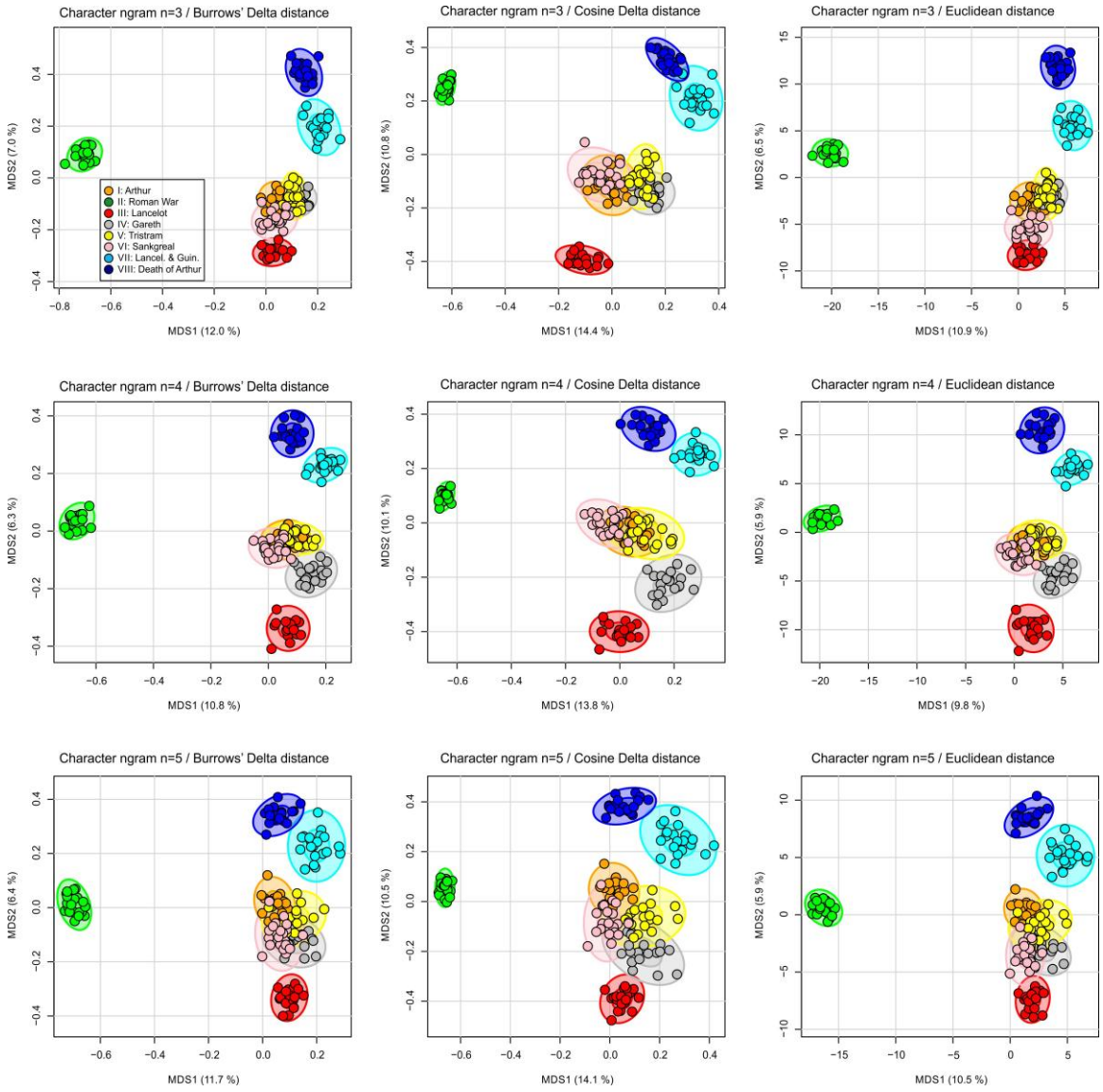

Supplementary Figure S2: The first two MDS dimensions of three different types of tokenization (character ngrams of length 3, 4, and 5) that were combined with three different types of distance functions. Colours and symbols have the same meaning as in Fig. 1.