components, enabling it to be trained to process bytes (Fig. 1).

The first component is a local encoder that reads input text and builds a mathematical representation of each byte. A boundary predictor then decides where one patch should end and the next begin, so that the bytes can be grouped and handed to the original model. A local decoder then turns that model's output back into bytes, one at a time. In this way, the subword model reads and writes patches even though the system as a whole reads and writes bytes.

The new components and the original model are then trained in two stages. In the first, the parameters of the original subword model are fixed but those of the added components can be adjusted. The aim of this stage is to output text that matches, as closely as possible, that of the original subword model. In the second stage, the original model and the extra components can be adjusted and they are trained together to predict the next byte in a sequence of text. The aim of this stage is to train the subword model to use the information encoded in bytes effectively.

Applying byteification to several existing LLMs with between 7 billion and 8 billion parameters produced byte-level models that outperformed published models of comparable size. The average performance of the byteified models also came close to that of the subword models from which they were derived. Byteified models also outperformed the original subword models on tasks that require the manipulation of individual characters.

Pretraining, which is the phase in which a model learns from a huge body of text, accounts for most of the cost of building an LLM. Byteification required less than 1% of a typical pretraining budget, turning the construction of a competitive byte-level model from an expensive restart into a comparatively inexpensive conversion of an existing model. Byteification also preserves a connection to the infrastructure and training algorithms that have been used to develop mature subword models. In one demonstration, the researchers transferred the instruction-following ability of a trained subword model to its byteified counterpart without extra training.

Byte-level models are not yet ready to replace subword models, however. They perform slightly worse in tests, and it is not yet clear whether the high-throughput inference infrastructures developed for subword models can be adapted to byte-level models. Subword tokenization is therefore likely to remain the default approach for most LLMs in the near future, although byte-level models could be useful for situations in which access to individual characters in text is important, such as analysing the spelling of words. Nevertheless, the authors have shown that it is possible to combine the strengths of byte-level models with the knowledge, engineering tools and specialized capabilities that have already been developed for subword LLMs.

Zhao Zhang and **Yingfei Xiong** are in the Key Lab of High Confidence Software Technologies, School of Computer Science, Peking University, Beijing 100871, China.
e-mails: zhangzhao2019@pku.edu.cn; xiongyf@pku.edu.cn

1. Minixhofer, B. *et al. Nature* **658**, 915–922 (2026).
2. Sennrich, R., Haddow, B. & Birch, A. in *Proc. 54th Ann. Meet. Assoc. Comput. Linguist.* 1715–1725 (2016).
3. Cosma, A., Ruseti, S., Radoi, E. & Dascalu, M. in *Proc. 2025 Conf. Empir. Meth. Nat. Lang. Proc.* 28252–28263 (2025).
4. Pagnoni, A. *et al. Proc. 63rd Ann. Meet. Assoc. Comput. Linguist.* 9238–9258 (2025).

The authors declare no competing interests.
This article was published online on 7 October 2026.

Microbiology

Gut microbes that trace ancient human journeys

Carina M. Schlebusch & Maja Vukovikj

The Indigenous African Hadza and South American Tsimane share gut microorganisms that shed light on human dispersals and microbial losses in industrialized societies. **See p.1008**

Humans have never travelled alone. When our ancestors spread out of Africa and across the other continents, they carried their microorganisms with them. On page 1008, Carter *et al.*[1] compare the community of gut microbes (also known as the microbiota) from the Indigenous Hadza people of Tanzania with that of the Tsimane of Bolivia, whose populations diverged during human migrations that occurred tens of thousands of years ago. The authors report that, despite these dispersals, hundreds of microbial species are present in both groups. Furthermore, genetic differences between strains of the same microbial species in Hadza and Tsimane populations were broadly consistent with the timescale of prehistoric human migrations. Many of these shared microbes are now rare or missing in populations from industrialized regions.

The gut microbiota is affected by a variety of factors, including diet, environment and transmission between people, with industrialized populations carrying markedly less microbial diversity than do non-industrialized ones. Analysis of ancient faecal samples suggests that the historical microbiota more closely resembled those of contemporary non-industrialized groups than those of modern industrialized populations[2]. However, such ancient samples are generally no more than 1,000 or 2,000 years old, whereas modern humans dispersed and established communities around the globe many tens of thousands of years ago. It has therefore remained unclear to what extent present-day gut microbial diversity traces the history of those ancient human migrations, rather than reflecting more modern ecological processes or the effects of demographic factors.

Carter and colleagues sequenced faecal samples from Tsimane forager-horticulturalists and compared them with existing data from East African Hadza hunter-gatherers. Using an approach called deep metagenomic sequencing enabled the authors to determine thousands of microbial genomes and to examine strain-level variation between the microbes in the two human populations.

Of the 1,408 microbial (bacterial and archaeal) species identified in the Tsimane microbiota, 1,231 (87.4%) were also detected in the Hadza data set, despite the populations' geographical separation and lack of direct historical contact. Of these shared species, more than two-thirds (848 in total) are rare or absent in industrialized microbiota. Many of these species might therefore have been associated with humans before the migrations that ultimately separated the ancestors of the two populations.

A microbial species can contain many genetically distinct strains and the Hadza and Tsimane populations mostly carried different strains of the microbial species they shared. Of 636 well-sampled species, 545 showed no evidence of a shared strain lineage in approximately the past 5,000 years. Thus, the presence of these species in both populations seems to reflect older associations, whereas the differences between the strains is consistent with their later geographical isolation.

The authors inferred that an onset of genetic isolation between the Hadza and Tsimane microbial populations occurred around 17,000 years ago. This is more recent than the out-of-Africa migration, which occurred around 60,000 years ago[3], and provides a minimum timescale for the separation of



L TO R: ALAMY; MICHAEL GURVEN

Figure 1 | Hadza people of Tanzania (left) and the Tsimane of Bolivia (right).

the human lineages represented here. If the microbes accompanied humans during that divergence, the estimate of 17,000 years is probably an underestimation, potentially resulting from uncertainty in microbial mutation rates, continued indirect exchange between intervening human populations or differences among the microbial species analysed. The estimated divergence was nevertheless much older than that obtained from comparisons between industrialized populations, for which contemporary mixing is likely to have been greater.

Microbes have been used to study human demographic history before. For example, geographical variation in the stomach bacterium *Helicobacter pylori* has been used to trace migrations out of Africa, as well as the peopling of the Americas[4,5]. African *H. pylori* lineages suggest an association with humans extending back approximately 100,000 years[6]. The deepest bacterial split involves a lineage associated with southern African San hunter-gatherers. Similarly, human genomic studies indicate that San populations represent one of the earliest population divergences among living humans[7]. Another striking microbial example is *Methanobrevibacter smithii*, the main methane-producing archaeal microbe in the human gut. Ancient genomes suggest that this microbe began diversifying around 85,000 years ago, which broadly coincides with early human migrations out of Africa that occurred before the main population dispersal[2].

Carter and colleagues' study expands on previous examinations of individual microbes by analysing hundreds of gut microbial species. Some of these species might eventually provide useful information about human migrations. If these bacterial lineages are stably associated with people, microbial genomes could be analysed alongside human genetic data. Comparing results across several microbes and their human hosts might provide a more detailed picture of migrations and contacts between populations. However, gut microbial variation is also shaped by diet and lifestyle, with some microbes possibly evolving alongside their hosts in response to these factors[8]. Accounting for these influences might help to distinguish a shared population history from patterns caused by microbial ecology or the transmission dynamics of a particular microorganism.

The dates of population splits are less certain than the evidence for long-term genetic isolation between the microbial populations. The rates at which DNA changes through processes such as mutation and recombination vary among species, and gene transfer directly from one species to another (known as horizontal gene transfer) produces mosaic genomes, complicating genomic analysis and estimates of divergence times. Only two non-industrialized populations were studied in depth by Carter and colleagues, therefore the evidence cannot capture the full range of human microbial diversity. Environmental and animal reservoirs of bacteria were also not sampled, so some microbes might have been acquired from similar habitats rather than being transmitted continuously between human generations.

> **"The authors' study expands on previous examinations of individual microbes."**

Many microbial lineages with long histories of human association are disappearing from industrialized populations. Antibiotic exposure might be an important cause of this, because repeated treatments can eliminate resident gut strains and interrupt their transmission. A study of ancient and modern faeces revealed antibiotic-resistance genes that are present in the modern gut microbiome, indicating that genomes have been shaped by antibiotic exposure in modern times[2]. Those ancient samples also contained more mobile genetic elements (DNA sequences that can move inside or between genomes) than were found in modern microbiomes. These elements might have helped gut microbes to adapt rapidly to local and seasonal changes[2]. Changes in diet, sanitation and other aspects of industrialized life probably also contribute to the shaping of the gut microbial community.

The loss of gut microbes in industrial societies correlates with lifestyles linked to chronic inflammatory and metabolic diseases[9]. However, a long-running association with humans does not by itself show that a given species is beneficial. Reintroducing lost microbes would require functional evidence, safety testing and genuine partnership with the communities whose microbiota made this research possible. Broader sampling of contemporary forager and other non-industrialized communities, combined with ancient microbiome data and functional experiments, are now needed to establish how much information about human history is retained in the microbiome and how much has already been lost.

Carina M. Schlebusch and **Maja Vukovikj** are in the Department of Organismal Biology, Uppsala University, SE-752 36 Uppsala, Sweden. C.M.S. is also at SciLifeLab, Uppsala and the Palaeo-Research Institute, University of Johannesburg, Johannesburg, South Africa.
e-mails: carina.schlebusch@ebc.uu.se; maja.vukovikj@ebc.uu.se

1. Carter, M. M. *et al. Nature* **658**, 1008–1013 (2026).
2. Wibowo, M. C. *et al. Nature* **594**, 234–239 (2021).
3. Bergström, A., Stringer, C., Hajdinjak, M., Scerri, E. M. L. & Skoglund, P. *Nature* **590**, 229–237 (2021).
4. Linz, B. *et al. Nature* **445**, 915–918 (2007).
5. Falush, D. *et al. Science* **299**, 1582–1585 (2003).
6. Moodley, Y. *et al. PLoS Pathog.* **8**, e1002693 (2012).
7. Schlebusch, C. M. *et al. Science* **358**, 652–655 (2017).
8. Otani, S. *et al. Gut Microbes* **17**, 2484385 (2025).
9. Sonnenburg, E. D. & Sonnenburg, J. L. *Nature Rev. Microbiol.* **17**, 383–390 (2019).

The authors declare no competing interests.
This article was published online on 7 October 2026.