FM IIA



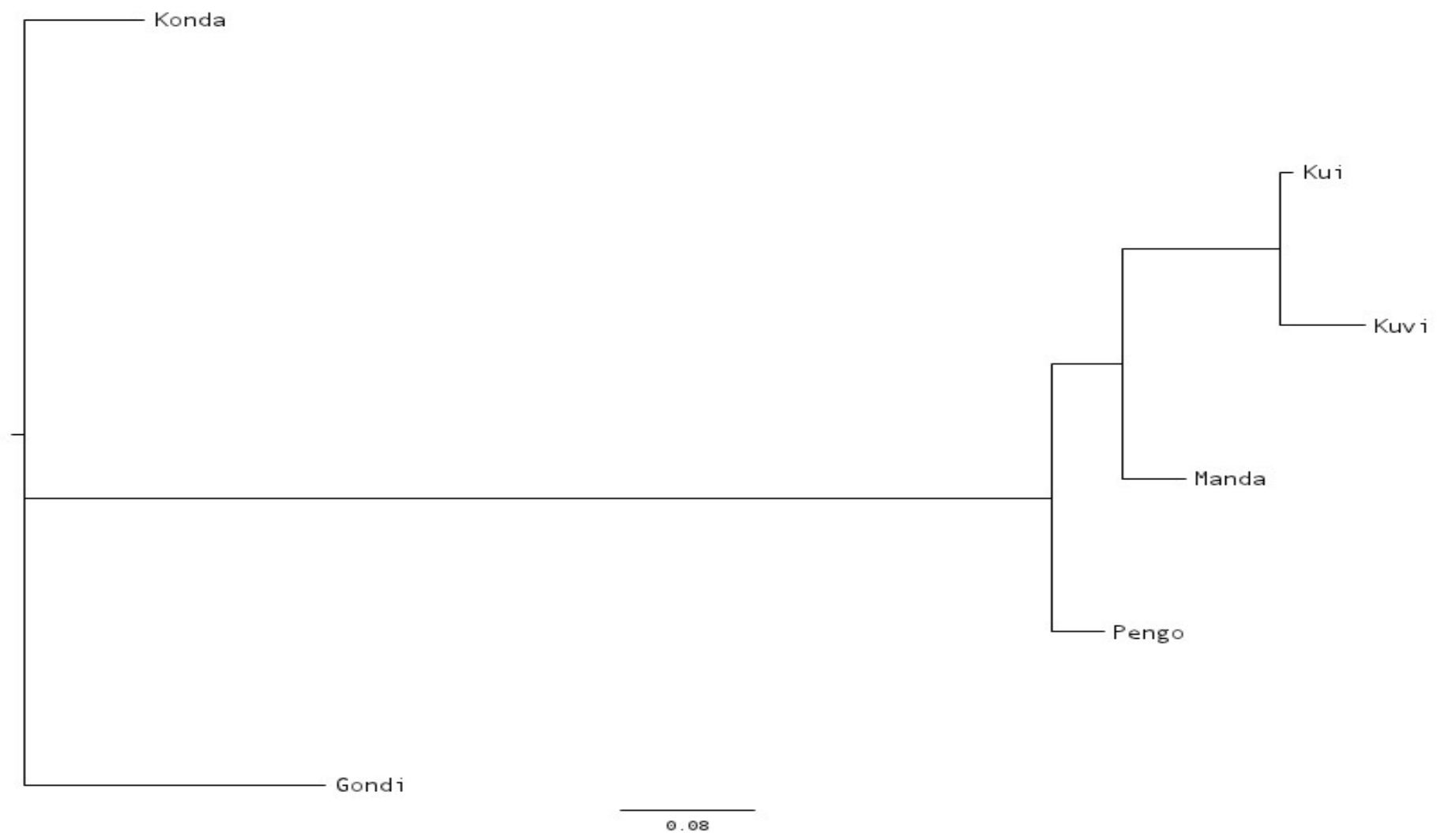
ME IIA



UPGMA IIA



NJ IIA



Comparison

- Tree comparison done automatically using Symmetric Difference (Felsenstein, 2003)
- Number of unshared splits between the two trees
- *Treedist* implementation of symmetric difference in PHYLIP

Pairwise tree distances

Fitch-Margoliash				Minimum Evolution			
I	IA	IB	IIA	I	IA	IB	IIA
0	6	2	4	0	4	4	6
6	0	4	4	4	0	2	6
2	4	0	2	4	2	0	6
4	4	2	0	6	6	6	0

NJ				UPGMA			
I	IA	IB	IIA	I	IA	IB	IIA
0	6	2	4	0	4	0	0
6	0	4	4	4	0	4	4
2	4	0	2	0	4	0	0
4	4	2	0	0	4	0	0

Observations

- Agreement between trees inferred from shared innovations datasets (I, IA, IB) and lexical diffusion dataset (IIA)
  - Worst in the case of Minimum Evolution (6, 6, 6)
  - Best in the case of UPGMA (0, 4, 0)
  - Similar for both FM and NJ (4, 4, 2)
- Agreement better between I, IB and IIA
  - Normalization 2 leads to better agreement
  - Tree inferred from 'proximity' values agrees equally well
- Summary: No 'perfect' agreement

Observations

- Agreement among trees inferred from shared innovations datasets (I, IA, IB)
  - Plenty of disagreement
  - Information contained in 'proximity' values not the same as that in distances
  - Again, agreement in the case of UPGMA better than the other methods

Fit Index : A Diagnostic Test

- Fit of the data to the tree structure evaluated using Least Squares
- Least Squares fit defined as
  - Ratio of *1-L* to the sum of the squared pairwise observed distances

    where, *L* = sum of the squares of the difference between the pairwise distance and the observed distances (Salemi, M. et al. 2010)
- Measure of the tree signal in the data

Fit Index

- Applied to results of Neighbor Joining and UPGMA for dataset IIA
- Implemented in Splitstree (Huson and Byrant 2006)
- Fit of the character data to the tree
  - Neighbor Joining : 99.492
  - UPGMA : 92.205
- Lexical diffusion data does contain information of a Tree

Conclusions

- No 'perfect' agreement between the trees inferred from the two datasets (result contrary to the perfect agreement reported by Krishnamurti)
- However, Fit index values suggest that lexical diffusion data does contain information about tree-like phylogeny
- Unchanged cognates (retentions) and their distributional relationship with changed cognates (innovations) in a lexcial diffusion scenario are useful at getting the subgrouping relations
- As always, desirable to experiment with more of such lexical diffusion data
- Tough nut: Identifying which changes are lexically diffused

Conclusions

- Normalization of the numbers of shared cognates-with-change an important factor
- Summary of our contributions
  - Application of distance-based phylogenetic inference methods to diachronic Dravidian datasets
  - Attempt to verify the usefulness of unchanged cognates in linguistic subgrouping claimed in previous work
  - Exploration of a specialized dataset such as the lexical diffusion data
  - Contribution to lexical diffusion studies???

Double Sound Change

- In a further application of MP-like postulates, Krishnamurti et al. (1983) study another dataset which contains a second sound change (word-initial consonant loss)
- Their inferred tree does not match with the traditional subgrouping or the tree inferred from single change data
  - No clear explanation provided
- Objection:
  - Not clear if the second sound change resulted from the first or affected the items independent of the first sound sound change (apical displacement)

Future Work

- Qualitative comparison of inferred trees with the standard tree
- Other possible normalizations while converting 'proximity' values to distances
- Include the double sound change dataset in the experiments
- As always, further experiments with more data
- Creation of specialized datasets such as the lexical diffusion data from the DEDR (Burrow, 1984)

References

- Campbell, L. 2004. Historical linguistics: an introduction. MIT Press.
- D'Andrade, R.G. 1978. U-statistic hierarchical clustering. Psychometrika 43(1):59–67.
- Felsenstein, J. Inferring phylogenies. Sinauer Associates Sunderland, Mass., USA
- Huson, D. H. and D. Bryant, Application of Phylogenetic Networks in Evolutionary Studies, Mol. Biol. Evol., 23(2):254-267, 2006..
- Krishnamurti, B. 1978. Areal and lexical diffusion of sound change. Language 54(1):1–20.
- Krishnamurti, B. 2003. The Dravidian languages. Cambridge University Press.
- Krishnamurti, B., L. Moses, and D. Danforth. 1983. Unchanged cognates as a criterion in linguistic subgrouping. Language 59(3):541–568.
- Nakhleh, Luay, Tandy, Warnow, Donald A., Ringe, Jr, and Steven N. Evans. 2005. A comparison of phylogenetic reconstruction methods on an IE dataset. Transactions of the Philological Society 103.171–92.
- Salemi, M. and Vandamme, A.M. and Lemey, P. 2010. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge University Press
- Wang, William, S-Y. 1969. Competing changes as a cause of residue. Language 45 9-25.
- Burrow, T. A Dravidian Etymological dictionary. 2 ed., Oxford: Clarendon press, 1984

Questions?

DOMO ARIGATO !!!