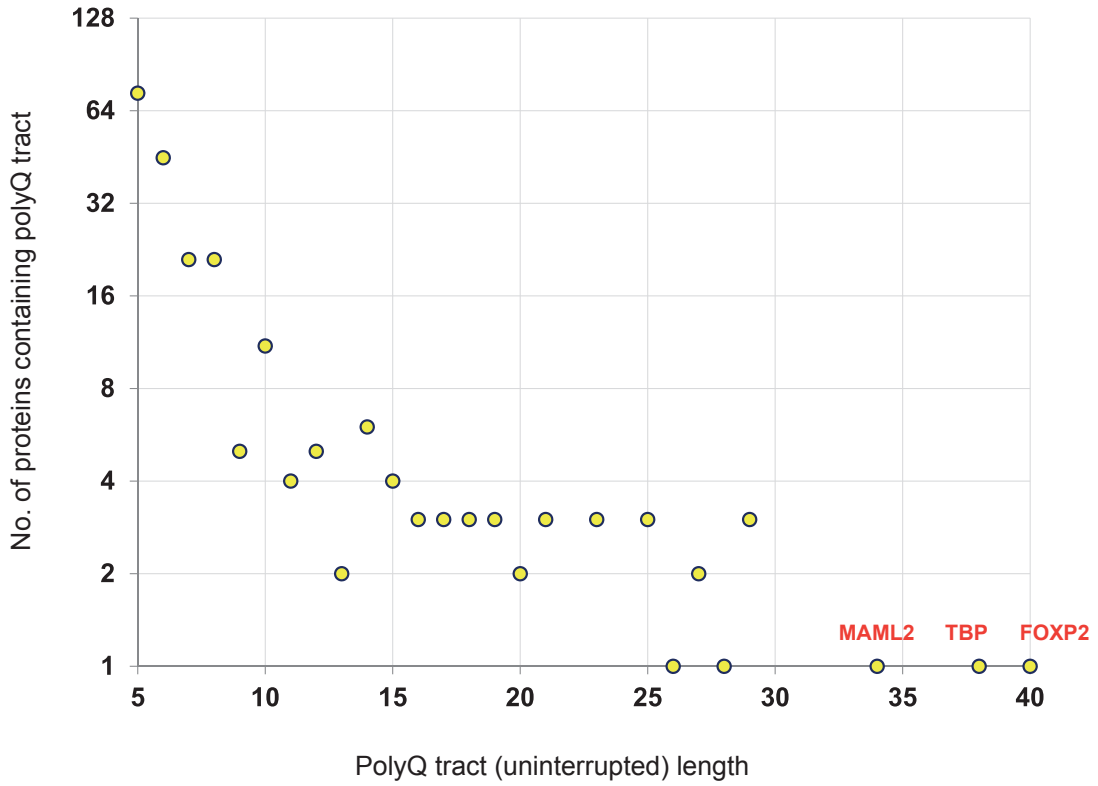
Trends in polyQ tracts across the human proteome in normal and disease biology

Figure 1. Survey of polyglutamine lengths across the human reference proteome. To determine the range and distribution of polyQ tracts across the human proteome, an in-house Java program was used to count the longest uninterrupted polyglutamine stretch in each protein sequence in the EBI release 2013_04 of the *Homo sapiens* reference proteome. Of the 20,249 proteins contained in the release, 230 proteins possessed a polyQ stretch of 5 or longer, with 40 glutamines being the longest length. Please note the logarithmic scale of the y-axis.

FOXP2 (Forkhead box protein P2), TBP (TATA-box-binding protein) and MAML2 (Mastermind-like protein 2) contained 40, 38 and 34 glutamines, respectively, which represent the top three proteins in the reference proteome in terms of polyQ length. The presence of such a long glutamine tract in FOXP2 is intriguing, given the protein's putative involvement in certain facets of the motor externalization of language.

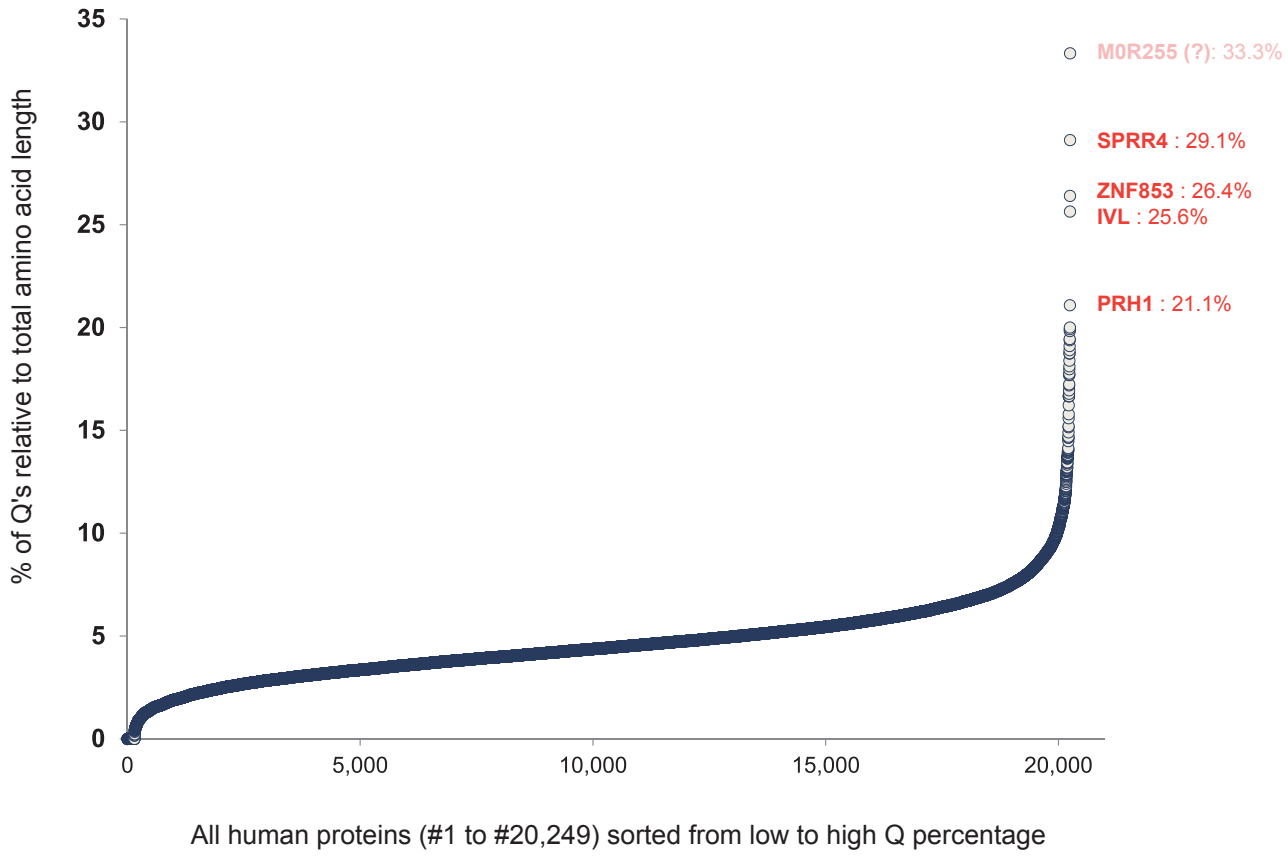


Trends in polyQ tracts across the human proteome in normal and disease biology

Figure 2. Background Q levels. To determine the distribution of glutamine residues across human proteins, the percentage of Q amino acids relative to the total length of each protein in the reference human proteome was calculated and is shown sorted from low to high values. The values range from **0% to 29.1%** (a value of 33.3% was obtained for 'M0R255'; however, this 12-AA protein does not appear to be a bona fide protein). The **average background Q level across all proteins was 4.6%**.

Outlier proteins containing a Q background level of more than 20% were Small Proline-Rich Protein 4 (SPRR4; 29.1%), Zinc Finger Protein 853 (ZNF853; 26.4%), Involucrin (IVL; 25.6%) and Salivary Acidic Proline-Rich Phosphoprotein 1 (PRH1; 21.1%).

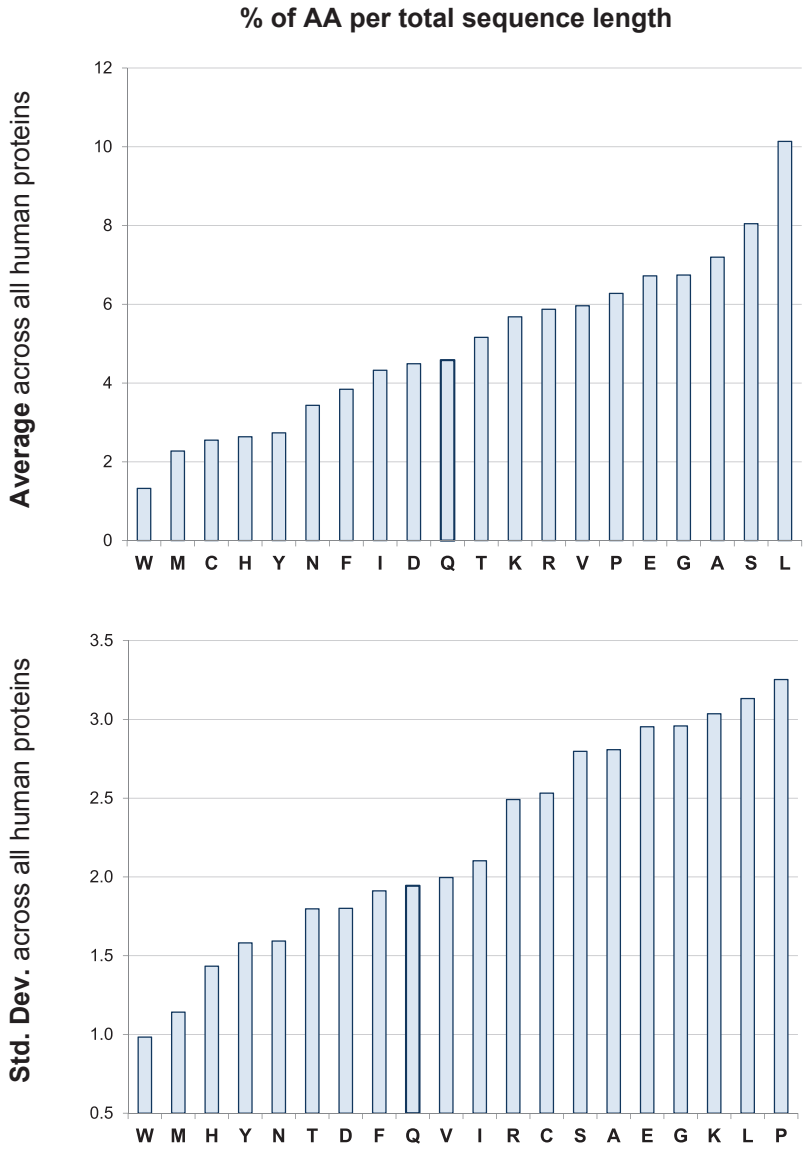
Interestingly, the three highest polyQ-containing proteins identified in **Figure 1** all contain a Q background level around **17%**: FOXP2 (17.8%), TBP (17.7%) and MAML2 (16.8%).



Trends in polyQ tracts across the human proteome in normal and disease biology

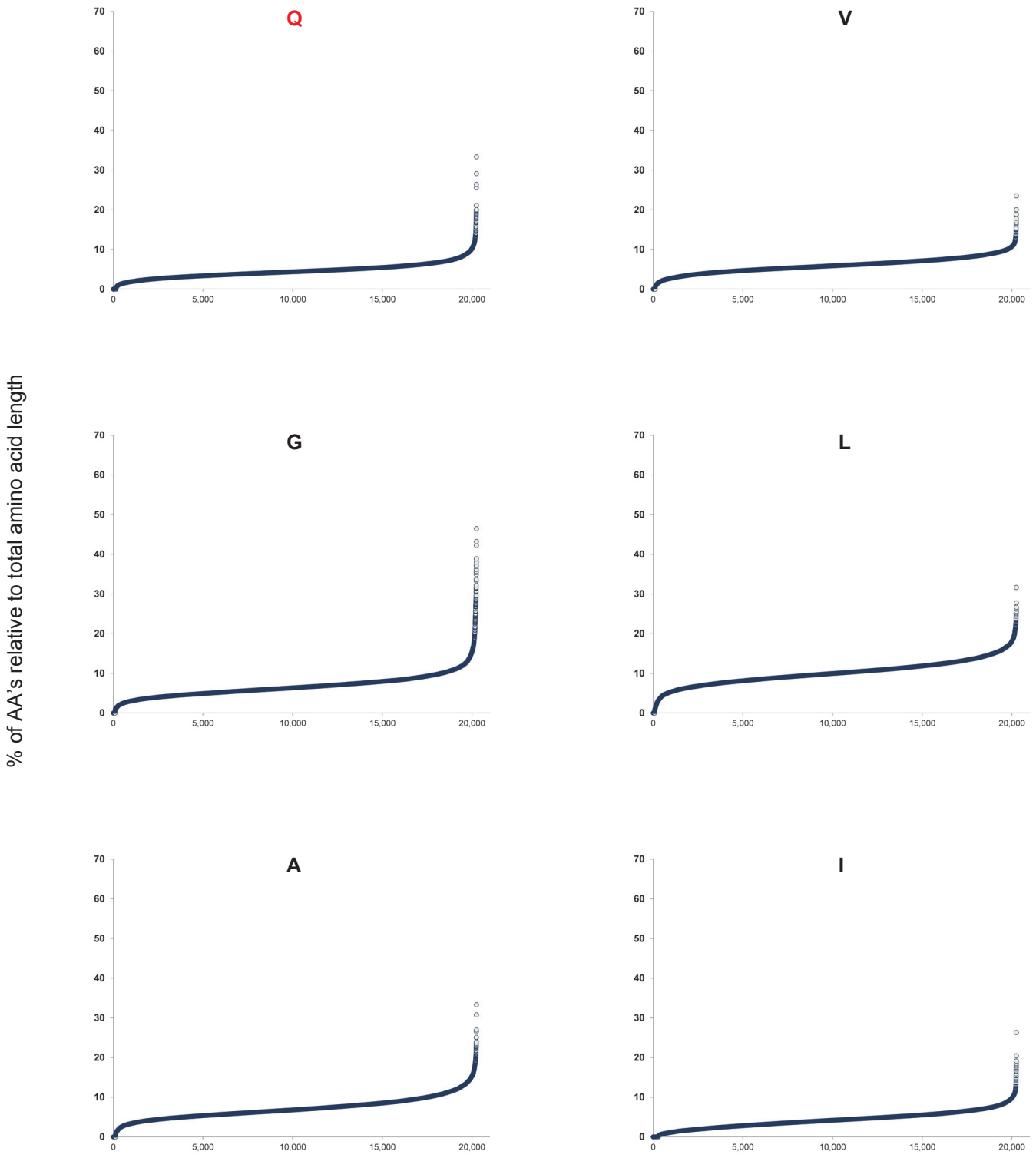
Figure 3. Comparison of Q background levels across human proteins to the other nineteen amino acids. The background levels of all amino acids are presented in panel **B**. The distribution of Q has been repeated on each page for comparison purposes. All amino acid distribution shapes appear similar to each other. The average and standard deviation of the distribution of each amino acid is summarized in panel **A**. Glutamine does not stand out in either of these measures.

A

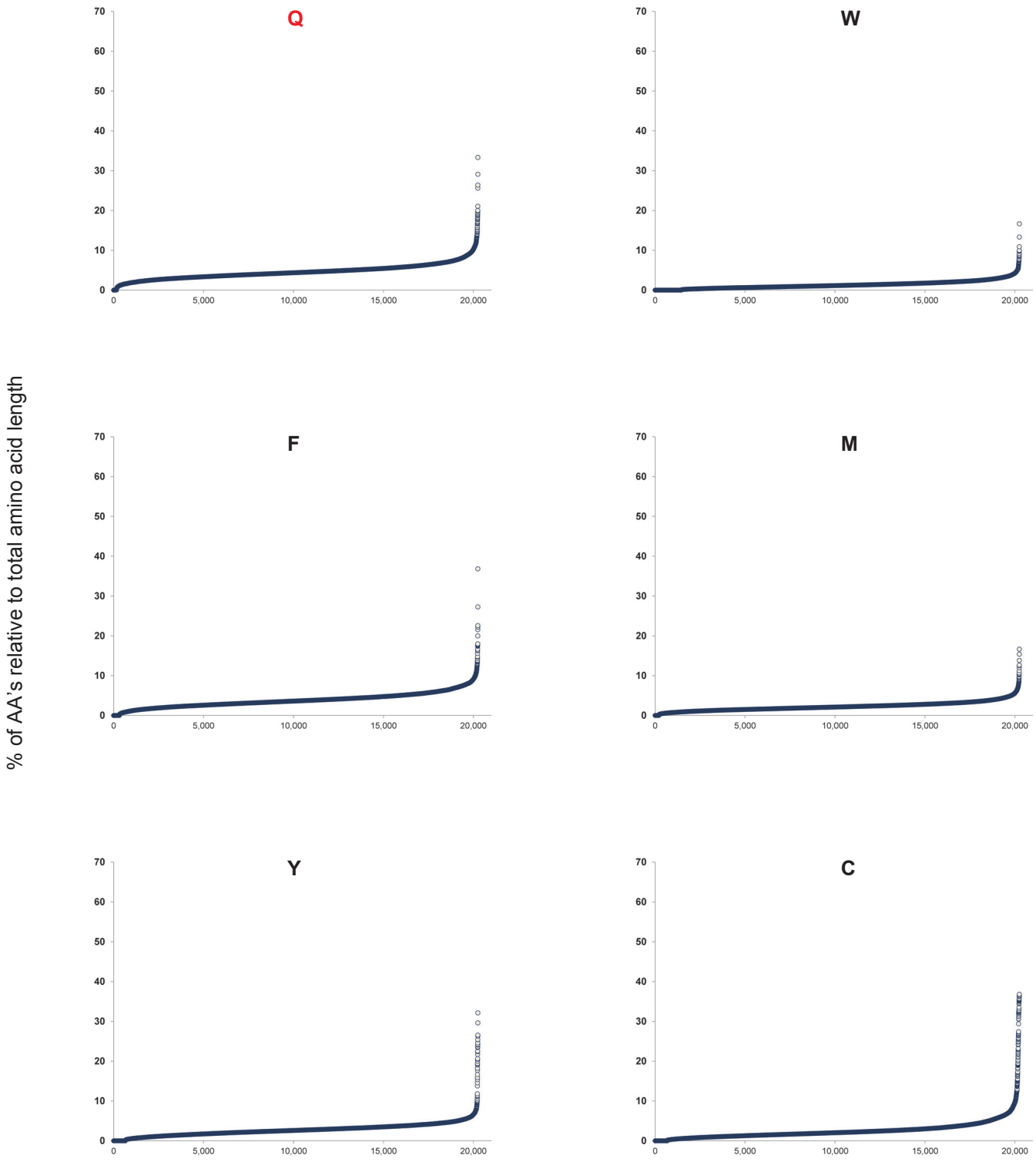


Trends in polyQ tracts across the human proteome in normal and disease biology

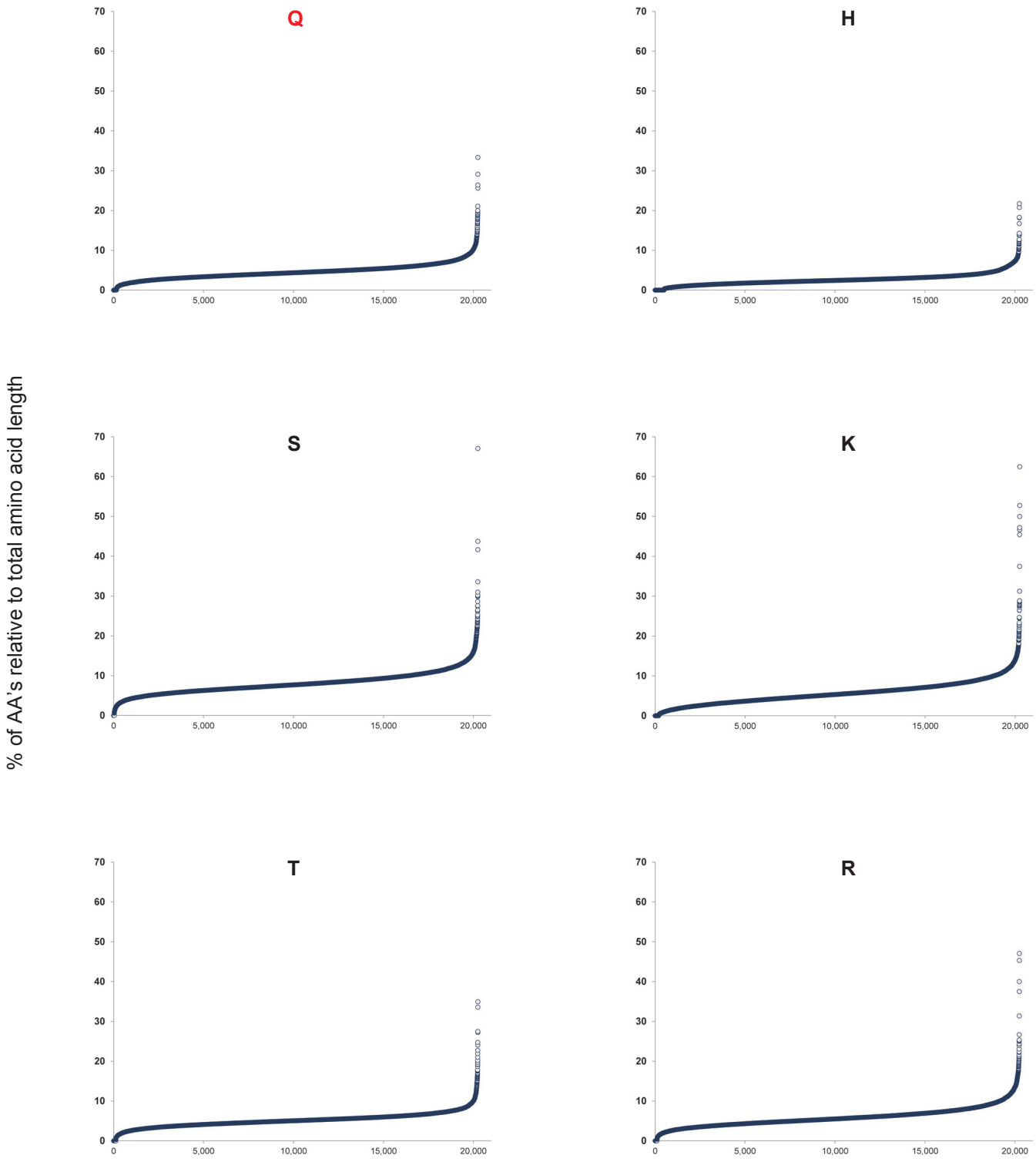
B



Trends in polyQ tracts across the human proteome in normal and disease biology



Trends in polyQ tracts across the human proteome in normal and disease biology



Trends in polyQ tracts across the human proteome in normal and disease biology