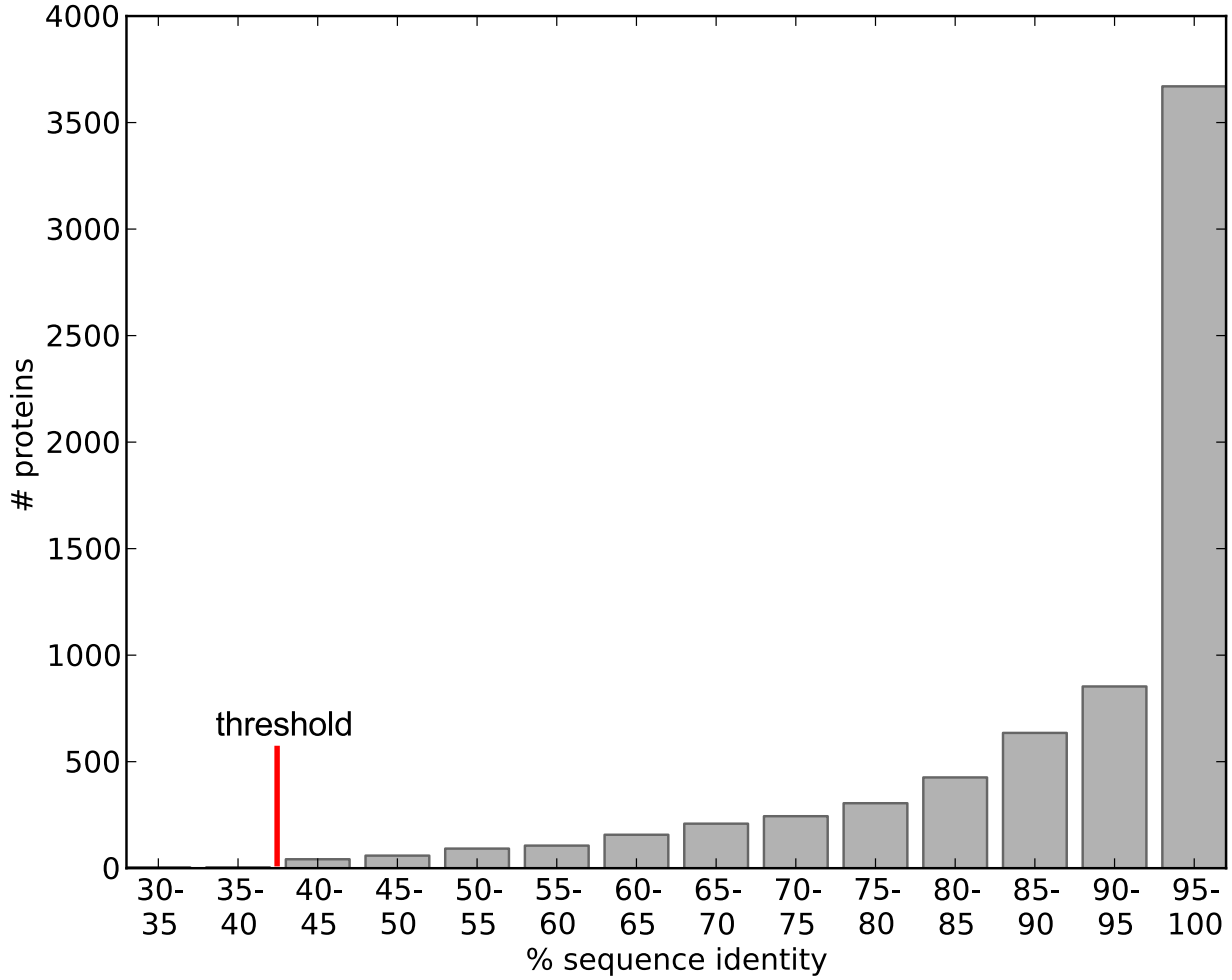

Figure S1. Sequence identity to best hit within same subfamily. Histogram of sequence identity of all sequences in our reference database to their respective best hit within the same subfamily (itself excluded). Subfamilies can contain sequences from organisms anywhere in the eukaryotic tree. The threshold is the minimal required identity for a sequence to be attributed to the subfamily of its best hit (see **Figure 1**). It is chosen to minimise the number of times a sequence is annotated as belonging to the unspecified subfamily RabX although it is a member of the same subfamily as its best hit.