Comparing genomes and languages using current knowledge

In a recent correspondence, Hong-Yu Zhang proposes an analogy between the evolutionary mechanisms of genomes and language, using the Chinese language as an example (Zhang, 2006). Although his idea—that the evolutionary strategies of both systems are themselves evolving—is interesting, we believe some of Zhang's analysis is based on incorrect information. This makes his arguments misleading, particularly to non-Chinese-speaking readers.

Zhang bases many of his arguments on the extent and speed with which Chinese characters were created during different historical periods, using data taken from Ji (1989) and references therein. We doubt the validity of these data, and believe that Zhang's interpretation is mistaken. The superficial precision in these values masks their inaccuracies, which we demonstrate by redrawing Zhang's Table 1 (Fig 1). According to these data, the largest increase in the number of Chinese characters occurred during the Qing Dynasty: 13,856 characters were invented at a rate of more than 50 per year (Fig 1). This unprecedented peak occurred despite China's turbulent history in the latter part of that dynasty, which ended with the establishment of the Republic of China in 1911. The Qing Dynasty was notorious for its widespread persecution of dissident intellectuals; it is difficult to imagine how so many new characters could be invented during such a period of oppression. We also wonder how the number of characters in each dynasty can be known so precisely.

Zhang points out that 'virtual', or modifying, elements were quite rare in Oracle bone inscriptions (before 11th century BC) but are more frequent in modern Chinese. In fact, virtual elements were frequently and elegantly used in ancient Chinese as far back as 3,000 years ago. Fig 2A shows two examples: the first character is equivalent to the one Zhang proposed, meaning "of". The

second has a structural role in a sentence. Virtual, or 'empty' characters such as these were well developed and widely used in scripts far before the Han Dynasty, which began in 206 BC.

We also disagree with Zhang's statement that "modern Chinese is definitely more powerful than its ancient predecessor, because it is able to describe a much more complex world." In fact, ancient Chinese is no less powerful. With the addition of few new characters—such as Chinese names for most chemical elements—ancient Chinese fits in well with modern society. For example, the characters used to describe inventions such as the computer (*ji suan ji*) and Internet (*hu lian wang*; Fig 2B) can all be found in dictionaries compiled before the Qing Dynasty. Ancient Chinese has its own merits; even in modern China, some official documents incorporate the concision and meaningfulness of the ancient style to convey formality and authority.

Zhang's analogy between virtual elements and non-coding RNA is based on a misunderstanding. By writing "Using RNA rather than proteins as regulators circumvents protein translation, and the average length of such RNA signals—22 nucleotides—is almost two orders of magnitude shorter than that required to encode an average protein," Zhang misinterprets John Mattick's statement that "A sequence-specific RNA signal in animals and plants can be just 22 nucleotides" (Mattick, 2004). The number of nucleotides can be as small as 22, but this is certainly not the average length.

It is possible to evaluate the similarities between genomes and languages, as both are involved in the process of encoding and decoding information, but only if the two are compared using similar foundations. Unfortunately, by basing his analysis on inaccurate information, Zhang fails to present a convincing argument. It is still too early to compare the evolution of languages and genomes, partly because we do not have enough knowledge about the initial events that triggered both processes. There are various theories to explain the origin of languages (Jespersen, 1922), but no unanimous conclusion has yet been reached (Beaken,

Fig 1 | Evolution of Chinese characters. (A) Number of Chinese characters invented cumulatively. (B) Average speed of character invention during different historical periods. (Redrawn from Table 1 in Zhang, 2006.)

Fig 2 | Examples of ancient Chinese characters. (A) Two virtual elements widely and frequently used in ancient Chinese. (B) Two Chinese words for modern inventions expressed with ancient Chinese characters, meaning "computer" and "Internet", respectively. The simplified forms are shown in brackets.

1996). For genomic evolution, there is an even larger information gap between the evolution of biological macromolecules and

the first functional cells. The initial processes might have profound impacts on the subsequent evolutionary strategies. Before these events are fully understood, tentative hypotheses on their similarities are welcome, but only if they are supported by current and well-founded knowledge.

REFERENCES
Beaken M (1996) *The Making of Language*. Edinburgh, UK: Edinburgh University Press
Jespersen O (1922) *Language: Its Nature, Development and Origin*. London, UK: Allen & Unwin
Ji XL (1989) *Encyclopedia of China: Linguistics and Philology*. Beijing, China: Publishing House of Encyclopedia of China
Mattick JS (2004) RNA regulation: a new genetics? *Nat Rev Genet* 5: 316–323
Zhang H-Y (2006) The evolution of genomes and language. *EMBO Rep* 7: 748–749

Zhen-Ling Sun is in the Department of Animal and Plant Sciences at the University of Sheffield, UK.
E-mail: z.l.sun@sheffield.ac.uk
John Zhu is in the KROTO Research Institute at the University of Sheffield, UK.

Response from the author

In this era of systems science, one of the greatest challenges is to find universal properties for evolving systems. As language and genomes are two of the best-studied systems, comparing their evolution might reveal some intrinsic properties. This was the motive behind my recent correspondence (Zhang, 2006), in which I argue that, in the evolution of genomes and language, not only is complexity enhanced, but also the strategies to enhance this complexity undergo paradigm shifts. As this work aimed to provoke discussions, I welcome Sun and Zhu's response (Sun & Zhu, 2006). Although their criticisms have little relevance to my overall conclusion, I will address them here.

First, information on the quantity of Chinese characters was taken from various ancient dictionaries (Table 1; Ji, 1989); as the source information is trivial to my conclusion, it was removed from my original correspondence but is listed here. The precise publication date of each dictionary cannot be traced in most cases; therefore, I used the time-scale of the corresponding dynasty merely to convey a sense of historical evolution. My aim was to demonstrate the clear increase in the quantity of Chinese characters, although the exact speed of character invention is unknown. This is similar to the situation in genomics: the fact that the exact number and evolutionary speed of genes is unknown does not prevent us from believing that there is an increase in gene numbers from prokaryotes to eukaryotes. Consequently, the speed of Chinese character invention is irrelevant to my arguments and conclusion.

Second, I stated that "virtual words were quite rare in Oracle" because most virtual characters were undergoing evolution from solid characters in Oracle (Jiang, 1994). It is interesting to note that one of the two characters shown by Sun and Zhu provides a good example that supports rather than refutes my statement. The character *zhi* (Fig 2A, left, from their correspondence) was not a preposition in Oracle, but instead served as a pronoun (sometimes as nominal and sometimes as predicate; Zhang, 2005). Now as a virtual character, it means "of", but in post-Oracle ancient Chinese it had many other meanings such as "go" or "leave for". Virtual characters were prevalent in post-Oracle ancient Chinese, but not in Oracle itself. By comparison, genome evolution exhibits a similar feature: although plants are not the most advanced organisms, they depend largely on non-coding RNA regulation, whereas prokaryotes do not.

Third, Sun and Zhu misunderstand my statement that "modern Chinese is definitely

more powerful than its ancient predecessor." My point was that the modern Chinese words (>70,000) that are derived from a combination of about 3,500 characters can cover the meaning of most ancient characters (>47,000). In fact, the examples (*ji suan ji* and *hu lian wang*) proposed by Sun and Zhu support my opinion that new meanings can often be readily expressed by a combination of old characters.

Fourth, it is well known that the most intensively studied regulatory RNAs—for example, microRNAs and short interfering RNAs—are generally 21–25 nucleotides long (Ambros, 2004; Bartel, 2004; He & Hannon, 2004; Mattick & Makunin, 2005; Shi, 2003). The inadvertent use of the term "average length" in my original correspondence does not negate my analogy between virtual elements and non-coding RNA.

In conclusion, the above explanations indicate that Sun and Zhu's criticisms have no impact on my original conclusion. I maintain that there are similar strategy shifts in the evolution of genomes and language. It will be interesting to investigate whether these features are also shared by other evolving systems, such as economic systems.

REFERENCES
Ambros V (2004) The functions of animal microRNAs. *Nature* 431: 350–355
Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116: 281–297
He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* 5: 522–531
Ji XL (1989) *Encyclopedia of China: Language and Character*. Beijing, China: Encyclopedia of China Publishing House
Jiang BC (1994) Analysis on virtual words in divination. In Cheng XQ (ed) *Research on Chinese of Pre-Qin Dynasty* pp 22–44. Jinan, China: Shandong Education Press
Mattick JS, Makunin IV (2005) Small regulatory RNAs in mammals. *Hum Mol Genet* 14 (Suppl 1): R121–R132
Shi Y (2003) Mammalian RNAi for the masses. *Trends Genet* 19: 9–12
Sun ZL, Zhu J (2006) Comparing genomes and languages using current knowledge. *EMBO Rep* 7: 954–955
Zhang HY (2006) The evolution of genomes and language. *EMBO Rep* 7: 748–749
Zhang YJ (2005) Origin of the particle "zhi" in Classical Chinese. *Yindu J* 2: 13–17

Hong-Yu Zhang is at the Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Shandong University of Technology, Zibo, People's Republic of China.
E-mail: zhanghy@sdut.edu.cn
doi:10.1038/sj.embor.7400809

Table 1 | Evolution of Chinese characters (Ji, 1989)

Era (oldest to most recent)	Quantity of characters	Source
Shang Dynasty	~5,000	Oracle bone inscriptions
Eastern Han Dynasty	9,353	<i>Shuo Wen Jie Zi</i> , by Xu Shen
Jin Dynasty	12,824	<i>Zi Lin</i> , by Lü Chen
Liang Dynasty	16,917	<i>Yu Pian</i> , by Gu Ye-Wang
Tang Dynasty	22,561	<i>Zeng Zi Ben Yu Pian</i> , by Sun Qiang
Northern Song Dynasty	31,319	<i>Lei Pian</i> , by Si-Ma Guang
Ming Dynasty	33,179	<i>Zi Hui</i> , by Mei Ying-Zuo
Qing Dynasty	47,035	<i>Kangxi Dictionary</i> , by Chen Ting-Jing