Supplementary Information for the Paper "Dated Language Phylogenies Shed Light on the ancestry of Sino-Tibetan languages"

Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin J. Ryder,
Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List

March 2019

Contents

1 Organization of the supplementary material **3**

2 Information on Sino-Tibetan **3**
- 2.1 Outline of Sino-Tibetan languages . . . 3
- 2.2 Information on different subgrouping hypotheses . . . 4

3 Language data and historical language comparison **5**
- 3.1 Languages in our sample . . . 5
  - 3.1.1 Criteria for language choice . . . 6
  - 3.1.2 Dealing with ancient languages . . . 7
- 3.2 Concepts in our sample . . . 7
- 3.3 Cognate coding . . . 14
  - 3.3.1 Avoiding concepts of high compoundhood . . . 15
  - 3.3.2 From partial cognates to root cognates . . . 15
  - 3.3.3 Supporting cognate judgments with alignment analyses . . . 15
- 3.4 Coverage and cognate statistics . . . 16

4 Phylogenetic analyses **17**
- 4.1 Clades discussed in the paper . . . 17
- 4.2 Phylogenetic constraints . . . 17
- 4.3 Subgrouping results . . . 17
- 4.4 Outgroups . . . 18
- 4.5 Root age . . . 20
- 4.6 Age of Old Burmese . . . 20
- 4.7 Analysis of a subset of the languages . . . 20
- 4.8 Densitrees . . . 22
- 4.9 Etymologies supporting several proposed subgroups . . . 24
  - 4.9.1 Tibeto-Dulong . . . 24
  - 4.9.2 Tibeto-Gyalrongic . . . 25
  - 4.9.3 Burmo-Gyalrongic . . . 26

www.pnas.org/cgi/doi/10.1073/pnas.1817972116

5 Homeland and domesticates **26**
5.1 Lexical sets . . . 27
5.2 Archaeological dates . . . 29
5.3 Foxtail millet (*Setaria italica*) . . . 29
5.4 Broomcorn millet (*Panicum miliaceum*) . . . 30
5.5 Rice . . . 30
5.6 Pig . . . 31
5.7 Field . . . 31
5.8 Sheep . . . 32
5.9 Horse . . . 32
5.10 Cattle . . . 33

1 Organization of the supplementary material

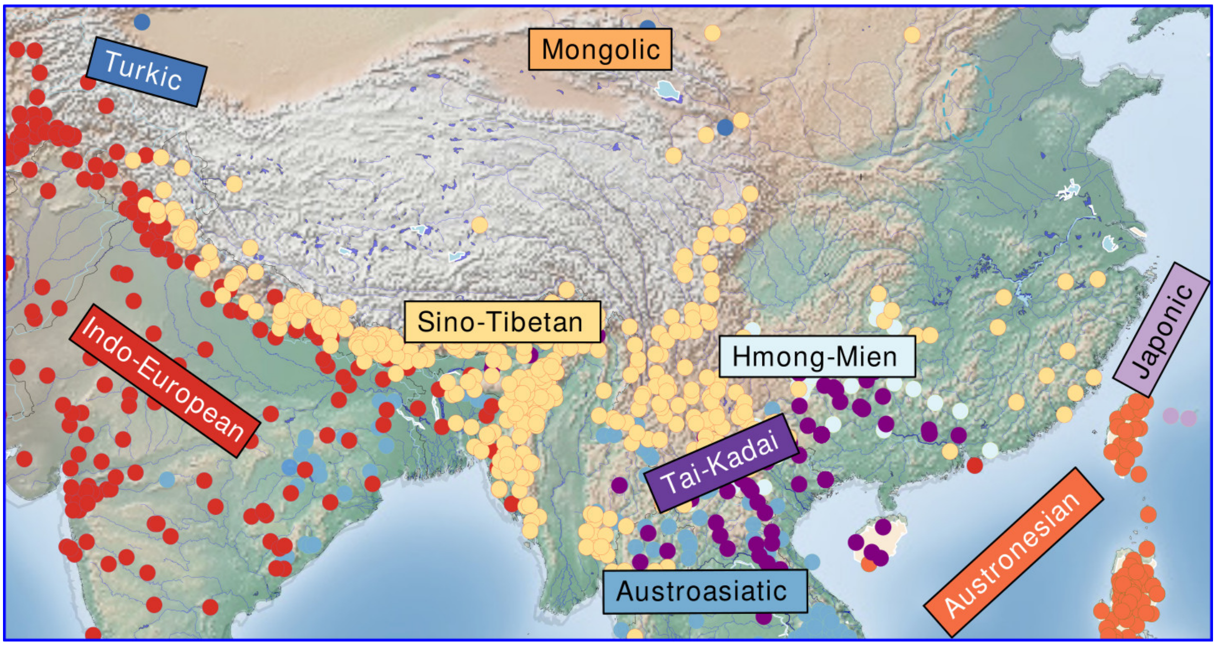
The data was curated with help of the EDICTOR (List 2017). We used a server-based version to ease collaboration. A link to the database can be found at `http://dighl.github.io/sinotibetan`, where languages and concepts can be selected and then browsed in the EDICTOR application. Since the database curation process was in flux for a long time, and may still change in the future, we provide a final stable dump of this database, including all concepts and the 50 languages that we collected before, in the repository accompanying this supplementary information. The data itself is curated on GitHub (`https://github.com/lexibank/sagartst`), while the versions underlying this draft are archived at Zenodo (`https://zenodo.org/record/1465485`). The version we used for the experiments reported here and in the paper is Version 1.0.0. The code to convert the database to Nexus format and to replicate the phylogenetic reconstruction analyses can be found on GitHub (`https://github.com/lingpy/sino-tibetan-paper`, version 1.0.3), and has been archived with Zenodo (`https://zenodo.org/record/2543222`).

2 Information on Sino-Tibetan

There is broad agreement on the existence of the Sino-Tibetan family (a.k.a "Trans-Himalayan"), including Chinese, Tibetan, Burmese, Tangut, Newari and several hundred related languages on and around the Tibetan plateau, but excluding Kra-Dai, Hmong-Mien, Austroasiatic or Austronesian. Earlier versions of Sino-Tibetan, still defended by certain Chinese scholars, were more inclusive: Li (1937 [1973]) and Shafer (1955) also included Kra-Dai, and Li (1937 [1973]) also Hmong-Mien. Previous attempts at reconstructing Proto-Sino-Tibetan include Coblin (1986), Gong (1995), and Peiros and Starostin (1996); reconstructions of Proto-Tibeto-Burman, the putative ancestor of the non-Chinese part of the family, Benedict (1972) and Matisoff (2003a). For methodological reservations on the Benedict-Matisoff reconstruction paradigm, see Hill (2009), Miller (1974), and Sagart (2006). Whether Sino-Tibetan is an isolated language family or whether it belongs to a larger macro-family is disputed: Sino-Tibetan has been linked, among others, with Yenissean and north Caucasian (Starostin 1984[1988]); Austronesian (Sagart 2005); Austroasiatic and Hmong-Mien (Starosta 2005); Indo-European (Chang 1988). Knowledge of Sino-Tibetan sound correspondences is improving, both within branches (Jacques 2014, Jacques 2017a, Joseph and Burling 2006, VanBik 2009) and across branches (Hill 2012, Hill 2014, Sagart 2017).

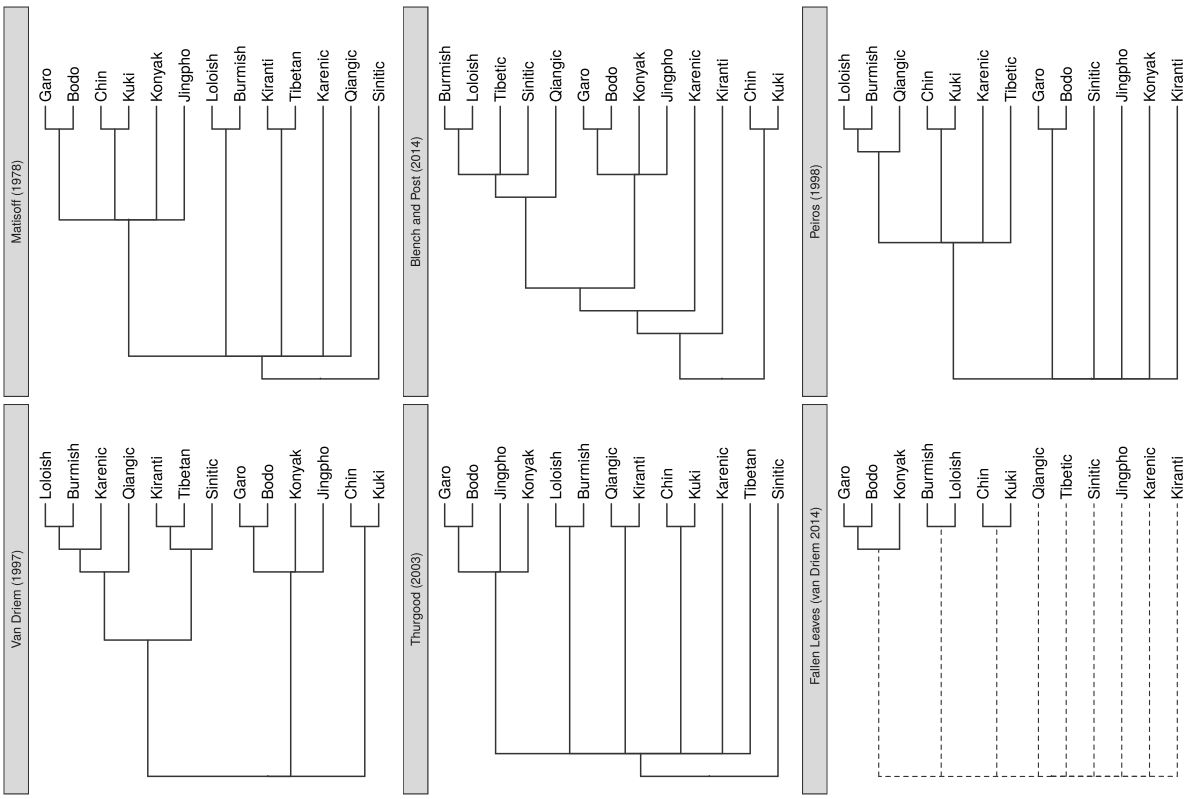
2.1 Outline of Sino-Tibetan languages

The figure below shows the general outline of Sino-Tibetan languages, along with surrounding languages from different families. The map lists the information provided in Glottolog.



2.2 Information on different subgrouping hypotheses

The figure below contrasts different subgrouping hypotheses for a small sample of subgroups occurring in all different hypotheses.



3 Language data and historical language comparison

3.1 Languages in our sample

The following is the list of languages in our sample. More information can be found in our Lexibank repository, from which this data was taken. Along with the subgroups as provided by Glottolog (Hammarström et al. 2017), we also list where the sources can be found in the STEDT database (Matisoff 2015) (if we used the digitized versions provided by STEDT for our study).

ID	Variety	Subgroup	Cov.	Glott.	STEDT	Source
01	Achang	Burmish	197	acha1249	TBL	Hill and List 2017, Huáng 1992
02	Atsi	Burmish	197	zaiw1241	TBL	Hill and List 2017, Huáng 1992
07	Bola	Burmish	214	pela1242	TBL	Hill and List 2017, Huáng 1992
03	Bahing	Kiranti	178	bahi1252	BM-Bah	Michailovsky 1989a
04	Bantawa	Kiranti	195	bant1281		Jongens 2009
05	Beijing	Sinitic	217	beij1235		Lai 2017a
06	Bokar	Tani	225	boka1249	TBL	Huáng 1992
08	Bunan	Tibeto-Kinauri	216	gahr1239		Widmer 2017
09	Byangsi	Tibeto-Kinauri	195	byan1241		Sharma 2003a
10	Chaozhou	Sinitic	205	chao1238		Lai 2017a
11	Chepang	Chepang	202	chep1245	RC-DOC	Caughley 2000
12	Daofu	rGyalrong	223	horp1240	TBL	Huáng 1992
13	Darang Taraon	Deng	217	diga1241	TBL	ibid.
14	Dulong	Nungic	224	drun1238	TBL	ibid.
15	Garo	Garo	206	garo1247	RB-LMMG	Burling 2003
16	Guangzhou	Sinitic	218	guan1279		Lai 2017a
17	Hakha	Chin	222	haka1240		VanBik 2014
18	Hayu	Kiranti	177	wayu1241	BM-Hay	Michailovsky 1989b
19	Japhug	rGyalrong	225	japh1234		Jacques 2015
20	Jieyang	Sinitic	213	chao1239		Lai 2017a
21	Jingpho	Jingpho	225	jing1260	TBL	Huáng 1992
22	Khaling	Kiranti	212	khal1275		Jacques 2017b
23	Kulung	Kiranti	195	kulu1253		Tolsma 1999
24	Lashi	Burmish	197	lash1243	TBL	Hill and List 2017, Huáng 1992
25	Limbu	Kiranti	198	limb1266		Jacques 2017b
26	Lisu	Loloish	224	lisu1250	TBL	Huáng 1992
27	Longgang	Sinitic	211	hakk1236		Lai 2017a
28	Mizo (Lushai)	Mizo	200	lush1249		Lorrain 1940
29	Maru	Burmish	213	maru1249	TBL	Huáng 1992, Hill and List 2017
30	Karbi (Mikir)	Mikir	247	karb1241		Konnerth forthcoming, Walker 1925

31	Motuo Menba	Bodic	217	tsha1245	TBL	Huáng 1992
32	Old Burmese	Burmish	214	oldb1235		Hill and List 2017, Luce 1985, Nishi 1999, Okell 1971
33	Old Chinese	Sinitic	221	oldc1244		Baxter and Sagart 2014a
34	Old Tibetan	Tibetan	216	clas1254	TBL	Takeuchi 2013
35	Rabha	Koch	171	rabh1238		Joseph 2007
36	Rangoon Burmese	Burmish	216	nucl1310	TBL	Huáng 1992, Hill and List 2017
37	Rongpo	Tibeto-Kinauri	190	rong1264		Sharma 2003b
38	Tangut	Tangut	236	tang1334		Lǐ 1997
39	Thulung	Kiranti	210	thul1246	NJA-Thulung	Allen 1975
40	Alike Tibetan	Tibetan	209	amdo1237	TBL	Huáng 1992
41	Batang Tibetan	Tibetan	225	kham1282	TBL	ibid.
42	Lhasa Tibetan	Tibetan	225	utsa1239	TBL	ibid.
43	Xiahe Tibetan	Tibetan	225	amdo1237	TBL	ibid.
44	Ukhrul	Naga	190	ukhr1238	DRM-Tk	Mortensen 2012
45	Wobzi Khroskyabs	rGyalrong	224	eree1240		Lai 2017b
46	Xiandao	Burmish	190	xian1249	TBL	Hill and List 2017, Huáng 1992
47	Xingning	Sinitic	212	hakk1236		Lai 2017a
48	Yidu	Deng	213	idum1241	TBL	Huáng 1992
49	Zhaba	Qiangic	224	zhab1238	TBL	ibid.
50	Maerkang rGyalrong	rGyalrong	225	situ1238	TBL	ibid.

3.1.1 Criteria for language choice

Well-sampled language data with few missing items – a.k.a. high coverage – plays an important role in our study. First, low-coverage languages may have unwanted effects on phylogenetic reconstruction by increasing topological and timing uncertainty (Wiens 2006, Wiens and Morrill 2011). Second, low-coverage languages deprive us from the chance of confirming cognate judgments by identifying regular sound correspondences. Therefore, our selection of languages could not take into account all the languages for which data are available, be it in the form of a glossary or of a dictionary. In fact, we tested many more languages for potential inclusion in our database, but then had to discard them, because the mutual coverage turned out to be far too low. Examples include Dolakha Newar by Genetti (2007), for which we identified less than 80% of our larger list of 250 items, Kathmandu Newar by Kölver and Shresthacarya (1994), where for less than 60% of our items a translation could be identified, Tangkhul by Bhat (1969), with less than 80% of coverage, Dumi by Driem (1993) with less than 60% of coverage, and many more languages we checked. The coverage problems we encountered also explain why our selection of Sinitic languages does not contain all of the traditionally mentioned major groups. As we could not (yet) acquire first-hand data on these varieties, and sources, such as Liú et al. (2007) or Běijīng Dàxué (CIHUI) would not provide sufficient coverage for our basic vocabulary sample, we decided to exclude these varieties from the current analysis, rather than adding them at the cost of producing a low-coverage dataset.

In making our final section of 50 languages, we gave preference to those languages where we have first-hand knowledge or contact to experts whom we could ask for advice when facing problems. We also decided to include slightly larger subgroups for Burmish, Sinitic, and Kiranti, in order to allow for an independent verification of our findings. Since the subgrouping of these groups is rather well-known, a comparison of the inferred topologies and divergence dates with our general analysis can help to avoid major model misspecifications. Given the problem of partial cognates, which is specifically prevalent in Sinitic and Burmish, we also decided to include closely related language varieties, such as Achang and Xiandao for Burmish, or Chaozhou and Jieyang for Sinitic (both Southern Min dialects), since these would allow us to see to which degree closely related languages can already differ with respect to partial cognates.

3.1.2 Dealing with ancient languages

Given the complexity of identifying cognates in the ST language family, and specifically the problem of working with ancient languages, we took great care of verifying entries in all datasets with additional sources. In the case of Tibetan, for example, the attestations of the words were rechecked by G. Jacques in the Old Tibetan Documents Online database (Takeuchi 2013), for Burmese we used Nathan W. Hill's data that was (as indicated in the article by Hill and List 2017) using three different main sources to verify the old forms for Old Burmese (Luce 1985, Nishi 1999, Okell 1971), for Tangut G. Jacques and Y.-F. Lai carefully compiled the list using in particular a searchable text database compiled by G. Jacques (partially available as the supplementary materials for (Jacques 2016)).

Our ancient languages for calibration were chosen with great care, giving preference specifically to those varieties where our group has first-hand expertise, or close collaboration with experts whom we trust would allow us direct access to the data. As a result of our high demands regarding quality and being able to work directly with first-hand experts in the fields, our current selection of languages is not complete, although we think that it is sufficient in that it contains the most prominent archaic languages of the ST family. Although philological evidence is rich for languages like Newar, for example, it is much less easy to verify that we could obtain a high-coverage wordlist with good translations of the concepts in our data. Even modern sources of Newar often omit many concepts in their glossaries (see Section 3.4), thus failing to pass our coverage tests. We hope that future work will allow us to successively add more languages to the sample, and we are currently trying to establish connections with more experts who could help in this endeavour in the future.

3.2 Concepts in our sample

All concepts in our sample were linked to the Concepticon (List et al. 2016a), to allow for an easy comparison across other resources. Below, we list the original list of 250 concepts, indicating with help of an asterisk in the ID column, which concepts were not retained for the phylogenetic study, because their coverage in terms of languages was too low, or for additional reasons mentioned in the main text. We also list the coverage across all 50 languages in our sample, since we used general coverage across the data as a criterion to successively reduce the concept list. In addition, we list the corresponding identifiers used for the concepts in the *Tibeto-Burman lexicon* (Huáng 1992), a very large resource on Sino-Tibetan languages, which was digitized during the STEDT project (Matisoff 2015). As can be easily seen from the table: the coverage drastically differs between the concepts we retained and the concepts we discarded. The lowest coverage we observe for the concepts we retained is 88% of all 50 languages ("early", "eight", "nine"), with an average coverage of 97%. In contrast, among the 70 concepts we discarded, the average language coverage is 46%.

ID	English	TBL	Conc. ID	Conc. Gloss	Cov.
*1	above	731	1741	ABOVE	0.46
*2	all	962	98	ALL	0.47
3	the ant	365	587	ANT	1
*4	the armpit	92	1886	ARMPIT	0.49
*5	bad	1053	1292	BAD	0.61
6	the bamboo	389	1927	BAMBOO	0.98
*7	the barley (tibetan or highland)	411	932	BARLEY	0.25
8	to be alive	1087	1422	BE ALIVE	0.96
9	the belly	96	1251	BELLY	0.98
*10	below, under	732	1485	BELOW OR UNDER	0.45
11	big	964	1202	BIG	1
12	the bird	326	937	BIRD	1
13	to bite	1753	1403	BITE	0.98
14	black	1005	163	BLACK	1
15	the blood	129	946	BLOOD	1
*16	to blow (of wind)	1738	175	BLOW (OF WIND)	0.56
*17	the body hair (hair or fur)	266	189	HAIR (BODY)	0.52
18	the bone	133	1394	BONE	0.98
19	the branch	374	1531	BRANCH	0.9
20	the breast (female)	94	1402	BREAST	0.98
21	to burn [intransitive]	1269	1428	BURNING	0.96
22	to buy	1516	1869	BUY	1
23	to chew	1424	321	CHEW	0.94
*24	the child (young human)	169	1304	CHILD (YOUNG HUMAN)	0.47
25	the cloud	7	1489	CLOUD	1
26	cold (of temperature)	1063	1287	COLD	1
27	to come	1491	1446	COME	0.98
*28	correct (right)	1045	1725	CORRECT (RIGHT)	0.4
29	to count	1640	1420	COUNT	0.98
30	to cry (weep)	1485	1839	CRY	0.98
*31	dark	1013	706	DARK	0.48
*32	the daughter	222	1357	DAUGHTER	0.53
33	the dew	15	1977	DEW	0.92
34	to die	1651	1494	DIE	1
35	to dig	1698	1418	DIG	0.96
36	dirty	1086	1230	DIRTY	0.92
37	the dog	289	2009	DOG	1
38	the dream	699	2374	DREAM	0.98
39	to drink	1370	1401	DRINK	0.98
40	dry	1028	1398	DRY	0.98
41	the dust	45	2	DUST	0.92
42	the ear	81	1247	EAR	0.96

43	early	1018	672	EARLY	0.88
44	the earth (soil)	40	1228	EARTH (SOIL)	0.94
*45	the earthworm	363	2350	EARTHWORM	0.49
46	to eat	1198	1336	EAT	1
47	the egg	450	744	EGG	1
48	eight	804	1705	EIGHT	0.88
49	the eye	79	1248	EYE	0.98
50	far	974	1406	FAR	1
*51	the father	218	1217	FATHER	0.59
52	the feather	299	1201	FEATHER	0.9
53	to fight	1234	1423	FIGHT	0.96
54	the fire	18	221	FIRE	1
*55	firewood	534	10	FIREWOOD	0.46
56	the fish	351	227	FISH	1
57	five	801	493	FIVE	0.94
58	the flea	355	232	FLEA	0.88
*59	to float	1553	1574	FLOAT	0.43
*60	to flow	1502	2003	FLOW	0.43
61	the flower	377	239	FLOWER	1
62	to fly (move through air)	1318	1441	FLY (MOVE THROUGH AIR)	1
*63	the fog	16	249	FOG	0.56
64	the foot	103	1301	FOOT	1
65	the forest	50	420	FOREST	0.96
66	to forget	1704	1523	FORGET	0.96
67	four	800	1500	FOUR	0.94
*68	the fox	325	1312	FOX	0.32
69	the frog	349	503	FROG	0.98
70	the front (front side)	712	2194	FRONT (PART)	0.92
*71	the frost	14	2034	FROST	0.48
72	the fruit	378	1507	FRUIT	1
73	full	984	1429	FULL	0.92
74	to give	1345	1447	GIVE	0.98
75	the goat	276	1502	GOAT	0.92
76	good	1052	1035	GOOD	1
77	the grass	436	606	GRASS	0.98
78	green	1009	1425	GREEN	0.96
79	the hail	13	609	HAIL	0.9
80	the hair (of the head)	75	1040	HAIR	1
81	the hand	107	1277	HAND	0.98
82	hard	1034	1884	HARD	0.98
83	he or she [third person singular]	934	262	HE OR SHE OR IT	0.94
84	the head	74	1256	HEAD	1
85	to hear	1682	1408	HEAR	0.98
86	the heart	144	1223	HEART	1
87	heavy	1014	1210	HEAVY	0.96
*88	here	944	136	HERE	0.53

89	to hide (conceal)	1169	602	HIDE (CONCEAL)	0.98
90	high / tall	968	1265	HIGH	0.98
91	to hold	1709	1448	HOLD	0.96
*92	the hoof	264	152	HOOF	0.48
*93	horizontal	1001	2376	HORIZONTAL	0.47
94	the horn (keratinized skin)	263	1393	HORN (ANATOMY)	0.98
*95	the horse	268	615	HORSE	0.45
96	hot	1062	1286	HOT	1
97	the house	494	1252	HOUSE	0.98
*98	hundred	824	1634	HUNDRED	0.51
99	to hunt	1230	1435	HUNT	0.98
*100	the husband	247	1200	HUSBAND	0.49
101	I [first person singular]	928	1209	I	0.98
*102	the ice	17	617	ICE	0.52
103	inside	715	1606	INSIDE	0.98
104	to kill	1602	1417	KILL	1
*105	to knead	1594	274	KNEAD	0.43
106	the knee	101	1371	KNEE	0.98
*107	knife	549	1352	KNIFE	0.43
*108	to know (something)	1798	1410	KNOW (SOMETHING)	0.55
*109	the lake	31	624	LAKE	0.57
110	late	1019	477	LATE	0.94
111	to laugh	1735	1355	LAUGH	1
112	the leaf	376	628	LEAF	1
*113	to learn	1742	504	LEARN	0.49
114	left	710	244	LEFT	0.96
115	to lick	1674	319	LICK	1
*116	to lie down	1661	215	LIE DOWN	0.48
117	light (of weight)	1015	1052	LIGHT (WEIGHT)	0.98
118	the lip (the lips)	85	478	LIP	0.98
119	the liver	145	1224	LIVER	1
120	long	972	1203	LONG	1
121	the louse	356	1392	LOUSE	1
122	the lung	143	688	LUNG	0.98
123	the man (male human)	173	2106	MALE PERSON	0.96
124	many	987	1198	MANY	0.96
*125	to marry (a man marries a woman)	1578	2164	MARRY (AS MAN)	0.42
126	the meat	443	634	MEAT	1
127	middle	708	1093	MIDDLE	0.92
128	the moon	4	1313	MOON	0.98
129	morning	749	1339	MORNING	0.96
*130	the mosquito	360	1509	MOSQUITO	0.58
*131	the mother	219	1216	MOTHER	0.64

132	the mountain	23	639	MOUNTAIN	1
133	the mouse or rat	320	2139	MUROID (MOUSE OR RAT)	1
134	the mouth	84	674	MOUTH	0.98
135	the mud	46	640	MUD	0.92
136	the nail (fingernail or claw)	113	2128	CLAW OR NAIL	1
137	the name	687	1405	NAME	1
138	narrow	977	1267	NARROW	0.9
139	near	975	1942	NEAR	0.98
140	the neck	89	1333	NECK	1
141	the needle (for sewing)	578	1382	NEEDLE (FOR SEWING)	0.96
142	new	1050	1231	NEW	0.98
143	nine	805	1483	NINE	0.88
*144	the nit	357	267	NIT	0.33
*145	noon	750	12	MIDDAY	0.49
146	the nose	80	1221	NOSE	1
147	old (of person)	1058	2112	OLD (AGED)	0.96
148	one	797	1493	ONE	0.98
*149	the otter	317	15	OTTER	0.35
150	outside	714	762	OUTSIDE	0.96
151	the pig	284	1337	PIG	0.98
*152	to plant (vegetals, rice)	1774	1486	PLANT (SOME-THING)	0.43
153	to play	1703	1413	PLAY	0.98
*154	to pull	1568	1455	PULL	0.59
155	to push	1689	1452	PUSH	0.96
156	the rain	10	658	RAIN (PRECIPI-TATION)	0.98
*157	the rainbow	11	1733	RAINBOW	0.49
158	red	1007	156	RED	0.96
159	to reside (live)	1452	1099	RESIDE	0.94
*160	the rice plant	439	2026	RICE PLANT	0.39
161	right	711	1019	RIGHT	1
*162	the river	30	666	RIVER	0.55
163	the road	38	667	ROAD	0.98
164	the root	375	670	ROOT	1
165	the rope	619	1218	ROPE	0.96
166	round	990	1395	ROUND	0.94
*167	to run	1544	1519	RUN	0.49
168	the salt	61	1274	SALT	0.98
*169	salty	1076	1091	SALTY	0.49
170	the sand	44	671	SAND	0.94
171	to scratch	1530	1436	SCRATCH	0.98
*172	the sea	32	1474	SEA	0.46
173	to see	1471	1409	SEE	0.98

174	the seed	405	714	SEED	0.96
175	seven	803	1704	SEVEN	0.88
176	sharp	1020	1396	SHARP	0.96
*177	the sheep	275	1331	SHEEP	0.49
178	to shoot (an arrow)	1611	1172	SHOOT	0.98
179	short	973	1645	SHORT	0.92
180	the shoulder I	90	1482	SHOULDER	0.98
*181	shy E	1365	487	SHY	0.43
*182	the sickle T	624	341	SICKLE	0.46
183	to sing C	1184	1261	SING	0.94
184	six C	802	1703	SIX	0.88
185	the skin C	120	763	SKIN	0.98
186	the sky	1	1732	SKY	0.96
187	to sleep	1646	1585	SLEEP	0.96
188	small	965	1246	SMALL	1
189	to smell (perceive odor) [transitive]	1707	1586	SMELL (PERCEIVE)	0.92
190	the smoke	19	778	SMOKE (EXHAUST)	1
*191	smooth	1037	1234	SMOOTH	0.56
192	the snake	347	730	SNAKE	1
193	the snow	12	784	SNOW	0.9
194	soft	1035	1856	SOFT	1
195	the son	220	1620	SON	1
*196	the sparrow	336	1854	SPARROW	0.42
197	the spider	361	843	SPIDER	0.96
198	to spit	1688	1440	SPIT	0.98
199	to stand	1784	1442	STAND	1
200	the star	5	1430	STAR	1
201	to steal	1686	713	STEAL	0.96
202	the stick	586	1295	STICK	0.94
203	the stone (a piece of)	43	857	STONE	1
204	straight	1003	1404	STRAIGHT	0.94
205	the sun	2	1343	SUN	1
206	the tail	267	1220	TAIL	1
207	ten	806	1515	TEN	0.9
*208	that	947	78	THAT	0.52
*209	there	950	1937	THERE	0.52
210	thick	980	1244	THICK	0.94
211	the thigh	100	800	THIGH	0.9
212	thin (object)	981	2307	THIN (OF HAIR AND LEAF)	0.94
213	to think (reflect)	1726	1415	THINK (REFLECT)	0.98
*214	this	942	1214	THIS	0.53
215	thou [second person singular]	931	1215	THOU	0.94
216	three	799	492	THREE	0.98

217	to throw	1687	1456	THROW	0.98
218	the thunder	8	1150	THUNDER	0.98
*219	the tiger	304	846	TIGER	0.54
*220	today	738	1283	TODAY	0.53
*221	tomorrow	742	1329	TOMORROW	0.54
222	the tongue	139	1205	TONGUE	1
223	the tooth (front)	137	1380	TOOTH	1
224	the tree	372	906	TREE	0.96
*225	twenty	816	1710	TWENTY	0.38
226	two	798	1498	TWO	0.98
227	to vomit	1535	1278	VOMIT	0.98
228	to walk	1815	1443	WALK	1
229	the water	47	948	WATER	1
*230	we [first person plural inclusive]	930	1131	WE (INCLUSIVE)	0.53
231	wet	1029	1726	WET	1
*232	what	954	1236	WHAT	0.52
*233	the wheat	410	1077	WHEAT	0.46
*234	where	955	1237	WHERE	0.53
235	white	1006	1335	WHITE	1
236	who	953	1235	WHO	1
*237	the wife	248	1199	WIFE	0.53
238	the wind	9	960	WIND	1
239	the wing	298	1257	WING	0.92
240	to wipe	1163	1454	WIPE	0.96
*241	the wolf	324	522	WOLF	0.42
242	the woman	174	962	WOMAN	0.98
*243	the wood (material)	511	1803	WOOD	0.5
*244	to sow (broadcast, scatter seeds)	1597	748	SOW SEEDS	0.32
245	the year	777	1226	YEAR	0.98
246	yellow	1008	1424	YELLOW	0.96
*247	yesterday	739	1174	YESTERDAY	0.54
*248	you [second person plural]	933	1213	YOU	0.52
*249	young	1059	1207	YOUNG	0.46
250	the shit	151	676	SHIT (DEFECATE)	0.92

In the table below, we provide a detailed comparison regarding the number of concepts shared in our list of 180 concepts and other popular concept lists, including the list by Matisoff (1978) for Tibeto-Burman languages, the list by Blust for Austronesian languages (Greenhill et al. 2008), the classical lists by Morris Swadesh (Swadesh 1952, Swadesh 1955), the Leipzig-Jakarta list (Tadmor 2009), the alternative 100-item list by Sergey Yakhontov (Starostin 1991), the list of stable concepts proposed by the ASJP project (Holman et al. 2008), and the list of stable Tibet-Burman concepts proposed by Satterthwaite-Phillips (2011). As these lists (as well as our 180 item list) are all linked by the Concepticon project (List et al. 2016a), a direct comparison of the number of shared concepts is easy to achieve, and the supplementary source code and data show how this can be done.

Concept list	Shared	Proportion	Concepts
Blust-2008-210	120	0.57	210
Holman-2008-40	36	0.9	40
Matisoff-1978-200	105	0.5	210
Swadesh-1952-200	116	0.58	200
Swadesh-1955-100	78	0.78	100
SatterthwaitePhillips-2011-50	32	0.64	50
Tadmor-2009-100	67	0.67	100
Yakhontov-1991-100	84	0.84	100

Comparing the proportion of concepts shared with our selection of 180 concepts.

As can be seen from the results in the table below, our concept list reflects largely those basic vocabulary items that were also employed in different analyses before. Of our 180 concepts, 32 (i.e., 18%) do not directly recur in any of the lists provided in the table. However, given that the Concepticon concept linking is very fine-grained, based on very detailed concept definitions, there are at least 5 concepts that can be found in the lists but show slightly different definitions there. All "unique" concepts are listed in the table below, along with their Concepticon definitions and their Concepticon gloss. The items which recur in the popular concept lists are further marked with an asterisk.

232	FLEA	477	LATE
478	LIP	609	HAIL
640	MUD	672	EARLY
762	OUTSIDE	1093	MIDDLE
1099	RESIDE	1265	HIGH
1286	*HOT	1339	MORNING
1415	*THINK (REFLECT)	1502	GOAT
1606	INSIDE	1620	*SON
1977	DEW	2112	*OLD (AGED)
2194	FRONT (PART)	2307	*THIN (OF HAIR AND LEAF)

Concepts with no direct counterparts in the popular concept lists.

3.3 Cognate coding

A full reconstruction of all the branches in the sample, and of the ancestor of the family as a whole could have helped refining many of the cognate judgments in our data. However, Sino-Tibetan comparative phonology and morphology present specific difficulties that make more challenging in many aspects than Indo-European or Austronesian. It is still unclear whether we will ever be able to build a reconstruction system of comparable rigor for Sino-Tibetan, due to the combined effect of intense language contact and typological upheaval in this family. Moreover, even if we knew the sound laws and the historical morphology as well as in IE, this would not be enough to completely rule out the existence of undetectable loanwords (see for instance some undecided cases in Armenian, (Hübschmann 1897: 16f), (Jacques and List forthcoming)).

The cognate judgments used in this study, while 'subjective" in some cases in the sense that we cannot account for all sound correspondences for all languages in the sample, are however firmly grounded in the accumulated research on each of the languages in the sample.

The concrete cognate coding was carried out with help of the EDICTOR tool, offering straightforward and convenient ways to annotate cognates. To allow for collaborative editing, a server interface for EDICTOR was setup, which can be easily accessed from `https://digl.github.io/sinotibetan/`. For scholars interested in inspecting our cognate judgments in detail, we recommend inspecting our data via this website.

Compoundhood is a well-known obstacle for cognate coding (Ben Hamed and Wang 2006, Satterthwaite-Phillips 2011, Starostin 2013), since compounding is a very frequently recurring process in the Sino-Tibetan language family (Matisoff 2003b). The problem for phylogenetic analyses is that compoundhood creates patterns of *partial cognacy* (List 2016, List et al. 2016b), which are still only poorly handled in computational historical linguistics (List et al. 2016c). While EDICTOR allows to annotate partial cognate relations consistently (Hill and List 2017), current phylogenetic software packages cannot handle the complexity of partial homologies (List 2016). For this reason, we had to take great care in coding the cognate sets very consistently, making sure to avoid arbitrary decisions. Our strategy to deal with the problem were three-fold, involving the avoiding of concepts with high compoundhood, a consistent way to generate root cognate decisions from partial cognates, and to support difficult cases by constructing multiple phonetic alignments from cognate sets.

3.3.1 Avoiding concepts of high compoundhood

Based on data inspection, we identified a larger number of concepts which are usually lexified by compounds in the Sino-Tibetan languages. Examples for obvious cases include 'armpit' (Concepticon 1886), 'noon' (Conc. 12), or 'firewood' (Conc. 10). By giving preference in our basic list of 180 concepts to concepts that are less frequently expressed by compound words in the ST languages, we could avoid a couple of notoriously difficult cases.

3.3.2 From partial cognates to root cognates

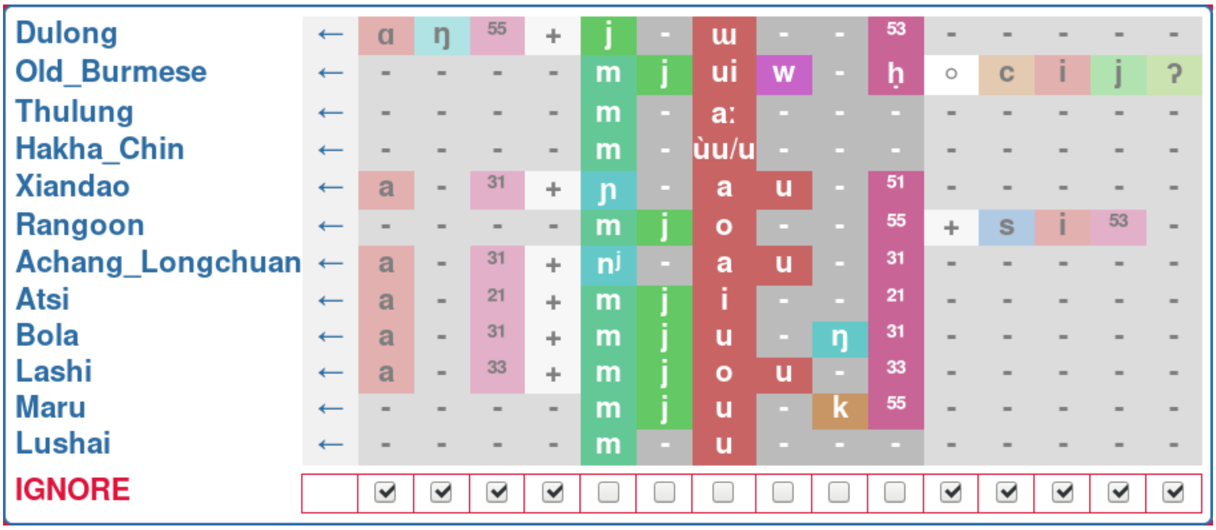
Subgroups of Sino-Tibetan which show a high degree of compounding, such as the Burmish languages, or the Chinese dialects (Sinitic) need a more specific treatment when carrying out cognate judgments for root cognates. Based on the morpheme-gloss annotation provided in Hill and List (2017), from whom we took the Burmish languages in our sample, we developed a way to allow for an unambiguous annotation of scholar's decisions regarding the main component of a compound word. This annotation adds an underscore (_) to all morpheme glosses of a words which are not considered to be central for the base meaning of a word. In EDICTOR, these cases are displayed by making the respective morphemes slightly transparent, and annotation is facilitated by allowing to toggle the central morphemes with a right mouse-click. Root cognates can be derived from this annotation by ignoring those partial cognate sets which are marked as not contributing to the main meaning. The following screenshot illustrates this annotation within the EDICTOR for the concept 'seed' (Conc. 714) and four Burmish languages.

ID	DOCULECT	CONCEPT	TOKENS	COGID	COGIDS	MORPHEME STRUCTURE
35559	Achang_Longchuan	the seed	a 31 + nʲ a u 31	127	1154 640	a-prefix + seed
31192	Old_Burmese	the seed	m j ui w ḥ ∘ c i j ?	127	640 1155	seed + s-seed
34807	Rangoon	the seed	m j o 55 + s i 53	127	640 1155	seed + s-seed
35564	Xiandao	the seed	a 31 + ɲ a u 51	127	1154 640	a-prefix + noun

Screenshot of the EDICTOR tool showing annotated lexemes for the concept 'seed'.

3.3.3 Supporting cognate judgments with alignment analyses

To make sure that we avoid intransitive cognate sets, in which one word A is partially cognate with a word B and the word B is partially cognate with a word C without being partially cognate with the word A, we used the alignment functionalities of the EDICTOR tool when preparing the data (for a general discussion of alignments and their importance for historical linguistics, see List et al. 2018b and List 2014). While we are often not (yet) in a stage where we can provide complete alignments of word forms, given that our field lacks the deeper understanding of many sound change processes involving the Sino-Tibetan language family, alignments for complex cognate sets containing words with many compounds make it much easier to make sure that no intransitive cognate sets have been annotated by the experts, since they allow to check quickly whether at least one morpheme is reflected in all words that are assigned to a cognate set. This is illustrated in the following screenshot for the full cognate set of 'seed', when inspected with help of the alignment editor provided by the EDICTOR tool.



A screenshot of the EDICTOR tool showing the fullly aligned cognate set for 'seed'.

While we cannot provide alignments for all cognate sets in our data at this stage, we made active use of the alignment function of the EDICTOR to verify that our cognate sets were transitive. In the future, we hope to edit the data further, to make sure that all cognate sets are also aligned.

3.4 Coverage and cognate statistics

One possible concern is that our data is insufficient for large-scale computational phylogenetic analysis, and researchers should wait until Sino-Tibetan is better understood. However, comparing the coverage and cognate statistics for common phylogenetic analyses of different language families allows us to locate the state of Sino-Tibetan reconstruction in comparison with other well-established language families.

We compare our data with cognate data for four different language families for which phylogenetic studies have been conducted: Austro-Asiatic (AA, Sidwell 2015), Austronesian (AN, Greenhill et al. 2008), Indo-European (IE, IELex), and Pama-Nyungan (PN, Bowern and Atkinson 2012). For our comparison, we use the data as provided by Rama et al. (2018).

Dataset	ST	AA	AN	IE	PN
Concepts	180	200	207	208	183
Languages	50	58	45	42	67
Words	9132	11827	9267	9854	12691
Coverage	0.94	0.90	0.79	0.95	0.89
Cognates	3501	3804	1872	2157	6495
Freq. Cognates	5	5	12	20	1

Coverage and cognate statistics comparison

The table above compares our dataset with these other datasets for a number of different statistics. These include the number of concepts, languages, and words in each datasets, the number of different cognate sets (*cognates*), the average mutual coverage (*coverage*), as defined by List et al. (2018a) and Rama et al. (2018), and the number of *frequent cognates* recurring in at least 10% of all languages in the sample.

This comparison shows that the Sino-Tibetan language data does not differ substantially in terms of coverage, number of cognates, and frequent words, from the data reported in these other datasets. Instructions on how the statistics can be calculated are provided in our supplementary data and code (see file `README.md` in folder `LexicalData`).

4 Phylogenetic analyses

4.1 Clades discussed in the paper

- Gyalrongic: Daofu, Wobzi, Maerkang, Japhug, Zhaba, Tangut
- Burmo-Gyalrongic: Lolo-Burmese, Gyalrongic
- Tibeto-Gyalrongic: Tibetan, Burmo-Gyalrongic
- Tibeto-Dulong: Tibeto-Gyalrongic, Dulong
- West-Himalayish: Rongpo, Byangsi, Bunan
- Kiranti: Kulung, Khaling, Thulung, Bahing, Limbu, Bantawa, Hayu
- Tani-Yidu: Yidu, Taraon, Bokar
- Kuki-Tangkhul: Lushai, Hakha, Ukhrul
- Kuki-Karbi: Kuki-Tangkhul, Karbi
- Sal: Rabha, Garo, Jinghpo

4.2 Phylogenetic constraints

Based on evidence from Old Chinese (Baxter and Sagart 2014a), we have a rather clear idea about the time when the Chinese dialects first separated. For this reason, we constrained the Chinese dialects to be a monophyletic group, with a MRCA in a uniform prior $[-2200\ -2000]$ YBP (including Beijing Chinese; Chaozhou Chinese, Guangzhou Chinese, Jieyang Chinese, Longgang Chinese, and Xingning Chinese).

4.3 Subgrouping results

The results obtained using the strict-clock covarion model and a Stochastic Dollo model are mostly compatible with those of the relaxed-clock covarion model discussed in the paper. Figures 1 and 2 give consensus trees for these two supplementary analyses.

The table below gives the posterior probability of each subgroups discussed in the paper. The following table shows the posterior probabilities of various subgroups under each model.[1]

Clade	Relaxed-clock	Strict-clock	Stochastic Dollo
Sal	0.96	1	1
Tibetan	1	1	1
Lolo-Burmese	1	1	1
Gyalrongic	1	1	1

[1] Tangkhul is represented by Ukhrul in our dataset, thus Kuki-Tangkhul refers to the branch consisting of Mizo, Karbi, and Tangkhul.

Burmo-Gyalrongic	0.98	0.99	1
Tibeto-Gyalrongic	0.98	0.97	0.57
Tibeto-Dulong	0.62	0.98	0.74
Kiranti	1	1	1
West-Himalayish	1	1	1
Tani-Yidu	0.99	1	1
Kuki-Tangkhul	1	1	0.89
Kuki-Tangkhul-Karbi	0.79	0.92	0.98

In particular, the following subgroups are overwhelmingly supported in all analyses (all posterior probabilities are > 0.95: Sal, Tibetan, Lolo-Burmese, Gyalrongic, Burmo-Gyalrongic, Kiranti, West-Himalayish, Tani-Yidu. We also have strong support for Kuki-Tangkhul and Kuki-Tangkul-Karbi subgroups in all analyses. There is also support for a hypothesized Tibeto-Dulong clade (and overwhelming support under the strict-clock model); note that in such a clade, the relaxed-clock and strict-clock models favour grouping the Burmo-Gyalrongic languages with the Tibetan languages, whereas the Stochastic Dollo model gives more uncertain output, with a 41% posterior probability that the Burmo-Gyalrongic languages group with Dulong first.

Note that the following subgroups, which have been proposed in the literature, have 0% posterior probability in all three analyses:

- Tibetan + Kiranti + Sinitic (Sino-Bodic hypothesis, Driem 1997),
- Sal+Kuki-Tangkhul (Central Trans-Himalayan hypothesis, DeLancey 2015),
- Gyalrongic+Kiranti+Dulong (Rungic hypothesis, Thurgood 2017),

On the Chinese side, although our tree shows close mutual genetic relations among Cantonese (Guangzhou), Hakka (Xingning, Longgang) and Min (Chaozhou, Jieyang), the probability for them to form a "Southern Chinese" clade, as Norman 1988 suggests, is rather low (0.42). Support for this alleged clade consists exclusively of words independently retained from Old Chinese, to the exclusion of shared innovations (Sagart 2011a).

As for Burmish languages, our result does not support Nishi 1999's classification, in which Burmese, Achang and Xiandao form the "Burmic" branch, and the other languages the Maruic branch. Our tree shows that Burmese languages alone are the first branch, as opposed to the rest of the Burmish languages.

4.4 Outgroups

In our main relaxed-clock analysis, the following clades are possible outgroups, with associated posterior probabilities of being the outgroup:

- Sinitic 33%
- West Himalayish 15%
- Tani-Yadu 9%
- Sinitic+Sal group 8%
- Sal 6%

Compared with the probabilities in our two alternative models: in the strict-clock model, the Sinitic group is the outgroup with 99% posterior probability. In the main TraitLab analysis, the possible outgroups are:

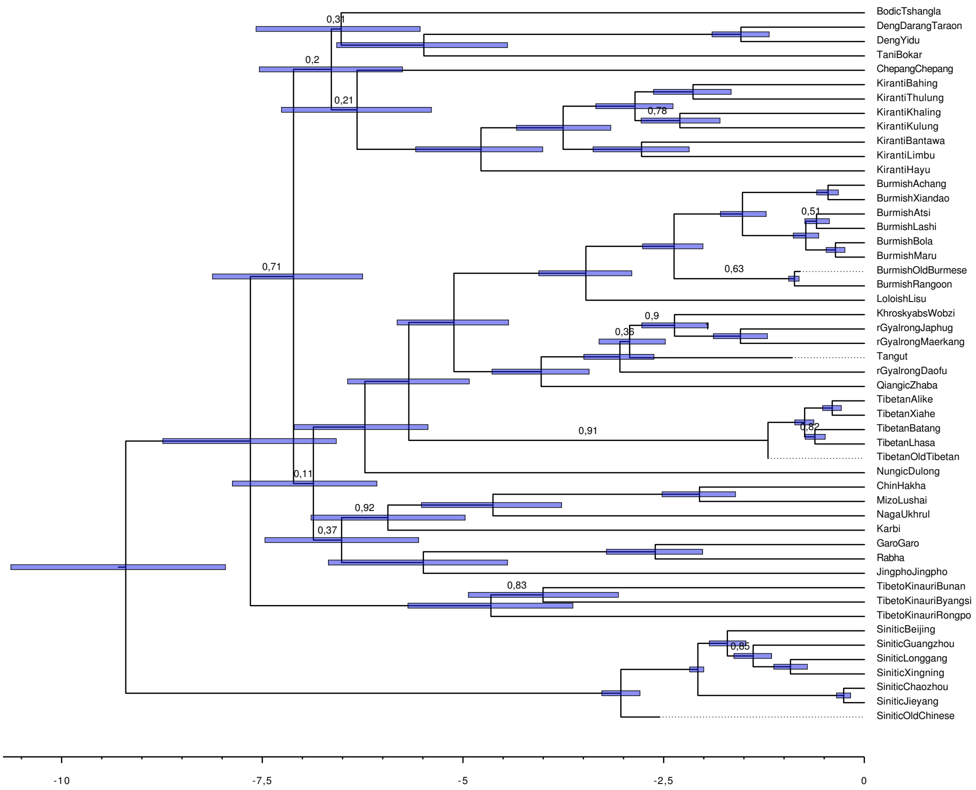


Figure 1: Consensus tree from the strict-clock covarion model analysis

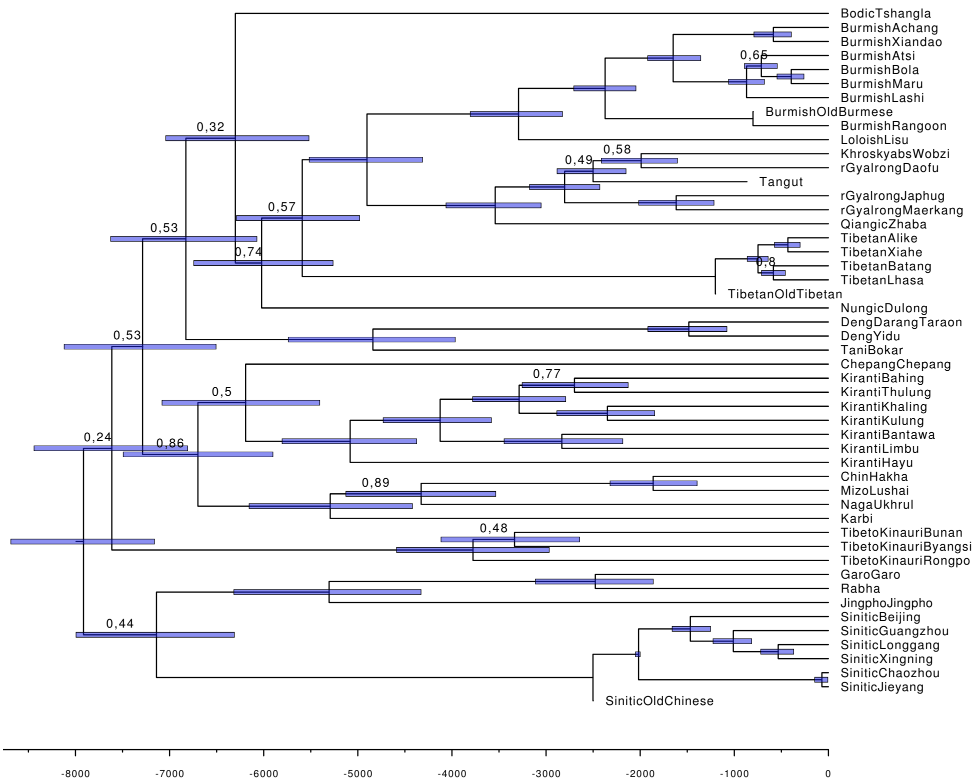


Figure 2: Consensus tree from the Stochastic Dollo model analysis

- Sinitic 43%
- West Himalayish 20%
- Sinitic+Sal group 16%
- West Himalayish + Sal 13%

The 54% probability for the Sinitic+Sal group in Fig. 2 of the main text indicates that Sinitic and Sal form a subgroup in 54% of the trees in the relaxed-clock analysis. This subgroup is the outgroup in only 8% of the cases. The subset of trees where Sinitic+Sal are a subgroup is mutually exclusive with that including trees with Sinitic as the outgroup. Therefore, 87% of the trees in the relaxed-clock analysis have either Chinese as the outgroup, or Chinese in a subgroup with Sal.

A possible way of interpreting these results in a way that is compatible with the *Sinitic outgroup scenario* is that the lexical commonalities supporting the Sinitic+Sal group in the relaxed-clock analysis are in fact common retentions exclusively shared by these two branches.

4.5 Root age

The root age estimated assuming a strict clock is at 9200 BP, with 95% HPD [8000 10600]. The root age estimated assuming a Stochastic Dollo model is at 7915 BP, with 95% HPD [7270 8650].

4.6 Age of Old Burmese

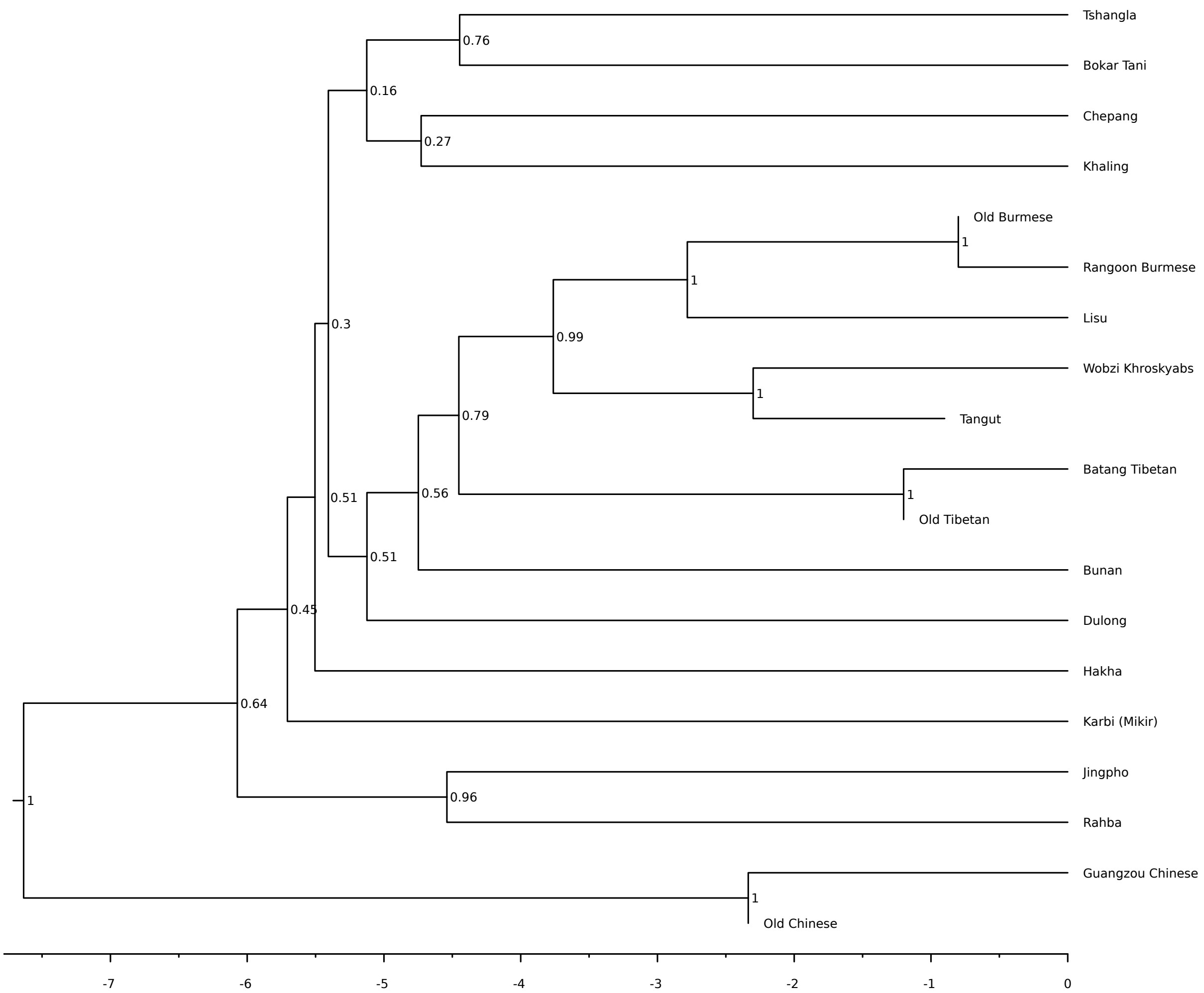
At the suggestion of an anonymous reviewer, we repeated the analysis under all 3 models with a different constraint for Old Burmese, allowing the age to be vary in the range 800 - 900 BP. This had no impact on the results: we observed no significant difference in the reconstructed topologies, and the reconstructed root ages are essentially unchanged. For example, under the Stochastic Dollo model, this analysis gives a root age of 7869 BP, with 95% HPD [7172 8509].

4.7 Analysis of a subset of the languages

Some subfamilies are represented by more languages than others, and the amount of missing data varies. To ensure that this does not bias our results, we repeated the analysis by including one representative of each subfamily (chosen to be the language with the least missing data in the subfamily), and all the ancient languages. We therefore used only the following 19 languages in this analysis: Rangoon (Burmish), Wobzi Khroskyabs (rGyalrong), Batang Tibetan (Tibetan), Dulong (Nungic), Bunan (Tibeto-Kinauri), Khaling (Kiranti), Bokar (Tani), Motuo Menba (Bodic), Chepang (Chepang), Karbi (Mikir), Guangzhou Chinese (Sinitic), Jingpho (Jingpho), Tangut (Tangut), Old Tibetan (Tibetan), Old Chinese (Sinitic), Old Burmese (Burmish), Lisu (Loloish), Hakha (Chin), and Rabha (Koch).

The results are substantially similar to those on the complete dataset. The posterior variances are higher, which is to be expected since we are using a smaller amount of data: we thus have larger HPDs in our age estimates, and more uncertainty in the inferred topologies.

The tree inferred from this analysis (relaxed clock model in BEAST2) is shown below.



BEAST2 tree produced with the subset of 19 languages and the relaxed-clock model.

The root age is estimated at :

- Stochastic Dollo model: mean 6930 BP and 95% HPD [6100 7710]
- Strict clock model: mean 8214 BP and 95% HPD [6745 9809]
- Relaxed clock model: mean 7617 BP and 95% HPD [4422 11051]

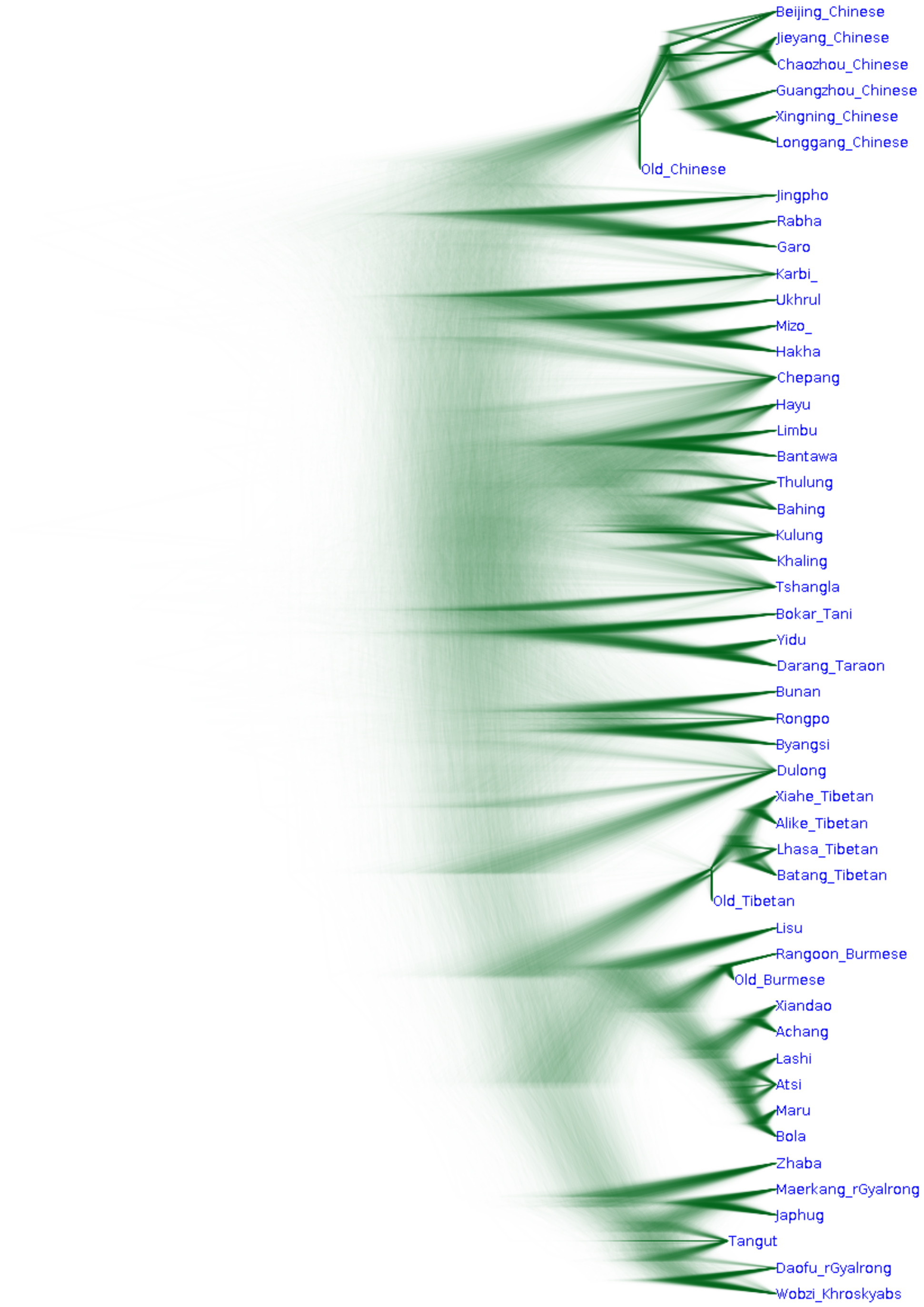
All HPDs have significant overlap with those of the main analyses.

This table summarizes the posterior probabilities of the subgroups of interest in the analyses of a subset of languages. The values in italics are those which differ by more than 0.10 from the main analyses.

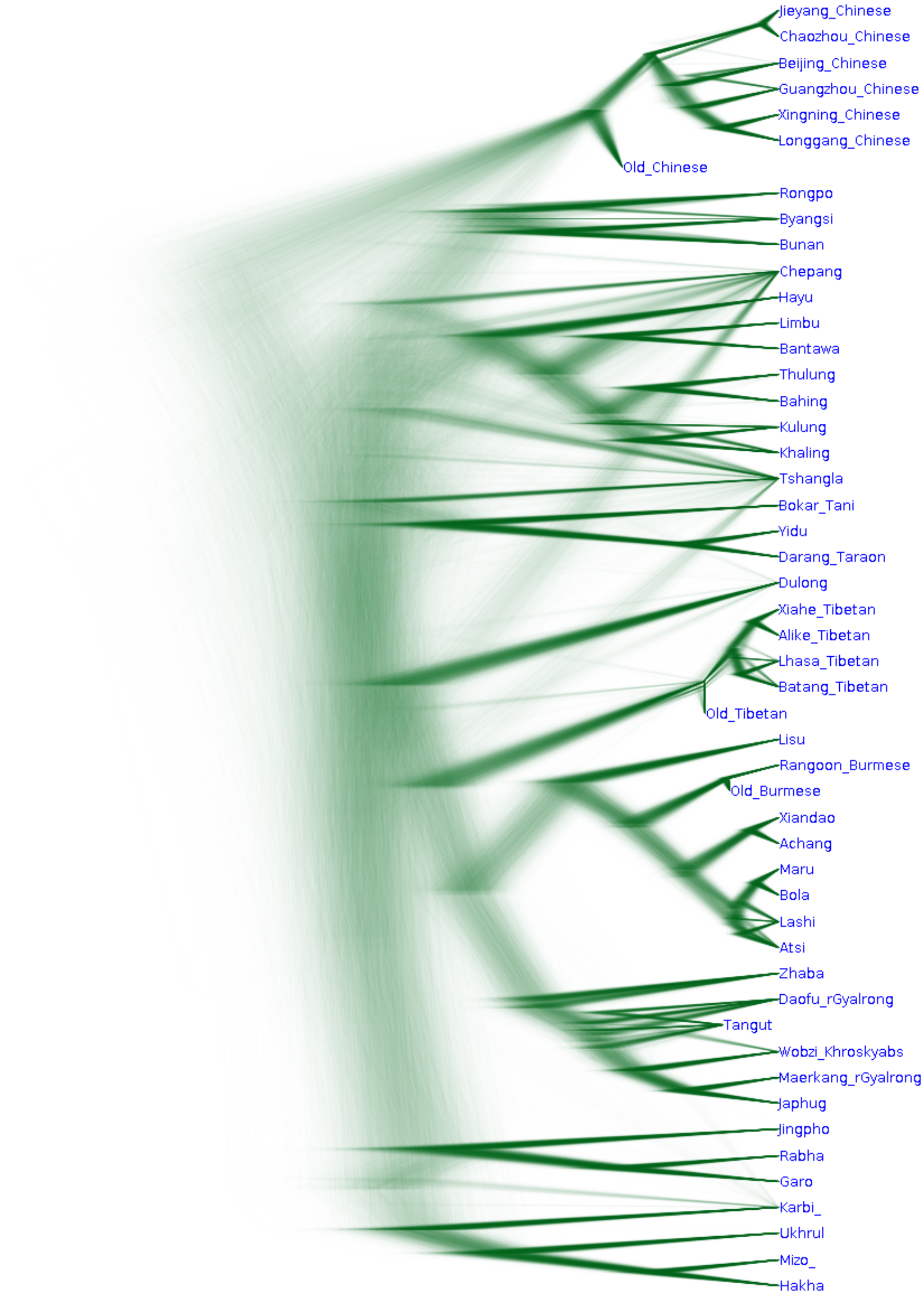
Clade	**Relaxed-clock**	**Strict-clock**	**Stochastic Dollo**
Sal	0.96	0.99	1
Tibetan	1	1	1
Lolo-Burmese	1	1	1
Burmo-Gyalrongic	0.99	0.99	1
Tibeto-Gyalrongic	*0.79*	0.94	*0.32*
Tibeto-Dulong	*0.27*	*0.63*	*0.44*
Kuki-Tangkhul-Karbi	*0.31*	*0.16*	*0.88*

4.8 Densitrees

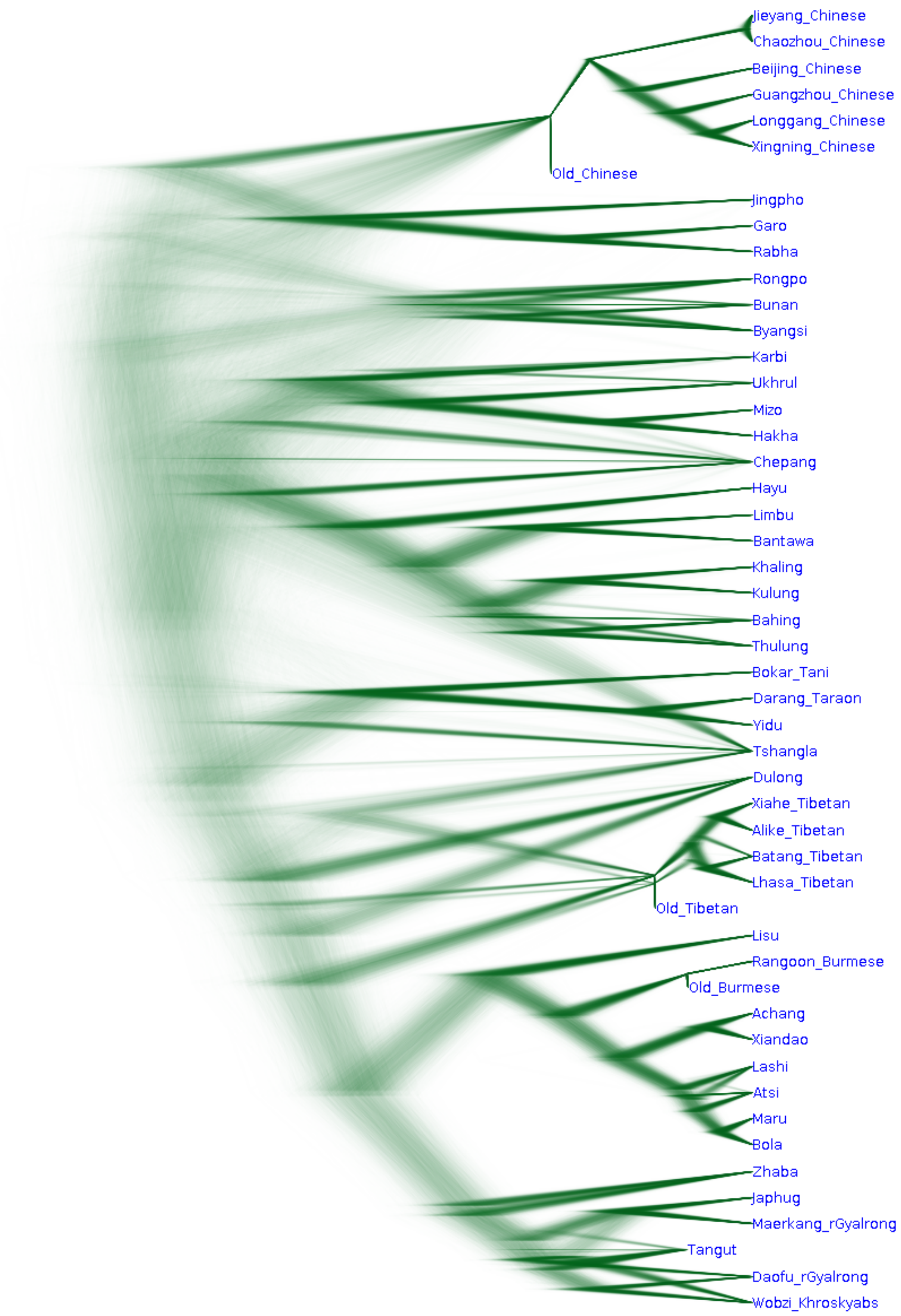
As an alternative representation of the reconstructed phylogenies and their uncertainty, we present Densitrees (Bouckaert 2010) for the three analyses in the following figures.



Densitree for the main analysis (relaxed-clock).



Densitree for the strict-clock analysis.



Densitree for the Stochastic Dollo analysis.

4.9 Etymologies supporting several proposed subgroups

4.9.1 Tibeto-Dulong

The ”Tibeto-Dulong” subgroup proposed in this paper is supported by a number of etyma shared by Dulong on the one hand, and at least language among Tibetan, Gyalrongic or Lolo-Burmese languages on the other hand. Among these etyma, some reflect roots attested elsewhere in the family, but not in the sample under investigation, such as the etyma “flower” (reflected by Dulong *ɕiŋ*^{55}*wɑt*^{55} and Lisu *si*^{35}*ve̱*^{33}), or “dry” (Dulong *kam*^{55} and Tibetan *skam.po* “dry”).

Potential cases of common innovations, exclusively shared by Dulong and Tibeto-Gyalrongic, include the following words:

1. Dulong *nɯi*^{55} "mouth" – Burmese *nhut*. This root is possibly shared with Karen languages (not included in our sample), as suggested by the STEDT (#471).

2. Dulong *tɯ*^{31}*wɑn*^{53} – Japhug *tɤjpa* (Japhug regularly loses final *-n and *-l, and Dulong -n can originate from either *-n or *-l). The STEDT (#471) contains comparisons with Tibetan, Chepang and Kaman words meaning "hail" and reconstruct a final *-l. The Tibetan forms *wal* (Amdo) and *kha*^{55}*wa*^{53} (Derge) included do not have final *-l: the former regularly come from Old Tibetan *bad* "frost" (with the sound changes -d > -l and b- > w-), and *kha.ba* "snow" respectively (note that *-ba* is a suffix). Tibetan *bad* "frost" is related to *ba.mo* "frost", and the coda -d is a nominal suffix. It is possible that Dulong *tɯ*^{31}*wɑn*^{53} "snow" and its cognates are related to the root of *bad/ba.mo* "frost", but with a different suffix. Schuessler 2007: 235 cites a comparison with Chinese 雰 p^hjun "mist", but in addition to the fact that Schuessler's gloss is probably wrong, as 雰 always appear reduplicated and is better analyzed as an ideophone meaning "fluttering", this word rather had final *-r in Old Chinese. Kaman *wɑ̆l*^{35} "hail" and Chepang *wer* "hail" have final consonants that rather originate from *-r (for instance Kaman săl^{53} "louse", cognate of Khaling *sēr* "louse", a language that preserves the contrast between the codas *-n, *-r and *-l), and since Dulong preserves *-r as -ɹ (see *ɑŋ*^{55}*ɕɑɹ*^{55} "new", cognate of Tibetan *gsar.ba* "new"), *tɯ*^{31}*wɑn*^{53} cannot phonologically correspond to either Kaman or Chepang (in addition to the meaning difference). Thus, the etymon "snow" is only found in Dulong and Gyalrongic, with a possible cognate in Tibetan; there are no cognates outside of the Tibeto-Dulong group.

3. Dulong *ɕin*^{55} "grass" –Japhug *xɕaj* "grass", Tangut *śjɨ* "grass" (*-n is lost in Japhug and Tangut, Jacques 2014)

Other possible Tibeto-Dulong exclusive cognates outside of our database include Dulong *dɯ*^{31}*gɹɯ*^{53} "sinew", Japhug *tɯ-ŋgru* "sinew", Tibetan *rgyus.pa* "sinew" (partially in STEDT #536, Jacques and Michaud 2011).

4.9.2 Tibeto-Gyalrongic

Potential innovations exclusively shared by Tibetan and Burmo-Gyalrongic (excluding borrowings) are the following:

1. Tibetan *sŋo* "blue, green" – Japhug *arŋi*, Burmese *ññuiv* (Jacques 2014: 163)
2. Tibetan *riŋ.po* "long" – Japhug *zri* (ibid.: 101)
3. Tibetan *ske* "neck" – Japhug *tɯ-mke* (Jacques 2004: 125)
4. Tibetan *gson.po* "alive" – Japhug *sɯsu* "alive"
5. Tibetan *rlon.po* "wet" – Wobzi *lú* "wet", Daofu *ɬəɬə* (ibid.: 148)

Additional exclusive Tibeto-Gyalrongic vocabulary not in the database include:

1. Tibetan *snas* "heddle", Japhug *ɕnat* "heddle", Burmese *hnat*. This highly technical weaving term is not borrowed from Tibetan into Japhug, as the cluster *ɕn-* is exclusively found in native words.

2. Tibetan *myong* "experience", Japhug *rŋo* "experience", (ibid.: 299)

3. Tibetan *phrin* "message", Japhug *tɯpri* "message". The prefix *tɯ-* in Japhug is a frozen indefinite possessor prefix.

4. Tibetan *gnyen* "relative, friend" (nominalized form of *nye* "near"), Japhug *tɯ-ɣɲi* "friend". This example is particularly significant, as reflects a morphologically complex etymon with similar structure and semantic specialization in both Tibetan and Japhug (*-n is lost in Japhug).

4.9.3 Burmo-Gyalrongic

The hypothesis of a genetic relationship between Lolo-Burmese and Gyalrongic languages has been previously discussed by several scholars, including Bradley (1997), Jacques and Michaud (2011), and Lǐ (1998).

The Burmo-Gyalrongic hypothesis is strongly supported by our analysis. Potential Burmo-Gyalrongic innovations in our database are the following:

1. Japhug *ɕʏɣ* "new" – Old Burmese *sac* 'new' (Jacques (2004: 190)). This etymon has cognates elsewhere in Sino-Tibetan, but in other branches the coda is nasal (for instance Chinese 新 sin < *siŋ)
2. Japhug *ɣɯrni* "red" – Old Burmese *nī* 'red' (Jacques (ibid.: 172))
3. Japhug *zdɯm* "cloud" – Old Burmese *tim* "cloud" (Jacques (ibid.: 185))
4. Japhug *ɯ-ʁʏri* "the front side" – Old Burmese *rheʔ* 'the front side' (Jacques (ibid.: 104)).
5. Japhug *tɯ-rtsʰʏz* "lung" – Old Burmese *ʔachut* 'lung' (Jacques (ibid.: 150)).
6. Japhug *tɯ-mɯ* "rain, sky, weather" – Rangoon *mo*[55] 'rain, sky'(Jacques (ibid.: 154))
7. Daofu *sme* "woman" – Old Burmese *minḥ-ma* 'woman'
8. Daofu *kvo*, Wobzi Khroskyabs *djú* "year" – Lisu *kho̱*[31] "year" (Jacques (2014: 101)).
9. Japhug *nɯqambɯmbjom*, Wobzi Khroskyabs *jmbjə̂m* "to fly" – Old Burmese *pjaṃ* 'to fly'. This root is related to Tibetan *'byam* "spread" and Chinese 泛 phjomH "float", with unidirectional semantic change "float" → "fly". The Japhug verb form has reduplication and additional prefixes, but related language have the simple verb root.

Jacques (ibid.: 305-306) lists a additional few phonetic and lexical innovations, notably the verb 'to be', *ŋu* in Japhug, *ŋǽ* in Wobzi Khroskyabs, corresponding to Proto-Burmese **ŋwa*[1] 'to be the case' (Bradley 1979 0698a). Jacques (2014: 305-306) suggests that the copular use of this verb is derived from an earlier meaning, namely 'to be true'.

5 Homeland and domesticates

There are two main traditions concerning the location of the ST homeland and the direction of ST expansions. One of them supposes a homeland west or southwest of the north China plain (Benedict 1975, Blench and Post 2014, Driem 2017, Haudricourt and Strecker 1991, Matisoff 1991, Starostin 2004). Authors within this tradition think early Sinitic speakers reached north China from the ST homeland in the west, superimposing themselves over indigenous population(s) who were the original domesticators of the main East Asian cereals: the pre-Hmong-Mien for rice (Driem 2017, Haudricourt and Strecker 1991), the pre-Altaics for millet (Starostin 2004). Accordingly Driem (2017) and Haudricourt and Strecker (1991) see the Chinese agricultural vocabulary as borrowed from Hmong-Mien, while Starostin (2004) envisions an Altaic contribution to the Chinese agricultural vocabulary. Some of the authors working within this paradigm envision a particularly close phylogenetic relationship between Sinitic and Tibetan (Blench and Post 2014, Driem 1997).

The present paper aligns with another tradition, which places the ST homeland in the eastern part of the ST domain (Bradley 2018, Janhunen 1996, LaPolla 2001, Sagart 2011c, Thurgood 2008). Being indigenous to north China, early Sinitic speakers did not have to migrate from a far-away location in the (south-)west, and no substratum language is assumed under Sinitic. Early Sinitic speakers are the only stay-at-home branch of the ST family. Like the rest of ST, their demographic expansion ultimately results from their possession around 7000 BP of the domesticated millets *Panicum miliaceum* and *Setaria italica*, pigs, and sheep, as well as morphologically wild, but managed taurine cattle. The non-Sinitic part of ST, generally viewed as monophyletic by these authors, did expand west and south, their progress perceptible in the archaeology through the progression of millet farming out of Majiayao culture (Guedes 2011). That the spread of millet is explained not culturally, but by a demic expansion, follows from the gradual north-to-south decrease in Y-chromosome O3-haplogroup diversity between present-day non-Sinitic ST speakers in northern Sichuan and ST speakers in the eastern Himalayas (Kang et al. 2011). Cline-like decrease in genetic diversity is ascribable to genetic drift caused by repeated founder effects in the course of a migration.

The eastern homeland hypothesis makes sense of the consilience between the domesticates foxtail millet, pigs and sheep, archaeologically attested at our tree's root date, and the cognate sets for the same species, described below. Although the western homeland hypothesis at first sight agrees better with the modern distribution of linguistic diversity in the family, we argue that diversity in the eastern Himalayan region is not original; not any more than the high linguistic diversity in the Austronesian languages of Melanesia and Eastern New Guinea, which led Dyen (1963: 83) to mistakenly place the Austronesian homeland there. High Austronesian linguistic diversity in that region most likely results from intimate contact between highly diverse preexisting Papuan languages and the incoming Austronesians. Similarly, high linguistic diversity among the ST languages of the eastern Himalayas is the fruit of intimate contact between expanding ST speakers and highly diverse languages of non-ST hunter-gatherers to whom the region had served as a linguistic refuge. Conversely, the low degree of diversity of ST languages in modern eastern China is recent: classical Chinese texts indicate a much higher degree of diversity in early historical times (Pulleyblank 1983); the present situation is the result of leveling caused by Chinese expansion and subsequent language shifts to Chinese. As to the appearance of a close phylogenetic relationship between Chinese and Tibetan, it is an illusion due to their being old literary languages, with a much more thorough documentation than all other ST languages; in addition they are the oldest documented ST languages: even though they belong to distinct primary branches of the family, the patristic distance between them on the ST tree is shorter than that between any other pair of ST languages. While Tibetan and Chinese do share more lexical material, that material consists entirely of shared retentions: it does not argue for a close genetic relationship.

Blench and Post (2014) cite the case of ST-speaking groups in the Eastern Himalayas who do not grow cereals but rely on sago and livestock. They take this as evidence that the ancestral ST speakers were not farmers. However, instances of East Asian farmer groups reverting to a non-agricultural life-style are documented (Oota et al. 2005, Pierron et al. 2014, Reid 1992): the principle "once a farmer, always a farmer" is not a reliable one.

5.1 Lexical sets

The following table lists etymologically relatable words for early terms of agriculture and domestication which are reflected in some major ST branches, although the rice plant and the horse clearly were not part of PST. We have left out the name of the dog, a paleolithic domesticate widespread in the old world at the time of millet domestication; the names of barley and wheat, which reached the Yellow River, Gansu and Tibet separately from the Eurasian steppes along parallel north-south routes between 4600 and 3600 BP (Long et al. 2018); and etyma for beans, tubers and the like, which may refer to plants collected in the wild by PST speakers. PST food production also relied on fishing, with a PST etymon for the fish-net (Sagart 2011c); and probably hunting too. We do not know a solid cognate set for *Panicum miliaceum*, one of the two millets we assume PST speakers grew. We believe this is due to the name of this cereal being under-recorded by fieldworkers.

Term	**Old Chinese**	**Tani-Yidu**	**Kiranti + Lhokpu**	**Sal**	**Kuki-Karbi**	**Tibeto-Dulong**
foxtail plant	稷 *[ts]ək		Lhokpu *cək*			Dulong *tɕaʔ*^{55}
rice plant		Bengni *am*		Sak *aŋ*		Dulong *am*^{55}
field	田 *lˤiŋ			Dimasa *ha-bliŋ*		Tibetan ʑiŋ.ka
pig I	豕 *l̥ajʔ	Bengni *rjuɯk*				
pig II	(富 *pək-s "wealth")		Limbu *phak*	Jingpo *waʔ*^{31}	Lushai *vok*	Japhug *paʁ*
sheep	羊 *ɢaŋ	Taraon *kɯ*^{31} *joŋ*^{35}				Japhug *qaʑo*
horse		Taraon *mɑ*^{31} *ɹoŋ*^{55}		Sak *mərâŋ*		Japhug *mbro*
cattle	牛 *[ŋ]ʷə		Limbu *saːŋwa*	Jingpo *ŋa*		Japhug *nɯŋa*

In the above table, the language Hlokpu of Nepal is tentatively placed together with Kiranti based on observations by Gerber, Gerber and Grollmann. OC forms are cited in the Baxter-Sagart system (Baxter and Sagart 2014b); Japhug forms are drawn from Jacques (2015–2016). The cognate set for "rice plant" was assembled in Sagart (2018a) where references can be found. The set for pig I is new. Many forms in the table, such as (Classical) Tibetan *ʑiŋ.ka* "field", Lushai *vok* "pig", Bengni *rjuɯk* "pig" are cited from standard dictionaries, or from monographs like Sun (1993); yet others others are drawn from the STEDT web site (Matisoff 2015), in particular from the inclusive STEDT sets #2406 PTB **b-liŋ* FOREST / FIELD, #1006 PTB **pʷak* PIG, #6028 PTB **g-ya(k/ŋ)* SHEEP / YAK, #1431 PTB **s/m-raŋ* HORSE, and #2538 PTB **ŋwa* CATTLE. By citing forms included in STEDT we are not necessarily expressing support for the validity of the relevant STEDT sets in their entirety, or for the reconstructions which accompany them.

To summarize, a late Cishan-early Yangshao homeland combined with the Chinese outgroup scenario largely agrees with the facts: cognate sets are attested in and outside of Sinitic when the corresponding domesticate is archeologically attested in the late Cishan-early Yangshao area; while those cognate sets found only outside of Sinitic are not attested archaeologically in the late Cishan-early Yangshao area.

	archaeologically present in Cishan/Yangshao	cognates inside and outside Sinitic
foxtail millet	+	+
broomcorn millet	+	(insufficient data)
pig	+	+
sheep	+	+
rice	-	-
horse	-	-
cattle	-	+
wheat	-	-
barley	-	-

Cognate sets for domesticated species, attested inside and outside of Sinitic

The main discrepancy concerns the term for cattle, found both inside and outside of Sinitic, even though domesticated cattle is not attested in late Cishan-early Yangshao. An explanation for this apparent anomaly is proposed in section 5.10.

5.2 Archaeological dates

Currently the earliest East Asian archaeological attestations for the domesticates in the table above are as follows:

- foxtail millet (Setaria italica): Cishan and Yangshao culture, 8500-5000 BP (Stevens and Fuller 2017);
- broomcorn millet (Panicum miliaceum): Cishan culture, 10,300-8700 BP (Lu et al. 2009);
- rice: Baligang, Henan province, 8700-8300 BP (Deng et al. 2015);
- sheep: Shihushan, Inner Mongolia, 6700-6400 BP (Dodson et al. 2014);
- horses: Qijia culture, 4200-3600 BP (Flad et al. 2007);
- pigs: Nanzhuangtou, Hebei province 10,500 BP (Xiang et al. 2017);
- cattle: Shizhao Village site, Tianshui city, Gansu Province, 5400-4700 BP (Cai et al. 2014).

We provide below some etymological and comparative notes on these etyma.

5.3 Foxtail millet (*Setaria italica*)

The term “millet” designates botanically disparate grain-bearing plants, belonging to several distinct genera within the family Poaceae, having in common to produce very small grains. Among the world’s oldest cereals are two millets, both domesticated in northern China: *Panicum miliaceum* and *Setaria italica*. These are visually and botanically very different plants: no confusion between them is possible. Each has synonyms which it is important to recognize. Common synonyms for *Panicum miliaceum* are “broomcorn millet”, “proso millet”, “common millet”, “panicled millet”. The main synonym of *Setaria italica* is “foxtail millet”, so called because its ear is shaped like the tail of a fox.

Much confusion exists in the literature on ST millet names. This is due in part to the fact that fieldworkers, comparative linguists and compilers of dictionaries or language atlases often take “millet” to be a meaningful taxonomic notion, and assume that the differences between different kinds are of little consequence. In fact, to those who cultivate them, foxtail and broomcorn millet are as different as are dogs and pigs. In effect, words glossed simply as “millet” are useless for comparative purposes. An example of a botanically (and linguistically) naive reconstruction of a millet term is the (thankfully provisional) STEDT etymon #5860 PLB *C-lu-k MILLET which draws together Lolo-Burmese forms like Written Burmese lu^3 “Panicum miliaceum” and the Chinese word 秫 *m.lut “glutinous foxtail millet”. The STEDT author added a *-k suffix of no particular function, apparently to explain the final consonant in Chinese: but Old Chinese *-t cannot be derived from an earlier *-k. In fact “glutinous”, not “millet”, is the semantically relevant part of the Chinese word: 秫 *m.lut is cognate with Written Tibetan lud “phlegm, mucus”, being a ST term meaning “sticky, mucilaginous”.

The cognate set for *Setaria italica* in 5.1 was first assembled in its outline in Sagart (2005). Sagart et al. (2017) added the Dulong form (widespread in Nungic: Rawang *saʔ*, Nung *tɕʰɛ³¹*). In that paper the Chinese names of the millets are discussed, and the clear evidence for 稷 *[ts]ək being the OC name of *Setaria italica* is for the first time laid out: despite claims in the literature from the 10th century CE until today that Old Chinese 稷 *[ts]ək referred to *Panicum miliaceum*, Sagart et al. (ibid.) show that this date corresponds to the phonological convergence as [tɕi] (pin-yin *jì*) of Old Chinese 稷 *[ts]ək *Setaria italica* and 穄 *[ts][a][t]-s *Panicum miliaceum* in a large area of northeastern China. Since 穄 *[ts][a][t]-s was undoubtedly a name of *Panicum miliaceum*, 稷 *[ts]ək is the only candidate for the name of archaeologically prominent *Setaria italica*.

Dulong *tɕɑʔ* and Hlokpu *cək*, both identified as foxtail millet by the relevant fieldworkers, are almost certainly cognates of 稷 *[ts]ək. All three items fit known sound correspondences and refer to precisely the same plant, genus and species. This is in all likelihood the PST word for *Setaria italica*.

Bradley (2011) argues that Old Burmese tɕhap "*Setaria italica*" is a cognate of OC 稷 *[ts]ək. However, as he notes, the correspondence between final -p in Old Burmese tɕhap and *-k in Chinese and Lhokpu, and with -ʔ in Dulong, is unexplained. Elsewhere he (Bradley 2017) supposes that PST final *-p changed to *-k in the Old Chinese word 稷 *[ts]ək–this shift is attested in a few forms of what appears to be a western subdialect of OC; however Lhokpu also has -k in its word for *Setaria italica*, and Dulong -ʔ reflects *-k, not *-p, which invalidates this idea. If the Burmic and Chinese words really are cognate, one must suppose an irregular change of PST *-k to *-p *inside Burmic*, for instance due to place assimilation on a lost suffix with labial initial.

5.4 Broomcorn millet (*Panicum miliaceum*)

Bradley (ibid.) compares the Chinese word 稻 *[l]ˤuʔ "rice plant" with Written Burmese lu^3 *Panicum miliaceum* (Bernot et al. 1998), claiming that this is the PST etymon for *Panicum miliaceum*. This is doubtful: first, ethnobotanists Watanabe et al. (2007) recorded *luu* as the word for finger millet (*Eleusine coracana*) in upper Burma, so this term's meaning in Burmese is not entirely clear; second, there is no evidence at all that the Chinese term ever referred to a millet. It is not even clear that 稻 *[l]ˤuʔ was originally the name of a plant; the oldest tokens of this character had the semantic determinative 米 "dehusked grain" instead of 禾 "cereal plant" in the modern character, so 稻 may have been a word for grain in storage in early Old Chinese. Moreover, although this comparison is phonologically regular, it is limited to a consonant and a vowel. If the resemblance is not accidental, the PST word was more likely a generic term for grain at a certain stage of processing than the name of any specific domesticated plant.

Owing to the dearth of unambiguously recorded names of *Panicum miliaceum*, we are still not able to identify a PST etymon for this plant.

5.5 Rice

Based on the comparison between Old Chinese 米 *C.mˤ[e]jʔ "millet or rice grains, dehusked and polished" and Proto-Bodo-Garo (Joseph and Burling 2006) *mai 1 "rice, paddy, cooked rice", Sagart (2011b: 124) argued that the speakers of PST were acquainted with rice. We now recognize that this is problematic, as the Chinese word is a general term for dehusked grains, not specifically rice. Neither, for the same reason, does the comparison between Old Burmese kok "rice" and OC 穀 *[k]ˤok "grain (in the husk)" (Bradley 2011) support the view that PST speakers knew rice. Cognates of the Burmese form outside of Chinese (STEDT #586) indicate that "husk" or "grain in the husk" was part of the etymon's PST meaning. There is therefore no linguistic evidence that the ancestral Sino-Tibetans knew rice. This is is parallel to the absence of rice archaeologically in the Cishan and early Yangshao cultures. The non-Sinitic (Bengni, Sak and Dulong) forms in 5.1 constitute a phonologically regular set of words specifically designating the rice plant, with other forms in rice-related meanings elsewhere in non-Sinitic ST (Sagart 2018b). This is clearly a specialized form of the unproblematic STEDT set #487 PTB **ʔam* EAT / DRINK, reflected as "eat" in Kiranti, Karenic and Dulong and as "drink" in Dhimal, an unclassified ST language of Nepal. The probable Chinese cognate: 飲 *q(r)[u]mʔ "to drink" shows the same semantics as Dhimal: together these forms suggest a PST verb "to eat liquid foods, such as gruel". We may thus be dealing with an innovation inside the non-Sinitic part of ST: this would reflect the establishment of rice consumption in the form of gruel. It does not signal the beginning of Sino-Tibetan acquaintance with rice, even less a *de novo* domestication.

Driem (2017: 204-207) presents a scenario in which rice was domesticated three times "in the region between the Brahmaputra river basin and the Yangtze river basin": *Aus* and *Indica* rices by the ancient Austroasiatics and Hmong-Miens, and *Japonica* by a group he calls "para-Austroasiatic" who "disseminated rice agriculture to the lower Yangtze". He accounts for the fact that all domesticated rice types by and large share the same set of domestication genes by supposing mutual transfer of useful genes between already domesticated rice varieties. He further claims (ibid.: 206) that the Hmong-Miens adopted rice agriculture from the Austroasiatics. There are several problems with this scenario. First, archaeologically, the region between the Brahmaputra and the Yangtze lies at the *recent* end of a clear cline of dates for neolithic transitions in East and south Asia (Cobo et al. 2019): the earliest archaeobotanical evidence for rice domestication in progress, at 8700-8300 BP is in Baligang, Henan, i.e., north, not south, of the Yangtze valley, in fact half-way between the Yangtze and Yellow river valleys (Deng et al. 2015). That rice was under domestication there in spite of its small grain size follows from the observation that 80 percent of spikelet bases were of the non-shattering type. Elimination of shattering is an important target of rice domestication. The second oldest domestication sequence, also observed through the spikelet base paradigm, took place in the period 6900-6600 BP (Fuller et al. 2009) in the *lower* Yangtze, very far from the Brahmaputra. There is no evidence at all of domestication in progress in van Driem's zone at comparable dates. Second, rice geneticists usually regard *Japonica* as the first domesticated rice variety and the main donor of domestication genes to *Indica* and *Aus*, although *Aus* rice also received some domestication genes from *Indica* (Choi et al. 2017). Third, linguistically, if the Hmong-Miens truly acquired rice cultivation from the Austroasiatics, the two groups should share at least some vocabulary of rice cultivation: but van Driem cites no such vocabulary. In fact Sagart (2011b) showed that the Austroasiatic vocabulary of rice is entirely independent from all other East Asian rice vocabularies, including Hmong-Mien. Fourth, agronomically, by van Driem's theory, the types of rice adopted by the Hmong-Miens from the Austroasiatics should be *Aus* and *Indica*: but evidence that this is the case is missing. It is generally assumed that traditional *Aus* and *Indica* landraces are not cultivated outside of south Asia and of the lowland regions of mainland and insular southeast Asia.

5.6 Pig

We provide two sets for "pig" in 5.1. Pig I includes 豕 *l̥ajʔ, the main OC word for "pig, swine", and Bengni *rjɯk*, from Proto-Tani *rjek "pig" (Sun 1993: 199). The Baxter-Sagart reconstruction OC *l̥ajʔ should properly have been formulated as *l̥[aj]ʔ, to allow for the alternative reconstruction *l̥eʔ, also admissible under the Chinese-internal evidence at hand. Proto-Sino-Tibetan *-q regularly gives Proto-Tani *-k and OC *-ʔ (Sagart 2017); and Proto-Tani merges *lj- and *rj- as *rj- (Sun 1993: 292), resulting in a sound correspondence between proto-Tani *rj- and OC laterals: e.g., "pig", "bow (weapon)", "fathom (n.)", "to lick", Proto-Tani (ibid.) *rjek, *rji, *rjam, *rjak, Old Chinese (Baxter and Sagart 2014a) 豕 *l̥ajʔ , 矢 *l̥i[j]ʔ ("arrow"), 尋 *sə-l[ə]m ("measure of eight feet"), 舐 *Cə.leʔ. Proto-Tani *e is relatively rare, so that examples supporting the correspondence of Proto-Tani *e to OC *e cannot be numerous, but one can cite Proto-Tani *ken "to know" : Old Chinese 見 *[k]ˤen-s "to see" and Proto-Tani *jem "satiated/tired of" : Old Chinese 猒 *ʔem "satiate, satisfy". The Tani word for "pig" is treated by Sun as a loan from Proto-Mon-Khmer, e.g. Old Mon clik "pig", but the existence of a Chinese cognate shows that Mon-Khmer is on the receiving side.

The pig II set in 5.1 is reflected as "pig" only in the non-Sinitic languages. Several authors, e.g. Schuessler (2007: 32) consider Chinese 豝 *pˤra "sow" to be cognate with the non-Sinitic forms. However, just as likely, 豝 *pˤra "sow" can be compared to Written Tibetan ba "cow" as the name of a large female mammal. Similarly we speak of whale "cows". Sagart (2011) argued that 富 *pək-s "rich; wealth" is cognate with the non-Sinitic words for "pig" in the pig II set. This comparison fits all known sound correspondences. Etymological contacts between words meaning "cattle" and "wealth" are cross-linguistically not rare (e.g. Latin *pecunia* "wealth" from *pecu* "cattle"; Arabic *maːl* "cattle, wealth"). It is undecidable *a priori* whether the relevant etymon meant "pig" or "wealth" in the ancestral language: however, considering that the pig I and pig II etyma coexist in Chinese, the semantic difference between them in Chinese probably reflects the PST semantics: the pig I etymon was then the animal's name, while the pig II etymon meant "wealth in pigs, pigs in one's possession". The pig II set then displaced pig I in some ST branches.

5.7 Field

The main part of the cognate set for "field" in 5.1 is drawn from the reliable set STEDT #2406, with members in Chinese, Sal, Tibeto-Dulong and Lepcha. The semantics oscillate between "forest" and "field", pointing to an earlier meaning of "forest swidden", appropriate for millet cultivation in East Asia. Driem (2011) claims this is a loan from Hmong-Mien to his "Sino-Bodic". We show in this paper that, with zero posterior probability, "Sino-Bodic" (Chinese + Tibetan + Kiranti) is not a plausible clade. Consequently, if van Driem is right about the ST forms being loanwords, a minimum of four distinct borrowing events out of Hmong-Mien are needed: to Sinitic, Tibeto-Dulong, Sal and Lepcha. Only Sinitic is in contact with Hmong-Mien. Second, Sinitic does not distinguish between different kinds of fields, the only word being 田 OC *lˤiŋ. Economically the main cereals in early China were the two millets: rice was marginal. Rice is possibly mentioned once in the Shang inscriptions (oracle bone inscription 13505 in the Jiaguwen Heji collection (yanjiusuo 1978), where the noun 秜 *nrəj > nrij > lí "perennial rice" is interpreted by Liu (2005: 441) as a verb meaning "to plow paddies". Verbal use of a cereal name as a verb meaning "to plant, cultivate X" has parallels with 黍 *s-tʰaʔ > syoX > shǔ "Panicum miliaceum"). The near-absence of terms for "rice" in Shang oracle bones contrasts with the many occurrences of words for the millets. It is not likely that early Sinitic speakers would have replaced their inherited word for "(millet) field" with a word designating the rice field in a neighboring language. Van Driem's proposal is part of a larger claim that the Chinese vocabulary of rice consists of borrowings from Hmong-Mien. These claims, and the similar claims of Haudricourt and Strecker (1991), were discussed and rejected in Sagart (1995) and Sagart (2011b).

5.8 Sheep

In addition to the forms cited in 5.1, the cognate set for "sheep" includes Written Tibetan g.yang (in g.yang dkar "sheep"; dkar means "white") and Dulong *ɑ*[31] *ɟɑ̆ŋ*[53] "sheep". Bradley (2016) takes the Chinese word 羊 *ɢaŋ to mean "goat" rather than "sheep", but text occurrences as "goat" are usually accompanied by a modifier such as 山 *s-ŋrar "mountain". This set appears to be phonologically regular, although the overall regularity of correspondences between Old Chinese initial *ɢ and the presyllabic formatives Japhug *qa-*, Taraon *kɯ*[31], Dulong *ɑ*[33] and Tibetan g. needs confirmation. With this important caveat, the etymon could be part of the Sino-Tibetan proto-language given the early date of its first appearance in the archaeological record (5.2) "in a domestic setting where millet was grown" (Dodson et al. 2014).

Bradley (2016) claims that "the goat was a local wild animal before its domestication, but the sheep (*Ovis aries*) was introduced from the Middle East fairly early, probably about the same time as the domestication of the goat". This statement implies that goats were domesticated locally. We do not know what the basis for this is. There is general agreement in the literature that both species were domesticated in the Middle East/western central Asia and later introduced to north China. Current evidence argues that sheep were present in northern Shaanxi, at the northern edge of the Yangshao area, in the period 4700-4300 BCE (Dodson et al. 2014). These dates are reliable, based as they are on three direct radiocarbon dates from a single location. The authors write: "Since the bones were found in association with other domestic species and in an archaeological setting of the Yangshao Culture it is a reasonable conclusion that the sheep were domesticated." In addition their analysis of bone collagen shows that these animals consumed some millet, suggestive of a domestic setting. We find this argument reasonable.

The first STs must have been familiar with the takin (Budorcas taxicolor), a wild member of the Caprinae subfamily, present in the Cishan-Yangshao area: however this animal's name has been sparsely recorded. We know of no evidence indicating that this animal's name in PST was the etymon we give in S5.1. One of course cannot specifically exclude that our etymon originally referred to a locally known wild animal before expanding its meaning to include domesticated sheep or goats, as Bradley (2016) states, but this conjecture is not necessary, given the new archaeological dates for Chinese sheep.

5.9 Horse

While the three items in 5.1 exhibit phonologically regular correspondences, suggestive of a prototype pre-reconstructible as *m-raŋ, other clearly related forms, such as Jingpo *kum*[31] *ʒa*[31] and OC 馬 *mˤraʔ lack the nasal ending; this irregularity is the sign of secondary spread of domesticated horses within the family, perhaps out of a ST language where the rhyme in [mraŋ] had changed to [ã]. Yet other ST forms, like Chepang *sĕraŋ*, Bunan *ʂaŋs* and Lai Hakha *ràŋ* point to *s-raŋ and *raŋ prototypes: this suggests we are in the presence of indigenous forms derived out of a verb root √raŋ by means of prefixes *m- and *s-. These elements conclusively indicate the absence in the ancestral ST language of a word for "horse", in full agreement with the late date of archaeological appearance of domesticated horses in East Asia (5.2).

5.10 Cattle

The cognate set for "cattle" in 5.1 presents an interesting riddle: it has all the appearances of phonological regularity, implying knowledge of cattle by PST speakers, as proposed by Bradley (2016); yet currently the first archaeological occurrence of domesticated cattle in East Asia is in far western Yangshao or Majiayao area at 5400-4700 BP, too late for PST by our dates. Supposing that non-Sinitic speakers first encountered domesticated cattle in that north-westerly region, and that their term for it was later transmitted to Sinitic through contact will not work either, because the Chinese loanword should then have the vowel *a, not *ə (while the correspondence between non-Sinitic *a and Old Chinese *ə is regular in cognate words). Zooarchaeology provides a solution: there is evidence that morphologically wild cattle was managed by humans in early Holocene northern China (Zhang et al. 2013). Presumably the cognate set in 5.1 is the PST term for early East Asian managed cattle; it was later applied by westward-expanding Sino-Tibetan groups to west Eurasian domesticated cattle that they encountered as they reached the western end of the loess plateau.

References

Allen, N. J. (1975). *Sketch of Thulung grammar*. Ithaca: Cornell University China-Japan Program.

Baxter, William H. and Laurent Sagart (2014a). *Old Chinese. A new reconstruction*. Oxford: Oxford University Press.

– (2014b). *The Baxter-Sagart reconstruction of Old Chinese (Version 1.1, 20 September 2014)*. Online document: `http://ocbaxtersagart.lsait.lsa.umich.edu/`. misc.

Běijīng Dàxué (1964). *Hànyǔ fāngyán cíhuì* 汉语方言词汇 [Chinese dialect vocabularies]. Běijīng: Wénzì Gǎigé.

Ben Hamed, Mahe and Feng Wang (2006). "Stuck in the forest: Trees, networks and Chinese dialects". *Diachronica* 23, 29–60.

Benedict, Paul K. (1972). *Sino-Tibetan: a conspectus*. Ed. by James A. Matisoff. Cambridge: Cambridge University Press.

– (1975). *Austro-Thai language and culture, with a glossary of roots*. With a forew. by Ward Goodenough. HRAF Press.

Bernot, Denise, Marie Yin Yin Myint, and Cristina Cramerotti (1998). *Dictionnaire français-birman*. L'Asiathèque.

Bhat, D. N. Shankara (1969). *Tankhur Naga vocabulary*. Poona: Deccan College Postgraduate and Research Institute.

Blench, Roger and Mark W. Post (2014). "Rethinking Sino-Tibetan phylogeny from the perspective of Northeast Indian languages". In: *Trans-Himalayan-Linguistics*. Ed. by Thomas Owen-Smith and Nathan W. Hill. Berlin: de Gruyter, 71–104.

Bouckaert, Remco R (2010). "DensiTree: making sense of sets of phylogenetic trees". *Bioinformatics* 26.10, 1372–1373.

Bowern, Claire and Quentin D. Atkinson (2012). "Computational phylogenetics of the internal structure of Pama-Nguyan". *Language* 88, 817–845.

Bradley, David (1979). *Proto-Loloish*. London: Curzon Press.

– (1997). "Tibeto-Burman languages and classification". In: *Papers in Southeast Asian Linguistics*. Ed. by David Bradley. Vol. 14. Canberra: Pacific Linguistics, 1–72.

– (2011). "Proto-Tibeto-Burman Grain Crops". *Rice* 4.3-4, 134–141.

– (2016). "Chinese calendar animals in 山海经 Shanhaijing and in Sino-Tibetan languages". In: *Shanhaijing world geography and ancient Chinese civilization*. Foreign Languages Teaching and Research Press.

– (2017). *Ancient history of Sino-Tibetan in China*. paper presented at the 9th International conference of Evolutionary Linguistics. Yunnan Minzu University, Kunming, China.

– (2018). *Subgrouping of the Sino-Tibetan Languages*. Paper presented at the 10th International Conference on Evolutionary Linguistics, Nanjing University, 27-28 October, 2018.

Burling, Robbins (2003). *The language of the Modhupur Mandi, Garo: Vol. III : Glossary*. Ann Arbor, Michigan: Univ. of Michigan, vii+232.

Cai, Dawei, Yang Sun, Zhuowei Tang, Songmei Hu, Wenying Li, Xingbo Zhao, Hai Xiang, and Hui Zhou (2014). "The origins of Chinese domestic cattle as revealed by ancient DNA analysis". *Journal of Archaeological Science* 41, 423–434.

Caughley, Ross. C (2000). *Dictionary of Chepang language: A Tibeto-Burman Language*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies. Australian National University Press.

Chang, Tsung-tung (1988). "Indo-European Vocabulary in Old Chinese". *Sino-Platonic papers* 7, 1–56.

Choi, Jae Young, Adrian E. Platts, Dorian Q. Fuller, Yue-Ie Hsing, Rod A. Wing, and Michael D. Purugganan (2017). "The rice paradox: Multiple origins but single domestication in Asian rice". *Molecular Biology and Evolution*.

Coblin, Weldon South (1986). *A Sinologist's Handlist of Sino-Tibetan Lexical Comparisons*. Steyler Verlag.

Cobo, José M., Joaquim Fort, and Neus Isern (2019). "The spread of domesticated rice in eastern and southeastern Asia was mainly demic". *Journal of Archaeological Science* 101, 123–130.

DeLancey, Scott (2015). "Morphological evidence for a central branch of Trans-Himalayan (Sino-Tibetan)". *Cahiers de Linguistique Asie Orientale* 44, 122 —149.

Deng, Zhenhua, Ling Qin, Yu Gao, Alison Ruth Weisskopf, Chi Zhang, and Dorian Q. Fuller (2015). "From Early Domesticated Rice of the Middle Yangtze Basin to Millet, Rice and Wheat Agriculture: Archaeobotanical Macro-Remains from Baligang, Nanyang Basin, Central China (6700–500 BC)". *PLOS ONE* 10.10. Ed. by Swarup Kumar Parida.

Dodson, John, Eoin Dodson, Richard Banati, Xiaoqiang Li, Pia Atahan, Songmei Hu, Ryan J. Middleton, Xinying Zhou, and Sun Nan (2014). "Oldest Directly Dated Remains of Sheep in China". *Scientific Reports* 4.1.

Driem, George van (1993). *A grammar of Dumi*. Grammar Library. Berlin and New York: Mouton de Gruyter.

– (1997). "Sino-Bodic". *Bulletin of the School of Oriental and African Studies* 60.3, 455–488.

– (2011). "Rice and the Austroasiatic and Hmong-Mien homelands". In: *Dynamics of Human Diversity: The Case of Mainland Southeast Asia (Pacific Linguistics, 627)*. The Australian National University, 361–389.

– (2014). "Trans-Himalayan". In: *Trans-Himalayan linguistics*. Ed. by Nathan W. Hill and Thomas Owen-Smith. Berlin: Mouton de Gruyter, 11–40.

Driem, George L. van (2017). "The domestications and the domesticators of Asian rice". In: *Language Dispersal Beyond Farming*. John Benjamins Publishing Company, 183–214.

Dunn, Michael (2012). *Indo-European lexical cognacy database (IELex)*. URL: `http://ielex.mpi.nl/`.

Dyen, Isidore (1963). *The Lexicostatistical classification of the Austronesian languages*. Yale University.

Flad, Rowan K., Yuan Jing 袁靖, and Li Shuicheng 李水城 (2007). "Zooarcheological evidence for animal domestication in northwest China". *Developments in Quaternary Science* 9, 167–203.

Fuller, Dorian Q., L. Qin, Y. Zheng, Z. Zhao, X. Chen, L. A. Hosoya, and G.-P. Sun (2009). "The Domestication Process and Domestication Rate in Rice: Spikelet Bases from the Lower Yangtze". *Science* 323.5921, 1607–1610.

Genetti, Carol (2007). *A grammar of Dolakha Newar*. Berlin and New York: Mouton de Gruyter.

Gong, Hwang-cherng (1995). "The System of Finals in Proto-Sino-Tibetan". In: *The Ancestry of Chinese*. Ed. by William Wang. Journal of Chinese Linguistics Monograph Series, 41–92.

Greenhill, Simon J., Robert Blust, and Russell D. Gray (2008). "The Austronesian Basic Vocabulary Database: From bioinformatics to lexomics". *Evolutionary Bioinformatics* 4, 271–283.

Guedes, Jade d'Alpoim (2011). "Millets, Rice, Social Complexity, and the Spread of Agriculture to the Chengdu Plain and Southwest China". *Rice* 4.3-4, 104–113.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath (2017). *Glottolog*. misc. Leipzig.

Haudricourt, André G. and David Strecker (1991). "Hmong-Mien (Miao-Yao) Loans in Chinese". *Toung Pao* 77, 335–341.

Hill, Nathan W. (2009). "Review of Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction. By James A. Matisoff". *Languages and Linguistics* 10.1, 173–195.

– (2012). "The six vowel hypothesis of Old Chinese in comparative context". *Bulletin of Chinese Linguistics* 6.2, 1–69.

– (2014). "Cognates of Old Chinese *-n, *-r, and *-j in Tibetan and Burmese". *Cah. Linguistique – Asie Orientale* 43.2, 91–109.

Hill, Nathan W. and Johann-Mattis List (2017). "Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages". *Yearbook of the Poznań Linguistic Meeting* 3.1, 47—76.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker (2008). "Explorations in automated lexicostatistics". *Folia Linguistica* 20.3, 116–121.

Huáng, Bùfán, ed. (1992). *Zàngmiǎn yǔzú yǔyán cíhuì*. Běijīng: Central Institute of Minorities.

Hübschmann, Heinrich (1897). *Armenische Grammatik. I. Theil. Armenische Etymologie*. Leipzig: Breitkopf & Härtel.
Jacques, Guillaume (2004). "Phonologie et morphologie du japhug (Rgyalrong)". PhD thesis. Université Paris VII - Denis Diderot.
– (2014). *Esquisse de phonologie et de morphologie historique du tangoute*. Leiden: Brill, 373.
– (2015). "On the cluster *sr- in Sino-Tibetan". *Journal of Chinese Linguistics* 43.1A, 215–223.
– (2015–2016). *Dictionnaire Japhug-Chinois-Français, version 1.1*. Paris: Projet HimalCo.
– (2016). "Tangut, Gyalrongic, Kiranti and the nature of person indexation in Sino-Tibetan/Trans-Himalayan". *Ling. Vanguard*.
– (2017a). "A reconstruction of Proto-Kiranti verb roots". *Folia Linguistica Historica* 38, 177–215.
– (2017b). *Fieldwork notes*. misc.
Jacques, Guillaume and Johann-Mattis List (forthcoming). "Save the trees: Why we need tree models in linguistic reconstruction (and when we should apply them)". *Journal of Historical Linguistics* 19.1, ??–??
Jacques, Guillaume and Alexis Michaud (2011). "Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze". *Diachronica* 28.4, 468–498.
Janhunen, Juha (1996). *Manchuria: An ethnic history*. Finno-Ugrian Society.
Jongens, M. (2009). *A grammar of Bantawa. Grammar, paradigm tables, glossary and texts of a Rai language of Eastern Nepal*. Utrecht: LOT.
Joseph, U. V. (2007). *Rabha*. Leiden and Boston: Brill.
Joseph, U. V. and Robbins Burling (2006). *The comparative phonology of the Boro-Garo Languages*. Includes bibliographical references (p. [149]-151). Mysore: Manasagangotri, Mysore, India: Central Institute of Indian Languages, 164.
Kang, Longli, Yan Lu, Chuanchao Wang, Kang Hu, Feng Chen, Kai Liu, Shilin Li, Li Jin, and Hui Li (2011). "Y-chromosome O3 Haplogroup Diversity in Sino-Tibetan Populations Reveals Two Migration Routes into the Eastern Himalayas". *Annals of Human Genetics* 76.1, 92–99.
Konnerth, Linda (forthcoming). *A grammar of Karbi*. Berlin: de Gruyter.
Kölver, U. and I. Shresthacarya (1994). *A dictionary of contemporary Newari*. Bonn: VGH Wissenschaftsverlag.
Lai, Yunfan (2017a). *Fieldwork notes*. misc.
– (2017b). "Grammaire du Khroskyabs de Wobzi". PhD thesis. Paris: Université Sorbonne Nouvelle - Paris III.
LaPolla, Randy (2001). "The Role of Migration and Language Contact in the Development of the Sino-Tibetan Language Family". In: *Areal Diffusion and Genetic Inheritance: Problems in Comparative Linguistics*. Ed. by Robert Malcolm Ward Dixon and Alexandra & A. Y. Aikhe nv Aikhenvald. OXFORD UNIV PR, 225–254.
Li, Fang-kuei (1937 [1973]). "Languages and dialects of China". *Journal Of Chinese Linguistics* 1.1, 1–13.
List, Johann-Mattis (2014). *Sequence comparison in historical linguistics*. Düsseldorf: Düsseldorf University Press.
– (2016). "Beyond cognacy: Historical relations between words and their implication for phylogenetic reconstruction". *Journal of Language Evolution* 1.2, 119–136.
– (2017). "A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Valencia: Association for Computational Linguistics, 9–12.
List, Johann-Mattis, Michael Cysouw, and Robert Forkel (2016a). "Concepticon. A resource for the linking of concept lists". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Ed. by Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. European Language Resources Association (ELRA), 2393–2400.
List, Johann-Mattis, Jananan Sylvestre Pathmanathan, Philippe Lopez, and Eric Bapteste (2016b). "Unity and disunity in evolutionary sciences: process-based analogies open common research avenues for biology and linguistics". *Biology Direct* 11.39, 1–17.
List, Johann-Mattis, Philippe Lopez, and Eric Bapteste (2016c). "Using sequence similarity networks to identify partial cognates in multilingual wordlists". In: *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Berlin, 599–605.
List, Johann-Mattis, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel (2018a). "CLICS^2. An improved database of cross-linguistic colexifications assembling lexical data with help of cross-linguistic data formats". *Linguistic Typology* 22.2, 277–306.
List, Johann-Mattis, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel (2018b). "Sequence comparison in computational historical linguistics". *Journal of Language Evolution* 3.2, 130—144.
Liu, Zhiji et al. (2005). *Han Ying duizhao jiaguwen jinyi leijian:* 广西教育出版社.

Liú, Lìlǐ, Hóngzhōng Wáng, and Yíng Bǎi (2007). *Xiàndài Hànyǔ fāngyán héxīncí, tèzhēng cíjí* 现代汉语方言核心词·特征词集 [Collection of basic vocabulary words and characteristic dialect words in modern Chinese dialects]. Nánjīng: Fènghuáng.
Long, Tengwen, Christian Leipe, Guiyun Jin, Mayke Wagner, Rongzhen Guo, Oskar Schröder, and Pavel E. Tarasov (2018). "The early history of wheat in China from 14C dating and Bayesian chronological modelling". *Nature Plants* 4.5, 272–279.
Lorrain, J. Herbert (1940). *Dictionary of the Lushai language (Bibliotheca Indica 261)*. Calcutta: Royal Asiatic Society of Bengal.
Lu, H. et al. (2009). "Earliest domestication of common millet (Panicum miliaceum) in East Asia extended to 10,000 years ago". *Proceedings of the National Academy of Sciences* 106.18, 7367–7372.
Luce, G. H. (1985). *Phases of Pre-Pagán Burma: Languages and history*. Oxford: Oxford University Press.
Lǐ, Fànwén (1997). *Xià-Hàn Zìdiǎn*. Běijīng: Zhōngguó Shèkēyuàn Chūbǎnshè.
Lǐ, Yǒngsuì (1998). "Qiāngmiǎn yǔqún chúyì". *Mízúyǔwén* 1, 16—28.
Matisoff, James A. (1978). *Variational semantics in Tibeto-Burman. The 'organic' approach to linguistic comparison*. Institute for the Study of Human Issues.
– (1991). "Sino-Tibetan Linguistics: Present State and Future Prospects". *Annual Review of Anthropology* 20.1, 469–504.
– (2003a). *Handbook of Proto-Tibeto-Burman*. Vol. 135. University of California Publications in Linguistics. Berkeley and Los Angeles: University of California press.
– ed. (2003b). *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University Presses of California, Columbia and Princeton.
– (2015). *The Sino-Tibetan Etymological Dictionary and Thesaurus project*. Berkeley: University of California.
Michailovsky, Boyd (1989a). *Bahing*. misc. Berkeley.
– (1989b). *Hayu*. misc. Berkeley.
Miller, Roy Andrew (1974). "Sino-Tibetan: Inspection of a Conspectus". *Journal of the American Oriental Society* 94.2, 195.
Mortensen, David R. (2012). *Database of Tangkhulic Languages*. misc. Berkeley.
Nishi, Yoshio (1999). *Four papers on Burmese: Toward the history of Burmese (the Myanmar language)*. Tokyo: Institute for the study of languages, cultures of Asia, and Africa, Tokyo University of Foreign Studies.
Norman, Jerry (1988). *Chinese*. Cambridge: Cambridge University Press.
Okell, J. (1971). "K Clusters in Proto-Burmese". In: *Papers presented at the Sino-Tibetan Conference*. (Bloomington, 10/08–10/09/1971). Bloomington.
Oota, Hiroki, Brigitte Pakendorf, Gunter Weiss, Arndt von Haeseler, Surin Pookajorn, Wannapa Settheetham-Ishida, Danai Tiwawech, Takafumi Ishida, and Mark Stoneking (2005). "Recent Origin and Cultural Reversion of a Hunter–Gatherer Group". *PLoS Biology* 3.3. Ed. by David Penny, e71.
Peiros, Ilia (1998). *Comparative linguistics in Southeast Asia (Pacific linguistics)*. Pacific Linguistics. Research School of Pacific and Asian Studies, Australian National University.
Peiros, Ilia and Sergei Starostin (1996). *A comparative vocabulary of five Sino-Tibetan languages*. Melbourne: University of Melbourne, Department of Linguistics.
Pierron, Denis et al. (2014). "Genome-wide evidence of Austronesian–Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar". *Proceedings of the National Academy of Sciences* 111.3, 936–941.
Pulleyblank, Edwin (1983). "The Chinese and their neighbors in prehistoric and early historic times". In: *The Origins of Chinese Civilization*. Ed. by David N. Keightley. University of California Press, 411–466.
Rama, Taraka, Johann-Mattis List, Johannes Wahle, and Gerhard Jäger (2018). "Are automatic methods for cognate detection good enough for phylogenetic reconstruction in historical linguistics?" In: *Proceedings of the North American Chapter of the Association of Computational Linguistics*. "NAACL 18" (New Orleans, 06/01–06/06/2018), 393–400.
Reid, Lawrence A. (1992). "The Tasaday language: a key to Tasaday prehistory". In: *The Tasaday Controversy: Assessing the Evidence*. Ed. by Thomas N. Headland. Amer Anthropological Assn, 180–193.
Sagart, Laurent (1995). "Chinese 'buy' and 'sell' and the direction of borrowings between Chinese and Miao-Yao". *T'oung Pao* LXXXI.4-5, 328–342.
– (2005). "Sino-Tibetan-Austronesian: an updated and improved argument. Putting together archaeology, linguistics and genetics". In: *The peopling of East Asia: Putting together Archaeology, Linguistics and Genetics*. Ed. by Laurent Sagart, Roger Blench, and Alicia Sanchez-Mazas. RoutledgeCurzon. Chap. 9, 161–176.
– (2006). "Review of Handbook of Proto-Tibeto-Burman: System and philosophy of Sino-Tibeto-Burman reconstruction. By James A. Matisoff". *Diachronica* 23.1, 206–223.

Sagart, Laurent (2011a). *Classifying Chinese dialects/Sinitic languages on shared innovations*. Paper, presented at the Séminaire Sino-Tibétain du CRLAO (2011-03-28). URL: https://www.academia.edu/19534510/Chinese_dialects_classified_on_shared_innovations.
– (2011b). "How Many Independent Rice Vocabularies in Asia?" *Rice* 4.3-4, 121–133.
– (2011c). "华澳语系发源于何时何地? The homeland of Sino-Tibetan-Austronesian: where and when ?" In: *Communication on Contemporary Anthropology*. Vol. 5, 143–147.
– (2017). "A candidate for a Tibeto-Burman innovation". *Cahiers de Linguistique Asie Orientale* 46.1, 101–119.
– (2018a). *The names of the rice plant. II 'Tibeto-Burman'*. https://stan.hypotheses.org/176. misc.
– (2018b). *The names of the rice plant. II 'Tibeto-Burman'*. URL: https://stan.hypotheses.org/176..
Sagart, Laurent, Tze-Fu Hsu, Yuan-Ching Tsai, and Yue-Ie C. Hsing (2017). "Austronesian and Chinese words for the millets". *Language Dynamics and Change* 7.2, 187–209.
Satterthwaite-Phillips, D. (2011). "Phylogenetic inference of the Tibeto-Burman languages or on the usefuseful of lexicostatistics (and "megalo"-comparison) for the subgrouping of Tibeto-Burman". thesis. Stanford: Stanford University.
Schuessler, Axel (2007). *ABC Etymological Dictionary of Old Chinese*. Honolulu: University of Hawaii Press.
Shafer, Robert (1955). "Classification of the Sino-Tibetan Languages". *Word* 11.1, 94–111.
Sharma, Suhnu Ram (2003a). "A Sketch of Byangsi Grammar". In: *Tibeto-Burman Languages of Uttar Pradesh*. Ed. by Randy J. LaPolla, 79–140.
– (2003b). "A Sketch of Rongpo Grammar". In: *Tibeto-Burman Languages of Uttar Pradesh*. Ed. by Randy J. LaPolla, 1–77.
Sidwell, Paul (2015). "Austroasiatic dataset for phylogenetic analysis: 2015 version". *Mon-Khmer Studies (Notes, Reviews, Data-Papers)* 44, lxviii–ccclvii.
Starosta, Stanley (2005). "Proto-East Asian and the original dispersal of the languages of east and southeast Asia and the Pacific". In: *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*. Routledge.
Starostin, George S. (2013). *Metodologija. Kojsanskie jazyki*. Vol. 1. Moscow: Jazyki Russkoj Kul'tury.
Starostin, Sergei (1984[1988]). "Gipoteza o genetičeskix svjazjax sinotibetzkix jazykov s enisejskimi i severnokavkazskimi jazykami , pp. 19-38. Moscou: Nauka." In: *Lingvističeskaja rekonstrukcija i drevnejšaja istorija vostoka: tezisy I doklady konferencii, čast' 4: Drevnejšaja jazykovaja situacija v vostočnoj Azii*, 19–38.
Starostin, Sergei A. (2004). "Altaic and Chinese". In: *Works in Linguistics*. Moscow: Yazyki slavyanskikh kul'tur, 850–853.
Starostin, Sergej Anatol'evic (1991). *Altajskaja problema i proischoždenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]*. Moscow: Nauka.
Stevens, Chris J. and Dorian Q Fuller (2017). "The spread of agriculture in eastern Asia". *Language Dynamics and Change* 7.2, 152–186.
Sun, Jackson T. S. (1993). "A historical-comparative study of the Tani (Mirish) branch in Tibeto-Burman". PhD thesis. University of California at Berkeley.
Swadesh, Morris (1952). "Lexico-statistic dating of prehistoric ethnic contacts. With special reference to North American Indians and Eskimos". *Proceedings of the American Philosophical Society* 96.4, 452–463.
– (1955). "Towards greater accuracy in lexicostatistic dating". *International Journal of American Linguistics* 21.2, 121–137.
Tadmor, Uri (2009). "Loanwords in the world' s languages. Findings and results". In: *Loanwords in the world's languages. A comparative handbook*. Ed. by Martin Haspelmath and Uri Tadmor. Berlin and New York: de Gruyter, 55–75.
Takeuchi, Tsuguhito, ed. (2013). *Old Tibetan Documents Online*. URL: https://otdo.aa-ken.jp/.
Thurgood, Graham (2003). "A subgrouping of the Sino-Tibetan languages: The interaction between language contact, change, and inheritance". In: *The Sino-Tibetan languages*. Ed. by Graham Thurgood and Randy J. LaPolla. London and New York: Routledge, 3–21.
– (2008). "Han-Zang Yuyan de Puxi [The Sino-Tibetan languages: Genetic inheritance, external contact and change] (in Chinese). 15." *Minzu Yuwen* (2), 3–15.
– (2017). "Sino-Tibetan: Genetic and areal subgroups". In: *The Sino-Tibetan Languages, 2nd Edition*. Ed. by Graham Thurgood and Randy LaPolla. London: London: Routledge, 3–39.
Tolsma, Gerard (1999). "A Grammar of Kulung". PhD thesis. University of Leiden.
VanBik, David (2014). *English to Chin (Hakha) dictionary*.
VanBik, Kenneth (2009). *Proto Kuki-Chin*. Berkeley: STEDT Monograph.
Walker, G. D. (1925). *A Dictionary of the Mikir Language*. Mittal Publications.

Watanabe, Kazuo, Ye Tint Tun, and Makoto Kawase (2007). “Field Survey and Collection of Traditionally Grown Crops in Northern Areas of Myanmar, 2006”. 植探報 23, 161–175.
Widmer, Manuel (2017). *A grammar of Bunan*. Berlin and New York: De Gruyter Mouton.
Wiens, J. J. (2006). “Missing data and the design of phylogenetic analyses”. *Journal of Biomedical Informatics* 39.1, 34–42.
Wiens, John J. and Matthew C. Morrill (2011). “Missing Data in Phylogenetic Analysis: Reconciling Results from Simulations and Empirical Data”. *Systematic Biology* 60.5, 719–731.
Xiang, Hai et al. (2017). “Origin and dispersal of early domestic pigs in northern China”. *Scientific Reports* 7.1.
yanjiusuo, Zhongguo shehui kexueyuan lishi (1978). *Jiaguwen heji*. Zhonghua shuju.
Zhang, Hucai et al. (2013). “Morphological and genetic evidence for early Holocene cattle management in northeastern China”. *Nature Communications* 4.1.