

## A Markov model of the Indus script

**Rajesh P. N. Rao<sup>a,1</sup>, Nisha Yadav<sup>b,c</sup>, Mayank N. Vahia<sup>b,c</sup>, Hrishikesh Joglekar<sup>d</sup>, R. Adhikari<sup>e</sup>, and Iravatham Mahadevan<sup>f</sup>**

<sup>a</sup>Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195; <sup>b</sup>Department of Astronomy and Astrophysics, Tata Institute of Fundamental Research, Mumbai 400005, India; <sup>c</sup>Centre for Excellence in Basic Sciences, Mumbai 400098, India; <sup>d</sup>14, Dhushwadi, Laxminiketan, Thakurdwar, Mumbai 400002, India; <sup>e</sup>Institute of Mathematical Sciences, Chennai 600113, India; and <sup>f</sup>Indus Research Centre, Roja Muthiah Research Library, Chennai 600113, India

Communicated by Roddam Narasimha, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore, India, June 15, 2009 (received for review December 26, 2008)

Although no historical information exists about the Indus civilization (flourished ca. 2600–1900 B.C.), archaeologists have uncovered about 3,800 short samples of a script that was used throughout the civilization. The script remains undeciphered, despite a large number of attempts and claimed decipherments over the past 80 years. Here, we propose the use of probabilistic models to analyze the structure of the Indus script. The goal is to reveal, through probabilistic analysis, syntactic patterns that could point the way to eventual decipherment. We illustrate the approach using a simple Markov chain model to capture sequential dependencies between signs in the Indus script. The trained model allows new sample texts to be generated, revealing recurring patterns of signs that could potentially form functional subunits of a possible underlying language. The model also provides a quantitative way of testing whether a particular string belongs to the putative language as captured by the Markov model. Application of this test to Indus seals found in Mesopotamia and other sites in West Asia reveals that the script may have been used to express different content in these regions. Finally, we show how missing, ambiguous, or unreadable signs on damaged objects can be filled in with most likely predictions from the model. Taken together, our results indicate that the Indus script exhibits rich syntactic structure and the ability to represent diverse content, both of which are suggestive of a linguistic writing system rather than a nonlinguistic symbol system.

ancient scripts | archaeology | linguistics | machine learning |  
statistical analysis

The Indus (or Harappan) civilization flourished from *ca.* 2600 to 1900 B.C. in a vast region spanning what is now Pakistan and northwestern India. Its trade networks stretched to the Persian Gulf and the Middle East. The civilization emerged from the depths of antiquity in the late 19th century, when General Alexander Cunningham (1814–1893) visited a site known as Harappa and published a description (1), including an illustration of a tiny seal with characters in an unknown script. Since then, much has been learned about the Indus civilization through the work of archaeologists (see refs. 2 and 3 for reviews), but the script still remains an enigma.

More than 3,800 inscriptions in the Indus script have been unearthed on stamp seals, sealings, amulets, small tablets, and ceramics (see Fig. 1A for examples). Although there have been >60 claimed decipherments (4), none of these has been widely accepted by the community. Several obstacles to decipherment have been identified (4), including the lack of any bilinguals, the brevity of the inscriptions (the average inscription is  $\approx 5$  signs long), and our almost complete lack of knowledge of the language(s) used in the civilization.

Given these formidable obstacles to direct decipherment, we propose instead the analysis of the script’s syntactical structure using techniques from the fields of statistical pattern analysis and machine learning (5). It is our belief that such an approach could provide valuable insights into the grammatical structure of the script, paving the way for a possible eventual decipherment. As a first step in this endeavor, we present here results obtained

from analyzing the sequential structure of the Indus script using a simple type of probabilistic graphical model (6) known as a Markov model (7). Markov models assume that the current “state” (e.g., symbol in a text) depends only on the previous state, an assumption that renders learning and inference over sequences tractable. Although this assumption is a simplification of the complex sequential structure seen in languages, Markov models have proved to be extremely useful in analyzing a range of sequential data, including speech (8) and natural language (9). Here, we apply them to the Indus script. A major goal of this article is to provide an exposition of Markov models for those unfamiliar with them and to illustrate their usefulness in the study of an undeciphered script. We leave the application of more advanced higher-order models and grammar induction methods to other papers (10) and future work.

**Markov Models for Analyzing the Indus Script.** A Markov model (also called a Markov chain) (7, 11, 12) consists of a finite set of  $N$  “states”  $s_1, s_2, \dots, s_N$  (e.g., the states could be the signs in the script) and a set of conditional (or transition) probabilities  $P(s_i|s_j)$  that determine how likely it is that state  $s_i$  follows state  $s_j$ . There is also a set of prior probabilities  $P(s_i)$  that denote the probability that state  $s_i$  starts a sequence. Fig. 1B shows an example of a “state diagram” for a Markov model with 3 states labeled A, B, and #, where A and B denote letters in a language, and # denotes the end of a text. Fig. 1B also shows the prior probabilities  $P(s_i)$  and the transition probabilities  $P(s_i|s_j)$ , picked arbitrarily here for the purposes of illustration. Some example sequences generated by this Markov model are BAAB, ABAB, B, etc. (the terminal sign # is not shown). Texts that are not generated by this Markov model include all texts that contain a repetition of B (...BB...) and all texts that end in A, because these are precluded by the transition probability table. A more realistic example would be a Markov model for English texts involving the 26 letters of the alphabet plus space. In this case, the transition probability table (or matrix) would be of size  $27 \times 27$ . In the matrix, we would expect, for example, higher probabilities for the letter “s” to be immediately followed by letters such as “e,” “o,” or “u,” than letters such as “x” or “z,” because of the morphological structure of words in English. A Markov model learned from a corpus of English texts would capture such statistical regularities.

Markov models [and their variants, hidden Markov models (HMMs)] are special cases of a more general class of probabilistic models known as graphical models (6). Graphical models can be used to model complex relationships between states, including higher-order dependencies such as the dependence of

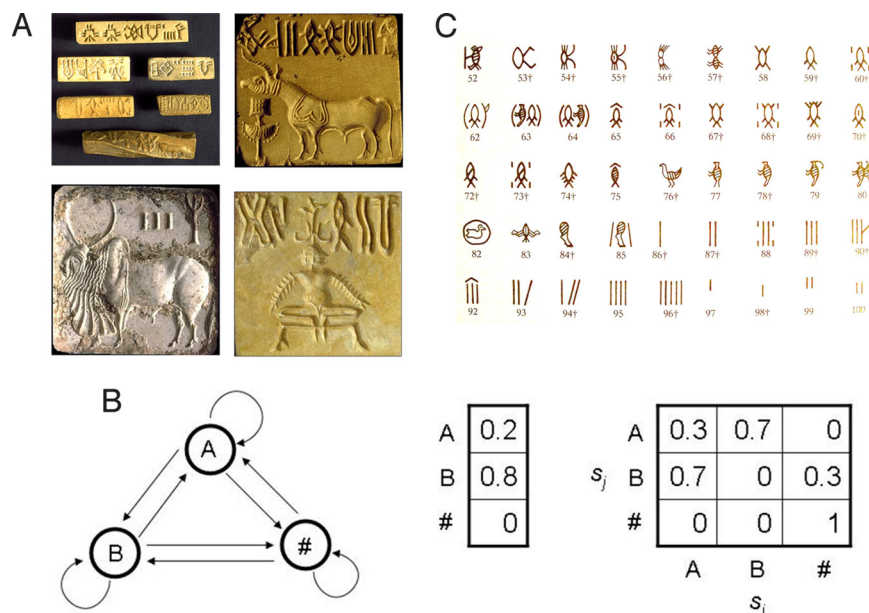
Author contributions: R.P.N.R., N.Y., M.N.V., H.J., and R.A. designed research; R.P.N.R., N.Y., H.J., and R.A. performed research; I.M. contributed new reagents/analytic tools; R.P.N.R., N.Y., M.N.V., H.J., and R.A. analyzed data; and R.P.N.R., M.N.V., and R.A. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence should be addressed. E-mail: rao@cs.washington.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0906237106/DCSupplemental](http://www.pnas.org/cgi/content/full/0906237106/DCSupplemental).



**Fig. 1.** The Indus script and Markov models. (A) Examples of Indus inscriptions on seals and tablets (Upper Left) and 3 square stamp seals (image credit: J. M. Kenoyer, Harappa.com). Note that inscriptions appear reversed on seals. (B) Example of a simple Markov chain model with 3 states. (C) Subset of Indus signs from Mahadevan's list of 417 signs (13).

a symbol on the past  $N-1$  symbols [equivalent to  $N$ -gram models in language modeling (9)]. In this article, we focus on first-order Markov models. In other work (10), we have compared  $N$ -gram models for Indus texts using the information-theoretic measure of “perplexity” (9). We have found that the bulk of the perplexity in the Indus corpus can be captured by a bigram ( $n = 2$ ) model (or equivalently, a first-order Markov model), supporting the usefulness of such models in analyzing the Indus script.

We focus in this article on 3 applications of Markov models to analyzing the Indus script: (i) Sampling: We show how new sample texts can be generated by randomly sampling from the prior probability distribution  $P(s_i)$  to obtain a starting sign (say  $S_1$ ) and then sampling from the transition probabilities  $P(s_i|S_1)$  to obtain the next sign  $S_2$ , and so on. The string generation process can be terminated by assuming an end of text (EOT) sign that transitions to itself with probability one. (ii) Likelihood computation: If  $x$  is a string of length  $L$  and is of the form  $x_1x_2...x_L$ , where each  $x_i$  is a sign from the list of signs, the likelihood that  $x$  was generated by a Markov model  $M$  can be computed as:  $P(x|M) = P(x_1x_2...x_L|M) = P(x_1)P(x_2|x_1)P(x_3|x_2)...P(x_L|x_{L-1})$ . The likelihood of a string under a particular model tells us how closely the statistical properties of the given string match those of the original strings used to learn the given model. Thus, if the original strings were generated according to the statistical properties of a particular language, the likelihood is useful in ascertaining whether a given string might have been generated by the same language. (iii) Filling in missing signs: Given a learned Markov model  $M$  and an input string  $x$  with some known signs and some missing signs, one can estimate the most likely complete string  $x^*$  by calculating the most probable explanation (MPE) for the unknown parts of the string (see *Methods* for details).

**The Indus Script.** A prerequisite for understanding the sequential structure of a script is to identify and isolate its basic signs. This task is particularly difficult in the case of an undeciphered script such as the Indus script, because there is considerable variability in the rendering of the signs, making it difficult to ascertain whether 2 signs that look different are stylistic variants of the

same sign or 2 independent signs. After an analysis of the positional statistics of variant signs in a corpus of known inscriptions (*ca.* 1977), Mahadevan arrived at a list of 417 independent signs in his concordance (13) [Parpola used similar methods to estimate a slightly shorter list of 386 signs (14)]. We used this list of 417 signs as the basis for our study. Fig. 1C shows a subset of signs from this list.

Barring a few exceptions (see p. 14 in ref. 13), the writing direction is generally accepted to be predominantly from right to left (i.e., left to right in seals and right to left in the impressions). There exists convincing external and internal evidence supporting this hypothesis (e.g., refs. 4, 13, and 14). We consequently learned the sequential structure of the Indus texts based on a right-to-left reading of their signs, although a Markov model could equally well be learned for the opposite direction.

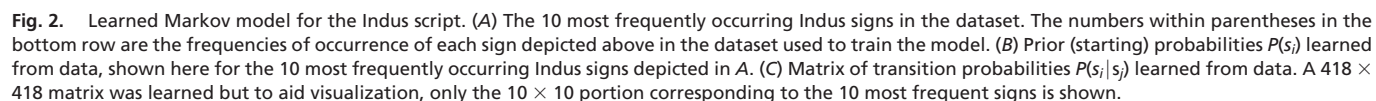
## Results

**Markov Model of Indus Texts.** The learned Markov model provides several interesting insights into the nature of the Indus script. First, examining the learned prior probabilities  $P(s_i)$  provides information about how likely it is that a particular sign  $s_i$  starts a text. Fig. 2B shows this probability for the 10 most frequently occurring signs (Fig. 2A) in the corpus of Indus inscriptions.

An examination of Fig. 2B reveals that certain frequently occurring signs such as  $\diamond$  and  $\otimes$  (signs numbered 3 and 10 in Fig. 2B) are much more likely to start a text than the others. On the other hand, certain signs such as  $\text{U}$  (the most frequent sign in the corpus) and  $\uparrow$  are highly unlikely to start a text (they are, in fact, highly likely to end a text; see Fig. 3C and Table 1). These observations have also been made by others (13–16).

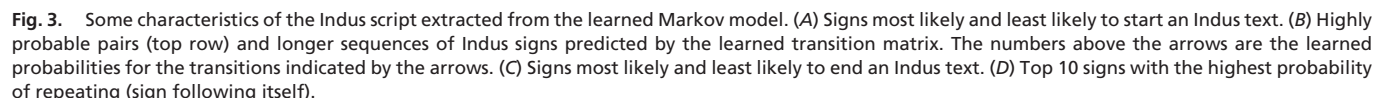
From the learned values for  $P(s_i)$ , one can also extract the 10 most likely signs to start a text and the 10 least likely signs to do so (among signs occurring at least 20 times in the dataset) (Fig. 3A). These results suggest that some signs that look similar, such as  $|$  and  $!$ , subserve different functions within the script and thus cannot be considered variants of the same sign.

The matrix of transition probabilities  $P(s_i|s_j)$  learned from data is shown in Fig. 2C (only the portion corresponding to the 10 most frequent signs is shown). The learned transition prob-



The diagonal of the transition matrix [representing  $P(s_i|s_i)$ ] tells us how likely it is that a given sign follows itself (i.e., repeats). Fig. 3D shows 10 signs with the highest probability of repeating. The sign  $\uparrow^*$  occurs in only one inscription where it occurs as a pair, whereas the sign  $\uparrow$  occurs in 8 inscriptions, 6 times as a pair. More interestingly, the frequently occurring sign  $\boxplus$  occurs as a pair in 33 of the 58 inscriptions in which it occurs. The presence of such repeating symbols in the Indus texts puts

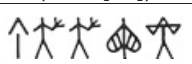




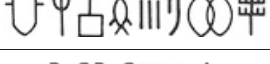
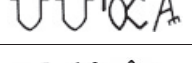

Table 1 provides additional support for the hypothesis that sequences of signs in the Indus texts may be governed by specific rules. Table 1 shows that for each of the 10 most frequently occurring signs in the dataset, only a small set of signs has a high











**Table 2. Likelihood of West Asian Texts compared with Indus valley texts**






West Asian Text (from [13])	Likelihood
	0
	$3.11 \times 10^{-10}$
	$7.09 \times 10^{-8}$
	$6.34 \times 10^{-14}$
	0
	$1.13 \times 10^{-11}$
	$1.22 \times 10^{-12}$
	$2.87 \times 10^{-17}$
Indus valley held-out texts (median)	$1.12 \times 10^{-7}$

Note: Only complete and unambiguous West Asian texts from ref. 13 are included in this table. Two texts have a likelihood of zero, because they each contain a symbol not occurring in the training dataset used to learn the model. The last row shows for comparison the median likelihood for a randomly selected set of 100 texts originating from within the Indus valley, which were held out and not used for training the Markov model (these 100 texts had the same average length as the West Asian texts).

text at the bottom. The closest matching text differs from the generated text in 2 ways: the “modifier”  has been inserted and the sign  is replaced by a visually similar sign . The text shown at the left of Fig. 4E is another actual Indus inscription with an unknown number of signs missing (from ref. 13). The 3 texts shown at the right are MPE outputs assuming 1, 2, or 3 signs are missing. The first and third MPE texts actually occur in the Indus corpus, whereas the middle text contains the frequently occurring pair . Additional examples of filling in of damaged texts are given in [supporting information \(SI\) Table S1](#).

**Testing the Likelihood of Indus Inscriptions.** The likelihood of a particular sequence of Indus signs with respect to the learned Markov model tells us how likely it is that the sign sequence belongs to the putative language encoded by the Markov model. We found that altering the order of signs in an existing Indus text typically caused the likelihood of the text to drop dramatically ([SI Text](#) and [Fig. S1](#)), supporting the hypothesis that the Indus texts are subject to specific syntactic rules determining the sequencing of signs.

Applying this analysis to Indus texts discovered outside the Indus valley, for example, in Mesopotamia and other sites in West Asia, we found that the likelihoods of most of these inscriptions are extremely low compared with their counterparts found in the Indus valley (Table 2). Indeed, the median value of likelihoods for the West Asian texts is  $6.40 \times 10^{-13}$ , which is  $\approx 100,000$  times less than the median value of  $1.12 \times 10^{-7}$  obtained for a random set of 100 texts of Indus valley origin that were excluded from the training set for comparison purposes.

These findings suggest the intriguing possibility that the Indus script may have been used to represent a different language or subject matter by Indus traders conducting business in West Asia or West Asian traders sending goods back to the Indus valley. Such a possibility was earlier suggested by Parpola, who noted that the West Asian texts often contain unusual sign combinations (14). Our results provide a quantitative basis for this possibility. The low likelihoods arise from the fact that many of the West Asian texts in Table 2 contain sign combinations such as , , , , and  that never appear in any texts found in the Indus valley, even though the signs themselves occur frequently in other combinations in the Indus valley texts.

## Discussion

A number of researchers have made observations regarding sequential structure in the Indus script, focusing on frequently occurring pairs, triplets, and other groups of signs (13–15, 20). Koskeniemi suggested the use of pairwise frequencies of signs to construct syntax trees and segment texts, with the goal of eventually deriving a formal grammar (20). More recently, Yadav, Vahia, and colleagues (15, 16) have performed statistical analyses of the Indus texts, including explicit segmentation of texts based on most frequent pairs, triplets, and quadruplets.

In this article, we provide an investigation of sequential structure in the Indus script based on Markov models. An analysis of the transition matrix learned from a corpus of Indus texts provided important insights into which signs tend to follow particular signs and which signs do not. The transition matrix also provides a probabilistic basis for extracting common sequences and subsequences of signs in the Indus texts. We demonstrated how the learned Markov model can be used to generate new sample texts, revealing groups of signs that tend to function similarly within a text. The approach can also be used to fill in missing portions of illegible and incomplete Indus inscriptions based on the corpus of complete inscriptions. Finally, a comparison of the likelihood of Indus inscriptions discovered in West Asian sites with those from the Indus valley suggests that many of the West Asian inscriptions may represent subject matter different from Indus valley inscriptions.

Our results appear to favor the hypothesis that the Indus script represents a linguistic writing system. Our Markov analysis of sign sequences, although restricted to pairwise statistics, makes it clear that the signs do not occur in a random manner within inscriptions but appear to follow certain rules: (i) some signs have a high probability of occurring at the beginning of inscriptions whereas others almost never occur at the beginning; and (ii) for any particular sign, there are signs that have a high probability of occurring after that sign and other signs that have negligible probability of occurring after the same sign. Furthermore, signs appear to fall into functional classes in terms of their position within an Indus text, where a particular sign can be replaced by another sign in its equivalence class. Such rich syntactic structure is hard to reconcile with a nonlinguistic system. Additionally, our finding that the script may have been versatile enough to represent different subject matter in West Asia argues against the claim that the script merely represents religious or political symbols. Other arguments in favor of the linguistic hypothesis for the Indus script are provided by Parpola (21).

Our study suffers from some shortcomings that could be addressed in future work. First, our first-order Markov model captures only pairwise dependencies between signs, ignoring important longer-range dependencies. Higher-order Markov models (10) and other types of probabilistic graphical models (6) would allow more accurate modeling of such dependencies. A second potential shortcoming is our use of an Indus corpus of texts from 1977 (13). New texts and signs have since been discovered, and new sign lists have been suggested with up to 650 signs (22). However, the types of new material that have been



# Supplementary Information

Rao et al. 10.1073/pnas.0906237106

## SI Text

**Examples of Damaged Texts from Mahadevan's Concordance Filled in Using MPE.** Table S1 shows complete texts predicted using MPE for a number of damaged texts given in Mahadevan's concordance [Mahadevan I (1977) *The Indus Script: Texts, Concordance, and Tables* (Memoirs of Archaeological Survey of India, New Delhi, India)]. The text numbers in Table S1 are from the concordance.  $\text{\textcircled{.}}$  denotes one or more missing signs. \* denotes a doubtfully read sign. Because the exact number of missing signs can be obscured by the damage to the object, filled-in texts are shown for the cases of 1 and 2 signs missing for each  $\text{\textcircled{.}}$ .

**Likelihood of Altered and West Asian Texts.** To test whether the model could be used to ascertain whether a text belongs to the

putative language encoded by the Markov model, we generated several example texts by altering existing Indus texts. Alterations involved randomly switching the positions of signs within a text. Fig. S1 shows an example of an Indus valley text, the same text altered (last sign moved to the beginning), and an Indus text discovered in West Asia. The likelihoods for each according to the learned model are given on the right. The altered text has a very small probability of  $5.2 \times 10^{-10}$  under the model, consistent with the observations that  $\text{\textcircled{.}}$  rarely begins a text and the sign pair  $\text{\textcircled{.}}$  almost never occurs. Interestingly, the West Asian text is  $\approx 100$  times less likely than even this altered text and  $\approx 10^6$  times less likely than the Indus text under the learned model, suggesting that the script may have been used to encode different subject matter in West Asia. Additional examples of West Asian texts and their likelihoods are given in Table 2.

		Likelihood
Indus text		$8.0 \times 10^{-6}$
Altered text		$5.2 \times 10^{-10}$
West Asian seal		$1.2 \times 10^{-12}$




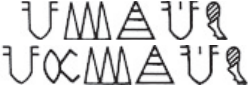


Fig. S1. Comparison of likelihoods of example Indus text, altered text, and West Asian text.



Table S1. Filling in of damaged texts from Mahadevan's concordance

Text No.	Text with Missing Signs	Filled-In Texts
1001	𑀩𑀓𑀭𑀮𑀭𑀮𑀭𑀮	𑀩𑀓𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀩𑀓𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1002	𑀭𑀮𑀭𑀮𑀭𑀮*	𑀩𑀭𑀮𑀭𑀮 𑀭𑀮𑀩𑀭𑀮𑀭𑀮
1017	𑀭𑀮𑀭𑀮*𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1020	𑀩𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮	𑀩𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀩𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1034	𑀩𑀭𑀮𑀭𑀮	𑀩𑀭𑀮𑀭𑀮𑀭𑀮 𑀩𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1059	𑀭𑀮𑀭𑀮𑀭𑀮*𑀭𑀮𑀭𑀮	𑀩𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀩𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1111	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1144	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1165	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1176	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮*	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮
1240	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮	𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮 𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮𑀭𑀮

Table S1. Cont'd

1265		
1289		
1343		
1408	