Supplemental material: Towards the Complete Characterization of Host Cell Proteins in Biotherapeutics via Affinity Depletions, LC-MS/MS, and Multivariate Analysis

James A. Madsen Victor Farutin Theresa Carbeau Steve Wudyka
Yan Yin Stephen Smith James Anderson Ishan Capila

Sample	Antibody	CellLine	Process	Lot	Comment	Batch
111.2	1	1	1	2		B
111.3	1	1	1	3		A
111.3v	1	1	1	3	Orbitrap Velos Pro	A
111.4	1	1	1	4		A
111.5	1	1	1	5		A
111.5c	1	1	1	5	different depletion	E
111.5r	1	1	1	5	different nano column	C
112.1	1	1	2	1		D
112.3	1	1	2	3		A
112.4L	1	1	2	4	Protein L depletion	F
112.5	1	1	2	5		A
112.7	1	1	2	7		D
134.1	1	3	4	1		B
135.1	1	3	5	1		B
136.1	1	3	6	1		D
137.1	1	3	7	1		D
138.1	1	3	8	1		E
223.1	2	2	3	1	Protein L depletion	G

Table S1: A description of each sample analyzed in this study and its associated nomenclature. The Antibody and Cell Line columns represent the mAb therapeutic present in the sample and the type of CHO cell line used to produce the mAb, respectively. The Process column represents different upstream process conditions and various combinations of protein purification steps (including anion/cation exchange chromatography, hydrophobic interaction chromatography, Protein A, etc.) used to produce final drug product. The Lot column represents arbitrary lot numbers that indicate different lots of material generated from the same combination of antibody, cell line and process conditions. The Batch column indicates groups of samples measured together. Unless explicitly stated, all depletions used Protein A for HCP enrichment, and all measurements were obtained using a Q Exactive mass spectrometer.

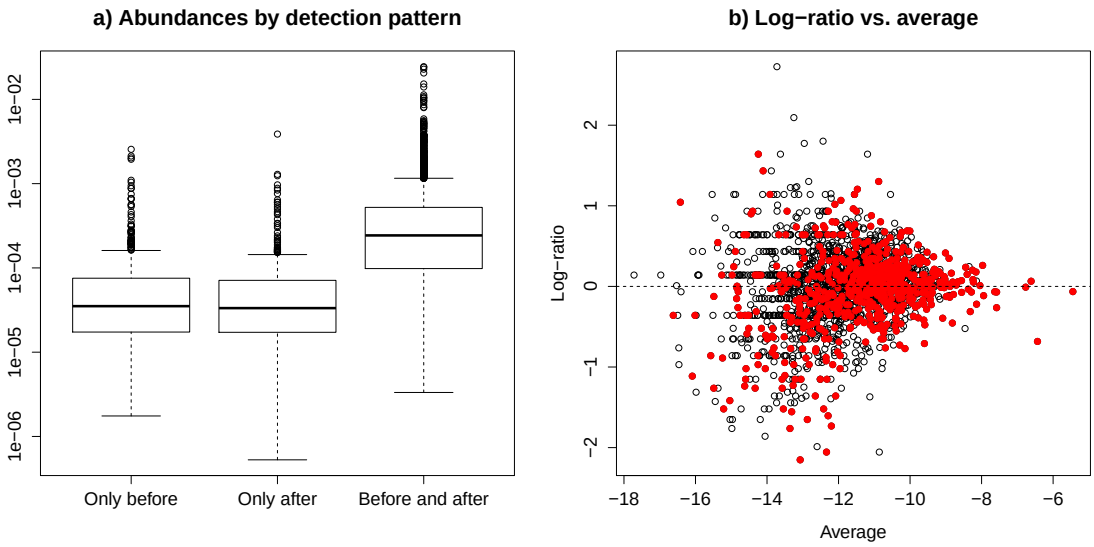

Figure S1: NSAF protein abundances for CHO cell lysates before and after protein A column: a) boxplots of abundances by consistency of their detection, b) plot of log-ratio of protein abundances (averaged over technical replicates) before and after Protein A column versus average abundance before and after protein A column (red color indicates HCPs that were detected in commercial grade drug products in this study).

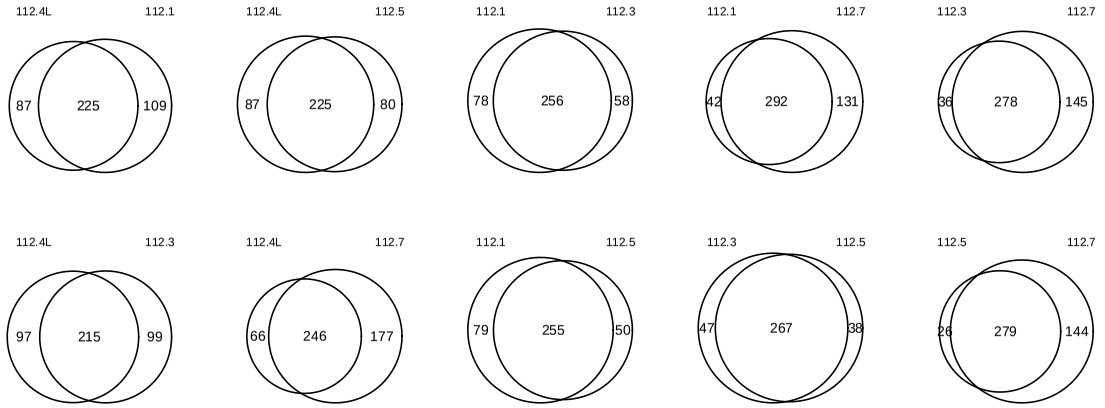

Figure S2: Overlaps between HCP identities for each pair of HCP profiles for commercial grade monoclonal antibody product lots produced using the same cell line and process (four of them were analyzed using Protein A depletion and one – 112.4L – using Protein L depletion). Left two columns represent comparisons of the HCP profile obtained using Protein L depletion to those obtained using Protein A depletion. The remaining columns represent comparisons of HCP profiles obtained using Protein A depletion among themselves. Areas are approximately proportional to the counts shown in the figures. Corresponding Jaccard distances can be found in Table S2.

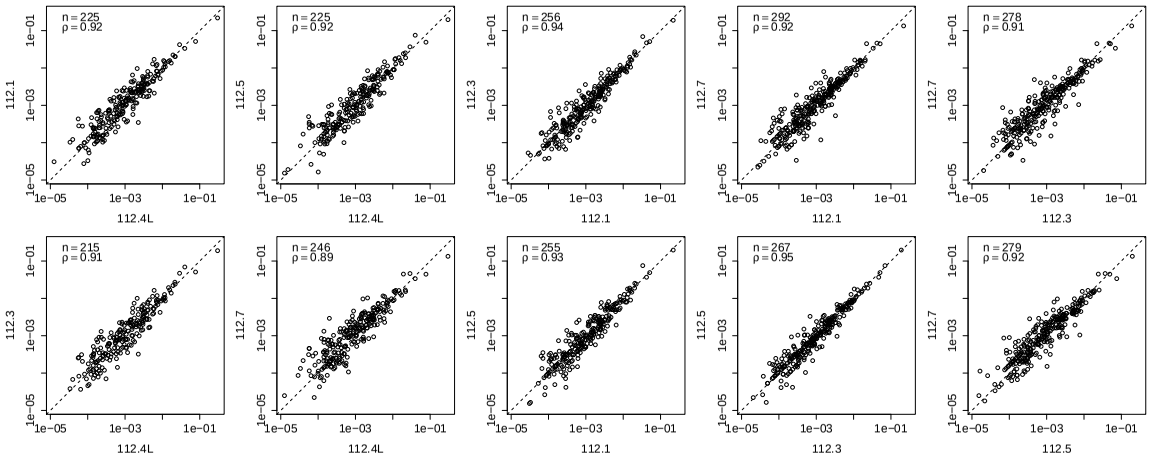

Figure S3: Scatterplots of NSAF abundances for each pair of HCP profiles for commercial grade monoclonal antibody product lots produced using the same cell line and process (four of them were analyzed using Protein A depletion and one – 112.4L – using Protein L depletion). Left two columns represent comparisons of the HCP profile obtained using Protein L depletion to those obtained using Protein A depletion. The remaining columns represent comparisons of HCP profiles obtained using Protein A depletion among themselves. The same (logarithmic) scale on horizontal and vertical axes are applied. Spearman correlation coefficients and numbers of HCPs detected in each pair of HCP profiles are shown in the legend of each figure. Dashed lines represent $x = y$ diagonal. All correlations shown are extremely unlikely to be encountered by chance: as a reference, statistical significance of observing $\rho = 0.85$ on $n = 200$ observations is estimated to be below 10^{-56} and all counts and correlations presented here are larger.

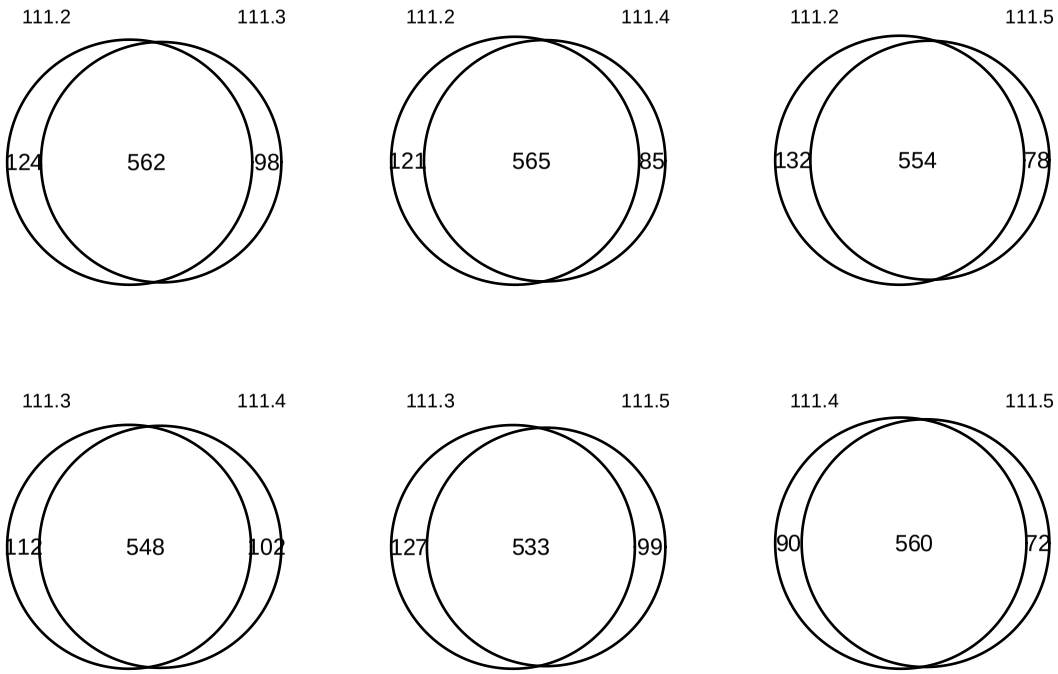

Figure S4: Overlaps between HCP identities for each pair of analyzed commercial grade drug product lots for the same monoclonal antibody produced using the same cell line and process. Areas are approximately proportional to the counts shown in the figures. Corresponding Jaccard distances can be found in Table S2.

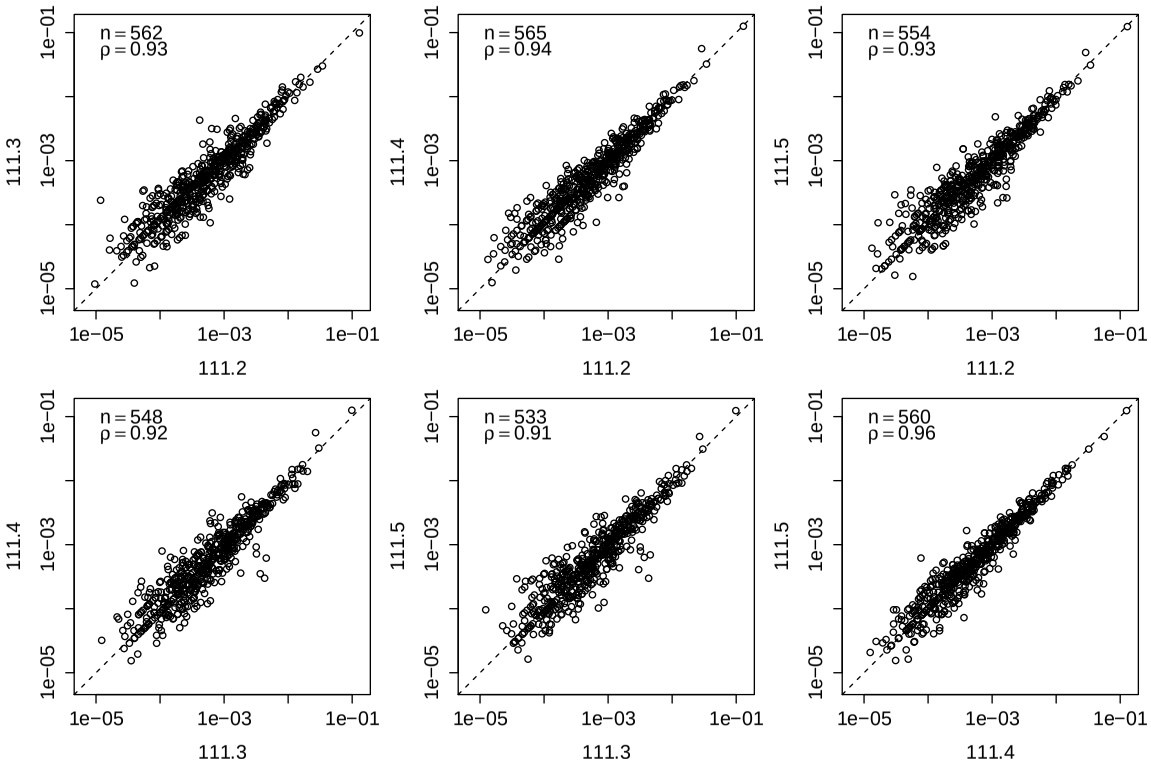

Figure S5: HCP NSAF abundances for each pair of analyzed commercial grade drug product lots for the same monoclonal antibody produced using the same cell line and process. The same (logarithmic) scale on horizontal and vertical axes are applied. Spearman correlation coefficients and numbers of HCPs detected in each pair of HCP profiles are shown in the legend of each figure. Dashed lines represent $x = y$ diagonal. All correlations shown are extremely unlikely to be encountered by chance with corresponding p-values below 10^{-100} (as a reference, statistical significance of observing $\rho = 0.85$ on $n = 450$ datapoints is estimated to be several orders of magnitude below 10^{-100} and all counts and correlations presented here are much larger).

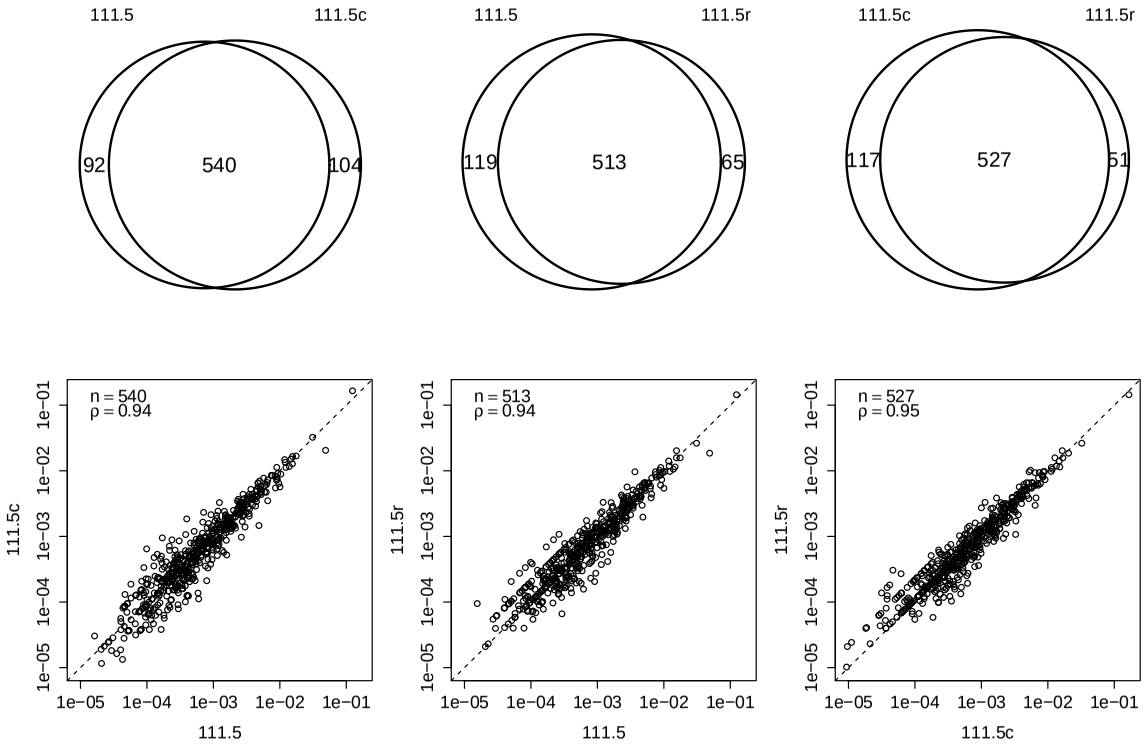

Figure S6: Overlaps of HCP identities (top row) and scatterplots of HCP NSAF abundances (bottom row) for each pair of analyses of commercial grade monoclonal antibody product lot (111.5) performed at different times: 111.5r indicates data obtained for the same protein A depletion as lot 111.5, but ran 20 days later on a different EASY spray nano column, 111.5c indicates a different protein A depletion column. The same (logarithmic) scale on horizontal and vertical axes are applied. Spearman correlation coefficients and numbers of HCPs detected in each pair of HCP profiles are shown in the legend of each figure. Dashed lines represent $x = y$ diagonal. All correlations are highly statistically significant with corresponding p-values below 10^{-200} (for reference, statistical significance of observing $\rho = 0.92$ on $n = 510$ datapoints is estimated to be below 10^{-200} and all counts and correlations presented here are larger).

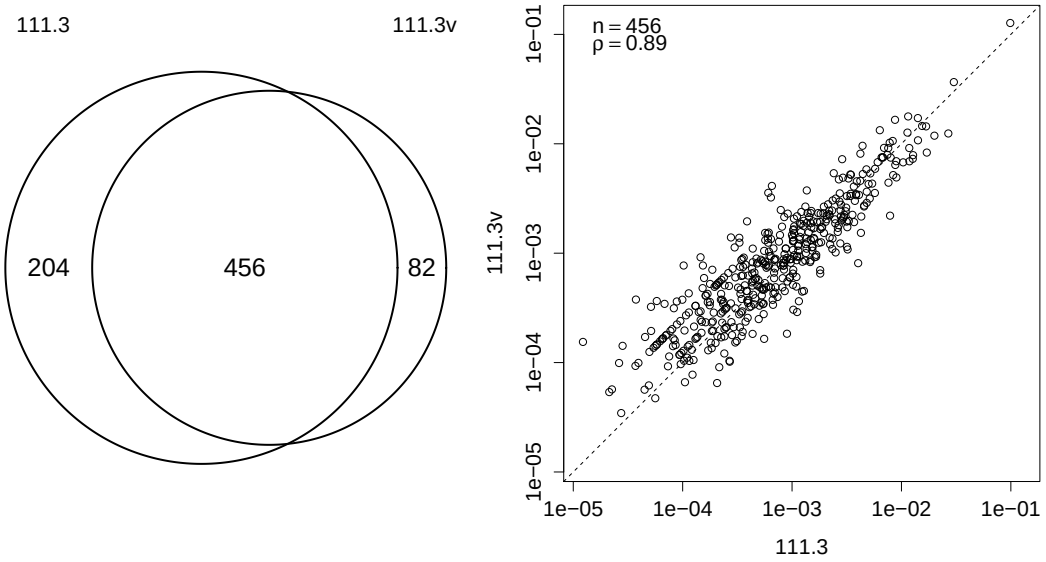

Figure S7: Overlap between protein identities and scatterplots of NSAF abundances for the HCP profiles obtained using Orbitrap Velos Pro and Q Exactive mass spectrometers. The resulting correlation between protein abundances is highly unlikely to be observed by random chance ($p < 10^{-159}$).

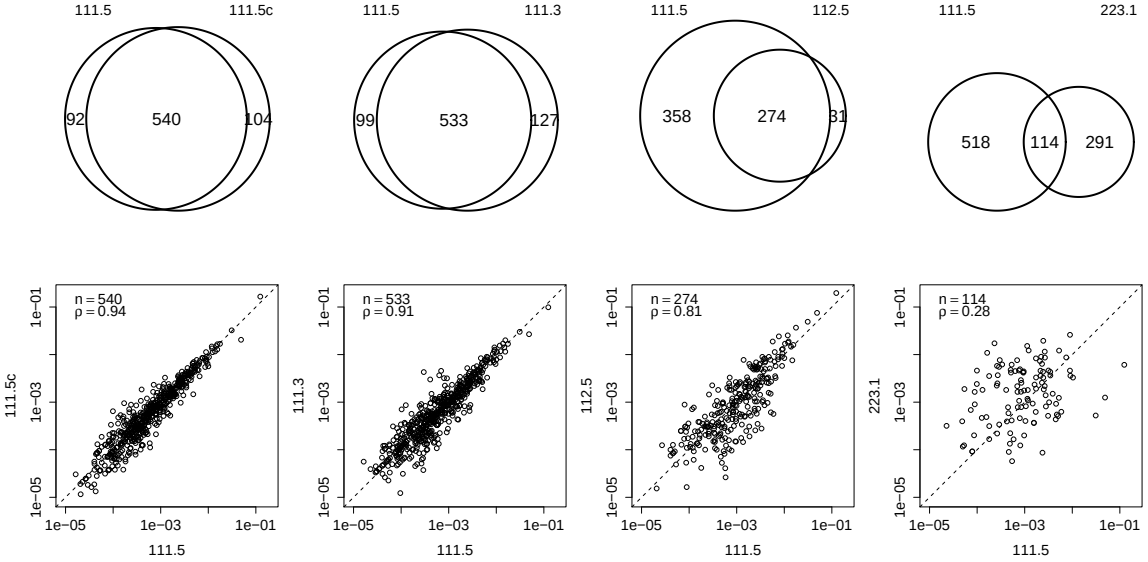

Figure S8: Overlaps of HCP identities and scatterplots of protein abundances for a selected set of drug product lots representing different analytical runs (111.5 and 111.5c), different lots of the same drug product (111.5 and 111.3), variants of manufacturing process (111.5 and 112.5) and entirely different monoclonal antibody (111.5 vs. 223.1).

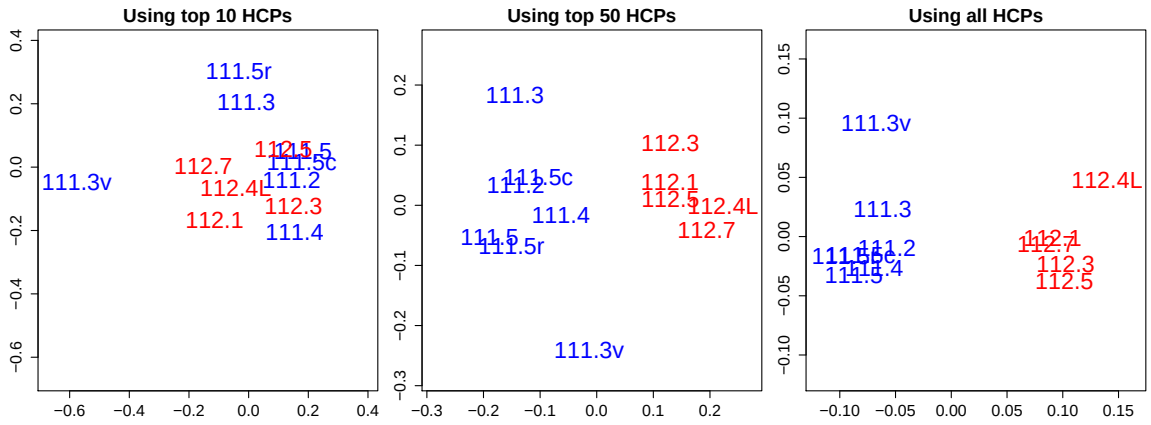

Figure S9: MDS plots of dissimilarity between HCP profiles quantified by one complement of Spearman correlation coefficient for NSAF abundances using several subsets of top most abundant HCPs in each profile. Color indicates HCP profiles obtained for the drug product lots manufactured using two different processes.

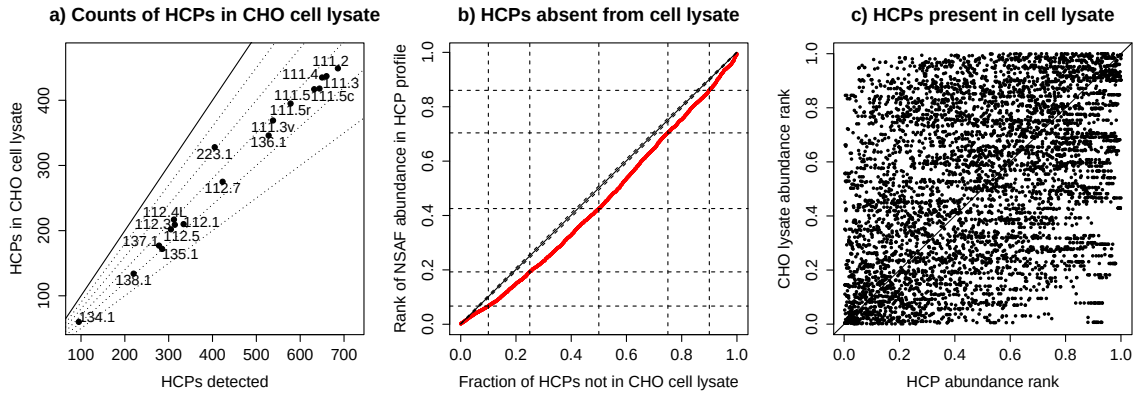

Figure S10: Comparisons of drug product HCP profiles to CHO cell lysate. a) Counts of HCP identities detected in CHO cell lysate versus total count of HCPs detected in each drug product HCP profile. The solid line represents $y = x$ diagonal and corresponds to a drug product HCP profile with all protein identities detected in the CHO cell lysate. Diagonal dashed lines correspond to 90%, 80%, 70%, 60% and 50% of protein identities in drug product HCP profiles that were also present in CHO cell lysate. b) Quantile plot of the ranks of NSAF abundances of HCPs that were *not* detected in CHO cell lysate (red curve). The diagonal line represents the average abundance ranks of random samples of HCPs. Vertical dashes indicate the top and bottom deciles, quartiles and median (sorted in ascending order by the ranks of NSAF abundances). Horizontal dashes indicate corresponding abundance ranks for respective drug product HCP profiles. c) Graphical comparison of the abundance ranks of the proteins detected *both* in drug product HCP profiles and CHO cell lysate. The diagonal line represents equal abundance ranks in the drug product HCP profile and in CHO cell lysate.

	111.2	111.3	111.3v	111.4	111.5	111.5c	111.5r	112.1	112.3	112.4L	112.5	112.7	134.1	135.1	136.1	137.1	138.1	223.1
111.2	686	-243.8	-159.5	-269.1	-238.3	-270.9	-233.3	-79.1	-72.9	-53.2	-71.0	-99.4	-2.9	-19.0	-43.9	-19.8	-13.0	-2.9
111.3	0.28	660	-159.6	-223.9	-207.6	-197.8	-186.8	-72.3	-69.1	-54.6	-63.6	-89.6	-3.3	-22.2	-47.2	-22.0	-18.4	-4.1
111.3v	0.38	0.39	538	-137.5	-136.3	-142.3	-140.0	-58.7	-54.1	-51.4	-51.6	-69.1	-3.3	-20.9	-35.3	-19.8	-15.3	-4.4
111.4	0.27	0.28	0.40	650	-304.6	-240.9	-231.3	-70.1	-73.3	-50.1	-70.0	-82.0	-1.6	-16.4	-41.5	-17.3	-11.6	-4.0
111.5	0.27	0.30	0.40	0.22	632	-251.9	-245.4	-67.2	-68.0	-46.9	-63.9	-79.7	-1.3	-15.7	-42.0	-18.2	-11.0	-2.9
111.5c	0.26	0.32	0.40	0.28	0.27	644	-278.8	-66.3	-61.7	-49.5	-64.3	-87.9	-1.8	-20.4	-43.6	-22.9	-12.9	-3.2
111.5r	0.30	0.33	0.41	0.29	0.26	0.24	578	-63.3	-55.9	-43.2	-61.0	-77.4	-2.1	-16.6	-38.7	-18.8	-13.3	-2.4
112.1	0.61	0.60	0.60	0.60	0.59	0.61	0.59	334	-119.0	-92.9	-115.3	-120.7	-2.0	-14.9	-31.6	-18.4	-13.8	-3.1
112.3	0.61	0.60	0.59	0.57	0.58	0.60	0.58	0.35	314	-85.4	-141.2	-105.4	-1.4	-12.7	-28.6	-14.0	-11.1	-3.4
112.4L	0.67	0.65	0.62	0.67	0.64	0.65	0.64	0.47	0.48	312	-93.0	-86.2	-2.0	-15.0	-28.2	-14.6	-13.3	-2.5
112.5	0.62	0.61	0.59	0.59	0.59	0.60	0.58	0.34	0.24	0.43	305	-111.0	-2.0	-10.7	-25.1	-13.8	-9.8	-4.0
112.7	0.54	0.53	0.53	0.52	0.50	0.52	0.52	0.37	0.39	0.50	0.38	423	-2.4	-16.6	-34.9	-21.5	-15.4	-3.0
134.1	0.92	0.91	0.91	0.92	0.92	0.93	0.92	0.89	0.89	0.89	0.88	0.90	95	-6.4	-5.7	-6.1	-5.5	-2.7
135.1	0.78	0.76	0.76	0.77	0.76	0.76	0.77	0.76	0.76	0.75	0.76	0.75	0.84	285	-41.6	-62.7	-53.1	-2.1
136.1	0.64	0.64	0.67	0.64	0.64	0.64	0.65	0.72	0.72	0.74	0.71	0.67	0.89	0.60	528	-42.1	-22.8	-5.3
137.1	0.76	0.76	0.75	0.77	0.76	0.74	0.75	0.72	0.75	0.76	0.73	0.73	0.82	0.41	0.60	278	-52.1	-1.7
138.1	0.81	0.80	0.80	0.81	0.80	0.80	0.79	0.74	0.75	0.77	0.74	0.76	0.80	0.43	0.68	0.44	220	-1.2
223.1	0.88	0.87	0.87	0.88	0.88	0.87	0.88	0.87	0.87	0.85	0.88	0.88	0.92	0.89	0.86	0.88	0.89	405

Table S2: Dissimilarity measures (lower values correspond to greater similarity) for each pair of HCP profiles for all samples analyzed in this study: Jaccard distances between HCP profiles are in the lower triangle, log base 10 of statistical significances of Spearman correlations between HCP abundances are in the upper triangle. Diagonal represents total count of HCPs detected in each lot.

1 Computational details

Similarity between each pair of HCP profiles herein is assessed both in terms of the overlap between sets of HCP identities detected for each sample analyzed as well as the correlation between abundances of HCPs identified in both samples being compared.

1.1 Jaccard distance

The overlap between two sets of HCP identities detected in two HCP profiles, A and B, is quantified by the Jaccard distance, $d_J(A,B) = 1 - J(A,B)$, where $J(A,B) = |A \cap B|/|A \cup B|$ is the Jaccard index, the ratio of the size of the overlap between the two sets to the size of their union. This measure is commonly used for similar purposes in the proteomics field [1, 2, 3, 4, 5], and is also closely related to Tanimoto distance (another commonly used measure of similarity between sets [6, 7]). The resulting value d_J ranges between zero when two sets are identical and unity for completely disjointed sets of HCP identities. Given two sets of HCP identities, Jaccard distance was calculated as one-complement of the ratio of the number of HCPs detected in each of the two samples to the total number of unique HCPs detected in at least one sample.

1.2 Spearman correlation

For the proteins that are common between two HCP profiles, further comparisons are made on the basis of their NSAF protein abundances. The Spearman rank correlation coefficient, commonly employed for proteomics data sets [8, 9, 10], is used to quantify similarity between the abundances of HCPs that were detected in two samples. It is calculated using conventional expression $\rho = \sum_i (x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum_i (x_i - \bar{x}) \sum_i (y_i - \bar{y})}$, where x_i and y_i are ranks of the corresponding protein abundances in HCP profiles X and Y, and $\bar{x}$ and $\bar{y}$ represent average ranks. It yields values between 1 and -1, the former indicating HCP abundances that appear in the same order in both samples, and the latter indicating the order of HCP abundances is reversed. The Spearman coefficient is a non-parametric statistic that is insensitive to the shape of the distribution that NSAF values are drawn from, only to their relative order. The corresponding estimate of statistical significance, $p_s = \Pr(\geq \rho)$, represents the probability of observing a Spearman correlation coefficient of this magnitude or larger for a given number of HCP identities upon random reordering.

1.3 Multivariate analysis

Results from hierachical clustering and multidimensional scaling presented below have been obtained using standard functionality available in statistical software R [11]. Specifically, heat map visualizations were obtained using `heatmap.2` from R package `gplots`, Ward clustering method (as implemented in `hclust`) was used in rendering heat maps and classical multidimensional scaling (MDS) was performed using `cmdscale`.

2 Analytical method development

2.1 CHO lysate and Protein A column

Effect of Protein A column depletion upon protein abundances (as estimated by shotgun proteomics NSAF) has been evaluated by analyzing CHO cell lysate via LC-MS before and after Protein A treatment. Two technical replicates have been analyzed in each case to evaluate technical variability. Overall, 3297 protein identities were detected in this experiment at least once before or after

Protein A depletion. The majority of them, 2506 proteins, ($\sim$ 76% of the total) have been detected both before and after Protein A column. Of the remaining HCP, 507 proteins have been detected before but not after Protein A depletion and 284 HCPs have been identified after but not before Protein A treatment. The increase in protein counts not detected after protein A column is fairly modest ($\sim$ 7% of the total count of proteins detected) and is likely due to experimental variability.

The overlap between HCP identities detected in the CHO cell lysate before and after the Protein A column, and scatterplot of their NSAF abundances are presented in Figure 3 in the main text (panels a and b, respectively). Further comparisons of the NSAF estimates of protein abundances in the CHO cell lysate as determined before and after Protein A treatment are presented in Figure S1. Panel a) demonstrates that proteins consistently detected before and after Protein A column tend to be more abundant than those detected only before or only after depletion, and the abundances of proteins detected either only before or only after Protein A column are roughly comparable (Wilcoxon test p-value of $p = 0.89$). Panel b) in Figure S1 compares abundances of proteins before and after Protein A column indicating that the abundances of proteins in CHO cell lysate after Protein A treatment are quantitatively representative of those prior to that. Furthermore, red color in panel b) indicates proteins identities that were among HCPs detected in the commercial grade drug product lots in this study and shows that the lower abundant proteins from CHO cell lysate that tend to be more affected by the Protein A treatment are relatively sparsely represented among HCPs.

2.2 Protein A vs. Protein L depletion

To assess the impact of the choice between Protein A vs. Protein L column depletion one of the lots of drug product "1" generated using cell line "1" and process "2" (112.4L) has been analyzed using Protein L column depletion and compared to the HCP profiles of the other lots of the same drug product that were obtained using depletion by Protein A column (112.1, 112.3, 112.5, 112.7). Figures S2 and S3 depict overlaps between resulting protein identities and correlations between HCP abundances estimated by NSAF. Their comparison for HCP profiles obtained using Protein L and Protein A depletions shows that the overlap between HCP identities detected using Protein L purification and those obtained using Protein A tends to be smaller than those observed between two different drug product lots both processed by Protein A column. However, the majority of the proteins detected using Protein L depletion is also detected using Protein A and vice versa. Additionally, protein abundances obtained using Protein A column are highly correlated with those obtained using Protein L indicating that their relative abundances are reasonably comparable and likely representative of those found in the drug product prior to the depletion.

2.3 Protein L depletion repeatability

Reproducibility and column-to-column variability of Protein L depleted samples was also assessed. Three 40mg aliquots of mAb1 were depleted on three different Protein L columns, and were then prepared and analyzed by LC-MS/MS using the same procedures outlined in the Materials and Methods section. The samples used had low levels of total HCP; therefore, lower numbers of individual HCPs were detected (169, 123 and 131 protein identities per replicate). For each pairwise combination of resulting HCP profiles (in total, three pairs for the three aliquots analyzed), the values for the overlaps of protein identities, as quantified by Jaccard distance, were 0.43, 0.37 and 0.33 (%RSD $\approx$ 14%), indicating that the majority of the protein identities were consistently detected across repeated Protein L depletions. Furthermore, NSAF abundances of HCPs repeatedly

identified in these replicates were highly correlated (Spearman $\rho \geq 0.96$, $p < 10^{-50}$), demonstrating reproducible detection of HCPs and estimation of their levels in Protein L depleted samples.

3 Effects of process differences

3.1 Selected examples

The sensitivity of this analytical methodology to the differences between the cell lines and processes used to produce drug product is illustrated in Figure S8, which compares HCP profiles for drug product lot 111.5 to a few selected examples that are expected to be very similar as well as more dissimilar.

HCP profile 111.3 was obtained for another lot of the same drug product (111). The overlap between HCP identities detected in this experiment and those found for 111.5 as well as the correlation between their abundances are very comparable to those from repeated analyses of drup product lot 111.5 (HCP profiles 111.5 and 111.5c).

Drug product lot 112.5 represents the same drug product produced by the same cell line with some process differences. This process results in HCP profile with lower total number of protein identities detected, the majority of which are also detected in 111.5. The correlation between their NSAF abundances ($\rho = 0.81$) is lower than that observed for repeated analyses of the same lot (111.5 and 111.5c: $\rho = 0.94$) or for the analyses of different lots of the same drug product (111.5 and 111.3: $\rho = 0.91$). It is still highly unlikely to be observed by random chance ($p < 10^{-60}$) thus suggesting that these NSAF abundances are representative of their relative levels in the drug product.

Lastly, the HCP profile for drug product 223.1 corresponds to an entirely different monoclonal antibody "2" (different amino acid sequence than that in monoclonal antibody "1") grown in a different cell line ("2" vs. "1") using different upstream/downstream setup ("3" vs. "1" or "2"). It can be seen that the majority of HCP identities detected in this case are unique to either 111.5 or 223.1 HCP profiles and that the correlation between NSAF abundances of those proteins that were detected in both samples is far lower than those observed for the same drug product lot and the same cell line.

3.2 Combined analysis of all HCP profiles

A more comprehensive evaluation of similarities and differences detected by this method was afforded by the analysis of all HCP profiles presented in this study together. Corresponding quantitative measures of their dissimilarity in the form of Jaccard distance and statistical significance of Spearman correlation between their abundances are presented in Table S2 and also graphically rendered in the Figure 4 of the main text.

The HCP profiles obtained for different lots of the same drug product as well as repeated analyses of the same drug product lots are most similar to each other both in terms of the overlaps between sets of HCP identities and the correlations between their abundances. For example, HCP profiles resulting from the repeated analyses of the same drug product lot multiple times changing nano and protein A column (e.g. 111.5, 111.5r and 111.5c respectively), on different model of mass spectrometer (e.g. 111.3 and 111.3v) or depleting drug product lot from the same subset of lots using protein L instead of protein A column (e.g. 112.4L and the rest of 112* lots) display greater similarity (by overlap between HCP identities or by correlation of NSAF abundances) with the other lots belonging to the same subset than with those from any other one.

It can be also seen from these comparisons that the effect of the process differences can impact the similarity of the resulting HCP profiles to varying degree. For instance, the impact of the differences between process "2" and process "1" on the HCP profiles, as illustrated by the analysis of multiple lots of drug product obtained from them (e.g. 111.x and 112.x) is fairly modest. The resulting HCP profiles for the multiple lots of drug product produced by each of these two processes are more similar among those produced by the same process than to the HCP profiles of the same drug product produced by another process, but altogether they can be distinguished from HCP profiles characterizing the same monoclonal antibody expressed in a distinct cell line (13x.x). By comparison the difference between process "4" (as exemplified by the HCP profile 134.1) and processes "5" through "8" (HCP profiles 135.1, 136.1, 137.1 and 138.1 respectively) is much more pronounced. In this case the difference between 134.1 and other HCP profiles for the same monoclonal antibody "1" and same cell line "3" is on par with the dissimilarity observed for an entirely difference monoclonal antibody expressed in a different cell line (223.1). Such robustness of the results of the comparison with respect to the measure of similarity (overlap of HCP identities and significance of non-parametric correlation between NSAF abundances of matching HCPs) suggests that differences in HCP profiles reflect inherent properties of the samples analyzed.

Additionally, as indicated by the Batch column in Table S1, some of the samples showing greater similarity (e.g. 111.2 and 111.3) were processed in distinct analysis batches, while some of the less similar HCP profiles (e.g. 111.2 and 134.1) were obtained from the same group of samples analyzed together. This demonstrates that the differences and similarities between HCP profiles presented here are not a direct consequence of batching these samples for analytical processing. Still, a study employing this analytical methodology for the evaluation of the similarity of HCP profiles for multiple samples of biotherapeutics must be properly designed to minimize possible (even if small) impact of batch effects [12].

Overall, from the comparison of HCP profiles as obtained by this approach for commercial grade monoclonal antibody drug product lots across multiple lots of the same drug product, repeated analyses of the same material, different monoclonal antibodies, various manufacturing processes and cell lines, the following conclusions can be drawn:

- the depth of characterization by the method presented here enables resolution of a sufficiently large number of HCPs present in commercial grade drug product lots that in combination with the follow-up quantitative treatment of the data detects reproducible differences in their identities and relative abundances that could be reflective of the differences in the type of monoclonal antibody produced, cell line expressing it, and other aspects of process differences
- these differences were sufficiently robust upon repeated analyses of the same material and with respect to the changes in the assay conditions in terms of whether protein A or L was used for depletion or which model of mass-spectrometer was used to acquire data

3.3 Impact of characterization depth

Large number of HCPs identified by the method presented herein allows evaluation of the sensitivity of the resulting conclusions to the depth of the characterization (in terms of the counts of HCPs detected in each sample). For instance, the same analyses can be repeated by using only top 10 (or, top 20, top 50, etc.) most abundant HCPs from each sample as a means of emulating experimental results obtained by the methods of lower resolution (in terms of the number of HCPs identified). Results of such analysis using (log base 10 of) the statistical significance of Spearman correlation between the abundances of the resulting HCP profiles for the top 10, 20, 50, 100, 200 and all HCPs

detected in each profile are shown graphically in the form of heat maps by Figure 5 in the main text. It is worth emphasizing here that the use of non-parametric Spearman correlation here makes these results insensitive to the renormalization of the subset of the HCPs abundances as long as it does not affect their relative order.

Although direct comparison of the statistical significances of correlation coefficients is difficult for the samples of different sizes, overall increase in the statistical significance of the correlation between HCP profiles with the increase in the number of most abundant HCPs used for their calculation is easy to notice from visual inspection of Figure 5 in the main text. It can be also seen that the higher counts of top most abundant proteins used for comparisons of HCP profiles correspond to clearer relationships between similarities based on HCP profiles and known properties of drug product lots analyzed. For instance, when only top 10 or top 20 most abundant HCPs are used for comparison, the distinctions between 111.x, 112.x and 13x.x groups of HCP profiles are not as clean as when larger numbers of HCPs are considered.

The last point is further illustrated by the Figure S9 that depicts results of the classical multidimensional scaling (MDS) on the dissimilarities between HCP profiles for few selected subsets of the most abundant HCPs. Dissimilarities between HCP profiles are represented as one complement of Spearman correlation between their protein abundances. As a result of MDS the distances between their positions in two-dimensional space reflect their relative dissimilarities – the points that are closer on the 2D plane are more similar to each other. It can be seen from this figure that with the increase in the number of HCPs used to estimate dissimilarity, the separation between 111.x and 112.x subsets of commercial grade drug product lots in the two-dimensional MDS rendering progressively improves.

Considered jointly, the results presented here demonstrate profound impact of the depth of characterization of HCP profiles upon the ability to draw conclusions about (dis-)similarities of HCP profiles. Furthermore, sufficiently deep resolution and quantitation of HCP abundances in monoclonal antibody drug product lots enabled the discrimination between HCP profiles of commercial grade drug product lots from different monoclonal antibodies, cell lines and manufacturing processes.

3.4 Comparison of HCPs and CHO cell lysate

For the comparison of the drug product HCP profiles (described in Section 3.2 above) to the proteins identified in CHO cell lysate (Section 2.1), the abundances of the cell lysate proteins have been averaged over all CHO cell profiles (two technical replicates before and after Protein A column). This yielded an average CHO cell lysate protein abundance profile with 3297 protein identities and their respective average NSAF abundances. Given that the majority of cell lysate proteins have been detected before and after Protein A column treatment and the strong correlation between their abundances before and after Protein A column (see Figure 3b in main text), the rank order of the protein abundances in the resulting average profile was very comparable to each of the individual CHO cell lysate profiles. The resulting average CHO cell lysate profile has been compared to each of the drug product HCP profiles in terms of overlap between protein identities and correlation between their abundances.

Figure S10a depicts overlaps between protein identities detected in each of the drug product HCP profiles and those in CHO cell lysate. Further discussion is provided in the main text.

Assessment of the abundances of HCPs that were not identified in CHO cell lysate is summarized in Figure S10b in the form of a quantile plot of the ranks of the NSAF abundances of the proteins detected only in HCP profiles. For each drug product HCP profile, every protein detected was

assigned a rank between zero and one (corresponding to the lowest and highest NSAF abundance in a given profile respectively). Figure S10b plots the resulting ranks of the proteins that were detected only in HCP profiles, but not in CHO cell lysate (pooled over all drug product HCP profiles and sorted in ascending order). The corresponding average of the sets of HCPs independently selected at random from drug product HCP profiles (with equal probability for every protein in HCP profile to be selected; $N = 1000$ simulations) closely follows $y = x$ diagonal. The standard deviation, indicated in Figure S10b by the dashes along the diagonal, is quite small (comparable to the line width). The top 10%, 25% and 50% of the most abundant HCPs in drug products that were *not* detected in the CHO cell lysate corresponded to the top 14%, 30% and 57% of the most abundant HCPs in drug products, *overall*, respectively.

The abundances of the proteins that were detected both in the final drug product HCP profiles *and* in CHO cell lysate were graphically compared in Figure S10c. For each HCP profile and for the average CHO cell lysate abundance profile every protein in the profile was assigned a rank between zero and one (corresponding to the least and most abundant proteins in a given abundance profile, respectively). For each HCP profile, abundance ranks of proteins that were detected both in HCP profile and in CHO cell lysate were pooled over all HCP profiles. As a result, proteins that were detected in multiple HCP profiles appear in Figure S10c multiple times with the same value for CHO cell lysate abundance rank remaining the same and different values of abundance ranks in HCP profile. Resulting correlation between abundance ranks in CHO cell lysate and HCP profiles (Spearman's $\rho = 0.4$) was highly unlikely to be observed by chance for two independent samples. For example, the probability of obtaining a correlation of $\rho = 0.3$ or greater for a sample size of $n = 3000$ by random chance is estimated to be below 10^{-60} and both the correlation coefficient and number of points ($n = 5250$) observed here are much higher. Approximately 40% of the points in Figure S10(c) represent 2x or greater difference between abundance ranks in CHO cell lysate and HCP profiles.

Supplemental References

[1] M. E. Sardiu, Y. Cai, J. Jin, S. K. Swanson, R. C. Conaway, J. W. Conaway, L. Florens, and M. P. Washburn, "Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics," *Proceedings of the National Academy of Sciences*, vol. 105, no. 5, pp. 1454–1459, 2008.

[2] B. Meunier, E. Dumas, I. Piec, D. Bechet, M. Hebraud, and J.-F. Hocquette, "Assessment of hierarchical clustering methodologies for proteomic data mining," *Journal of proteome research*, vol. 6, no. 1, pp. 358–366, 2007.

[3] M. Hauskrecht, R. Pelikan, M. Valko, and J. Lyons-Weiler, "Feature selection and dimensionality reduction in genomics and proteomics," in *Fundamentals of Data Mining in Genomics and Proteomics*, pp. 149–172, Springer, 2007.

[4] M. A. Grobei, E. Qeli, E. Brunner, H. Rehrauer, R. Zhang, B. Roschitzki, K. Basler, C. H. Ahrens, and U. Grossniklaus, "Deterministic protein inference for shotgun proteomics data

provides new insights into arabidopsis pollen development and function," *Genome research*, vol. 19, no. 10, pp. 1786–1800, 2009.

[5] J. Bruand, T. Alexandrov, S. Sistla, M. Wisztorski, C. Meriaux, M. Becker, M. Salzet, I. Fournier, E. Macagno, and V. Bafna, "Amass: algorithm for msi analysis by semi-supervised segmentation," *Journal of proteome research*, vol. 10, no. 10, pp. 4734–4743, 2011.

[6] L. M. Akella, T. Rejtar, C. Orazine, M. Hincapie, and W. S. Hancock, "Clue-tips, clustering methods for pattern analysis of lc-ms data," *Journal of proteome research*, vol. 8, no. 10, pp. 4732–4742, 2009.

[7] P. Zerefos, J. Prados, S. Kossida, A. Kalousis, and A. Vlahou, "Sample preparation and bioinformatics in maldi profiling of urinary proteins," *Journal of Chromatography B*, vol. 853, no. 1, pp. 20–30, 2007.

[8] K. Baerenfaller, J. Grossmann, M. A. Grobei, R. Hull, M. Hirsch-Hoffmann, S. Yalovsky, P. Zimmermann, U. Grossniklaus, W. Gruissem, and S. Baginsky, "Genome-scale proteomics reveals arabidopsis thaliana gene models and proteome dynamics," *Science*, vol. 320, no. 5878, pp. 938–941, 2008.

[9] T. Kislinger, A. O. Gramolini, D. H. MacLennan, and A. Emili, "Multidimensional protein identification technology (mudpit): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue," *Journal of the American Society for Mass Spectrometry*, vol. 16, no. 8, pp. 1207–1220, 2005.

[10] N. M. Griffin, J. Yu, F. Long, P. Oh, S. Shore, Y. Li, J. A. Koziol, and J. E. Schnitzer, "Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis," *Nature biotechnology*, vol. 28, no. 1, pp. 83–89, 2010.

[11] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[12] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[13] F. Leisch, "Sweave: Dynamic generation of statistical reports using literate data analysis," in *Compstat 2002 — Proceedings in Computational Statistics* (W. Härdle and B. Rönz, eds.), pp. 575–580, Physica Verlag, Heidelberg, 2002. ISBN 3-7908-1517-9.

[14] M. Goossens, F. Mittelbach, A. Samarin, and E. M. Souidi, *The LATEX companion*. Addison-Wesley Reading, Massachusetts, 1994.

Session Information

All computational results presented in this Supplemental Material have been obtained using standard facilities in statistical programming environment R [11]. Numerical and graphical results included in this document have been compiled using literate programming capacity supported in R by `Sweave` [13] and further converted into Portable Document Format (PDF) with LaTeXsuite of tools [14]. This section contains detailed info about the version of R software and additional packages that have been used throughout this analysis.

- R version 3.1.2 (2014-10-31), `i386-w64-mingw32`
- Locale: `LC_COLLATE=English_United States.1252`, `LC_CTYPE=English_United States.1252`, `LC_MONETARY=English_United States.1252`, `LC_NUMERIC=C`, `LC_TIME=English_United States.1252`
- Base packages: base, datasets, graphics, grDevices, methods, stats, utils
- Other packages: gplots 2.16.0, SuppDists 1.1-9.1, xtable 1.7-4
- Loaded via a namespace (and not attached): bitops 1.0-6, caTools 1.17.1, gdata 2.13.3, gtools 3.4.1, KernSmooth 2.23-13, tools 3.1.2