

ASJP World Language Tree of Lexical Similarity: Version 3 (July 2010)

by

**André Müller, Søren Wichmann, Viveka Velupillai, Cecil H. Brown,
Pamela Brown, Sebastian Sauppe, Eric W. Holman, Dik Bakker, Johann-
Mattis List, Dmitri Egorov, Oleg Belyaev, Robert Mailhammer, Matthias
Urban, Helen Geyer, and Anthony Grant**

The World Language Tree graphically illustrates relative degrees of lexical similarity holding among 4350 of the world's languages and dialects (henceforth, languages) currently found in the ASJP database (ASJP stands for Automated Similarity Judgment Program). Languages branched more closely together on the ASJP tree are lexically more similar than those branched less closely together. While most lexical resemblance charted in the tree almost certainly is related to genetic affiliation, closely branched languages cannot routinely be assumed to be closely genetically associated since lexical resemblance among languages can be due to factors other than genetic relatedness (see below).

The tree is generated through use of the neighbour-joining computer algorithm originally designed to depict phylogenetic relationships in biology (Saitou & Nei 1987). This is implemented in MEGA 4 (Kumar et al. 2008),¹ the software that we use. The algorithm is applied to a matrix of lexical similarity scores based on Levenshtein (or edit) distances holding between all possible pairs of the 4350 languages (for details about this, including how we modify the Levenshtein distances for our purposes, see Bakker et al. 2009: 169). All languages of the database are compared to one another with respect to lexical similarity relating to their words for 40 referents determined statistically in Holman et al. (2008) to be most stable among core vocabulary items commonly used in lexicostatistical analysis. The tree is unrooted, but organized around a midpoint, i.e., the point which is equidistant between the two most lexically dissimilar languages in the network. Finally, the tree is annotated to show how it corresponds to the classification used in the latest version of the online World Atlas of Language Structures (Haspelmath et al. 2008),² with some updates from Dryer (personal communication). This annotation is presented for ease of orientation, not necessarily because ASJP agrees with it. The language names used are normally simply those

¹ <http://www.megasoftware.net/>

² <http://wals.info/>

of the sources consulted. The sources, as well as corresponding language names of *Ethnologue*, are provided in a continuously updated wiki.³

Four factors influence lexical similarity registered in the tree: (1) genetic or genealogical relationship of languages, (2) diffusion (language borrowing), (3) universal tendencies for lexical similarity such as onomatopoeia, and (4) random variation (chance).

Languages branched closely together on the tree may be so because of strong lexical similarity produced by any one or a combination of the four factors. Genetic relationship would appear to be the most dominant factor accounting for close branching, followed next by diffusion. Universal tendencies and chance are less significant contributors to close branching than either genetic relationship or diffusion, but nonetheless clearly contribute to the overall structure of the tree.

Typically, all languages of non-controversial language families such as Austro-Asiatic, Uralic, or Mayan, are respectively branched together on the tree. When some languages of a non-controversial family are not found branched together, this is because they are substantially lexically different from other members of their family despite unambiguously belonging to that family. Occasionally, a language can be so lexically different from co-members of its family that it is found branched more closely with some language or languages with which it is not genetically related at all, usually because of chance lexical similarity or similarity due to borrowing. (When such languages are geographically remote from one another, chance usually explains close branching.)

Typically, branching accords closely with genetic subgroups recognized by experts within non-controversial language families. When branching is not isomorphic with genealogical subgrouping, this often reflects diffusion among languages of the family promoted by language contact. Thus, when used in conjunction with expert classifications of non-controversial language families, the tree can be helpful in calling attention to historical relationships (contact) among genetically related languages that sometimes might not be otherwise apparent.

The tree may also suggest relationships heretofore not noticed among languages that may be profitably investigated. For example, if two languages not known to be related in any way are found together on a terminal branch, this may indicate a relationship between them entailing either inheritance or contact, especially if they are not geographically remote from one another. If the two languages are geographically distant, their close lexical similarity is

³ <http://lingweb.eva.mpg.de/asjp/index.php/ASJP>

more likely explained by chance than by either inheritance or diffusion. Also, language isolates may join one another on a terminal branch because they have nowhere else to go in the tree, creating the illusion that exciting, new far-flung relations may be in evidence. One should be cautious in the interpretation of these cases.

Earlier versions of the ASJP World Language Tree did not include languages regarded as creoles and pidgins, while this version does. Excluding creoles and pidgins would allow human judgments to intrude into the classification (or, in this case non-classification). It is of general interest to show how such languages pattern in the classification when preconceived notions of how they should be treated are avoided.

References

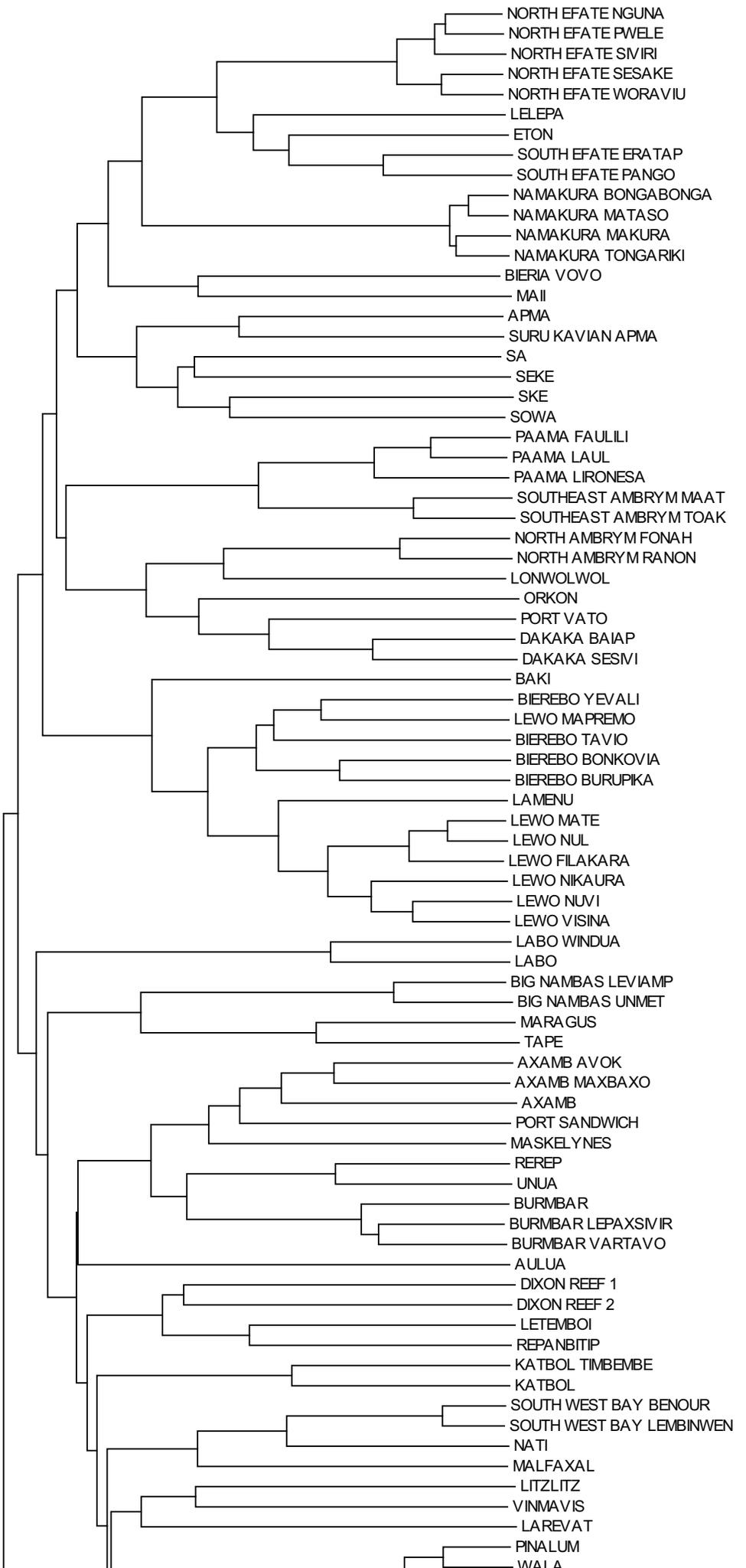
- Bakker, Dik, André Müller, Viveka Velupillai, Søren Wichmann, Cecil H. Brown, Pamela Brown, Dmitry Egorov, Robert Mailhammer, Anthony Grant, and Eric W. Holman. 2009. Adding typology to lexicostatistics: a combined approach to language classification. *Linguistic Typology* 13: 167-179.
- Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker. 2008. Explorations in automated lexicostatistics. *Folia Linguistica* 42.2: 331-354.
- Haspelmath, Martin, Matthew S. Dryer, David Gil and Bernard Comrie. 2008. *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library.
- Kumar S., J. Dudley, M. Nei, and K. Tamura K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. *Briefings in Bioinformatics* 9: 299-306.
- Saitou, Naruya and Masatoshi Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4: 406-425.

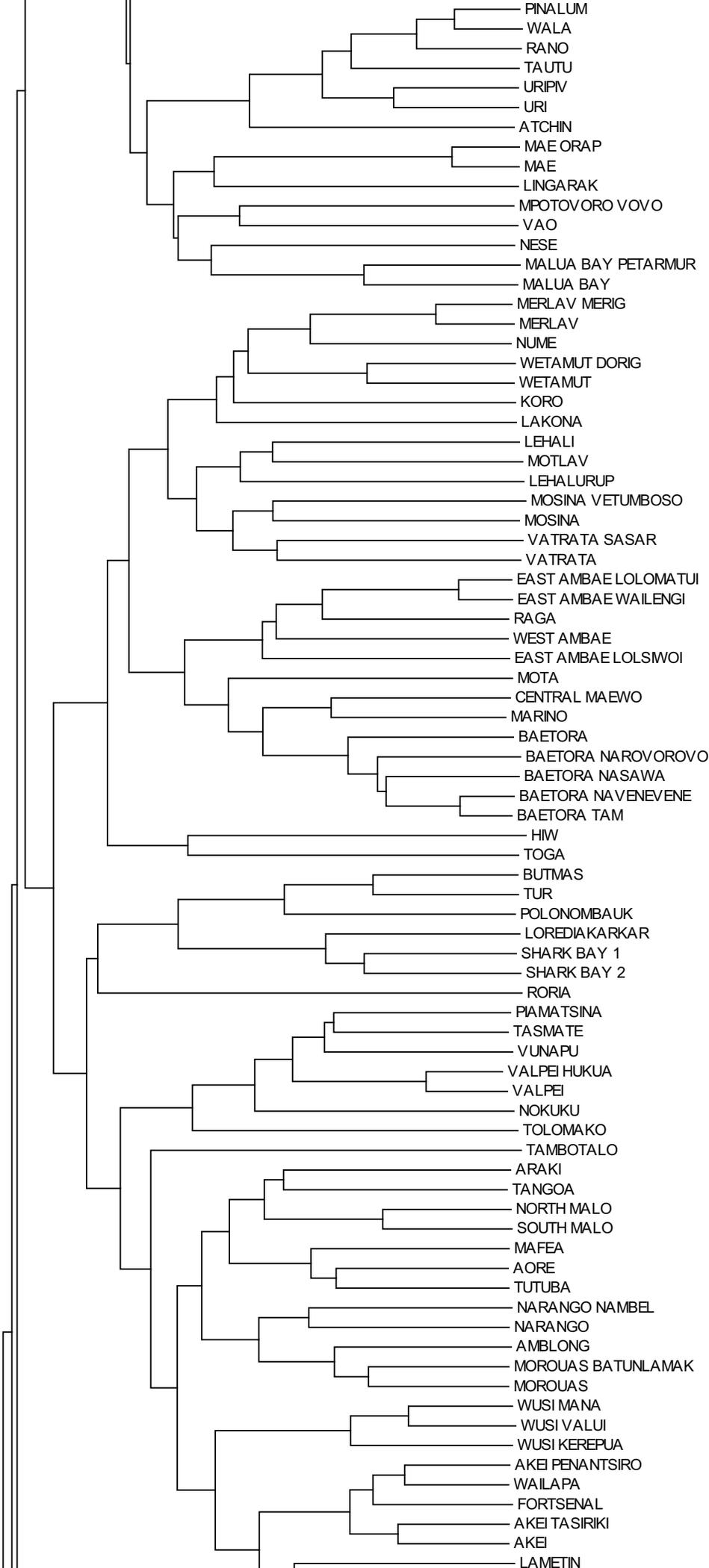
Language family abbreviations

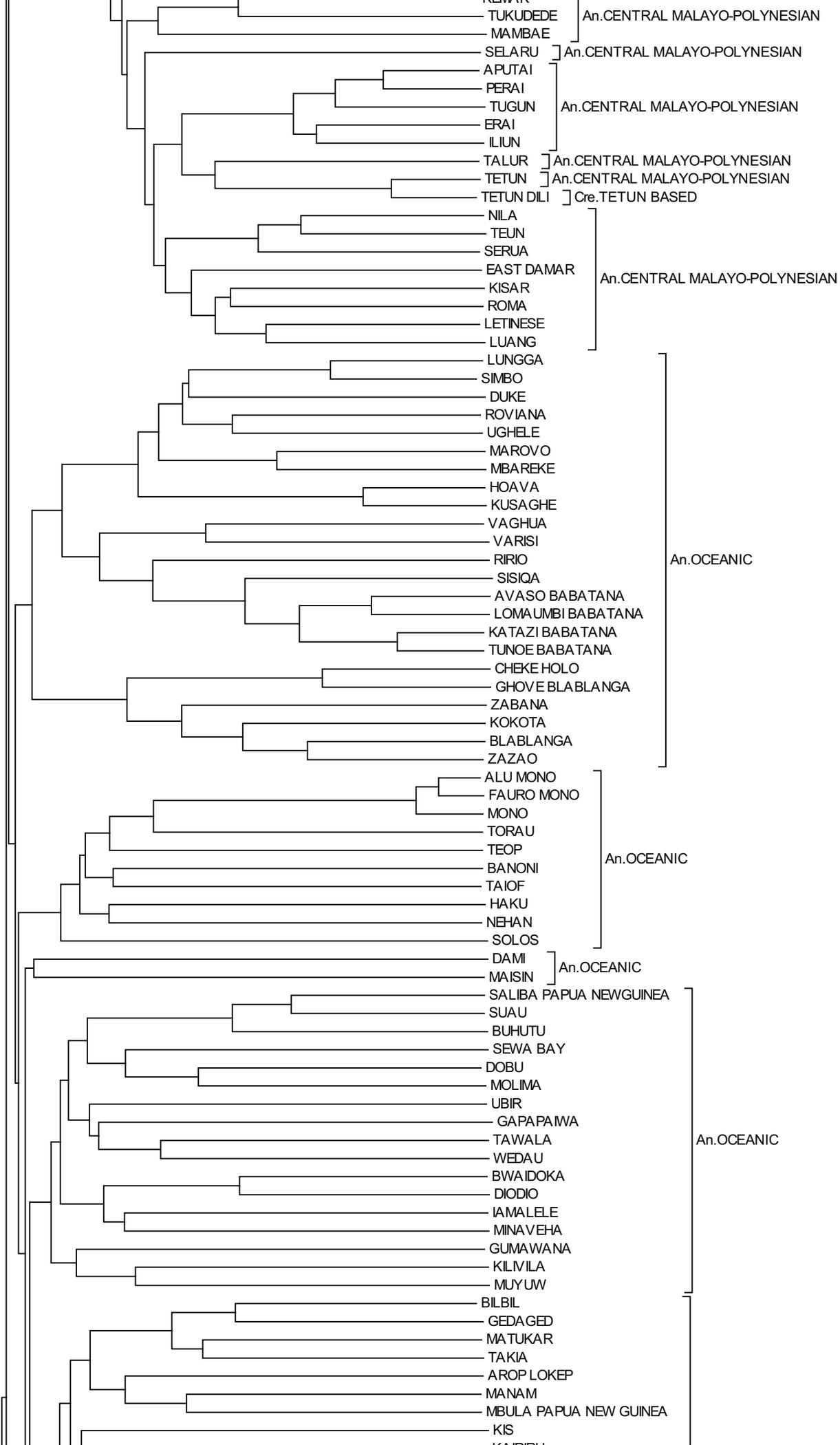
AA	Afro-Asiatic	Chn	Chon
Aik	Aikana	Cho	Choco
Ain	Ainu	Chq	Chiquito
Ala	Alacalufan	Cht	Chitimacha
Alg	Algic	Chu	Chumash
Alt	Altaic	CK	Chukotko-Kamchatkan
AM	Amto-Musan	Cmu	Chimúan
An	Austronesian	CN	Cacua-Nukak
AP	Awin-Pare	Cnd	Candoshi
Ara	Arafundi	Cof	Cofán
Arc	Araucanian	Com	Comecrudan
Art	Arutani	Cre	Creoles & Pidgins
Aru	Arauan	Cui	Cuitlatec
Arw	Arawakan	CW	Chapacura-Wanhan
Ata	Atakapa	Dos	Doso
AuA	Austro-Asiatic	Dra	Dravidian
Aus	Australian	EA	Eskimo-Aleut
Aym	Aymaran	EB	East Bougainville
Ban	Bangi Me	EBH	East Bird's Head
Bar	Barbacoan	EGB	East Geelvink Bay
Bas	Basque	Ele	Eleman
Beo	Beothuk	ES	East Strickland
Bil	Bilua	GA	Great Andamanese
Bor	Border	Gcu	Guaicuruan
Bos	Bosavi	GS	Gogodala-Suki
Brs	Burushaski	Gua	Guahiban
Bul	Bulaka River	Had	Hadza
Bur	Burmeso	Hai	Haida
Cad	Caddoan	Har	Harakmbet
Cah	Cahuapanan	HM	Hmong-Mien
Cam	Camsá	Hok	Hokan
Car	Cariban	Hua	Huavean
Cay	Cayuvava	Hui	Huitotoan
Chi	Chibchan	IE	Indo-European
Chk	Chimakuan	IG	Inland Gulf
Chl	Cholón	Ira	Irantxe
Chm	Chimila	Iro	Iroquoian

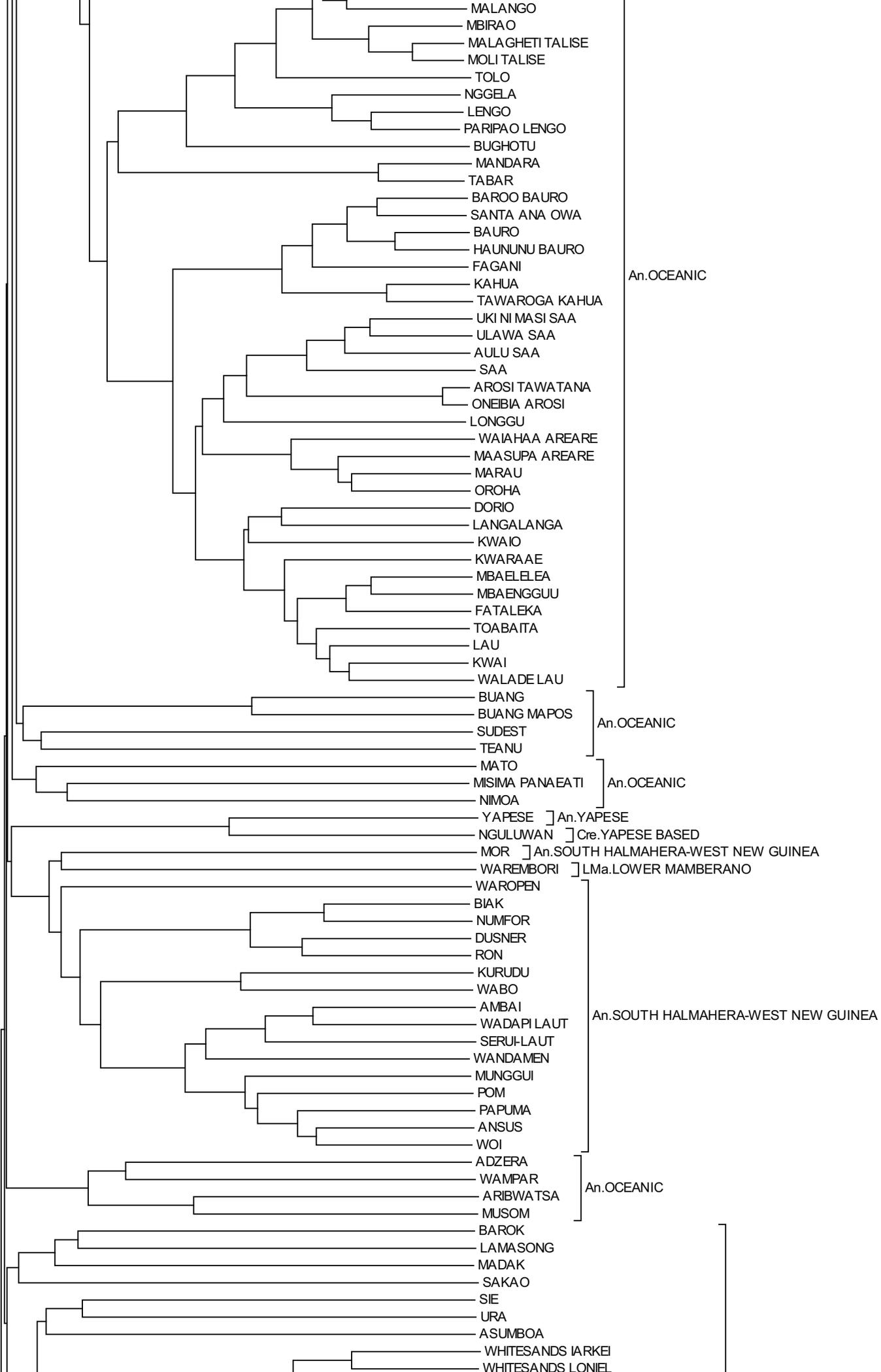
Ito	Itonama	May	Mayan
Jab	Jabuti	MGe	Macro-Ge
Jap	Japanese	Mis	Misumalpan
Jiv	Jivaroan	Mol	Molof
Kad	Kadugli	Mom	Mombum
Kam	Kamula	Mon	Monumbo
Kap	Kapixana	Mos	Mosetenan
Kar	Karok	Mov	Movima
Kat	Katukinan	Mrw	Morwap
Kau	Kaure	MUM	Morehead and Upper Maro Rivers
Kay	Kayagar	Mur	Mura
Ken	Kenaboi	Mus	Muskogean
KF	Kwomtari-Fas	MZ	Mixe-Zoque
Kho	Khoisan	Nah	Nahali
Kiw	Kiwaian	Nam	Nambikuaran
Kol	Kolopom	Nat	Natchez
Kor	Korean	NC	Niger-Congo
Krt	Kartvelian	NDa	Nakh-Daghestanian
KT	Kiowa Tanoan	NDe	Na-Dene
Ktn	Kutenai	Nim	Nimboran
Kuj	Kujarge	Niv	Nivkh
Kun	Kunza	NS	Nilo-Saharan
Kus	Kusunda	NWC	Northwest Caucasian
Kut	Kuto	OC	Oregon Coast
Kwa	Kwalean	Odi	Odiai
Kwe	Kwerba	Oks	Oksapmin
Kwz	Kwaza	OM	Oto-Manguean
Lav	Lavukaleve	Pae	Paezan
Lec	Leco	Pan	Panoan
LeM	Left May	Pat	Pataxo
Len	Lencan	Pau	Pauwasi
LMa	Lower Mamberamo	Pen	Penutian
LP	Lakes Plain	Pui	Puinave
LS	Leonhard Schultze	PY	Peba-Yaguan
LSR	Lower Sepik-Ramu	Que	Quechuan
Mai	Mairasi	Sal	Salishan
Mar	Marind	SAn	South Andamanese
Mas	Mascoian	Sav	Savosavo
Mat	Matacoan	Sen	Senagi

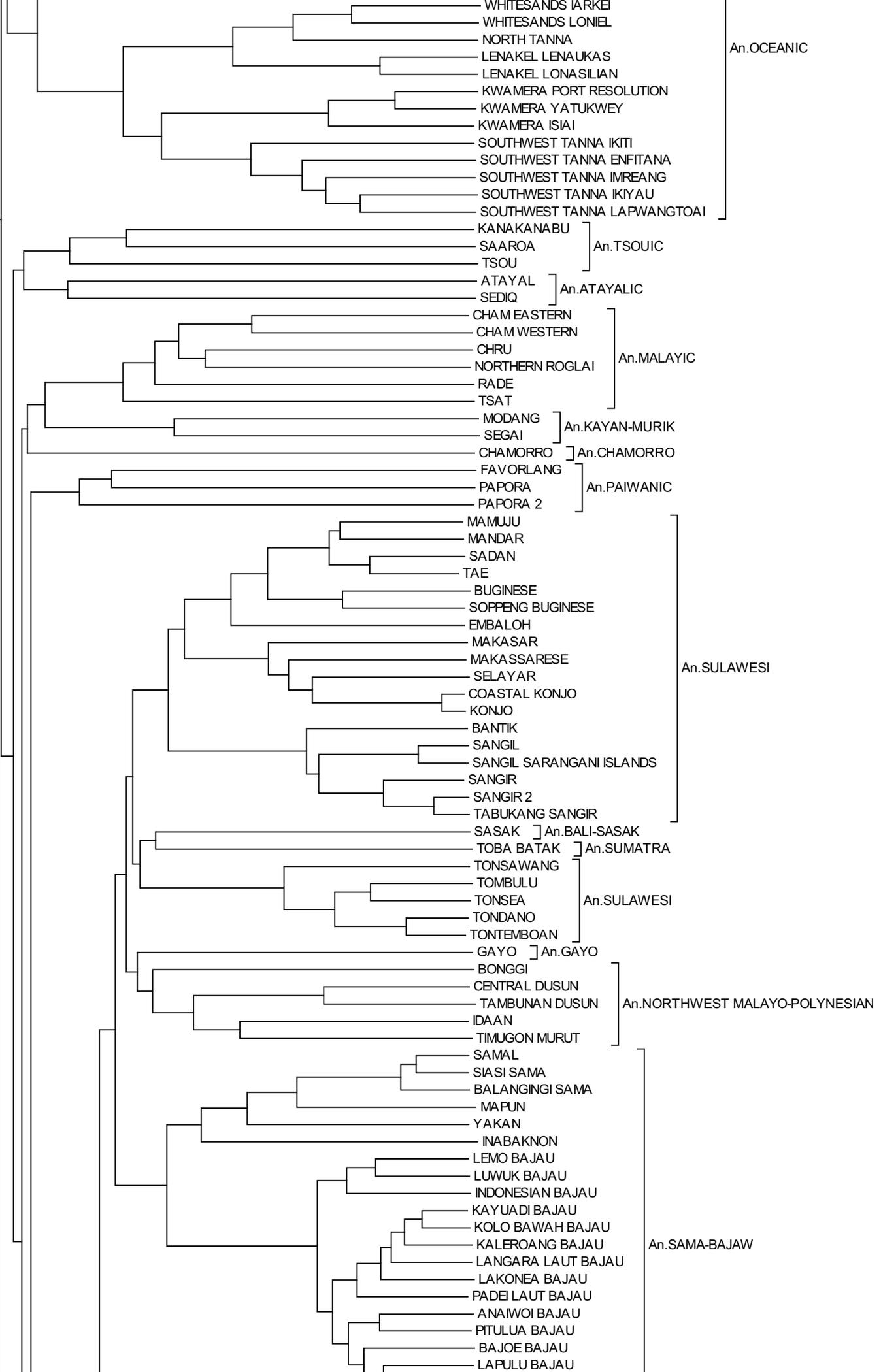
Sep	Sepik	War	Warao
Sho	Shom Peng	Was	Wasi
Sio	Siouan	WBg	West Bougainville
Sko	Sko	WBm	West Bomberai
Sln	Salinan	WF	Western Fly
Slv	Sáliban	WP	West Papuan
Snt	Sentani	Wsh	Washo
ST	Sino-Tibetan	WY	Wappo-Yukian
Tac	Tacanan	Xin	Xincan
Tak	Takelma	Yal	Yale
Tar	Tarascan	Yam	Yamana
Tau	Taushiro	Yan	Yanomam
Teb	Teberan-Pawaiian	Yaw	Yawa
Teq	Tequistlatecan	Yel	Yele
Tic	Ticuna	Yen	Yeniseian
Tim	Timucua	Yka	Yukaghir
TK	Tai-Kadai	Yrb	Yareban
TNG	Trans-New Guinea	Yrr	Yaruro
TO	Tor-Orya	Yua	Yuat
Tof	Tofanma	Yuc	Yuchi
Tol	Tol	Yur	Yuracare
Ton	Tonkawa	Yuw	Yuwana
Tor	Torricelli	Zam	Zamucoan
Tot	Totonacan	Zap	Zaparoan
Tou	Touo	Zun	Zuni
Tru	Trumai		
Tuc	Tucanoan		
TuK	Turama-Kikorian		
Tup	Tupian		
UA	Uto-Aztecan		
UC	Uru-Chipaya		
Un	Unknown		
Ura	Uralic		
Urr	Urarina		
Usk	Usku		
UY	Upper Yuat		
VJ	Vaupés-Japurá		
Wak	Wakashan		
Wao	Waorani		

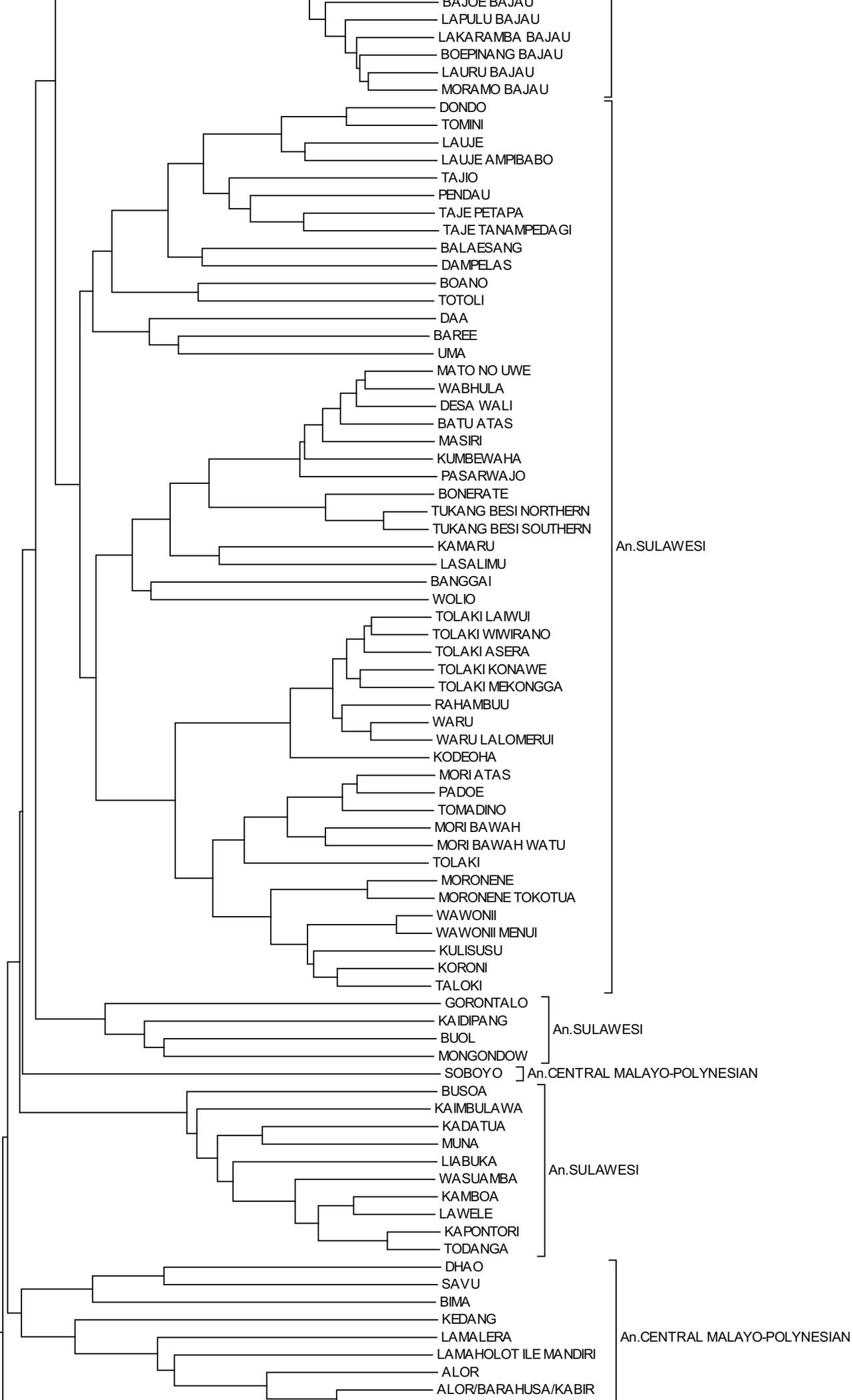












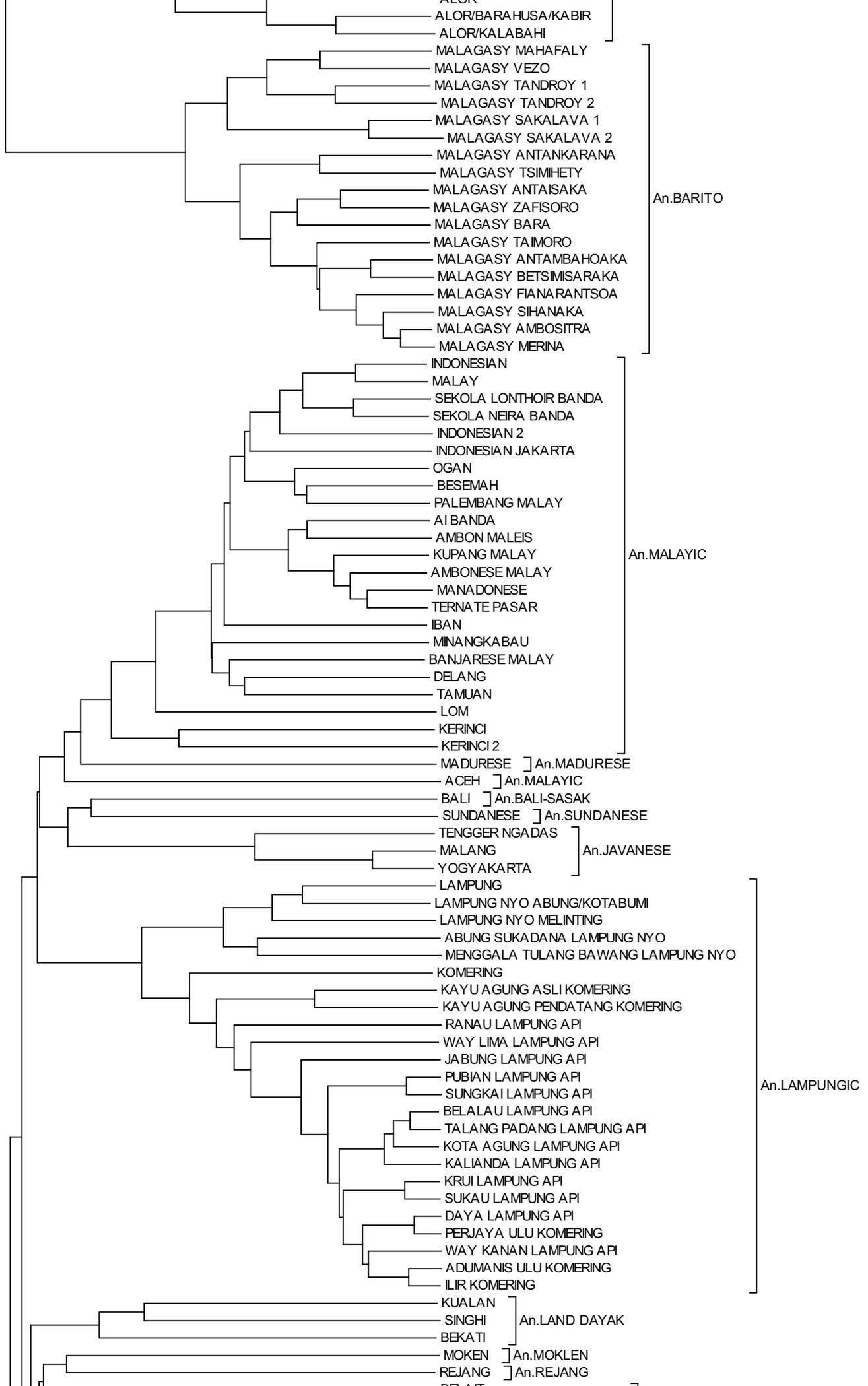
An.SULAWESI

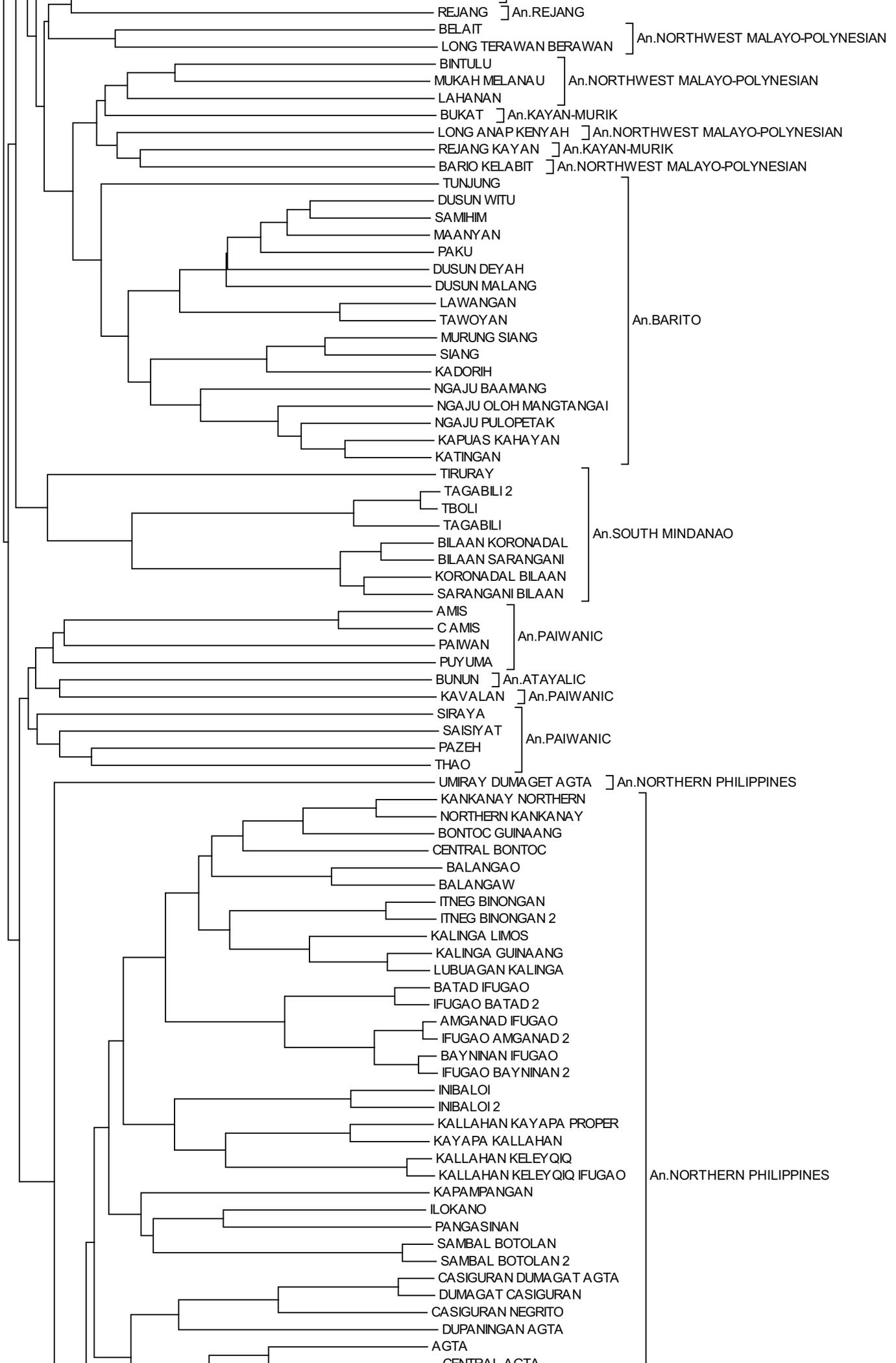
An.SULAWESI

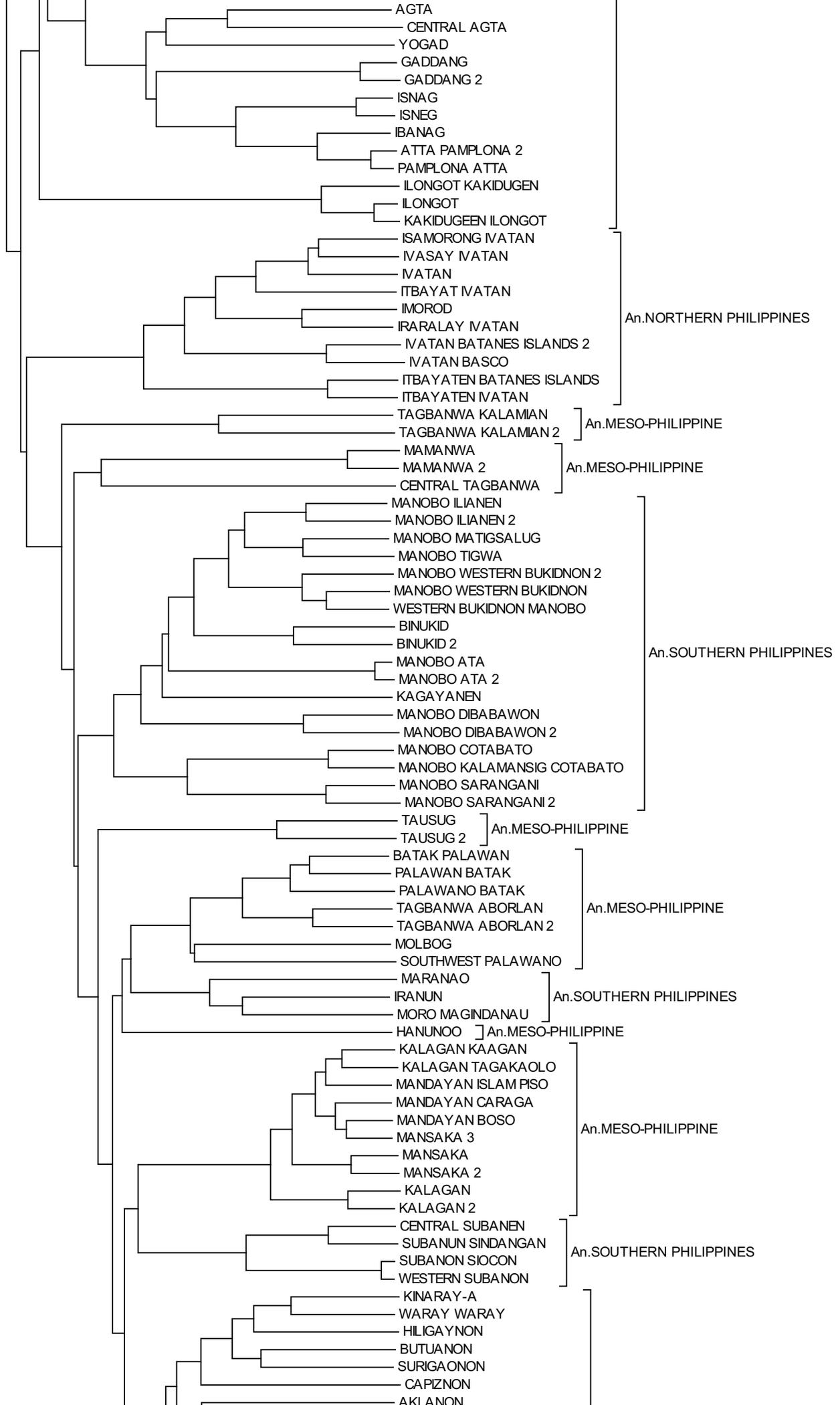
An.CENTRAL MALAYO-POLYNESIAN

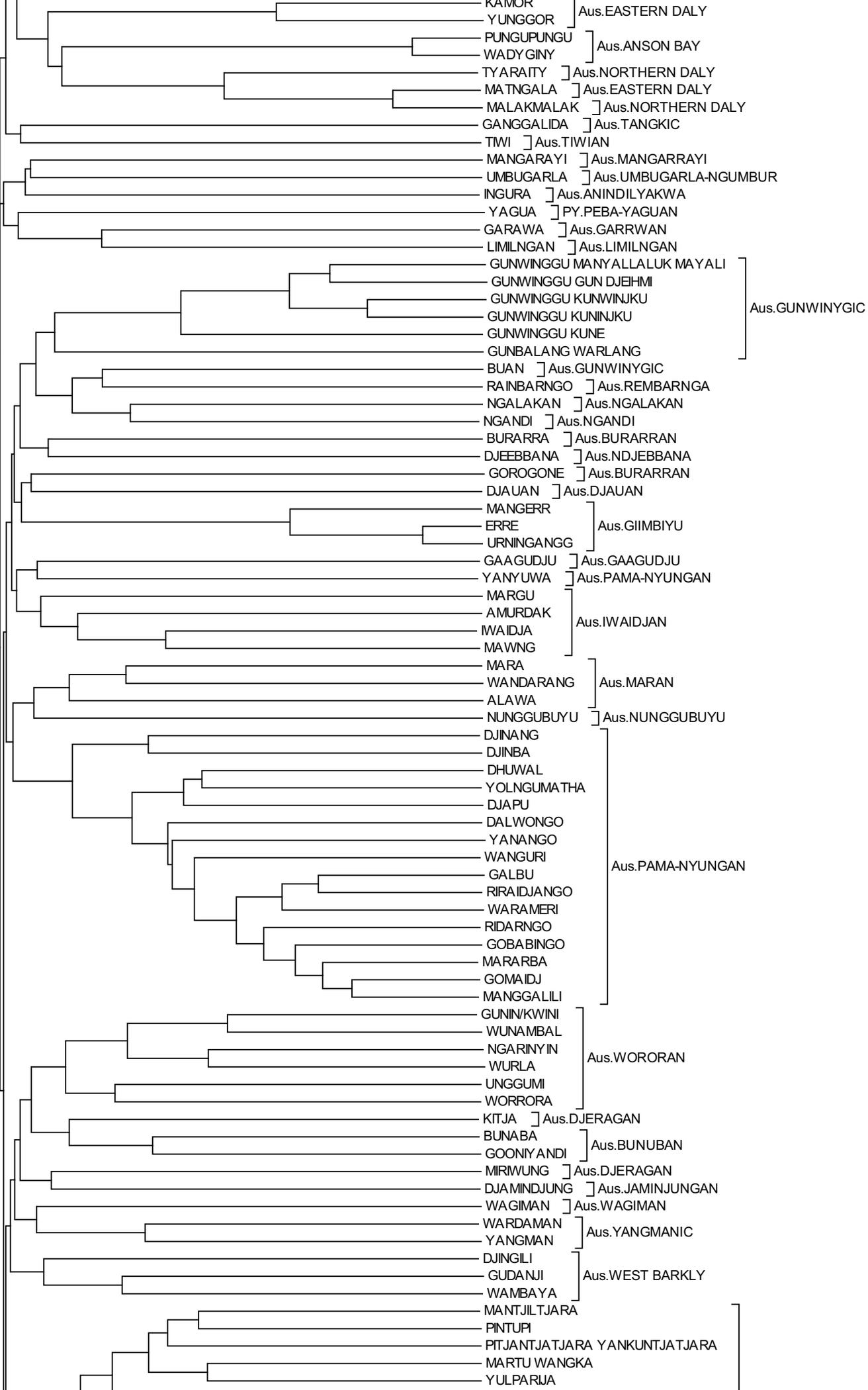
An.SULAWESI

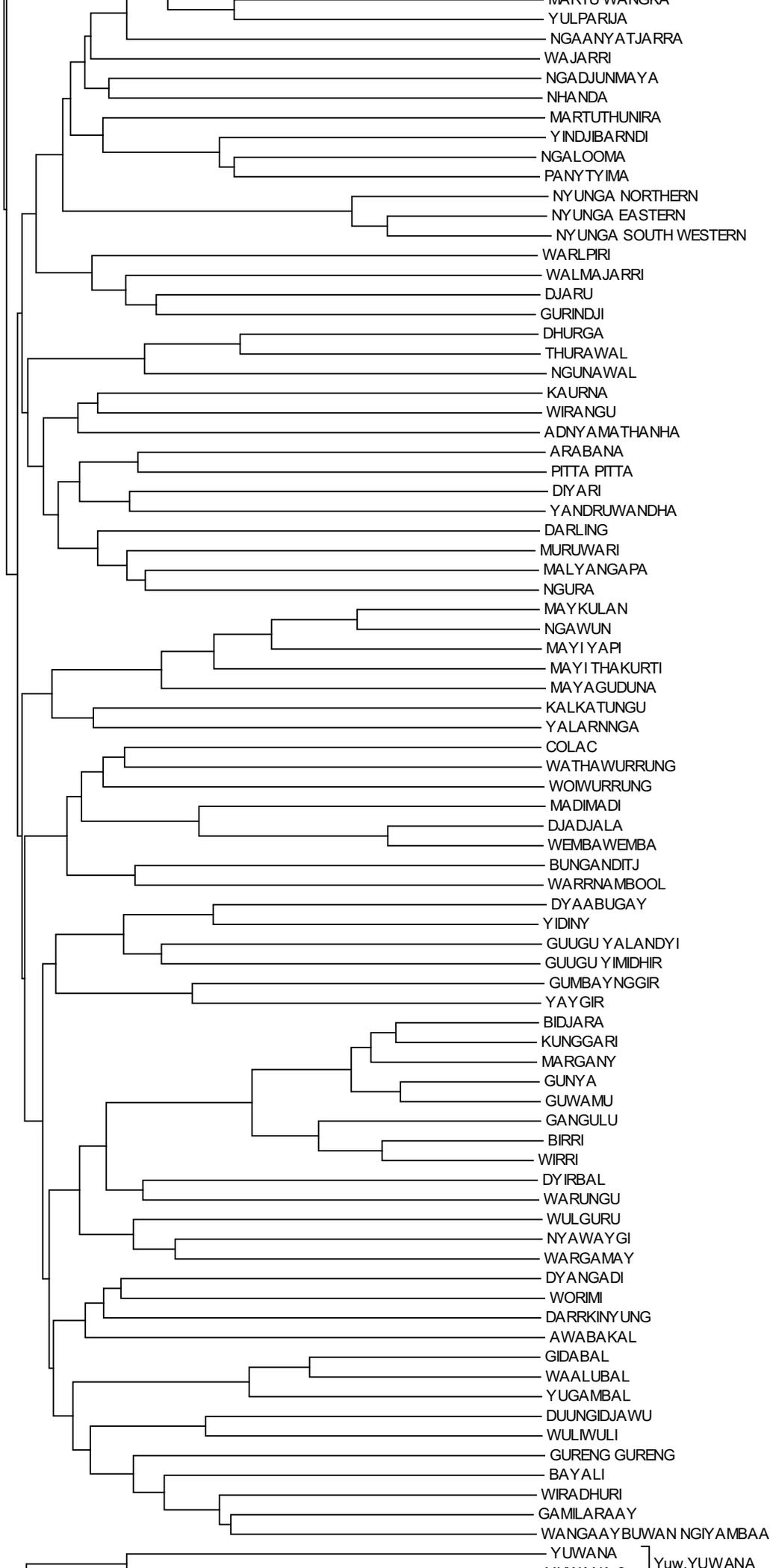
An.CENTRAL MALAYO-POLYNESIAN





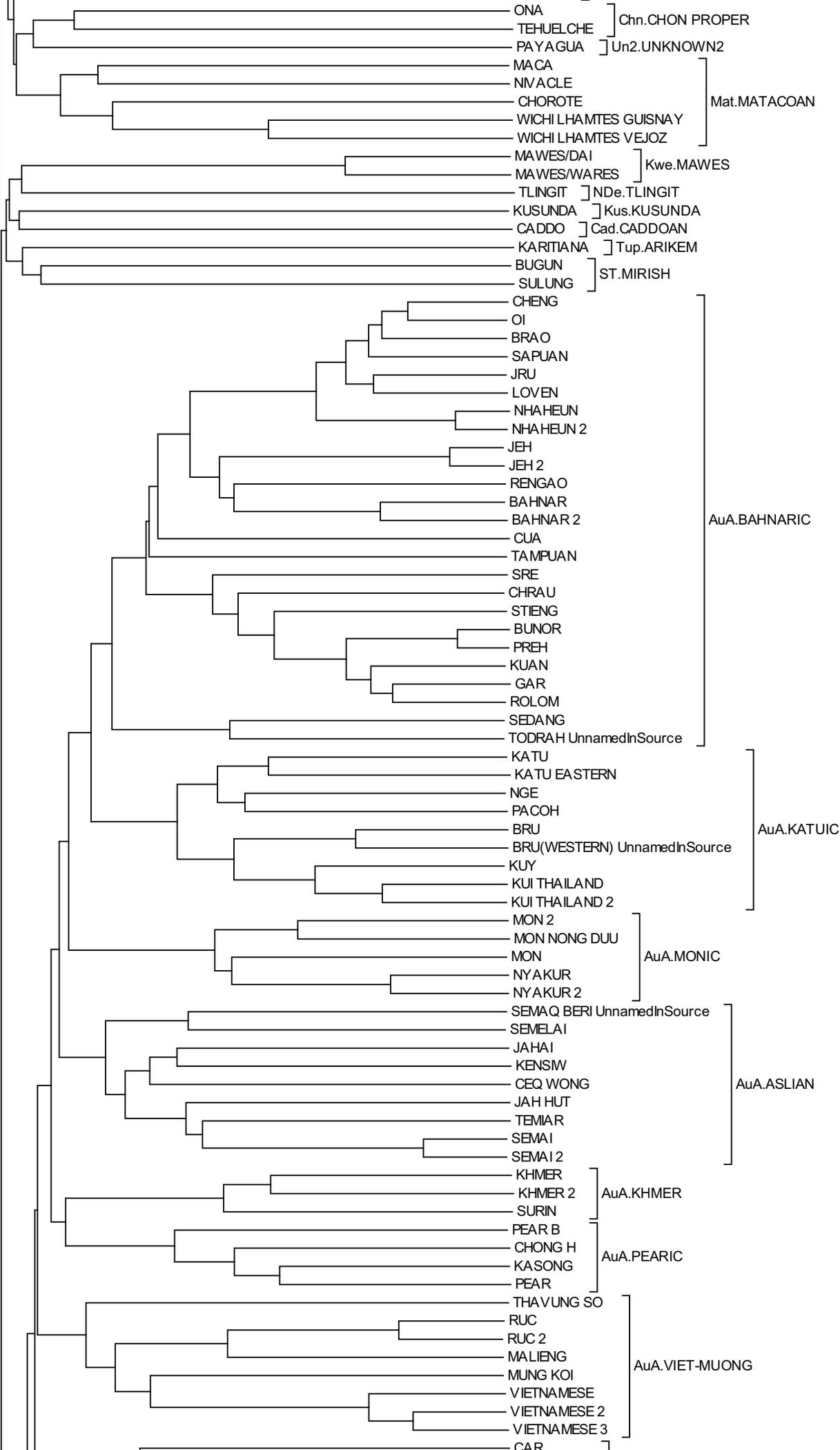


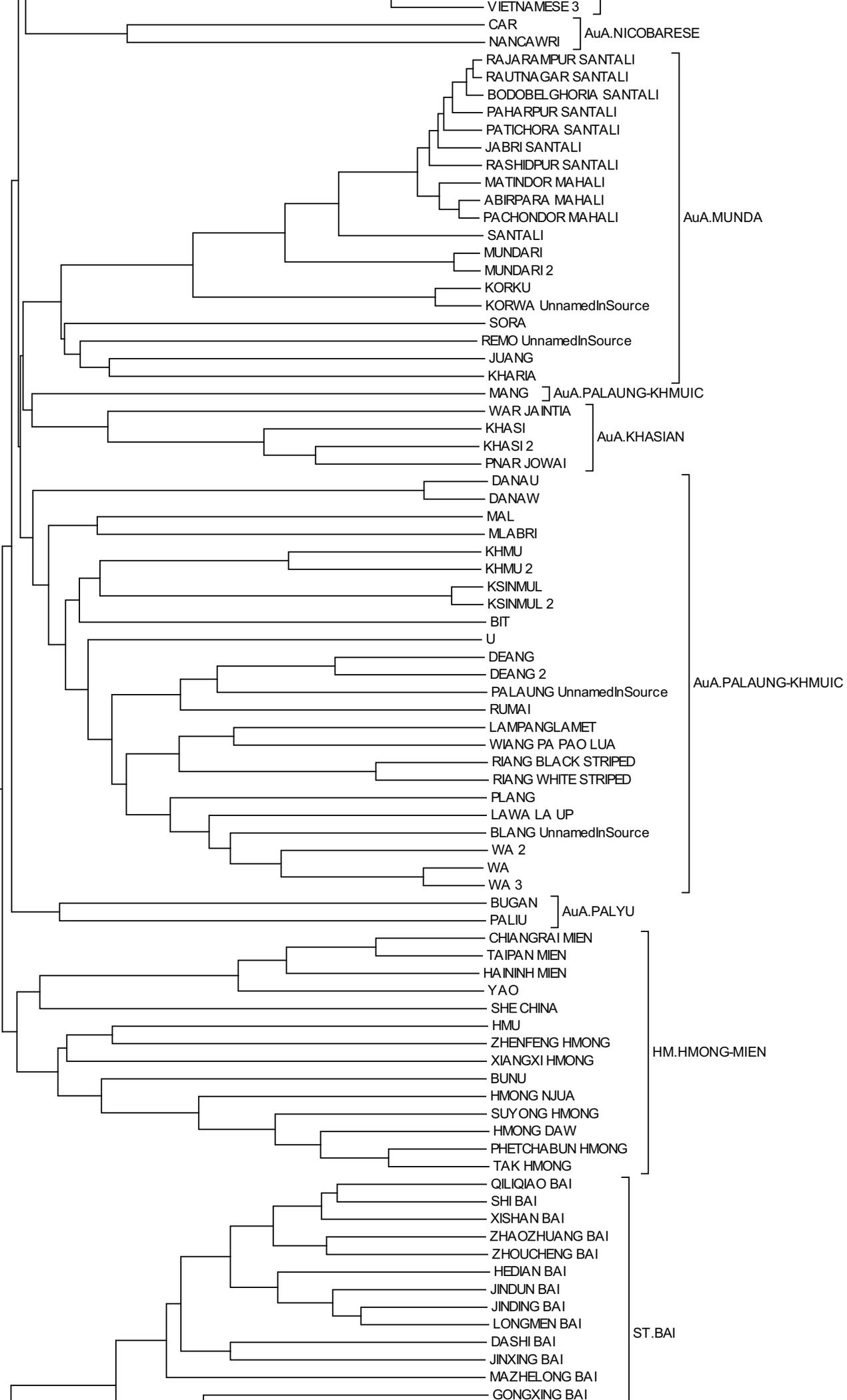


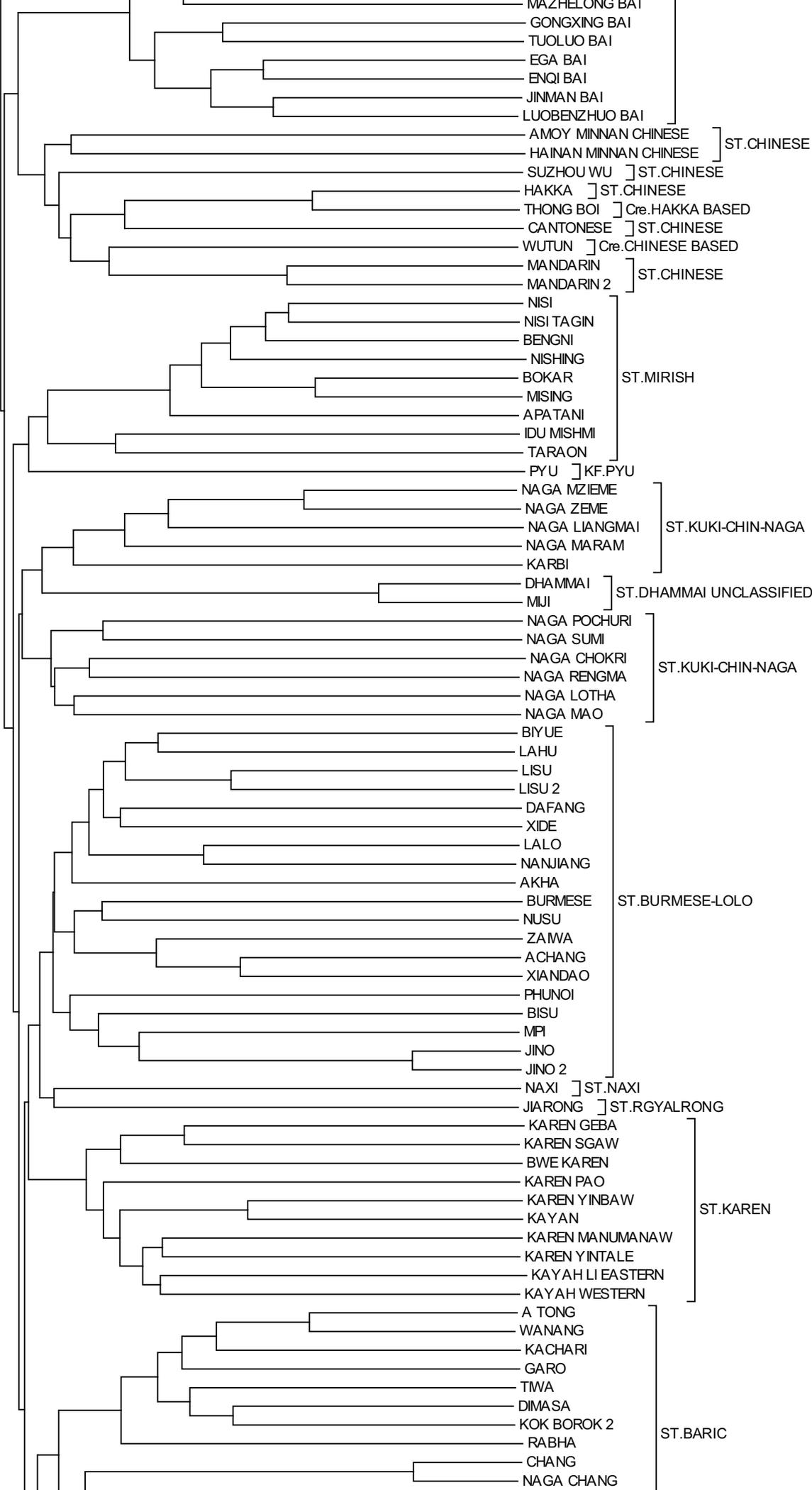


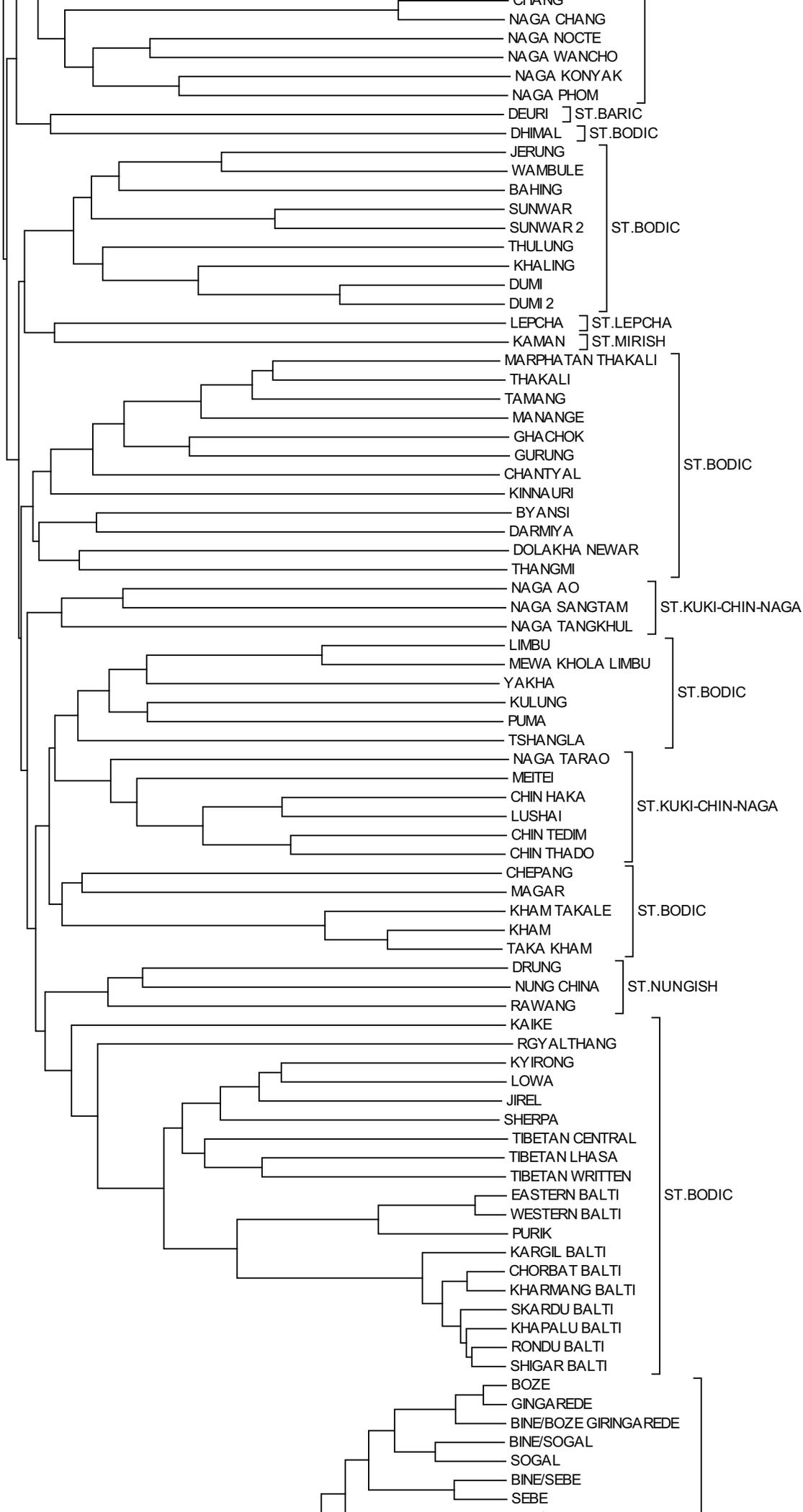
Aus.PAMA-NYUNGAN

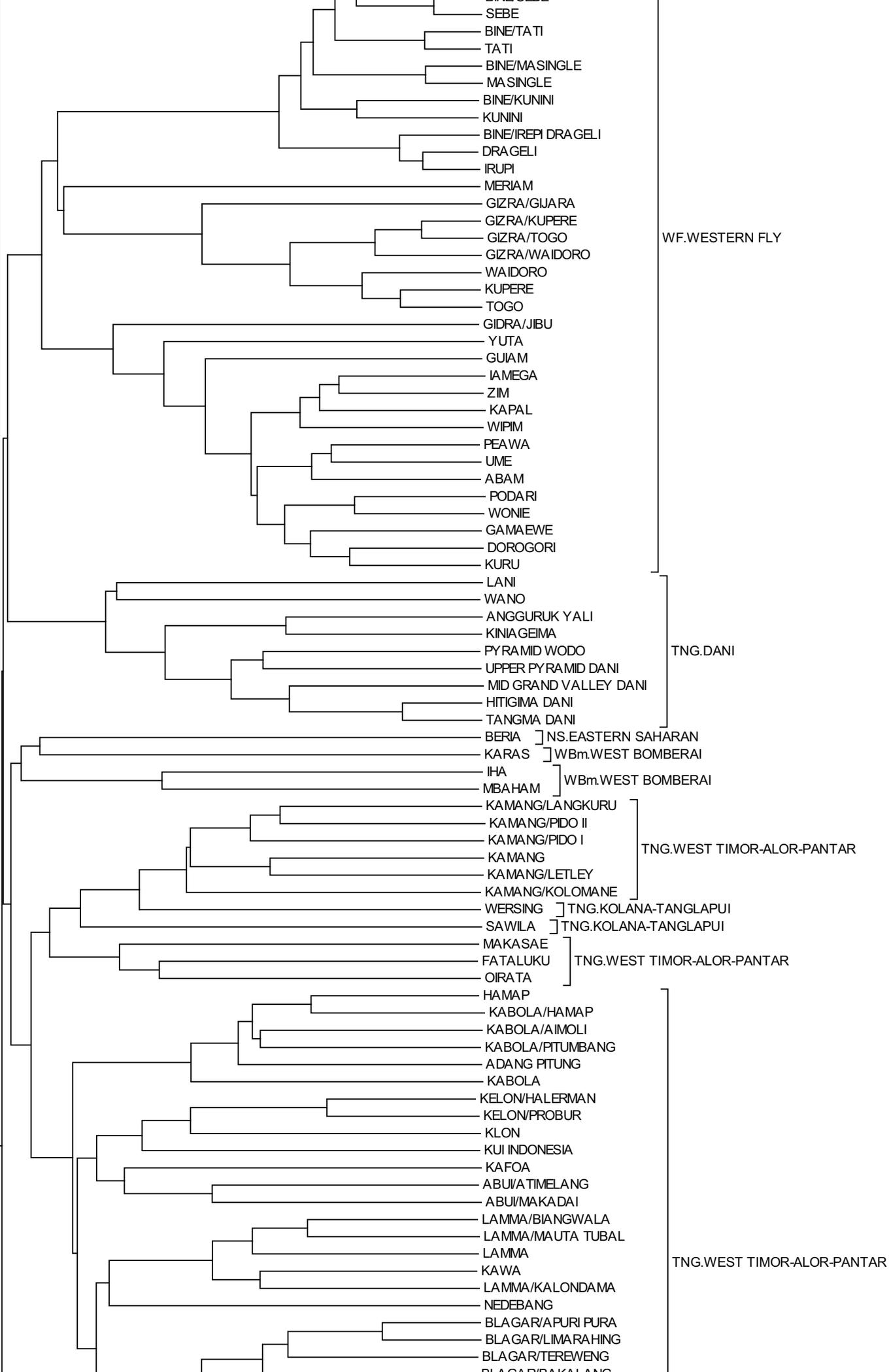
Yuw.YUWANA

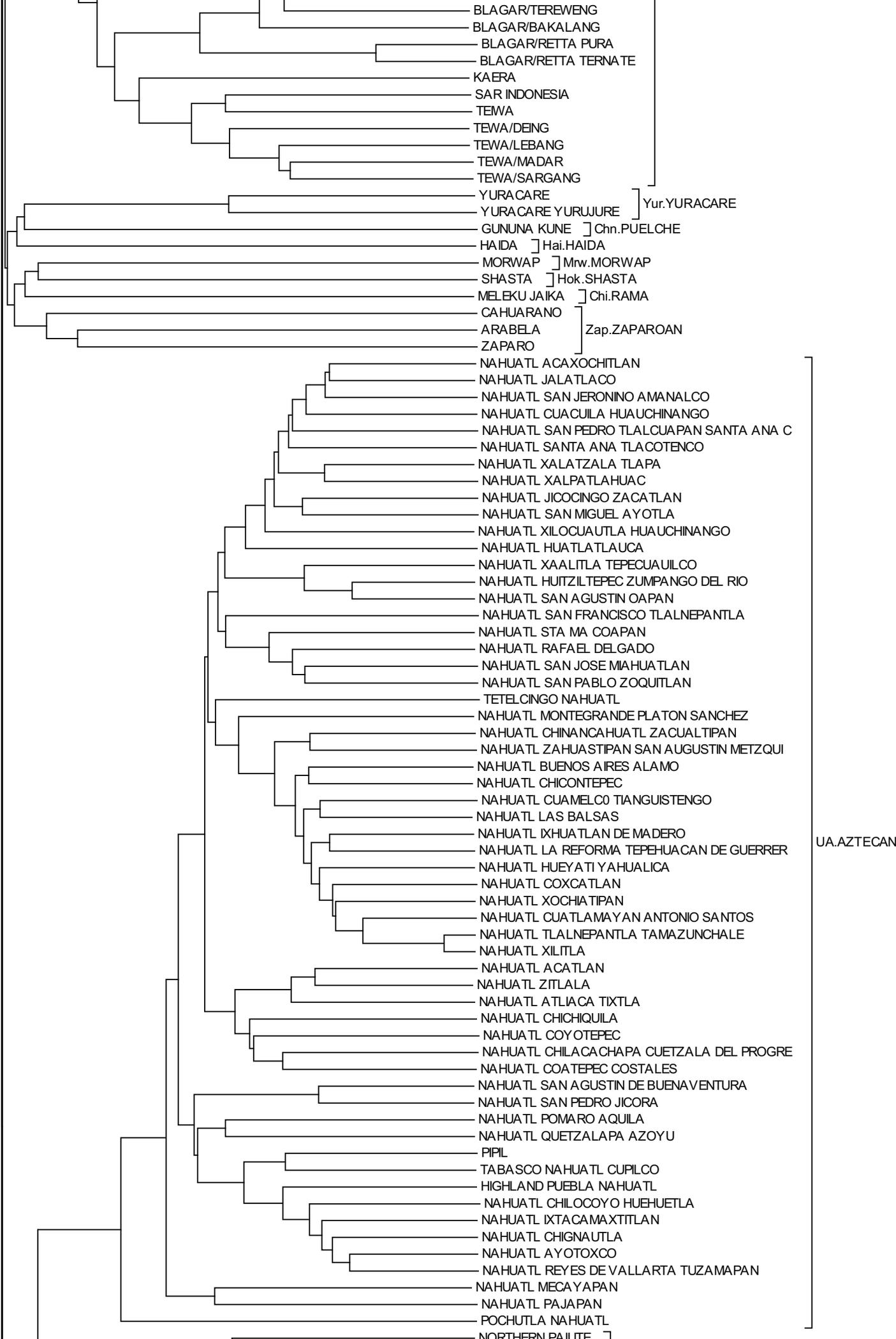




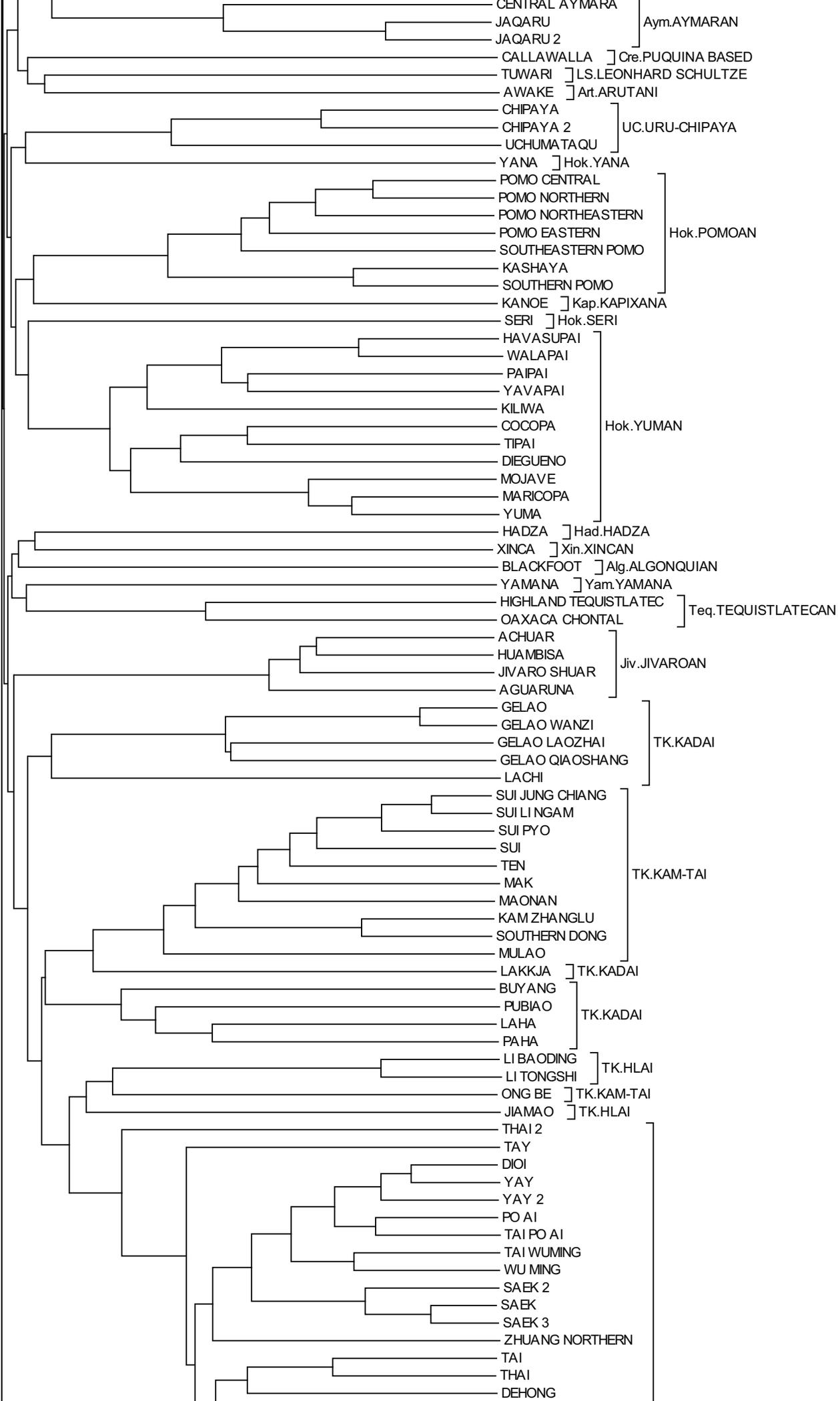


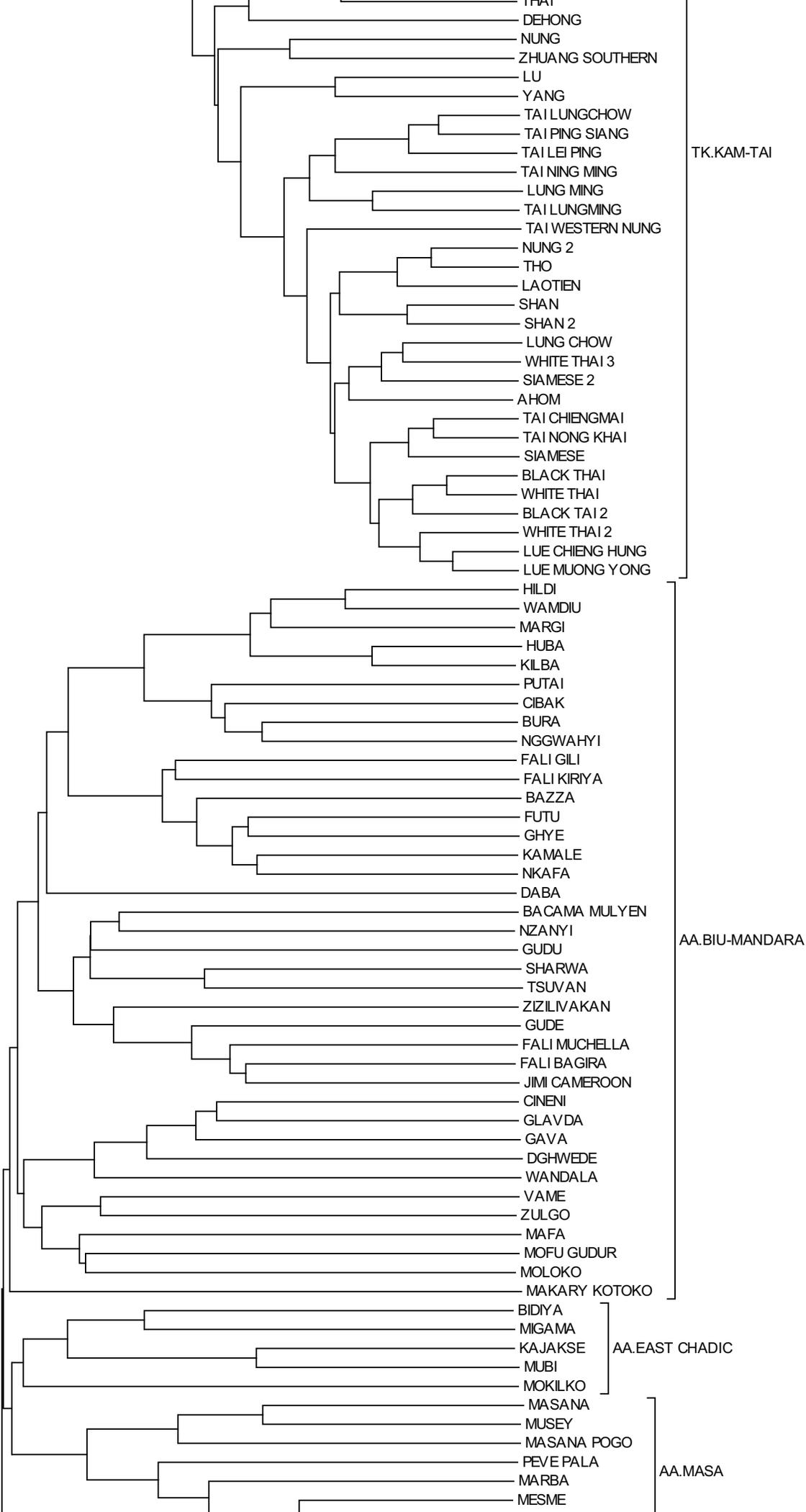


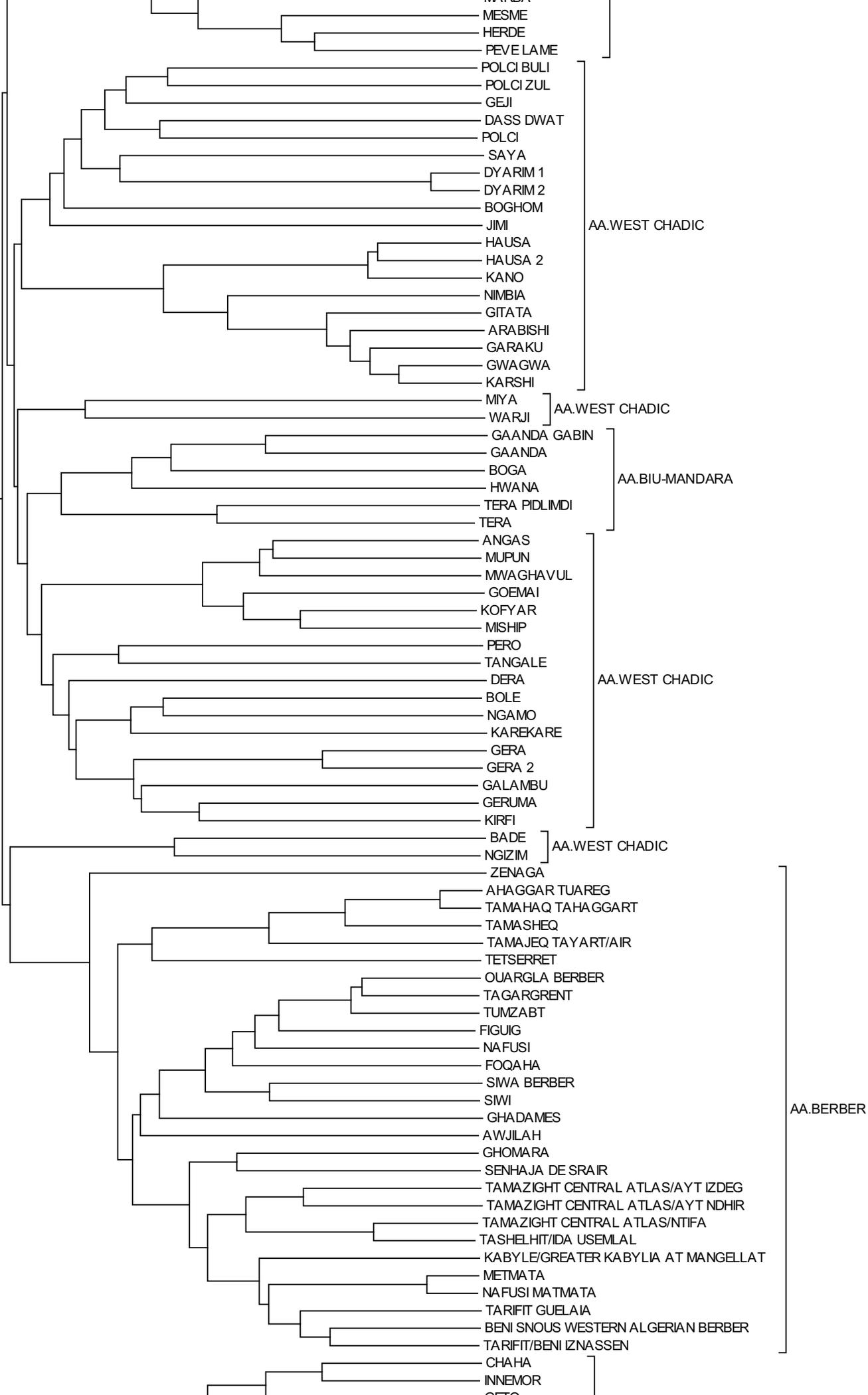












AA.WEST CHADIC

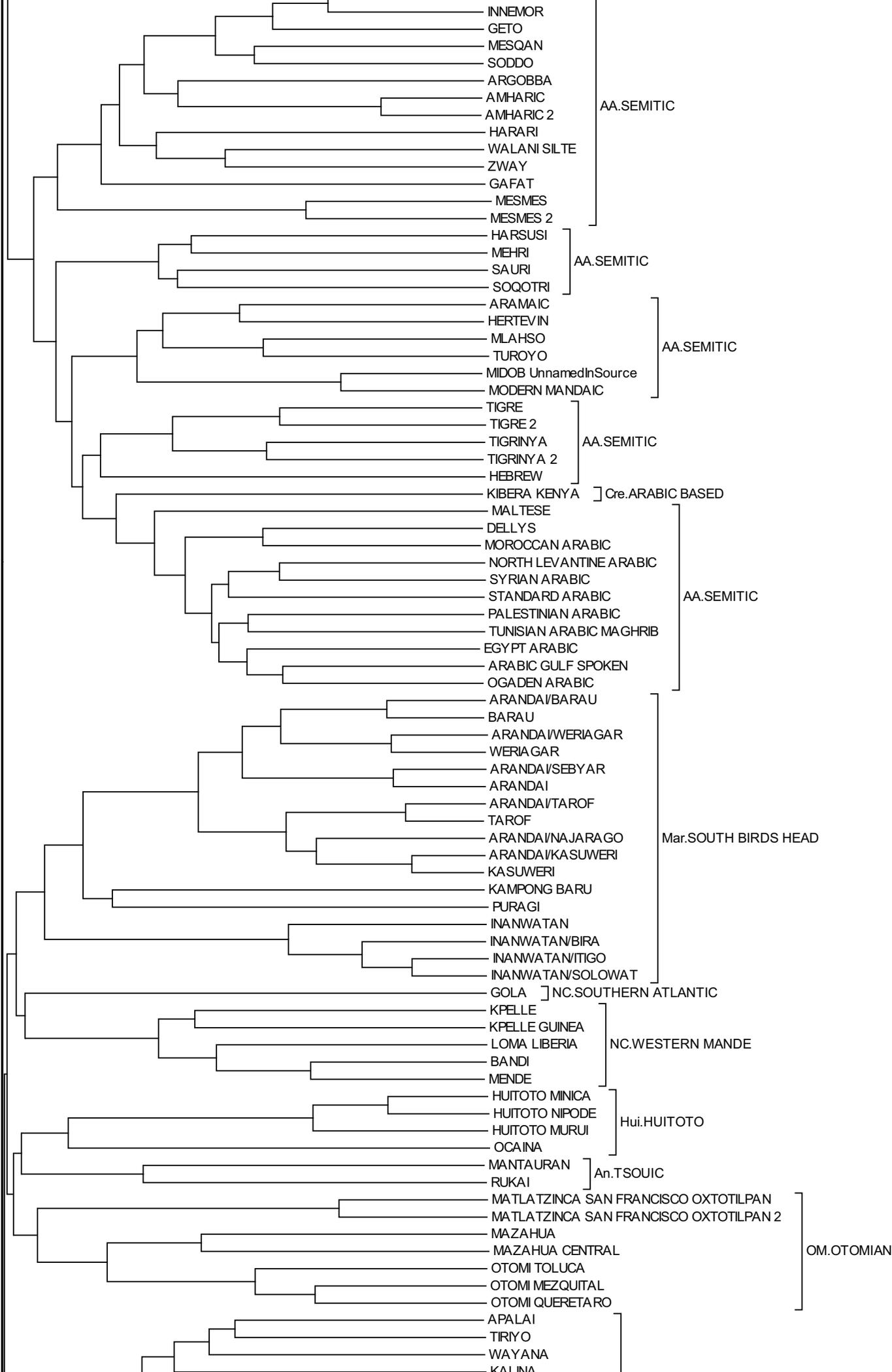
AA.WEST CHADIC

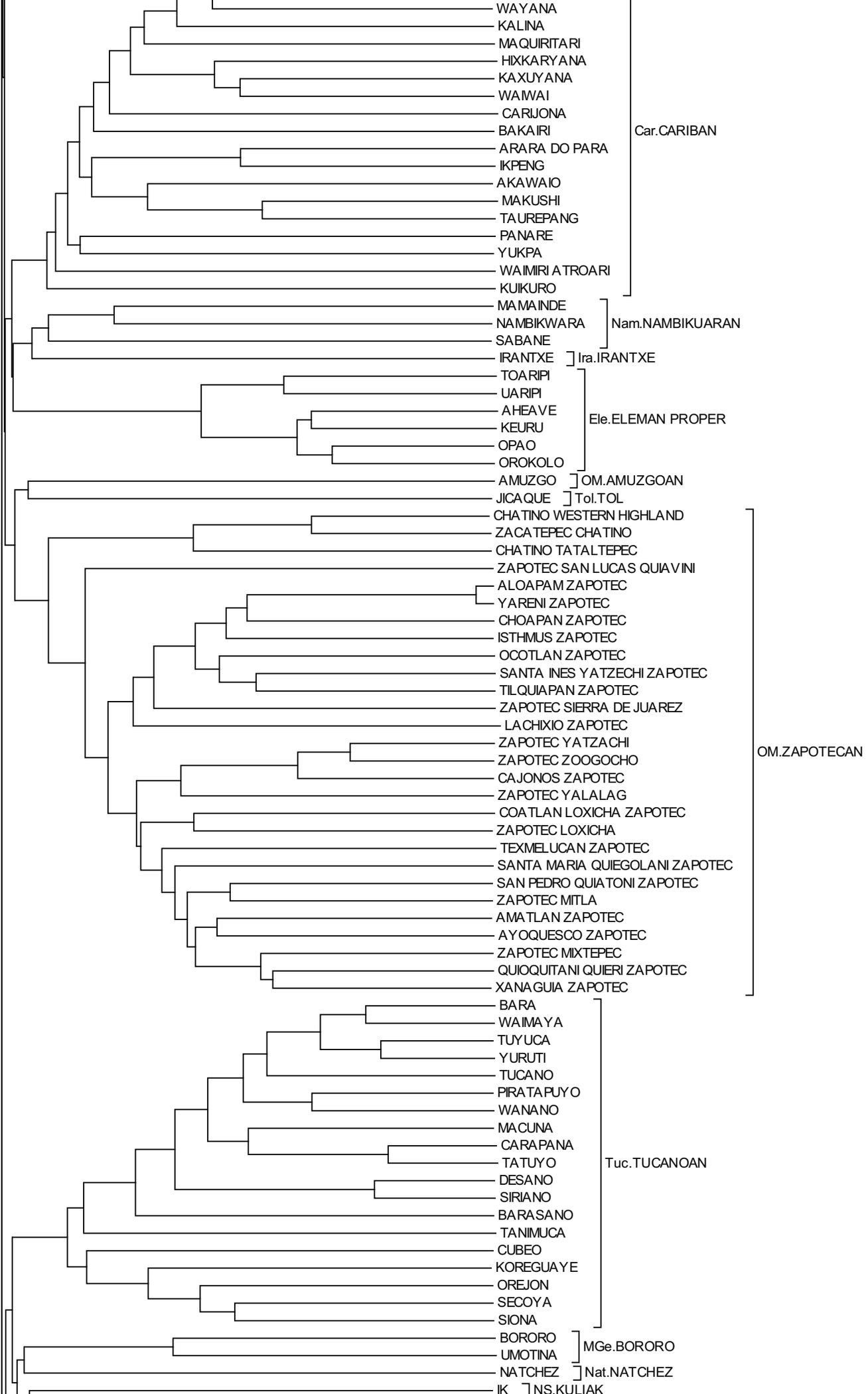
AA.BIU-MANDARA

AA.WEST CHADIC

AA.WEST CHADIC

AA.BERBER





WAYANA

KALINA

MAQUIRITARI

HIXKARYANA

KAXUYANA

WAIWAI

CARJONA

BAKAIRI

ARARA DO PARA

KPENG

AKAWAIO

MAKUSHI

TAUREPANG

PANARE

YUKPA

WAIMIRI A TROARI

KUIKURO

MAMA INDE

NAMBIKWARA

SABANE

IRANTXE

TOARIPI

UARIPI

AHEAVE

KEURU

OPAO

OROKOLO

AMUZGO

JICAQUE

CHATINO WESTERN HIGHLAND

ZACA TEPEC CHATINO

CHATINO TATALTEPEC

ZAPOTEC SAN LUCAS QUIAVINI

ALOAPAM ZAPOTEC

YARENI ZAPOTEC

CHOAPAN ZAPOTEC

ISTHMUS ZAPOTEC

OCOTLAN ZAPOTEC

SANTA INES YATZECHI ZAPOTEC

TILQUIAPAN ZAPOTEC

ZAPOTEC SIERRA DE JUAREZ

LACHIXIO ZAPOTEC

ZAPOTEC YATZACHI

ZAPOTEC ZOOGOCHO

CAJONOS ZAPOTEC

ZAPOTEC YALALAG

COATLAN LOXICHA ZAPOTEC

ZAPOTEC LOXICHA

TEXMELUCAN ZAPOTEC

SANTA MARIA QUIEGOLANI ZAPOTEC

SAN PEDRO QUIATONI ZAPOTEC

ZAPOTEC MITLA

AMATLAN ZAPOTEC

AYOQUESCO ZAPOTEC

ZAPOTEC MIXTEPEC

QUIOQUITANI QUIERI ZAPOTEC

XANAGUIA ZAPOTEC

BARA

WAIMAYA

TUYUCA

YURUTI

TUCANO

PIRATA PUYO

WANANO

MACUNA

CARAPANA

TATUYO

DESANO

SIRIANO

BARASANO

TANIMUCA

CUBEO

KOREGUA YE

OREJON

SECOYA

SIONA

BORORO

UMOTINA

NATCHEZ

IK

Car. CARIBAN

Nam. NAMBIKUIARAN

Ira. IRANTXE

Ele. ELEMEN PROPER

OM. AMUZGOAN

Tol. TOL

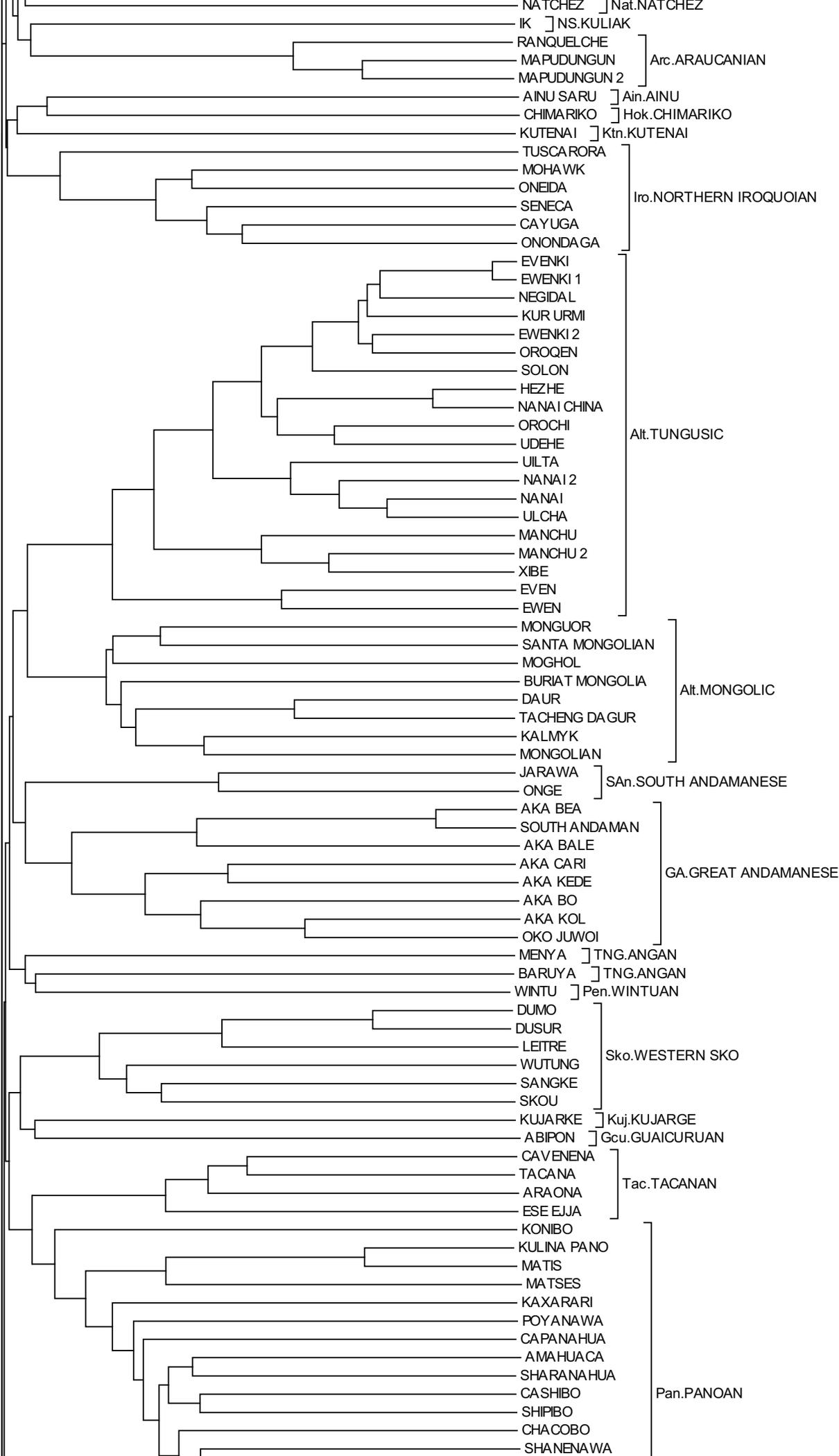
OM. ZAPOTECAN

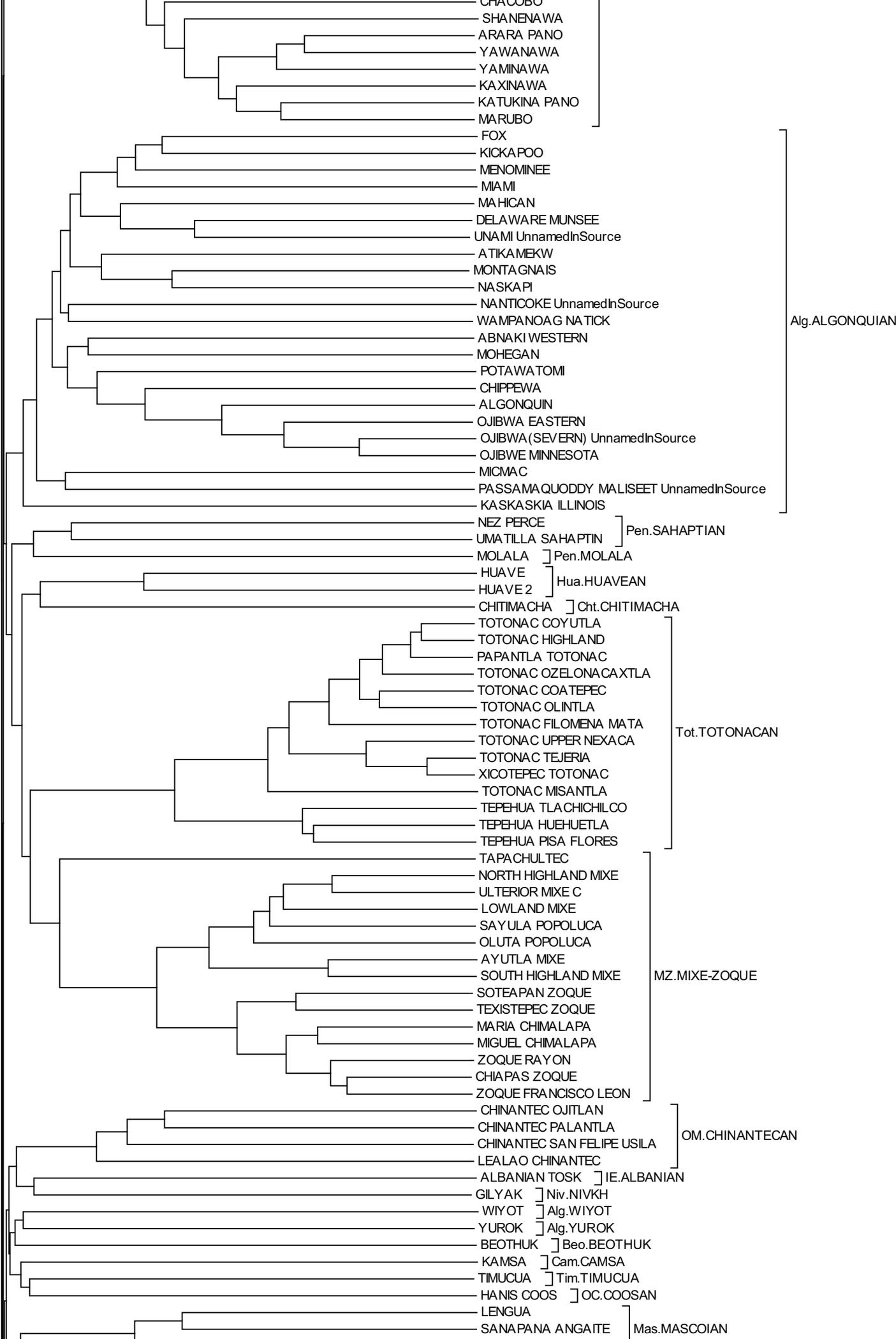
Tuc. TUCANOAN

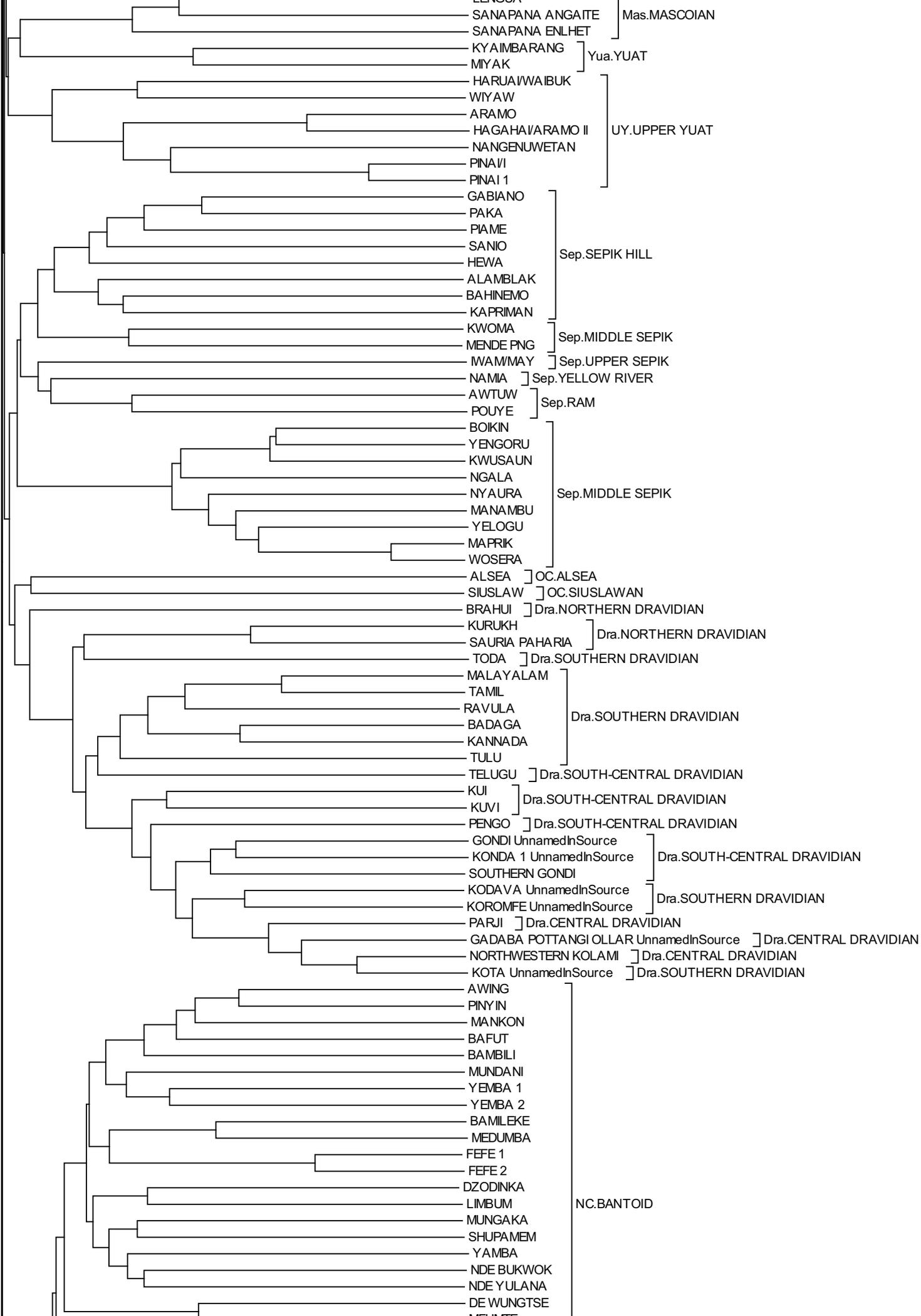
MGe. BORORO

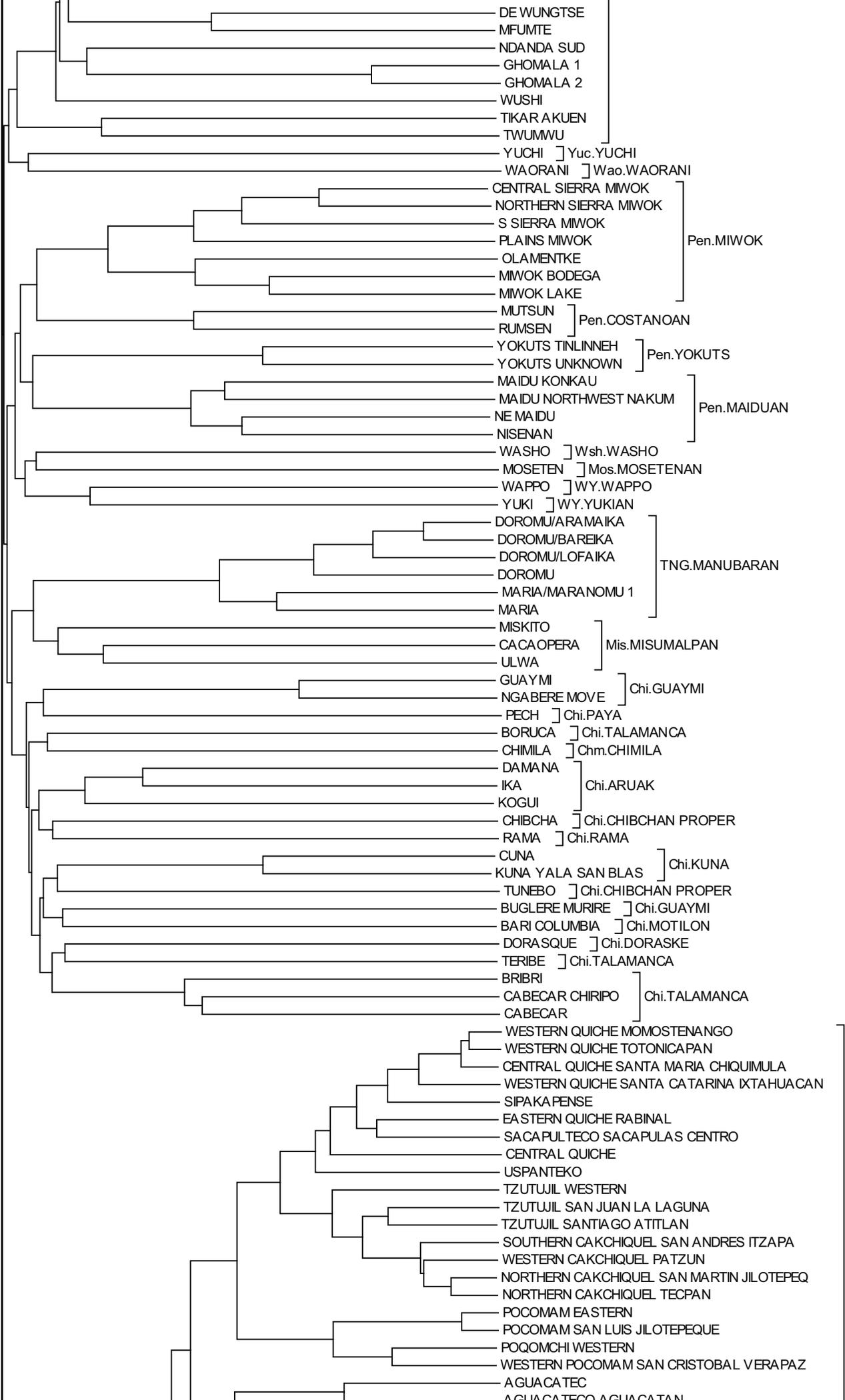
Nat. NATCHEZ

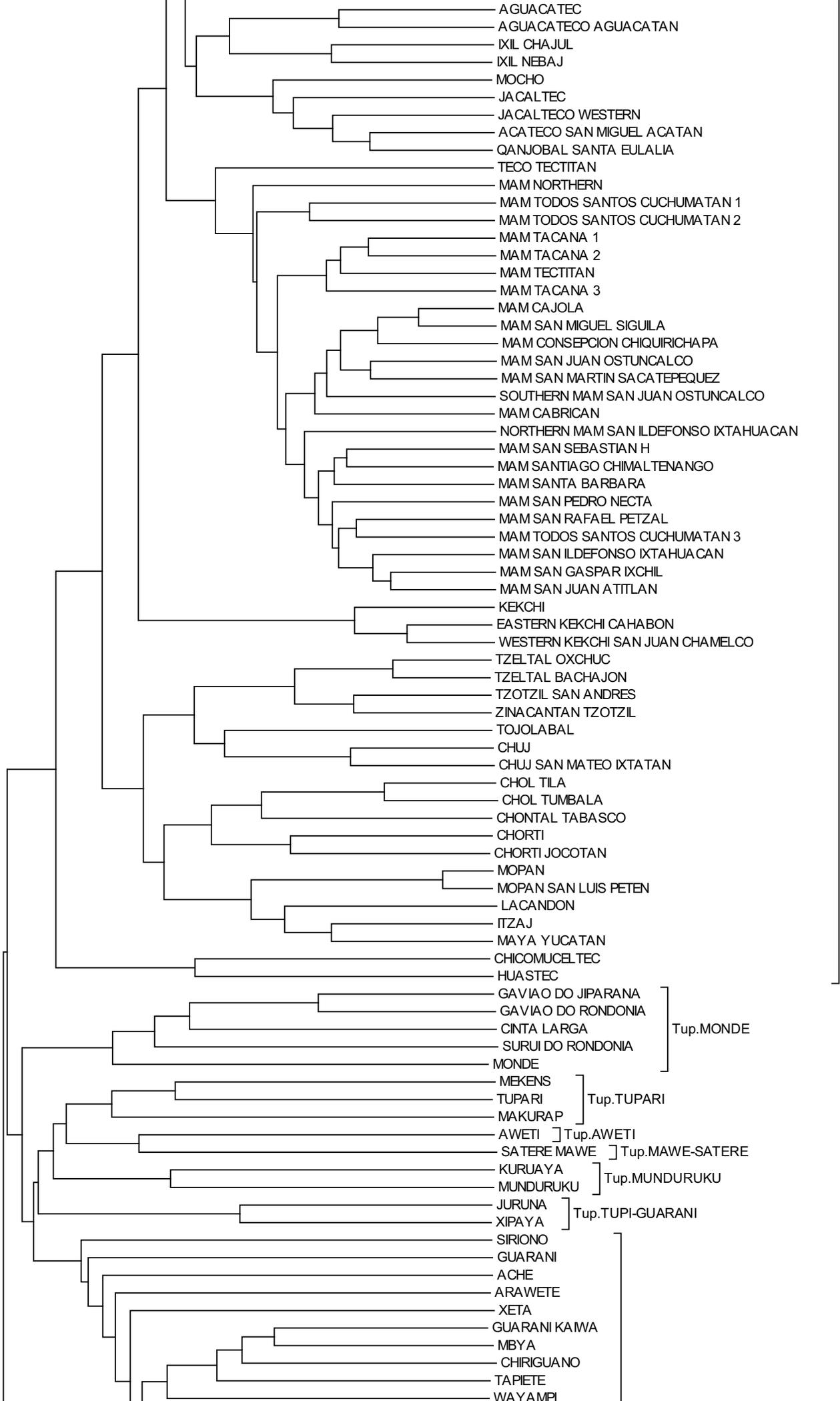
NS. KULIAK











May.MAYAN

Tup.MONDE

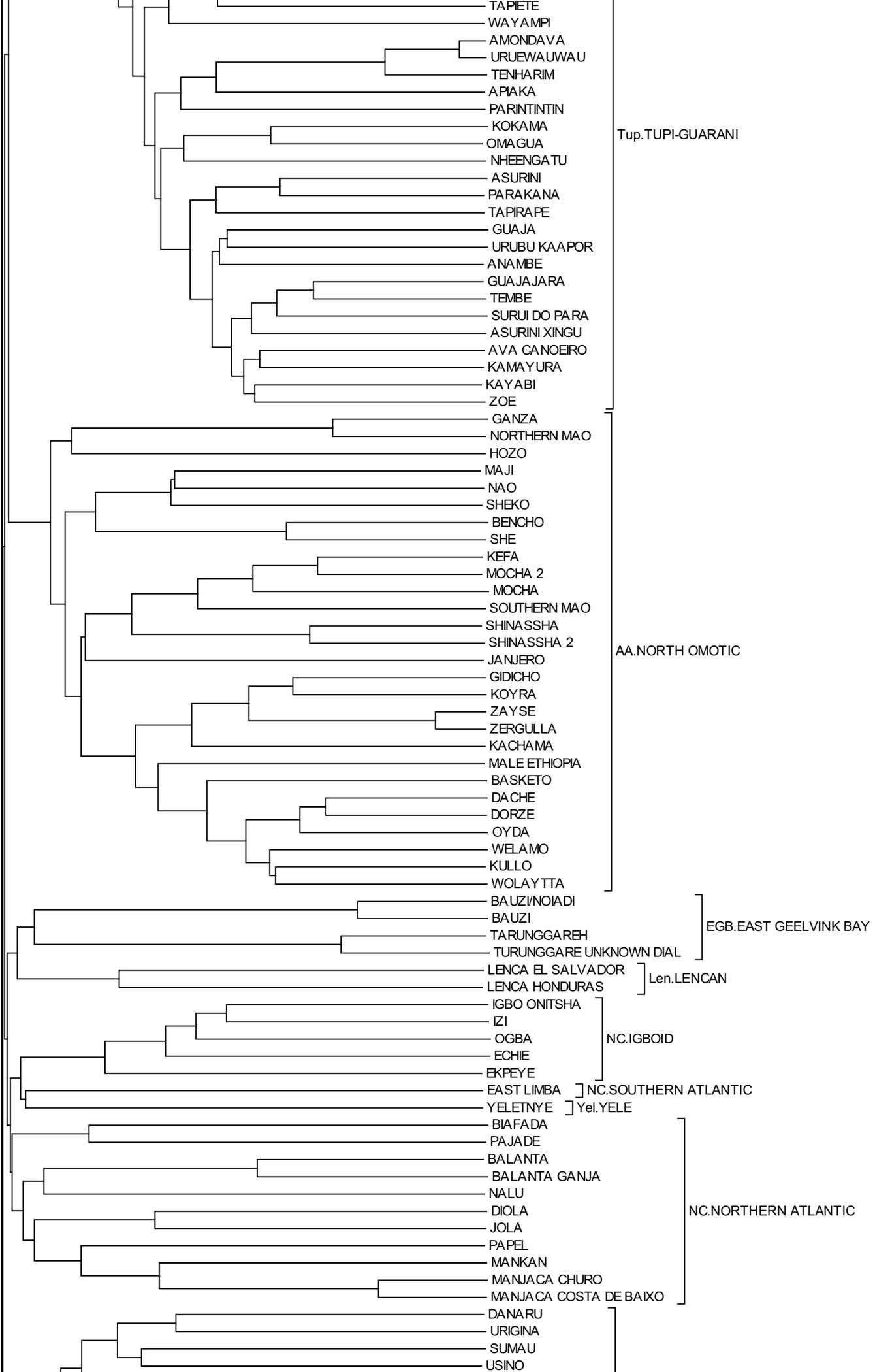
Tup.TUPARI

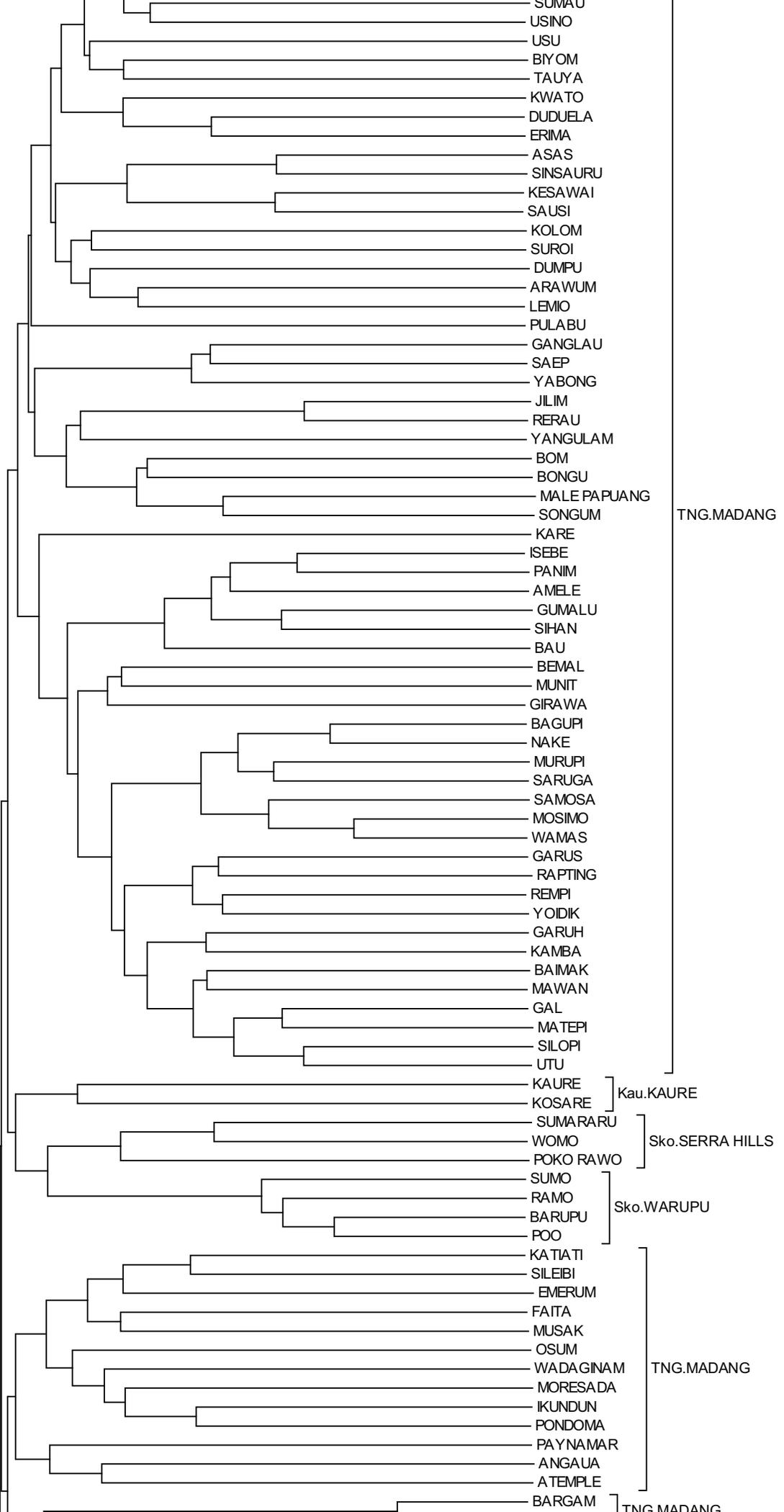
Tup.AWETI

Tup.MAWE-SATERE

Tup.MUNDURUKU

Tup.TUPI-GUARANI





SUMA

USINO

USU

BIYOM

TAUYA

KWATO

DUDUELA

ERIMA

ASAS

SINSAURU

KESAWAI

SAUSI

KOLOM

SUROI

DUMPU

ARAWUM

LEMIO

PULABU

GANGLAU

SAEP

YABONG

JILIM

RERAU

YANGULAM

BOM

BONGU

MALE PAPUANG

SONGUM

TNG.MADANG

KARE

ISEBE

PANIM

AMELE

GUMALU

SIHAN

BAU

BEMAL

MUNIT

GIRAWA

BAGUPI

NAKE

MURUPI

SARUGA

SAMOSI

MOSIMO

WAMAS

GARUS

RAPTING

REMPI

YOIDIK

GARUH

KAMBA

BAIMAK

MAWAN

GAL

MATEPI

SILOPI

UTU

KAURE

KOSARE

Kau.KAURE

SUMARARU

WOMO

POKO RAWO

Sko.SERRA HILLS

SUMO

RAMO

BARUPU

POO

Sko.WARUPU

KATIATI

SILEIBI

EMERUM

FAITA

MUSAK

OSUM

WADAGINAM

MORESADA

IKUNDUN

PONDOMA

PAYNAMAR

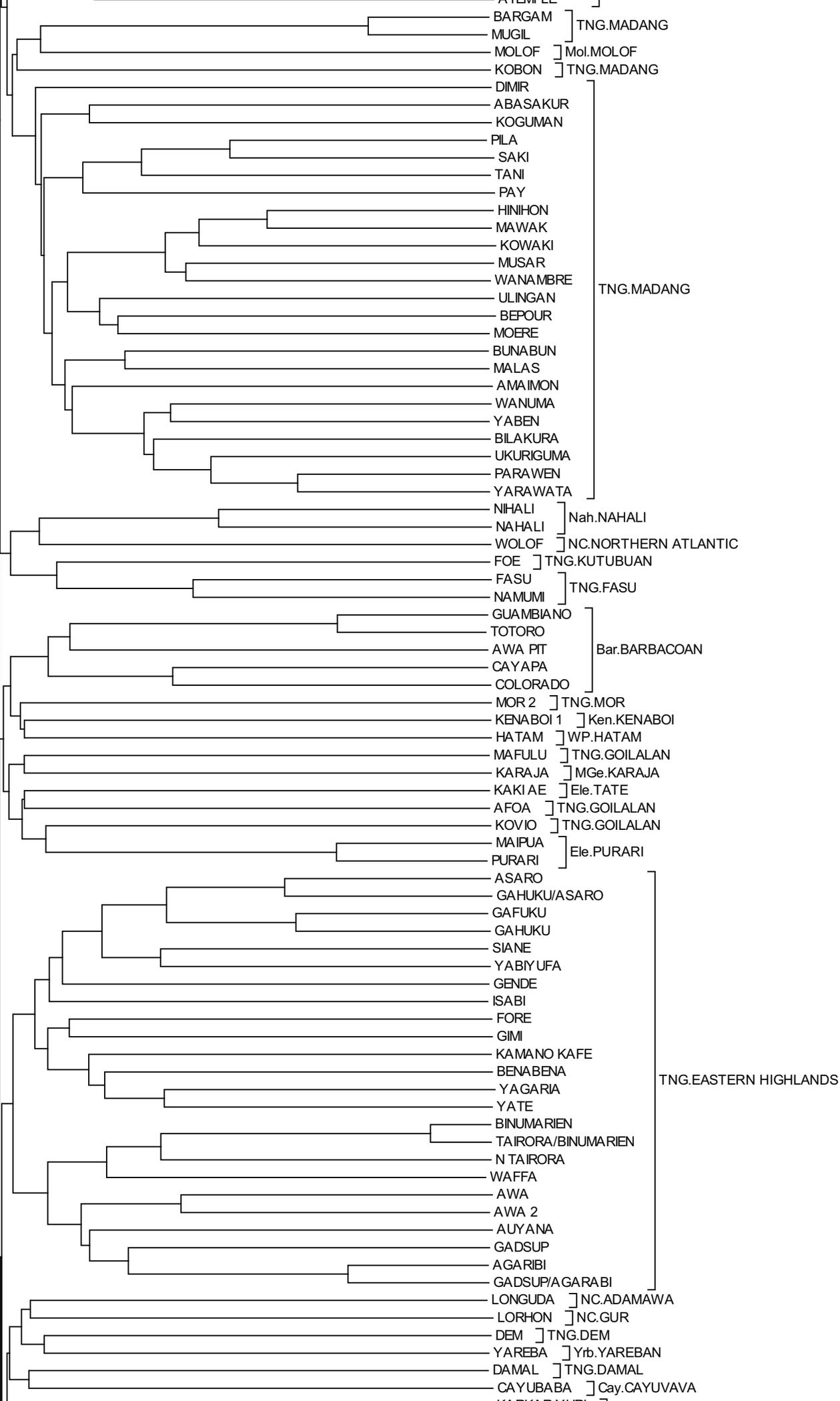
ANGAUA

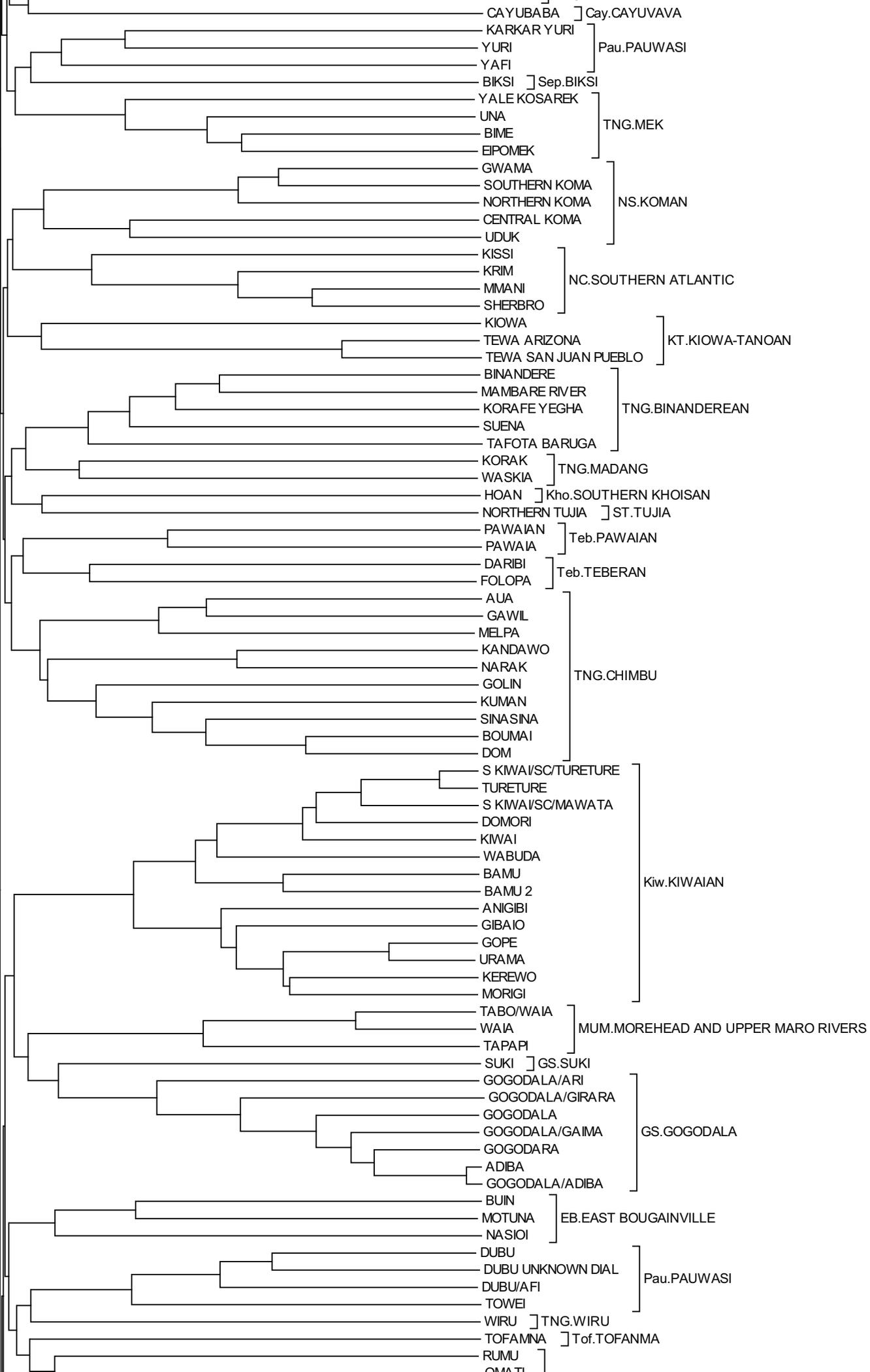
ATEMPLE

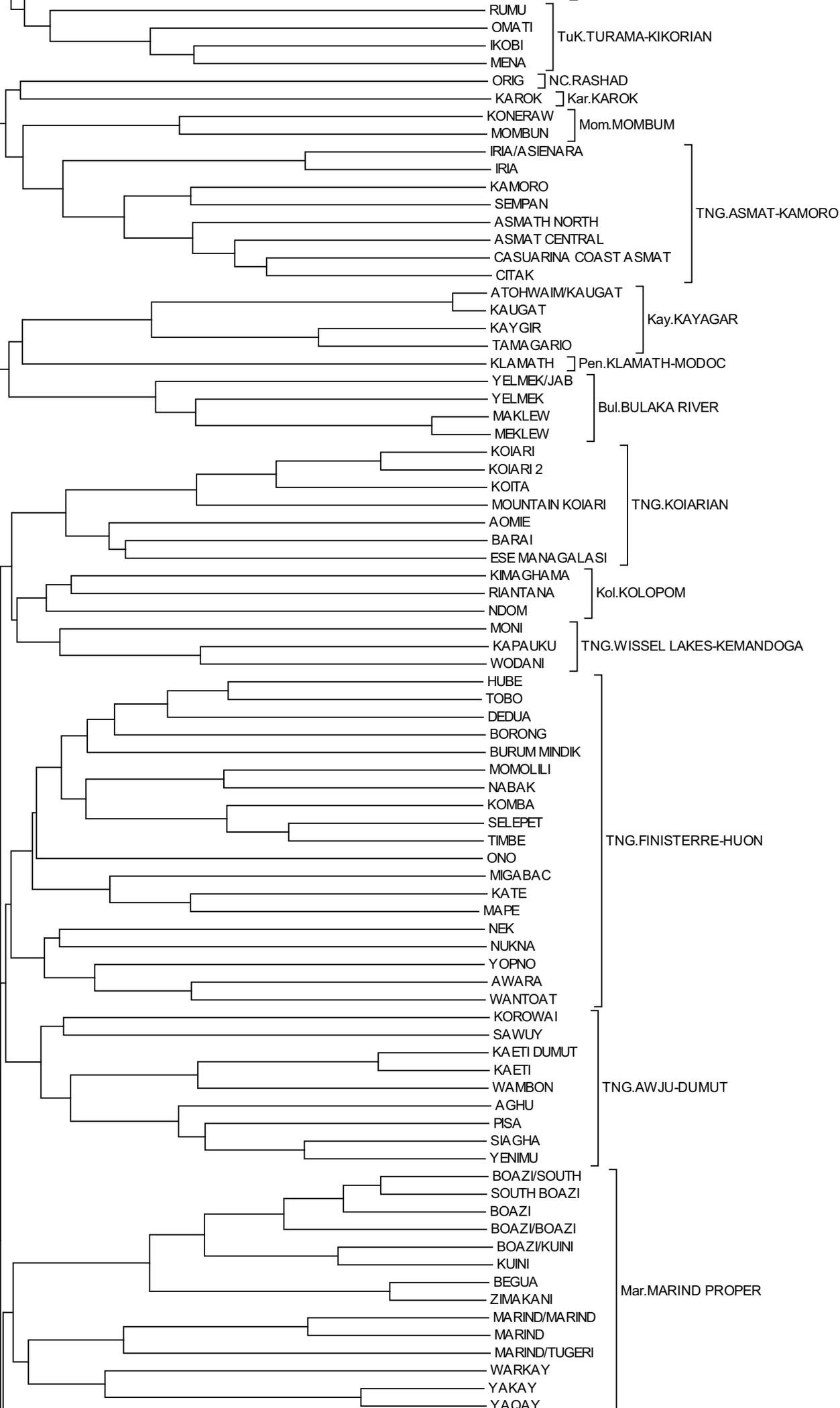
BARGAM

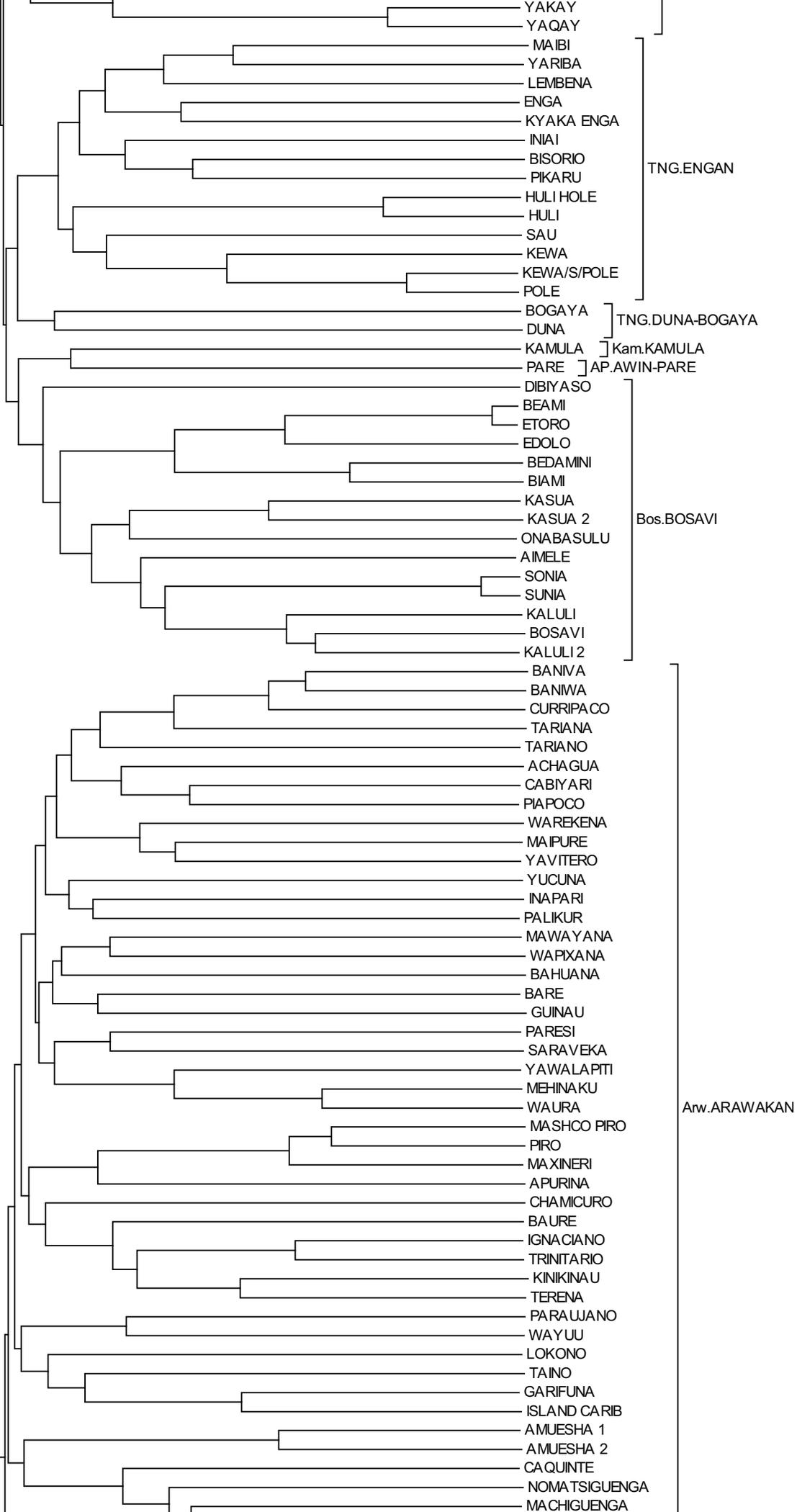
TNG.MADANG

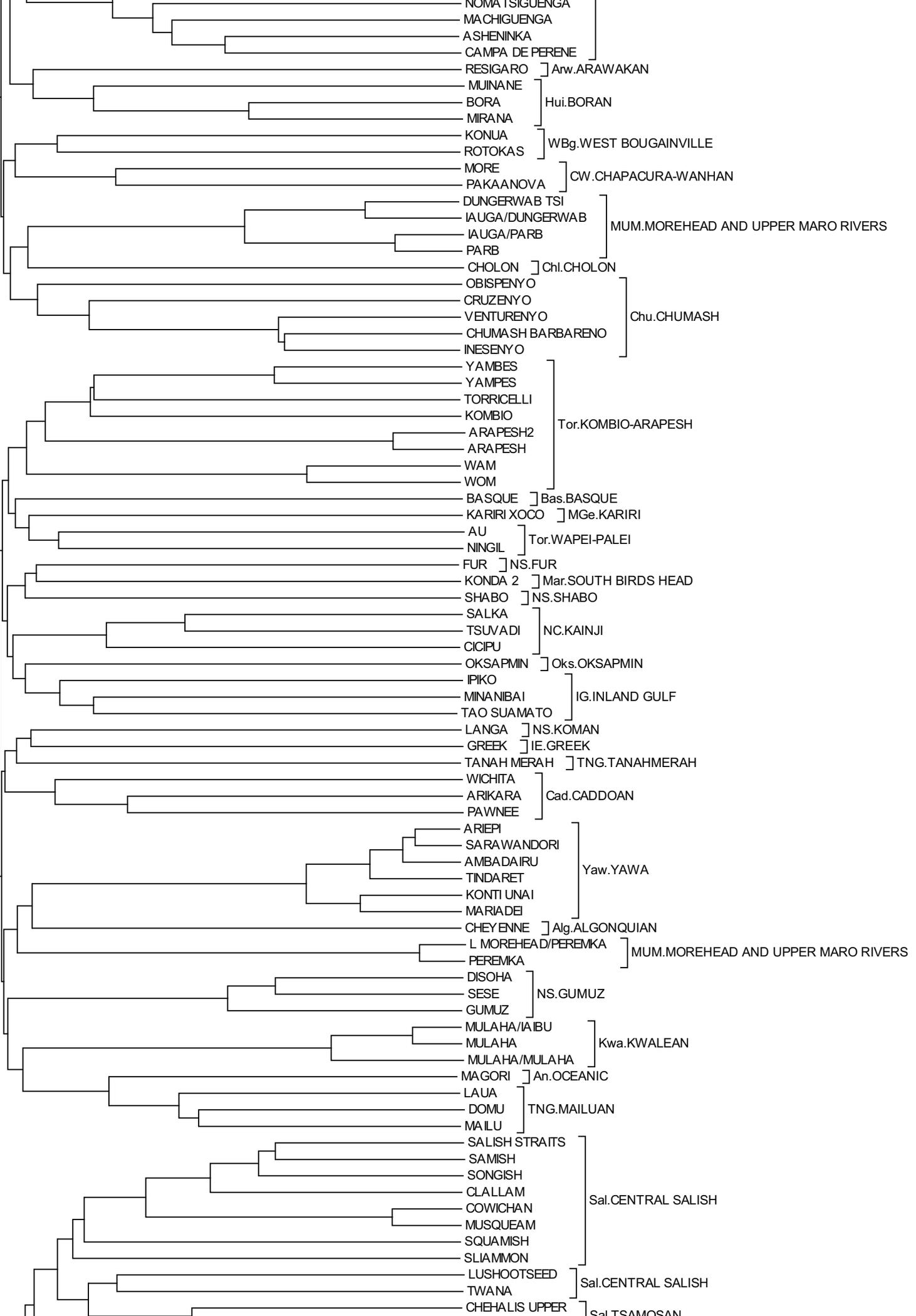
TNG.MADANG

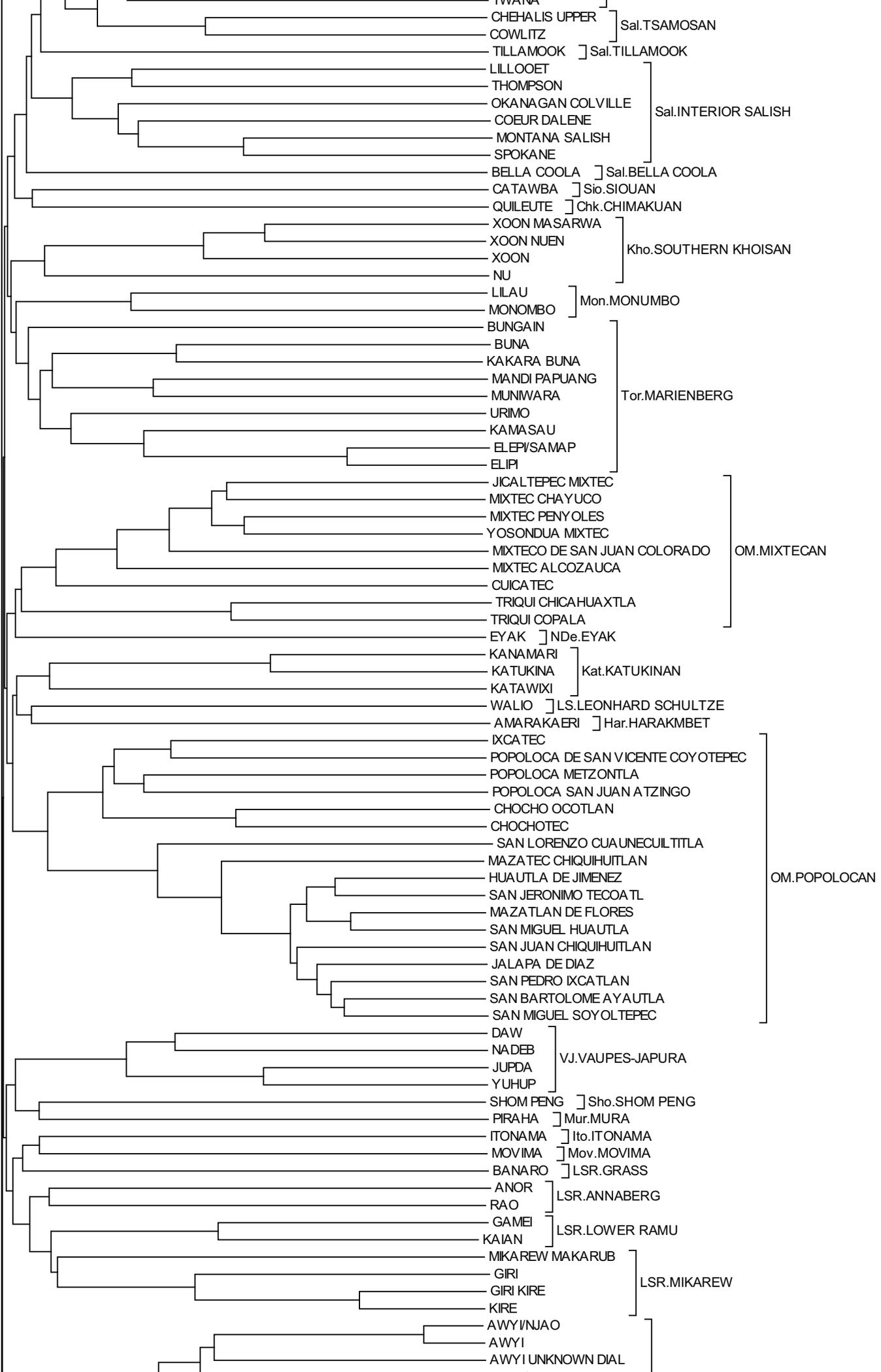


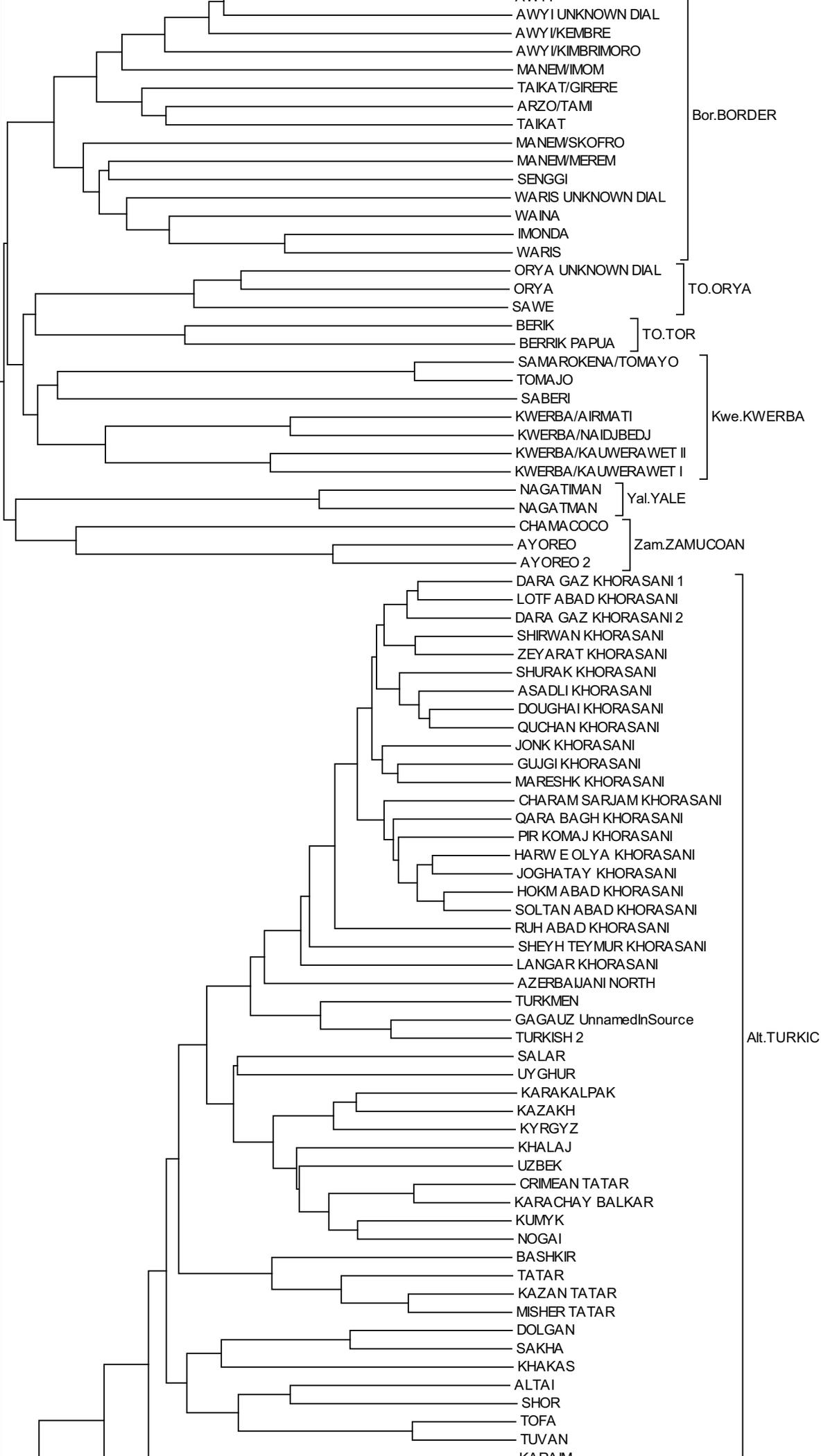


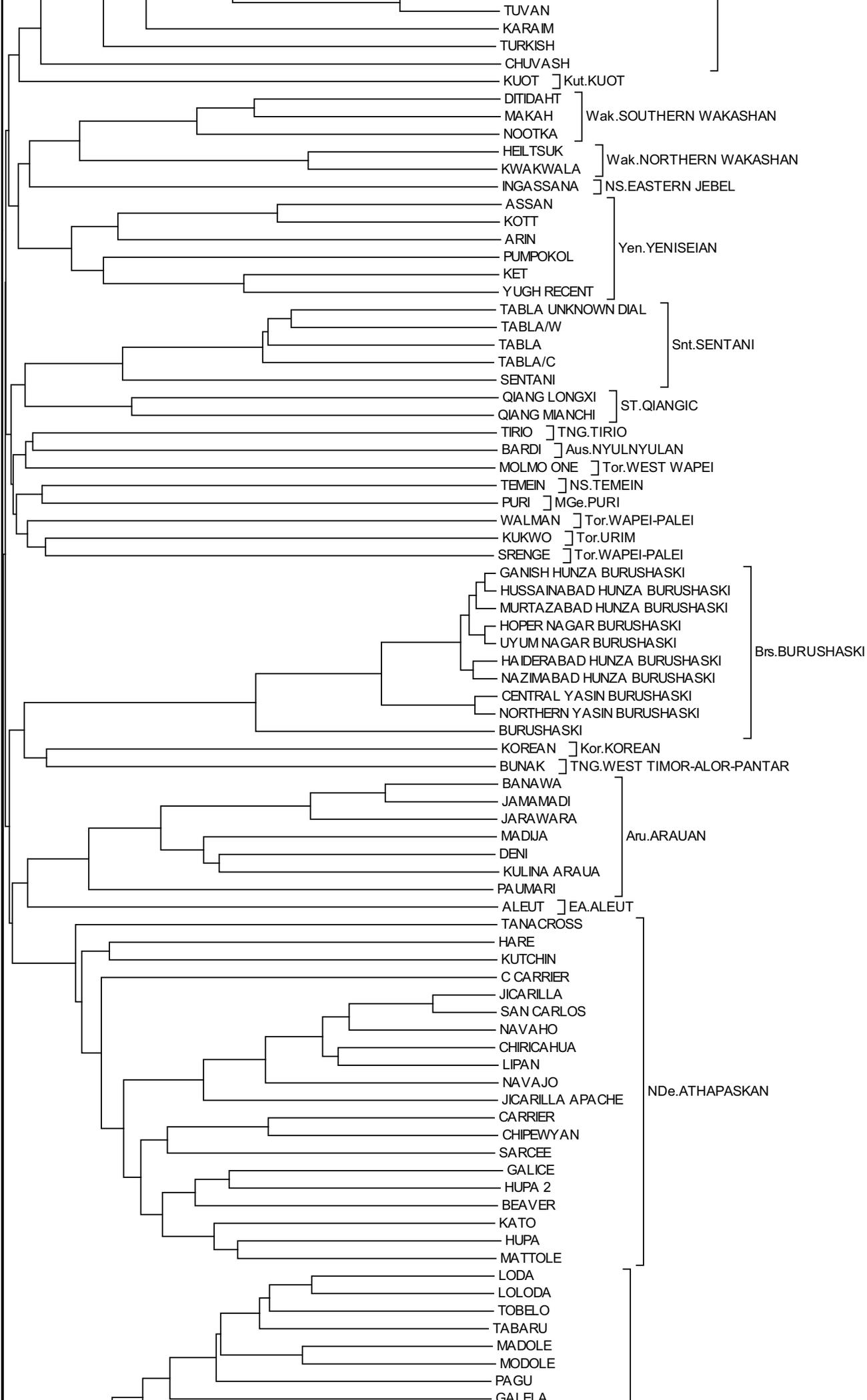


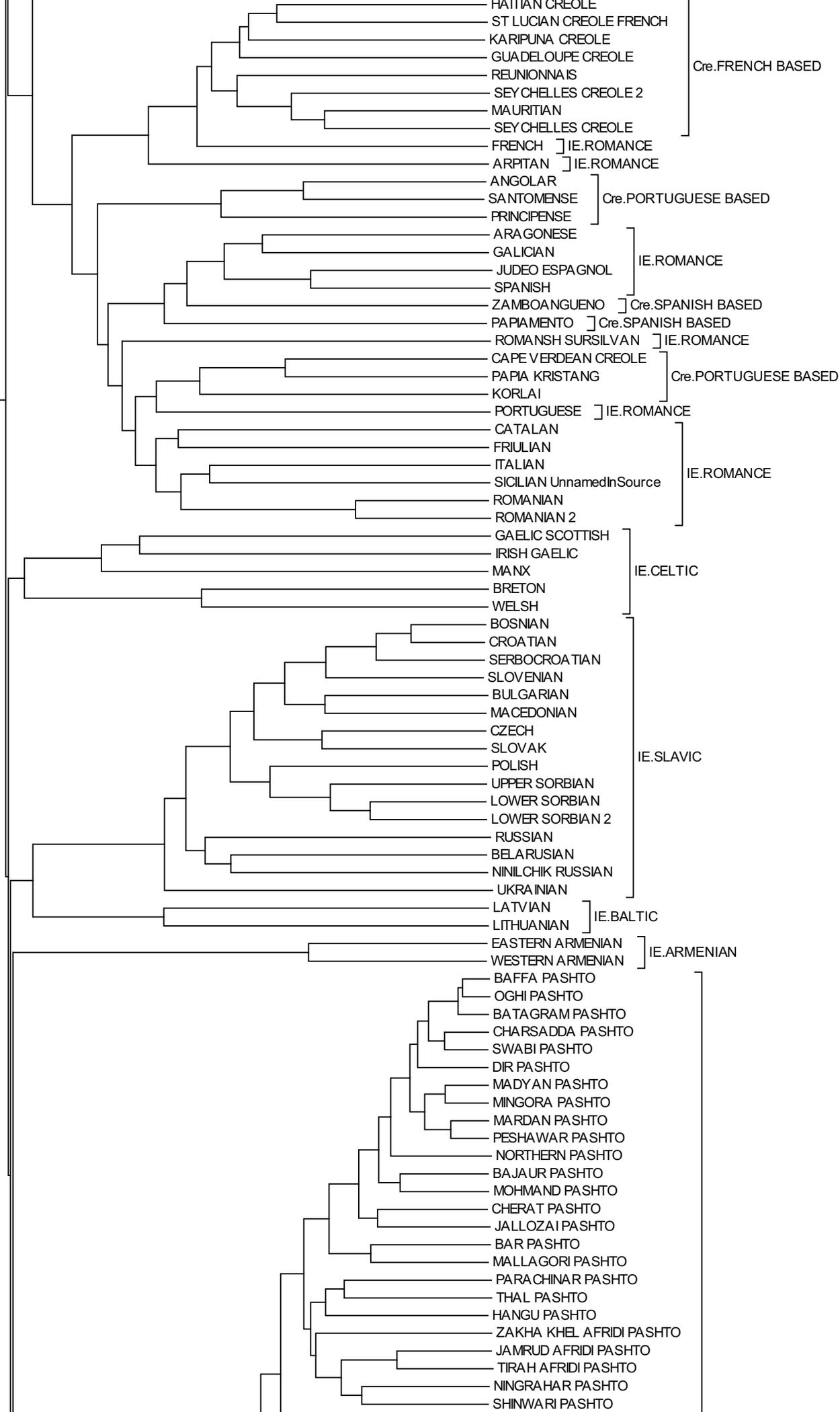


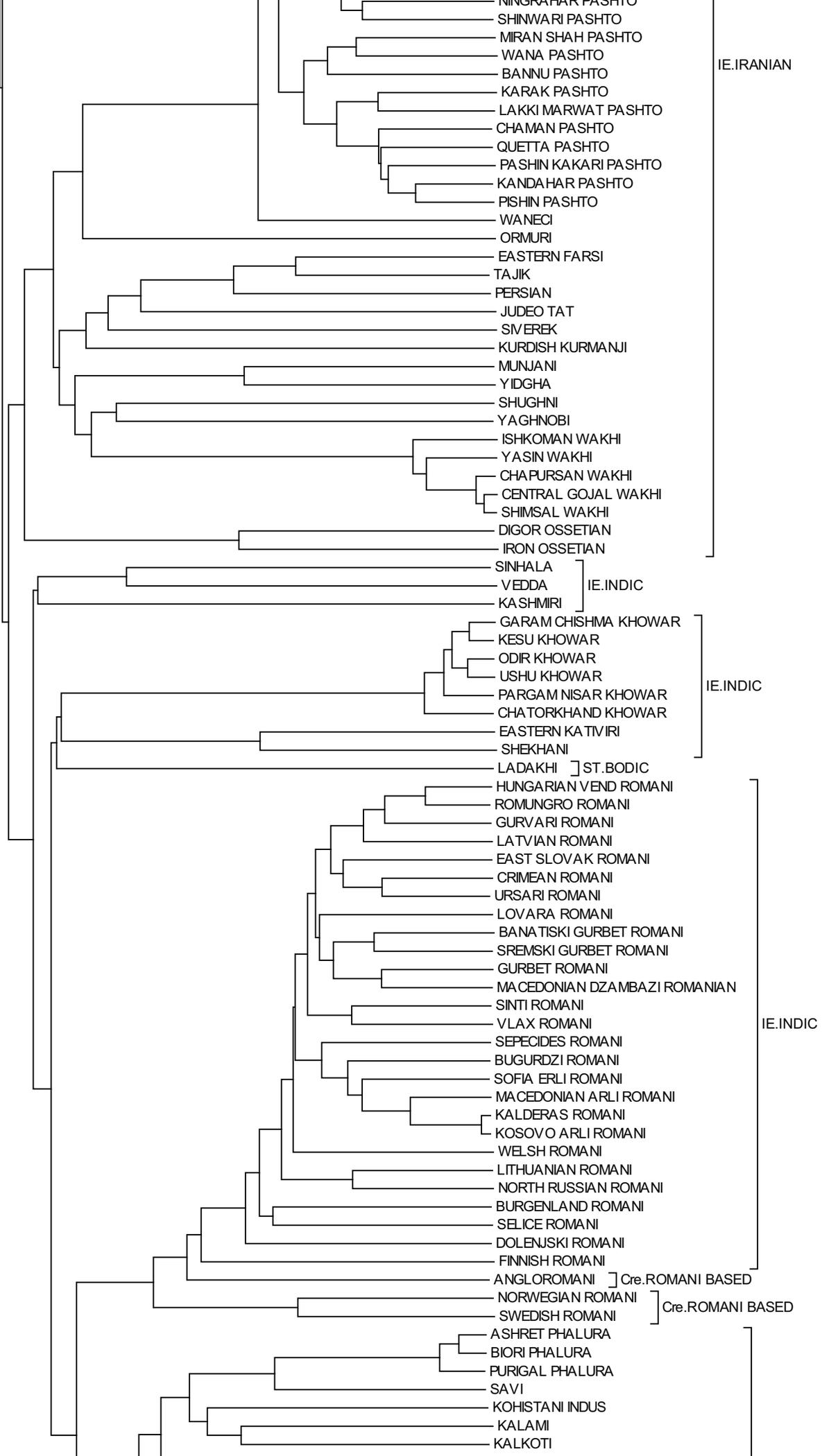


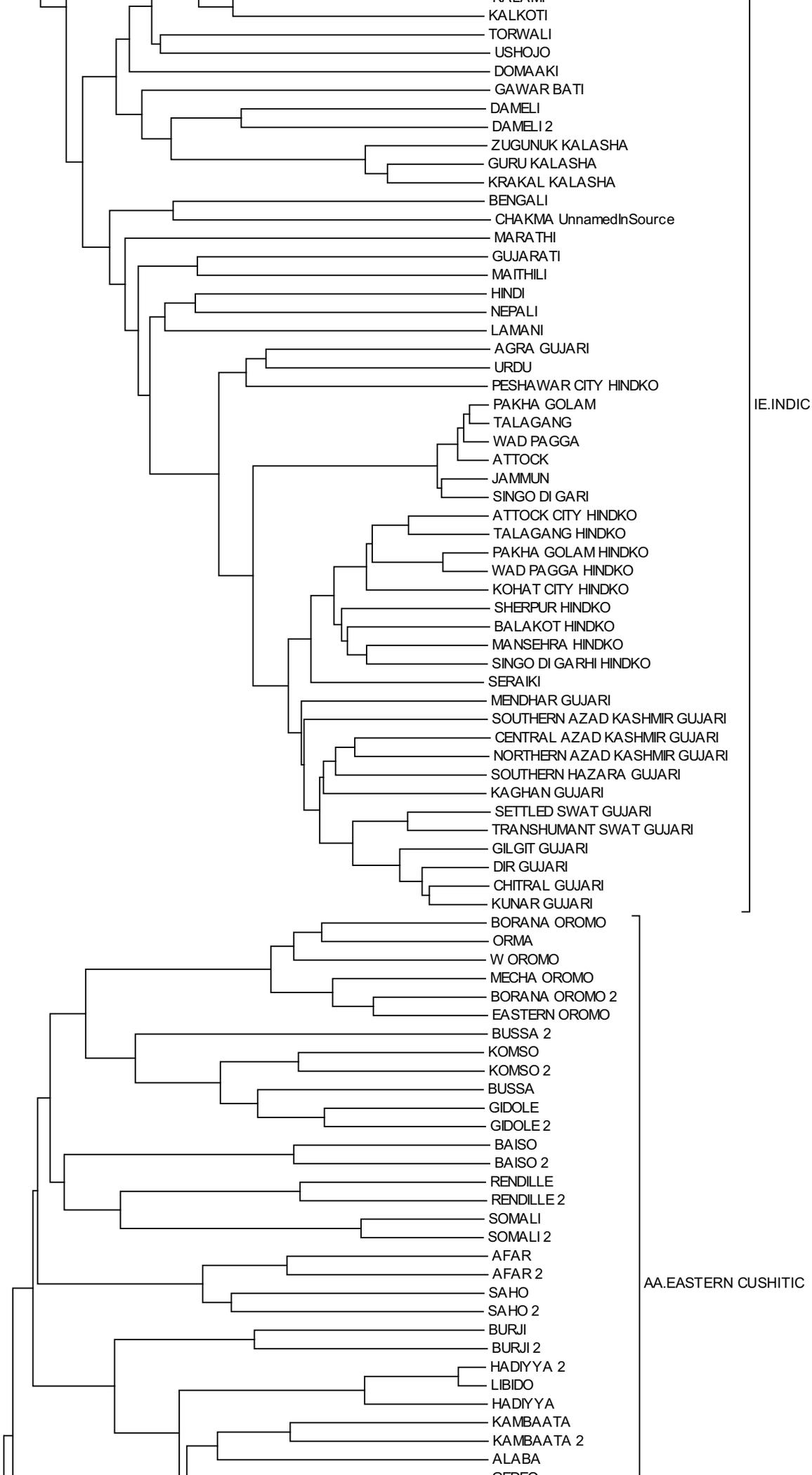


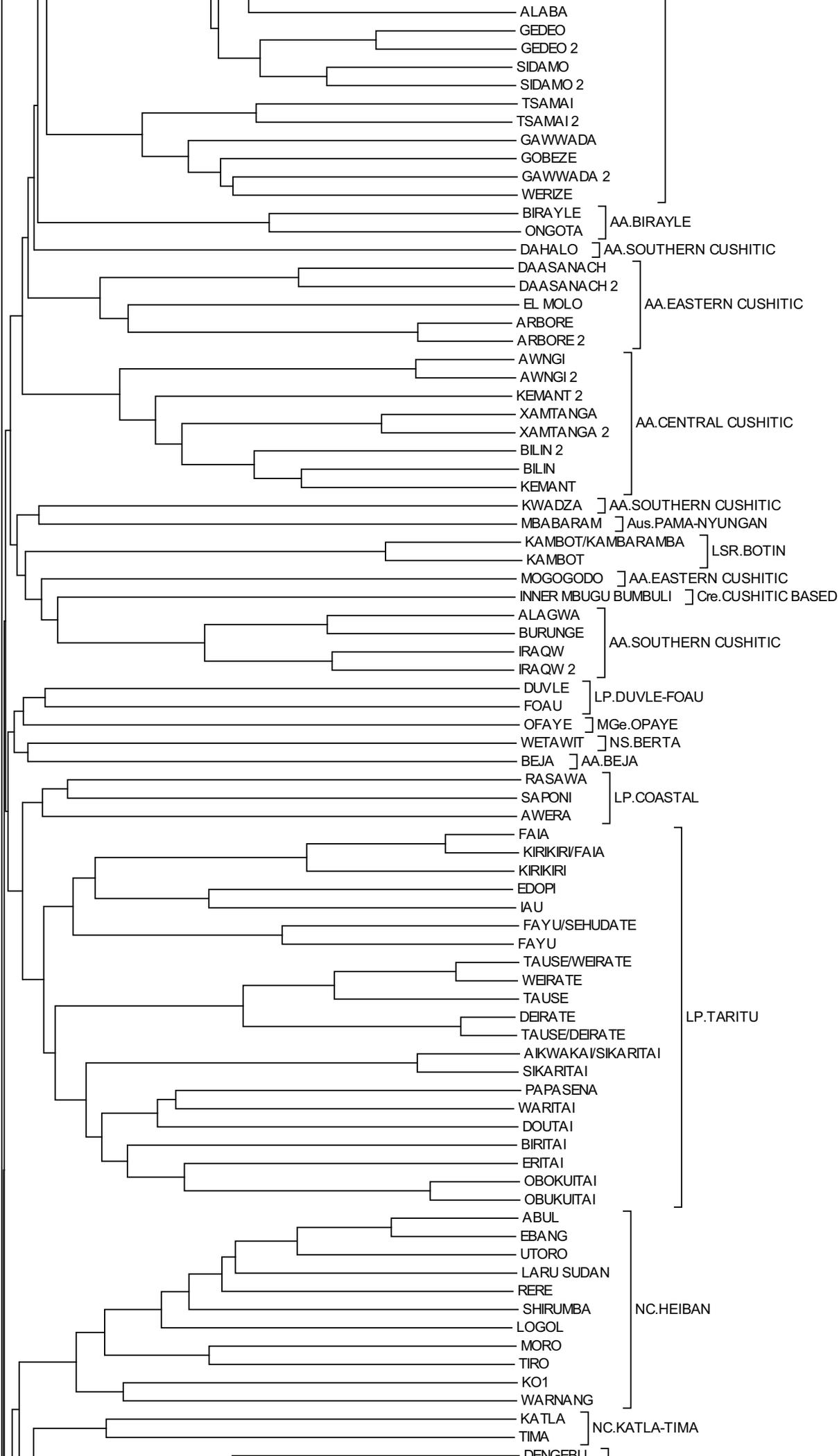












ALABA

GEDEO

GEDEO 2

SIDAMO

SIDAMO 2

TSAMAI

TSAMAI 2

GAWWADA

GOBEZE

GAWWADA 2

WERIZE

BIRAYLE

ONGOTA

AA.BIRAYLE

DAHALO

AA.SOUTHERN CUSHITIC

DAASANACH

DAASANACH 2

EL MOLO

AA.EASTERN CUSHITIC

ARBORE

ARBORE 2

AWNGI

AWNGI 2

KEMANT 2

XAMTANGA

XAMTANGA 2

AA.CENTRAL CUSHITIC

BILIN 2

BILIN

KEMANT

KWADZA

AA.SOUTHERN CUSHITIC

MBABARAM

Aus.PAMA-NYUNGAN

KAMBOT/KAMBARAMBA

KAMBOT

LSR.BOTIN

MOGOGODO

AA.EASTERN CUSHITIC

INNER MBUGU BUMBULI

Cre.CUSHITIC BASED

ALAGWA

BURUNGE

IRAQW

IRAQW 2

AA.SOUTHERN CUSHITIC

DUVLE

FOAU

LP.DUVLE-FOAU

OFAYE

MGe.OPAYE

WETAWIT

NS.BERTA

BEJA

AA.BEJA

RASAWA

SAPONI

AWERA

LP.COASTAL

FAIA

KIRIKIR/FAIA

KIRIKIRI

EDOPI

IAU

FAYU/SEHUDATE

FAYU

TAUSE/WEIRATE

WEIRATE

TAUSE

DEIRATE

TAUSE/DEIRATE

AIKWAKA/SIKARITAI

SIKARITAI

PAPASENA

WARITAI

DOUTAI

BIRITAI

ERITAI

OBOKUITAI

OBUKUITAI

LP.TARITU

ABUL

EBANG

UTORO

LARU SUDAN

RERE

SHIRUMBA

NC.HEIBAN

LOGOL

MORO

TIRO

KO1

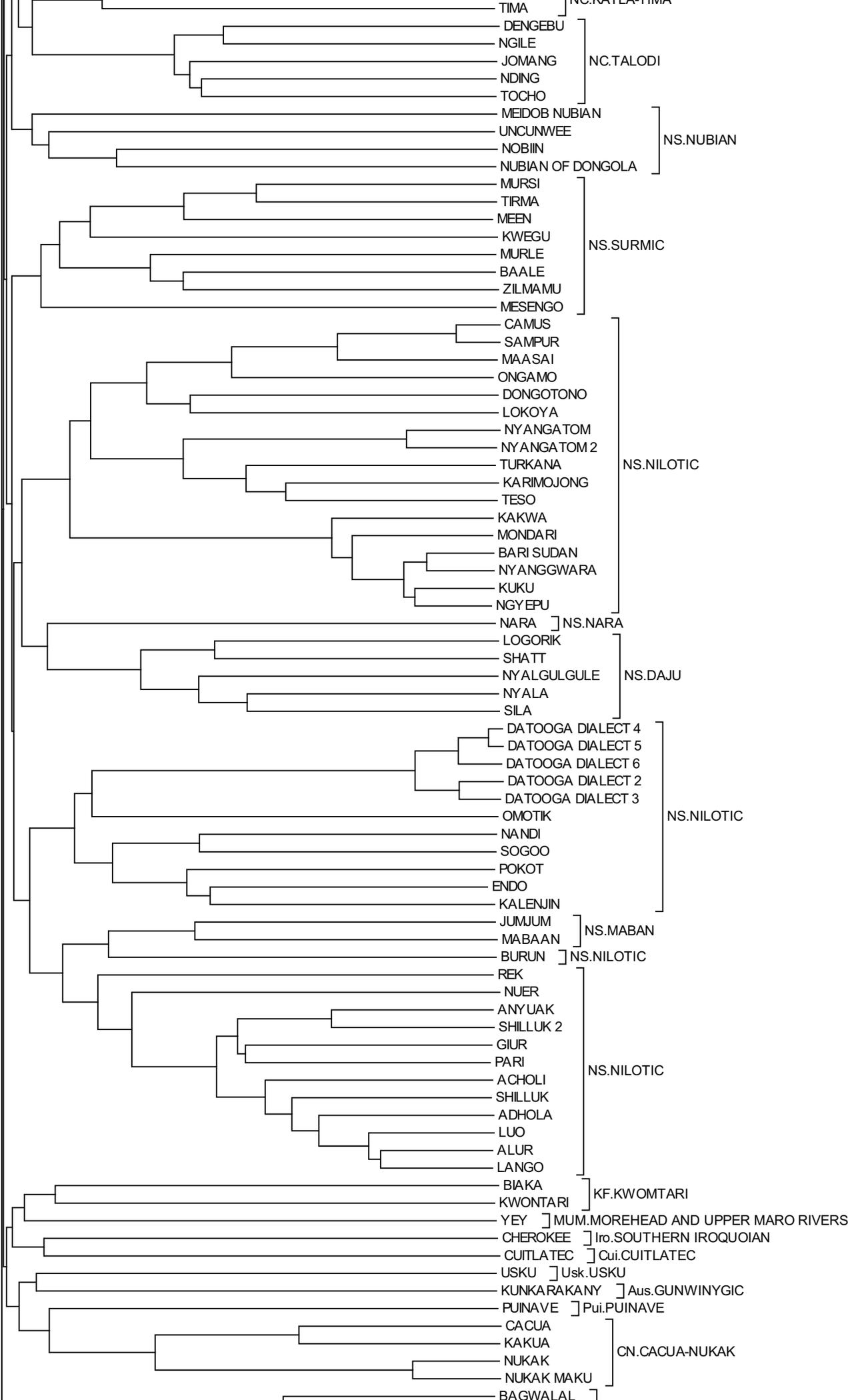
WARNANG

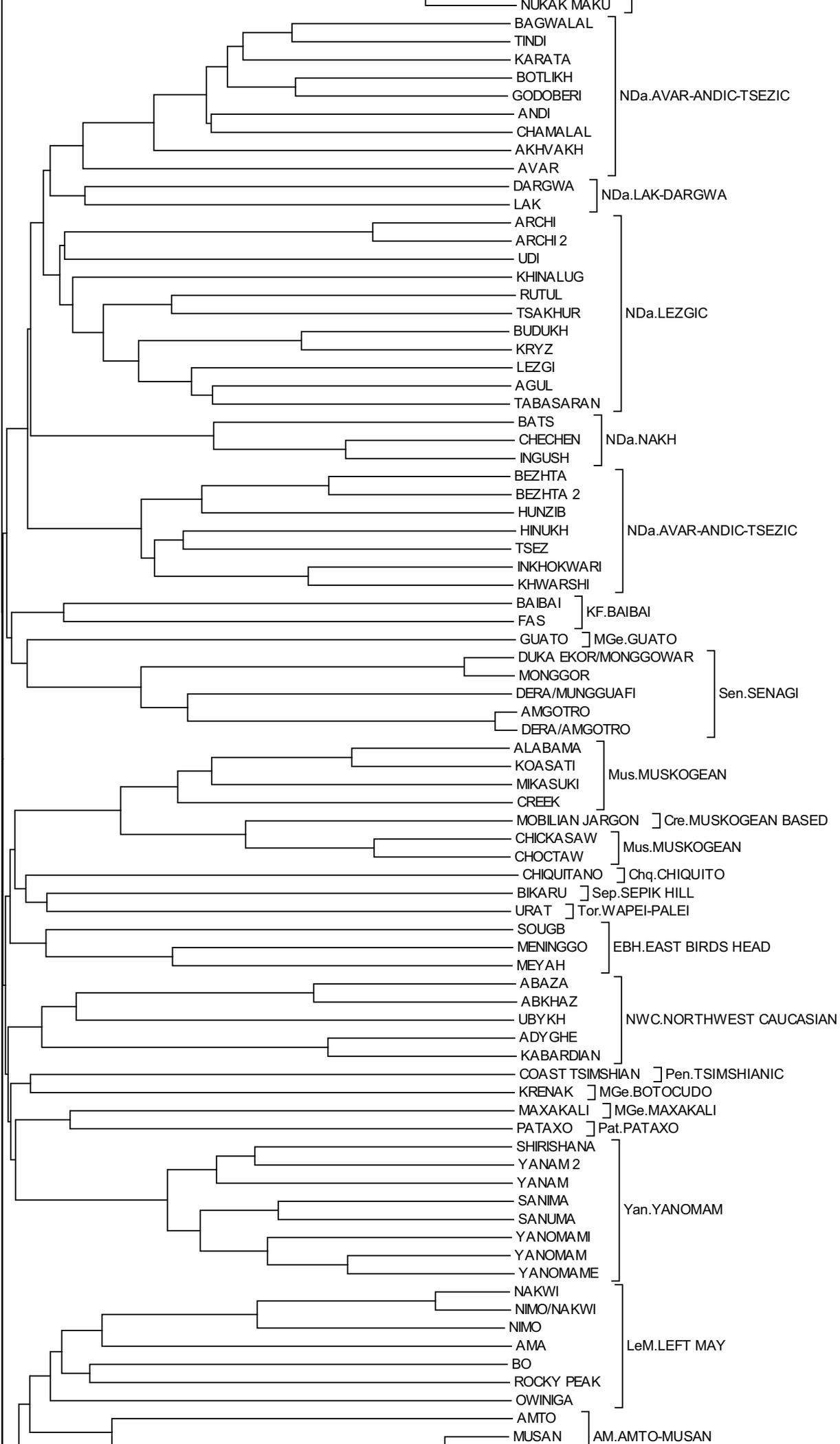
KATLA

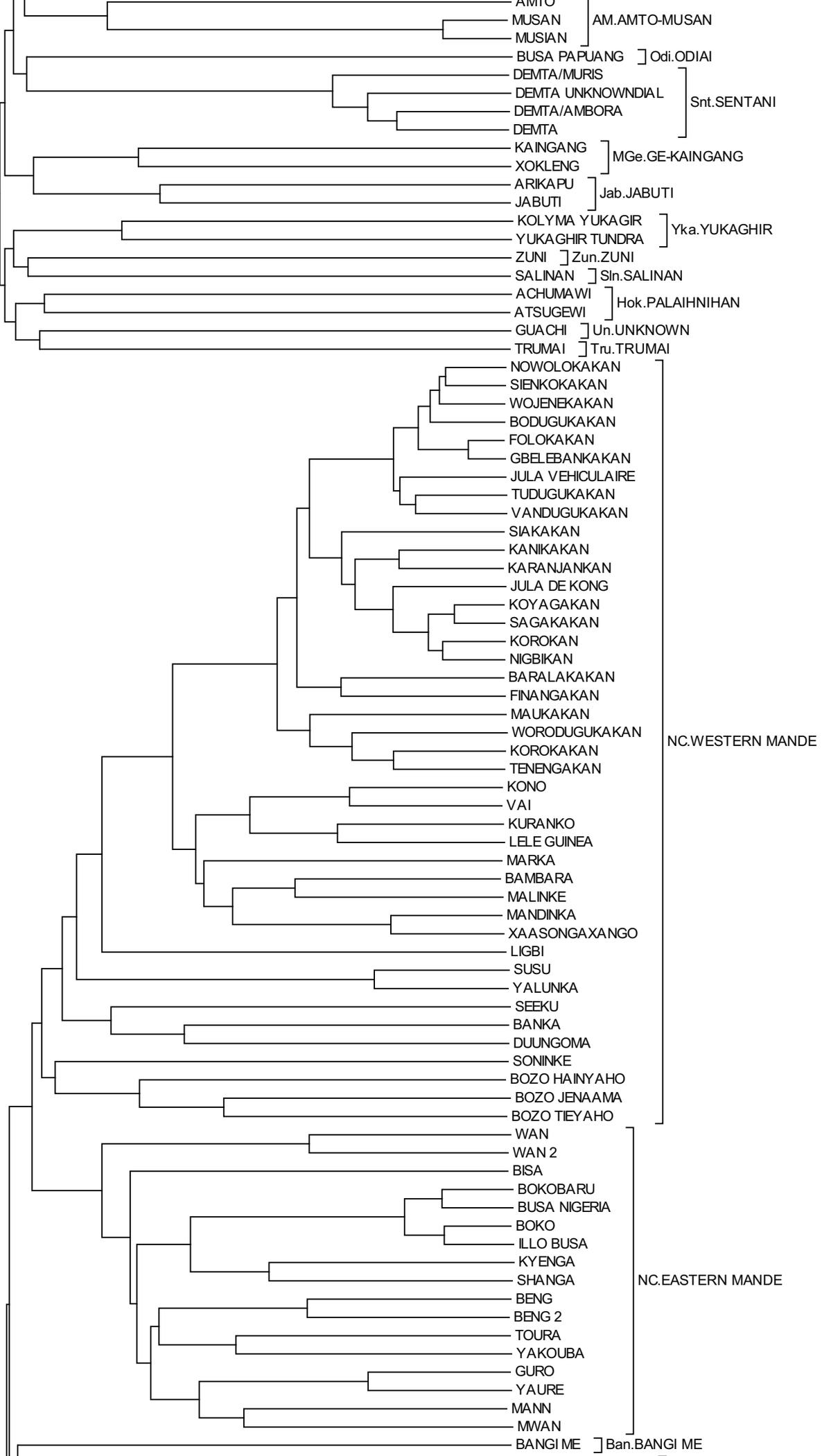
TIMA

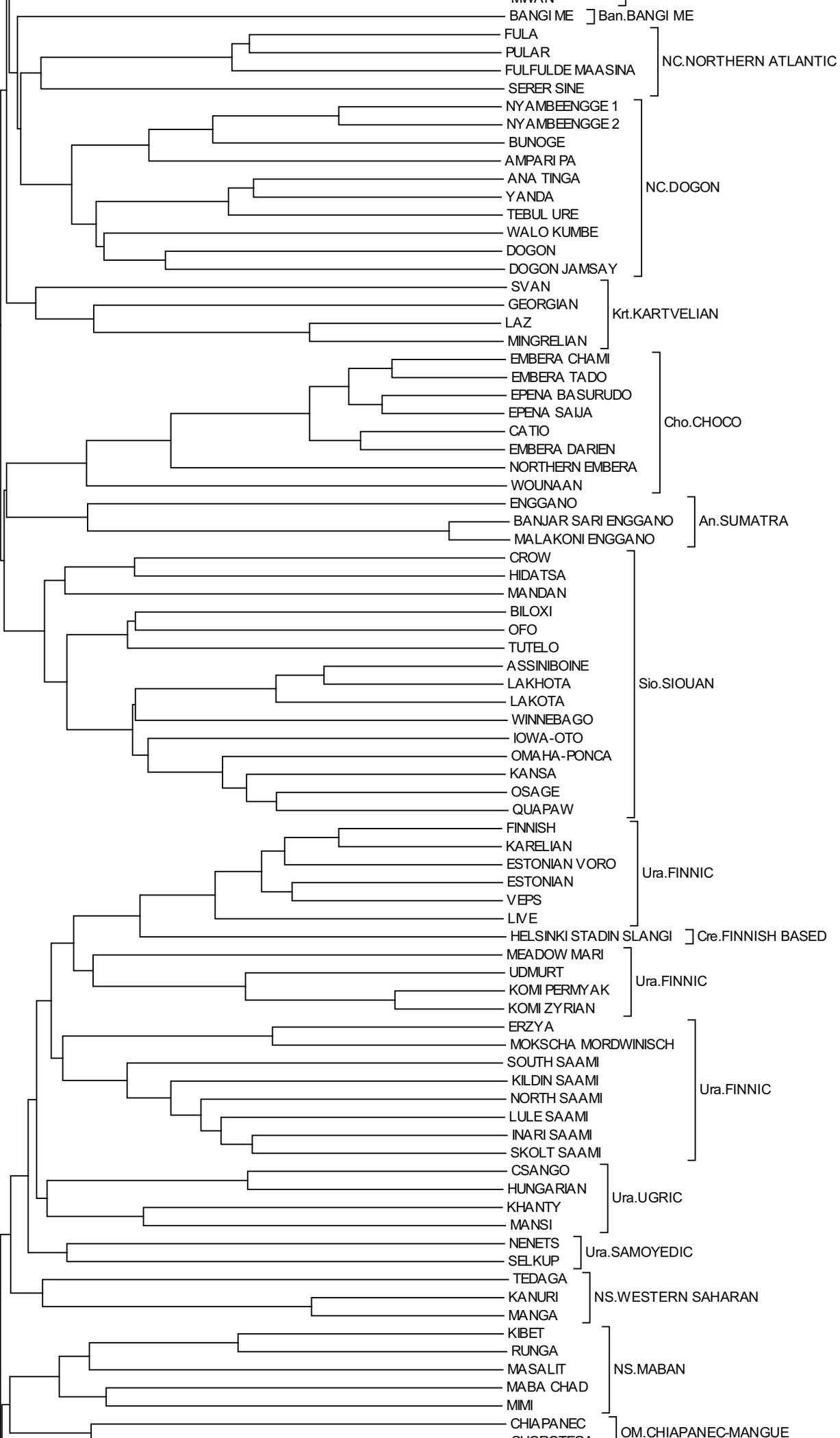
NC.KATLA-TIMA

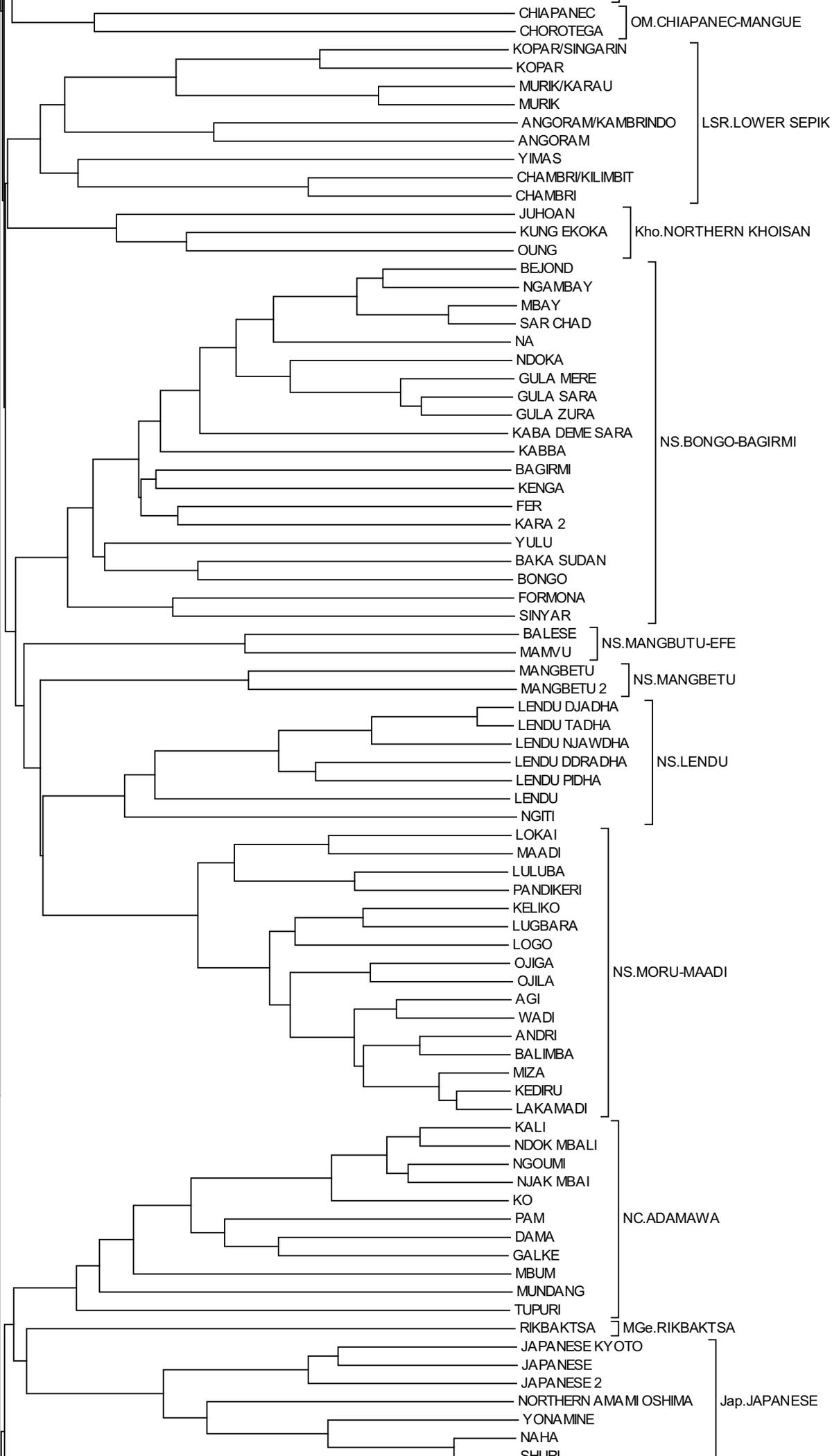
DENGERU

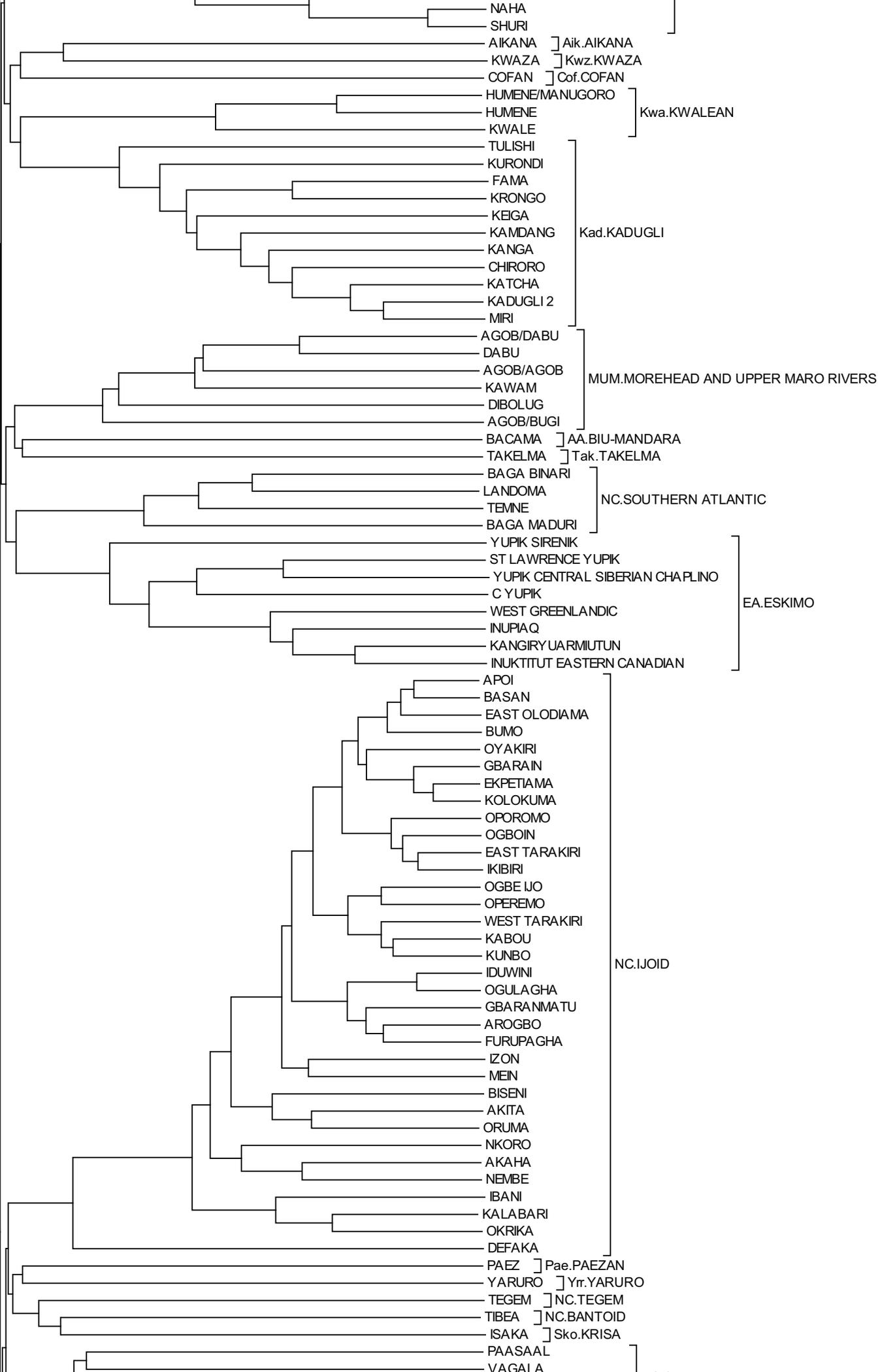


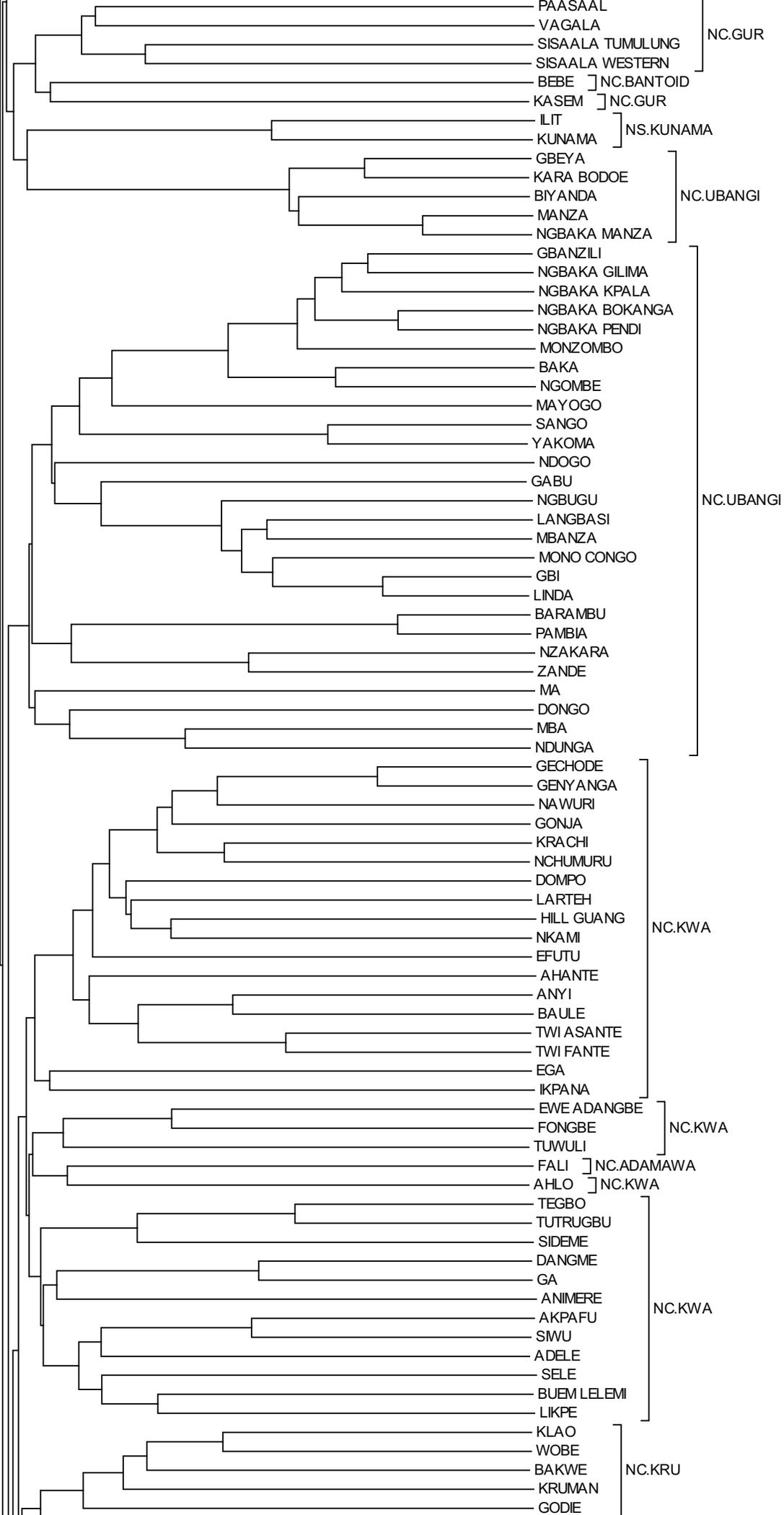


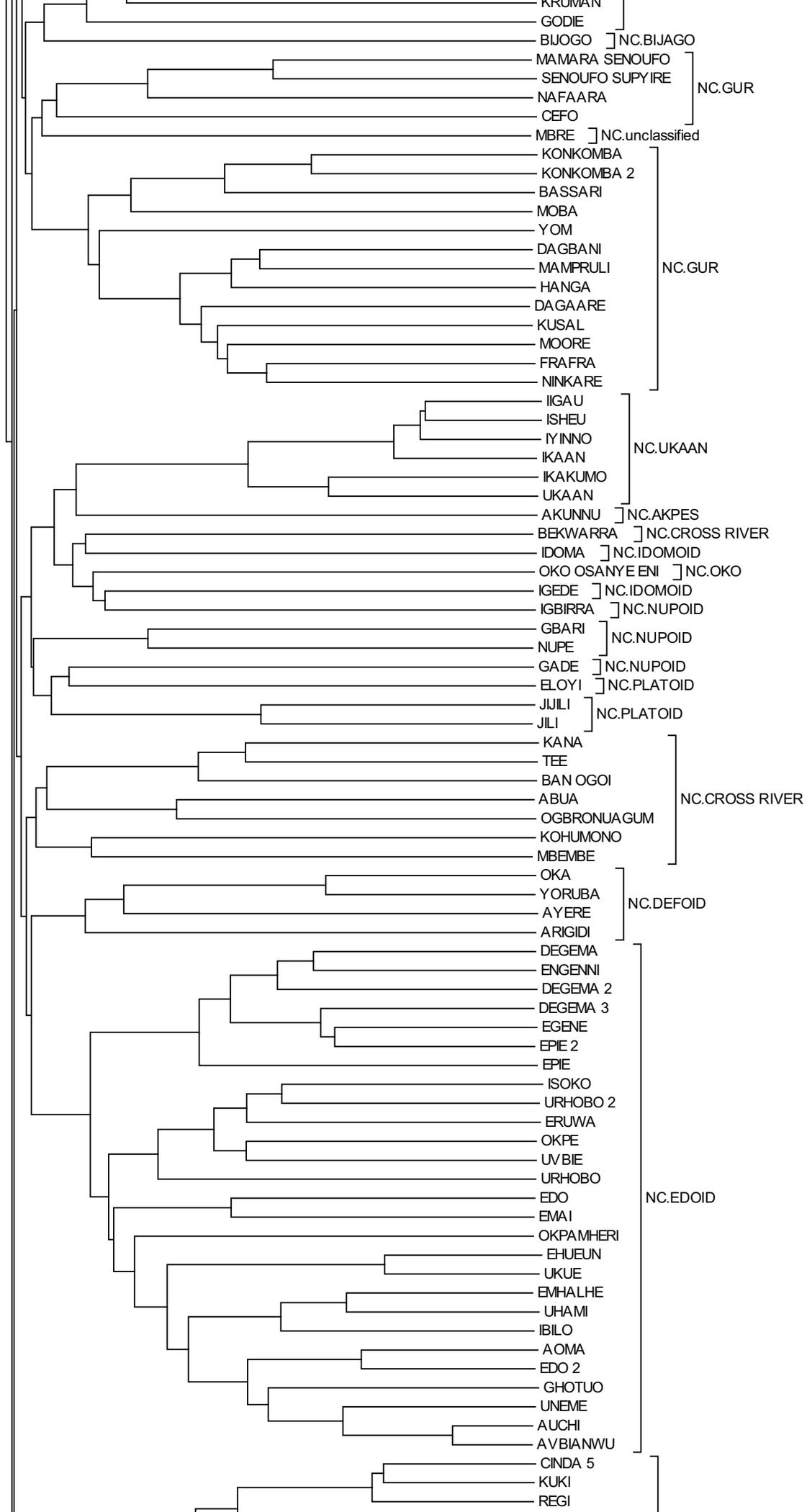


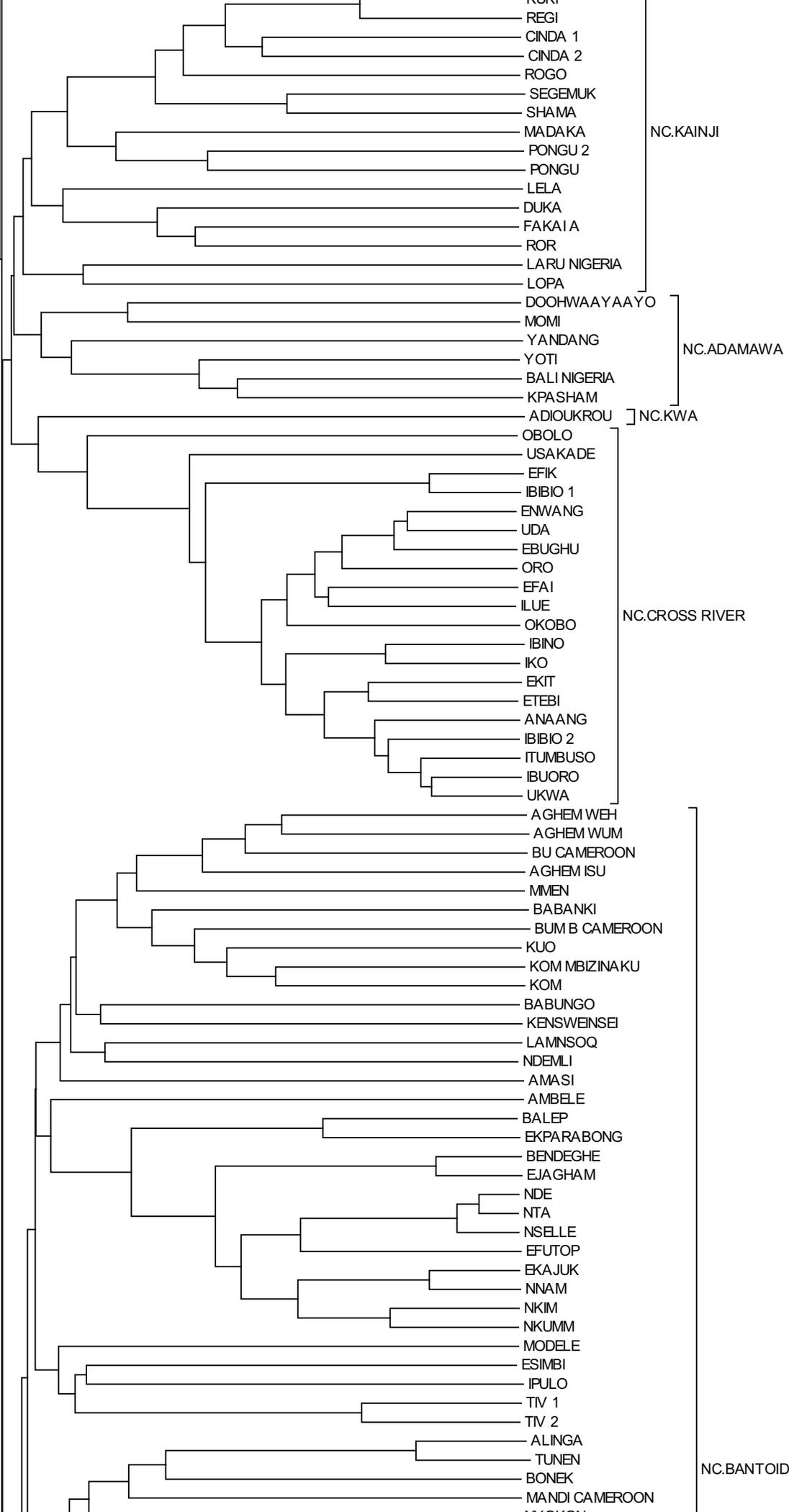


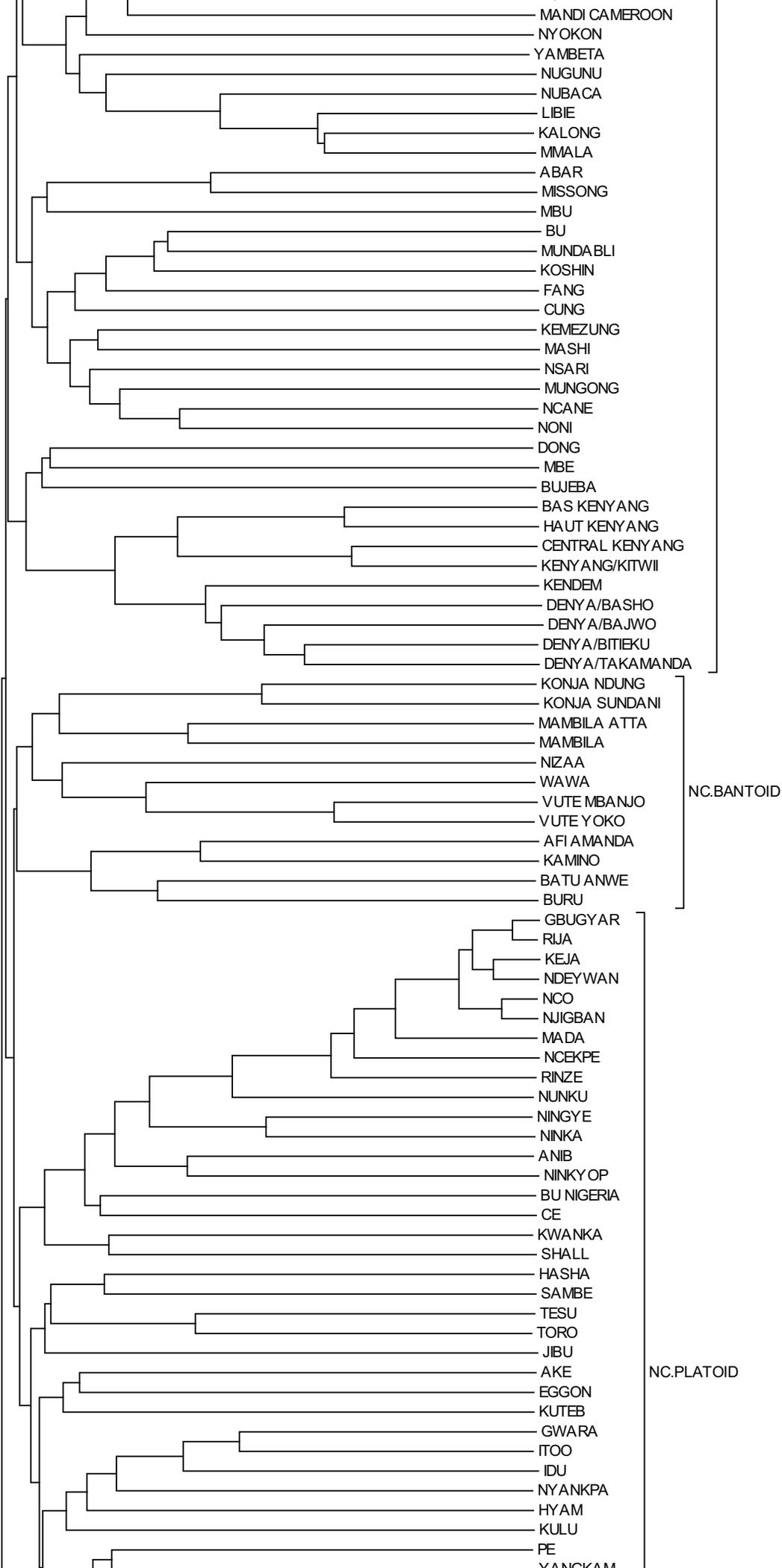


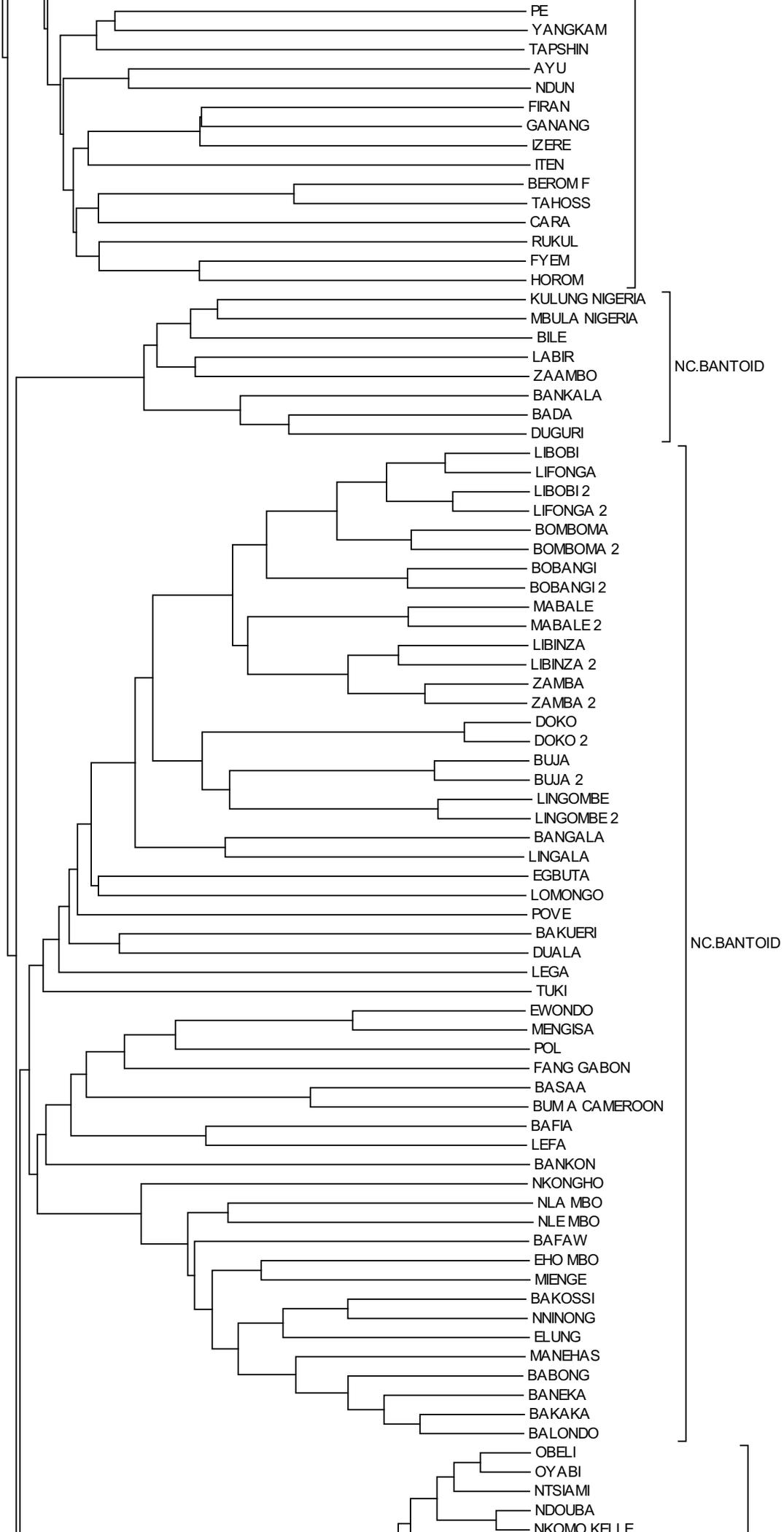


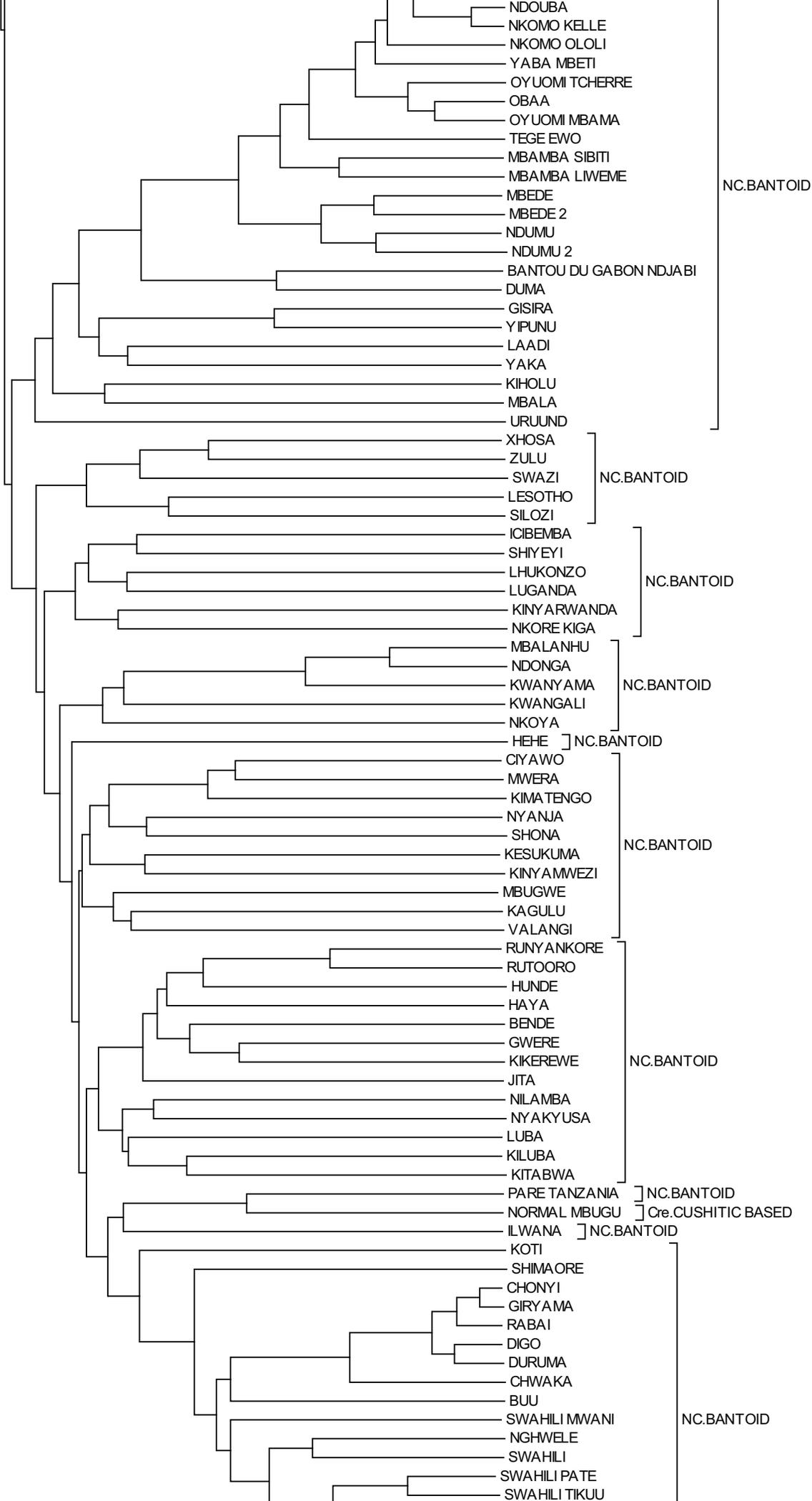


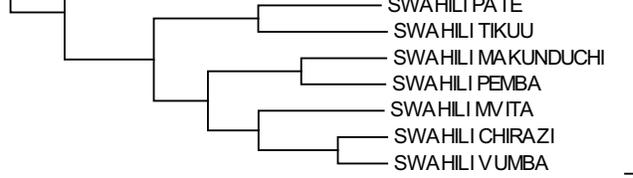












1000