Supplementary Information Appendix for:

Proteomic analysis of monolayer-integrated proteins on lipid droplets identifies amphipathic interfacial α-helical membrane anchors

Camille I. Pataki[1] João Rodrigues[2], Lichao Zhang[3], Junyang Qian[4], Bradley Efron[4], Trevor Hastie[4], Joshua Elias[3], Michael Levitt[2] and Ron R. Kopito[5]*.

Departments of Biochemistry[1], Structural Biology[2], Chemical and Systems Biology[3], Statistics[4], and Biology[5], Stanford University, Stanford, California, 94305

* Address correspondence to Ron Kopito
Phone: 650 723-7581
email: Kopito@stanford.edu

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Datasets S1 to S3 (XLS)

www.pnas.org/cgi/doi/10.1073/pnas.1807981115

Supporting Materials and Methods

Reagents

Sodium carbonate (Fisher), sodium chloride (Fisher), sodium iodide (Sigma), sodium chlorate (J.T. Baker Chemical Co), Tween20 (Fisher), bovine serum alumin (BSA) (Sigma), dithiothreitol (DTT), iodoacetamide (Sigma).

Antibodies

The following antibodies were used: anti-ATGL (Cell Signaling), anti-AUP1 (Proteintech Group Inc.), anti-CGI-58/ABHD5 (Proteintech Group Inc.), anti-GAPDH (Chemicon), anti-GFP (Clontech Laboratories, Inc.), anti-Plin2/ADFP (Novus Biologicals), anti-P97 (Novus Biologicals), anti-UBXD8 (Proteintech Group Inc.).

Plasmids

The GFP-tagged Rab5 and Rab7 constructs were kind gifts from Suzanne Pfeffer, Stanford University, Stanford, CA. Lipid-deficient mutants were made by site-directed mutagenesis with the following primers: Rab5(C212S, C213S): GTGGATCCTTTAGTTACTAGAAGACTCATTCCTGGTTG and Rab7(C205S, C207S): GTACCTATCAGGAACTGGAGCTTTCCGCTG. The C-terminally GFP-tagged DHRS3 constructs were generated by PCR amplification of the indicated residues and ligated into the HindIII/EcoRI sites of the pcDNA3.1-GFP-tag construct (1). The cysteine mutant library was generated by site-directed mutagenesis of the DHRS3(1-60)-GFP construct (General Biosystems).

Cell culture and transfection

HEK293 cells were maintained at 37°C in DMEM (Mediatech) supplemented with 10% animal serum complex (ASC) (Gemini Bio-Products) using standard cell culture techniques. Cells were transfected using lipofectamine LTX (Roche) according to manufacturer's instructions. In brief, DNA:PLUS reagent:Lipofectamine reagent was mixed in a 1:1:3 ratio in DMEM. After a 5-min incubation, complexes were added directly to cells in complete growth medium. LDs were inducted with 200μM oleic acid in complex with 0.2% bovine serum albumin in standard medium for 16 hours.

Immunoblotting

Proteins were separated by SDS-PAGE, followed by wet-transfer onto PDF membranes. Skimmed milk (5%) in PBS with 0.1% TritonX-100 was used to block nonspecific binding. Antibodies were diluted with BSA (4%) with sodium azide [0.2%(w/v)]. IRDye secondary antibodies (LiCor) were used for signal detection by Odyssey imaging (LiCor). Band intensities were quantified by densitometry using Image Studio Lite software (LiCor).

Negative stain transmission electron microscopy

4uL of purified LD fractions were placed on a 300 mesh Carbon/Formvar coated copper grid-grid. Fixations were performed with glutaraldehyde (1% w/v), paraformaldehyde (2% w/v) in sodium cacodylate (NaCaC) buffer (0.1M, pH 7.2). Contrasting was performed with a 9:1 mix of 2% methyl cellulose and 4% uranyl

acetate for 5min. Specimens were examined with a Gatan Orius 10.7 megapixel CCD camera at 120kV.

Cellular fractionation

As described in ref (2) with the following exceptions: 1) each fractionation was done with 2 sub-confluent 15cm plates of HEK293 cells, 2) LD fractions were not TCA precipitated, and 3) LD, cytosolic, and membrane fractions were loaded by the following volume ratios: 1:0.15:0.25.

Solvent accessibility assays

LDs and ER fractions were purified by the cellular fractionation method described above for each DHRS3(1-60)-GFP cysteine mutant. The PEGylation assay was adapted from (3). The ER pellet was washed three times in mPEG reaction buffer (50mM HEPEs-KOH, pH 7.2, 250mM sorbitol, 70mM potassium acetate, 5mM sodium EGTA, and 1.5mM magnesium acetate) and resuspended in 500µL of mPEG reaction buffer. 25µL of the ER pellet fraction was used in a final reaction volume of 40µL with 1.0mM tris(20carboxyelthyl)phosphine (TCEP) (Sigma). ER pellet samples were treated with or without Triton X-100 (1% v/v) as indicated. 25ug TAG [measured by Serum Triglyceride Determination Kit (Sigma)] was used for each LD reaction at a final concentration of 0.5µg/µl with 1.5mM TCEP. LD samples were treated with or without Triton X-100 (1% v/v) and SDS (0.5% v/v) as indicated. ER pellet and LD samples were treated with maleimide-PEG (mPEG) (2mM) (Quanta Biodesign, catalog number 10406). All reactions were allowed to proceed for 30min

at room temperature and quenched with 10mM dithiothreitol (DTT) on ice for 10min. 5X laemli buffer was added to all reactions and 15µL of each reaction were run on SDS-PAGE gels.

TMT-MS sample preparation

Samples were sequentially reduced and alkylated with 5mM DTT and 14mM iodoacetamide, respectively. Samples were sequentially digested with LysC (Wako) and Trypsin (Promega) at 4M and 1M urea, respectively. All dilutions were with 100mM ammonium bicarbonate (pH=8). RapiGest was denatured by manufacturer's instructions. Sample stagetip, TMT label ligation, were done as described in (4). Experiment 1 was mixed at equal volume ratios for MS/MS analysis. Experiments 2 and 3 were ratio checked and remixed as described in (4) for MS/MS analysis.

Structural modeling of predicted TMDs

None of the predicted TMDs in our dataset has a structure deposited in the RCSB PDB database. We submitted the sequence of each of our predicted TMDs to the CABS-fold server (5) to perform template-free structure predictions. We used the *de novo* modeling setting, which defines a large range for the acceptance ratio parameter in the Replica Exchange Monte Carlo scheme and therefore allows sampling of a larger fraction of the conformational landscape. We took the densest cluster of each prediction run as a representative model for our simulations.

Molecular dynamics simulations of predicted TMDs

Simulations were performed with GROMACS 2016.1 (6), using the MARTINI coarse-grained model for proteins (7) and lipids (8). For each predicted TMD, the representative model produced by CABS was coarse-grained using the *martinize.py* script with default settings, except for uncharged termini, and energy minimized in vacuum using the steepest descent algorithm for 5,000 steps. The peptide was then placed in a hexagonal periodic box and inserted horizontally at the mid-point of a POPC bilayer using the *insane.py* script (9). The system was then solvated and neutralized (0.15M of sodium and chlorine ions). Anti-freeze particles replaced 10% of the water molecules. The entire system was energy minimized using the steepest descent algorithm for 10,000 steps and equilibrated for 23 ns applying position restraints on the peptide backbone beads. All simulations were run in the isothermal-isobaric (NpT) ensemble, performed at 310 K to maintain the POPC membrane in the fluid phase. Temperature was controlled using the V-rescale thermostat with a coupling constant of τ_t=1 ps, while pressure was semi-isotropically coupled to an external bath of ρ=1 bar with a coupling constant of τ_p=12 ps and a compressibility of 3.0^{-4}.bar^{-1} using the Parrinello-Rahman barostat. We used a timestep of 10 fs for equilibration and 20 fs for all other simulations. The neighbor list was updated using the Verlet neighbor search algorithm, with the neighbor list length being automatically determined. Lennard-Jones (LJ) and Coulomb potentials and forces were cut off at 1.1 nm. The LJ potential was shifted to zero at the cutoff and electrostatic interactions were calculated using a reaction-field potential with a $\varepsilon_{RF}=\infty$ and a relative dielectric constant of ε=15 nm.

Umbrella sampling simulations

Starting structures for the umbrella sampling simulations were generated by pulling the center of mass (COM) of each peptide along the Z direction (v=0,0,1) from the bilayer center into water at a pulling rate of 0.06 nm/ns with a force constant of 1000 kJ.mol^{-1}.nm^{-2} for 100 ns. Additionally, we enforced rotational constraints to keep the peptide parallel to the bilayer plane and ensure reliable COM measurements. Umbrella sampling was performed in the range of 0 to 5 nm of separation along the membrane normal (z) between the peptide COM and the bilayer COM. Starting configurations were selected from the pulling simulations at a spacing of 0.1 nm, resulting in 51 windows for each peptide. Each window was then equilibrated for 10 ns, removing the rotational constraints and enforcing position restraints on the backbone beads. Umbrella sampling production runs were then performed on each window for 250 ns, with a harmonic restraint on the distance between COM of the peptide and the bilayer COM in the z dimension with a force of 1000 kJ.mol^{-1}.nm^{-2}. Energy profiles were calculated using weighted histogram analysis method (WHAM) (10).

Unrestrained simulations and solvent accessible surface area calculations

To calculate solvent accessible surface areas for each residue, we simulated each peptide without any restraints for 1 μs. The starting structure was taken from the last frame of the umbrella sampling production runs. Solvent accessible surface areas (SASA) were calculated for each coarse-grained bead using the Shrake-Rupley algorithm implemented in the MDTraj (11) and a probe radius of 0.23 nm (diameter of a standard MARTINI water molecule). Accessibility values per bead were

averaged after discarding the first 200 ns of simulation and then summed and averaged over all side-chain beads for each individual residue (except alanine and glycine that are represented only by one backbone bead).

SI Appendix References

1. Christianson JC*, et al.* (2011) Defining human ERAD networks through an integrative mapping strategy. *Nat Cell Biol* 14(1):93-105.
2. Schrul B & Kopito RR (2016) Peroxin-dependent targeting of a lipid-droplet-destined membrane protein to ER subdomains. *Nat Cell Biol* 18(7):740-751.
3. Sun LP, Li L, Goldstein JL, & Brown MS (2005) Insig required for sterol-mediated inhibition of Scap/SREBP binding to COPII proteins in vitro. *J Biol Chem* 280(28):26483-26490.
4. Zhang L & Elias JE (2017) Relative Protein Quantification Using Tandem Mass Tag Mass Spectrometry. *Methods Mol Biol* 1550:185-198.
5. Blaszczyk M, Jamroz M, Kmiecik S, & Kolinski A (2013) CABS-fold: Server for the de novo and consensus-based prediction of protein structure. *Nucleic Acids Res* 41(Web Server issue):W406-411.
6. Barneda D*, et al.* (2015) The brown adipocyte protein CIDEA promotes lipid droplet fusion via a phosphatidic acid-binding amphipathic helix. *Elife* 4:e07485.
7. de Jong DH*, et al.* (2013) Improved Parameters for the Martini Coarse-Grained Protein Force Field. *J Chem Theory Comput* 9(1):687-697.
8. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, & de Vries AH (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B* 111(27):7812-7824.
9. Wassenaar TA, Ingolfsson HI, Bockmann RA, Tieleman DP, & Marrink SJ (2015) Computational Lipidomics with insane: A Versatile Tool for Generating Custom Membranes for Molecular Simulations. *J Chem Theory Comput* 11(5):2144-2155.
10. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, & Kollman PA (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of computational chemistry* 13(8):1011-1021.
11. McGibbon RT*, et al.* (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* 109(8):1528-1532.

Fitting Mixtures of Compositional Data

1 Overview

From our exploratory analysis, the protein-level data seem to fall into three categories, each presenting a different pattern in the composition. Mixture model is often used to capture data of this kind, where the marginal density of each observation is assumed to be a mixture of the densities for each class. Due to the special structure of compositional data, including nonnegative coordinate values and unit sum, there are some common classes of distributions for choice. We decide to use Dirichlet distribution for each class and will discuss other options later.

Once we determine the parametric model, we will fit the model and estimate the parameters based on our data. A constrained Expectation-Maximization (EM) algorithm is used, given some specific structure assumed for the mixture model.

After estimation, we want to make proper statistical inference for the discoveries to be made. We focus on the criterion of false discovery rate (FDR) control. The empirical bayes framework provides a foundation for valid inference based on estimated parameters. Details will be presented in the following sections.

2 Modeling

Dirichlet Distribution Parameterized by positive parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$, the density of Dirichlet distribution at $\boldsymbol{x} = (x_1, \ldots, x_J)$

$$f(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{J} x_j^{\alpha_j - 1},$$

where the beta function $B(\boldsymbol{\alpha}) = \prod_{j=1}^{J} \Gamma(\alpha_j) / \Gamma\left(\sum_{j=1}^{J} \alpha_j\right)$. Notice that the density introduces strong independence among the coordinates by taking the product form. This implies some special properties of the family.

Proposition 1 (Almost Independence) *Dirichlet distribution can be derived from independent Gamma distributions. Given J independent Gamma random variables $Y_j \sim Gamma(\alpha_j, \theta), j = 1, \ldots, J$, we have*

$$X = (X_1, \ldots, X_J) = \left(\frac{Y_1}{S}, \ldots, \frac{Y_J}{S}\right)$$

follows Dirichlet distribution with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_J)$*, where* $S = \sum_{j=1}^{J} Y_j$*.*

Proposition 2 (Convex Contours [Ait86]) *If* $\alpha_j > 1, j = 1, \ldots, J$*, then the isoprobability contours of Dirichlet distribution must be convex.*

Other Choices of Distribution Logistic normal distribution is also a common choice of model for compositional data. It assumes the log-ratio of the data follows a multivariate normal distribution. The number of free parameters - J for Dirichlet and $(J-1)(J+2)/2$ for logistic normal - seems to suggest that logistic normal is a more flexible class. Indeed, it has been suggested [Ait86] that logistic normal is generally a preferred class for compositional data due to its flexibility and ability to model non-convex pattern. For more details about those distributions and their comparison, refer to [Ait86]. Nevertheless, the choice is still problem-dependent. Here, for example, Dirichlet distribution is a reasonable choice considering the following reasons.

- Mixture modeling provides another layer of flexibility on top of single Dirichlet. It is thus able to express more complex data patterns.
- From the scatterplot, each component does seem to satisfy the convexity constraint.
- The inherent independence structure is indeed aligned with the data normalization process. We normalized the data in a similar way as in Proposition 1 to get to the compositional data we start with.

We also looked at the scatterplot of the log-ratio transformed data. The clouds of data tend to collapse together and it loses clear cluster separation. For the reasons above, we did not pursue logistic normal further and use Dirichlet as our base class instead.

Mixture Model When we model the population using mixture model with K components ($K = 3$ in our analysis), the density is then assumed to be a weighted average of the densities of the base classes. The generative model assumes a latent variable $z \in \{1, \ldots, K\}$, and the weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ are the prior distribution of z. In our case, given $z = k$, the conditional probability

$$p(\boldsymbol{x}|z = k) = f(\boldsymbol{x}; \boldsymbol{\alpha}^{(k)})$$

being the Dirichlet density for class k. When we use superscript to denote each mixture component, the marginal (overall) density has the form

$$g(\boldsymbol{x}; \boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K)}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}; \boldsymbol{\alpha}^{(k)}).$$

For more coverage of mixture models, see [MP04].

Special Considerations There are some other information available that enables us to improve statistical efficiency of the inference.

Center Component Recall that the normalization process implies the existence of a center component, of which the observed data in that component have roughly same coordinates. In fact, such structure can be characterized by equivalent parameters in the Dirichlet density. In other words, the distribution of the center component has only one free parameter. Without loss of generality, we assume the first component is the center one, i.e. $\alpha_1^{(1)} = \cdots = \alpha_J^{(1)}$.

Known Proteins Several proteins such as AUP1, FAF2 have been well studied and already identified as MIPs. For such proteins, the marginal density is assume to only take the density from the center component $f(\boldsymbol{x}; \boldsymbol{\alpha}^{(1)})$.

Therefore, we are able to write out the (log-)likelihood function of the observed data in terms of Dirichlet parameters

$$\ell(\boldsymbol{\alpha}^{(1)}, \ldots, \boldsymbol{\alpha}^{(K)}, \boldsymbol{\pi}) = \sum_{i \in \text{REF}} \log f\left(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(1)}\right) + \sum_{i \notin \text{REF}} \log \left(\sum_{k=1}^{K} \pi_k f(\boldsymbol{x}_i; \boldsymbol{\alpha}^{(k)}) \right), \tag{1}$$

where REF is the collection of proteins already identified as MIPs.

3 Parameter Estimation

Expectation-Maximization (EM) [DLR77] algorithm is a standard method for estimating parameters in mixture models. If we suppress all the parameters into $\boldsymbol{\theta}$, including the Dirichlet parameters $\boldsymbol{\alpha}$ and prior parameters $\boldsymbol{\pi}$, the EM in our case iterates through the following steps until convergence:

1. E-step: for $i \notin$ REF, compute posterior probabilities

   $$Q_i^{(t)}(k) = p(k|\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}^{(t)}), \quad k = 1, \ldots, K.$$

   For $i \in$ REF, set $Q_i^{(t)}(1) = 1$ and $Q_i^{(t)}(k) = 0, k = 2, \ldots, K$.

2. M-step:

   $$\begin{aligned} \hat{\boldsymbol{\theta}}^{(t+1)} &= \underset{\theta \in \boldsymbol{\Theta}}{\text{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} Q_i^{(t)}(k) \log f(\boldsymbol{x}_i; \boldsymbol{\theta}_k), \\ \hat{\pi}_k^{(t+1)} &= \frac{\sum_{i=1}^{n} Q_i^{(t)}(k)}{\sum_{i=1}^{n} \sum_{j=1}^{K} Q_i^{(t)}(j)}, \quad k = 1, \ldots, K, \end{aligned}$$

   where $\boldsymbol{\Theta}$ is a parameter space that also takes into account the range of Dirichlet parameters and the special constraint on the center component.

3. Exit if the change of value of the likelihood function (1) from last step is below some threshold $\epsilon > 0$.

It can be shown that the likelihood function (1) increases with iteration t. The argument is similar to that in [FHT01] and we don't expand it here. That being said, if the likelihood function is bounded, the algorithm will terminate in finite steps. Also, since EM algorithm subjects to local suboptimal, we run multiple times with random initialization, and use the one with the maximum log-likelihood value.

4 Multiple Experiments

We have collected data for $M = 3$ experiments, and would like to aggregate the results to improve the quality of inference. According to the setup of experiments, we do not assume uniform parameters across experiments but assume independence of the observations given the class label. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M$ be observed values of the same protein in the M experiments. The assumption means

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M | z = k) = \prod_{m=1}^{M} p(\boldsymbol{x}_m | z = k).$$

Using Bayes' formula, we compute posterior distribution of the class variable

$$p(z = k | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M) \;\; = \;\; \frac{p(z = k, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)}{p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)} \propto \pi_k \cdot \prod_{m=1}^{M} f(\boldsymbol{x}_m; \boldsymbol{\alpha}_m^{(k)}),$$

where $\boldsymbol{\alpha}_m^{(k)}$ is the Dirichlet parameter for the kth component of experiment m, and π_k the mixing proportions for the kth class. With multiple experiments, the above aggregation enables us to have higher confidence about assigning one protein to a particular class than with single experiment.

The estimates obtained from last section are then plugged into the formula, i.e. $\hat{p}(z = k | \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M) \propto \hat{\pi}_k \cdot \prod_{m=1}^{M} f(\boldsymbol{x}_m; \hat{\boldsymbol{\alpha}}_m^{(k)})$. Here $\hat{\pi}_k$ can be obtained via a pooled estimator $\hat{\pi}_k = (1/M) \sum_{m=1}^{M} \hat{\pi}_{k,m}$, where each $\hat{\pi}_{k,m}$ is experiment-specific estimate of prior of class k.

5 False Discovery Rate and Empirical Bayes

With the soft assignment rules above, we are also able to obtain an estimate of the false discovery rate (FDR) when the class of interest is in the center. Formally, if we have N null hypotheses $\mathcal{H}_0 = \{H_{01}, \ldots, H_{0N}\}$ to test and a decision rule $\mathcal{D}$, FDR is defined as

$$\text{FDR} = \mathbb{E}\left[\frac{\text{number of false positives}}{\max(\text{total number of discoveries}, 1)}\right] = \mathbb{E}\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{\max(|\mathcal{R}|, 1)}\right],$$

where $\mathcal{R}$ is the set of rejected hypotheses or discoveries.

Local FDR Under the framework of mixture model, it is convenient to talk about local FDR [ETST01] [ET02], a localized version of the traditional "tail area" FDR. When we have specified a set of classes $\mathcal{C}_0$ as null - in our application, $\mathcal{C}_0 = \{2, 3\}$, the non-central classes - local FDR is defined as

$$\text{fdr}(\boldsymbol{x}) = P(z \in \mathcal{C}_0 | \boldsymbol{x}) = \frac{\sum_{k \in \mathcal{C}_0} \pi_k f_k(\boldsymbol{x})}{f(\boldsymbol{x})},$$

where f_k is the density function of class k and $f = \sum_{k=1}^{K} \pi_k f_k$ is the marginal density function. When we consider multiple independent experiments, f_k is the joint probability density of the observations from the experiments.

Bayes FDR For a rejection region $\mathcal{X}$, Bayes FDR can be defined in terms of local FDR as

$$\text{FDR}(\mathcal{X}) = P(z \in \mathcal{C}_0 | \boldsymbol{x} \in \mathcal{X}) = \mathbb{E}[\text{fdr}(\boldsymbol{x}) | \boldsymbol{x} \in \mathcal{X}]$$

Optimal FDR Control We can also define a sequence of regions that each contains all possible values of $\boldsymbol{x}$ at which the local FDR is below a specific threshold. Given threshold $t \geq 0$, let $\mathcal{X}_t = \{\boldsymbol{x} : \text{fdr}(\boldsymbol{x}) \leq t\}$. Then the optimal rejection region that controls Bayes FDR at level q is

$$\mathcal{X}^*(q) = \{\boldsymbol{x} : \text{fdr}(\boldsymbol{x}) \leq \gamma(q)\},$$

where $\gamma(q) = \max\{t : \text{FDR}(\mathcal{X}_t) \leq q\}$.

Empirical Bayes Estimator The inference above goes through if we know the true parameters of the model. When we don't as in our current analysis, one way is to use estimated ones from empirical observations. This approach is also called empirical bayes. For more details, one can refer to [Efr12]. In particular,

$$\widehat{\text{fdr}}(\boldsymbol{x}) = \frac{\sum_{k \in \mathcal{C}_0} \hat{\pi}_k \cdot \hat{p}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M | z_i = k)}{\hat{f}(\boldsymbol{x})} = \frac{\sum_{k \in \mathcal{C}_0} \hat{\pi}_k \cdot \prod_{m=1}^{M} f(\boldsymbol{x}_m; \hat{\boldsymbol{\alpha}}_m^{(k)})}{\sum_{k=1}^{K} \hat{\pi}_k \cdot \prod_{m=1}^{M} f(\boldsymbol{x}_m; \hat{\boldsymbol{\alpha}}_m^{(k)})},$$

and for any region $\mathcal{X}$,

$$\widehat{\text{FDR}}(\mathcal{X}) = \frac{1}{|\{i : \boldsymbol{x}_i \in \mathcal{X}\}|} \sum_{\boldsymbol{x}_i \in \mathcal{X}} \widehat{\text{fdr}}(\boldsymbol{x}_i).$$

Define $\hat{\mathcal{X}}_t$ similarly as above that $\hat{\mathcal{X}}_t = \{\boldsymbol{x} : \widehat{\text{fdr}}(\boldsymbol{x}) \leq t\}$. Given a desired level q, we find $\hat{\gamma}(q) = \max\{t : \widehat{\text{FDR}}(\hat{\mathcal{X}}_t) \leq q\}$ and identify the set of proteins $\hat{\mathcal{I}} = \{i : \widehat{\text{fdr}}(\boldsymbol{x}_i) \leq \hat{\gamma}(q)\}$.

References

[Ait86] J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London, UK, UK, 1986.

[DLR77] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[Efr12] Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

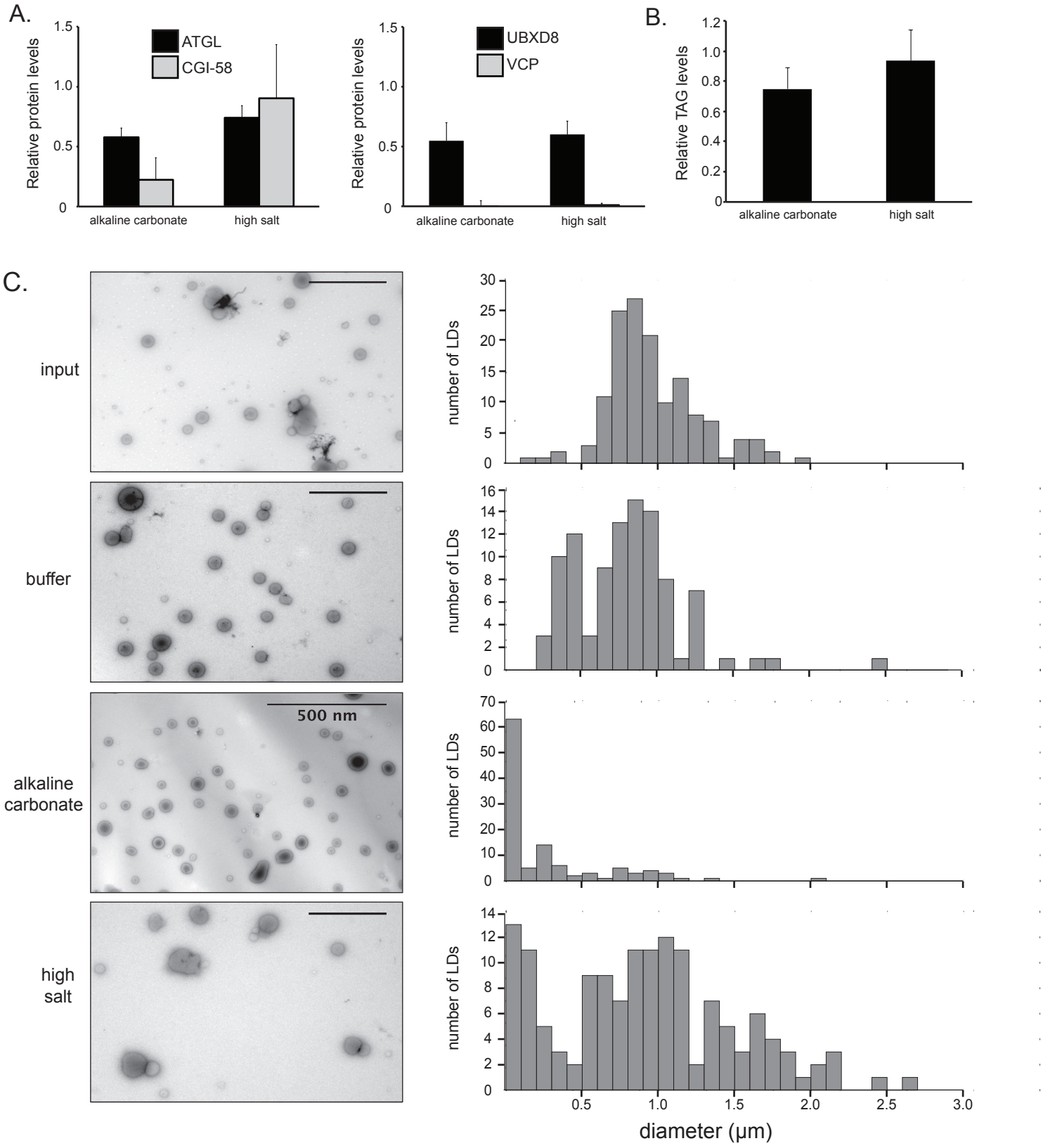
[ET02] Bradley Efron and Robert Tibshirani. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86, 2002.

[ETST01] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American statistical association*, 96(456):1151–1160, 2001.

[FHT01] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

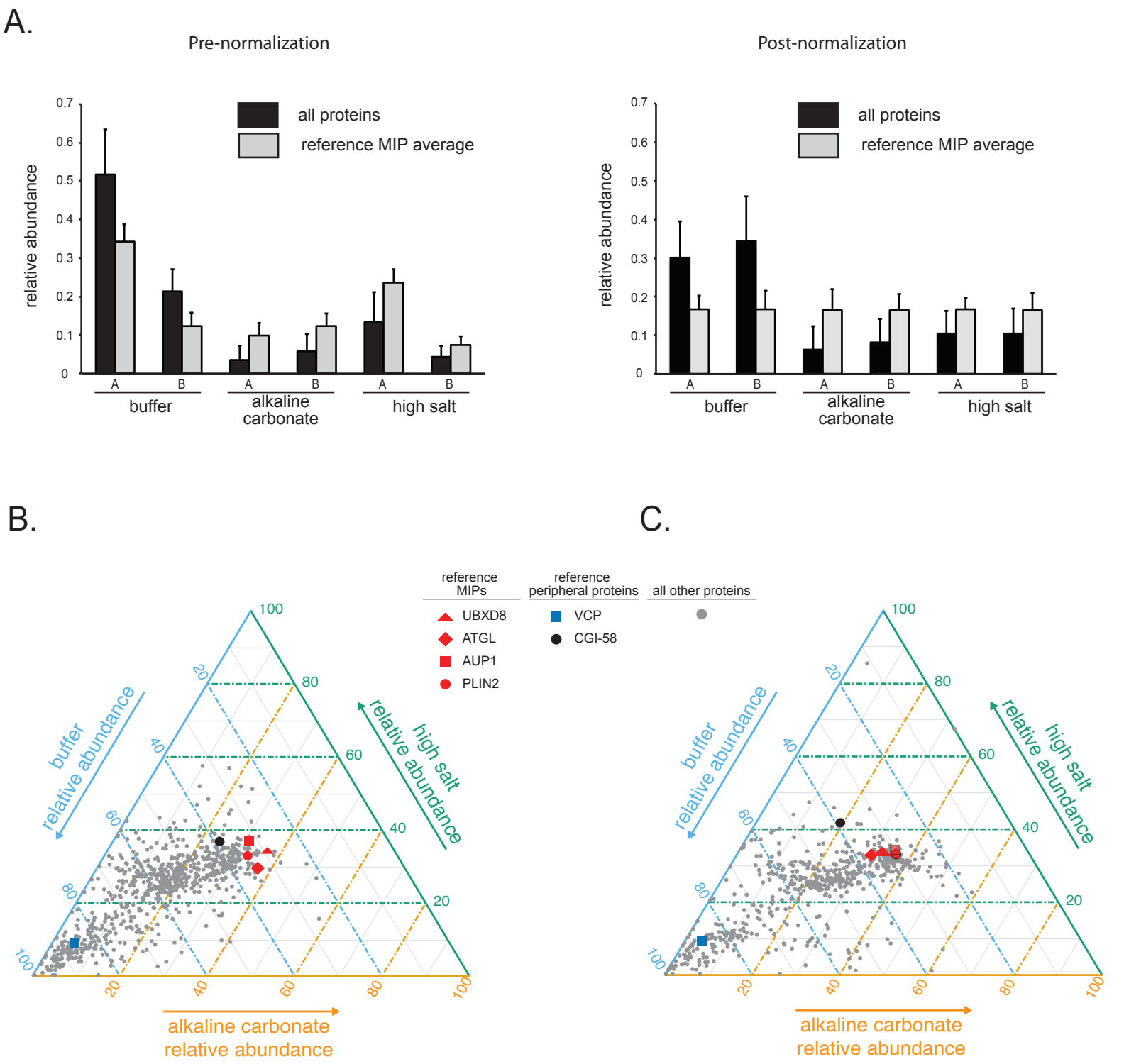
[MP04] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Supplementary Figure 1



Supplementary Figure 1. (related to Fig 1B). Quantification of recovery of MIPs and their binding partners from chaotrope-treated repurified LDs. **A**) Relative amounts of indicated proteins in repurified LDs were quantified by densitometry from immunoblots and normalized to the buffer-treated values. Each bar represents mean + SD from n=5 independent experiments. **B**) Relative triacylglycerol levels were quantified and normalized to buffer-treated values. Each bar represents mean+SD from n=5 independent experiments. **C**) *left,* Representative transmission electron micrographs of chaotrope-treated repurified LDs. Scale bars, 5μm unless indicated. *RIght*, Histograms show the size distribution of repurified chaotrope-treated LDs. n= 142, 99, 112, and 142 LDs were measured from input, buffer, alkaline carbonate, and high salt images, respectively

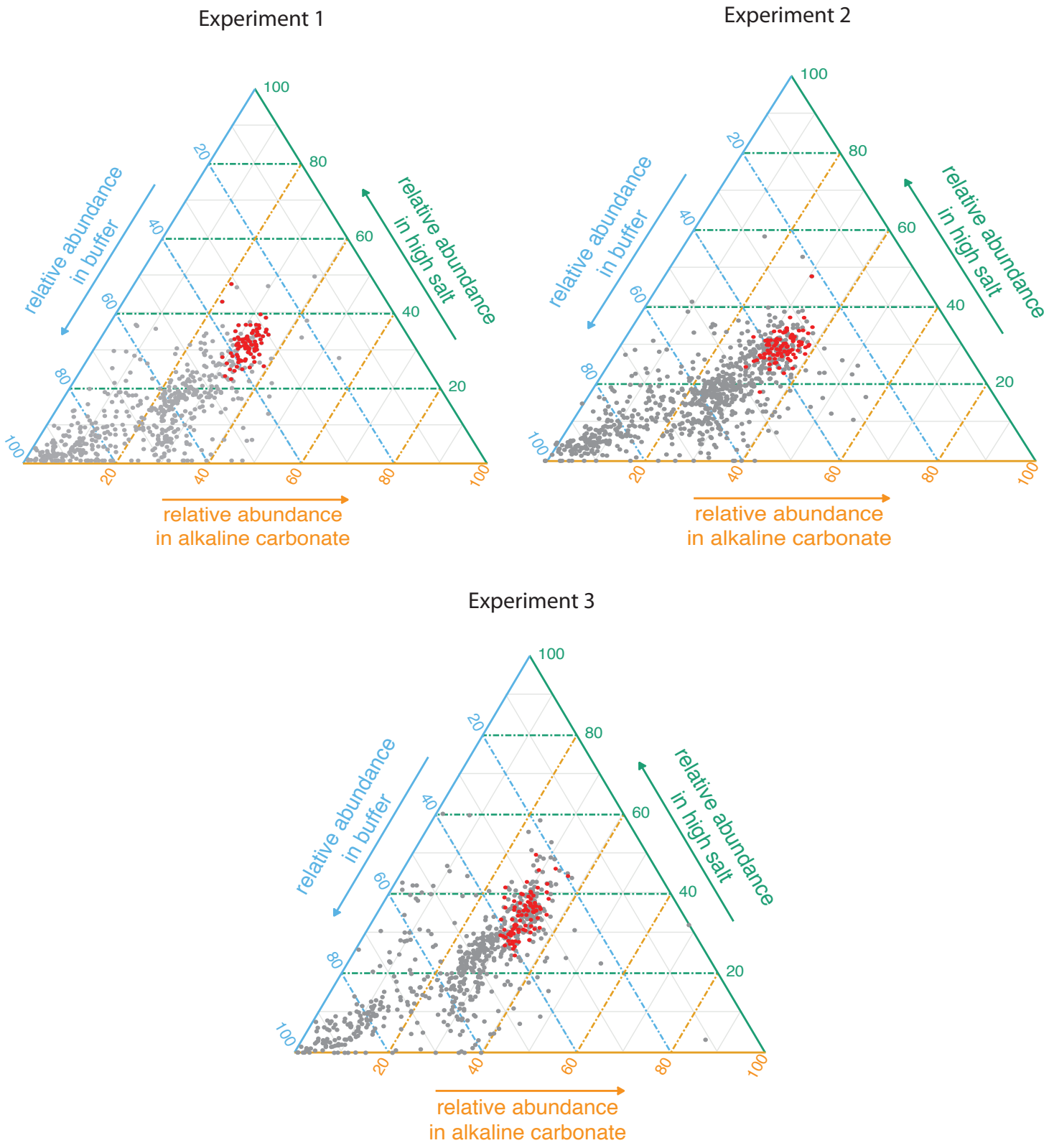
Supplementary Figure 2



Supplementary Figure 2. (related to Fig 1C-F). Effect of normalization on all six relative abundance values and ternary plots for TMT-MS replicates 2 and 3.

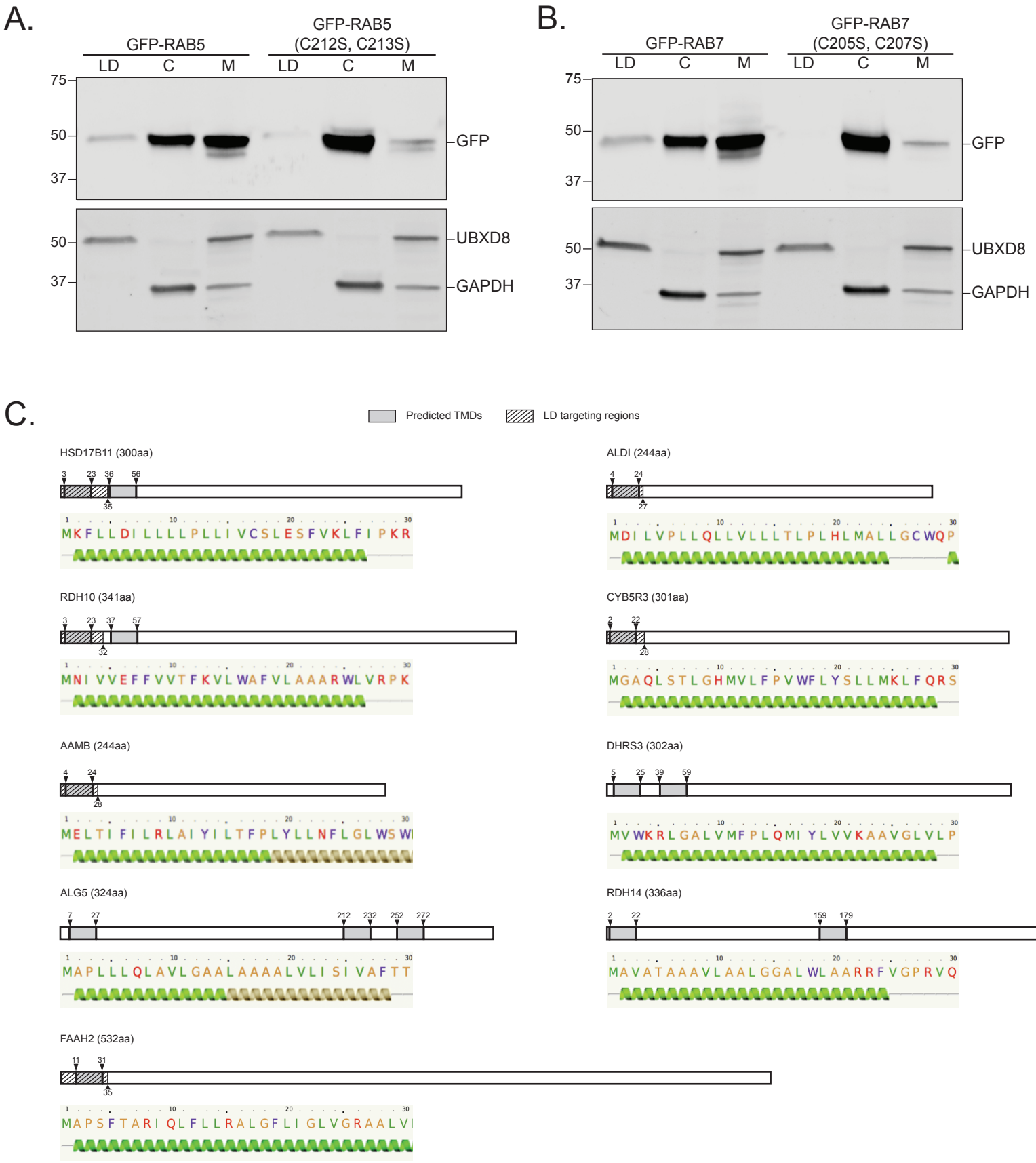
A) Average of non-normalized (left panel) and normalized (right panel) relative abundances of all peptides for the indicated groups of proteins. Reference MIPs are UBXD8, ATGL, AUP1, and PLIN2. Each bar represents mean+SD. **B-C**) Ternary plots for biological replicate 2 (B) and 3 (C).

Supplementary Figure 3



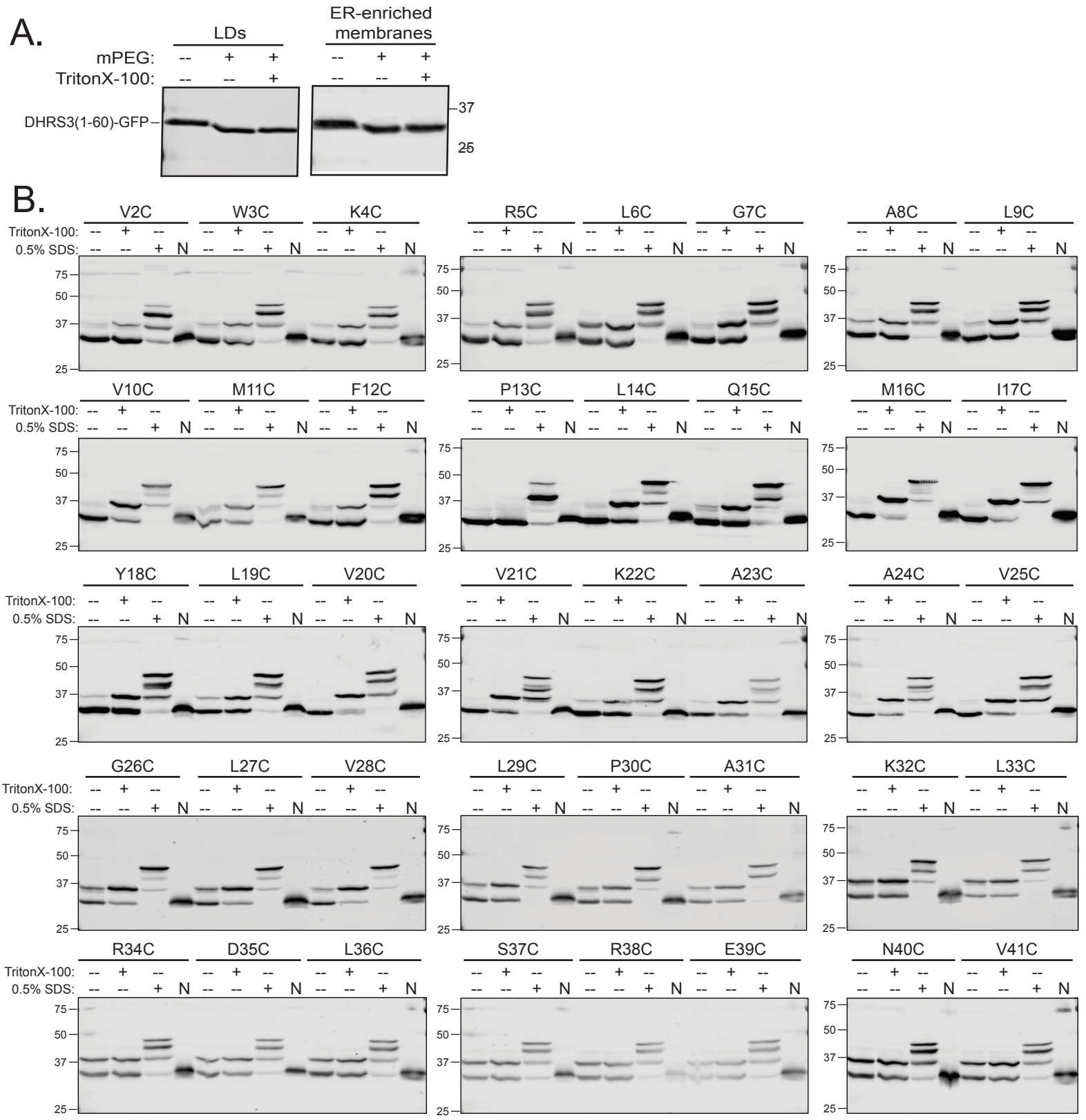
Supplementary Figure 3. (related to Fig 2A). Ternary plots of all proteins from three experiments, plotted as as in Fig 1F. Solid red symbols denote the relative abundance values of the 87 high confidence candidate MIPS.

Supplementary Figure 4



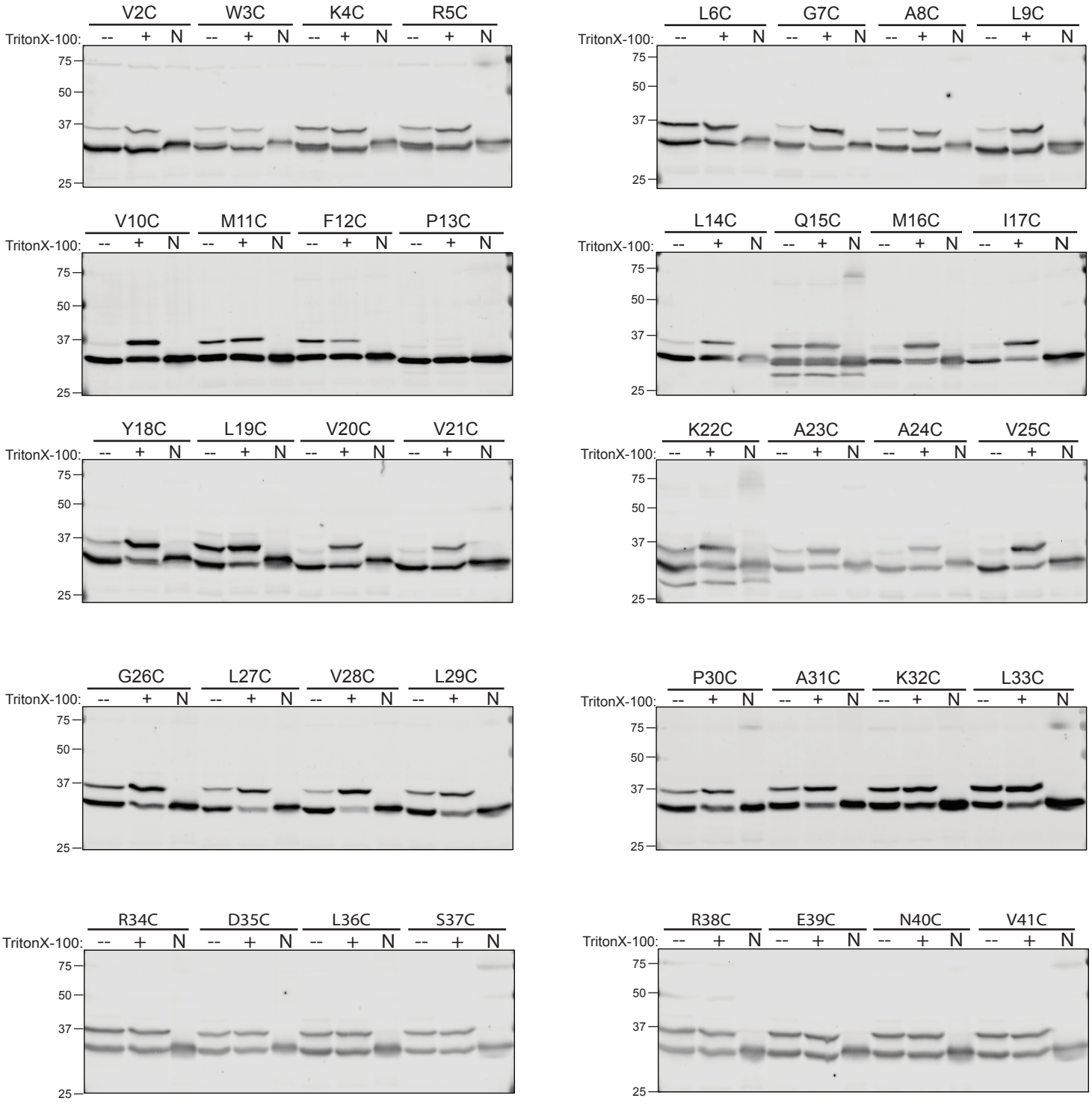
Supplementary Figure 4. (related to Figure 2) Association of Rab5 (**A**) and Rab7 (**B**) with LDs requires lipid anchors. Oleate-treated HEK cells expressing the indicated constructs were separated into lipid droplet (LD), cytosol (C), and membrane (M) fractions and analyzed by immunoblot for the indicated proteins. Representative immunoblots are shown from a single experiment out of n=2. **C**) MIPs with predicted N-terminal TMDs. SPOCTPUS predicts N-terminal TMDs in nine protein sequences from the 87 candidate MIPs. Predicted TMDs are indicated by grey shading and regions for which there is experimental evidence for LD targeting are textured with hatched shading (**Dataset S1**). The secondary structure was predicted using Phyre2 and displayed below each cartoon of the full-length protein.

Supplementary Figure 5



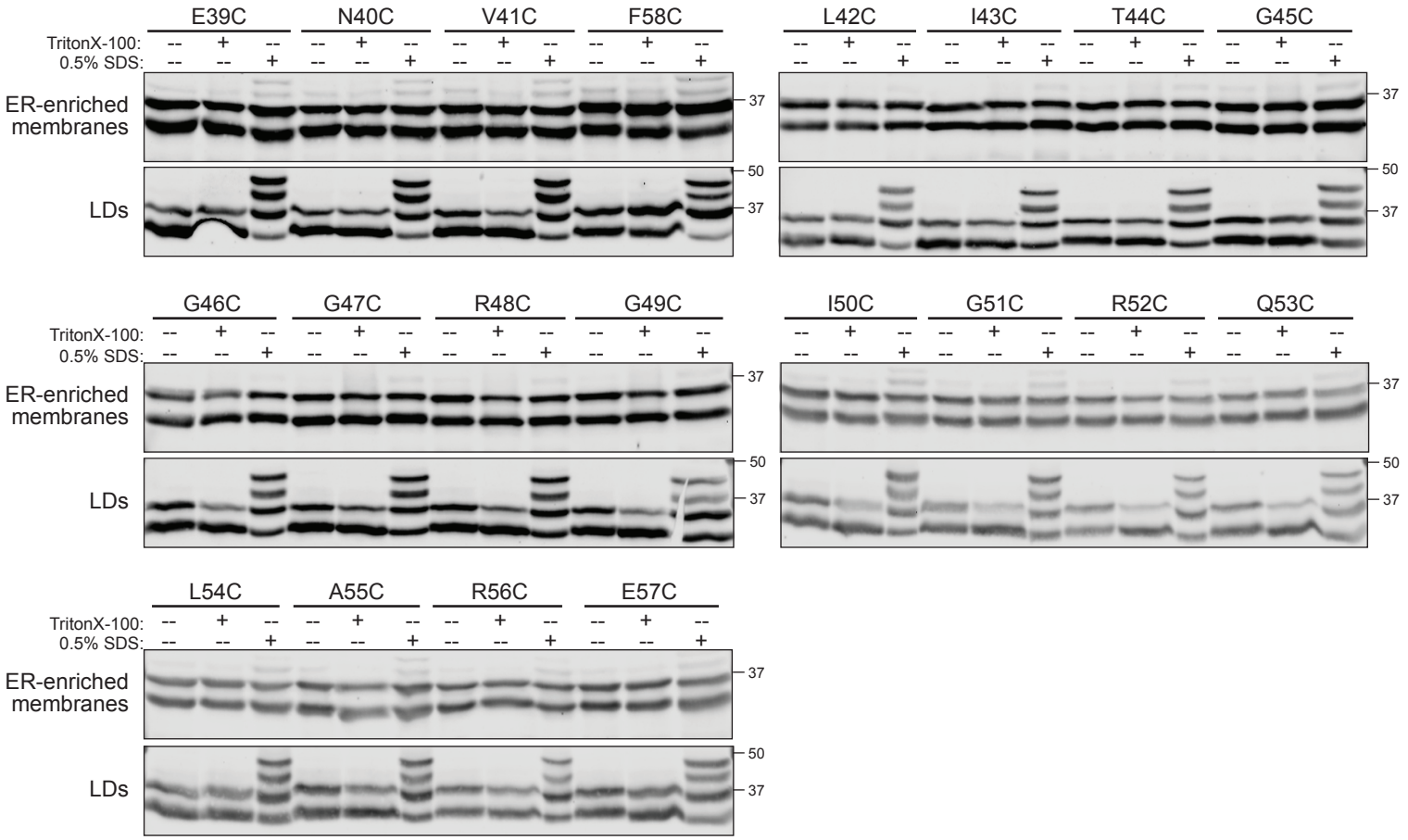
Supplementary Figure 5 (related to Fig 4). mPEG reactivity of DHRS3(1-60)GFP single Cys mutants in LDs. **A**) Wild-type DHRS3(1-60) does not react with mPEG. LD and ER fractions were purified from oleate-treated HEK cells expressing DHRS3(1-60)-GFP and treated with or without mPEG. **B**) mPEG reactivity of DHRS3(1-60)GFP single Cys mutants in LDs. Purified LDs were reacted with mPEG with and without TritonX-100 or SDS for 30min at room temperature and quenched with DTT for 10min at room temperature. Non-reduced samples("N") were included for every construct to ensure that intermolecular disulfide bonds were not formed. Proteins were separated by 15% SDS-PAGE and immunoblotted with anti-GFP antibody. The denaturing effect of SDS exposed the two cysteines of GFP, which provided additional sites for mPEG reactivity and were detected by higher molecular weight bands. Representative membranes are shown for one of three replicates for each residue.

Supplementary Figure 6



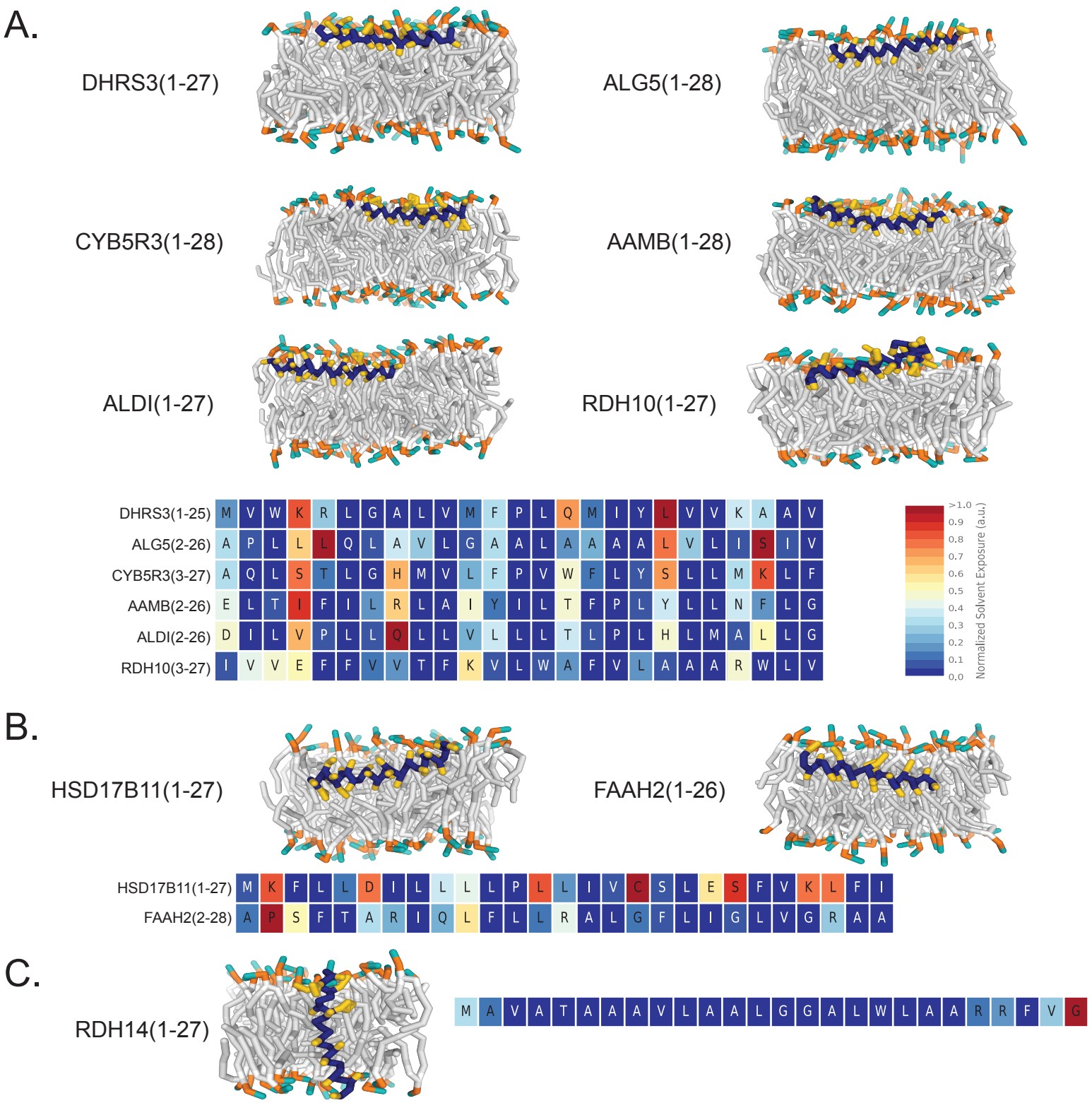
Supplementary Figure 6 (related to Figure 4). mPEG reactivity of DHRS3(1-60)-GFP single Cys mutants in ER. Purified ER fractions were reacted with mPEG as described in Figure S4.

Supplementary Figure 7



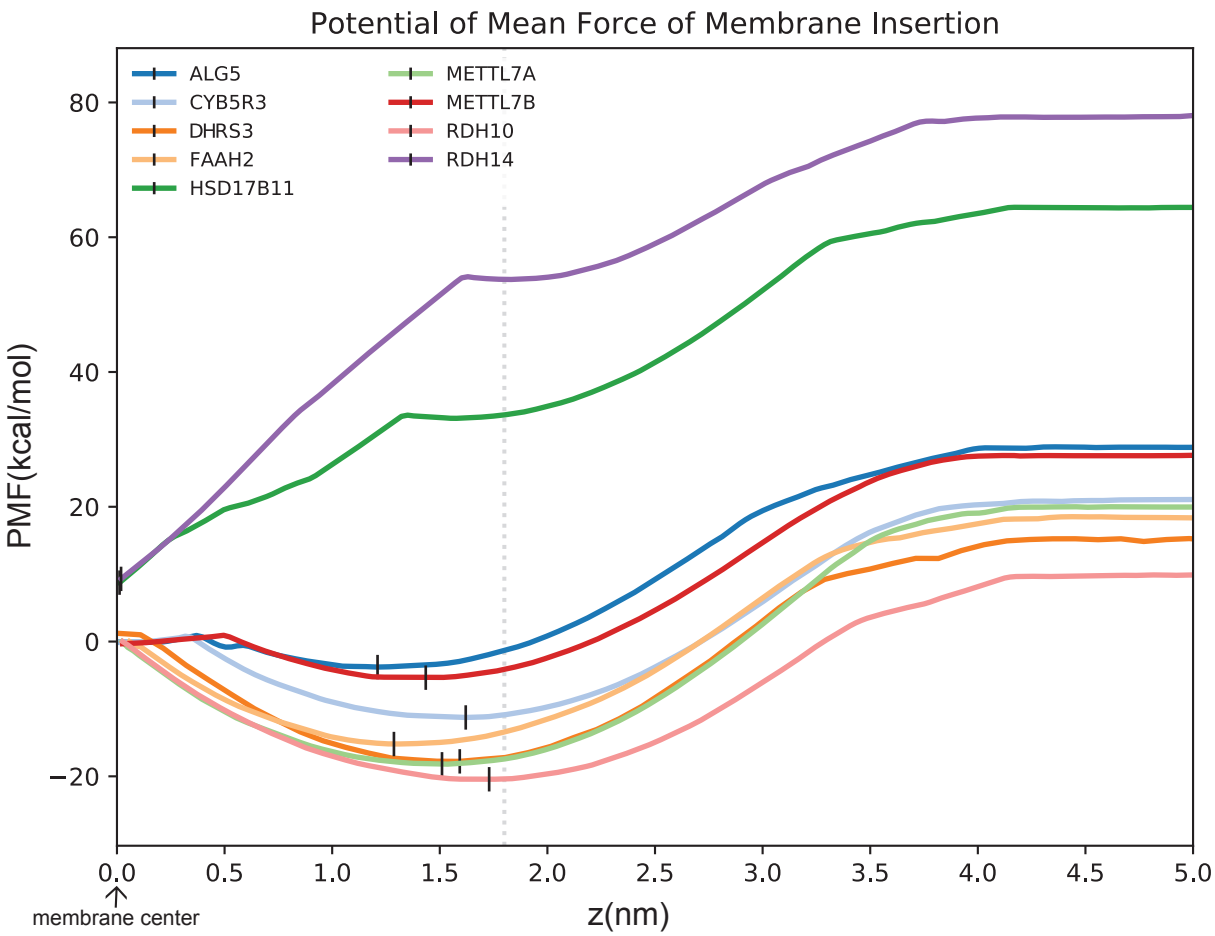
Supplementary Figure 7 (related to Figure 4). DHRS3(40-60) are maximally reactive with mPEG. Purified ER and LD fractions were reacted with mPEG as described in Figure S4.

Supplementary Figure 8



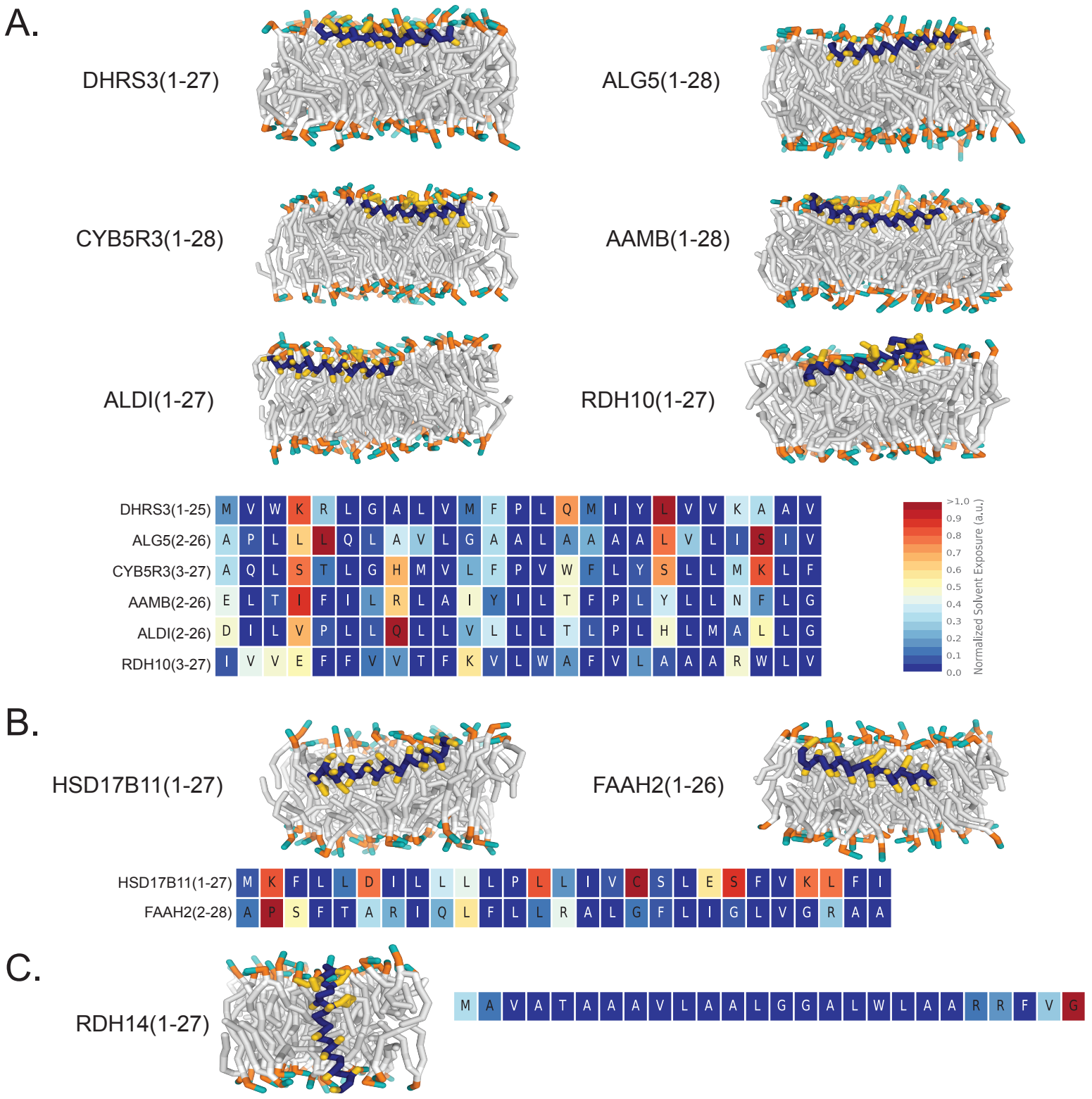
Supplementary Figure 8: Molecular dynamics predict N-terminal amphipathic helices in nine candidate MIPs. Cross-sectional views of models of the indicated protein segments from the unrestrained simulations and resulting side-chain accessibilities for each indicated protein segment rendered in heatmaps as in Figure 5C. Each unrestrained simulation was started at the location indicated by their respective PMF minimum for each indicated peptide (Figure S8). **A**) Simulation results of six candidate MIPs with N-terminal predicted TMDs that resulted in a similar interfacial amphipathic topology as DHRS3. **B**) Simulation of two candidate MIPs with N-terminal predicted TMDs that resulted in a more buried topology that interacts with the bottom of the phospholipid headgroups. **C**) Simulation of the RDH14 N-terminus predicted a TMD topology, consistent with its PMF minima being at the center of the membrane (Figure S8)

Supplementary Figure 9



Supplementary Figure 9. Potential mean force (PMF) measurements of nine candidate MIPs with N-terminal predicted TMDs. Plots of PMF measurements for the indicated proteins as they are pulled from the middle of the membrane to bulk solvent. Each colored line corresponds to a different candidate MIP. The black vertical lines denote the minimum PMF force for each protein, and thus, the most stable location of the peptide in the membrane. The grey vertical line denotes top of the membrane.

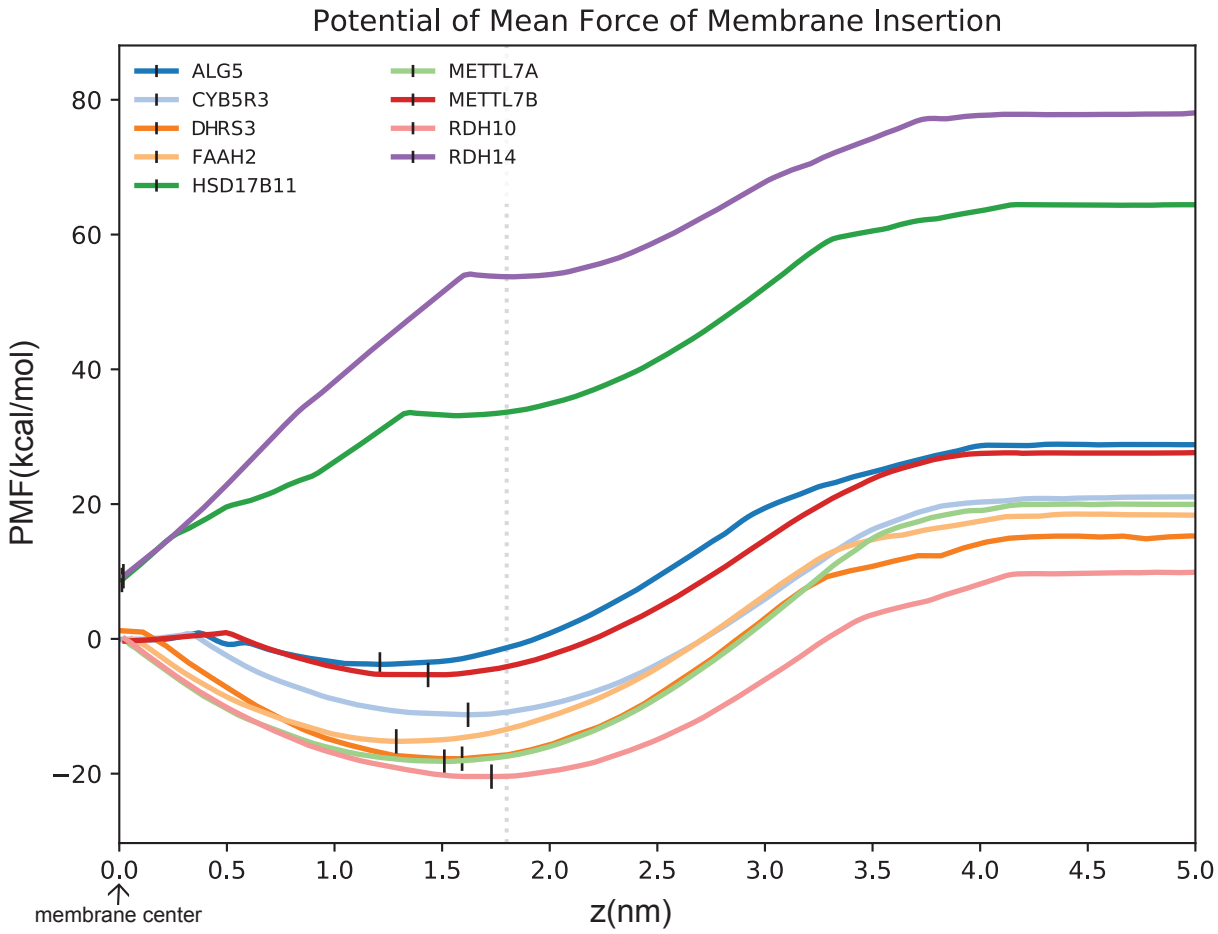
Supplementary Figure 10



Supplementary Figure 10: Molecular dynamics predict other N-terminal amphipathic helices in nine candidate MIPs.

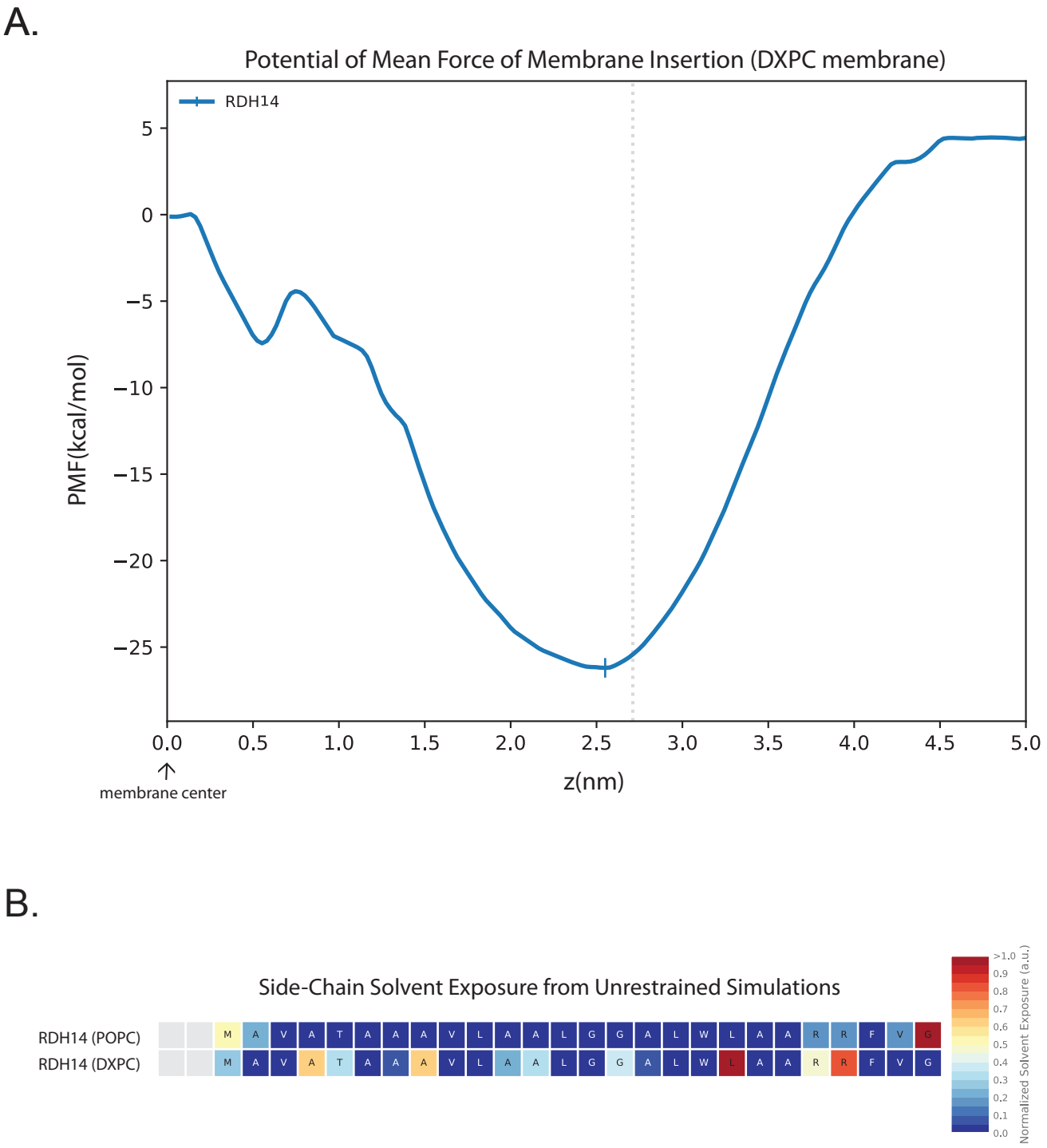
Cross-sectional views of models of the indicated protein segments from the unrestrained simulations and resulting side-chain accessibilities for each indicated protein segment rendered in heatmaps as in Figure 6d. Each unrestrained simulation was started at the location indicated by their respective PMF minimum for each indicated peptide (Supplementary figure 10). a) Simulation results of six candidate MIPs with N-terminal predicted TMDs that resulted in a similar interfacial amphipathic topology as DHRS3. b) Simulation of two candidate MIPs with N-terminal predicted TMDs that resulted in a more buried topology that interacts with the bottom of the phospholipid headgroups. c) Simulation of the RDH14 N-terminus predicted a TMD topology, consistent with its PMF minima being at the center of the membrane (Figure S10)

Supplementary Figure 11



Supplementary Figure 11. Potential mean force (PMF) measurements of all nine candidate MIPs with N-terminal predicted TMDs. Plots of PMF measurements for the indicated proteins as they are pulled from the middle of the membrane to bulk solvent. Each colored line corresponds to a different candidate MIP. The black vertical lines denote the minimum PMF force for each protein, and thus, the most stable location of the peptide in the membrane. The grey vertical line denotes top of the membrane.

Supplementary Figure 12



Supplementary Figure 12. Simulation of RDH14 in membrane of DXPC lipids. **(A)** Potential mean force (PMF) of membrane insertion calculated for the RDH14 N-terminal. Plot indicates the PMF as a function of the distance to the membrane center (0.0 nm), up to bulk solvent (5.0 nm). The vertical line denotes the minimum of the PMF and thus, the most stable location of the peptide. The grey vertical line denotes the average distance of phospholipid heads to the membrane center. **(B)** Side-chain solvent accessibility calculations from unrestrained simulations of RDH14 in POPC and DXPC lipids started from the configuration corresponding to the minimum of the respective PMF profiles. In POPC lipid, the peptide adopts a transmembrane topology (termini accessible, otherwise buried), while in thicker DXPC membrane it adopts an interfacial topology with alternating side-chains breaching the phospholipid heads and being solvent exposed.