Supplementary Figure 1 Distribution of five biological factors. A: Family history; B: TMN stage; C: Tumor length; D: Tumor location; E: Urban/rural in two patient groups. χ^2 test was used for these factors.

Supplementary Figure 2 Variant allele frequency, mutually exclusive and co-occurring. A and B: Distribution of variant allele frequency (VAF) for TOP 20 genes; C and D: Mutually exclusive and co-occurring gene pairs in the ESCC are displayed as a triangular matrix. Green indicates co-occurring tendencies, while yellow represents mutual exclusivity trends.

Supplementary Figure 3 The rainfall plot (rainfall/raindrop plot) for individual ESCC patients. A: All Han patients; B: All Kazakh patients. Each data point is color-coded based on the category of SNV. The horizontal axis represents the chromosome number. Ultra-highly mutated genomic segments identified via the change-point method are highlighted with black arrows. By calculating the distance between adjacent SNVs, these regions can be identified to reveal the clustering of mutations. The core statistical method of the tumor genomic rainfall plot is the change-point detection algorithm, which is used to identify local hypermutated regions in the genome. It is supplemented by logarithmic transformation of genomic distance and visualization techniques for spatial distribution. This method is implemented by bioinformatics tools (rainfallPlot and maftools), and its theoretical basis belongs to non-parametric statistical modeling of mutation patterns in genomic data.
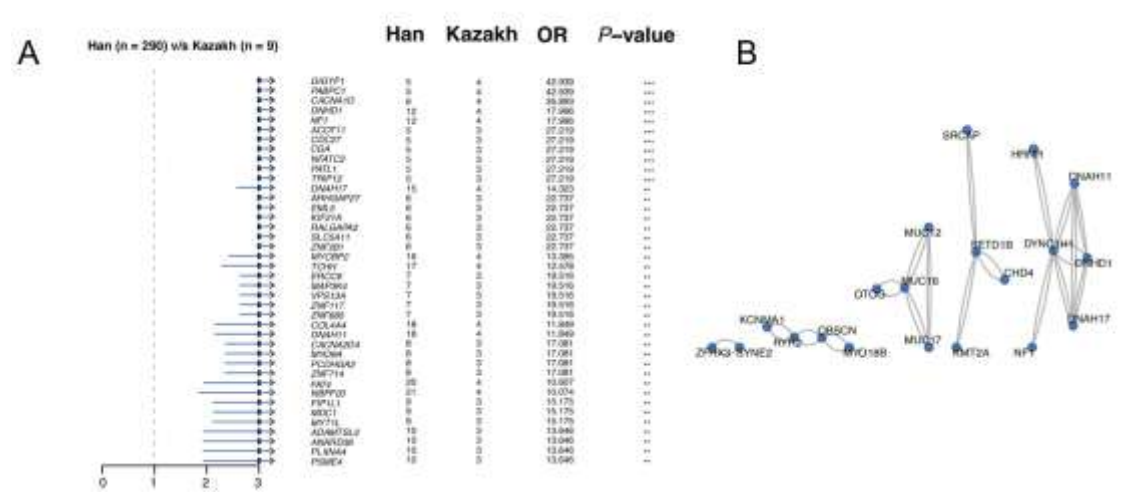
Supplementary Figure 4 Difference gene between the two ethnics groups. (A) Difference gene between the two ethnics groups (fisher's exact test, all patients,

P < 0.005). MafCompare was used to compare the two cohort difference genes. (B) PPI diagram of difference and significant mutation genes(P < 0.05, and the mutation frequency ⩾ 5%). The STRINGdb is used for the analysis of protein interactions. At the same time, igraph and ggraph are employed to visualize the interaction network.
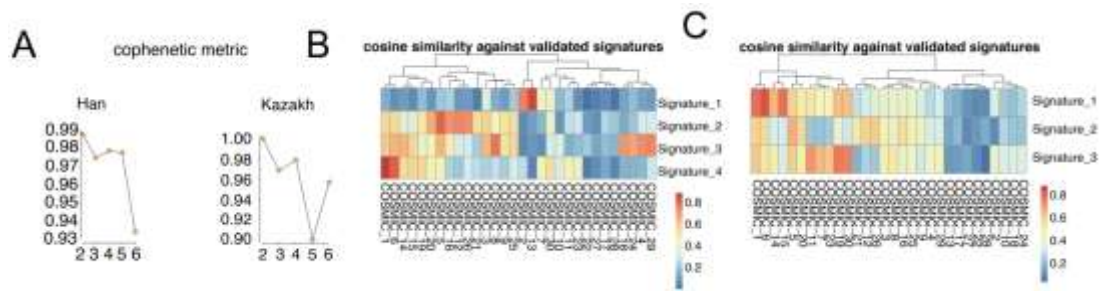
Supplementary Figure 5 Comparison of similarities of detected signatures against validated signatures. A: Best possible value is the one at which the correlation value on the y-axis drops significantly. In Han patients it appears to be at n = 5(signature 2 and 3 are same as SBS5, hence we select n = 4, and same to Kazakh group), and in Kazakh patients it appears to be at n = 4. B and C: Comparison of similarities of detected signatures against validated signatures (B for Han patients, and C for Kazakh patients).