Supplementary Information for "mapDIA: Model-based Analysis of Quantitative Proteomics Data in Data Independent Acquisition Mode"

Prior distributions

The prior distribution for μ_{qg}, the average of the all intensities in peptide q group $g \in \{i, j\}$ and μ_q, or the average of the all intensities in peptide q group i and j, is conditional on the variance parameter σ_q^2 and is the Gaussian distribution with mean 0 and variance $(\sigma_q^2 \cdot V)$.
The hyperparameter V is set to 1000 to render this prior to be least subjective.

$$\mu_{qg}|\sigma_q^2 \sim \mathcal{N}(0, \sigma_q^2 \cdot V)$$
$$\mu_q|\sigma_q^2 \sim \mathcal{N}(0, \sigma_q^2 \cdot V)$$

The prior distribution for σ_q^2, the variance of the all intensities in peptide q group i and j, is the inverse gamma distribution with hyperparameters (a, b).
The hyperparameters (a, b) is set to the method of moments estimates of the gamma distribution based on the sample variance calculated assuming equal means across the two groups (i.e. assuming EE)

$$\sigma_q^2 \sim IG(a, b)$$

Closed form expression of the marginal likelihood

The closed form expression of $\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p)$ for the EE case.

$$
\begin{aligned}
&\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p = 0) \\
=&\int_0^\infty \int_{-\infty}^\infty \varphi(\mathbf{y}_q^i, \mathbf{y}_q^j | \mu_q, \sigma_q^2) \varphi(\mu_q | 0, \sigma_q^2 V) \Gamma^{-1}(\sigma_q^2 | a, b)\, d\mu_q d\sigma_q^2 \\
=&\frac{1}{\sqrt{(n_i+n_j)V+1}} \frac{\Gamma(a+(n_i+n_j)/2)}{\Gamma(a)} \frac{1}{(2\pi)^{(n_i+n_j)/2}} \\
&\times \frac{b^a}{\left[b + \frac{1}{2}\left(\sum_{y \in \mathbf{y}_q^i, \mathbf{y}_q^j} y^2 - (\frac{1}{n_i+n_j+1/V})(\sum_{y \in \mathbf{y}_q^i, \mathbf{y}_q^j} y)^2\right)\right]^{a+(n_i+n_j)/2}}
\end{aligned}
$$

The closed form expression of $\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p)$ for the DE case.

$$
\begin{aligned}
&\pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p = 1) \\
=&\int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \varphi(\mathbf{y}_q^i | \mu_{qi}, \sigma_q^2) \varphi(\mathbf{y}_q^j | \mu_{qj}, \sigma_q^2) \\
&\times \varphi(\mu_{qi} | 0, \sigma_q^2 V) \varphi(\mu_{qj} | 0, \sigma_q^2 V) \mathcal{IG}(\sigma_q^2 | a, b)\, d\mu_{qi} d\mu_{qj} d\sigma_q^2 \\
=&\frac{1}{\sqrt{n_i V+1}} \frac{1}{\sqrt{n_j V+1}} \frac{\Gamma(a+(n_i+n_j)/2)}{\Gamma(a)} \frac{1}{(2\pi)^{(n_i+n_j)/2}} \\
&\times \frac{b^a}{\left[b + \frac{1}{2}\left(\sum_{y \in \mathbf{y}_q^i} y^2 - (\frac{1}{n_i+1/V})(\sum_{y \in \mathbf{y}_q^i} y)^2 + \sum_{y \in \mathbf{y}_q^j} y^2 - (\frac{1}{n_j+1/V})(\sum_{y \in \mathbf{y}_q^j} y)^2\right)\right]^{a+(n_i+n_j)/2}}
\end{aligned}
$$

where $n_g = \sum_{y \in y_q^g} I\{\text{y observed}\}$ is the number of observed intensities in peptide q group g,

Estimation

The model parameters $\Phi = (\gamma, \beta)$ are estimated by the iterative conditional maximization (ICM) algorithm [1] as follows:

1. Obtain an initial estimate $\hat{\mathbf{Z}}$ of the true state $\mathbf{Z}_*$, using simple two sample t-tests.

2. Estimate Φ by the value $\hat{\Phi}$ which maximizes the pseudo-likelihood $\prod_p \pi(\{z_p\}_{i<j} | \{z_{(\partial p)}\}_{i<j}, \Phi)$.

3. Carry out a single cycle of ICM based on the current $\hat{\mathbf{Z}}, \hat{\theta}, \hat{\Phi}$, to obtain a new $\hat{\mathbf{Z}}$. For $p = 1, \ldots, P$, update z_p which maximizes

$$\pi(z_p|\mathbf{y}, \hat{z}_{(\Omega/p)}) \propto \left[\prod_{q \in \mathcal{I}_p} \pi(\mathbf{y}_q^i, \mathbf{y}_q^j | z_p, \hat{\theta})\right] \pi(z_p|\hat{z}_{(\partial p)}, \hat{\Phi}). \quad (1)$$

4. Go to step 3 until $\hat{\mathbf{Z}}$ converges.

This estimation is performed simultaneously for all pairwise comparisons specified by the user $\{(i, j)\}$ and a single set of MRF coefficients is applied to all the comparisons.

DEPs on the scale-free network

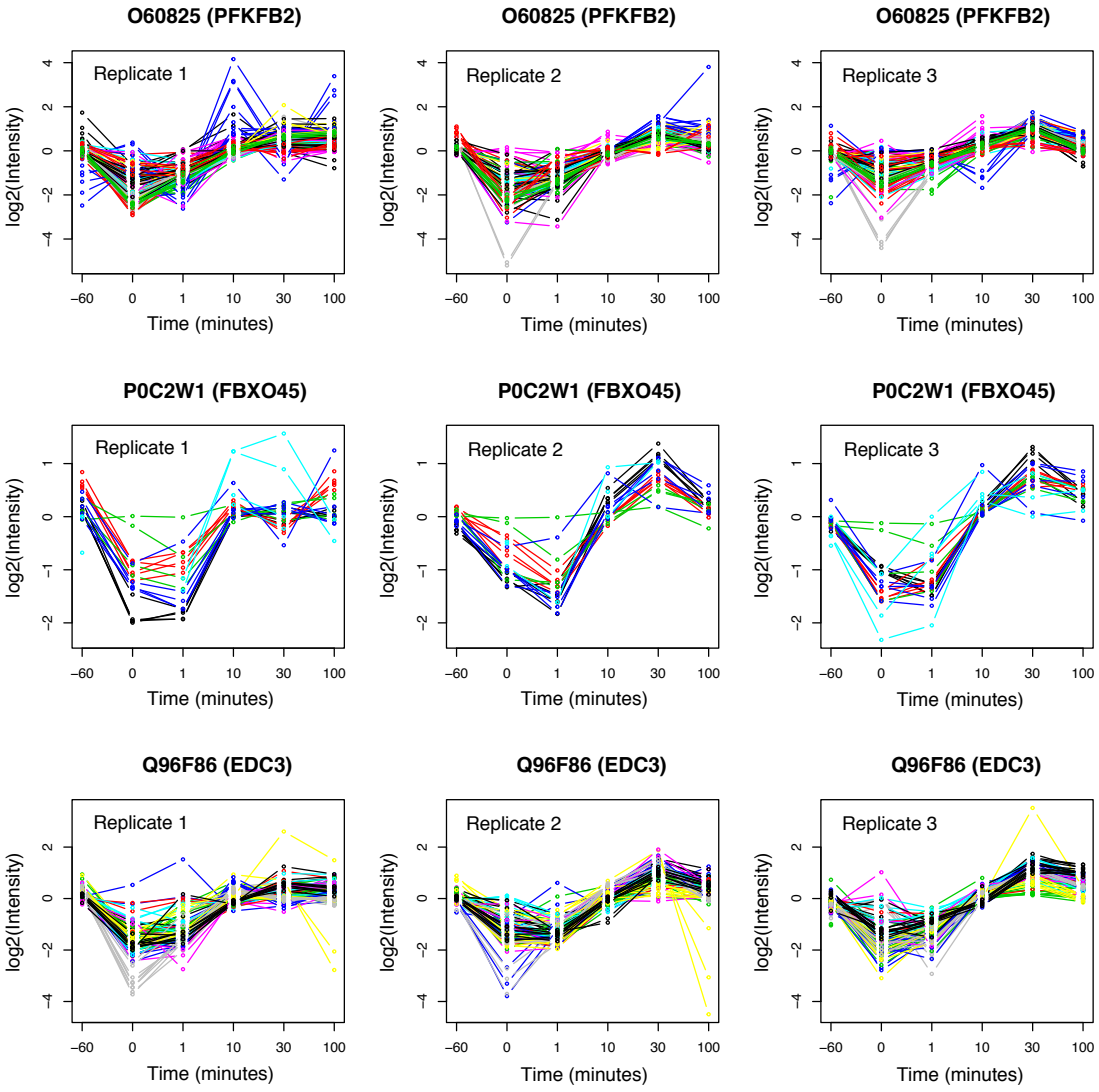
Using the algorithm of Herrera and Zufiria [2], the generation of the 1,500 node network starts with a circular graph of 11 nodes. Most these 11 nodes are highly connected and play the role of hubs in the protein interactome. 2 neighbouring nodes from these 11 nodes were arbitrarily selected as DEPs. Next, the neighbors of these 2 nodes were also set as DEPs. This process was repeated until 150 DEPs were produced.

References

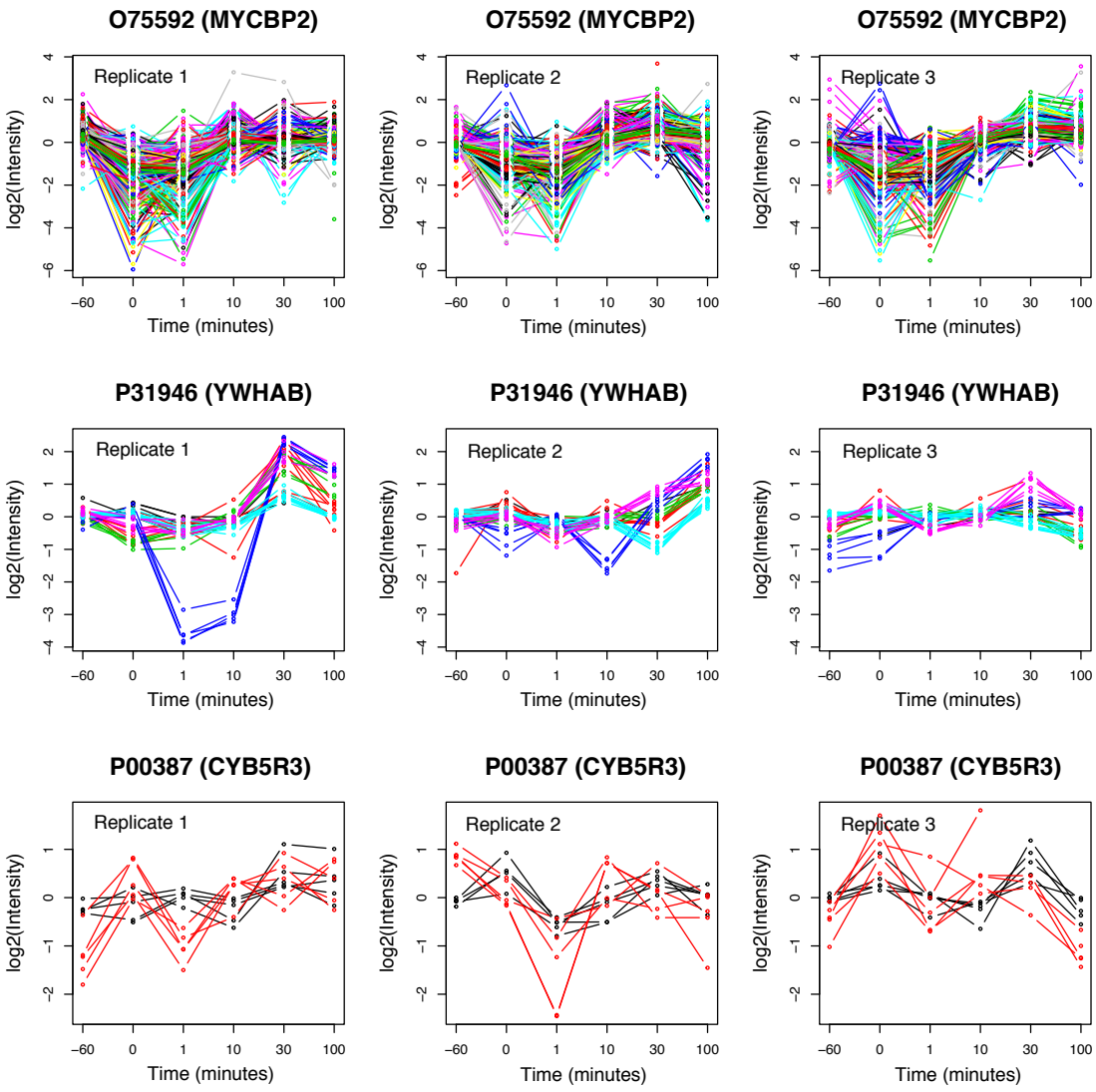
[1] J. Besag. On the statistical analysis of dirty pictures. *J. Royal Statist. Soc. B*, 48:259–302, 1986.

[2] C. Herrera and P.J. Zufiria. Generating scale-free networks with adjustable clustering coefficient via random walks. *arXiv*, 1105.3447, 2011.

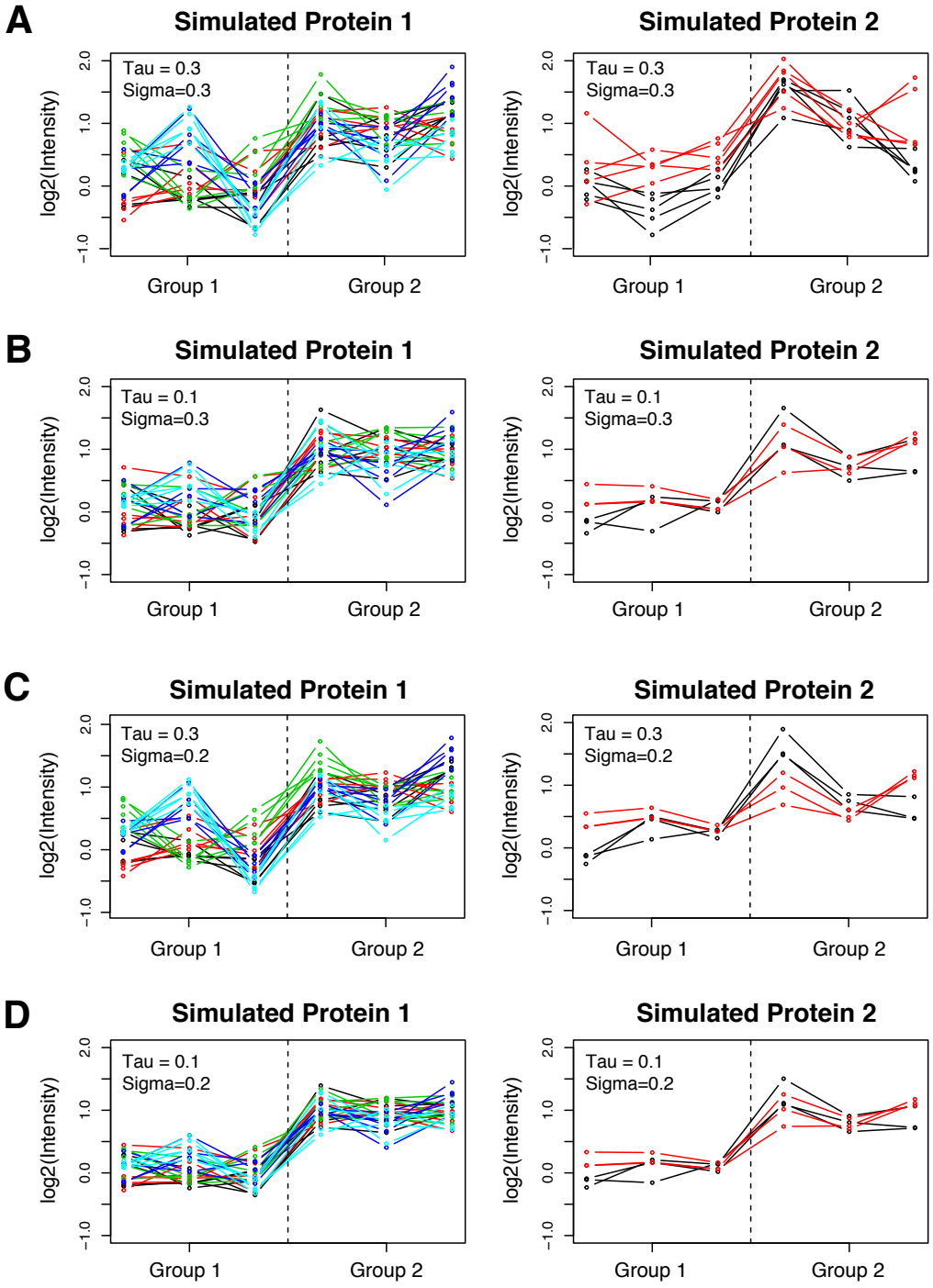
Supplementary Figures



Supplementary Figure 1 (14-3-3β interactome data set). The example of three proteins in which fragment-level intensity data are highly consistent within each peptide and peptide-level abundances are highly consistent within the same protein. In each protein, the preprocessed intensity data are log2 transformed and centered by median in each replicate within each fragment. Each line represents a time-course trajectory of log2 intensities of one fragment, with the lines of the same color corresponding to the same peptide.

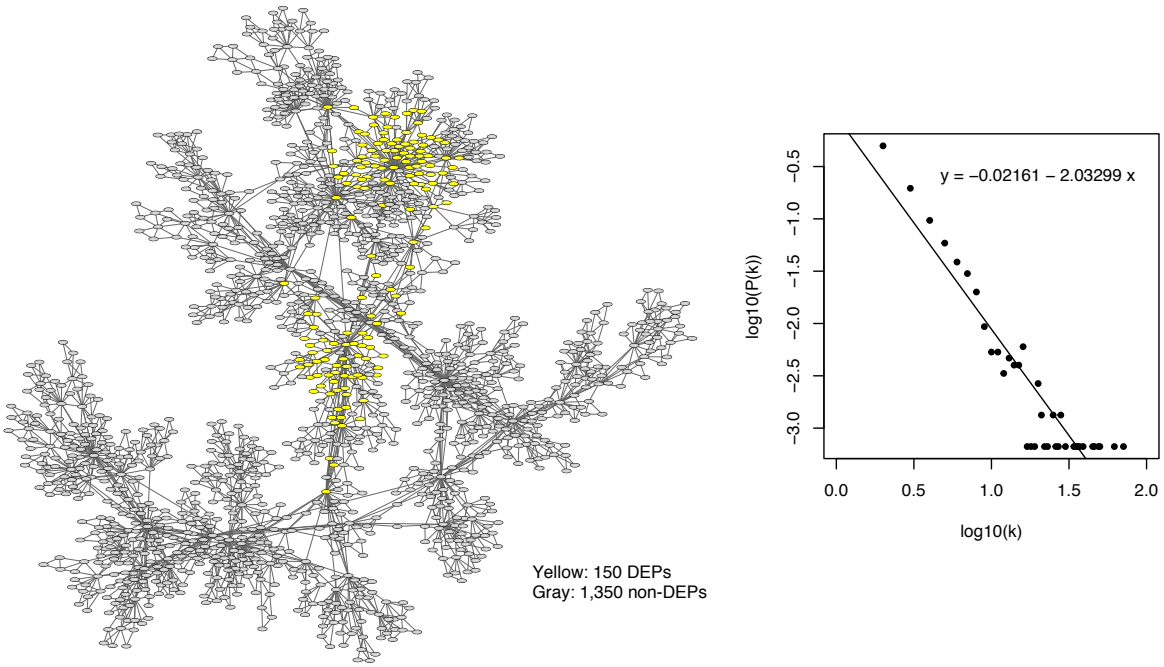


Supplementary Figure 2 (14-3-3β interactome data set). The example of three proteins in which peptide-level abundances are highly inconsistent within the same protein. Although there is poor correlation between peptides, fragments within the same peptide tend to show consistent log2 intensities. The scale of horizontal and vertical axes and the coloring scheme remain the same as Supplementary Figure 1.

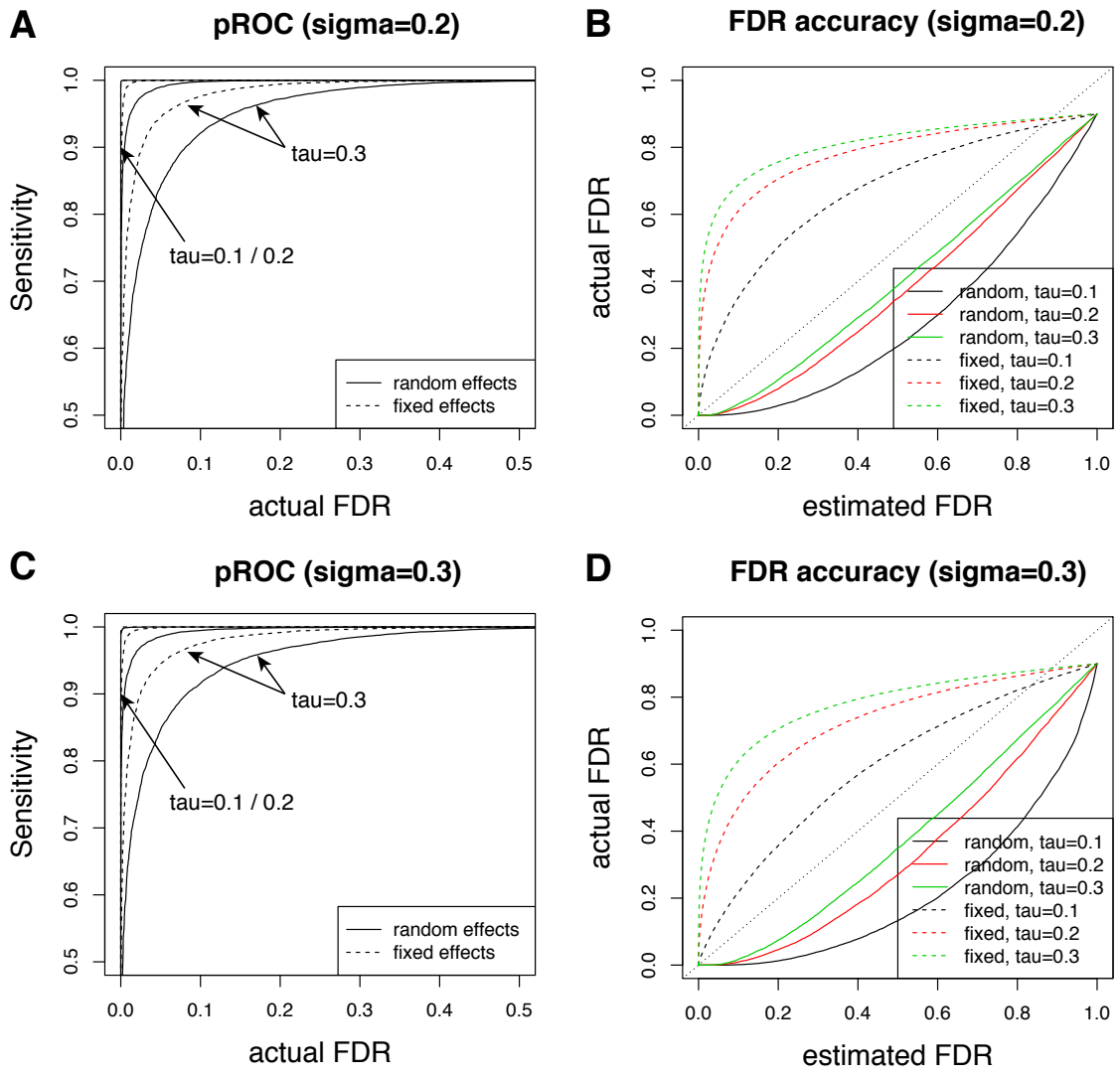


Supplementary Figure 3 (Simulation data set). Two example proteins across different simulation setting in terms of the peptide deviation from protein abundance τ and fragment intensity measurement error σ. In each panel, the log2 fragment intensities of each peptide were visualised by

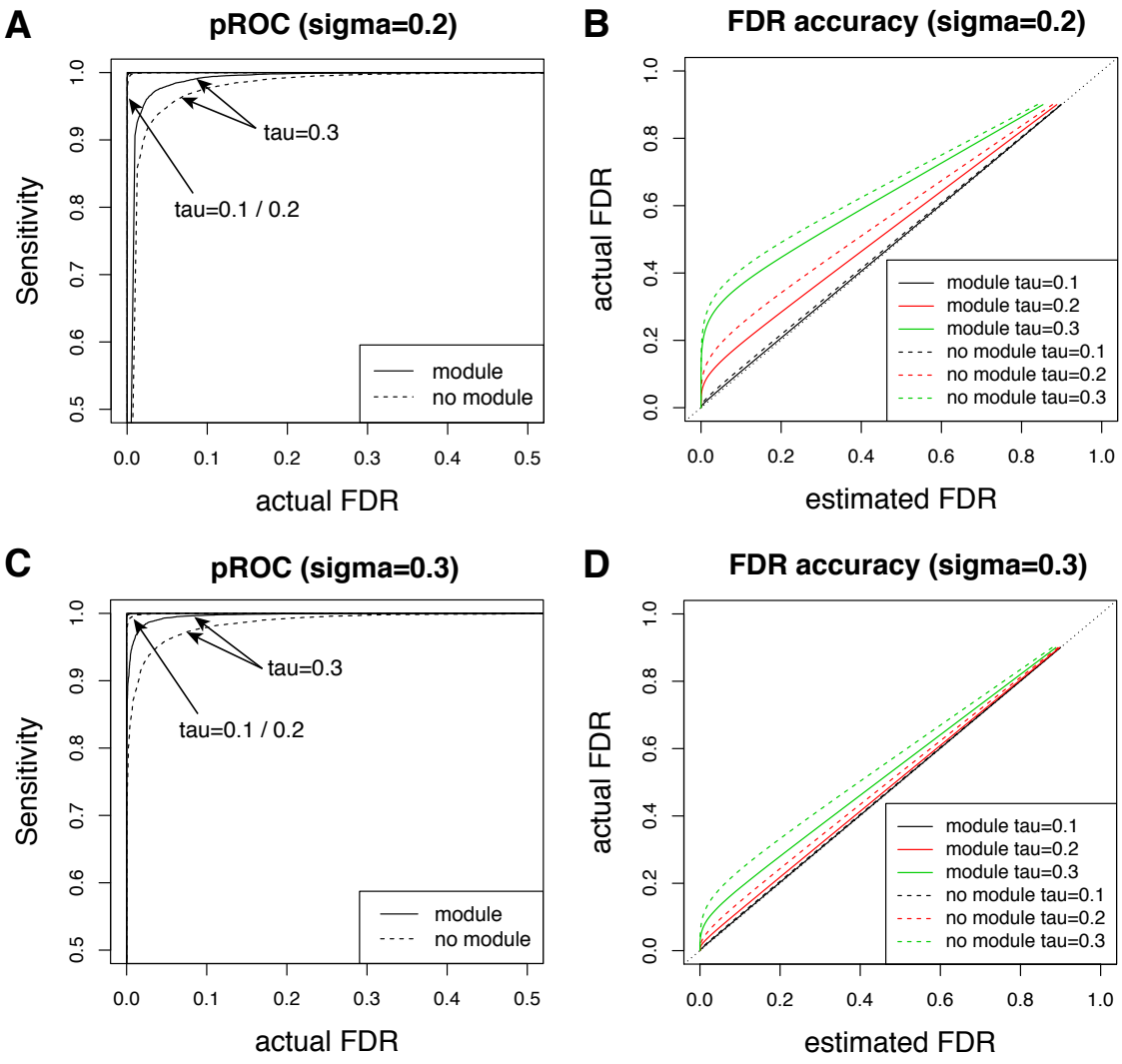
the dots of the same color, with additional lines connecting them across the samples. Simulation parameter settings as as follows. (A) $(\tau, \sigma) = (0.3, 0.3)$. (B) $(\tau, \sigma) = (0.1, 0.3)$. (C) $(\tau, \sigma) = (0.3, 0.2)$. (D) $(\tau, \sigma) = (0.1, 0.2)$. The scale of horizontal and vertical axes and the coloring scheme remain the same across all panels.



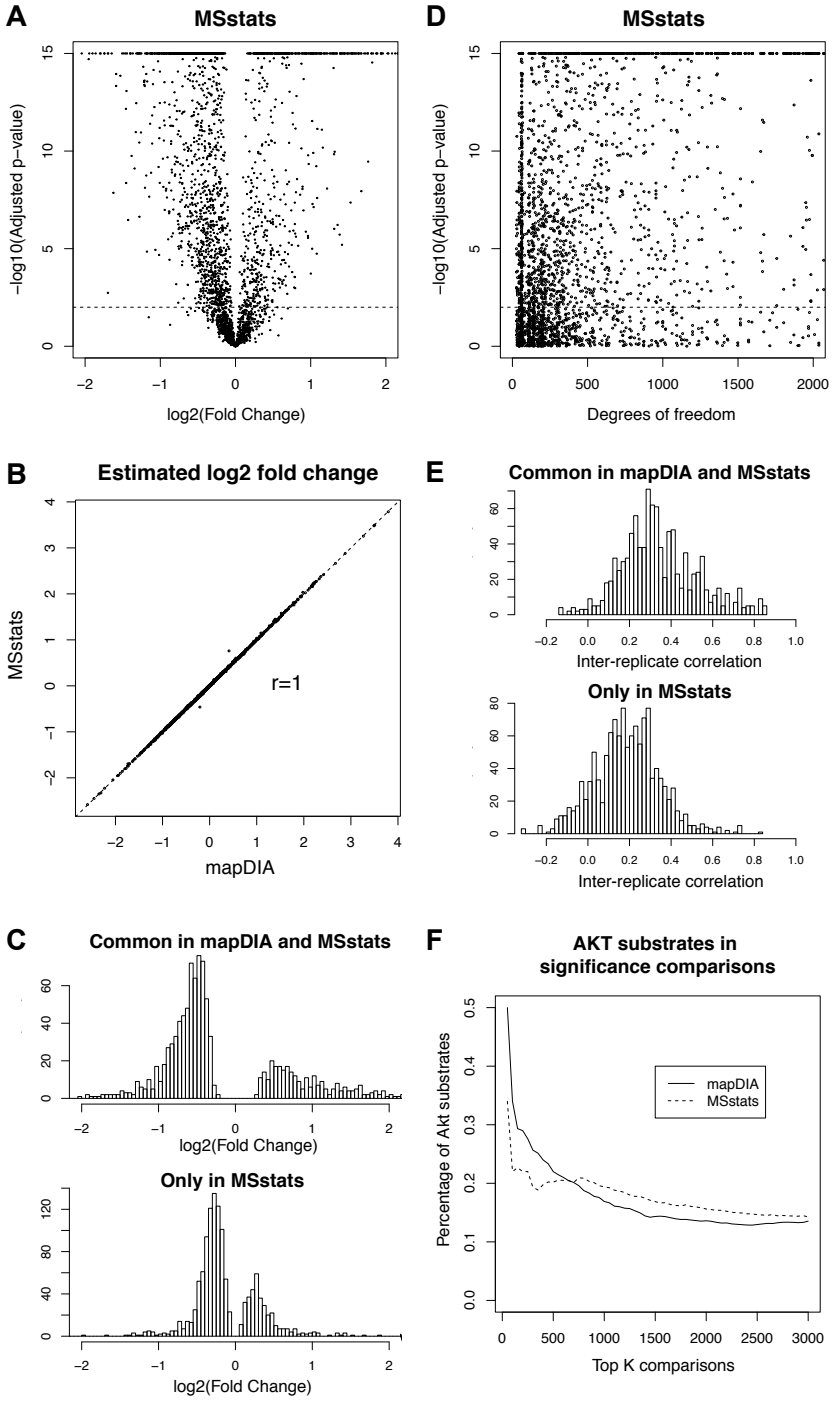
Supplementary Figure 4 (Simulation data set). The scale-free network of 1,500 proteins with 150 DEPs concentrated in localized subnetworks (yellow).



Supplementary Figure 5 (Simulation data set). Classification performance and FDR accuracy in MSstats. In each plot, the fragment intensity measurement error σ ("sigma") was fixed and the peptide deviation from protein abundance τ ("tau") was varied. (A, C) Sensitivity versus FDR (pseudo-ROC curve) plot and (B, D) FDR accuracy plot in MSstats with fixed effects model (dashed) and random effects model (solid) for different values of τ ranging from 0.1 to 0.3 at fixed value of σ at 0.2 or 0.3.

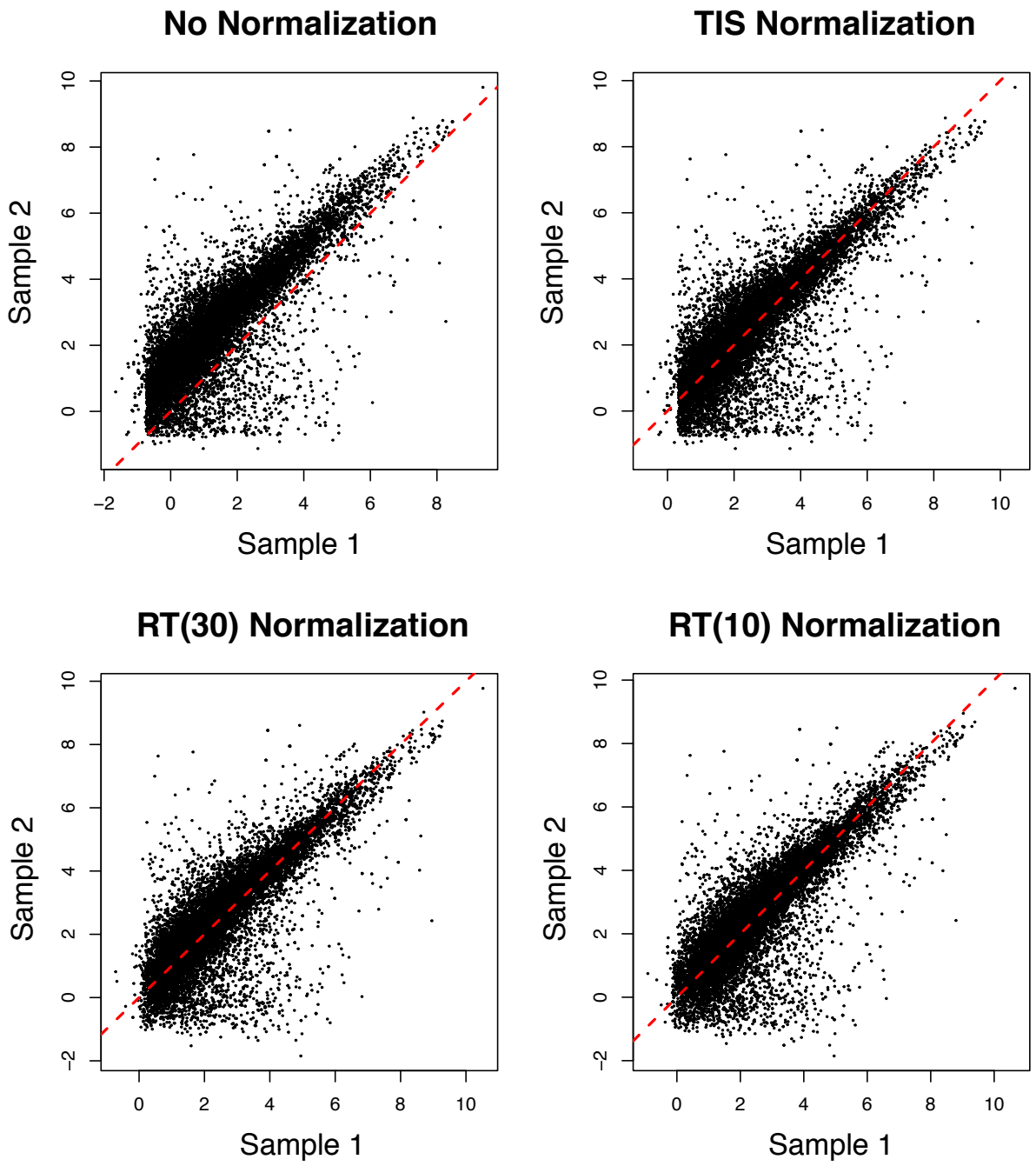


Supplementary Figure 6 (Simulation data set). Classification performance and FDR accuracy in mapDIA. In each plot, the fragment intensity measurement error σ ("sigma") was fixed and the peptide deviation from protein abundance τ ("tau") was varied. (A, C) Sensitivity versus FDR (pseudo-ROC curve) plot and (B, D) FDR accuracy plot in mapDIA with module information (solid) and mapDIA without module information (dashed) for different values of τ ranging from 0.1 to 0.3 at fixed value of σ at 0.2 or 0.3.



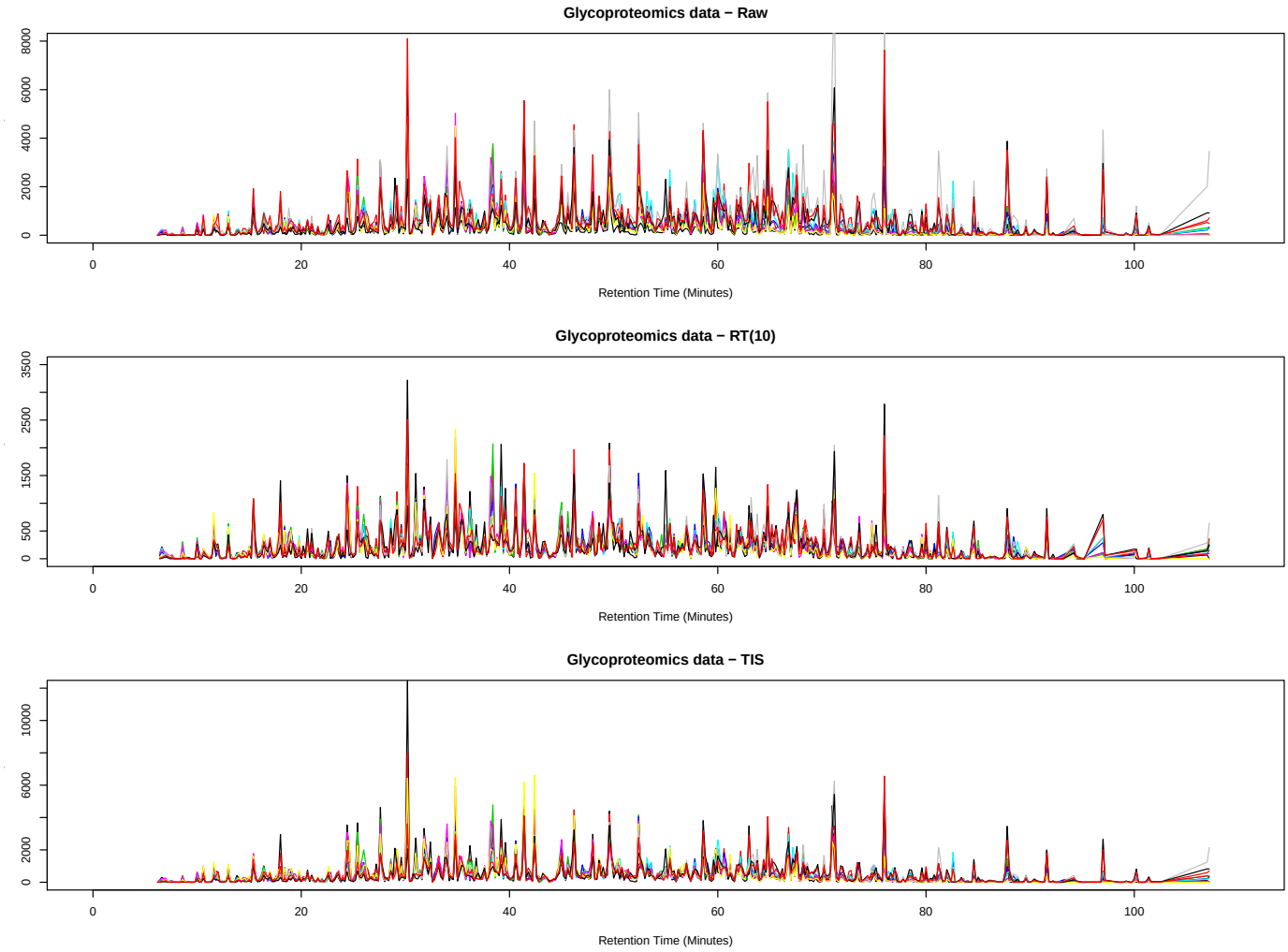
Supplementary Figure 7 (14-3-3β interactome data set). (A) Significance score versus log2 fold change in `MSstats`. (B) The reported log2 fold changes from `mapDIA` and `MSstats`. (C) Log2 fold changes for the comparisons found significant in both softwares (top) and those found

significant only in `MSstats` (bottom). (D) The trend in significance scores along the number of fragments (represented by the degrees of freedom in the regression model with fixed effects). (E) Inter-replicate correlation for the comparisons found significant in both softwares versus those significant in `MSstats` only. (F) Akt substrate enrichment in the top K comparisons in `mapDIA` and `MSstats`.

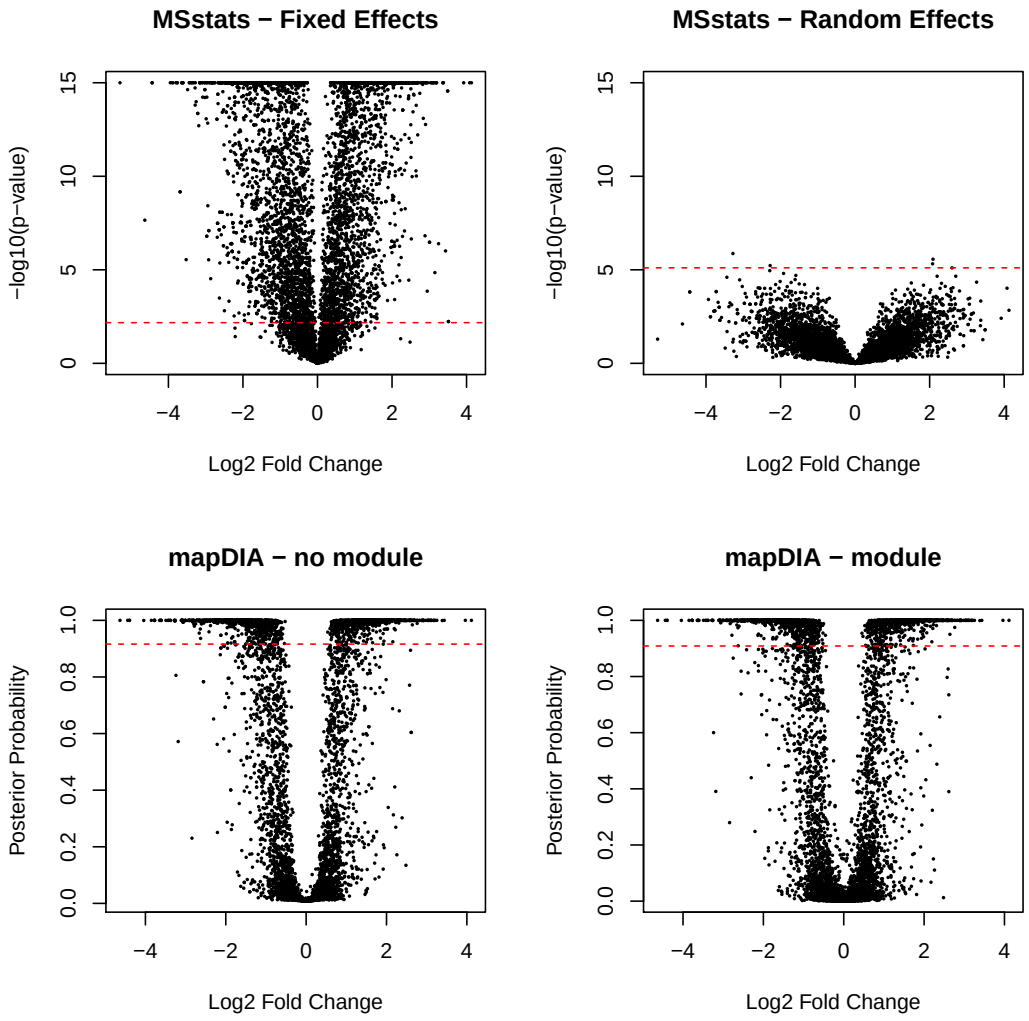


Supplementary Figure 8 (Glycoproteomics data set). Within-group pairwise scatter plot of fragment-level intensity data using four different normalization options: no normalization, TIC normalization, RT(30) and RT(10) normalization in prostate cancer glycoproteomics data (control groups). The trend and improvement was observed in the other three groups, which are not shown

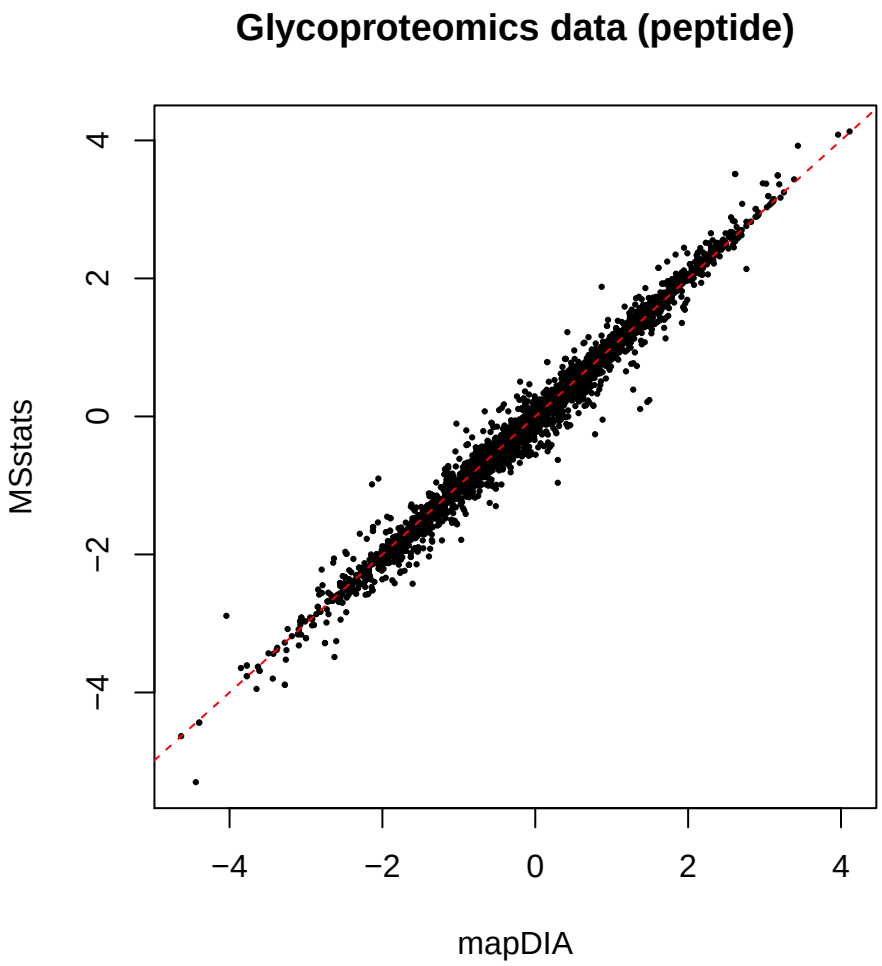
due to large file sizes.



Supplementary Figure 9 (Glycoproteomics data set). The TIC profiles of 10 samples with no intensity normalization (raw), RT(10) normalization, and TIS normalization in `mapDIA`.



Supplementary Figure 10 (Glycoproteomics data set). Statistical significance scores against log2 fold change in the fixed effects model (upper left) and random effects model (upper right) in `MSstats`, and in the models without module information (lower left) and with module information (lower right) in `mapDIA`. The scores on the y-axis of the top two panels are negative log p-value (based 10), where a small quantity 10^{-15} was added to the p-value to avoid infinite values (log of zeros).



Supplementary Figure 11 (Glycoproteomics data set). The reported log2 fold changes from `mapDIA` and `MSstats`. The Pearson correlation was 0.992.