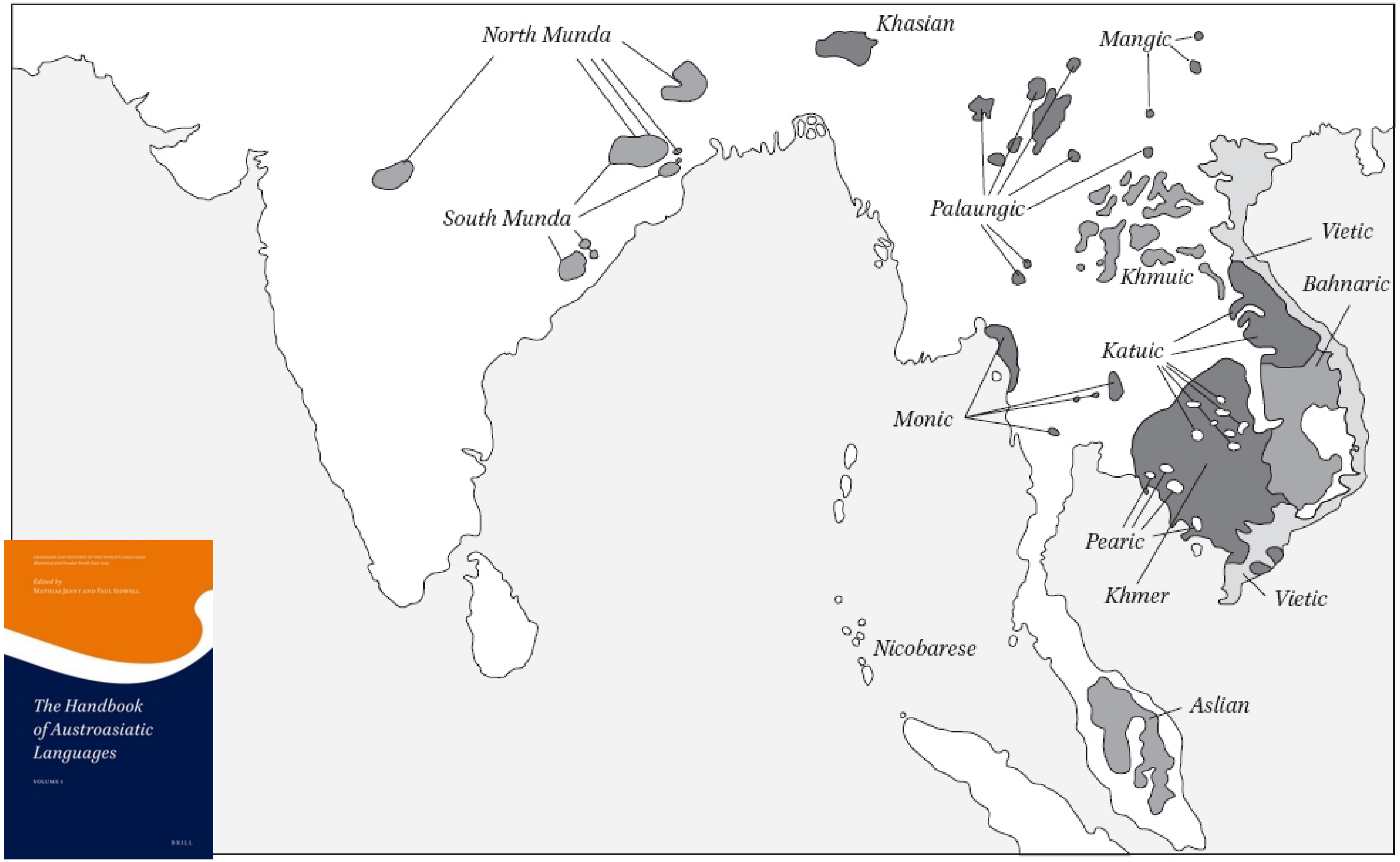
A comprehensive phylogenetic analysis of the Austroasiatic languages

Diversity Linguistics:
Retrospect and Prospect

Paul Sidwell paul.sidwell@anu.edu.au

Closing Conference of the Department of Linguistics at the MPI EVA May 1-3, 2015

Austroasiatic



GRAMMARS AND SKETCHES OF THE WORLD'S LANGUAGES
Mainland and Insular South East Asia

Edited by
Mathias Jenny and Paul Sidwell

The Handbook of Austroasiatic Languages

VOLUME 1

BRILL

Background

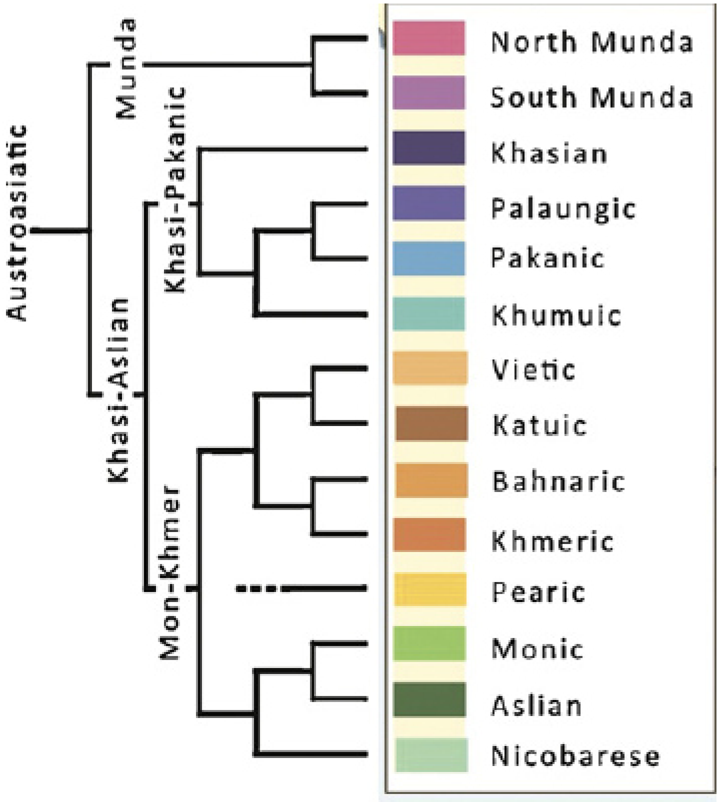
- Austroasiatic is the principal linguistic substrate of MSEAsia (130+ languages, many more named lects).
- All other language families are later intrusions.
- Various efforts and methods produced very different AA classifications over 100+ years.
- A small number of widely cited classifications give an illusion of consensus.
- Generally agreed that there are ~13 branches, but fundamental questions remain:
  - Do branches form nested sub-groupings implying a deep history?,
  - Are they coordinated in a radial or rake pattern, suggesting rapid dispersal?
  - Is there a centre of diversity / apparent homeland ?

Conflicting claims: e.g. homeland…..

- **Northern India** (Vedic substrate?): Berger & Mayrhofer, Levi (1923), Przyluski (1922, 1923), Bloch (1930), Kuiper (1948, etc.), Fuller 2010…
- **Western India** (Indus Valley): Witzel (1999)
- **Eastern India**: Pinnow (1963)
- **Shores of Bay of Bengal**: van Driem (2001), Diffloth (2011)
- **Southern China**: Nagaraja (2011)
- **Central China/Yangtze River**: Norman & Mei (1976), Haudricourt (1966), Jakhontov (1977), ….
- **Eastern China**/**Shandong**: Schuessler (2007)
- **Southeast Asia**: von Heine-Geldern (1928, 1932), Shorto (1979), Belwood (2001), Sidwell & Blench (2011), etc.

○ Generally poor arguments along the lines of,
"The name of the middle stretch of one river in China resembles …..."
"Prefixes with ka- are found in the AV, YV and the Brahmanas….."

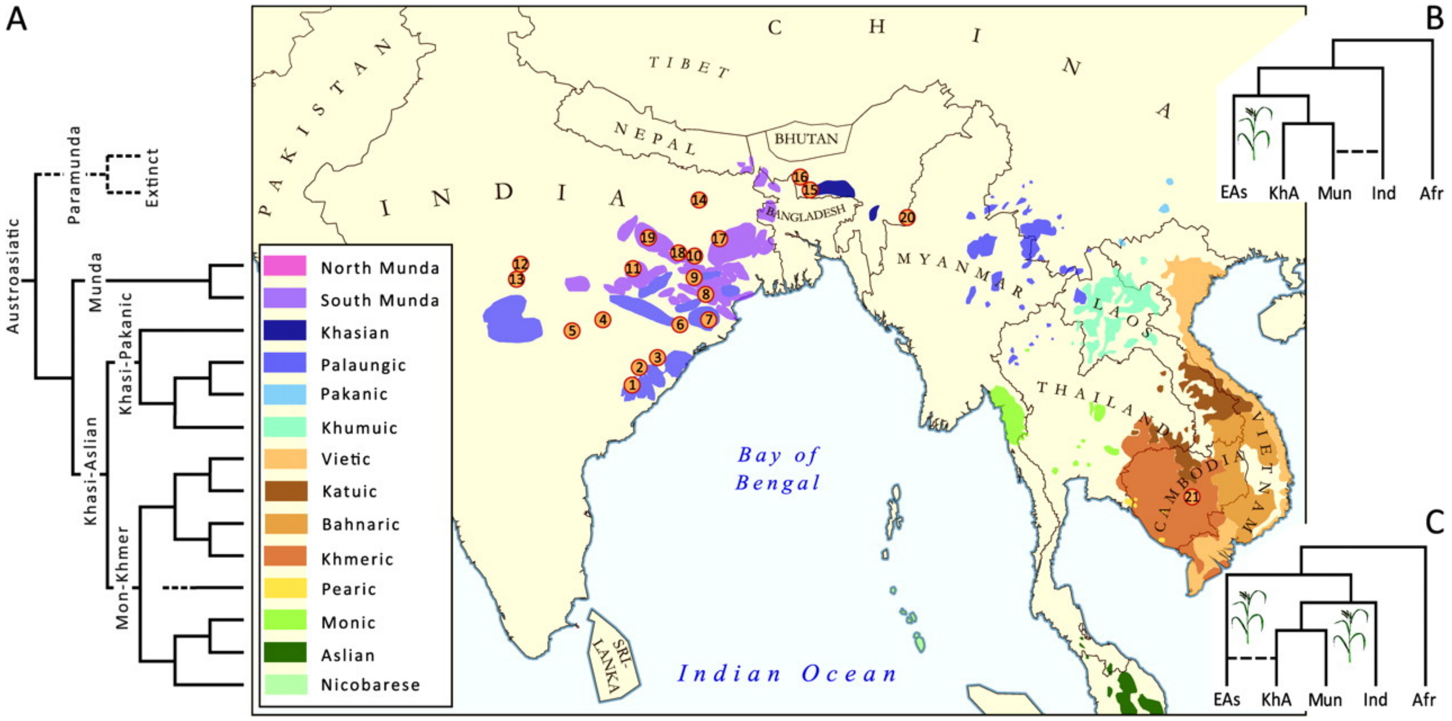
Classifications cited uncritically…



Diffloth, G., 2009. More on Dvaravati Old Mon. Paper presented at the Fourth International Conference on Austroasiatic Linguistics. Mahidol University, Salaya.

- van Driem (2012) cites Diffloth (2009) as the source of this classification.
- I am not picking on George, he has to use what he can piece together from scraps. The diagram on the left does not appear in Diffloth's unpublished conference talk, which was about the etymology for 'wife' in Mon and the *ua > a sound change it demonstrates.

Classifications cited uncritically…#2



- The above, credited to Diffloth (2009) but obviously lifted from van Driem, appears in a *Molecular Biology and Evolution* paper canvassing possible AA homeland in India.
Notice augmentation with *Paramunda* branch
(Gyaneshwer Chaubey et al. Mol Biol Evol 2011; 28:1013-1024)

We need assessable studies/methods

- The modelling of language history changed with the introduction of computational phylogenetics, e.g.:

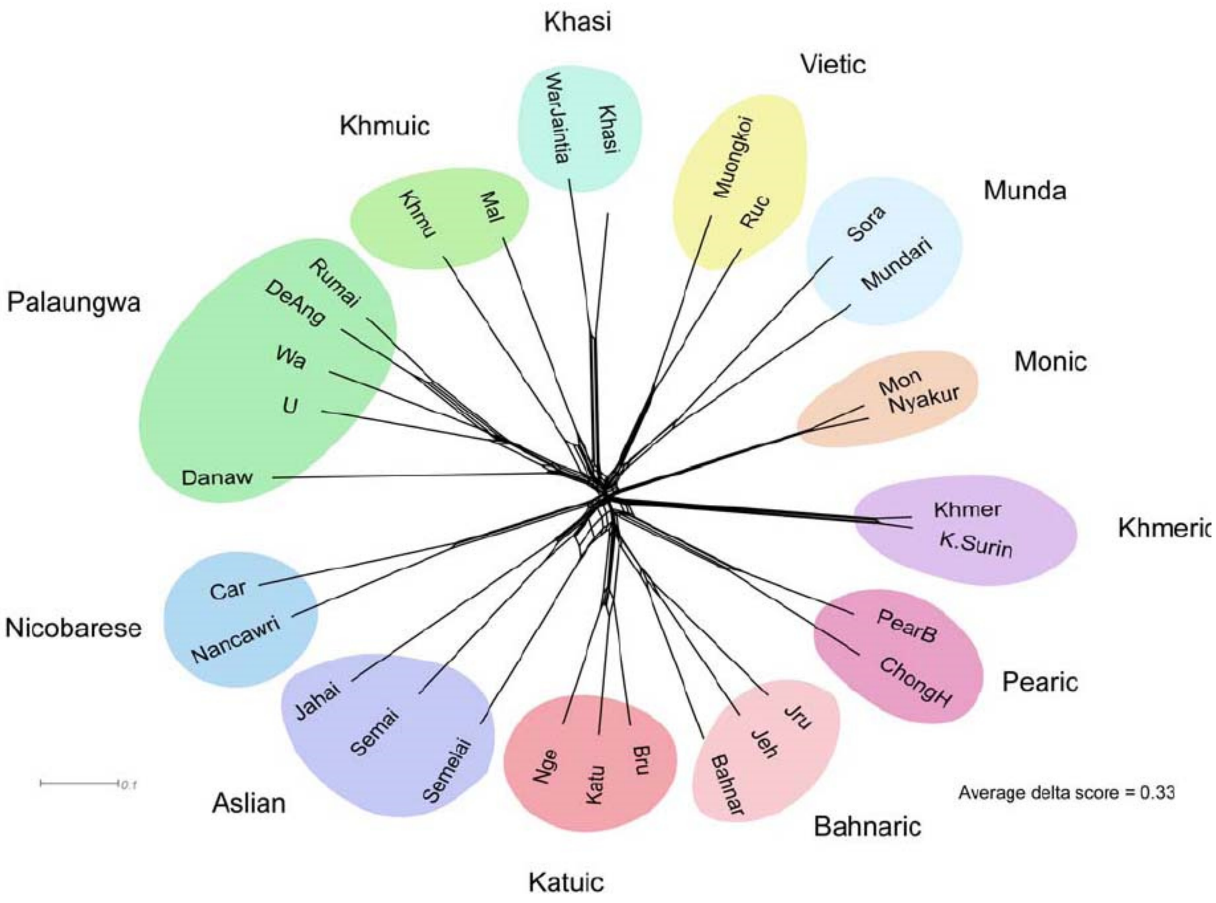
  Indo-European (Gray & Atkinson 2003, etc.)
  Austronesian (Gray & Jordan 2000 etc.)
  and many more studies since….

- With support from Russell Gray and Simon Greenhill,
  since 2009 I have been trialling the phylogenetic methods with AA.

- I tried lexicostatistics, so I had well organised data to start with.

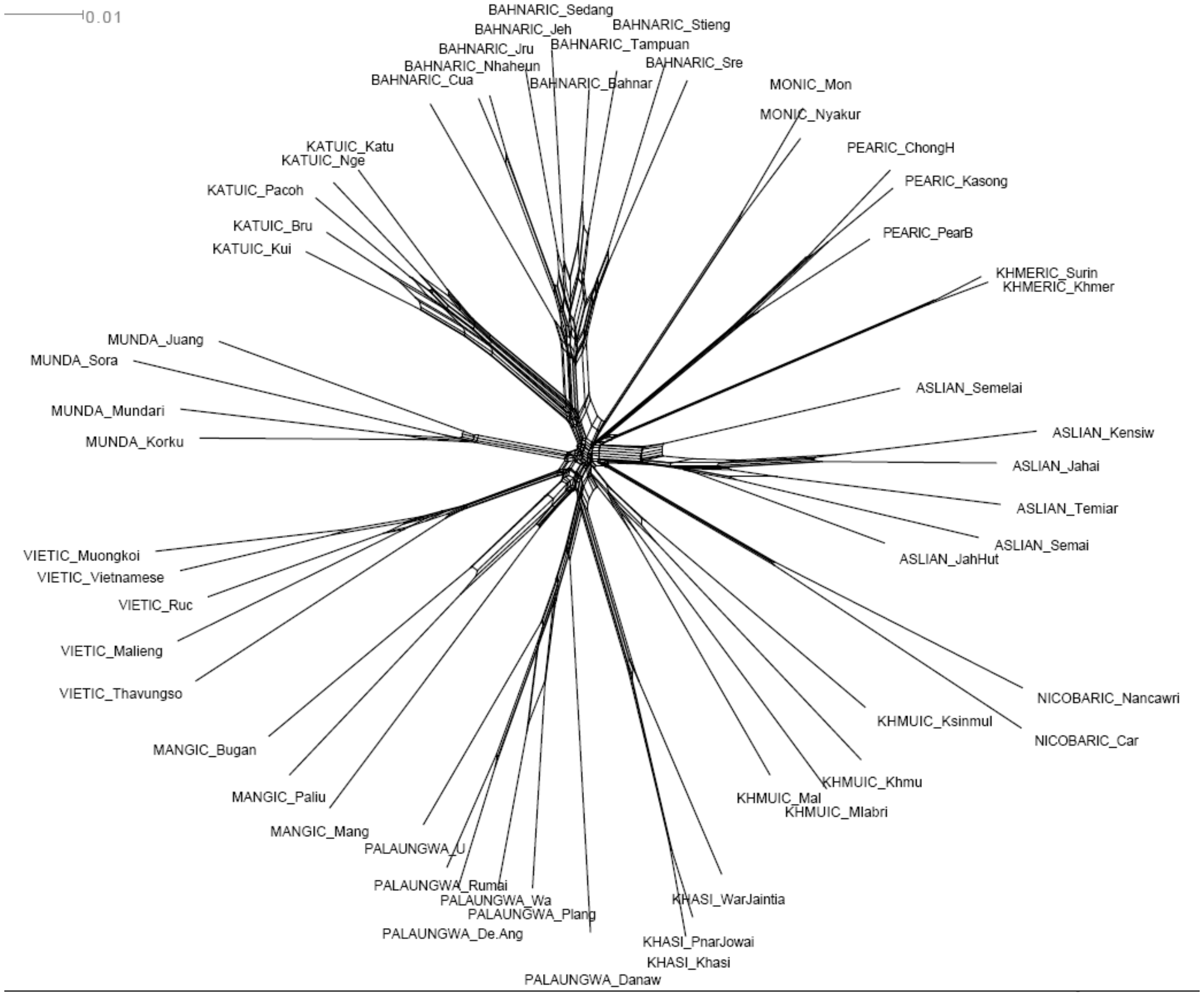
- The first results were presented as NeighbourNet analyses at the 2009 ICAAL meeting published as Sidwell and Blench (2011)

2009 NeighborNet:
30 languages 100 words



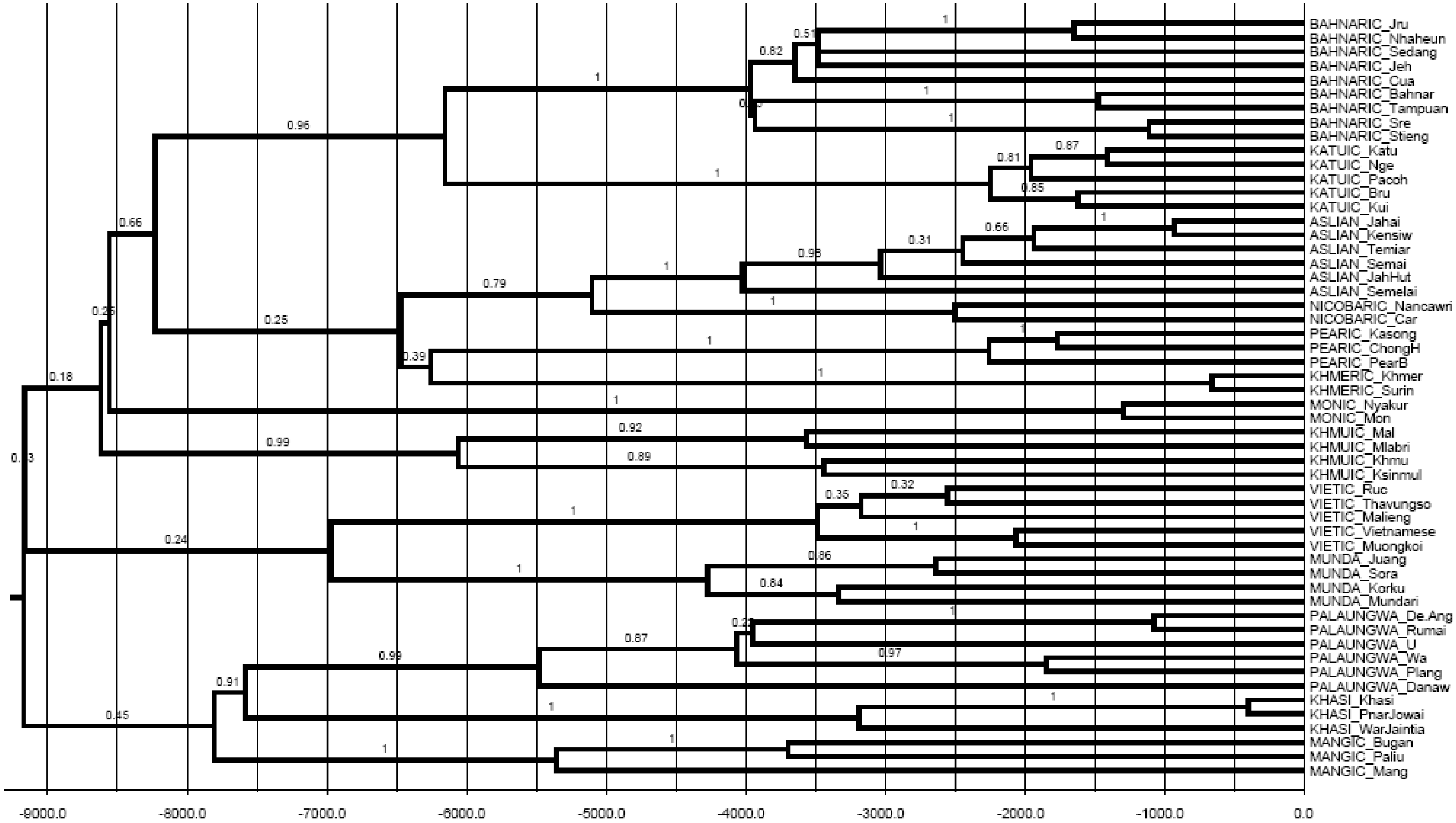
Strongly branching,
suggests radical dispersal,
Not progressive branching.
No strong support for
existing classifications.

2010 NeighborNet:
54 languages 100 words



Nearly doubled the amount of data giving more detail to larger branches.
Overall pattern is the same, plus beginning to get lower level grouping structure.

2010 tree analysis: covarion-relaxed clock
54 languages 100 words



Family tree results considered:

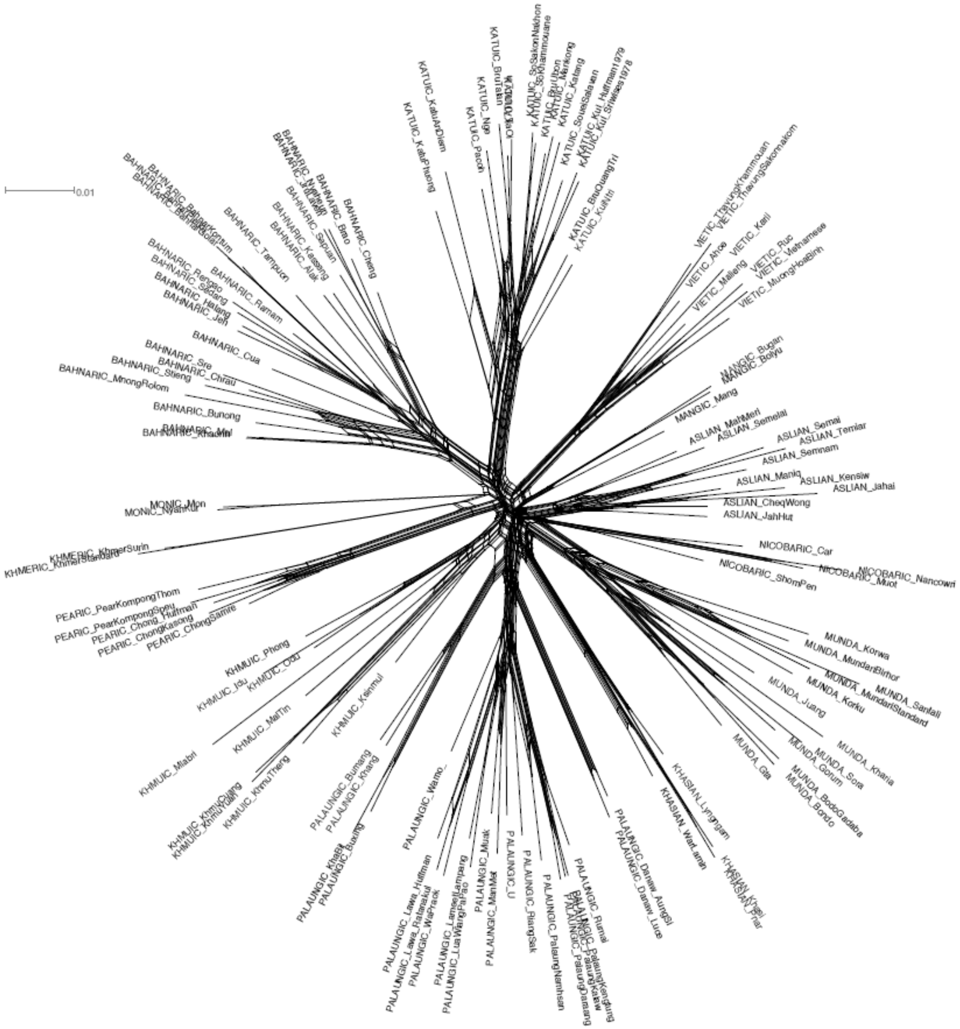
- The 54 langs./100 words covarion-relaxed clock:
  - Very old overall dating estimate: 9000+ BP, and a mix of old and young branch-level estimates;
  - Suggestions of nested branching with some unexpected groupings, e.g. Viet-Munda, Katuic-Bahnaric; but mixed levels of statistical confidence.
  - High certainly found with the 13 principal branches.

2014 began new phase:

- Expand no. of langs to every ISO with useful data: 120+
- Increase list size to 200 words
- Mix with some money, blind optimism and sleepless nights:

2015 NeighborNet:
122 sources, 200 words

- Now we are getting somewhere.



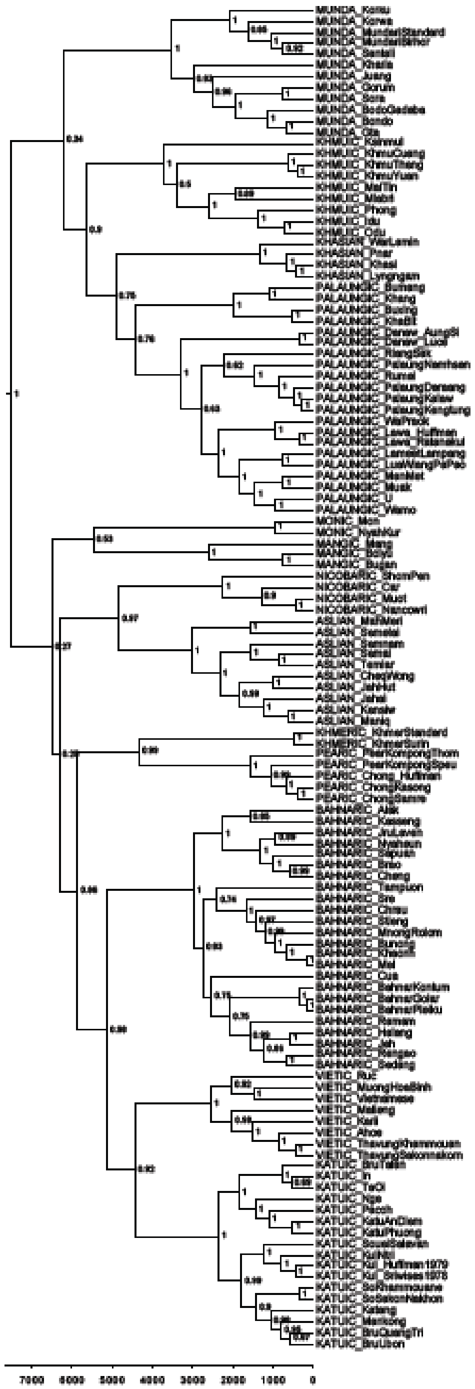
2015 tree analysis
122 sources, 200 words

- Maximum Clade Credibility Tree of the CTMC + Gamma Relaxed
  Analysis was run by Greenhill

- Data complied into a spreadsheet and coded for cognates by Sidwell

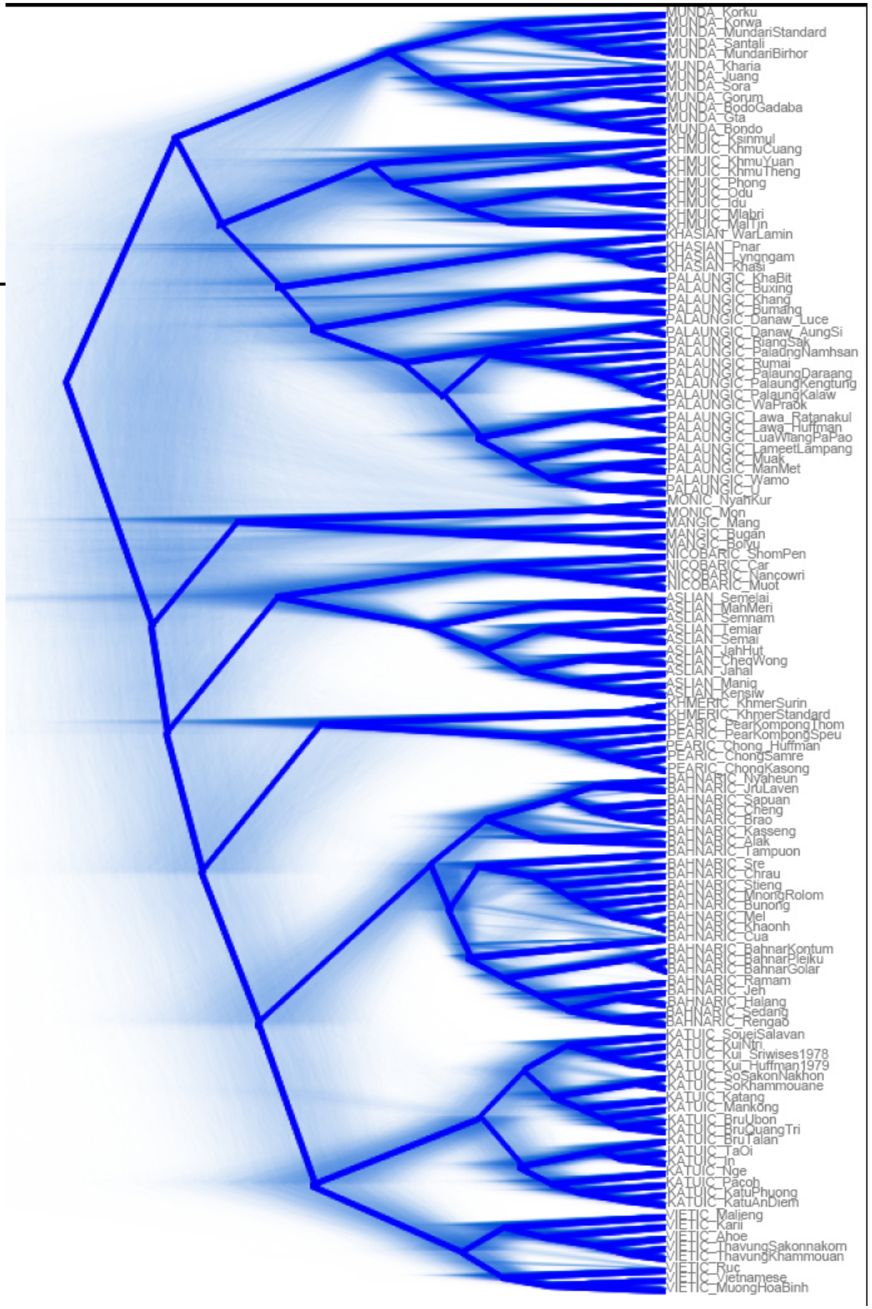
- Data coverage is more than 80%, it is extremely problematic to get coverage from extant sources

- The tree has characteristics that suggest a high degree of usefulness.



2015 tree analysis Densitree

The Densitree indicates conflicting signal at depth: There are many trees in a tree!

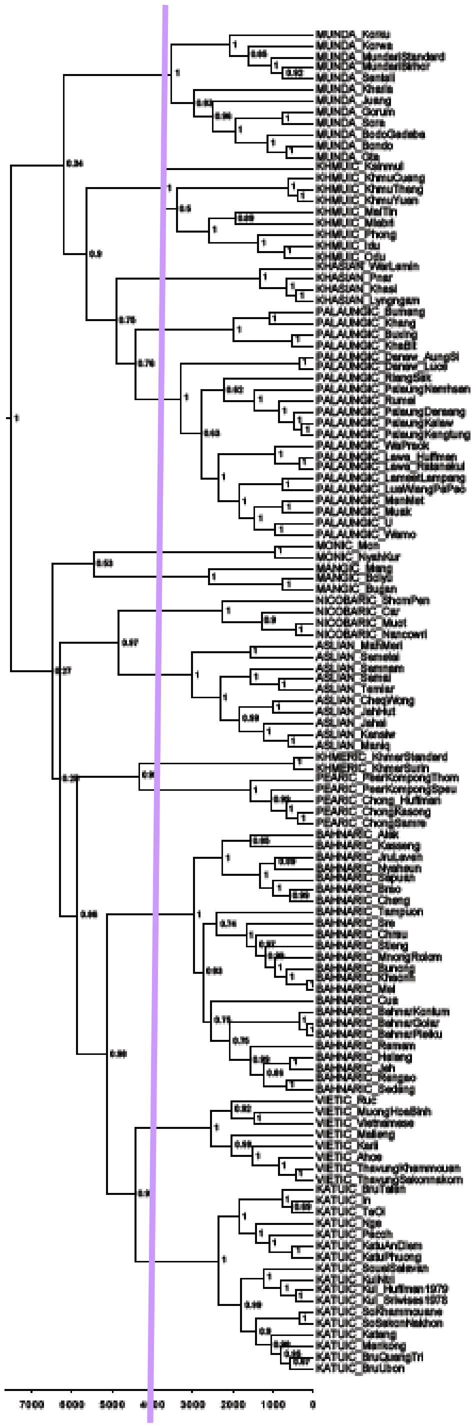


2015 tree analysis discussion

- ○ Calibrated dating estimates are much tighter;
  there is a basic east-west split ~7000BP
  all branches established by 4000BP
  rapid branch internal growth from ~2500BP

- ○ Some coordination only weakly supported:
  - Munda-WestAA
  - Monic-Mangic
  - Nico-Asli with EastAA

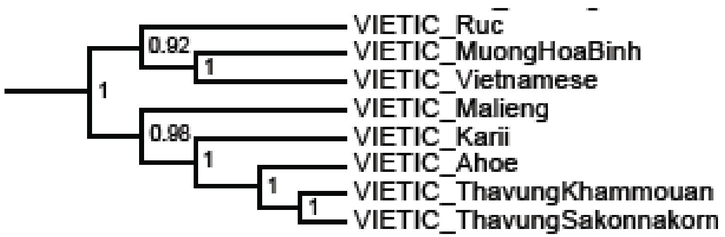
- ○ Surprising/doubtful strong groupings:
  - Katuic-Vietic
  - Khmer-Pearic
  - primary E-W split



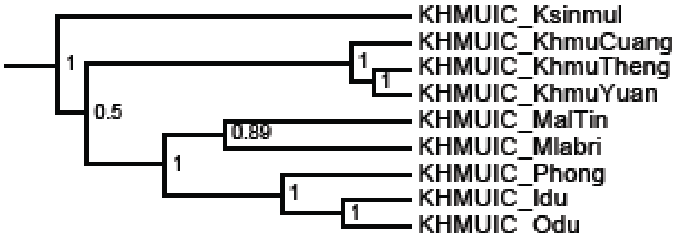
Branch internal results #1

Some odd results are achieve for single languages

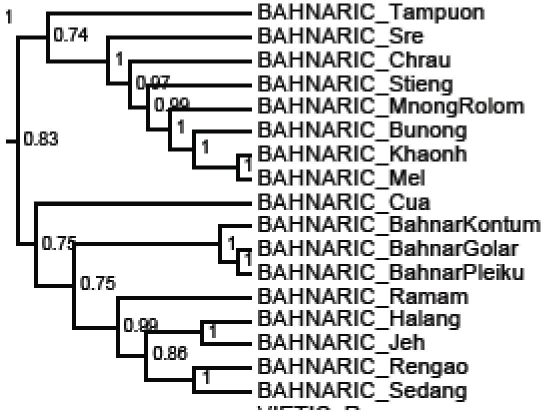
Vietic is good, except **Ruc**, but the list has only about 50% coverage so Viet cognates/loans dominate skewing the result



Khmuic is good, except for **Ksinmul**, not clear why.



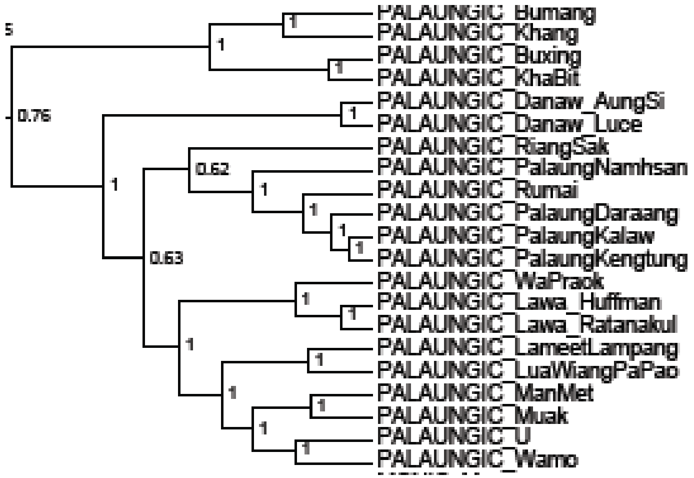
Bahnaric is good but **Bahnar** should pair with **Tampuon**, but it is not possible to separate all **Rengao** loans, but NeighborNet groups them.



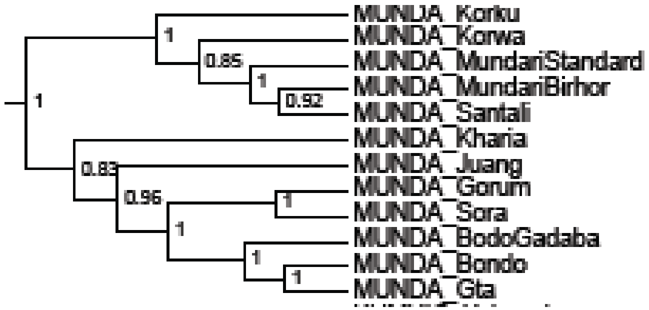
Branch internal results #2

Some odd results are achieve for single languages

Palaungic is good, except the **Bit-Khang** sub-group should coordinate with East Palaungic, but ½ lexicon is replaced with with mostly Tai loans.



Munda is good except that we expect **Kharia-Juang** to pair, but this is seen in NeighborNet.



Provisional assessment

Computational phylogenetic methods perform very well with good data coverage (> 80%) and good cognate recognition

Sensitive to data density / loans, factor that skew wordlists towards basal forms, but needs formal testing

Need to quantify reliability and conduct well structured experiments to test practical thresholds, different wordlists, semantic fields.

Cannot be pursued blindly/dumbly, expect phylum specific knowledge is needed on the part of investigators.

Experiments in automated cognate recognition would be interesting.

thanks