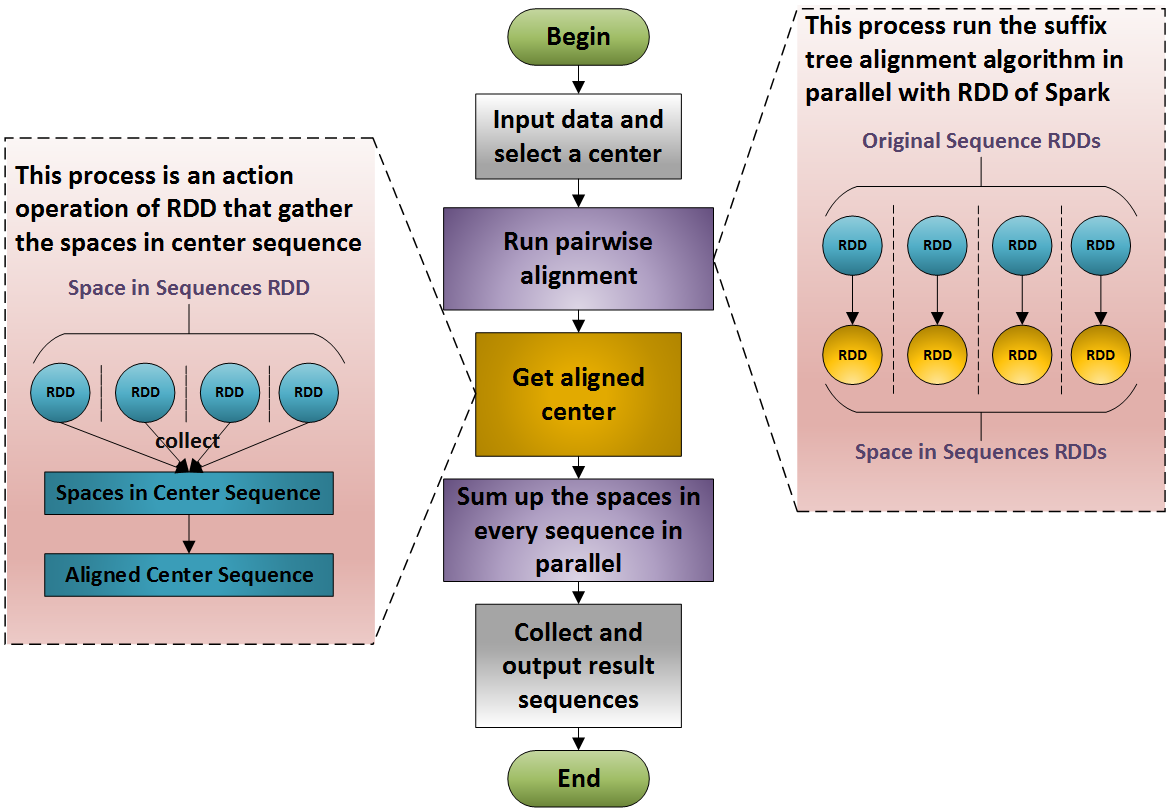
**Supplementary Method**

**Spark** Spark is a MapReduce type framework that can carry on all computation in memory, without needing to save intermediate results. The main abstraction Spark provides is a resilient distributed dataset (RDD), which is a collection of elements partitioned across the nodes of the cluster that can be operated on in parallel. RDDs are created by starting with a file in the file system, or an existing collection in the driver program, and transforming it. The collection means a set of data which is distributed on different computers in a cluster.

**Design of Spark version of MASC** Our Spark center-star MSA has two stages which is shown in Supplementary Figure 2.



**Supplementary Figure 2** Implementation of MASC on Spark

Initially, data is input from the local file system or an Apache Hadoop Distributed File System (HDFS), and a format examination is performed. Next, the sequence array is parallelized, which consists of converting the string lists to sequence string RDDs. Concurrently, the program needs to choose a center sequence, which, as previously explained, is randomly chosen, because any sequence can be regarded as average, if all are sufficiently similar. This randomly selected sequence is then used to construct a center sequence suffix tree, and pairwise alignments are run between that suffix tree and all other sequences, using Spark, making use of the algorithm previously illustrated. In this step, the suffix tree is used to get matching substrings in each sequence serially, because this step runs extremely fast, and wasting memory can be avoided by keeping a single copy of the center suffix tree. When the information regarding all the matched substrings is obtained, a parallel process of aligning the unmatched substrings is run by passing the function to the Spark transformation, which implements the Needleman–Wunsch algorithm. In the previous step, the sequence string RDDs are transformed into pair RDDs of spaces in the center sequence and spaces in all other sequences. Next, the RDDs of spaces in the center sequence are collected to make the center aligned. Then the aligned center sequence is broadcast to all executors, and align all other sequences. In this step the pair RDDs of spaces are transformed into string RDDs of aligned sequences, and the results are stored. Finally, the aligned string RDDs are collected and are output to the local file system. The data flow and operations are shown in Supplementary Figure 3.



**Supplementary Figure 3** Data flow and operations

In Figure 6, all the nodes represent data and the edges represent the operations. The cycle elements are the RDDs in Spark that are distributed in the executors, the rectangle elements are datasets in the driver’s memory. The operations signed by solid lines represent the transformations in Spark that transfer RDDs to subsequent RDDs and the dotted line edges represent the action operations that convert RDDs into dataset in driver memory.

**Process of Experiment** A dataset contains 600 homologous sequences selected randomly from mt genome (Tanaka *et al.*, 2004) dataset is used to execute 600 center-star MSA, which use every sequence in the set as center. Then the results of these experiments are quantitatively assessed as raw score, here, SP score (Zou *et al.*, 2009) is used.

**Transformation of SP score to Z score** SP score was normalized and transformed into a z-score. The formula is as following:

Given,

[Sample SP score mean]

[Sample SP score std. dev]

Then,

The distribution of these normalized SP scores is shown in Supplementary Figure 1.

Distribution%20of%20SP.pdf

**Supplementary Figure 2** The distribution of normalized SP score

When a MSA is carried out with a random selected center sequence, the probability of Z-score less than 4, which means the SP score is less than 17063, is 0.9983. The SP of the result of MAFFT and KAlign is between 16500 and 17000.

**Bibliography**

Tanaka,M. *et al.* (2004) Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan. *Genome Res.*, 1832–1850.

Zou,Q. *et al.* (2009) An Algorithm for DNA Multiple Sequence Alignment Based on Center Star Method and Keyword Tree. *Acta Electron. Sin.*, **38**, 1746–1750.