studies (9, 10). Genes were considered to have evidence of protein presence from the Human Protein Atlas only if the antibody reliability was annotated as “medium” or “high”, from Uniprot if the protein had the “Evidence at protein level” attribute and from proteogenomics if either of the two studies mapped at least two non-discriminating peptides to the protein. The remaining genes, lacking protein evidence from any source, was divided into genes with RNA-evidence, i.e. genes with “Evidence at transcript level” according to UniProt or an FPKM > 1 in at least one of our 32 tissues (RNA evidence), and genes lacking any evidence (No evidence).

Prediction methods for membrane protein topology

The membrane protein analysis was performed as previously described (46) using seven methods for membrane protein topology based on different underlying algorithms, such as hidden Markov models (HMMs), neural networks and support vector machines (SVMs): MEMSAT3 (47), MEMSAT-SVM (48), Phobius version 1.01 (49), SCAMPI multi-sequence-version (50), SPOCTOPUS (25), TMHMM (51) and THUMBUP (52). The resulting predicted transmembrane (TM) regions from these selected methods were used to construct a majority-decision based method (MDM) (46) to discriminate transmembrane proteins from soluble proteins and determine the potential TM positions in the protein sequence. Briefly, the proteins predicted as transmembrane by the MDM consists of all proteins where there is at least one predicted transmembrane region overlapping by four out of the seven methods.

Prediction of the human secretome

For the secretome analysis, three methods for the prediction of signal peptides (SP) were used: SignalP4.0 (23), Phobius (24) and SPOCTOPUS (25). SignalP4.0 is an update of the widely used SignalP3.0 method (53), and is focused on the prediction of signal peptides, whereas the other two algorithms combine the prediction of transmembrane (TM) regions with the prediction of SPs. All three methods, however, are trained to discriminate between the two types of protein features and are based on Neural Networks and Hidden Markov models. Stand-alone applications were acquired for each of the methods and run on a computer cluster as previously described for SPOCTOPUS and Phobius (46). For SignalP4.0, all predicted amino acid cleavage positions <11 were considered false positives and removed from the results in accordance to instructions from the authors.

Similarly to the construction of the MDM for membrane protein prediction, a majority-decision approach was used to predict the human secretome using a new method called MDSEC. All human proteins for which at least two out of the three methods predicted a signal peptide were considered potentially secreted by MDSEC. However, as certain types of membrane proteins also contain an N-terminal signal peptide, the proteins with both a predicted SP and at least one TM region predicted by the MDM were classified as membrane-spanning and removed from the set of secreted proteins. Thus, the resulting list of predicted secreted proteins consists of all protein with a predicted signal peptide by two out of three methods and without a predicted MDM region. As this method requires a predicted signal peptide, proteins secreted using the