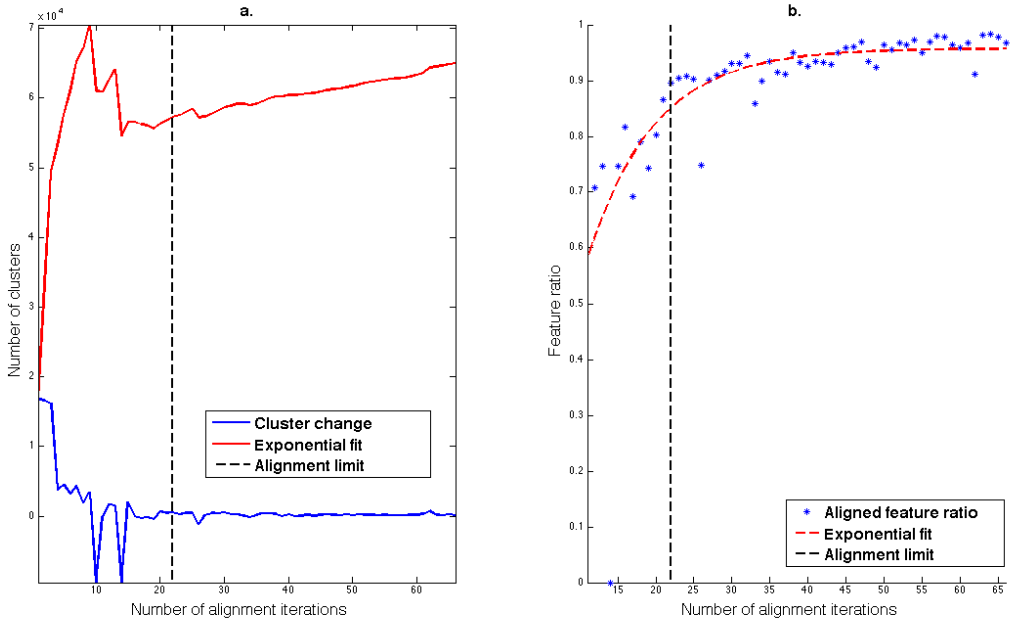
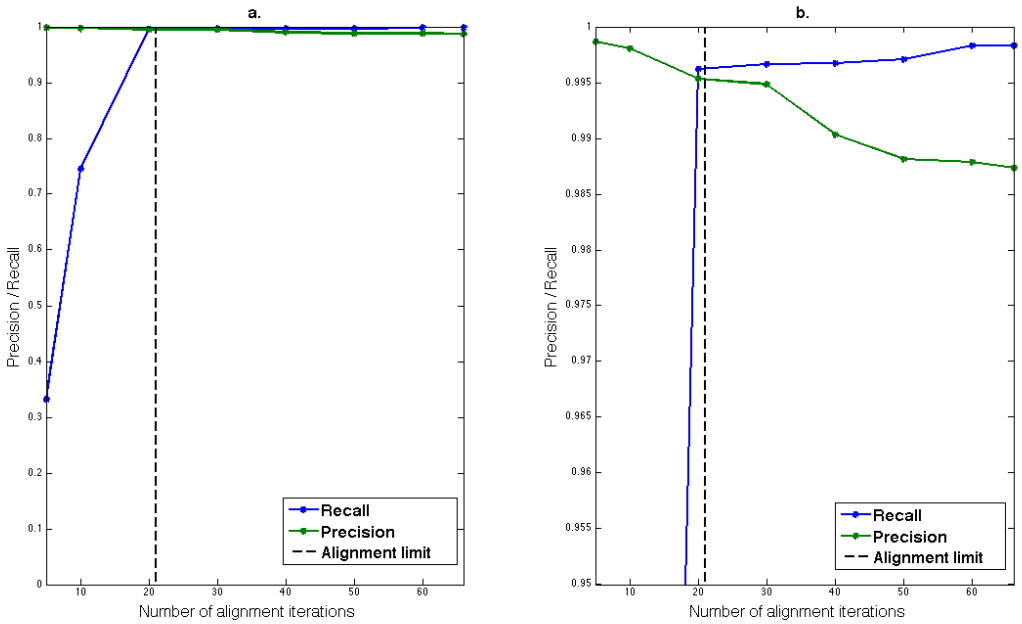
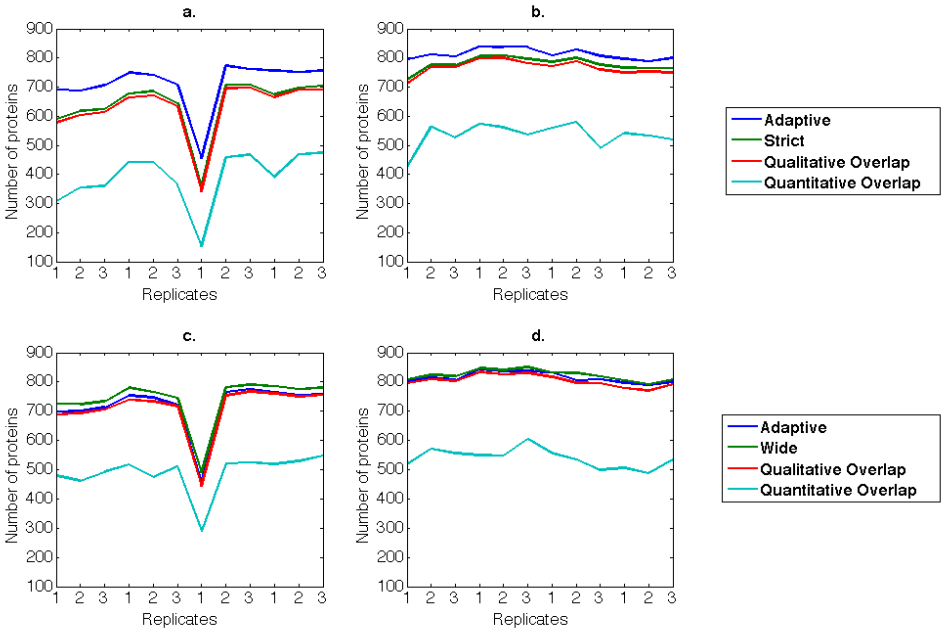


Supplementary Figure 1: (a) The change in number of clusters as file pairs selected for alignment are processed in the Clone dataset. The blue line represents the number of clusters added or combined during alignment whereas the red line corresponds to the total number of clusters, i.e. the cumulative sum of the blue line. The dashed black line marks the alignment limit cutoff in both plots, in this case after 21 matched file pairs. During alignment in Proteios, a ratio of features that are re-aligned to the total number of aligned features for a file pair is computed. This ratio is illustrated in figure (b) with an exponential curve fit to the nonzero ratios. As shown, they follow an increasing function. The alignment cutoff corresponds to a ratio of around 85%. By file pair 45, the ratio stabilizes around 95%. A high consecutive ratio indicates that subsequent matches add very little improvement to the alignment of the file set.

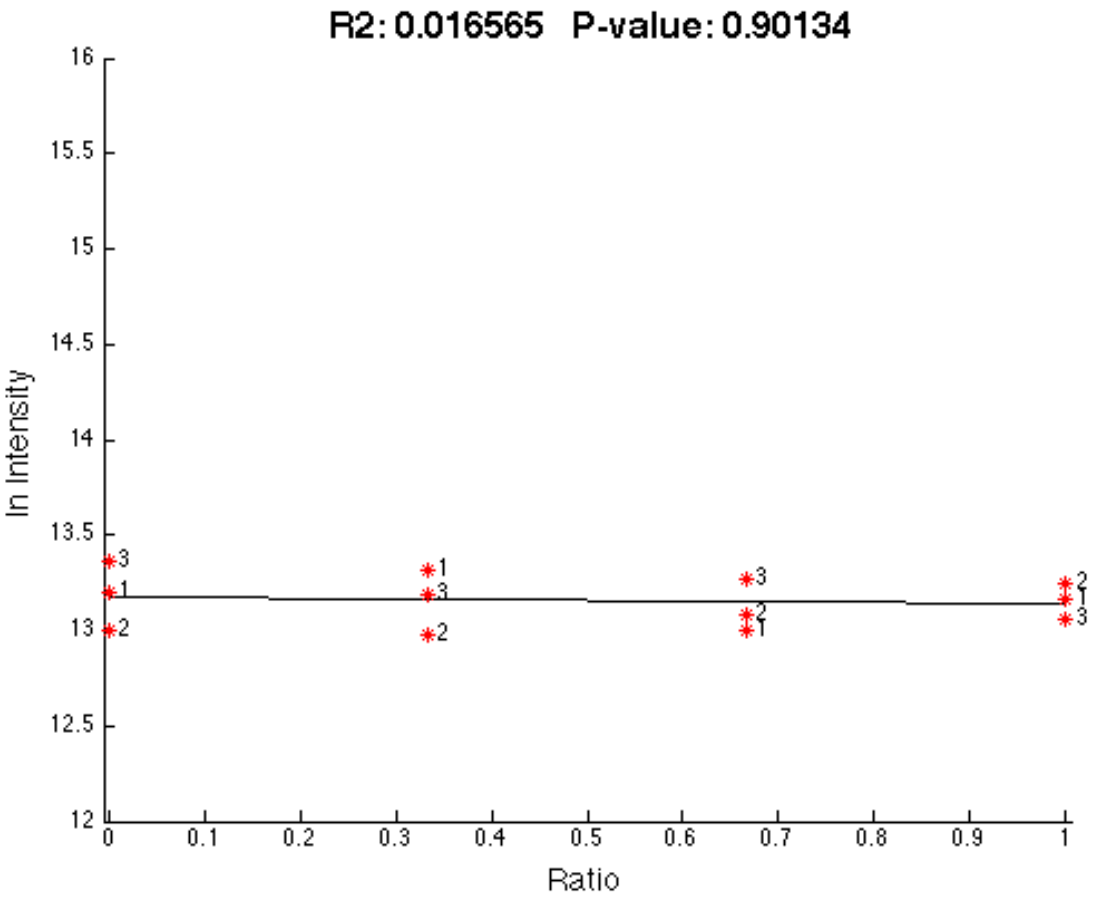


Supplementary Figure 2: (a) The change in global identity-based precision and recall for the Clone dataset when the alignment is terminated at increasing number of matched file pairs. The dashed black line marks the alignment limit. The recall increases rapidly at first as the number of alignment iterations increases. (b) A zoom in of the top part of (a) to allow distinction of the smaller recall changes seen after a certain number of alignments. As seen, aligning a larger number of file pairs would result in relatively little increase in recall, compared to the decrease in precision. Indeed, the absolute ratio of recall gain to precision loss when matching all files compared to aborting the alignment at the set limit is roughly 1/4.

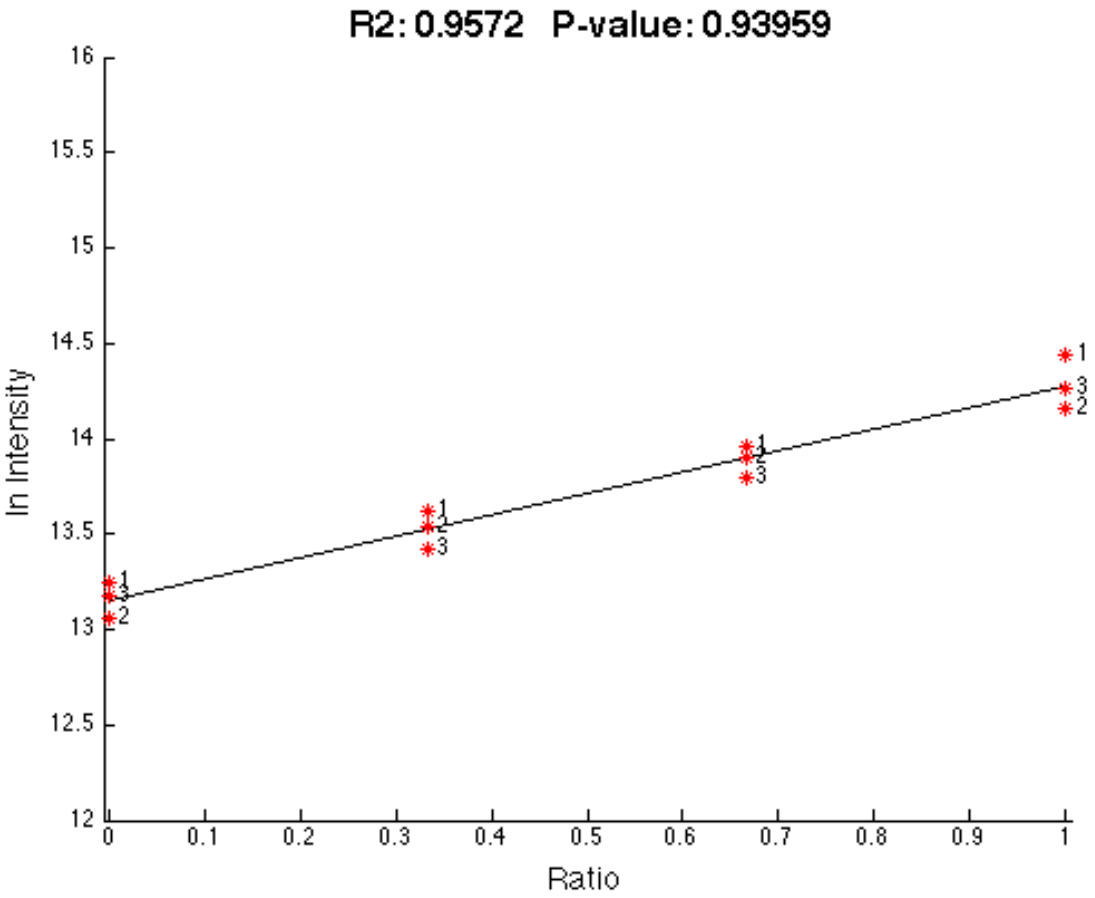


Supplementary Figure 3: Protein level evaluation. Peptides identified with FDR below 0.01 (including those propagated by alignment) were used to generate protein groups using the minimal protein set approach (Occam's razor). The minimum number of peptides required per protein group was set to one. This was performed for both the TimePoint (a, c) and Clone (b, d) datasets run with adaptive, strict and wide parameters, respectively. Every technical replicate was subsequently compared to itself regarding unique protein content and quantity. For the comparison, the leading protein (first in alphabetical order) of each protein group was extracted. Displayed for the different parameter settings is the number of proteins that are shared (qualitative overlap) within a replicate run and the number of proteins with the same derived quantity (quantitative overlap). The four ratio mixes are displayed in the same order and the replicates marked 1-3 as seen in Table 1 and Table 2. One deviating file can clearly be seen for the TimePoint dataset. (a) and (b) The datasets run with strict parameters. As can be expected, the number of detected proteins in every replicate is lower than for the adaptive algorithm. The opposite can be seen in (c) and (d), where the wide parameter settings have been used, although the difference is not as large. However, the quantitative overlap between the different parameter settings is relatively low, showing the impact of differing peptide quantities at the protein level. It

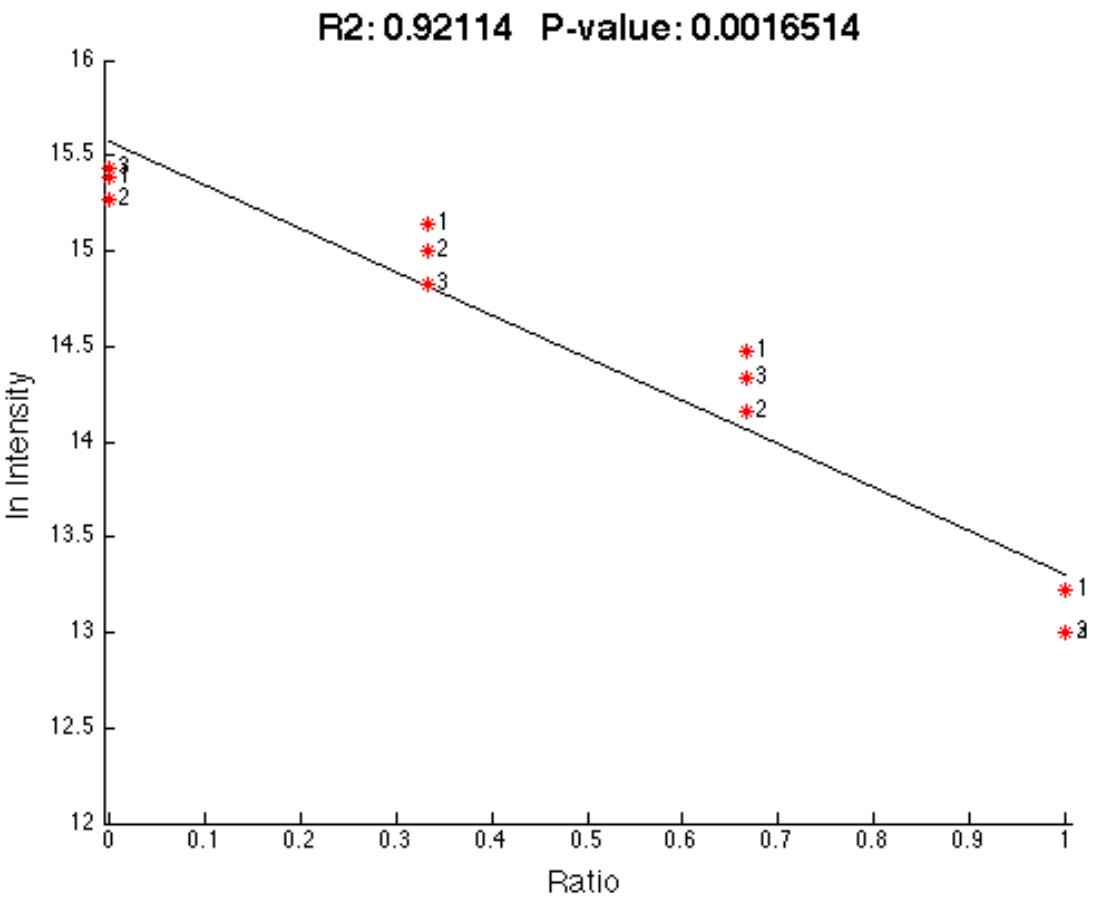
should however be noted that the quantitative overlap number is computed by requiring exactly the same protein quantity.



Supplementary Figure 4: Example of successful quantification and associated p-value. R2 (coefficient of determination) is included for completeness. If a peptide exists in both mixed samples in equal amounts, a linear response similar to the one above is expected and correct. As can be seen from the p-value for the lack-of-fit F-test, the linear model cannot be discarded at a significance level of 0.05. Further discussion on R2 compared to the lack-of-fit F-test can be found in the Supplementary Information of reference [5].



Supplementary Figure 5: Another example of successful quantification and associated p-value. R2 (coefficient of determination) is included for completeness.



Supplementary Figure 6: Example of an unsuccessful quantification and associated p-value. R2 (coefficient of determination) is included for completeness. The systematic deviation from linearity is clearly seen in the two intermediate ratio cohorts.