Current Biology, Volume *25*

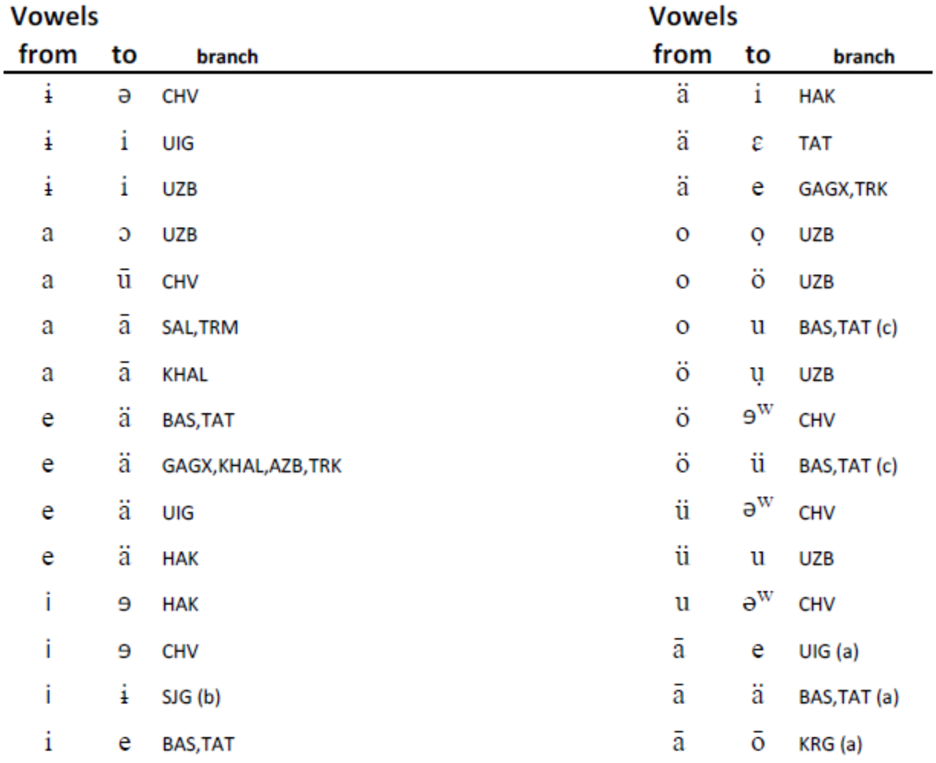
Supplemental Information

Detecting Regular Sound Changes in Linguistics as Events of Concerted Evolution

Daniel J. Hruschka, Simon Branford, Eric D. Smith, Jon Wilkins, Andrew Meade, Mark Pagel, and Tanmoy Bhattacharya

Supplemental Information

Table S1. List of estimated regular changes with support in > 90 trees.



Vowels			Vowels		
from	to	branch	from	to	branch
ɨ	ə	CHV	ä	i	HAK
ɨ	i	UIG	ä	ɛ	TAT
ɨ	i	UZB	ä	e	GAGX,TRK
a	ɔ	UZB	o	ọ	UZB
a	ū	CHV	o	ö	UZB
a	ā	SAL,TRM	o	u	BAS,TAT (c)
a	ā	KHAL	ö	ụ	UZB
e	ä	BAS,TAT	ö	ɘ^{w}	CHV
e	ä	GAGX,KHAL,AZB,TRK	ö	ü	BAS,TAT (c)
e	ä	UIG	ü	ə^{w}	CHV
e	ä	HAK	ü	u	UZB
i	ɘ	HAK	u	ə^{w}	CHV
i	ɘ	CHV	ā	e	UIG (a)
i	ɨ	SJG (b)	ā	ä	BAS,TAT (a)
i	e	BAS,TAT	ā	ō	KRG (a)

a. rare starting sound

b. language without sound correspondence proposal in Starostin et al.

c. phoneme swap (e.g. o to u and u to o).

Consonants

from	to	branch
d	d	TOF,TUV,DOLG,JAK (a)
y	v	CHV
y	w	QUM,KLPX,BLKX,BAS,NOGX,KAZ,KRG,TAT
č	d́	ALT
č	ś	CHV
č	š	KLPX,NOGX,KAZ
č	h	DOLG,JAK
č	š	TOF,TUV,DOLG,JAK,ALT,SHR,HAK
č	s	BAS
b	p	SHR,HAK
b	p	CHV
d	č	CHV
š	ź	CHV
š	s	KLPX,NOGX,KAZ
š	s	DOLG,JAK
š	s	SJG (b)
š	s	HAK
h	s	JAK
ž	č	TOF,TUV,DOLG,JAK,ALT,SHR,HAK
ž	h	DOLG
ž	h	JAK
ž	ś	CHV

Consonants

from	to	branch
ž	j	BAS,TAT
ž	j	SAL,KRMX,GAGX,KHAL,TRM,AZB,UIG,UZB,TRK
ž	j	SJG (b)
ž	j	NOGX
ž	j	QUM
ž	z	HAK
ž	ǯ	KRG
q	χ	SAL (b)
q	G	SAL,TRM
q	G	AZB
q	k	TRK
q	x	DOLG,JAK
q	x	HAK
q	x	CHV
s	h	BAS
ŋ	m	CHV
ŋ	n	GAGX,AZB,TRK
x	k	DOLG (b)
z	ž	DOLG,JAK
z	r	CHV
z	δ	BAS

Table S2 Sound correspondences compared to historical linguistic proposals

Each pair of rows of the table records (upper row) linguists' proposals as to regular sound changes in twenty-three Turkic languages and (lower row) corresponding predictions from the model of regular sound changes. Items coloured red correspond to regular sound changes, uncoloured items refer to ancestral phonemes being retained. In each case the upper row records linguists' proposals and the lower row reports the phonemes to which the model assigns its highest probabilities. The left columns identify the proto-phonemes favoured by linguistic analysis and the corresponding phoneme favoured by the model of regular change. The rightmost column records the count of the number of proposed changes under the two approaches. A blue border around a cell denotes a change that can be attributed to one or more of the regular sound changes inferred on the phylogeny of Figure 2 (main text)

	Model Ancestor	Linguists' proposal	Yakut	Tuvan	Tofalar	Khakass	Shor	Altai	Kirghiz	Uighur	Uzbek	Kazak	Karakalpak	Nogay	Bashkir	Tatar	Kumyk	Balkar	Karaim	Turkmen	Azerbaijanian	Gagauz	Turkish	Khalaj	Chuvash	Phonological innovations
1		t	t	t d	t d	t d	t d	t d	t	t	t	t	t	t	t	t	t d	t	t	t d	t d	t d	t d	t	t d	11
2	t		t	t d	t d	t d	t d	t d	t	t	t	t	t	t	t	t	t d	t	t	t d	t d	t d	t d	t	t č	11
1		r	r	r	r	r	r	r	r	r j	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	1
2	r		r	r	r	r	r	r	r	r j	r	r	r	r	r	r	r	r	r	r	r	r	r	r	r	1
1		k	k x g ɣ	k x g	k x h q g	k x g ɣ	k q g ɣ	k g	k g	k q	k q	k q g ɣ	k q g ɣ	k g	k q g ɣ	k g	k g q ɣ	k g q ɣ	k g	k G q g ɣ	k G x g ɣ	k g	k ɣ	k q	k x g	50
2	k		k g x q	k g x	k g x q j	k g x j	k g x q	k g	k g	k g	k g	k g x q	k g x q	k g	k g x q	k g	k g x q	k g x q	k g	k g x j q	k g x G j	k g	k g	k g	k g x	50
1		l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	0
2	l		l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	l	0
1		k	k x g ɣ	k x g	k x h q g	k x g ɣ	k q g ɣ	k g	k g	k q	k q	k q g ɣ	k q g ɣ	k g	k q g ɣ	k g	k g q ɣ	k g q ɣ	k g	k G q g ɣ	k G x g ɣ	k g	k ɣ	k q	k x g	62
2	q		x k q ɣ	q x G	q x G ɣ k	x k G q	q x G ɣ	q x	q x	q x	q x	q x G ɣ	q x G ɣ	q x	q x G ɣ	q x	q x G ɣ	q x G ɣ	q x	G q ɣ x g	G x q ɣ k	q x	q x	q x	x k q	50
1		ń	j j̃	j j̃	j̃	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j v	4
2	j		j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j	j m	1
1		n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n m	1
2	n		n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n m	1
1		s	h s t	s z	s z	s z	s z	s z	s	s	s	s	s	s	h ϑ	s	s	s	s	ϑ s	s	s	s	s	s	10
2	s		s h z	s z	s z	s z	s z	s z	s	s	s	s	s	s	h ϑ	s	s	s	s	s z	s	s	s	s	s	10
1		b	b	b p v g	b p g	p b g	p g b	b	b j	b p v g j	b v g j	b w j	b w j	b w j	b w j	b w j	b p w j	b w j	b w j	b p w j	b p v j	b p v	b p v j	b v	p v	48
2	b		b	b p m v	b p m	p b m	p b m	b	b p	b p m v j	b p m v	b p m	b p m	b p m	b p m	b p m	b p m v	b p m	b p m	b p m v	b p m v	b p m	b p m v	b p	p b	47
1		m	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	m n	23
2	m		m b	m b	m b	m p	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	m b	23
1		ĺ	h s t	ž š	ǯ š	z s	ž š	ž š	š	š	š	s	s	s	š	š	š	š	š	š	š	š	š	š	l š	13
2	š		s č z	š č	š č	s č	š č	š s	š	š	š	s	s	s	š	š	š	š	š	š	š	š	š	š	l ź	14
1		č	s h	š ž	š ǯ	s ǯ	š ǯ	č	č	č	č	š	š	š	s	č	č	č	č	č ǯ	č ǯ	č ǯ	č ǯ	č	ś ź	20
2	č		s č	š č	š č	š č	š	š	č	č	č	š	š	š	s	č	č	č	č	č š	č š	č š	č š	č	ś č	15
1		ŕ	h s t	z s	z s	z s	z s	z s	z	z	z	z	z	z	z	z	z	z	z	z	z	z	z	z	r s	10
2	z		h s ǯ	z s	z s	z s	z s	z s	z	z	z	z	z	z	ϑ	z	z	z	z	z	z	z	z	z	r s	11
1		d	t	t d	t d	t z s	t z s	t d j	t j	t j	t j	t j	t j	t j	t j	t j	t d j	t j	t j	d j	d t j	d t j	d t j	t d	t r	44
2	d		t	d t	d t	d t j	d t j	d t j	d t	d t	d t	d t	d t	d t	d t	d t	d t j	d t	d t	d t	d t j	d t j	d t j	d t	č t	31
1		p	b p	v p	p b	b p	b p	b p	p b	p	p v b	b p	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	b p	b p	23
2	p		p b	p b	p b	p b	p b	p b	p b	p	p b v	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	p b	23
1		g	k	k x g ɣ	k x g ɣ	k g ɣ	k g ɣ	k j g	k j g	k g ɣ q	k g ɣ q	k w j	k w j	k w j	k w j	k w j	k g w j	k g w j	k g w j	g ɣ w j	g ɣ w j	g ɣ w j	g ɣ w j	k ɣ w j	k v	66
2	ɣ		ɣ	ɣ q G g	ɣ q G g	ɣ x G	ɣ q G	ɣ q G	w v ɣ	ɣ q G g	ɣ q G g	w v ɣ	w v ɣ	w v q	w v ɣ	w v ɣ	w v q ɣ	w v ɣ q	ɣ q G g	ɣ G q g	ɣ G g w	ɣ q G g	ɣ k G g	ɣ q G g	v x	57
1		g	k	k x g ɣ	k x g ɣ	k g ɣ	k g ɣ	k j g	k j g	k g ɣ q	k g ɣ q	k w j	k w j	k w j	k w j	k w j	k g w j	k g w j	k g w j	g ɣ w j	g ɣ w j	g ɣ w j	g ɣ w j	k ɣ w j	k v	62
2	g		g	g k ɣ j	g k ɣ j	g k ɣ	g k ɣ	g k ɣ	g k ɣ	g k ɣ j	g k ɣ j	g k ɣ	g k ɣ	g k ɣ	g k ɣ	g k ɣ	g k ɣ w	g k ɣ w	g k ɣ j	q k ɣ G	q k ɣ G	g k ɣ j	g k ɣ j	g k ɣ j	g k	57
1		ŋ	ŋ	ŋ	ŋ	g ŋ	g ŋ	ŋ	ŋ	ŋ n	ŋ j n	ŋ j n	ŋ j n	ŋ j n	ŋ j g n	ŋ j g n	ŋ j n	ŋ j g n	ŋ j g n	ŋ	n j	n m j	n m	ŋ j g n	n m	37
2	ŋ		ŋ	ŋ	ŋ	ŋ n	ŋ n	ŋ	ŋ	ŋ n	ŋ n j	ŋ n j	ŋ n j	ŋ n j	ŋ n j m	ŋ n j m	ŋ n j	ŋ n j m	ŋ n j m	ŋ	n j	n j m	n j	ŋ n j m	m n	37
1		j	s j t	č j d	č j d	č j z	č j z	d' j	ǯ j	j	j	ž j	ž j	j	j	j ǯ	j ǯ	ǯ j	j	j	j	j	j	j d	ś j v	41
2	ž		s j ǯ	č š ǯ	č š ǯ	č s ǯ	č š ǯ	d' j	ǯ š	j	j	ž ǯ	ž ǯ	j	j	j ǯ	j ǯ	ž ǯ	j	j	j	j	j	j ǯ	s č r	40

Total linguists' phonological innovations	526
Total model's phonological innovations	479

	Model Ancestor	Linguists' proposal	Yakut	Tuvan	Tofalar	Khakass	Shor	Altai	Kirghiz	Uighur		Uzbek		Kazak	Karakalpak	Nogay	Bashkir	Tatar	Kumyk	Balkar	Karaim	Turkmen	Azerbaijanian		Gagauz	Turkish	Khalaj		Chuvash			Phonological innovations
1		a	a	a	a	a	a	a	a	a	e	a	ɔ	a	a	a	a	a	a	a	a	a	a		a	a	a		o			3
2	a		a	a	a	a	a	a	a	a	e	ɔ	a	a	a	a	a	a	a	a	a	a	a		a	a	ā		o			4
1		u	u	u	u	u	u	u	u	u		u		u	u	u	o	o	u	u	u	u	u		u	u	u		əʷ			3
2	u		u	u	u	u	u	u	u	u		u		u	u	u	o	o	u	u	u	u	u		u	u	u		əʷ			3
1		ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	i		i		ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ		ɨ	ɨ	i		ə	ɘ		5
2	ɨ		ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	i		i		ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ		ɨ	ɨ	ɨ		ə	ɘ		4
1		i	i	i	i	ə	i	i	i	i		i		i	i	i	e	e	i	i	i	i	i		i	i	i		ɘ	ə		5
2	i		i	i	i	ə	i	i	i	i		i		i	i	i	e	e	i	i	i	i	i		i	i	i		ɘ	a		5
1		e	e	e	e	i	e	e	e	e	ä	e	a	e	e	e	i	i	e	e	e	e	ä	e	e	e	ä		a			8
2	e		e	e	e	i	e	e	e	ä	i	e	i	e	e	e	ä	ɛ	e	e	e	e	ä	i	e	e	ä		e			9
1		o	o	o	o	o	o	o	o	o		ŭ		o	o	o	u	u	o	o	o	o	o		o	o	o		vɨ	u		5
2	o		o	o	o	o	o	o	o	o		ụ		o	o	o	u	u	o	o	o	o	o		o	o	o		o	u		4
1		ü	ü	ü	ü	ü	ü	ü	ü	ü		u		ü	ü	ü	ö	ö	ü	ü	ü	ü	ü		ü	ü	ü	i	ɘʷ			5
2	ü		ü	ü	ü	ü	ü	ü	ü	ü		u		ü	ü	ü	ö	ö	ü	ü	ü	ü	ü		ü	ü	ü	i	əʷ			5
1		ö	ö	ö	ö	ö	ö	ö	ö	ö		ŭ		ö	ö	ö	ü	ü	ö	ö	ö	ö	ö		ö	ö	ö	e	vi	u	ü	7
2	ö		ö	ö	ö	ö	ö	ö	ö	ö		ụ		ö	ö	ö	ü	ü	ö	ö	ö	ö	ö		ö	ö	ö	ü	ɘʷ	e	ü	7
1		ā	ā	a	a	a	a	a	a	a	e	a	ɔ	a	a	a	a	a	a	a	a	ā	a		a	a	āa		o			23
2	ā		a	a	a	a	a	a	a	a	e	ɔ	ā	a	a	a	a	a	a	a	a	a	a		a	a	ā		a			23
1		ū	ū	u	u	u	u	u	u	u		u		u	u	u	o	o	u	u	u	ū	u		u	u	ūu		əʷ			21
2	ū		u	u	u	u	u	u	u	u		u		u	u	ū	u	u	u	u	u	u	u		u	u	u		o			22
1		ī	ī	i	i	ɘ	i	i	i	i		i		i	i	i	e	e	i	i	i	i	i		i	i	īi		ɘ	ə		23
2	ī		i	i	i	ɘ	i	i	i	i		i		i	i	ī	e	ī	i	i	i	i	i		i	i	i		i	ɘ		22

Total linguists' phonological innovations	**108**
Total model's phonological innovations	**108**

Table S3. Translation of dataset transcription to International Phonetic Alphabet

		v	v	ī	īi:	ū	u:
Dataset	IPA	r	r	c	c	w	w
q	q	n	n	ɛ	ɛ	ō	o:
a	a	e	e	ē	e:	h	h
b	b	m	m	g	g	ẹ	ɪ
ɨ	ɨ	d	d	ä	æ	ụ	ŭ
k	k	ü	y	ǟ	æ:	ọ	ʊ
u	u	δ	ð	ɣ	ɣ	əʷ	əʷ
p	p	j	j	ś	ɕ	d́	ɟ
ɔ	ɔ	z	z	č	ʧ	f	f
i	i	ö	ø	ǯ	ʤ	ɨ:	ɨ:
o	o	ə	ə	ž	ʒ	ń	ɲ
G	ɢ	ǖ	y:	ŋ	ŋ	ȫ	ø:
ā	a:	š	ʃ	χ	χ	ź	ʑ
x	x	s	s	l	l	ü̆	y̆
ə	ə	t	t	ɘʷ	ɘʷ	ŭ	ŭ

Supplemental Data

Data: Etymologically aligned and phonetically coded data for the Turkic languages is available from the Evolution of Human Languages project at: http://starling.rinet.ru/downl.php?lan=en. The Turkic languages are contained inside the altaic.exe file (filename = turcet). This study used the first 225 words.

Table S1, related to Figure 2. List of 73 regular sound changes that were detected in at least 90% of the trees in the posterior sample**.** Note: The model detected an average of 74.27 regular sound changes across the set of trees in the posterior sample. The extra 1.27 events per tree (74.27-73) are regular sound changes that occur infrequently in the posterior sample, and so fail to make the list in Table S1.

See Table S1

Table S2, related to Figure 3. Comparison of inferred regular sound changes to historical linguistic inferences.

See Table S2

Table S3, related to Figure 2 and Figure 3. Translation of dataset transcription to International Phonetic Alphabet

See Table S3

Supplemental Experimental Procedures

Concerted evolution. Concerted evolution is a term normally used to describe the process by which copies of duplicated genes, alleles or repetitive elements within a species come to be similar at many different sites of their sequences. But concerted change might be a general feature of evolving systems, including language and culture. In language evolution, concerted or regular change affects copies of the same sound (or phoneme) that appear in different words. Linguists have suggested that this regular change occurs because to fulfill their function of word discrimination, instances of the same phoneme must be both independent occurrences and also repeated realizations of the *same* functional element within the language's sound system. The resulting changes, of many instances of one phoneme in parallel to some other phoneme, yield regular sound correspondences between pairs or groups of languages.

The complexity in an instance of concerted evolution can range from a simple covariance of a few elements, such as amino acids that jointly determine a protein fold, to patterns of sound change in language that might affect groups of phonemes, to rules for conditional change involving many segments across a syllable or a word. We model regular sound change in a language family at the level of independent sounds (the most common kind), to demonstrate the large effect that the change model has on the interpretation of patterns in data. Some sound changes in our data set cannot be incorporated in this simple form of regular shifts, either because different instances are truly independent or because they depend on multi-phoneme conditions for which our model does not test. For instance, some kinds of change depend upon ‘context’ defined as other features that might or might not be present in a word (see Box and Discussion in main text). If the contextual conditions are rare, our model will likely treat these as sporadic changes. If they are common, however, our model will likely treat them as a regular change but not label the context of that change. The statistical modelling of complex concerted evolution brings many challenges of representation as well as simulation, and is left to future work.

The model assumes that a language's lexicon evolves primarily from changes to form-meaning links, from sporadic sound changes that occur independently within words, or from concerted, regular sound changes that occur across all words in a lexicon simultaneously. The model also assumes that when a language diverges into two languages, the two languages begin to change independently. These elements of the model— changes to form-meaning links, sporadic and regular sound changes, and the pattern of historical divergence—are treated as independent free parameters to be estimated. To assess how adding regular, concerted sound changes improve model fit and prediction, we fit the model in two ways: (1) permitting regular, concerted sound changes in addition to sporadic changes (henceforth, regular sound change model) and (2) permitting only sporadic changes (henceforth, sporadic sound change model). In addition, languages may lose entire words used to express a particular sense—with words of different origins used to express the original meaning—and these changes in cognacy are modeled as an independent stochastic process.

Likelihood modelling We model the sporadic sound changes as a continuous-time Markov process, widely used in models of DNA or protein sequence evolution, where in place of the usual 4×4 or 20×20 matrices of nucleotide or amino acid transitions we erect a 62×62 sound transition rate matrix, denoted $\boldsymbol{Q_s}$. As in DNA sequence and protein models, $\boldsymbol{Q_s}$ is an instantaneous rate matrix whose rows sum to zero and whose main diagonal elements for row i are equal to $-\sum_{j \neq i} q_{ij}$.

We estimate the elements of $\boldsymbol{Q_s}$ from the data employing a reversible-jump Markov chain Monte Carlo (RJ-MCMC) procedure described elsewhere (*S1*) that allows the large number of potential parameters to be reduced to a potentially far smaller set of statistically distinct parameters, and without loss of statistical accuracy or prior knowledge on the part of investigators.

Regular sound changes of the general form denoting the i^{th} sound changing to the j^{th} ($i \neq j$) are modeled in a stochastic matrix $\boldsymbol{Q_r}$ that takes the form of an identity matrix with an i^{th} diagonal element interchanged with the off diagonal position (ji). Pre-multiplication of any stochastic matrix $\boldsymbol{Q}$ (e.g., $P(D \mid Q_s, t)$) by $\boldsymbol{Q_r}$ is equivalent to adding all elements $q_{i1}, q_{i2}, \ldots q_{ik}$ to the corresponding values of $q_{j1}, q_{j2} \ldots q_{jk}$, and then zeroing out the $q_{i1}, q_{i2}, \ldots q_{ik}$.

We then use a different RJ-MCMC procedure to propose possible $\boldsymbol{Q_r}$ matrices in branches of the phylogenetic tree, thereby allowing regular changes to occur or not on a branch-specific basis. Candidate $\boldsymbol{Q_r}$ matrices are chosen by randomly selecting a pair of sounds (*i,j*) as above. These matrices are then proposed to land with uniform probability anywhere in the tree, allowing the model to estimate both the number of regular changes along branches and the position or timing of regular sound changes along a given branch. We do not place a prior probability distribution on the number of regular sound changes in the tree, but to be accepted in the Markov chain, new $\boldsymbol{Q_r}$ matrices must satisfy the requirements of the reversible-jump procedure.

The likelihood of the data for *n* languages given the model (*M*) of phoneme evolution is calculated by first assigning possible reconstructions (or phoneme assignments) at each of the *n-1* ancestral or internal nodes of the tree (*T*). Where *k* regular sound changes have occurred along a branch, the contribution to the likelihood of this branch is computed by multiplying together the site likelihoods that are appropriate elements of the total transition matrix

$$\exp[Q_s(t_1 - t_0)]Q_{r_1}\exp[Q_s(t_2 - t_1)]Q_{r_2}\ldots\exp[Q_s(t_k - t_{k-1})]Q_{r_k}\exp[Q_s(t - t_k)]$$

where *t* is the total time for the branch, and t_i represents the times of the *k* regular changes. In branches where no regular sound changes occur, the likelihood is found from $P(D \mid Q_s, t)$. Then the probabilities of these possible ancestral assignments are calculated by multiplying the likelihoods of the evolutions represented by each of the *2n-2* branches. Finally, the total likelihood is obtained by summing over all possible reconstructions at each of the *n-1* internal or ancestral nodes in the tree. In other words,

$$L(D \mid M,T) = \sum_{\{D_{\text{int}}\}} \int \prod_{t=1}^{2n-2} \left(P(D_t \mid D_{a(t)}, Q_s, Q_{r_1}, Q_{r_2}, \ldots, Q_{r_m}) dQ_{r_1} dQ_{r_2} \ldots dQ_{r_m} \right) dQ_s$$

, where D_t and $D_{a(t)}$ represent the possibly reconstructed phoneme sequence at the tip and ancestor of the t^{th} branch, the sum runs over all possible assignments of reconstructions, the integral over all possible assignments of regular (Q_r)and sporadic (Q_s)matrices, and the product runs over the branches in the tree. Likelihoods are combined at the root of the tree weighted by their empirical phoneme frequencies. In the above notation, all priors are implicitly set to the uniform non-informative distribution. This is the standard approach to calculating the likelihood of tree topologies from discrete data such as genes, proteins or sounds, but with the addition of integrating over the collection of regular-change matrices in addition to the usual integration over the sporadic matrix.

Parameters of the models are then estimated in the usual way [e.g., (*S2*)] by Markov chain Monte Carlo procedures that numerically explore the posterior distribution of model parameters, given the data and the assumed prior distributions of the parameters. In practice, to improve the rate of convergence of the Markov chains, we augmented the likelihood of the sound-change model with that obtained from cognacy data for the same set of 225 words, following methods described elsewhere (*S3, S4*).

This model does not assume *a priori* any pattern in sound correspondence, categories of sounds (*e.g.,* vowels and consonants), or similarity between sounds. Though it would have been straightforward to incorporate known facts, this linguistic intuition is currently not stated in a probabilistic way. This allows us to treat the raw data as the only evidence to support or deny the standard linguistic knowledge about common sound changes.

Simplifications. While the current model incorporates regular and sporadic sound changes, it also makes several simplifying assumptions. Sound changes in different positions in the same word are treated as independent, thus ignoring conditioning contexts and vowel harmony. Second, all sporadic processes are assumed to work sufficiently independently and uniformly that they can be expressed by a single probability function across the entire family. It allows us to explore the feasibility of inferring sound correspondences and sound change with no additional linguistic prior knowledge. Third, the alignment permits metathesis, but the full model explicitly ignores this and other rare processes (*S5*), replacing them with independent idiosyncratic changes in nearby sounds. The effects of non-independent evolution on various branches, *e.g.,* due to processes like contact phenomena, are ignored. Adding such within-word correlations might help resolve the discrepancies between model predictions and linguistic descriptions of sound correspondences for uvulars and velars such as *k*, *q* and *x* (Tables S1,S2).

Supplemental References

S1. M. Pagel, A. Meade, Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* 167, 808 (2006).

S2. W. R. Gilks, S. Richardson, D. J. Spiegelhalter, in *Markov chain Monte Carlo in practice,* W. R. Gilks, S. Richardson, D. J. Spiegelhalter, Eds. (Chapman and Hill, London, 1996), pp. 1-19.

S3. M. Pagel, New approaches to lexicostatistics and glottochronology. *Time Depth in Historical Linguistics. Cambridge: McDonald Institute for Archaeological Research*, 209 (2000).

S4. M. Pagel, A. Meade, in *Phylogenetic methods and the prehistory of languages,* P. Forster, C. Renfrew, Eds. (McDonald institute Monographs, 2006), pp. 173-182.

S5. J. Blevins, A. Garrett, The origins of consonant-vowel metathesis. *Language*, 508 (1998).