Protein Probability Model for High-Throughput Protein Identification by Mass Spectrometry-Based Proteomics

Gorka Prieto*,† and Jesús Vázquez*,‡

†*Department of Communications Engineering, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain*

‡*Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), 28049 Madrid, Spain*

E-mail: gorka.prieto@ehu.eus; jesus.vazquez@cnic.es

Phone: +34 94 601 3994

Supporting Information

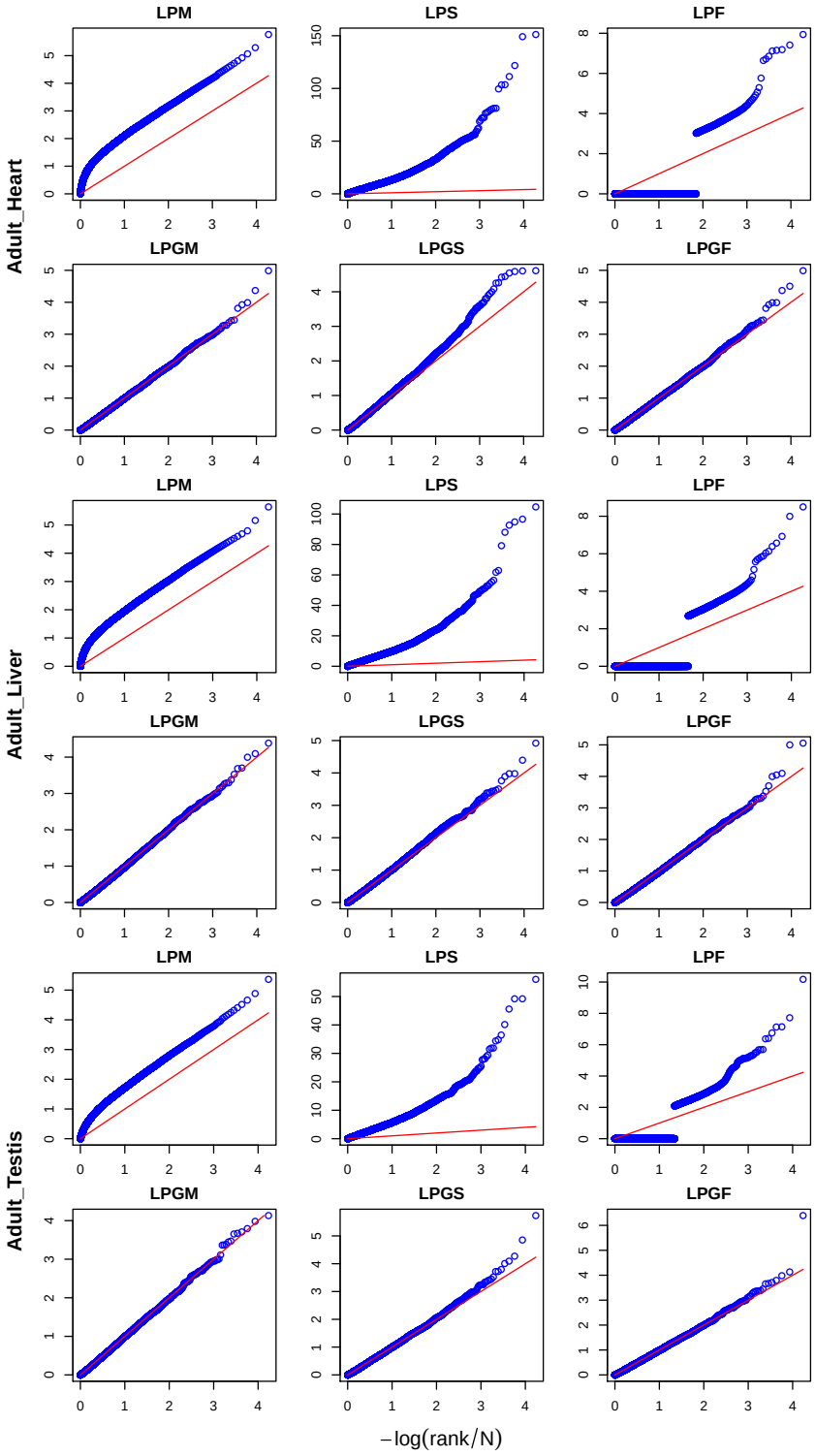

Figure S1: Distribution of different decoy protein scores derived from Comet after searching three tissues of the Human Protein Map as separated test data sets. The y-axis represents the cologarithm of protein probabilities calculated by the methods of Table 1B as indicated by the title of each graph. The x-axis represents the cologarithm of the expected uniform protein probabilities. Deviations from the identity line (drawn in red) mean that the calculated probabilities are inaccurate.

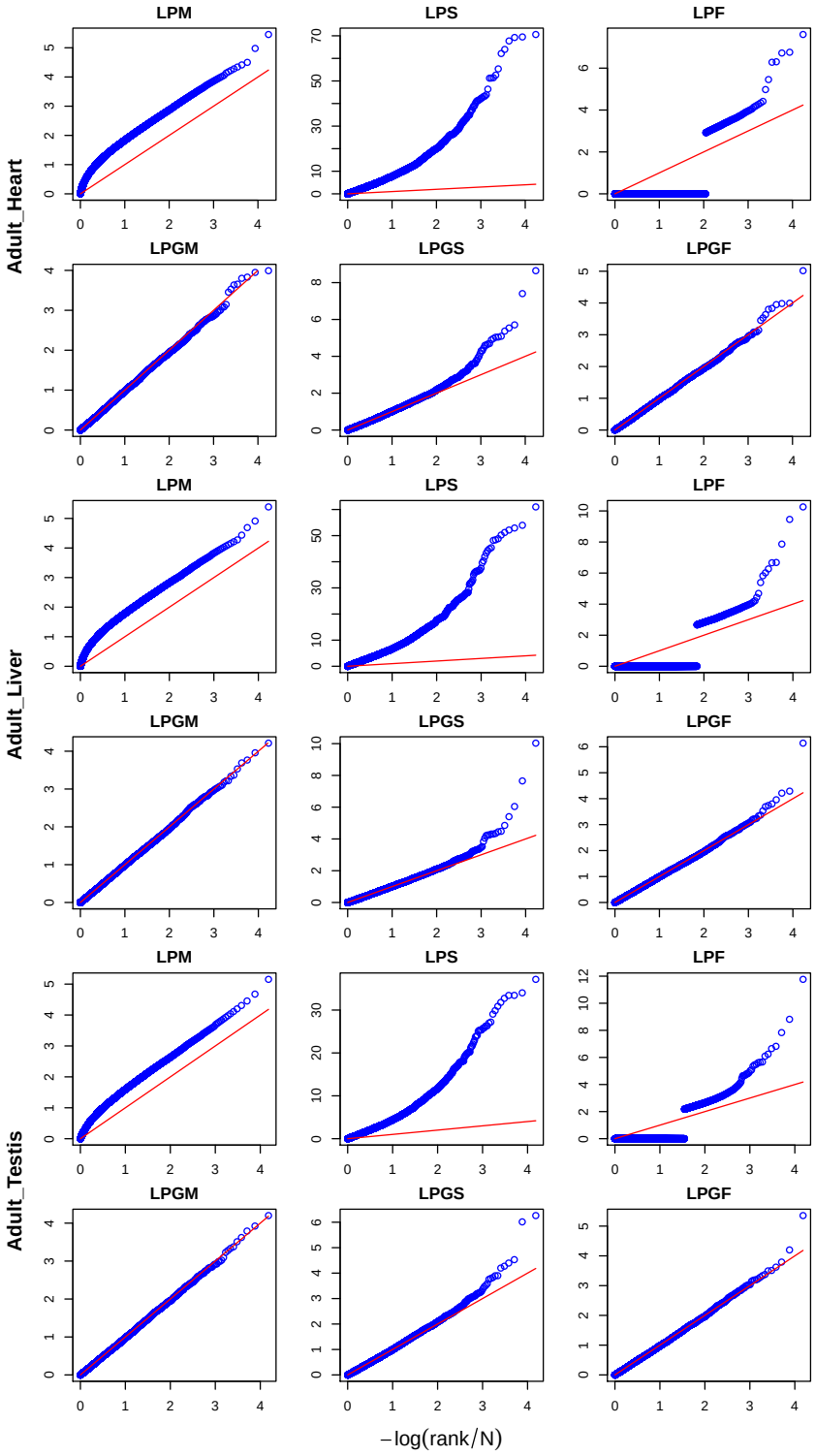

Figure S2: Distribution of different decoy protein scores derived from MSFragger after searching three tissues of the Human Protein Map as separated test data sets. The y-axis represents the cologarithm of protein probabilities calculated by the methods of Table 1B as indicated by the title of each graph. The x-axis represents the cologarithm of the expected uniform protein probabilities. Deviations from the identity line (drawn in red) mean that the calculated probabilities are inaccurate.

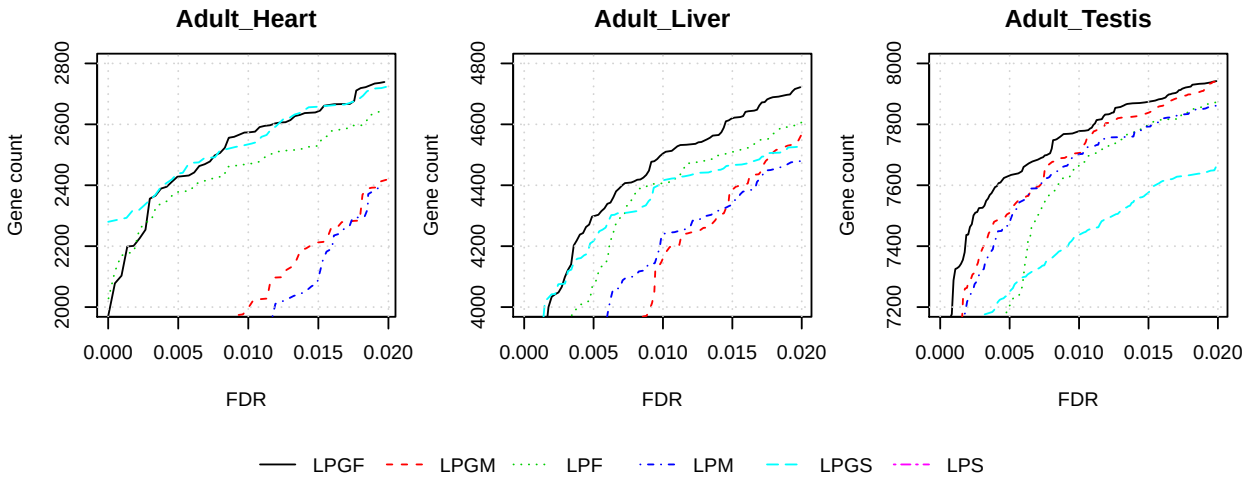

Figure S3: Number of identified genes as a function of the FDR threshold for different protein scores types derived from Comet after searching three tissues from the Human Protein Map. The number of identifications provided by LPS is so small (less than 300) that it is not depicted in the figure. For these calculations the FDR was calculated as the fraction of decoy proteins divided by the number of target proteins that pass the protein score threshold (i.e. the $FDRn$ method defined in eq 11).

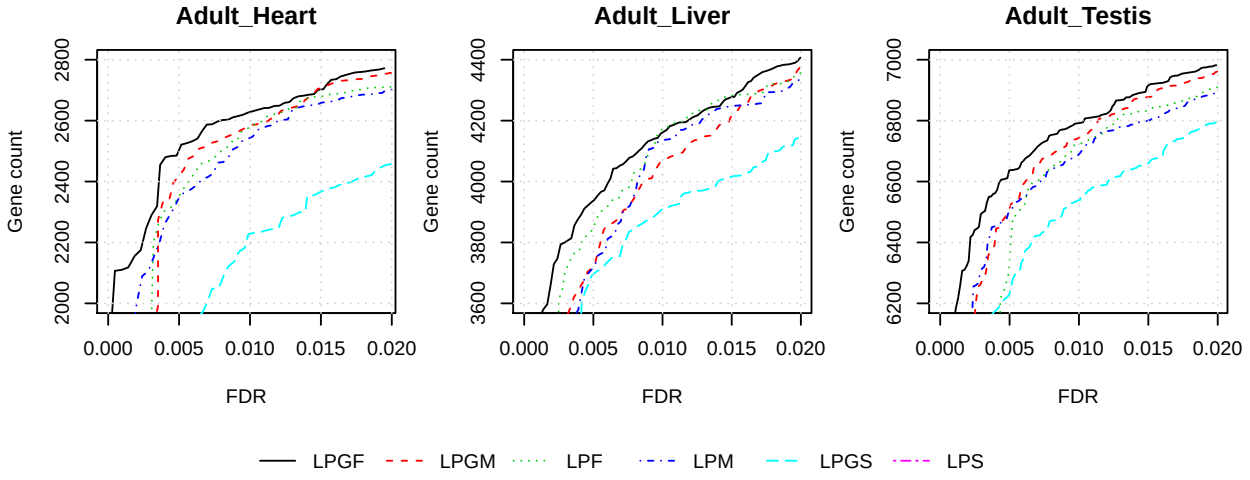

Figure S4: Number of identified genes as a function of the FDR threshold for different protein scores types derived from MSFragger after searching three tissues from the Human Protein Map. The number of identifications provided by LPS is so small (less than 300) that it is not depicted in the figure. For these calculations the FDR was calculated as the fraction of decoy proteins divided by the number of target proteins that pass the protein score threshold (i.e. the $FDRn$ method defined in eq 11).

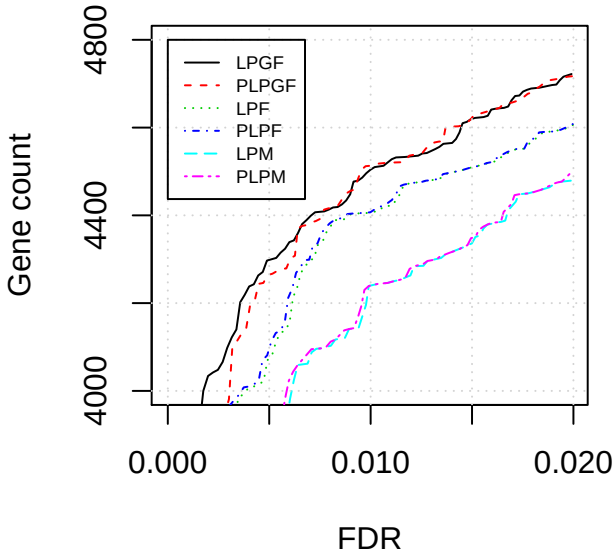

Figure S5: Comparison of the number of identified genes as a function of the FDR threshold when using parametric peptide scores. These results are derived from Comet after searching the `Adult_Liver` tissue from the Human Protein Map data set. A gamma distribution has been used for modeling the peptide probabilities from the `Xcorr` values. Gene-level scores $PLPM$, $PLPF$ and $PLPGF$ have been calculated using these parametric peptide probabilities in comparison to the corresponding LPM, LPF and $LPGF$ scores calculated using eq 1.

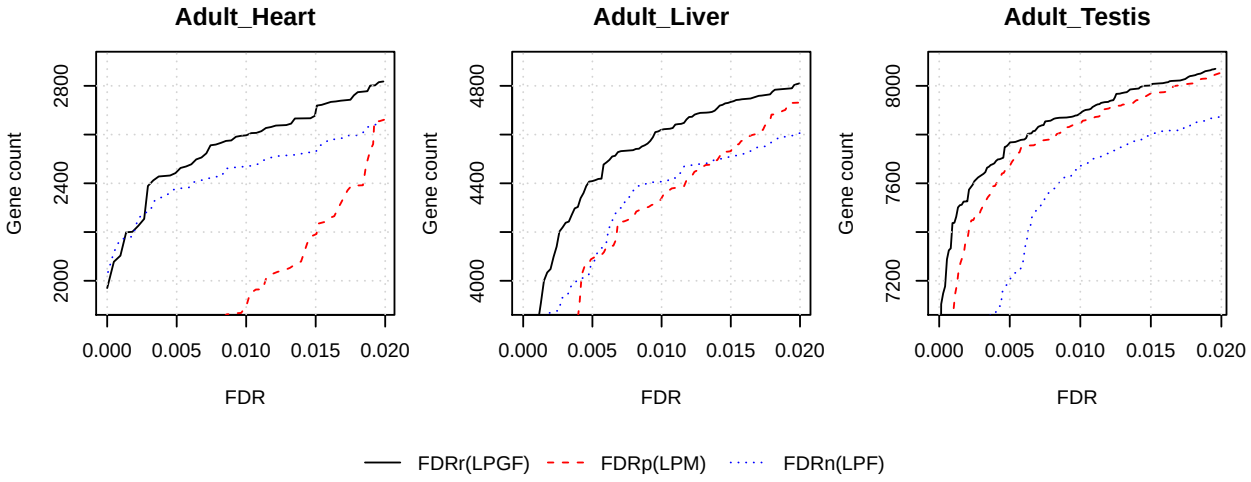

Figure S6: Number of identified genes as a function of the FDR threshold for different protein identification workflows using as separated tests three tissues from the Human Protein Map searched with Comet.

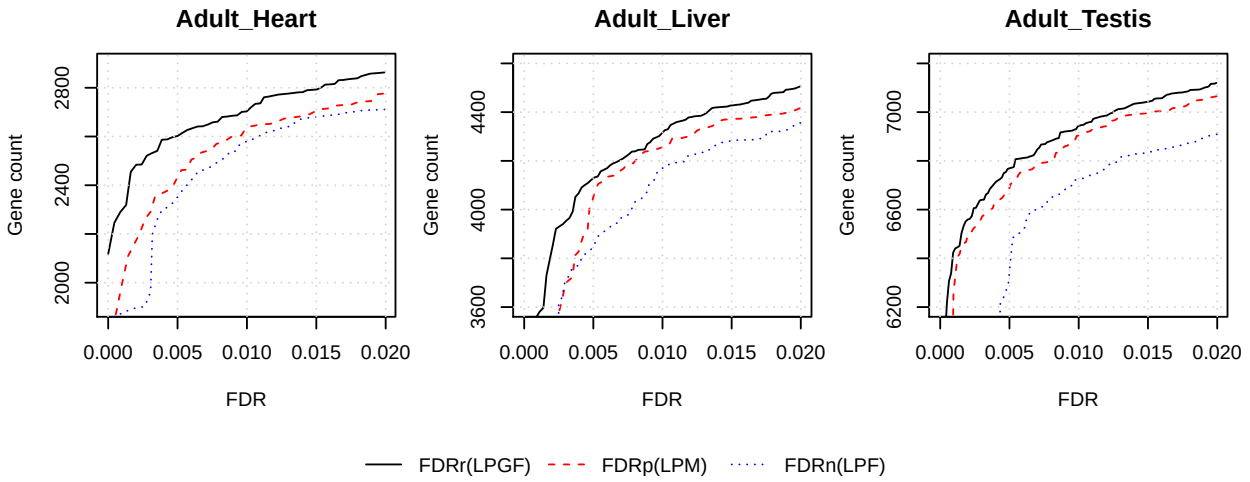

Figure S7: Number of identified genes as a function of the FDR threshold for different protein identification workflows using as separated tests three tissues from the Human Protein Map searched with MSFragger.

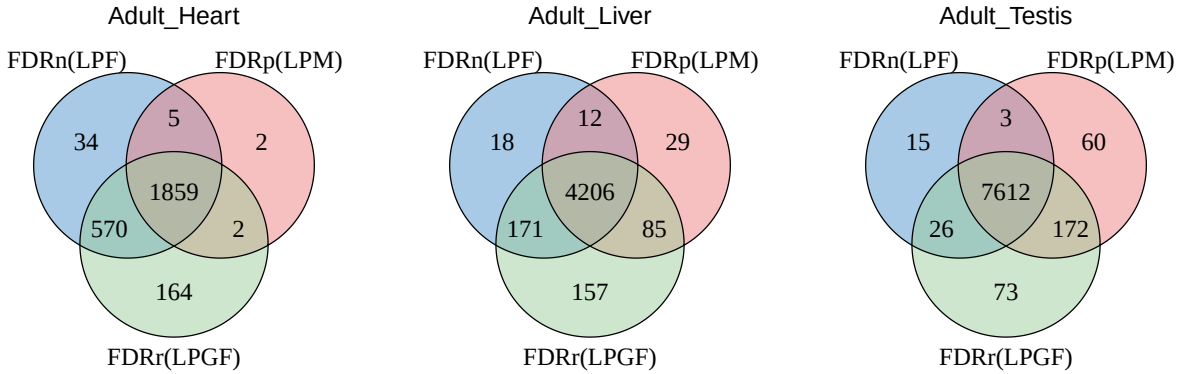

Figure S8: Venn diagrams with the number of identified genes using different protein identification workflows in three tissues of the Human Protein Map searched with Comet.

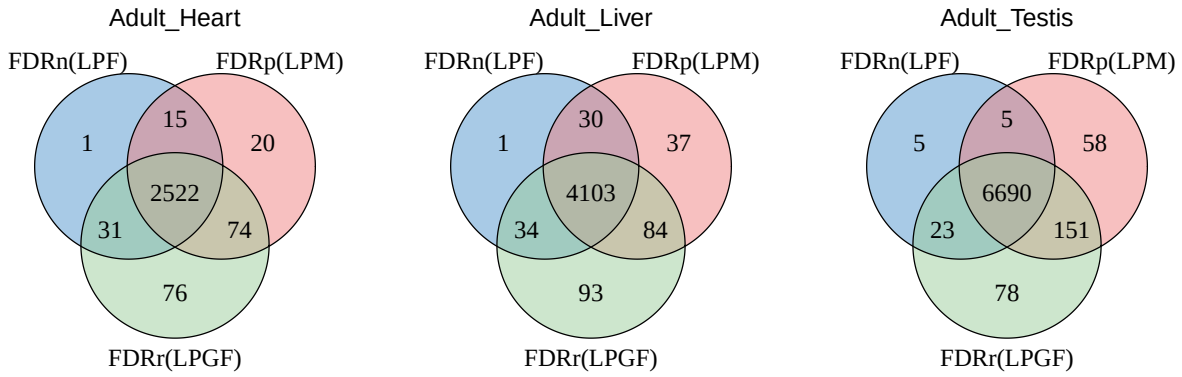

Figure S9: Venn diagrams with the number of identified genes using different protein identification workflows in three tissues of the Human Protein Map searched with MSFragger.

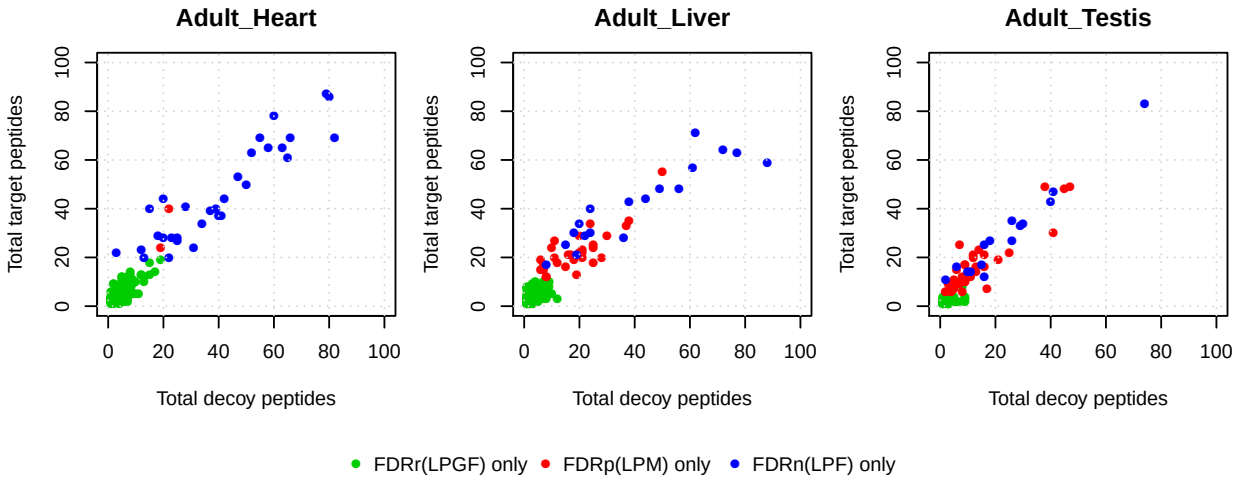

Figure S10: Comparison of target versus decoy peptides for each gene identified exclusively by any of the three different protein identification workflows discussed using Comet. The total number of peptides is considered, without filtering by FDR. Each point corresponds to a target-decoy pair. The comparison has been carried out in three tissues of the Human Protein Map

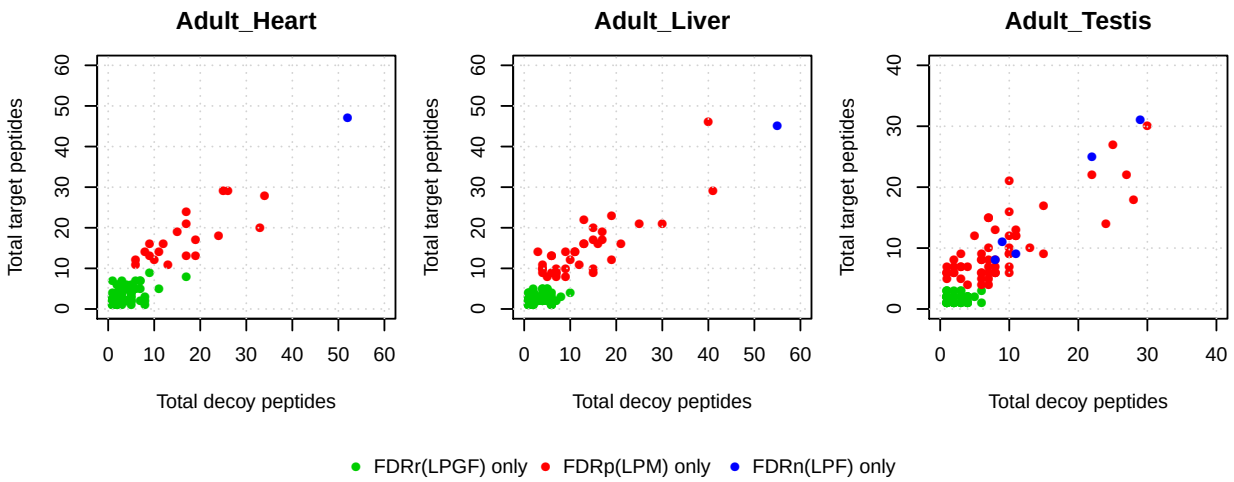

Figure S11: Comparison of target versus decoy peptides for each gene identified exclusively by any of the three different protein identification workflows discussed using MSFragger. The total number of peptides is considered, without filtering by FDR. Each point corresponds to a target-decoy pair. The comparison has been carried out in three tissues of the Human Protein Map.

Table S1: Sequences of the best three peptides for top-scoring decoy proteins using *LPGS* as the protein probability after searching with MSFragger the Adult_Heart tissue of the HPM.

Rank	Best peptide	Second peptide	Third peptide
1	GGEEGAPGPAGQPGAPGPEK	GPGLPGTAGQPGPPR	GAKGPGKDGSPGPPGPPGPNGNSGPPGR
2	GGGGGGGGWGGGGGGGGGWGGGGGGWGR-GGGGGGGGRGGGGGGGGGWGR	GGGGGGGGWGGGGGGGGGWGGGGGGWGR	GGGGGGGGRGGGGGGGGGWGRGGGGGER
3	GAGPAGVNGLPGPPGAPGAPGPPGADKGAGADGPEK	GGGKEGAPGPPGPPGANGSPGPPGRVGAAGPFG-TAGPPGASR	DGPAGASGPPR
4	GAGMLGPAGPLGAPGMEKGGGPSGNLGPPK	GPGPSGLGRGEGPMGQLGPTGAAGPPGPPGAAK	GPGSVGPSGPAGTNGAPGPPGK
5	GAGALGAHK	GAGAPGPAGRVRGEGPLR	GAGTAGAAGVPGPLR
6	FGGGFGGGGFGGGSFGSGGGFSSGGGFGGR	SSGFGGGGFGGGGGGR	YGGSGGGSGYGGGSLSGGGGSGRGSGSR
7	SSSSSSSSSSSSSSSSSSSSR	SRSSSSSSSSSSSSSSSSSSSSRR	SSSSSSSSSSSAESRTNSGSSVRSSR
8	FAAAVALVAAWR	QPASLLQEPLLSTSGDQLGK	LEVEVELELTEEDAEVDLQQMLEEK
9	GDKGLGTPGPDGPGK	GPGPPGQVGPLGQDK	KGGPPGPAGPEGSAGPPGPGGVDGPDK
10	GPGPPGGEGPLGPEGAEGPEK	GPGPPGAAGSPGSEK	GPGSDKGLGPLGPPGMPGPPGDPGPSGPLGQEGVPGPLR
11	KTEVSLAR	EAQFLFTLALTTVTFDWLR	GAGPSGQTGPLRGDK
12	GPRGERGPGQSGPPGAPGRPR	KGGEEGPPGLPGQLGHPGPEK	GAGPTGDTGLAGPLGPSGVPGPPGTERGQGQPGEPR
13	YSGQVPVGGPGTVNAAEAGQHVFVDK	YSGQVPVGGPGTVNAAEAK	EAGGQEAEPASSNEAPPDAKATK
14	GAEPAQGADAGEAEVQNKGGK	GAEPAQGADAGEAEVQNKGGKK	GAEPAQGADAGEAEVQNK
15	EGNLAEVLALAFEFSTGPR	KDLVLPK	KDLVLPKAVQAAVK
16	AGAAGAVGGPGYGPK	AAAAAAAAAAAPTGVGALGGLGPLAGPVGPVGA-GGVGPLVGAAGAGK	AAAAAAAAPGAGVVGGPLGVGGLAGLGGLVGPVAAGK
17	GLGPWGPARGLGRPSSGPSGPVGPPGSPGQPR	GPGDLGENGPTGPLGPLGPLGLEGHEGSNRGEGP-FGPTGPGGPMAVAPEK	GTGPFGQDKGPGPPGPFGLGPLGPEGDPGPLGPDGK
18	SESSSSSSDSYSSR	SSSSGSESSSSSSSSRSESSSSSSDSYSSR	SSSSGSESSSSSSSSR