

TOOLBOX

New macaque reference genome fills sequence gaps

BY CHLOE WILLIAMS

4 MARCH 2021

Listen to this story:

Researchers have cataloged genetic variation among hundreds of rhesus macaques and created the **most complete macaque reference genome** to date. The resources could help researchers use the animals to **study genes that underlie autism**.

The first macaque reference genome **debuted in 2007**, and most of the DNA was from a single monkey of Indian origin. Scientists have since updated the reference, but the genome still contains gaps.

Part of the problem is that researchers assembled the genome by sequencing fragments of DNA, each about 100 base pairs, or 'letters,' in length, and piecing them together. These 'short reads' make it difficult to reconstruct repetitive sequences. "Think of it as a puzzle where you have [pieces of] blue sky, and you don't know where they go," says **Wesley Warren**, professor of genomics at the University of Missouri in Columbia, who co-led the new work.

Warren and his colleagues sequenced DNA fragments upwards of 10,000 letters long. The result is the most contiguous macaque reference genome yet.

The team extracted long DNA fragments from the cells of a female macaque of Indian origin and sequenced them using existing methods. They then used software to assemble the sequences and map them onto chromosomes. The researchers also used a sequencing technique that tracks the direction of a stretch of DNA to try to correct segments that were inverted.

The new reference, which is **available online**, bridges more than 99.7 percent of the gaps in the

previous Indian-origin macaque genome, the researchers reported in December in *Science*. In addition, they estimate that fewer than 4 million letters are misoriented in the new genome, compared with more than 130 million in the previous reference.

Monkey mutations:

To identify genes among the DNA stretches, the researchers sequenced RNA — the intermediary between genes and proteins — from macaque brain tissue, testes and stem cells.

The team identified tens of thousands of genes. They also identified thousands of new **noncoding RNAs**, which do not code for proteins but instead regulate gene expression, and discovered novel isoforms — alternate versions of RNA encoded by the same gene — that are specific to macaques.

To assess genetic variation among macaques, the researchers sequenced blood, DNA or tissue samples from 850 captive macaques in the United States and three wild animals from China. They scanned the sequences for alterations, such as changes to single DNA letters or ‘indels’ — small insertions or deletions of letters.

The team cataloged 85.7 million single-letter changes and 10.5 million ‘indels,’ including 824,200 variants that fall within genes that code for proteins. After analyzing these variants, the researchers identified nine macaque genes — including **SHANK3**, **CHD8** and **ARID1B** — that harbor potentially harmful mutations and are linked to autism and other neurodevelopmental conditions in people.

Scientists could study macaques with these naturally occurring mutations to find out if they have characteristics similar to those seen in people with autism, the researchers say. These animals could form the basis for new macaque models of autism.